GENOMICS AND BIOINFORMATICS ENABLED INSIGHTS INTO LIFESTYLE AND AQUATIC ADAPTATION OF A FOOD-BORNE PATHOGEN, *VIBRIO PARAHAEMOLYTICUS*

SUMA TIRUVAYIPATI

FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2015

GENOMICS AND BIOINFORMATICS ENABLED INSIGHTS INTO LIFESTYLE AND AQUATIC ADAPTATION OF A FOOD-BORNE PATHOGEN, *VIBRIO PARAHAEMOLYTICUS*

SUMA TIRUVAYIPATI

THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2015

UNIVERSITI MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: (I.C/Passport No:)

Registration/Matric No:

Name of Degree:

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

Field of Study:

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date

Subscribed and solemnly declared before,

Witness's Signature Date

Name:
Designation:

ABSTRACT

The onset of Next-generation sequencing (NGS) invoked the development of a large number of quality control (QC), assembly, annotation, and alignment tools. Bacterial genome data when processed with NGS tools have individual outputs. This effects further analysis and research related to host-pathogen studies. A defined set of bacterial genomes along with *Vibrio parahaemolyticus* PCV08-7 (VPPCV08-7) (an important bacterial isolate from prawn (*Macrobrachium rosenbergii*) in the current thesis) were selected for further bioinformatic analysis and experimentation. Strategies of High quality filtering, Trimming, Adaptor filtering and N's removal based on which seven different pipelines (Pipeline 1, 2, 3, 4, 5, 6 & Trim-Adaptor-Highquality (TAH) pipeline) made, were used. Pre-existing QC tools have been utilized in the current study. The TAH pipeline had been used on VPPCV08-7 genome data to complete the draft genome constituting of two chromosomes. The genome adaptation of *Vibrio parahaemolyticus* strains to prawn could lead to host requisition towards pathogenic infections which can be further understood by gleaning into the organization of pathogenicity islands, virulence genes and the evolution of house-keeping genes among the different strains of the bacteria. *V. parahaemolyticus* phylogeny with *gbpA* in comparison to the 16S rRNA gene as well as the house-keeping *dnaB*, *deaD* genes has been done. The interesting factor of *gbpA* (N-acetylglucosamine-binding protein A) gene, of its interaction with the epithelial chitin surfaces was observed in the present study in response to magnesium sulphate ($MgSO_4.7H_2O$) environment in the presence of carapace and commercial chitin flakes. Our results showed the importance of the QC strategy "Trimming" with a cut-off value ≥30 phred score, applying it as the first QC tool followed by adaptor and high quality filtering could help use maximum percentage of raw read data. On application to different NGS data of VPPCV08-7 showed a good percentage of usage of read data. This helped analyse the draft genome which adds to an

important gastro-intestinal pathogen repertoire. Followed by the phylogenetic relations between the 52 different *V. parahaemolyticus* strains with 16S rRNA, *deaD*, *dnaB* and *gbpA* which portray how randomly phylogenetic branches are formed. This showed selecting a house-keeping gene to relate to the evolutionary past of any strain specific to a country or location could play a major role in phylogeny. Lastly, our *in vitro* experiment and RT-PCR findings showed an increase in expression of *gbpA* in the presence of magnesium compound while significantly lower expression on addition with chitin and carapace. This explains that during the "molting" of prawn the *gbpA* expression could increase, therefore increasing the chances of bacterial interactions with the host epithelial chitin. Our study supports the use of phred quality ≥30, QC's significance in dealing with sequencing data and how the roles of host-pathogen-environment together need to be dealt to elucidate any further possibility of entry of the pathogen into the host. The data could give insights to the scientific community about precautious handling of NGS data for precision towards host-pathogen-environment studies.

ABSTRAK

Permulaan penjujukan generasi-seterusnya (NGS) telah menggunakan pembangunan sebilangan besar kawalan mutu (QC), perhimpunan, penamaan dan alat-alat penjajaran. Pelbagai jenis data genom bakteria apabila diproses menggunakan alat-alat NGS menghasilkan output individu pada peringkat pemprosesan yang berbeza. Oleh itu, genom bakteria daripada pelbagai jenis kandungan genom dan *Vibrio parahaemolyticus* PCV08-7 (VPPCV08-7), bakteria yang membawa dua kromosom, satu bakteria penting yang diambil daripada udang (*Macrobrachium rosenbergii*) telah dipilih untuk analisis bioinformatik selanjutnya dan eksperiment. Strategi penapisan kualiti tinggi, perapian, penapisan adaptor dan penyingkiran N berdasarkan tujuh aliran berbeza (Aliran 1, 2, 3, 4, 5, 6 & aliran Trim-Adaptor-Highquality (TAH)) telah digunakan. Alat-alat QC yang sedia ada telah digunakan dalam kajian semasa. Saluran paip TAH telah digunakan pada data genom VPPCV08 7 untuk menyempurnakan draf genom yang mengandungi dua kromosom. Penyesuaian genom dari strain *Vibrio parahaemolyticus* terhadap udang boleh membawa kepada permintaan tuan rumah terhadap jangkitan patogenik yang boleh difahami selanjutnya dengan memasuki ke dalam organisasi kepulauan kepatogenan, gen-gen kebisaan dan evolusi gen-gen *housekeeping* diantara jenis bakteria berbeza. Filogeni *V. parahaemolyticus* dengan gbpA berbanding gen 16S rRNA serta gen *housekeeping* dnaB dan deaD telah dilakukan. Satu faktor menarik daripada gen gbpA (N- asetil glukosamin A) ialah interaksi gen gbpA dengan permukaan epitelium kitin telah diperhatikan dalam kajian ini dengan tindak balas terhadap persekitaran magnesium sulfat (MgSO4.7H2O) dengan kehadiran karapas dan chitin emping komersil. Keputusan kami menunjukkan kepentingan strategi QC "Trimming" dengan satu nilai pemotongan ≥ 30 mata phred. Penggunaannya sebagai alat QC yang pertama diikuti dengan adaptor dan penapisan berkualiti tinggi boleh membantu menggunakan peratusan maksimum bacaan data mentah. Penggunaan data

NGS yang berbeza VPPCV08-7 telah menunjukkan peratusan penggunaan bacaan data yang baik. Ini membantu menganalisis draf genome yang menambah kepada koleksi patogen gastro-intestinal penting. Diikuti dengan hubungan-hubungan filogenetik antara 52 *V. parahaemolyticus* yang berbeza dengan 16S rRNA, deaD, dnaB dan gbpA telah mengambarkan betapa rawaknya cabang filogenetik ini terbentuk. Ini menunjukkan pemilihan gen *housekeeping* yang berkaitan dengan evolusi masa lalu pelbagai jenis strain yang spesifik kepada negara atau lokasi boleh memainkan peranan utama dalam filogeni. Akhir sekali, eksperimen in vitro dan penemuan RT-PCR menunjukkan peningkatan ekspresi gbpA dengan kehadiran sebatian magnesium manakala penurunan ekspresi yang ketara apabila ditambah dengan kitin dan karapas. Ini menjelaskan bahawa semasa ekdisis berlaku, ekspresi gbpA udang boleh meningkat, lalu meningkatkan peluang interaksi bakteria dengan kitin epitelium tuan rumah. Kajian kami menyokong penggunaan kualiti phred $\geq$ 30, kepentingan QC dalam menangani data penjujukan dan bagaimana peranan tuan rumah-patogen-alam sekitar bersama perlu ditangani untuk menjelaskan apa-apa kemungkinan selanjutnya kemasukan patogen ke dalam tuan rumah. Data ini boleh memberi wawasan kepada masyarakat saintifik mengenai pengendalian mengenai langkah berjaga-jaga mengendalikan data NGS demi ketepatan terhadap kajian host- patogen - persekitaran.

ACKNOWLEDGEMENTS

God I thank you from the bottom of my heart for always being with me. I would like to express my heartfelt gratitude and thanks to my supervisor Assoc Prof. Dr. Subha Bhassu for her continuous guidance, encouragement and appreciation. I would like to express my sincere gratitude and thanks to my co-supervisor Dr. Niyaz Ahmed for his all-time motivation and support. I would like to thank Prof. Thong Kwai Lin for providing me with the bacteria. Thanks to Genotypic Pvt. Ltd for helping me receive the data related to my work on time and for their services. Special thanks to Mr. Ng, Ms. Li Nah, Mr. KC and Mr. Foo for helping me receive chemicals and kits on time. Thank you so much "Bright Spark Unit" team for the support provided under Bright spark fellowship. Thanks to Postgraduate Research Fund (PPP). I would like to thank UM for giving me PhD opportunity and each and every faculty and staff who has assisted me in helping with my queries at every stage of my PhD tenure. Thanks to each and every one of my lab mates (AGAGEL members) who have built a friendly environment in the lab and have made research a wonderful experience. Thank you very much one and all. Special thanks to my sis and family for always making me feel positive. I would like to even thank PBL lab members for having been an encouragement whenever I have been in India. Thanks to each and every friend of mine in Malaysia and India for boosting up my confidence whenever I needed it. Thanks to my roommate Shaz who always was excited to listen to my research which helped my thinking process. From the bottom of my heart I express my fullest gratitude to my Dad, Mamma, Potti, Sushu for being my confidence, strength, and believing in me always. Thank you very much for all the love and support. I thank all my family members especially my father-in-law and mother-in-law for being a constant support at the time I most need. Last but not least thank you Deen Bava for installing in me positivity, zeal in being my strong pillar of happiness & support in my life. DEDICATED TO MY HUSBAND, DAD, MOM & BROTHER.

TABLE OF CONTENTS

		Page Number
ABSTRACT		**iii - iv**
ABSTRAK		**v - vi**
ACKNOWLEDGEMENTS		**vii**
TABLE OF CONTENTS		**viii - xii**
LIST OF FIGURES		**xiii - xix**
LIST OF TABLES		**xx - xxi**
LIST OF SYMBOLS & ABBREVIATIONS		**xxii - xxv**
LIST OF APPENDICES		**xxvi - xxvii**
CHAPTER 1		
INTRODUCTION		**1-3**
CHAPTER 2		
LITERATURE REVIEW		**4-18**
2.1	Next Generation Sequencing	**4-5**
2.2	Bacterial evolution and their survival in aquatic environments	**6-8**
2.3	Vibrio – As an aquatic micro-organism	**8-10**
2.4	*Vibrio parahaemolyticus* and its significance	**10-16**
	Temperatures & growth	**10-11**
	Salinity	**11**
	Metals	**11-12**
	Virulence factors	**12-15**
	Serotyping	**15-16**

Food effect **16**

Symptoms **16**

2.5 *Macrobrachium rosenbergii* **17-18**

CHAPTER 3 – UNDERSTANDING A DEFINED SET OF BACTERIAL GENOMES & OPTIMIZING A STANDARD QUALITY CONTROL PROTOCOL FOR BACTERIA USING NEXT-GENERATION SEQUENCING RAW READS

3.1 Introduction **19-20**

3.2 Review of the literature **21-22**

3.3 Materials & Methods **23-31**

3.3.1 Next-generation sequencing raw reads of bacterial genomes **23**

3.3.2 Quality control **23-24**

3.3.3 High Quality filtering **24**

3.3.4 Read Trimming **24-25**

3.3.5 Adaptor file **25-26**

3.3.6 N's removal **26**

3.3.7 Adaptor removal **26**

3.3.8 Assembly **26**

3.3.9 Alignment **26**

3.3.10 rRNA and GC content **26**

3.3.11 Customized pipelines **27-31**

3.4 Results & Discussion **32-55**

3.4.1 Next-generation sequencing & finishing bacterial genomes **32-46**

3.4.2 Quality control protocol **46-55**

3.5 Conclusion **56-57**

Achievement **57**

CHAPTER 4 – GENOME ANATOMY OF THE GASTROINTESTINAL PATHOGEN, VIBRIO PARAHAEMOLYTICUS OF CRUSTACEAN ORIGIN

4.1 Introduction **58-59**

4.2 Review of the literature **60-62**

4.3 Materials and Methods/Methodology **63-66**

4.3.1 Source, isolation and culture of *V. parahaemolyticus* PCV08-7 **63**

4.3.2 Genomic DNA isolation and Next-Generation sequencing **63**

4.3.3 Assembly and alignment **64**

Velvet **64**

OSLAY **64-65**

SSPACE **65**

Mauve **65-66**

4.4 Results & Discussion **67-75**

4.4.1 Genome assembly **67-68**

4.4.2 Genome statistics and annotation **68-70**

4.4.3 Identification of novel gene content and comparative analysis **70-75**

4.5 Conclusion **76-77**

Achievement **77**

CHAPTER 5 – VIBRIO PARAHAEMOLYTICUS PCV08-7 PHYLOGENETIC RELATIONSHIPS WITH CO-EVOLVED STRAINS AND ITS STUDY ON SYMBIOSIS WITH HOST AND ENVIRONMENT

5.1 Introduction **78-79**

5.2 Review of the literature **80-85**

5.3 Materials and Methods **86-94**

5.3.1 Next-generation sequencing **86**

5.3.2 Quality control and Assembly of *V. parahaemolyticus* PCV08-7 NGS data **86-87**

5.3.3 Alignment and Annotation of VPPCV08-7 contigs **87**

5.3.4 Protein-protein and Protein-chemical interactions **87**

5.3.5 Alignment and Phylogenetic analysis of *V. parahaemolyticus gbpA* gene **88**

5.3.6 Polymerase chain reaction **88-90**

5.3.7 Cloning of *gbpA* gene **90-91**

5.3.8 *V. parahaemolyticus* PCV08-7 culture and bacterial plate count **91**

5.3.9 RNA isolation and cDNA synthesis **92**

5.3.10 Real Time Polymerase chain reaction using SYBR green **92-94**

Absolute quantification **92-94**

Relative quantification **94**

5.4 Results & Discussion **95-112**

5.4.1 *Vibrio parahaemolyticus* PCV08-7 draft genome completion **95-99**

5.4.2 Bioinformatic analysis of host-pathogen genes **99-101**

5.4.3 Phylogenetic analysis of VPPCV08-7 **101-107**

5.4.4 Host-Pathogen-Environment interaction studies **107-112**

5.5 Conclusion **113-115**

Achievement **115**

CHAPTER 6

CONCLUSION **116-119**

INNOVATION **119**

SIGNIFICANCE **119**

REFERENCES **120-136**

SUPPLEMENTARY DATA

LIST OF PUBLICATIONS & PAPERS PRESENTED **138-205**

APPENDIX A **206-232**

APPENDIX B **233-235**

APPENDIX C **236-305**

LIST OF FIGURES

CHAPTER 2		**Page Number**
Figure 2.1	FASTQ format	**5**
CHAPTER 3		
Figure 3.3.1	Strategies of quality control used for QC of bacterial genomes	**24**
Figure 3.3.2	Flowchart showing the process of Pipeline 1	**27**
Figure 3.3.3	Flowchart showing the process of Pipeline 2	**28**
Figure 3.3.4	Flowchart showing the process of Pipeline 3	**28**
Figure 3.3.5	Flowchart showing the process of Pipeline 4	**29**
Figure 3.3.6	Flowchart showing the process of Pipeline 5	**30**
Figure 3.3.7	Flowchart showing the process of Pipeline 6	**30**
Figure 3.3.8	Flowchart showing the process of TAH pipeline	**31**
Figure 3.4.1	SeqQC quality graph representing average phred quality score per base position (A) Read 1 of *H. pylori* 908 (B) Read 2 of *H. pylori* 908	**32**
Figure 3.4.2	SeqQC quality graph representing average phred quality score per base position (A) Read 1 of *E. coli* NA114 (B) Read 2 of *E. coli* NA114	**33**
Figure 3.4.3	SeqQC quality graph representing average phred quality score per base position (A) Read 1 of *P. denitrificans* SD1 (B) Read 2 of *P. denitrificans* SD1	**33**
Figure 3.4.4	SeqQC quality graph representing average phred quality score per base position (A) Read 1 of *M. tuberculosis* NA-A0008 (B) Read 2 of *M. tuberculosis* NA-A0008	**34**
Figure 3.4.5	SeqQC quality graph representing average phred quality	**34**

score per base position (A) Read 1 of *V. parahaemolyticus* PCV08-7 (B) Read 2 of *V. parahaemolyticus* PCV08-7

Figure 3.4.6 SolexaQA heatplots showing average per base quality of HP908 (wherein, white represents an error rate of 0 ranging through yellow and orange to pure black representing an error rate of 0.75) (A) Read 1 (B) Read 2 **35**

Figure 3.4.7 SolexaQA heatplots showing average per base quality of ECNA114 (wherein, white represents an error rate of 0 ranging through yellow and orange to pure black representing an error rate of 0.75) (A) Read 1 (B) Read 2 **36**

Figure 3.4.8 SolexaQA heatplots showing average per base quality of PDSD1 (wherein, white represents an error rate of 0 ranging through yellow and orange to pure black representing an error rate of 0.75) (A) Read 1 (B) Read 2 **37**

Figure 3.4.9 SolexaQA heatplots showing average per base quality of MTBNA-A0008 (wherein, white represents an error rate of 0 ranging through yellow and orange to pure black representing an error rate of 0.75) (A) Read 1 (B) Read 2 **38**

Figure 3.4.10 SolexaQA heatplots showing average per base quality of VPPCV08-7 (wherein, white represents an error rate of 0 ranging through yellow and orange to pure black representing an error rate of 0.75) (A) Read 1 (B) Read 2 **39**

Figure 3.4.11 RNAmmer 1.2 Server results for rRNA genes of HP908 in (A) Contig assembly by Velvet (B) Contig assembly by Edena **41**

Figure 3.4.12 RNAmmer 1.2 Server results for rRNA genes of ECNA114 in (A) Contig assembly by Velvet (B) Contig assembly by Edena **41**

Figure 3.4.13 RNAmmer 1.2 Server results for rRNA genes of PDSD1 in (A) Contig assembly by Velvet (B) Contig assembly by Edena **42**

Figure 3.4.14 RNAmmer 1.2 Server results for rRNA genes of MTBNA-A0008 in (A) Contig assembly by Velvet (B) Contig assembly by Edena **42**

Figure 3.4.15 RNAmmer 1.2 Server results for rRNA genes of VPPCV08-7 in (A) Contig assembly by Velvet (B) Contig assembly by Edena **43**

Figure 3.4.16 M-GCAT alignment showing similarity between velvet and edena assembled contigs of HP908 **44**

Figure 3.4.17 M-GCAT alignment showing similarity between velvet and edena assembled contigs of ECNA114 **44**

Figure 3.4.18 M-GCAT alignment showing similarity between velvet and edena assembled contigs of PDSD1 **45**

Figure 3.4.19 M-GCAT alignment showing similarity between velvet and edena assembled contigs of MTBNA-A0008 **45**

Figure 3.4.20 M-GCAT alignment showing similarity between velvet and edena assembled contigs of VPPCV08-7 **45**

Figure 3.4.21 Graphs representing effect of QC strategies on percentage of reads produced in (A) PDSD1, (B) VPPCV08-7 and (C) MTBNA-A0008 data (with phred quality < 30) **47**

Figure 3.4.22 Graphs representing effect of QC strategies on percentage of reads produced with (A) HP908, (B) ECNA114 data (with phred score > 30) **48**

Figure 3.4.23 Result of QC strategies on (A) MAX contig length, (B) Total number of contigs and (C) Percentage of reads used in HP908 **49**

raw read data

Figure 3.4.24 Result of QC strategies on (A) MAX contig length, (B) Total number of contigs and (C) Percentage of reads used in ECNA114 raw read data **50**

Figure 3.4.25 Result of QC strategies on (A) MAX contig length, (B) Total number of contigs and (C) Percentage of reads used in PDSD1 raw read data **51**

Figure 3.4.26 Result of QC strategies on (A) MAX contig length, (B) Total number of contigs and (C) Percentage of reads used in MTBNA-A0008 raw read data **52**

Figure 3.4.27 Result of QC strategies on (A) MAX contig length, (B) Total number of contigs and (C) Percentage of reads used in VPPCV08-7 raw read data **53**

Figure 3.4.28 Comparison of the percentage of reads used by the customized pipelines in PSD1,MTBNA-A0008, VPPCV08-7, HP908 and ECNA114 (P1-P6, TAH pipeline) **54**

Figure 3.4.29 Graphical representation of patterns of output data (total number of contigs, percentage of reads used and MAX length of contig) across pipelines (1,2,3,4,5,6,7(TAH)) corresponding to PDSD1, MTBNAA-0008, HP908, ECNA114 and VPPCV08-7 raw read data respectively. **55**

CHAPTER 4

Figure 4.4.1 Circular view of *V. parahaemolyticus* PCV08-7 draft genome. Diagrammatic representation of major genes carried by the two chromosomes of *V. parahaemolyticus* PCV08-7 genome using CGview **68**

Figure 4.4.2 Alignment of the genome of strain RIMD2210633 against that of isolate PCV08-7 and strain O1:K33. **(a)** Comparison of chromosomes of strain RIMD2210633 (VPRIMD2210633 chr1, VPRIMD2210633 chr2) with the draft chromosomes of PCV08-7 (VPPCV08-7 draft chr1, VPPCV08-7 draft chr2) using M-GCAT. **(b)** Comparison of chromosomes of strain O1:K33 (VPO1:K33 chr1, VPO1:K33 chr2) with the draft chromosomes of PCV08-7 (VPPCV08-7 draft chr1, VPPCV08-7 draft chr2) using M-GCAT **69**

Figure 4.4.3 Alignment of a unique PCV08-7 protein sequence similar to *Photobacterium damselae subsps. damselae*. A unique sequence from PCV08-7 genome showed similarity with putative uncharacterized proteins of *V. anguillarum* 775 (F7YI77), *V. cholera* MJ-1236 (C3NPG0) and *Vibrio sp.* RC586 (D0IJW6) and similarity to a phage integrase of *Photobacterium damselae subsps. damselae* (H1A9J8) **71**

Figure 4.4.4 (a) *V. parahaemolyticus* AQ3810 alignment against PCV08-7 genome: Concatenated chromosome 1 and chromosome 2 of *V. parahaemolyticus* AQ3810 (AQ3810.fasta) against *V. parahaemolyticus* PCV08-7 (PCV08-7.fasta) (b) *V. parahaemolyticus* O1:K33 alignment against PCV08-7. Concatenated chromosome 1 and chromosome 2 of *V. parahaemolyticus* O1:K33 (O1_K33.fasta) against *V. parahaemolyticus* PCV08-7 (PCV08-7.fasta). **74**

Figure 4.4.5 Comparison of whole genome sequences of strains RIMD2210633, PCV08-7 and O1:K33. Alignment of complete genomes of *V. parahaemolyticus* RIMD2210633, **74**

V. parahaemolyticus PCV08-7 and *V. parahaemolyticus* O1:K33, showing PCV08-7 being more similar to RIMD2210633.

CHAPTER 5

Figure 5.2.1 Minimum spanning tree of 174 global strains **81**

Figure 5.2.2 Hypothetical representation of the possible host-pathogen-environment interaction **85**

Figure 5.3.1 Agarose gel electrophoresis (1% w/v) showing the PCR products *dnaB*, *deaD* and *gbpA* genes at their corresponding gene lengths **89**

Figure 5.3.2 Agarose gel electrophoresis (1% w/v) showing the PCR product with the *gbpA* real time primers on the 16 isolated colonies after transformation of the *gbpA* clone **90**

Figure 5.3.3 Absolute quantitative standard curve for DNA concentrations (0.005µg, 0.05µg, 0.5µg, 5µg and 50µg) vs Average Ct values for (A) primer concentration of 0.15nm (B) primer concentration of 0.3nm and (C) primer concentration of 0.6nm respectively **93**

Figure 5.4.1 Subsystem category distribution of the parsed draft chromosome 1 of VPPCV08-7 **96**

Figure 5.4.2 Subsystem category distribution of the parsed draft chromosome 2 of VPPCV08-7 **97**

Figure 5.4.3 Subsystem category distribution of the parsed probable plasmid of VPPCV08-7 **97**

Figure 5.4.4 STRING v9.1 protein-protein interactions of *gbpA* protein of *V. parahaemolyticus*. **100**

Figure 5.4.5 STITCH 3 protein-chemical interactions of *gbpA* protein of *V. parahaemolyticus* **101**

Figure 5.4.6 Phylogenetic tree representing the selected 52 *V. parahaemolyticus* genomes in relation with the 16srRNA gene **103**

Figure 5.4.7 Phylogenetic tree representing the selected 52 *V. parahaemolyticus* genomes in relation with the house-keeping *dnaB* (DNA helicase) gene **104**

Figure 5.4.8 Phylogenetic tree representing the selected 52 *V. parahaemolyticus* genomes in relation with the *deaD* (RNA helicase) gene **105**

Figure 5.4.9 Phylogenetic tree representing the selected 52 *V. parahaemolyticus* genomes in relation with the *gbpA* gene **106**

Figure 5.4.10 Graphs representing *gbpA* gene expression (2-delta delta Ct values) against (A) the concentrations of $Mg_2SO_4.7H_2O$ (B) presence of Chitin with $Mg_2SO_4.7H_2O$ and (C) presence of carapace with $Mg_2SO_4.7H_2O$ **108**

Figure 5.4.11 Graphs showing *gbpA* gene expression (2-delta delta Ct values) against the concentrations of $Mg_2SO_4.7H_2O$ and presence of chitin and carapace **109**

LIST OF TABLES

CHAPTER 3		**Page Number**
Table 3.3.1	Next-generation sequencing data of selected bacterial Genomes	**23**
Table 3.3.2	Trimming values of reads based on phred quality scores of bacterial genomes	**25**
CHAPTER 4		
Table 4.4.1	Table representing pathogenicity related clusters and other VP clusters in *V. parahaemolyticus* PCV08-7	**73**
CHAPTER 5		
Table 5.3.1	Details of the Next-generation sequencing output raw read data	**86**
Table 5.3.2	Table with details of the PCR and Real time PCR primers	**89**
Table 5.4.1	Number of ORFs distributed into subsystems in the draft of chromosome 1, chromosome 2 and plasmid of VPPCV08-7	**96**
Table 5.4.2	Important genes of VPRIMD2210633 and EMS strain and their presence in VPPCV08-7	**98**
Table 5.4.3	Analysis of Variance to identify the level of significance for gene expression with the presence of $Mg_2SO_4.7H_2O$	**110**
Table 5.4.4	Analysis of Variance to identify the level of significance for gene expression with the presence of $Mg_2SO_4.7H_2O$ and chitin	**110**
Table 5.4.5	Analysis of Variance to identify the level of significance for gene expression with the presence of $Mg_2SO_4.7H_2O$ and carapace	**111**
Table 5.4.6	Student two tail paired t-test of equal variance across	**111**

gene expression samples of samples treated with only $Mg_2SO_4.7H_2O$, with $Mg_2SO_4.7H_2O$ and Chitin, with $Mg_2SO_4.7H_2O$ and carapace

Table 5.4.7 Table represents the bacterial count in the presence of only $Mg_2SO_4.7H_2O$ (NONE). In the presence of $Mg_2SO_4.7H_2O$ with Chitin and in the presence of $Mg_2SO_4.7H_2O$ with carapace (concentrations represented in ppm) **112**

LIST OF SYMBOLS AND ABBREVIATIONS

API	-	Analytical Profile index
AT	-	adenine thymine
bp	-	base pairs
BWA	-	Burrow Wheeler's alignment
cDNA	-	complementary deoxy ribonucleic acid
clpB	-	caseinolytic peptidase B homolog
CHR1	-	Chromosome 1
CHR2	-	Chromosome 2
dATP	-	deoxy adenosine triphosphate
dCTP	-	deoxy cytidine triphosphate
dGTP	-	deoxy guanosine triphosphate
deaD	-	RNAhelicase
dnaB	-	DNAhelicase
dTTP	-	thymidine triphosphate
DNA	-	Deoxy ribonucleic acid
EC	-	*Escherichia coli*
FPs	-	Fingerprint patterns
g	-	gram
gbpA	-	N-acetylglucosamine-binding protein A
GC	-	guanine cytosine
HGT	-	Horizontal gene transfer
HP	-	*Helicobacter pylori*
HP908	-	*Helicobacter pylori* 908
KCl	-	Potassium chloride
LB	-	Luria-Bertani

min	-	minutes
ml	-	millilitre
mM	-	millimolar
Mg++	-	Magnesium ion
Mg	-	Magnesium
$MgCl_2$	-	Magnesium chloride
MR	-	*Macrobrachium rosenbergii*
MTB	-	*Mycobacterium tuberculosis*
ng	-	nanogram
NaCl	-	Sodium Chloride
NGS	-	Next-generation sequencing
ORFs	-	Open reading frames
ppm	-	parts per million
ppt	-	parts per thousand
PCR	-	Polymerase chain reaction
PD	-	*Paracoccus denitrificans*
QC	-	Quality control
RACE	-	Random amplification of cDNA ends
RAST	-	Rapid Annotation using Subsystem Technology
rRNA	-	Ribosomal Ribonucleic acid
sec	-	seconds
STEC	-	Shiga-toxin producing *Escherichia coli*
STs	-	Sequence types
tdh+	-	TDH positive
tdh-	-	TDH negative
trh+	-	TRH positive

trh- - TRH negative

tRNA - Transfer Ribonucleic acid

TAH - Trimming-Adaptor-Highquality

TCBS - Thiosulfate-citrate-bile salts sucrose

TDH - Thermostable direct hemolysin

TRH - TDH related hemolysin

TTSS - Type III secretion system

T3SS1 - Type III secretion system 1

T3SS2 - Type III secretion system 2

ure - Urease

VP - *Vibrio parahaemolyticus*

VPaI - Vibrio pathogenicity island-like element

VPI - Vibrio pathogenicity island

µg - microgram

µl - microliter

µM - micromolar

^{0}C - degree centigrade

AP-PCR - arbitrary primer PCR

BLAST - Basic local alignment search tool

ECNA114 - *Escherichia coli* NA114

GS-PCR - group specific PCR

M-GCAT - Multiple genome comparative analysis tool

$MgSO_4$ - Magnesium sulfate

$MgSO_4.7H_20$ - Magnesium sulfate hepta hydrate

NGSQC - Next-generation sequencing quality control

OSLAY - Optimal syntenic layout of unfinished assemblies

PDSD1	-	*Paracoccus denitrificans* SD1
qRT-PCR	-	Quantitative real time polymerase chain reaction
VPPCV08-7	-	*Vibrio parahaemolyticus* PCV08-7
β – hemolysis	-	beta hemolysis
MTBNA-A0008	-	*Mycobacterium tuberculosis* NA-A0008

LIST OF APPENDICES

SUPPLEMENTARY DATA		**Page Number**
Paper/ Poster presentation		**138**
Publications		**145-205**
APPENDICES		
Appendix A		
(a)	Adaptor file	**206-214**
(b)	Quality control data	**215-232**
Appendix B		
(a)	Genomic DNA isolation	**233-235**
Appendix C		
(a)	Plasmid DNA isolation	**236-237**
(b)	Data corresponding to assembly of *Vibrio parahaemolyticus* PCV08-7 sequencing data	**238-240**
(c)	Heatmap	**241-250**
(d)	*Vibrio parahaemolyticus*_Genomes	**251-252**
(e)	ClustalW command-line parameters	**253-254**
(f)	PCR sequencing results	**255-262**
(g)	pGEM T-easy vector system	**263**
(h)	Plasmid isolation sequencing results from 16 selected colonies	**264-288**
(i)	*Vibrio parahaemolyticus* PCV08-7 bacterial plate count	**289-291**
(j)	*Vibrio parahaemolyticus* PCV08-7 RNA isolation and cDNA synthesis concentrations and Nanodrop quality readings	**292-294**
(k)	Absolute quantification	**295**

(l) Relative quantification **296-300**

(m) Phylogenetic tree **301-305**

CHAPTER 1

INTRODUCTION

It was in the year 2005 that the Next generation sequencing (NGS) was introduced into genomic research. Until now there has been terra bytes of data generated for the scientific community to deal with. During the onset of NGS, it was compared to Sanger sequencing in being cost-effective, producing high throughput data and less time consuming in the process of generating sequencing reads. By the new sequencing technologies such as the Illumina, it has become really easy to generate large amounts of sequencing data with higher depths and coverage (Baker, 2010). Ion torrent sequencing followed by the upgraded Ion proton sequencing have shown up as advanced technologies which take as less as two hours to generate gigabytes of data. Apart from these, there have been various third generation sequencing technologies which have emerged such as the 454 pyrosequencing, PacBio, Helicos, etc (Metzker, 2005). Further technologies which are currently not on the run publicly such as the newest Oxford Nanopore (Maitra, Kim, & Dunbar, 2012) could compete in the near future in generating more accurate sequencing data from the sequencers.

With higher amounts of data, it becomes very tedious to handle, incorporate data and extract valuable information from any particular sequence without losing sensitive information related to the DNA/RNA. This primarily is applicable to smaller genomes such as those of bacteria. Among any sequencing theories bacterial genome sequencing is the need of the hour, as there have been many multi drug resistant, extremely multi drug resistant cases being identified (Mellmann et al., 2011; Rohde et al., 2011). Apart from such bacterial genomes, specifically enterobacteria are one of the most important lower organisms to deal with, as they reside within the human gut flora, which not only

play an important role in maintaining human health but, also can affect the human system as zoonotic pathogens.

There are several lower organisms which can be categorized based on the number of chromosomes they carry, high AT content, high GC content, bearing numerous resistant genes, etc. Based on the aspects of assembly, alignment and annotation, sometimes dealing with these bacteria could differ in order of avoiding loss of key information related to the bacterial sequence. There have been many tools which developed to identify and retain maximum sequence to further annotate bacterial whole genomes. The critical detail lies in how to deal with all these kind of data and complete a bacterial genome in a single viewpoint. For this there is an utmost need to prepare a reliable stepwise pipeline at the level of quality control of raw sequencing reads and deal with a unique and reliable identifier for PCR amplification in bacterial genomes. This pipeline could be applied to our genome *Vibrio parahaemolyticus* PCV08-7 (VPPCV08-7) also to retain and use as many percent of reads as possible for further bioinformatics analysis. The need for a reference dataset including the usage of assemblers, aligners and annotation tools has its importance so as to avoid the scientific community from working on these aspects to obtain false positive or negative results.

Apart from using the above reliable pipeline for the VPPCV08-7 genome completion, it is needed to extract out important genes without losing any crucial information in the bacterial genome. Such crucial information could lead to information related to host-pathogen interactions studies. *V. parahaemolyticus* is one such bacteria which invades the gut or rather the hepatopancreas of the prawn, *Macrobrachium rosenbergii.* Selecting an important gene from the genome information of VPPCV08-7 which might participate in the interactions with the host component, and this gene in correlation with the house-keeping genes as well as the 16S rRNA to check the closely related species based on phylogeny could help us understand the evolutionary origin of

a range of *V. parahaemolyticus* genomes. It has been reported that this pathogen stays with the host during its life cycle, but causes infection during the older stages.

Applying all the data and information retrieved from a next-generation sequenced bacterial genome to host-pathogen studies is the aim of our study. The need to tackle a pathogen in response to the surrounding environment while at the same time considering the host and pathogen interactions is an important aspect involved in our study.

Based on the above description the three primary objectives of our study are as follows.

1. Understanding a defined set of bacterial genomes and optimizing a standard quality control protocol for bacteria using next-generation sequencing raw reads.
2. To glean into the genome anatomy of the gastrointestinal pathogen, *V. parahaemolyticus* of crustacean origin.
3. To elucidate *V. parahaemolyticus* PCV08-7 phylogenetic relationships with co-evolved strains and study its symbiosis with host and environment.

CHAPTER 2

LITERATURE REVIEW

2.1 Next Generation Sequencing

The genomic era in bacteriology started in 1995 when Sanger sequencing was applied for the first time on a bacterial genome. In 2005, the first of high-throughput (next-generation) sequencing technologies started off with a significant cost-effective advancement in sequencing. A variety of approaches have been introduced based on the conventional Sanger sequencing which includes sequencing based on template amplification technologies, based on difference in sequencing chemistry and single-molecule sequencing (Loman, Constantinidou, et al., 2012; Metzker, 2005). Apart from this, choosing a specific platform for high-throughput bacterial genome sequencing has become a challenge not only to the sequencing market, but also to microbiologists. With the availability of high-end instruments there has been an increased amount of sequencing data generated. Bench top instruments (Loman, Misra, et al., 2012) come with an advantage of sequencing bacterial genomes within a very limited span of time, lower running costs and lower set-ups. The most recent example is the Ion PGM launched in 2011 which has shown its immediate effect in the crowd-sourcing of the outbreak of the Shiga-toxin producing *Escherichia coli* (STEC) in Germany (Mellmann et al., 2011; Rohde et al., 2011). The presence of such sophisticated high-end sequencers (Quail et al., 2012) appear to make bacteriology closer to explore for microbiologists by making it cost effective and making genomic epidemiology a reality by keeping a track on the infections caused by pathogens worldwide. The challenge comes in coping with huge amounts of raw read data generated by next-generation sequencers every year.

The major task in NGS revolves around the following format which the primary user information is given, that is the FASTQ file (Figure 2.1) constituting of the following format.

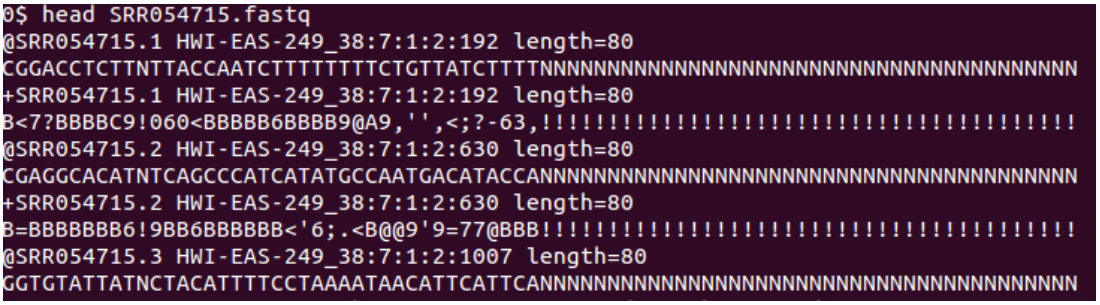


Figure 2.1: FASTQ format

Whole genome bacterial sequencing has been identified as one reliable way to analyse the pathogenicity, virulence and evolutionary history of the pathogens. While the need for high-quality sequencing has been introduced as more approachable for bioinformatic analysis than high-throughput sequencing (Alkan, Sajjadian, & Eichler, 2011). The need for tools which improve the quality of the raw reads and help interpret sequencing data of bacterial genomes play an important role in the current scenario. The analysis pipeline of a bacterial genome sequenced data constitutes of mainly quality control, assembly, alignment and annotation. There are several quality control tools which have been used widely. These have even been tested on different sets of bacterial data to develop and improve the already existing methods. The assembly, alignment and annotation processes follow the same kind of trial and error testing for proper execution of tools and generation of reliable and accurate data of bacterial genomes. While there are pipelines for de novo assembly of microbial genomes, there are studies on automated finishing of bacterial genomes utilizing hybrid approach (Jiang et al., 2012).

2.2 Bacterial evolution and their survival in aquatic environments

A largely complicated open question till today in bacteria is the gene content and the patterns of its genome evolution. There are various processes such as the loss of genes, gene genesis, duplication of genes and horizontal gene transfer which have reasonable quantitative importance to reconstruct a specific aspect of ancestral genome apart from specific aspects of genome evolution (Aravind, Watanabe, Lipman, & Koonin, 2000; Bruccoleri, Dougherty, & Davison, 1998; Kyrpides, Overbeek, & Ouzounis, 1999; Ochman & Jones, 2000)

Many studies on the presence, absence and shared genes of bacterial genomes show that the variation in the size of genomes effect the number of shared genes between genomes (Snel, Bork, & Huynen, 1999) and their evolutionary distance (Gaasterland & Ragan, 1998; Tekaia, Lazcano, & Dujon, 1999). Functional interactions between genes (Huynen & Snel, 2000; Pellegrini, Marcotte, Thompson, Eisenberg, & Yeates, 1999) were predicted by the technique of correlation in the presence of genes.

From all the above we can understand that the genome size, evolutionary history and functional selection together help determine the gene content. Some studies have even emphasized on the role of important processes which show involvement in the evolution of gene content of genomes such as in the ancestor strains of *Escherichia coli* K12 and *E. coli* 0157:H7 through horizontal gene transfer (Perna et al., 2001), gene loss in the ancestor of Buchnera (Shigenobu, Watanabe, Hattori, Sakaki, & Ishikawa, 2000) and the postulated gene duplication in the ancestor of *Vibrio cholerae* (Perna et al., 2001). Compared to the bacteria *Helicobacter pylori* and Chlamydia, horizontal gene transfer appear to be much more in *E. coli* genomes as per some studies (Perna et al., 2001).

Studies show that horizontal gene transfer (HGT) occurs mainly through biased gene transfer that is gene transfer which rarely occurs between distant relatives

and which occurs more frequently between closely related species and individuals. Such as, certain genes derived from Chlamydia species found in red and green algae and in plants (Becker, Hoef-Emden, & Melkonian, 2008; Horn et al., 2004; Huang & Gogarten, 2007; Moustafa, Reyes-Prieto, & Bhattacharya, 2008), the marine gammaproteobacteria constituting a threonyl-tRNA synthetase to which the enzymes from cyanobacterium genus Prochlorococcus show a close relation (Zhaxybayeva, Gogarten, Charlebois, Doolittle, & Papke, 2006). Hence, the ancestry of bacteria and the Archea can be well studied using both cladistics and phenetics (Andam, Williams, & Gogarten, 2010) in the form of a pluralistic approach (Bapteste et al., 2009).

The several other processes on which gene content of bacteria depend on other than HGT are duplication, gene fusion and vertical inheritance, though gene genesis and ancestral genome size do depend on HGT (Snel et al., 1999).

Every bacterial genome has its own pattern of gene loss and gene gain through evolution which could even be niche specific as some bacteria are well known to thrive in extreme conditions of acid, heat, brine water, chemicals and aquatic environments. This is followed by their individual adaptation to the living organisms in those environments. Among all the above environments prokaryotic adaptation to aquatic organisms is one of the primary interest of research in seafood consuming countries. Whereby, diseases caused in humans and aquatic organisms observed presently are caused even due to consumption of contaminated seafood.

There have been a wide range of bacteria reported from contaminated lake water such as *V. cholerae*, *E. coli* 0157:H7, Shigella, *Campylobacter jejuni*, Leptospirosis, Salmonella and Typhoid fever, *H. pylori*, Legionella, *Mycobacterium avium* complex (http://www.cdc.gov/healthyswimming/) ("Centers for Disease Control. Vaccines for selected use in international travel: cholera vaccine," 1978; Craun, Craun, Calderon, & Beach, 2006)

The major factors affecting bacterial survival in sea water are: absence of required nutrients, presence of toxic substances in sea water, presence of bacteriophages, adsorption of bacteria and their sedimentation, the harmful action of sunlight, utilization of bacteria as food by not only protozoa but other predators and competitive and antagonistic effects of the microorganism (Carlucci & Pramer, 1959).

2.3 Vibrio – As an aquatic microorganism

16S rRNA sequence is the basis on which the Vibrio species is largely classified within the Vibrionaceae family, the *V. cholerae*, *V. parahaemolyticus* and *Vibrio vulnificus* are closely related in terms of their clades and distant from Photobacteriaceae and Eterovibrionaceae families encompassed by the *Vibrio fischeri* group. This has been demonstrated by the phylogenetic analysis of gene sequences of *recA*, *rpoA* and the 16S rRNA as one concatenated sequence. *V. cholerae* gene content has been shown to be more closely related to that of *Vibrio mimicus* and slightly distant from all other species (F. L. Thompson, Iida, & Swings, 2004) based on DNA-DNA hybridization studies.

To establish the patterns of epidemiological interest which are associated with the pathogenicity of the strain and to record correlation of diseases with specific strains, serotyping has been identified as one of the useful markers. The cause for the recent marked emergence of pathogenic bacterial strains and prediction of events (Faruque et al., 2005; Reen, Almagro-Moreno, Ussery, & Boyd, 2006) through construction of models and identification of evolutionary relationships were done by using multi-locus sequence typing/analysis, serogroup association and comparative genomics. With the potential pathogenicity of *V. cholerae, V. parahaemolyticus*, and association of their serogroups, the specificity of the serogroups has been correlated. However, a wide range of species of Vibrio are pathogenic to marine organisms. Among the vast number of Vibrio species the pathogenicity effects on population of fish and other marine

organisms such as eel are due to *Vibrio anguillarum*, *Vibrio anginolyticus*, *Vibrio panaeicida*, *V. vulnificus*, *Vibrio harveyi* and *Vibrio salmonicida* and those associated with coral reef bleaching are *Vibrio campbellii*, *Vibrio shiloi*, *V. harveyi* and *Vibrio fortis* which can have a severe effect on the aquaculture industry.

Medicinal relevance to humans of some species of Vibrio has also been shown (Igbinosa & Okoh, 2008). Vibrios associated with infections which are extra-intestinal such as meningitis, wound infections and with gastroenteritis are *V. mimicus*, *Vibrio fernissi*, *Vibrio metehnikovii*, *Vibrio caincinnatiensis* and *V. alginolyticus*. *Vibrio fluvialis* is also associated with infections in the gastrointestinal region in a majority of countries. In terms of food borne diseases various factors highlight *V. vulnificus*, *V. parahaemolyticus* and *V. cholerae* as considerable importance. While *V. parahaemolyticus* and *V. cholerae* are mostly related to oysters, causing gastroenteritis. *V. vunificus* has been observed to be a causative of primary septicemia not only in marine populations (Gulig, Bourdage, & Starks, 2005) but also in humans.

V. cholerae O1 causes cholera. Most cases of infection with these bacteria were reported through eating seafood (Shigenobu et al., 2000) and especially consumption of shellfish (Besser, Feikin, Eberhart-Phillips, Mascola, & Griffin, 1994; Finelli, Swerdlow, Mertz, Ragazzoni, & Spitalny, 1992; Hayat et al., 1993; Hollis, Weaver, Baker, & Thornsberry, 1976; Klontz, Tauxe, Cook, Riley, & Wachsmuth, 1987; Lowry et al., 1989; Pavia et al., 1987).

Non-O1 *V. cholerae* are strains of *V. cholerae* which do not show agglutination in O1 antiserum, infection through this bacterium is said to be caused due to consumption of shellfish. *V. vulnificus* has been reported to have caused high fatality rates due to its invasiveness associated with infections. This species has been reported an encapsulated form which most commonly occurs in clinical isolates rather than environmental isolates (Hollis et al., 1976).

Other species such as *V. fluvialis*, *V. mimicus*, *V. alginolyticus*, *Photobacterium damsel* (*Vibrio damsela*), *V. metschnikovii*, *Vibrio cincimnatiensis*, *Vibrio fuenisii* and *Vibrio hollisae* are also known to be pathogenic which can cause severe infections among which are environmental as well as human specimens. *V. parahaemolyticus* was identified as a causative of food-borne illness and associated with consumption of crab (Dadisman, Nelson, Molenda, & Garber, 1972). It has also been very well defined as being associated with seafood ranging from crustacean, molluscan shellfish to even giant water prawn. In comparison to other species of the Vibrio genus, *V. parahaemolyticus* has emerged from a pre-pandemic to a pandemic clone which will be dealt later on. Its emergence has shown a drastic change in the presence and absence of certain pathogenicity islands which define these clones. The interesting factor about this *V. parahaemolyticus* is whether all bacteria cause infection or live symbiotically within the prawn.

2.4 *V. parahaemolyticus* and its significance

Temperatures & growth

Gram-negative bacterium, *V. parahaemolyticus* is curved rod-shaped and non-spore forming bacterium whose high motility is due to a polar flagellum. By a mechanism called swarming these microorganisms migrate across semi-solid surfaces (Barrow & Miller, 1974a) with the help of their lateral flagella. Throughout the world inshore marine waters are the primary area where distribution of *V. parahaemolyticus* is in abundance as it is said to be an inhabitant mostly of estuarine marine water. The effect of seasons on *V. parahaemolyticus* has reported (Colwell et al., 1984) that *V. parahaemolyticus* in a small number has been isolated from among sediment samples of marine water and was not detected during the period of winter (November - March) in Chesapeake Bay seawater. *V. parahaemolyticus* is proposed to multiply when there is

an increase in temperature that is by re-introduction of the microorganism into the sea water and to live in the marine sediments throughout the winter.

The temperature ranging from 35^0C to 39^0C (Jackson, 1974) are the optimal conditions for the growth of *V. parahaemolyticus*. Though the doubling time of *V. parahaemolyticus* is as little as 5 min (Barrow & Miller, 1974b) in certain conditions, under optimal conditions this organism has a generation time of less than 20 min. Hence, *V. parahaemolyticus* is most prevalently observed to be in a suitable environment in the course of the warm season in summer being a mesophillic bacterium and cause food borne outbreaks which peaks in summer (Daniels, MacKinnon, et al., 2000; Daniels, Ray, et al., 2000). Though the range of *V. parahaemolyticus* in seafood which is freshly harvested are rather lower than the dose of infection which has been predicted (Sanyal & Sen, 1974) the rapid multiplying ability of this bacteria at suitable temperatures shows its presence in food enough to cause a disease.

Salinity

V. parahaemolyticus has an important requirement for its multiplication and living conditions that is salinity. *V. parahaemolyticus* encounter salinity concentrations in marine environment typically ranging between 0.8% and 3% (DePaola, Kaysner, Bowers, & Cook, 2000). With optimal levels ranging between 1% and 3%, *V. parahaemolyticus* can thrive very well in a range of concentrations of sodium chloride that is, between 0.5% and 10% based on laboratory studies.

Metals

Apart from salinity, the capacity of the organism to utilize, tolerate and thrive in marine conditions is affected by various concentrations of metal ions present. *V. parahaemolyticus* isolates are found to survive in 300 mM magnesium a condition which is considered as toxic to various other microorganisms which is one of the example from severely polluted coastal waters in some parts of India (Bhattacharya,

Roy, Biswas, & Kumar, 2000). *V. parahaemolyticus* survival under several conditions can be improved by the increase in the ability to utilize magnesium. A 5.5 kb plasmid is said to carry the genes responsible for high resistance to increased magnesium concentrations (Bhattacharya et al., 2000). Injured or thermally treated *V. parahaemolyticus* cells show increased uptake of magnesium which indicates a possible increased requirement for magnesium not only for the stability and repair (Heinis, Beuchat, & Boswell, 1978) of ribosome, but also the membrane.

V. parahaemolyticus capability to survive magnesium or any metal ion at high concentrations out-competes other basic flora of seawater for its own survival and growth in the presence of the ions.

Virulence factors

The mechanisms specific to mount an infection which contribute to the *V. parahaemolyticus* strain's ability which lack the virulence factors, well recognized as the TRH (TDH related hemolysin) and the TDH (Thermostable direct hemolysin) are still unknown. The following factors described below show association with virulence properties exhibited in *V. parahaemolyticus*. Clinically *V. parahaemolyticus* exhibits beta hemolysis (β-hemolysis) on Wagatsuma agar through TDH production (Yeung & Boor, 2004) and is characterized as "Kanagawa phenomenon" positive with only less than 5% of isolates from the environment producing TDH (Miyamoto et al., 1969), it has been shown to be associated with clinical isolates, describing TDH production as a virulence factor. When compared to TDH positive (*tdh*+) wildtype *V. parahaemolyticus*, a *tdh* mutant was observed in a model of a ligated rabbit ileal loop with nil fluid accumulation (Nishibuchi, Fasano, Russell, & Kaper, 1992) hence describing TDH to be a virulence factor of considerable importance in strains which are *tdh*+. Resulting in altered ion flux TRH induces calcium activated chloride channels. Showing a decreased accumulation of fluid in rabbit ileal loops (Xu, Yamamoto,

Honda, & Ming, 1994) is a genetically modified TRH in a KP-negative strain which gives support to TRH's role in *V. parahaemolyticus.* As gastroenteritis is mostly being linked to the involvement of *trh* and *tdh* genes (Shirai et al., 1990), both these genes have enough background to qualify as virulence factors of *V. parahaemolyticus*. Neither of these hemolysin genes has been possessed by most environmental isolates though their presence in few clinical isolates has been studied (Shirai et al., 1990). This shows that a bacterial strain's capacity to elicit an infection does not depend only on the presence of the *tdh* gene (Hoashi et al., 1990) and suggest that expression of TDH may vary among different strains. Hence, results suggest that TDH is not the one and only factor contributing to the pathogenicity of *V. parahaemolyticus,* but there are many additional virulence factors which could be contributing to the pathogenesis of the bacteria.

The pathogenic strains of *V. parahaemolyticus* constituting the chromosome of its subset (Iida et al., 1995) carries the *trh* and *ure* (Urease) genes in close proximity which can be correlated to the urea hydrolysis being shown as a decent marker for the strains *V. parahaemolyticus* in the Pacific Northwest (Kaysner et al., 1994) associated with sporadiac illness. From a number of patients as urease negative strains have also been isolated including the new O3:K6 strains, its specific role in the pathogenesis of *V. parahaemolyticus* is still not clear and the urease activity cannot be defined as a universal marker for its virulence.

Bacteria are well known to carry with them iron storage proteins (Andrews, 1998) that are heme free ferritins and heme-containing bacterioferritins. A novel siderophore called vibrioferrin (Yamamoto et al., 1994) is produced by *V. parahaemolyticus* under iron limiting conditions. For example, in the human host with conditions where it is iron-limiting environments, strains could be provided with an advantage of surviving with an increase in the production of vibrioferrin. Increased

haemolytic activities, higher adherence and greater proliferation rate was demonstrated through culture of *V. parahaemolyticus* in iron limiting conditions. The possible relation between the concentration of iron present in the growth media of bacteria and the *V. parahaemolyticus* virulence was observed in previous studies (H. C. Wong & Lee, 1994). Hence, contribution towards the pathogenesis of *V. parahaemolyticus* may also be due to the production of vibrioferrin. The emergence of unique variety of strains with enhanced virulence among the existing strains of *V. parahaemolyticus* is due to the incorporation of new genetic elements within its genome. In *V. parahaemolyticus*, while the rest of the genome has 45.4% as an average GC content, the pathogenicity island carries a GC content of 39.8%. It is the smaller chromosome where this pathogenicity island is located which suggests the island as being acquired due to horizontal transfer.

Two genes of *tdh* and some other genes are borne by the pathogenicity island which has association with the virulence properties of other organisms which include the genes encoding the Pseudomonasexoenzyme T and the cytotoxic necrotizing factor of *E. coli* homologues. As the secretion mechanism is triggered on the movement of a bacterium into close contact with the host cells the Type III secretion system (TTSS) is also known as a contact-dependent secretion. *V. cholerae* does not show the presence of TTSS. While *V. cholerae* causes non-inflammatory secretory diarrhoea, inflammatory diarrhoea is considered to be shown by *V. parahaemolyticus* even though some clinical manifestations could be common between the infections caused due to *V. parahaemolyticus* and *V. cholerae.*

V. parahaemolyticus with the presence of TTSS brings out the major differences in the pathogenicity between closely related organisms. The *tdh* and the *trh* genes are suggested to be important related to *V. parahaemolyticus* in terms of virulence factors based on epidemiological studies, but these would be well supported by further research studies.

Serotyping

It is based on the antigens defined as O and K that the strains of the *V. parahaemolyticus* are typically serotyped. 13 O groups and 71 K types of strains can be marked as identified by the use of commercially available antisera. Since 1996 it is the serovar O3:K6 (*tdh*+ and TRH negative (*trh*-)) which is highly predominant. It is to this serovar that the outbreaks related to the bacteria *V. parahaemolyticus* are linked to (H. Han et al., 2008). There has been a recent emergence of a new O3:K6 serovar which has been studied (Chowdhury, Stine, Morris, & Nair, 2004; H. C. Wong et al., 2007). The derivatives of this new O3:K6 serovar that are O1:KUT, O4:K68 and the O1:K65 isolated since the same year 1996 have similarity in terms of sequence types (STs) or fingerprint patterns (FPs) constituting a clonal complex collectively together defined as the "pandemic group". These pandemic groups are said to possess a number of unique DNA markers such as ORF8 in phage f237, toxRS/new sequence (GS-PCR), an ORF VP2905, AP-PCR fragment of 932bp and a sequence of insertion within the gene Hu-alpha (Hu alpha-insertion) (Okura et al., 2003; Okura et al., 2004). Two genome sequences that is, one complete sequence of non-pandemic strain of O3:K6 AQ3810 and that of a pandemic strain O3:K6 RIMD2210633 (Boyd et al., 2008; Makino et al., 2003) were determined. Studies showed that none of the ORF8, Hu-alpha insertion and toxRS/new sequence were suitable to identify strains as being pandemic by PCR (Meador et al., 2007). Further work showed that the cluster VP2900, VP2910 with genomic island VPaI-5 and GS-PCR sequence were shown to be specific genomic features to the pandemic group of strains. Studies (H. Han et al., 2008) show that the present pandemic clone (*trh*-, GS-PCR+ and *tdh*+) has emerged from that of its previous O3:K6 clone (TRH positive (*trh*+), GS-PCR and TDH negative (*tdh*-)). The pandemic clone has been shown to consist of five serovars, O:KUT (1strain), O1:K25(3 strains), O6:K18 (1 strain), O4:K68 (5 strains) and O3:K6 (29 strains). Apart from

arising from old O3:K6, it has been recently reported that the pandemic clone acquired genomic regions including type IV secretion system and VPaI-1 to VpI-7. The pandemic clone microevolution has been hence not only an acquisition of new genomic islands, but also differentiation in genes of O/K antigen providing us with the pandemic group's evolutionary history.

Food effect

The concentration of *V. parahaemolyticus* goes to higher levels in shellfish than those in the surrounding water as they obtain food by filter feeding. The raw and under-cooked sea-food consumption such as the molluscan shellfish that are clams, mussels and oysters therefore cause illness due to *V. parahaemolyticus*. Cooked products being contaminated or improper cooking could show the presence of *V. parahaemolyticus* as it is extremely sensitive to heat.

Symptoms

The consumption of raw seafood causes major infections of *V. parahaemolyticus* infections and are considerably important in countries like Japan, but in countries wherein seafood is consumed after being well-cooked, *V. parahaemolyticus* associated diseases occur less frequently. In general, reports of dosage of infection are considered to be approximately 10 to the power of 5 per organism (Zhaxybayeva et al., 2006). *V. parahaemolyticus* linked to watery diarrhoea causes gastroenteritis and is very often followed by nausea, fever, vomiting and even abdominal cramps. These symptoms appear within a day after consuming under-cooked or raw seafood such as fish and shellfish. It usually resolves in 3-7 days in case it is a self-limiting infection. The presence of blood and mucus are observed in the watery stool in cases of severity.

2.5 *Macrobrachium rosenbergii*

The farming in modern times of *M. rosenbergii* started in early 1960's. It was proposed to require brackish conditions for its survival though found even in freshwater (New, 2004).

M. rosenbergii is one freshwater prawn species of which there is considerable production compared to that of *Macrobrachium nipponense* which appears less in official details of prawn production. Though it is influenced by the areas of brackish water this species resides in tropical environments of freshwater. The female body bears a gelatinous mass underneath between the fourth pair of its walking legs. It is here that the male deposits the sperm. After a few hours of mating, eggs are laid and fertilized by the sperm. "Berried Females" is the terminology used for females carrying the eggs (D'Abramo & Brunson, 1996). During the course of embryo development the eggs remain constantly adhered to the female. It is during this time that the females migrate towards estuaries as the larvae cannot survive in freshwater for more than two days. The eggs hatch in brackish water whose salinity ranges from approximately 9 parts per thousand (ppt) to 19 ppt (D'Abramo & Brunson, 1996). They exist as free-swimming larvae at this stage.

The change in phase from a larval to a post larval stage is very crucial in a prawn's life cycle as it grows through a process of moulting (Saravanan, Biju Sam Kamalam, Stephen, & Kumar, 2008). It undergoes around almost 11 molts to transform into post larvae. These molts represent a process of metamorphosis. This stage can be explained as a critical part of a prawn's life cycle as the old exoskeleton is replaced by a new soft exoskeleton underneath. It is here that the *M. rosenbergii* absorbs water into the tissue to increase in size. Hence, the environmental condition to which the *M. rosenbergii* is exposed to, further enhances its ability to grow into a healthy adult or decreases its chances of survival. These physiological traits of both *V.*

parahaemolyticus and *M. rosenbergii* could be used to further study the host-pathogen interactions.

CHAPTER 3

UNDERSTANDING BACTERIAL GENOMES & OPTIMIZING A STANDARD QUALITY CONTROL PROTOCOL FOR BACTERIA USING NEXT-GENERATION SEQUENCING RAW READ DATA

3.1 Introduction

Bacterial whole genome sequencing requires a number of quality controls entailing the assembly, annotation and alignment processes to herald a bacterial genome from the test tube on to the computer. The primary issues with next-generation sequencing data is the raw read data constituting contamination in terms of sequencing adaptors and low quality reads in terms of per base quality. Any bacterial genome raw read sequencing data processed without a proper quality control procedure could affect further steps that are assembly, annotation and alignment. There are several bacterial genomes that differ from each other in a number of ways such as bearing a small genome size, presence of two chromosomes, a completely new genome/ a *de-novo* assembled genome, etc. Such as these a number of genomes have been quality filtered, assembled and analysed till date with various different approaches to finish bacterial genomes. In our present work we have concentrated upon the contigs obtained by one simple assembly process and validated backwards to define the importance of reads and the data which we obtain from the contigs. Quality control (QC) tools were developed previously on the basis of four main strategies that are: high quality filtering, adaptor filtering, N's removal and trimming. The arrangement of these strategies while processing raw reads under quality control could have an effect on the resulting output read data. A several number of QC tools have been made since the start of Next-generation sequencing which had made its entry in 2005 (Cox, Peterson, & Biggs, 2010; Dai et al., 2010; Dodt, Roehr, Ahmed, & Dieterich, 2012; Lassmann, Hayashizaki, & Daub, 2009; Patel & Jain, 2012; Schmieder

& Edwards, 2011; Yang et al., 2013). Usage of the pre-existing tools with application to the above defined strategy is the aim of our study. The validation results obtained for the selected five bacterial genome data while processing for GC content, RNA information, QC and assembly hybrid data could serve as a reference to have one common protocol to deal with bacterial genomes. Hence, the problems faced can therefore be prevented by the first step of bioinformatics in Next-generation sequencing analysis that is the Quality control.

3.2 Review of the literature

With the generation of high amounts of next-generation sequencing data the need to develop reliable tools for every step of analysing data has been very important. Apart from each tool for analysis, there has been much difference in handling data generated from several different sequencing platforms (Kircher & Kelso, 2010). The value of sequencing a new bacterial genome has been increasing as a rapid rise in the number of sequencing datasets of genomes could affect the loss of important biological information (Barbosa et al., 2014). There have been various studies on how to comply with the usage of the best possible information from the sequencing reads. These were studied in detail as applications of next generation sequencing to bacterial genetics (Hall, 2007; MacLean, Jones, & Studholme, 2009; Mardis, 2008; Morozova & Marra, 2008; Shendure & Ji, 2008). A guide to selecting appropriate read lengths across prokaryotic genomes as an algorithm has also been reported (Cahill, Koser, Ross, & Archer, 2010). A fixed assembly strategy regardless of the emergence of the sequencing reads has been reported recently which involved the stages of pre-filtering processes (El-Metwally, Hamza, Zakaria, & Helmy, 2013). Quality control for non-model organisms such as those without a defined reference so as to carry ahead for assembly process have also been built (Trivedi et al., 2014). Several quality control tools have been developed which are directed towards high quality read filtering, trimming of sequencing read data, adaptor removal, in short cleaning up of artefacts. These tools were developed based on both depending on graphic user interface as well as command-line usage. SeqQC (http://genotypic.co.in/SeqQC.html) has been available for both windows and Linux operating systems to process high quality filtering, inbuilt adaptor filtering as well as trimming. Tools such as FASTX (Taylor, Schenck, Blankenberg, & Nekrutenko, 2007), NGStoolkit (Patel & Jain, 2012) , PRINSEQ (Schmieder &

Edwards, 2011), SolexaQA (Cox et al., 2010) are highly used in recent times for quality control.

Out of all the strategies mentioned above a study has shown that across different types of NGS data, trimming quality with a cut off of thirty could help increase the quality and reliability of the analysis (Del Fabbro, Scalabrin, Morgante, & Giorgi, 2013). Few tools were made to specifically target the sequencing data of Illumina read information (Bolger, Lohse, & Usadel, 2014; Yang et al., 2013). Few others show quality control pipelines constituting simulated data and real time metagenomics datasets which reported lesser processing time as well as the identification of contamination identification ability (Zhou, Su, & Ning, 2014). NGSQC has been one cross-platform analysis pipeline which has utilized read data regardless of the assay technologies which shows that it can help researchers to understand the quality issues with sequencing (Dai et al., 2010). A study based on three stage quality control strategy was also reported considering the raw data, alignment and variant calling which explains the importance of control strategies at three different levels of DNA sequencing analysis (Guo, Ye, Sheng, Clark, & Samuels, 2014).

The primary aim of this objective is to identify the best strategy among the quality control steps and create a pipeline with the order of these strategies so as to utilize maximum percentage of reads. With the increase in the number of tools for quality control and data generation from next generation sequencers, it could be important to utilize the pre-existing tools. Utilizing these routinely used tools to facilitate access to reliable data generation could be currently very significant in the next-generation sequencing market.

3.3 Materials and Methods

3.3.1 Next – generation sequencing raw reads of bacterial genomes

Five different bacterial genome next – generation sequencing data were selected for our studies (Avasthi, Kumar, et al., 2011; Devi et al., 2010; Majid et al., 2014; Siddavattam et al., 2011; Tiruvayipati et al., 2013). The genomes selected were namely *Helicobacter pylori* 908 (HP), *Escherichia coli* NA114 (EC), *Mycobacterium tuberculosis* NA1 (MTB), *Vibrio parahaemolyticus* PCV08-7 (VP) and *Paracoccus denitrificans* SD (PD) corresponding to sequencing data as mentioned in Table 3.3.1. Among which HP, EC, VP and PD correspond to Illumina sequencing data, while MTB correspond to ion torrent data, respectively. These bacterial genomes were selected based on their differences entailing high GC content, de-novo assembly, high AT content, presence of two chromosomes and ion torrent data.

Table 3.3.1: Next-generation sequencing data of selected bacterial genomes

Bacterial genome	Strain name	Accession number	Total number of reads	Read length (bp)	Coverage	Source
Helicobacter pylori	908	CP002184	3,015,238	54	~100X	From African patient suffering from duodenal ulcer, France
Escherichia coli	NA114	NC_017644	45,440,998	54	~50X	Urine of 70-year-old male patient, India
Paracoccus denitrificans	SD1	CP002897	3,951,522	72	~30X	Coal mine leftovers, India
Mycobacterium tuberculosis	NA-A0008	ALGY00000000	15,839,552	72	~200X	Patients in rural Odisha, India
Vibrio parahaemolyticus	PCV08-7	AOCL00000000	4,356,146	100	~80X	From prawn, Malaysia

3.3.2 Quality control

The next – generation quality control tools in our study were SeqQC, FASTX, NGStoolkit, PRINSEQ, SolexaQA, Flexbar_v2.33 (Dodt et al., 2012) and Tagdust (Lassmann et al., 2009). Among these the SeqQC tool has been used to plot the phred quality scores of the raw read data and Solexa QA to generate heatmap to identify the quality level per base.

The tools were used based on different quality control categories: High quality filtering, Read trimming, N's removal and Adaptor removal as in Figure 3.3.1 to validate sequencing data. The quality control parameters were made in reference to our filtering protocols which were published (Baddam, Kumar, et al., 2012; Baddam, Thong, et al., 2012; Kumar et al., 2012)

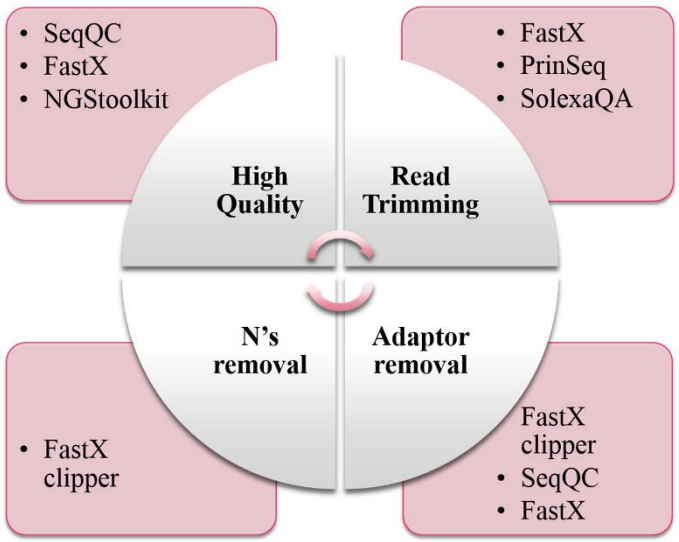


Figure 3.3.1: Strategies of quality control used for QC of bacterial genomes

3.3.3 High Quality filtering

SeqQC, FASTX and NGStoolkit were utilized directly on the raw sequencing data of the five selected bacterial genomes HP, EC, VP, PD, MTB. The cut-off value for high quality filtering of raw sequencing reads was set to greater than equal to 20 (>=20) to remove all low quality reads.

3.3.4 Read Trimming

As mentioned in the previous studies that trimming with a quality score cutoff of 30 has been good in filtering data (Del Fabbro et al., 2013), we have used three different tools that are FASTX trim, PRINSEQ trim and SolexaQA trim for checking the effects of trimming utilizing a common quality trimming cutoff of 30. Obtained from the quality

graph of SeqQC we have input values for phred score cutoff of greater than equal to 30 for all the bacterial sequencing reads as shown in Table 3.3.2.

Table 3.3.2: Trimming values of reads based on phred quality scores of bacterial genomes

	Read length (bp)	Read 1	Read 2
Helicobacter pylori 908	54	Phred quality > 35	Phred quality > 34
Escherichia coli NA114	54	Phred quality > 34	Phred quality > 33
Paracoccus dentrificans SD1	72	65	63
Mycobacterium tuberculosis NA-A0008	72	54	54
Vibrio parahaemolyticus PCV08-7	100	88	94

3.3.5 Adaptor File

The adaptor file had been made with a combination of two files which were taken from the Flexbar_v2.33 as well as Tagdust quality control tools constituting of the following sequences Illumina Single end adapter 1, Illumina Single end adapter 2, Illumina Single end PCR primer 1, Illumina Single end PCR primer 2, Illumina Paired end adapter 1, Illumina Paired end adapter 2, Illumina Paired end PCR primer 1, Illumina Paired end PCR primer 2, Illumina multiplexing adapter 1, Illumina multiplexing adapter 2, Illumina multiplexing PCR primer 1.01, Illumina multiplexing PCR primer 2.01, Illumina multiplexing read1 primer, Illumina multiplexing index primer, Illumina multiplexing read2 primer, Illumina PCR primer index (1 – 12), TruSeq universal adapter, TruSeq Adapter Index (1 – 12), Solexa 5' adapter, Solexa 3' adapter, PE adapters1, PE adapters1_1, PE PCR primers1, PE PCR primers1_1, PE sequencing primer, PE sequencing primer 1, Gex adapter 1, Gex adapter 1_2, Gex adapter 2, Gex

adapter 2_2, Gex PCR primer 1, Gex PCR primer 2, Gex sequencing primer, adapters1, adapters1_1, PCR primers1, PCR primers1_1, Genomic DNA sequencing primer.

All mentioned adapter sequences were maintained at a uniform minimum length of 12 following length maintained in Tagdust adaptor file. The adaptor file is as mentioned in Appendix A (a).

3.3.6 N's removal

To remove all the raw sequencing reads containing at least one "N" instead of bases A, T, G, C, FASTX clipper has been used.

3.3.7 Adaptor removal

Adaptor filtering has been done with two different methods combining High quality and adaptor filtering as well as only adaptor filtering using the tools FASTX and SeqQC.

3.3.8 Assembly

The results of the quality control tools were validated using one step assembly that is by Velvet version 1.0.12a (Zerbino & Birney, 2008) assembly tool. Apart from this the bacterial genomes were compared of their contig information with two assembly tools: one based on the principle of overlapping reads of assembly "Edena" (Hernandez, Francois, Farinelli, Osteras, & Schrenzel, 2008) and the second based on the de-bruijn method "Velvet" version 1.0.12a.

3.3.9 Alignment

The contigs generated by Edena were re-aligned according to the contig arrangement of contigs generated by Velvet for individual bacterial genomes using Mauve (Darling, Mau, & Perna, 2010) alignment tool. The alignments of Edena re-arranged contigs with Velvet contigs were plotted using M-GCAT (Treangen & Messeguer, 2006) multiple genome alignment tool.

3.3.10 rRNA and GC content

rRNA copies in the contigs were determined using RNAmmer 1.2 Server (Lagesen et al., 2007). GC content has been determined using the Artemis tool (Rutherford et al., 2000).

3.3.11 Customized pipelines

Based on the usage of quality control strategies as mentioned above, 7 pipelines of difference in order of usage of the tools were formed to process the bacterial data for any difference in total percent of raw reads used. The last step of all these pipelines was assembly by the Velvet tool. The Pipeline 1 begins with N's removal with FASTX with adaptor removal by FLEXBAR followed by SeqQC to generate quality control graph, trimming by PRINSEQ and lastly High quality filtering by SeqQC and assembly by Velvet. (Figure 3.3.2)

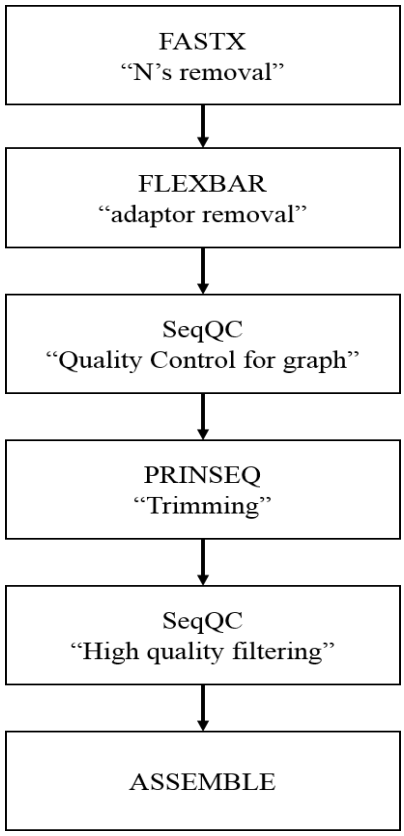


Figure 3.3.2: Flowchart showing the process of Pipeline 1

The Pipeline 2 follows trimming by PRINSEQ firstly then followed by adaptor removal by FLEXBAR by generating quality control graph through SeqQC with assembly by Velvet (Figure 3.3.3).

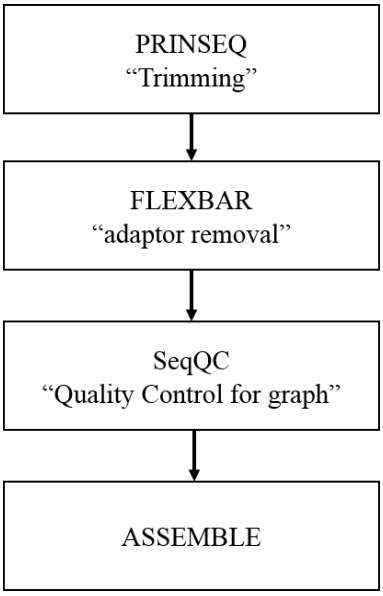


Figure 3.3.3: Flowchart showing the process of Pipeline 2

Pipeline 3 consists of trimming with PRINSEQ followed by adaptor removal by FLEXBAR and high quality filtering by SeqQC and finally assembly by Velvet (Figure 3.3.4).

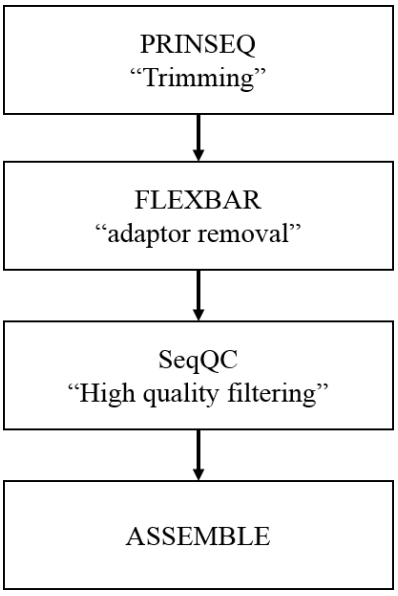


Figure 3.3.4: Flowchart showing the process of Pipeline 3

Pipeline 4 involves adaptor removal by FLEXBAR followed by quality control graph generation with trimming by PRINSEQ and high quality filtering by SeqQC followed with assembly by Velvet (Figure 3.3.5).

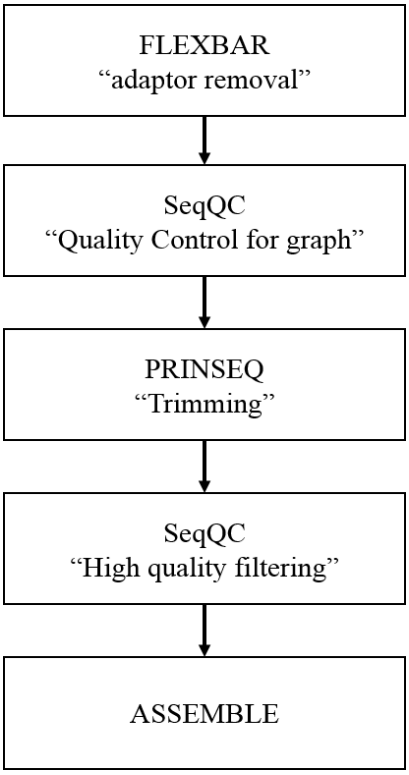


Figure 3.3.5: Flowchart showing the process of Pipeline 4

Pipeline 5 is ordered to filter adaptors by FLEXBAR followed by high quality filtering using SeqQC and trimming by PRINSEQ with assembly by Velvet (Figure 3.3.6).

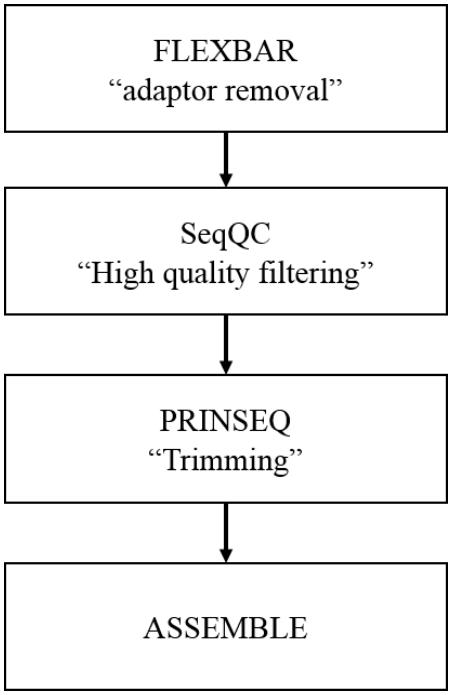


Figure 3.3.6: Flowchart showing the process of Pipeline 5

Adaptor removal with FASTX followed by high quality filtering SeqQC and assembly by Velvet has been the Pipeline 6 strategy as used for checking the best pipeline to use (Figure 3.3.7).

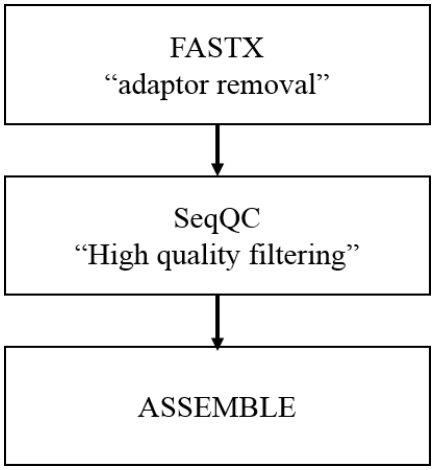


Figure 3.3.7: Flowchart showing the process of Pipeline 6

The 7th pipeline named as the Trimming-Adaptor-High quality (TAH) pipeline made was with trimming by NGStoolkit followed by adaptor removal using FLEXBAR with FASTX for high quality filtering and assembly by Velvet (Figure 3.3.8).

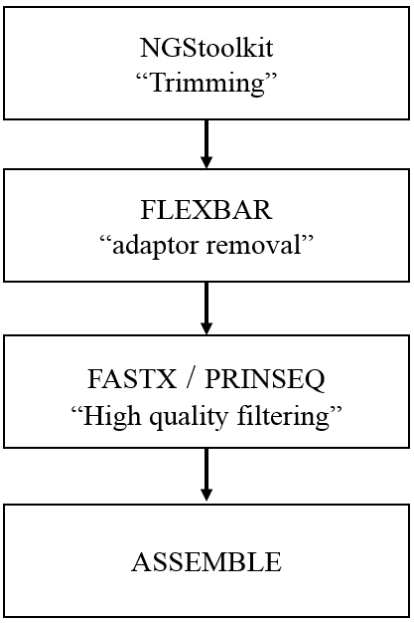


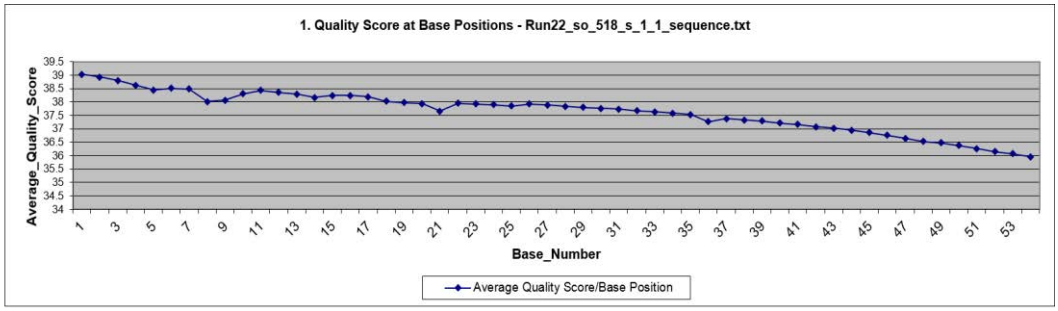
Figure 3.3.8: Flowchart showing the process of Pipeline 7 (TAH pipeline)

3.4 Results and Discussion

3.4.1 Next-generation sequencing & finishing bacterial genomes

The first step of processing raw read sequencing information has been the QC to avoid possible contamination by sequencing primers and adaptors. To evaluate the probable erroneous information which any bacterial raw read data could provide us with, we have performed a set of experiments. Five bacterial genomes HP908, ECNA114, PDSD1, MTBNA-A0008 and VPPV08-7 were selected which differ in their physiology and occurrence in the environment for this study. The analysis include checking raw reads of their phred quality scores and running heat plots to check the probable error rate per average base of any read length. The quality control graphs generated by SeqQC (Figures 3.4.1, 3.4.2, 3.4.3, 3.4.4 and 3.4.5) show how different any sequencing data could be which needs a standard protocol to be dealt with as a single moiety which in the present study as a "bacterial dataset".

A



B

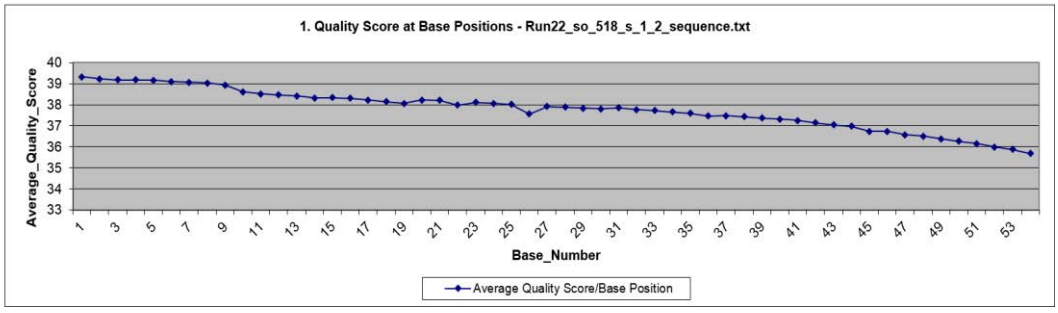


Figure 3.4.1: SeqQC quality graph representing average phred quality score per base position (A) Read 1 of *H. pylori* 908 (B) Read 2 of *H. pylori* 908
The above plot shows that the average phred quality score for total number of reads generated per base position in the *H. pylori* 908 genome sequencing in Read 1 (Run22_so_518_s_1_1_sequence.txt) is greater than 35.5 and in Read 2 (Run22_so_518_s_1_2_sequence.txt) is greater than 35.

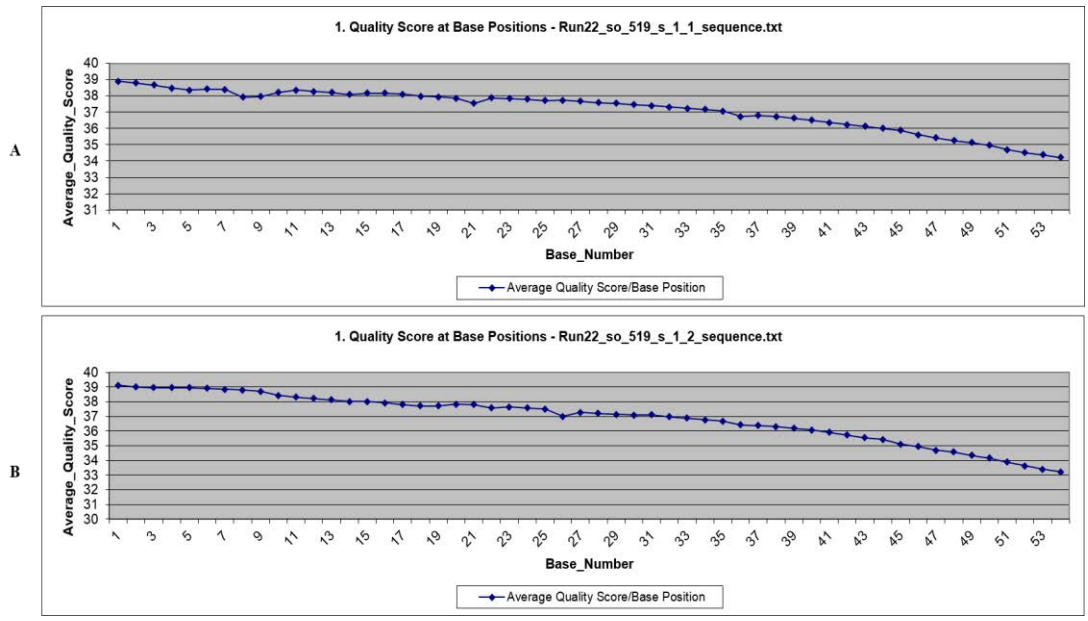


Figure 3.4.2: SeqQC quality graph representing average phred quality score per base position (A) Read 1 of *E. coli* NA114 (B) Read 2 of *E. coli* NA114
The above plot shows that the average phred quality score for total number of reads generated per base position in the *E. coli* NA114 genome sequencing in Read 1 (Run22_so_519_s_1_1_sequence.txt) is greater than 34 and in Read 2 (Run22_so_519_s_1_2_sequence.txt) is greater than 33.

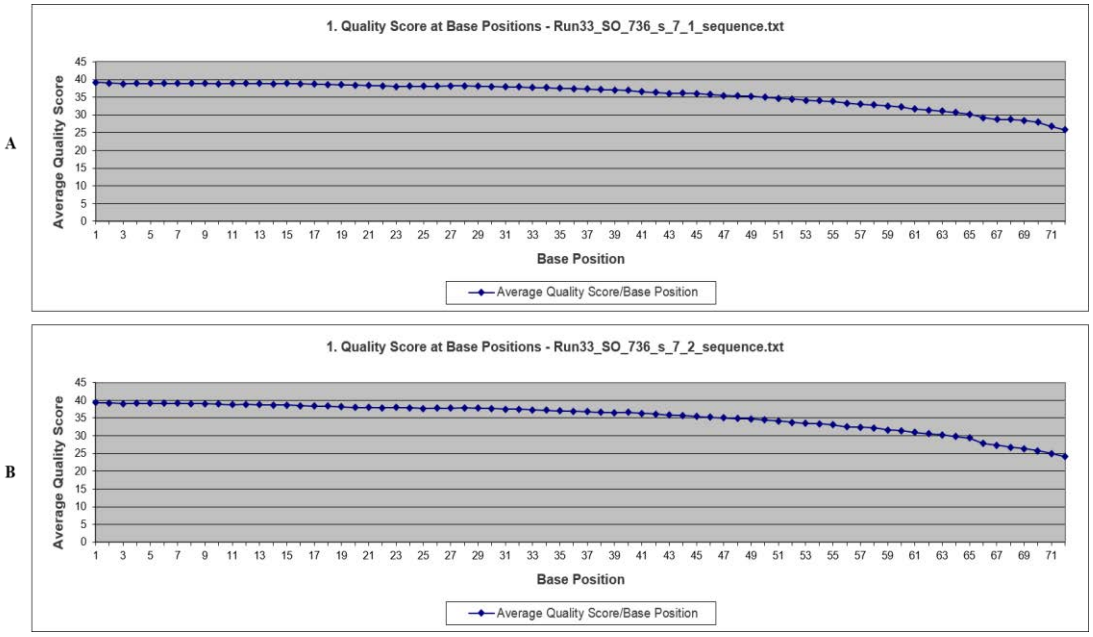


Figure 3.4.3: SeqQC quality graph representing average phred quality score per base position (A) Read 1 of *P. denitrificans* SD1 (B) Read 2 of *P. denitrificans* SD1
The above plot shows that the average phred quality score for total number of reads generated per base position in the *P. denitrificans* SD1 genome sequencing in Read 1 (Run33_SO_736_s_7_1_sequence.txt) is greater than 25 and in Read 2 (Run33_SO_736_s_7_2_sequence.txt) is greater than 20.

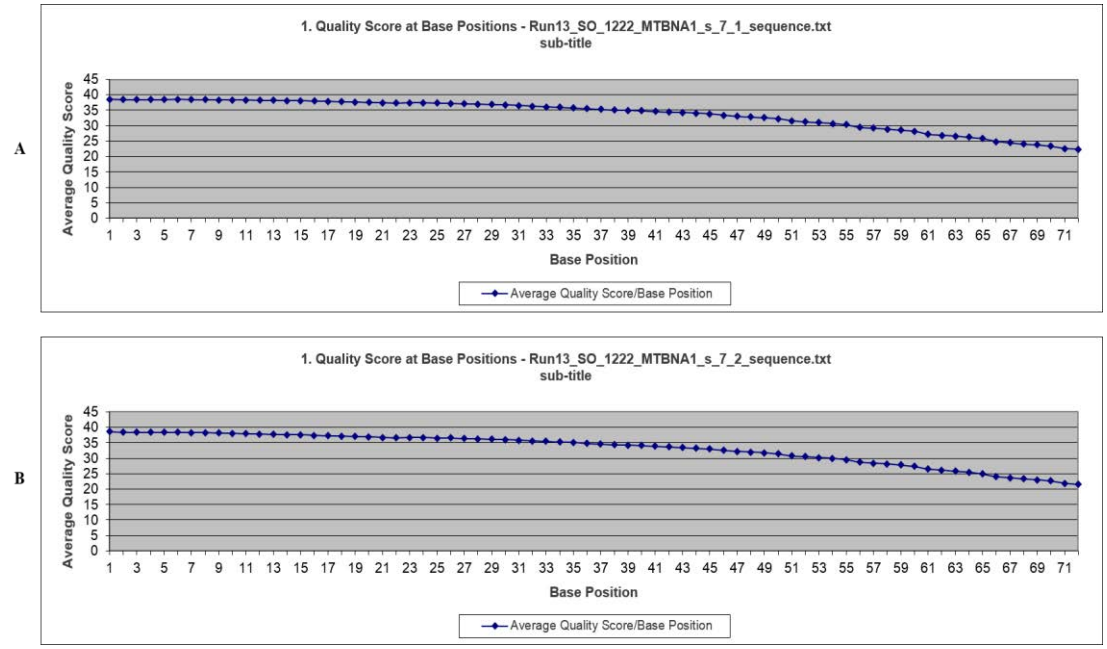


Figure 3.4.4: SeqQC quality graph representing average phred quality score per base position (A) Read 1 of *M. tuberculosis* NA-A0008 (B) Read 2 of *M. tuberculosis* NA-A0008

The above plot shows that the average phred quality score for total number of reads generated per base position in the *M. tuberculosis* NA-A0008 genome sequencing in Read 1 (Run13_SO_1222_MTBNA1_s_7_1_sequence.txt) is greater than 20 and in Read 2 (Run13_SO_1222_MTBNA1_s_7_2_sequence.txt) is greater than 20.

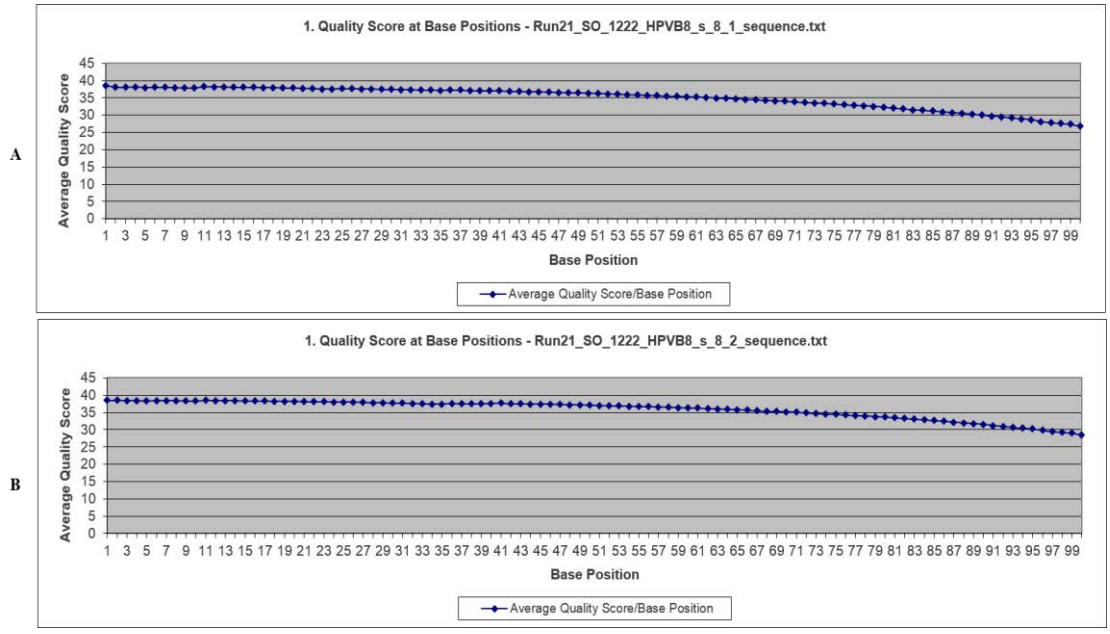


Figure 3.4.5: SeqQC quality graph representing average phred quality score per base position (A) Read 1 of *V. parahaemolyticus* PCV08-7 (B) Read 2 of *V. parahaemolyticus* PCV08-7

The above plot shows that the average phred quality score for total number of reads generated per base position in the *V. parahaemolyticus* PCV08-7 genome sequencing in Read 1 (Run21_SO_1222_HPVB8_s_8_1_sequence.txt) is greater than 25 and in Read 2 (Run21_SO_1222_HPVB8_s_8_2_sequence.txt) is greater than 25.

The heat plots generated by SolexaQA shows the average quality of a base of various read lengths (Figures 3.4.6, 3.4.7, 3.4.8, 3.4.9 and 3.4.10) in terms of 120 tiles in a flowcell.

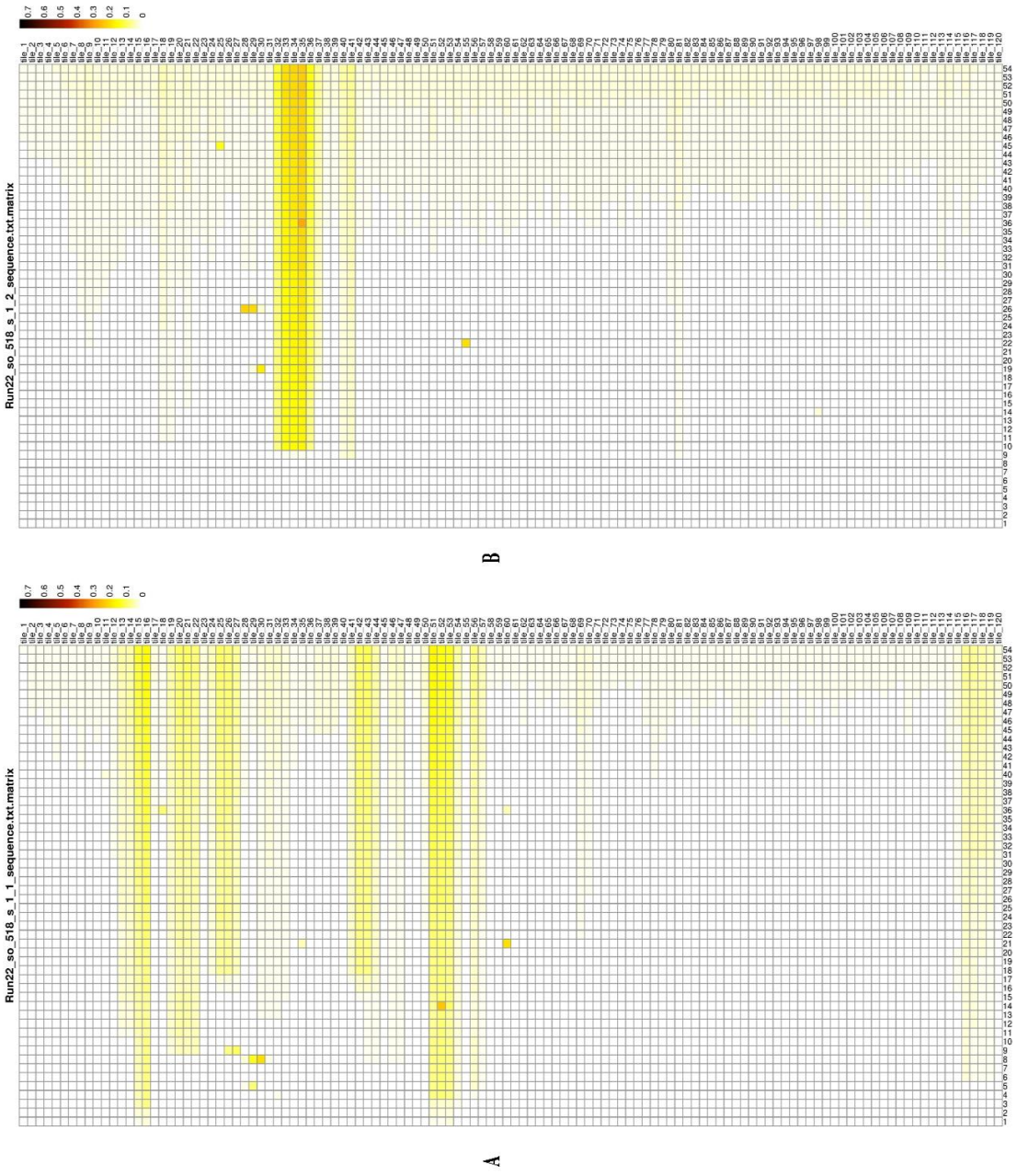


Figure 3.4.6: SolexaQA heatplots showing average per base quality of HP908 (wherein, white represents an error rate of 0 ranging through yellow and orange to pure black representing an error rate of 0.75) (A) Read 1 (B) Read 2

The above heat plot represents the average per base quality of HP908 sequencing reads of both Read 1 (Run22_so_518_s_1_1_sequence.txt) and Read 2 (Run22_so_518_s_1_2_sequence.txt) in terms of color coded error rates. The heat maps show a wide range of difference in the distribution of quality scores through the 120 tiles of a flow cell within the reads.

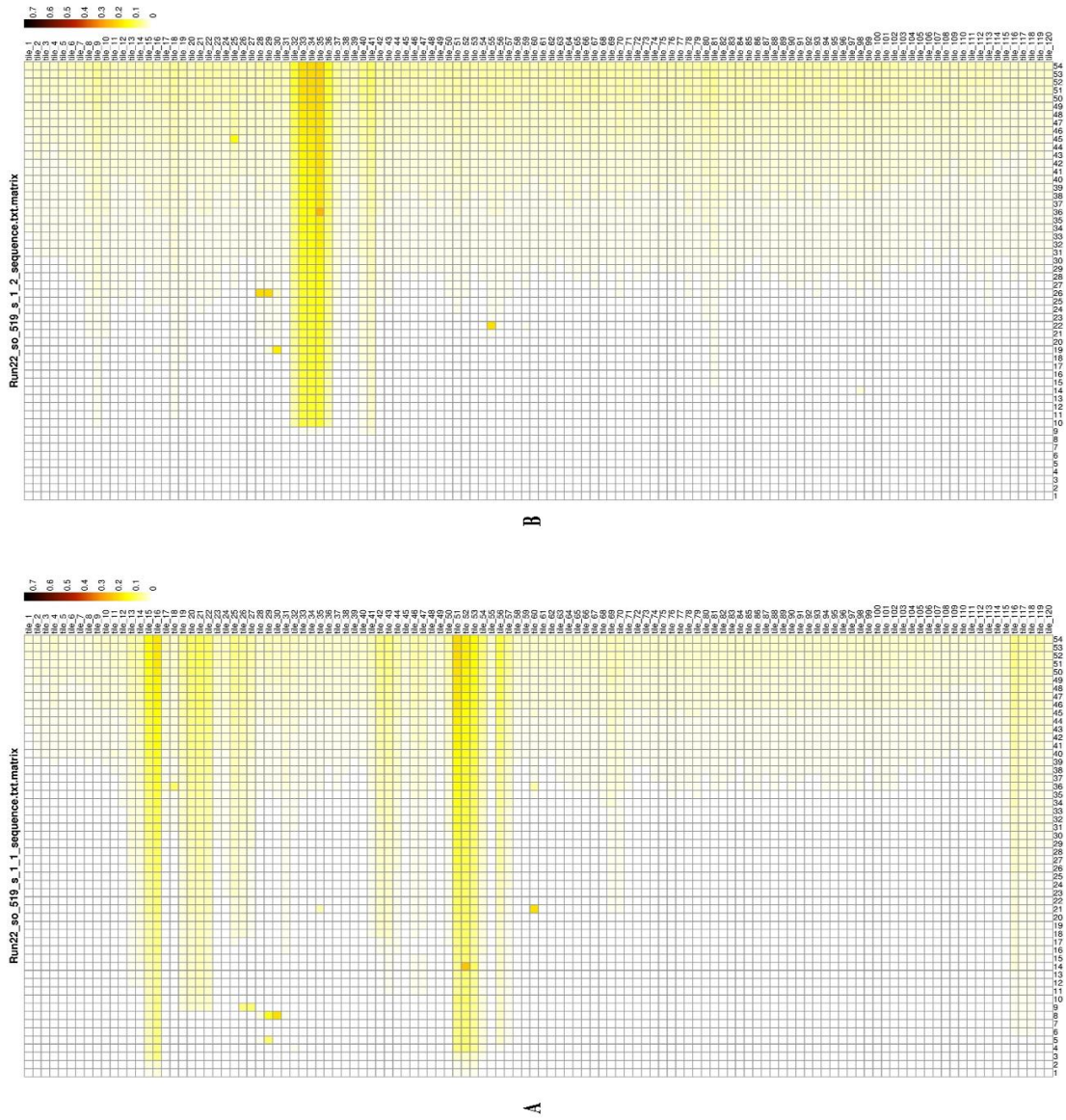


Figure 3.4.7: SolexaQA heatplots showing average per base quality of ECNA114 (wherein, white represents an error rate of 0 ranging through yellow and orange to pure black representing an error rate of 0.75) (A) Read 1 (B) Read 2

The above heat plot represents the average per base quality of ECNA114 sequencing reads of both Read 1 (Run22_so_519_s_1_1_sequence.txt) and Read 2 (Run22_so_519_s_1_2_sequence.txt) in terms of color coded error rates. The heat maps show a wide range of difference in the distribution of quality scores through the 120 tiles of a flow cell within the reads.

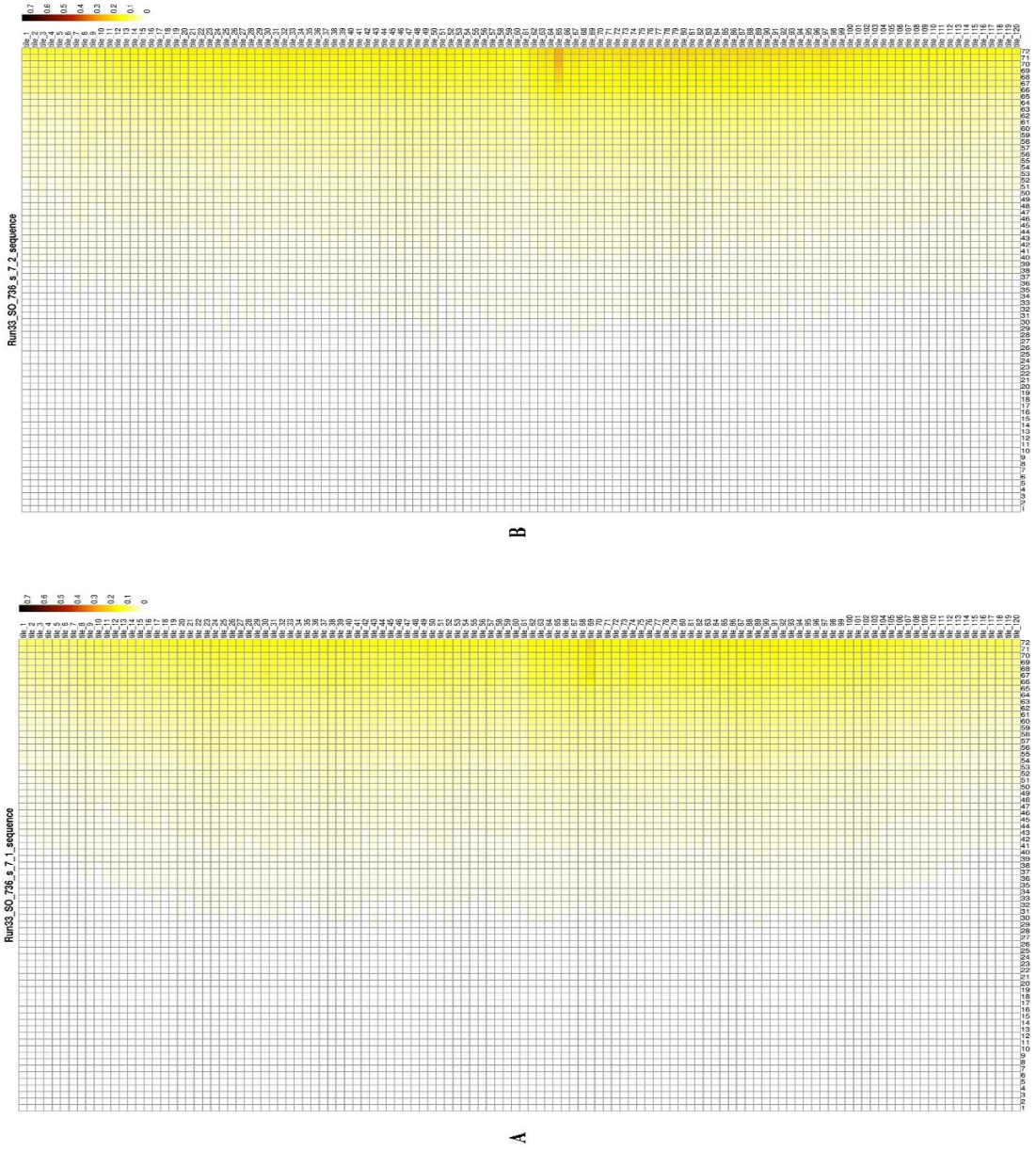


Figure 3.4.8: SolexaQA heatplots showing average per base quality of PDSD1 (wherein, white represents an error rate of 0 ranging through yellow and orange to pure black representing an error rate of 0.75) (A) Read 1 (B) Read 2

The above heat plot represents the average per base quality of PDSD1 sequencing reads of both Read 1 (Run33_SO_736_s_7_1_sequence.txt) and Read 2 (Run33_SO_736_s_7_2_sequence.txt) in terms of color coded error rates. The heat maps show a wide range of difference in the distribution of quality scores through the 120 tiles of a flow cell within the reads.

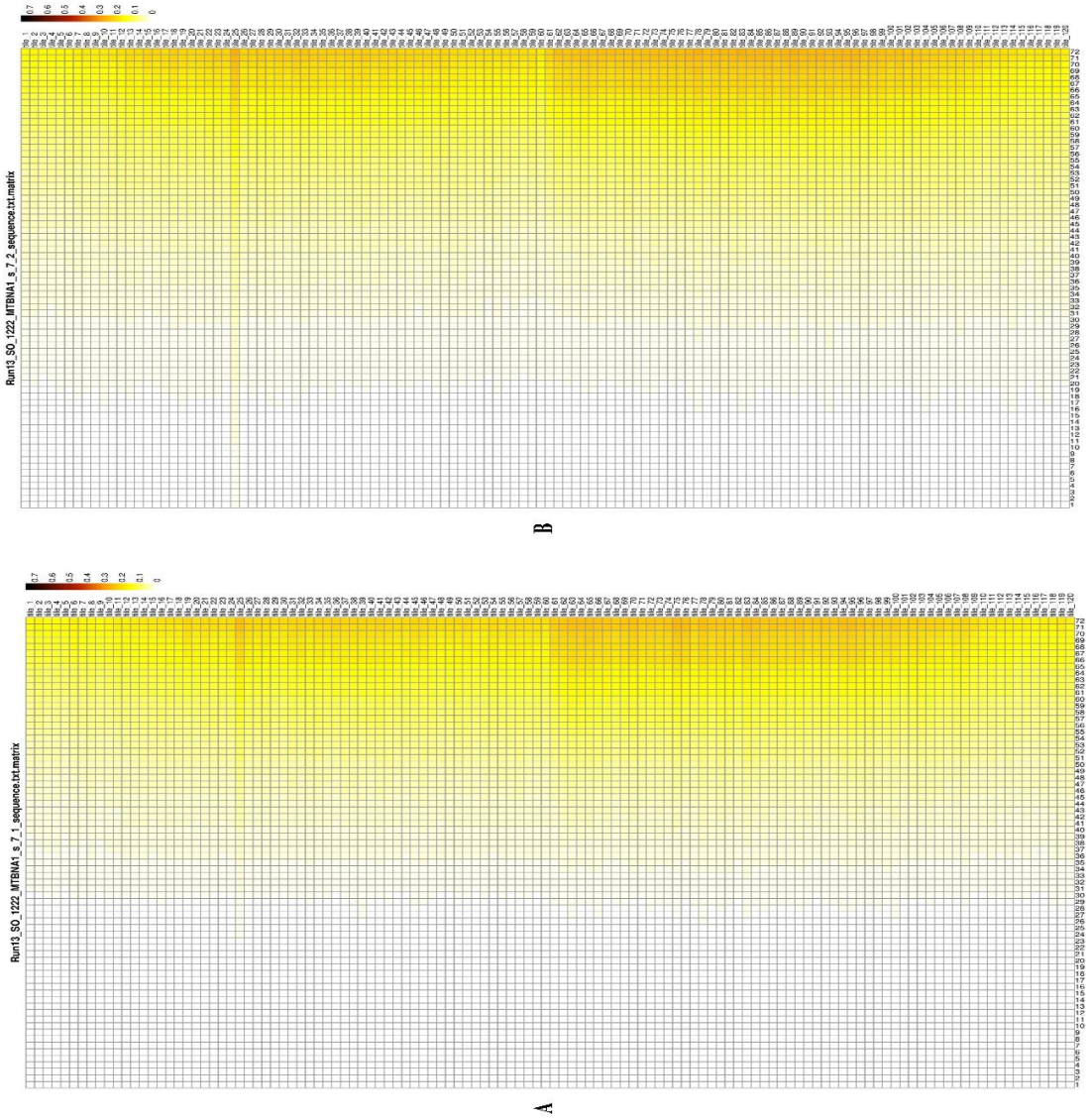


Figure 3.4.9: SolexaQA heatplots showing average per base quality of MTBNA-A0008 (wherein, white represents an error rate of 0 ranging through yellow and orange to pure black representing an error rate of 0.75) (A) Read 1 (B) Read 2

The above heat plot represents the average per base quality of MTBNA-A0008 sequencing reads of both Read 1 (Run13_SO_1222_MTBNA1_s_7_1_sequence.txt) and Read 2 (Run13_SO_1222_MTBNA1_s_7_2_sequence.txt) in terms of color coded error rates. The heat maps show a wide range of difference in the distribution of quality scores through the 120 tiles of a flow cell within the reads.

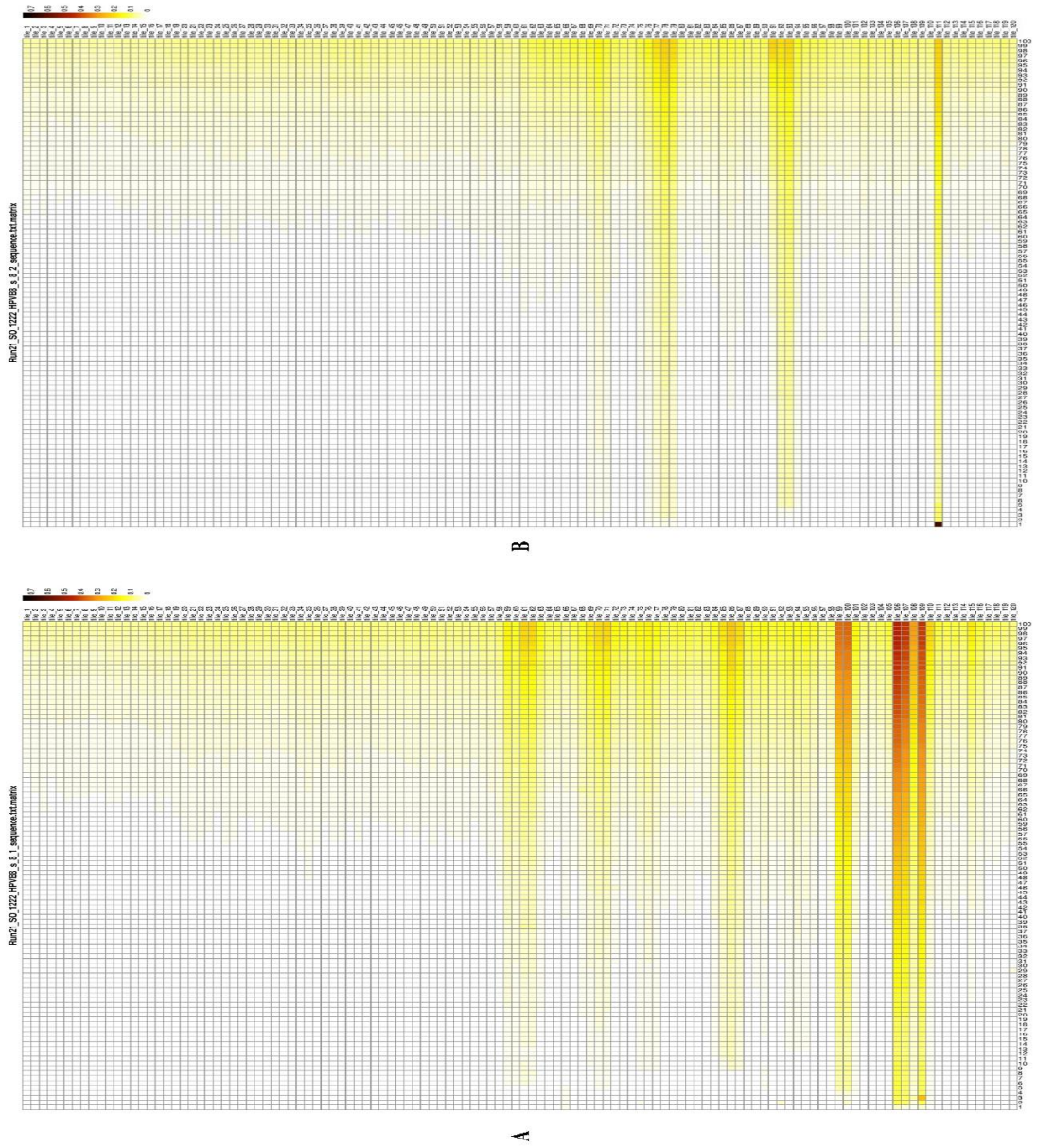


Fig 3.4.10: SolexaQA heatplots showing average per base quality of VPPCV08-7 (wherein, white represents an error rate of 0 ranging through yellow and orange to pure black representing an error rate of 0.75) (A) Read 1 (B) Read 2

The above heat plot represents the average per base quality of VPPCV08-7 sequencing reads of both Read 1 (Run21_SO_1222_HPVB8_s_8_1_sequence.txt) and Read 2 (Run21_SO_1222_HPVB8_s_8_2_sequence.txt) in terms of color coded error rates. The heat maps show a wide range of difference in the distribution of quality scores through the 120 tiles of a flow cell within the reads.

The plots represent a range of different read lengths of bacterial genomes consisting of randomly different qualities ranging from white, explaining the occurrence of an error to be zero to black, being the highest error prone base. This when checked along all the selected genomes show how the plots differ completely having a zero error to possibly highly error prone bases in sequencing data. This could give us an insight on how these slightly error prone bases could also incorporate change in the actual genome assembly of a bacteria with newer gene information altogether.

In continuation we have concentrated on the primary information which a bacterial genome can give the user. That is in terms of uniqueness of the 16S rRNA genes which are well known to be conserved hence, the rRNA genes (5S, 16S, and 23S) were selected. Along with this two different strategies of assemblies were employed: one being the de-brujin method (Velvet) and the overlapping method (Edena) to obtain contigs. The contigs generated were checked for alignments among data of these two assemblers a method that we would like to name "Self-assembly alignment" which differs from previous studies which were performed among different next – generation sequencing platform data to obtain a hybrid assembly (Koren et al., 2012; Utturkar et al., 2014). We have used the Velvet and Edena contig information to scan the number of rRNA in the individual genomes (Figures 3.4.11, 3.4.12, 3.4.13, 3.4.14 and 3.4.15) which had shown different number of copies in each bacteria.

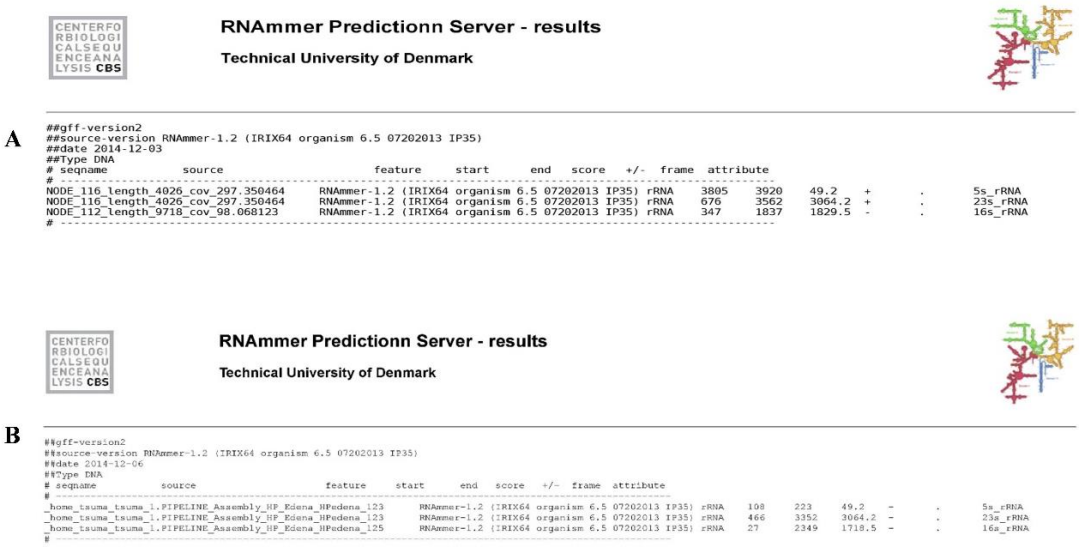


Figure 3.4.11: RNAmmer 1.2 Server results for rRNA genes of HP908 in (A) Contig assembly by Velvet (B) Contig assembly by Edena

RNA prediction results for HP908 from RNAmmer prediction server showing the detection of one copy of 5S rRNA, 23S rRNA, 16S rRNA each respectively in Velvet assembled contigs and one copy of 5S rRNA, 23S rRNA, 16S rRNA each respectively in Edena assembled contigs.

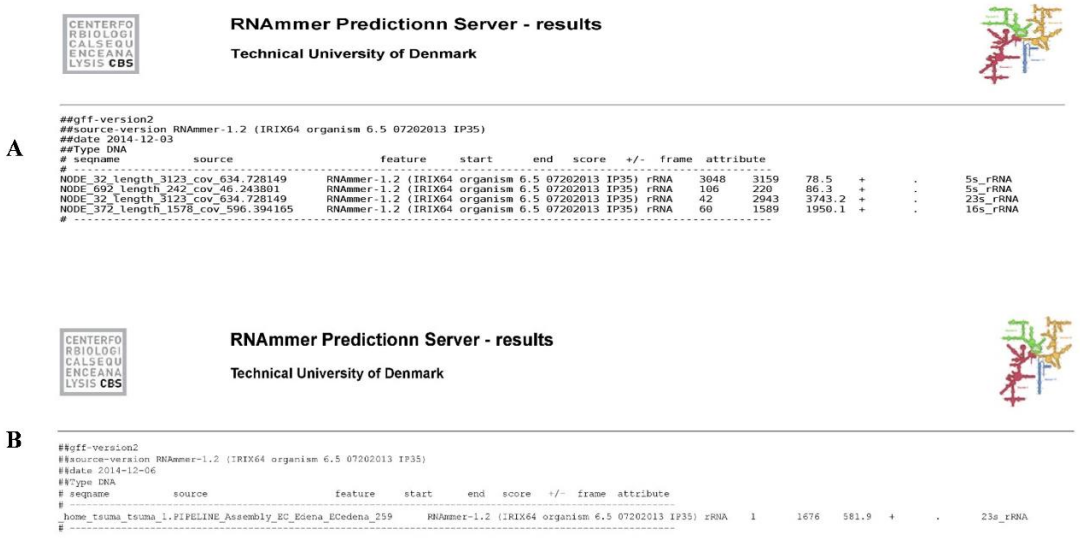


Figure 3.4.12: RNAmmer 1.2 Server results for rRNA genes of ECNA114 in (A) Contig assembly by Velvet (B) Contig assembly by Edena

RNA prediction results for ECNA114 from RNAmmer prediction server showing the detection of two copies of 5S rRNA, one copy of 23S rRNA and 16S rRNA respectively in Velvet assembled contigs and one copy of 23S rRNA in Edena assembled contigs.

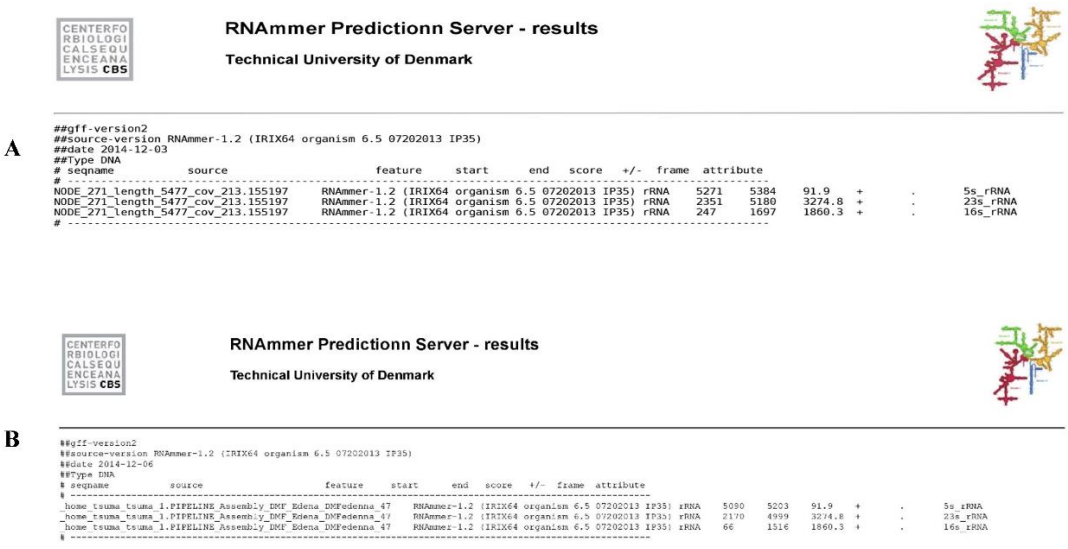


Figure 3.4.13: RNAmmer 1.2 Server results for rRNA genes of PDSD1 in (A) Contig assembly by Velvet (B) Contig assembly by Edena

RNA prediction results for PDSD1 from RNAmmer prediction server showing the detection of one copy of 5S rRNA, 23S rRNA, 16S rRNA each respectively in Velvet assembled contigs and one copy of 5S rRNA, 23S rRNA, 16S rRNA each respectively in Edena assembled contigs.

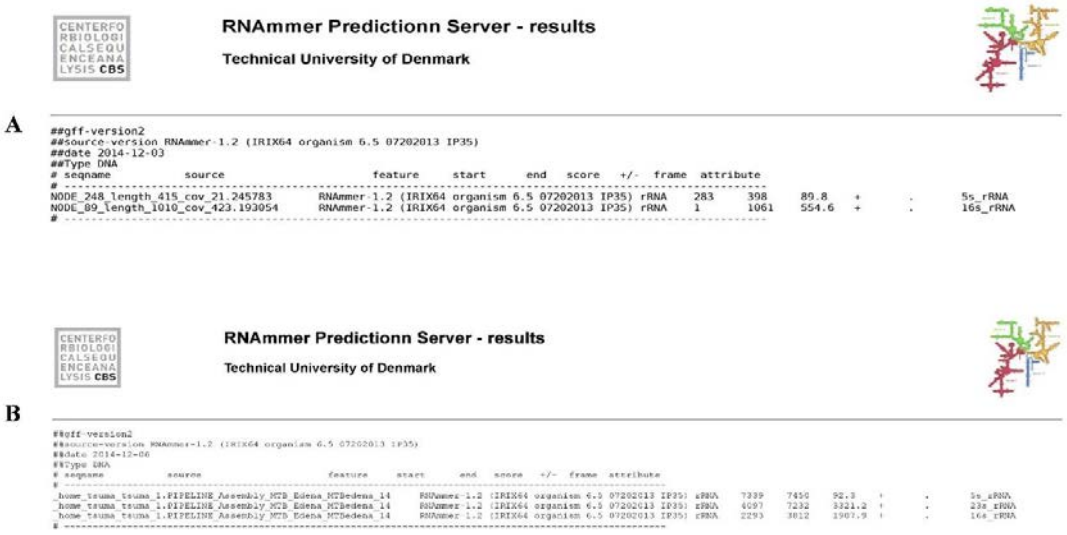


Figure 3.4.14: RNAmmer 1.2 Server results for rRNA genes of MTBNA-A0008 in (A) Contig assembly by Velvet (B) Contig assembly by Edena

RNA prediction results for MTBNA-A0008 from RNAmmer prediction server showing the detection of one copy of 5S rRNA and 16S rRNA respectively in Velvet assembled contigs and one copy of 5S rRNA, 23S rRNA, 16S rRNA each respectively in Edena assembled contigs.

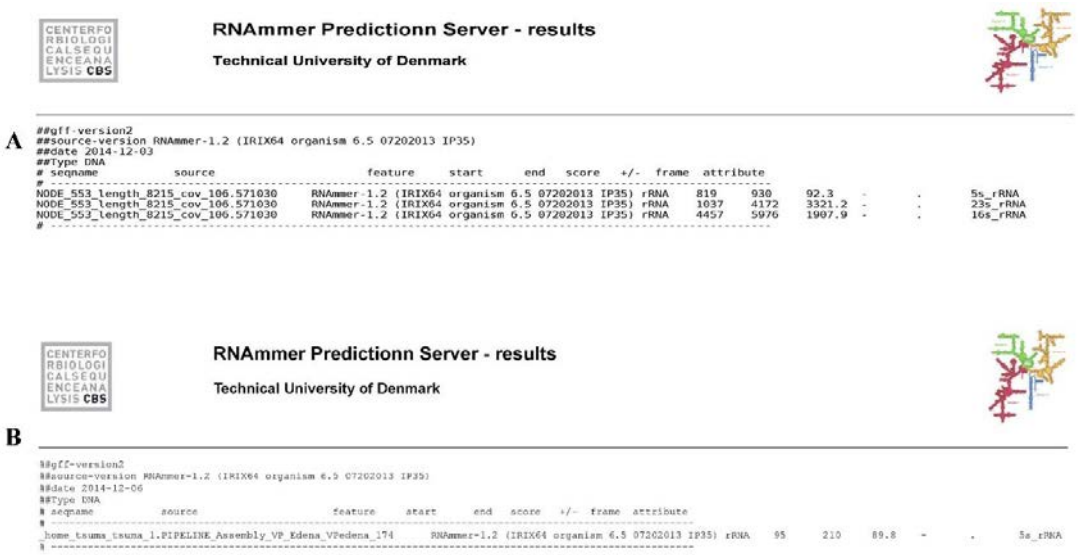


Figure 3.4.15: RNAmmer 1.2 Server results for rRNA genes of VPPCV08-7 in (A) Contig assembly by Velvet (B) Contig assembly by Edena

RNA prediction results for VPPCV08-7 from RNAmmer prediction server showing the detection of one copy of 5S rRNA, 23S rRNA, 16S rRNA respectively in Velvet assembled contigs and one copy of 5S rRNA in Edena assembled contigs.

While the Velvet contigs showed 3, 4, 3, 2, 3 number of rRNA genes in HP908, ECNA114, PDSD1, MTBNA-A0008 and VPPCV08-7 respectively (Figures 3.4.11A, 3.4.12A, 3.4.13A, 3.4.14A and 3.4.15A). The Edena contig information showed 3, 1, 3, 3, 1 rRNA genes in HP908, ECNA114, PDSD1, MTBNA-A0008 and VPPCV08-7 (Figures 3.4.11B, 3.4.12B, 3.4.13B, 3.4.14B and 3.4.15B). The GC content of these genomes were as follows HP908 – 39.39 %, ECNA114 – 51.15%, PDSD1 – 65.24%, MTBNA-A0008 – 64.13% and VPPCV08-7 – 45.3% respectively. While some studies report the high GC bias in sequencing could probably effect the read information lacking the lower GC regions, some explained that the GC bias and assembly obtained could depend on the organism (Y. C. Chen, Liu, Yu, Chiang, & Hwang, 2013). The above information shows that not only the genomes PDSD1 and MTBNA-A0008 with higher GC content have been detected of their 5S, 16S and 23S when compared among the dual assembly contigs but, the genomes HP908, ECNA114 and VPPCV08-7 have shown the copies of the rRNA genes too. This shows not only that a combination of

these two assemblers could solve the problem of detection of rRNA genes but, these genes could be of utmost importance in obtaining genome information. It can be explained as that when the contigs from both the assemblies were aligned (Figures 3.4.16, 3.4.17, 3.4.18, 3.4.19 and 3.4.20) by M-GCAT, there was a proper alignment of genome information in both assembled contigs.

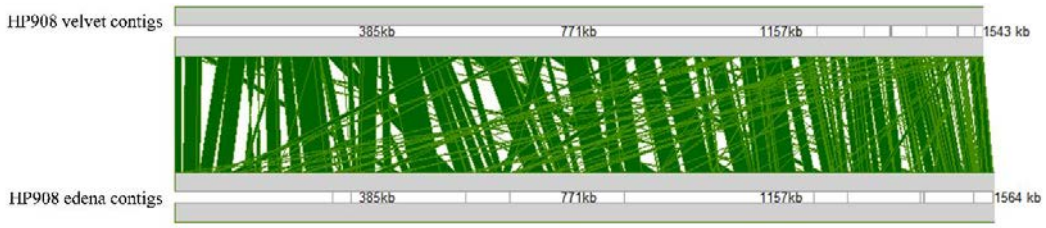


Figure 3.4.16: M-GCAT alignment showing similarity between Velvet and Edena assembled contigs of HP908

Alignment of HP908 Velvet and Edena assembled contigs show high similarity among both de-bruijn generated and overlap based assemblies with Velvet contigs amounting to a total of 1543 kb and Edena contigs to a total of 1564 kb, respectively.

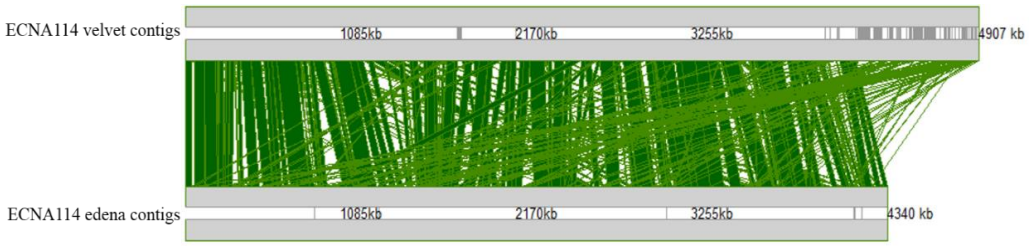


Figure 3.4.17: M-GCAT alignment showing similarity between Velvet and Edena assembled contigs of ECNA114

Alignment of ECNA114 Velvet and Edena assembled contigs show high similarity among both de-bruijn generated and overlap based assemblies with Velvet contigs amounting to a total of 4907 kb and Edena contigs to a total of 4340 kb, respectively.

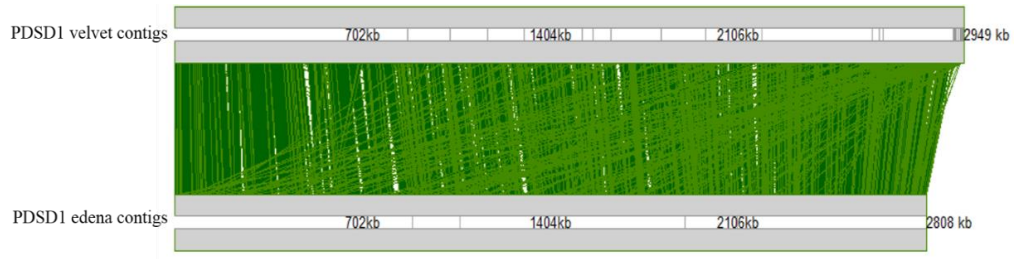


Figure 3.4.18: M-GCAT alignment showing similarity between Velvet and Edena assembled contigs of PDSD1

Alignment of PDSD1 Velvet and Edena assembled contigs show high similarity among both de-bruijn generated and overlap based assemblies with Velvet contigs amounting to a total of 2949 kb and Edena contigs to a total of 2808 kb, respectively.

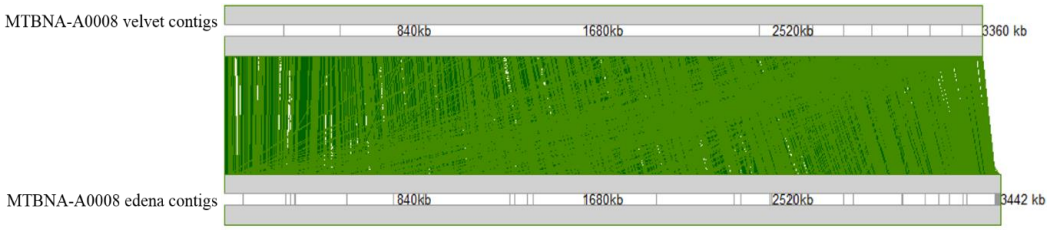


Figure 3.4.19: M-GCAT alignment showing similarity between Velvet and Edena assembled contigs of MTBNA-A0008

Alignment of MTBNA-A0008 Velvet and Edena assembled contigs show high similarity among both de-bruijn generated and overlap based assemblies with Velvet contigs amounting to a total of 3360 kb and Edena contigs to a total of 3442 kb, respectively.

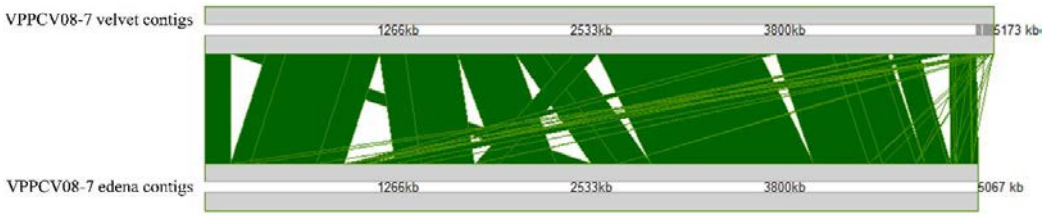


Figure 3.4.20: M-GCAT alignment showing similarity between Velvet and Edena assembled contigs of VPPCV08-7

Alignment of VPPCV08-7 Velvet and Edena assembled contigs show high similarity among both de-bruijn generated and overlap based assemblies with Velvet contigs amounting to a total of 5173 kb and Edena contigs to a total of 5067 kb, respectively.

The interesting factor could be that utilizing the rRNA gene information we could further perform Inverse polymerase chain reaction (Inverse PCR) for each rRNA gene as in 5S, 16S and 23S rRNA. As these exist in multiple copies throughout the bacterial genome it could help us obtain the sequence information of probable flanking regions and check the overlapping sequences among the contigs to incorporate the correct sequence among the assembly.

3.4.2 Quality control protocol

Trimming has been one of the best method used on NGS data at phred score 30 as reported in an earlier study (Del Fabbro et al., 2013). We have used a range of QC tools to check which strategy could give the best possible data in terms of maximum percentage of usage of raw read data. This had not been reported yet in recent years. The data we generated from all five different bacteria datasets with the QC tools (Figure 3.4.21) show that trimming was one QC strategy which retains 100% of raw reads and could be maintained as the first step of QC.

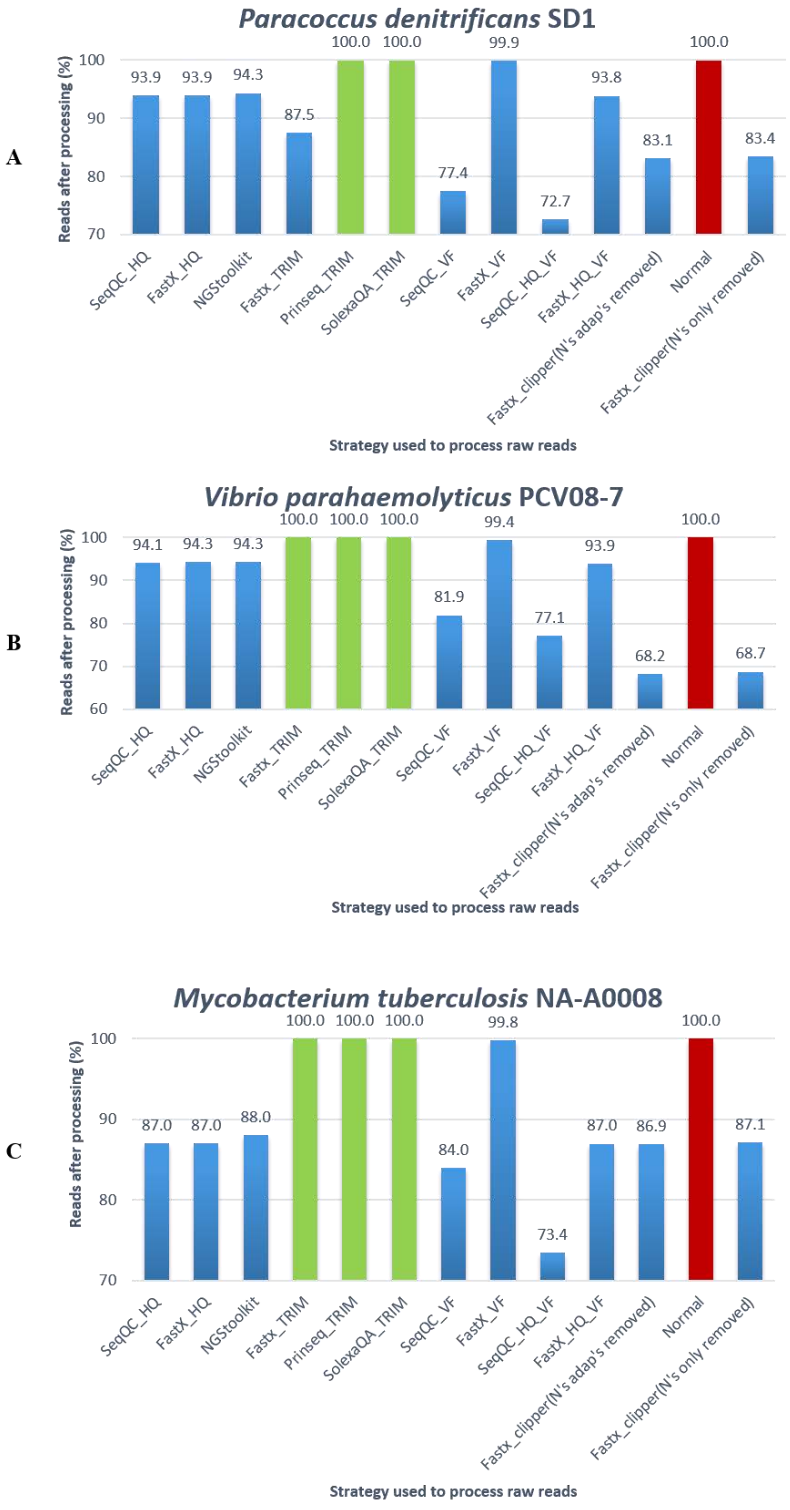


Figure 3.4.21: Graphs representing effect of QC strategies on percentage of reads produced in (A) PDSD1, (B) VPPCV08-7 and (C) MTBNA-A0008 data (with phred quality < 30)

The above graph explains the effect of trimming as the first QC process which has no drastic change in retaining the total number of reads for further QC processing. This can be observed as percentage of reads retained by usage of FASTX_trim, PRINSEQ_trim, Solexa_trim as 92.5%, 100%, 100%, respectively in PDSD1, 100%, 100%, 100%, respectively in VPPCV08-7 and 100%, 100%, 100%, respectively in MTBNA-A0008.

The genomes PDSD1, MTBNA-A0008 and VPPCV08-7 have lower phred scores (Figures 3.4.3, 3.4.4 and 3.4.5) for which trimming had to be run (Figure 3.4.21) while, HP908 and ECNA114 have phred scores greater than 30 for which no trimming has been represented on the graph (Figure 3.4.22) as by default it showed the usage of 100% of raw reads.

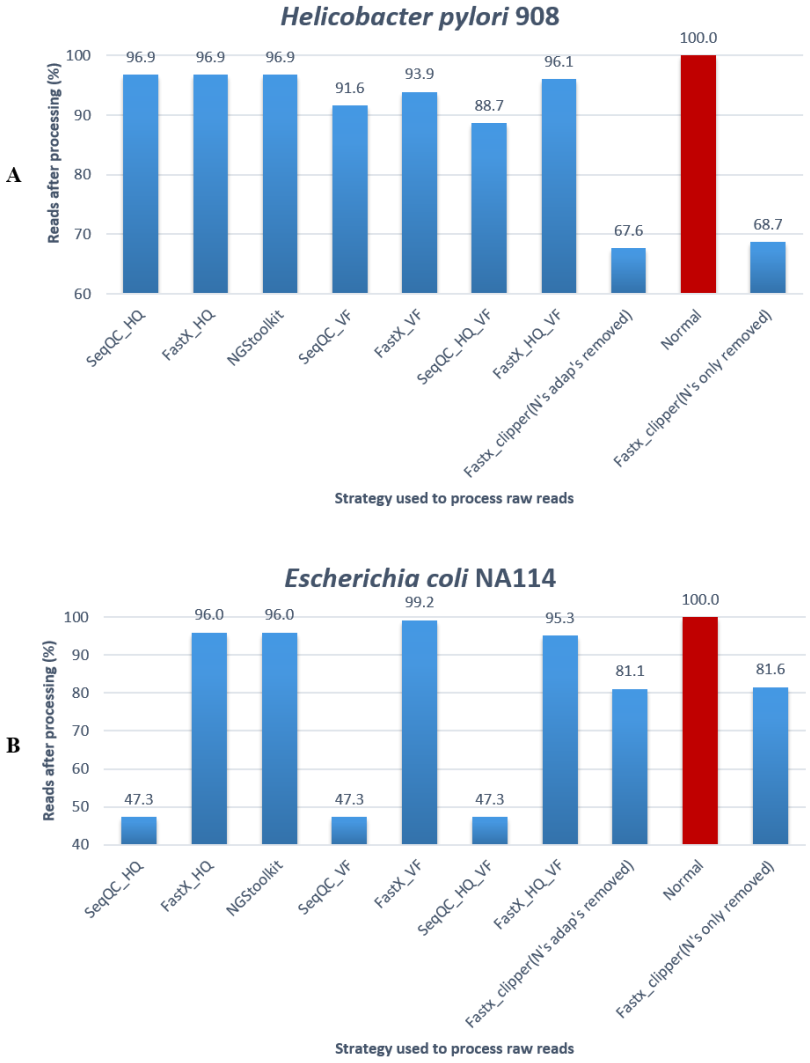


Figure 3.4.22: Graphs representing effect of QC strategies on percentage of reads produced with (A) HP908, (B) ECNA114 data (with phred score > 30)

The above graph explains the effect of using raw read sequencing data with high phred quality scores which do not require trimming in retaining the total number of reads for further QC processing. This can be observed as percentage of reads retained by usage of raw reads as 100% in HP908 and 100% in ECNA114.

The best QC strategy among the four (Figure 3.3.1) High quality, Trimming, Adaptor filtering and N's removal seems to be trimming as it shows a good MAX length of

contig, good percentage of reads used as well as fewer contigs as possible (which makes it easier to handle genome information) (Figures 3.4.23, 3.4.24, 3.4.25, 3.4.26 and 3.4.27) (Appendix A (b)).

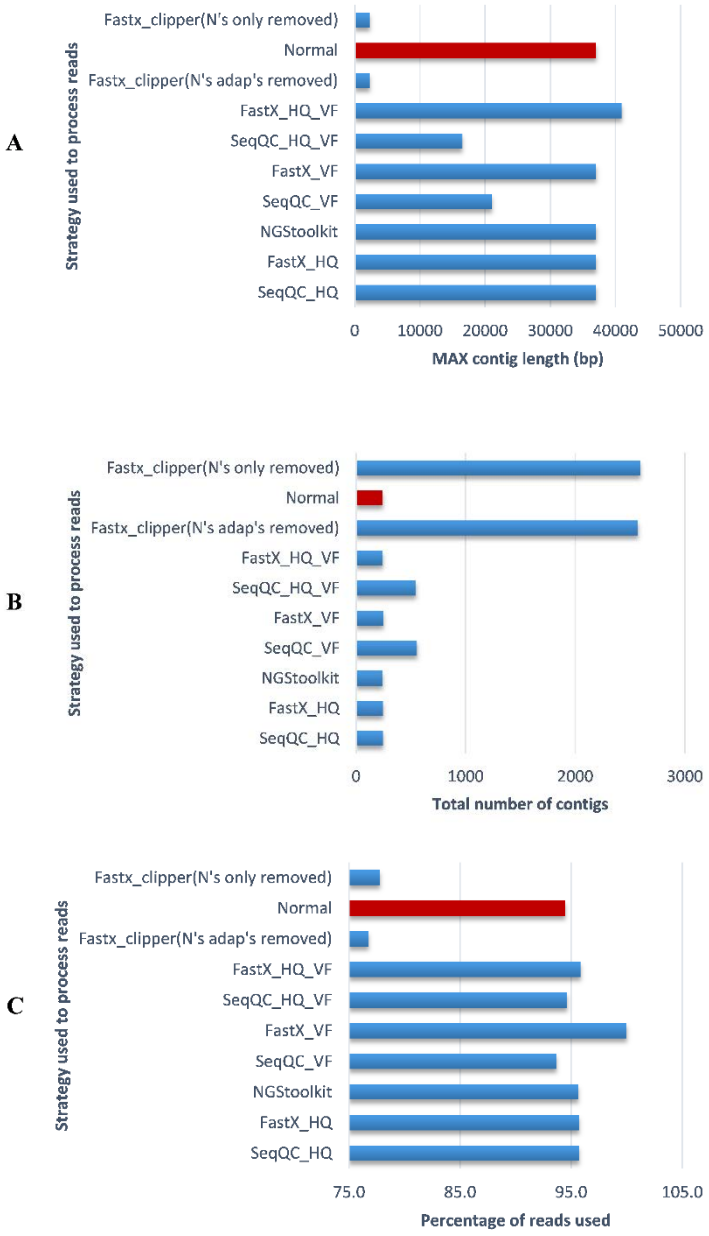


Figure 3.4.23: Result of QC strategies on (A) MAX contig length, (B) Total number of contigs and (C) Percentage of reads used in HP908 raw read data

In HP908 by QC processing utilizing various QC tools it has been observed that due to high phred quality scores of HP908 raw reads the process of assembly directly from the raw reads, “Normal” shows comparatively good MAX contig length, lower number of contigs and good percentage of reads which is considered precise in assembly of bacterial genomes.

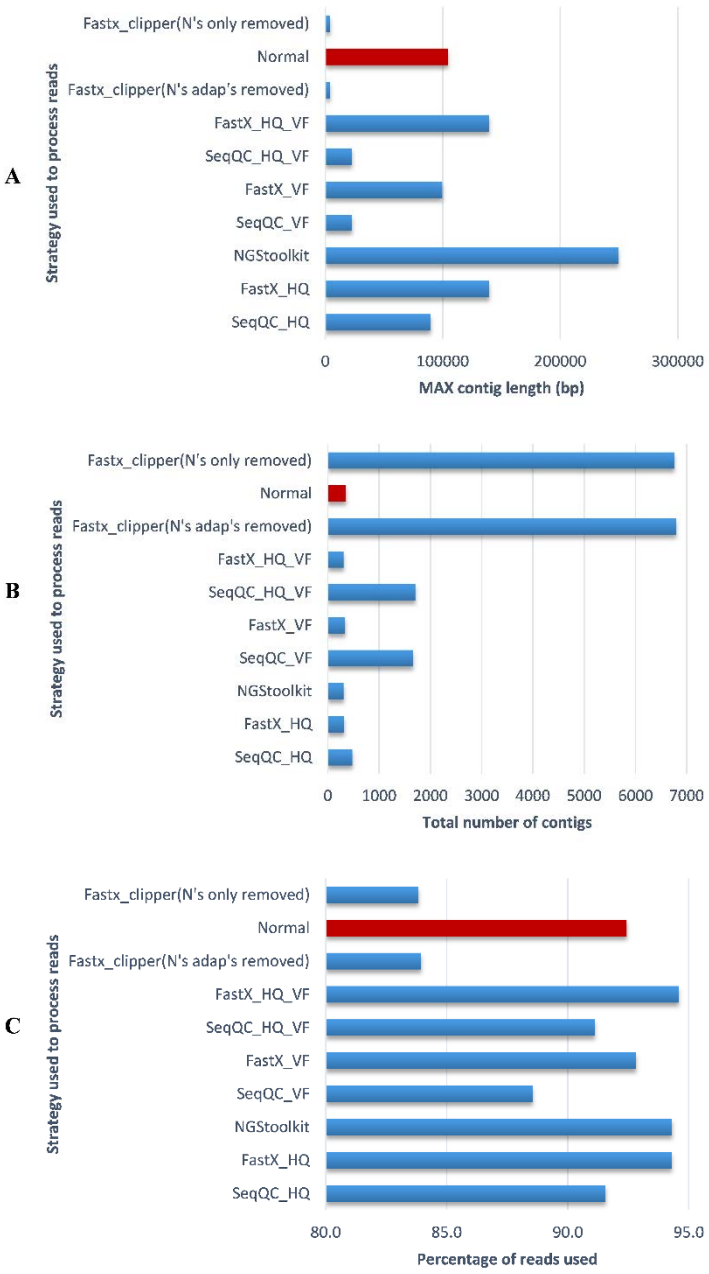


Figure 3.4.24: Result of QC strategies on (A) MAX contig length, (B) Total number of contigs and (C) Percentage of reads used in ECNA114 raw read data

In ECNA114 by QC processing utilizing various QC tools it has been observed that due to high phred quality scores of ECNA114 raw reads the process of assembly directly from the raw reads, "Normal" shows comparatively good MAX contig length, lower number of contigs and good percentage of reads which is considered precise in assembly of bacterial genomes.

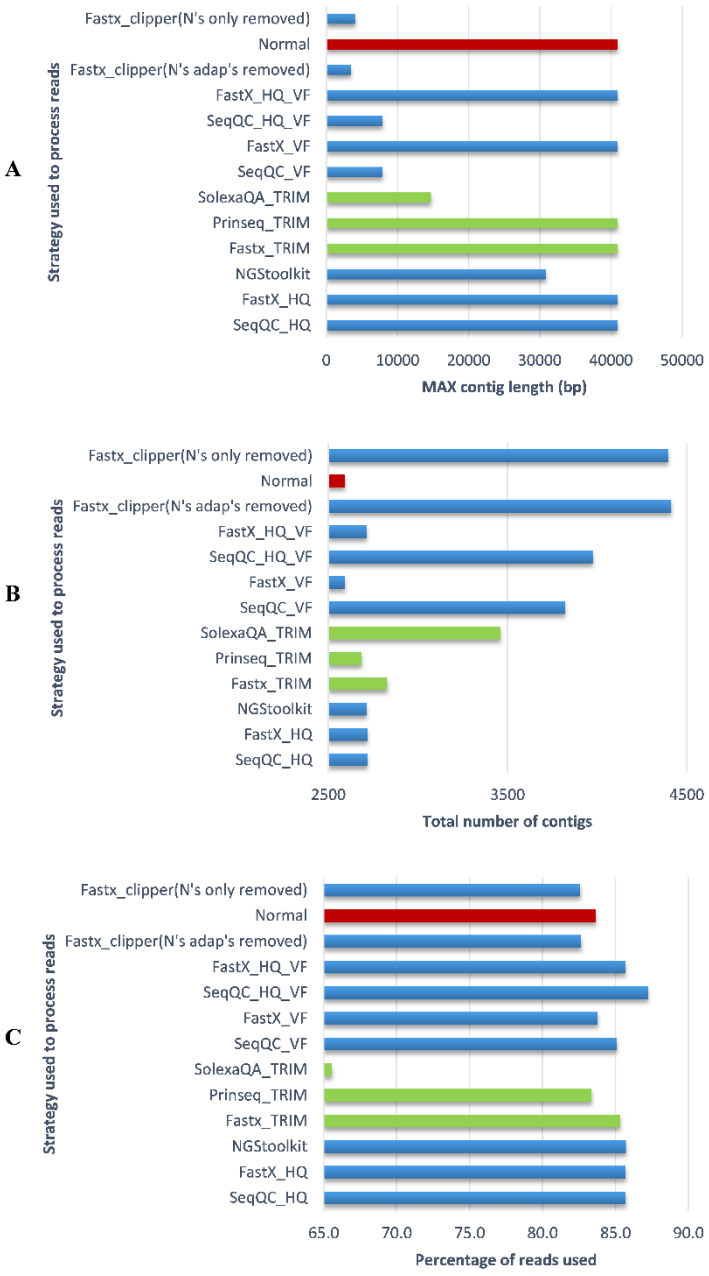


Figure 3.4.25: Result of QC strategies on (A) MAX contig length, (B) Total number of contigs and (C) Percentage of reads used in PDSD1 raw read data

In PDSD1 by QC processing utilizing various QC tools it has been observed that the process of assembly from "Trimming" shows comparatively good MAX contig length, lower number of contigs and good percentage of reads which is considered precise in assembly of bacterial genomes.

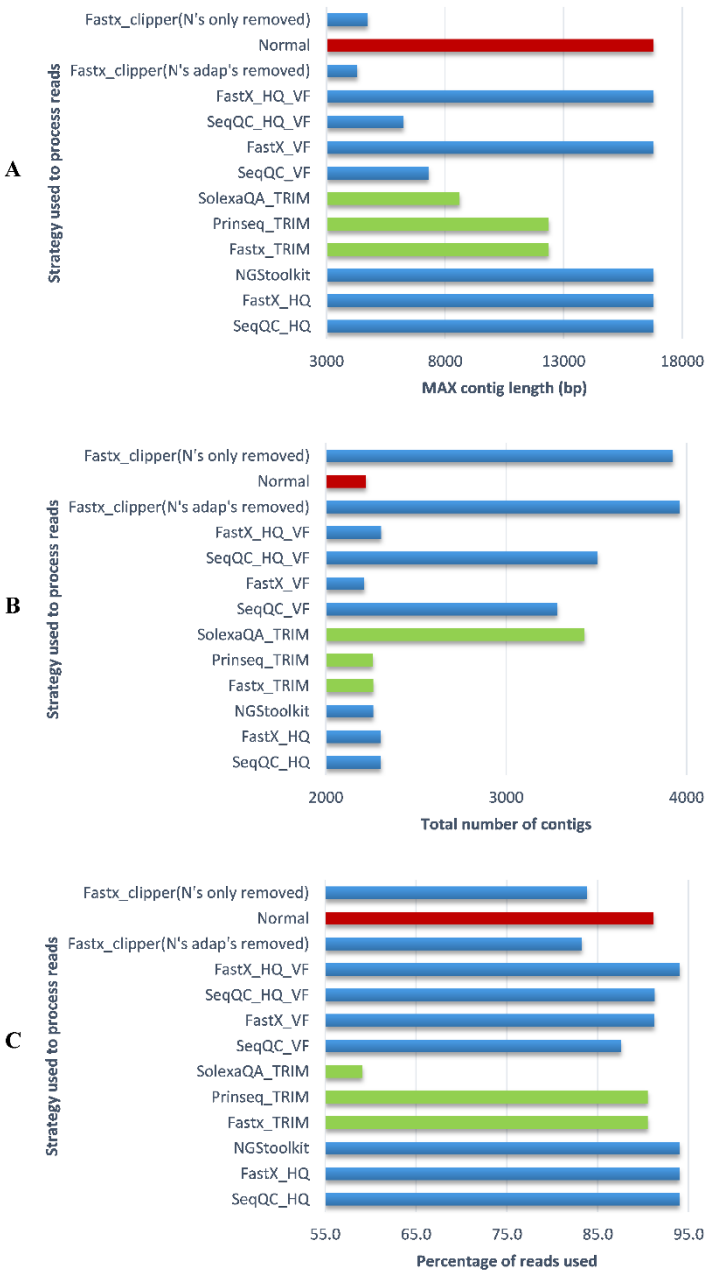


Figure 3.4.26: Result of QC strategies on (A) MAX contig length, (B) Total number of contigs and (C) Percentage of reads used in MTBNA-A0008 raw read data

In MTBNA-A0008 by QC processing utilizing various QC tools it has been observed that the process of assembly from "Trimming" shows comparatively good MAX contig length, lower number of contigs and good percentage of reads which is considered precise is considered good in assembly of bacterial genomes.

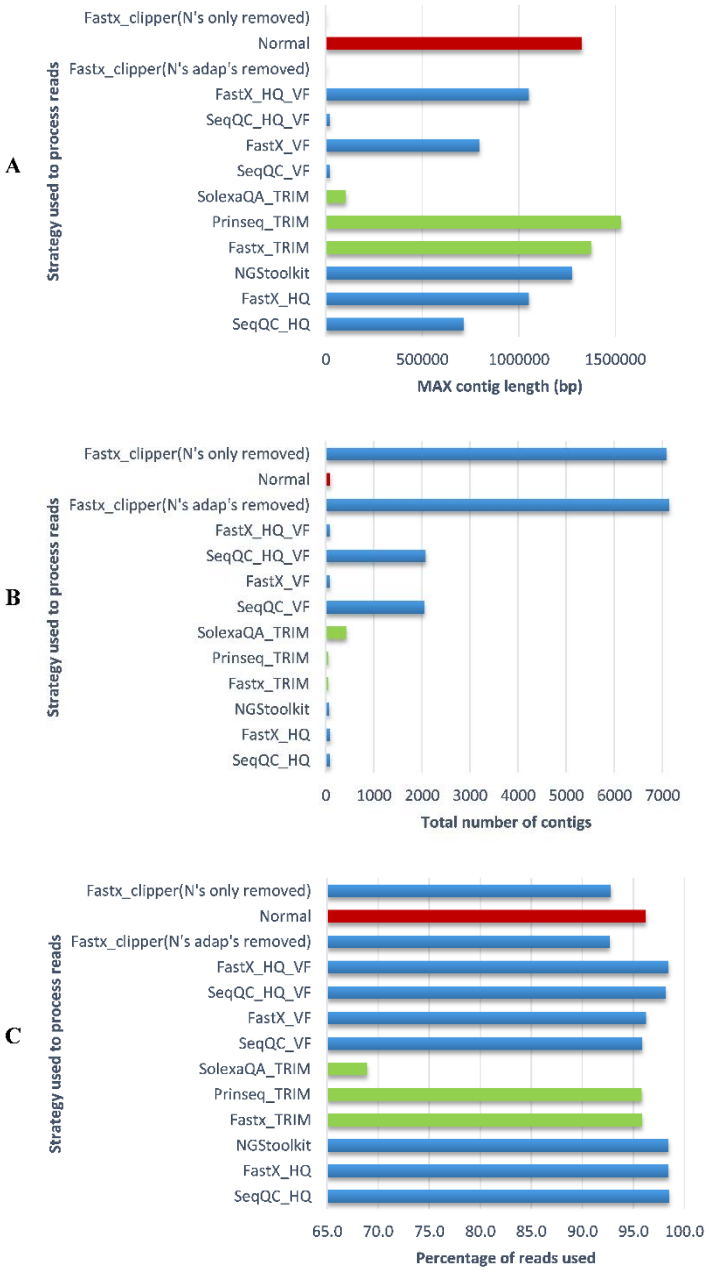


Figure 3.4.27: Result of QC strategies on (A) MAX contig length, (B) Total number of contigs and (C) Percentage of reads used in VPPCV08-7 raw read data

In VPPCV08-7 by QC processing utilizing various QC tools it has been observed that the process of assembly from "Trimming" shows comparatively good MAX contig length, lower number of contigs and good percentage of reads which is considered precise in assembly of bacterial genomes.

A number of pipelines were reported previously which dealt with NGS information in order to QC as well as assemble (Dai et al., 2010; Tritt, Eisen, Facciotti, & Darling, 2012). Here we have dealt a newer approach to QC of NGS data that is, we

have constructed seven different pipelines with difference in order of strategies to select the one which utilizes the best percentage of read data in all bacterial genomes in common. We have observed and named one TAH pipeline [Trimming by NGStoolkit, Adaptor filtering (with FLEXBAR and customized adap.txt file (Appendix A (a)) and High quality filtering by FASTX] which has shown to utilize the best percentage of reads among all bacterial genomes (Figure 3.4.28).

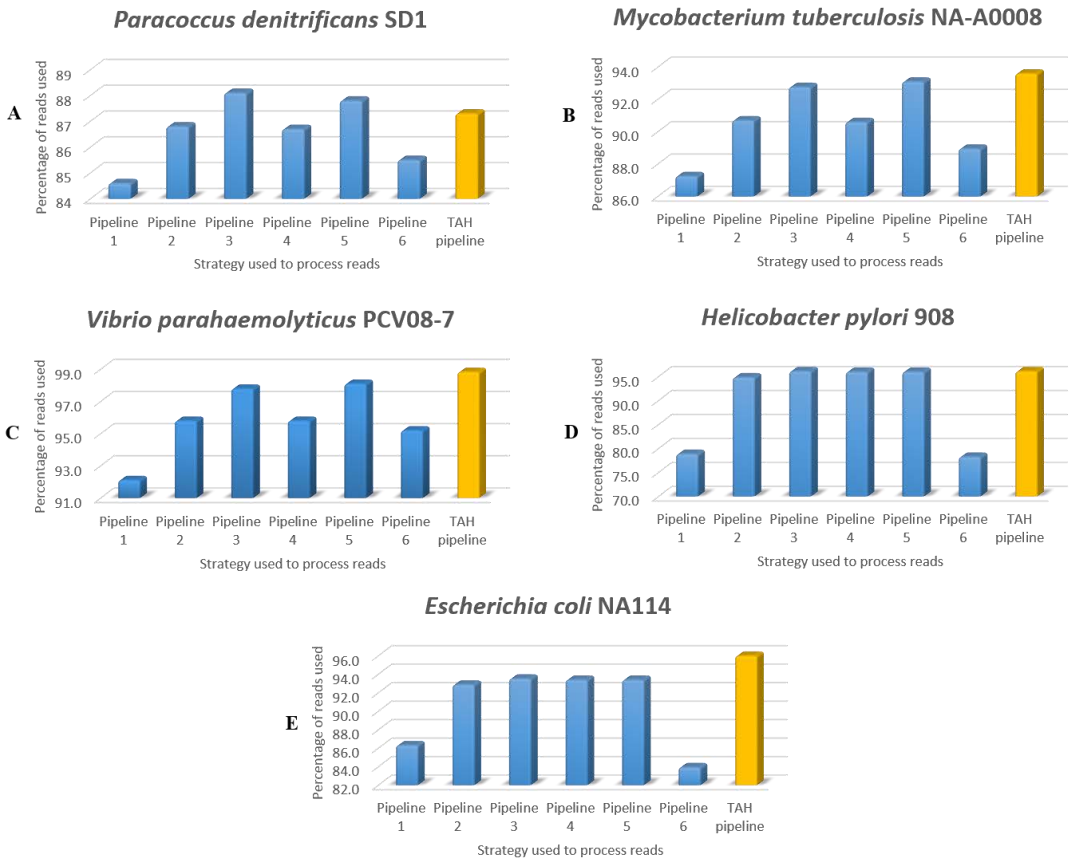


Figure 3.4.28: Comparison of the percentage of reads used by the customized pipelines in PDSD1, MTBNA-A0008, VPPCV08-7, HP908 and ECNA114 (P1-P6, TAH pipeline)

The graphs represent the output of the 7 pipelines in the PDSD1, MTBNA-A0008, VPPCV08-7, HP908 and ECNA114 bacterial genomes in terms of usage of total percentage of reads. The yellow highlighted bar represents the Trim – Adaptor – High Quality pipeline (TAH) based on the usage of reads which has a uniformly occurring higher value among the 6 pipelines in all the graphs from A to E corresponding to PDSD1, MTBNA-A0008, VPPCV08-7, HP908 and ECNA114 respectively.

Meanwhile, the HP908 and ECNA114 were omitted from trimming as they already had higher phred quality score that was above 30. Figure 3.4.29 shows the patterns of output

data for are MAX length of contig, total number of contigs and percentage of reads used for all the bacterial genomes where the TAH pipeline shows comparatively good statistics of data output.

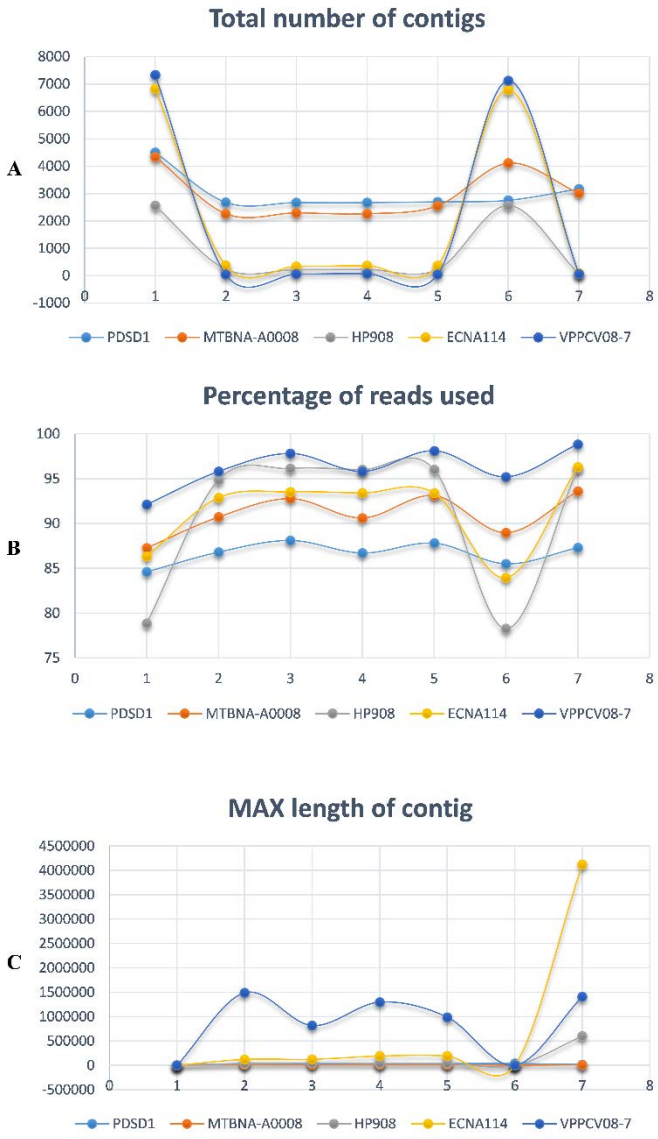


Figure 3.4.29: Graphical representation of patterns of output data (total number of contigs, percentage of reads used and MAX length of contig) across pipelines (1, 2, 3, 4, 5, 6, 7 (TAH)) corresponding to PDSD1, MTBNAA-0008, HP908, ECNA114 and VPPCV08-7 raw read data, respectively

The patterns represented under (A) Total number of contigs show a comparison of the effect of the TAH pipeline "P7" exhibiting uniformly less number of contigs being produced after the assembly. (B) Percentage of reads used shows a comparison of effect of TAH pipeline to the other pipelines exhibiting uniformly good usage of reads (C) MAX length of contig shows significantly good max length of contig among all the genomes with the application of TAH pipeline.

3.5 Conclusion

For each and every base produced through NGS the quality factor is of great significance as any error in any base could alter the genome information completely. While concentrating on the bacterial genome as a whole, considering each and every gene would create a huge amount of data for analysis and completion of a bacterial genome. Targeting in particular the basic machinery such as the rRNA genes by performing Inverse PCR could help fix the completion of a bacterial genome. This could be furthered to tRNA too as there would be multiple copies. Hence, the presence of multiple copies of these RNA genes could help simplify the experimental time and could help set up Inverse PCR more efficiently.

Trimming appears to be best as the first step of QC as it does not allow us to lose any information in terms of raw reads. An adaptor file modified according to the necessity of the NGS platforms could help resolve at least few problems concerned with primer and adaptor filtering. The TAH pipeline has been the best compared to all the pipelines and in common at the level of working as a QC step among all the bacterial data.

Our work could help researchers understand that NGS raw read data need to be dealt carefully with their quality scores which probably could be one of the reasons behind altered genome information and could even further lead to erroneous outputs. Such data could cause problems at the level of assembling a drug resistant, virulent or a highly pathogenic bacteria as per the case in the current study. To finish a bacterial genome, studies showed PCR filling between contigs (Chibana et al., 2005), but the most basic genes of importance that are rRNA could possibly be targeted for finishing a bacterial genome followed by the tRNA genes too.

While trimming reads has been applied in most NGS reads to obtain best data for further genome analysis. Our study shows that trimming could be applied as the first

step of QC followed by adaptor filtering and high quality filtering. The TAH pipeline has been formed using the pre-existing tools which make it even more unique as using the best quality of a tool could provide us with accurate data.

These details could help researchers not only to understand the importance of dealing with NGS data but, the problems which could be encountered with even one base change due to a low quality base. The actual genome in the bacterial cell can never be viewed microscopically hence, caution in dealing NGS data is indeed very important with drastic rise in data generation each hour.

Achievement

The four quality control strategies selected as follows: N's removal, high quality filtering, adaptor filtering and trimming were successfully applied on the selected bacterial genome raw read data. The various QC tools were employed and a standard quality control TAH pipeline was successfully established from the data generated.

CHAPTER 4

GENOME ANATOMY OF THE GASTROINTESTINAL PATHOGEN, VIBRIO PARAHAEMOLYTICUS OF CRUSTACEAN ORIGIN

4.1 Introduction

The assembly of draft genomes of bacteria had been of utmost importance with the increase in next-generation sequencing data per hour. While working with draft bacterial genomes has been a challenge, dealing with bacteria consisting of two chromosomes has been a tedious task to accomplish. Every bacterial genome has its role to play in the environment among which *Vibrio parahaemolyticus* has been known to cause infection in aquatic organisms. The need to glean into the statistics of the bacterial genome of aquatic origin is of utmost importance due to the increase in seafood consumption countries. Previously studied *V. parahaemolyticus* genomes were studied of their virulence factors and their possible pathogenic roles in zoonotic infection spreading through undercooked or raw seafood to humans. Annotation for only one genome (*Vibrio parahaemolyticus* RIMD2210633) was available on Genbank during the beginning of our study which has been selected as our reference genome for our study. There were cases reported of pathogenic as well as non-pathogenic strains worldwide in North America, South East Asia, Japan and in some parts of East Asia. These genomes consisted of pathogenic islands which differ among serovar and from genome to genome. The presence, absence and co-existence of two genes TDH (thermostable direct haemolysin) and TRH (TDH - related haemolysin) have also been considered to be a reference for the evolution of *V. parahaemolyticus*. Herein we deal with one *Vibrio parahaemolyticus* PCV08-7[1] strain isolated from *Macrobrachium rosenbergii*, a freshwater prawn in Malaysia whose whole genome study could give us

[1] This chapter was published as (Tiruvayipati et al., 2013) in Gut Pathogens.

interesting information about whether this bacteria causes infection or survives within the prawn.

4.2 Review of the literature

V. parahaemolyticus inhabits the estuarine, marine and brackish water ecosystems. It is an important human pathogen associated with gastroenteritis linked to contaminated seafood consumption. Since this species is abundant in marine products, it has become a significant concern in the production and trade of seafood worldwide (DePaola et al., 2003). In Southeast Asian countries, including Malaysia, virulent *V. parahaemolyticus* in raw seafood have been reported (Paydar, Teh, & Thong, 2013; Sujeewa, Norrakiah, & Laina, 2009). Numerous cases of *V. parahaemolyticus* infection were reported in North America, South East Asia and Japan including some places in East Asia (Bag et al., 1999; "Guidelines for national human immunodeficiency virus case surveillance, including monitoring for human immunodeficiency virus infection and acquired immunodeficiency syndrome. Centers for Disease Control and Prevention," 1999; H. Han et al., 2008; Nair & Hormazabal, 2005; Nair et al., 2007; Nishibuchi & Kaper, 1995; Okuda et al., 1997) giving the illness a pandemic status affecting thousands of people. Thus, the prevalence of pathogenic Vibrios in seafood is of public health concern and is an open ended issue. The pathogenic *V. parahaemolyticus* strains are differentiated from non-pathogenic ones by their ability to cause β-hemolysis on Wagatsuma agar, an activity known as 'Kanagawa phenomenon'. This effect is mediated by the activity of thermostable direct hemolysin (TDH) encoded by the tdh genes (Nishibuchi & Kaper, 1995). A pandemic clone of *V. parahaemolyticus* can broadly be defined as the one that is positive for TDH and exhibits the Kanagawa phenomenon (H. Han et al., 2008). *V. parahaemolyticus* strains are classified based on the types and variants of their O antigen and flagellar antigen (K). There are 13 O-serogroups and 71 K antigens and various combinations of these give rise to a wide variety of serovars which have been recognized as the causative agents of the disease. A clone of serovar O3:K6 has recently emerged and was associated with outbreaks in

India and Japan (Nair et al., 2007). Frequent recombination events that promote clonal diversification suggest a scenario whereby a subset of O3:K6 strains might continue to evolve (Gonzalez-Escalona et al., 2008). Consequently, different groups of related O3:K6 clonal strains have now been globally disseminated in Asia, North and South America, Africa and Europe (Nair et al., 2007). The genomes of *V. parahaemolyticus* strains are said to have undergone a number of recombination events that could have been the reason for serotype conversion from O3:K6 to O4:K68 (Y. Chen et al., 2011). Regions of recombination likely involve a genetic element larger than the gene clusters encoding O and K-antigens. More than 20 serovariants which include O3:K6, O4:K68, O1:K25, O6:K18 and O1: KUT (Chowdhury et al., 2000; Chowdhury et al., 2004) emerged from an original pandemic strain, O3:K6. The pandemic group of these bacteria has evolved through a number of deletions, substitutions and acquisitions of regions primarily corresponding to TDH or a TDH-related hemolysin (TRH). It is the presence of either of these two virulence factors that confer potential to cause gastroenteritis in human populations. The pandemic clone is said to have emerged from a pre-pandemic clone which was positive for TRH and negative for TDH genes and harboured a new sequence of *toxR* (GS-PCR). The intermediate clone is described as being negative for both TRH and TDH, but positive for GS-PCR. It has been observed that *V. parahaemolyticus* contains two chromosomes; *V. parahaemolyticus* RIMD2210633 has 3.2 Mb and 1.8 Mb of genome sizes for chromosome1 and 2 respectively (Makino et al., 2003). There are several *V. parahaemolyticus* genomes which were sequenced and deposited in Genbank as whole genomes or shotgun submissions (WGS) and sequence read archives (SRA). The only fully annotated submissions entail *V. parahaemolyticus* RIMD2210633 and *V. parahaemolyticus* BB220P. The *V. parahaemolyticus* RIMD2210633 genome harbors a Type III secretion system as a central virulence factor which is found in most diarrhea-causing bacteria

(Makino et al., 2003). As mentioned above, many studies link to the evolutionary aspects of the present pandemic clone formed from a pre- pandemic clone with a drastic change in its gene content that is, the evolution from a TDH negative (*tdh*-)/TRH positive (*trh*+) to a TDH positive (*tdh*+)/TRH negative (*trh*-) strain and the occurrence of several serovariants in the *V. parahaemolyticus* species. The present isolate (*V. parahaemolyticus* PCV08-7) has been recovered from seafood (prawn) in 2008 which were purchased from a wet market in Selangor, Malaysia. The main purpose of this study was to analyse the PCV08-7 genome that originates from Malaysia, a large peninsular as well as archipelagic country having a thriving seafood business and that it experiences several food borne outbreaks each season. Unfortunately, there are no markers based on native genome(s) to guide detection of *V. parahaemolyticus* in wet market, in the aquaculture farms and from human excreta and blood. We hope that this genome sequence will be helpful in identifying markers relevant in diagnostic development and molecular epidemiology/transmission dynamics of this significant bacterium in Malaysia and elsewhere.

4.3 Materials and Methods/ Methodology

4.3.1 Source, isolation and culture of *V. parahaemolyticus* PCV08-7

The *V. parahaemolyticus* PCV08-7 (VPPCV08-7) isolate was identified and characterized by obtaining pure cultures on selective media followed by analysis through biochemical tests, Analytical Profile Index (API) tests and genetic confirmation by polymerase chain reaction (PCR). The bacterial culture was maintained by streak plate on Thiosulfate-Citrate- Bile-Sucrose (Difco, France) agar plates. After incubation at 37°C for 21 – 24 hr, characteristic bacterial colonies appeared with blue-green colored boundaries. An isolated bacterial colony was cultured in Luria-Bertani (LB) broth with 2% Sodium Chloride (NaCl) and incubated overnight at 37°C for 16 – 18 hr. This bacterial culture was further maintained as glycerol stocks at −80°C in 20% glycerol. The genomic DNA was isolated from a pure, single colony. The bacterial identity was confirmed by sequence analysis of the 16S rRNA.

4.3.2 Genomic DNA isolation and Next-Generation Sequencing

The genomic DNA was isolated using Qiagen DNeasy Blood & Tissue kit (Qiagen, Germany) (Appendix B) and the genome sequence was determined by Illumina genome analyzer at the Genotypic Technology Pvt. Ltd. Bengaluru, India (GA2x, pipeline version 1.6). The sequencing data comprised of 100 bp paired-end reads with an insert size corresponding to approximately 240 bp. The genome coverage obtained was approximately about 80X with per base quality of reads in the range of 25 – 40. A total of 3.8 million reads were generated. Bioinformatic analysis was carried out with the help of protocols, algorithms and scripts developed previously (Avasthi & Ahmed, 2011; Avasthi, Devi, et al., 2011; Avasthi, Kumar, et al., 2011; Baddam, Kumar, et al., 2012; Baddam, Thong, et al., 2012; Kumar et al., 2012; Tiruvayipati et al., 2013)

4.3.3 Assembly and alignment

Various strategies were applied to resolve the difficulties in dealing with the two chromosomes to be assembled from the sequence reads. The following main approaches were adopted:

Velvet (Zerbino & Birney, 2008): Contigs were generated using the sequence reads which consisted of information from both the chromosomes of the isolate PCV08-7. This was checked by manually comparing contigs against the NCBI database by BLAST to check the highest similarity hit. *V. parahaemolyticus* RIMD2210633 was found to be the closest match in each search. The contigs showed unique hits to chromosome 1 (CHR1) and chromosome 2 (CHR2) as well as few common hits in both the chromosomes. The strategy of using the contigs together representing a whole genome (CHR1 and CHR2 together) or using the contigs separately as CHR1 and CHR2 was found to be challenging for further analysis to assemble them separately into two chromosomal sequences.

OSLAY (Richter, Schuster, & Huson, 2007): All the contigs were compared against both the chromosomes of the genome of RIMD2210633 individually and were then used to form supercontigs for both the chromosomes separately. This procedure was found to be problematic as the supercontig files generated from CHR1 and CHR2 (separately) revealed that the preliminary contigs mapped to sequences in both the supercontig files. This was perhaps due to the input file comprising assembled whole genome contigs used against CHR1 and CHR2. The second strategy under OSLAY was to attach CHR1 and CHR2 of the reference genome RIMD2210633 as follows: CHR1 and CHR2 were concatenated (as a 'whole genome stretch') and then further used as one full length single sequence. Using this whole genome stretch for BLAST analysis, supercontigs were generated using Velvet contigs and the BLAST results. This also eventually

proved inefficient since the supercontigs contained some sequences with several "N" representing a gap in this case and such supercontigs had to be sorted to their own positions on the genome.

SSPACE (Boetzer, Henkel, Jansen, Butler, & Pirovano, 2011): Scaffolding was performed on Velvet assembled contigs. As explained above, scaffolds were obtained separately from both CHR1 and CHR2 as well as with the whole genome stretch. All the scaffolds were then BLAST analysed against both CHR1 and CHR2 of the reference genome individually, as well as at the level of the whole genome stretch. The difficulty faced with scaffolding was similar to that of OSLAY. Hence, the option of separately identifying the scaffolds with respect to CHR1 and CHR2 and dealing with them separately remained a problem.

Mauve (Darling et al., 2010): Velvet assembled contigs were used at this step and exported as sorted contigs by performing an alignment against the whole genome stretch. The results obtained as aligned sorted contigs were taken through a stand-alone BLAST protocol against the whole genome of RIMD2210633. Then the BLAST results were carefully checked for their positions corresponding to both CHR1 and CHR2. The contigs were carefully divided as belonging to CHR1 and CHR2 sequences of PCV08-7 draft genome. The issues faced here were limited to identifying and dealing with the sequences other than those present in the contigs, but which were common to both RIMD2210633 and PCV08-7 genomes. While working on the above strategies, BWA alignment (Li & Durbin, 2010) was performed using sequence reads against the whole genome stretch of VPRIMD2210633. Using SAMTOOLS (Li et al., 2009) a .sam file was generated with which the whole genome of RIMD2210633/FASTA sequence was loaded on Tablet viewer (Milne et al., 2010) to manually inspect the presence of

common genes and to position the draft genome of PCV08-7. The sequencing reads obtained by us were primarily passed through a quality control step using FASTX toolkit (Taylor et al., 2007) to obtain high quality reads free from adaptor and primer contamination which was further standardized to an optimal parameter p value of 70. High quality reads thus obtained were assembled de-novo (Milne et al., 2010; Taylor et al., 2007) using the Velvet assembly tool which produced 83 contigs with a hash length optimized to 71. These contigs were used to run OSLAY to form supercontigs with the reference genome RIMD2210633. Alignment of the reads against the reference genome was performed using BWA. The preassembled reads were also formed into scaffolds using SSPACE. Perl scripts written in house and modified after Baddam et al. (Baddam, Kumar, et al., 2012) were used to re-order the contigs, supercontigs and scaffolds into their individual files. These approaches were put together to finalize the draft genome of *V. parahaemolyticus* PCV08-7.

The *V. parahaemolyticus* PCV08-7 whole genome shotgun project was deposited in Genbank under the accession AOCL00000000. The version described in this paper is the first version, AOCL01000000. This consists of sequences from AOCL01000000 – AOCL01000083 (http://www.ncbi.nlm.nih.gov/nuccore/AOCL00000000).

4.4 Results and Discussion

4.4.1 Genome assembly

The 100 bp paired end reads were assembled using Velvet assembly tool that effectively utilized approximately 3.7 million reads. The N50 value observed was 261989 bp. The contig with the maximum length was 704232 bp and the total number of bases in the genome were 5184164 bp. The genome was artificially closed. The genomes with multiple chromosomes pose technical difficulties during assembly. It is a known fact that Vibrios – *V. cholerae*, *V. parahaemolyticus* and *V. vulnificus* contain two circular chromosomes (Yamaichi, Iida, Park, Yamamoto, & Honda, 1999). The reference genome used in this study, *V. parahaemolyticus* RIMD2210633 also consists of two chromosomes (Chowdhury et al., 2000). As studied previously (Chowdhury et al., 2000), the origin of replication in chromosome 1 with the presence of dnaA gene shows its similarity to many genomes of prokaryotic origin and the origin of replication of chromosome 2 shows homology with that present on *V. cholerae* chromosome 2. The identification of distinct replication sites is of utmost importance for assembling bacterial genomes with two chromosomes which in the case of *V. cholerae* were studied earlier (Egan & Waldor, 2003). Previous studies explain need for a more accurate procedure to handle data to correctly assemble two chromosomes and assign gene locations. The reads were assembled into a total of 83 contigs which were separated based on the assembly strategy as explained in the materials and methods section. Dealing with the present data, we observed that many of the genes of significant virulence or fitness importance were located on the chromosomes rather than showing any significant homology to the Vibrio plasmids. The presence of the Phd-Doc toxin antitoxin gene in our genome (Figure 4.4.1) makes it interesting as the antitoxin gene has been previously reported related to plasmids (McKinley & Magnuson, 2005) while

a recent study (Guerout et al., 2013) described its occurrence on the chromosome of Vibrio species.

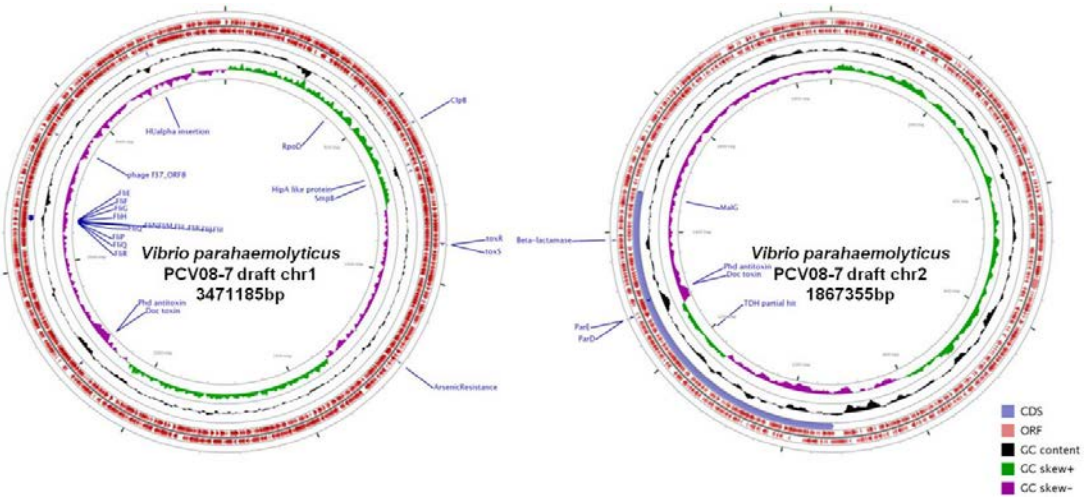


Figure 4.4.1: Circular view of *V. parahaemolyticus* PCV08-7 draft genome. Diagrammatic representation of major genes carried by the two chromosomes of *V. parahaemolyticus* PCV08-7 genome using CGview (Stothard & Wishart, 2005).

However, we agree that the exact source of these genes can be mapped only when the plasmids will be sequenced and or analysed separately.

4.4.2 Genome statistics and annotation

The draft assembled genome was annotated using the RAST server (Aziz et al., 2008). Statistics of the *V. parahaemolyticus* PCV08-7 draft genome were derived using Artemis (Rutherford et al., 2000), RNAmmer (Lagesen et al., 2007) and tRNAscanSE (Schattner, Brooks, & Lowe, 2005): the sizes of chromosome 1 and chromosome 2 of the isolate were 3471185 bp and 1867355 bp respectively with G + C content of 45.35%. The number of tRNA and rRNA genes were 102 and 31 for chromosome 1, and 13 and 3 for chromosome 2, respectively. The chromosome 1 revealed a coding percentage of 85 with an average gene length of 943 bp while the chromosome 2 had a coding percentage of 86.2 with an average gene length of 950 bp. The alignment of *V. parahaemolyticus* PCV08-7 genome with that of the *V. parahaemolyticus* RIMD2210633 genome using M-GCAT (Treangen & Messeguer, 2006) showed visible

rearrangements in the sequences of the two chromosomes of PCV08-7 isolate (Figure 4.4.2).

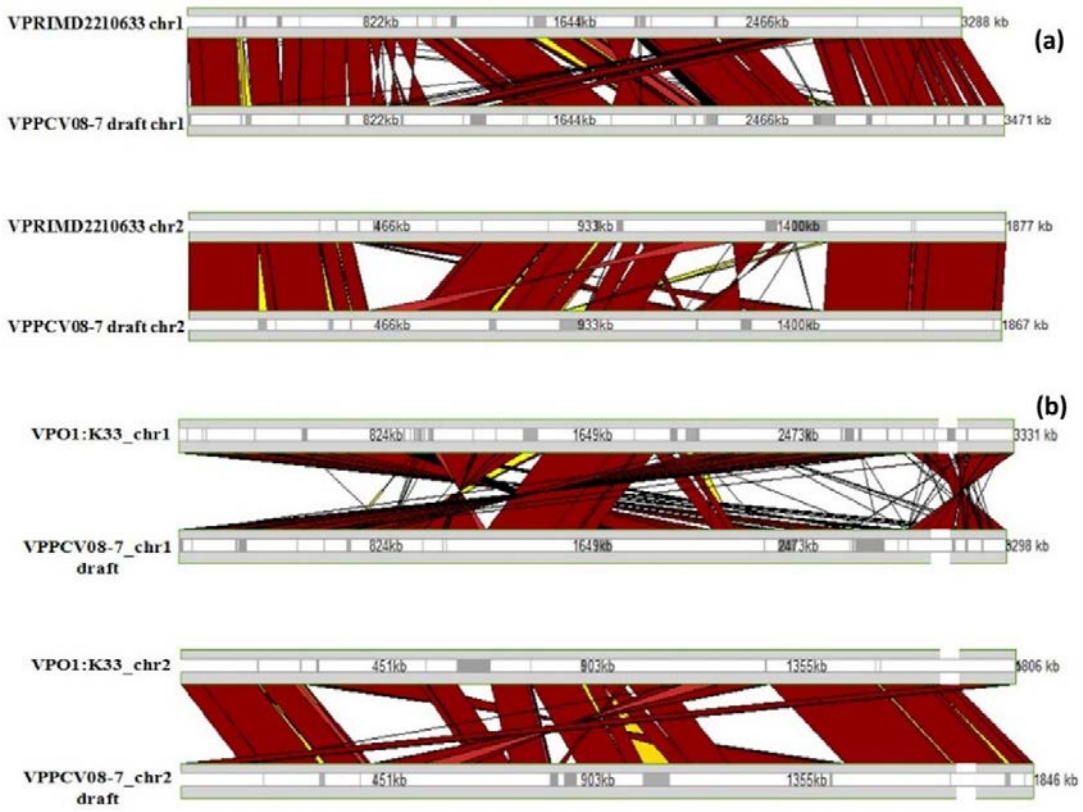


Figure 4.4.2: Alignment of the genome of strain RIMD2210633 against that of isolate PCV08-7 and strain O1:K33. **(a)** Comparison of chromosomes of strain RIMD2210633 (VPRIMD2210633 chr1, VPRIMD2210633 chr2) with the draft chromosomes of PCV08-7 (VPPCV08-7 draft chr1, VPPCV08-7 draft chr2) using M-GCAT. **(b)** Comparison of chromosomes of strain O1:K33 (VPO1:K33 chr1, VPO1:K33 chr2) with the draft chromosomes of PCV08-7 (VPPCV08-7 draft chr1, VPPCV08-7 draft chr2) using M-GCAT.

The chromosome 1 of the draft genome carried phage shock proteins A, B and C, and bacteriophage f237 ORF8. It contained an integrated *tmRNA* gene with the closest element encoding the ribonuclease H. A site-specific recombinase *IntI4* and a gene encoding beta-lactamase were present. The draft genome also revealed genes responsible for fatty acid and amino acid metabolism. An important outer membrane protein OmpU was also identified. Genes coding for gyrase B (gyrB), HU-alpha insertion and putative sigma factors such as rpoD, rpoE, rpoS, rpoN and rpoH were also found in our analysis. The chromosome 2 carried a TDH pathogenicity island with

many deletions and substitutions and displayed a *malG* gene on one of the flanking regions of the pathogenicity island. This region also contained genes coding for nutrient uptake and metabolism. We documented the presence of vibrioferrin receptor *pvuA* and ferrichrome ABC transport *pvuB*, *pvuC*, *pvuD* and *pvuE* encoding genes, and the related *pvsA*, *pvsB*, *pvsC*, *pvsD* and *pvsE* genes. The analysis of the genome further revealed presence of a cobalt-zinc-cadmium resistance protein and a Rhodanese related sulfur transferase (as also present in RIMD2210633 genome) and a lead-cadmium-zinc-mercury transporting ATPase enzyme (as seen in the *V. parahaemolyticus* BB220P genome). Phd antitoxin and Doc toxin (McKinley & Magnuson, 2005) which fall under the programmed cell death systems were also uniquely identified. Studies in *E. coli* have shown the presence of a stress related protein *clpB* (caseinolytic peptidase B homolog) along with *rpoS* and a few other genes (Wang & Wood, 2011) which help cope with stress conditions and help in survival. Our analysis detected the presence of *clpB*, *rpoS* and *hipA* genes in the present genome as was also seen in the reference genome of RIMD2210633. There were two types of Type III secretion systems observed in *V. parahaemolyticus* RIMD2210633 (Park et al., 2004); T3SS1 and T3SS2. Our genome analysis remains open ended with respect to the presence of such type III secretion systems.

4.4.3 Identification of novel gene content and comparative analysis

Our genome analysis revealed some unique sequences which have good similarity to hypothetical proteins of other Vibrio species such as *Vibrio anguillarum* and *Vibrio cholerae*. A 6315 bp nucleotide sequence showed identity to a *V. anguillarum* hypothetical protein and a *V. cholerae* hypothetical protein on NCBI-BLASTN. One of the coding proteins in this stretch revealed similarity to the annotated phage integrase encoding gene of *Photobacterium damselae subsp. damselae* plasmid pAQU1 DNA (Figure 4.4.3).

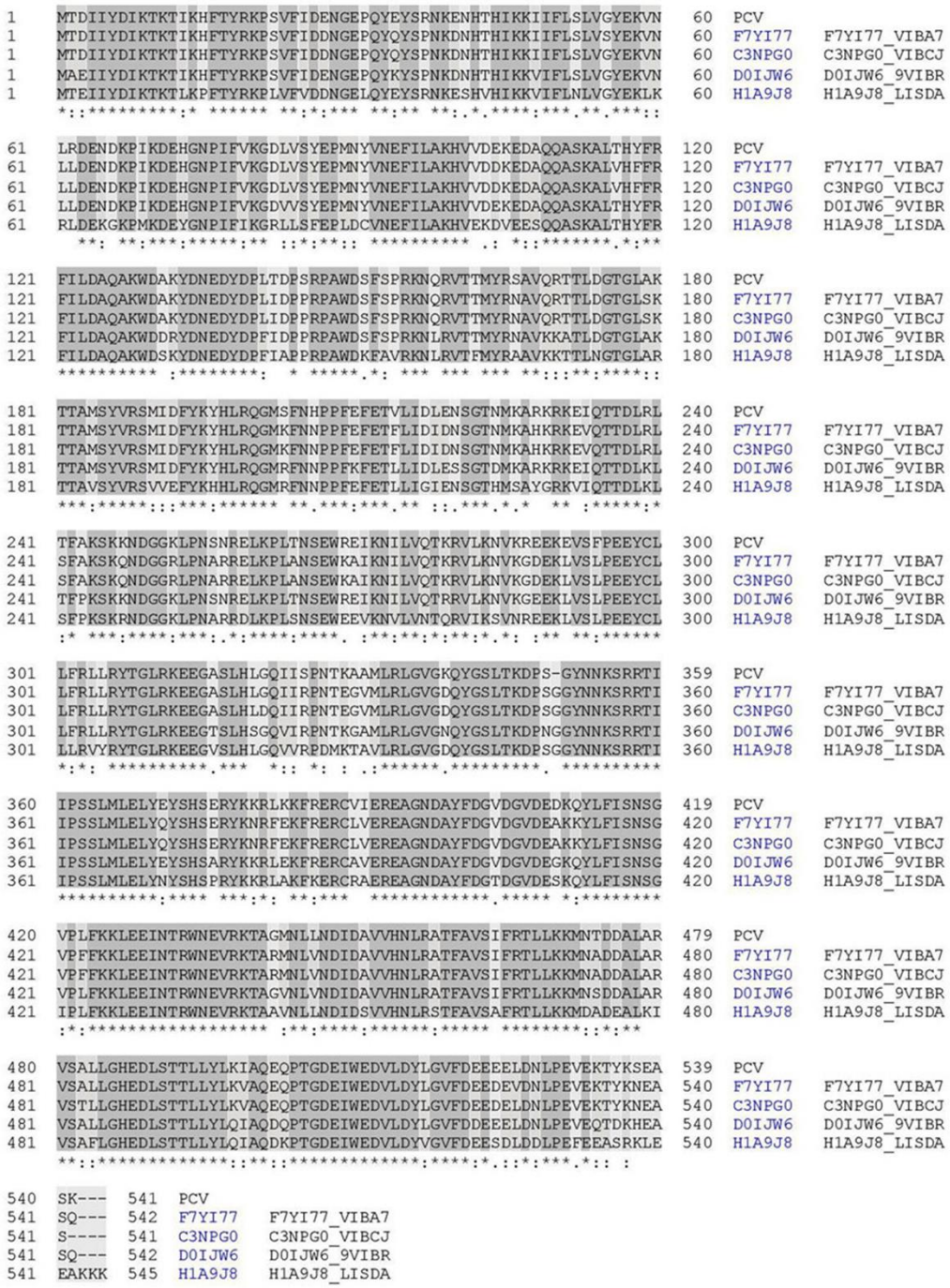


Figure 4.4.3: Alignment of a unique PCV08-7 protein sequence similar to *Photobacterium damselae subsps. damselae*. A unique sequence from PCV08-7 genome showed similarity with putative uncharacterized proteins of *V. anguillarum* 775 (F7YI77), *V. cholerae* MJ-1236 (C3NPG0) and *Vibrio sp.* RC586 (D0IJW6) and similarity to a phage integrase of *Photobacterium damselae subsps. damselae* (H1A9J8).

A *parD* gene (antitoxin to *parE*) was also found which showed closer identity to other Vibrio species such as *Vibrio vulnificus*, *Vibrio mimicus* and *Vibrio orientalis*. *parD*

when aligned against *V. vulnificus* and *V. mimicus* revealed an identity of 76 bp out of 80 bp (95%) (e-value 2e-48) and with *V. orientalis* an identity of 72 bp out of 80 bp (90%) (e-value 2e-45) on NCBI-BLASTN. A few newer hypothetical proteins with no reported annotation were identified. The genome also contained a gene relevant to arsenic resistance, possibly important in the adaptation of the bacterium to a high arsenic environment. Our analysis of the genome revealed presence of a partially similar sequence of TDH Pathogenicity Island, as compared to *V. parahaemolyticus* RIMD2210633. This island revealed genetic instability due to various insertion/ deletion and substitution events which were documented. The presence of *toxS* and *toxR* genes was also observed.

The old pandemic O3:K6 strain of *V. parahaemolyticus* is said to have gained gene clusters VPaI1-VPaI7 (Hurley, Quirke, Reen, & Boyd, 2006) to develop into a new pandemic clone of which VPaI4-VPaI6 are said to be putative virulence factors and may be potential pathogenicity islands. These regions are said to carry along with them a type VI secretion system (VP1386-1420). Our PCV08-7 genome analysis revealed that only one cluster, VPaI2, was detected completely, whereas VPaI3 and VPaI7 were partially present (Table 4.1).

Table 4.4.1: Table representing pathogenicity related clusters and other VP clusters in *V. parahaemolyticus* PCV08-7: (1) pathogenicity related clusters (VPaI1-VPaI7) in the genome of strain RIMD2210633 that signify it to be a pandemic O3:K6 strain and their presence or absence in the genome of PCV08-7 isolate, (2) various other VP clusters and their occurrence in the genome of PCV08-7.

(1) *Vibrio parahaemolyticus* RIMD2210633	***V. parahaemolyticus* PCV08-7**
VPaI1 (VP0380-0403)	Absent
VpaI2 (VP0635-0643)	Present
VpaI3 (VP1071-1094)	Partially present
VpaI4 (VP2131-2144)	Absent
VpaI5 (VP2900-2910)	Absent
VpaI6 (VPA1254-1270)	Absent
VpaI7 (VP1312-1398)	Present
Type VI secretion system (VP1386-1420)	Absent
(2) Other VP clusters	***V. parahaemolyticus* PCV08-7**
VP1355-1368	Partially present
VPA0074-0089	Present
VPA0713-0732	Present
VPA1194-1210	Present

This perhaps shows that our strain could be possibly a new serovariant of a non-pandemic O3:K6 strain like the *V. parahaemolyticus* AQ3810 (Nishibuchi & Kaper, 1995). While variability of different gene clusters (Table 4.4.1) portrays a probably novel serovariant of *V. parahaemolyticus* with the presence of ribonuclease H encoding element (previously thought to be present only in *V. parahaemolyticus* RIMD2210633 and absent in *V. parahaemolyticus* AQ3810 (Y. Chen et al., 2011). A further comparative study between the *V. parahaemolyticus* PCV08-7 and the non-pandemic *V. parahaemolyticus* AQ3810 (O3:K6 strain) and the newest *V. parahaemolyticus* O1:K33 (*trh*+/ *tdh*+ genotype) strain showed that *V. parahaemolyticus* PCV08-7 has more genetic relatedness towards a *trh*+/ *tdh*+ strain (Figure 4.4.4).

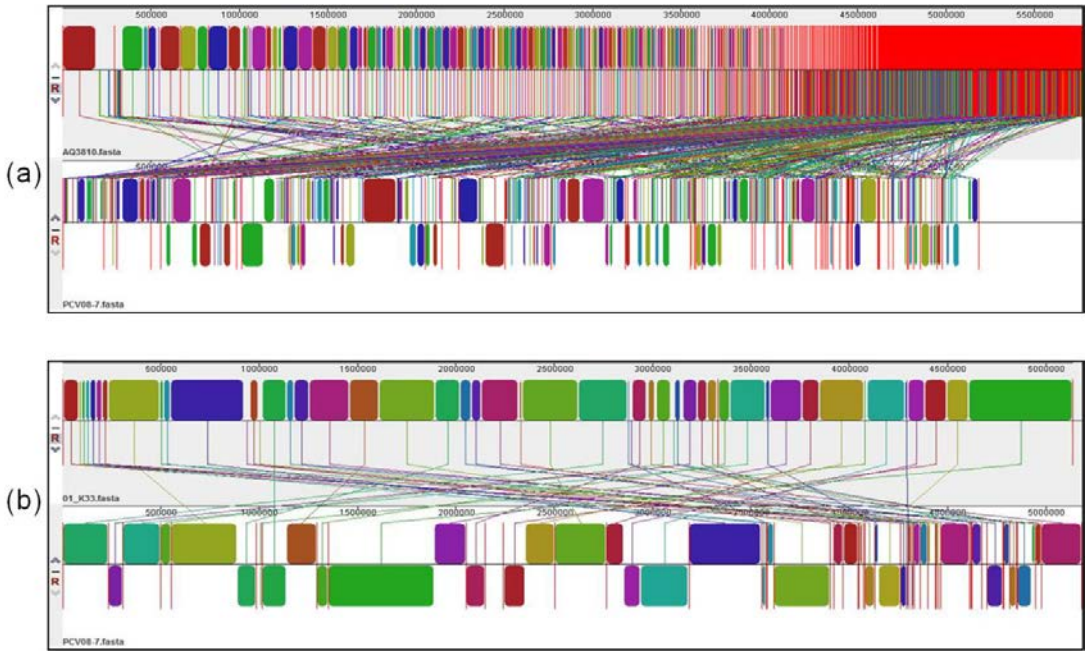


Figure 4.4.4: (a) *V. parahaemolyticus* AQ3810 alignment against PCV08-7 genome: Concatenated chromosome 1 and chromosome 2 of *V. parahaemolyticus* AQ3810 (AQ3810.fasta) against *V. parahaemolyticus* PCV08-7 (PCV08-7.fasta) **(b)** *V. parahaemolyticus* O1:K33 alignment against PCV08-7. Concatenated chromosome 1 and chromosome 2 of *V. parahaemolyticus* O1:K33 (O1_K33.fasta) against *V. parahaemolyticus* PCV08-7 (PCV08-7.fasta).

But, alignments of the *V. parahaemolyticus* PCV08-7 contig data against the *V. parahaemolyticus* O1:K33 and *V. parahaemolyticus* RIMD2210633 (Figure 4.4.2) strains show that it is closer to O3:K6 serotype (Figure 4.4.5).

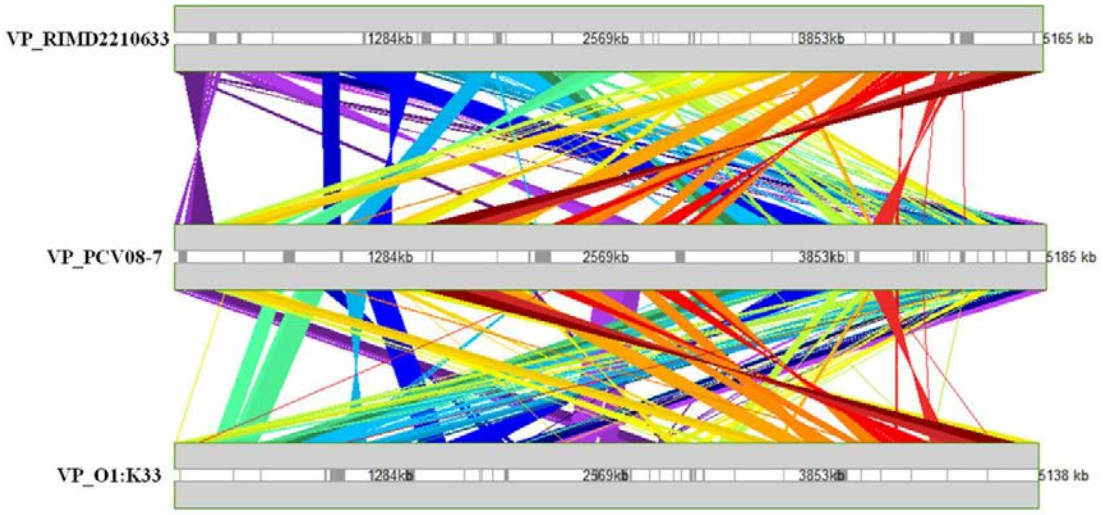


Figure 4.4.5: Comparison of whole genome sequences of strains RIMD2210633, PCV08-7 and O1:K33. Alignment of complete genomes of *V. parahaemolyticus* RIMD2210633, *V. parahaemolyticus* PCV08-7 and *V. parahaemolyticus* O1:K33, showing PCV08-7 being more similar to RIMD2210633.

From the above data, it becomes probably apparent that the genome of *V. parahaemolyticus* PCV08-7 meaningfully adds to the battery of important genomic sequences representing enteropathogenic bacteria. The genome of an arthropod derived, foodborne Vibrio should be important to understand adaptation to a crustacean host and a human host.

4.5 Conclusion

A first account of the genome of *V. parahaemolyticus* PCV08-7 has been presented. The draft genome and its annotation as described would be able to explain the lifestyle of pathogenic Vibrio species. The experience of assembling this genome and the difficulties associated with separating the data with respect to two chromosomes would certainly be helpful to the community in the follow-up studies. Further, a host of new molecular markers as gleaned by our analysis would be relevant in the diagnostic development and molecular epidemiology.

The present genome and the ensuing comparative genomics would be able to rekindle our thoughts on the survival and virulence as well as transmission potentials of *V. parahaemolyticus* and also on their adaptation to different hosts and the niches thereof. Our results clearly reveal a significantly novel gene content which could presumably have been acquired through a horizontal gene transfer mechanism. Our analysis revealed the presence of not only the conserved genomic regions among different *V. parahaemolyticus* bacteria, but also dissects some of the unique sets of genes that hold relevance to virulence. We propose to finish and polish the genome in the near future also with the help of further coverage using alternative sequencing platforms and by employing a hybrid assembly approach. Also, it will be possible to determine the true extent of the diversity of *V. parahaemolyticus* strains obtained from seafood as compared to those isolated from human cases. Such a diversity analysis would focus on 1) genomic coordinates relevant to colonization of and adaptation to different hosts in different ecosystems; 2) genome dynamics relative to bacterial fitness shaping over time and with transmission across different hosts; and 3) profile of genomic rearrangements including additive and reductive genome evolution and their significance in the evolution of pathogenic Vibrio species. Presently, the epidemiology of *V. parahaemolyticus* infection in resource-poor countries largely entails a classical

serology concocted with guess work as to the type of strain involved and its source. Our genomic data would hopefully contribute to this situation also.

Achievement

The draft genome of VPPCV08-7 was completed using the Illumina sequencing raw read data. A customized assembly strategy was employed in our work utilizing SSPACE, OSLAY, Velvet together. The draft genome was successfully gleaned of its various important genes and compared to other closely related *Vibrio parahaemolyticus* genomes for meaningful comparative data.

CHAPTER 5

VIBRIO PARAHAEMOLYTICUS PCV08-7 PHYLOGENETIC RELATIONSHIPS WITH CO-EVOLVED STRAINS AND ITS STUDY ON SYMBIOSIS WITH HOST AND ENVIRONMENT

5.1 Introduction

Draft genome of bacteria helps extract information from sequencing raw reads which further guides in annotation and genome analysis. Apart from this, the huge data generated can be used in a wide range of studies such as designing experiments for completion of a genome and to study the evolution within the species. If the bacteria is pathogenic the data could be quite useful in studying the host-pathogen interactions. Our study deals with bacteria with two chromosomes. There are numerous studies on bacteria containing multiple chromosomes and the presence of plasmids, for the integration of assembled data into individual components. The foremost issue with multiple chromosomes is the placement of genomic information into individual chromosomes in comparison to a reference genome. Herein retaining the uniqueness of a bacterial genome could be lost as we use the order of a reference and construct a synteny among the genomes. As the order would be more similar to the reference than that of the new genome, the data generated as above has chances of affecting further analysis. Any evolutionary studies conducted with data following the above aligned data could make the genomes lie more closer or distant depending upon the alignment and completion. Further, such data can even affect any host-pathogen interaction studies as researchers tend to select the data which contains sequences arranged according to the reference genome data. *V. parahaemolyticus* PCV08-7 (VPPCV08-7) as mentioned in Chapter 4 was isolated from *Macrobrachium rosenbergii* (MR), a freshwater aquatic organism. The presence of an external environment, in this case freshwater and its

constituents as the environment cannot be ruled out from host-pathogen interaction studies.

In our study, as dealt earlier in Chapter 4 we deal with a bacterial genome with two chromosomes. Our primary aim in the current chapter is to deal with separation of contig information into two chromosomes and a plasmid. To deal with the genome information at the contig level as it provides with maximum information related to virulence, pathogenicity and survival. Further, we would apply this extracted information to choose a few important genes one-by-one while tracing back their evolution. These genes would contain the house-keeping gene, 16S rRNA (Janda & Abbott, 2007) and the host-pathogen related genes corresponding to interactions with the component of the host (*M. rosenbergii*). The genes related to interactions with the host would be selected bioinformatically. This gene will be checked with Real Time Polymerase chain reaction studies for their responses in levels of gene expression through an *in vitro* stimulated experiment. The experiment would be setup based on the *gbpA* gene from *V. parahaemolyticus* PCV08-7 in relation to the host component under specific external environment conditions that is in the presence of varied magnesium concentrations. This study could give us an insight on how undoubtedly it is more advisable to use contig data for completing a bacterial genome with the discussion as mentioned in Chapter 3. Apart from this the tracing back of a number of strains of a bacteria genome with the house-keeping gene and the 16S rRNA as a reference for phylogeny, would help us identify the effects on host-pathogen interaction related genes. Real time studies of gene expression could even help us understand possible activation mechanism the environment follows to keep the host-pathogen intact. Further studies could help in identifying the core responsible factor to control any possible entry conditions of pathogen into the host.

5.2 Review of the literature

V. parahaemolyticus PCV08-7 draft genome has been reported in December 2013 (Tiruvayipati et al., 2013), as a partial Thermostable direct hemolysin pathogenicity island carrying *V. parahaemolyticus* strain. Chromosome 1 and chromosome 2 bear a size of 3471185 bp and 1867355 bp respectively in the *V. parahaemolyticus* PCV08-7 draft genome with G + C content of 45.35%. The number of tRNA and rRNA genes were 102 and 31 in chromosome 1, and 13 and 3 in chromosome 2, respectively. A number of annotation tools were developed among which Genemark (Besemer, Lomsadze, & Borodovsky, 2001), Glimmer (Delcher, Harmon, Kasif, White, & Salzberg, 1999), Easygene (Larsen & Krogh, 2003) are widely used. Rapid Annotation using Subsystem Technology (RAST) (Aziz et al., 2008) is an online GUI based server which provides with high quality annotations for prokaryotic genomes. It provides the user with downloadable annotated formats such as genbank, fasta, embl, etc files for further gleaning into the statistics of the bacterial genomes. Open Reading Frames (ORFs) both present in the subsystem as well as not present within the subsystem with both hypothetical as well non-hypothetical categories are also represented as subsystem figures by RAST. Important genes obtained from the annotations with their possible roles in interaction with any particular host components can be determined through the protein-protein interactions and protein-chemical interactions. STITCH (Kuhn, von Mering, Campillos, Jensen, & Bork, 2008) database for protein-chemical interactions and STRING (von Mering et al., 2003) database for protein-protein interactions are used widely for predicting information among chemicals and protein.

The *V. parahaemolyticus* has undergone a number of evolutionary changes, of which few strains have even been labelled as a pandemic clone. Based on different pathogenicity islands of the bacteria a haplotyping study has been done to study the grouping of the different existing strains of *V. parahaemolyticus*. The pathogenicity

islands mainly considered as being VPA1322-1369, VPaI5, TDH Pathogenicity Island, TRH Pathogenicity Island, GS-PCR, *toxS* and *toxR* genes. It has been shown how the old-pandemic clone has formed a completely new strain through microevolution over time. The genome plasticity and the "microevolution" of this bacteria has been explained in previous publications as serovariants due to their acquisition or loss of gene regions (Figure 5.2.1).

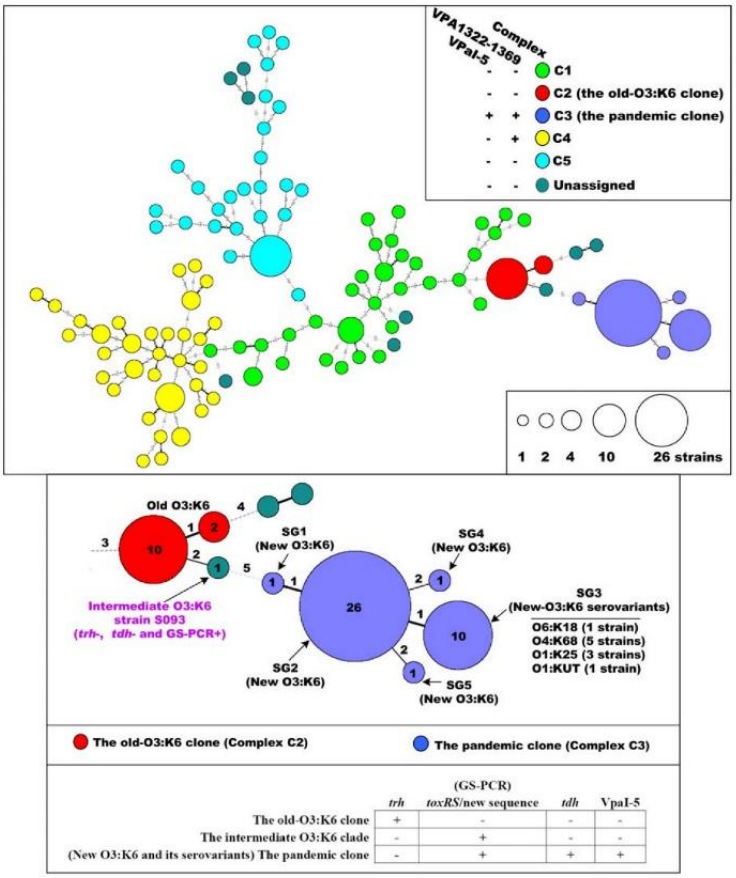


Figure 5.2.1: Minimum spanning tree of 174 global strains: each circle indicates a haplotype (node), and a larger size of the circle corresponds to a larger number of strains included. The number along each edge reflects the phylogenic distance between each two neighbouring nodes. In addition, a thicker edge corresponds to a shorter phylogenic distance. 165 (98 nodes) of the 174 strains are assigned into five complexes, C1 to C5, while the remaining nine strains are assigned into five sub-groups, SG1 to SG5 (different SGs distribute in different nodes). The presence or absence of *trh*, *tdh*, *toxRS*/new sequence and VpaI-5 in various groups of strain are shown as well (H. Han et al., 2008).

Studies have shown how a selected number of *V. parahaemolyticus* strains can acquire completely new genes over a period of time and which could be by horizontal gene transfer (Y. Chen et al., 2011). Previous reports have compared six different strains that are *V. parahaemolyticus* RIMD2210633, *V. parahaemolyticus* Peru466, *V. parahaemolyticus* AQ4037, *V. parahaemolyticus* AQ3810, *V. parahaemolyticus* AN5034 and *V. parahaemolyticus* K5030 for the presence of common and unique number of genes. Other genes related to the interaction of pathogen with the host which have not been studied evolutionarily could help us understand the specific niche to which the strains could belong to worldwide.

M. rosenbergii is a freshwater aquatic organism. The optimal range for *M. rosenbergii* larvae to survive is 28^0C – 31^0C. It has been observed that a salinity of <10% parts per thousand (ppt) would be ideal for hatcheries for freshwater prawn (New, 2002). Though calcium shows its importance in the formation of exoskeleton (Saravanan et al., 2008) it is the conditions which are favourable for the "survival" of larvae which stands of primary importance. There are previous reports which describe Magnesium as an important component in the environment for prawn survival. One such previous literature explains the requirement of the magnesium in juvenile prawns (Akio, Teshima, & Sasaki, 1984). A recent article (Hangsapreurke et al., 2008) describes the study of effects of salinity with the use of artificial sea water. Here it clearly explains the role of magnesium (Mg) in the survival rates of post larvae. Taking an example of the effect of an acidic environment with the presence of aluminium, magnesium ion (Mg^{++}) in increased levels has been observed of importance in survival stages of the post larvae (Rejeki, 2003). The characteristics of water which are good for prawn hatcheries are said to be 10-27 parts per million (ppm) Mg in freshwater, 1250-1345 ppm Mg in seawater and 460-540ppm Mg in brackish water (New, 2002). These features and conditions show how important is the Magnesium ion for the survival of

larvae which undergo a very critical “Molting stage” before reaching the post-larval stage.

Most Vibrio as shown are adapted to aquatic organisms and cause severe infections on consumption by humans. The *V. parahaemolyticus* has several virulence, pathogenicity and antibiotic resistance factors with which it can survive aquatic organisms especially the giant fresh water prawn, *M. rosenbergii* (Hameed, Rahaman, Alagan, & Yoganandhan, 2003).

The growth conditions of *M. rosenbergii* in the environment can be studied in depth to understand the adaptation correlation of *V. parahaemolyticus* to *M. rosenbergii*. Studies show that *M. rosenbergii* survival in different media compositions has been observed with variations in NaCl, KCl and $MgCl_2$ + $MgSO_4$ (Daniels, MacKinnon, et al., 2000). The fertilization envelope of shrimp eggs has been observed to thin when there is depletion of calcium and magnesium contents (Clark & Lynn, 1977). Embryos at early stages has been shown to require optimum levels of medium including $MgCl_2$ + $MgSO_4$ for their proper development (Damrongphol, Jaroensastraraks, & Poolsanguan, 2001).

The role of magnesium ion in the normal hatching rate or the newly hatched larvae is not shown to be significant (Damrongphol et al., 2001). The importance of Magnesium in survival mechanism has been observed (Hangsapreurke et al., 2008). There are various resistance factors which *V. parahaemolyticus* carry such as the Cobalt, Zinc, Cadmium and Chromium resistance genes which can also explain the possible link to *M. rosenbergii* which could have been exposed to toxic substances during its life cycle. During the course of evolution the bacteria must have acquired these resistance genes on prolonged exposure while surviving together with the host that is *M. rosenbergii*. Antitoxin genes such as the Phd antitoxin and the Doc antitoxin genes (Guerout et al., 2013) have also been identified in *V. parahaemolyticus* PCV08-7

isolated from prawn which strengthens the explanation on host-pathogen co-existence. The most interesting factor is the tolerance of *V. parahaemolyticus* unlike other bacteria to higher concentrations of Magnesium as mentioned in Chapter 2 and growth under iron-limiting conditions directly proportional to conditions of prawn larval survival.

Apart from this, another important factor is the N-acetylglucosamine-binding protein (*gbpA*) that has been studied in one Vibrio species, *Vibrio cholerae* (Jude, Martinez, Skorupski, & Taylor, 2009; E. Wong et al., 2012) to have the property to bind to epithelial cell surfaces and chitin on the host surface (Bhowmick et al., 2008). There are no reports yet on previous studies of host-pathogen interactions on *gbpA* gene of *V. parahaemolyticus* with *M. rosenbergii* as the host bearing chitin as a component in its carapace.

The evolutionary strength of this bacteria having acquired virulence could make the *V. parahaemolyticus* a strong host-pathogen partner to *M. rosenbergii* and therefore explain the host-pathogen co-existence and the presence of Magnesium ion in the environment to be hypothesized as one of the important link between *V. parahaemolyticus* and *M. rosenbergii*.

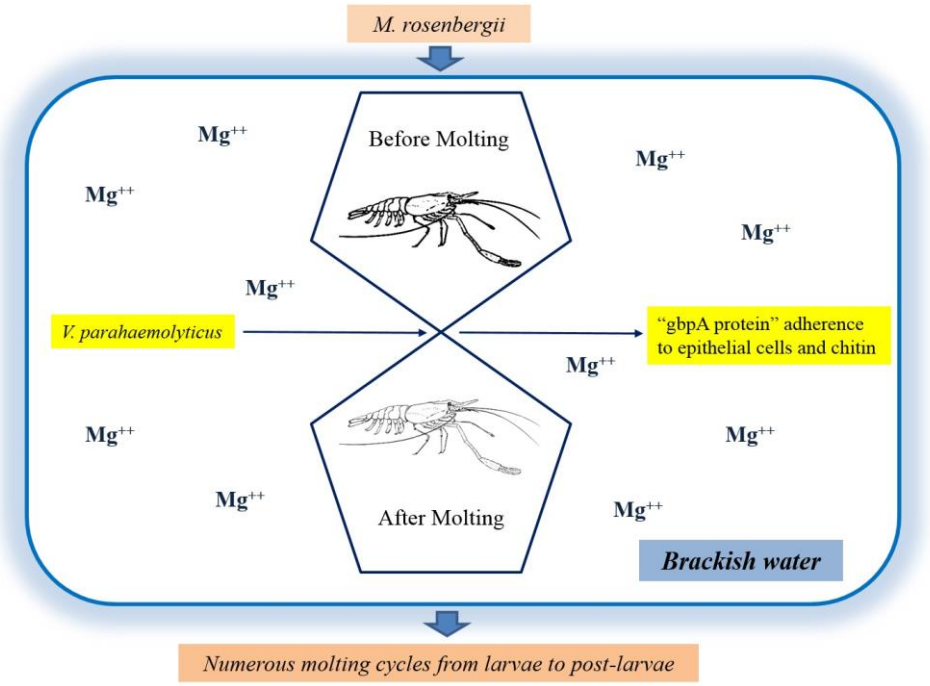


Figure 5.2.2: Hypothetical representation of the possible check point where the host-pathogen interaction of *V. parahaemolyticus* and *M. rosenbergii* can be targeted to deal with bacteria infecting seafood.

5.3 Materials and Methods

5.3.1 Next-generation sequencing

V. parahaemolyticus PCV08-7 genome has been sequenced by Illumina GAIIX analyzer as well as Ion Proton technology by Genotypic Pvt. Ltd, Bangalore, India. The plasmid has been isolated by the Hi Yield plasmid extraction kit (BioAmerica) (Appendix C (a)). The plasmid has been sent for sequencing through Illumina (GA2x, pipeline version 1.6). The genomic DNA was isolated using Qiagen DNeasy Blood & Tissue kit (Qiagen, Germany) and the genome sequence was determined by Ion proton. The details of the sequencing are shown in Table 5.3.1.

Table 5.3.1: Details of the Next-generation sequencing output raw read data

	VPPCV08-7 Illumina sequencing	VPPCV08-7 plasmid Illumina sequencing	VPPCV08-7 Ion torrent sequencing
Read Length (bp)	100	101	286
Read 1	2178073	30260889	81219173
Read 2	2178073	30260889	-
Total reads	4356146	60521778	-

5.3.2 Quality control and Assembly of *V. parahaemolyticus* PCV08-7 NGS data

The Trimming-Adaptor-High quality (TAH) pipeline customized quality control pipeline has been used for the filtering of raw sequencing read data of VPPCV08-7. The high quality step has been tweaked so as to compliment the quality score format in the plasmid Illumina and Ion torrent sequencing reads as per the syntax mentioned below. The Illumina sequencing reads were high quality filtered using the FASTX toolkit while the Ion torrent sequencing reads were processed using PRINSEQ.

VPPCV08-7 Illumina sequencing data high quality filtering step

./fastq_quality_filter -t 21 -l 100 -i /PATH/xx.fastq -o /path/output*VPPCV08-7 plasmid Illumina sequencing high quality filtering step*

./fastq_quality_filter -t 10 -l 100 –i /PATH/xx.fastq -o /path/output

VPPCV08-7 Ion torrent sequencing data high quality filtering step

perl prinseq-lite.pl -verbose -fastq /PATH/xx.fastq -out_good null -out_bad null -min_len 54 -trim_to_len 54 -out_format 3

All the three datasets corresponding to Illumina and Ion proton sequenced genomic DNA as well as the Illumina sequenced plasmid DNA were assembled by Velvet-1.0.12a and Edena (Appendix C (b)).

5.3.3 Alignment and Annotation of VPPCV08-7 contigs

The contigs produced from the VPPCV08-7 plasmid Illumina sequencing data using Edena assembler were used for alignment against the reference genome *V. parahaemolyticus* RIMD2210633 as in Chapter 4. The methodology as in Chapter 4 has been followed to re-order contigs aligning with Mauve to the concatenated reference genome. The re-ordered contigs (Appendix C (c)) were once again aligned against the reference to label the contigs to their respective chromosomes, represented as a heatmap. The contigs which show match in both the chromosomes were retained in a separate file. The contigs with no match were checked against the Genbank database and then placed into a file named plasmid.fa. Contigs matching with only RNA and with match in both the chromosomes were separated into two different files. The annotation of the parsed components was done using the online Rapid Annotation using Subsystems Technology (RAST). tRNA-scanSE (Lowe & Eddy, 1997) has been used for analysing the tRNA genes.

5.3.4 Protein-protein and Protein-chemical interactions

Databases STITCH v1.9 for protein-protein interactions and STRING 3 for protein-chemical interactions were used to check the interactions of the *gbpA* gene of the VPPCV08-7 with other proteins and chemical molecules.

5.3.5 Alignment and Phylogenetic analysis of *V. parahaemolyticus gbpA* gene

With publicly available annotated *V. parahaemolyticus* genomes as well as Sequence Read Archive data on Genbank, 52 different genomes (Appendix C (d)) were selected to check the evolutionary relationship of each genome with reference to our selected gene of interest *gbpA*, two house-keeping genes of *V. parahaemolyticus*, *dnaB* (DNA helicase), *deaD* (RNA helicase) and 16S rRNA sequences. Chitinase gene isoforms found in the genome of *V. parahaemolyticus* have also been used for the study. Standalone BLASTN has been used to obtain sequences from all the 52 individual genome files. ClustalW (J. D. Thompson, Higgins, & Gibson, 1994) has been used on the command-line to produce multiple alignments (Appendix C (e)). Splitstree (Huson, 1998) has been used to run Neighbour-joining with a bootstrap value of 1000 and Figtree v1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/) for plotting the phylogenetic tree in order of increasing clades with a scalebar of 2.0.

5.3.6 Polymerase chain reaction

Polymerase chain reaction has been performed on the *gbpA* (N-acetylglucosamine-binding protein) gene for the whole length of 1464 bp gene. The tools used for primer designing were AmliFX , DNA star (Burland, 2000), Primer3 (Untergasser et al., 2012; Untergasser et al., 2007) and NCBI primer Blast (Ye et al., 2012). The house keeping genes that are *dnaB* (DNA helicase) and *deaD* (RNA helicase) have also been selected to perform PCR for the whole length of the genes (Table 5.3.2) and for their selected domain lengths (Figure 5.3.1).

Table 5.3.2: Table with details of the PCR and Real time PCR primers.

S.No	GENE	FORWARD PRIMER	LENGTH (bp)	REVERSE PRIMER	LENGTH (bp)
1	*gbpA*	ATGAAATCATTCCCTAATAAATCGCTC	27	TTAGTTCAGTTTATCCCAAGCCATT	25
2	*dnaB*	CCACTACCACTCTCATATGC	20	GGTACTAAATGGCTGACATC	20
3	*deaD*	GGCTCAAAATGGTTAAGCG	19	CATTTCCGCTCTCATATGC	19
REAL	TIME PRIMERS				
4	*gbpA*	CTCGTTCGCTCTCAACCCTT	20	CACAGGGTCGTCACCATCAA	20
5	*dnaB*	CGTGAGCACGGAGGTCTATC	20	GTCGTTTATCCGCTCGTTGC	20
6	*deaD*	GTGCACACGTTGTTGTTGGT	20	AGAACGCGTTGTGCTGATTC	20

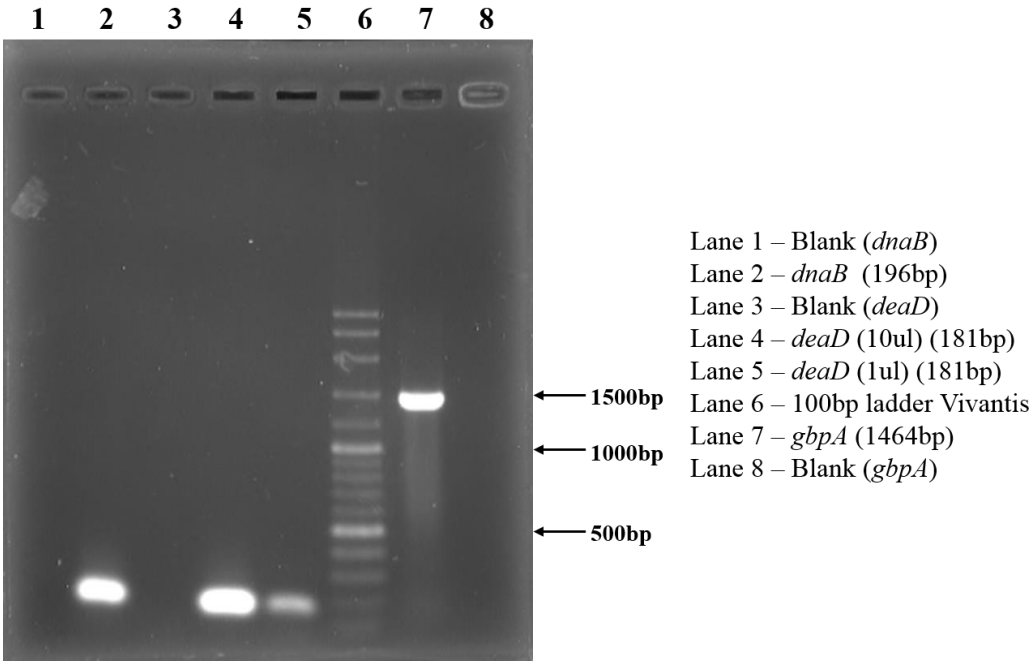


Figure 5.3.1: Agarose gel electrophoresis (1% w/v) showing the PCR products *dnaB*, *deaD* and *gbpA* genes at their corresponding gene lengths

The gel picture shows 100 bp ladder (Vivantis) in Lane 6 with 1464 bp amplified *gbpA* gene product in Lane 7, 196 bp amplified *dnaB* gene product (corresponding to its domain region) in Lane 2 and 181 bp amplified *deaD* gene product (corresponding to its domain region) in Lane 4. The negative control or blank can be observed in Lane1, Lane3 and Lane8 for *dnaB*, *deaD* and *gbpA* respectively.

The domains of the three genes were selected for amplifying (Table 5.3.2). The PCR working concentrations for all the three genes *deaD*, *dnaB* and *gbpA* were standardized as follows: 1X Gotaq green buffer, 1.5 mM Magnesium chloride ($MgCl_2$), 0.2 mM

dATP, 0.2 mM dCTP, 0.2 mM dTTP, 0.2 mM dGTP, 0.25 µM forward primer, 0.25 µM reverse primer, 0.025 u/ul Taq polymerase, 5 µg/ul genomic DNA template and made up to 10 µl reaction using ultrapure water.

The PCR reactions were run on Biorad Thermal C1000 machine with the standardized reaction as follows: Initial denaturation of 95^0C/5 min, Denaturation 95^0C, Annealing 60^0C and Extension 72^0C for 1 min and 40 cycles, Final extension 72^0C/5 min and 40^0C forever. PCR amplified products were sent to First Base Laboratories (Malaysia) for confirming the gene (Appendix C (f)).

5.3.7 Cloning of gbpA gene

gbpA gene has been cloned into pGEM vector and has been transformed into JM109 *Escherichia coli* bacterial cells for isolation of cloned plasmid DNA. After transformation by the standard pGEM T-easy vector system I kit protocol (Promega), colony PCR for 190 bp product length of gbpA gene with real time PCR primers (Table 5.4) has been done. Plasmid isolation (Appendix C (g)) was performed to check the presence of the gene by PCR (Figure 5.3.2).

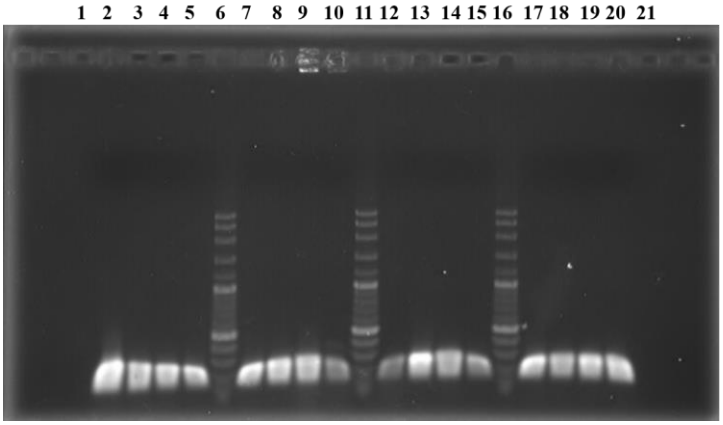


Figure 5.3.2: Agarose gel electrophoresis (1% w/v) showing the PCR product with the *gbpA* real time primers (190 bp length corresponding to its domain region) on the 16 isolated colonies after transformation of the *gbpA* clone.

Selected number of plasmid isolations were sent for DNA sequencing to FirstBase Laboratories (Malaysia) to confirm the clone (Appendix C (h)). The concentration of

the *gbpA* gene was obtained after PCR amplification which as per the Nanodrop 2000/2000c Spectrophotomer readings (Thermo-Scientific) was 37.9 ng/µl.

5.3.8 *V. parahaemolyticus* PCV08-7 culture and bacterial plate count

Luria-Bertani Broth with 2% NaCl (Sodium chloride) was prepared for culturing VPPCV08-7 with Magnesium as the environment and in the presence of carapace of prawn and commercial grade chitin (chitin from shrimp shells, SIGMA). Six concentrations of Magnesium sulphate hepta hydrate ($MgSO_4.7H_2O$) had been used (Stock prepared was 500 mM). The three sets of experiments were performed as follows. Six different concentrations of $MgSO_4.7H_2O$ were selected corresponding to 0 ppm (0 mM), 300 ppm (1 mM), 18500 ppm (75 mM), 34000 ppm (137 mM), 55500 ppm (225 mM), 73941 ppm (300 mM) respectively for each set. The first set was added with varied concentrations of $MgSO_4.7H_2O$. 0.05 g/3 ml by dry weight of carapace of the prawn was added per tube to all the six concentrations of $MgSO_4.7H_2O$ in the second set. While the third set of six concentrations of $MgSO_4.7H_2O$ were added with 0.05 g/3 ml of chitin flakes (Sigma).

1. (Without carapace)

   0 ppm, 300 ppm, 18500 ppm ,34000 ppm , 55500 ppm ,73941 ppm

2. (With carapace)

   0 ppm, 300 ppm, 18500 ppm ,34000 ppm , 55500 ppm ,73941 ppm

3. (With chitin flakes (Sigma))

   0 ppm, 300 ppm, 18500 ppm ,34000 ppm , 55500 ppm ,73941 ppm

The bacterial plate count has been done in triplicates by counting on Difco TCBS agar by incubating the spread plates overnight at 37^0C for 16-20 hours (Appendix C (i)).

5.3.9 RNA isolation and cDNA synthesis

RNA has been isolated from the triplicates of each of the samples above using Promega Total RNA isolation kit and converted to cDNA using the reverse transcription PCR protocol (Appendix C (j)). 4 µl of each isolated RNA sample from each of the triplicates was added with 1 µl Random primer to make a reaction mixture of 5 µl for an initial incubation at 70^0C/5 mins followed by 4^0C/5 mins in Biorad PCR machine. A reverse transcription mix of 15 µl each (6.1 µl Nuclease free water, 4 µl of 5X reaction buffer, 2.4 µl $MgCl_2$, 1 µl dNTPs, 0.5 µl ribonuclease inhibitor and 1 µl reverse transcriptase) has been added to the former mixture for annealing at 25^0C/5 min, extension at 42^0C/60min and heat inactivation of reverse transcriptase at 70^0C/15 min for cDNA synthesis.

5.3.10 Real Time Polymerase chain reaction using SYBR green

Absolute quantification

The standardization of absolute quantification of cDNA was performed on the Applied Biosystems 7500 Real time PCR system with varied DNA concentrations of 0.005 µg, 0.05 µg, 0.5 µg, 5 µg and 50 µg. While the forward primer and the reverse primers were standardized with 0.6 nm, 0.3 nm and 0.15 nm concentrations, respectively (Figure 5.3.3)

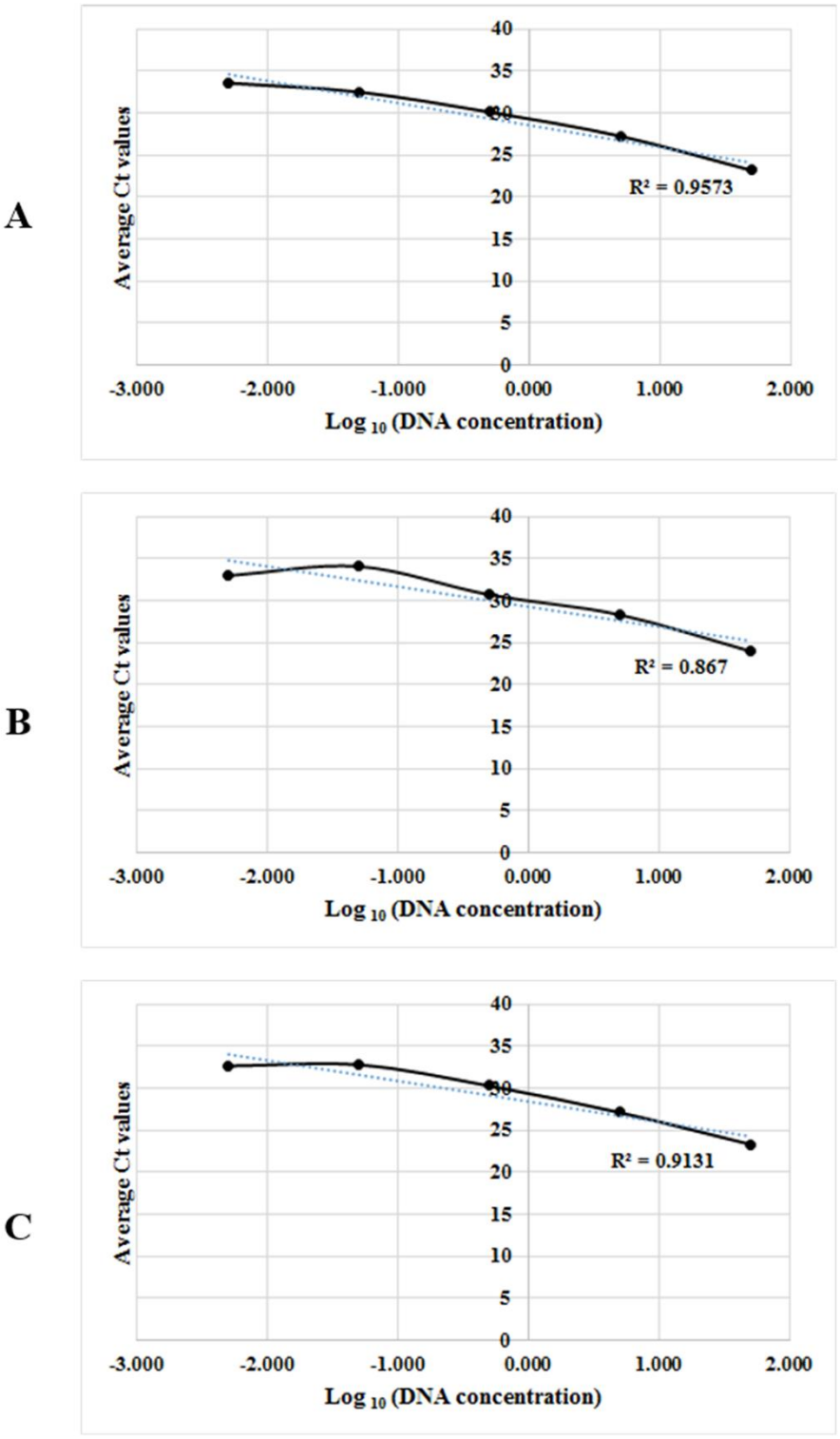


Figure 5.3.3: Absolute quantitative standard curve for DNA concentrations (0.005µg, 0.05µg, 0.5µg, 5µg and 50µg) vs Average Ct values for (A) primer concentration of 0.15nm (B) primer concentration of 0.3nm and (C) primer concentration of 0.6nm respectively

R^2 values for primer concentrations of 0.15nm, 0.3nm and 0.6nm are 0.9573, 0.867 and 0.9131 respectively.

The default thermal cycling conditions were used for the run with stage 1 of 50^0C/2 min and stage 2 95^0C/10 min for one cycle, stage 3 with 40 repetitions with 95^0C/15 sec followed by 60^0C/1 min (Appendix C (k)).

Relative quantification

Relative quantification of the cDNA samples was performed on an Applied Biosystems 7500 Real Time PCR system with each qRT-PCR reaction amounting to 10 µl Power SYBR green PCR master mix, 0.6 µl forward primer, 0.6 µl reverse primer 1 µl DNA template and 7.8 µl ultrapure water. The default thermal cycling conditions are used for the run with stage 1 of 50^0C/2 min and stage 2 95^0C/10 min for one cycle, stage3 with 40 repetitions with 95^0C/15 sec followed by 60^0C/1 min (Appendix C (l)).

5.4 Results and Discussion

5.4.1 *Vibrio parahaemolyticus* PCV08-7 draft genome completion

Applying the TAH pipeline as in the discussion from Chapter 3 showed reasonably good usage of reads corresponding to the VPPCV08-7 sequencing reads. Using the pre-existing tools the data from all sequencing data has been brought to one common format to deal with the accuracy of the TAH pipeline (Appendix C (b)). The VPPCV08-7 plasmid Illumina sequencing generated contigs were aligned against the concatenated reference genome of VPRIMD2210633. The contigs were separated into chromosome 1, chromosome 2 and the plasmid depending on the alignment hits obtained (Appendix C (c)). The presence of RNA genes in all these files hence, support the fact that PCR reactions for rRNA and tRNA can help support the completion of a bacterial genome as it could be one common factor among all bacterial genomes. Previously, there were studies reported on optically mapping bacterial genomes for completion of genomes experimentally (Riley, Lee, Lesho, & Kirkup, 2011).

The individual files were annotated on the RAST annotation system which showed a total of 1676, 546 and 1 ORFs in the subsystem in VPPCV08-7 chr1, VPPCV08-7 chr2 and VPPCV08-7 plasmid, respectively (Table 5.4.1).

Table 5.4.1: Number of ORFs distributed into subsystems in the draft of chromosome 1, chromosome 2 and plasmid of VPPCV08-7

Table shows ORF's distributed under subsytems annotated by RAST server as 1676, 546, 1 for VPPCV08-7 chr1, VPPCV08-7 chr2 and VPPCV08-7 plasmid, respectively. The number of ORF's not present within the subsystem were 1382, 1066, 12 for VPPCV08-7 chr1, VPPCV08-7 chr2 and VPPCV08-7 plasmid, respectively.

	VPPCV08-7 chr1	VPPCV08-7 chr2	VPPCV08-7 plasmid
SUBSYSTEM			
Total	1676	546	1
Non-hypothetical	1569	510	1
Hypothetical	107	36	0
NOT IN SUBSYSTEM			
Total	1382	1066	12
Non-hypothetical	715	629	11
Hypothetical	667	437	1

Among the above, 1569, 510 and 1 number of ORFs were non-hypothetical and 107, 36 and 0 were hypothetical corresponding to the individual components. A total of 1382, 1066 and 12 ORFs were found in chr1, chr2 and the plasmid respectively which are not present within the subsystem (Figures 5.4.1, 5.4.2 and 5.4.3).

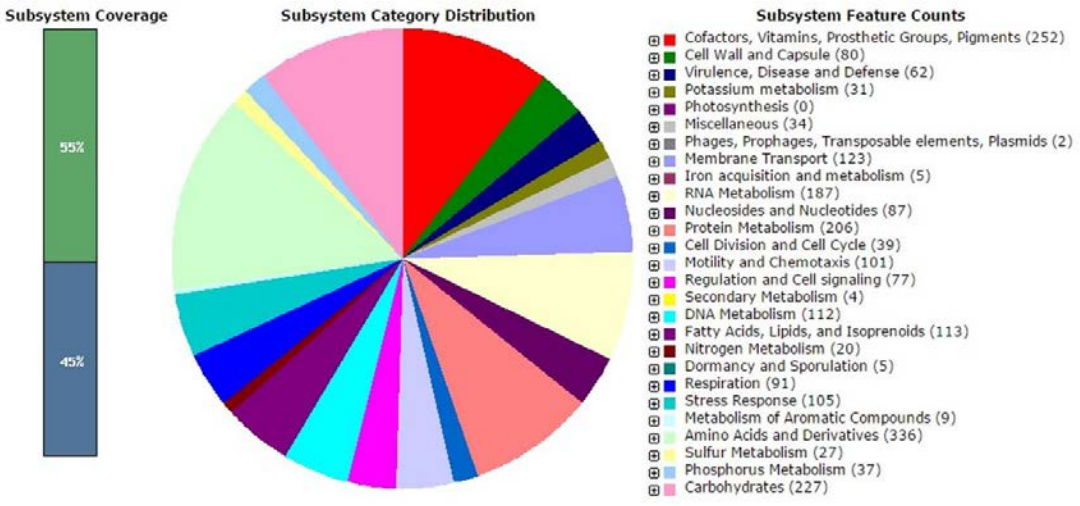


Figure 5.4.1: Subsystem category distribution of the parsed draft chromosome 1 of VPPCV08-7

The RAST subsystem category distribution shows highest number of feature counts under Amino acids and derivatives with observable counts under phages, prophages, transposable elements and plasmids.

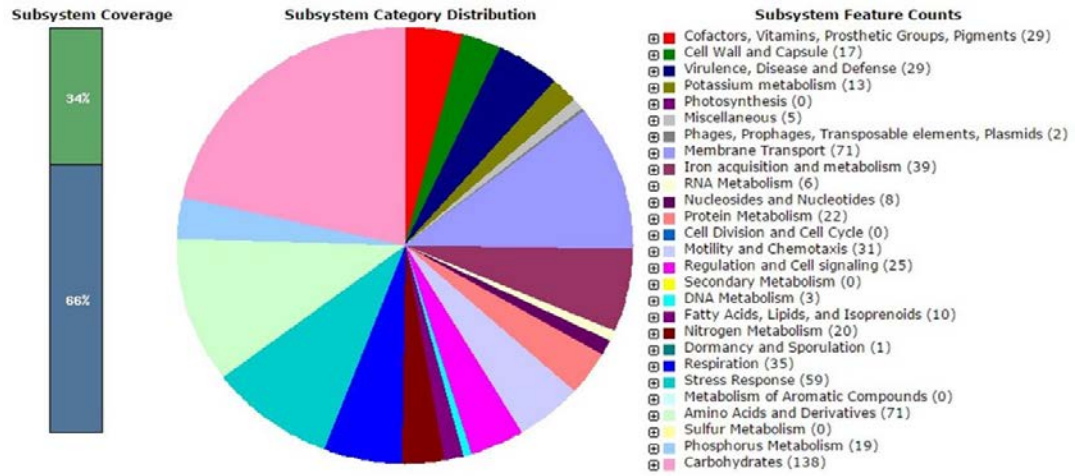


Figure 5.4.2: Subsystem category distribution of the parsed draft chromosome 2 of VPPCV08-7

The RAST subsystem category distribution shows highest number of feature counts under Carbohydrates with observable counts under phages, prophages, transposable elements and plasmids.

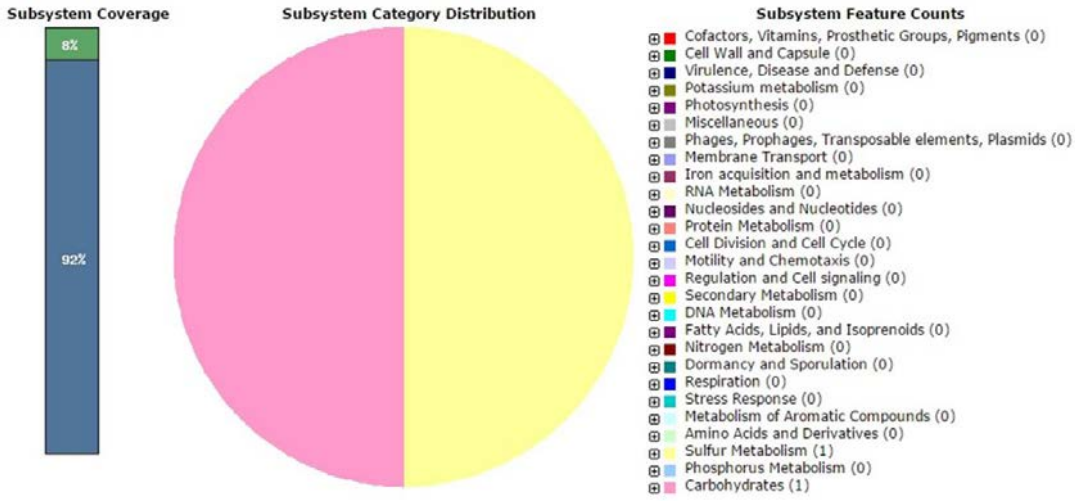


Figure 5.4.3: Subsystem category distribution of the parsed probable plasmid of VPPCV08-7

The RAST subsystem category distribution shows feature counts in sulphur metabolism and carbohydrates.

The VPPCV08-7 chr1, chr2 and plasmid data has been used to glean into the important genes in similarity with the VPRIMD2210633 genome. Alignments matching chromosome 1 of VPRIMD2210633 were genes coding for lipopolysaccharides, phagef37 like genes type 3 secretion system 1 and the integron class 1 which are partially present. While the hypothetical proteins and osmotolerance genes are completely present. Alignments against chromosome 2 of VPRIMD2210633 showed partial presence of genes responsible for degradation processes, phagef37 like genes, biofilm formation and multidrug efflux systems. The genes coding for gametolysin (peptidases, amylases, lipases), osmotolerance, capsule lipopolysaccharide, type 1 secretion, type 1 pilus, ferric uptake system and a 140kb region not found in Early Mortality syndrome (EMS) strains were present (Table 5.4.2).

Table 5.4.2: Important genes of VPRIMD2210633 and EMS strain and their presence in VPPCV08-7

The list of genes in the table represent important selected genes of VPRIMD2210633 chr1 (Lipopolysaccharides, Phagef37, T3SS-1, Osmotolerance and Integronclass-1), VPRIMD2210633 chr2 (Genes in degradation processes, Phagef37-like genes, Biofilm, Gametolysin, Osmotolerance, Capsule lipopolysaccharide, Type 1 secretion, Type 1 pilus, Multidrug efflux system, Ferrics uptake system) and VPTUMSATD06_S3 (63kb gene stretch in EMS strain) and their presence or absence in VPPCV08-7.

Serial No	ORF/ contig	Product name/ Importance	Position/ Contig	VPPCV08-7
VPRIMD2210633 chr1				
1	VP0081-VP0092	Hypothetical proteins	92601 – 102006	PRESENT
2	VP0218-VP0234	Lipopolysaccharides	229457 – 247387	PARTIAL
3	VP1549-VP1590	Phagef37	1658782 1683479	PARTIAL
4	VP1658-VP1702	T3SS-1	1777109 – 1812764	PARTIAL
5	VP1719-VP1728	Osmotolerance	1832356 – 1844597	PRESENT
6	VP1787-VP1865	Integronclass-1	1900885 – 1949243	PARTIAL
VPRIMD2210633 chr2				
1	VPA0434-VPA0458	Genes in degradation processes	423787 – 452216	PARTIAL
2	VPA0887-VPA0914	Phagef37-like genes	932561 – 948136	PARTIAL
3	VPA0950-VPA0962	Biofilm	985238 – 1005188	PARTIAL
4	VPA0989-VPA0999	Gamctolysin	1035685 – 1052235	PRESENT
5	VPA1102-VPA1115	Osmotolerance	1162547 – 1179213	PRESENT
6	VPA1403-VPA1412	Capsule lipopolysaccharide	1475072 – 1486668	PRESENT
7	VPA1440-VPA1444	Type 1 secretion	1518749 – 1538480	PRESENT
8	VPA1503-VPA1521	Type 1 pilus	1607136 – 1625613	PRESENT
9	VPA1559-VPA1583	Multidrug efflux system	1658285 – 1680999	PARTIAL
10	VPA1652-VPA1679	Ferric uptake system	1767656 – 1800892	PRESENT
11	140kb	Not present in EMS strain	1387700 – 1467700	PRESENT
VPTUMSATD06_S3	63kb	Present in EMS strain	Contig 4	ABSENT

The 140kb region constitutes thermostable direct hemolysin A and S, transposases, cytotoxic necrotizing factor, transcriptional regulator, putative type III secretion proteins, membrane proteins and hypothetical proteins. A 63kb stretch of sequence present in *Vibrio parahaemolyticus* TUMSATD06_03, a recently reported EMS strain from Thailand was not present in our genome. The 63kb stretch constitutes conjugal transfer proteins (Trb C, D, E, F, G, H, I, L), Type II and III secretion proteins and matches a *Vibrio parahaemolyticus strain* 13-028/A3 *plasmid* pVPA3-1 containing Photorhabdus insect-related (Pir) genes (J. E. Han, Tang, Tran, & Lightner, 2015). This explains that the current genome has no common features with that of the EMS strain ruling out the fact that our genome could be fatal for prawn. The online BLASTN analysis with the contigs separated into the plasmid file against the Genbank database corresponded to genes from *Vibrio parahaemolyticus* strain vp plasmid pVP1 and *Vibrio parahaemolyticus* UCM-V493 chr1 which correspond to conjugative transfer protein Tra1, relaxase and mobile elements.

The above data extracted from the draft completion of genome showed that though the number of contigs generated might be comparatively more during assembly, it indeed helps not to lose important gene information. It supports in identifying the undetectable gene features which might as well would have been lost existing within huge contigs during the assembly process.

5.4.2 Bioinformatic analysis of host-pathogen genes

Previous studies have shown that the gbpA gene product of *Vibrio cholerae* interacts with the intestinal epithelial chitinous membrane (Bhowmick et al., 2008). As *M. rosenbergii* is known to be infected by *V. parahaemolyticus* we have selected this gene for probable interactions with any other proteins or chemicals (Frederiksen et al., 2013). The results of protein-protein interactions and protein-chemical interactions of *gbpA* of *Vibrio parahaemolyticus* showed interactions with various genes and carbohydrates.

Among which the chitinase and chitin were observed to be closer to our research area. The results obtained from the STRING v9.1 interactions show the interaction of *gbpA* with various other gene products of VPRIMD2210633 and also its possible interaction with chitinase (Figure 5.4.4).

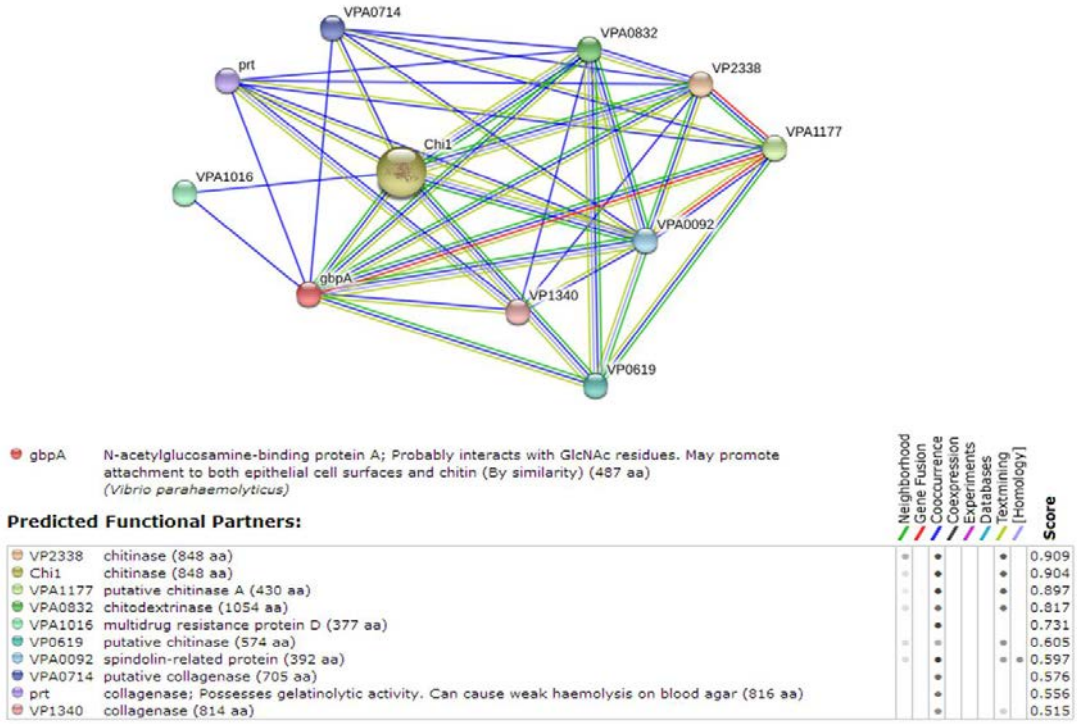
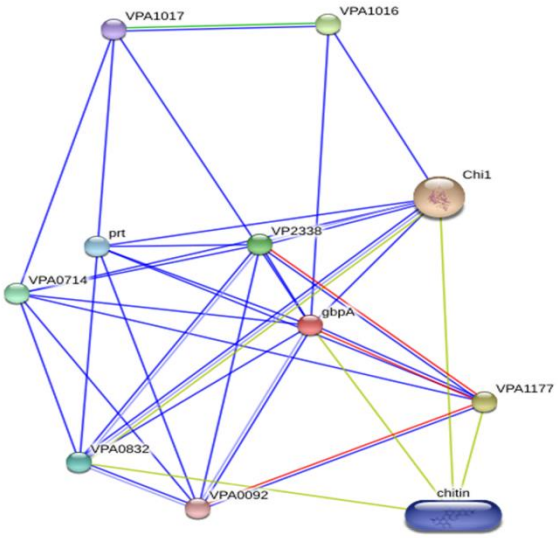


Figure 5.4.4: STRING v9.1 protein-protein interactions of gbpA protein of *V. parahaemolyticus*

Protein-protein interactions of gbpA (N-acetylglucosamine-binding protein A) on STRING v9.1 database shows interactions primarily with gene proteins of *Vibrio parahaemolyticus* (VPA0714 – putative collagenase, VPA0832 - chitodextrinase, VP2338 - chitinase, VPA1177 – putative chitinase A, VPA0092 – spindolin-related protein, VP1340 - collagenase, VP0619 – putative chitinase, VPA1016 – multidrug resistance protein D, prt - collagenase, Chi1 - chitinase).

Figure 5.4.5 obtained from STITCH 3 shows the probable interaction with chitin which explains that the gbpA protein could participate in epithelial cell surface binding whose component could be chitin, a major complex carbohydrate constituent of the prawn shell.



gbpA N-acetylglucosamine-binding protein A; Probably interacts with GlcNAc residues. May promote attachment to both epithelial cell surfaces and chitin (By similarity) (487 aa)
(Vibrio parahaemolyticus)

Predicted Functional Partners:

		Score
Chi1	chitinase (848 aa)	0.745
VPA1177	putative chitinase A (430 aa)	0.744
VPA1016	multidrug resistance protein D (377 aa)	0.731
VP2338	chitinase (848 aa)	0.685
VPA0714	putative collagenase (705 aa)	0.576
VPA0832	chitodextrinase (1054 aa)	0.575
prt	collagenase; Possesses gelatinolytic activity. Can cause weak haemolysis on blood agar (816 aa)	0.556
chitin	Polysaccharide. (627.6 g/mol)	0.503
VPA1017	AraC/XylS family transcriptional regulator (259 aa)	0.488
VPA0092	spindolin-related protein (392 aa)	0.463

Figure 5.4.5: STITCH 3 protein-chemical interactions of gbpA protein of *V. parahaemolyticus*

Protein – chemical interactions of *gbpA* (N-acetylglucosamine-binding protein A) on STITCH 3 database shows interactions primarily with proteins of *V. parahaemolyticus* (VPA1016 – multidrug resistance protein D, VPA1017 – AraC/XylS family transcriptional regulator, VP2338 – chitinase, VPA1177 – putative chitinase A, VPA0092 – spindolin-related protein, VPA0832 – chitodextrinase, VPA0714 – putative collagenase, prt – collagenase, Chi1 – chitinase) and chitin (one of the main constituents of the shell of prawn (*M. rosenbergii*)).

5.4.3 Phylogenetic analysis of VPPCV08-7

Phylogenetic analysis has been performed on 52 genomes of *V. parahaemolyticus* selected from Genbank (Appendix C (d)). Among which the VPPCV08-7, VPRIMD2210633, VPO1:K33 strains and recent EMS strains (VPBAVF, VPBAHG, VPBAHH) were of our interest. The genes selected for phylogenetic analysis were DNA helicase (*dnaB*) and RNA helicase (*deaD*) which represent the basic machinery of any bacterial genome, the 16S rRNA which has been used as the blueprint for

identifying the genus of bacterial genomes and the *gbpA* gene which has been used in our study as a possible interacting agent with the host from the *V. parahaemolyticus*. Among all the phylogenetic trees the genome VPPCV08-7 lies closer to the Thailand EMS strains over the phylogenies plotted for 16S rRNA (Figure 5.4.6), DNA helicase (Figure 5.4.7), RNA helicase (Figure 5.4.8) and gbpA (Figure 5.4.9) genes representing a closer proximity in relation to the basic machinery of *V. parahaemolyticus*.

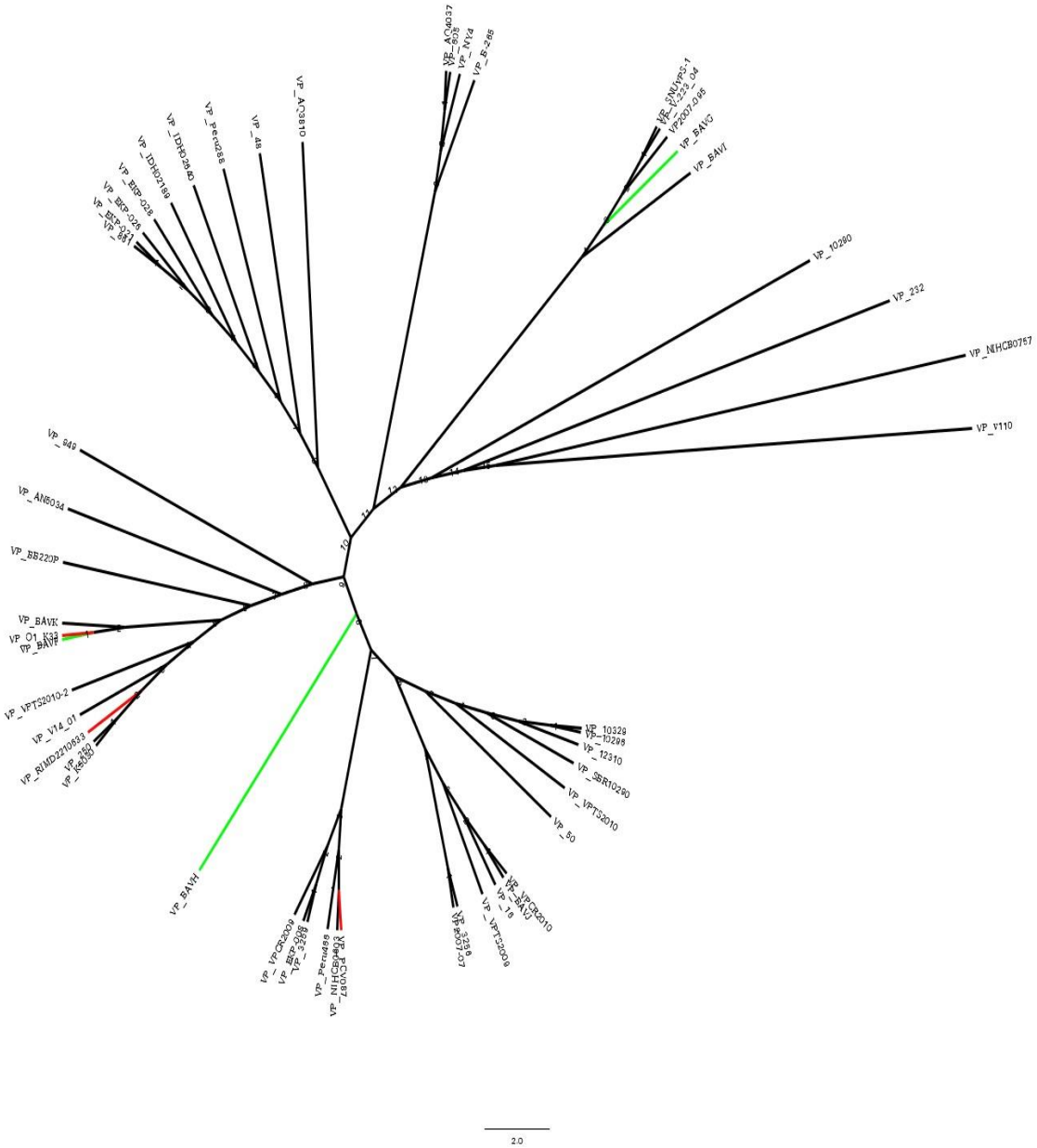


Figure 5.4.6: Phylogenetic tree representing the selected 52 *V. parahaemolyticus* genomes in relation with the 16S rRNA gene. Marked in RED are the three genomes VPPCV08-7, VPRIMD2210633 and VP01:K33 genomes. In GREEN are the three EMS strains from Thailand VPBAVF (*V. parahaemolyticus* TUMST_DE1_S1), VPBAVG (*V. parahaemolyticus* TUMSAT_DE2_S2) and VPBAVH (*V. parahaemolyticus* TUMSAT_D06_S3)

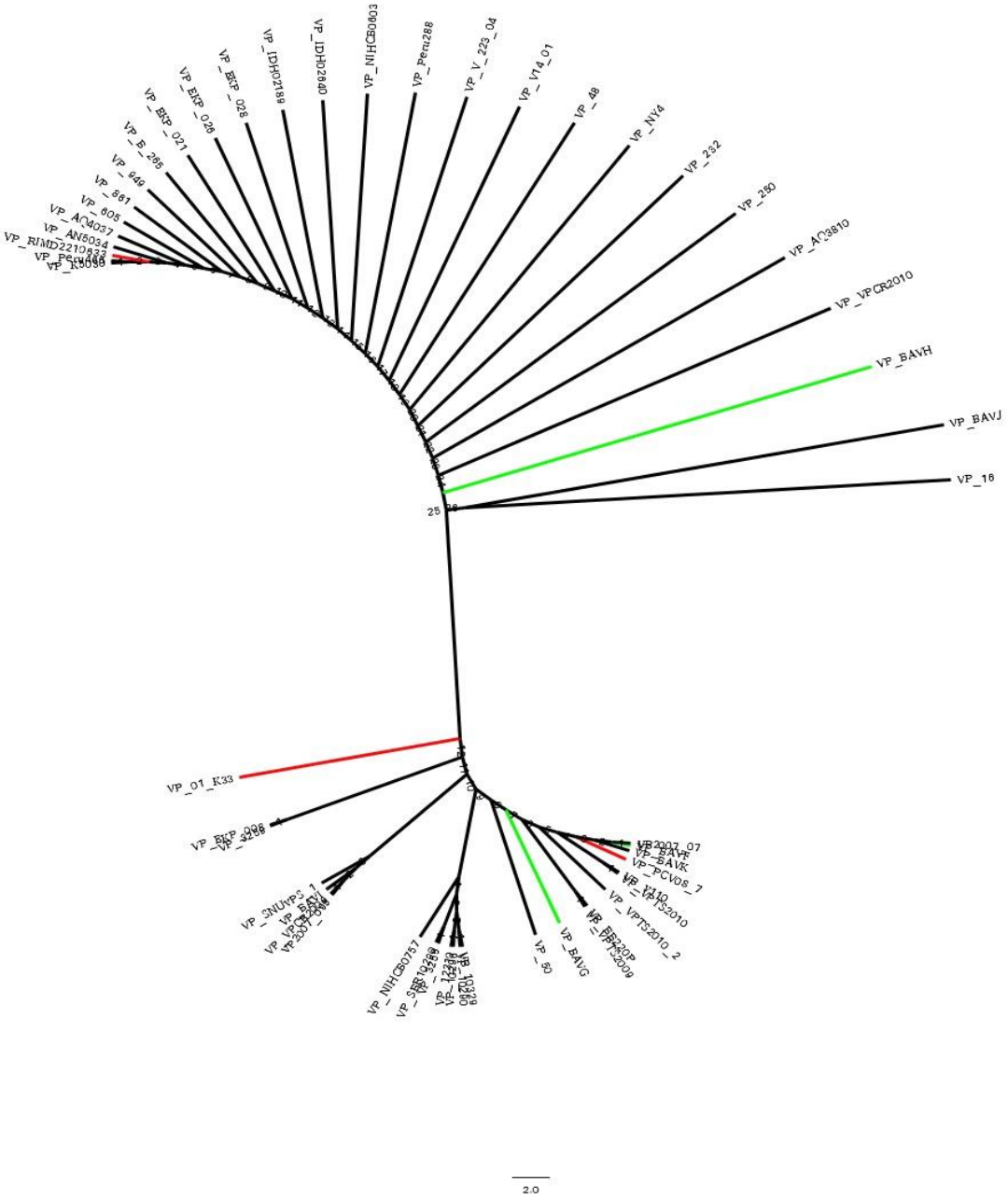


Figure 5.4.7: Phylogenetic tree representing the selected 52 *V. parahaemolyticus* genomes in relation with the house-keeping *dnaB* (DNA helicase) gene. Marked in RED are the three genomes VPPCV08-7, VPRIMD2210633 and VP01:K33 genomes. In GREEN are the three EMS strains from Thailand VPBAVF (*V. parahaemolyticus* TUMST_DE1_S1), VPBAVG (*V. parahaemolyticus* TUMSAT_DE2_S2) and VPBAVH (*V. parahaemolyticus* TUMSAT_D06_S3)

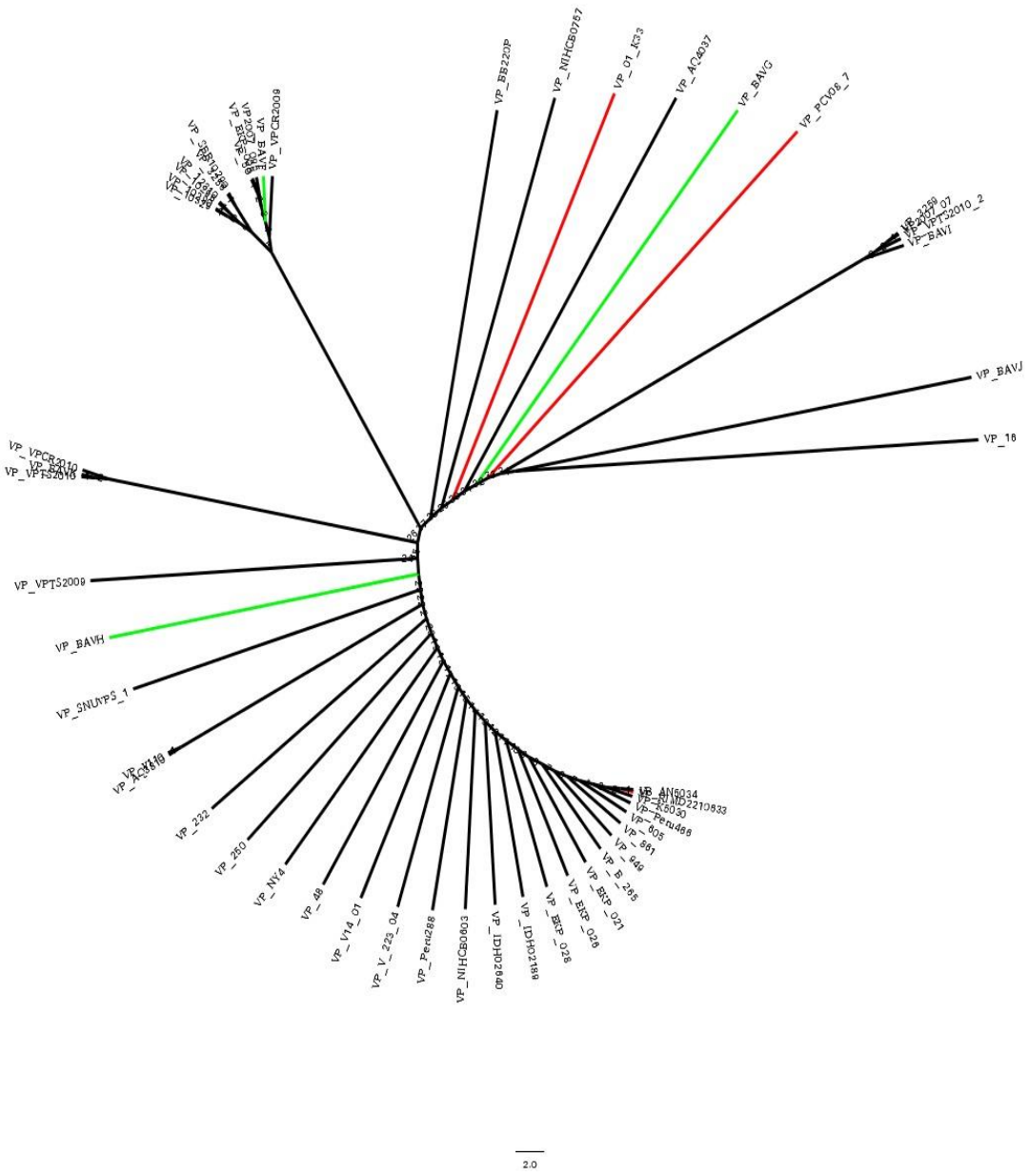


Figure 5.4.8: Phylogenetic tree representing the selected 52 *V. parahaemolyticus* genomes in relation with the *deaD* (RNA helicase) gene. Marked in RED are the three genomes VPPCV08-7, VPRIMD2210633 and VP01:K33 genomes. In GREEN are the three EMS strains from Thailand VPBAVF (*V. parahaemolyticus* TUMST_DE1_S1), VPBAVG (*V. parahaemolyticus* TUMSAT_DE2_S2) and VPBAVH (*V. parahaemolyticus* TUMSAT_D06_S3)

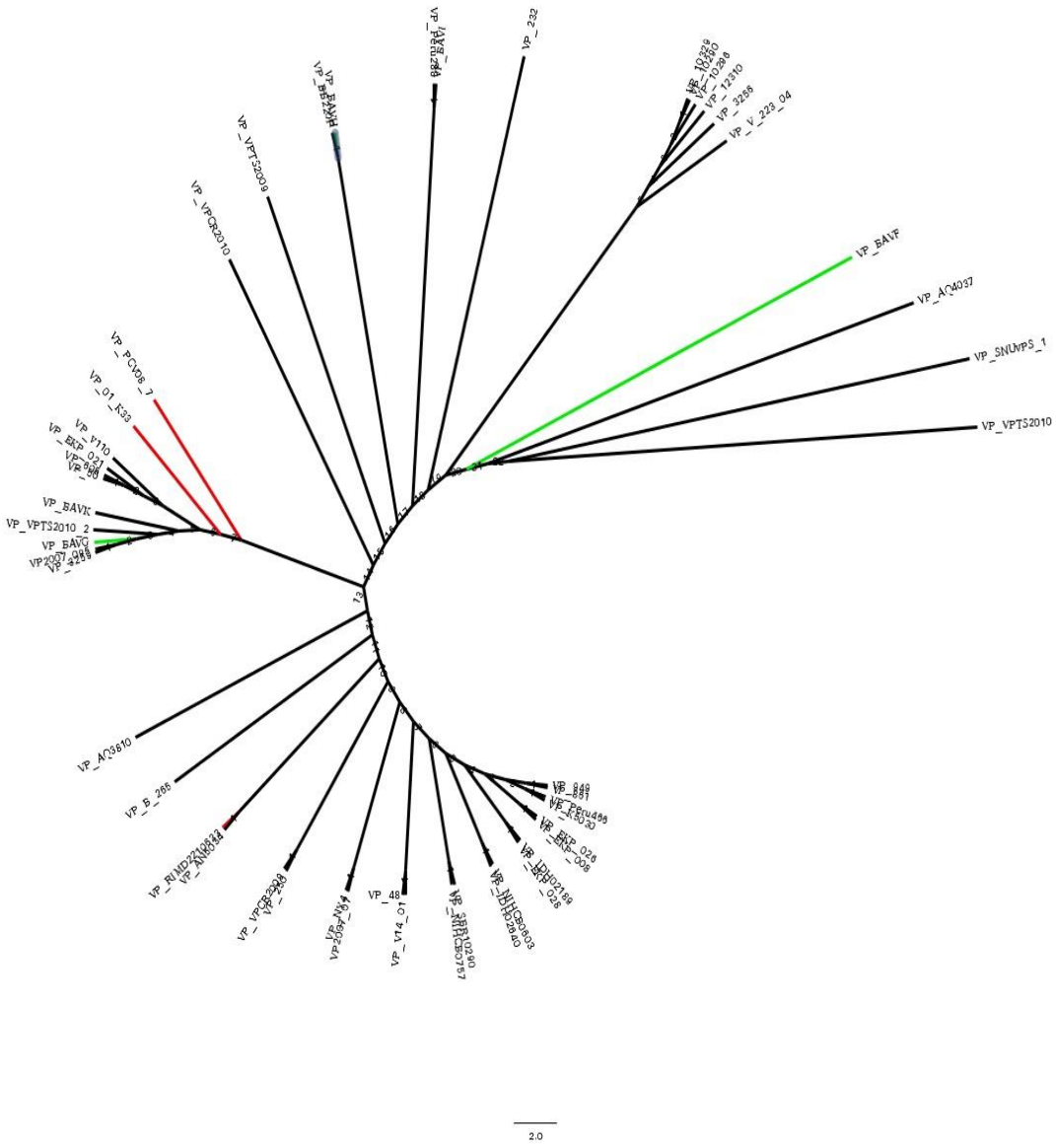


Figure 5.4.9: Phylogenetic tree representing the selected 52 *V. parahaemolyticus* genomes in relation with the *gbpA* gene. Marked in RED are the three genomes VPPCV08-7, VPRIMD2210633 and VP01:K33 genomes. In GREEN are the three EMS strains from Thailand VPBAVF (*V. parahaemolyticus* TUMST_DE1_S1), VPBAVG (*V. parahaemolyticus* TUMSAT_DE2_S2) and VPBAVH (*V. parahaemolyticus* TUMSAT_D06_S3)

VPRIMD2210633 taken as the reference genome for VPPCV08-7 is observed to have a distant relation in terms of *dnaB*, *deaD* and *gbpA* genes from VPPCV08-7. The results showed that though VPRIMD2210633 being a pandemic strain could not possibly evolve into a newer pathogenic or virulent strain in the present scenario that is the Early Mortality Syndrome strain. On the other hand, the phylogenetic tree of all the genes considered together explain that the difference in the genomic content of the *V. parahaemolyticus* might be expanded worldwide with lesser chances to identify closer strains among countries based only upon 16S rRNA genes. The five different chitinase genes present in genome have also been aligned by ClustalW to plot the phylogeny (Appendix C (m)). The phylogeny patterns of chitinase 1 and 3 genes as shown in the Appendix C (m (i) & (iii)) show more similarity in the arrangements closer to the house-keeping genes *dnaA* (Figure 5.4.7) and *deaD* (Figure 5.4.8).

5.4.4 Host-Pathogen-Environment interaction studies

In the presence of $MgSO_4.7H_2O$, a uniform increase in the gene expression of *gbpA* was observed (Figures 5.4.10 and 5.4.11).

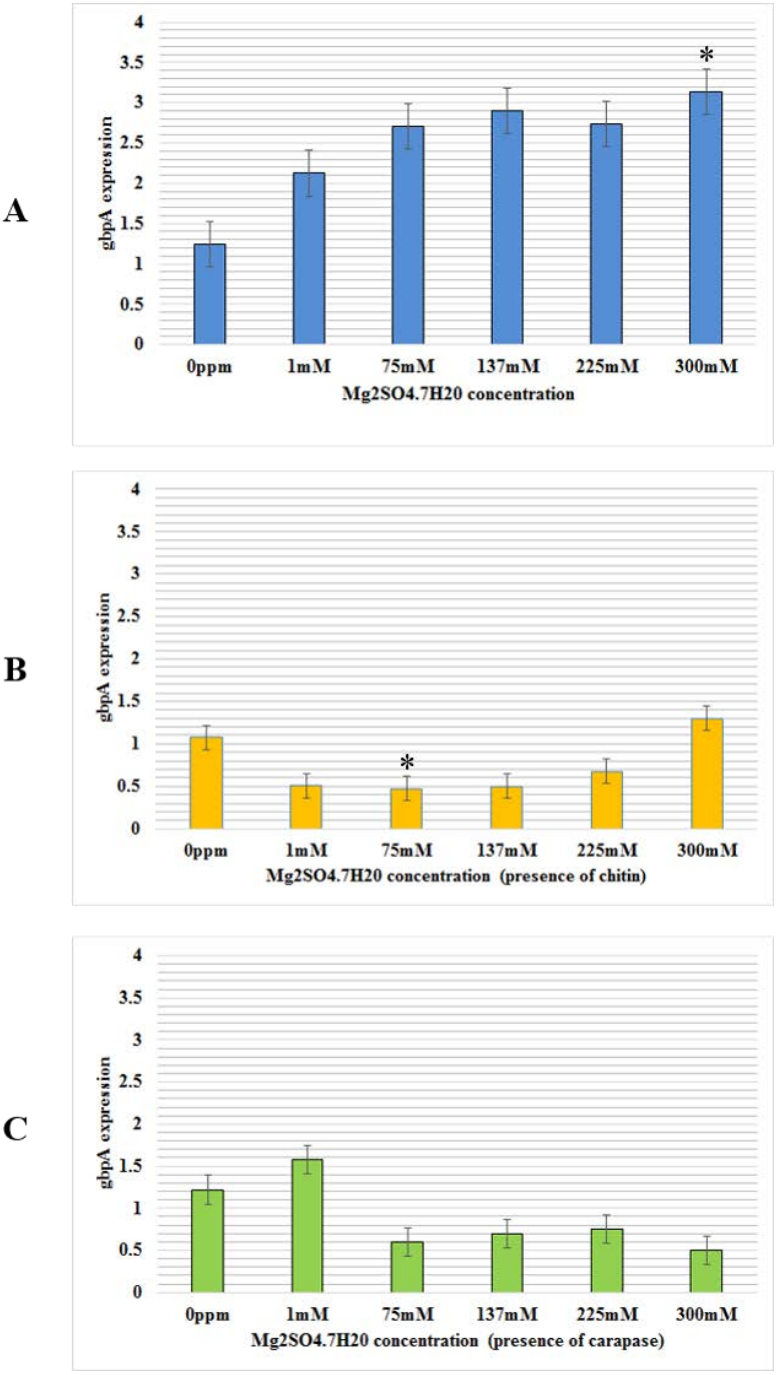


Figure 5.4.10: Graphs representing *gbpA* gene expression (2-delta delta Ct values) against (A) the concentrations of $MgSO_4.7H_2O$ (B) presence of chitin with $MgSO_4.7H_2O$ and (C) presence of carapace with $MgSO_4.7H_2O$

Increased gene expression can be observed in the presence of $MgSO_4.7H_2O$ in (A) while a decrease in gbpA gene expression can be observed in the presence of chitin (B) and carapace (C). The graphical representation has been made using the Livak method for real time PCR calculations. The statistical highly significant values ($p<0.05$) are represented with an asterisk.

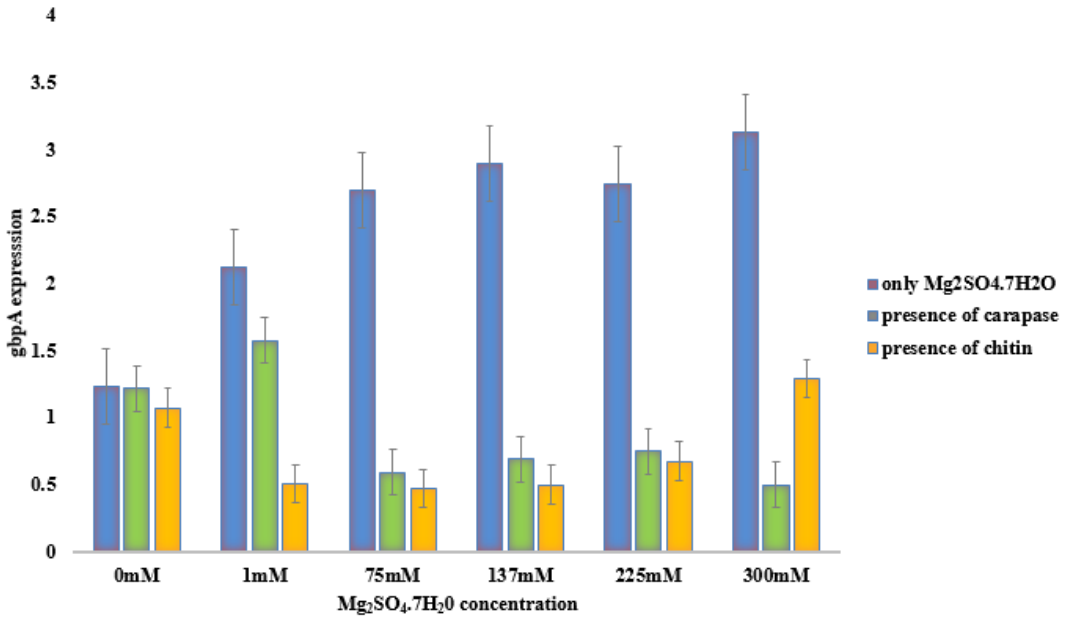


Figure 5.4.11: Graphs showing *gbpA* gene expression (2-delta delta Ct values) against the concentrations of Magnesium sulphate and with presence of chitin and carapace

Graph representing comparative *gbpA* expression among the three treatment groups of *V. parahaemolyticus* PCV08-7 culture. The graphical representation has been made using the Livak method for real time PCR calculations. The statistical highly significant values ($p<0.05$) are represented with an asterisk.

$MgSO_4.7H_2O$ with the presence of chitin showed a gradual increase in gene expression of *gbpA*. $MgSO_4.7H_2O$ with carapace showed an increase and then a slow decrease in the expression of the *gbpA* gene. In common to all the three experiments in the presence of 0 ppm $MgSO_4.7H_2O$, a standard level of expression was observed among all the three experiments. The results of the experiment checking the host-pathogen-environment interaction was significant in the *V. parahaemolyticus* culture treated with $MgSO_4.7H_2O$ only with a P-value of 0.0154 (Table 5.4.3) and also in the presence of $MgSO_4.7H_2O$ with chitin, P-value of 0.000441 (Table 5.4.4).

Table 5.4.3: Analysis of Variance to identify the level of significance for gene expression with the presence of $MgSO_4.7H_2O$

ANNOVA: Single Factor

Groups	Count	Sum	Average	Variance
0mM	3	3.7246	1.241533	0.559281
1mM	3	6.3717	2.1239	0.000633
75mM	3	8.1098	2.703267	0.056954
137mM	3	8.6902	2.896733	0.431421
225mM	3	8.2288	2.742933	1.154946
300mM	3	9.3924	3.1308	0.370459

ANNOVA

Source of Variation	SS	df	MS	F	P-value	F-crit
Between groups	7.129555	5	1.425911	3.324198	0.040941	3.105875
Within groups	5.147387	12	0.428949			
Total	12.27694	17				

Table 5.4.4: Analysis of Variance to identify the level of significance for gene expression with the presence of $MgSO_4.7H_2O$ and chitin

ANNOVA: Single Factor

Groups	Count	Sum	Average	Variance
0mM	3	3.23354	1.077847	0.012896
1mM	3	1.51932	0.50644	0.043052
75mM	3	1.43062	0.476873	0.004015
137mM	3	1.51015	0.503383	0.014588
225mM	3	2.03331	0.67777	0.074795
300mM	3	3.89621	1.298737	0.056636

ANNOVA

Source of Variation	SS	df	MS	F	P-value	F-crit
Between groups	1.824813	5	0.364963	10.63094	0.000442	3.105875
Within groups	0.411963	12	0.03433			
Total	2.236776	17				

There was no significant effect in treating *V. parahaemolyticus* culture with $MgSO_4.7H_2O$ with the presence of carapace (Table 5.4.5).

Table 5.4.5: Analysis of Variance to identify the level of significance for gene expression with the presence of $MgSO_4.7H_2O$ and carapace

ANNOVA: Single Factor

Groups	Count	Sum	Average	Variance
0mM	3	3.66068	1.220227	0.516649
1mM	3	4.73671	1.578903	0.15404
75mM	3	1.79222	0.597407	0.001811
137mM	3	2.08237	0.694123	0.002649
225mM	3	2.25824	0.752747	0.005014
300mM	3	1.50081	0.50027	0.023958

ANNOVA

Source of Variation	SS	df	MS	F	P-value	F-crit
Between groups	2.635025	5	0.527005	4.490744	0.0154	3.105875
Within groups	1.408243	12	0.117354			
Total	4.043268	17				

Concentrations of 300 mM (P-value = 0.02744) and 75 mM (P-value = 0.00132) were highly significant values in $MgSO_4.7H_2O$ treated and $MgSO_4.7H_2O$ with chitin treated sample respectively (Table 5.4.6).

Table 5.4.6: Student two tail paired t-test of equal variance across gene expression samples of samples treated with only $MgSO_4.7H_2O$, with $MgSO_4.7H_2O$ and chitin, with $MgSO_4.7H_2O$ and carapace

Concentrations of $MgSO_4.7H_2O$	Samples with $MgSO_4.7H_2O$	Samples with $MgSO_4.7H_2O$ + Chitin	Samples with $MgSO_4.7H_2O$ + carapase
1mM	0.11064	0.01387	0.49034
75mM	0.03212	0.00132	0.20845
137mM	0.045	0.00388	0.27471
225mM	0.11796	0.07937	0.32502
300mM	0.02744	0.22042	0.16513

The colony count of VPPCV08-7 in the presence of $MgSO_4.7H_2O$, $MgSO_4.7H_2O$ with chitin and $MgSO_4.7H_2O$ with carapace has been observed to show a significant decrease in fold change with increase in the treatment of concentrated $MgSO_4.7H_2O$ (Table 5.4.7).

Table 5.4.7: Table represents the bacterial count in the presence of only $MgSO_4.7H_2O$ (NONE). In the presence of $MgSO_4.7H_2O$ with chitin and in the presence of $MgSO_4.7H_2O$ with carapace (concentrations represented in ppm).

Table shows colony forming units at different concentrations of 0 ppm, 300 ppm, 18500 ppm, 34000 ppm, 55000 ppm and 73941 ppm in the presence of chitin (0.05 gm/3 ml) carapace (0.05 gm/3 ml) and in the presence of only $MgSO_4.7H_2O$

Concentrations of MgSO4.7H20	NONE	CHITIN (0.05gm/3ml)	CARAPASE (0.05gm/3ml
0ppm	148.845 x 10^3 cfu/ml	74 x 10^3 cfu/ml	137.94 x 10^3 cfu/ml
300ppm	117.6 x 10^3 cfu/ml	43.3 x 10^3 cfu/ml	103.89 x 10^3 cfu/ml
18500ppm	87.795 x 10^3 cfu/ml	17.9 x 10^3 cfu/ml	90.69 x 10^3 cfu/ml
34000ppm	60 x 10^3 cfu/ml	15.9 x 10^3 cfu/ml	89.04 x 10^3 cfu/ml
55500ppm	57.9 x 10^3 cfu/ml	5.6 x 10^3 cfu/ml	24 x 10^3 cfu/ml
73941ppm	56.55 x 10^3 cfu/ml	3.7 x 10^3 cfu/ml	14.94 x 10^3 cfu/ml

Hence, this explains that in the presence of chitin and carapace, the levels of *gbpA* expression has been lowered significantly when compared to the absence of chitin and carapace. This shows that the *gbpA* expression has been regulated both by the host component (chitin/ carapace) as well as by the presence of an external source magnesium.

5.5 Conclusion

The TAH pipeline shows comparatively good usage of percentage of reads for assembly. The need to get all the sequencing read data into a similar format had been important to process the next annotation process. This shows that any sequencing data has to be dealt with post processing to one common format of FASTQ file. The presence of more number of contigs can help in identifying unique sequences as less number of contigs could increase the probability of occurrence of a gene inside a huge contig. This could interfere with the probable alignments against genes on databases such a Genbank. The usage of two assembly tools Velvet and Edena could give a better result as two different algorithms based on de-bruijn graphs and overlaps were used. This procedure could probably help in filling the gaps between contigs which has not been covered by the other assembly tool. Our genomes do not show any similarity to EMS strain in terms of its specific sequences but, shows similarity to VPRIMD2210633 and VPO1:K33.

The bioinformatics analysis on the *gbpA* gene shows probable interactions with chitin which is an important component of the outer carapace content of the prawn or shrimp. The 16S rRNA has been used to categorize bacteria into the niche specific regions worldwide. This has been critical in the case of *V. parahaemolyticus* as it is an aquatic organism which has a probability of being exported and imported across the world with its presence in the prawn/ shrimp. Hence, it could be of utmost importance in such conditions to trace back the similarity among the strains based on the basic machinery while utilizing 16S rRNA and the functional gene of interest as the reference in any phylogenetic analysis. Therefore, the importance of utilizing genes such as *dnaB*, *deaD* cannot be ruled out from phylogenetic analysis as these could probably help us understand the ancestral strains of the current *V. parahaemolyticus* strain. Meanwhile, in comparison to these basic machinery understanding the evolution of individual genes

such as *gbpA* and chitinase in the current scenario which are related to the interactions with the host is significant. This significance is in terms of extracting information of the evolutionary trend followed by these host related genes within the pathogen to overlay and understand its relationship with the already existing patterns of *V. parahaemolyticus* principle gene features through phylogeny.

The *in vitro* experiment on the *gbpA* expression of *V. parahaemolyticus* in the presence of *M. rosenbergii* component (chitin/ carapace) and presence of an external environment (Magnesium) has shown how *gbpA* gene expression has been regulated. Our results show no definite change in levels of *gbpA* expression in the absence of magnesium as well as absence of chitin and carapace but, showed an increase in *gbpA* expression of magnesium and decreased *gbpA* expression in presence of both magnesium as well as with chitin and carapace. Hence this explains the importance of the presence of an environment (Magnesium) in regulating the gene expression of the bacterial *gbpA*. The patterns of gene expression of *gbpA* observed could possibly help understand the roles of magnesium as the environment and the chitin as the component of the host which are indeed necessary for a pathogenic gene to respond. In reference to molting phase of prawn wherein thick or mature carapace of the prawn is lost and the new carapace replaces as a thin chitinous layer, our study could support that an increased *gbpA* expression could be observed when compared to the stage in the presence of the mature carapace before molting. Hence, this supports the fact that it could be probably during the molting stage that the pathogen shows higher *gbpA* expression. This could further be studied with introduction of various pathogen related genes and carried ahead for research on possible interaction studies between *gbpA* and chitin (pathogen and host) with magnesium as one of the very important component in the environment.

Achievement

The genome of VPPCV08-7 was once again assembled using Ion proton and Illumina raw read data successfully applying our TAH pipeline from objective 1. 52genes of *Vibrio parahaemolyticus* were downloaded as Sequence Read Archives (SRA) from Genbank, assembled preceded by phylogenetic analysis. Phylogenetic analysis was performed with house-keeping genes in view and 16srRNA and gbpA gene. Real time PCR was performed for gbpA gene expression and successfully an expression pattern was observed in the presence and absence of $Mg_2SO_4.7H_2O$, chitin and carapace.

CHAPTER 6

CONCLUSION

A defined set of bacterial next-generation sequencing raw read data was used in the current research for establishing an accurate quality control pipeline which was successfully achieved. Followed by the usage of retrieved sequencing data of various *V. parahaemolyticus* genomes for analysis of the sequenced *V. parahaemolyticus* PCV08-7 genome. This was further targeted at studying not only the host-pathogen-environment interactions but, even the phylogenetic relationships with the co-evolved strains in terms of house-keeping genes and basic machinery genes. These ultimate goals are summarized below in order of the successfully achieved objectives and challenging innovative goals for further research.

"Trimming" has shown to retain maximum percentage of reads for further assembly processing among high quality filtering, adaptor filtering, N's removal and trimming strategies in quality control (QC) of next-generation sequencing of bacterial data. Keeping this "Trimming" strategy as the first step of QC pipeline in common among the five different bacterial genomes showed a constancy in maintaining the total percentage of used reads. Apart from trimming, strategies of high quality filtering and adaptor filtering among these genomes was observed to be quite important to remove artefacts from sequencing raw read data. Hence, the QC pipeline has been assembled with "trimming – adaptor filtering – high quality filtering" as a good application to the QC flow to refine the output data which further will be used in bioinformatic analysis. Our pipeline named TAH pipeline includes the usage of pre-existing QC control tools by improvising and tweaking their parameters in order to make the best use of and produce as output the best QC filtered data for more reliable and accurate NGS genome analysis.

Finishing a draft genome of bacteria is quite necessary as it is important to keep the genome information intact. Dealing with a genome with two chromosomes is quite tedious with assembled sequencing data being a challenge lying in separating the genome data into two chromosomes. The challenge is two sided: One, to separate data into two chromosomes based on a reference genome already present in Genbank. Two, is working with data which belongs to a *de-novo* assembled genome. In the present work we dealt with the availability of a reference genome VPRIMD2210633 but, for this we have completely dealt unmatched sequences between both chromosomes as separate sequences during the process of annotation. Our study mentions that dealing with bacterial genomes needs utmost support and evidence through completing the bacterial genome with amplifications by PCR reactions. Inverse PCR could possibly be one such strategy which can be used on rRNA's which exist in multiple copies in bacteria to fill up the gaps between the assembled contigs. Applying a similar procedure to tRNA's could be an added advantage.

In addition to this, we have worked on the current genome *V. parahaemolyticus* PCV08-7 and have worked on assembly-assembly (Overlap-*De-bruijn*) hybrid approach with two different assembly (Edena and Velvet) tools which does indeed explain the importance of the generation of more number of contigs in comparison to larger contigs in the assembly process. Herein, we have shown that this strategy helps the user to identify few important gene features of the bacteria as they exist separately within short contigs, while their existence in larger contigs blurs the identification of short genes which could be of significant importance in microbiology research.

Apart from this, we have used VPPCV08-7 to focus on using one single gene for phylogenetic analysis with reference to the house-keeping genes as well as 16S rRNA which could give more in depth information about the existence and evolutionary trace back of bacterial strains. This can be explained as that the basic machinery genes

are the preliminary processed genes of any bacterial genome. Avoiding these genes of bacteria could be critical in our understanding of a bacterial pathogen, especially to identify the evolution of a certain virulence trait of a bacteria. Hence, our work shows how distantly a phylogenetic relationship exists among genomes when compared between pathogenic gene traits (*gbpA*, chitinase) and basic machinery (*deaD*, *dnaB*) and 16S rRNA genes. Considering these genes together could help understand bacterial genomes with in-depth understanding towards evolutionary emergence of bacterial genomes based on location or region of origin. The application of a *gbpA* gene to study its response in the presence of a host component (chitin/ carapace) and environmental factor (magnesium) with an *in vitro* experiment showed how the gene was regulated in their presence and absence. It even could probably support the fact of higher expression of pathogenic gene *gbpA* during the sensitive phase of prawn during molting with a declined chitinous layer of its carapace and *vice versa*. This data could help us understand host-pathogen-environment play a role together and cannot be neglected in order to further studies in understanding the mechanisms of interactions between the host-pathogen components of *V. parahaemolyticus* in *M. rosenbergii*. Along with this our research shows how important is an external environment factor (Magnesium) is in regulating gene expression of the pathogen in accordance with the presence of host component.

Our research on the whole deals with data that supports further work on making one markedly significant tool based on the strategies applied to the QC pipeline. The current work could help gradually lessen the difficulty in dealing with erroneous raw read sequencing data, to deal with bacterial genome assembly hybrid data as well as concentrate on rRNA's and tRNA's for completion of a draft bacterial genome. Apart from this, further work could be held on dealing with the *M. rosenbergii* itself and

checking its infection with *V. parahaemolyticus* with the source of magnesium as an external environment *in-vitro*. This could help us understand possible roles of interactions between the host and the pathogen and the probable phase of entry of the pathogen into the host.

INNOVATION

1. Establishing the importance of trimming short reads in bacterial genome datasets generated by next-generation sequencing technique.
2. Developing and evaluating several novel next-generation sequencing data analysis pipelines with TAH pipeline identified as best suited to the problem at hand.
3. Successfully applying the TAH pipeline to the VPPCV08-7 bacterial draft genome.
4. The data analysis performed was used to identify a single gene to track the host-pathogen response under different environmental conditions.

SINGIFICANCE

The findings reported in the current research thesis are significant, with the defined objectives set out being achieved. More importantly, mechanism by which the bacterium infects prawns has been proposed, which is essential for developing strategies to prevent infection by *V. parahaemolyticus*.

REFERENCES

Akio, K., Teshima, S., & Sasaki, M. (1984). Requirements of the juvenile prawn for calcium, phosphorous, magnesium, potassium, copper, manganese and iron. *Mem Fac Fish, 33*(1), 63 - 71.

Alkan, C., Sajjadian, S., & Eichler, E. E. (2011). Limitations of next-generation genome sequence assembly. *Nat Methods, 8*(1), 61-65. doi:10.1038/nmeth.1527

Andam, C. P., Williams, D., & Gogarten, J. P. (2010). Natural taxonomy in light of horizontal gene transfer. *Biology & Philosophy, 25*(4), 589-602. doi:DOI 10.1007/s10539-010-9212-8

Andrews, S. C. (1998). Iron storage in bacteria. *Adv Microb Physiol, 40*, 281-351.

Aravind, L., Watanabe, H., Lipman, D. J., & Koonin, E. V. (2000). Lineage-specific loss and divergence of functionally linked genes in eukaryotes. *Proc Natl Acad Sci U S A, 97*(21), 11319-11324. doi:10.1073/pnas.200346997

Avasthi, T. S., & Ahmed, N. (2011). Helicobacter pylori and type 1 diabetes mellitus: possibility of modifying chronic disease susceptibility with vaccinomics at the anvil. *OMICS, 15*(9), 589-596. doi:10.1089/omi.2010.0138

Avasthi, T. S., Devi, S. H., Taylor, T. D., Kumar, N., Baddam, R., Kondo, S., . . . Ahmed, N. (2011). Genomes of two chronological isolates (Helicobacter pylori 2017 and 2018) of the West African Helicobacter pylori strain 908 obtained from a single patient. *J Bacteriol, 193*(13), 3385-3386. doi:10.1128/JB.05006-11

Avasthi, T. S., Kumar, N., Baddam, R., Hussain, A., Nandanwar, N., Jadhav, S., & Ahmed, N. (2011). Genome of multidrug-resistant uropathogenic Escherichia coli strain NA114 from India. *J Bacteriol, 193*(16), 4272-4273. doi:10.1128/JB.05413-11

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., . . . Zagnitko, O. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics, 9*, 75. doi:10.1186/1471-2164-9-75

Baddam, R., Kumar, N., Shaik, S., Suma, T., Ngoi, S. T., Thong, K. L., & Ahmed, N. (2012). Genome sequencing and analysis of Salmonella enterica serovar Typhi strain CR0063 representing a carrier individual during an outbreak of typhoid fever in Kelantan, Malaysia. *Gut Pathog, 4*(1), 20. doi:10.1186/1757-4749-4-20

Baddam, R., Thong, K. L., Avasthi, T. S., Shaik, S., Yap, K. P., Teh, C. S., . . . Ahmed, N. (2012). Whole-genome sequences and comparative genomics of Salmonella

enterica serovar Typhi isolates from patients with fatal and nonfatal typhoid fever in Papua New Guinea. *J Bacteriol, 194*(18), 5122-5123. doi:10.1128/JB.01051-12

Bag, P. K., Nandi, S., Bhadra, R. K., Ramamurthy, T., Bhattacharya, S. K., Nishibuchi, M., . . . Nair, G. B. (1999). Clonal diversity among recently emerged strains of Vibrio parahaemolyticus O3:K6 associated with pandemic spread. *J Clin Microbiol, 37*(7), 2354-2357.

Baker, M. (2010). Next-generation sequencing: adjusting to data overload. *Nat Methods, 7*(7), 495-499.

Bapteste, E., O'Malley, M. A., Beiko, R. G., Ereshefsky, M., Gogarten, J. P., Franklin-Hall, L., . . . Martin, W. (2009). Prokaryotic evolution and the tree of life are two different things. *Biol Direct, 4*, 34. doi:10.1186/1745-6150-4-34

Barbosa, E. G., Aburjaile, F. F., Ramos, R. T., Carneiro, A. R., Le Loir, Y., Baumbach, J., . . . Azevedo, V. (2014). Value of a newly sequenced bacterial genome. *World J Biol Chem, 5*(2), 161-168. doi:10.4331/wjbc.v5.i2.161

Barrow, G. I., & Miller, D. C. (1974a). Genus 1. Vibrio. In: Bergey's Manual of Systematic Bacteriology *Williams and Wilkins, Baltimore, pp., Volume 1.*, 205-210.

Barrow, G. I., & Miller, D. C. (1974b). Growth studies on Vibrio parahaemolyticusin relation to pathogenicity. *International Symposium of Vibrio parahaemolyticus. Saikon, Tokyo, pp*, 205-210.

Becker, B., Hoef-Emden, K., & Melkonian, M. (2008). Chlamydial genes shed light on the evolution of photoautotrophic eukaryotes. *BMC Evol Biol, 8*, 203. doi:10.1186/1471-2148-8-203

Besemer, J., Lomsadze, A., & Borodovsky, M. (2001). GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res, 29*(12), 2607-2618.

Besser, R. E., Feikin, D. R., Eberhart-Phillips, J. E., Mascola, L., & Griffin, P. M. (1994). Diagnosis and treatment of cholera in the United States. Are we prepared? *JAMA, 272*(15), 1203-1205.

Bhattacharya, M., Roy, S. S., Biswas, D., & Kumar, R. (2000). Effect of Mg(2+) ion in protein secretion by magnesium-resistant strains of Pseudomonas aeruginosa and Vibrio parahaemolyticus isolated from the coastal water of Haldia port. *FEMS Microbiol Lett, 185*(2), 151-156.

Bhowmick, R., Ghosal, A., Das, B., Koley, H., Saha, D. R., Ganguly, S., . . . Chatterjee, N. S. (2008). Intestinal adherence of Vibrio cholerae involves a coordinated interaction between colonization factor GbpA and mucin. *Infect Immun, 76*(11), 4968-4977. doi:10.1128/IAI.01615-07

Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., & Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics, 27*(4), 578-579. doi:10.1093/bioinformatics/btq683

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114-2120. doi:10.1093/bioinformatics/btu170

Boyd, E. F., Cohen, A. L., Naughton, L. M., Ussery, D. W., Binnewies, T. T., Stine, O. C., & Parent, M. A. (2008). Molecular analysis of the emergence of pandemic Vibrio parahaemolyticus. *BMC Microbiol, 8*, 110. doi:10.1186/1471-2180-8-110

Bruccoleri, R. E., Dougherty, T. J., & Davison, D. B. (1998). Concordance analysis of microbial genomes. *Nucleic Acids Res, 26*(19), 4482-4486.

Burland, T. G. (2000). DNASTAR's Lasergene sequence analysis software. *Methods Mol Biol, 132*, 71-91.

Cahill, M. J., Koser, C. U., Ross, N. E., & Archer, J. A. (2010). Read length and repeat resolution: exploring prokaryote genomes using next-generation sequencing technologies. *PLoS One, 5*(7), e11518. doi:10.1371/journal.pone.0011518

Carlucci, A. F., & Pramer, D. (1959). Factors affecting the survival of bacteria in sea water. *Appl Microbiol, 7*, 388-392.

Centers for Disease Control. Vaccines for selected use in international travel: cholera vaccine. (1978). *MMWR, 27*, 173-174.

Chen, Y., Stine, O. C., Badger, J. H., Gil, A. I., Nair, G. B., Nishibuchi, M., & Fouts, D. E. (2011). Comparative genomic analysis of Vibrio parahaemolyticus: serotype conversion and virulence. *BMC Genomics, 12*, 294. doi:10.1186/1471-2164-12-294

Chen, Y. C., Liu, T., Yu, C. H., Chiang, T. Y., & Hwang, C. C. (2013). Effects of GC bias in next-generation-sequencing data on de novo genome assembly. *PLoS One, 8*(4), e62856. doi:10.1371/journal.pone.0062856

Chibana, H., Oka, N., Nakayama, H., Aoyama, T., Magee, B. B., Magee, P. T., & Mikami, Y. (2005). Sequence finishing and gene mapping for Candida albicans

chromosome 7 and syntenic analysis against the Saccharomyces cerevisiae genome. *Genetics, 170*(4), 1525-1537. doi:10.1534/genetics.104.034652

Chowdhury, N. R., Chakraborty, S., Ramamurthy, T., Nishibuchi, M., Yamasaki, S., Takeda, Y., & Nair, G. B. (2000). Molecular evidence of clonal Vibrio parahaemolyticus pandemic strains. *Emerg Infect Dis, 6*(6), 631-636. doi:10.3201/eid0606.000612

Chowdhury, N. R., Stine, O. C., Morris, J. G., & Nair, G. B. (2004). Assessment of evolution of pandemic Vibrio parahaemolyticus by multilocus sequence typing. *J Clin Microbiol, 42*(3), 1280-1282.

Clark, J. W. H., & Lynn, J. W. (1977). A Mg ++ dependent cortical reaction in the eggs of Penaeid shrimp. *J Exp Zool, 200*, 177 - 183.

Colwell, R. R., West, P. A., Maneval, D., Remmers, E. F., Elliot, E. L., & Carlson, N. E. (1984). Ecology of pathogenic vibrios in Chesapeake Bay. In: Vibrios in the Environment. *Wiley, New York, pp.*, 367-387.

Cox, M. P., Peterson, D. A., & Biggs, P. J. (2010). SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics, 11*, 485. doi:10.1186/1471-2105-11-485

Craun, M. F., Craun, G. F., Calderon, R. L., & Beach, M. J. (2006). Waterborne outbreaks reported in the United States. *J Water Health, 4 Suppl 2*, 19-30.

D'Abramo, L. R., & Brunson, M. W. (1996). Biology and Life History of freshwater prawns. *SRAC publications, No: 483*.

Dadisman, T. A., Jr., Nelson, R., Molenda, J. R., & Garber, H. J. (1972). Vibrio parahaemolyticus gastroenteritis in Maryland. I. Clinical and epidemiologic aspects. *Am J Epidemiol, 96*(6), 414-426.

Dai, M., Thompson, R. C., Maher, C., Contreras-Galindo, R., Kaplan, M. H., Markovitz, D. M., . . . Meng, F. (2010). NGSQC: cross-platform quality analysis pipeline for deep sequencing data. *BMC Genomics, 11 Suppl 4*, S7. doi:10.1186/1471-2164-11-S4-S7

Damrongphol, P., Jaroensastraraks, P., & Poolsanguan, B. (2001). Effect of various medium compositions on survival and hatching rates of embryos of the giant freshwater prawn Macrobrachium rosenbergii cultured in vitro. *Fisheries Science, 67*(1), 64-70. doi:DOI 10.1046/j.1444-2906.2001.00200.x

Daniels, N. A., MacKinnon, L., Bishop, R., Altekruse, S., Ray, B., Hammond, R. M., . . . Slutsker, L. (2000). Vibrio parahaemolyticus infections in the United States, 1973-1998. *J Infect Dis, 181*(5), 1661-1666. doi:10.1086/315459

Daniels, N. A., Ray, B., Easton, A., Marano, N., Kahn, E., McShan, A. L., 2nd, . . . Angulo, F. J. (2000). Emergence of a new Vibrio parahaemolyticus serotype in raw oysters: A prevention quandary. *JAMA, 284*(12), 1541-1545.

Darling, A. E., Mau, B., & Perna, N. T. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One, 5*(6), e11147. doi:10.1371/journal.pone.0011147

Del Fabbro, C., Scalabrin, S., Morgante, M., & Giorgi, F. M. (2013). An extensive evaluation of read trimming effects on Illumina NGS data analysis. *PLoS One, 8*(12), e85024. doi:10.1371/journal.pone.0085024

Delcher, A. L., Harmon, D., Kasif, S., White, O., & Salzberg, S. L. (1999). Improved microbial gene identification with GLIMMER. *Nucleic Acids Res, 27*(23), 4636-4641.

DePaola, A., Kaysner, C. A., Bowers, J., & Cook, D. W. (2000). Environmental investigations of Vibrio parahaemolyticus in oysters after outbreaks in Washington, Texas, and New York (1997 and 1998). *Appl Environ Microbiol, 66*(11), 4649-4654.

DePaola, A., Nordstrom, J. L., Dalsgaard, A., Forslund, A., Oliver, J., Bates, T., . . . Gulig, P. A. (2003). Analysis of Vibrio vulnificus from market oysters and septicemia cases for virulence markers. *Appl Environ Microbiol, 69*(7), 4006-4011.

Devi, S. H., Taylor, T. D., Avasthi, T. S., Kondo, S., Suzuki, Y., Megraud, F., & Ahmed, N. (2010). Genome of Helicobacter pylori strain 908. *J Bacteriol, 192*(24), 6488-6489. doi:10.1128/JB.01110-10

Dodt, M., Roehr, J. T., Ahmed, R., & Dieterich, C. (2012). FLEXBAR-Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms. *Biology (Basel), 1*(3), 895-905. doi:10.3390/biology1030895

Egan, E. S., & Waldor, M. K. (2003). Distinct replication requirements for the two Vibrio cholerae chromosomes. *Cell, 114*(4), 521-530.

El-Metwally, S., Hamza, T., Zakaria, M., & Helmy, M. (2013). Next-generation sequence assembly: four stages of data processing and computational challenges. *PLoS Comput Biol, 9*(12), e1003345. doi:10.1371/journal.pcbi.1003345

Faruque, S. M., Islam, M. J., Ahmad, Q. S., Faruque, A. S., Sack, D. A., Nair, G. B., & Mekalanos, J. J. (2005). Self-limiting nature of seasonal cholera epidemics: Role of host-mediated amplification of phage. *Proc Natl Acad Sci U S A, 102*(17), 6119-6124. doi:10.1073/pnas.0502069102

Finelli, L., Swerdlow, D., Mertz, K., Ragazzoni, H., & Spitalny, K. (1992). Outbreak of cholera associated with crab brought from an area with epidemic disease. *J Infect Dis, 166*(6), 1433-1435.

Frederiksen, R. F., Paspaliari, D. K., Larsen, T., Storgaard, B. G., Larsen, M. H., Ingmer, H., . . . Leisner, J. J. (2013). Bacterial chitinases and chitin-binding proteins as virulence factors. *Microbiology, 159*(Pt 5), 833-847. doi:10.1099/mic.0.051839-0

Gaasterland, T., & Ragan, M. A. (1998). Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes. *Microb Comp Genomics, 3*(4), 199-217.

Gonzalez-Escalona, N., Martinez-Urtaza, J., Romero, J., Espejo, R. T., Jaykus, L. A., & DePaola, A. (2008). Determination of molecular phylogenetics of Vibrio parahaemolyticus strains by multilocus sequence typing. *J Bacteriol, 190*(8), 2831-2840. doi:10.1128/JB.01808-07

Guerout, A. M., Iqbal, N., Mine, N., Ducos-Galand, M., Van Melderen, L., & Mazel, D. (2013). Characterization of the phd-doc and ccd Toxin-Antitoxin Cassettes from Vibrio Superintegrons. *J Bacteriol, 195*(10), 2270-2283. doi:10.1128/JB.01389-12

Guidelines for national human immunodeficiency virus case surveillance, including monitoring for human immunodeficiency virus infection and acquired immunodeficiency syndrome. Centers for Disease Control and Prevention. (1999). *MMWR Recomm Rep, 48*(RR-13), 1-27, 29-31.

Gulig, P. A., Bourdage, K. L., & Starks, A. M. (2005). Molecular Pathogenesis of Vibrio vulnificus. *J Microbiol, 43 Spec No*, 118-131.

Guo, Y., Ye, F., Sheng, Q., Clark, T., & Samuels, D. C. (2014). Three-stage quality control strategies for DNA re-sequencing data. *Brief Bioinform, 15*(6), 879-889. doi:10.1093/bib/bbt069

Hall, N. (2007). Advanced sequencing technologies and their wider impact in microbiology. *J Exp Biol, 210*(Pt 9), 1518-1525. doi:10.1242/jeb.001370

Hameed, A. S. S., Rahaman, K. H., Alagan, A., & Yoganandhan, K. (2003). Antibiotic resistance in bacteria isolated from hatchery-reared larvae and post-larvae of Macrobrachium rosenbergii. *Aquaculture, 217*(1-4), 39-48.

Han, H., Wong, H. C., Kan, B., Guo, Z., Zeng, X., Yin, S., . . . Zhou, D. (2008). Genome plasticity of Vibrio parahaemolyticus: microevolution of the 'pandemic group'. *BMC Genomics, 9*, 570. doi:10.1186/1471-2164-9-570

Han, J. E., Tang, K. F., Tran, L. H., & Lightner, D. V. (2015). Photorhabdus insect-related (Pir) toxin-like genes in a plasmid of Vibrio parahaemolyticus, the causative agent of acute hepatopancreatic necrosis disease (AHPND) of shrimp. *Dis Aquat Organ, 113*(1), 33-40. doi:10.3354/dao02830

Hangsapreurke, K., Thamrongnawasawat, T., Powtongsook, S., Tabthipwon, P., Lumubol, P., & Pratoomchat, B. (2008). Embryonic development, hatching, mineral consumption, and survival of Macrobrachium rosenbergii (de Man)reared in artificial seawater in closed recirculating water system at different levels of salinity. *Mj Int J Sci Tech, 2*(3), 471 - 482.

Hayat, U., Reddy, G. P., Bush, C. A., Johnson, J. A., Wright, A. C., & Morris, J. G., Jr. (1993). Capsular types of Vibrio vulnificus: an analysis of strains from clinical and environmental sources. *J Infect Dis, 168*(3), 758-762.

Heinis, J. J., Beuchat, L. R., & Boswell, F. C. (1978). Antimetabolite sensitivity and magnesium uptake by thermally stressed Vibrio parahaemolyticus. *Appl Environ Microbiol, 35*(6), 1035-1040.

Hernandez, D., Francois, P., Farinelli, L., Osteras, M., & Schrenzel, J. (2008). De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Res, 18*(5), 802-809. doi:10.1101/gr.072033.107

Hoashi, K., Ogata, K., Taniguchi, H., Yamashita, H., Tsuji, K., Mizuguchi, Y., & Ohtomo, N. (1990). Pathogenesis of Vibrio parahaemolyticus: intraperitoneal and orogastric challenge experiments in mice. *Microbiol Immunol, 34*(4), 355-366.

Hollis, D. G., Weaver, R. E., Baker, C. N., & Thornsberry, C. (1976). Halophilic Vibrio species isolated from blood cultures. *J Clin Microbiol, 3*(4), 425-431.

Horn, M., Collingro, A., Schmitz-Esser, S., Beier, C. L., Purkhold, U., Fartmann, B., . . . Wagner, M. (2004). Illuminating the evolutionary history of chlamydiae. *Science, 304*(5671), 728-730. doi:10.1126/science.1096330

Huang, J., & Gogarten, J. P. (2007). Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? *Genome Biol, 8*(6), R99. doi:10.1186/gb-2007-8-6-r99

Hurley, C. C., Quirke, A., Reen, F. J., & Boyd, E. F. (2006). Four genomic islands that mark post-1995 pandemic Vibrio parahaemolyticus isolates. *BMC Genomics, 7*, 104. doi:10.1186/1471-2164-7-104

Huson, D. H. (1998). SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics, 14*(1), 68-73.

Huynen, M. A., & Snel, B. (2000). Gene and context: integrative approaches to genome analysis. *Adv Protein Chem, 54*, 345-379.

Igbinosa, E. O., & Okoh, A. I. (2008). Emerging Vibrio species: an unending threat to public health in developing countries. *Res Microbiol, 159*(7-8), 495-506. doi:10.1016/j.resmic.2008.07.001

Iida, T., Tang, G. Q., Suttikulpitug, S., Yamamoto, K., Miwatani, T., & Honda, T. (1995). Isolation of mutant toxins of Vibrio parahaemolyticus hemolysin by in vitro mutagenesis. *Toxicon, 33*(2), 209-216.

Jackson, H. (1974). Temperature relationships of Vibrio parahaemolyticus. In: International Symposium of Vibrio parahaemolyticus. *Saikon, Tokyo, pp.*, 139-145.

Janda, J. M., & Abbott, S. L. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *J Clin Microbiol, 45*(9), 2761-2764. doi:10.1128/JCM.01228-07

Jiang, J., Li, J., Kwan, H. S., Au, C. H., Wan Law, P. T., Li, L., . . . Leung, F. C. (2012). A cost-effective and universal strategy for complete prokaryotic genomic sequencing proposed by computer simulation. *BMC Res Notes, 5*, 80. doi:10.1186/1756-0500-5-80

Jude, B. A., Martinez, R. M., Skorupski, K., & Taylor, R. K. (2009). Levels of the secreted Vibrio cholerae attachment factor GbpA are modulated by quorum-sensing-induced proteolysis. *J Bacteriol, 191*(22), 6911-6917. doi:10.1128/JB.00747-09

Kaysner, C. A., Abeyta, C., Trost, P. A., Wetherington, J. H., Jinneman, K. C., Hill, W. E., & Wekell, M. M. (1994). Urea Hydrolysis Can Predict the Potential Pathogenicity of Vibrio-Parahaemolyticus Strains Isolated in the Pacific-Northwest. *Appl Environ Microbiol, 60*(8), 3020-3022.

Kircher, M., & Kelso, J. (2010). High-throughput DNA sequencing--concepts and limitations. *Bioessays, 32*(6), 524-536. doi:10.1002/bies.200900181

Klontz, K. C., Tauxe, R. V., Cook, W. L., Riley, W. H., & Wachsmuth, I. K. (1987). Cholera after the Consumption of Raw Oysters. *Annals of Internal Medicine, 107*(6), 846-848.

Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., . . . Adam, M. P. (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat Biotechnol, 30*(7), 693-700. doi:10.1038/nbt.2280

Kuhn, M., von Mering, C., Campillos, M., Jensen, L. J., & Bork, P. (2008). STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res, 36*(Database issue), D684-688. doi:10.1093/nar/gkm795

Kumar, N., Mukhopadhyay, A. K., Patra, R., De, R., Baddam, R., Shaik, S., . . . Ahmed, N. (2012). Next-generation sequencing and de novo assembly, genome organization, and comparative genomic analyses of the genomes of two Helicobacter pylori isolates from duodenal ulcer patients in India. *J Bacteriol, 194*(21), 5963-5964. doi:10.1128/JB.01371-12

Kyrpides, N., Overbeek, R., & Ouzounis, C. (1999). Universal protein families and the functional content of the Last Universal Common Ancestor. *Journal of Molecular Evolution, 49*(4), 413-423. doi:Doi 10.1007/Pl00006564

Lagesen, K., Hallin, P., Rodland, E. A., Staerfeldt, H. H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res, 35*(9), 3100-3108. doi:10.1093/nar/gkm160

Larsen, T. S., & Krogh, A. (2003). EasyGene--a prokaryotic gene finder that ranks ORFs by statistical significance. *BMC Bioinformatics, 4*, 21. doi:10.1186/1471-2105-4-21

Lassmann, T., Hayashizaki, Y., & Daub, C. O. (2009). TagDust--a program to eliminate artifacts from next generation sequencing data. *Bioinformatics, 25*(21), 2839-2840. doi:10.1093/bioinformatics/btp527

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics, 26*(5), 589-595. doi:10.1093/bioinformatics/btp698

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and

SAMtools. *Bioinformatics, 25*(16), 2078-2079. doi:10.1093/bioinformatics/btp352

Loman, N. J., Constantinidou, C., Chan, J. Z., Halachev, M., Sergeant, M., Penn, C. W., . . . Pallen, M. J. (2012). High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nat Rev Microbiol, 10*(9), 599-606. doi:10.1038/nrmicro2850

Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J., & Pallen, M. J. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol, 30*(5), 434-439. doi:10.1038/nbt.2198

Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res, 25*(5), 955-964.

Lowry, P. W., Pavia, A. T., Mcfarland, L. M., Peltier, B. H., Barrett, T. J., Bradford, H. B., . . . Blake, P. A. (1989). Cholera in Louisiana - Widening Spectrum of Seafood Vehicles. *Archives of Internal Medicine, 149*(9), 2079-2084. doi:DOI 10.1001/archinte.149.9.2079

MacLean, D., Jones, J. D., & Studholme, D. J. (2009). Application of 'next-generation' sequencing technologies to microbial genetics. *Nat Rev Microbiol, 7*(4), 287-296. doi:10.1038/nrmicro2122

Maitra, R. D., Kim, J., & Dunbar, W. B. (2012). Recent advances in nanopore sequencing. *Electrophoresis, 33*(23), 3418-3428. doi:10.1002/elps.201200272

Majid, M., Kumar, N., Qureshi, A., Yerra, P., Kumar, A., Kumar, M. K., . . . Ahmed, N. (2014). Genomes of Two Clinical Isolates of Mycobacterium tuberculosis from Odisha, India. *Genome Announc, 2*(2). doi:10.1128/genomeA.00199-14

Makino, K., Oshima, K., Kurokawa, K., Yokoyama, K., Uda, T., Tagomori, K., . . . Iida, T. (2003). Genome sequence of Vibrio parahaemolyticus: a pathogenic mechanism distinct from that of V cholerae. *Lancet, 361*(9359), 743-749. doi:10.1016/S0140-6736(03)12659-1

Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends Genet, 24*(3), 133-141. doi:10.1016/j.tig.2007.12.007

McKinley, J. E., & Magnuson, R. D. (2005). Characterization of the Phd repressor-antitoxin boundary. *J Bacteriol, 187*(2), 765-770. doi:10.1128/JB.187.2.765-770.2005

Meador, C. E., Parsons, M. M., Bopp, C. A., Gerner-Smidt, P., Painter, J. A., & Vora, G. J. (2007). Virulence gene- and pandemic group-specific marker profiling of clinical Vibrio parahaemolyticus isolates. *J Clin Microbiol, 45*(4), 1133-1139. doi:10.1128/JCM.00042-07

Mellmann, A., Harmsen, D., Cummings, C. A., Zentz, E. B., Leopold, S. R., Rico, A., . . . Karch, H. (2011). Prospective genomic characterization of the German enterohemorrhagic Escherichia coli O104:H4 outbreak by rapid next generation sequencing technology. *PLoS One, 6*(7), e22751. doi:10.1371/journal.pone.0022751

Metzker, M. L. (2005). Emerging technologies in DNA sequencing. *Genome Res, 15*(12), 1767-1776. doi:10.1101/gr.3770505

Milne, I., Bayer, M., Cardle, L., Shaw, P., Stephen, G., Wright, F., & Marshall, D. (2010). Tablet--next generation sequence assembly visualization. *Bioinformatics, 26*(3), 401-402. doi:10.1093/bioinformatics/btp666

Miyamoto, Y., Kato, T., Obara, Y., Akiyama, S., Takizawa, K., & Yamai, S. (1969). In vitro hemolytic characteristic of Vibrio parahaemolyticus: its close correlation with human pathogenicity. *J Bacteriol, 100*(2), 1147-1149.

Morozova, O., & Marra, M. A. (2008). Applications of next-generation sequencing technologies in functional genomics. *Genomics, 92*(5), 255-264. doi:10.1016/j.ygeno.2008.07.001

Moustafa, A., Reyes-Prieto, A., & Bhattacharya, D. (2008). Chlamydiae has contributed at least 55 genes to Plantae with predominantly plastid functions. *PLoS One, 3*(5), e2205. doi:10.1371/journal.pone.0002205

Nair, G. B., & Hormazabal, J. C. (2005). The Vibrio parahaemolyticus pandemic. *Rev Chilena Infectol, 22*(2), 125-130. doi:/S0716-10182005000200002

Nair, G. B., Ramamurthy, T., Bhattacharya, S. K., Dutta, B., Takeda, Y., & Sack, D. A. (2007). Global dissemination of Vibrio parahaemolyticus serotype O3:K6 and its serovariants. *Clin Microbiol Rev, 20*(1), 39-48. doi:10.1128/CMR.00025-06

New, M. B. (2002). Farming freshwater prawns

A manual for the culture of the giant river

prawn (Macrobrachium rosenbergii). Retrieved from http://www.fao.org/docrep/005/y4100e/y4100e04.htm#P193_35649

New, M. B. (2004, 1st January 2004). Cultured Aquatic Species Information Programme. Macrobrachium rosenbergii. *Cultured Aquatic Species Information*

Programme. Retrieved from http://www.fao.org/fishery/culturedspecies/Macrobrachium_rosenbergii/en

Nishibuchi, M., Fasano, A., Russell, R. G., & Kaper, J. B. (1992). Enterotoxigenicity of Vibrio parahaemolyticus with and without genes encoding thermostable direct hemolysin. *Infect Immun, 60*(9), 3539-3545.

Nishibuchi, M., & Kaper, J. B. (1995). Thermostable direct hemolysin gene of Vibrio parahaemolyticus: a virulence gene acquired by a marine bacterium. *Infect Immun, 63*(6), 2093-2099.

Ochman, H., & Jones, I. B. (2000). Evolutionary dynamics of full genome content in Escherichia coli. *EMBO J, 19*(24), 6637-6643. doi:10.1093/emboj/19.24.6637

Okuda, J., Ishibashi, M., Hayakawa, E., Nishino, T., Takeda, Y., Mukhopadhyay, A. K., . . . Nishibuchi, M. (1997). Emergence of a unique O3:K6 clone of Vibrio parahaemolyticus in Calcutta, India, and isolation of strains from the same clonal group from Southeast Asian travelers arriving in Japan. *J Clin Microbiol, 35*(12), 3150-3155.

Okura, M., Osawa, R., Iguchi, A., Arakawa, E., Terajima, J., & Watanabe, H. (2003). Genotypic analyses of Vibrio parahaemolyticus and development of a pandemic group-specific multiplex PCR assay. *J Clin Microbiol, 41*(10), 4676-4682.

Okura, M., Osawa, R., Iguchi, A., Takagi, M., Arakawa, E., Terajima, J., & Watanabe, H. (2004). PCR-based identification of pandemic group Vibrio parahaemolyticus with a novel group-specific primer pair. *Microbiol Immunol, 48*(10), 787-790.

Park, K. S., Ono, T., Rokuda, M., Jang, M. H., Okada, K., Iida, T., & Honda, T. (2004). Functional characterization of two type III secretion systems of Vibrio parahaemolyticus. *Infect Immun, 72*(11), 6659-6665. doi:10.1128/IAI.72.11.6659-6665.2004

Patel, R. K., & Jain, M. (2012). NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One, 7*(2), e30619. doi:10.1371/journal.pone.0030619

Pavia, A. T., Campbell, J. F., Blake, P. A., Smith, J. D., McKinley, T. W., & Martin, D. L. (1987). Cholera from raw oysters shipped interstate. *JAMA, 258*(17), 2374.

Paydar, M., Teh, C. S., & Thong, K. L. (2013). Prevalence and characterisation of potentially virulent Vibrio parahaemolyticus in seafood in Malaysia using conventional methods, PCR and REP-PCR. *Food Control, 32*, 13-18.

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., & Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A, 96*(8), 4285-4288.

Perna, N. T., Plunkett, G., 3rd, Burland, V., Mau, B., Glasner, J. D., Rose, D. J., . . . Blattner, F. R. (2001). Genome sequence of enterohaemorrhagic Escherichia coli O157:H7. *Nature, 409*(6819), 529-533. doi:10.1038/35054089

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., . . . Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics, 13*, 341. doi:10.1186/1471-2164-13-341

Reen, F. J., Almagro-Moreno, S., Ussery, D., & Boyd, E. F. (2006). The genomic code: inferring Vibrionaceae niche specialization. *Nat Rev Microbiol, 4*(9), 697-704. doi:10.1038/nrmicro1476

Rejeki, S. (2003). Accumulation of aluminium in the tissue of giant fresh water prawn (Macrobrachium rosenbergii de Man) exposed to acidic water contaminated with aluminium salt. *J Coast Dev, 6*(2), 83 - 95.

Richter, D. C., Schuster, S. C., & Huson, D. H. (2007). OSLay: optimal syntenic layout of unfinished assemblies. *Bioinformatics, 23*(13), 1573-1579. doi:10.1093/bioinformatics/btm153

Riley, M. C., Lee, J. E., Lesho, E., & Kirkup, B. C., Jr. (2011). Optically mapping multiple bacterial genomes simultaneously in a single run. *PLoS One, 6*(11), e27085. doi:10.1371/journal.pone.0027085

Rohde, H., Qin, J., Cui, Y., Li, D., Loman, N. J., Hentschke, M., . . . Consortium, E. c. O. H. G. A. C.-S. (2011). Open-source genomic analysis of Shiga-toxin-producing E. coli O104:H4. *N Engl J Med, 365*(8), 718-724. doi:10.1056/NEJMoa1107643

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A., & Barrell, B. (2000). Artemis: sequence visualization and annotation. *Bioinformatics, 16*(10), 944-945.

Sanyal, S. C., & Sen, P. C. (1974). Human volunteer study on the pathogenicity of Vibrio parahaemolyticus. In: International Symposium of Vibrio parahaemolyticus. *International Symposium of Vibrio parahaemolyticus. Saikon, Tokyo, pp.*, 227-230.

Saravanan, S., Biju Sam Kamalam, J., Stephen, J., & Kumar, S. (2008). Moulting and Behaviour Changes in Freshwater Prawn. Issue 25 Spring/Summer 2008.

Retrieved from http://www.thefishsite.com/articles/464/moulting-and-behaviour-changes-in-freshwater-prawn/

Schattner, P., Brooks, A. N., & Lowe, T. M. (2005). The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res, 33*(Web Server issue), W686-689. doi:10.1093/nar/gki366

Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics, 27*(6), 863-864. doi:10.1093/bioinformatics/btr026

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nat Biotechnol, 26*(10), 1135-1145. doi:10.1038/nbt1486

Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y., & Ishikawa, H. (2000). Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp. APS. *Nature, 407*(6800), 81-86. doi:10.1038/35024074

Shirai, H., Ito, H., Hirayama, T., Nakamoto, Y., Nakabayashi, N., Kumagai, K., . . . Nishibuchi, M. (1990). Molecular epidemiologic evidence for association of thermostable direct hemolysin (TDH) and TDH-related hemolysin of Vibrio parahaemolyticus with gastroenteritis. *Infect Immun, 58*(11), 3568-3573.

Siddavattam, D., Karegoudar, T. B., Mudde, S. K., Kumar, N., Baddam, R., Avasthi, T. S., & Ahmed, N. (2011). Genome of a novel isolate of Paracoccus denitrificans capable of degrading N,N-dimethylformamide. *J Bacteriol, 193*(19), 5598-5599. doi:10.1128/JB.05667-11

Snel, B., Bork, P., & Huynen, M. A. (1999). Genome phylogeny based on gene content. *Nat Genet, 21*(1), 108-110. doi:10.1038/5052

Stothard, P., & Wishart, D. S. (2005). Circular genome visualization and exploration using CGView. *Bioinformatics, 21*(4), 537-539. doi:10.1093/bioinformatics/bti054

Sujeewa, A. K. W., Norrakiah, A. S., & Laina, M. (2009). Prevalence of toxic genes of Vibrio parahaemolyticus in shrimps (Penaeus monodon) and culture environment. *International Food Research Journal, 16*, 89-95.

Taylor, J., Schenck, I., Blankenberg, D., & Nekrutenko, A. (2007). Using galaxy to perform large-scale interactive data analyses. *Curr Protoc Bioinformatics, Chapter 10*, Unit 10 15. doi:10.1002/0471250953.bi1005s19

Tekaia, F., Lazcano, A., & Dujon, B. (1999). The genomic tree as revealed from whole proteome comparisons. *Genome Res, 9*(6), 550-557.

Thompson, F. L., Iida, T., & Swings, J. (2004). Biodiversity of vibrios. *Microbiol Mol Biol Rev, 68*(3), 403-431, table of contents. doi:10.1128/MMBR.68.3.403-431.2004

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res, 22*(22), 4673-4680.

Tiruvayipati, S., Bhassu, S., Kumar, N., Baddam, R., Shaik, S., Gurindapalli, A. K., . . . Ahmed, N. (2013). Genome anatomy of the gastrointestinal pathogen, Vibrio parahaemolyticus of crustacean origin. *Gut Pathog, 5*(1), 37. doi:10.1186/1757-4749-5-37

Treangen, T. J., & Messeguer, X. (2006). M-GCAT: interactively and efficiently constructing large-scale multiple genome comparison frameworks in closely related species. *BMC Bioinformatics, 7*, 433. doi:10.1186/1471-2105-7-433

Tritt, A., Eisen, J. A., Facciotti, M. T., & Darling, A. E. (2012). An integrated pipeline for de novo assembly of microbial genomes. *PLoS One, 7*(9), e42304. doi:10.1371/journal.pone.0042304

Trivedi, U. H., Cezard, T., Bridgett, S., Montazam, A., Nichols, J., Blaxter, M., & Gharbi, K. (2014). Quality control of next-generation sequencing data without a reference. *Front Genet, 5*, 111. doi:10.3389/fgene.2014.00111

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen, S. G. (2012). Primer3--new capabilities and interfaces. *Nucleic Acids Res, 40*(15), e115. doi:10.1093/nar/gks596

Untergasser, A., Nijveen, H., Rao, X., Bisseling, T., Geurts, R., & Leunissen, J. A. (2007). Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Res, 35*(Web Server issue), W71-74. doi:10.1093/nar/gkm306

Utturkar, S. M., Klingeman, D. M., Land, M. L., Schadt, C. W., Doktycz, M. J., Pelletier, D. A., & Brown, S. D. (2014). Evaluation and validation of de novo and hybrid assembly techniques to derive high-quality genome sequences. *Bioinformatics, 30*(19), 2709-2716. doi:10.1093/bioinformatics/btu391

von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., & Snel, B. (2003). STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res, 31*(1), 258-261.

Wang, X., & Wood, T. K. (2011). Toxin-antitoxin systems influence biofilm and persister cell formation and the general stress response. *Appl Environ Microbiol, 77*(16), 5577-5583. doi:10.1128/AEM.05068-11

Wong, E., Vaaje-Kolstad, G., Ghosh, A., Hurtado-Guerrero, R., Konarev, P. V., Ibrahim, A. F., . . . van Aalten, D. M. (2012). The Vibrio cholerae colonization factor GbpA possesses a modular structure that governs binding to different host surfaces. *PLoS Pathog, 8*(1), e1002373. doi:10.1371/journal.ppat.1002373

Wong, H. C., & Lee, Y. S. (1994). Regulation of iron on bacterial growth and production of thermostable direct hemolysin by Vibrio parahaemolyticus in intraperitoneal infected mice. *Microbiol Immunol, 38*(5), 367-371.

Wong, H. C., Liu, S. H., Chiou, C. S., Nishibuchi, M., Lee, B. K., Suthienkul, O., . . . Taniguchi, H. (2007). A pulsed-field gel electrophoresis typing scheme for Vibrio parahaemolyticus isolates from fifteen countries. *International Journal of Food Microbiology, 114*(3), 280-287. doi:DOI 10.1016/j.ijfoodmicro.2006.09.024

Xu, M., Yamamoto, K., Honda, T., & Ming, X. (1994). Construction and characterization of an isogenic mutant of Vibrio parahaemolyticus having a deletion in the thermostable direct hemolysin-related hemolysin gene (trh). *J Bacteriol, 176*(15), 4757-4760.

Yamaichi, Y., Iida, T., Park, K. S., Yamamoto, K., & Honda, T. (1999). Physical and genetic map of the genome of Vibrio parahaemolyticus: presence of two chromosomes in Vibrio species. *Mol Microbiol, 31*(5), 1513-1521.

Yamamoto, S., Okujo, N., Matsuura, S., Fujiwara, I., Fujita, Y., & Shinoda, S. (1994). Siderophore-mediated utilization of iron bound to transferrin by Vibrio parahaemolyticus. *Microbiol Immunol, 38*(9), 687-693.

Yang, X., Liu, D., Liu, F., Wu, J., Zou, J., Xiao, X., . . . Zhu, B. (2013). HTQC: a fast quality control toolkit for Illumina sequencing data. *BMC Bioinformatics, 14*, 33. doi:10.1186/1471-2105-14-33

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., & Madden, T. L. (2012). Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics, 13*, 134. doi:10.1186/1471-2105-13-134

Yeung, P. S., & Boor, K. J. (2004). Epidemiology, pathogenesis, and prevention of foodborne Vibrio parahaemolyticus infections. *Foodborne Pathog Dis, 1*(2), 74-88. doi:10.1089/153531404323143594

Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res, 18*(5), 821-829. doi:10.1101/gr.074492.107

Zhaxybayeva, O., Gogarten, J. P., Charlebois, R. L., Doolittle, W. F., & Papke, R. T. (2006). Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. *Genome Res, 16*(9), 1099-1108. doi:10.1101/gr.5322306

Zhou, Q., Su, X., & Ning, K. (2014). Assessment of quality control approaches for metagenomic data analysis. *Sci Rep, 4*, 6957. doi:10.1038/srep06957

SUPPLEMENTARY DATA

List of Paper / Poster presentations

1. “A customized pipeline for quality control of Next-generation sequencing raw read data of bacterial genomes” at the International Conference on “Advances in Biotechnology and Bioinformatics” (ICABB2013) held during 25th – 27th November, 2013.

2. “The genome anatomy of *Vibrio parahaemolyticus* PCV08-7 of crustacean origin” at the 20th MSMBB Annual Scientific Meeting, 26th – 27th June, 2013, University of Malaya, Kuala Lumpur, Malaysia.

3. “An efficient and reliable workflow environment for the genome mapping of bacteria using High-throughput Next-Generation Sequencing data” at the 31st Symposium of the Malaysian Society for Microbiology, Kota Kinabalu, Sabah, Malaysia on 13th -15th December, 2012.

SIMA
731

INTERNATIONAL CONFERENCE ON ADVANCES IN

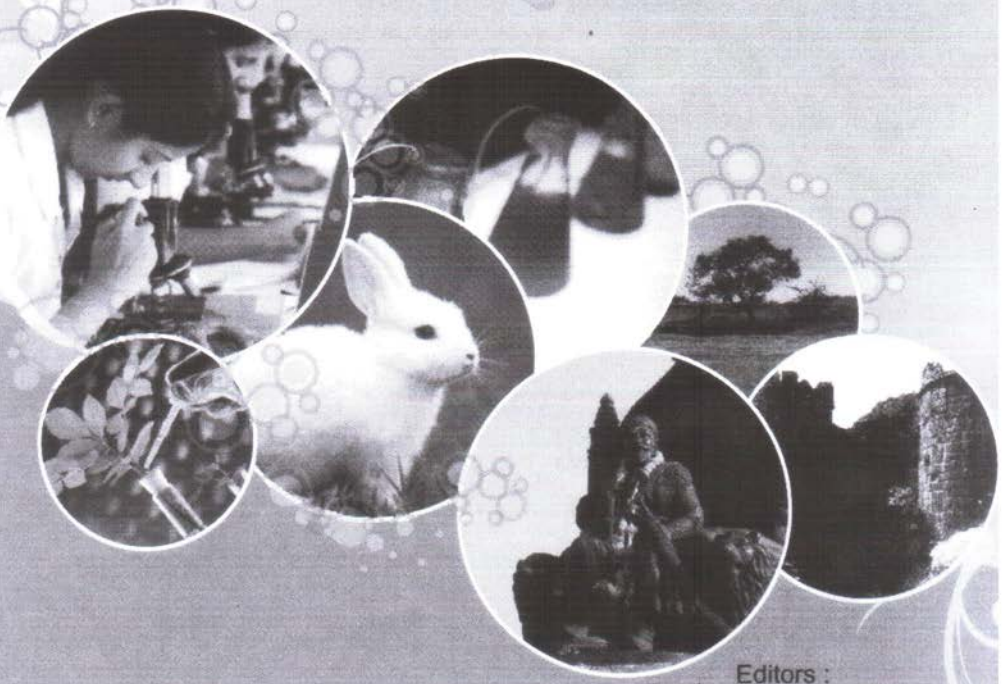
BIOTECHNOLOGY AND BIOINFORMATICS

ICABB 2013

&

X CONVENTION OF THE BIOTECH RESEARCH SOCIETY, INDIA

ABSTRACTS



Editors :
Madhukar Khetmalas
Ashok Pandey
P.Gunasekaran
Neelu Nawani
Shuchi Nagar

Dr.D.Y.Patil Vidyapeeth, Pune
&
The Biotech Research Society, India

DPU
PUBLICATIONS

Department of Biotechnology
Ministry of Science & Technology,
Government of India

NABAR

Subha Bhassu, PhD
Assoc. Prof.
Division of Genetics & Molecular Biology
Institute of Biological Sciences,
Faculty of Science, University of Malaya,
50603 Kuala Lumpur, Malaysia.

Laccase, GLOX (glyoxidase), and aryl alcohol oxidases (AAO), were observed. Highest reducing sugar production was recorded from sawdust pretreated with *Phanerochate sordida* (793.4 mg/ml/2 g) followed by *Trichoderma harzianum* treated sawdust (725.2 mg/ml/2 g) and Byssochlamys nivea treated sawdust (713.9 mg/ml/2 g) after 19th, 15th and 17th day of fungal treatment respectively. The study presents a potential alternative cost effective method for producing reducing sugars.

Abstract Code : MIB-204 **Reg No :** 731

A customized pipeline for quality control of next generation sequencing raw read data of bacterial genomes

Suma Tiruvayipati[1,2], Sabiha Shaik[2], Ramani Baddam[2], Narender Kumar [2], Subha Bhassu[1,3], Niyaz Ahmed[1,2]

[1]*Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia.*
[2] *Pathogen Biology Laboratory, Department of Biotechnology and Bioinformatics, School of Life Sciences, University of Hyderabad, Hyderabad, India.*
[3]*Genomics and Evolutionary Biology Laboratory, CEBAR, Center of Biotechnology for Agriculture, Universiti Malaya, Kuala Lumpur, Malaysia*
[4]*Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia.*
Corresponding E-mail: sumatiruvayipati@gmail.com

Advances in Next generation sequencing technology has not only increased data generation to above a terabyte per run, but even enhanced the development of many strategies of quality control to deal with sequencing data. Various quality control methods have been published within the past few years. To obtain maximum number of total bases, while maintaining fewer counts of contigs as possible, as a result of the first step of assembly is the present aim. To make this strategy successful it is really important to deal with the aspects of proper adaptor filtering, trimming low quality read ends and high quality read filtering. We present a customized pipeline which integrates few aspects from previously published filtering algorithms along with few new algorithms to simplify the entire filtering pipeline with quality control steps at every stage. With the present work we demonstrate dealing stepwise with raw sequencing read data to clean adaptors, trim reads to make good quality read data sets. Especially to avoid low quality read information and reads carrying less probability of the existing base nucleotide data. Our customized pipeline can filter reads to fulfil all criteria of a whole quality control procedure depending upon the number of reads and sequencing quality to obtain the best quality sequencing read dataset which further can be better processed through assembly tools.

Abstract Code : MIB-205 **Reg No :** 734

Production of Asparaginase by Endophytic bacteria *Pseudomonas oryzihabitans*

Jyoti Bhagat, Amarjeet Kaur, B.S. Chadha

Department of Micribiology, Guru Nanak Dev University, Amritsar-143005, India.
Corresponding E-mail: jots_005@yahoo.com

L-Asparaginase is used as a therapeutic molecule for the treatment of acute lymphoblastic leukaemia (ALL). It is also used in the food industry to remove acrylamide from fried starch-based foods. In an attempt to identify novel organisms as sources of this enzyme, different endophytic cultures were isolated and screened from plant *Hibiscus rosa-sinensis*. In total 27 endophytic cultures were isolated from medicinal plant *H. rosa* (China rose) procured from Guru Nanak Dev University, Amritsar Punjab. Out of the 27 screened cultures, twelve asparaginase positive isolates were obtained with a rapid plate assay method (primary screening).Whereas from these 12, only 4 were found positive in secondary screening. The maximal L-asparaginase producing strain was identified as Pseudomonas

Subha Bhassu, PhD
Assoc. Prof.
Division of Genetics & Molecular Biology,
Institute of Biological Sciences,
Faculty of Science, University of Malaya,
50603 Kuala Lumpur, Malaysia.

20th MSMBB Annual Scientific Meeting June 26-27 2013 Kuala Lumpur

Abstracts

MSMBB

PERSATUAN BIOLOGI MOLEKUL DAN BIOTEKNOLOGI MALAYSIA

25 years

Malaysian Society for Molecular Biology and Biotechnology

With the support of :

UMBIO

Health and Translational Medicine Cluster

UNIVERSITI MALAYA

In this study, the different phases of female flower development and subsequent mesocarp maturation were defined by over 13,000 of differentially expressed transcripts, which formed 22 clusters of distinct expression profiles. Several orthologs of key genes in the circadian rhythm pathway were differentially expressed throughout the development from immature female inflorescence to the fruit (mesocarp), showing consistency with reports in *Arabidopsis*, while comparisons of male to female inflorescences showed female-specific upregulation of the carotenoid pathway. The biological functions of the remaining unannotated transcripts, which exhibited differential expression patterns, were predicted via domain analysis.

The Genome Anatomy of *Vibrio parahaemolyticus* PCV08-7 of Crustacean Origin

Suma Tiruvayipati[1*], Subha Bhassu [1, 2], Thong Kwai Lin [2, 4], Niyaz Ahmed [3, 2]
[1]*Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia*
[2] *Genomics and Evolutionary Biology Laboratory, CEBAR, Center of Biotechnology for Agriculture, University of Malaya, Kuala Lumpur, Malaysia*
[3]*Pathogen Biology Laboratory, Department of Biotechnology and Bioinformatics, School of Life Sciences, University of Hyderabad, Hyderabad, India*
[4]*Laboratory of Biomedical Science and Molecular Microbiology, UMBIO Research Cluster, University of Malaya, Kuala Lumpur, Malaysia*

Vibrio parahaemolyticus is a Gram-negative halophilic bacterium, which has been a major cause of common foodborne illnesses. It has been reported as being one of the major causes of outbreaks due to the consumption of raw, undercooked seafood. Here we describe the sequence of a *Vibrio parahaemolyticus* isolate (PCV08-7) which has been cultured from a crustacean sample in 2008 sourced at Selangor, Malaysia. Our genome analysis revealed that the closest similarity was to the reference genome *Vibrio parahaemolyticus* RIMD2210633. The preliminary appraisal of our Next-Generation Sequencing data generated on the Illumina/ Solexa platform revealed homology to several conserved regions corresponding to the above reference strain, while a few regions, such as the integrated element target tmRNA, toxR, toxS and genes for fatty acid and amino acid metabolism, were present. Regions such as the Thermostable Direct Hemolysin (TDH) Pathogenicity Island, type IV secretion system and the phage f237 genes were partially discerned. The absence of the TDH-related direct hemolysin (TRH) Pathogenicity Island yet presence of the HU-alpha insertion element shows that the genome has a closer proximity to an O3:K6 type strain. Further investigation of the sequence data corresponding to chromosomal as well as plasmid components of the genome would add to our understanding of the novel virulence mechanisms and other genetic attributes of this pathogen at the population level.

Subha Bhassu, PhD
Assoc. Prof.
Division of Genetics & Molecular Biology,
Institute of Biological Sciences,
Faculty of Science, University of Malaya,
50603 Kuala Lumpur, Malaysia

31st SYMPOSIUM OF THE MALAYSIAN SOCIETY FOR MICROBIOLOGY

"Microbiology Research in the Omics Era"

13-15 December 2012

Promenade Hotel, Kota Kinabalu,
Sabah, Malaysia

CO-ORGANISED BY:

Biotechnology Research Institute,
Universiti Malaysia Sabah

and

Malaysian Society for Microbiology

In collaboration with

UKM, UM, UPM, USM, SIRIM, MARDI, NUCLEAR MALAYSIA,
ASM, BEP

MO05

AN EFFICIENT AND RELIABLE WORKFLOW ENVIRONMENT FOR THE GENOME MAPPING OF BACTERIA USING HIGH-THROUGHPUT NEXT-GENERATION SEQUENCING DATA

Suma Tiruvayipati[2,3], Narender Kumar[3], Ramani Baddam[3], Subha Bhassu[1,2*], and Niyaz Ahmed[2,3*]

[1]CEBAR, Center of Biotechnology for Agriculture, Universiti Malaya, 50603 Kuala Lumpur, Malaysia. [2]Institut Sains Biologi, Fakulti Sains, Universiti Malaya, 50603 Kuala Lumpur, Malaysia. [3]Pathogen Biology Laboratory, Department of Biotechnology, School of Life Sciences, University of Hyderabad, Hyderabad – 500046, Andhra Pradesh, India

*Corresponding author's e-mail: subhabhassu@um.edu.my

With the advent of the Next-Generation Sequencing (NGS) platforms for genome sequencing, the scientific community is confronted by the challenge to evolve innovative technologies for data management. The NGS technologies not only help in real-time sequencing of various genomes, but they also have huge application in heralding the benefits of OMICS sciences. As said "in great power lies great responsibility", with massive data churned by NGS, the tasks to accurately assemble, analyse and sort such an information is the need of the hour. While bacterial genomes are known to be smaller, as compared to their eukaryotic counterparts, it is quite difficult to assemble microbial genomes especially in cases where repetitive regions are abundant. We have been handling bacterial genome data of *Escherichia coli*, *Helicobacter pylori*, *Mycobacterium tuberculosis*, *Salmonella enterica*, *Paracoccus denitrificans* and *Vibrio parahaemolyticus* for some time and these genomes are analysed for different gene densities, coding patterns and G+C content etc. It has been very crucial to handle these genomes so as to conserve their fullest nucleotide content throughout different analysis phases so that the important co-ordinates, such as the 'intact' pathogenicity islands, virulence genes and niche specific genes are retained, clearly identified and annotated in the final assembly. There are lessons learnt from each of these important genome projects and this presentation would therefore deal with various approaches and strategies adopted to achieve complete, accurate and presentable genomic blueprints of these pathogens.

LIST OF PUBLICATIONS

1. Vibrio parahaemolyticus PCV08-7 gbpA gene expression studies in response to prawn chitin and external environment magnesium. **Suma Tiruvayipati**, Subha Bhassu *(to communicate)*

2. Understanding bacterial genomes and optimizing a standard quality control protocol for bacteria using Next-generation sequencing raw reads **Suma Tiruvayipati**, Subha Bhassu, Niyaz Ahmed *(to communicate)*

3. REVIEW: Impact of magnesium on Vibrio parahaemolyticus and Macrobrachium rosenbergii: A holistic approach. **Suma Tiruvayipati** & Subha Bhassu *(UNDER REVIEW)*

4. Characterization and gene expression of a putative clotting protein in giant freshwater prawn. Nazanin Ziba, **Suma Tiruvayipati**, Subha Bhassu *(UNDER REVIEW)*

5. RNA-seq analysis of Macrobrachium rosenbergii hepatopancreas in response to Vibrio parahaemolyticus infection. (2015) R Rama, Zhu YB, Alinejad T, **Tiruvayipati S**, Wang J, Thong KL, Bhassu S. *Gut Pathogens.* doi:10.1186/s13099-015-0052-6

6. Genomes of Two Clinical Isolated of Mycobacterium tuberculosis from Odisha, India. (2014) Majid M, Kumar N, Qureshi A, Yerra P, Kumar A, Kumar MK, **Tiruvayipati S**, Baddam R, Shaik S, Srikantam A, Ahmed N. *Genome Announc.* Mar 20;2(2). pii: e00199-14. doi: 10.1128/genomeA.00199-14.

7. Genome anatomy of the gastrointestinal pathogen, Vibrio parahaemolyticus of crustacean origin. (2013) **Tiruvayipati S**, Bhassu S, Kumar N, Baddam R, Shaik S, Gurindapalli K, Thong KL, Ahmed N. *Gut Pathogens*. 5(1):37. doi:10.1186/1757-4749-5-37

8. Genome sequencing and analysis of Salmonella enterica serovar Typhi strain CR0063 representing a carrier individual during an outbreak of typhoid fever in Kelatan, Malaysia. (2012) Baddam R, Kumar N, Shaik S, **Suma T**, Ngoi ST, Thong KL, Ahmed N. Gut Pathogens. 4(1):20. doi:10.1186/1757-4749-4-20

9. Genome Sequence and Comparative Pathogenomics Analysis of a Salmonella enterica Serovar Typhi Strain Associated with a Typhoid Carrier in Malaysia. (2012) Yap KP, Gan HM, Teh CS, Baddam R, Chai LC, Kumar N, **Tiruvayipati SA**, Ahmed N, Thong KL. *Journal of Bacteriology*. 194(21):5970-1

10. Next generation sequencing and analysis of the genomes of two Helicobacter pylori isolates from duodenal ulcer patients in India: de-novo assembly, genome organization and comparative genomics. (2012) Kumar N, Mukhopadhyay AK, Patra R, De R, Baddam R, Shaik S, Alam J, **Tiruvayipati S**, Ahmed N. *Journal of Bacteriology.* 194(21):5963-4

11. Insights from the genome sequence of Salmonella enterica serovar typhi strain associated with a sporadic case of typhoid fever in Malaysia. (2012) Yap KP, Teh CS, Baddam R, Chai LC, Kumar N, **Avasthi TS**, Ahmed N, Thong KL. *Journal of Bacteriology.* 194(18):5124-5

12. Whole Genome Sequences and Comparative Genomics of Salmonella enterica Serovar Typhi isolates from patients with fatal and non-fatal typhoid fever in Papua New Guinea. (2012) Baddam R, Thong KL, **Avasthi TS**, Shaik S, Yap KP, Teh CS, Chai LC, Kumar N, Ahmed N. *Journal of Bacteriology.* 194(18):5122-3

13. Crowd Sourcing a new paradigm for Interactome Driven Drug Target Identification in Mycobacterium tuberculosis. (2012) Vashisht R, Mondal AK, Jain A, Shah A, Vishnoi P, Priyadarshini P, Bhattacharyya K, Rohira H, Bhat AG, Passi A, Mukherjee K, Choudhary KS, Kumar V, Arora A, Munusamy P, Subramanian A, Venkatachalam A, S G, Raj S, Chitra V, Verma K, Zaheer S, J B, Gurusamy M, Razeeth M, Raja I, Thandapani M, Mevada V, Soni R, Rana S, Ramanna GM, Raghavan S, Subramanya SN, Kholia T, Patel R, Bhavnani V, Chiranjeevi L, Sengupta S, Singh PK, Atray N, Gandhi S, **Avasthi TS**, Nisthar S, Anurag M, Sharma P, Hasija Y, Dash D, Sharma A, Scaria V, Thomas Z; OSDD Consortium, Chandra N, Brahmachari SK, Bhardwaj A. *PlosOne*. 7(7):e39808. Epub 2012 Jul 11

14. Genetic Fine Structure of a Salmonella enterica Serovar Typhi Strain Associated with the 2005 Outbreak of Typhoid Fever in Kelantan, Malaysia. (2012) Baddam R, Kumar N, Thong KL, Ngoi ST, Teh CS, Yap KP, Chai LC, **Avasthi TS**, Ahmed N. *Journal of Bacteriology*. 194(13):3565-6

Rao et al. Gut Pathogens (2015) 7:6
DOI 10.1186/s13099-015-0052-6

Gut Pathogens

RESEARCH Open Access

RNA-seq analysis of *Macrobrachium rosenbergii* hepatopancreas in response to *Vibrio parahaemolyticus* infection

Rama Rao[1], Ya Bing Zhu[2], Tahereh Alinejad[1], Suma Tiruvayipati[1], Kwai Lin Thong[3], Jun Wang[2] and Subha Bhassu[1*]

Abstract

Background: The Malaysian giant freshwater prawn, *Macrobrachium rosenbergii*, is an economically important crustacean worldwide. However, production of this prawn is facing a serious threat from Vibriosis disease caused by *Vibrio* species such as *Vibrio parahaemolyticus*. Unfortunately, the mechanisms involved in the immune response of this species to bacterial infection are not fully understood. We therefore used a high-throughput deep sequencing technology to investigate the transcriptome and comparative expression profiles of the hepatopancreas from this freshwater prawn infected with *V. parahaemolyticus* to gain an increased understanding of the molecular mechanisms underlying the species' immune response to this pathogenic bacteria.

Result: A total of 59,122,940 raw reads were obtained from the control group, and 58,385,094 reads from the *Vibrio*-infected group. Via *de novo* assembly by Trinity assembler, 59,050 control unigenes and 73,946 *Vibrio*-infected group unigenes were obtained. By clustering unigenes from both libraries, a total of 64,411 standard unigenes were produced. The standard unigenes were annotated against the NCBI non-redundant, Swiss-Prot, Kyoto Encyclopaedia of Genes and Genome pathway (KEGG) and Orthologous Groups of Proteins (COG) databases, with 19,799 (30.73%), 16,832 (26.13%), 14,706 (22.83%) and 7,856 (12.19%) hits respectively, giving a final total of 22,455 significant hits (34.86% of all unigenes). A Gene Ontology (GO) analysis search using the Blast2GO program resulted in 6,007 unigenes (9.32%) being categorized into 55 functional groups. A differential gene expression analysis produced a total of 14,569 unigenes aberrantly expressed, with 11,446 unigenes significantly up-regulated and 3,103 unigenes significantly down-regulated. The differentially expressed immune genes fall under various processes of the animal immune system.

Conclusion: This study provided an insight into the antibacterial mechanism in *M. rosenbergii* and the role of differentially expressed immune genes in response to *V. parahaemolyticus* infection. Furthermore, this study has generated an abundant list of transcript from *M.rosenbergii* which will provide a fundamental basis for future genomics research in this field.

Keywords: Transcriptomics, *Macrobrachium rosenbergii*, *Vibrio parahaemolyticus*, *de novo* assembly, Immune genes, Host-pathogen interaction

* Correspondence: subhabhassu@gmail.com
[1]Genomic Research and Breeding Laboratory and Centre for Research in Biotechnology for Agriculture (CEBAR), Institute of Biological Sciences, Faculty of Science, University of Malaya, 50603 Kuala Lumpur, Malaysia
Full list of author information is available at the end of the article

BioMed Central

© 2015 Rao et al.; licensee BioMed Central. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

Background

The Malaysian giant freshwater prawn, *Macrobrachium rosenbergii* (locally known as 'udang galah'), belongs to the genus *Macrobrachium,* which is the largest genus of the family *Palaemonidae* [1]. They are found in most inland freshwater areas, including lakes, rivers, swamps, estuarine areas, ponds, canals as well as in irrigation ducts [2]. *M. rosenbergii* spends its adult life in fresh water, but requires brackish water during the initial stages of its life cycle [3]. High demand from the aquaculture industry has led to large-scale farming of this prawn in many countries; the major producers being Bangladesh, Brazil, China, Ecuador, India, Thailand, Taiwan Province of China, and Malaysia [4].

The global production of this prawn had increased to over 200 000 tonnes/year by 2002, and income in Asia alone is now worth US$1 billion per annum [5,6]. In Malaysia, the production of cultured *M. rosenbergii* reached 281 metric tonnes by 1998 [4]. Generally, *M. rosenbergii* is assumed to be less resistant towards diseases than penaeid shrimp [7]. However, with the rise of large-scale high density prawn aquaculture techniques, production of this prawn worldwide is facing a serious threat from fatal diseases caused by nodaviruses and bacteria, particularly from the *Vibrio* species [8,9]. The emergence of these pathogens has had a detrimental impact on the *M. rosenbergii* farming industry, causing considerable economic losses.

Vibrio is a Gram-negative halophilic bacterium found abundantly in marine and estuarine environments [10,11]. Among the different species, *Vibrio parahaemolyticus* has emerged as an important pathogen for *M. rosenbergii* [12]. Several other marine shrimps such as *Penaeus monodon, Penaeus japonicas* and *Litopenaeus vannamei* have also been found to be susceptible to *Vibrio* infection [13]. Severe *V. parahaemolyticus* infection in prawns leads to a disease known as 'Vibriosis' [14,15]. *M. rosenbergii* suffering from vibriosis may appear black in colour on the carapace, with red discolouration of the exoskeleton and loss of appendages within six days, leading to an 80% mortality rate [12].

Acquiring and establishing knowledge regarding host pathogen interactions is necessary to unlock the pathogenesis of a particular disease. Host pathogen interactions can result in acute and adaptive immune responses against an invader; however, this has been lacking in *M. rosenbergii* [16]. The species defends itself against pathogen invasion using an innate immune system involving the cellular and humoral mechanisms [17,18]. Recently, some progress has been made in analysing the molecular mechanisms of shrimp-pathogen interactions, and several immune genes from shrimp have been discovered such as lectins, antimicrobial peptides, prophenoloxidase and manganese superoxide dismutase, using methods such as suppression subtractive hybridization (SSH) and expressed sequence tags (EST) [19-21]. However, these two methods have been found to be laborious and costly, which limits their use for the production of large-scale transcripts [22].

A cutting edge technology has emerged recently, known as Next Generation Sequencing technology (NGS). Currently, there are four established platforms which uses NGS technology: the Illumina Genome Analyzer, the Roche/454 Genome Sequencer FLX Instrument, and the ABI SOLiD System [23,24]. These platforms have proven versatile and cost-effective tools for advanced research in various genomic areas, such as genome sequencing and re-sequencing, DNA methylation analysis, miRNA expression profiling, and also in non-model organisms as the *de novo* transcriptome sequencing [25]. By using the NGS platform, transcriptome analysis can be performed faster and more easily, because it does not require any bacterial cloning of cDNAs [26]. NGS sequencing has the further advantage of generating greater depth of short reads with minimum error rates [27]. Moreover, it is more reliable and efficient than previous methods in measuring transcriptome composition, revealing RNA expression patterns, and discovering new genes on a larger scale [28]. The superiority of this technology also lies in its sensitivity, which allows the detection of low-abundance transcripts [29].

Previous studies have been performed on whole transcriptome sequencing of the hepatopancreas, gill and muscle tissues of *M. rosenbergii* using the Illumina Genome Analyzer IIx platform (Illumina). They successfully produced a comprehensive transcript data for this freshwater prawn, leading to the discovery of new genes [30]. This present study utilised a similar approach to analyse transcriptome data obtained from the hepatopancreas of *M. rosenbergii* experimentally infected with *V. parahaemolyticus*. The aim was to discover, and determine the role of, immune genes in *M. rosenbergii* involved in *V. parahaemolyticus* infection, which in turn could provide insights into the host-pathogen interactions between these two organisms.

Material and methods

M. rosenbergii and *V. parahaemolyticus* PCV08-7 challenge

M. rosenbergii prawns (5-8 g body weight) purchased from a local hatchery (Kuala Kangsar, Perak, Malaysia) were acclimatized at 28 ± 1°C in aerated and filtered freshwater for one week prior to challenge with *V. parahaemolyticus*. During the challenge experiment, the prawns (n = 10) were intramuscularly injected with 100 μl $1X10^5$ cfu cultured *V. parahaemolyticus* [31] whereas another batch of prawns (n = 10) were injected with 100 μl 2% NaCl (1:10, w/v) solution which serves as negative control group. The hepatopancreas tissues of the prawns were dissected at 12 hours post-infection. The tissues were rapidly frozen in liquid nitrogen and stored at −80°C until total

RNA extraction. The 12 hour time point was chosen based on our previous work regarding immune related genes from *M. rosenbergii* in response to pathogen such as viruses showing significant gene expression at this time point [32-35].

Total RNA extraction and next-generation sequencing

Total RNA (~20 mg) was isolated from both the *V. parahaemolyticus*-challenged and negative control group hepatopancreases. The RNA extraction process was performed by using the Macherey-Nagel NucleoSpin RNA II extraction kit in accordance with the manufacturer's protocols and stored at −80°C prior to RNA sequencing. The purity and integrity of the RNA was assessed by using the Bioanalyzer 2100 (Agilent technologies, USA). In each group, the total RNA samples were pooled from 10 prawns after which cDNA was synthesized followed by sequencing. The sequencing run was conducted on an Illumina HiSeq™ 2000 platform at the Beijing Genome Institute, Shenzhen, China. The sequencing data constituted 90 bp paired end read data, with ~117 million raw reads.

Assembly and functional annotation

The raw reads were primarily quality filtered to remove adaptor sequences followed by removal of ambiguous 'N' nucleotides (with a ratio of 'N' more than 10%) and sequences with a phred quality score of less than 20 before proceeding to *de novo* assembly by using the Trinity software [36]. The Trinity programme assembles the reads into contigs and these contigs were assembled to unigenes. Finally, the TIGR Gene Indices clustering tools (TGICL) [37] with default parameters was applied to cluster the unigenes from both groups which produces non-redundant unigenes.

The non-redundant unigene sequences were aligned to databases which included NCBI non-redundant (Nr), Swissprot [38], Cluster of Orthologous Groups (COG) [39] and Kyoto Encyclopaedia of Genes and Genome (KEGG) [40] using BLASTX [41] with an E-value cut-off of 10^{-5}. Gene Ontology (GO) was conducted utilizing default parameters using the BLAST2GO software [42,43]. It was from the above mentioned databases that the gene direction of the unigenes which were annotated and the coding sequence were determined from the BLAST results. The prediction for the coding sequence and the gene direction was performed by ESTscan [44] for those sequences with no defined annotation by using BLAST predicted coding sequence data as the training set.

Identification of differentially expressed unigenes

The FPKM method (Fragments Per kb per Million fragments) was used to calculate the transcript expression levels [45]. An FDR (false discovery rate) of <0.001 was used as the threshold p-value in multiple tests to judge the degree of differences in gene expression [46]. In a given library when the p-value was less than 0.001 and when the expression level showed greater than two-fold change between two groups genes were considered as differentially expressed.

Quantitative RT-PCR analysis

We selected seven differentially expressed *M. rosenbergii* unigenes (arginine kinase 1, anti-lipopolysaccharide factor, inhibitor of apoptosis protein, caspase, heat shock protein 21, lectin 1, and NF-kappa B inhibitor alpha) for quantitative RT-PCR analysis (qRT-PCR) to evaluate our Illumina sequencing result. The primer design for the seven unigenes was performed by using Primer3 software http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi/ and listed in Additional file 1. Using 1 μg of RNA, first strand cDNA synthesis was carried out (similar to the sample used for transcriptome sequencing) by using the ImProm-II™ Reverse Transcription System (Promega). The qRT-PCR reaction (20 μl) consisted of a 10 μl TaqMan Universal RT-PCR Master Mix (Applied Biosystems, Foster City, CA, USA), a 1 μl of primers/probe set containing 900 nM of forward reverse primers, a 300 nM probe and 2 μl of template cDNA. The qRT-PCR program was set with an incubation step at 50°C for 2 min, 40 cycles at 95°C for 10 min, 95°C for 15 sec, and 60°C for 1 min, carried out by using Step One Plus Real-Time PCR System® (Applied Biosystems). Similar qRT-PCR cycle profile was applied for the internal control gene, Elongation factor 1-alpha (primer sequences are listed in Additional file 1). The expression level of the seven immune genes were analysed by using the comparative CT method ($2^{-\Delta\Delta CT}$ method) [47].

Results

Illumina sequencing and assembly

The task of profiling all the immune-related genes involved in *V .parahaemolyticus* infection began with sequencing the two cDNA libraries prepared from pooled mRNAs obtained from the hepatopancreases of the control and infected groups using the Illumina HiSeq™ 2000 platform. A total of 59,122,940 raw reads were obtained from the control group, and 58,385,094 reads from the *Vibrio*-infected group. The raw reads were further filtered to remove adaptor sequences, ambiguous reads and low quality reads, thereby generating 90-bp of 54,708,014 and 54,295,342 clean reads for the control and infected groups respectively (Q_{20} ~ 98% and percentage of unknown nucleotide is 0%). All sequencing reads were deposited into the Short Read Archive (SRA) of the National Centre for Biotechnology Information (NCBI), and can be accessed under the accession number

SRR1424572 for control and SRR1424574 for *Vibrio*-infected ones.

All the clean reads were subjected to *de novo* assembly using the Trinity program which uses three independent software modules – Inchworm, Chrysalis, and Butterfly – applied sequentially to process the huge sequencing data of RNA-seq reads. The assembly of the reads produced 95,645 contigs (with an N_{50} of 467 bp and mean length of 313 bp) for the control group and 123,141 contigs (with an N_{50} of 482 bp and mean length of 318 bp) for *Vibrio*-infected group. These contigs were further assembled into unigenes, producing 59,050 control unigenes (with an N_{50} of 685 bp and mean length of 479 bp) and 73,946 infected group unigenes (with an N_{50} of 829 bp and mean length of 532 bp). The length distribution of control and *Vibrio*-infected contigs and unigenes are shown in Additional file 2. By clustering unigenes from both libraries, a total of 64,411 standard unigenes were produced, with a mean size of 698 bp and an N_{50} of 1137 bp. An overview of the sequencing and assembly is shown in Table 1.

The standard unigenes were annotated by searching the sequences using BLASTX against the NCBI non-redundant, Swiss-Prot, Kyoto Encyclopaedia of Genes and Genome pathway (KEGG) and Orthologous Groups of Proteins (COG) databases, which produced 19,799 (30.73%), 16,832 (26.13%), 14,706 (22.83%) and 7,856 (12.19%) hits respectively, giving a final total of 22,455 significant hits (34.86% of all unigenes). The size distribution profile for the coding sequences (CDS) and identified proteins are shown in Additional file 2. A Gene Ontology (GO) analysis search using the Blast2GO program resulted in 6,007 unigenes (9.32%) being categorized into 55 functional groups. The unigenes without hits using the BLASTX analysis were subjected to an ESTScan, producing 4,977 unigenes (7.82%) predicted to contain coding sequences. The size distribution of the ESTs and proteins are shown in Additional file 2.

Table 1 Summary of the control and *V. parahaemolyticus* infected transcriptome sequencing

	Control	*V. parahaemolyticus* infected
Total number of clean reads	54,708,014	54,295,342
Total base pairs (bp)	4,923,721,260	4,886,580,780
Q20 value	97.73%	97.77%
Total number of contigs	95,645	123,141
Mean length of contigs (bp)	313	318
Total number of unigenes	59,050	73,946
Mean length of unigenes (bp)	479	532
NCBI Nr annotated	19,799	
Swiss-Prot annotated	16,832	
KEGG annotated	14,706	
COG annotated	7,856	
GO annotated	6,007	

The species distribution of the unigenes using the BLASTX results is shown in Figure 1. The *M. rosenbergii* unigenes were matched against *Daphnia pulex* sequences (10.3%), *Tribolium castaneum* (6.1%) and *Pediculus humanus corporis* (4.2%). The unigenes showed a match with those of *D. pulex* and *T. castaneum* probably because of their closer phylogenetic relationship and the availability of vast genomic information. The remaining unigenes (66.5%) which matched were similar to other species due to limited genome information in crustaceans.

Functional annotation

The standard unigenes were analysed using the COG database to classify them and predict their functions. A total of 7,856 unigenes were assigned to COG classifications and functionally classified into 25 protein families which mainly were involved in cellular structure, biochemistry metabolism, molecular processing, and signal transduction (Figure 2). The cluster predicted for general function (3,431, 43.67%) emerged as the largest group, followed by translation, ribosomal structure, biogenesis cluster (1,760, 22.40%) and the replication, recombination, repair clusters (1,413, 17.98%). The clusters with the lowest number of unigenes were nuclear structure and extracellular structures (<1% in each cluster).

The standard unigenes with Nr annotations were subjected to Gene Ontology (GO) analysis, which provides a dynamic controlled vocabulary and hierarchical relationship for the representation of information on molecular functions, cellular components and biological processes, allowing a coherent annotation of gene products. The GO annotations produced 6,007 unigenes (biological process: 2,177 unigenes; cellular component: 1,563 unigenes; and molecular function: 2,267 unigenes) which were assigned to 55 GO ontology sub-categories (Figure 3). In the molecular function category, most of the genes fell into the 'binding' and 'catalytic activity' groups, whereas in the biological process category, the majority of the genes fell into the categories of 'cellular processes', 'metabolic processes' and 'single-organism processes'. Finally, in the cellular component category, a high percentage of the genes fell into the 'cell', 'cell part' and 'organelle' sub-categories.

The Kyoto Encyclopaedia of Genes and Genomes (KEGG) Pathway is a collection of manually drawn pathway maps representing knowledge on molecular interactions and reaction networks. Pathway-based analysis are helpful in identifying biological functions and gene interactions in the pathway. Using the KEGG database, 14,706 unigenes were grouped into 253 pathways. The majority of the unigenes fell into the categories of "metabolic

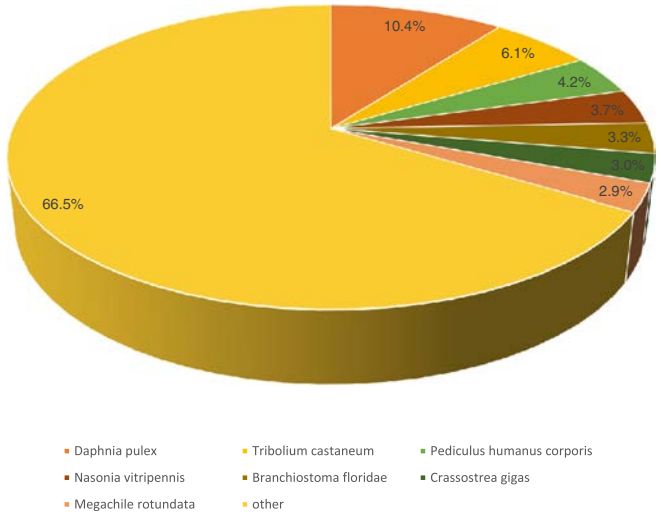


Figure 1 Species distribution of the BLASTX matches of the transcriptome unigenes. This figure shows the species distribution of unigene BLASTX matches against the nr protein database (cutoff value $E<10^{-5}$) and the proportions for each species.

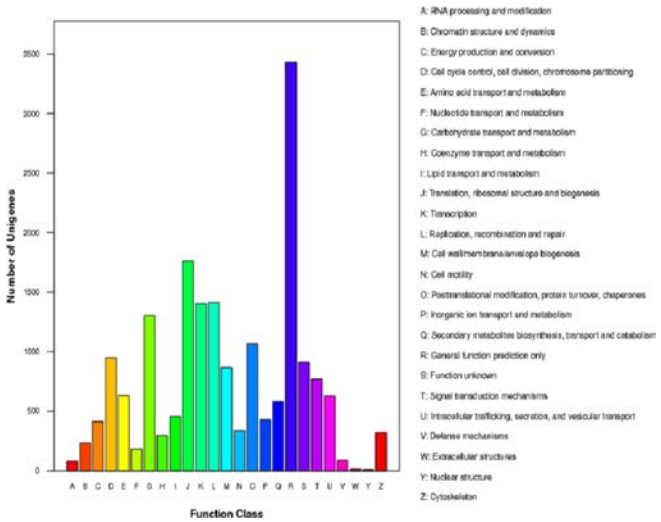


Figure 2 Histogram presentation of Cluster of Orthologus Groups (COG) classification of 7,856 known protein annotated unigenes. Each bar represents the number of unigenes classified into each of the 25 COG functional categories.

pathways" (2192 members, 14.91%), "regulation of actin cytoskeleton" (557 members, 3.79%), "spliceosome" (509 members, 3.46%), "RNA transport" (498 members, 3.38%) and "focal adhesion" (475 members, 3.22% each). The least represented pathways, with less than 10 unigenes categorized in each pathway, were "biotin metabolism", "phenylalanine, tyrosine and tryptophan biosynthesis", "vitamin B6 metabolism", "lipoic acid metabolism" and "thiamine metabolism". The top twenty of these KEGG biological pathway classifications are shown in Figure 4.

Identification of aberrantly expressed genes

We identified differentially expressed genes between these two groups by comparing the relative transcript abundance in each unigene by using the FPKM method (Fragments Per kb per Million fragments). A total of 14,569 unigenes were found to be aberrantly expressed; 11,446 of these were significantly up-regulated, whereas 3,103 unigenes were significantly down-regulated (Figure 5). The differentially expressed genes were annotated against the NR, Swiss-Prot, GO, COG and KEGG databases by BLASTX with a cut-off E-value of 10^{-5}. The annotation analysis is presented in Additional file 3. The 9,469 (65%) unigenes containing low sequence homology to known sequences in public databases could represent non-coding RNA, misassembled unigenes or unknown genes of *M. rosenbergii* which responded to *V. parahaemolyticus*-infection.

For validation of the Illumina sequencing result, seven unigenes were chosen randomly for quantitative real time-PCR (qRT-PCR) analysis. The qRT-PCR results showed similar trends for all the genes to the sequencing data (Figure 6). For example, based on the Illumina sequencing analysis, arginine kinase 1, anti-lipopolysaccharide factor, inhibitor of apoptosis protein, caspase, heat shock protein 21, lectin 1 and NF-kappa B inhibitor alpha were up-regulated 4.67, 4.13, 1.02, 3.08, 4.61, 4.4 and 3.11 log2-fold respectively; and the same elements showed 2.5, 4.4, 1.5, 2.1, 4.8, 3.6 and 2.7 log2-fold change respectively in the qRT-PCR analysis. While the results from these two analyses did not match perfectly, perhaps due to sequencing biases, the qRT-PCR analysis broadly confirmed the direction of change obtained from the Illumina sequencing analysis.

Potential immune-related genes involved in *M. rosenbergii* immune response

Many of the aberrantly expressed genes found in the *V. parahaemolyticus*-infected groups compared to the control group are known to belong to various processes clustered under the animal immune system (Table 2). These immune genes are grouped into 11 functions, including antimicrobial proteins, proteases and proteinases, signal transduction, blood clotting system, cell death, cytoskeletal, heat shock proteins, oxidative stress, pathogen recognition immune receptors, prophenoloxidase system and other immune genes. The role of these groups is described in more detail in the next (Discussion) section.

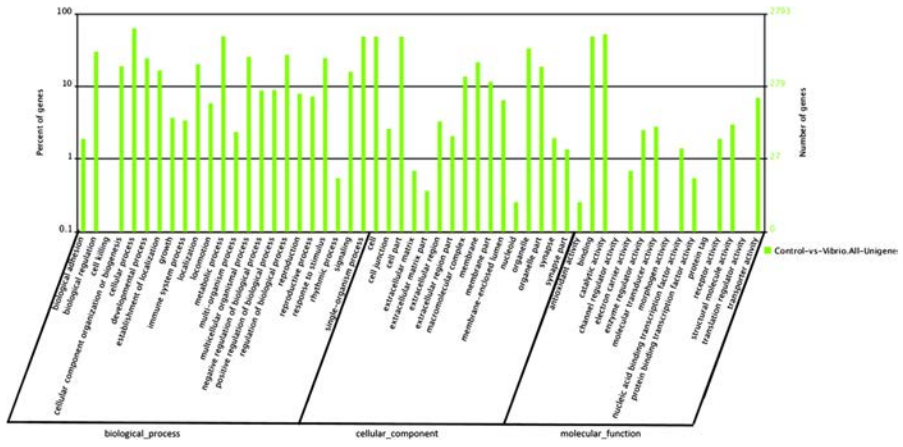


Figure 3 Gene ontology (GO) classification of the 6,007 protein annotated unigenes. Unigenes sequences were systematically classified into GO sub-categories under the biological process, cellular component and molecular function gene ontology catalogue system. Each bar represents the relative abundance of unigenes classified under each sub-category.

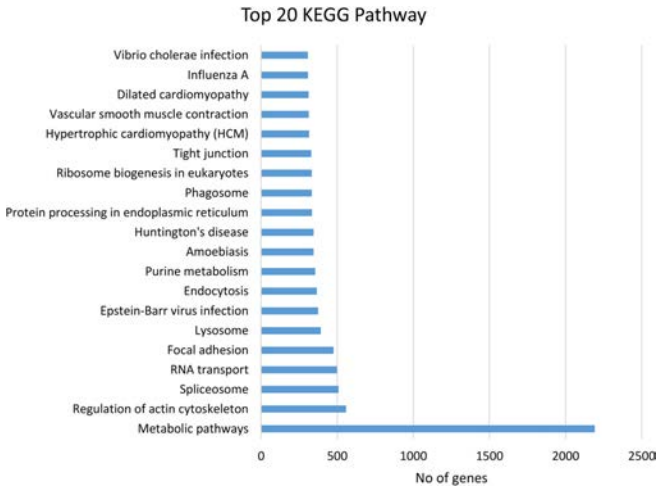


Figure 4 Top 20 KEGG biological pathway classification histograms for annotated unigenes.

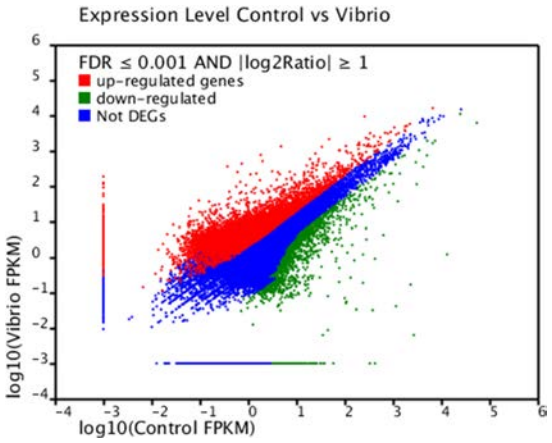


Figure 5 Digital gene expression between control group and *V. parahaemolyticus* infected group. Each point represents a unigene. The x- and y-axis are the log10 of the normalized expression level (FPKM) of unigene between the two groups. Red and green points indicate significant change at the absolute value of log2 (FPKM ratio in two groups) ≥1 and fdr =0.001. Red points indicate up-regulated unigenes and green points indicate down-regulated unigenes in the two groups which its expression level is represented by the y-axis. Blue points indicate insignificant differentially expressed unigenes.

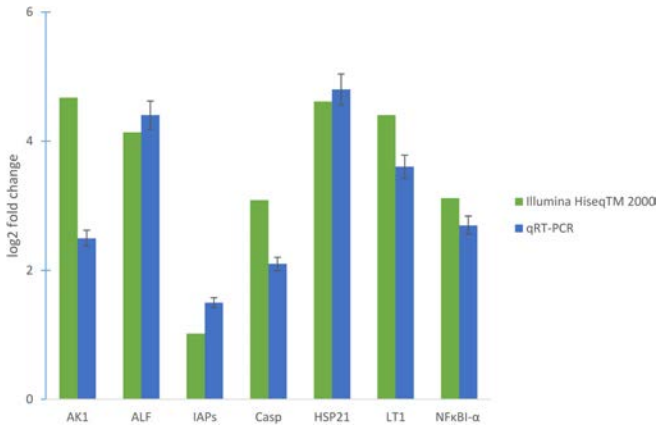


Figure 6 Comparison of the expression profiles of selected genes as determined by Illumina Hiseq ^{TM} 2000 sequencing (green) and qRT-PCR (blue). Target gene abbreviations are as follows: AK1 - arginine kinase, ALF- anti-lipopolysaccharide factor, IAPs- apoptosis inhibitor, Casp- caspase, HSP21- heat shock protein 21, LT1- lectin 1, NFκBI-α - NF-kappa B inhibitor alpha. Error bars indicated standard deviations of averages from three replicates.

Discussion

Apart from viral diseases, *Vibrio* infections causing Vibriosis is another factor hindering the shrimp aquaculture industry worldwide [9]. This fatal disease has contributed to mass mortality and severe economic losses in India, Thailand, Philippines, Japan, Ecuador, Peru, Colombia and Central America [13]. Knowledge about the interaction between *M. rosenbergii* and *Vibrio* species is in its infancy, and in-depth study is urgently needed to address this issue. Discovery of the molecular mechanisms surrounding the innate immune system against *Vibrio* infection in freshwater prawns should be beneficial to both scientific research and the aquaculture industry. Identifying and quantifying immune-related gene expression on a large scale is a promising method to investigate the host response against pathogens and provide a platform for further studies in this area.

Microarray and EST analyses have long been used to study the molecular mechanisms underlying the innate immune system and to identify genes aberrantly expressed during infection [48-50]. However, the most recent NGS platforms such as the Illumina HiSeq™ 2000 appear much better at quantifying transcripts expressed at low levels than microarrays or EST analysis [29]. This is because this revolutionary technique verifies direct transcript profiling without compromise or bias, unlike previous methods [27]. Furthermore, this technology has been successfully used in transcriptome profiling studies on non-model organisms where there is no complete genome database [51-53]. The introduction of NGS technology has led to various studies on host-viral interactions in shrimps to identify potential immune-related genes [54-56] – but not so far on the interaction between the freshwater prawn and *Vibrio* species. To our knowledge, this study could be the very first to use the Illumina HiSeq™ 2000 platform to explore the immune-related gene response in *M. rosenbergii* against *V. parahaemolyticus*.

Taking advantage of the Illumina HiSeq™ 2000 platform's capability to sequence with a high throughput data providing more candidate genes, the total RNA extracted at the 12th hour time point from a pool of control and infected hepatopancreases was sequenced and assembled using the Trinity assembler. The overall analysis yielded 14,549 differentially expressed unigenes, with 11,446 unigenes significantly up-regulated and 3,103 unigenes significantly down-regulated. The sequencing data analyses obtained clearly showed a significant impact of *V. parahaemolyticus* infection on the *M. rosenbergii* transcriptome.

M. rosenbergii possesses an innate immune system, consisting of cellular and humoral components which work individually or cooperatively to protect the species from invading pathogens such as *V. parahaemolyticus* [17,18]. This immune response is activated when the animal detects the invading pathogen through pattern recognition proteins (PRPs) [57]. Two important PRPs molecules identified in shrimp are lectins and the beta-1,3- glucan binding protein (LGBP) [58-61]. Besides pathogen recognition, lectins are also involved in phagocytosis through opsonisation in crustaceans [62]. Our data

Table 2 Candidate genes involved in *M. rosenbergii* immune response against *V. parahaemolyticus*

Category or gene id	Homologues function	Species	FC*
Antimicrobial			
Unigene4120_all	Anti-lipopolysaccharide factor	*Macrobrachium rosenbergii*	4.13
Unigene23546_all	Anti-lipopolysaccharide factor 2	*Macrobrachium rosenbergii*	2.84
CL1276.Contig1_all	Anti-lipopolysaccharide factor 3	*Macrobrachium rosenbergii*	2.15
Unigene37309_all	Crustin	*Macrobrachium rosenbergii*	5.13
Unigene26338_all	Lysozyme	*Portunus trituberculatus*	1.92
Unigene13073_all	NF-kappa B inhibitor alpha	*Macrobrachium rosenbergii*	3.11
Blood clotting system			
Unigene13048_all	Clottable protein	*Marsupenaeus japonicus*	1.42
Unigene34308_all	Transglutaminase	*Macrobrachium rosenbergii*	11.34
Unigene36567_all	Proclotting enzyme	*Harpegnathos saltator*	11.22
Unigene41253_all	Coagulation factor XII	*Hageman factor*	10.61
PRPs			
Unigene10978_all	Lectin 1	*Macrobrachium rosenbergii*	4.4
CL4516.Contig1_all	Lectin 2	*Macrobrachium rosenbergii*	1.05
Unigene7825_all	Lectin 3	*Macrobrachium rosenbergii*	2.54
CL3039.Contig4_all	Lectin 4	*Macrobrachium rosenbergii*	4.55
Unigene9635_all	C-type lectin	*Penaeus monodon*	1.14
CL1600.Contig1_all	C-type lectin-2	*Litopenaeus vannamei*	1.69
CL5230.Contig3_all	C-type lectin 5	*Fenneropenaeus chinensis*	2
Unigene1391_all	Hemolectin	*Papilio xuthus*	5.81
Unigene25701_all	M-type lectin	*Marsupenaeus japonicus*	1.69
Unigene23215_all	Perlucin-like protein	*Mytilus galloprovincialis*	−1.92
CL1075.Contig4_all	Tachylectin	*Macrobrachium rosenbergii*	−1.71
Unigene28944_all	Ficolin	*Branchiostoma floridae*	1.31
Unigene19793_all	Lectin B isoform 2	*Marsupenaeus japonicus*	3.55
Unigene37041_all	Mannose-binding protein	*Procambarus clarkii*	11.69
CL1124.Contig1_all	Glucan pattern-recognition lipoprotein	*Fenneropenaeus chinensis*	1.12
Unigene23671_all	lipopolysaccharide and beta-1,3-glucan binding protein	*Macrobrachium rosenbergii*	1.71
Proteinases and proteinases inhibitors			
Unigene1509_all	Cathepsin B	*Pandalus borealis*	1.87
Unigene22859_all	Cathepsin C	*Fenneropenaeus chinensis*	2.33
CL1587.Contig2_all	Cathepsin D	*Penaeus monodon*	−1.56
Unigene26111_all	Cathepsin L	*Pandalus borealis*	1.41
CL5824.Contig1_all	Serpin serine protease inhibitor	*Fenneropenaeus chinensis*	12.24
CL6038.Contig1_all	26S protease regulatory	*Nasonia vitripennis*	1.44
Unigene12869_all	Alpha-2-macroglobulin	*Macrobrachium rosenbergii*	5.09
CL2365.Contig1_all	Caspase	*Marsupenaeus japonicus*	1.04
Unigene13293_all	Astacin	*Strongylocentrotus purpuratus*	1.64
Unigene13633_all	Serine protease	*Fenneropenaeus chinensis*	3.08
Unigene17259_all	Caspase 8	*Ictalurus punctatus*	2.41
CL2565.Contig1_all	Serine proteinase inhibitor 6	*Penaeus monodon*	4.51
CL1118.Contig1_all	Serpin B	*Marsupenaeus japonicus*	7.02
Unigene26757_all	Hemocyte kazal-type proteinase inhibitor	*Penaeus monodon*	3.87

Table 2 Candidate genes involved in *M. rosenbergii* immune response against *V. parahaemolyticus* (Continued)

CL1127.Contig2_all	Kazal-type serine proteinase inhibitor 4	*Procambarus clarkii*	−6.01
CL5487.Contig2_all	Aminopeptidase N	*Camponotus floridanus*	1.4
Unigene5621_all	Masquerade-like serine proteinase-like protein 2	*Penaeus monodon*	5.39
Unigene21512_all	Serine proteinase inhibitor	*Macrobrachium rosenbergii*	6.91
Unigene28987_all	CUB-serine protease	*Panulirus argus*	2.92
Heat shock proteins			
Unigene3736_all	Heat shock protein 21	*Macrobrachium rosenbergii*	4.61
Unigene23034_all	Heat shock protein 40	*Frankliniella occidentalis*	1.8
Unigene14757_all	Heat shock protein 70	*Portunus trituberculatus*	2.26
CL4309.Contig3_all	Heat shock protein 90	*Scylla paramamosain*	1.93
Unigene16858_all	Small heat shock protein	*Fenneropenaeus chinensis*	2.41
Oxidative stress			
CL353.Contig2_all	Glutathione S transferase	*Procambarus clarkii*	1.01
Unigene19405_all	Cu/Zn superoxide dismutase	*Bombyx mori*	1.61
CL5590.Contig3_all	Glutamine synthetase	*Panulirus argus*	−3.03
CL559.Contig1_all	Farnesoic acid O-methyltransferase	*Nilaparvata lugens*	1.81
CL2477.Contig1	Catalase	*Litopenaeus vannamei*	1.6
CL2787.Contig2	Thioredoxin reductase	*Branchiostoma floridae*	1.09
Cytoskeletal			
Unigene10471	Chitinase	*Trichomonas vaginalis*	6.16
CL6026.Contig5_all	Actin	*Marsupenaeus japonicus*	2.92
Unigene13021_all	Calponin	*Chironomus riparius*	2.32
CL5710.Contig1_all	Profilin	*Fenneropenaeus chinensis*	1.45
CL461.Contig8_all	Tubulin	*Homarus americanus*	1.39
Unigene19173_all	Beta-integrin	*Fenneropenaeus chinensis*	2.78
CL2046.Contig1_all	Integrin	*Litopenaeus vannamei*	3.03
Signal transduction			
Toll pathway			
CL1915.Contig5_all	Toll interacting protein	*Marsupenaeus japonicus*	1.73
CL5402.Contig1_all	Toll receptor 2	*Marsupenaeus japonicus*	−2.46
Unigene25507_all	Toll-like receptor 3	*Daphnia pulex*	2.59
CL2421.Contig1_all	Toll-like receptor 8	*Ixodes scapulari*	1.48
Unigene25799_all	Myd88 protein	*Artemia sinica*	3.25
Unigene9776_all	Caspase	*Fenneropenaeus merguiensis*	3.08
Unigene14481_all	Dorsal	*Fenneropenaeus chinensis*	3.19
Wnt signaling pathway			
Unigene36938_all	Lipoprotein receptor	*Callinectes sapidus*	10.94
Unigene15960_all	Prickle	*Saccoglossus kowalevskii*	1.71
Unigene15970_all	Low-density lipoprotein receptor-related protein 6	*Daphnia pulex*	2.05
Unigene12737_all	Glypican 4	*Daphnia pulex*	1.14
Unigene25938_all	Secreted frizzled-related protein 5	*Daphnia pulex*	1.82
MAPK signaling pathway			
Unigene26817_all	Mitogen-activated protein kinase kinase kinase 1	*Oryctolagus cuniculus*	2.79
Unigene110_all	Mitogen-activated protein kinase 8 interacting protein 3	*Bombus terrestris*	2.94

Table 2 Candidate genes involved in *M. rosenbergii* immune response against *V. parahaemolyticus* *(Continued)*

CL726.Contig3_all	Mitogen-activated protein kinase kinase kinase kinase 4	*Bombus terrestris*	1.65
CL3174.Contig1_all	Mitogen-activated protein kinase kinase kinase 7	*Danaus plexippus*	1.41
Unigene29569_all	Mitogen-activated protein kinase kinase kinase 13	*Bombus terrestris*	4.03
CL4432.Contig2_all	Max protein	*Daphnia pulex*	−3.11
Unigene14757_all	Heat shock protein 70	*Portunus trituberculatus*	2.26
Unigene3873_all	Protein phosphatase 5	*Bombus terrestris*	1.21
Unigene102_all	Map kinase-interacting serine/threonine	*Scylla paramamosain*	2.48
Unigene11596_all	Raf homolog serine/threonine-protein kinase phl-like	*Nasonia vitripennis*	2.41
JAK/STAT pathway			
Unigene25562_all	STAT long form	*Penaeus monodon*	2.84
Unigene25762_all	Signal transducing adaptor molecule	*Megachile rotundata*	1.98
Unigene16474_all	Tumor susceptibility gene 101 protein	*Harpegnathos saltator*	2.46
CL1133.Contig6_all	Domeless	*Tribolium castaneum*	2.94
Unigene28876_all	Thyroid peroxidase	*Pediculus humanus corporis*	3.66
Other signal transduction genes			
CL1409.Contig1_all	Innexin 3	*Cancer borealis*	10.29
Unigene30035_all	cAMP-dependent protein kinase type II regulatory subunit	*Camponotus floridanus*	1.05
CL3406.Contig2_all	Casein kinase II subunit alpha	*Danaus plexippus*	1.75
CL4247.Contig1_all	Rab-protein 14	*Tribolium castaneum*	1.85
Unigene37520_all	Afadin	*Pediculus humanus corporis*	2.89
Unigene34256_all	TBC1 domain family member 10B	*Crassostrea gigas*	11.52
Unigene33128_all	REM2- and Rab-like small GTPase 1-like	*Cavia porcellus*	3.21
ProPO system			
Unigene12734_all	Prophenoloxidase	*Macrobrachium rosenbergii*	3.15
Unigene7353_all	Prophenoloxidase activating factor	*Fenneropenaeus chinensis*	3.79
Unigene18337_all	Prophenoloxide activating enzyme III	*Macrobrachium rosenbergii*	4.99
Cell death			
Unigene2555_all	Beclin 1	*Megachile rotundata*	11.89
CL1843.Contig2_all	ALG-2 interacting protein x *Penaeus monodon*	*Penaeus monodon*	3.2
Unigene14045_all	Program cell death 5-like	*Penaeus monodon*	1.61
CL2477.Contig1_all	Catalase	*Litopenaeus vannamei*	1.6
Unigene3163_all	inhibitor of apoptosis protein	*Litopenaeus vannamei*	1.02
Unigene9776_all	Caspase	*Fenneropenaeus merguiensis*	3.08
Unigene36052_all	DNA fragmentation factor subunit beta	*Danio rerio*	3.26
Other immune genes			
CL4920.Contig1_all	Calmodulin	*Procambarus clarkii*	1.92
CL1939.Contig4_all	Ferritin	*Litopenaeus vannamei*	−2.87
CL4071.Contig2_all	Peritrophin	*Macrobrachium nipponense*	−2.18
Unigene10835_all	Selenoprotein W	*Cricetulus griseus*	−1.98
Unigene26739_all	Metallothionein I	*Macrobrachium rosenbergii*	−2.75
Unigene29279_all	Selenoprotein L	*Ciona intestinalis*	1.06
CL6324.Contig2_all	Arginine kinase 1	*Macrobrachium rosenbergii*	4.67
CL5294.Contig2_all	Hemocyanin	*Macrobrachium nipponense*	1.21
Unigene26326_all	Tetraspanin-like protein CD9	*Fenneropenaeus chinensis*	1.02
CL4316.Contig1_all	Crustacyanin-like lipocalin	*Macrobrachium rosenbergii*	−4.12

Table 2 Candidate genes involved in *M. rosenbergii* immune response against *V. parahaemolyticus* (Continued)

CL6297.Contig2_all	E cadherin	*Tenebrio molitor*	3.19
CL2339.Contig2_all	Adenosine deaminase	*Caligus rogercresseyi*	−1.35
CL3196.Contig2_all	Calnexin	*Penaeus monodon*	1.57
CL2701.Contig1_all	Ubiquitin-conjugating enzyme	*Macrobrachium nipponense*	1.33

*Fold changes (Log2 ratio) in gene expression. PRPs- pattern recognition proteins, ProPO- prophenoloxidase.

showed that challenge by *V. parahaemolyticus* greatly affects the expression of PRPs, as observed in previous studies using different *Vibrio* strains [63-67]. The activation of LGBP molecules triggers the melanisation process, a prophenoloxidase-activating system (proPO-AS) which is an enzymatic cascade involving several enzymes, including the key enzyme phenoloxidase (PO) [68-70]. The active PO converts phenols into quinones. These build a non-specific crosslink between neighbouring molecules to form melanin, which provides defence against invading micro-organisms [71]. Increased activity of prophenoloxidase against *Vibrio* species has also been noted in *Fenneropenaeus indicus* [72], *L. vannamei* [73] and *P. monodon* [74].

Antimicrobial peptides (AMPs) play a pivotal role in killing or clearing infected pathogens, especially *Vibrio* species [75]. Notable shrimp AMPs, such as penaeidins, lysozymes, crustins, anti-lipopolysaccharide factors (ALFs) and stylicins have been identified and characterized previously in shrimps [76-80]. However, only four types of AMPs – lysozyme, crustin, NF-kappa B inhibitor alpha and ALF (3 isoforms) – were detected in our transcriptome data and found to be highly expressed. The up-regulation of these AMPs correlated with previous studies showing their antimicrobial properties against *Vibrio* species and other bacteria [81-84]. Blood clotting is vital in crustaceans to prevent excess blood loss from a wound and prevent micro-organisms from invading the wound [85]. We found four molecules of the blood clotting system – transglutaminase, clottable protein, proclotting enzyme and coagulation factor XII – to be highly induced in our transcriptome data after challenge by *V. parahaemolyticus*. A similar expression of these molecules after bacterial challenge has been reported in previous studies [72,86,87].

Stress conditions such as bacterial infections lead to an accumulation of reactive oxygen species (ROS) in a cell [88]. Increased levels of ROS causes oxidative damage to important cellular macromolecules (lipids, proteins, carbohydrates and nucleotides) which are components of the membranes, cellular enzymes and DNA [89]. In order to restrict the production of ROS, antioxidant genes are activated to produce antioxidant enzymes which eliminate ROS. Several antioxidant enzymes have been isolated and characterised in the penaeid shrimp in previous studies [90-92]. In this study, we found six antioxidant unigenes to be over-expressed after *V. parahaemolyticus* challenge – the exception being glutamine synthetase. High expression levels of these genes have similarly been observed in other shrimps and scallops after *Vibrio* challenge [93-96]. The up-regulation of actin and tubulin genes play a crucial role in a wide range of cellular functions such as nodule formation, phagocytosis, encapsulation, as well as cell shape change, cell motility and adhesion, all of which may aid in clearing the pathogen [97].

Heat shock proteins (HSPs) are highly generated when induced by stress. They are known to play a major role in protein folding, the protection of proteins from denaturation or aggregation, and aiding protein transport through membrane channels [98,99]. In addition to molecular chaperones, HSPs have been reported to play important roles in innate immune responses, and have been well studied in crustaceans [100-102]. In this study, we noted higher levels of expression of all heat shock proteins in *M. rosenbergii* when challenged by *V. parahaemolyticus*. The increased expression of these HSPs is in line with previous reports, which tends to confirm the important role of these proteins in protecting this species from the stress induced by *Vibrio* challenge [97,103,104]. The general higher expression of proteinases and their inhibitors was to be expected in our data, as these are known to modulate elements of the innate immune system such as haemolymph coagulation, antimicrobial peptide synthesis, cell adhesion, and melanisation [105].

Bacteria like *Vibrio* are known to induce cell apoptosis through a variety of mechanisms such as pore-forming proteins, secretion of protein synthesis inhibitors, molecules activating the endogenous death machinery in an infected cell, lipopolysaccharides, and other superantigens [106]. Increased levels of apoptosis contribute to the degradation of DNA and RNA, which may contribute to shrimp mortality [107]. In our transcriptome data, an up-regulation of genes involved in apoptosis was observed, similar to the trends reported in previous studies [74,108,109]. Apoptosis may also serve as host defence against bacterium by allowing other healthy cells to phagocytise apoptotic bodies containing bacteria from target cells, effectively clearing the pathogen [110].

The signalling pathways involved in the *M. rosenbergii* innate immune response against *V. parahaemolyticus*

were observed to be highly induced in our transcriptome data. The Toll protein initially identified in *Drosophila* had been reported to play a key role in the anti-fungal and anti-Gram-positive bacterial responses of flies in the Toll pathway [111]. Other Toll components have also been found to be activated in penaeid shrimps when challenged with *Vibrio* species [109,112,113]. In *Caenorhabditis elegans*, the mitogen-activated protein kinase (MAPK) pathways are transcriptionally up-regulated by the pore-forming toxin released by *Bacillus thuringiensis*, which provide a cellular defence against this toxin [114]. This could explain the higher expression of this signalling pathway in our data, as *V. parahaemolyticus* is known to release a thermostable direct haemolysin, which is a pore-forming toxin [115]. The Janus kinase (JAK) and signal transducer and activator of transcription (STAT) pathways have been reported to be activated when *Fenneropenaeus chinensis* is challenged with *Vibrio anguillarum,* which suggest that these pathways are important for immune responses against bacteria [116]. In addition, Rab-related proteins have been reported to regulate the hemocytic phagocytosis of bacteria in *Marsupenaeus japonicas* [117].

Several genes in the other immune gene group were found to be aberrantly expressed in our transcriptome data. Calmodulin, which plays an important role in calcium-dependent signal transduction pathways, was over-expressed in our data – as was the case in *L. vannamei* when challenged with *V. parahaemolyticus* [118]. Ferritin, an iron storage protein crucial for the metabolism of iron and maintaining iron homeostasis in a cell, was found to be down-regulated. The reduced expression of this gene could possibly lead to prawn mortality, as increased expression of this gene has been found to protect *P. monodon* from *Vibrio harveyi* [119]. Arginine kinase (AK), a phosphagen kinase in the invertebrate energy metabolism, has previously been reported to play an immune role against viral infection [32]. However, we observed a higher expression of this gene in *M. rosenbergii* after challenge by *V. parahaemolyticus,* which could suggest that AK plays a similar role in bacterial infection. Haemocyanin, an important immune gene in crustaceans, is involved in prophenoloxidase activity [120]. It has antiviral properties against WSSV [121], and its increased expression in our transcriptome data tends to bear out its importance as a defence molecule against challenge by *V. parahaemolyticus*. Finally, metallothioneins, a metal-binding protein, was found to be highly expressed in our transcriptome data. The increased expression of this gene was to be expected, as it is known as a scavenger of reactive oxygen intermediates and generally shows higher expression levels during immune responses in invertebrates against pathogens [49].

Conclusion

We utilised the Illumina HiSeq™ 2000 platform and Trinity assembler package to perform a *de novo* transcriptome profiling of the hepatopancreases isolated from *M. rosenbergii* challenged with *V. parahaemolyticus.* The differential expression analysis between *V. parahaemolyticus*-infected and control groups revealed significant differences in the gene expression with 11,446 unigenes found to be significantly up-regulated and 3,103 unigenes observed to be significantly down-regulated. This study provided a valuable insight into antibacterial mechanisms of freshwater prawn against *V. parahaemolyticus* with majority of the differentially expressed unigenes were grouped into 11 animal immune system categories. Furthermore, this study has generated an abundant list of transcript from *M.rosenbergii* which will provide a fundamental basis for future genomics research in this field.

Additional files

Additional file 1: Table S1. List of genes, primers and probes used in real-time TaqMan PCR assay.

Additional file 2: Figure S1. Overview of the control and *V. parahaemolyticus* infected transcriptome assembly. **(A)** The size distribution of the contigs obtained from our *denovo* assembly of high-quality clean reads. **(B)** The size distribution of the unigenes produced from further assembly of contigs (i.e., contig joining, gap filling, and scaffold clustering). **(C)** Size distributions of the coding sequences (CDS) and identified proteins and **(D)** Size distributions of the ESTs and proteins obtained from the ESTScan results. For unigene CDS that had no hits in the databases (Nr, SwissProt, KEGG and COG), the BLAST results were subjected to ESTScans and then translated into peptide sequences.

Additional file 3: Table S2. List of differentially expressed genes with Nr, Nt, Swissprot, GO, COG, KEGG.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
Conceived and designed the experiments: JW SB. Performed the experiments: RR TA. Analysed the data: RR ST YBZ. Contributed reagents/materials/analysis tools: SB and JW. Wrote the paper: RR, JW, SB and KLT. All authors read and approved the final manuscript.

Acknowledgements
This work was supported by a Flagship Grant No: FL002-201 granted by Cluster-Bio, University of Malaya and Postgraduate Research Grant (PPP) of the University of Malaya, Malaysia (PG088-2012B) and publication assisted by Grant No 57-02-03-1015.

Author details
[1]Genomic Research and Breeding Laboratory and Centre for Research in Biotechnology for Agriculture (CEBAR), Institute of Biological Sciences, Faculty of Science, University of Malaya, 50603 Kuala Lumpur, Malaysia. [2]Beijing Genomics Institute, Shenzhen, 11th Floor, Main Building, Beishan, Industrial Zone, Yantian District, Shenzhen 518083, China. [3]Microbiology Unit, Institute of Biological Sciences, Faculty of Science, University of Malaya, 50603 Kuala Lumpur, Malaysia.

Received: 13 November 2014 Accepted: 13 February 2015
Published online: 14 March 2015

References

1. De Grave S, Cai Y, Anker A. Global diversity of shrimps (Crustacea: Decapoda: Caridea) in freshwater. Hydrobiologia 2008 (595):287–293.
2. New MB. Farming freshwater prawns: a manual for the culture of the giant river prawn (Macrobrachium rosenbergii): Food & Agriculture Org. 2002;(148):1-11.
3. Wowor D, Muthu V, Meier R, Balke M, Cai Y, Ng PK. Evolution of life history traits in Asian freshwater prawns of the genus Macrobrachium(Crustacea: Decapoda: Palaemonidae) based on multilocus molecular phylogenetic analysis. Mol Phylogenet Evol. 2009;52(2):340–50.
4. New MB. History and global status of freshwater prawn farming. *Freshwater Prawns: Biology and Farming* 2009(194):16–40.
5. Schwantes VS, Diana JS, Yi Y. Social, economic, and production characteristics of giant river prawn Macrobrachium rosenbergii culture in Thailand. Aquaculture. 2009;287(1):120–7.
6. New MB. Freshwater prawn farming: global status, recent research and a glance at the future. Aquacult Res. 2005;36(3):210–30.
7. Nash G, Chinabut S, Limsuwan C. Idiopathic muscle necrosis in the freshwater prawn, Macrobrachium rosenbergii de Man, cultured in Thailand. J Fish Dis. 1987;10(2):109–20.
8. Bonami J-R, Sri Widada J. Viral diseases of the giant fresh water prawn Macrobrachium rosenbergii: a review. J Invertebr Pathol. 2011;106(1):131–42.
9. Tonguthai K. Diseases of the freshwater prawn, Macrobrachium rosenbergii, The Aquat. Anim Health Res Inst Newsl. 1995;4:1–4.
10. Ramesh A, Loganathan B, Venkateswaran K. Ecological dynamics of marine luminous bacteria. J Basic Microbiol. 1990;30(9):689–703.
11. Thompson FL, Iida T, Swings J. Biodiversity of vibrios. Microbiol Mol Biol Rev. 2004;68(3):403–31.
12. Khuntia CP, Das BK, Samantaray BR, Samal SK, Mishra BK. Characterization and pathogenicity studies of Vibrio parahaemolyticus isolated from diseased freshwater prawn, Macrobrachium rosenbergii (de Man). Aquacult Res. 2008;39(3):301–10.
13. Lightner DV. A handbook of shrimp pathology and diagnostic procedures for diseases of cultured penaeid shrimp. 1996.
14. Ruangpan L, Kitao T. Vibrio bacteria isolated from black tiger shrimp, Penaeus monodon Fabricius. J Fish Dis. 1991;14(3):383–8.
15. Xu B, Xu H, Ji W, Shi J. Pathogens and pathogenicity to Penaeus orientalis Kishinouye. Acta Oceanol Sinic. 1994;13(2):297–304.
16. Kimbrell DA, Beutler B. The evolution and genetics of innate immunity. Nat Rev Genet. 2001;2(4):256–67.
17. Jiravanichpaisal P, Lee BL, Söderhäll K. Cell-mediated immunity in arthropods: hematopoiesis, coagulation, melanization and opsonization. Immunobiology. 2006;211(4):213–36.
18. Young Lee S, Söderhäll K. Early events in crustacean innate immunity. Fish Shellfish Immunol. 2002;12(5):421–37.
19. Zhao Z-Y, Yin Z-X, Weng S-P, Guan H-J, Li S-D, Xing K, et al. Profiling of differentially expressed genes in hepatopancreas of white spot syndrome virus-resistant shrimp (Litopenaeus vannamei) by suppression subtractive hybridisation. Fish Shellfish Immunol. 2007;22(5):520–34.
20. Pan D, He N, Yang Z, Liu H, Xu X. Differential gene expression profile in hepatopancreas of WSSV-resistant shrimp (Penaeus japonicus) by suppression subtractive hybridization. Dev Comp Immunol. 2005;29(2):103–12.
21. Leu J-H, Chen S-H, Wang Y-B, Chen Y-C, Su S-Y, Lin C-Y, et al. A review of the major penaeid shrimp EST studies and the construction of a shrimp transcriptome database based on the ESTs from four penaeid shrimp. Marine Biotechnol. 2011;13(4):608–21.
22. Morozova O, Hirst M, Marra MA. Applications of new sequencing technologies for transcriptome analysis. Annu Rev Genomics Hum Genet. 2009;10:135–51.
23. Metzker ML. Sequencing technologies—the next generation. Nat Rev Genet. 2010;11(1):31–46.
24. Ansorge WJ. Next-generation DNA sequencing techniques. N Biotechnol. 2009;25(4):195–203.
25. Varshney RK, Nayak SN, May GD, Jackson SA. Next-generation sequencing technologies and their implications for crop genetics and breeding. Trends Biotechnol. 2009;27(9):522–30.
26. Martin JA, Wang Z. Next-generation transcriptome assembly. Nat Rev Genet. 2011;12(10):671–82.
27. Reis-Filho JS. Next-generation sequencing. Breast Cancer Res. 2009;11 Suppl 3:S12.
28. Mutz K-O, Heilkenbrinker A, Lönne M, Walter J-G, Stahl F. Transcriptome analysis using next-generation sequencing. Curr Opin Biotechnol. 2013;24(1):22–30.
29. Asmann YW, Wallace MB, Thompson EA. Transcriptome profiling using next-generation sequencing. Gastroenterology. 2008;135(5):1466–8.
30. Mohd-Shamsudin MI, Kang Y, Lili Z, Tan TT, Kwong QB, Liu H, et al. In-depth tanscriptomic analysis on giant freshwater prawns. PLoS One. 2013;8(5):e60839.
31. Tiruvayipati S, Bhassu S, Kumar N, Baddam R, Shaik S, Gurindapalli AK, et al. Genome anatomy of the gastrointestinal pathogen, Vibrio parahaemolyticus of crustacean origin. Gut Pathog. 2013;5(1):37.
32. Arockiaraj J, Vanaraja P, Easwvaran S, Singh A, Alinejaid T, Othman RY, et al. Gene profiling and characterization of arginine kinase-1 (MrAK-1) from freshwater giant prawn (Macrobrachium rosenbergii). Fish Shellfish Immunol. 2011;31(1):81–9.
33. Arockiaraj J, Easwvaran S, Vanaraja P, Singh A, Othman RY, Bhassu S. Molecular cloning, characterization and gene expression of an antioxidant enzyme catalase (Mr Cat) from Macrobrachium rosenbergii. Fish Shellfish Immunol. 2012;32(5):670–82.
34. Arockiaraj J, Easwvaran S, Vanaraja P, Singh A, Othman RY, Bhassu S. Immunological role of thiol-dependent peroxiredoxin gene in Macrobrachium rosenbergii. Fish Shellfish Immunol. 2012;33(1):121–9.
35. Arockiaraj J, Vanaraja P, Easwvaran S, Singh A, Othman RY, Bhassu S. Bioinformatic characterization and gene expression pattern of apoptosis inhibitor from Macrobrachium rosenbergii challenged with infectious hypodermal and hematopoietic necrosis virus. Fish Shellfish Immunol. 2011;31(6):1259–67.
36. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013;8(8):1494–512.
37. Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, et al. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics. 2003;19(5):651–2.
38. Boeckmann B, Bairoch A, Apweiler R, Blatter M-C, Estreicher A, Gasteiger E, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 2003;31(1):365–70.
39. Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 2000;28(1):33–6.
40. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30.
41. Mount DW. Using the basic local alignment search tool (BLAST). Cold Spring Harb Protoc. 2007;2007(7):pdb. top17.
42. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000;25(1):25–9.
43. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005;21(18):3674–6.
44. Iseli C, Jongeneel CV, Bucher P. ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. ISMB. 1999;1999:138–48.
45. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5(7):621–8.
46. Reiner A, Yekutieli D, Benjamini Y. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics. 2003;19(3):368–75.
47. Schmittgen TD, Livak KJ. Analyzing real-time PCR data by the comparative CT method. Nat Protoc. 2008;3(6):1101–8.
48. Aoki T, Wang H-C, Unajak S, Santos MD, Kondo H, Hirono I. Microarray analyses of shrimp immune responses. Marine Biotechnol. 2011;13(4):629–38.
49. Gross P, Bartlett T, Browdy C, Chapman R, Warr G. Immune gene discovery by expressed sequence tag analysis of hemocytes and hepatopancreas in the Pacific White Shrimp, Litopenaeus vannamei, and the Atlantic White ShrimpL. setiferus. Dev Comp Immunol. 2001;25(7):565–77.
50. Supungul P, Klinbunga S, Pichyangkura R, Jitrapakdee S, Hirono I, Aoki T, et al. Identification of immune-related genes in hemocytes of black tiger shrimp (Penaeus monodon). Marine Biotechnol. 2002;4(5):487–94.
51. Xu J, Ji P, Wang B, Zhao L, Wang J, Zhao Z, et al. Transcriptome Sequencing and Analysis of Wild Amur Ide (Leuciscus waleckii) Inhabiting an Extreme Alkaline-Saline Lake Reveals Insights into Stress Adaptation. PloS One. 2013;8(4):e59703.

52. Feldmeyer B, Wheat CW, Krezdorn N, Rotter B, Pfenninger M. Short read Illumina data for the de novo assembly of a non-model snail species transcriptome (Radix balthica, Basommatophora, Pulmonata), and a comparison of assembler performance. BMC Genomics. 2011;12(1):317.
53. Garg R, Patel RK, Tyagi AK, Jain M. De novo assembly of chickpea transcriptome using short reads for gene discovery and marker identification. DNA Res. 2011;18(1):53–63.
54. Zeng D, Chen X, Xie D, Zhao Y, Yang C, Li Y, et al. Transcriptome analysis of pacific white shrimp (Litopenaeus vannamei) hepatopancreas in response to Taura Syndrome Virus (TSV) experimental infection. PLoS One. 2013;8(2):e57515.
55. Chen X, Zeng D, Chen X, Xie D, Zhao Y, Yang C, et al. Transcriptome analysis of Litopenaeus vannamei in response to white spot syndrome virus infection. PLoS One. 2013;8(8):e73218.
56. Xue S, Liu Y, Zhang Y, Sun Y, Geng X, Sun J. Sequencing and De Novo Analysis of the Hemocytes Transcriptome in Litopenaeus vannamei Response to White Spot Syndrome Virus Infection. PLoS One. 2013;8(10):e76718.
57. Akira S, Uematsu S, Takeuchi O. Pathogen recognition and innate immunity. Cell. 2006;124(4):783–801.
58. Kumar H, Kawai T, Akira S. Pathogen recognition by the innate immune system. Int Rev Immunol. 2011;30(1):16–34.
59. Liu Y-C, Li F-H, Dong B, Wang B, Luan W, Zhang X-J, et al. Molecular cloning, characterization and expression analysis of a putative C-type lectin (Fclectin) gene in Chinese shrimp Fenneropenaeus chinensis. Mol Immunol. 2007;44(4):598–607.
60. Ma THT, Tiu SHK, He J-G, Chan S-M. Molecular cloning of a C-type lectin (LvLT) from the shrimp Litopenaeus vannamei: Early gene down-regulation after WSSV infection. Fish Shellfish Immunol. 2007;23(2):430–7.
61. Sritunyalucksana K, Lee SY, Söderhäll K. A β-1, 3-glucan binding protein from the black tiger shrimp, Penaeus monodon. Dev Comp Immunol. 2002;26(3):237–45.
62. Marques MRF, Barracco MA. Lectins, as non-self-recognition factors, in crustaceans. Aquaculture. 2000;191(1):23–44.
63. Adams A. Response of penaeid shrimp to exposure to Vibrio species. Fish Shellfish Immunol. 1991;1(1):59–70.
64. Sun Y-D, Fu L-D, Jia Y-P, Du X-J, Wang Q, Wang Y-H, et al. A hepatopancreas-specific C-type lectin from the Chinese shrimp Fenneropenaeus chinensis exhibits antimicrobial activity. Mol Immunol. 2008;45(2):348–61.
65. Cheng W, Liu C-H, Tsai C-H, Chen J-C. Molecular cloning and characterisation of a pattern recognition molecule, lipopolysaccharide-and β-1, 3-glucan binding protein (LGBP) from the white shrimp Litopenaeus vannamei. Fish Shellfish Immunol. 2005;18(4):297–310.
66. Liu F, Li F, Dong B, Wang X, Xiang J. Molecular cloning and characterisation of a pattern recognition protein, lipopolysaccharide and β-1, 3-glucan binding protein (LGBP) from Chinese shrimp Fenneropenaeus chinensis. Mol Biol Rep. 2009;36(3):471–7.
67. Soonthornchai W, Rungrassamee W, Karoonuthaisiri N, Jarayabhand P, Klinbunga S, Söderhäll K, et al. Expression of immune-related genes in the digestive organ of shrimp, Penaeus monodon, after an oral infection by Vibrio harveyi. Dev Comp Immunol. 2010;34(1):19–28.
68. Vargas-Albores F, Yepiz-Plascencia G. Beta glucan binding protein and its role in shrimp immune response. Aquaculture. 2000;191(1):13–21.
69. Soltanian S, Stuyven E, Cox E, Sorgeloos P, Bossier P. Beta-glucans as immunostimulant in vertebrates and invertebrates. Crit Rev Microbiol. 2009;35(2):109–38.
70. Amparyup P, Sutthangkul J, Charoensapsri W, Tassanakajon A. Pattern recognition protein binds to lipopolysaccharide and β-1, 3-glucan and activates shrimp prophenoloxidase system. J Biol Chem. 2012;287(13):10060–9.
71. Perazzolo LM, Barracco MA. The prophenoloxidase activating system of the shrimp Penaeus paulensis and associated factors. Dev Comp Immunol. 1997;21(5):385–95.
72. Sarathi M, Ahmed V, Venkatesan C, Balasubramanian G, Prabavathy J, Hameed A. Comparative study on immune response of Fenneropenaeus indicus to Vibrio alginolyticus and white spot syndrome virus. Aquaculture. 2007;271(1):8–20.
73. Liu C-H, Yeh S-T, Cheng S-Y, Chen J-C. The immune response of the white shrimp Litopenaeus vannamei and its susceptibility to Vibrio infection in relation with the moult cycle. Fish Shellfish Immunol. 2004;16(2):151–61.
74. Nayak S, Singh S, Ramaiah N, Sreepada R. Identification of upregulated immune-related genes in Vibrio harveyi challenged Penaeus monodon postlarvae. Fish Shellfish Immunol. 2010;29(3):544–9.
75. Bachère E. Anti-infectious immune effectors in marine invertebrates: potential tools for disease control in larviculture. Aquaculture. 2003;227(1):427–38.
76. Rolland J-L, Abdelouahab M, Dupont J, Lefevre F, Bachère E, Romestand B. Stylicins, a new family of antimicrobial peptides from the Pacific blue shrimp Litopenaeus stylirostris. Mol Immunol. 2010;47(6):1269–77.
77. Tassanakajon A, Amparyup P, Somboonwiwat K, Supungul P. Cationic antimicrobial peptides in penaeid shrimp. Marine Biotechnol. 2011;13(4):639–57.
78. Destoumieux D, Bulet P, Loew D, Van Dorsselaer A, Rodriguez J, Bachère E. Penaeidins, a new family of antimicrobial peptides isolated from the shrimp Penaeus vannamei (Decapoda). J Biol Chem. 1997;272(45):28398–406.
79. Zhang J, Li F, Wang Z, Xiang J. Cloning and recombinant expression of a crustin-like gene from Chinese shrimp, Fenneropenaeus chinensis. J Biotechnol. 2007;127(4):605–14.
80. De la Vega E, O'Leary NA, Shockey JE, Robalino J, Payne C, Browdy CL, et al. Anti-lipopolysaccharide factor in Litopenaeus vannamei(Lv ALF): a broad spectrum antimicrobial peptide essential for shrimp immunity against bacterial and fungal infection. Mol Immunol. 2008;45(7):1916–25.
81. Burge EJ, Madigan DJ, Burnett LE, Burnett KG. Lysozyme gene expression by hemocytes of Pacific white shrimp, Litopenaeus vannamei, after injection with Vibrio. Fish Shellfish Immunol. 2007;22(4):327–39.
82. Arockiaraj J, Gnanam AJ, Muthukrishnan D, Gudimella R, Milton J, Singh A, et al. Crustin, a WAP domain containing antimicrobial peptide from freshwater prawn Macrobrachium rosenbergii: Immune characterization. Fish Shellfish Immunol. 2013;34(1):109–18.
83. Somboonwiwat K, Marcos M, Tassanakajon A, Klinbunga S, Aumelas A, Romestand B, et al. Recombinant expression and anti-microbial activity of anti-lipopolysaccharide factor (ALF) from the black tiger shrimp Penaeus monodon. Dev Comp Immunol. 2005;29(10):841–51.
84. Arockiaraj J, Avin FA, Vanaraja P, Easwvaran S, Singh A, Othman RY, et al. Immune role of Mr NFκBI-α, an IκB family member characterized in prawn M. rosenbergii. Fish Shellfish Immunol. 2012;33(3):619–25.
85. Maningas MBB, Kondo H, Hirono I, Saito-Taki T, Aoki T. Essential function of transglutaminase and clotting protein in shrimp immunity. Mol Immunol. 2008;45(5):1269–75.
86. Yeh M-S, Liu C-H, Hung C-W, Cheng W. cDNA cloning, identification, tissue localisation, and transcription profile of a transglutaminase from white shrimp, Litopenaeus vannamei, after infection by Vibrio alginolyticus. Fish Shellfish Immunol. 2009;27(6):748–56.
87. Liu Y-C, Li F-H, Wang B, Dong B, Zhang Q-L, Luan W, et al. A transglutaminase from Chinese shrimp (Fenneropenaeus chinensis), full-length cDNA cloning, tissue localization and expression profile after challenge. Fish Shellfish Immunol. 2007;22(5):576–88.
88. Bandyopadhyay U, Das D, Banerjee RK. Reactive oxygen species: oxidative damage and pathogenesis. Curr Sci. 1999;77(5):658–66.
89. Yu BP. Cellular defenses against damage from reactive oxygen species. Physiol Rev. 1994;74(1):139.
90. Zhang Q, Li F, Zhang X, Dong B, Zhang J, Xie Y, et al. cDNA cloning, characterization and expression analysis of the antioxidant enzyme gene, catalase, of Chinese shrimp Fenneropenaeus chinensis. Fish Shellfish Immunol. 2008;24(5):584–91.
91. Zhou J, Wang W-N, Wang A-L, He W-Y, Zhou Q-T, Liu Y, et al. Glutathione S-transferase in the white shrimp Litopenaeus vannamei: characterization and regulation under pH stress. Comp Biochem Physiol C Toxicol Pharmacol. 2009;150(2):224–30.
92. Cheng W, Tung Y-H, Liu C-H, Chen J-C. Molecular cloning and characterisation of copper/zinc superoxide dismutase (Cu, Zn-SOD) from the giant freshwater prawn Macrobrachium rosenbergii. Fish Shellfish Immunol. 2006;21(1):102–12.
93. Tian J, Chen J, Jiang D, Liao S, Wang A. Transcriptional regulation of extracellular copper zinc superoxide dismutase from white shrimp Litopenaeus vannamei following Vibrio alginolyticus and WSSV infection. Fish Shellfish Immunol. 2011;30(1):234–40.
94. Ren Q, Sun R-R, Zhao X-F, Wang J-X. A selenium-dependent glutathione peroxidase (Se-GPx) and two glutathione S-transferases (GSTs) from Chinese shrimp (Fenneropenaeus chinensis). Comp Biochem Physiol C Toxicol Pharmacol. 2009;149(4):613–23.
95. Li D-X, Du X-J, Zhao X-F, Wang J-X. Cloning and expression analysis of an o-methyltransferase (OMT) gene from Chinese shrimp, Fenneropenaeus chinensis. Fish Shellfish Immunol. 2006;21(3):284–92.
96. Li C, Ni D, Song L, Zhao J, Zhang H, Li L. Molecular cloning and characterization of a catalase gene from Zhikong scallop Chlamys farreri. Fish Shellfish Immunol. 2008;24(1):26–34.

97. He N, Liu H, Xu X. Identification of genes involved in the response of haemocytes of Penaeus japonicus by suppression subtractive hybridization (SSH) following microbial challenge. Fish Shellfish Immunol. 2004;17(2):121–8.
98. Horváth I, Multhoff G, Sonnleitner A, Vígh L. Membrane-associated stress proteins: more than simply chaperones. Biochim Biophys Acta. 2008;1778(7):1653–64.
99. Vabulas RM, Raychaudhuri S, Hayer-Hartl M, Hartl FU. Protein folding in the cytoplasm and the heat shock response. Cold Spring Harb Perspect Biol. 2010;2(12):a004390.
100. Jiang S, Qiu L, Zhou F, Huang J, Guo Y, Yang K. Molecular cloning and expression analysis of a heat shock protein (Hsp90) gene from black tiger shrimp (Penaeus monodon). Mol Biol Rep. 2009;36(1):127–34.
101. Lo W-Y, Liu K-F, Liao I-C, Song Y-L. Cloning and molecular characterization of heat shock cognate 70 from tiger shrimp (Penaeus monodon). Cell Stress Chaperones. 2004;9(4):332.
102. Cui Z, Liu Y, Luan W, Li Q, Wu D, Wang S. Molecular cloning and characterization of a heat shock protein 70 gene in swimming crab (Portunus trituberculatus). Fish Shellfish Immunol. 2010;28(1):56–64.
103. Zhou J, Wang W-N, He W-Y, Zheng Y, Wang L, Xin Y, et al. Expression of HSP60 and HSP70 in white shrimp, Litopenaeus vannamei in response to bacterial challenge. J Invertebr Pathol. 2010;103(3):170–8.
104. Rungrassamee W, Leelatanawit R, Jiravanichpaisal P, Klinbunga S, Karoonuthaisiri N. Expression and distribution of three heat shock protein genes under heat shock stress and under exposure to Vibrio harveyi in Penaeus monodon. Dev Comp Immunol. 2010;34(10):1082–9.
105. Robalino J, Almeida JS, McKillen D, Colglazier J, Trent HF, Chen YA, et al. Insights into the immune transcriptome of the shrimp Litopenaeus vannamei: tissue-specific expression profiles and transcriptomic responses to immune challenge. Physiol Genomics. 2007;29(1):44–56.
106. Weinrauch Y, Zychlinsky A. The induction of apoptosis by bacterial pathogens. Annu Rev Microbiol. 1999;53(1):155–87.
107. Chen Y, Zychlinsky A. Apoptosis induced by bacterial pathogens. Microb Pathog. 1994;17(4):203–12.
108. Chang C-C, Yeh M-S, Lin H-K, Cheng W. The effect of Vibrio alginolyticus infection on caspase-3 expression and activity in white shrimp Litopenaeus vannamei. Fish Shellfish Immunol. 2008;25(5):672–8.
109. Fall J, Kono T, Tanekhy M, Itami T, Sakai M. Expression of innate immune-related genes of Kuruma shrimp, Marsupenaeus japonicus, after challenge with Vibrio nigripulchritudo. Afr J Microbiol Res. 2010;4(22):2426–33.
110. Finlay BB, McFadden G. Anti-immunology: evasion of the host immune system by bacterial and viral pathogens. Cell. 2006;124(4):767–82.
111. Belvin MP, Anderson KV. A conserved signaling pathway: the Drosophila toll-dorsal pathway. Annu Rev Cell Dev Biol. 1996;12(1):393–416.
112. Yang C, Zhang J, Li F, Ma H, Zhang Q, Jose Priya T, et al. A Toll receptor from Chinese shrimp Fenneropenaeus chinensis is responsive to Vibrio anguillarum infection. Fish Shellfish Immunol. 2008;24(5):564–74.
113. Li F, Wang D, Li S, Yan H, Zhang J, Wang B, et al. A Dorsal homolog (FcDorsal) in the Chinese shrimp Fenneropenaeus chinensis is responsive to both bacteria and WSSV challenge. Dev Comp Immunol. 2010;34(8):874–83.
114. Huffman DL, Abrami L, Sasik R, Corbeil J, van der Goot FG, Aroian RV. Mitogen-activated protein kinase pathways defend against bacterial pore-forming toxins. Proc Natl Acad Sci U S A. 2004;101(30):10995–1000.
115. Honda T, Ni Y, Miwatani T, Adachi T, Kim J. The thermostable direct hemolysin of Vibrio parahaemolyticus is a pore-forming toxin. Can J Microbiol. 1992;38(11):1175–80.
116. Sun C, Shao H-L, Zhang X-W, Zhao X-F, Wang J-X. Molecular cloning and expression analysis of signal transducer and activator of transcription (STAT) from the Chinese white shrimp Fenneropenaeus chinensis. Mol Biol Rep. 2011;38(8):5313–9.
117. Zong R, Wu W, Xu J, Zhang X. Regulation of phagocytosis against bacterium by Rab GTPase in shrimp Marsupenaeus japonicus. Fish Shellfish Immunol. 2008;25(3):258–63.
118. Ji P-F, Yao C-L, Wang Z-Y. Two types of calmodulin play different roles in Pacific white shrimp (Litopenaeus vannamei) defenses against Vibrio parahaemolyticus and WSSV infection. Fish Shellfish Immunol. 2011;31(2):260–8.
119. Maiti B, Khushiramani R, Tyagi A, Karunasagar I, Karunasagar I. Recombinant ferritin protein protects Penaeus monodon infected by pathogenic Vibrio harveyi. Dis Aquat Organ. 2010;88(2):99.
120. YAN F, ZHANG Y-l, LUO H-q, HU Z, HUANG T-w, YE X-q. The phenoloxidase activity of hemocyanin from white leg shrimp Litopenaeus vannamei. Fisheries Sci. 2008;1:001.
121. Zhang X, Huang C, Qin Q. Antiviral properties of hemocyanin isolated from shrimp Penaeus monodon. Antiviral Res. 2004;61(2):93–9.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

genomeA
Journals.ASM.org

Genomes of Two Clinical Isolates of *Mycobacterium tuberculosis* from Odisha, India

Mohammad Majid,[a,d] Narender Kumar,[a] Asifa Qureshi,[d] Priyadarshini Yerra,[a] Ashutosh Kumar,[a] Mandala Kiran Kumar,[a] Suma Tiruvayipati,[a,c] Ramani Baddam,[a] Sabiha Shaik,[a] Aparna Srikantam,[b] Niyaz Ahmed[a,c]

Pathogen Biology Laboratory, Department of Biotechnology and Bioinformatics, University of Hyderabad, Hyderabad, India[a]; Blue Peter Research Centre (Lepra Society), Cherlapally, Hyderabad, India[b]; Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia[c]; Environmental Genomics Division, CSIR–National Environmental Engineering Research Institute (NEERI), Nagpur, India[d]

We report whole-genome sequences of two clinical isolates of *Mycobacterium tuberculosis* isolated from patients in Odisha, India. The sequence analysis revealed that these isolates are of an ancestral type and might represent some of the "pristine" isolates in India that have not admixed with other lineages.

Received 20 February 2014 **Accepted** 27 February 2014 **Published** 20 March 2014

Citation Majid M, Kumar N, Qureshi A, Yerra P, Kumar A, Kumar MK, Tiruvayipati S, Baddam R, Shaik S, Srikantam A, Ahmed N. 2014. Genomes of two clinical isolates of *Mycobacterium tuberculosis* from Odisha, India. Genome Announc. 2(2):e00199-14. doi:10.1128/genomeA.00199-14.

Copyright © 2014 Majid et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported license.

Address correspondence to Narender Kumar, kumarnaren13@gmail.com, or Niyaz Ahmed, niyaz.ahmed@uohyd.ac.in.

Tuberculosis caused by *Mycobacterium tuberculosis* is a chronic infectious disease that is often fatal if not effectively treated. Every year, 8 to 9 million new infections and a death toll of 1.5 million are recorded worldwide. It is estimated that about one-third of the human population is infected with *M. tuberculosis* (1). Comparative genomic studies have provided deeper insights into the genetic diversity and clonal architecture of *M. tuberculosis* (2). Recent studies conducted on isolates from India have shown that highly concentrated reservoirs of the ancestral *M. tuberculosis* lineages prevail in South and Central India (3–6). Only a limited number of *M. tuberculosis* genomes from India are sequenced. The whole-genome analysis of ancestral and modern lineages would facilitate deciphering of the genetic variability and evolutionary mechanisms of this obligate parasite. We describe the whole-genome sequences of two *M. tuberculosis* strains, NA-A0008 and NA-A0009, isolated in 2008 from patients in rural Odisha, India.

Genomic DNAs of both strains were isolated using the Qiagen kit method. Whole-genome sequencing was carried out on an Ion Torrent sequencing platform (Life Technologies). The process generated 3 million and 2.9 million reads amounting to 89× and 93× genome coverage for NA-A0008 and NA-A0009, respectively, with a mean read length of 250 bp. The reads after filtration were assembled into 280 and 310 contigs for NA-A0008 and NA-A0009, respectively, using the MIRA v.2 *de novo* assembler. These contigs were ordered and reoriented according to the *M. tuberculosis* CCDC 5180 genome using in-house written scripts. The resulting draft genomes were annotated using the RAST annotation server (7), and CDSs were validated by comparing outputs from EasyGene (8) and Glimmer (9), as done previously (10–13). The number of rRNA operons were predicted in both strains using RNAmmer (14), while tRNAscan-SE (15) was used to identify tRNA sequences. Artemis (16) was used to glean the genome statistics of both the strains. The genome sizes of NA-A0008 and NA-A0009 were 4,259,206 and 4,271,739 bp, with coding percentages of 89.4% and 89.3%, respectively. The G+C contents of both strains were high, as usually observed for the *M. tuberculosis* complex, 65.31% (NA-A0008) and 65.28% (NA-A0009). The two genomes, NA-A0008 and NA-A0009, were predicted to encode 4,400 and 4,453 CDSs with average lengths of 866 and 857 bp, respectively. Both of them contained a single rRNA operon and 45 tRNA genes.

The availability of these genome sequences would definitely complement the gene pool analysis of the Indian strains from different parts of the country. Besides this, comparative genomic analysis and phylogenetic study of these isolates with other *M. tuberculosis* strains might give us important insights into the biology and molecular epidemiology of this organism.

Nucleotide sequence accession numbers. The *M. tuberculosis* NA-A0008 and NA-A0009 whole-genome shotgun projects have been deposited in the GenBank database under the accession numbers ALYG00000000 and ALYH00000000, respectively. The BioProject designations for these projects are PRJNA168604 and PRJNA168605, respectively.

ACKNOWLEDGMENTS

We acknowledge the DFG program GRK1673 and financial support under the DBT-COE grant (BT/01/COE/07/02) from the Department of Biotechnology, Government of India. Mohammad Majid is grateful to Vijay N. Charde, M. Uma, and Mumtaz Baig for their help and support.

REFERENCES

1. **Frieden TR, Sterling TR, Munsiff SS, Watt CJ, Dye C.** 2003. Tuberculosis. Lancet **362:**887–899. http://dx.doi.org/10.1016/S0140-6736(03)14333-4.
2. **Ilina EN, Shitikov EA, Ikryannikova LN, Alekseev DG, Kamashev DE, Malakhova MV, Parfenova TV, Afanas'ev MV, Ischenko DS, Bazaleev NA, Smirnova TG, Larionova EE, Chernousova LN, Beletsky AV, Mardanov AV, Ravin NV, Skryabin KG, Govorun VM.** 2013. Comparative genomic analysis of *Mycobacterium tuberculosis* drug resistant strains from Russia. PLoS One **8:**e56577. http://dx.doi.org/10.1371/journal.pone.0056577.
3. **Narayanan S, Gagneux S, Hari L, Tsolaki AG, Rajasekhar S, Narayanan**

PR, Small PM, Holmes S, Deriemer K. 2008. Genomic interrogation of ancestral *Mycobacterium tuberculosis* from south India. Infect. Genet. Evol. **8:**474–483. http://dx.doi.org/10.1016/j.meegid.2007.09.007.

4. **Rao KR, Kauser F, Srinivas S, Zanetti S, Sechi LA, Ahmed N, Hasnain SE.** 2005. Analysis of genomic downsizing on the basis of region-of-difference polymorphism profiling of *Mycobacterium tuberculosis* patient isolates reveals geographic partitioning. J. Clin. Microbiol. **43:**5978–5982. http://dx.doi.org/10.1128/JCM.43.12.5978-5982.2005.
5. **Ahmed N, Saini V, Raghuvanshi S, Khurana JP, Tyagi AK, Tyagi AK, Hasnain SE.** 2007. Molecular analysis of a leprosy immunotherapeutic bacillus provides insights into *Mycobacterium* evolution. PLoS One **2:**e968. http://dx.doi.org/10.1371/journal.pone.0000968.
6. **Thomas SK, Iravatham CC, Moni BH, Kumar A, Archana BV, Majid M, Priyadarshini Y, Rani PS, Valluri V, Hasnain SE, Ahmed N.** 2011. Modern and ancestral genotypes of *Mycobacterium tuberculosis* from Andhra Pradesh, India. PLoS One **6:**e27584. http://dx.doi.org/10.1371/journal.pone.0027584.
7. **Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O.** 2008. The RAST server: rapid annotations using subsystems technology. BMC Genomics **9:**75. http://dx.doi.org/10.1186/1471-2164-9-75.
8. **Larsen TS, Krogh A.** 2003. EasyGene—a prokaryotic gene finder that ranks ORFs by statistical significance. BMC Bioinformatics **4:**21. http://dx.doi.org/10.1186/1471-2105-4-21.
9. **Delcher AL, Harmon D, Kasif S, White O, Salzberg SL.** 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res. **27:**4636–4641. http://dx.doi.org/10.1093/nar/27.23.4636.
10. **Avasthi TS, Kumar N, Baddam R, Hussain A, Nandanwar N, Jadhav S, Ahmed N.** 2011. Genome of multidrug-resistant uropathogenic *Escherichia coli* strain NA114 from India. J. Bacteriol. **193:**4272–4273. http://dx.doi.org/10.1128/JB.05413-11.
11. **Avasthi TS, Devi SH, Taylor TD, Kumar N, Baddam R, Kondo S, Suzuki Y, Lamouliatte H, Mégraud F, Ahmed N.** 2011. Genomes of two chronological isolates (*Helicobacter pylori* 2017 and 2018) of the West African *Helicobacter pylori* strain 908 obtained from a single patient. J. Bacteriol. **193:**3385–3386. http://dx.doi.org/10.1128/JB.05006-11.
12. **Baddam R, Thong KL, Avasthi TS, Shaik S, Yap KP, Teh CS, Chai LC, Kumar N, Ahmed N.** 2012. Whole-genome sequences and comparative genomics of *Salmonella enterica* serovar. *Typhi* isolates from patients with fatal and nonfatal typhoid fever in Papua New Guinea. J. Bacteriol. **194:**5122–5123. http://dx.doi.org/10.1128/JB.01051-12.
13. **Kumar N, Mukhopadhyay AK, Patra R, De R, Baddam R, Shaik S, Alam J, Tiruvayipati S, Ahmed N.** 2012. Next-generation sequencing and de novo assembly, genome organization, and comparative genomic analyses of the genomes of two *Helicobacter pylori* isolates from duodenal ulcer patients in India. J. Bacteriol. **194:**5963–5964. http://dx.doi.org/10.1128/JB.01371-12.
14. **Lagesen K, Hallin P, Rødland EA, Staerfeldt HH, Rognes T, Ussery DW.** 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. **35:**3100–3108. http://dx.doi.org/10.1093/nar/gkm160.
15. **Lowe TM, Eddy SR.** 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. **25:**955–964. http://dx.doi.org/10.1093/nar/25.5.0955.
16. **Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B.** 2000. Artemis: sequence visualization and annotation. Bioinformatics **16:**944–945. http://dx.doi.org/10.1093/bioinformatics/16.10.944.

Tiruvayipati *et al. Gut Pathogens* 2013, **5**:37
http://www.gutpathogens.com/content/5/1/37

Gut Pathogens

RESEARCH Open Access

Genome anatomy of the gastrointestinal pathogen, Vibrio parahaemolyticus of crustacean origin

Suma Tiruvayipati[1,2], Subha Bhassu[1,3*], Narender Kumar[2], Ramani Baddam[2], Sabiha Shaik[2], Anil Kumar Gurindapalli[2], Kwai Lin Thong[1,4] and Niyaz Ahmed[1,2*]

Abstract

Vibrio parahaemolyticus, an important human pathogen, is associated with gastroenteritis and transmitted through partially cooked seafood. It has become a major concern in the production and trade of marine food products. The prevalence of potentially virulent and pathogenic *V. parahaemolyticus* in raw seafood is of public health significance. Here we describe the genome sequence of a *V. parahaemolyticus* isolate of crustacean origin which was cultured from prawns in 2008 in Selangor, Malaysia (isolate PCV08-7). The next generation sequencing and analysis revealed that the genome of isolate PCV08-7 has closest similarity to that of *V. parahaemolyticus* RIMD2210633. However, there are certain unique features of the PCV08-7 genome such as the absence of TDH-related hemolysin (TRH), and the presence of HU-alpha insertion. The genome of isolate PCV08-7 encodes a thermostable direct hemolysin (TDH), an important virulence factor that classifies PCV08-7 isolate to be a serovariant of O3:K6 strain. Apart from these, we observed that there is certain pattern of genetic rearrangements that makes *V. parahaemolyticus* PCV08-7 a non-pandemic clone. We present detailed genome statistics and important genetic features of this bacterium and discuss how its survival, adaptation and virulence in marine and terrestrial hosts can be understood through the genomic blueprint and that the availability of genome sequence entailing this important Malaysian isolate would likely enhance our understanding of the epidemiology, evolution and transmission of foodborne Vibrios in Malaysia and elsewhere.

Keywords: *Vibrio parahaemolyticus*, Genomics, Malaysia, Seafood, Comparative genomics

Background

Vibrio parahaemolyticus inhabits the estuarine, marine and brackish water ecosystems. It is an important human pathogen associated with gastroenteritis linked to contaminated seafood consumption. Since this species is abundant in marine products, it has become a significant concern in the production and trade of seafood worldwide [1]. In Southeast Asian countries, including Malaysia, virulent *V. parahaemolyticus* in raw seafood have been reported [2,3]. Numerous cases of *V. parahaemolyticus* infection were reported in North America, South East Asia and Japan including some places in East Asia [4-10] giving the illness a pandemic status affecting thousands of people. Thus, the prevalence of pathogenic Vibrios in seafood is of public health concern and is an open ended issue.

The pathogenic *V. parahaemolyticus* strains are differentiated from non-pathogenic ones by their ability to cause beta-haemolysis on Wagatsuma agar, an activity known as 'Kanagawa phenomenon'. This effect is mediated by the activity of thermostable direct hemolysin (TDH) encoded by the *tdh* genes [8]. A pandemic clone of *V. parahaemolyticus* can broadly be defined as the one that is positive for TDH and exhibits the Kanagawa phenomenon [10].

V. parahaemolyticus strains are classified based on the types and variants of their O antigen and flagellar antigen (K). There are 13 O-serogroups and 71 K antigens and various combinations of these give rise to a wide variety of serovars which have been recognized as the causative agents of the disease. A clone of serovar O3:K6 has recently emerged and was associated with outbreaks

* Correspondence: subhabhassu@um.edu.my; niyaz.ahmed@uohyd.ac.in
[1]Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia
[2]Pathogen Biology Laboratory, Department of Biotechnology and Bioinformatics, University of Hyderabad, Hyderabad, India
Full list of author information is available at the end of the article

BioMed Central

© 2013 Tiruvayipati et al.; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

in India and Japan [7]. Frequent recombination events that promote clonal diversification suggest a scenario whereby a subset of O3:K6 strains might continue to evolve [11]. Consequently, different groups of related O3:K6 clonal strains have now been globally disseminated in Asia, North and South America, Africa and Europe [7].

The genomes of *V. parahaemolyticus* strains are said to have undergone a number of recombination events that could have been the reason for serotype conversion from O3:K6 to O4:K68 [12]. Regions of recombination likely involve a genetic element larger than the gene clusters encoding O and K-antigens. More than 20 serovariants which include O3:K6, O4:K68, O1:K25, O6:K18 and O1:KUT [13,14] emerged from an original pandemic strain, O3:K6. The pandemic group of these bacteria has evolved through a number of deletions, substitutions and acquisitions of regions primarily corresponding to TDH or a TDH-related hemolysin (TRH). It is the presence of either of these two virulence factors that confer potential to cause gastroenteritis in human populations. The pandemic clone is said to have emerged from a prepandemic clone which was positive for TRH and negative for TDH genes and harbored a new sequence of toxR (GS-PCR). The intermediate clone is described as being negative for both TRH and TDH, but positive for GS-PCR.

It has been observed that *V. parahaemolyticus* contains two chromosomes; *V. parahaemolyticus* RIMD2210633 has 3.2 Mb and 1.8 Mb of genome sizes for chromosome1 and 2 respectively [15]. There are several *V. parahaemolyticus* genomes which have been sequenced and deposited in Genbank as whole genomes or shotgun submissions (WGS) and sequence read archives (SRA). The only fully annotated submissions entail *V. parahaemolyticus* RIMD2210633 and *V. parahaemolyticus* BB220P. The *V. parahaemolyticus* RIMD2210633 genome harbors a Type III secretion system as a central virulence factor which is found in most diarrhea-causing bacteria [15]. As mentioned above, many studies link to the evolutionary aspects of the present pandemic clone formed from a prepandemic clone with a drastic change in its gene content i.e., the evolution from a TDH negative/TRH positive to a TDH positive/TRH negative strain and the occurrence of several serovariants in the *V. parahaemolyticus* species. The present isolate (*V. parahaemolyticus* PCV08-7) has been recovered from seafood (prawn) in 2008 which were purchased from a wet market in Selangor, Malaysia.

The main purpose of this study was to analyze the PCV08-7 genome that originates from Malaysia, a large peninsular as well as archipelagic country having a thriving seafood business and that it experiences several food borne outbreaks each season. Unfortunately, there are no markers based on native genome(s) to guide detection of *V. parahaemolyticus* in wet market, in the aquaculture farms and from human excreta and blood. We hope that this genome sequence will be helpful in identifying markers relevant in diagnostic development and molecular epidemiology/transmission dynamics of this significant bacterium in Malaysia and elsewhere.

Methods

Source, isolation and culture of V. parahaemolyticus PCV08-7

The *V. parahaemolyticus* PCV08-7 (VPPCV08-7) isolate was identified and characterized by obtaining pure cultures on selective media followed by analysis through biochemical tests, Analytical Profile Index (API) tests and genetic confirmation by PCR. The bacterial culture was maintained by streak plate on a Thiosulfate-Citrate-Bile-Sucrose (Difco, France) agar plates. After incubation at 37°C for 21 – 24 hr, characteristic bacterial colonies appeared with blue-green colored boundaries. An isolated bacterial colony was cultured in Luria-Bertani (LB) broth with 2% Sodium Chloride (NaCl) and incubated overnight at 37°C for 16 – 18 hr. This bacterial culture was further maintained as glycerol stocks at −80°C in 20% glycerol. The genomic DNA was isolated from a pure, single colony. The bacterial identity was confirmed by sequence analysis of the 16S rRNA.

Genomic DNA isolation and Next-Generation Sequencing

The genomic DNA was isolated using Qiagen DNeasy Blood & Tissue kit (Qiagen, Germany) and the genome sequence was determined by Illumina genome analyzer at the Genotypic Technology Pvt. Ltd. Bengaluru, India (GA2x, pipeline version 1.6). The sequencing data comprised of 100 bp paired-end reads with an insert size corresponding to approximately 240 bp. The genome coverage obtained was approximately about 80X with per base quality of reads in a range of 25 – 40. A total of 3.8 million reads were generated. Bioinformatics analysis was carried out with the help of protocols, algorithms and scripts developed, customized and tested in Ahmed Labs.

Assembly and alignment

Various strategies were applied to resolve the difficulties in dealing with the two chromosomes to be assembled from the sequence reads. The following main approaches were adopted:

1. *Velvet* [16]: Contigs were generated using the sequence reads which consisted of information from both the chromosomes of the isolate PCV08-7. This was checked by manually comparing contigs against the NCBI database by BLAST to check the highest similarity hit. *V. parahaemolyticus* RIMD2210633 was found to be the closest match in each search. The contigs showed unique hits to chromosome 1

(CHR1) and chromosome 2 (CHR2) as well as few common hits at both the chromosomes. The strategy of using the contigs together representing a whole genome (i.e., CHR1 and CHR2 together) or using the contigs separately as CHR1 and CHR2 was found to be challenging for further analysis to assemble them separately into two chromosomal sequences.

2. *OSLAY* [17]: All the contigs were compared against both the chromosomes of the genome of RIMD2210633 individually and were then used to form supercontigs for both the chromosomes separately. This procedure was found to be problematic as the supercontig files generated from CHR1 and CHR2 (separately) revealed that the preliminary contigs mapped to sequences in both the supercontig files. This was perhaps due to the input file comprising assembled whole genome contigs used against CHR1 and CHR2. The second strategy under OSLAY was to attach CHR1 and CHR2 of the reference genome RIMD2210633 as follows: CHR1 and CHR2 were concatenated (as a 'whole genome stretch') and then further used as one full length single sequence. Using this whole genome stretch for BLAST analysis, supercontigs were generated using Velvet contigs and the BLAST results. This also eventually proved inefficient since the supercontigs contained some sequences with several 'N' representing a gap in this case and such supercontigs had to be sorted to their own positions on the genome.
3. *SSPACE* [18]: Scaffolding was performed on velvet assembled contigs. As explained above, scaffolds were obtained separately from both CHR1 and CHR2 as well as with the whole genome stretch. All the scaffolds were then BLAST analyzed against both CHR1 and CHR2 of the reference genome individually, as well as at the level of the whole genome stretch. The difficulty faced with scaffolding was similar to that of OSLAY. Hence, the option of separately identifying the scaffolds with respect to CHR1 and CHR2 and dealing with them separately remained a problem.
4. *Mauve* [19]: Velvet assembled contigs were used at this step and exported as sorted contigs by performing an alignment against the whole genome stretch. The results obtained as aligned sorted contigs were taken through a stand-alone BLAST protocol against the whole genome of RIMD2210633. Then the BLAST results were carefully checked for their positions corresponding to both CHR1 and CHR2. The contigs were carefully divided as belonging to CHR1 and CHR2 sequences of PCV08-7 draft genome. The issues faced here were limited to identifying and dealing with the sequences other than those present in the contigs, but which were common to both RIMD2210633 and PCV08-7 genomes. While working on the above strategies, BWA alignment [20] was performed using sequence reads against the whole genome stretch of VPRIMD2210633. Using SAMTOOLS [21] a *.sam* file was generated with which the whole genome of RIMD2210633/FASTA sequence was loaded on Tablet viewer [22] to manually inspect the presence of common genes and to position the draft genome of PCV08-7.

The sequencing reads obtained by us were primarily passed through a quality control step using FASTX toolkit [23] to obtain high quality reads free from adaptor and primer contamination which was further standardized to an optimal parameter p value of 70. High quality reads thus obtained were assembled de-novo [22,23] using the Velvet assembly tool which produced 83 contigs with a hash length optimized to 71. These contigs were used to run OSLAY to form supercontigs with the reference genome RIMD2210633. Alignment of the reads against the reference genome was performed using BWA. The pre-assembled reads were also formed into scaffolds using SSPACE. Perl scripts written in house and modified after Baddam *et al.* [24] were used to re-order the contigs, supercontigs and scaffolds into their individual files. These approaches were put together to finalize the draft genome of *V. parahaemolyticus* PCV08-7 (Figure 1).

Results and discussion

Genome assembly

The 100 bp paired end reads were assembled using Velvet assembly tool that effectively utilized approximately 3.7 million reads. The N50 value observed was 261989 bp. The contig with the maximum length was 704232 bp and the total number of bases in the genome were 5184164 bp. The genome was artificially closed.

The genomes with multiple chromosomes pose technical difficulties during assembly. It is a known fact that Vibrios – *V. cholerae, V. parahaemolyticus* and *V. vulnificus* contain two circular chromosomes [26]. The reference genome used in this study, *V. parahaemolyticus* RIMD2210633 also consists of two chromosomes [13]. As studied previously [13], the origin of replication in chromosome 1 with the presence of *dnaA* gene shows its similarity to many genomes of prokaryotic origin and the origin of replication of chromosome 2 shows homology with that present on *V. cholerae* chromosome 2. The identification of distinct replication sites is of utmost importance for assembling bacterial genomes with two chromosomes which in the case of *V. cholerae* have been studied earlier [27]. Previous studies explain need for a more accurate procedure to handle data to correctly assemble two chromosomes and assign gene

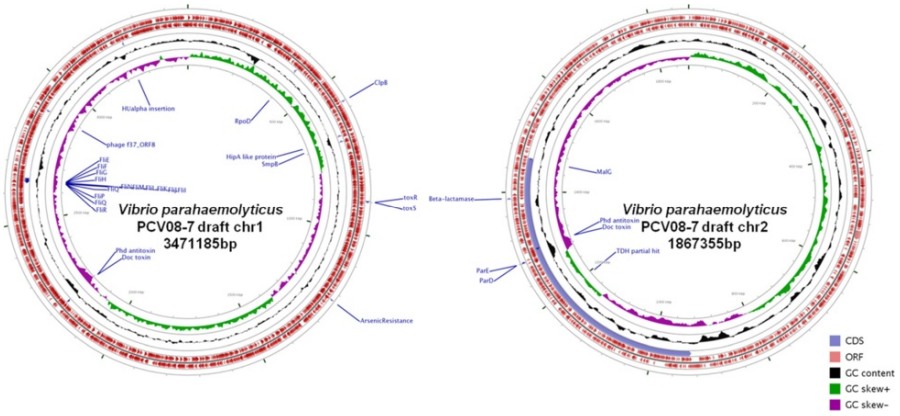


Figure 1 Circular view of *Vibrio parahaemolyticus* PCV08-7 draft genome. Diagrammatic representation of major genes carried by the two chromosomes of *Vibrio parahaemolyticus* PCV08-7 genome using CGview [25].

locations. The reads were assembled into a total of 83 contigs which were separated based on the assemble strategy as explained in the materials and methods section. Dealing with the present data, we observed that many of the genes of significant virulence or fitness importance were located on the chromosomes rather than showing any significant homology to the *Vibrio* plasmids. The presence of the Phd-Doc toxin antitoxin gene in our genome makes it interesting as the antitoxin gene has been previously reported related to plasmids [28] while a recent study [29] described

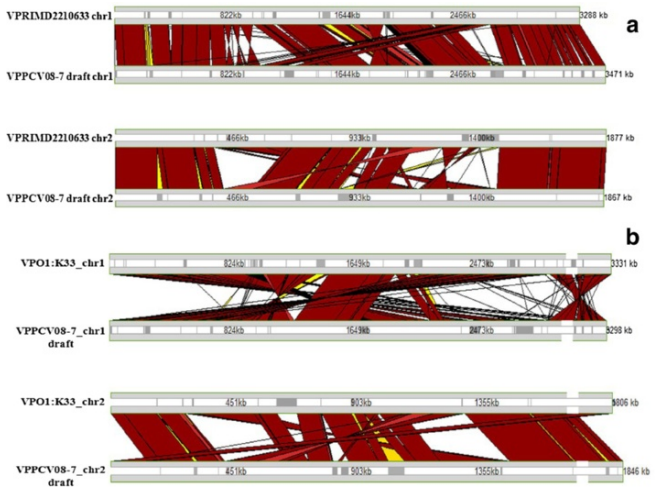


Figure 2 Alignment of the genome of strain RIMD2210633 against that of isolate PCV08-7 and strain O1:K33. (a) Comparison of chromosomes of strain RIMD2210633 (VPRIMD2210633 chr1, VPRIMD2210633 chr2) with the draft chromosomes of PCV08-7 (VPPCV08-7 draft chr1, VPPCV08-7 draft chr2) using M-GCAT. **(b)** Comparison of chromosomes of strain O1:K33 (VPO1:K33 chr1, VPO1:K33 chr2) with the draft chromosomes of PCV08-7 (VPPCV08-7 draft chr1, VPPCV08-7 draft chr2) using M-GCAT.

its occurrence on the chromosome of *Vibrio* species. However, we agree that the exact source of these genes can be mapped only when the plasmids will be sequenced and or analyzed separately.

Genome statistics and annotation

The draft assembled genome was annotated using the RAST server [30]. Statistics of the *V. parahaemolyticus* PCV08-7 draft genome were derived using Artemis [31], RNAmmer [32] and tRNAscanSE [33]: the sizes of chromosome 1 and chromosome 2 of the isolate were 3471185 bp and 1867355 bp respectively with G + C content of 45.35%. The tRNA and rRNA genes were 102 and 31 for chromosome 1, and 13 and 3 for chromosome 2, respectively. The chromosome 1 revealed a coding percentage of 85 with an average gene length of 943 bp while the chromosome 2 had a coding percentage of 86.2 with an average gene length of 950 bp.

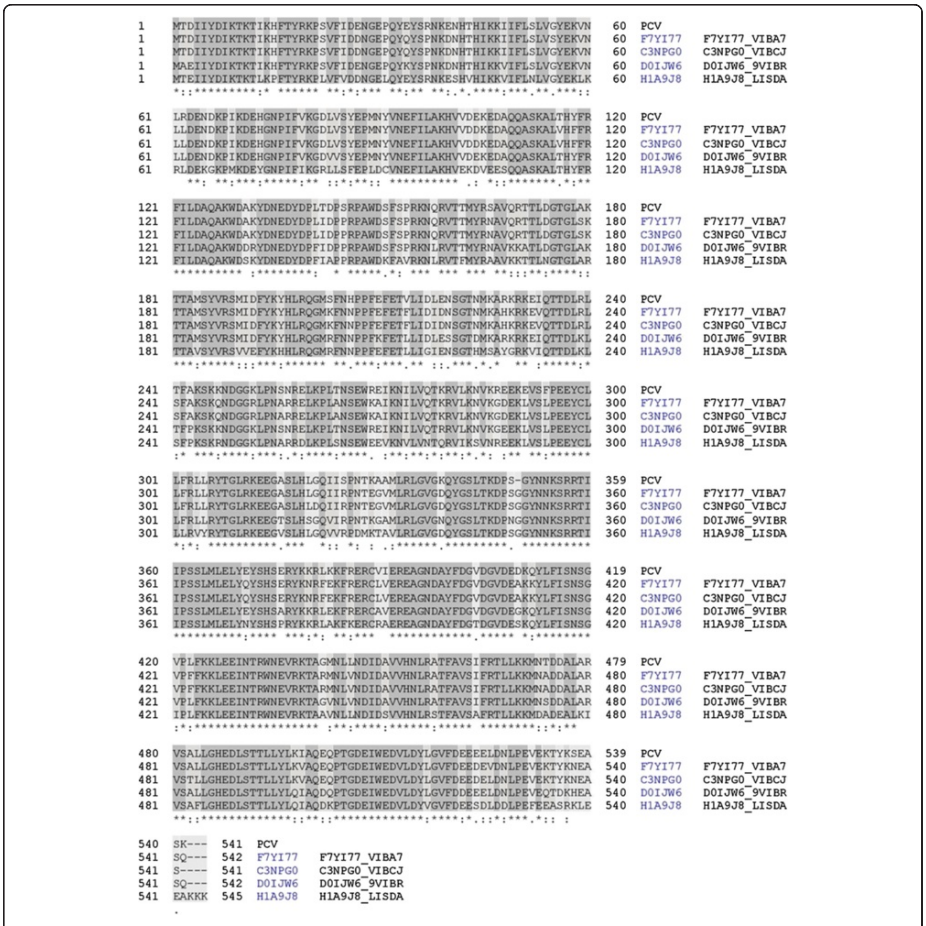


Figure 3 Alignment of a unique PCV08-7 protein sequence similar to *Photobacterium damselae subsps. damselae*. A unique sequence from PCV08-7 genome showed similarity with putative uncharacterized proteins of *V. anguillarum* 775 (F7YI77), *V. cholera* MJ-1236 (C3NPG0) and *Vibrio sp.* RC586 (D0IJW6) and similarity to a phage integrase of *Photobacterium damselae subsps. damselae* (H1A9J8).

The alignment of *V. parahaemolyticus* PCV08-7 genome with that of the *V. parahaemolyticus* RIMD2210633 genome using M-GCAT [34] showed visible rearrangements in the sequences of the two chromosomes of PCV08-7 isolate (Figure 2). The chromosome 1 of the draft genome carried phage shock proteins A, B and C, and bacteriophage f237 ORF8. It contained an integrated *tmRNA* gene with the closest element encoding the ribonuclease H. A site-specific recombinase *IntI4* and a gene encoding beta-lactamase were present. The draft genome also revealed genes responsible for fatty acid and amino acid metabolism. An important outer membrane protein *OmpUI* was also identified. Genes coding for gyrase B (*gyrB*), HU-alpha insertion and putative sigma factors such as *rpoD, rpoE, rpoS, rpoN and rpoH* were also found in our analysis. The chromosome 2 carried a TDH pathogenicity island with many deletions and substitutions and displayed a *malG* gene on one of the flanking regions of the pathogenicity island. This region also contained genes coding for nutrient uptake and metabolism. We documented the presence of vibrio ferrin receptor *pvuA* and ferrichrome ABC transport *pvuB*, *pvuC*, *pvuD* and *pvuE* encoding genes, and the related *pvsA*, *pvsB*, *pvsC*, *pvsD* and *pvsE* genes. The analysis of the genome further revealed presence of a cobalt-zinc-cadmium resistance protein and a Rhodanese related sulfur transferase (as also present in RIMD2210633 genome) and a lead-cadmium-zinc-mercury transporting ATPase enzyme (as seen in the *V. parahaemolyticus* BB220P genome). Phd antitoxin and Doc toxin [28] which fall under the programmed cell death systems were also uniquely identified. Studies in *E. coli* have shown the presence of a stress related protein *clpB* along with *rpoS* and a few other genes [35] which help cope with stress conditions and help in survival. Our analysis detected the presence of *clpB, rpoS* and *hipA* genes in the present genome as was also seen in the reference genome of RIMD2210633. There were two types of Type III secretion systems observed in *V. parahaemolyticus* RIMD2210633 [36]; T3SS1 and T3SS2. Our genome analysis remains open ended with respect to the presence of such type III secretion systems.

Identification of novel gene content and comparative analysis

Our genome analysis revealed some unique sequences which have good similarity to hypothetical proteins of other Vibrio species such as *Vibrio anguillarum* and *Vibrio cholerae*. A 6315 bp nucleotide sequence showed identity to a *V. anguillarum* hypothetical protein and a *V. cholera* hypothetical protein on NCBI-BLASTN. One of the coding proteins in this stretch revealed similarity to the annotated phage integrase encoding gene of *Photobacterium damselae subsp. damselae* plasmid pAQU1 DNA (Figure 3). A parD gene (antitoxin to parE) was also found which showed closer identity to other *Vibrio* species such as *Vibrio vulnificus*, *Vibrio mimicus* and *Vibrio orientalis*. parD when aligned against *V. vulnificus* and *V. mimicus* revealed an identity of 76 bp out of 80 bp (95%) (e-value 2e-48) and with *V. orientalis* an identity of 72 bp out of 80 bp (90%) (e-value 2e-45) on NCBI-BLASTN. A few newer hypothetical proteins with no reported annotation were identified. The genome also contained a gene relevant to arsenic resistance, possibly important in the adaptation of the bacterium to a high arsenic environment. Our analysis of the genome revealed presence of a partially similar sequence of TDH Pathogenicity Island, as compared to *V. parahaemolyticus* RIMD2210633. This island revealed genetic instability due to various insertion/deletion and substitution events we documented. The presence of toxS and toxR genes was also observed.

The old pandemic O3:K6 strain of *V. parahaemolyticus* is said to have gained gene clusters VPaI1-VPaI7 [37] to develop into a new pandemic clone of which VPaI4-VPaI6 are said to be putative virulence factors and may be potential pathogenicity islands. These regions are said to carry along with them a type VI secretion system (VP1386-1420). Our PCV08-7 genome analysis revealed that only one cluster, VPaI2, was detected completely, whereas VPaI3 and VPaI7 were partially present (Table 1). This perhaps shows that our strain could be possibly a new serovariant of a non-pandemic O3:K6 strain like the *V. parahaemolyticus* AQ3810 [8]. While

Table 1 Table representing pathogenicity related clusters and other VP clusters in *V. parahaemolyticus* PCV08-7: (1) pathogenicity related clusters (VPaI1-VPaI7) in the genome of strain RIMD2210633 that signify it to be a pandemic O3:K6 strain and their presence or absence in the genome of PCV08-7 isolate, (2) various other VP clusters and their occurrence in the genome of PCV08-7

(1) *Vibrio parahaemolyticus* RIMD2210633	***V. parahaemolyticus* PCV08-7**
VPaI1 (VP0380-0403)	Absent
VpaI2 (VP0635-0643)	Present
VpaI3 (VP1071-1094)	Partially present
VpaI4 (VP2131-2144)	Absent
VpaI5 (VP2900-2910)	Absent
VpaI6 (VPA1254-1270)	Absent
VpaI7 (VP1312-1398)	Present
Type VI secretion system (VP1386-1420)	Absent
(2) Other VP clusters	***V. parahaemolyticus* PCV08-7**
VP1355-1368	Partially present
VPA0074-0089	Present
VPA0713-0732	Present
VPA1194-1210	Present

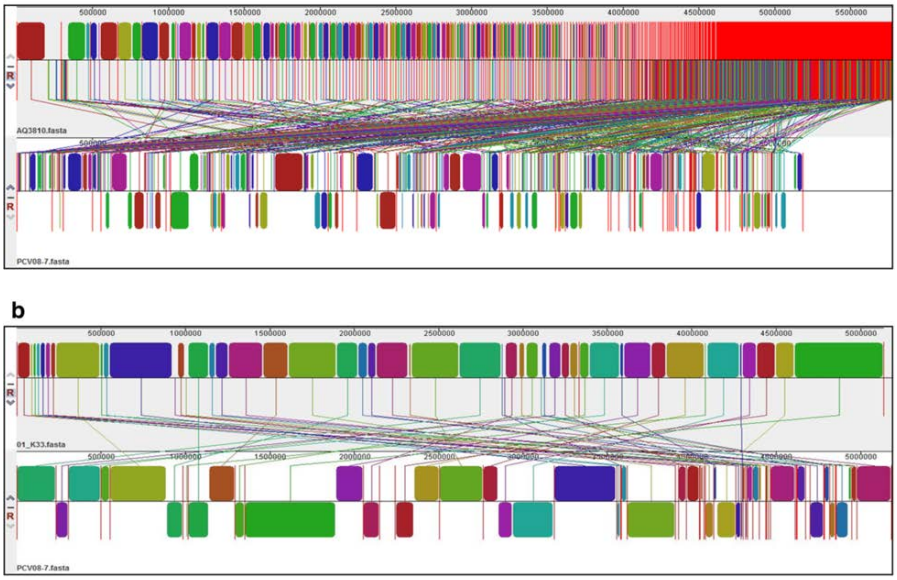


Figure 4 (a) ***V. parahaemolyticus*** **AQ3810 alignment against PCV08-7 genome: Concatenated chromosome 1 and chromosome 2 of *V. parahaemolyticus* AQ3810 (AQ3810.fasta) against *V. parahaemolyticus* PCV08-7 (PCV08-7.fasta) (b) *V. parahaemolyticus* O1:K33 alignment against PCV08-7.** Concatenated chromosome 1 and chromosome 2 of *V. parahaemolyticus* O1:K33 (O1_K33.fasta) against *V. parahaemolyticus* PCV08-7 (PCV08-7.fasta).

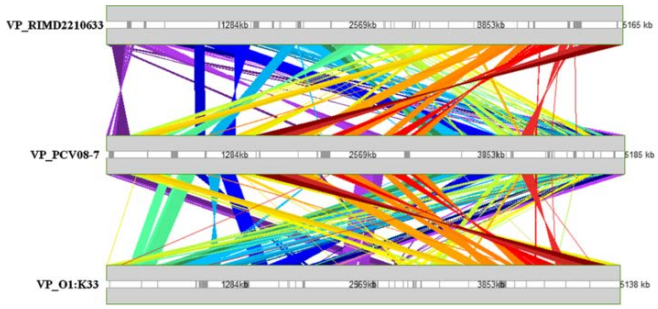


Figure 5 Comparison of whole genome sequences of strains RIMD2210633, PCV08-7 and O1:K33. Alignment of complete genomes of *V. parahaemolyticus* RIMD2210633, *V. parahaemolyticus* PCV08-7 and *V. parahaemolyticus* O1:K33, showing PCV08-7 being more similar to RIMD2210633.

variability of different gene clusters (Table 1) portrays a probably novel serovariant of *V. parahaemolyticus* with the presence of ribonuclease H encoding element (previously thought to be present only in *V. parahaemolyticus* RIMD2210633 and absent in *V. parahaemolyticus* AQ3810 [12]). A further comparative study between the *V. parahaemolyticus* PCV08-7 and the non-pandemic *V. parahaemolyticus* AQ3810 (O3:K6 strain) and the newest *V. parahaemolyticus* O1:K33 (trh+/ tdh + genotype) strain showed that *V. parahaemolyticus* PCV08-7 has more genetic relatedness towards a trh+/ tdh + strain (Figure 4). But, alignments of the *V. parahaemolyticus* PCV08-7 contig data against the *V. parahaemolyticus* O1:K33 and *V. parahaemolyticus* RIMD2210633 (Figure 2) strains show that it is closer to O3:K6 serotype (Figure 5).

From the above thesis, it becomes probably apparent that the genome of *V. parahaemolyticus* PCV08-7 meaningfully adds to the battery of important genomic sequences representing enteropathogenic bacteria. The genome of an arthropod derived, foodborne *Vibrio* should be important to understand adaptation to a crustacean host and a human host.

Epilogue and future directions

A first account of the genome of *V. parahaemolyticus* PCV08-7 has been presented. The draft genome and its annotation as described would be able to explain the lifestyle of pathogenic *Vibrio* species. The experience of assembling this genome and the difficulties associated with separating the data with respect to two chromosomes would certainly be helpful to the community in the follow-up studies. Further, a host of new molecular markers as gleaned by our analysis would be relevant in the diagnostic development and molecular epidemiology. The present genome and the ensuing comparative genomics would be able to rekindle our thoughts on the survival and virulence as well as transmission potentials of *V. parahaemolyticus* and also on their adaptation to different hosts and the niches thereof. Our results clearly reveal a significantly novel gene content which could presumably have been acquired through a horizontal gene transfer mechanism. Our analysis revealed the presence of not only the conserved genomic regions among different *V. parahaemolyticus* bacteria, but also dissects some of the unique sets of genes that hold relevance to virulence. We propose to finish and polish the genome in the near future also with the help of further coverage using alternative sequencing platforms and by employing a hybrid assembly approach. Also, it will be possible to determine the true extent of the diversity of *V. parahaemolyticus* strains obtained from seafood as compared to those isolated from human cases. Such a diversity analysis would focus on 1) genomic coordinates relevant to colonization of and adaptation to different hosts in different ecosystems; 2) genome dynamics relative to bacterial fitness shaping over time and with transmission across different hosts; and 3) profile of genomic rearrangements including additive and reductive genome evolution and their significance in the evolution of pathogenic *Vibrio* species. Presently, the epidemiology of *V. parahaemolyticus* infection in resource-poor countries largely entails a classical serology concocted with guess work as to the type of strain involved and its source. Our genomic data would hopefully contribute to this situation also.

Availability of supporting data

The *Vibrio parahaemolyticus* PCV08-7 whole genome shotgun project was deposited in Genbank under the accession AOCL00000000. The version described in this paper is the first version, AOCL01000000. This consists of sequences from AOCL01000000 – AOCL01000083 (http://www.ncbi.nlm.nih.gov/nuccore/AOCL00000000).

Competing interests

NA and TKL are the editors of Gut Pathogens.

Authors' contributions

NA and SB: Designed and supervised the study and written and edited the manuscript, TS: performed genomic DNA preparation, sequencing analysis, annotation and comparative genomics, AKG: performed initial bioinformatics analysis, RB and SS: provided tools and IT support for the study, NK: contributed to quality control of the NGS data and assembly. TKL: isolated and maintained the strain and provided inputs on lifestyle and evolution of the organism. All authors read and approved the final manuscript.

Acknowledgements

TS was supported by a doctoral fellowship from University of Malaya under the Bright Sparks program (BSP 226(3)-12). SB would like to thank University of Malaya for the support from the PPP grant PG088-2012B. SB and KLT would like to acknowledge research support received from University of Malaya under different funding instruments. NA would like to acknowledge partial support from the UM-HIR project of the University of Malaya. NA is an Academy Professor (Adjunct) of the Academy of Scientific and Innovative Research, India and visiting Professor at the Institute of Biological Sciences, University of Malaya, Kuala Lumpur.

Author details

[1]Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia. [2]Pathogen Biology Laboratory, Department of Biotechnology and Bioinformatics, School of Life Sciences, University of Hyderabad, Hyderabad, India. [3]Center of Biotechnology for Agriculture (CEBAR), University of Malaya, Kuala Lumpur, Malaysia. [4]Laboratory of Biomedical Science and Molecular Microbiology, UMBIO Research Cluster, University of Malaya, Kuala Lumpur, Malaysia.

Received: 3 October 2013 Accepted: 29 November 2013
Published: 11 December 2013

References

1. DePaola A, Nordstrom JL, Dalsgaard A, Forslund A, Oliver J, Bates T, Bourdage KL, Gulig PA: **Analysis of Vibrio vulnificus from market oysters and septicemia cases for virulence markers.** *Appl Environ Microbiol* 2003, **69**(7):4006–4011.
2. Sujeewa AKW, Norrakiah AS, Laina M: **Prevalence of toxic genes of Vibrio parahaemolyticus in shrimps (Penaeus monodon) and culture environment.** *International Food Research Journal* 2009, **16**:89–95.

3. Paydar M, Teh CS, Thong KL: **Prevalence and characterisation of potentially virulent Vibrio parahaemolyticus in seafood in Malaysia using conventional methods, PCR and REP-PCR.** *Food Control* 2013, **32**:13–18.
4. **Guidelines for national human immunodeficiency virus case surveillance, including monitoring for human immunodeficiency virus infection and acquired immunodeficiency syndrome. Centers for Disease Control and Prevention.** *MMWR Recomm Rep: Morbidity and mortality weekly report Recommendations and reports/Centers for Disease Control* 1999, **48**(RR-13):1–27. 29–31.
5. Bag PK, Nandi S, Bhadra RK, Ramamurthy T, Bhattacharya SK, Nishibuchi M, Hamabata T, Yamasaki S, Takeda Y, Nair GB: **Clonal diversity among recently emerged strains of Vibrio parahaemolyticus O3:K6 associated with pandemic spread.** *J Clin Microbiol* 1999, **37**(7):2354–2357.
6. Nair GB, Hormazabal JC: **The Vibrio parahaemolyticus pandemic.** *Revista chilena de infectologia: organo oficial de la Sociedad Chilena de Infectologia* 2005, **22**(2):125–130.
7. Nair GB, Ramamurthy T, Bhattacharya SK, Dutta B, Takeda Y, Sack DA: **Global dissemination of Vibrio parahaemolyticus serotype O3:K6 and its serovariants.** *Clin Microbiol Rev* 2007, **20**(1):39–48.
8. Nishibuchi M, Kaper JB: **Thermostable direct hemolysin gene of Vibrio parahaemolyticus: a virulence gene acquired by a marine bacterium.** *Infect Immun* 1995, **63**(6):2093–2099.
9. Okuda J, Ishibashi M, Hayakawa E, Nishino T, Takeda Y, Mukhopadhyay AK, Garg S, Bhattacharya SK, Nair GB, Nishibuchi M: **Emergence of a unique O3: K6 clone of Vibrio parahaemolyticus in Calcutta, India, and isolation of strains from the same clonal group from Southeast Asian travelers arriving in Japan.** *J Clin Microbiol* 1997, **35**(12):3150–3155.
10. Han H, Wong HC, Kan B, Guo Z, Zeng X, Yin S, Liu X, Yang R, Zhou D: **Genome plasticity of Vibrio parahaemolyticus: microevolution of the 'pandemic group'.** *BMC genomics* 2008, **9**:570.
11. Gonzalez-Escalona N, Martinez-Urtaza J, Romero J, Espejo RT, Jaykus LA, DePaola A: **Determination of molecular phylogenetics of Vibrio parahaemolyticus strains by multilocus sequence typing.** *J Bacteriol* 2008, **190**(8):2831–2840.
12. Chen Y, Stine OC, Badger JH, Gil AI, Nair GB, Nishibuchi M, Fouts DE: **Comparative genomic analysis of Vibrio parahaemolyticus: serotype conversion and virulence.** *BMC Genomics* 2011, **12**:294.
13. Chowdhury NR, Chakraborty S, Ramamurthy T, Nishibuchi M, Yamasaki S, Takeda Y, Nair GB: **Molecular evidence of clonal Vibrio parahaemolyticus pandemic strains.** *Emerg Infect Dis* 2000, **6**(6):631–636.
14. Chowdhury NR, Stine OC, Morris JG, Nair GB: **Assessment of evolution of pandemic Vibrio parahaemolyticus by multilocus sequence typing.** *J Clin Microbiol* 2004, **42**(3):1280–1282.
15. Makino K, Oshima K, Kurokawa K, Yokoyama K, Uda T, Tagomori K, Iijima Y, Najima M, Nakano M, Yamashita A, *et al*: **Genome sequence of Vibrio parahaemolyticus: a pathogenic mechanism distinct from that of V cholerae.** *Lancet* 2003, **361**(9359):743–749.
16. Zerbino DR, Birney E: **Velvet: algorithms for de novo short read assembly using de Bruijn graphs.** *Genome Res* 2008, **18**(5):821–829.
17. Richter DC, Schuster SC, Huson DH: **OSLay: optimal syntenic layout of unfinished assemblies.** *Bioinformatics* 2007, **23**(13):1573–1579.
18. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W: **Scaffolding pre-assembled contigs using SSPACE.** *Bioinformatics* 2011, **27**(4):578–579.
19. Darling AE, Mau B, Perna NT: **ProgressiveMauve: multiple genome alignment with gene gain, loss and rearrangement.** *PLoS One* 2010, **5**(6):e11147.
20. Li H, Durbin R: **Fast and accurate long-read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2010, **26**(5):589–595.
21. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **Genome project data processing S: the sequence alignment/ Map format and SAMtools.** *Bioinformatics* 2009, **25**(16):2078–2079.
22. Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, Marshall D: **Tablet–next generation sequence assembly visualization.** *Bioinformatics* 2010, **26**(3):401–402.
23. Taylor J, Schenck I, Blankenberg D, Nekrutenko A: **Using galaxy to perform large-scale interactive data analyses.** *Curr Protoc Bioinformatics* 2007, **10**:10.5.
24. Baddam R, Kumar N, Shaik S, Suma T, Ngoi ST, Thong KL, Ahmed N: **Genome sequencing and analysis of Salmonella enterica serovar Typhi strain CR0063 representing a carrier individual during an outbreak of typhoid fever in Kelantan, Malaysia.** *Gut Pathogens* 2012, **4**(1):20.
25. Stothard P, Wishart DS: **Circular genome visualization and exploration using CGView.** *Bioinformatics* 2005, **21**:537–539.
26. Yamaichi Y, Iida T, Park KS, Yamamoto K, Honda T: **Physical and genetic map of the genome of Vibrio parahaemolyticus: presence of two chromosomes in Vibrio species.** *Mol Microbiol* 1999, **31**(5):1513–1521.
27. Egan ES, Waldor MK: **Distinct replication requirements for the two Vibrio cholerae chromosomes.** *Cell* 2003, **114**(4):521–530.
28. McKinley JE, Magnuson RD: **Characterization of the Phd repressor-antitoxin boundary.** *J Bacteriol* 2005, **187**(2):765–770.
29. Guerout AM, Iqbal N, Mine N, Ducos-Galand M, Van Melderen L, Mazel D: **Characterization of the phd-doc and ccd Toxin-Antitoxin Cassettes from Vibrio Superintegrons.** *J Bacteriol* 2013, **195**(10):2270–2283.
30. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, *et al*: **The RAST Server: rapid annotations using subsystems technology.** *BMC Genomics* 2008, **9**:75.
31. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B: **Artemis: sequence visualization and annotation.** *Bioinformatics* 2000, **16**(10):944–945.
32. Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW: **RNAmmer: consistent and rapid annotation of ribosomal RNA genes.** *Nucleic Acids Res* 2007, **35**(9):3100–3108.
33. Schattner P, Brooks AN, Lowe TM: **The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs.** *Nucleic Acids Res* 2005, **33**(Web Server issue):W686–W689.
34. Treangen TJ, Messeguer X: **M-GCAT: interactively and efficiently constructing large-scale multiple genome comparison frameworks in closely related species.** *BMC Bioinformatics* 2006, **7**:433.
35. Wang X, Wood TK: **Toxin-antitoxin systems influence biofilm and persister cell formation and the general stress response.** *Appl Environ Microbiol* 2011, **77**(16):5577–5583.
36. Park KS, Ono T, Rokuda M, Jang MH, Okada K, Iida T, Honda T: **Functional characterization of two type III secretion systems of Vibrio parahaemolyticus.** *Infect Immun* 2004, **72**(11):6659–6665.
37. Hurley CC, Quirke A, Reen FJ, Boyd EF: **Four genomic islands that mark post-1995 pandemic Vibrio parahaemolyticus isolates.** *BMC Genomics* 2006, **7**:104.

doi:10.1186/1757-4749-5-37
Cite this article as: Tiruvayipati *et al.*: Genome anatomy of the gastrointestinal pathogen, Vibrio parahaemolyticus of crustacean origin. *Gut Pathogens* 2013 **5**:37.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

Baddam et al. Gut Pathogens 2012, 4:20
http://www.gutpathogens.com/content/4/1/20

Gut Pathogens

GENOME ANNOUNCEMENT Open Access

Genome sequencing and analysis of *Salmonella enterica* serovar Typhi strain CR0063 representing a carrier individual during an outbreak of typhoid fever in Kelantan, Malaysia

Ramani Baddam[1], Narender Kumar[1], Sabiha Shaik[1], Tiruvayipati Suma[1,2], Soo Tein Ngoi[2,3], Kwai-Lin Thong[2,3] and Niyaz Ahmed[1,2*]

Abstract

Salmonella Typhi is a human restricted pathogen with a significant number of individuals as asymptomatic carriers of the bacterium. *Salmonella* infection can be effectively controlled if a reliable method for identification of these carriers is developed. In this context, the availability of whole genomes of carrier strains through high- throughput sequencing and further downstream analysis by comparative genomics approaches is very promising. Herein we describe the genome sequence of a *Salmonella* Typhi isolate representing an asymptomatic carrier individual during a prolonged outbreak of typhoid fever in Kelantan, Malaysia. Putative genomic coordinates relevant in pathogenesis and persistence of this carrier strain are identified and discussed.

Background

Salmonella enterica serovar Typhi, the aetiologic agent of typhoid fever is still posing a major health problem for the developing world, as about 16 million new cases are reported each year [1]. *S.* Typhi causes systemic infections (typhoid fever) as well as chronic infections (asymptomatic carriers) in humans, the latter serve as the source of infection [2]. The transmission of *S.* Typhi is primarily through faecal-oral route and a significant number of infected individuals become chronic asymptomatic carriers and keep shedding *S.* Typhi in faeces for decades [3]. This results in endemicity of *S.* Typhi in regions of the world with underdeveloped sanitation and community hygiene [4].

Carrier identification becomes extremely important as some of the ancestral haplotypes were observed in recent isolates suggesting their persistence in these asymptomatic carriers [5]. Traditional methods such as culturing of bacteria from faecal samples are not fool proof as the carriers shed bacteria intermittently. Serological tests to detect specific antibodies such as anti-H and anti-O are unable to differentiate between carriers and individuals who have recovered from the infection [6]. Especially, in areas endemic for *S.* Typhi, due to high background levels of these antibodies, serological tests cannot be adopted for the identification of a carrier [7]. Thus, there is an urgent need for inexpensive and efficient detection methods for the establishment of carrier state, perhaps based on genomic markers.

The genetic typing tools such as PFGE, AFLP, ribotyping etc. can resolve limited genetic variation occurring within specific sites, and therefore are incapable of differentiating highly clonal strains such as outbreak related strains from the ones not associated with the outbreak (carrier isolates) [8-10]. High-throughput sequencing technologies have already been employed as a high resolution molecular epidemiologic tool to discern microevolution of highly related strains [11].

In this study, we attempted to determine if whole genome sequencing of *S.* Typhi isolated from a carrier individual can provide insights related to persistence and or adaptation mechanisms. We describe the genome

* Correspondence: ahmed.nizi@gmail.com
[1]Pathogen Biology Laboratory, Department of Biotechnology and Bioinformatics, School of Life Sciences, University of Hyderabad, Hyderabad, India
[2]Institute of Biological sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia
Full list of author information is available at the end of the article

BioMed Central

© 2012 Baddam et al.; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

sequence of a *Salmonella enterica* serovar Typhi strain (ST CR0063) isolated from a carrier individual during a prolonged outbreak of typhoid fever in Kelantan, Malaysia.

Results and discussion

Genome statistics

The size of the draft genome of *Salmonella* Typhi (ST CR0063) is 4,585,851 bp with a coding percentage of 86.1%. The G + C content of this strain is about 51.71%. The total number of CDS determined are 4946 with an average length of gene about 798 nucleotides. The genome of ST CR0063 revealed 77 tRNA and 22 rRNA genes. The subsystems distribution of basic metabolic machinery of this strain is represented in Figure 1. The assembled draft genome shows high degree of similarity and shared core genome regions with *Salmonella* Typhi ST BL196 [12], the one identified as associated with a typhoid outbreak in Kelantan during the same period (Figure 2).

Virulence factors

The gene *shdA*, a key factor predicted to be involved in persistence of the bacterium in the intestines [14] by binding to its extracellular matrix, was identified and annotated. This gene, by mimicking the host heparin, is able to bind to the extracellular matrix proteins, fibronectin and collagen, and probably plays an important role in carriers by contributing to prolonged faecal shedding [15]. The *fim* gene cluster [16] of chaperone –usher family involved in adhesion to non-phagocytic cells was detected along with its negative regulator fimW. Type IV pili and *agf* operon [17,18] encoding curli fimbriae which aid in attachment of the bacterium to intestinal villi and also with each other, were found in the genome. These adherence factors determine the sites of bacterial colonisation and thereby adaptation and pathogenicity of a particular strain [19,20].

The *S.* Typhi strain ST CR0063 genome also revealed *viaA* and *viaB* loci, the prime regulators of Vi antigen expression. The *viaB* locus contains all genes for the biosynthesis (*tviA-E*) and export (*vexA-E*) of the Vi antigen, a well-known virulence factor [21,22]. The *mgtC* gene involved in Magnesium uptake and ferric uptake regulators (fur) [23] were also identified in ST CR0063. The PhoPQ regulon [24], which induces cytokine secretion and cationic antimicrobial peptide resistance, was also found to be conserved in our carrier strain. The RpoS sigma factor needed to cope up with external stress and nutrient depletion conditions [25] was also identified and annotated. The co-ordinates of these virulence factors in the genome of ST CR0063 are depicted in Figure 3.

Phages and pathogenicity islands (PAIs)

The phages gifsy-1 and fels-2 [27] together with many phage proteins and a few hypothetical proteins were identified in the genome of ST CR0063 by various algorithms (See Methods for details). It is expected that these phages are acquired by horizontal gene transfer (HGT) events as they were embedded in some of the genomic islands recognized. The phage encoding SopE effector protein of SPI-1 (Salmonella Pathogenicity Island) was present in ST CR0063 as recognized in other Typhi genomes [28,29].

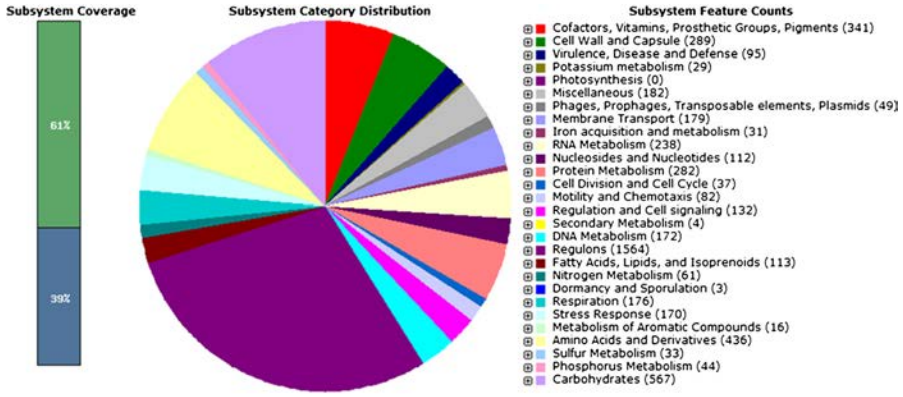


Figure 1 Subsystem distribution of ST CR0063. The subsystem statistics of ST CR0063 based on genome annotations performed according to RAST conventions.

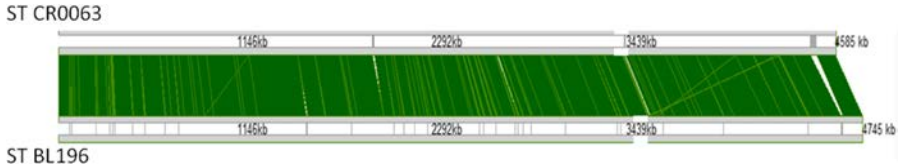


Figure 2 Comparison of *Salmonella* Typhi strains ST CR0063 and ST BL196. Comparison of whole genome sequences of *S. Typhi* strains using MG-CAT – one strain was isolated from a carrier individual (ST CR0063) and another from an infected individual (ST BL196) during a prolonged outbreak of Typhoid fever in Kelantan [13].

More than 15 PAIs that encode clusters of virulence associated genes have been identified across various serovars of *Salmonella enterica*. Ten pathogenicity islands have been identified by us in ST CR0063 and as expected [30], they were characterised by different G + C content and bounded by t-RNA genes. The SPI-1 type III secretion system (TTSS) structural genes spaMNOPQRS and invABCEFGH and their regulatory proteins HilA, HilC, HilD [31] were also identified and annotated. The SPI-1 secreted effector proteins SopE,

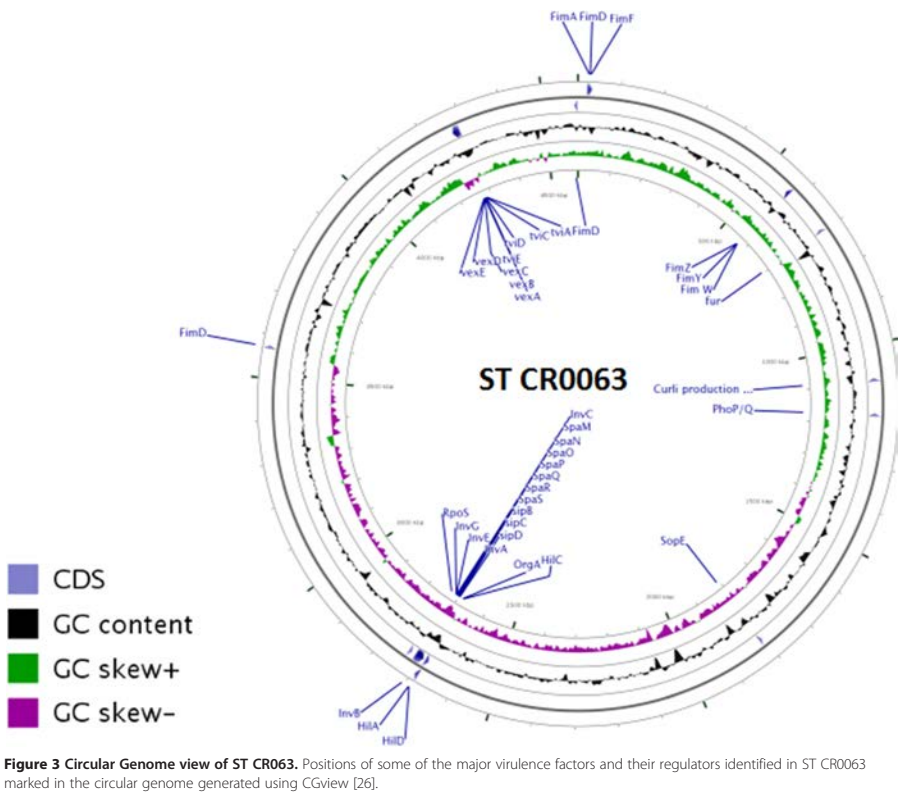


Figure 3 Circular Genome view of ST CR063. Positions of some of the major virulence factors and their regulators identified in ST CR0063 marked in the circular genome generated using CGview [26].

SopE2, SipA, SipB, SipC and SptP required for endothelial uptake and invasion [32] are also present. The genes SpiC, SseF, SseG, SifA, SifB secreted by SPI-2 TTSS and that are needed for survival in macrophages and colonisation of host organs [33] were also recognised in the present genome. The known regulators of SPI-2, OmpR-EnvZ and PhoP-PhoQ [34] were present. SPI-3, identified by us, contained magnesium transport genes *mgtC* and *marT* which are required for survival in macrophages [35]. Type I secretion system and its associated proteins encoded by SPI-4, and that are involved in the invasion of the intestinal epithelium [36], were also located in the present genome. The SPI-1 effector proteins SopB and PipB associated with enteritis and coded by SPI-5 [37] were also detected and annotated. The chaperone-usher fimbrial operons carried by SPI-6, SPI-10 and bacteriocin immunity proteins carried by SPI-8 [38] were identified. The SPI-7 and SPI-9 were identified in the ST CR0063 genome and were found to encode *viaB* locus, type IV pili formation proteins and TISS [38,39].

Conclusions and prospective

The genomic blueprint of *Salmonella* Typhi isolate ST CR0063 was elucidated in this study. The genome sequence information presented herein may be harnessed to guide comparative genomics and identification of novel and specific diagnostic markers. However, further studies involving large scale genome sequencing of the strains from several of the endemic countries and especially those from carrier individuals of different socio-economical settings is needed to develop a reliable approach to decipher the characteristics of a carrier state. Also, it will be required to determine the true extent of the diversity of carrier strains as juxtaposed to their acutely pathogenic forms in terms of 1) gene gain/loss during colonization and adaptation; 2) dynamics of virulence acquisition/attenuation; 3) possible genomic rearrangements; and 4) the relative preponderance of carrier and virulent strains circulating in different endemic regions of the world. Finally, an in-depth analysis of the host-pathogen interactions and their influence on gut microbiota can only explain the adaptation and persistence mechanisms of the (asymptomatic) carrier strains.

Methods

Genome sequencing

DNA was isolated from the stool sample of an asymptomatic carrier individual from Kelantan, Malaysia in 2007 during a prolonged outbreak. The draft genome sequence of this strain (STCR0063) was determined by Illumina Genome Analyzer (GAIIx, pipe- line ver l.6). The 100 bp paired-end sequencing was done with an insert size of 300 bp. About 67X genome coverage was achieved and 1.9 gigabytes of data were obtained.

Assembly and annotation

The sequence data were assembled *de-novo* in the same way as described previously [40-45] into 538 contigs using Velvet [46] at optimal hash length 39. SSPACE [47] was used for scaffolding the pre-assembled contigs using paired-end data. The gaps within these scaffolds were filled using Gapfiller by aligning the reads against already generated Scaffolds by SSPACE [48].

A reference guided assembly was generated by aligning reads to *Salmonella* Typhi str. CT18 [GenBank: AL513382.1] using bwa tools [49]. This reference guided assembly was used to re-order the scaffolds generated in de-novo way. In-house written Perl scripts were used for this re-ordering process and to finalize the gaps. The de novo and reference guided approaches were used to finalize the consensus draft genome. The reference guided assembly and reordered scaffolds were loaded on to Tablet – NGS data visualisation tool, to visualise the repeats, insertions and deletions [50].

The final draft nucleotide sequence after manual curation was annotated in our laboratory using RAST [51] and ISGA pipeline [52]. The genome statistics were gleaned using Artemis [53]. The data were further validated using gene prediction tools such as Glimmer [54] and EasyGene [55]. The RNAmmer [56] and tRNAscan-SE [57] were used to identify rRNA and tRNA respectively.

Phages and PAIs

Prophages and putative phage like elements in the genome were identified using PhiSpy [58] and Prophage Finder [59]. The putative HGT events were determined using Alien Hunter tool [60]. An integrated interface Island Viewer was used to predict putative genomic islands within the genome [61].

Sequence data access

The Salmonella enterica subsp. enterica serovar Typhi str. CR0063 whole genome shotgun (WGS) project has been submitted to the GenBank and has the project accession AKIC00000000. The project version entailing draft assembly described herein has the accession number AKIC01000000, and consists of sequences AKIC01000001-AKIC01000538.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

NA designed the study, interpreted the results and edited the manuscript. RB and NK managed Illumina sequencing, made the assemblies, analyzed the genome, and performed annotations. SS and TS provided computational tools and contributed to automation of the analysis process. KT provided

inputs related to the outbreak and the strain features, characterized the strain and maintained it in pure cultures. STN contributed to microbiology of the strain and prepared high molecular weight DNA for genome sequencing. All the authors read and approved the manuscript prior to submission.

Acknowledgements
We thankfully acknowledge support received from the University of Malaya High Impact Research Grant (Ref. UM.C/625/1HIR/MOHE/02 [A000002-5000 1]) - MOLECULAR GENETICS.

Author details
[1]Pathogen Biology Laboratory, Department of Biotechnology and Bioinformatics, School of Life Sciences, University of Hyderabad, Hyderabad, India. [2]Institute of Biological sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia. [3]Laboratory of Biomedical Science and Molecular Microbiology, UMBIO Research Cluster, University of Malaya, Kuala Lumpur, Malaysia.

Received: 14 November 2012 Accepted: 29 November 2012
Published: 13 December 2012

References

1. Parry CM, Hien TT, Dougan G, White NJ, Farrar JJ: **Typhoid fever.** *N Engl J Med* 2002, **347:**1770–1782.
2. Boyle EC, Bishop JL, Grassl GA, Finlay BB: **Salmonella: from pathogenesis to therapeutics.** *J Bacteriol* 2007, **189:**1489–1495.
3. Gonzalez-Escobedo G, Marshall JM, Gunn JS: **Chronic and acute infection of the gall bladder by Salmonella Typhi: understanding the carrier state.** *Nat Rev Microbiol* 2011, **9:**9–14.
4. Gopinath S, Carden S, Monack D: **Shedding light on Salmonella carriers.** *Trends Microbiol* 2012, **20:**320–327.
5. Roumagnac P, Weill FX, Dolecek C, Baker S, Brisse S, Chinh NT, Le TA, Acosta CJ, Farrar J, Dougan G, Achtman M: **Evolutionary history of Salmonella Typhi.** *Science* 2006, **314:**1301–1304.
6. Olopoenia LA, King AL: **Widal agglutination test – 100 years later: still plagued by controversy.** *Postgrad Med J* 2000, **76:**80–84.
7. Gupta A, My Thanh NT, Olsen SJ, Sivapalasingam S, My Trinh TT, Phuong Lan NT, Hoekstra RM, Bibb W, Minh NT, Danh TP, Cam PD, Mintz ED: **Evaluation of community-based serologic screening for identification of chronic Salmonella Typhi carriers in Vietnam.** *Int J Infect* 2006, **10:**309–314.
8. Nair S, Schreiber E, Thong KL, Pang T, Altwegg M: **Genotypic characterization of Salmonella typhi by amplified fragment length polymorphism fingerprinting provides increased discrimination as compared to pulsed-field gel electrophoresis and ribotyping.** *J Microbiol Methods* 2000, **41:**35–43.
9. Thong KL, Puthucheary S, Yassin RM, Sudarmono P, Padmidewi M, Soewandojo E, Handojo I, Sarasombath S, Pang T: **Analysis of *Salmonella* typhi isolates from southeast Asia by pulsed-field gel electrophoresis.** *J Clin Micro boil* 1995, **33:**1938–1941.
10. Baddam R, Thong KL, Avasthi TS, Shaik S, Yap KP, Teh CS, Chai LC, Kumar N, Ahmed N: **Whole-genome sequences and comparative genomics of *Salmonella enterica* serovar Typhi isolates from patients with fatal and nonfatal typhoid fever in Papua New Guinea.** *J Bacteriol* 2012, **194:**5122–5123.
11. Aziz RK, Nizet V: **Pathogen microevolution in high resolution.** *Sci Transl Med* 2010, **2**(16):16–4.
12. Baddam R, Kumar N, Thong KL, Ngoi ST, Teh CS, Yap KP, Chai LC, Avasthi TS, Ahmed N: **Genetic fine structure of a Salmonella enterica serovar Typhi strain associated with the 2005 outbreak of typhoid fever in Kelantan, Malaysia.** *J Bacteriol* 2012, **194:**3565–3566.
13. Treangen T, Messeguer X: **M-GCAT: Interactively and efficiently constructing large-scale multiple genome comparison frameworks in closely related species.** *BMC Bioinformatics* 2006, **7:**433.
14. Kingsley RA, Santos RL, Keestra AM, Adams LG, Bäumler AJ: **Salmonella enterica serotype Typhimurium ShdA is an outer membrane fibronectin-binding protein that is expressed in the intestine.** *Mol Microbiol* 2002, **43:**895–905.
15. Kingsley RA, van Amsterdam K, Kramer N, Bäumler AJ: **The *shdA* gene is restricted to serotypes of Salmonella enterica subspecies I and contributes to efficient and prolonged fecal shedding.** *Infect Immun* 2000, **68:**2720–2727.
16. Muscas P, Rossolini GM, Chiesurin A, Santucci A, Satta G: **Purification and characterization of type 1 fimbriae of Salmonella typhi.** *Microbiol Immunol* 1994, **38:**353–358.
17. Craig L, Pique ME, Tainer JA: **Type IV pilus structure and bacterial pathogenicity.** *Nat Rev Microbiol* 2004, **2:**363–378.
18. Collinson SK, Clouthier SC, Doran JL, Banser PA, Kay WW: **Salmonella enteritidis agfBAC operon encoding thin, aggregative fimbriae.** *J Bacteriol* 1996, **178:**662–667.
19. Duncan MJ, Mann EL, Cohen MS, Ofek I, Sharon N, Abraham SN: **The distinct binding specificities exhibited by enterobacterial Type 1 fimbriae are determined by their fimbrial shafts.** *J Biol Chem* 2005, **280:**37707–37716.
20. Guo A, Cao S, Tu L, Chen P, Zhang C, Jia A, Yang W, Liu Z, Chen H, Schifferli DM: **FimH alleles direct preferential binding of Salmonella to distinct mammalian cells or to avian cells.** *Microbiology* 2009, **155:**1623–1633.
21. Virlogeux I, Waxin H, Ecobichon C, Popoff MY: **Role of the viaB locus in synthesis, transport and expression of Salmonella typhi Vi antigen.** *Microbiology* 1995, **141:**3039–3047.
22. Robbins JD, Robbins JB: **Re examination of the protective role of the capsular polysaccharide (Vi antigen) of Salmonella typhi.** *J Infect Dis* 1984, **150:**436–449.
23. Moncrief MB, Maguire ME: **Magnesium transport in prokaryotes.** *J Biol Inorg Chem* 1999, **4:**523–527.
24. Guo L, Lim KB, Gunn JS, Bainbridge B, Darveau RP, Hackett M, Miller SI: **Regulation of lipid A modifications by Salmonella typhimurium virulence genes phoP-phoQ.** *Science* 1997, **276:**250–253.
25. Chen CY, Eckmann L, Libby SJ, Fang FC, Okamoto S, Kagnoff MF, Fierer J, Guiney DG: **Expression of Salmonella typhimurium rpoS and rpoS-dependent genes in the intracellular environment of eukaryotic cells.** *Infect Immun* 1996, **64:**4739–4743.
26. Stothard P, Wishart DS: **Circular genome visualization and exploration using CGView.** *Bioinformatics* 2005, **21:**537–539.
27. Stanley TL, Ellermeier CD, Slauch JM: **Tissue-specific gene expression identifies a gene in the lysogenic phage Gifsy-1 that affects Salmonella enterica serovar typhimurium survival in Peyer's patches.** *J Bacteriol* 2000, **182:**4406–4413.
28. Mirold S, Rabsch W, Rohde M, Stender S, Tschape H, Russmann H, Igwe E, Hardt WD: **Isolation of a temperate bacteriophage encoding the type III effector protein SopE from an epidemic Salmonella typhimurium strain.** *Proc Natl Acad Sci U S A* 1999, **96:**9845–9850.
29. Hardt WD, Urlaub H, Galan JE: **A substrate of the centisome 63 type III protein secretion system of Salmonella typhimurium is encoded by acryptic bacteriophage.** *Proc Natl Acad Sci U S A* 1998, **95:**2574–2579.
30. Marcus SL, Brumell JH, Pfeifer CG, Finlay BB: **Salmonella pathogenicity islands: big virulence in small packages.** *Microbes Infect* 2000, **2:**145–156.
31. Lee VT, Schneewind O: **Type III secretion machines and the pathogenesis of enteric infections caused by Yersinia and Salmonella spp.** *Immunol Rev* 1999, **168:**241–255.
32. McGhie EJ, Brawn LC, Hume PJ, Humphreys D, Koronakis V: **Salmonella takes control: effector driven manipulation of the host.** *Curr Opin Microbiol* 2009, **12:**117–124.
33. Waterman SR, Holden DW: **Functions and effectors of the Salmonella pathogenicity island 2 type III secretion system.** *Cell Microbiol* 2003, **5:**501–511.
34. Garmendia J, Beuzón CR, Ruiz-Albert J, Holden DW: **The roles of SsrA-SsrB and OmpR-EnvZ in the regulation of genes encoding the Salmonella typhimurium SPI-2 type III secretion system.** *Microbiology* 2003, **149:**2385–2396.
35. Blanc-Potard AB, Solomon F, Kayser J, Groisman EA: **The SPI-3 Pathogenicity Island of Salmonella enterica.** *J Bacteriol* 1999, **181:**998–1004.
36. Gerlach RG, Claudio N, Rohde M, Jackel D, Wagner C, Hensel M: **Cooperation of Salmonella pathogenicity islands 1 and 4 is required to breach epithelial barriers.** *Cell Microbiol* 2008, **10:**2364–2376.
37. Wood MW, Jones MA, Watson PR, Hedges S, Wallis TS, Galyov EE: **Identification of a pathogenicity island required for Salmonella enteropathogenicity.** *Mol Microbiol* 1998, **29:**883–891.
38. Parkhill J, Dougan G, James KD, Thomson NR, Pickard D, Wain J, Churcher C, Mungall KL, Bentley SD, Holden MT, Sebaihia M, Baker S, Basham D, Brooks K, Chillingworth T, Connerton P, Cronin A, Davis P, Davies RM, Dowd L, White N, Farrar J, Feltwell T, Hamlin N, Haque A, Hien TT, Holroyd S, Jagels K, Krogh A, Larsen TS, Leather S, Moule S, O'Gaora P, Parry C, Quail M, Rutherford K, Simmonds M, Skelton J, Stevens K, Whitehead S, Barrell BG: **Complete genome sequence of a multiple drug resistant Salmonella enterica serovar TyphiCT18.** *Nature* 2001, **413:**848–852.
39. Pickard D, Wain J, Baker S, Line A, Chohan S, Fookes M, Barron A, Gaora PO, Chabalgoity JA, Thanky N, Scholes C, Thomson N, Quail M, Parkhill J,

Dougan G: **Composition, acquisition, and distribution of the Vi exopolysaccharide-encoding Salmonella enterica pathogenicity island SPI-7.** *J Bacteriol* 2003, **185:**5055–5065.
40. Avasthi TS, Devi SH, Taylor TD, Kumar N, Baddam R, Kondo S, Suzuki Y, Lamouliatte H, Mégraud F, Ahmed N: **Genomes of two chronological isolates (Helicobacter pylori 2017 and 2018) of the West African Helicobacter pylori strain 908 obtained from a single patient.** *J Bacteriol* 2011, **193:**3385–3386.
41. Avasthi TS, Kumar N, Baddam R, Hussain A, Nandanwar N, Jadhav S, Ahmed N: **Genome of multidrug-resistant uropathogenic Escherichia coli strain NA114 from India.** *J Bacteriol* 2011, **193:**4272–4273.
42. Devi SH, Taylor TD, Avasthi TS, Kondo S, Suzuki Y, Megraud F, Ahmed N: **Genome of Helicobacter pylori strain 908.** *J Bacteriol* 2010, **192:**6488–6489.
43. Siddavattam D, Karegoudar TB, Mudde SK, Kumar N, Baddam R, Avasthi TS, Ahmed N: **Genome of a novel isolate of Paracoccus denitrificans capable of degrading N, N-dimethylformamide.** *J Bacteriol* 2011, **193:**5598–5599.
44. Yap KP, Teh CS, Baddam R, Chai LC, Kumar N, Avasthi TS, Ahmed N, Thong KL: **Insights from the genome sequence of a Salmonella enterica serovar Typhi strain associated with a sporadic case of typhoid fever in Malaysia.** *J Bacteriol* 2012, **194:**5124–5125.
45. Yap KP, Gan HM, Teh CS, Baddam R, Chai LC, Kumar N, Tiruvayipati SA, Ahmed N, Thong KL: **Genome sequence and comparative pathogenomics analysis of a Salmonella enterica serovar Typhi strain associated with a typhoid carrier in Malaysia.** *J Bacteriol* 2012, **194:**5970–5971.
46. Zerbino DR, Birney E: **Velvet: algorithms for de novo short read assembly using de Bruijn graphs.** *Genome Res* 2008, **18:**821–829.
47. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W: **Scaffolding pre-assembled contigs using SSPACE.** *Bioinformatics* 2011, **27:**578–579.
48. Nadalin F, Vezzi F, Policriti A: **GapFiller: a de novo assembly approach to fill the gap within paired reads.** *BMC Bioinformatics* 2012, **13**(14):8. doi:10.1186/1471-2105-13-S14-S8.
49. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler Transform.** *Bioinformatics* 2009, **25:**1754–1760.
50. Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, Marshall D: **Tablet–next generation sequence assembly visualization.** *Bioinformatics* 2010, **26:**401–402.
51. Aziz RK, Devoid S, Disz T, Edwards RA, Henry CS, *et al*: **SEED Servers: high-performance access to the SEED genomes, annotations, and metabolic models.** *PLoS ONE* 2012, **7**(10):48053. doi:10.1371/journal.pone.0048053.
52. Hemmerich C, Buechlein A, Podicheti R, Revanna KV, Dong Q: **An Ergatis-based prokaryotic genome annotation web server.** *Bioinformatics* 2010, **26:**1122–1124.
53. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B: **Artemis: sequence visualization and annotation.** *Bioinformatics* 2000, **16:**944–945.
54. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL: **Improved microbial gene identification with GLIMMER.** *Nucleic Acids Res* 1999, **27:**4636–4641.
55. Larsen TS, Krogh A: **EasyGene—a prokaryotic gene finder that ranks ORFs by statistical significance.** *BMC Bioinformatics* 2003, **4:**21.
56. Lagesen K, Hallin PF, Rødland E, Stærfeldt HH, Ussery DW RT: **RNammer: consistent annotation of rRNA genes in genomic sequences.** *Nucleic Acids Res* 2007, **35:**3100–3108.
57. Schattner P, Brooks AN, Lowe TM: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.** *Nucleic Acids Res* 2005, **33:**W686–W689.
58. Akhter S, Aziz RK, Edwards RA: **PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies.** *Nucleic Acids Res* 2012, **40:**e126.
59. Bose M, Barber RD: **Prophage Finder: a prophage loci prediction tool for prokaryotic genome sequences.** *In Silico Biol* 2006, **6:**223–227.
60. Vernikos GS, Parkhill J: **Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands.** *Bioinformatics* 2006, **22:**2196–2203.
61. Langille MG, Brinkman FS: **IslandViewer: an integrated interface for computational identification and visualization of genomic islands.** *Bioinformatics* 2009, **25:**664–665.

doi:10.1186/1757-4749-4-20
Cite this article as: Baddam *et al.*: Genome sequencing and analysis of *Salmonella enterica* serovar Typhi strain CR0063 representing a carrier individual during an outbreak of typhoid fever in Kelantan, Malaysia. *Gut Pathogens* 2012 **4**:20.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

Journal of Bacteriology

Genome Sequence and Comparative Pathogenomics Analysis of a Salmonella enterica Serovar Typhi Strain Associated with a Typhoid Carrier in Malaysia

Kien-Pong Yap, Han Ming Gan, Cindy Shuan Ju Teh, Ramani Baddam, Lay-Ching Chai, Narender Kumar, Suma Avasthi Tiruvayipati, Niyaz Ahmed and Kwai-Lin Thong

J. Bacteriol. **2012, 194(21):5970. DOI: 10.1128/JB.01416-12.**

Updated information and services can be found at:
http://jb.asm.org/content/194/21/5970

These include:

REFERENCES This article cites 26 articles, 9 of which can be accessed free at:
http://jb.asm.org/content/194/21/5970#ref-list-1

CONTENT ALERTS Receive: RSS Feeds, eTOCs, free email alerts (when new articles cite this article), more»

Information about commercial reprint orders: **http://journals.asm.org/site/misc/reprints.xhtml**
To subscribe to to another ASM Journal go to: **http://journals.asm.org/site/subscriptions/**

Journals.ASM.org

Downloaded from http://jb.asm.org/ on April 30, 2014 by UNIVERSITY OF MALAYA

JB Journals.ASM.org

GENOME ANNOUNCEMENT

Genome Sequence and Comparative Pathogenomics Analysis of a *Salmonella enterica* Serovar Typhi Strain Associated with a Typhoid Carrier in Malaysia

Kien-Pong Yap,[a,b] Han Ming Gan,[c] Cindy Shuan Ju Teh,[a,b] Ramani Baddam,[d] Lay-Ching Chai,[a,b] Narender Kumar,[d] Suma Avasthi Tiruvayipati,[a,d] Niyaz Ahmed,[a,d] and Kwai-Lin Thong[a,b]

Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia[a]; Laboratory of Biomedical Science and Molecular Microbiology, UMBIO Research Cluster, University of Malaya, Kuala Lumpur, Malaysia[b]; ScienceVision SB, Setia Alam, SeksyenU13, Shah Alam, Selangor, Malaysia[c]; and Pathogen Biology Laboratory, Department of Biotechnology, School of Life Sciences, University of Hyderabad, Hyderabad, India[d]

***Salmonella enterica* serovar Typhi is a human pathogen that causes typhoid fever predominantly in developing countries. In this article, we describe the whole genome sequence of the *S.* Typhi strain CR0044 isolated from a typhoid fever carrier in Kelantan, Malaysia. These data will further enhance the understanding of its host persistence and adaptive mechanism.**

Typhoid fever caused by human-specific *Salmonella enterica* serovar Typhi (*S.* Typhi) remains a major health problem that affects 21.7 million people, with 217,000 deaths worldwide annually (6). *S.* Typhi is transmitted through the oral-fecal route and sometimes persists in the body, establishing an asymptomatic chronic carrier (10, 12). The risk of developing gallbladder diseases, including carcinoma, is also higher among typhoid carriers (5, 10, 21).

Although typhoid fever is endemic in many countries, including Malaysia, little is known about the mechanism of survival and persistence of *S.* Typhi in the host. Therefore, the genome sequence and comparative pathogenomics analysis of carrier strain will provide in-depth understanding of its persistence and adaptive mechanism within its host.

S. Typhi CR0044 was isolated from stool sample of a typhoid carrier in Kelantan, Malaysia, in 2007. This strain was subtyped as ST1 by multilocus sequence typing (14) and was highly similar to the outbreak strain in 2005 by pulsed field gel electrophoresis (PFGE) (2). Genome sequencing of *S.* Typhi strain CR0044 was performed using the Ilumina Genome Analyzer (GA2x, pipeline version 1.60, insert size 300), which generated 1.0 gigabyte of data with a 90× depth coverage and a 73-bp read length. Genome assembly was constructed with Velvet (26) using the *de novo* approach, which generated 201 contigs with a minimum contig length of more than 200 bp and an average size of 23,367 bp. The open reading frames (ORFs) of the resultant contigs were predicted using RAST (1) and Prodigal (13) and subsequently annotated using Blast2GO (4), whereas tRNA and rRNA genes were identified with tRNAscan-SE (17) and RNAmmer (15), respectively. The predicted genome size is approximately 4,769,054 bp, with an average GC content of 52.1% and coding percentage of 85.8. The genome revealed approximately 4,884 coding sequences (CDS) with an average length of 825 bp. The genome also contains predicted 69 tRNA and 22 rRNA genes.

The genome revealed a type III secretion system and flagellum subsystem as reported in *S.* Typhi strains Ty2 and CT18 (7, 12). The genome contains genes reported in Ty2 and CT18, such as the gene coding for type 4 fimbrial assembly protein, the *yjbEFGH* locus, *yhjD* conserved clusters, and *wca* genes, which are related to cell wall and biofilm formation and host persistence (3, 8, 7, 12, 18, 25). It is noteworthy that the genome sequence also revealed the presence of the GGDEF family protein YeaJ, which is associated with cell surface adhesion and biofilm formation, which was not identified in Ty2 and CT18 (9, 19). The gene encoding the rhamnogalacturonide transporter RhiT for rhamnose utilization was also found adjacent to a transposase gene in CR0044 (20). Interestingly, the genome also revealed a zonular occludens toxin family protein that was not previously reported in *Salmonella* spp.

S. Typhi in Southeast Asia is genetically diverse, with genome variations and clonal expansion reported (2, 11, 15, 16, 21, 22, 23, 24). The dynamic nature of the *S.* Typhi chromosome greatly enhances its persistence and adaptation within the host, which allows the pathogen to survive and thrive in typhoid carriers. The genomic information obtained here could unveil the genome evolution and mechanism involved in carrier-state transformation.

Nucleotide sequence accession numbers. This Whole Genome Shotgun project has been deposited in GenBank under accession no. AKZO00000000. The version described in this paper is the first version, AKZO01000000. The Bioproject designation for this project is PRJNA160187.

ACKNOWLEDGMENTS

This research is supported by a University of Malaya High Impact Research Grant—Molecular Genetics (reference no. UM.C/625/1HIR/MOHE/-02 [A000002-5000 1]), for which K.-L. Thong is the grant holder for the high-impact research project under the title "Pathogenomic and Phenomic of Food-Borne Disease." We also acknowledge Indo-German International Research Training Group—Internationales Graduiertenkolleg (GRK1673)—Functional Molecular Infection Epidemiology, an initiative of the German Research Foundation (DFG) with N.A. from the University of Hyderabad (India). We are also grateful to M/s Genotypic Technology Pvt., Ltd., Bengaluru, India, for support with Illumina sequencing.

We acknowledge Safwan Jusoh from the ICT department, University of Malaya, for assisting us with computing solutions and allowing us to

Received 6 August 2012 Accepted 17 August 2012
Address correspondence to Kwai-Lin Thong, thongkl@um.edu.my.
Copyright © 2012, American Society for Microbiology. All Rights Reserved.
doi:10.1128/JB.01416-12

Downloaded from http://jb.asm.org/ on April 30, 2014 by UNIVERSITY OF MALAYA

use their servers and computational facilities and Soo Tein Ngoi from LBSMM, IPS, for technical assistance with DNA preparation.

REFERENCES

1. **Aziz RK, et al.** 2008. The RAST server: rapid annotations using subsystems technology. BMC Genomics **9**:75. doi:10.1186/1471-2164-9-75.
2. **Baddam R, et al.** 2012. Genetic fine structure of a *Salmonella enterica* serovar Typhi strain associated with the 2005 outbreak of typhoid fever in Kelantan, Malaysia. J. Bacteriol. **194**:3565–3566.
3. **Cano DA, Bernal GD, Tierrez A, Portillo FG, Casadesus J.** 2002. Regulation of capsule synthesis and cell motility in *Salmonella enterica* by the essential gene igaA. Genetics **162**:1513–1523.
4. **Conesa A, et al.** 2005. Blast2Go: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics **21**: 3674–3676.
5. **Crawford RW, et al.** 2010. Gallstones play a significant role in Salmonella spp. gallbladder colonization and carriage. Proc. Natl. Acad. Sci. U. S. A. **107**:4353–4358.
6. **Crump J, Mintz E.** 2010. Global trends in typhoid and paratyphoid fever. Clin. Infect. Dis. **50**:241–246.
7. **Deng W, et al.** 2003. Comparative genomics of *Salmonella enterica* serovar Typhi strains Ty2 and CT18. J. Bacteriol. **185**:2330–2337.
8. **Ferrieres L, Aslam SN, Cooper RM, Clarke DJ.** 2007. The yjbEFGH locus in Escherichia coli K-12 is an operon encoding proteins involved in exopolysaccharide production. J. Microbiol. **153**:1070–1080.
9. **Garcia B, et al.** 2004. Role of the GGDEF protein family in Salmonella cellulose biosynthesis and biofilm formation. J. Mol. Microbiol. **54**:264–277.
10. **Gonzalez-Escobedo G, Marshall JM, Gunn JS.** 2011. Chronic and acute infection of the gall bladder by *Salmonella* Typhi: understanding the carrier state. Nat. Rev. Microbiol. **9**:9–14.
11. **Holt KE, et al.** 2011. Temporal fluctuation of multidrug resistant *Salmonella* Typhi haplotypes in the Mekong River delta region of Vietnam. PLoS Negl. Trop. Dis. **5**:e929. doi:10.1371/journal.pntd.0000929.
12. **Holt KE, et al.** 2008. High throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. Nat. Genet. **40**:987–993.
13. **Hyatt D, et al.** 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics **11**:119. doi: 10.1186/1471-2105-11-119.
14. **Kidgell C, et al.** 2002. Salmonella typhi, the causative agent of typhoid fever, is approximately 50,000 years old. Infect. Genet. Evol. **2**:39–45.
15. **Lagesen K, et al.** 2007. RNammer: consistent annotation of rRNA genes in genomic sequences. Nucleic Acids Res. **35**:3100–3108.
16. **Le TA, et al.** 2007. Clonal expansion and microevolution of quinolone-resistant *Salmonella enterica* serotype Typhi in Vietnam from 1996 to 2004. J. Clin. Microbiol. **45**:3485–3492.
17. **Lowe TM, Eddy SR.** 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. **25**: 955–964.
18. **Mamat U, et al.** 2008. Single amino acid substitutions in either YhjD or MsbA confer viability to 3-deoxy-D-manno-oct-2-ulosonic acid-depleted Escherichia coli. J. Mol. Microbiol. **67**:633–648.
19. **Parkhill J, et al.** 2001. Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. Nature **413**:848–852.
20. **Rodionov DA, Gelfand MS, Pattat NHC.** 2004. Comparative genomics of the KdgR regulon in Erwinia chrysanthemi 3937 and other gamma-proteobacteria. J. Microbiol. **150**:3571–3590.
21. **Shukla VK, Singh H, Pandey M, Upadhyay SK, Nath G.** 2000. Carcinoma of the gallbladder—is it a sequel of typhoid? Dig. Dis. Sci. **45**:900–903.
22. **Thong KL, Cheong YM, Puthucheary SD, Koh CL, Pang T.** 1994. Epidemiologic analysis of sporadic *Salmonella* Typhi isolates and those from outbreaks by pulsed-field gel electrophoresis, J. Clin. Microbiol. **32**: 1135–1141.
23. **Thong KL, Puthucheary SD, Pang T.** 1996. Genome size variation among recent human isolates of *Salmonella* Typhi. Res. Microbiol. **148**:229–235.
24. **Thong KL, et al.** 1995. Analysis of *Salmonella* Typhi from Southeast Asia by pulsed-field gel electrophoresis. J. Clin. Microbiol. **33**:1938–1941.
25. **Townsend SM, et al.** 2001. *Salmonella enterica* serovar Typhi possesses a unique repertoire of fimbrial gene sequences. J. Infect. Immun. **69**:2894–2901.
26. **Zerbino DR, Birney E.** 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. **18**:821–829.

Downloaded from http://jb.asm.org/ on April 30, 2014 by UNIVERSITY OF MALAYA

Journal of Bacteriology

Next-Generation Sequencing and *De Novo* Assembly, Genome Organization, and Comparative Genomic Analyses of the Genomes of Two Helicobacter pylori Isolates from Duodenal Ulcer Patients in India

Narender Kumar, Asish K. Mukhopadhyay, Rajashree Patra, Ronita De, Ramani Baddam, Sabiha Shaik, Jawed Alam, Suma Tiruvayipati and Niyaz Ahmed
J. Bacteriol. **2012, 194(21):5963. DOI: 10.1128/JB.01371-12.**

Updated information and services can be found at:
http://jb.asm.org/content/194/21/5963

These include:

REFERENCES This article cites 19 articles, 7 of which can be accessed free at:
http://jb.asm.org/content/194/21/5963#ref-list-1

CONTENT ALERTS Receive: RSS Feeds, eTOCs, free email alerts (when new articles cite this article), more»

Information about commercial reprint orders: **http://journals.asm.org/site/misc/reprints.xhtml**
To subscribe to to another ASM Journal go to: **http://journals.asm.org/site/subscriptions/**

Journals.ASM.org

Downloaded from http://jb.asm.org/ on April 30, 2014 by UNIVERSITY OF MALAYA

JB Journals.ASM.org

GENOME ANNOUNCEMENT

Next-Generation Sequencing and *De Novo* Assembly, Genome Organization, and Comparative Genomic Analyses of the Genomes of Two *Helicobacter pylori* Isolates from Duodenal Ulcer Patients in India

Narender Kumar,[a] Asish K. Mukhopadhyay,[b] Rajashree Patra,[b] Ronita De,[b] Ramani Baddam,[a] Sabiha Shaik,[a] Jawed Alam,[b] Suma Tiruvayipati,[a,c] and Niyaz Ahmed[a,c,d]

Pathogen Biology Laboratory, Department of Biotechnology, School of Life Sciences, University of Hyderabad, Gachibowli, Hyderabad, India[a]; Division of Bacteriology, National Institute of Cholera and Enteric Diseases (Indian Council of Medical Research), Kolkata, India[b]; Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia[c]; and Institute of Life Sciences, University of Hyderabad Campus, Gachibowli, Hyderabad, India[d]

The prevalence of different *H. pylori* genotypes in various geographical regions indicates region-specific adaptations during the course of evolution. Complete genomes of *H. pylori* from countries with high infection burdens, such as India, have not yet been described. Herein we present genome sequences of two *H. pylori* strains, NAB47 and NAD1, from India. In this report, we briefly mention the sequencing and finishing approaches, genome assembly with downstream statistics, and important features of the two draft genomes, including their phylogenetic status. We believe that these genome sequences and the comparative genomics emanating thereupon will help us to clearly understand the ancestry and biology of the Indian *H. pylori* genotypes, and this will be helpful in solving the so-called Indian enigma, by which high infection rates do not corroborate the minuscule number of serious outcomes observed, including gastric cancer.

Helicobacter pylori's coevolution with its host (10, 11, 16) and its tight compartmentalization (13, 16, 18, 19) into several different populations and subpopulations have delivered an excellent premise to pursue the idea of geographic evolution/spread of humans and their pathogens from Africa and to gain insights into pathogen adaptation mechanisms (1, 3). Based partly on these conventions, Indian *H. pylori* isolates have shown to have European origins (9) and are widely held as mostly innocuous or only mildly pathogenic. The severity of *H. pylori*-induced gastro-duodenal diseases and their outcomes vary in different geographic regions and populations, which may be significantly attributable to different genetic compositions of the underlying bacterial strains. More data based on genome sequences from many of strains from different countries are needed to clearly establish the genetic makeup, colonization potential, and virulence characteristics of a particular strain or genotype. In view of this, genome sequence-based characterizations of strains prevalent in different locales is necessary (2).

We describe genomes of *H. pylori* strains NAB47 (Bangalore) and NAD1 (Delhi) from duodenal ulcer patients. Illumina sequencing was performed as described previously (4, 8); briefly, about 3 gigabytes and 1.8 gigabytes of data comprising 72-bp paired-end reads (insert size, 300 bp) provided genome coverages of approximately 300× and 200×, respectively. The raw reads were filtered using the FASTX tool kit (17) and assembled using Velvet (20); the reads yielded 107 (NAB47) and 103 (NAD1) contigs with a hash length set to 37. These contigs were joined into 34 (NAB47) and 48 (NAD1) scaffolds by using SSPACE (6). The scaffolds were aligned and ordered according to their closest reference genome and confirmed using BLAST (12) and Mummer (14). The draft genomes were submitted to RAST (5) for annotation, and the output was validated by using Glimmer (7) and EasyGene (15).

The draft genomes of *H. pylori* NAB47 and NAD1 had sizes of about 1,590,862 bp and 1,588,938 bp, respectively, with G+C contents of 39.17 and 39.03%, respectively. The genomes revealed coding percentages of 91.5% (NAB47) and 91.3% (NAD1) and encoded 1,572 and 1,567 proteins, respectively; each of the genomes contained 36 tRNA genes and 6 rRNA genes. The average lengths for protein-coding genes were found to be 929 bp and 922 bp, respectively. Major virulence markers, such as *cagA*, *vacA*, the whole *cag* pathogenicity island, and several outer membrane proteins of the Hop family, were annotated. In addition, NAD1 harbored two plasmids of 16 kb and 10 kb each that carried genes for transposase, IS*606*, and mobilization proteins, together with replication protein A. CagA protein in both of the strains contained EPIYA D-type motifs, which are typical of Indo-European strains. Important plasticity region genes, such as jhp0940, jhp0947, and *dupA*, were absent, and hp0986 was detected only in NAB47. Finally, whole-genome phylogeny incorporating all the available genomes reconfirmed an Indo-European ancestry (HpEurope).

We believe that the genomes described herein are likely to rekindle our knowledge of the genetic makeup and evolutionary relationships of *H. pylori* in India. Comparative genomic analyses extending out to other unexplored strains from the tribal and mainstream populations will facilitate understanding of the true pathogenic potential (amid adaptive evolution) of the Indian *H. pylori*. Furthermore, they will be immensely helpful in global epidemiological studies and also for the development of diagnostic tools tailored to a particular host population.

Received 31 July 2012 Accepted 17 August 2012
Address correspondence to Niyaz Ahmed, niyazSL@uohyd.ernet.in.
N.K. and A.K.M. contributed equally to this work.
Copyright © 2012, American Society for Microbiology. All Rights Reserved.
doi:10.1128/JB.01371-12

Downloaded from http://jb.asm.org/ on April 30, 2014 by UNIVERSITY OF MALAYA

Nucleotide sequence accession numbers. The genome sequences of *H. pylori* NAB47 and NAD1 have been deposited with GenBank and assigned accession numbers AJFA00000000 and AJGJ00000000, respectively. The updated sequences/contigs are also available for download from the International Society for Genomic and Evolutionary Microbiology (ISOGEM) server (http://isogem.org/HPNAB47.txt and http://isogem.org/HPNAD1.txt).

ACKNOWLEDGMENTS

We acknowledge support from the University of Malaya High Impact Research Grant (UM.C/625/1HIR/MOHE/CHAN-02)-Molecular Genetics. A.K.M. acknowledges support from the Department of Biotechnology (BT/PR10407/BRB/10/604/2008) and Indian Council of Medical Research. These genomes were completed under the wider umbrella of the Indo-German International Research Training Group, Internationales Graduiertenkolleg (GRK1673), Functional Molecular Infection Epidemiology, an initiative of the German Research Foundation (DFG) and the University of Hyderabad (India). N.K. would like to acknowledge a Junior Research Fellowship received from the Council of Scientific and Industrial Research (CSIR), India, and J.A. acknowledges ICMR for a Senior Research Fellowship.

We are also grateful to M/s Genotypic Technology Pvt. Ltd., Bengaluru, India, for their efforts with the Illumina sequencing. We acknowledge the Bioinformatics Facility (BIF) at the Department of Biotechnology, University of Hyderabad, for use of their computational infrastructure. Further, we thank Akash Ranjan for helpful discussions and for enabling access to the SUN Microsystems CDFD Centre of Excellence for some of our data analyses.

REFERENCES

1. **Ahmed N.** 2011. Coevolution and adaptation of *Helicobacter pylori* and the case for 'functional molecular infection epidemiology.' Med. Princ. Pract. **20**:497–503.
2. **Ahmed N.** 2009. A flood of microbial genomes: do we need more? PLoS One **4**:e5831. doi:10.1371/journal.pone.0005831.
3. **Atherton JC, Blaser MJ.** 2009. Coadaptation of *Helicobacter pylori* and humans: ancient history, modern implications. J. Clin. Invest. **119**:2475–2487.
4. **Avasthi TS, et al.** 2011. Genomes of two chronological isolates (*Helicobacter pylori* 2017 and 2018) of the West African *Helicobacter pylori* strain 908 obtained from a single patient. J. Bacteriol. **193**:3385–3386.
5. **Aziz RK, et al.** 2008. The RAST server: rapid annotations using subsystems technology. BMC Genomics **9**:75. doi:10.1186/1471-2164-9-75.
6. **Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W.** 2011. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics **5**:578–579.
7. **Delcher AL, et al.** 1999. Improved microbial gene identification with Glimmer. Nucleic Acids Res. **27**:4636–4641.
8. **Devi SH, et al.** 2010. Genome of *Helicobacter pylori* strain 908. J. Bacteriol. **192**:6488–6489.
9. **Devi SM, et al.** 2007. Ancestral European roots of *Helicobacter pylori* in India. BMC Genomics **8**:184. doi:10.1186/1471-2164-8-184.
10. **Devi SM, et al.** 2006. Genomes of *Helicobacter pylori* from native Peruvians suggest admixture of ancestral and modern lineages and reveal a Western type *cag*-pathogenicity island. BMC Genomics **7**:191. doi:10.1186/1471-2164-7-191.
11. **Falush D, et al.** 2003. Traces of human migrations in *Helicobacter* pylori populations. Science **299**:1582–1585.
12. **Kent WJ.** 2002. BLAT: the BLAST-like alignment tool. Genome Res. **12**:656–664.
13. **Kersulyte D, et al.** 2010. Helicobacter pylori from Peruvian Amerindians: traces of human migrations in strains from remote Amazon, and genome sequence of an Amerind strain. PLoS One **5**:e15076. doi:10.1371/journal.pone.0015076.
14. **Kurtz S, et al.** 2004. Versatile and open software for comparing large genomes. Genome Biol. **5**:R12. doi:10.1186/gb-2004-5-2-r12.
15. **Larsen TS, Krogh A.** 2003. EasyGene: a prokaryotic gene finder that ranks ORFs by statistical significance. BMC Bioinformatics **4**:21. doi:10.1186/1471-2105-4-21.
16. **Linz B, et al.** 2007. An African origin for the intimate association between humans and *Helicobacter pylori.* Nature **445**:915–918.
17. **Taylor J, Schenck I, Blankenberg D, Nekrutenko A.** 2007. Using Galaxy to perform large-scale interactive data analyses. Curr. Protoc. Bioinformatics **Chapter 10:**Unit 10.5.
18. **Wirth T, et al.** 2004. Distinguishing human ethnic groups by means of sequences from *Helicobacter pylori*: lessons from Ladakh. Proc. Natl. Acad. Sci. U. S. A. **101**:4746–4751.
19. **Yamaoka Y.** 2009. Helicobacter pylori typing as a tool for tracking human migration. Clin. Microbiol. Infect. **15**:829–834.
20. **Zerbino DR, Birney E.** 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. **18**:821–829.

Downloaded from http://jb.asm.org/ on April 30, 2014 by UNIVERSITY OF MALAYA

Journal of Bacteriology

Insights from the Genome Sequence of a Salmonella enterica Serovar Typhi Strain Associated with a Sporadic Case of Typhoid Fever in Malaysia

Kien-Pong Yap, Cindy Shuan Ju Teh, Ramani Baddam, Lay-Ching Chai, Narender Kumar, Tiruvayipati Suma Avasthi, Niyaz Ahmed and Kwai-Lin Thong

J. Bacteriol. **2012, 194(18):5124. DOI: 10.1128/JB.01062-12.**

Updated information and services can be found at:
http://jb.asm.org/content/194/18/5124

These include:

REFERENCES This article cites 25 articles, 14 of which can be accessed free at: http://jb.asm.org/content/194/18/5124#ref-list-1

CONTENT ALERTS Receive: RSS Feeds, eTOCs, free email alerts (when new articles cite this article), more»

Information about commercial reprint orders: **http://journals.asm.org/site/misc/reprints.xhtml**
To subscribe to to another ASM Journal go to: **http://journals.asm.org/site/subscriptions/**

Journals.ASM.org

Downloaded from http://jb.asm.org/ on April 30, 2014 by UNIVERSITY OF MALAYA

JB Journals.ASM.org

GENOME ANNOUNCEMENT

Insights from the Genome Sequence of a *Salmonella enterica* Serovar Typhi Strain Associated with a Sporadic Case of Typhoid Fever in Malaysia

Kien-Pong Yap,[a,b] Cindy Shuan Ju Teh,[a,b] Ramani Baddam,[c] Lay-Ching Chai,[a,b] Narender Kumar,[c] Tiruvayipati Suma Avasthi,[a,c] Niyaz Ahmed,[a,c] and Kwai-Lin Thong[a,b]

Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia[a]; Laboratory of Biomedical Science and Molecular Microbiology, UMBIO Research Cluster, University of Malaya, Kuala Lumpur, Malaysia[b]; and Pathogen Biology Laboratory, Department of Biotechnology, School of Life Sciences, University of Hyderabad, Hyderabad, India[c]

***Salmonella enterica* serovar Typhi is the causative agent of typhoid fever, which causes nearly 21.7 million illnesses and 217,000 deaths globally. Herein, we describe the whole-genome sequence of the *Salmonella* Typhi strain ST0208, isolated from a sporadic case of typhoid fever in Kuala Lumpur, Malaysia. The whole-genome sequence and comparative genomics allow an in-depth understanding of the genetic diversity, and its link to pathogenicity and evolutionary dynamics, of this highly clonal pathogen that is endemic to Malaysia.**

Salmonella enterica serovar Typhi (*S.* Typhi) is a human intracellular pathogen of global importance, infecting 21.7 million people and causing 217,000 deaths annually (4). The challenges of higher incidence of typhoid fever in developing countries have led to an increased health burden (4).

In 2008, 201 cases of sporadic outbreaks were reported in Malaysia (18). The genome sequence of this sporadically associated strain will provide insights into possible genetic events that would confer a fitness advantage.

S. Typhi strain ST0208 was isolated from the stool sample of a typhoid fever patient admitted to University Malaya Medical Centre (UMMC), Kuala Lumpur, Malaysia, in 2008. The strain was characterized by pulsed-field gel electrophoresis (PFGE), repetitive extragenic palindromic (REP)-PCR, and antimicrobial susceptibility profiling (25). The genome sequence of *S.* Typhi ST0208 was determined using an Illumina genome analyzer (GA2x, pipeline version 1.60) with an insert size of 300 bp, which generated 1.83 gigabytes of data with an average coverage of 165× and yielded 1,499,986 paired-end reads with a 100-bp read length. Genome assembly was constructed *de novo* using Velvet (26), which generated 222 contigs. The resultant contigs were uploaded into the RAST server (2, 17) to predict the open reading frames (ORFs) by using Glimmer3 (5) and validated with the ISGA integrated system (8). In brief, the predicted ORFs were annotated by searching against clusters of orthologous group (21) and SEED (7) databases, whereas tRNA and rRNA genes were identified by using tRNAscan-SE (15) and RNAmmer (12), respectively. The draft genome size is approximately 4,798,272 bp in length, with an average GC content of 52.0%, and is composed of 4,890 predicted coding sequences with an average length of 810 bp. A mean percentage of 83.7% of nucleotides of the genome are predicted to encode proteins. The genome reveals 71 tRNA and 22 rRNA predicted genes.

The genome contains several monosaccharide and polysaccharide metabolism-related genes, which were not reported in *S.* Typhi strains Ty2 and CT18 (6, 11, 18), such as D-galactarate permease, gluconate permease, tagatose-6-phosphate kinase, trehalase, and arabinose-proton transporter, that could be associated with host persistence (20). The genome sequence revealed four multidrug resistance clusters, *mdtABCR* and *marABCR* proteins (1, 3, 19), DNA gyrase subunit A and B, and topoisomerase subunit (IV) A and B (9), which were also identified in *S.* Typhi Ty2 and *S.* Typhi CT18 (6, 11, 19). Hypothetical proteins and pathogenicity islands were annotated.

High genetic diversity of *S.* Typhi was detected among human strains in Malaysia and Southeast Asia (10, 14, 22, 23, 24). Variation in the genome sequence revealed by the strain ST0208 is consistent with the proposed key theory of persistent adaptation and optimization of function (13, 16, 19). Genomic information from locality-specific strains associated with clinical manifestation allows genome evolution and endemicity to be studied.

Nucleotide sequence accession numbers. This whole-genome shotgun project has been deposited in DDBJ/EMBL/GenBank under the accession number AJXA00000000. The version described in this paper is the first version, AJXA01000000. The BioProject designation for this project is PRJNA160181.

ACKNOWLEDGMENTS

This research is supported by the University of Malaya High Impact Research Grant (reference UM.C/625/1HIR/MOHE/-02 [A000002-50001]), Molecular Genetics, in which K.-L. Thong is the grant holder for the high-impact research project under the title of Pathogenomics and Phenomics of Food-borne Bacterial Diseases.

We are also grateful to M/s Genotypic Technology Pvt. Ltd., Bengaluru, India, for their assistance with Illumina sequencing. We acknowledge Safwan Jusoh from the ICT department, University of Malaya, for assisting us with computing solutions and allowing us to use the servers and computational facilities and S. T. Ngoi from LBSMM, IPS, for technical assistance with DNA preparation.

REFERENCES

1. **Alekshun MN, Levy SB.** 1999. The mar regulon: multiple resistance to antibiotics and other toxic chemicals. Trends Microbiol. 7:410–413.

Received 15 June 2012 Accepted 29 June 2012
Address correspondence to Kwai-Lin Thong, thongkl@um.edu.my.
Copyright © 2012, American Society for Microbiology. All Rights Reserved.
doi:10.1128/JB.01062-12

Downloaded from http://jb.asm.org/ on April 30, 2014 by UNIVERSITY OF MALAYA

2. **Aziz RK, et al.** 2008. The RAST server: rapid annotations using subsystems technology. BMC Genomics **9**:75.
3. **Baddam R, et al.** 2012. Genetic fine structure of a *Salmonella enterica* serovar Typhi strain associated with the 2005 outbreak of typhoid fever in Kelantan, Malaysia. J. Bacteriol. **194**:3565–3566.
4. **Crump J, Mintz E.** 2010. Global trends in typhoid and paratyphoid fever. Clin. Infect. Dis. **50**:241–246.
5. **Delcher AL, Bratke KA, Powers EC, Salzberg SL.** 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics **23**:673–679.
6. **Deng W, et al.** 2003. Comparative genomics of *Salmonella enterica* serovar Typhi strains Ty2 and CT18. J. Bacteriol. **185**:2330–2337.
7. **Disz T, et al.** 2010. Accessing the SEED genome database via web services API: tools for programmers. BMC Bioinformatics **11**:319.
8. **Hemmerich C, Buechlein A, Podicheti R, Revanna KV, Dong Q.** 2010. An Ergatis-based prokaryotic genome annotation web server. Bioinformatics **26**:1122–1124.
9. **Hirose K, et al.** 2002. DNA sequence analysis of DNA gyrase and DNA topoisomerase IV quinolone resistance-determining regions of *Salmonella enterica* serovar Typhi and serovar Paratyphi A. Antimicrob. Agents Chemother. **46**:3249–3252.
10. **Holt KE, et al.** 2011. Temporal fluctuation of multidrug resistant *Salmonella* Typhi haplotypes in the Mekong River delta region of Vietnam. PLoS Negl. Trop. Dis. **5**:e929.
11. **Holt KE, et al.** 2008. High throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. Nat. Genet. **40**:987–993.
12. **Lagesen K, et al.** 2007. RNAmmer: consistent annotation of rRNA genes in genomic sequences. Nucleic Acids Res. **35**:3100–3108.
13. **Lapierre P, Gogarten JP.** 2009. Estimating the size of the bacterial pan-genome. Trends Genet. **25**:107–110.
14. **Le TA, et al.** 2007. Clonal expansion and microevolution of quinolone-resistant *Salmonella enterica* serotype Typhi in Vietnam from 1996 to 2004. J. Clin. Microbiol. **45**:3485–3492.
15. **Lowe TM, Eddy SR.** 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. **25**: 955–964.
16. **Matthews TD, Rabsch W, Maloy S.** 2011. Chromosomal rearrangements in *Salmonella enterica* serovar Typhi strains isolated from asymptomatic human carriers. mBio **2**:e00060-11. doi:10.1128/mBio.00060-11.
17. **Meyer F, et al.** 2008. The metagenomic RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics **9**:386.
18. **Ministry of Health Malaysia.** 2008. Epidemiology of foodborne diseases in Malaysia, p 89. Director General Ministry of Health Malaysia technical report. Ministry of Health Malaysia, Putrajaya, Malaysia.
19. **Parkhill J, et al.** 2001. Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. Nature **413**:848–852.
20. **Shelburne SA, et al.** 2008. A direct link between carbohydrate utilization and virulence in the major human pathogen group A Streptococcus. Proc. Natl. Acad. Sci. U. S. A. **105**:1698–1703.
21. **Tatusov RL, Koonin EV, Lipman DJ.** 1997. A genomic perspective on protein families. Science **278**:631–637.
22. **Thong KL, Cheong YM, Puthucheary SD, Koh CL, Pang T.** 1994. Epidemiologic analysis of sporadic *Salmonella* Typhi isolates and those from outbreaks by pulsed-field gel electrophoresis. J. Clin. Microbiol. **32**: 1135–1141.
23. **Thong KL, Puthucheary SD, Pang T.** 1996. Genome size variation among recent human isolates of *Salmonella* Typhi. Res. Microbiol. **148**:229–235.
24. **Thong KL, et al.** 1995. Analysis of *Salmonella* Typhi from Southeast Asia by pulsed-field gel electrophoresis. J. Clin. Microbiol. **33**:1938–1941.
25. **Tiong V, et al.** 2010. Macrorestriction analysis and antimicrobial susceptibility profiling of *Salmonella enterica* at a university teaching hospital, Kuala Lumpur. Jpn. J. Infect. Dis. **63**:317–322.
26. **Zerbino DR, Birney E.** 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. **18**:821–829.

Downloaded from http://jb.asm.org/ on April 30, 2014 by UNIVERSITY OF MALAYA

Journal of Bacteriology

Whole-Genome Sequences and Comparative Genomics of Salmonella enterica Serovar Typhi Isolates from Patients with Fatal and Nonfatal Typhoid Fever in Papua New Guinea

Ramani Baddam, Kwai-Lin Thong, Tiruvayipati Suma Avasthi, Sabiha Shaik, Kien-Pong Yap, Cindy Shuan Ju Teh, Lay-Ching Chai, Narender Kumar and Niyaz Ahmed

J. Bacteriol. **2012, 194(18):5122. DOI: 10.1128/JB.01051-12.**

Updated information and services can be found at:
http://jb.asm.org/content/194/18/5122

These include:

REFERENCES This article cites 23 articles, 15 of which can be accessed free at: http://jb.asm.org/content/194/18/5122#ref-list-1

CONTENT ALERTS Receive: RSS Feeds, eTOCs, free email alerts (when new articles cite this article), more»

Information about commercial reprint orders: **http://journals.asm.org/site/misc/reprints.xhtml**
To subscribe to to another ASM Journal go to: **http://journals.asm.org/site/subscriptions/**

Journals.ASM.org

Downloaded from http://jb.asm.org/ on April 30, 2014 by UNIVERSITY OF MALAYA

JB Journals.ASM.org

GENOME ANNOUNCEMENT

Whole-Genome Sequences and Comparative Genomics of *Salmonella enterica* Serovar Typhi Isolates from Patients with Fatal and Nonfatal Typhoid Fever in Papua New Guinea

Ramani Baddam,[a] Kwai-Lin Thong,[b,c] Tiruvayipati Suma Avasthi,[a,b] Sabiha Shaik,[a] Kien-Pong Yap,[b,c] Cindy Shuan Ju Teh,[c] Lay-Ching Chai,[b,c] Narender Kumar,[a] and Niyaz Ahmed[a,b,d]

Pathogen Biology Laboratory, Department of Biotechnology, School of Life Sciences (Centre for Advanced Studies—UGC-SAP-CAS-I), University of Hyderabad, Gachibowli, Hyderabad, India[a]; Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia[b]; Laboratory of Biomedical Science and Molecular Microbiology, UMBIO Research Cluster, University of Malaya, Kuala Lumpur, Malaysia[c]; and Institute of Life Sciences, University of Hyderabad Campus, Gachibowli, Hyderabad, India[d]

Many of the developing countries of the Southeast Asian region are significantly affected by endemic typhoid fever, possibly as a result of marginal living standards. It is an important public health problem in countries such as Papua New Guinea, which is geographically close to some of the foci of endemicity in Asia. The severity of the disease varies in different regions, and this may be attributable to genetic diversity among the native strains. Genome sequence data on strains from different countries are needed to clearly understand their genetic makeup and virulence potential. We describe the genomes of two *Salmonella* Typhi isolates from patients with fatal and nonfatal cases of typhoid fever in Papua New Guinea. We discuss in brief the underlying sequencing methodology, assembly, genome statistics, and important features of the two draft genomes, which form an essential step in our functional molecular infection epidemiology program centering on typhoid fever. The comparative genomics of these and other isolates would enable us to identify genetic rearrangements and mechanisms responsible for endemicity and the differential severity of pathogenic salmonellae in Papua New Guinea and elsewhere.

Typhoid fever is a major pestilence in the developing world (20), and its prevalence is significant (1,000 cases per 100,000 individuals per year) in Papua New Guinea, which is geographically close to Southeast Asia (15). DNA profiling, which was used previously (6, 14, 20, 21, 22), could not fully explore genome diversity, as *Salmonella enterica* serovar Typhi isolates from Papua New Guinea showed limited heterogeneity, perhaps because of recent clonal expansion from a single endemic/ancestral strain (the disease was rarely seen before 1985) (17, 20). Minimal selection pressure and confinement to a specific geographical region might explain this limited genetic diversity (21, 22) despite horizontal gene transfer (13).

We hypothesized that the genome sequences of *Salmonella* Typhi isolates from patients with typhoid fever due a fatal strain (UJ308A) or a nonfatal strain (UJ816A) would provide significant insights into the association among disease phenotypes and strain characteristics. Two such strains were isolated from blood samples from patients and were found to be sensitive to common antibiotics. Strain UJ308A (phage type VS1) was obtained from a patient who died of typhoid, while UJ816A (phage type DI) was from a patient who recovered.

The 73-bp paired-end sequence data (insert size, 300 bp) were determined with an Illumina Genome Analyzer (GA2x, pipeline version l.6). About 95× and 105× coverage was achieved for strains UJ308A and UJ816A, respectively, comprising 1.9 and 2.0 Gb of data, respectively. *De novo* assembly was done as described previously (1, 2, 4, 8, 19); initial assembly generated 416 and 335 contigs for UJ308A and UJ816A, respectively, using Velvet (23) with a hash length of 39. The scaffolds were generated from contigs by using SSPACE (5) and further assembled and curated to give a consensus draft. The following statistics were gleaned upon analysis at RAST (3). The sizes of the chromosomes for UJ308A and UJ816A were approximately 4,724,875 and 4,736,723 bp, respectively, with a G+C contents of 51.89 and 51.94%, respectively. The coding percentage for both strains was ~86.8%; UJ308A and UJ816A contained approximately 4,720 and 4,710 protein coding sequences with average lengths of 869 and 873 bp, respectively. The data were further validated by Glimmer (7) and EasyGene (12). RNAmmer (11) revealed that the genome of UJ308A has 78 tRNA and 21 rRNA genes and the genome of UJ816A contains 77 tRNA and 22 rRNA genes. All of the major virulence markers encoded by pathogenicity islands and the genes relevant to the assembly of a type III secretion system (16) were identified in both the strains. The Vi antigen (10, 18), which plays major role in immune evasion, was present in both strains, as in *Salmonella* Typhi CT18 (16). The homologues of *Campylobacter* toxin *cdtB* and *Bordetella* pertussis toxin (9) were also present.

In view of this, further efforts are needed to determine the true extent of strain diversity in terms of (i) gene gains and losses over an evolutionary time scale, (ii) geographic gene flow, (iii) core versus accessory genome dynamics, (iv) virulence acquisition and attenuation, and (v) the preponderance of highly virulent versus "docile" strains across the regions of Asia where typhoid fever is endemic.

Nucleotide sequence accession numbers. The GenBank accession numbers for the genomes reported here are AJTD00000000 (UJ308A) and AJTE00000000 (UJ816A).

Received 12 June 2012 Accepted 29 June 2012
Address correspondence to Niyaz Ahmed, niyazSL@uohyd.ernet.in.
Copyright © 2012, American Society for Microbiology. All Rights Reserved.
doi:10.1128/JB.01051-12

Downloaded from http://jb.asm.org/ on April 30, 2014 by UNIVERSITY OF MALAYA

ACKNOWLEDGMENTS

We thankfully acknowledge support received from the University of Malaya High Impact Research Grant (UM.C/625/1HIR/MOHE/02 [A000002-5000 1]) Molecular Genetics. This genome program was completed under the wider umbrella of the Indo-German International Research Training Group, Internationales Graduiertenkolleg (GRK1673), Functional Molecular Infection Epidemiology, an initiative of the German Research Foundation (DFG) and the University of Hyderabad (India), of which N.A. is a speaker. N.A. is an Adjunct Professor of Molecular Biosciences at the University of Malaya, Kuala Lumpur, Malaysia, and an Adjunct Professor of Chemical Biology at the Institute of Life Sciences, Hyderabad, India. We are also grateful to M/s Genotypic Technology Pvt. Ltd., Bengaluru, India, for their untiring efforts with Illumina sequencing. We acknowledge the Bioinformatics Facility (BIF) at the Department of Biotechnology, University of Hyderabad, for the use of its computational infrastructure. We are thankful to Akash Ranjan for enabling access to the high-speed computing infrastructure at CDFD, Hyderabad.

All of the members of the Ahmed and Thong labs are gratefully acknowledged for their help and support.

REFERENCES

1. **Avasthi TS, et al.** 2011. Genomes of two chronological isolates (*Helicobacter pylori* 2017 and 2018) of the West African *Helicobacter pylori* strain 908 obtained from a single patient. J. Bacteriol. **193**:3385–3386.
2. **Avasthi TS, et al.** 2011. Genome of multidrug-resistant uropathogenic *Escherichia coli* strain NA114 from India. J. Bacteriol. **193**:4272–4273.
3. **Aziz RK, et al.** 2008. The RAST server: rapid annotations using subsystems technology. BMC Genomics **9**:75. doi:10.1186/1471-2164-9-75.
4. **Baddam R, et al.** 2012. Genetic fine structure of a *Salmonella enterica* serovar Typhi strain associated with the 2005 outbreak of typhoid fever in Kelantan, Malaysia. J. Bacteriol. **194**:3565–3566.
5. **Boetzer M, et al.** 2011. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics **27**:578–579.
6. **Combs BG, et al.** 2005. Ribotyping of *Salmonella enterica* serovar Typhi isolates from Papua New Guinea over the period 1977 to 1996. P. N. G. Med. J. **48**:158–167.
7. **Delcher AL, et al.** 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res. **27**:4636–4641.
8. **Devi SH, et al.** 2010. Genome of *Helicobacter pylori* strain 908. J. Bacteriol. **192**:6488–6489.
9. **Haghjoo E, Galan JE.** 2004. Salmonella typhi encodes a functional cytolethal distending toxin that is delivered into host cells by a bacterial-internalization pathway. Proc. Natl. Acad. Sci. U. S. A. **101**:4614–4619.
10. **Hirose K, et al.** 1997. Survival of Vi-capsulated and Vi-deleted *Salmonella typhi* strains in cultured macrophage expressing different levels of CD14 antigen. FEMS Microbiol. Lett. **147**:259–265.
11. **Lagesen K, et al.** 2007. RNammer: consistent annotation of rRNA genes in genomic sequences. Nucleic Acids Res. **35**:3100–3108.
12. **Larsen TS, Krogh A.** 2003. EasyGene—a prokaryotic gene finder that ranks ORFs by statistical significance. BMC Bioinformatics **4**:21. doi: 10.1186/1471-2105-4-21.
13. **Liu SL, Sanderson KE.** 1995. Rearrangements in the genome of the bacterium Salmonella typhi. Proc. Natl. Acad. Sci. U. S. A. **92**:1018–1022.
14. **Nair S, Schreiber E, Thong KL, Pang T, Altwegg M.** 2000. Genotypic characterization of *Salmonella typhi* by amplified fragment length polymorphism fingerprinting provides increased discrimination as compared to pulsed-field gel electrophoresis and ribotyping. J. Microbiol. Methods **41**:35–43.
15. **Pang T, Bhutta ZA, Finlay BB, Altwegg M.** 1995. Typhoid fever and other salmonelloses: a continuing challenge. Trends Microbiol. **3**:253–255.
16. **Parkhill J, et al.** 2001. Complete genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi CT18. Nature **413**:848–852.
17. **Passey M, et al.** 1995. Highly endemic typhoid fever in Papua New Guinea. Southeast Asian J. Trop. Med. Public Health **26**:83–84.
18. **Raffatellu M, et al.** 2006. Capsule-mediated immune evasion: a new hypothesis explaining aspects of typhoid fever pathogenesis. Infect. Immun. **74**:19–27.
19. **Siddavattam D, et al.** 2011. Genome of a novel isolate of *Paracoccus* denitrificans capable of degrading *N*,*N*-dimethylformamide. J. Bacteriol. **193**:5598–5599.
20. **Thong KL, et al.** 1995. Analysis of *Salmonella typhi* isolates from Southeast Asia by pulsed-field gel electrophoresis. J. Clin. Microbiol. **33**:1938–1941.
21. **Thong KL, et al.** 1996. Molecular analysis of isolates of *Salmonella typhi* obtained from patients with fatal and nonfatal typhoid fever. J. Clin. Microbiol. **34**:1029–1033.
22. **Thong KL, et al.** 2002. Increasing genetic diversity of *Salmonella enterica* serovar Typhi isolates from Papua New Guinea over the period from 1992 to 1999. J. Clin. Microbiol. **40**:4156–4160.
23. **Zerbino DR, Birney E.** 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. **18**:821–829.

Downloaded from http://jb.asm.org/ on April 30, 2014 by UNIVERSITY OF MALAYA

OPEN ACCESS Freely available online

PLoS one

Crowd Sourcing a New Paradigm for Interactome Driven Drug Target Identification in *Mycobacterium tuberculosis*

Rohit Vashisht[1,3☯], Anupam Kumar Mondal[2☯], Akanksha Jain[1☯], Anup Shah[2☯], Priti Vishnoi[2☯], Priyanka Priyadarshini[1☯], Kausik Bhattacharyya[2☯], Harsha Rohira[4], Ashwini G. Bhat[3], Anurag Passi[1], Keya Mukherjee[2], Kumari Sonal Choudhary[2], Vikas Kumar[5], Anshula Arora[4], Prabhakaran Munusamy[6], Ahalyaa Subramanian[7], Aparna Venkatachalam[7], Gayathri S[8], Sweety Raj[4], Vijaya Chitra[9], Kaveri Verma[10], Salman Zaheer[11], Balaganesh J[12], Malarvizhi Gurusamy[13], Mohammed Razeeth[13], Ilamathi Raja[13], Madhumohan Thandapani[13], Vishal Mevada[14], Raviraj Soni[14], Shruti Rana[14], Girish Muthagadhalli Ramanna[15], Swetha Raghavan[15], Sunil N. Subramanya[15], Trupti Kholia[16], Rajesh Patel[17], Varsha Bhavnani[18], Lakavath Chiranjeevi[19], Soumi Sengupta[20], Pankaj Kumar Singh[21], Naresh Atray[22], Swati Gandhi[23], Tiruvayipati Suma Avasthi[24,29], Shefin Nisthar[25], Meenakshi Anurag[2], Pratibha Sharma[26], Yasha Hasija[27], Debasis Dash[2], Arun Sharma[28], Vinod Scaria[2], Zakir Thomas[1], OSDD Consortium[1], Nagasuma Chandra[3]*, Samir K. Brahmachari[1,2*¶], Anshu Bhardwaj[1]*

1 Council of Scientific and Industrial Research (CSIR), Delhi, India, **2** Institute of Genomics and Integrative Biology, CSIR, Delhi, India, **3** Department of Biochemistry, Indian Institute of Science, Bangalore, Karnataka, India, **4** Acharya Narendra Dev College, University of Delhi, India, **5** Goethe University, Frankfurt, Germany, **6** PSG College of Technology, Peelamedu, Coimbatore, Tamil Nadu, India, **7** SASTRA University, Tirumalaisamudram, Thanjavur, Tamilnadu, India, **8** SDM College, Ujire, Karnataka, India, **9** Sree Narayan Guru College, Coimbatore, Tamil Nadu, India, **10** Maharshi Dayanand University, Rohtak, Haryana, India, **11** Amity Institute of Biotechnology, Amity University, Lucknow, Uttar Pradesh, India, **12** Bharathiar University, Coimbatore, Tamil Nadu, India, **13** Bharathidasan University, Palkalaiperur, Tiruchirappall, Tamil Nadu, India, **14** Bitvirtual patan Node, Hem. North Gujarat University, Patan, Gujarat, India, **15** Business Intelligence Technologies India Pvt Ltd., Bangalore, Karnataka, India, **16** Christ College, Vidya Niketan, Saurashtra University, Rajkot, Gujarat, India, **17** Department of Life Sciences, Hemchandracharya North Gujarat University, Patan, Gujarat, India, **18** Department of Biotechnology, University of Pune, Maharashtra State, India, **19** Indian Institute of Toxicology Research, CSIR, Lucknow, Uttar Pradesh, India, **20** Indian Statistical Institute, Kolkata, West Bengal, India, **21** Maulana Azad National Institute of Technology, Bhopal, Madhya Pradesh, India, **22** Shri Ram College of Pharmacy, Karnal, Haryana, India, **23** The Maharaj Sayajirao University of Baroda, Gujarat, India, **24** Pathogen Biology Laboratory, Department of Biotechnology, School of Life Sciences, University of Hyderabad, Hyderabad, Andhra Pradesh, India, **25** University of Kerala, Thiruvananthapuram, Kerala, India, **26** All India Institute of Medical Sciences, New Delhi, India, **27** Department of Biotechnology, Delhi Technological University, Shahbad Daulatpur, Delhi, India, **28** Bioinformatics Centre, Institute of Microbial Technology, CSIR, Chandigarh, India, **29** Faculty of Science, Institute of Biological Sciences, University of Malaya, Kuala Lumpur, Malaysia

Abstract

A decade since the availability of *Mycobacterium tuberculosis* (Mtb) genome sequence, no promising drug has seen the light of the day. This not only indicates the challenges in discovering new drugs but also suggests a gap in our current understanding of Mtb biology. We attempt to bridge this gap by carrying out extensive re-annotation and constructing a systems level protein interaction map of Mtb with an objective of finding novel drug target candidates. Towards this, we synergized crowd sourcing and social networking methods through an initiative 'Connect to Decode' (C2D) to generate the first and largest manually curated interactome of Mtb termed 'interactome pathway' (IPW), encompassing a total of 1434 proteins connected through 2575 functional relationships. Interactions leading to gene regulation, signal transduction, metabolism, structural complex formation have been catalogued. In the process, we have functionally annotated 87% of the Mtb genome in context of gene products. We further combine IPW with STRING based network to report central proteins, which may be assessed as potential drug targets for development of drugs with least possible side effects. The fact that five of the 17 predicted drug targets are already experimentally validated either genetically or biochemically lends credence to our unique approach.

Citation: Vashisht R, Mondal AK, Jain A, Shah A, Vishnoi P, et al. (2012) Crowd Sourcing a New Paradigm for Interactome Driven Drug Target Identification in *Mycobacterium tuberculosis*. PLoS ONE 7(7): e39808. doi:10.1371/journal.pone.0039808

Editor: Manfred Jung, Albert-Ludwigs-University, Germany

Received November 4, 2011; **Accepted** May 30, 2012; **Published** July 11, 2012

Copyright: © 2012 Vashisht et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by the Council of Scientific and Industrial Research, India, Funding (Grant No. HCP0001). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: Authors Dr. Ramanna, Dr. Raghavan and Dr. Subramanya are employed by Business Intelligence Technologies Pvt Ltd. This does not alter the authors' adherence to all the PLoS ONE policies on sharing data and materials.

* E-mail: skb@igib.res.in (SKB); nchandra@biochem.iisc.ernet.in (NC); anshu@csir.res.in (AB)

☯ These authors contributed equally to this work.

¶Lead author for the OSDD Consortium.

Introduction

Proclaimed a global health emergency by the World Health Organization (WHO) in 1993, Tuberculosis (TB) still remains the leading cause of mortality and affects approximately 32% of the world population [1]. The emergence of multi-drug-resistant strains of *Mycobacterium tuberculosis*, the causative agent of TB, and the vulnerability of the patients infected with HIV to tuberculosis have not only fuelled the spread of the disease but also present a challenging task of understanding the disease physiology and discovering new drug targets. In this quest, Mtb was sequenced and annotated in 1998 [2]. A subsequent re-annotation in 2002 successfully assigned functions to almost half of the approximately 4000 genes [3]. More recently, 20 more ORFs have been added to this list and the annotations updated [4,5]. However a huge gap in information exists between published literature and the genome databases. The existing annotations in these databases are thus insufficient to generate the protein interaction map or the interactome, pivotal to understanding Mtb biology and identification of novel drug targets. To this end, Open Source Drug Discovery (OSDD) project (www.osdd.net) [6,7] launched the Connect to Decode (C2D) program (http://c2d.osdd.net), an innovative blend of crowd sourcing and social networking in a virtual cloud space for a comprehensive collaborative re-annotation of Mtb which is the primer for generating the interactome. The ultimate objective is to identify drug targets based on better understanding of the complex interactions of various biological macromolecules in the pathogen.

Systems biology-based approaches have been applied to obtain better insights into the pathogen biology [8]. This strategy may help in identifying more than one potential drug targets and these can be utilized as sets of targets for a polypharmacology approach. A promising candidate in this category is bi-substrate acyl-sulfamoyl analogues that simultaneously disrupt crucial nodes in biosynthetic network of virulent lipid with dramatic effect on the cell surface architecture of Mtb [9]. Also, a recent study on genome-wide siRNA experiment has identified host factors that regulate Mtb load in human macrophages and are crucial to understand the dynamic interplay of molecular components of the pathogen and the host [10]. There are many such studies that try to capture the snapshots of the molecular interactions in Mtb in different conditions. It is therefore imperative to capture and curate data on experimentally validated interactions lying scattered in diverse sources in the literature to generate a genome scale network. This was achieved through the C2D program. The C2D community started with initial registration of more than 800 researchers, which largely consisted of research scholars, graduate students and under-graduate students. The participants were trained, evaluated and filtered at various stages of online training and assignments (https://sites.google.com/a/osdd.net/c2d-01/pathwayannotationproject/results-of-the-exercise). More than 100 researchers were selected as curators to obtain the final annotations (https://sites.google.com/a/osdd.net/c2d-01/pathwayannotationproject).

Here we describe how C2D has implemented a community annotation approach in a distributed co-creation mode for mining literature and how the accuracies and scope of assigning functions were enhanced using combined evidence approach. We have enriched the annotations of the Mtb genome both in terms of coverage and details (**Table 1**). Web2.0 collaborative online tools enabled voluntary community participation for implementing this task. An important part of the project was creating self-organized communities to collectively learn and share the process and the standards for reporting annotations. As per published estimates, this innovative approach packed nearly 300 man-years into 4 months [11] and it has also established a novel way of collective problem solving on a voluntary basis in a sustainable manner [12]. This is, to the best of our knowledge, lead to the creation of the largest manually curated interactome of Mtb. Based on the varied nature of interactions among proteins in vivo, we propose a new network definition called "Protein-Protein Functional Network" (PPFN). This network encompasses a total of 1434 proteins connected through 2575 functional relationships. In this paper, we detail how the Interactome - PathWay (IPW), an open collaborative platform was used to generate and analyze potential drug targets. Using betweenness centrality [13] as a first indicator to shortlist candidate drug targets, we zeroed into 73 proteins. We have in the process also created a sustainable open innovation platform.

Results and Discussion

C2D Annotation

An overview of the approach followed in *'Connect to Decode'* (C2D) exercise is as illustrated in **Figure 1**. Broadly the approach was designed based on the principles of the fourth paradigm of science, encompassing data collation, curation and analysis [14]. Roughly ~4.4 Mbp genome of Mtb was re-annotated manually. To streamline the annotation process and select a community of researchers competent to implement this project, a series of online assignments and training modules were assigned (see methods). These steps ensured the selection of serious and dedicated contributors thereby assuring the quality of data collation, curation and analysis. Various standard operating protocols (SOPs) were designed and shared with the participants for the consistency in the steps followed for the annotation of genes (https://sites.google.com/a/osdd.net/c2d-01/pathwayannotationproject/instructions-for-annotation and https://sites.google.com/a/osdd.net/c2d-01/pathwayannotationproject/example-annotation and https://sites.google.com/a/osdd.net/c2d-01/pathwayannotationproject/stepsforproteinannotation). Given the exponential increase in the number of publications from about 300 per year since 1990's to a staggering 2000 per year in 2010, the challenging task of collating and curating data was achieved through the formulation of community editable interactive platform designed to facilitate real time annotations and continuous updates. The community scanned and retrieved information from nearly 10,000 published studies in addition to extracting information from databases and transferred annotations using sequence and structure analyses based approaches. The community has cited more than 3000 papers in annotation process as on an average 3–4 manuscripts were referred or read in order to get the relevant information to annotate a given protein.

The Mtb Genome Annotation and Interactome Curation

IPW has resulted in annotation of 87% of the genome in the context of reporting gene products as compared to 52% in the re-annotation reported in 2002. Moreover, less than 5% of the interactions in IPW (Table S1) exist in other manually curated interaction databases such as BIND [15], APID [16], IntAct [17], DIP [18] and MINT [19] (**Figure 2(b)**). Thus, to the best of our knowledge, *Connect to Decode's* Interactome Pathway Annotation (IPW) has generated the largest data set of manually curated interactions in Mtb. These interactions not only include data from large interaction databases such as IntAct, BIND, MINT, APID,

Table 1. The data structure that was used to capture the interactome data.

Field	Description
GeneID	The unique identifier for a given gene (Rv ID and NCBI Gene ID)
Gene Name	Assigned name of the gene
Pathway	Biological role of the gene
Gene function	The biological function of the gene
Interacting Partners	All the interacting partners for a given gene
Type of Interaction	Type of interaction (protein-protein [p-p], protein-nucleotide [p-n])
Nature of Interaction	This field contains nature of interaction, such as structural complex, regulatory, signaling etc.
Method of Inferring Interaction	Contains information about the experimental or computational methods used for the inference of interacting partners
Type of Evidence	Type of evidence, adopted from Gene Ontology (IDA, IPI, ISO, TAS, etc)
PUBMED/Link of source	PubMed ID or any web based link from where the interaction and other annotation details were inferred
Email of author	E-mail address of curator

There were 11 annotation fields for reporting annotations. The data is available in PSI MITAB format.
doi:10.1371/journal.pone.0039808.t001

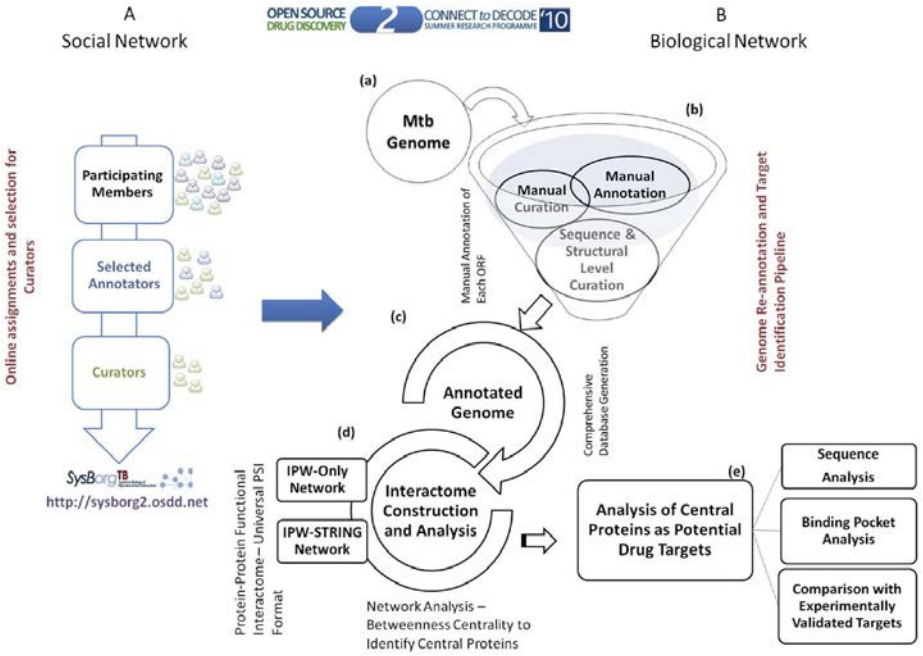


Figure 1. From Social Network to Biological Network. The C2D annotation approach for manual annotation and curation of Mtb interactome followed by network analysis to predict potential drug targets reported at various sequence and structural level filters. (A) Illustrates the overall approach of crowd sourcing through social network implemented in C2D exercise (B)(a) Mtb Genome (b) Manual collation and sequence/structure based curation for gene annotation (c) Collation of re-annotated genome into comprehensive data structure (d) Construction of protein-protein interaction network based on the annotated data (e) Target identification using network analysis; Sequence level comparison of selected proteins with that of human homologs, human gut flora and human oral flora; systems, sequence and structure level analysis of shortlisted proteins and experimentally validated drug targets.
doi:10.1371/journal.pone.0039808.g001

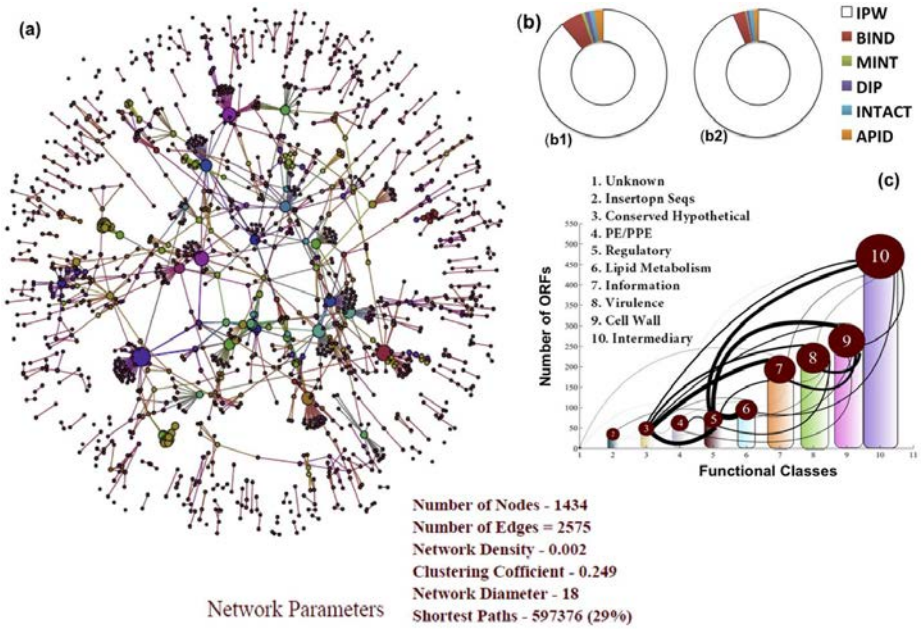


Figure 2. IPW interactome and comparison with existing annotation databases (a) IPW-Only protein-protein functional interaction network, (b) Comparative analysis of IPW-Only proteins and interaction with existing manually curated databases, Ring represents all interactions and proteins in IPW displaying the subsets which are obtained from other manually curated databases (b1) Comparative analysis of IPW-Only interactions to that of existing manually curated databases (b2) comparative analysis of protein as curated in IPW-Only to that of proteins presents in other manually curated databases (c) TubercuList functional class interaction relation based on the interactions as obtained from IPW-Only. The connectivity (lines) represents the interacting proteins within these classes.
doi:10.1371/journal.pone.0039808.g002

DIP, etc but also include a large amount of manually curated information from literature.

Of the 1193 hypothetical proteins from TubercuList [4], the IPW based annotations identify gene products for 770 proteins. Of the 1480 hypothetical proteins reported in KEGG [20] database, functional associations have been made to 1055 proteins, clearly showing how IPW has bridged the wide gap that existed between information captured in databases and that available in literature. To ensure that IPW remains up to date, the data from IPW is shared with members of the OSDD community in an 'edit' mode, through which new interactions can be added using the SOP that includes a rigorous quality check phase, specifically designed for community contribution.

Interactome Construction: IPW and Combined Network with STRING

Interactome as a whole constitutes various biological interactions belonging to both structural and functional type of protein-protein associations. To have an encyclopedic view of various interactions that take place at protein functional level, we report the construction of two types of networks. The first network, termed IPW only (**Figure 2(a)**), was constructed on the basis of the IPW curated data alone. The nodes in the network represent the proteins whereas the edges represent the functional interactions among those proteins. The nodes were scaled and color coded in proportion to their degrees. Also, based on the common interactions we derived a connectivity relationship between various TubercuList functional classes [4]. **Figure 2(c)** shows the connectivity among 10 broad functional classes of TubercuList. The edge thickness was taken to be directly proportional to the number of common proteins between the two TubercuList functional classes for the given pair. Significant functional dependencies are seen among the 'Lipid Metabolism, Cell Wall, Intermediary metabolism and Regulatory systems' functional classes, reflected in their edge thicknesses in the network. Disruption of such linkages can lead to breakdown of crosstalk between these biological processes and thus could be exploited to identify new drug targets.

Secondly, in order to obtain insights on the complete functional organization among all the possible proteins of Mtb, a combined network termed, IPW-STRING (IPWSI), was constructed by overlaying STRING network on the IPW network. The STRING based network of Mtb was derived from STRING 8.0 [21] database consisting of various interactions among proteins as derived on the basis of extensive computational and limited experimentally inferred interactions. Computational predictions

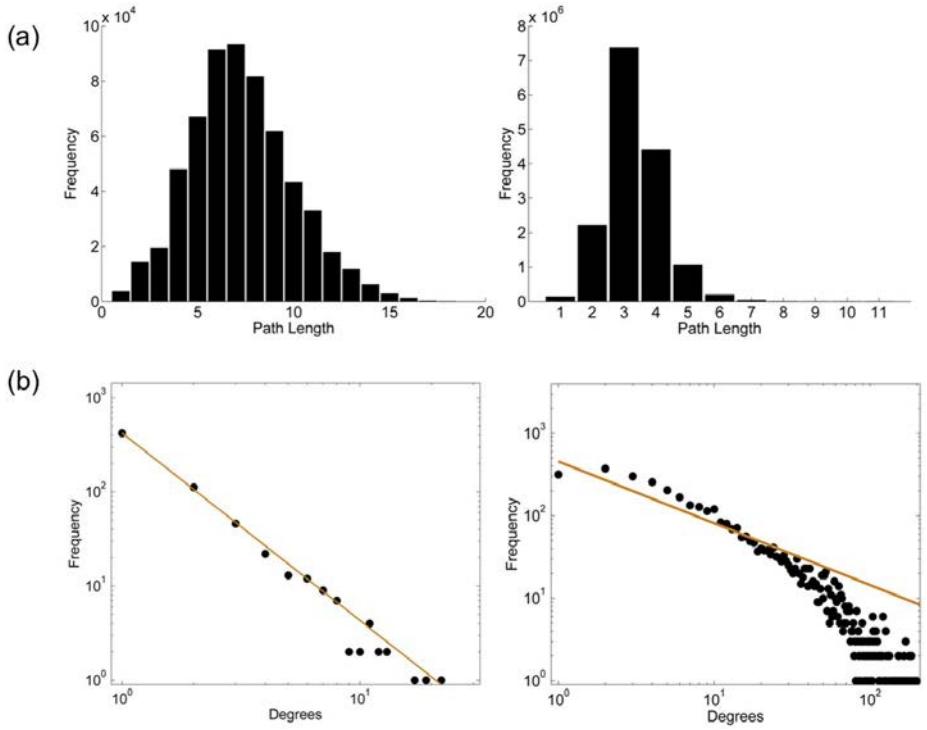


Figure 3. Network parameters (a) Characteristic path length of IPW-Only network and IPWSI network. In both the graphs the x-axis represents the path length whereas the y-axis represents the frequency. 3(b) Log-Log plot of degree distribution of IPW network, the solid line was obtained by fitting the power law for $\gamma = 1.99$ and Log-Log plot of degree distribution of IPWSI network, the solid line represents the power law fit with $\gamma = 2.01$.
doi:10.1371/journal.pone.0039808.g003

have been based on established methods such as phylogenetic profiling, domain fusion, common gene neighborhood and operon criteria. However, computational models over predicts interactions since they do not account for spatio-temporal separation of the interacting partners. Thus, in the combined network the accuracy of interaction decreases whereas the coverage increases. It should also be noted that there is an inherent bias for well-studied proteins in IPW. A simple comparison shows that nearly 60% of IPW interactions have experimental evidence codes as compared to 2% existing in STRING. Also, about 29 additional proteins and 1762 new functional interactions apart from that reported by STRING were included in the new IPW-STRING combined interactome.

The combined IPW-STRING interactome was further used to decipher various possible drug targets using the concepts of graph theory. The network analysis of these networks provides a means to understand the functional organization of the organism from the network topology point of view [22,23]. Various network properties as computed for both the networks and their biological relevance are discussed below.

Topological Organization of Interactome

In order to understand the functional organization of constructed interactome we further assessed the fundamental properties of this network from the graph theoretic point of view. Given a vast interaction space encompassing the interactome as whole, where the nodes represents proteins and interaction represents a functional relation between them, it becomes imperative to understand the functional organization of the network from its topology. The most fundamental characteristic of a graph is the connectivity of its constituent nodes as represented by the degree. Degree, being a measure of interconnectedness of nodes highlight the importance of a node (protein in this case) with respected to other nodes in the network. A maximum degree of 44 and 289 was observed for the IPW and IPWSI networks, respectively, suggesting the level of maximum number of functional relation of a given protein in both the networks.

Clustering coefficient for a node indicates the connectivity of the neighbours of a given node to the other nodes in the network [24]. This parameter was computed to elucidate the dependencies of two or more proteins with respect to each other and to rest of the proteins in the network. The clustering coefficient for

IPW and IPWSI networks was observed to be 0.249 and 0.377, respectively. The high clustering coefficient of both the networks suggests the presence of well-connected hubs within the network, which are important from the functional crosstalk between the proteins of Mtb. Further, the characteristic path length of both the networks was computed in order to comprehend the extent of functional relation between any two given proteins in the network. The characteristic path length of both the networks is as shown in **Figure 3(a)**. The characteristic path length in IPW network was observed to be 7.2 whereas for IPWSI it was observed to be 3.13. From the network navigability point of view the characteristic path length can be inferred as the number of steps that one has to take traversing from one node to other, which from biological point of view could be inferred as the amount of communication that is possible between any two proteins. Pertaining to the high characteristic path length of IPW alone, the absence of functional relation between any two proteins can be inferred; however, the functional relation between any two proteins increase when the IPW alone was clubbed with STRING based network. The characteristic path length, thus, can be used to understand the functional gap that possibly exists in the protein-protein interaction network. Emphasizing on the network communication further, the network diameter was computed representing the length of the 'longest' shortest path in the network. The network diameter of IPW and IPWSI networks was observed to be 18 and 10, respectively. Akin to characteristic path length, the network diameter can be used to interpret the overall navigability of the network, the higher the diameter, the more distantly two nodes are related and *vice versa*.

As discussed, understanding the topological organization of the network could lead to better understanding of its underlying principles. The network topology could also be used to understand the number of possible modules (hubs) in the network, which may help in identifying potential drug targets. In order to obtain such insights, we tested the existence of power law distribution on IPW and IPWSI networks, respectively. The power law distribution can also be used to understand the scale free nature of a network [23]. There is extensive literature that reports the existence of scale free nature of biological networks. The power law distribution on the node degree distribution of IPW and IPWSI networks is shown in **Figure 3(b)**. The value of γ was observed to be 1.99 for IPW and 2.01 for IPW-STRING combined node degree distribution.

Target Identification

Apart from inferring fundamental principles of network properties the availability of an interactome also enables prediction of essential proteins from the network structure point of view. The protein lethality within a network is usually obtained from the degree distribution of the nodes in the networks. The nodes with high degree are considered important and hence regarded as probable drug targets. The degree distribution alone could lead to improper putative drug target identification as it does not capture the alternate routes in the network. Most of the biological networks possess large number of shortest paths [25]. The large number of shortest paths also suggests the availability of alternate routes within the network which could be used to achieve a certain biological objective. Removing such nodes from the network could lead to maximum disruption in the network. In order to capture these properties, important nodes as well as important edges, we used betweenness centrality [24,26] as a metric system to infer putative drug targets. The node betweenness centrality at a threshold of ≥ 0.2 lead to identification of 17 and 64 central proteins from IPW and IPWSI networks, respectively (Table S2).

Analysis of Putative Drug Targets: Identifying Probable Non-toxic Targets

To design a viable drug it is essential to ensure least probability of off-target interactions. A sequence, structure and systems based analysis was performed in order to predict the druggability of the shortlisted central proteins from the two networks so as to reduce the chances of off-target interactions.

The list of 17 and 64 proteins (73 unique proteins as eight are common in the two lists) was first filtered against human homologs and human oral and gut flora [27]. Of the 17 targets identified by IPW, none had a homolog in human proteome and in human oral and gut flora. In the combined network IPWSI, 53 such targets were identified (**Figure 4**). There are 62 unique central proteins without any significant homology to human proteome, gut and oral flora from IPW and IPWSI. We further analyzed this list of 62 proteins for absence of small peptides (octamer) since it has been reported that a small fraction of peptide sequences are evolutionarily conserved and invariant across several organisms [28]. These peptide sequences can adopt similar conformation in different protein structures [28]. A comparative analysis shows that one protein Rv3221A does not share any common octapeptide with human proteome, gut or oral flora. However, a closer and detailed analysis needs to be performed for proteins sharing octapeptide with human proteome and human microbiome in order to evaluate their status for off-target binding. In order to understand the binding pockets, an independent analysis has been performed to predict and match binding pockets of central proteins with human proteome. Of the 73 central proteins, 57 have either PDB or ModBase [29] structure making them amenable to structural analysis for druggability. We analyzed these proteins as per the targetTB [30] pipeline where the top 10 binding sites for each of these 57 proteins were identified using PocketDepth algorithm [31]. The binding pockets of these proteins were then compared with human proteome using PocketMatch [32]. Of the 57 proteins, 31 proteins have high structural similarity with human proteome at binding site level whereas 26 proteins which do not have binding site similarity with human proteome. It is interesting to note that seven of these are experimentally validated drug targets.

Rv3221A (RshA) (**Figure 4 List d**), an anti-sigma factor to the primary stress response sigma factor SigH, passed all filters but is neither reported as a potential drug target in literature nor in targetTB predictions. The gene encoding RshA lies in the same operon as SigH and is co-expressed with the same [33]. It has a strong affinity to bind with SigH and attenuates its ability to bind to the RNA polymerase holoenzyme under normal growth conditions. Under conditions of oxidative stress, phosphorylation of RshA by Rv0014c (PknB) abolishes its binding to SigH, which in turn leads to the cascade of expression of several stress response proteins [34] (**Figure 5**). SigH causes increased expression of two other sigma factors Rv2710 (SigB) and Rv1221 (SigE), which are also known to be stress related sigma factors that assist Mtb in its survival during several stress conditions and are also central proteins. The other interacting partners of RshA include heat shock proteins and chaperones like Rv0384c (ClpB) and Rv0350 (DnaK), enzymes for oxidative stress response Rv1471 (TrxB1), Rv3913 (TrxB2) and Rv3914 (TrxC) which are also part of the sigH regulon. sigH also induces enzymes involved in cysteine biosynthesis and in the metabolism of ribose and glucose for the production of mycothiol precursors, which assist in cellular protection under oxidative stress. The SigB and SigE signaling

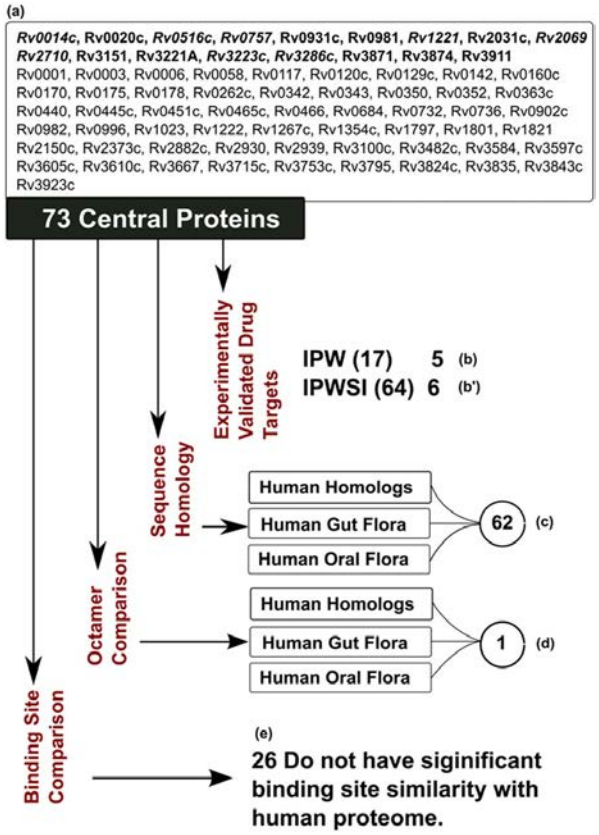


Figure 4. Illustrates the comprehensive analyses of central proteins as potential drug targets. The various filters include comparison with validated drug targets, sequence and structural level comparison with Human proteome, gut and oral flora (a) The list of 73 central ORFs wherein Rv Ids in bold represent IPW central ORFs, Rv IDs in regular font represents IPWSI central ORFs and the italicized-bold represent common Rv Ids to both IPW and IPWSI. (b & b') Five of the 17 IPW and six of 64 central ORFs with experimental validation as drug targets. (c) Sequence homology comparison with human proteome and human microbiome results in 62 ORFs with no significant similarity (d) Octamer analyses against human proteome and human microbiome results in one ORF with no hits (e) Comparative binding site analysis with human proteome results in 26 ORFs with no significant similarity (lists b, b', c, d and e available in Table S2).
doi:10.1371/journal.pone.0039808.g004

cascade downstream interacts and regulate other central proteins (**Figure 5**). RshA is at the beginning of this cascade and seems to play in crucial role in regulating the stress response proteins, starting with sigH.

The objective of the interactome construction and analyses was to identify central proteins, which have significant roles in maintaining growth and survival of the bacterial pathogen. We identified 17 such central proteins (**Table 2**), five of which (*PknB*, *NuoG*, *PhoP*, *EccCb1*, *HspX*) have been previously functionally characterized and shown to be essential by gene deletion and mutation and thus are considered as validated drug targets. The target gets further validated if there are inhibitors which inhibit the function of the target enzyme or protein as well. *PknB* (Rv0014c) is an essential serine-threonine protein kinase in Mtb and has role in a number of signaling pathways in cell division and metabolism. Several inhibitors have been reported for this kinase [35] and is also one of the targets being pursued by Working Group on New TB Drugs (http://www.newtbdrugs.org/project.php?id = 81). *NuoG* (Rv3151) is a subunit of type I NADH dehydrogenase playing an important role in growth in macrophage and pathogenesis in animal models [36]. *PhoP* (Rv0757), a response regulator and part of the two component system, when mutated leads to growth defects in macrophages and in mouse models [37]. *eccCb1* (Rv3871) is a part of the RD1

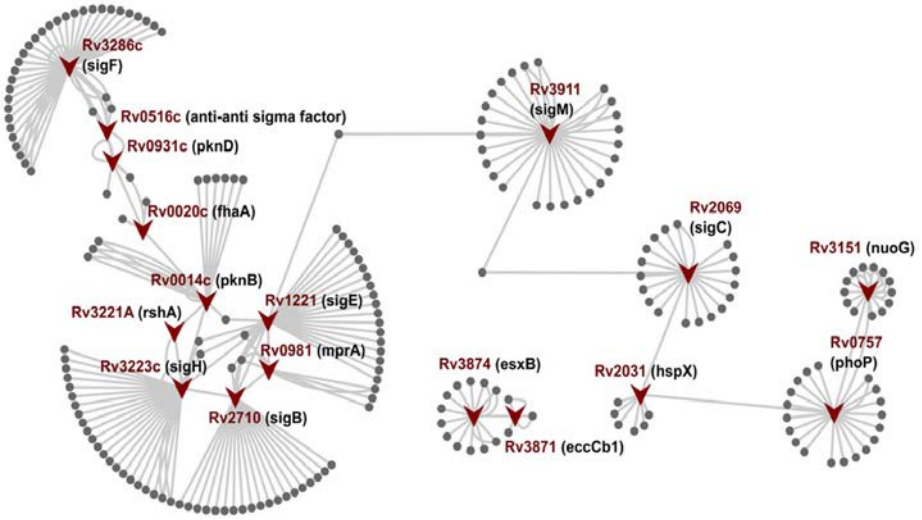


Figure 5. Illustration of 17 putative drug target interaction from IPW interactome depicting the cascade of how the central proteins interact with each other in a spatio-temporal manner under different conditions like growth, stress and survival in macrophages including virulence. Under normal conditions, PknB phosphorylates RshA which inhibits SigH. However, under oxidative stress, RshA is not phosphorylated and this abolishes its binding to SigH, rendering it free. SigH in turn upregulates expression of SigE and SigB which regulates MprA (bacterial persistence regulator). MprA also regulates SigB and SigE. SigB plays important role in adaptation to stationary phase and nutritionally poor conditions and SigE is upregulated in mycobacterial growth within human macrophages and is transcribed from three different promoters under different conditions. sigB is also regulated by SigF, which regulates the expression of genes involved in the biosynthesis and structure of the mycobacterial cell envelope, including complex polysaccharides and lipids, particularly virulence- related sulfolipids and several transcription factors. Rv0516c is an anti-anti sigma factor and regulates anti-sigma factor SigF (upregulated during infection culture of human macrophages and in nutrient starvation condition; regulates transcription of genes involved in cell wall biosynthesis, sulfolipid metabolism, nucleotide metabolism, energy metabolism and several transcription factors) on getting phosphorylated by PknD which in turn is regulated by Rv0020c phosphorylated by PknB and PknE. SigF also regulates sigC and regulates hspx that is also regulated by dosR regulon. dosR regulon in turn is again regulated by PhoP which is a transcription factor for nuoG, eccCb1, esxb/cfp10 [48].
doi:10.1371/journal.pone.0039808.g005

region and mutation leads to attenuated growth and toxicity in THP-1 cells. The mutants cannot export CFP-10 and are avirulent [38]. *hspX* (Rv2031c) encodes for a alpha-crystallin-like protein and plays a significant role in retaining a non-replicating state in latency [39,40]. The fact that five of the 17 putative drug targets from IPW are already validated drug targets, lends credence to our approach of annotating the genome and interactome construction of Mtb for system level understanding towards novel drug target identification.

Despite the efforts over a number of years, discovering novel, fast acting drugs for TB has been a major challenge. However, recently introduced combination drug Risorine designed on the principles of Ayurveda has been shown to cut down rifampicin use leading to very high compliance [41]. Understanding the biology of the pathogen through a systems level approach can help in identifying the Achilles heel for Mtb. Towards this, Interactome Pathway annotation has captured the updated relevant information on Mtb genes and has tried to unravel the puzzle. We have amalgamated crowd sourcing with social networking to comprehensively reannotate the Mtb genome, generated its largest ever interactome and propose potentially efficacious drug targets. In the process, we have set up an open collaborative platform and a dynamic community to ensure regular updates.

Materials and Methods

Crowd Sourcing for Data Curation

Data capture. Two annotation standard operating protocols (SOPs), in the presence of literature and in the absence of literature, were designed in order to capture the maximum amount of relevant data. Wherever the protein was not studied in Mtb, the annotations were transferred from other organisms based on conservative statistical measures in sequence and structure-based analysis as discussed below (i and ii). To ensure consistency and integrity of the data added to the resource, Standard Operating Protocols (SOPs) were created and followed by the community. These SOPs and tutorials may be accessed at (http://c2d.osdd.net and https://sites.google.com/a/osdd.net/c2d-01/pathwayannotationproject).

Annotation SOP in presence of literature. The first step was to retrieve information on Mtb proteins with experimental evidence from literature. PubMed and Google based literature searches were carried out using suitable keywords, such as the respective Rv number, gene name, *Mycobacterium tuberculosis*, along with suitable Boolean expressions, such as AND and OR (for example, [Rv1018c] AND [mycobacterium tuberculosis], [epoxide hydrolase] AND [mycobacterium tuberculosis]). While

Table 2. The list of all the 17 central proteins as predicted from the betweenness centrality of >0.2 through IPW network with their gene products.

Accession	Gene Name	Description (Gene Product)
Rv0014c*[35]	pknB	TRANSMEMBRANE SERINE/THREONINE-PROTEIN KINASE B PKNB (PROTEIN KINASE B)
Rv0020c	fhaA	CONSERVED HYPOTHETICAL PROTEIN WITH FHA DOMAIN, TB39.8
Rv0516c	Rv0516c	ANTI-ANTI SIGMA FACTOR
Rv0757*[37]	phoP	POSSIBLE TWO COMPONENT SYSTEM RESPONSE TRANSCRIPTIONAL POSITIVE REGULATOR
Rv0931c	pknD	TRANSMEMBRANE SERINE/THREONINE-PROTEIN KINASE D PKND (PROTEIN KINASE D)
Rv0981	mprA	MYCOBACTERIAL PERSISTENCE REGULATOR MRPA (TWO COMPONENT RESPONSE TRANSCRIPTIONAL REGULATORY PROTEIN)
Rv1221	sigE	ALTERNATIVE RNA POLYMERASE SIGMA FACTOR SIGE
Rv2031c*[39,40]	hspX	HEAT SHOCK PROTEIN HSPX (ALPHA-CRSTALLIN HOMOLOG) (14 kDa ANTIGEN) (HSP16.3)
Rv2069	sigC	PROBABLE RNA POLYMERASE SIGMA FACTOR, ECF SUBFAMILY
Rv2710	sigB	RNA POLYMERASE SIGMA FACTOR
Rv3151*[36]	nuoG	PROBABLE NADH DEHYDROGENASE I (CHAIN G) NUOG (NADH-UBIQUINONE OXIDOREDUCTASE CHAIN G)
Rv3221A	Rv3221A	POSSIBLE ANTI-SIGMA FACTOR RSHA
Rv3223c	sigH	ALTERNATIVE RNA POLYMERASE SIGMA-E FACTOR (SIGMA-24)
Rv3286c	sigF	ALTERNATE RNA POLYMERASE SIGMA FACTOR
Rv3871*[38]	Rv3871	ESX CONSERVED COMPONENT ECCCB1 (ATPase activity)
Rv3874	esxB	10 KDA CULTURE FILTRATE ANTIGEN ESXB (LHP) (CFP10)
Rv3911	sigM	RNA POLYMERASE SIGMA FACTOR

RvIDs superscripted with asterisk are essential proteins as evidenced by genetic and biochemical studies.
doi:10.1371/journal.pone.0039808.t002

manually scanning the available literature, emphasis was placed on the references, which dealt with *Mycobacterium tuberculosis* H37Rv followed, by evidence in other mycobacterial species. If the protein was an enzyme, the corresponding reaction, along with the EC number, and the pathway(s) in which the protein participates was also recorded.

SOP for annotation in the absence of direct information from literature. In absence of direct literature information, annotations were derived based on sequence, structure and profile based information and analyses. To begin with, NCBI-BLAST [42] was used to obtain homology information of the query protein. Hits with e-value of $\leq$0.0001 and identity of $\geq$35%, with $\geq$75% sequence coverage were considered as significant hits. Annotations of the closest homologue were transferred and recorded in the template against each annotation. Similarity search in the Pfam [43] database was carried out to support BLAST results and also to annotate in the absence of high query coverage with BLAST analysis. If both BLAST and Pfam similarity search failed to give a significant hit, Phyre [44], an automatic fold recognition tool was used for predicting the function of the Mtb proteins through high confidence fold associations. Appropriate evidence codes have been used to distinguish between transferred annotations and experimental based annotations.

Data curation. Multiple rounds of collaborative data quality checks were carried out to ensure that the data has been correctly captured and reported. A set of instructions (SOPs) were devised for the same (https://sites.google.com/a/osdd.net/c2d-01/pathwayannotationproject/data-qc-guide) where the annotations curated by the members were systematically crosschecked iteratively by other members. It was interesting to note that high quality curation was achieved by this approach of 'many eyeballs make the bug shallow', a common phenomenon in open source software projects.

Data organization. The collated data was organized into a defined data structure as depicted in **Table 1** with two columns, field and description. The PSI MI (Proteomics Standards Initiative Molecular Interactions) was used as the community standard for reporting protein-protein interactions in MITAB format (Table S3). This helps in improving the representation of molecular interaction data and its accessibility to the user community.

Interactome Construction and Network Parameter Estimation

IPW and IPWSI network. The IPW-only network was constructed based on the annotations and curation of the data from IPW. The combined network of IPW and STRING termed, IPWSI, was constructed by combining the IPW network with that from STRING. All the interactions in STRING with high and medium level confidence score (above 400) were used to construct STRING based protein-protein interaction network. Methods used to compute network parameters are discussed below.

Network properties. To understand the functional organization of interacting proteins in both the networks, an analysis of various network topologies was performed. These network properties were computed using Boost Graph library in MATLAB (David Gleich; http://www.stanford.edu/~dgleich/programs/matlab_bgl/).

Connectivity or degree. The most elementary characteristic of a node in the network is its degree k, which represents, for a given node the number of other nodes it is connected to.

Clustering coefficient. The clustering coefficient was first defined by Watts and Strogatz [24]. The clustering coefficient, *C*, for a node is a notion of how connected the neighbours of a given node are to the other nodes (*cliquishness*) [45]. The average clustering coefficient for all nodes in a network is taken to be the network clustering coefficient. In an undirected graph, if a vertex *vi* has *ki* neighbors, $k\,i\,(k\,i - 1)/2$ edges could exist among the vertices

within the neighbourhood (N_i). The clustering coefficient for an undirected graph **G**(**V**, **E**) (where **V** represents the set of vertices in the graph **G** and **E** represents the set of edges) can then be defined as

$$C_i = \frac{2|\{e_{jk}\}|}{k_i(k_i-1)}; v_j, v_k \in N_i, e_{jk} \in E$$

The average clustering coefficient characterizes the overall tendency of nodes to form clusters or groups. $C(k)$ is defined as the average clustering coefficient for all nodes with k links.

Characteristic path length. The characteristic path length, L, is defined as the number of edges in the shortest path between two vertices, averaged over all pairs of vertices. It measures the typical separation between two vertices in the network. Intuitively, it represents the network's overall navigability [45].

Network diameter. The network diameter d is the greatest distance (shortest path, or geodesic path) between any two nodes in a network. It can also be viewed as the length of the 'longest' *shortest path* in the network.

$$d = \max_{u,v \in G} d_G(u, v)$$

where $d_{\mathbf{G}}(u, v)$ is the shortest path between u and v in **G** [45].

Power law distribution. For a given network the power law distribution states the probability that a given node has k links, which is given by equation $p(k) \sim k^{-\gamma}$, where γ is degree exponent. For smaller values of γ, the role of the 'hubs', or highly connected nodes, in the network becomes more important. For $\gamma > 3$, hubs are not relevant, while for $2 < \gamma < 3$, there is a hierarchy of hubs, with the most connected hub being in contact with a small fraction of all nodes. Scale-free networks have a high degree of robustness against random node failures, although they are sensitive to the failure of hubs [23]. The probability that a node is highly connected is statistically more significant than in a random graph [45].

Betweenness centrality. The betweenness centrality is the measure of vertex within a graph. For a given graph G(V,E), with n vertices, the betweenness CB (v) of a vertex v is defined as.

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where σ_{st} is the number of shortest path from s to t, and σ_{st} (v) is the number of shortest paths from s to t that passes from vertex v. The betweenness centrality analysis was performed for both the networks [45–46].

Drug Target Identification

Sequence homology with human proteome, oral and gut flora. The complete human proteome was downloaded from NCBI and BLAST was used to filter out the proteins, which had homology of greater than 45% with human protein. Human gut and oral flora constitutes the microbes that are considered to influence the physiology, nutrition, immunity and development of host. The complete proteome of 8-gut flora and 27 oral floras were downloaded. CD-HIT with similarity of 60% and a word size of 4 was used to compare the predicted proteins [27].

Binding site similarity with human proteome, oral and gut flora. We analyzed these proteins as reported in targetTB [30] pipeline where the top 10 binding sites for each protein was identified using PocketDepth algorithm [31]. The binding pockets of these proteins were then compared with human proteome using PocketMatch [32].

Peptide level conformation comparison with human proteome, oral and gut flora. We analyzed the proteins for absence of small peptides (octamer) [28] across human proteome, gut or oral flora using in house PERL scripts.

Literature based target validation. The predicted targets were further validated based on presence of existing functional evidence in literature. Data-mining and manual curation was performed to identify and document validated drug targets in Mtb. In addition to this, it was also documented whether the central protein is also reported to be essential or non-essential in context of Mtb growth and survival.

Web Server for Accessing and Searching IPW

The IPW data has been posted on http://sysborg2.osdd.net, the semantic web-based platform of Open Source Drug Discovery (OSDD) project [47]. For ease of access and search, the data is provided through a web-based server available at http://crdd.osdd.net/servers/ipw built using PHP and Mysql. This also works as the annotation and curation interface for the community. Any new submission to this web servers requires http://sysborg2.osdd.net open ID for authentication so that appropriate credits may be given to the members submitting updated information.

Supporting Information

Table S1 The annotations in the data structure format described in Table 1. This data may be searched in customized manner using the IPW web-interface (http://crdd.osdd.net/servers/ipw).
(XLSX)

Table S2 Rv Ids in lists b, b', c, d and e as obtained from various sequence and structural level analysis of central proteins as potential drug targets from IPW and IPWSI as depicted in Figure 2.
(XLSX)

Table S3 Central proteins predicted from analysis of the IPW interactome with details of interacting partners in PSI MITAB format.
(XLSX)

Acknowledgments

We would like to thank Dr. Vipin Singh, Assistant Professor, Amity University, Noida, for very constructive and detailed comments on the manuscript. We also thank Dr. TS Balganesh, CSIR, and Dr. Vani Brahmachari, ACBR, for fruitful discussions on the manuscript. We acknowledge India 800 foundation for support towards rewarding the best contributors with net books, Hewlett-Packard and Sun Microsystems for providing financial and logistics support for the on-site phase of C2D. We would also like to thank Mahanagar Telephone Nigam Limited, Delhi, India, for providing connectivity and the National Knowledge Network (NKN) for providing high-bandwidth support for C2D on-site phase activities. We also acknowledge School of Information Technology, Jawaharlal Nehru University, Delhi, India, for hosting the on-site phase and Dr. Andrew Michael Lynn and his group, School of Information Technology, Jawaharlal Nehru University, and Dr. S Ramachandran, CSIR-Institute of Genomics and Integrative Biology and his group, Delhi, for providing logistics support. We also thank Dr. GPS Raghava, CSIR-Institute of Microbial Technology, Chandigarh, for hosting the IPW web server. The authors thank all the OSDD members for their active participation.

Author Contributions

Wrote the paper: RV NC AB SKB. Project Management: AB VS ZT Project Conceptualization: SKB. Project Design: AB VS. Standard Operating Protocols: RV AGB SNS A. Shah KB GMR SZ AKM NC AB. IPW web server: SNS A. Sharma AB. Network Analysis: RV NC AB SKB. Analysis of Drug targets: AJ AKM A. Shah PV PP KSC AP VK KB KM PS MA DD RV AB SKB. Annotators: A. Subramanian AJ AA A. Shah AKM AV BJ GS GMR HR IR KB KV LC MT MG MR NA PS PM PV PP RP RS SZ SN S. Rana SS SA SNS SG S. Raj S. Raghavan TK VB VC VM AB. Curators: AJ AKM PP KB A. Shah PV HR VC A. Subramanium SZ YH GS AV KV AGB RV AB.

References

1. World Health Organization (WHO) (2010) Global tuberculosis control.
2. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, et al. (1998) Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature 393: 537–544.
3. Camus J-C, Pryor MJ, Médigue C, Cole ST (2002) Re-annotation of the genome sequence of Mycobacterium tuberculosis H37Rv. Microbiology 148: 2967–2973.
4. Lew JM, Kapopoulou A, Jones LM, Cole ST (2011) TubercuList –10 years after. Tuberculosis 91: 1–7.
5. Reddy TBK, Riley R, Wymore F, Montgomery P, DeCaprio D, et al. (2009) TB database: an integrated platform for tuberculosis research. Nucleic Acids Research 37: D499-D508.
6. Bhardwaj A, Scaria V, Raghava GPS, Lynn AM, Chandra N, et al. (2011) Open source drug discovery– A new paradigm of collaborative research in tuberculosis drug development. Tuberculosis 91: 479–486.
7. Singh S (2008) India Takes an Open Source Approach to Drug Discovery. Cell 133: 201–203.
8. Chandra N, Kumar D, Rao K (2011) Systems biology of tuberculosis. Tuberculosis 91: 487–496.
9. Arora P, Goyal A, Natarajan VT, Rajakumara E, Verma P, et al. (2009) Mechanistic and functional insights into fatty acid activation in Mycobacterium tuberculosis. Nat Chem Biol 5: 166–173.
10. Kumar D, Nath L, Kamal MA, Varshney A, Jain A, et al. (2010) Genome-wide Analysis of the Host Intracellular Network that Regulates Survival of Mycobacterium tuberculosis. Cell 140: 731–743.
11. Munos B (2010) Can Open-Source Drug R&D Repower Pharmaceutical Innovation? Clin Pharmacol Ther 87: 534–536.
12. Kitano H, Ghosh S, Matsuoka Y (2011) Social engineering for virtual 'big science' in systems biology. Nat Chem Biol 7: 323–326.
13. Freeman LC (1977) A set of measures of centrality based on betweenness. Sociometry 40: 35–41.
14. Hey AJG (2009) The fourth paradigm:data intensive scientific discovery: Microsoft Research.
15. Bader GD, Donaldson I, Wolting C, Ouellette BFF, Pawson T, et al. (2001) BIND–The Biomolecular Interaction Network Database. Nucleic Acids Research 29: 242–245.
16. Prieto C, De Las Rivas J (2006) APID: Agile Protein Interaction DataAnalyzer. Nucleic Acids Research 34: W298–W302.
17. Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, et al. (2010) The IntAct molecular interaction database in 2010. Nucleic Acids Research 38: D525–D531.
18. Xenarios I, Fernandez E, Salwinski L, Duan XJ, Thompson MJ, et al. (2001) DIP: The Database of Interacting Proteins: 2001 update. Nucleic Acids Research 29: 239–241.
19. Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, et al. (2007) MINT: the Molecular INTeraction database. Nucleic Acids Research 35: D572–D574.
20. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, et al. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research 27: 29–34.
21. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, et al. (2009) STRING 8–a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Research 37: D412–D416.
22. Mason O, Verwoerd M (2007) Graph theory and networks in Biology. IET Systems Biology 1: 89–119.
23. Barabasi A-L, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5: 101–113.
24. Watts DJ, Strogatz SH (1998) Collective dynamics of '/small-world/' networks. Nature 393: 440–442.
25. Grigorov MG (2005) Global properties of biological networks. Drug Discovery Today 10: 365–372.
26. Barabasi AL, Albert R (1999) Emergence of Scaling in Random Networks. Science 286: 509–512.
27. Anurag M, Dash D (2009) Unraveling the potential of intrinsically disordered proteins as drug targets: application to Mycobacterium tuberculosis. Molecular BioSystems 5: 1752–1757.
28. Prakash T, Ramakrishnan C, Dash D, Brahmachari SK (2005) Conformational Analysis of Invariant Peptide Sequences in Bacterial Genomes. Journal of Molecular Biology 345: 937–955.
29. Pieper U, Webb BM, Barkan DT, Schneidman-Duhovny D, Schlessinger A, et al. (2011) ModBase, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Research 39: D465–D474.
30. Raman K, Yeturu K, Chandra N (2008) targetTB: A target identification pipeline for Mycobacterium tuberculosis through an interactome, reactome and genome-scale structural analysis. BMC Systems Biology 2: 109.
31. Kalidas Y, Chandra N (2008) PocketDepth: A new depth based algorithm for identification of ligand binding sites in proteins. Journal of Structural Biology 161: 31–42.
32. Yeturu K, Chandra N (2008) PocketMatch: A new algorithm to compare binding sites in protein structures. BMC Bioinformatics 9: 543.
33. Song T, Dove SL, Lee KH, Husson RN (2003) RshA, an anti-sigma factor that regulates the activity of the mycobacterial stress response sigma factor SigH. Molecular Microbiology 50: 949–959.
34. Greenstein AE, MacGurn JA, Baer CE, Falick AM, Cox JS, et al. (2007) M. tuberculosis Ser/Thr protein kinase D phosphorylates an anti-anti-sigma factor homolog. PLoS Pathog 3: e49.
35. Magnet S, Hartkoorn RC, Székely R, Pató J, Triccas JA, et al. (2010) Leads for antitubercular compounds from kinase inhibitor library screens. Tuberculosis 90: 354–360.
36. Velmurugan K, Chen B, Miller JL, Azogue S, Gurses S, et al. (2007) *Mycobacterium tuberculosis* nuoG Is a Virulence Gene That Inhibits Apoptosis of Infected Host Cells. PLoS Pathog 3: e110.
37. Menon S, Wang S (2011) Structure of the Response Regulator PhoP from Mycobacterium tuberculosis Reveals a Dimer through the Receiver Domain. Biochemistry 50: 5948–5957.
38. Guinn KM, Hickey MJ, Mathur SK, Zakel KL, Grotzke JE, et al. (2004) Individual RD1-region genes are required for export of ESAT-6/CFP-10 and for virulence of Mycobacterium tuberculosis. Molecular Microbiology 51: 359–370.
39. Hu Y, Coates ARM (2011) *Mycobacterium tuberculosis* acg Gene Is Required for Growth and Virulence In Vivo. PLoS ONE 6: e20958.
40. Hu Y, Movahedzadeh F, Stoker NG, Coates ARM (2006) Deletion of the Mycobacterium tuberculosis α-Crystallin-Like hspX Gene Causes Increased Bacterial Growth In Vivo. Infection and Immunity 74: 861–868.
41. Sharma S, Kumar M, Sharma S, Nargotra A, Koul S, et al. (2010) Piperine as an inhibitor of Rv1258c, a putative multidrug efflux pump of Mycobacterium tuberculosis. Journal of Antimicrobial Chemotherapy 65: 1694–1701.
42. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) BASIC LOCAL ALIGNMENT SEARCH TOOL. Journal of Molecular Biology 215: 403–410.
43. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, et al. (2002) The Pfam Protein Families Database. Nucleic Acids Research 30: 276–280.
44. Kelley LA, Sternberg MJE (2009) Protein structure prediction on the Web: a case study using the Phyre server. Nat Protocols 4: 363–371.
45. Raman K (2010) Construction and analysis of protein-protein interaction networks. Automated Experimentation 2: 2.
46. Newman MEJ (2005) A measure of betweenness centrality based on random walks. Social Networks 27: 39–54.
47. Bhardwaj A, Scaria V, Thomas Z, Adayikkoth S, Open Source Drug Discovery (OSDD) Consortium, et al. (2011) Collaborative Tools to Accelerate Neglected Disease Research: the Open Source Drug Discovery Model Sean Ekins MAZH, Antony J Williams, editor: John Wiley & Sons, Inc. 576 p.
48. Sachdeva P, Misra R, Tyagi AK, Singh Y (2010) The sigma factors of Mycobacterium tuberculosis: regulation of the regulators. FEBS Journal 277: 605–626.

Journal of Bacteriology

Genetic Fine Structure of a Salmonella enterica Serovar Typhi Strain Associated with the 2005 Outbreak of Typhoid Fever in Kelantan, Malaysia

Ramani Baddam, Narender Kumar, Kwai-Lin Thong, Soo-Tein Ngoi, Cindy Shuan Ju Teh, Kien-Pong Yap, Lay-Ching Chai, Tiruvayipati Suma Avasthi and Niyaz Ahmed

J. Bacteriol. **2012, 194(13):3565. DOI: 10.1128/JB.00581-12.**

Updated information and services can be found at:
http://jb.asm.org/content/194/13/3565

These include:

REFERENCES This article cites 25 articles, 13 of which can be accessed free at: http://jb.asm.org/content/194/13/3565#ref-list-1

CONTENT ALERTS Receive: RSS Feeds, eTOCs, free email alerts (when new articles cite this article), more»

Information about commercial reprint orders: **http://journals.asm.org/site/misc/reprints.xhtml**
To subscribe to to another ASM Journal go to: **http://journals.asm.org/site/subscriptions/**

Journals.ASM.org

Downloaded from http://jb.asm.org/ on April 30, 2014 by UNIVERSITY OF MALAYA

JB Journals.ASM.org

GENOME ANNOUNCEMENT

Genetic Fine Structure of a *Salmonella enterica* Serovar Typhi Strain Associated with the 2005 Outbreak of Typhoid Fever in Kelantan, Malaysia

Ramani Baddam,[a] Narender Kumar,[a] Kwai-Lin Thong,[b,c] Soo-Tein Ngoi,[b,c] Cindy Shuan Ju Teh,[c] Kien-Pong Yap,[b,c] Lay-Ching Chai,[b,c] Tiruvayipati Suma Avasthi,[a,b] and Niyaz Ahmed[a,b]

Pathogen Biology Laboratory, Department of Biotechnology, School of Life Sciences, University of Hyderabad, Hyderabad, India[a]; Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia[b]; and Laboratory of Biomedical Science and Molecular Microbiology, UMBIO Research Cluster, University of Malaya, Kuala Lumpur, Malaysia[c]

Among enteric pathogens, *Salmonella enterica* serovar Typhi is responsible for the largest number of food-borne outbreaks and fatalities. The ability of the pathogen to cause systemic infection for extended durations leads to a high cost of disease control. Chronic carriers play important roles in the evolution of *Salmonella* Typhi; therefore, identification and in-depth characterization of isolates from clinical cases and carriers, especially those from zones of endemicity where the pathogen has not been extensively studied, are necessary. Here, we describe the genome sequence of the highly virulent *Salmonella* Typhi strain BL196/05 isolated during the outbreak of typhoid in Kelantan, Malaysia, in 2005. The whole-genome sequence and comparative genomics of this strain should enable us to understand the virulence mechanisms and evolutionary dynamics of this pathogen in Malaysia and elsewhere.

Salmonella enterica serovar Typhi and other pathogenic salmonellae are endemic in some countries (8, 15), and outbreaks occur due to unhygienic conditions (16), leading to alarming morbidity and mortality figures (9, 18). This results in a huge burden on the public health machinery. *Salmonella* Typhi persists for an extended duration in its host (18) due to modulation of the host immune responses together with genetic rearrangements that confer fitness advantages (12, 13, 17).

The incidence of typhoid in Kelantan had always been higher than in other states of Malaysia. In the 2005 outbreak (April to June 2005), 735 cases and 2 deaths occurred (18). We hypothesized that the genome sequences of the underlying strains would provide more insights to enhance understanding of endemicity or persistence of typhoid in Kelantan.

Salmonella Typhi BL196/05 was isolated from blood samples of a severe typhoid case in Kelantan during the notorious outbreak of 2005. The strain was characterized by PCR to determine the presence of many different virulence genes. The genome sequence was analyzed and annotated exactly as described previously (4, 5, 11, 21). Briefly, the 73-bp paired-end Illumina sequence reads, amounting to 1.7 gigabytes of data (insert size, 300 bp), were generated with 80× genome coverage. Velvet (26), with the hash length set to 37, was used to assemble sequence reads into 191 contigs; these were further assembled into a draft genome and were submitted to RAST (6) to determine the following data. The size of the single chromosome was approximately 4,744,056 bp, with a G+C content of 53.21% and a coding percentage of 87.1. There were 4,875 protein coding sequences found, with an average length of 875 bp. The genome revealed 76 tRNA and 22 rRNA genes. Our strain did not harbor any plasmid, and its phage typing revealed Vi phage type B1. The genome sequence revealed two *mar* regulons, *marRAB* and *marC*, as reported also in *Salmonella* Typhi strains CT18 and Ty2, and these were homologous to *Escherichia coli mar* (multiple antibiotic resistance) regulon members (2, 3, 10, 20, 25). The genes encoding a melittin resistance protein, PqaB, and a polymyxin resistance protein, PmrD (7), were located in the genome as also seen in the *Salmonella* Typhi CT18 and Ty2 genomes. The methyl viologen resistance gene (14) *smvA* was also identified. Several pathogenicity islands as well as a pool of hypothetical proteins were annotated.

Considerable genetic diversity exists among human isolates of *Salmonella* Typhi in Malaysia and Southeast Asia (22, 23), and the heterogeneity in genome sizes (24) points to a high degree of plasticity in the *Salmonella* pan-genome. The genome sequence of strain BL196/05 also revealed rearrangements putatively relevant to virulence optimization, persistence, and adaptation within the host. Such information would be harnessed to improve understanding of genome evolution and adaptation dynamics of *Salmonella* in Malaysia and to develop point-of-care diagnostics. Further, future efforts are needed to evaluate the significance of comparative genomics/genotypic data in juxtaposition with host genetic polymorphisms to herald the beginning of the "functional molecular infection epidemiology" (1) of typhoid. Finally, we recommend greater use of data determined in studies of clinical isolates and strains specific to outbreaks and countries rather than use of the reference (type) strains alone.

Nucleotide sequence accession numbers. This genome project has been deposited at GenBank under accession no. AJGK00000000. The version described herein is the first version, AJGK01000000. The Bioproject designation for this project is PRJNA85621.

Received 9 April 2012 Accepted 10 April 2012
Address correspondence to Niyaz Ahmed, niyazSL@uohyd.ernet.in.
Copyright © 2012, American Society for Microbiology. All Rights Reserved.
doi:10.1128/JB.00581-12

Downloaded from http://jb.asm.org/ on April 30, 2014 by UNIVERSITY OF MALAYA

ACKNOWLEDGMENTS

We thankfully acknowledge support received from the University of Malaya (High Impact Research Grant; reference UM.C/625/1HIR/MOHE/-02 [A000002-5000 1])—Molecular Genetics. This genome program was completed under the wider umbrella of the Indo-German International Research Training Group—Internationales Graduiertenkolleg (GRK1673)—Functional Molecular Infection Epidemiology, an initiative of the German Research Foundation (DFG) and the University of Hyderabad (India), of which N.A. is a Speaker. N.A. is an Adjunct Professor at the Institute of Biological Sciences at the Universiti Malaya, Kuala Lumpur, Malaysia, and a Visiting Adjunct Professor at the Universiti Brunei—Darussalam, Brunei. N.K. acknowledges the Junior Research Fellowship received from the Council of Scientific and Industrial Research (CSIR), India.

We are also grateful to M/s Genotypic Technology Pvt. Ltd., Bengaluru, India, for their measureless efforts with respect to Illumina sequencing. We acknowledge the Bioinformatics Facility (BIF) at the Department of Biotechnology, University of Hyderabad, for the use of servers and computational infrastructure. We thank Dipshikha Chakravortty for critical reading of the manuscript. All other members of the Ahmed and Thong laboratories are gratefully acknowledged for their help and support.

REFERENCES

1. **Ahmed N.** 2011. Coevolution and adaptation of *Helicobacter pylori* and the case for 'functional molecular infection epidemiology'. Med. Princ. Pract. **20**:497–503.
2. **Alekshun MN, Levy SB.** 1997. Regulation of chromosomally mediated multiple antibiotic resistance: the mar regulon. Antimicrob. Agents Chemother. **41**:2067–2075.
3. **Alekshun MN, Levy SB.** 1999. The mar regulon: multiple resistance to antibiotics and other toxic chemicals. Trends Microbiol. **7**:410–413.
4. **Avasthi TS, et al.** 2011. Genomes of two chronological isolates (Helicobacter pylori 2017 and 2018) of the West African Helicobacter pylori strain 908 obtained from a single patient. J. Bacteriol. **193**:3385–3386.
5. **Avasthi TS, et al.** 2011. Genome of multidrug-resistant uropathogenic Escherichia coli strain NA114 from India. J. Bacteriol. **193**:4272–4273.
6. **Aziz RK, et al.** 2008. The RAST server: rapid annotations using subsystems technology. BMC Genomics **9**:75. doi:10.1186/1471-2164-9-75.
7. **Baker SJ, Gunn JS, Morona R.** 1999. The *Salmonella typhi* melittin resistance gene *pqaB* affects intracellular growth in PMA-differentiated U937 cells, polymyxin B resistance and lipopolysaccharide. Microbiology **145**(Pt. 2):367–378.
8. **Crump JA, Mintz ED.** 2010. Global trends in typhoid and paratyphoid fever. Clin. Infect. Dis. **50**:241–246.
9. **Crump JA, Luby SP, Mintz ED.** 2004. The global burden of typhoid fever. Bull. World Health Organ. **82**:346–353.
10. **Deng W, et al.** 2003. Comparative genomics of Salmonella enterica serovar Typhi strains Ty2 and CT18. J. Bacteriol. **185**:2330–2337.
11. **Devi SH, et al.** 2010. Genome of Helicobacter pylori strain 908. J. Bacteriol. **192**:6488–6489.
12. **Everest P, Wain J, Roberts M, Rook G, Dougan G.** 2001. The molecular mechanisms of severe typhoid fever. Trends Microbiol. **9**:316–320.
13. **Holt KE, et al.** 2008. High-throughput sequencing provides insights into genome variation and evolution in Salmonella Typhi. Nat. Genet. **40**:987–993.
14. **Hongo E, et al.** 1994. The methyl viologen-resistance-encoding gene smvA of Salmonella typhimurium. Gene **148**:173–174.
15. **House D, Bishop A, Parry C, Dougan G, Wain J.** 2001. Typhoid fever: pathogenesis and disease. Curr. Opin. Infect. Dis. **14**:573–578.
16. **Jain S, et al.** 2009. Multistate outbreak of Salmonella Typhimurium and Saintpaul infections associated with unpasteurized orange juice—United States, 2005. Clin. Infect. Dis. **48**:1065–1071.
17. **Jones BD, Falkow S.** 1996. Salmonellosis: host immune responses and bacterial virulence determinants. Annu. Rev. Immunol. **14**:533–561.
18. **Ministry of Health Malaysia.** 2006. Epidemiology of foodborne diseases in Malaysia, chapter 2, p 86–87. Director General Ministry of Health Malaysia Technical Report. Ministry of Health Malaysia, Putrajaya, Malaysia.
19. **Monack DM.** 2012. Salmonella persistence and transmission strategies. Curr. Opin. Microbiol. **15**:100–107.
20. **Parkhill J, et al.** 2001. Complete genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi CT18. Nature **413**:848–852.
21. **Siddavattam D, et al.** 2011. Genome of a novel isolate of Paracoccus denitrificans capable of degrading N,N-dimethylformamide. J. Bacteriol. **193**:5598–5599.
22. **Thong KL, Cheong YM, Puthucheary SD, Koh CL, Pang T.** 1994. Epidemiologic analysis of sporadic Salmonella Typhi isolates and those from outbreaks by pulsed-field gel electrophoresis. J. Clin. Microbiol. **32**:1135–1141.
23. **Thong KL, et al.** 1995. Analysis of *Salmonella* Typhi from Southeast Asia by pulsed-field gel electrophoresis. J. Clin. Microbiol. **33**:1938–1941.
24. **Thong KL, Puthucheary SD, Pang T.** 1996. Genome size variation among recent human isolates of *Salmonella typhi*. Res. Microbiol. **148**:229–235.
25. **Winfield MD, Groisman EA.** 2004. Phenotypic differences between Salmonella and Escherichia coli resulting from the disparate regulation of homologous genes. Proc. Natl. Acad. Sci. U. S. A. **101**:17162–17167.
26. **Zerbino DR, Birney E.** 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. **18**:821–829.

Downloaded from http://jb.asm.org/ on April 30, 2014 by UNIVERSITY OF MALAYA

APPENDIX A

(a) Adaptor File

>Illumina Single End Adapter 1
GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG
>Illumina Single End Adapter 1
GATCGGAAGAGCTCGTATGCCGTCTTCTGCTT
>Illumina Single End Adapter 1
GATCGGAAGAGCTCGTATGCCGTCTTCTGCT
>Illumina Single End Adapter 1
GATCGGAAGAGCTCGTATGCCGTCTTCTGC
>Illumina Single End Adapter 1
GATCGGAAGAGCTCGTATGCCGTCTTCTG
>Illumina Single End Adapter 1
GATCGGAAGAGCTCGTATGCCGTCTTCT
>Illumina Single End Adapter 1
GATCGGAAGAGCTCGTATGCCGTCTTC
>Illumina Single End Adapter 1
GATCGGAAGAGCTCGTATGCCGTCTT
>Illumina Single End Adapter 1
GATCGGAAGAGCTCGTATGCCGTCT
>Illumina Single End Adapter 1
GATCGGAAGAGCTCGTATGCCGTC
>Illumina Single End Adapter 1
GATCGGAAGAGCTCGTATGCCGT
>Illumina Single End Adapter 1
GATCGGAAGAGCTCGTATGCCG
>Illumina Single End Adapter 1
GATCGGAAGAGCTCGTATGCC
>Illumina Single End Adapter 1
GATCGGAAGAGCTCGTATGC
>Illumina Single End Adapter 1
GATCGGAAGAGCTCGTATG
>Illumina Single End Adapter 1
GATCGGAAGAGCTCGTAT
>Illumina Single End Adapter 1
GATCGGAAGAGCTCGTA
>Illumina Single End Adapter 1
GATCGGAAGAGCTCGT
>Illumina Single End Adapter 1
GATCGGAAGAGCTCG
>Illumina Single End Adapter 1
GATCGGAAGAGCTC
>Illumina Single End Adapter 1
GATCGGAAGAGCT
>Illumina Single End Adapter 1
GATCGGAAGAGC
>Illumina Single End Adapter 2
CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT
>Illumina Single End Adapter 2
CAAGCAGAAGACGGCATACGAGCTCTTCCGATC
>Illumina Single End Adapter 2

CAAGCAGAAGACGGCATACGAGCTCTTCCGAT
>Illumina Single End Adapter 2
CAAGCAGAAGACGGCATACGAGCTCTTCCGA
>Illumina Single End Adapter 2
CAAGCAGAAGACGGCATACGAGCTCTTCCG
>Illumina Single End Adapter 2
CAAGCAGAAGACGGCATACGAGCTCTTCC
>Illumina Single End Adapter 2
CAAGCAGAAGACGGCATACGAGCTCTTC
>Illumina Single End Adapter 2
CAAGCAGAAGACGGCATACGAGCTCTT
>Illumina Single End Adapter 2
CAAGCAGAAGACGGCATACGAGCTCT
>Illumina Single End Adapter 2
CAAGCAGAAGACGGCATACGAGCTC
>Illumina Single End Adapter 2
CAAGCAGAAGACGGCATACGAGCT
>Illumina Single End Adapter 2
CAAGCAGAAGACGGCATACGAGC
>Illumina Single End Adapter 2
CAAGCAGAAGACGGCATACGAG
>Illumina Single End Adapter 2
CAAGCAGAAGACGGCATACGA
>Illumina Single End Adapter 2
CAAGCAGAAGACGGCATACG
>Illumina Single End Adapter 2
CAAGCAGAAGACGGCATAC
>Illumina Single End Adapter 2
CAAGCAGAAGACGGCATA
>Illumina Single End Adapter 2
CAAGCAGAAGACGGCAT
>Illumina Single End Adapter 2
CAAGCAGAAGACGGCA
>Illumina Single End Adapter 2
CAAGCAGAAGACGGC
>Illumina Single End Adapter 2
CAAGCAGAAGACGG
>Illumina Single End Adapter 2
CAAGCAGAAGACG
>Illumina Single End Adapter 2
CAAGCAGAAGAC
>Illumina Single End PCR Primer 1
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC
CGATCT
>Illumina Single End PCR Primer 2
CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT
>Illumina Single End Sequencing Primer
ACACTCTTTCCCTACACGACGCTCTTCCGATCT
>Illumina Paired End Adapter 1
ACACTCTTTCCCTACACGACGCTCTTCCGATCT
>Illumina Paired End Adapter 1
ACACTCTTTCCCTACACGACGCTCTTCCGATC

>Illumina Paired End Adapter 1
ACACTCTTTCCCTACACGACGCTCTTCCGAT
>Illumina Paired End Adapter 1
ACACTCTTTCCCTACACGACGCTCTTCCGA
>Illumina Paired End Adapter 1
ACACTCTTTCCCTACACGACGCTCTTCCG
>Illumina Paired End Adapter 1
ACACTCTTTCCCTACACGACGCTCTTCC
>Illumina Paired End Adapter 1
ACACTCTTTCCCTACACGACGCTCTTC
>Illumina Paired End Adapter 1
ACACTCTTTCCCTACACGACGCTCTT
>Illumina Paired End Adapter 1
ACACTCTTTCCCTACACGACGCTCT
>Illumina Paired End Adapter 1
ACACTCTTTCCCTACACGACGCTC
>Illumina Paired End Adapter 1
ACACTCTTTCCCTACACGACGCT
>Illumina Paired End Adapter 1
ACACTCTTTCCCTACACGACGC
>Illumina Paired End Adapter 1
ACACTCTTTCCCTACACGACG
>Illumina Paired End Adapter 1
ACACTCTTTCCCTACACGAC
>Illumina Paired End Adapter 1
ACACTCTTTCCCTACACGA
>Illumina Paired End Adapter 1
ACACTCTTTCCCTACACG
>Illumina Paired End Adapter 1
ACACTCTTTCCCTACAC
>Illumina Paired End Adapter 1
ACACTCTTTCCCTACA
>Illumina Paired End Adapter 1
ACACTCTTTCCCTAC
>Illumina Paired End Adapter 1
ACACTCTTTCCCTA
>Illumina Paired End Adapter 1
ACACTCTTTCCCT
>Illumina Paired End Adapter 1
ACACTCTTTCCC
>Illumina Paired End Adapter 2
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
>Illumina Paired End Adapter 2
GATCGGAAGAGCGGTTCAGCAGGAATGCCGA
>Illumina Paired End Adapter 2
GATCGGAAGAGCGGTTCAGCAGGAATGCCG
>Illumina Paired End Adapter 2
GATCGGAAGAGCGGTTCAGCAGGAATGCC
>Illumina Paired End Adapter 2
GATCGGAAGAGCGGTTCAGCAGGAATGC
>Illumina Paired End Adapter 2
GATCGGAAGAGCGGTTCAGCAGGAATG

>Illumina Paired End Adapter 2
GATCGGAAGAGCGGTTCAGCAGGAAT
>Illumina Paired End Adapter 2
GATCGGAAGAGCGGTTCAGCAGGAA
>Illumina Paired End Adapter 2
GATCGGAAGAGCGGTTCAGCAGGA
>Illumina Paired End Adapter 2
GATCGGAAGAGCGGTTCAGCAGG
>Illumina Paired End Adapter 2
GATCGGAAGAGCGGTTCAGCAG
>Illumina Paired End Adapter 2
GATCGGAAGAGCGGTTCAGCA
>Illumina Paired End Adapter 2
GATCGGAAGAGCGGTTCAGC
>Illumina Paired End Adapter 2
GATCGGAAGAGCGGTTCAG
>Illumina Paired End Adapter 2
GATCGGAAGAGCGGTTCA
>Illumina Paired End Adapter 2
GATCGGAAGAGCGGTTC
>Illumina Paired End Adapter 2
GATCGGAAGAGCGGTT
>Illumina Paired End Adapter 2
GATCGGAAGAGCGGT
>Illumina Paired End Adapter 2
GATCGGAAGAGCGG
>Illumina Paired End Adapter 2
GATCGGAAGAGCG
>Illumina Paired End Adapter 2
GATCGGAAGAGC
>Illumina Paried End PCR Primer 1
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC
CGATCT
>Illumina Paired End PCR Primer 2
CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCT
CTTCCGATCT
>Illumina Paried End Sequencing Primer 1
ACACTCTTTCCCTACACGACGCTCTTCCGATCT
>Illumina Paired End Sequencing Primer 2
CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT
>Illumina Multiplexing Adapter 1
GATCGGAAGAGCACACGTCT
>Illumina Multiplexing Adapter 1
GATCGGAAGAGCACACGTC
>Illumina Multiplexing Adapter 1
GATCGGAAGAGCACACGT
>Illumina Multiplexing Adapter 1
GATCGGAAGAGCACACG
>Illumina Multiplexing Adapter 1
GATCGGAAGAGCACAC
>Illumina Multiplexing Adapter 1
GATCGGAAGAGCACA

>Illumina Multiplexing Adapter 1
GATCGGAAGAGCAC
>Illumina Multiplexing Adapter 1
GATCGGAAGAGCA
>Illumina Multiplexing Adapter 1
GATCGGAAGAGC
>Illumina Multiplexing Adapter 2
ACACTCTTTCCCTACACGACGCTCTTCCGATCT
>Illumina Multiplexing PCR Primer 1.01
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC
CGATCT
>Illumina Multiplexing PCR Primer 2.01
GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
>Illumina Multiplexing Read1 Sequencing Primer
ACACTCTTTCCCTACACGACGCTCTTCCGATCT
>Illumina Multiplexing Index Sequencing Primer
GATCGGAAGAGCACACGTCTGAACTCCAGTCAC
>Illumina Multiplexing Read2 Sequencing Primer
GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
>Illumina PCR Primer Index 1
CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTC
>Illumina PCR Primer Index 2
CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTC
>Illumina PCR Primer Index 3
CAAGCAGAAGACGGCATACGAGATGCCTAAGTGACTGGAGTTC
>Illumina PCR Primer Index 4
CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTC
>Illumina PCR Primer Index 5
CAAGCAGAAGACGGCATACGAGATCACTGTGTGACTGGAGTTC
>Illumina PCR Primer Index 6
CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTC
>Illumina PCR Primer Index 7
CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTC
>Illumina PCR Primer Index 8
CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGAGTTC
>Illumina PCR Primer Index 9
CAAGCAGAAGACGGCATACGAGATCTGATCGTGACTGGAGTTC
>Illumina PCR Primer Index 10
CAAGCAGAAGACGGCATACGAGATAAGCTAGTGACTGGAGTTC
>Illumina PCR Primer Index 11
CAAGCAGAAGACGGCATACGAGATGTAGCCGTGACTGGAGTTC
>Illumina PCR Primer Index 12
CAAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAGTTC
>TruSeq Universal Adapter
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC
CGATCT
>TruSeq Universal Adapter
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC
CGATC
>TruSeq Universal Adapter
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC
CGAT

>TruSeq Universal Adapter
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC
CGA
>TruSeq Universal Adapter
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC
CG
>TruSeq Universal Adapter
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC
C
>TruSeq Universal Adapter
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC
>TruSeq Universal Adapter
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTT
>TruSeq Universal Adapter
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCT
>TruSeq Universal Adapter
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTC
>TruSeq Universal Adapter
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCT
>TruSeq Universal Adapter
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGC
>TruSeq Universal Adapter
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACG
>TruSeq Universal Adapter
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGAC
>TruSeq Universal Adapter
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGA
>TruSeq Universal Adapter
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACG
>TruSeq Universal Adapter
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACAC
>TruSeq Universal Adapter
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACA
>TruSeq Universal Adapter
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTAC
>TruSeq Universal Adapter
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTA
>TruSeq Universal Adapter
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCT
>TruSeq Universal Adapter
AATGATACGGCGACCACCGAGATCTACACTCTTTCCC
>TruSeq Universal Adapter
AATGATACGGCGACCACCGAGATCTACACTCTTTCC
>TruSeq Universal Adapter
AATGATACGGCGACCACCGAGATCTACACTCTTTC
>TruSeq Universal Adapter
AATGATACGGCGACCACCGAGATCTACACTCTTT
>TruSeq Universal Adapter
AATGATACGGCGACCACCGAGATCTACACTCTT
>TruSeq Universal Adapter
AATGATACGGCGACCACCGAGATCTACACTCT
>TruSeq Universal Adapter

AATGATACGGCGACCACCGAGATCTACACTC
>TruSeq Universal Adapter
AATGATACGGCGACCACCGAGATCTACACT
>TruSeq Universal Adapter
AATGATACGGCGACCACCGAGATCTACAC
>TruSeq Universal Adapter
AATGATACGGCGACCACCGAGATCTACA
>TruSeq Universal Adapter
AATGATACGGCGACCACCGAGATCTAC
>TruSeq Universal Adapter
AATGATACGGCGACCACCGAGATCTA
>TruSeq Universal Adapter
AATGATACGGCGACCACCGAGATCT
>TruSeq Universal Adapter
AATGATACGGCGACCACCGAGATC
>TruSeq Universal Adapter
AATGATACGGCGACCACCGAGAT
>TruSeq Universal Adapter
AATGATACGGCGACCACCGAGA
>TruSeq Universal Adapter
AATGATACGGCGACCACCGAG
>TruSeq Universal Adapter
AATGATACGGCGACCACCGA
>TruSeq Universal Adapter
AATGATACGGCGACCACCG
>TruSeq Universal Adapter
AATGATACGGCGACCACC
>TruSeq Universal Adapter
AATGATACGGCGACCAC
>TruSeq Universal Adapter
AATGATACGGCGACCA
>TruSeq Universal Adapter
AATGATACGGCGACC
>TruSeq Universal Adapter
AATGATACGGCGAC
>TruSeq Universal Adapter
AATGATACGGCG
>TruSeq Universal Adapter
AATGATACGGC
>TruSeq Adapter, Index 1
GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCC
GTCTTCTGCTTG
>TruSeq Adapter, Index 2
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGCC
GTCTTCTGCTTG
>TruSeq Adapter, Index 3
GATCGGAAGAGCACACGTCTGAACTCCAGTCACTTAGGCATCTCGTATGCC
GTCTTCTGCTTG
>TruSeq Adapter, Index 4
GATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGTATGCC
GTCTTCTGCTTG
>TruSeq Adapter, Index 5

GATCGGAAGAGCACACGTCTGAACTCCAGTCACACAGTGATCTCGTATGCC
GTCTTCTGCTTG
>TruSeq Adapter, Index 6
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGCCAATATCTCGTATGCC
GTCTTCTGCTTG
>TruSeq Adapter, Index 7
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAGATCATCTCGTATGCC
GTCTTCTGCTTG
>TruSeq Adapter, Index 8
GATCGGAAGAGCACACGTCTGAACTCCAGTCACACTTGAATCTCGTATGCC
GTCTTCTGCTTG
>TruSeq Adapter, Index 9
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGATCAGATCTCGTATGCC
GTCTTCTGCTTG
>TruSeq Adapter, Index 10
GATCGGAAGAGCACACGTCTGAACTCCAGTCACTAGCTTATCTCGTATGCC
GTCTTCTGCTTG
>TruSeq Adapter, Index 11
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGGCTACATCTCGTATGCC
GTCTTCTGCTTG
>TruSeq Adapter, Index 12
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCTTGTAATCTCGTATGCC
GTCTTCTGCTTG
>Solexa 5' adapter
AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAG
>Solexa 3' adapter
TTTTCGTATGCCGTCTTCTGCTTG
>PE Adapters1
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
>PE Adapters1_!
ACACTCTTTCCCTACACGACGCTCTTCCGATCT
>PE PCR Primers1
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC
CGATCT
>PE PCR Primers1_1
CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCT
CTTCCGATCT
>PE Sequencing Primer
ACACTCTTTCCCTACACGACGCTCTTCCGATCT
>PE Sequencing Primer_1
CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT
>Gex Adapter 1
GATCGTCGGACTGTAGAACTCTGAAC
>Gex Adapter 1_2
ACAGGTTCAGAGTTCTACAGTCCGAC
>Gex Adapter 2
CAAGCAGAAGACGGCATACGANN
>Gex Adapter 2_2
TCGTATGCCGTCTTCTGCTTG
>Gex PCR Primer 1
CAAGCAGAAGACGGCATACGA
>Gex PCR Primer 2

AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGA
>Gex Sequencing Primer
CGACAGGTTCAGAGTTCTACAGTCCGACGATC
>Adapters1
GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG
>Adapters1_1
ACACTCTTTCCCTACACGACGCTCTTCCGATCT
>PCR Primers1
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC
CGATCT
>PCR Primers1_1
CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT
>Genomic DNA Sequencing Primer
ACACTCTTTCCCTACACGACGCTCTTCCGATCT

(b) QUALITY CONTROL DATA

OUTPUT DATA GENERATED FOR THE FIVE DIFFERENT BACTERIAL GENOMES USED

HQ High Quality
HQ_VF High Quality & Vector filtered
VF Vector filtered

Paracoccus denitrificans SD1 **(Accession: CP002897)**

SEQQC HASH	HQ NODES	N50	MAX	TOTAL	CONTIGS	READS USED (3709241)	PERCENTAGE
37	3031	1521	40897	2944856	2717	3178904	85.7022771
39	2920	1503	40897	2873050	2647	3142134	84.71096917
41	2996	1429	40897	2879215	2760	3142575	84.7228584
43	2954	1402	40897	2815256	2726	3109219	83.82359086
45	3047	1306	40897	2800927	2828	3102000	83.62896884
47	3139	1259	40897	2785848	2931	3093376	83.39646844
49	3245	1187	40897	2776891	3065	3089377	83.28865663
51	3164	1163	40897	2677954	2995	3037777	81.89753645
53	3272	1096	21194	2653194	3126	3028805	81.65565408
55	3447	1005	21194	2606899	3303	2999893	80.87619543
57	3682	935	21194	2569027	3519	2969091	80.04578295
59	3500	903	14857	2395757	3358	2864508	77.22625734
61	3784	795	14593	2310819	3627	2808098	75.70546104
63	4147	670	14593	2154545	3981	2700518	72.80513722

SEQQC HASH	HQ_VF NODES	N50	MAX	TOTAL	CONTIGS	READS USED (2871161)	PERCENTAGE
37	4572	809	7881	2673869	3975	2503969	87.21102718
39	4493	797	6869	2617177	3935	2476904	86.26837715
41	4634	767	6886	2613119	4085	2475590	86.22261169
43	4564	739	6869	2536479	4065	2447420	85.24147549
45	4735	707	6797	2526200	4209	2441920	85.04991535
47	4623	691	6797	2444563	4158	2409572	83.9232631
49	4808	669	5763	2430598	4330	2401635	83.6468244
51	4613	652	5761	2319177	4202	2352630	81.94002357
53	4796	620	5759	2293081	4390	2341881	81.5656454
55	4999	578	5757	2264086	4580	2321285	80.84830492
57	4799	559	4132	2121477	4438	2243935	78.15427278
59	5006	500	4954	2058193	4654	2204159	76.76890986
61	5221	446	3994	1942363	4809	2120926	73.86997803
63	4828	399	3273	1665182	4533	1946909	67.80911973

SEQQC HASH	VF NODES	N50	MAX	TOTAL	CONTIGS	READS USED (3059168)	PERCENTAGE
37	4381	835	7881	2646212	3821	2603324	85.09908576
39	4487	815	6869	2651917	3918	2595682	84.84927928
41	4449	789	6886	2574473	3913	2556032	83.55317524
43	4577	761	5769	2566236	4039	2546150	83.23014624
45	4725	724	6797	2553254	4194	2534810	82.85945721
47	4601	705	6797	2466135	4143	2495443	81.57260405
49	4803	675	5763	2448148	4307	2475043	80.90575607
51	4597	665	5761	2336800	4182	2419836	79.10111507
53	4775	627	5759	2311476	4367	2403607	78.57061136
55	4992	585	5757	2280485	4550	2372200	77.54395966
57	4795	570	4132	2133913	4414	2286796	74.75222021
59	4996	511	4954	2068918	4624	2239469	73.20516559
61	5195	454	3994	1951858	4781	2153869	70.40701916
63	4828	404	3295	1676459	4527	1974185	64.53339601

FASTX HASH	HQ NODES	N50	MAX	TOTAL	CONTIGS	READS USED (3709241)	PERCENTAGE
37	3031	1521	40897	2944856	2717	3178904	85.7022771
39	2920	1503	40897	2873050	2647	3142134	84.71096917
41	2996	1429	40897	2879215	2760	3142575	84.7228584
43	2954	1402	40897	2815256	2726	3109219	83.82359086
45	3047	1306	40897	2800927	2828	3102000	83.62896884
47	3139	1259	40897	2785848	2931	3093376	83.39646844
49	3245	1187	40897	2776891	3065	3089377	83.28865663
51	3164	1163	40897	2677954	2995	3037777	81.89753645
53	3272	1096	21194	2652194	3126	3028805	81.65565408
55	3447	1005	21194	2606899	3303	2999893	80.87619543
57	3682	935	21194	2569027	3519	2969091	80.04578295
59	3500	903	14857	2395757	3358	2864508	77.22625734
61	3784	795	14593	2310819	3627	2808098	75.70546104
63	4147	670	14593	2154545	3981	2700518	72.80513722

FASTX HASH	HQ_VF NODES	N50	MAX	TOTAL	CONTIGS	READS USED (3706894)	PERCENTAGE
37	3013	1523	40897	2940185	2714	3175823	85.67342363
39	2933	1502	40897	2870471	2661	3139416	84.69128062
41	2992	1428	40897	2879465	2755	3140280	84.71458855
43	3108	1344	40897	2877791	2868	3136961	84.62505267
45	3054	1303	40897	2801904	2835	3100897	83.6521627
47	3118	1259	40897	2785035	2930	3094034	83.46702118
49	3247	1187	40897	2777501	3068	3088510	83.31800154
51	3156	1167	40897	2678722	2992	3036436	81.9132136
53	3281	1095	21194	2652246	3126	3026085	81.63397712
55	3455	1005	21194	2606154	3300	2996359	80.83206588
57	3668	933	21194	2567941	3518	2968081	80.06921698
59	3502	904	14857	2395641	3358	2864242	77.26797691
61	3778	795	14593	2309654	3625	2806906	75.72123724
63	4150	670	14593	2154140	3980	2698681	72.80167709

FASTX HASH	VF NODES	N50	MAX	TOTAL	CONTIGS	READS USED (3946398)	PERCENTAGE
37	2916	1580	40897	2903631	2592	3306175	83.77702908
39	2924	1551	40897	2912437	2644	3298593	83.58490451
41	3000	1474	40897	2914827	2751	3286958	83.2900787
43	2955	1427	30843	2840853	2720	3242112	82.15370067
45	3038	1323	40897	2834936	2825	3229917	81.8446847
47	3111	1293	40897	2812821	2901	3208919	81.31260456
49	3237	1225	40897	2797186	3043	3192308	80.89168908
51	3156	1188	40897	2707378	2987	3136028	79.46557848
53	3262	1129	21194	2671739	3103	3108436	78.76640927
55	3421	1036	21194	2628439	3275	3074159	77.89784507
57	3659	957	21194	2586562	3500	3031959	76.82851552
59	3496	919	16291	2416750	3357	2923143	74.07116566
61	3753	805	14585	2322799	3596	2854261	72.32572589
63	4126	682	14587	2167919	3951	2738892	69.40232587

NGStoolkit HASH	HQ NODES	N50	MAX	TOTAL	CONTIGS	READS USED (3726381)	PERCENTAGE
37	3011	1531	30843	2947508	2713	3194015	85.71359182
39	2927	1503	40897	2881892	2652	3158360	84.75676534
41	3010	1437	40897	2887279	2756	3155980	84.69289641
43	3101	1356	40897	2884551	2871	3155094	84.66911999
45	3051	1314	40897	2803930	2824	3114317	83.57484111
47	3138	1255	40897	2790646	2938	3109497	83.44549309
49	3256	1187	40897	2779142	3065	3102002	83.24435961
51	3167	1167	40897	2685382	3002	3052727	81.92203105
53	3283	1097	21194	2656833	3132	3039861	81.57676308
55	3454	1011	21194	2612107	3302	3011482	80.81519308
57	3661	938	21194	2570900	3514	2979240	79.94995681
59	3507	906	14857	2397377	3356	2872391	77.08259032
61	3760	797	14593	2311501	3618	2815462	75.55486141
63	4165	672	14593	2157139	3975	2704455	72.57591212

NORMAL RUN HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (3951522)	PERCENTAGE
37	2942	1579	40897	2896523	2592	3305578	83.653286
39	2935	1534	40897	2911391	2654	3300101	83.51468118
41	3010	1465	40897	2921986	2749	3293040	83.33599054
43	2960	1426	30843	2838758	2721	3244549	82.10884312
45	3028	1349	40897	2837389	2814	3233419	81.8271795
47	3129	1295	40897	2812867	2908	3211141	81.26339674
49	3059	1265	40897	2728229	2872	3159207	79.94911834
51	3153	1187	40897	2707172	2990	3139007	79.43792291
53	3266	1125	21194	2672266	3105	3111022	78.72971478

55	3424	1035	21194	2628027	3277	3075262	77.82474702
57	3651	957	21194	2586465	3500	3035174	76.81025185
59	3513	919	16291	2417538	3359	2925620	74.03780113
61	3752	805	14585	2323044	3595	2855226	72.25636097
63	4132	682	14587	2168430	3952	2740753	69.35942657

FASTX Trim

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (3457858)	PERCENTAGE
37	3135	1399	40897	2881513	2828	2949992	85.31269936
39	3032	1397	40897	2823613	2786	2923405	84.54381296
41	3138	1310	40897	2821705	2888	2919639	84.43490161
43	3074	1277	40897	2763373	2852	2894300	83.70210691
45	3194	1214	40897	2757850	2990	2892771	83.65788879
47	3323	1158	40897	2747751	3125	2887769	83.51323276
49	3237	1119	34563	2653860	3066	2846241	82.31225805
51	3421	1037	19703	2633724	3228	2835422	81.99937649
53	3571	968	19701	2608216	3394	2829519	81.82866387
55	3468	929	16599	2470768	3298	2759846	79.81374597
57	3677	857	14483	2429441	3502	2738984	79.21042449
59	3986	765	14481	2369474	3794	2710552	78.38818135
61	4383	663	11946	2269257	4181	2658851	76.89300717

Prinseq Trim

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (3951522)	PERCENTAGE
37	3022	1478	40897	2897228	2685	3294052	83.36160092
39	3101	1406	30843	2887565	2780	3279893	83.00328329
41	3016	1379	40897	2798453	2749	3234479	81.85400461
43	3118	1285	40897	2786828	2879	3223279	81.57056952
45	3209	1218	24229	2762788	2987	3205098	81.11046832
47	3381	1129	24229	2731085	3156	3180541	80.48901158
49	3285	1060	24229	2594749	3108	3104450	78.56339912
51	3469	973	21194	2541828	3288	3071070	77.71866132
53	3748	860	21194	2453749	3532	3012282	76.23093077
55	4184	730	14756	2317035	3817	2909713	73.63524738
57	4688	567	9173	2046449	4129	2693347	68.15973693
59	4436	387	7915	1446487	3768	2129674	53.89503083
61	2108	208	2873	439523	1594	786455	19.90258437
63	24	302	616	6267	21	44881	1.135790209

SolexaQA Trim

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (3951522)	PERCENTAGE
37	3841	929	14681	2602175	3460	2590777	65.56402824
39	3745	894	11912	2486893	3400	2471114	62.53575205
41	4042	815	11755	2464488	3631	2393297	60.56646021
43	4346	751	11753	2453278	3884	2320140	58.71509763
45	4225	716	11751	2321141	3808	2194128	55.52614916
47	4590	651	11260	2297571	4102	2100884	53.1664508
49	4460	607	11258	2139118	3990	1954205	49.45448868
51	4924	534	6049	2107032	4328	1845662	46.70762304
53	4768	498	5252	1925404	4162	1681232	42.5464416
55	5255	417	3717	1850267	4458	1549637	39.21620581
57	4969	375	3580	1613675	4200	1354396	34.27529949
59	4475	338	2824	1342677	3839	1154553	29.21793172
61	4177	287	2822	1105775	3564	952128	24.09522204
63	3098	248	2868	746442	2715	698984	17.68898161

FASTX clipper (N's only removed)

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (3294457)	PERCENTAGE
37	5184	739	4034	2708167	4396	2720680	82.58356385
39	5294	716	4098	2713243	4514	2710191	82.26518057
41	5187	701	4150	2652730	4508	2682539	81.42583133
43	5316	677	4094	2641270	4615	2671577	81.09309061
45	5169	666	3420	2555674	4541	2629682	79.82140911
47	5261	638	3418	2539735	4654	2609600	79.21183977
49	5082	624	3416	2429105	4549	2553237	77.50099637
51	5155	600	3414	2408880	4667	2541476	77.14400279
53	5292	568	3328	2384448	4806	2522806	76.57729331
55	5120	551	3182	2253158	4696	2445674	74.23602736
57	5252	518	3180	2205749	4864	2407362	73.07310431
59	5398	484	3178	2144844	5007	2361974	71.69539624
61	5019	455	3176	1924037	4731	2228734	67.65102716
63	5106	401	2666	1760546	4848	2107843	63.98149983

REQUIRED	(cutoff = 30)	(cutoff = 35)	
Read1	Read2	Read1	Read2
65 (minus 7)	63 (minus 9)	50 (minus 22)	48 (minus 24)

FASTX clipper (N's adap's removed)

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (3282666)	PERCENTAGE
37	5201	730	3405	2700978	4413	2712929	82.64407649
39	5295	709	3403	2707452	4530	2703861	82.36783761
41	5220	695	3401	2647618	4526	2671736	81.3892123
43	5323	674	3399	2632561	4622	2660854	81.05771346
45	5198	661	3397	2548159	4552	2617578	79.73939475
47	5258	635	3395	2532191	4653	2601804	79.25887069
49	5096	621	3332	2422777	4556	2543448	77.48116927
51	5172	596	3330	2399475	4663	2527622	76.99906113
53	5293	566	3328	2376113	4803	2509212	76.43823648
55	5119	549	3181	2246543	4696	2436481	74.2226288
57	5262	517	3179	2199087	4863	2397079	73.02232393
59	5398	482	3177	2137701	5005	2351626	71.63768717
61	5011	453	3175	1916668	4724	2216736	67.52852712
63	5107	400	2660	1755406	4839	2099754	63.96489926

***Mycobacterium tuberculosis* NA-A0008 (Accession: ALYG00000000)**

SEQQC HQ

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (13787485)	PERCENTAGE
37	2729	2181	16781	3238077	2306	12959870	93.99734614
39	2663	2180	12147	3209818	2272	12852248	93.21676869
41	2542	2120	13230	3245317	2285	12874790	93.38026478
43	2580	2035	12143	3246849	2379	12852140	93.21598537
45	2560	1956	12801	3206171	2386	12768333	92.60813702
47	2655	1856	12052	3178914	2485	12702597	92.13135681
49	2730	1750	10675	3140054	2566	12615897	91.50252566
51	2859	1636	10654	3100495	2689	12524630	90.84057027
53	2969	1514	10656	3056611	2794	12423971	90.11049513
55	3065	1412	10163	2997286	2907	12272174	89.00951841
57	3187	1304	10163	2927173	3048	12083444	87.64066833
59	3424	1195	10163	2881812	3281	11937320	86.58083762
61	3668	1047	10165	2796760	3522	11690855	84.79323822
63	3890	941	8896	2696262	3764	11407915	82.7410873

SEQQC HQ_VF

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (11627720)	PERCENTAGE
37	4125	1137	6242	2986069	3506	10613757	91.27977798
39	4111	1126	6014	2975950	3546	10576249	90.95720399
41	4004	1085	6012	2953270	3586	10553511	90.76165405
43	4083	1029	5987	2946503	3715	10527294	90.53618422
45	4176	996	6421	2921340	3815	10482289	90.149135
47	4210	939	6421	2871723	3886	10400731	89.44772492
49	4296	904	6421	2825970	3986	10315444	88.71424493
51	4352	871	6421	2763878	4047	10212979	87.83303176
53	4494	826	5228	2726452	4191	10089558	86.77159409
55	4570	790	3947	2651823	4244	9937787	85.4663425
57	4725	731	3928	2590656	4423	9791109	84.20489141
59	4872	688	3716	2525816	4581	9605889	82.6119738
61	4909	636	3523	2407941	4660	9336661	80.29657577
63	5071	579	3503	2309829	4877	9103668	78.29280375

SEQQC VF

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (13297453)	PERCENTAGE
37	3976	1240	7295	2966933	3282	11644082	87.56625799
39	3883	1252	7293	2985927	3297	11588195	87.1459745
41	3814	1202	6099	2952012	3377	11497555	86.4643402
43	3867	1112	6097	2939248	3494	11408080	85.79146698
45	3962	1067	6330	2908149	3606	11294400	84.93656642
47	4012	1020	6324	2865605	3687	11157272	83.90533134
49	4081	972	6324	2814891	3769	10991111	82.65576122
51	4148	931	6331	2758242	3823	10785418	81.10890108
53	4277	877	6331	2708294	3977	10613681	79.81739811
55	4436	826	5231	2658434	4118	10419372	78.35614835
57	4522	768	5231	2569590	4257	10148373	76.31817161
59	4682	712	3716	2501920	4416	9889206	74.36917431
61	4817	659	3528	2414827	4580	9644575	72.52949117
63	4982	601	3503	2319125	4786	9331733	70.17684514

FASTX HQ

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (13787485)	PERCENTAGE
37	2729	2181	16781	3238077	2306	12959870	93.99734614
39	2663	2180	12147	3209818	2272	12852248	93.21676869
41	2542	2120	13230	3245317	2285	12874790	93.38026478
43	2580	2035	12143	3246849	2379	12852140	93.21598537
45	2560	1956	12801	3206171	2386	12768333	92.60813702
47	2655	1856	12052	3178914	2485	12702597	92.13135681
49	2730	1750	10675	3140054	2566	12615897	91.50252566
51	2859	1636	10654	3100495	2689	12524630	90.84057027
53	2969	1514	10656	3056611	2794	12423971	90.11049513
55	3065	1412	10163	2997286	2907	12272174	89.00951841
57	3187	1304	10163	2927173	3048	12083444	87.64066833
59	3424	1195	10163	2881812	3281	11937320	86.58083762
61	3668	1047	10165	2796760	3522	11690855	84.79323822
63	3890	941	8896	2696262	3764	11407915	82.7410873

FASTX HQ_VF

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (13773306)	PERCENTAGE
37	2732	2151	16781	3239003	2307	12949124	94.01609171
39	2668	2168	12802	3228759	2274	12875114	93.47874795
41	2544	2153	13230	3246044	2281	12863837	93.39687218
43	2572	2017	12143	3240629	2368	12836517	93.19851748
45	2563	1950	12801	3204272	2388	12765891	92.68574299
47	2636	1859	12052	3188016	2471	12715362	92.31888117
49	2748	1741	10675	3158037	2587	12644033	91.80100261
51	2846	1624	10654	3104060	2682	12529062	90.96626474
53	2969	1517	10656	3058374	2798	12424186	90.20482083
55	3064	1412	10163	2998828	2907	12276581	89.13314639
57	3193	1304	10163	2927123	3049	12075637	87.67420836
59	3431	1195	10163	2881068	3283	11931549	86.62806882
61	3662	1047	10165	2796310	3520	11688664	84.86462146
63	3888	941	8896	2695678	3763	11403686	82.79556121

FASTX VF

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (15810332)	PERCENTAGE
37	2723	2378	16781	3269493	2213	14425763	91.24263172
39	2591	2449	12695	3266331	2147	14311119	90.51751095
41	2467	2433	16913	3295095	2163	14250453	90.13379985
43	2433	2283	16911	3275218	2203	14124738	89.33865525
45	2442	2157	12808	3246249	2259	13979760	88.42167261
47	2518	2051	12808	3199994	2341	13791037	87.22800381
49	2572	1970	12803	3169791	2389	13611399	86.09179744
51	2678	1818	12801	3128267	2504	13411708	84.82875628
53	2798	1654	12275	3076688	2637	13197279	83.4724976
55	2933	1538	10658	3012917	2774	12951314	81.91677442
57	3106	1380	10161	2969105	2968	12732755	80.53439359
59	3268	1276	10161	2881905	3124	12413726	78.51654222
61	3472	1111	10161	2794028	3338	12098577	76.52323177
63	3696	997	10155	2688230	3554	11738208	74.24390582

NGStoolkit HQ

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (13940064)	PERCENTAGE
37	2702	2228	16781	3256680	2263	13105772	94.01514943
39	2649	2193	12802	3225513	2250	13002305	93.27292185
41	2558	2173	13230	3254840	2282	13016642	93.37576929
43	2542	2069	12143	3231000	2320	12944212	92.85618775
45	2552	1958	12704	3217867	2379	12918349	92.67065775
47	2596	1871	12801	3177809	2435	12824883	92.00017303
49	2699	1794	10652	3153136	2545	12757408	91.51613651
51	2823	1641	10654	3108670	2659	12645831	90.71573129
53	2945	1523	10656	3064593	2784	12535732	89.9259286
55	3062	1417	10658	3003097	2906	12366903	88.71482226
57	3156	1327	10163	2929363	3023	12173055	87.32424041
59	3402	1205	10163	2885068	3273	12007165	86.13421717
61	3642	1064	10165	2799604	3499	11750901	84.29588989
63	3867	946	8896	2700682	3745	11467550	82.26325216

NORMAL RUN

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (15839552)	PERCENTAGE
37	2741	2342	16781	3268643	2222	14436774	91.14382781
39	2579	2459	12695	3274532	2140	14342069	90.54592579
41	2447	2391	16913	3267589	2137	14213193	89.73229167
43	2427	2278	16911	3257246	2189	14107810	89.06697614
45	2458	2140	13239	3240506	2269	13981206	88.26768585
47	2529	2026	12808	3212738	2357	13835088	87.34519764
49	2610	1949	12803	3169738	2421	13619236	85.98245708
51	2682	1840	12801	3130854	2506	13425053	84.75651963
53	2811	1656	12275	3074217	2642	13209600	83.39629808
55	2936	1530	10658	3012983	2778	12960802	81.82555921
57	3101	1379	10161	2969902	2967	12748457	80.48495942
59	3271	1278	10161	2883425	3125	12423280	78.43201626
61	3471	1113	10161	2794055	3339	12106446	76.43174504
63	3689	997	10155	2688282	3556	11747636	74.16646632

FASTX Trim

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (15839552)	PERCENTAGE
27	3215	2360	13616	3256087	2237	14652946	92.50858863
29	2915	2508	15212	3314692	2130	14603170	92.19433731
31	2640	2585	15137	3359623	2055	14638187	92.41541049
33	2595	2474	12896	3360231	2112	14559760	91.92027653
35	2631	2242	12362	3343968	2180	14446580	91.20573612
37	2677	2090	12372	3320908	2263	14337502	90.51709291
39	2729	1972	12149	3272674	2340	14191159	89.5931842
41	2804	1799	12147	3239439	2512	14103247	89.0381685
43	2895	1597	10359	3161058	2674	13926148	87.92008764
45	3172	1413	10357	3099523	2960	13790842	87.06585893
47	3390	1216	10161	2963644	3204	13525720	85.39206159
49	3970	981	10155	2843019	3737	13267073	83.75914294
51	4656	703	9780	2516306	4296	12501238	78.92418927
53	3954	203	1645	770607	2348	4250849	26.83692695

Prinseq Trim

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (15839552)	PERCENTAGE
27	3215	2360	13616	3256087	2237	14652946	92.50858863
29	2915	2508	15212	3314692	2130	14603170	92.19433731
31	2633	2595	15137	3359963	2052	14638235	92.41571353
33	2596	2474	12896	3360227	2113	14559760	91.92027653
35	2631	2242	12362	3343759	2180	14446600	91.20586239
37	2673	2092	12372	3322269	2261	14337502	90.51709291
39	2718	1975	12149	3273017	2334	14191241	89.59370189
41	2797	1800	12147	3240809	2509	14103248	89.03817482
43	2895	1597	10359	3161058	2674	13926148	87.92008764
45	3173	1410	10357	3099934	2961	13791096	87.06746251
47	3390	1216	10161	2963644	3204	13525720	85.39206159
49	3970	981	10155	2843019	3737	13267073	83.75914294
51	4650	703	9780	2516340	4292	12501640	78.92672722
53	3954	203	1645	770607	2348	4250849	26.83692695

SolexaQA Trim

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (15839552)	PERCENTAGE
27	3719	1507	10615	3051730	2866	11329969	71.52960513
29	3643	1434	10615	3046991	2944	11038754	69.6910746
31	3546	1339	9957	3005161	3018	10638208	67.16230358
33	3660	1241	9226	3000802	3160	10283792	64.92476555
35	3740	1170	8643	2957343	3273	9851570	62.19601413
37	3853	1073	8619	2909800	3433	9362203	59.10648862
39	3993	1007	8617	2847113	3588	8925513	56.34952933
41	4141	927	6601	2788950	3796	8505453	53.69756039
43	4354	836	5948	2720940	4017	8105267	51.17106216
45	4561	774	5723	2647593	4218	7701407	48.62136884
47	4789	704	5223	2559552	4383	7207822	45.50521378
49	5057	634	5223	2482717	4633	6794264	42.89429398
51	5351	568	4582	2406151	4858	6292552	39.72683066
53	5407	521	4566	2265611	4915	5765166	36.39727942

FASTX clipper (N's only removed)

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (13798166)	PERCENTAGE
37	4901	935	4730	2910657	3925	11560137	83.78024297

REQUIRED	(cutoff = 30)	(cutoff = 35)	
Read1	Read2	Read1	Read2
54 (minus18)	54 (minus18)	40 (minus 32)	37 (minus 35)

39	4810	951	4540	2901001	3936	11482906	83.22052366
41	4706	933	4494	2895389	4010	11459426	83.05035611
43	4666	902	4494	2855977	4049	11320260	82.0417728
45	4679	887	5120	2852986	4144	11303644	81.921351
47	4682	857	5118	2803338	4179	11176193	80.99767027
49	4734	837	4809	2773863	4248	11021863	79.87918829
51	4772	803	4809	2710985	4290	10838231	78.54834476
53	4813	763	4451	2676642	4400	10695291	77.51240998
55	4874	727	4449	2594494	4432	10439491	75.65854042
57	4914	690	4446	2534764	4529	10233162	74.16320401
59	4996	654	4119	2460061	4665	9972151	72.27156855
61	5116	602	2779	2374494	4824	9706392	70.34552273
63	5218	561	2790	2290285	4991	9448336	68.47530317

FASTX clipper (N's adap's removed)

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (13766430)	PERCENTAGE
37	4975	917	4301	2902527	3963	11461625	83.25778724
39	4859	937	4438	2899457	3958	11409021	82.87566929
41	4709	922	4493	2896839	3998	11399017	82.80299976
43	4663	901	4493	2860735	4043	11302224	82.09989082
45	4696	879	5119	2851932	4170	11269571	81.86269788
47	4756	847	4813	2815037	4239	11138071	80.90747565
49	4750	831	4809	2767085	4255	11015087	80.01411404
51	4744	798	4809	2702171	4274	10808002	78.50983879
53	4759	763	4450	2644502	4338	10638494	77.27852464
55	4838	730	4448	2585273	4416	10422778	75.71155339
57	4901	687	4446	2528563	4534	10205978	74.13670792
59	4987	651	4118	2454662	4659	9954169	72.30755541
61	5110	602	2778	2370080	4827	9698515	70.45047264
63	5226	558	2790	2284219	4997	9421728	68.43987875

***Helicobacter pylori 908* (Accession: CP002184)**

SEQQC HQ

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (2920392)	PERCENTAGE
27	599	7502	28748	1484196	359	2786683	95.4215393
29	457	10678	52999	1510341	268	2770625	94.87168161
31	402	12063	37045	1541862	245	2794615	95.69314667
33	431	9925	32162	1539067	269	2774662	95.00991648
35	459	8599	26069	1524636	304	2750141	94.17026892
37	500	6752	23457	1494421	358	2700971	92.48659084
39	554	5235	18239	1460403	420	2675807	91.62492569
41	689	3718	14481	1426352	549	2645060	90.57208758
43	842	2637	13603	1336968	708	2580847	88.37330742
45	1252	1623	11748	1273840	1077	2531973	86.69976496
47	1756	1029	8351	1145981	1461	2390439	81.85336078
49	2455	584	5082	926578	1727	2100298	71.91835891
51	1831	324	3078	461295	1081	1337076	45.78412761
53	80	238	1002	15864	41	161397	5.5265526

SEQQC HQ_VF

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (2675157)	PERCENTAGE
27	816	4150	18216	1453280	562	2538350	94.88601977
29	758	4374	21047	1473804	538	2526058	94.42653272
31	723	4374	16438	1484064	547	2531181	94.6180355
33	763	3936	16609	1466045	595	2507653	93.73853572
35	786	3320	14552	1407890	626	2456676	91.83296532
37	890	2718	10076	1398838	743	2429728	90.82562257
39	1004	2373	8779	1353087	849	2394759	89.51844695
41	1234	1699	7814	1341682	1066	2385606	89.17629881
43	1503	1276	6918	1273855	1311	2337882	87.39232875
45	1908	931	5546	1169567	1613	2246848	83.98938829
47	2398	638	5371	1040068	1866	2081593	77.81199384
49	2764	408	3610	817328	1838	1776750	66.41666265
51	1731	258	2365	383690	1003	1075713	40.21121003
53	68	194	568	11528	33	106973	3.998755961

SEQQC VF

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (2762088)	PERCENTAGE
27	837	4141	18216	1462291	569	2598381	94.07307081
29	760	4428	21047	1484833	534	2583339	93.5284828
31	739	4374	21047	1494666	555	2587276	93.6710199
33	753	4010	16609	1444083	578	2534264	91.75174723
35	802	3315	14552	1409385	630	2495075	90.33292929
37	895	2822	10076	1407549	736	2471863	89.49255056
39	1013	2369	8779	1361559	846	2434802	88.15077579
41	1258	1719	6802	1348993	1059	2418819	87.57211935
43	1512	1297	6800	1285393	1311	2374478	85.96677586
45	1884	943	5546	1174030	1599	2276133	82.40624484
47	2401	641	5371	1048691	1871	2111268	76.43739084
49	2766	414	3610	824736	1852	1803687	65.3015762
51	1750	260	2365	388606	1020	1096604	39.70199356
53	68	194	568	11534	33	107870	3.905378829

FASTX HQ

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (2920392)	PERCENTAGE
27	599	7502	28748	1484196	359	2786683	95.4215393
29	457	10678	52999	1510341	268	2770625	94.87168161
31	402	12063	37045	1541862	245	2794615	95.69314667
33	431	9925	32162	1539067	269	2774662	95.00991648
35	459	8599	26069	1524636	304	2750141	94.17026892
37	500	6752	23457	1494421	358	2700971	92.48659084
39	554	5235	18239	1460403	420	2675807	91.62492569
41	689	3718	14481	1426352	549	2645060	90.57208758
43	842	2637	13603	1336968	708	2580847	88.37330742
45	1252	1623	11748	1273840	1077	2531973	86.69976496
47	1756	1029	8351	1145981	1461	2390439	81.85336078
49	2455	584	5082	926578	1727	2100298	71.91835891
51	1831	324	3078	461295	1081	1337076	45.78412761
53	80	238	1002	15864	41	161397	5.5265526

FASTX HQ_VF

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (2897981)	PERCENTAGE
27	596	7534	28748	1479841	357	2763131	95.34676038
29	467	10200	52999	1506064	277	2753808	95.02505365
31	407	11987	40952	1545486	243	2777746	95.85107701
33	420	10184	32162	1536162	265	2759210	95.21145929
35	456	8029	26069	1518744	303	2731294	94.24816795
37	501	6556	23457	1488950	358	2686777	92.71202951
39	563	5271	26598	1457190	418	2659594	91.77403164
41	690	3717	13713	1423124	545	2632386	90.83517111
43	849	2599	14324	1335846	711	2569921	88.67970494
45	1253	1623	11748	1271277	1075	2521374	87.00450417
47	1757	1027	8351	1143977	1461	2383572	82.24940053
49	2446	584	5082	924257	1725	2094936	72.28950086
51	1826	324	3078	460216	1080	1333191	46.00413184
53	81	238	1002	15980	41	161090	5.55869759

FASTX VF

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (2991111)	PERCENTAGE
27	607	7575	31492	1484644	359	2825176	94.45239578
29	491	9482	45142	1510031	287	2810438	93.95966917
31	424	11730	37047	1547416	249	2831298	94.6570689
33	421	10748	31769	1543953	261	2811976	94.01108819
35	472	8813	26069	1532416	304	2785122	93.11329469
37	512	6760	23457	1503898	357	2735906	91.46788601
39	570	5292	18239	1461190	413	2695576	90.11955758
41	685	3801	14481	1430464	537	2669651	89.25282278
43	860	2667	14324	1342592	708	2603909	87.05491037
45	1246	1672	11749	1280747	1070	2556116	85.45707598
47	1754	1031	8392	1151131	1457	2412449	80.6539443
49	2445	590	5082	933589	1744	2128427	71.15840903
51	1837	332	3078	465796	1090	1355677	45.32352694
53	83	238	1002	16254	41	162376	5.428618329

NGStoolkit HQ

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (2920352)	PERCENTAGE
27	599	7502	28748	1484196	359	2786679	95.42270932
29	454	10754	57544	1527201	268	2769665	94.84010832
31	398	12695	37045	1547586	241	2793029	95.64014886

33	433	9777	32162	1534954	272	2771704	94.90992867
35	458	8813	26069	1526436	302	2751100	94.20439728
37	505	6556	23457	1491396	360	2699271	92.42964547
39	565	5271	26598	1458480	420	2672023	91.49660726
41	693	3709	13713	1425864	551	2644767	90.56329511
43	840	2637	14324	1336597	708	2580314	88.35626664
45	1249	1623	11748	1273797	1076	2532101	86.70533552
47	1756	1029	8351	1145983	1461	2390440	81.85451617
49	2458	584	5082	926627	1727	2099830	71.9033185
51	1831	324	3078	461296	1081	1337076	45.78475471
53	80	238	1002	15864	41	161397	5.526628297

NORMAL RUN

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (3015238)	PERCENTAGE
27	610	7584	28748	1486341	358	2847574	94.43944392
29	474	10200	52999	1516831	275	2832581	93.9422029
31	409	12063	37047	1546563	244	2848099	94.45685548
33	439	10300	32162	1548638	266	2829530	93.84101686
35	469	8029	26069	1530662	302	2797814	92.7891596
37	515	6760	20649	1504274	359	2747407	91.1174176
39	567	5292	18924	1462394	412	2709016	89.84418477
41	691	3885	13713	1431938	537	2680714	88.9055524
43	844	2685	13603	1345899	702	2617358	86.8043584
45	1238	1677	11749	1281170	1068	2565847	85.09600237
47	1756	1037	8392	1154844	1461	2421626	80.31293052
49	2451	590	5082	935633	1745	2134872	70.80276913
51	1845	329	3078	466983	1091	1359298	45.08095215
53	78	255	1002	15617	41	162695	5.395759804

FASTX Trim

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED

Not required

Read1 quality	Read2 quality
36	35

Not required

Read1 quality	Read2 quality
36	35

Prinseq Trim

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED

SolexaQA Trim

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED

FASTX clipper (N's only removed)

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (2071499)	PERCENTAGE
27	4127	407	2271	1180726	2458	1632339	78.79989322
29	4159	394	2269	1181912	2548	1629378	78.65695325
31	4233	374	2267	1171869	2592	1611272	77.78290021
33	4080	364	2265	1114459	2521	1567805	75.68456466
35	4091	347	1644	1090301	2522	1546952	74.67790233
37	4123	329	1642	1062468	2530	1504608	72.63377873
39	4101	314	1469	1029657	2529	1477009	71.30145851
41	4107	298	1467	992326	2521	1438211	69.42851529
43	3795	280	1282	889843	2350	1367092	65.99530099
45	3762	251	1113	818777	2213	1282318	61.9029022
47	3169	229	1104	652908	1808	1125813	54.34774528
49	2563	194	1102	474505	1315	897614	43.33161638
51	1171	168	697	194546	520	491283	23.71630399
53	51	175	390	7056	24	64180	3.098239487

FASTX clipper (N's adap's removed)

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (2039702)	PERCENTAGE
27	4202	394	2269	1160641	2449	1583869	77.65198053
29	4238	373	2267	1157561	2535	1581634	77.5424057
31	4258	365	2265	1150852	2572	1565604	76.75650659
33	4084	354	2263	1092890	2499	1529776	74.99997549
35	4092	339	1642	1073217	2498	1507263	73.89623582
37	4103	322	1640	1043670	2501	1472161	72.17529816
39	4074	310	1451	1011015	2502	1445713	70.87863815
41	4079	293	1290	975414	2492	1410173	69.13622676
43	3757	276	1261	873513	2315	1342120	65.79980801
45	3716	250	1113	804715	2187	1259060	61.72764453
47	3134	227	1104	641282	1779	1104052	54.12810303
49	2523	193	1102	465551	1293	881367	43.21057684
51	1145	168	697	190175	512	481444	23.60364406
53	51	174	390	6999	23	61521	3.016175892

Escherichia coli NA114 **(Accession: NC_017644)**

SEQQC HQ

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (21500807)	PERCENTAGE
27	1793	14761	72096	4749076	716	20579050	95.71291905
29	1551	19733	107129	4810558	587	20359446	94.69154344
31	1104	34618	141903	4975046	380	20425831	95.00029929
33	1234	31439	151904	4904437	464	20268577	94.26891279
35	1090	33063	122744	4920074	428	20075756	93.37210459
37	1087	29220	89683	4914090	476	19685807	91.55845639
39	1072	27084	81660	4907158	506	19610575	91.20855324
41	1039	25661	81657	4903030	519	19451398	90.46822289
43	1063	20768	79559	4887951	564	19348205	89.98827346
45	1090	17927	79517	4860437	646	19154877	89.08910721
47	1293	12929	63942	4825065	808	18552165	86.2859008
49	1636	7378	27059	4706564	1180	18297235	85.10022438
51	3836	2224	20008	4452015	3245	17729662	82.46044904
53	5994	192	13386	1145570	3097	5157193	23.98604387

SEQQC HQ_VF

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (21497116)	PERCENTAGE
27	2523	5929	38416	4696207	1401	20512159	95.41819005
29	2438	5839	29767	4742213	1422	20261304	94.25126608
31	2404	5475	28781	4784519	1456	20294283	94.40467735
33	2358	5265	28779	4776592	1518	20127269	93.62776384
35	2398	4947	28793	4786166	1627	19953719	92.82044624
37	2406	4687	22609	4789788	1707	19586603	91.11270089
39	2407	4326	20157	4783234	1797	19535678	90.87580864
41	2527	4064	22532	4777981	1904	19377905	90.14188229
43	2591	3676	22530	4745793	2021	19249750	89.54573255
45	2764	3300	16156	4715287	2227	19042934	88.58366862
47	3078	2770	15974	4681365	2544	18518120	86.14234579
49	3768	2142	12831	4546589	3187	18202193	84.67272075
51	7105	941	7205	4295417	6036	17491899	81.36858451
53	6242	181	4105	1132520	3146	5128316	23.85583257

SEQQC VF

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (21497321)	PERCENTAGE
27	2579	5643	33861	4684147	1454	20212833	94.02489268
29	2464	6020	41007	4740324	1417	19923713	92.67998091
31	2381	5906	41005	4766830	1425	19881844	92.48521711
33	2350	5607	41003	4791746	1495	19730686	91.78206903
35	2337	5159	41001	4786639	1566	19482317	90.62672042
37	2350	4859	22609	4784172	1660	19035198	88.546838
39	2405	4540	22534	4776609	1752	18939037	88.0995218
41	2472	4173	22532	4773960	1861	18741737	87.18173302
43	2569	3876	22530	4750956	1983	18606036	86.55048692
45	2731	3366	15394	4726540	2201	18400819	85.59587029
47	2997	2791	15303	4658173	2488	17775708	82.68801494
49	3874	2102	11878	4577947	3312	17562615	81.69676119
51	7172	905	7205	4232298	6106	16661479	77.50490863
53	5621	176	4105	999186	2726	4374355	20.34837271

FASTX HQ

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (43607176)	PERCENTAGE
27	2077	10793	46822	4690334	938	41688633	95.60039614
29	1725	13753	58111	4746275	739	41481625	95.12568528
31	1325	26832	118529	4887677	500	41644491	95.49916968
33	934	36469	153992	5018761	371	41730239	95.69580704
35	838	41282	136320	5056522	335	41502920	95.17451898
37	722	39131	139561	5037632	315	41118707	94.29344152
39	669	40969	152955	5002074	312	40974891	93.96364259
41	660	40963	172080	5011194	326	40789452	93.53839377
43	660	39244	110289	5038716	331	40571710	93.0390677
45	860	31017	110506	4930880	453	40214295	92.21944342
47	886	27974	110256	4941100	494	39519901	90.62705872
49	998	19492	68897	4898353	600	39238969	89.9828253
51	1467	9259	40292	4792253	1026	38754737	88.87238421
53	8699	685	17051	3854770	6790	33398367	76.58915358

FASTX HQ_VF

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (43283560)	PERCENTAGE
27	2095	10596	46936	4683509	954	41395713	95.63842022
29	1721	13753	58111	4753904	729	41254569	95.31232875
31	1400	23391	89695	4846264	535	41396701	95.64070284
33	959	37230	112960	5017635	359	41493676	95.86474865
35	833	41728	114776	5028405	329	41283756	95.37976081
37	723	41427	139561	5019845	308	40936740	94.57803378
39	677	42179	170876	5022776	317	40853934	94.38672327
41	667	39203	149865	5023266	324	40612049	93.82788523
43	665	38489	125605	5043111	332	40442456	93.43606672
45	925	28791	84436	4926086	486	40078255	92.59463639
47	940	24968	110256	4915980	521	39430825	91.09884908
49	1067	18011	68897	4880515	636	39148674	90.44698264
51	1466	9247	40292	4791219	1026	38681354	89.36731175
53	8711	685	17051	3849733	6783	33300417	76.93548544

FASTX VF

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (45084047)	PERCENTAGE
27	2272	9018	44641	4683883	1058	42751900	94.82711257
29	1859	11609	56135	4736010	846	42471646	94.20548692
31	1540	16448	59060	4816041	645	42502279	94.27343335
33	1022	31426	93141	4997185	402	42557283	94.39543659
35	940	31474	92873	4977764	385	42256258	93.72773922
37	809	37651	99638	5035806	337	41846339	92.81850629
39	715	38412	164684	5001790	338	41637037	92.35425782
41	686	42120	175774	5039957	310	41319968	91.65097357
43	679	37912	125605	5030108	340	41104643	91.1733656
45	863	31380	110506	4934556	447	40663985	90.19595113
47	915	29277	110262	4934129	485	39935980	88.58117817
49	1018	19492	68897	4888790	604	39655442	87.95892259
51	1422	9459	40292	4796584	999	39161749	86.86387227
53	8543	717	17051	3880887	6638	33843674	75.06795918

NGStoolkit HQ

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (43607077)	PERCENTAGE
27	2088	10743	53828	4687812	942	41684661	95.59150456
29	1718	13910	58111	4760374	735	41523205	95.22125273
31	1372	27458	95356	4862021	527	41633178	95.47344345
33	929	36469	123662	5051867	363	41761970	95.76879001
35	827	39624	130967	5026613	337	41510609	95.19236751
37	705	39497	249390	5074611	308	41127793	94.31449166
39	681	43771	147651	5029873	313	40989227	93.99673131
41	656	39430	149865	5038757	319	40775821	93.50734744
43	689	38257	135764	5089454	333	40545200	92.97848604
45	885	30281	110506	4937265	458	40198646	92.18376641
47	895	27974	110256	4937654	503	39531403	90.65364092
49	1006	19492	68897	4901975	602	39230717	89.96410606
51	1459	9259	40292	4792161	1024	38762614	88.89064956
53	8699	685	17051	3854926	6790	33399893	76.59282689

NORMAL RUN

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (45440998)	PERCENTAGE
27	2292	8585	44641	4677152	1085	43036559	94.70865715
29	1887	11669	56135	4736002	851	42709308	93.98849031
31	1490	19206	87210	4828542	612	42775351	94.13382822
33	1046	29979	102042	5026097	411	42810331	94.21080717
35	950	28765	102510	4979062	393	42420060	93.35195499
37	812	36847	104347	5027252	350	42006443	92.4417263
39	705	37916	218214	4995119	326	41821170	92.03400418
41	678	41059	175866	5011659	325	41492440	91.31058257
43	661	38395	125605	5035428	334	41266721	90.81385272
45	878	30307	110506	4937812	455	40766794	89.71368543
47	918	29699	110262	4936958	493	40032948	88.0987429
49	1034	19307	68897	4891434	611	39720580	87.41132842
51	1438	9537	40292	4797861	998	39225636	86.32212699
53	8556	717	17051	3886490	6642	33928981	74.66601196

FASTX Trim

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED	Not required Read1 quality	Read2 quality	Not required Read1 quality	Read2 quality
Prinseq Trim										
HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED	34	33	34	33

SolexaQA Trim

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED

FASTX clipper (N's only removed)

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (37060498)	PERCENTAGE
27	8176	895	5562	4245275	5739	32385792	87.38628391
29	8254	859	4910	4216955	5930	31855509	85.95542618
31	8386	812	4743	4205447	6162	31564399	85.16992675
33	8594	775	4464	4206493	6402	31413149	84.76181027
35	8680	749	4001	4208033	6605	31328244	84.53271189
37	8748	725	4021	4192037	6755	31064360	83.82067613
39	8719	711	4019	4165763	6853	31012729	83.68136068
41	8735	699	4017	4152642	6953	30871401	83.30001664
43	8698	689	4005	4105589	6993	30615377	82.60918944
45	8754	671	3991	4091133	7132	30460000	82.18993711
47	8672	658	3425	4025406	7173	29841720	80.52163789
49	8739	637	3423	3991468	7325	29627474	79.94353988
51	9005	576	3301	3838553	7621	28955650	78.13076338
53	11119	261	1492	2575984	7476	20184120	54.46262487

FASTX clipper (N's adap's removed)

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (36872384)	PERCENTAGE
27	8440	857	5582	4224726	5901	32147204	87.18504342
29	8552	808	4855	4197748	6091	31580618	85.64843
31	8608	786	4742	4193672	6306	31417782	85.20680952
33	8740	751	4463	4175509	6450	31128862	84.42324207
35	8800	730	3470	4185272	6662	31189893	84.58876161
37	8784	710	4004	4170773	6795	30950040	83.93826664
39	8748	704	4012	4148244	6865	30850038	83.66705554
41	8793	689	4010	4130422	6962	30685138	83.21983737
43	8711	684	3998	4090234	6995	30488146	82.68558388
45	8772	665	3994	4071128	7127	30250365	82.0407083
47	8679	654	3424	4004493	7171	29685257	80.5081033
49	8707	634	3422	3968644	7314	29477504	79.94466536
51	8981	574	3299	3821286	7610	28828173	78.1836428
53	11081	260	1492	2555503	7437	20017764	54.28931311

Vibrio parahaemolyticus PCV08-7 **(Accession: AOCL00000000)**

SEQQC HQ

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (4099136)	PERCENTAGE
51	323	294704	941090	5225650	90	4065146	99.17080087
53	317	293611	717264	5214709	102	4057948	98.99520289
55	293	358836	713199	5215187	91	4046927	98.72634136
57	292	593300	713199	5216704	88	4037171	98.48833998
59	311	517174	1051802	5214016	103	4026421	98.2260896
61	272	517178	1018596	5220991	93	4017672	98.01265437
63	279	264374	920037	5221419	101	4004992	97.70332089
65	268	262022	704110	5214884	102	3996329	97.49198368
67	260	262022	920043	5222658	98	3982529	97.15532737
69	245	262022	704099	5213485	97	3972129	96.90161536
71	236	262022	704099	5217294	96	3962511	96.66698055
73	259	256751	508753	5211758	110	3947801	96.30812444
75	307	131450	458341	5180326	141	3933668	95.96334447
77	249	138543	420223	5215742	112	3918240	95.58697247
79	279	124948	420221	5237061	149	3902922	95.21328397
81	340	93064	420219	5208307	206	3887416	94.83500913
83	508	34819	158486	5190443	384	3868940	94.38428
85	868	15209	118198	5177354	732	3848628	93.88876095

SEQQC HQ_VF

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (3358162)	PERCENTAGE
51	2166	4326	20121	5089590	1925	3317499	98.78912929
53	2218	4315	20119	5092655	1984	3316881	98.77072637
55	2273	4190	20117	5092386	2042	3305516	98.43229719
57	2281	4131	20115	5082672	2078	3296647	98.16819439
59	2318	4025	20113	5080871	2136	3290769	97.99315816
61	2379	3920	16578	5077037	2181	3280891	97.69900916
63	2390	3818	16576	5069972	2231	3272894	97.46087294
65	2429	3705	16574	5065760	2286	3266740	97.27761794
67	2481	3589	16572	5055305	2344	3253908	96.89550415
69	2567	3418	16570	5050270	2427	3239082	96.45401264
71	2642	3277	16568	5042739	2516	3233416	96.28528939
73	2713	3119	16093	5030199	2599	3215250	95.74433872
75	2846	2921	16091	5020707	2738	3202221	95.35635863
77	3056	2643	12208	5001832	2907	3188937	94.9607851
79	3296	2397	12206	4985869	3168	3175785	94.56914229
81	3654	2095	12204	4943173	3534	3150539	93.817362
83	4245	1744	12202	4904357	4144	3129772	93.19895824
85	5232	1330	12200	4829951	5140	3089748	92.0071158

SEQQC VF

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (3566875)	PERCENTAGE
51	2151	4376	20121	5094303	1901	3457947	96.94612231
53	2190	4334	20119	5093598	1957	3449072	96.69730506
55	2219	4247	20117	5085123	1997	3429147	96.13869283
57	2259	4171	20115	5083754	2051	3419622	95.87165236
59	2305	4068	20113	5083566	2105	3407131	95.52145786
61	2310	4004	16578	5073838	2130	3396014	95.20978448
63	2388	3865	16576	5073662	2193	3381625	94.80637813
65	2412	3852	16574	5070054	2248	3371234	94.5150587
67	2455	3652	16572	5058864	2308	3352883	94.00057473
69	2537	3496	16570	5051494	2397	3332716	93.43517785
71	2599	3324	16568	5040258	2468	3318651	93.04085509
73	2697	3158	16566	5031832	2571	3295982	92.40531277
75	2820	2973	16564	5022504	2702	3279005	91.92934992
77	3000	2720	12210	5006561	2855	3265178	91.54169967
79	3234	2460	12208	4991276	3105	3247036	91.03307517
81	3576	2142	12206	4952060	3459	3214556	90.12247415
83	4149	1792	12202	4912918	4045	3190231	89.44050464
85	5142	1367	12200	4842586	5039	3144315	88.15321535

FASTX HQ

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (4109236)	PERCENTAGE
51	321	349657	941090	5221142	93	4074384	99.15186181
53	307	299350	941090	5204352	88	4066598	98.9623862
55	291	704087	1275359	5217025	87	4055994	98.70433336
57	299	562056	1051722	5216094	91	4045720	98.45431122
59	272	704087	1275363	5216952	86	4037556	98.25563681
61	274	517178	941669	5216333	95	4026208	97.97947842
63	274	284355	704110	5223270	97	4013160	97.66194981
65	262	284355	704093	5207158	102	4004735	97.45692387
67	266	275809	704116	5218196	96	3990327	97.10629908
69	248	260927	644778	5213440	97	3979700	96.84768653

FASTX HQ_VF

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (4091357)	PERCENTAGE
51	320	332991	941090	5220554	92	4056900	99.15780999
53	308	412751	941090	5215692	91	4049027	98.96537995
55	286	704103	1275359	5222808	86	4038270	98.70245984
57	275	607660	1051556	5221877	82	4028342	98.45980197
59	263	1275363	1422862	5233170	84	4020015	98.25627536
61	264	352161	820394	5226937	86	4008314	97.97028223
63	271	260927	966250	5232621	96	3995607	97.65970068
65	253	259933	704110	5222748	98	3987227	97.45487866
67	247	275809	823518	5222695	89	3973724	97.12484146
69	248	385774	704116	5210822	98	3962345	96.84671858

FASTX VF

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (4331866)	PERCENTAGE
51	315	481615	850103	5219692	87	4214152	97.28260292
53	300	593296	850293	5221017	79	4200727	96.97269029
55	287	777777	1147248	5226740	81	4184503	96.59816347
57	270	655046	795162	5216148	79	4169562	96.25325437
59	254	566598	713132	5221440	78	4155396	95.92623595
61	249	407078	713132	5222160	85	4139615	95.56193566
63	248	607652	742737	5228386	87	4120881	95.12946615
65	259	713132	941553	5229786	92	4103787	94.7348556
67	253	425833	742737	5218016	93	4084503	94.28968948
69	247	293709	762737	5223232	96	4066503	93.87416416

71	261	275805	762322	5222437	106	3969607	96.6020691	
73	241	249729	742887	5218303	102	3955052	96.247867	
75	267	208191	742721	5217535	118	3940741	95.89960275	
77	323	124464	458343	5180228	155	3923458	95.47901362	
79	348	84391	353659	5180247	190	3907756	95.09689879	
81	403	64619	353659	5179585	244	3889429	94.65090348	
83	513	41336	158486	5195430	384	3874091	94.27764675	
85	854	16132	118198	5177430	721	3853764	93.78298058	

71	249	275805	704116	5221884	101	3952353	96.60249643	
73	250	275803	633790	5221877	108	3937511	96.23973171	
75	272	199916	458341	5222773	119	3923315	95.89275636	
77	301	131450	489289	5260517	144	3906475	95.48115699	
79	337	93083	406063	5186528	184	3890502	95.09074862	
81	387	85290	406061	5185619	237	3872415	94.64867036	
83	552	34802	158760	5175515	406	3856292	94.25459572	
85	865	16125	118198	5177391	725	3836261	93.76500266	

71	238	293711	743166	5227774	98	4050681	93.5089174	
73	243	275929	638660	5239395	101	4029077	93.01019468	
75	286	199916	742721	5198565	128	4008582	92.53707294	
77	313	131450	458343	5180510	148	3988770	92.07971807	
79	322	104686	353659	5180670	171	3967678	91.59281474	
81	392	74392	353659	5179831	228	3940666	90.96924974	
83	504	46868	180133	5189013	362	3920042	90.49315007	
85	818	19234	118198	5178453	682	3895703	89.93129058	

NGStoolkit HQ

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (4109233)	PERCENTAGE
51	356	269926	704081	5202036	112	4073565	99.13200347
53	336	284034	704081	5213765	108	4065984	98.94751648
55	287	704087	1275359	5212031	84	4055781	98.69922197
57	276	343066	1275361	5223639	78	4045703	98.45396939
59	273	713199	1275363	5206120	86	4037380	98.25142551
61	292	329813	560174	5209657	100	4025362	97.95896217
63	290	284355	566592	5216127	108	4013371	97.66715589
65	262	260927	566594	5217540	100	4004577	97.45315002
67	271	259931	869614	5206900	102	3990406	97.10829247
69	248	260927	1022212	5223212	97	3979772	96.84950939
71	256	260927	644791	5211812	108	3969677	96.6038431
73	248	259925	638856	5212466	108	3955010	96.24691518
75	301	131450	458341	5180334	139	3940374	95.89074165
77	320	124464	458343	5180158	155	3923449	95.4788643
79	348	84391	353659	5180248	190	3907760	95.09706556
81	403	64619	353659	5179587	244	3889439	94.65121593
83	547	34802	158760	5175563	402	3873496	94.26323599
85	856	16132	118198	5182294	721	3853624	93.77964209

NORMAL RUN

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (4356146)	PERCENTAGE
51	313	751788	1275355	5220008	85	4237603	97.27871839
53	310	481619	893189	5220427	86	4223814	96.96217712
55	306	592237	893622	5214570	95	4207289	96.58282803
57	291	290107	1326693	5204014	92	4192070	96.23345958
59	263	396298	1275363	5223419	79	4177790	95.90564687
61	257	566590	1026951	5227324	90	4162415	95.55269727
63	254	566609	966263	5217666	89	4143431	95.1168992
65	293	284368	694426	5204339	110	4125792	94.71197705
67	271	284368	591529	5204937	101	4106080	94.25946697
69	263	293709	566604	5206121	103	4088670	93.85980176
71	253	284355	644774	5199292	106	4072472	93.48795931
73	262	308949	508737	5200487	112	4051029	92.99571226
75	271	284355	597060	5204048	126	4031416	92.54547483
77	280	179302	508737	5211965	129	4010835	92.07301592
79	294	137467	508737	5199410	157	3989485	91.58290379
81	382	76051	353659	5180123	223	3962515	90.96377853
83	498	49639	180133	5197190	356	3941940	90.49145736
85	822	19129	118198	5170793	677	3916620	89.91020962

FASTX Trim

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (4356146)	PERCENTAGE
51	222	696535	2061671	5302223	52	4219486	96.86282324
53	256	507308	1053095	5234395	52	4204299	96.51418938
55	235	720428	1506407	5251935	52	4189255	96.16883823
57	214	613972	1373857	5301904	48	4174726	95.83530947
59	226	635702	1543986	5271062	60	4160139	95.50044925
61	220	547468	798244	5294909	55	4140613	95.05220899
63	209	307811	996008	5281125	61	4126100	94.71904753
65	212	453374	959297	5295336	65	4109762	94.34399123
67	231	311348	639120	5293281	67	4089205	93.87208326
69	220	187893	623289	5292164	76	4073049	93.50120496
71	259	174658	563836	5287643	105	4054668	93.07924941
73	348	99130	580604	5277113	188	4033823	92.60073009
75	556	33283	183087	5180655	382	4013807	92.14124136
77	967	16753	147721	5173876	788	3991133	91.62073539
79	1960	6115	51040	5137286	1781	3958462	90.87073757
81	3837	2363	19431	4984019	3691	3863441	88.68942868
83	7357	844	9288	4480889	7255	3563644	81.80726725
85	8009	384	3951	2836154	7955	2465763	56.60423227

REQUIRED	(cutoff = 30)	(cutoff = 35)	
Read1	Read2	Read1	Read2
88 (minus12)	94 (minus6)	64 (minus 36)	70 (minus 30)

Prinseq Trim

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (4356146)	PERCENTAGE
51	259	1307085	1459681	5298886	57	4217591	96.81932148
53	267	837168	1052871	5301825	55	4203833	96.50349185
55	246	785406	2290732	5222823	70	4189711	96.1793062
57	236	1157576	1528682	5362967	53	4174196	95.82314275
59	260	713451	1014052	5260655	77	4159423	95.4840127
61	241	264897	990656	5279586	63	4140098	95.04038662
63	227	619016	779248	5275824	67	4125289	94.70043015
65	203	484791	794246	5301639	60	4109377	94.33515314
67	258	350886	742327	5284818	81	4087952	93.8433193
69	226	244869	709692	5300373	80	4072870	93.49709583
71	261	206878	634914	5302898	107	4054655	93.07895098
73	313	100288	515373	5280388	172	4035540	92.64014567
75	539	35384	183087	5198955	372	4013863	92.1425269
77	965	17212	147721	5174655	785	3991154	91.62121747
79	1960	6115	51040	5137286	1781	3958462	90.87073757
81	3837	2363	19431	4984019	3691	3863441	88.68942868
83	7357	844	9288	4480889	7255	3563644	81.80726725
85	8009	384	3951	2836154	7955	2465763	56.60423227

SolexaQA Trim

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (4356146)	PERCENTAGE
51	498	52179	196946	5274259	240	3202610	73.51934485
53	559	37005	115841	5223676	295	3138050	72.03730086
55	602	29940	142379	5202033	350	3074184	70.57118839
57	665	26505	104707	5199048	427	3002418	68.92372294
59	791	19216	77495	5177453	548	2935656	67.39112968
61	914	13952	72999	5178474	680	2861521	65.68928131
63	1102	10709	72997	5169202	881	2788092	64.00363991
65	1351	8344	62419	5154214	1138	2712871	62.27686124
67	1695	6086	56194	5146794	1482	2622358	60.19903832
69	2118	4656	23921	5119255	1930	2532371	58.1332903
71	2812	3258	20512	5101258	2624	2442013	56.05902557
73	3581	2357	20510	5038061	3441	2340161	53.72090375
75	4602	1653	9790	4929418	4477	2229002	51.16912978
77	6157	1151	8284	4838472	6040	2111090	48.46233345
79	7480	819	6634	4583121	7386	1953497	44.84461724
81	8432	602	4520	4113223	8368	1726313	39.62936504
83	9331	431	3795	3583551	9280	1475706	33.87641277
85	8329	340	4022	2690352	8293	1117234	25.64730383

FASTX clipper (N's only removed)

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (2993545)	PERCENTAGE
51	7356	886	4677	4593897	6749	2820681	94.22544174
53	7393	864	4675	4542921	6867	2803260	93.64348958
55	7573	842	4673	4537285	7066	2794920	93.36489012
57	7551	822	4671	4485963	7085	2778437	92.81427204
59	7663	801	4669	4476489	7251	2771260	92.57452285
61	7606	792	4667	4418762	7279	2755731	92.05577334
63	7719	775	4128	4406850	7457	2748783	91.82367394
65	7617	768	4126	4342352	7461	2726069	91.06490799
67	7706	751	3679	4327294	7649	2720039	90.86347458
69	7604	739	3677	4255902	7547	2689197	89.83319108
71	7662	724	3675	4235034	7621	2672226	89.26627126
73	7565	718	3258	4165818	7535	2639399	88.16967842
75	7650	703	3256	4141081	7612	2619425	87.50244276
77	7535	690	3254	4054886	7494	2583357	86.29758363
79	7642	667	2968	4020441	7600	2559503	85.50073575
81	7724	639	2953	3964832	7701	2526534	84.39939937
83	7689	615	2946	3836487	7670	2464935	82.34167183
85	7898	568	2566	3718655	7882	2403165	80.278232

FASTX clipper (N's adap's removed)

HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS USED (2972526)	PERCENTAGE
51	7487	854	4672	4549023	6865	2799182	94.16846144
53	7490	841	4670	4504034	6934	2783760	93.64964343
55	7629	823	4668	4502313	7126	2778723	93.48019159
57	7597	806	4666	4450171	7143	2754518	92.66590099
59	7711	791	4664	4441056	7285	2741967	92.24366751
61	7649	783	4127	4384798	7308	2730907	91.87159339
63	7769	766	4125	4375076	7485	2728198	91.78045877

65	7643	760	4123	4309234	7488	2706825	91.06144067
67	7712	740	3678	4292858	7659	2699943	90.82992041
69	7620	732	3676	4225113	7564	2669761	89.81455503
71	7672	719	3674	4204200	7628	2653504	89.26764644
73	7574	711	3091	4134332	7537	2619945	88.13867398
75	7645	696	3089	4110376	7611	2599198	87.44071541
77	7513	685	2970	4023736	7481	2558877	86.08425965
79	7602	663	2968	3989140	7572	2536178	85.32063302
81	7698	637	2953	3935274	7674	2505139	84.27643694
83	7654	612	2946	3806958	7636	2448545	82.37253434
85	7869	564	2565	3691744	7855	2387419	80.31616881

(b) QUALITY CONTROL DATA

	Paracoccus denitrificans	HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS	Total	Percentage	Actual Total	Percent in tools
High Quality	SeqQC_HQ	37	3031	1521	40897	2944856	2717	3178904	3709241	85.7	3951522	93.9
	FastX_HQ	37	3031	1521	40897	2944856	2717	3178904	3709241	85.7	3951522	93.9
	NGStoolkit	37	3011	1531	30843	2947508	2713	3194015	3726381	85.7	3951522	94.3
Read Trimming	Fastx_TRIM	37	3135	1399	40897	2881513	2828	2949992	3457858	85.3	3951522	87.5
	Prinseq_TRIM	37	3022	1478	40897	2897228	2685	3294052	3951522	83.4	3951522	100.0
	SolexaQA_TRIM	37	3841	929	14681	2602175	3460	2590777	3951522	65.6	3951522	100.0
Adaptor removal	SeqQC_VF	37	4381	835	7881	2646212	3821	2603324	3059168	85.1	3951522	77.4
	FastX_VF	37	2916	1580	40897	2903631	2592	3306175	3946398	83.8	3951522	99.9
HQ + Adaptor removal	SeqQC_HQ_VF	37	4572	809	7881	2673869	3975	2503969	2871161	87.2	3951522	72.7
	FastX_HQ_VF	37	3013	1523	40897	2940185	2714	3175823	3706894	85.7	3951522	93.8
	Fastx_clipper(N's adap's remove	37	5201	730	3405	2700978	4413	2712929	3282666	82.6	3951522	83.1
Normal	Normal	37	2942	1579	40897	2896523	2592	3305578	3951522	83.7	3951522	100.0
N's removed	Fastx_clipper(N's only removed	37	5184	739	4034	2708167	4396	2720680	3294457	82.6	3951522	83.4

	Mycobacterium tuberculosis	HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS	Total	Percentage	Actual Total	Percent in tools
High Quality	SeqQC_HQ	37	2729	2181	16781	3238077	2306	12959870	13787485	94.0	15839552	87.0
	FastX_HQ	37	2729	2181	16781	3238077	2306	12959870	13787485	94.0	15839552	87.0
	NGStoolkit	37	2702	2228	16781	3256680	2263	13105772	13940064	94.0	15839552	88.0
Read Trimming	Fastx_TRIM	37	2677	2090	12372	3320908	2263	14337502	15839552	90.5	15839552	100.0
	Prinseq_TRIM	37	2673	2092	12372	3322269	2261	14337502	15839552	90.5	15839552	100.0
	SolexaQA_TRIM	37	3853	1073	8619	2909800	3433	9362203	15839552	59.1	15839552	100.0
Adaptor removal	SeqQC_VF	37	3976	1240	7295	2966933	3282	11644082	13297453	87.6	15839552	84.0
	FastX_VF	37	2723	2378	16781	3269493	2213	14425763	15810332	91.2	15839552	99.8
HQ + Adaptor removal	SeqQC_HQ_VF	37	4125	1137	6242	2986069	3506	10613757	11627720	91.3	15839552	73.4
	FastX_HQ_VF	37	2732	2151	16781	3239003	2307	12949124	13773306	94.0	15839552	87.0
	Fastx_clipper(N's adap's remove	37	4975	917	4301	2902527	3963	11461625	13766430	83.3	15839552	86.9
Normal	Normal	37	2741	2342	16781	3268643	2222	14436774	15839552	91.1	15839552	100.0
N's removed	Fastx_clipper(N's only removed	37	4901	935	4730	2910657	3925	11560137	13798166	83.8	15839552	87.1

	Helicobacter pylori	HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS	Total	Percentage	Actual Total	Percent in tools
High Quality	SeqQC_HQ	31	402	12063	37045	1541862	245	2794615	2920392	95.7	3015238	96.9
	FastX_HQ	31	402	12063	37045	1541862	245	2794615	2920392	95.7	3015238	96.9
	NGStoolkit	31	398	12695	37045	1547586	241	2793029	2920352	95.6	3015238	96.9

Adaptor removal	SeqQC_VF	31	739	4374	21047	1494666	555	2587276	2762088	93.7	3015238	91.6
	FastX_VF	31	424	11730	37047	1547416	249	2831298	2831298	100.0	3015238	93.9
HQ + Adaptor removal	SeqQC_HQ_VF	31	723	4374	16438	1484064	547	2531181	2675157	94.6	3015238	88.7
	FastX_HQ_VF	31	407	11987	40952	1545486	243	2777746	2897981	95.9	3015238	96.1
	Fastx_clipper(N's adap's remove	31	4258	365	2265	1150852	2572	1565604	2039702	76.8	3015238	67.6
Normal	Normal	31	409	12063	37047	1546563	244	2848099	3015238	94.5	3015238	100.0
N's removed	Fastx_clipper(N's only removed	31	4233	374	2267	1171869	2592	1611272	2071499	77.8	3015238	68.7

	Escherichia coli	HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS	Total	Percentage	Actual Total	Percent in tools
High Quality	SeqQC_HQ	37	1087	29220	89683	4914090	476	19685807	21500807	91.6	45440998	47.3
	FastX_HQ	37	722	39131	139561	5037632	315	41118707	43607176	94.3	45440998	96.0
	NGStoolkit	37	705	39497	249390	5074611	308	41127793	43607077	94.3	45440998	96.0
Adaptor removal	SeqQC_VF	37	2350	4859	22609	4784172	1660	19035198	21497321	88.5	45440998	47.3
	FastX_VF	37	809	37651	99638	5035806	337	41846339	45084047	92.8	45440998	99.2
HQ + Adaptor removal	SeqQC_HQ_VF	37	2406	4687	22609	4789788	1707	19586603	21497116	91.1	45440998	47.3
	FastX_HQ_VF	37	723	41427	139561	5019845	308	40936740	43283560	94.6	45440998	95.3
	Fastx_clipper(N's adap's remove	37	8784	710	4004	4170773	6795	30950040	36872384	83.9	45440998	81.1
Normal	Normal	37	812	36847	104347	5027252	350	42006443	45440998	92.4	45440998	100.0
N's removed	Fastx_clipper(N's only removed	37	8748	725	4021	4192037	6755	31064360	37060498	83.8	45440998	81.6

	Vibrio parahaemolyticus	HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS	Total	Percentage	Actual Total	Percent in tools
High Quality	SeqQC_HQ	57	292	593300	713199	5216704	88	4037171	4099136	98.5	4356146	94.1
	FastX_HQ	57	299	562056	1051722	5216094	91	4045720	4109236	98.5	4356146	94.3
	NGStoolkit	57	276	343066	1275361	5223639	78	4045703	4109233	98.5	4356146	94.3
Read Trimming	Fastx_TRIM	57	214	613972	1373857	5301904	48	4174726	4356146	95.8	4356146	100.0
	Prinseq_TRIM	57	236	1157576	1528682	5362967	53	4174196	4356146	95.8	4356146	100.0
	SolexaQA_TRIM	57	665	26505	104707	5199048	427	3002418	4356146	68.9	4356146	100.0
Adaptor removal	SeqQC_VF	57	2259	4171	20115	5083754	2051	3419622	3566875	95.9	4356146	81.9
	FastX_VF	57	270	655046	795162	5216148	79	4169562	4331866	96.3	4356146	99.4
HQ + Adaptor removal	SeqQC_HQ_VF	57	2281	4131	20115	5082672	2078	3296647	3358162	98.2	4356146	77.1
	FastX_HQ_VF	57	275	607660	1051556	5221877	82	4028342	4091357	98.5	4356146	93.9
	Fastx_clipper(N's adap's remove	57	7597	806	4666	4450171	7143	2754518	2972526	92.7	4356146	68.2
Normal	Normal	57	291	290107	1326693	5204014	92	4192070	4356146	96.2	4356146	100.0
N's removed	Fastx_clipper(N's only removed	57	7551	822	4671	4485963	7085	2778437	2993545	92.8	4356146	68.7

(b) QUALITY CONTROL DATA

	Paracoccus denitrificans	HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS	Total	Percentage	Actual Total	Percentage
Pipeline 1	FasxN-flexOWN-trim-HQ	37	5341	668	4154	2597393	4502	2490434	2942693	84.6	3951522	63.0246776
Pipeline 2	Trim-flexOWN	37	3025	1474	40897	2892018	2684	3281416	3778468	86.8	3951522	83.0418254
Pipeline 3	Trim-flexOWN-HQ	37	2968	1472	40897	2874950	2675	3186330	3616478	88.1	3951522	80.6355121
Pipeline 4	FlexCON-trim-HQ	37	3012	1480	40897	2889767	2674	3279569	3780956	86.7	3951522	82.9
Pipeline 5	FlexCON-HQ-trim	37	3027	1455	40897	2846829	2705	3117533	3546772	87.8	3951522	78.8
Pipeline 6	Fastxadap-HQ	37	3111	1496	40897	2939862	2758	3172511	3706895	85.5	3951522	80.2857987
TAH pipeline		37	3511	1082	17891	2673937	3174	1882230	2155489	87.3	3951522	47.6330386

	Mycobacterium tuberculosis	HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS	Total	Percentage	Actual Total	Percentage
Pipeline 1	FasxN-flexOWN-trim-HQ	37	5181	861	4465	3018499	4345	12009671	13763730	87.2559328	15839552	75.8207745
Pipeline 2	Trim-flexOWN	37	2670	2102	12372	3322430	2266	14314447	15780884	90.7075104	15839552	90.3715395
Pipeline 3	Trim-flexOWN-HQ	37	2680	2041	12874	3297760	2294	13846224	14925801	92.7670414	15839552	87.4155027
Pipeline 4	FlexCON-trim-HQ	37	2676	2101	12372	3321481	2263	14300306	15780292	90.6	15839552	90.2
Pipeline 5	FlexCON-HQ-trim	37	2938	1742	12149	3234281	2552	12808758	13750403	93.1	15839552	80.8
Pipeline 6	Fastxadap-HQ	37	5007	899	4363	2944462	4113	10646332	11968147	88.9555585	15839552	67.2135929
TAH pipeline	NGStoolkit-flexOWN-HQ	37	3434	1336	10725	2995462	3015	7388938	7888233	93.6	15839552	46.6486552

	Helicobacter pylori	HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS	Total	Percentage	Actual Total	Percentage
Pipeline 1	FasxN-flexOWN-trim-HQ	31	4160	373	2264	1155377	2564	1541957	1955266	78.8617508	3015238	51.1388156
Pipeline 2	Trim-flexOWN	31	400	12063	37047	1548510	242	2815234	2968426	94.8392852	3015238	93.3668918
Pipeline 3	Trim-flexOWN-HQ	31	364	13110	47263	1560575	227	2764965	2876427	96.1249842	3015238	91.6997265
Pipeline 4	FlexCON-trim-HQ	31	375	13793	40952	1555186	232	2763204	2876427	96	3015238	91.6
Pipeline 5	FlexCON-HQ-trim	31	375	13793	40952	1555186	232	2763204	2876427	96	3015238	91.6
Pipeline 6	Fastxadap-HQ	31	4232	364	2262	1144246	2567	1542933	1971741	78.2523161	3015238	51.1711845
TAH pipeline	NGStoolkit-flexOWN-HQ	31	185	411830	602993	1551704	69	2611204	2714926	96.1	3015238	86.6002617

	Escherichia coli	HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS	Total	Percentage	Actual Total	Percentage
Pipeline 1	FasxN-flexOWN-trim-HQ	37	8859	719	4092	4224460	6830	23621270	27364482	86.3209104	45440998	51.982287
Pipeline 2	Trim-flexOWN	37	840	39168	122365	5026198	382	41627007	44819384	92.8772403	45440998	91.6067182
Pipeline 3	Trim-flexOWN-HQ	37	788	38148	125596	5022721	345	25598199	27369235	93.5290994	45440998	56.3328275
Pipeline 4	FlexCON-trim-HQ	37	844	37439	192125	4996285	375	25585939	27369235	93.4	45440998	56.3
Pipeline 5	FlexCON-HQ-trim	37	844	37439	192125	4996285	375	25585939	27369235	93.4	45440998	56.3

Pipeline 6	Fastxadap-HQ		37	8784	710	4004	4170773	6795	30950040	36872384	83.9382666	45440998	68.1103879
TAH pipeline	NGStoolkit-flexOWN-HQ		31	678	4121844	4121844	4928019	100	36828395	38238190	96.3	45440998	81.0466245
	Vibrio parahaemolyticus	HASH	NODES	N50	MAX	TOTAL	CONTIGS	READS	Total	Percentage	Actual Total	Percentage	
Pipeline 1	FasxN-flexOWN-trim-HQ		57	7855	771	3877	4415194	7337	2754681	2990434	92.1164286	4356146	63.2366546
Pipeline 2	Trim-flexOWN		57	207	934721	1489867	5300549	50	4129058	4307124	95.8	4356146	94.7
Pipeline 3	Trim-flexOWN-HQ		57	281	349789	819033	5229615	64	4026405	4114370	97.8	4356146	92.4
Pipeline 4	FlexCON-trim-HQ		57	282	319437	1297290	5260587	83	4126372	4306698	95.8	4356146	94.7
Pipeline 5	FlexCON-HQ-trim		57	229	494114	990781	5299735	53	3981341	4057916	98.1	4356146	91.3
Pipeline 6	Fastxadap-HQ		57	7564	803	4666	4440009	7129	2661732	2795859	95.2	4356146	61.1
TAH pipeline	NGStoolkit-flexOWN-HQ		57	266	1236609	1410318	5206616	69	2797610	2830497	98.8381192	4356146	62.1

(b) QUALITY CONTROL DATA

CONTIGS

	PDSD1	MTBNA-A0008	HP908	ECNA114	VPPCV08-7
Pipeline 1	4502	4345	2564	6830	7337
Pipeline 2	2684	2266	242	382	50
Pipeline 3	2675	2294	227	345	64
Pipeline 4	2674	2263	232	375	83
Pipeline 5	2705	2552	232	375	53
Pipeline 6	2758	4113	2567	6795	7129
TAH pipeline	3174	3015	69	100	69

MAX

	PDSD1	MTBNA-A0008	HP908	ECNA114	VPPCV08-7
Pipeline 1	4154	4465	2264	4092	3877
Pipeline 2	40897	12372	37047	122365	1489867
Pipeline 3	40897	12874	47263	125596	819033
Pipeline 4	40897	12372	40952	192125	1297290
Pipeline 5	40897	12149	40952	192125	990781
Pipeline 6	40897	4363	2262	4004	4666
TAH pipeline	17891	10725	602993	4121844	1410318

PERCENTAGE OF READS USED

	PDSD1	MTBNA-A0008	HP908	ECNA114	VPPCV08-7
Pipeline 1	84.6	87.3	78.9	86.3	92.1
Pipeline 2	86.8	90.7	94.8	92.9	95.8
Pipeline 3	88.1	92.8	96.1	93.5	97.8
Pipeline 4	86.7	90.6	96.0	93.4	95.8
Pipeline 5	87.8	93.1	96.0	93.4	98.1
Pipeline 6	85.5	89.0	78.3	83.9	95.2
TAH pipeline	87.3	93.6	96.1	96.3	98.8

(b) QUALITY CONTROL DATA

PDSD1

Hash	Nodes	N50	MAX	Total	Used reads	Total	% used	CONTIGS
37	3511	1082	17891	2673937	1882230	2155489	87.32	3174
39	3411	1071	17889	2618123	1868383	2155489	86.68	3128
41	3608	1024	17887	2638927	1875828	2155489	87.03	3321
43	3578	991	17885	2564317	1851456	2155489	85.89	3261
45	3477	987	17883	2483743	1831432	2155489	84.97	3214
47	3706	909	17881	2506861	1840731	2155489	85.4	3461
49	3623	887	14478	2413069	1815469	2155489	84.23	3411
51	3982	803	14476	2428354	1820234	2155489	84.45	3714
53	3879	766	14474	2310173	1783080	2155489	82.72	3630
55	4363	681	14285	2312741	1785153	2155489	82.82	4053
57	4268	638	11560	2167367	1738973	2155489	80.68	4002
59	4135	590	10486	1975966	1667297	2155489	77.35	3897
61	4616	495	10484	1877333	1624526	2155489	75.37	4316
63	4321	421	5242	1563932	1476518	2155489	68.5	4074
65	4104	318	2823	1189727	1239235	2155489	57.49	3854

HP908

Hash	Nodes	N50	MAX	Total	Used reads	Total	% used	CONTIGS
27	342	359004	511529	1502280	2606265	2714926	96	165
29	259	386201	572342	1527365	2595855	2714926	95.61	106
31	185	411830	602993	1551704	2611204	2714926	96.18	69
33	447	9486	32162	1539140	2592979	2714926	95.51	287
35	487	7273	23459	1510679	2559390	2714926	94.27	325
37	556	5980	20649	1500367	2532769	2714926	93.29	401
39	623	4679	18954	1458034	2502458	2714926	92.17	473
41	720	3392	12973	1389435	2455379	2714926	90.44	580
43	884	2359	11414	1296469	2392660	2714926	88.13	742
45	1311	1463	9300	1238688	2355772	2714926	86.77	1135
47	1813	945	8351	1111445	2220555	2714926	81.79	1511
49	2460	548	5082	886638	1931142	2714926	71.13	1729
51	1739	302	3078	421161	1183143	2714926	43.58	1019
53	65	251	829	13343	141997	2714926	5.23	38

MTBNA-A0008

Hash	Nodes	N50	MAX	Total	Used reads	Total	% used	CONTIGS
37	3434	1336	10725	2995462	7388938	7888233	93.67	3015
39	3424	1313	10723	2994753	7369679	7888233	93.43	3052
41	3339	1293	8935	2974210	7357114	7888233	93.27	3086
43	3408	1254	8933	2961519	7350492	7888233	93.18	3195
45	3417	1194	10165	2907551	7303680	7888233	92.59	3232
47	3548	1133	10165	2895208	7283823	7888233	92.34	3346
49	3684	1057	10165	2864805	7267651	7888233	92.13	3496
51	3837	1004	10165	2838101	7233815	7888233	91.7	3655
53	3902	938	10165	2760753	7160457	7888233	90.77	3730
55	4096	866	10165	2719949	7127359	7888233	90.35	3926
57	4279	811	5853	2663169	7074364	7888233	89.68	4104
59	4505	737	5558	2594787	7009593	7888233	88.86	4333
61	4768	666	5556	2517819	6945524	7888233	88.05	4602
63	5029	598	5554	2419838	6844975	7888233	86.77	4863
65	5340	509	5543	2272695	6685880	7888233	84.76	5206

ECNA114

Hash	Nodes	N50	MAX	Total	Used reads	Total	% used	CONTIGS
27	2365	9649	36631	4724859	36575849	38238190	95.65	993
29	789	1953528	2110120	4881444	36690082	38238190	95.95	124
31	678	4121844	4121844	4928019	36828395	38238190	96.31	100
33	1513	22945	84695	4914125	36606466	38238190	95.73	564

35	1390	22756	91593	4921399	36457047	38238190	95.34	545
37	1317	24242	91593	4934630	36131581	38238190	94.49	546
39	1249	24809	90152	4935886	36100872	38238190	94.41	567
41	1174	24046	90150	4935329	35936050	38238190	93.98	554
43	1167	23272	84434	4936986	35866947	38238190	93.8	562
45	429	562402	904051	4960745	35871087	38238190	93.81	131
47	376	750176	1557802	4953109	35404495	38238190	92.59	106
49	1369	11996	45149	4877312	35143862	38238190	91.91	881
51	1936	5490	27111	4751461	34728142	38238190	90.82	1455
53	10788	455	17051	3589485	28028320	38238190	73.3	8126

VPPCV08-7

Hash	Nodes	N50	MAX	Total	Used reads	Total	% used	CONTIGS
51	593	93616	353866	5173283	2809503	2830497	99.26	216
53	577	89794	353868	5173506	2804795	2830497	99.09	210
55	524	92594	353870	5174326	2800459	2830497	98.94	204
57	266	1236609	1410318	5206616	2797610	2830497	98.84	69
59	196	1363224	2506561	5251991	2794268	2830497	98.72	37
61	488	73983	353876	5174614	2786182	2830497	98.43	207
63	189	2805973	2805973	5252562	2783302	2830497	98.33	41
65	249	506952	1106043	5201674	2777978	2830497	98.14	87
67	181	2824303	2824303	5198729	2772327	2830497	97.94	41
69	173	1217792	1835687	5264435	2766902	2830497	97.75	41
71	182	2810319	2810319	5263919	2762093	2830497	97.58	54
73	178	1871980	2490340	5190297	2755503	2830497	97.35	61
75	195	1809160	2225334	5191086	2749377	2830497	97.13	72
77	920	15796	78642	5171146	2740310	2830497	96.81	687
79	1167	10850	78640	5164125	2733156	2830497	96.56	966
81	1607	6940	57421	5155082	2725114	2830497	96.28	1418
83	2306	4364	57419	5124281	2712246	2830497	95.82	2143
85	3514	2406	31044	5038676	2682787	2830497	94.78	3380

APPENDIX B

(a) Genomic DNA isolation

Qiagen DNeasy Blood & Tissue kit (Bacterial genomic DNA isolation protocol)

Protocol: Pretreatment for Gram-Negative Bacteria

This protocol is designed for purification of total DNA from Gram-negative bacteria, such as *E. coli*. The protocol describes the preliminary harvesting of bacteria before DNA purification.

Procedure

1. Harvest cells (maximum 2 x 109 cells) in a microcentrifuge tube by centrifuging for 10 min at 5000 x g (7500 rpm). Discard supernatant.

2. Resuspend pellet in 180 µl Buffer ATL.

3. Continued by following the protocol for Purification of total DNA from Animal tissues (Spin-Column protocol).

Purification of Total DNA from Animal Tissues (Spin-Column Protocol)

4. Add 20 µl proteinase K. Mix thoroughly by vortexing, and incubate at 56°C until the tissue is completely lysed. Vortex occasionally during incubation to disperse the sample, or place in a thermomixer, shaking water bath, or on a rocking platform. Lysis time varies depending on the type of tissue processed. Lysis is usually complete in 1–3 hr or, for rodent tails, 6–8 hr. If it is more convenient, samples can be lysed overnight; this will not affect them adversely. After incubation the lysate may appear viscous, but should not be gelatinous as it may clog the DNeasy Mini spin column.

Optional: If RNA-free genomic DNA is required, add 4 µl RNase A (100 mg/ml), mix by vortexing, and incubate for 2 min at room temperature before continuing with step 5. Transcriptionally active tissues such as liver and kidney contain high levels of RNA, which will copurify with genomic DNA. For tissues that contain low levels of RNA, such as rodent tails, or if residual RNA is not a concern, RNase A digestion is not necessary.

5. Vortex for 15 sec. Add 200 µl Buffer AL to the sample, and mix thoroughly by vortexing. Then add 200 µl ethanol (96–100%), and mix again thoroughly by vortexing. It is essential that the sample, Buffer AL, and ethanol are mixed immediately and thoroughly by vortexing or pipetting to yield a homogeneous solution. Buffer AL and ethanol can be premixed and added together in one step to save time when processing multiple samples.

A white precipitate may form on addition of Buffer AL and ethanol. This precipitate does not interfere with the DNeasy procedure. Some tissue types (e.g., spleen, lung) may form a gelatinous lysate after addition of Buffer AL and ethanol. In this case, vigorously shaking or vortexing the preparation is recommended.

6. Pipet the mixture from step 3 (including any precipitate) into the DNeasy Mini spin column placed in a 2 ml collection tube (provided). Centrifuge at 6000 x g (8000 rpm) for 1 min. Discard flow-through and collection tube.

7. Place the DNeasy Mini spin column in a new 2 ml collection tube (provided), add 500 µl Buffer AW1, and centrifuge for 1 min at 6000 x g (8000 rpm). Discard flow-through and collection tube.

8. Place the DNeasy Mini spin column in a new 2 ml collection tube (provided), add 500 µl Buffer AW2, and centrifuge for 3 min at 20,000 x g (14,000 rpm) to dry the DNeasy membrane. Discard flow-through and collection tube. It is important to dry the membrane of the DNeasy Mini spin column, since residual ethanol may interfere with subsequent reactions.

This centrifugation step ensures that no residual ethanol will be carried over during the following elution. Following the centrifugation step, remove the DNeasy Mini spin column carefully so that the column does not come into contact with the flow-through, since this will result in carryover of ethanol. If carryover of ethanol occurs, empty the collection tube, then reuse it in another centrifugation for 1 min at 20,000 x g (14,000 rpm).

9. Place the DNeasy Mini spin column in a clean 1.5 ml or 2 ml microcentrifuge tube (not provided), and pipet 200 µl Buffer AE directly onto the DNeasy membrane. Incubate at room temperature for 1 min, and then centrifuge for 1 min at 6000 x g (8000 rpm) to elute. Elution with 100 µl (instead of 200 µl) increases the final DNA concentration in the eluate, but also decreases the overall DNA yield.

10. Recommended: For maximum DNA yield, repeat elution once as described in step 9. This step leads to increased overall DNA yield. A new microcentrifuge tube can be used for the second elution step to prevent dilution of the first eluate. Alternatively, to combine the eluates, the microcentrifuge tube from step 9 can be reused for the second elution step. Note: Do not elute more than 200 µl into a 1.5 ml microcentrifuge tube because the DNeasy Mini spin column will come into contact with the eluate.

APPENDIX C

(a) Plasmid DNA isolation

High Copy number plasmid isolation protocol

Harvesting

1. Transfer 1.5 ml of bacterial culture to a microcentrifuge tube.

2. Centrifuge for 1 min at full speed (14,000 rpm) in a microcentrifuge and discard the supernatant (If more than 1.5 ml of bacterial culture is used, repeat the Harvesting Step).

Resuspension

3. Add 200 µl of PD1 Buffer (RNase A added) to the tube and re-suspend the cell pellet by vortexing or pipetting.

Lysis

4. Add 200 µl of PD2 Buffer and mix gently by inverting the tube 10 times. Do not vortex, avoid shearing genomic DNA.

5. Stand for 2 minutes at room temperature until lysate clears.

Neutralization

6. Add 300 µl of PD3 Buffer and mix immediately by inverting the tube 10 times. Do not vortex.

7. Centrifuge for 3 minutes at full speed (14,000 rpm).

DNA Binding

8. Place a PD Column in a Collection Tube.

9. Apply the clear lysate (supernatant) from Step 4 to the PD Column.

10. Centrifuge at 6,000 x g (8,000 rpm) for 30 seconds.

11. Discard the flow-through and place the PD Column back in the Collection Tube.

Wash

12. Add 400 µl of W1 Buffer in the PD column.

13. Centrifuge at 6,000 x g (8,000 rpm) for 30 seconds.

14. Discard the flow-through and place the PD Column back in the Collection Tube.

15. Add 600 µl of Wash Buffer (ethanol added) in the PD column.

16. Centrifuge at 6,000 x g (8,000 rpm) for 30 seconds.

17. Discard the flow-through and place the PD Column back in the Collection Tube.

18. Centrifuge again for 2 minutes at full speed to dry the Column matrix.

DNA Elution

19. Transfer dried PD Column onto a clean micro centrifuge tube (Not provided).

20. Add 50 µl of Elution Buffer or water into the centre of the column matrix.

21. Stand for 2 minutes until Elution Buffer or water is absorbed by the matrix

22. Centrifuge for 2 minutes at full speed to elute purified DNA.

(b) Data corresponding to assembly of Vibrio parahaemolyticus PCV08-7 sequencing data

	VP_ILLUMINA_GENOME	VP_ILLUMINA_PLASMID	VP_ION PROTON_GENOME
Read 1	643.1MB	8GB	21.5GB
Read 2	643.1MB	8GB	

	VP_ILLUMINA_GENOME	VP_ILLUMINA_PLASMID	VP_ION PROTON_GENOME
Read length	100	101	286
read1	2178073	30260889	81219173
read2	2178073	30260889	
Total reads	4356146	60521778	
quality			
r1 (to retain phred 30 n above only)	Cut 12 make 88		
r2 (to retain phred 30 n above only)	Cut 6 make 94		

	TRIMMING	FLEXBAR	FASTX
VP_Illumina_genome			
Total- 4356146			
reads			used to trim and keep sequences to defined size
2178073	2171997	2100663	1362036
2178073	2169739	2121256	1468461
4356146	4341736	4221919	2830497
VP_Illumina_plasmid			
Total- 60521778			
reads			
30260889	30258237	29841362	20332806
30260889	29930256	29086809	19015334
60521778	60188493	58928171	39348140
VP_IonTorrent_genome			
Total- 81219173			
split_files			
15130444	9784617	5808671	5205226
15130444	9692074	5737656	5115010
15130444	9783497	5799653	5201145
15130444	9825746	5810367	5194040
15130444	9445368	5542307	4939213
5566953	2404017	1285624	1110857
81219173	50935319	29984278	26765491

ASSEMBLIES

VP_Illumina_genome

Hash	Nodes	N50	MAX	TOTAL BASES	READS USED	TOTAL READS	% used	CONTIGS
51	593	93616	353866	5173283	2809503	2830497	99.26	216
53	577	89794	353868	5173506	2804795	2830497	99.09	210
55	524	92594	353870	5174326	2800459	2830497	98.94	204
57	266	1236609	1410318	5206616	2797610	2830497	98.84	69
59	196	1363224	2506561	5251991	2794268	2830497	98.72	37
61	488	73983	353876	5174614	2786182	2830497	98.43	207
63	189	2805973	2805973	5252562	2783302	2830497	98.33	41
65	249	506952	1106043	5201674	2777978	2830497	98.14	87
67	181	2824303	2824303	5198729	2772327	2830497	97.94	41
69	173	1217792	1835687	5264435	2766902	2830497	97.75	41
71	182	2810319	2810319	5263919	2762093	2830497	97.58	54
73	178	1871980	2490340	5190297	2755503	2830497	97.35	61
75	195	1809160	2225334	5191086	2749377	2830497	97.13	72
77	920	15796	78642	5171146	2740310	2830497	96.81	687
79	1167	10850	78640	5164125	2733156	2830497	96.56	966
81	1607	6940	57421	5155082	2725114	2830497	96.28	1418
83	2306	4364	57419	5124281	2712246	2830497	95.82	2143
85	3514	2406	31044	5038676	2682787	2830497	94.78	3380

VP_PLASMID

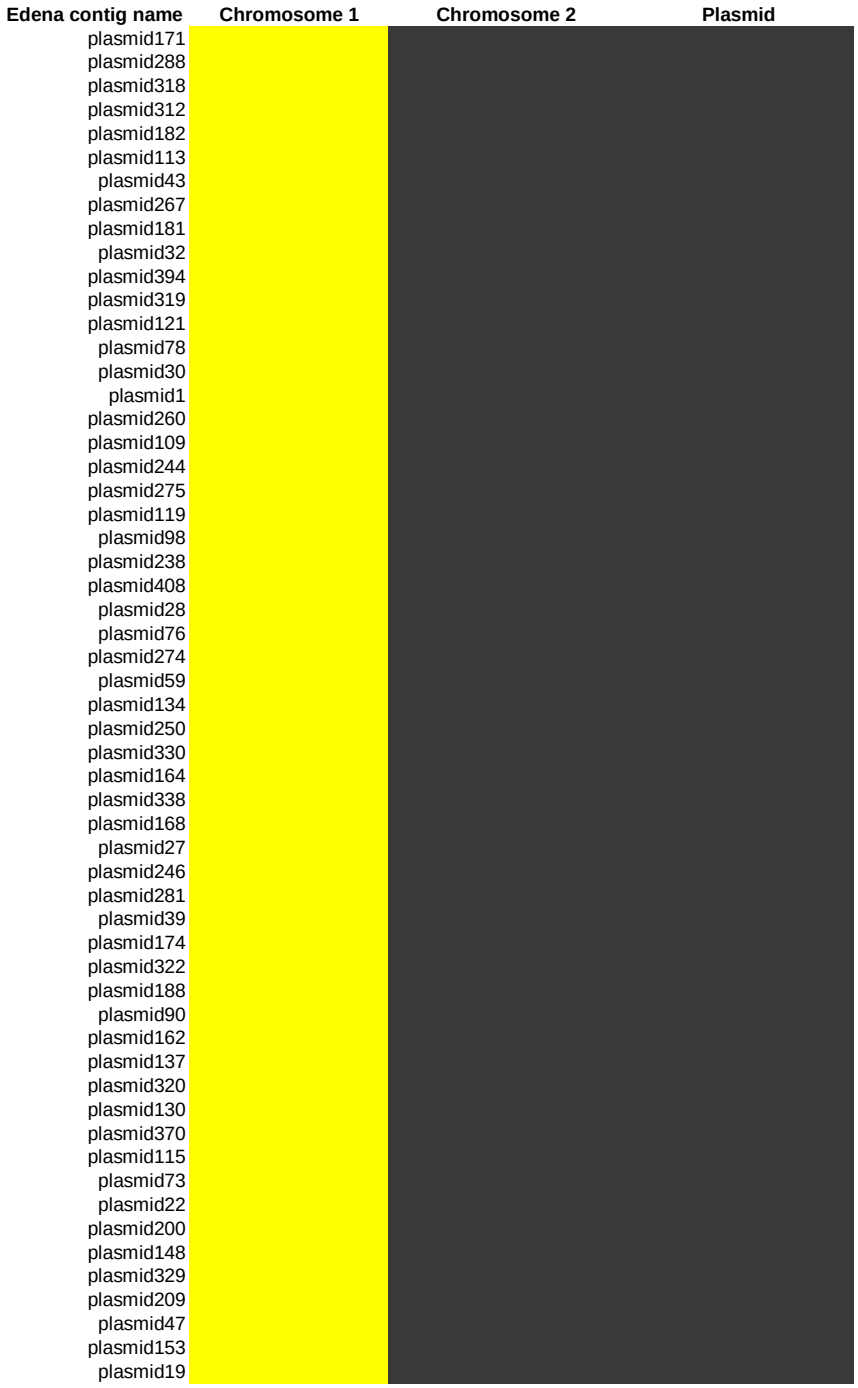
Hash	Nodes	N50	MAX	TOTAL BASES	READS USED	TOTAL READS	% used	CONTIGS
51	634755	42	2081	14448175	832660	39348140	2.11614	830
53	18842	1739	11380	5421132	37645146	39348140	95.672	4234
55	529	261534	912254	5363310	38772765	39348140	98.5377	192
57	479	181016	349205	5488434	38613894	39348140	98.134	161
59	907	53542	207595	5136826	38360541	39348140	97.4901	406
61	763	62761	311110	5140322	38191866	39348140	97.0614	329
63	249	1352636	2218988	6058656	38107491	39348140	96.847	74
65	632	77061	405453	5147653	37847433	39348140	96.1861	270
67	183	671267	1347815	5635904	37790795	39348140	96.0421	64
69	172	4314736	4314736	5208360	37626859	39348140	95.6255	61
71	149	1171944	1397891	5613125	37444968	39348140	95.1632	53
73	140	2623634	2946939	6031740	37261940	39348140	94.6981	45
75	141	4055079	4055079	5624067	37102863	39348140	94.2938	46
77	126	2448332	2755467	5657653	36934939	39348140	93.8671	41
79	134	1633116	2472921	5222960	36759083	39348140	93.4201	43
81	106	5169475	5169475	5621274	36591916	39348140	92.9953	38
83	109	5166428	5166428	5628398	36410555	39348140	92.5344	36
85	115	5167063	5167063	5624275	36217746	39348140	92.0444	40

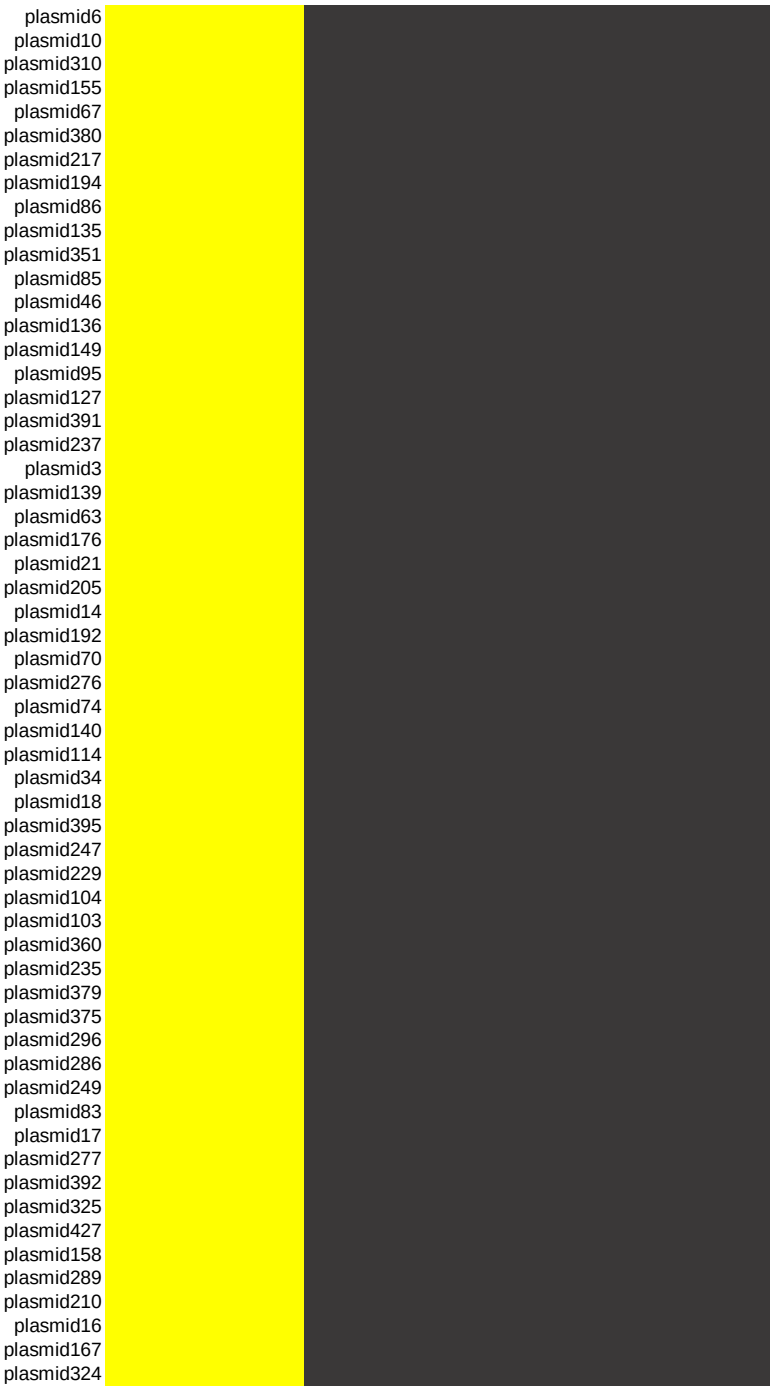
VP_IonTorrent_genome

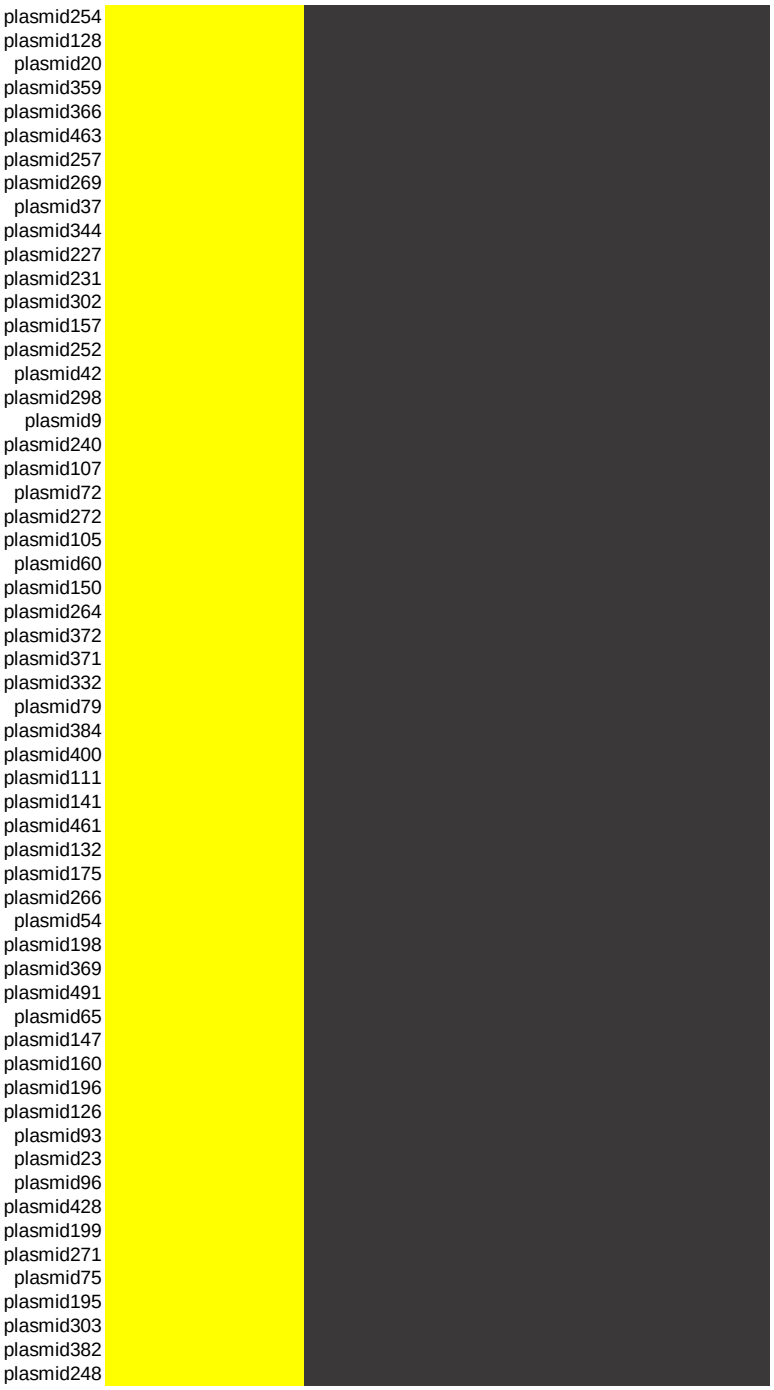
Hash	Nodes	N50	MAX	TOTAL BASES	READS USED	TOTAL READS	% used	CONTIGS
27	8197	2716	13419	5073781	25549532	26765491	95.457	2970
29	6684	2901	19891	5024350	25335172	26765491	94.6561	2773
31	5964	2922	17075	4986323	24999450	26765491	93.4018	2729
33	5270	2964	16013	4937943	24557102	26765491	91.7491	2719
35	4749	2809	17627	4857464	24094346	26765491	90.0202	2742
37	4325	2918	17629	4825867	23639079	26765491	88.3192	2639
39	3816	3100	18101	4795411	23204277	26765491	86.6948	2455
41	3220	3692	21263	4806384	22813949	26765491	85.2364	2071
43	2696	4654	21465	4815908	22383458	26765491	83.628	1708
45	2452	5395	30320	4844711	21963383	26765491	82.0586	1511
47	2944	3932	22973	4796875	21374949	26765491	79.8601	1922
49	4020	2206	13808	4634174	20633131	26765491	77.0886	3088
51	8441	7285	7285	4164787	18935695	26765491	70.7467	6883
53	4649	1357	1357	774878	3888964	26765491	14.5298	2093

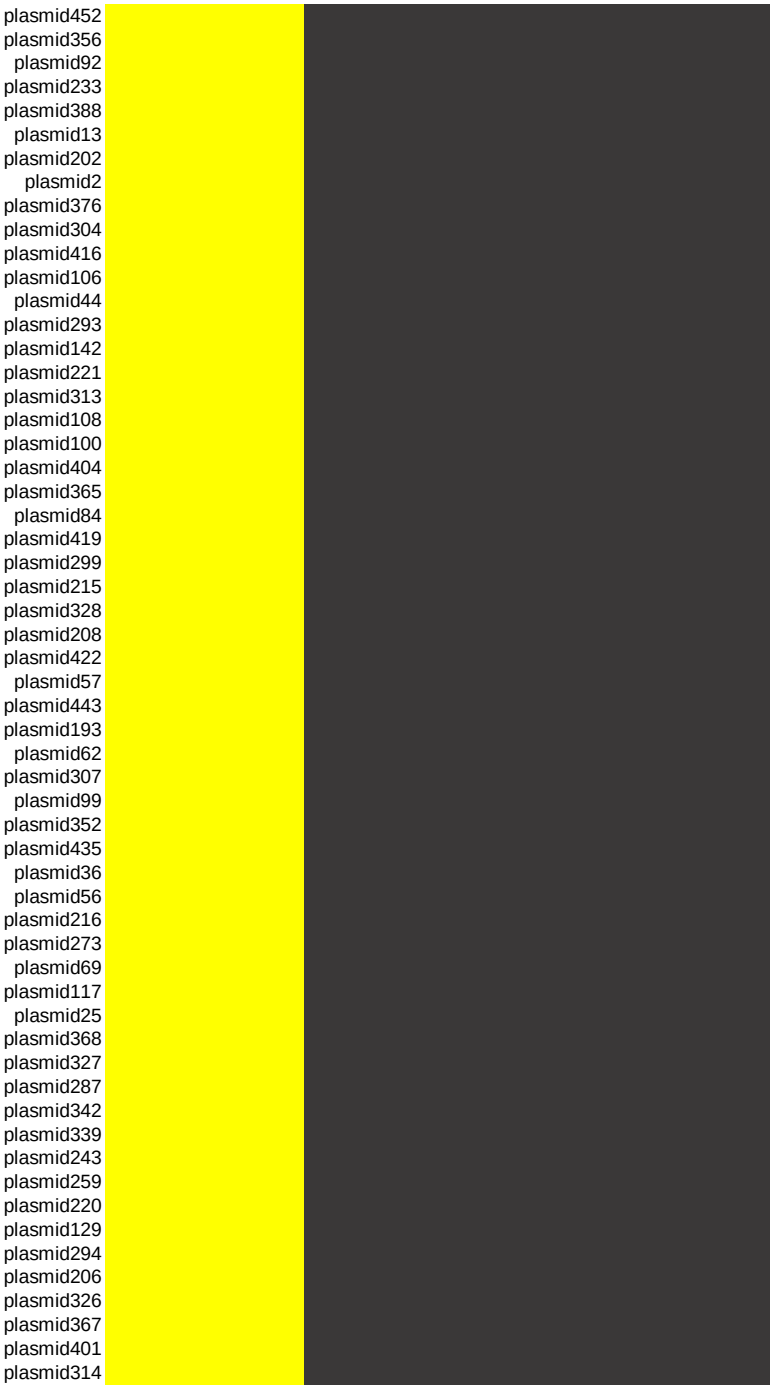
	EDENA
ILLUMINA	225
ILLUMINA_PLASMID	559
ION TORRENT	8451

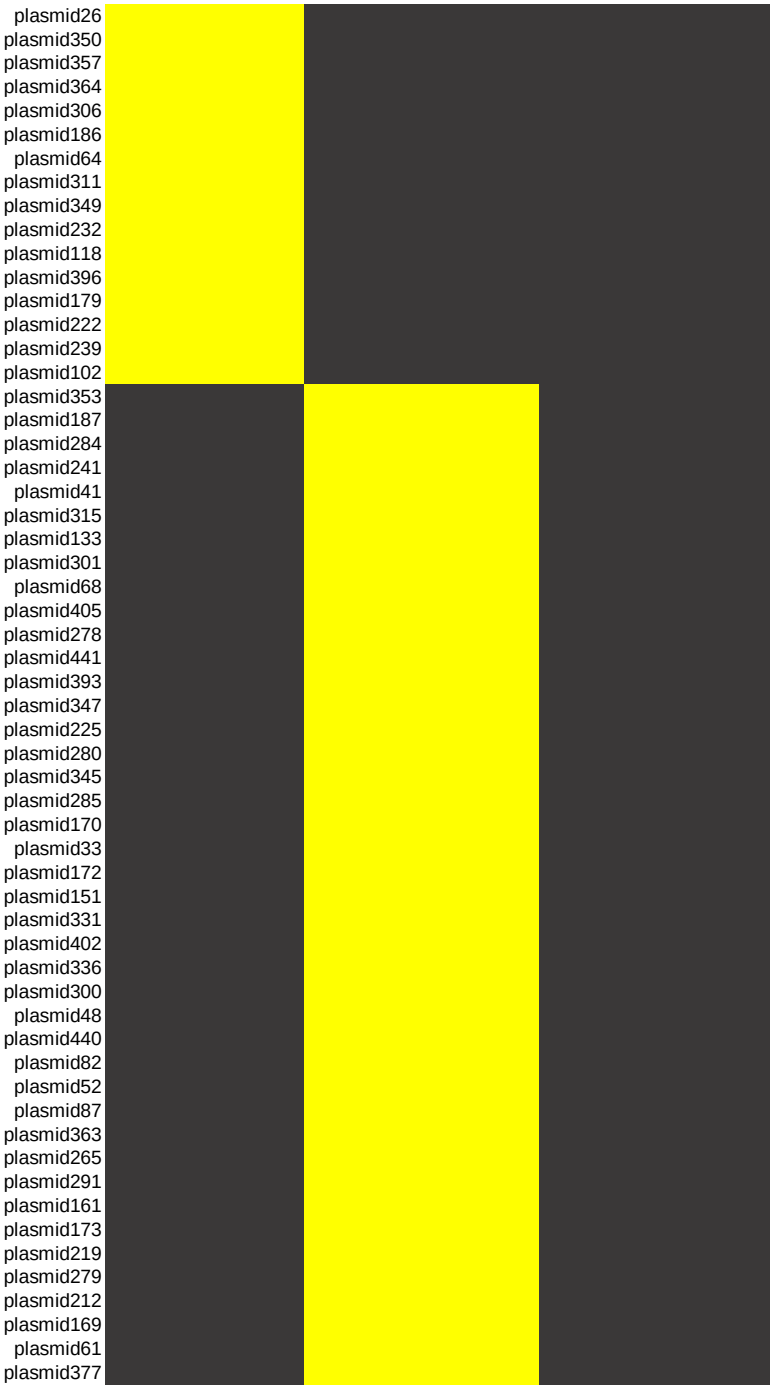
(c) Heatmap

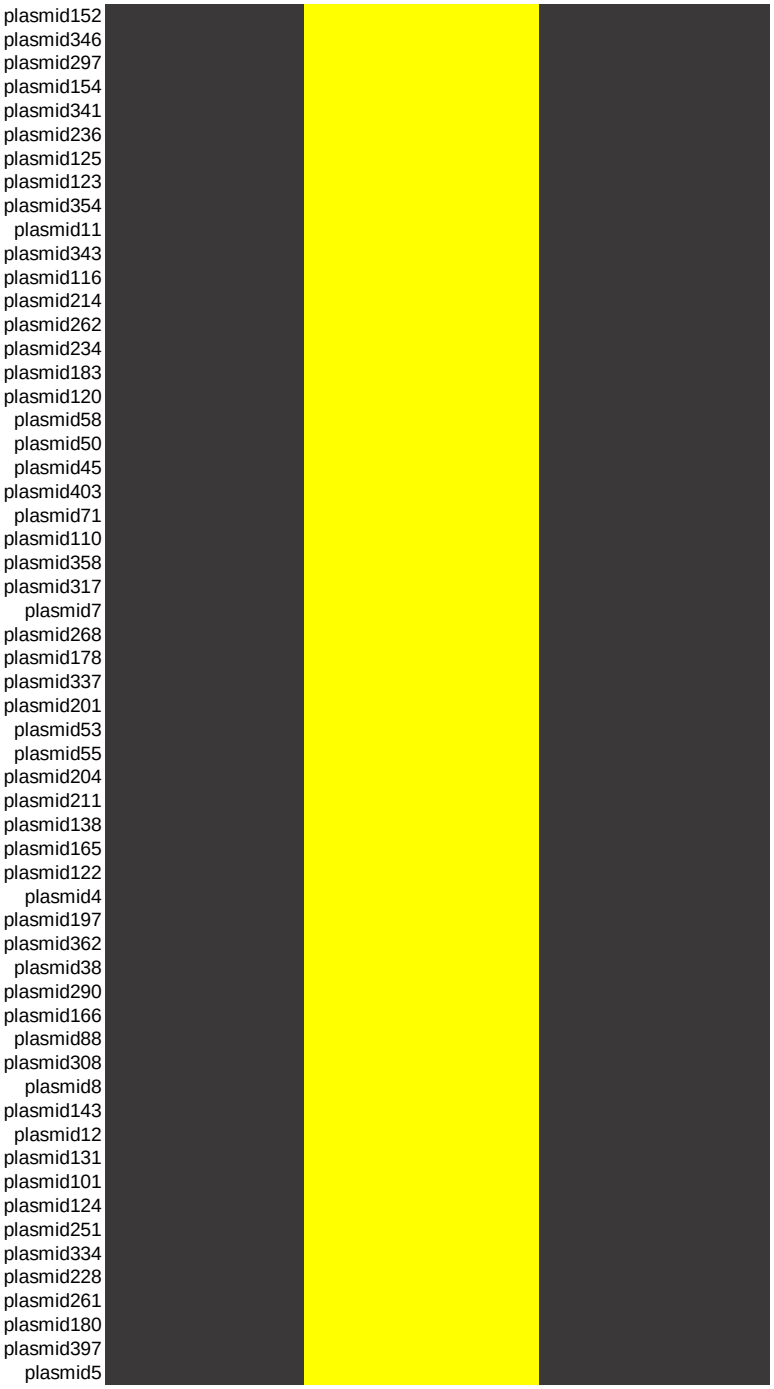

Edena contig name
Chromosome 1
Chromosome 2
Plasmid
plasmid171
plasmid288
plasmid318
plasmid312
plasmid182
plasmid113
plasmid43
plasmid267
plasmid181
plasmid32
plasmid394
plasmid319
plasmid121
plasmid78
plasmid30
plasmid1
plasmid260
plasmid109
plasmid244
plasmid275
plasmid119
plasmid98
plasmid238
plasmid408
plasmid28
plasmid76
plasmid274
plasmid59
plasmid134
plasmid250
plasmid330
plasmid164
plasmid338
plasmid168
plasmid27
plasmid246
plasmid281
plasmid39
plasmid174
plasmid322
plasmid188
plasmid90
plasmid162
plasmid137
plasmid320
plasmid130
plasmid370
plasmid115
plasmid73
plasmid22
plasmid200
plasmid148
plasmid329
plasmid209
plasmid47
plasmid153
plasmid19

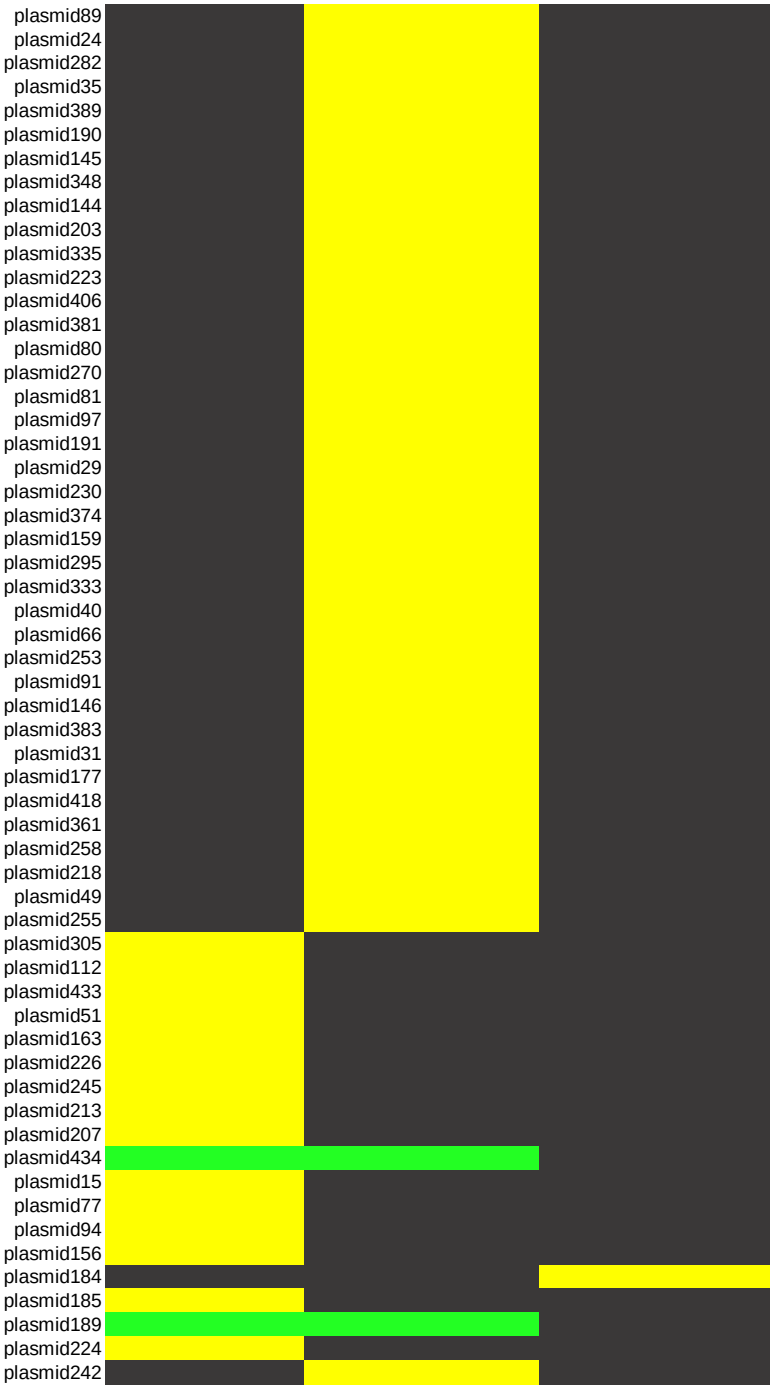

plasmid6
plasmid10
plasmid310
plasmid155
plasmid67
plasmid380
plasmid217
plasmid194
plasmid86
plasmid135
plasmid351
plasmid85
plasmid46
plasmid136
plasmid149
plasmid95
plasmid127
plasmid391
plasmid237
plasmid3
plasmid139
plasmid63
plasmid176
plasmid21
plasmid205
plasmid14
plasmid192
plasmid70
plasmid276
plasmid74
plasmid140
plasmid114
plasmid34
plasmid18
plasmid395
plasmid247
plasmid229
plasmid104
plasmid103
plasmid360
plasmid235
plasmid379
plasmid375
plasmid296
plasmid286
plasmid249
plasmid83
plasmid17
plasmid277
plasmid392
plasmid325
plasmid427
plasmid158
plasmid289
plasmid210
plasmid16
plasmid167
plasmid324


plasmid254
plasmid128
plasmid20
plasmid359
plasmid366
plasmid463
plasmid257
plasmid269
plasmid37
plasmid344
plasmid227
plasmid231
plasmid302
plasmid157
plasmid252
plasmid42
plasmid298
plasmid9
plasmid240
plasmid107
plasmid72
plasmid272
plasmid105
plasmid60
plasmid150
plasmid264
plasmid372
plasmid371
plasmid332
plasmid79
plasmid384
plasmid400
plasmid111
plasmid141
plasmid461
plasmid132
plasmid175
plasmid266
plasmid54
plasmid198
plasmid369
plasmid491
plasmid65
plasmid147
plasmid160
plasmid196
plasmid126
plasmid93
plasmid23
plasmid96
plasmid428
plasmid199
plasmid271
plasmid75
plasmid195
plasmid303
plasmid382
plasmid248


plasmid452
plasmid356
plasmid92
plasmid233
plasmid388
plasmid13
plasmid202
plasmid2
plasmid376
plasmid304
plasmid416
plasmid106
plasmid44
plasmid293
plasmid142
plasmid221
plasmid313
plasmid108
plasmid100
plasmid404
plasmid365
plasmid84
plasmid419
plasmid299
plasmid215
plasmid328
plasmid208
plasmid422
plasmid57
plasmid443
plasmid193
plasmid62
plasmid307
plasmid99
plasmid352
plasmid435
plasmid36
plasmid56
plasmid216
plasmid273
plasmid69
plasmid117
plasmid25
plasmid368
plasmid327
plasmid287
plasmid342
plasmid339
plasmid243
plasmid259
plasmid220
plasmid129
plasmid294
plasmid206
plasmid326
plasmid367
plasmid401
plasmid314

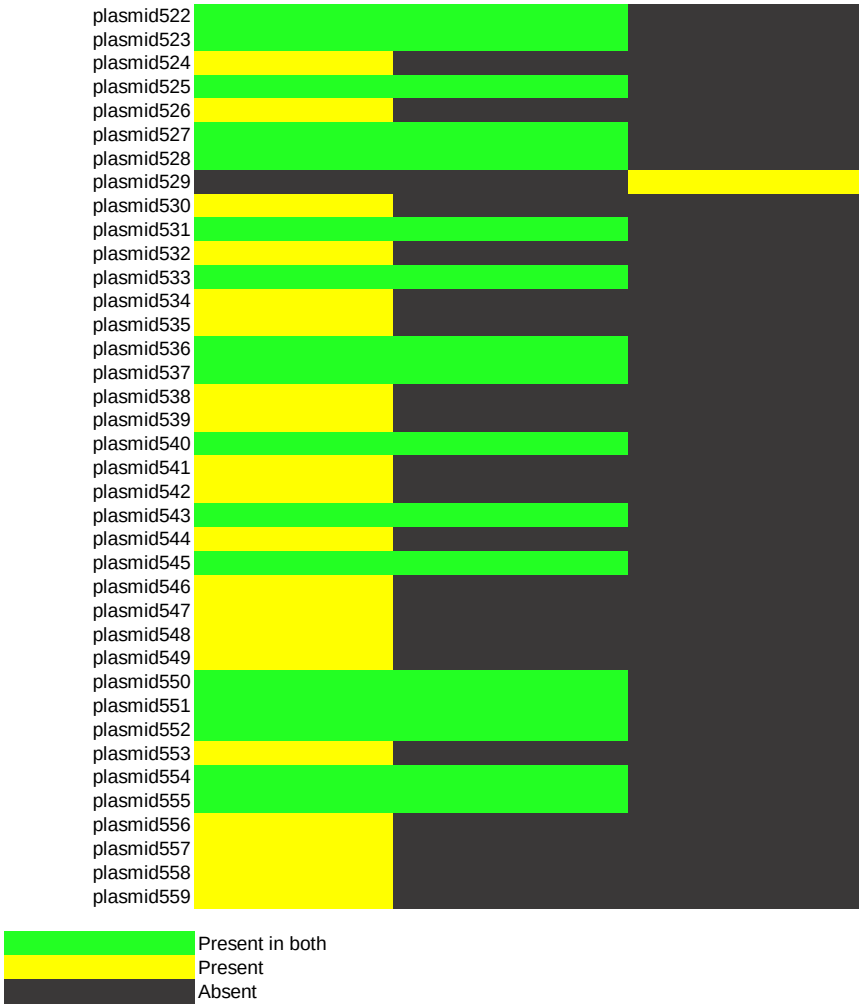

plasmid26
plasmid350
plasmid357
plasmid364
plasmid306
plasmid186
plasmid64
plasmid311
plasmid349
plasmid232
plasmid118
plasmid396
plasmid179
plasmid222
plasmid239
plasmid102
plasmid353
plasmid187
plasmid284
plasmid241
plasmid41
plasmid315
plasmid133
plasmid301
plasmid68
plasmid405
plasmid278
plasmid441
plasmid393
plasmid347
plasmid225
plasmid280
plasmid345
plasmid285
plasmid170
plasmid33
plasmid172
plasmid151
plasmid331
plasmid402
plasmid336
plasmid300
plasmid48
plasmid440
plasmid82
plasmid52
plasmid87
plasmid363
plasmid265
plasmid291
plasmid161
plasmid173
plasmid219
plasmid279
plasmid212
plasmid169
plasmid61
plasmid377


plasmid152
plasmid346
plasmid297
plasmid154
plasmid341
plasmid236
plasmid125
plasmid123
plasmid354
plasmid11
plasmid343
plasmid116
plasmid214
plasmid262
plasmid234
plasmid183
plasmid120
plasmid58
plasmid50
plasmid45
plasmid403
plasmid71
plasmid110
plasmid358
plasmid317
plasmid7
plasmid268
plasmid178
plasmid337
plasmid201
plasmid53
plasmid55
plasmid204
plasmid211
plasmid138
plasmid165
plasmid122
plasmid4
plasmid197
plasmid362
plasmid38
plasmid290
plasmid166
plasmid88
plasmid308
plasmid8
plasmid143
plasmid12
plasmid131
plasmid101
plasmid124
plasmid251
plasmid334
plasmid228
plasmid261
plasmid180
plasmid397
plasmid5


plasmid89
plasmid24
plasmid282
plasmid35
plasmid389
plasmid190
plasmid145
plasmid348
plasmid144
plasmid203
plasmid335
plasmid223
plasmid406
plasmid381
plasmid80
plasmid270
plasmid81
plasmid97
plasmid191
plasmid29
plasmid230
plasmid374
plasmid159
plasmid295
plasmid333
plasmid40
plasmid66
plasmid253
plasmid91
plasmid146
plasmid383
plasmid31
plasmid177
plasmid418
plasmid361
plasmid258
plasmid218
plasmid49
plasmid255
plasmid305
plasmid112
plasmid433
plasmid51
plasmid163
plasmid226
plasmid245
plasmid213
plasmid207
plasmid434
plasmid15
plasmid77
plasmid94
plasmid156
plasmid184
plasmid185
plasmid189
plasmid224
plasmid242

plasmid256
plasmid263
plasmid283
plasmid292
plasmid309
plasmid316
plasmid321
plasmid323
plasmid340
plasmid355
plasmid373
plasmid378
plasmid385
plasmid386
plasmid387
plasmid390
plasmid398
plasmid399
plasmid407
plasmid409
plasmid410
plasmid411
plasmid412
plasmid413
plasmid414
plasmid415
plasmid417
plasmid420
plasmid421
plasmid423
plasmid424
plasmid425
plasmid426
plasmid429
plasmid430
plasmid431
plasmid432
plasmid436
plasmid437
plasmid438
plasmid439
plasmid442
plasmid444
plasmid445
plasmid446
plasmid447
plasmid448
plasmid449
plasmid450
plasmid451
plasmid453
plasmid454
plasmid455
plasmid456
plasmid457
plasmid458
plasmid459
plasmid460

plasmid462
plasmid464
plasmid465
plasmid466
plasmid467
plasmid468
plasmid469
plasmid470
plasmid471
plasmid472
plasmid473
plasmid474
plasmid475
plasmid476
plasmid477
plasmid478
plasmid479
plasmid480
plasmid481
plasmid482
plasmid483
plasmid484
plasmid485
plasmid486
plasmid487
plasmid488
plasmid489
plasmid490
plasmid492
plasmid493
plasmid494
plasmid495
plasmid496
plasmid497
plasmid498
plasmid499
plasmid500
plasmid501
plasmid502
plasmid503
plasmid504
plasmid505
plasmid506
plasmid507
plasmid508
plasmid509
plasmid510
plasmid511
plasmid512
plasmid513
plasmid514
plasmid515
plasmid516
plasmid517
plasmid518
plasmid519
plasmid520
plasmid521


plasmid522
plasmid523
plasmid524
plasmid525
plasmid526
plasmid527
plasmid528
plasmid529
plasmid530
plasmid531
plasmid532
plasmid533
plasmid534
plasmid535
plasmid536
plasmid537
plasmid538
plasmid539
plasmid540
plasmid541
plasmid542
plasmid543
plasmid544
plasmid545
plasmid546
plasmid547
plasmid548
plasmid549
plasmid550
plasmid551
plasmid552
plasmid553
plasmid554
plasmid555
plasmid556
plasmid557
plasmid558
plasmid559
Present in both
Present
Absent

(d) Vibrio parahaemolyticus_Genomes

Organism	COUNTRY	DATA	HASH	Read Length	Read1 (TOTAL)	Read2 (TOTAL)	Contigs	BioProject
Vibrio parahaemolyticus RIMD 2210633	JAPAN	2chr					1	PRJNA57969, PRJNA360
Vibrio parahaemolyticus BB22OP	BANGLADESH	2chr					1	PRJNA184822, PRJNA170885
Vibrio parahaemolyticus 10329	US WEST COAST CLONAL	FASTA					33	PRJNA179736, PRJNA46533
Vibrio parahaemolyticus AN-5034	BANGLADESH	FASTA					54	PRJNA53489, PRJNA33625
Vibrio parahaemolyticus AQ3810	SINGAPORE	FASTA					1073	PRJNA54593, PRJNA18601
Vibrio parahaemolyticus AQ4037	MALDIVES	FASTA					164	PRJNA53507, PRJNA33623
Vibrio parahaemolyticus K5030	INDIA	FASTA					164	PRJNA53505, PRJNA36311
Vibrio parahaemolyticus PCV08-7	MALAYSIA	FASTA					83	PRJNA185026
Vibrio parahaemolyticus Peru-466	PERU	FASTA					149	PRJNA53491, PRJNA33621
Vibrio parahaemolyticus SNUVpS-1	SOUTH KOREA	FASTA					60	PRJNA199808, PRJNA175762
Vibrio parahaemolyticus v110	CHINA	FASTA					366	PRJNA202774, PRJNA192175
Vibrio parahaemolyticus 10290	USA	FQ	215	251	5815973	5815973	44	PRJNA176660
Vibrio parahaemolyticus 10296	USA	FQ	71	101	30452278	30452278	63	PRJNA176629
Vibrio parahaemolyticus 12310	USA	FQ	71	101	7733248	7733248	65	PRJNA176630
Vibrio parahaemolyticus 3256	USA	FQ	71	101	19990442	19990442	69	PRJNA176631
Vibrio parahaemolyticus 16		FASTA					179	PRJNA54741
Vibrio parahaemolyticus 3259	USA	FASTA					6	PRJNA176632
Vibrio parahaemolyticus 50	USA	FQ	71	101	13656176	13656176	75	PRJNA176636
Vibrio parahaemolyticus 605	USA	FQ	71	101	7682226	7682226	81	PRJNA176634
Vibrio parahaemolyticus 861	USA	FQ	71	101	14635449	14635449	82	PRJNA176635
Vibrio parahaemolyticus 949	USA	FASTA					7	PRJNA176633
Vibrio parahaemolyticus B-265	MOZAMBIQUE	FQ	71	101	14346823	14346823	95	PRJNA176641
Vibrio parahaemolyticus EKP-008	BANGLADESH	FQ	71	101	16167933	16167933	78	PRJNA176649
Vibrio parahaemolyticus EKP-021	BANGLADESH	FQ	71	101	4266320	4266320	94	PRJNA176651
Vibrio parahaemolyticus EKP-026	BANGLADESH	FQ	71	101	3126126	3126126	107	PRJNA176652
Vibrio parahaemolyticus EKP-028	BANGLADESH	FQ	71	101	8501574	8501574	82	PRJNA176647
Vibrio parahaemolyticus IDH02189	INDIA	FQ	71	101	17622605	17622605	76	PRJNA176646
Vibrio parahaemolyticus IDH02640	INDIA	FQ	71	101	12440606	12440606	79	PRJNA176645
Vibrio parahaemolyticus NIHCB0603	BANGLADESH	FASTA					13	PRJNA176648
Vibrio parahaemolyticus NIHCB0757	BANGLADESH	FASTA					4	PRJNA176650
Vibrio parahaemolyticus Peru-288	PERU	FQ	71	101	23874524	23874524	100	PRJNA176642
Vibrio parahaemolyticus SBR10290	USA	FQ	71	101	29536369	29536369	55	PRJNA176628
Vibrio parahaemolyticus V-223/04	CHILE	FQ	71	101	10975023	10975023	103	PRJNA176654
Vibrio parahaemolyticus V14/01	CHILE	FQ	71	101	17218094	17218094	72	PRJNA176640

Vibrio parahaemolyticus VP-48		FQ	71	101	23536100	23536100	77	PRJNA176639
Vibrio parahaemolyticus VP-NY4	INDIA	FASTA					14	PRJNA176653
Vibrio parahaemolyticus VP2007-007	USA	FQ	71	101	7263074	7263074	92	PRJNA176626
Vibrio parahaemolyticus VP2007-095	USA	FASTA					21	PRJNA176627
Vibrio parahaemolyticus VP232	INDIA	FQ	215	251	4977092	4977092	160	PRJNA176644
Vibrio parahaemolyticus VP250	INDIA	FASTA					5	PRJNA176643
Vibrio parahaemolyticus VPCR-2009		FQ	71	101	12855709	12855709	82	PRJNA176658
Vibrio parahaemolyticus VPCR-2010		FASTA					29	PRJNA176659
Vibrio parahaemolyticus VPTS-2009		FQ	71	101	13041083	13041083	74	PRJNA176656
Vibrio parahaemolyticus VPTS-2010		FQ	71	101	5794596	5794596	69	PRJNA176657
Vibrio parahaemolyticus VPTS-2010_2		FQ	71	101	19749696	19749696	82	PRJNA176655
Vibrio parahaemolyticus O1:K33	USA	FASTA					2	PRJNA203445
Vibrio parahaemolyticus TUMSAT_DE1_S1	THAILAND	FASTA					128	PRJDB1576
Vibrio parahaemolyticus TUMSAT_DE2_S2	THAILAND	FASTA					97	PRJDB1577
Vibrio parahaemolyticus TUMSAT_DO6_S3	THAILAND	FASTA					70	PRJDB1578
Vibrio parahaemolyticus TUMSAT_H01_S4	THAILAND	FASTA					71	PRJDB1579
Vibrio parahaemolyticus TUMSAT_H03_S5	THAILAND	FASTA					59	PRJDB1580
Vibrio parahaemolyticus TUMSAT_H10_S6	THAILAND	FASTA					65	PRJDB1581

(e) ClustalW command-line parameters

MULTIPLE ALIGNMENT VIEW

1. Do complete multiple alignment now Slow/Accurate
2. Produce guide tree file only
3. Do alignment using old guide tree file
4. Toggle Slow/Fast pairwise alignments = SLOW
5. Pairwise alignment parameters
6. Multiple alignment parameters
7. Reset gaps before alignment? = OFF
8. Toggle screen display = ON
9. Output format options

I. Iteration = NONE

S. Execute a system command

H. HELP

or press [RETURN] to go back to main menu

********* MULTIPLE ALIGNMENT PARAMETERS *********

1. Gap Opening Penalty :15.00
2. Gap Extension Penalty :6.66
3. Delay divergent sequences :30 %
4. DNA Transitions Weight :0.50
5. Protein weight matrix :Gonnet series
6. DNA weight matrix :IUB
7. Use negative matrix :OFF
8. Protein Gap Parameters

H. HELP

********* Format of Alignment Output *********

F. Toggle FASTA format output = OFF

1. Toggle CLUSTAL format output = ON

2. Toggle NBRF/PIR format output = OFF

3. Toggle GCG/MSF format output = OFF

4. Toggle PHYLIP format output = ON

5. Toggle NEXUS format output = ON

6. Toggle GDE format output = OFF

7. Toggle GDE output case = LOWER

8. Toggle CLUSTALW sequence numbers = OFF

9. Toggle output order = ALIGNED

0. Create alignment output file(s) now?

T. Toggle parameter output = OFF

R. Toggle sequence range numbers = OFF

H. HELP

(f) PCR sequencing results

(Highlighted area represents matching sequence of the gene)

gbpA – 1464bp

Forward sequencing product

CTGAGTGTCTTAGCACACGGTTACGTATCGGAATCCAACGATGGCGTCGCA

GCAAGTCGTGCAGCGCTGTGTAAGTACCCAACATCGGATACCAATGAAAGA

AACACTAACTGTGGCGCAATTCAATATGAACCACAAAGCGTGGAAGGCCCA

GATGGCTTCCCAGAGACAGGCCCTCGCGACGGGAAAATTGCCAGCGCAGAA

ACAGCTCGAGCAGCTGCACTAGACGAACAAACCGCCGACCGCTGGGTAAA

ACGTCCGATAAAATCTGGTACACAAACGTTTGAATGGACTTTTACAGCAAA

CCACGTGACTCGTGATTGGAAATACTACATCACAAAACCGAATTGGAACCC

GAACGCTTCATTGTCACGCGACTCGTTCGCTCTCAACCCTTTCTGTGTGGTT

GACGGCAATATGGTTCAACCACCAAAACAGATGAGCCACCAATGTAACGTG

CCTGAACGCGAAGGGTACCATGTCATCCTTGCGGTATGGGATGTGGGTGAT

ACCGCAGCATCATTCTACAACGTTATCGATGTCAAATTTGATGGTGACGACC

CTGTGATTCCTGAGTGGACTCAAGGTGGCCAAATCATCCCTACTATGAATCT

AAAGGTTGGCGATTCGGTTTACACTCGCGTGTTTGATCAATCTGGAGAGAA

CGTTGCATACCGCACTGAACTAGCAATTAGTAACGACGCACTCACTCAAGC

GAAAAACTGGTCTTACGCCTTGGCGAGTAAAGTCAACCAAGAGCAAACTAA

GCTTCAAGCTGGTCAATATTCGGAAGACAAGTTCACGCCAGTGTACGGCAC

GAACCCAATTTACCTGCAATCAAACAGTGGCCTTGAGCGTGTTGAAATTGG

CTACAATATTGAAACGCCAGTACCTGAGTATTCACTCACGGTTGATGGTCTT

GCCAGTGAGTACATAATTGCTACAGAACCAACCTCTTTAGACTCACTCTTAC

CGCAGAGGCGACTTAACCGCCGAGTACGGTTTACATCATCATCGCGACCAC

TAGCAGTGACGGTTCATTCAGATGTG

Reverse sequencing product

CTGGTTGCACTCTCACTCCACTGTACGCAGTAGCCAGAGTAAGGGAATGGT

TTACATTGATAAATCGCGCCGTCGCTTGCTAACACTTTTGTACCTGCTGTGT

AGCTAGAAAGCCCCTCTGGGAATACAAAGTCGTATTCGCCTGATGGTGGTG

GGGTGACTTCCTCACTTTTTAGATGGAAATCCAGAGTTTGCTGATCGACAAG

GTTGCCATCGGTATCTTTAATGCGGGTTACTAACATGTGGTGCCCCGGTTCT

GATTTGCTCAGCGTGAGTTCTACTTGTTCGCTTGCACCATCTTGAATTGAAC

CCGTCCAACTTGCTAGTGGTTCGCGATGATGATTGTAAACCGTTAACTCGGC

GGTTAAGTCGCCCTCTGCGGTAAGAGTGAGGTCTAAAGAGGTTGGTTCTGT

AGCAATTATGTACTCACTGGCAAGACCATCAACCGTGAGTGAATACTCAGG

TACTGGCGTTTCAATATTGTAGCCAATTTCAACACGCTCAAGGCCACTGTTT

GATTGCAGGTAAATTGGGTTCGTGCCGTACACTGGCGTGAACTTGTCTTCCG

AATATTGACCAGCTTGAAGCTTAGTTTGCTCTTGGTTGACTTTACTCGCCAA

GGCGTAAGACCAGTTTTTCGCTTGAGTGAGTGCGTCGTTACTAATTGCTAGT

TCAGTGCGGTATGCAACGTTCTCTCCAGATTGATCAAACACGCGAGTGTAA

ACCGAATCGCCAACCTTTAGATTCATAGTAGGGATGATTTGGCCACCTTGAG

TCCACTCAGGAATCACAGGGTCGTCACCATCAAATTTGACATCGATAACGTT

GTAGAATGATGCTGCGGTATCACCCACATCCCATACCGCAAGGATGACATG

GTACCCTTCGCGTTCATGCACGTTACATTGGTGGCTCATCTGTTTTGGTGGTT

GACATATTGCGTCACACACAGAAAGGGTGAGAGCGACGAGTCGCGTGACAT

GAAGCGTCGGGTTCATTCGTTTTGTGATGTAGTATTTCATCACGAGTCACGT

GATTGCTGTAAAGGTCA

dnaB – 1401bp

Forward sequencing product

ATAGGGTGGTCTGCGCTCGCAAGCGTGCCACCTCCCTCGGTAGAAGCACAG

GCTTCAGATGATTGGTGGTCTCTTACTGGACAATGAACGGAGGGATACCGT

TGCTGAGCGAGTTGTCGCCAGTGATTTCTATAGCCGTCCGCACCGTCTTATT

TTTGAAGGTGTGAAAACCATTCTTGAGTCGGGCAAACCGCTTGACCTTATCA

CTCTGTCTGAGCACCTAGAGCGACACGAGCAACTAGAAGATGTGGGTGGTT

TTGCTTACCTTGCTGACTTAGCGAAAAACACGCCCAGTGCTGCCAACATTAA

TGCCTACGCAGATATCGTGGCTGAACGTGCGATCGTCCGTGGTTTGATTGGT

GTTGCGAACGAAATTGCGGACGCGGGTTACGATCCTCAAGGTCGCAGTTCT

GAAGATTTGATAGACATGGCGGAGAGTAAAGTTTTTGCTATCGCCGAGTCT

CGCACCAGTGAAAACGAAGGTCCACAAAACGTCGACAACATCCTAGAAAA

GACGCTAGAACGTATCGAACTGCTGTACAAAACGCCGCAAGATGGCGTAAC

AGGTGTCGATACTGGTTTTACTGACCTCAACAAAAAGACCGCGGGTCTCCA

AGGTTCTGACTTAGTTATCGTGGCAGCGCGTCCATCGATGGGTAAAACCAC

CTTTGCGATGAACTTGTGTGAAAACGCGGCAATGCAGCAAGACAAACCCGT

TTTGATCTTCTCGCTAGAGATGCCGGCGGAACAGTTGATGATGCGTATGTTG

GCATCCTTGTCTCGCGTAGACCAAACCAAGATCCGTACGGGTCAGTTGGAT

GATGAAGACTGGGCACGTATTTCTTCTACCATGGGTATCCTGATGCAGAAG

AAAAATATGTACATCGATGACAGTTCAGGCTTAACGCCAACCGAAGTGCGT

TCACGTGCCCGTCGTGTTGCTCGTGAGCACGGAGGTCTATCAATGATCATGG

TCGACTACCTTCAGTTGATGCGCGTACCTTCGCTAACCGATAACCGTACTCT

TGAGATTGCCGAGATTTCTCGCTC

Reverse sequencing product

GGGGCCCGTATTTCAAGCGCGAGTGTTGTCCTTGGAATGTTAAACGCACCG

AGCCAATAGGACCGTTACGTTGTTTACCGATTTCTATTTCCGCCGTGCCTTT

GTATGGGCTATCTGGGTGATAGACCTCATCACGGTAGATGAACATGATCAA

GTCGGCATCCTGCTCGATGGAACCCGATTCACGCAAGTCCGAGTTAACAGG

TCGTTTATCCGCTCGTTGCTCTAGAGAACGGTTTAGCTGCGATAGTGCAACG

ACAGGCACGTTCAACTCTTTCGCCAGTGCTTTAAGCGAGCGAGAAATCTCG

GCAATCTCAAGAGTACGGTTATCGGTTAGCGAAGGTACGCGCATCAACTGA

AGGTAGTCGACCATGATCATTGATAGACCTCCGTGCTCACGAGCAACACGA

CGGGCACGTGAACGCACTTCGGTTGGCGTTAAGCCTGAACTGTCATCGATG

TACATATTTTTCTTCTGCATCAGGATACCCATGGTAGAAGAAATACGTGCCC

AGTCTTCATCATCCAACTGACCCGTACGGATCTTGGTTTGGTCTACGCGAGA

CAAGGATGCCAACATACGCATCATCAACTGTTCCGCCGGCATCTCTAGCGA

GAAGATCAAAACGGGTTTGTCTTGCTGCATTGCCGCGTTTTCACACAAGTTC

ATCGCAAAGGTGGTTTTACCCATCGATGGACGCGCTGCCACGATAACTAAG

TCAGAACCTTGGAGACCCGCGGTCTTTTTGTTGAGGTCAGTAAAACCAGTAT

CGACACCTGTTACGCCATCTTGCGGCGTTTTGTACAGCAGTTCGATACGTTC

TAGCGTCTTTTCTAGGATGTTGTCGACGTTTTGTGGACCTTCGTTTTCACTGG

TGCGAGACTCGGCGATAGCAAAAACTTTACTCTCCGCCATGTCTATCAAATC

TTCAGAACTGCGACCTTGAGGATCGTAACCCGCGTCCGCAATTTCGTTCGCA

ACACCAATCAAACCACGGACGATCGCACGTTCAGCCACGATATCTGCGTAG

GCATTAATGTTGGCAGCAC

deaD – 1930bp

Forward sequencing product

CTGATGAATCTATCCTTTCTGCTCTAGACGGAATGGGTTTCGTGTCACCAAA

CGGCCGTGGCCGTCGCACTCAAGACGCATAATTGGAAGGTGCTGACGCGCT

GGGTAAAGCACAGACGGGTACAGGTAAAACGGCAGCATTCTCTCTGCCTCT

TCTAAACAAGCTAGATCTTGATCAACGTAAACCACAAGCAATCGTTCTTGCT

CCAACTCGTGAGCTAGCGATCCAGGTTGCGGCTGAAATGAAAAACCTTGGT

AAAAACATCAAGGGTCTTAAAGTTCTAGAAATCTACGGTGGTGCATCTATC

GTAGATCAAATGCGTGCGCTTAAAAACGGTGCACACGTTGTTGTTGGTACTC

CTGGCCGTGTTCAAGACCTTATCAACCGTGAACGTCTACATCTTGACGAAGT

GAATACTTTCGTACTTGATGAAGCGGACGAAATGCTAAACATGGGTTTTGT

GGACGACGTAACTGCAATCATGGAGCACGCTCCTGAATCAGCACAACGCGT

TCTATTTTCTGCAACTATGCCACCAATGCTGAAAAACATTGTTGAGCGCTTC

CTACGTGATCCTGTAACTGTTGACGTTGCTGGTAAAAACCACACTGTAGACA

AAGTTCAACAACAGTTCTGGGTAGTGAAAGGTGTAGAGAAAGATGAAGCTA

TGTCTCGTCTTCTTGAAACTGAAGAAACTGACGCGTCAATCGTATTCGTACG

TACTCGTCAAGACACCGAGCGTCTAGCTGACTGGCTATCTGCACGTGGCTTC

AAAGCTGCTGCACTGCACGGTGACATTCCTCAGTCCCTACGTGAGCGTACTG

TTGATCACATCAAACAAGGTGTTATCGACATCCTAGTGGCAACTGACGTTGT

TGCTCGTGGTCTTGATGTTCCACGTATCACTCACGTATTTAACTACGACATC

CCATTCGATGTTGAGTCTTACATCCACCGTATCGGTCGTACTGGCCGTGCCG

GACGTAAAGGTAAAGCGATCCTTCTAGTTCGCACTAACCAATCCGTATGCTT

CGCACTATCGAGCGCGT

Reverse sequencing product

TGGGTACGCAACCCATTCCGTCTAGAGCAGAAAGGATAGAATCATTCTGAG

ATAAATCGCTGAATTGAATAACAGAATCTTGCATAGTTGCCGTCACGGTTAC

CACGGAAGCCACCTTCACGACGACCGCCTTCACGACGCTCACCACCACGGT

TGCCACCTTCGCGGTTACCACGGAAACCACCTTCACGACGACCGCCATCGC

GACGACCACCGTCACGGCGACCACCACGAGACTCGCGGAAATCGTCGAAGT

CACATACTACAGCACCAACATCTTGTTGACGGATACGCAGTTTGCTTAGCTT

GCTCGCAGCTTCAGAAGTCATTGCTTTTGGTAGCTGAACGAATGTGTGACCT

TGCGCTAGCTTGATCGCACCGATAGAACCTTTACCTAGGCCTAGTTCGTTTG

CTAGAGCGCCAACGATATCTTTAACTTGAACGCCTTGCTCGCGACCAACTTG

AAGCTGGTAAGTATCCCAGTCTTGGTTGTTGAAGCTACGACCGCCTTCACGA

CCACCTTCACGGCGTTCTTTACGACGCTGCTTGTCACGCTCAATTGCTTCAA

CCATTGGGTCTTCGCCAATGTAGAACAGAGGACGTTTACCTTGTTGACGCTT

AAGAAGCATTGCAGCAAGCATCGCAGGGTCGATTTCTAGAGAAGTTTGTAG

TTTCTCTACTAGCTCTGCGAACTTGTCTAGTGCTTTGTGCTCTTTCTCAGCTT

CCAGTTCAGCCGCTAGCTTGTTTAGACGAGCTTCTGCAACTTGGTCACGTAG

AGGAAGTTGGATTTCTTCCATAGATGACTTAGTCACGCGCTCGATAGTGCGA

AGCATACGGATTTGGTTAGTGCGAACTAGAAAGATCGCTTTACCTTTACGTC

CGGCACGGCCAGTACGACCGATACGGTGGATGTAAGACTCAACATCGAATG

GGATGTCGTAGTTAAATACGTGAGTGATACGTGGAACATCAAGACCACGAG

CAACAACGTCAGTTGCCACTAGGATGTCGATAACACCTTGTTTGATGTGATC

AACAGTACGCTCACGT

Sequencing lengths corresponding to gene products used for Real Time PCR

gbpA RT – 190bp

Forward sequencing product

GGGCATATGGGTTCACCACCAAAACAGATGAGCCACCAATGTAACGTGCCT
GAACGCGAAGGGTACCATGTCATCCTTGCGGTATGGGATGTGGGTGATACC
GCAGCATCATTCTACAACGTTATCGATGTCAAATTTGATGGTGACGACCCTG
TGA

Reverse sequencing product

TGCATGCTGCGGTATCACCCACATCCCATACCGCAAGGATGACATGGTACC
CTTCGCGTTCAGGCACGTTACATTGGTGGCTCATCTGTTTTGGTGGTTGAAC
CATATTGCCGTCAACCACACAGAAAGGGTTGAGAGCGAACGAGA

dnaB RT – 196bp

Forward sequencing product

GACTTCGTTGATGCGCGTACCTTCGCTAACCGATAACCGTACTCTTGAGATT
GCCGAGATTTCTCGCTCGCTTAAAGCACTGGCGAAAGAGTTGAACGTGCCT
GTCGTTGCACTATCGCAGCTAAACCGTTCTCTAGAGCAACGAGCGGATAAA
CGAC

Reverse sequencing product

ATCAGAGGACGTGCACTCAGGCACGTTCAACTCTTTCGCCAGTGCTTTAAGC
GAGCGAGAAATCTCGGCAATCTCAAGAGTACGGTTATCGGTTAGCGAAGGT
ACGCGCATCAACTGAAGGTAGTCGACCATGATCATTGATAGACTCCCGTGC
TCACGAAG

deaD RT – 181bp

Forward sequencing product

ATACCTTATCACCGTGAACGTCTACATCTTGACGAAGTGAATACTTTCGTAC

TTGATGAAGCGGACGAAATGCTAAACATGGGTTTTGTGGACGACGTAACTG

CAATCATGGAGCACGCTCCTGAATCAGCACAACGCGTTCTTGA

Reverse sequencing product

TGTCGTCCACAAAACCCATGTTTAGCATTTCGTCCGCTTCATCAAGTACGAA

AGTATTCACTTCGTCAAGATGTAGACGTTCACGGTTGATAAGGTCTTGAACA

CGGCCAGGAGTACCAACAACAACGTGTGCAC

(g) pGEM T-easy vector system

Ligation Protocol

1. Briefly centrifuge the pGEM®-T or pGEM®-T Easy Vector and Control Insert DNA tubes to collect the contents at the bottom of the tubes.

2. Set up ligation reactions as described below. Note: Use 0.5 ml tubes known to have low DNA-binding capacity. Vortex the 2X Rapid Ligation Buffer vigorously before each use.

3. Mix the reactions by pipetting. Incubate the reactions for 1 hour at room temperature. Alternatively, if maximum number of transformants are required, incubate the reactions overnight at 4°C.

Reaction Component	Standard Reaction	Positive Control	Background Control
2X Rapid Ligation Buffer, T4 DNA Ligase	5µl	5µl	5µl
pGEM®-T or pGEM®-T Easy Vector (50ng)	1µl	1µl	1µl
PCR product	Xµl*	–	–
Control Insert DNA	–	2µl	–
T4 DNA Ligase (3 Weiss units/µl)	1µl	1µl	1µl
nuclease-free water to a final volume of	**10µl**	**10µl**	**10µl**

*Molar ratio of PCR product:vector may require optimization.

4. An aliquot of the PCR reaction should be analysed on an agarose gel before use in the ligation reaction to verify that the reaction produced the desired product. The PCR product to be ligated can be gel-purified or purified directly from the PCR amplification. Clean-up of reactions prior to ligation is recommended to remove primer dimers or other undesired reaction products, and to improve ligation efficiency. Exposure of PCR products to shortwave ultraviolet light should be minimized in order to avoid the formation of pyrimidine dimers.

(h) Plasmid isolation sequencing results from 16 selected colonies

The sequencing has been done using the primers for the complete gene stretch of gbpA (1464bp) (Highlighted area represents matching sequence of the gene)

Colony 1

Forward sequencing product

AGACGGCCCGTGAATTGTAATACGACTCACTATAGGGCGAATTGGGCCCGA

CGTCGCATGCTCCCGGCCGCCATGGCGGCCGCGGGAATTCGATTTTAGTTCA

GTTTATCCCAAGCCATTTCCCAGTGTGAGCCTGTCGCTGGCTCAAATTGTGT

TGCACTCTCACTCCACTGTACGCAGTAGCCAGAGTAAGGGAATGGTTTACA

TTGATAAATCGCGCCGTCGCTTGCTAACACTTTTGTACCTGCTGTGTAGCTA

GAAAGCCCCTCTGGGAATACAAAGTCGTATTCGCCTGATGGTGGTGGGGTG

ACTTCCTCACTTTTTAGATGGAAATCCAGAGTTTGCTGATCGACAAGGTTGC

CATCGGTATCTTTAATGCGGGTTACTAACATGTGGTGCCCCGGTTCTGATTT

GCTCAGCGTGAGTTCTACTTGTTCGCTTGCACCATCTTGAATTGAACCCGTC

CAACTTGCTAGTGGTTCGCGATGATGATTGTAAACCGTTAACTCGGCGGTTA

AGTCGCCCTCTGCGGTAAGAGTGAGGTCTAAAGAGGTTGGTTCTGTAGCAA

TTATGTACTCACTGGCAAGACCATCAACCGTGAGTGAATACTCAGGTACTG

GCGTTTCAATATTGTAGCCAATTTCAACACGCTCAAGGCCACTGTTTGATTG

CAGGTAAATTGGGTTCGTGCCGTACACTGGCGTGAACTTGTCTTCCGAATAT

TGACCAGCTTGAAGCTTAGTTTGCTCTTGGTTGACTTTACTCGCCAAGGCGT

AAGACCAGTTTTTCGCTTGAGTGAGTGCGTCGTTACTAATTGCTAGTTCAGT

GCGGTATGCAACGTTCTCTCCAGATTGATCAAACACGCGAGTGTAAACCGA

ATCGCCAACCTTTAGATTCATAGTAGGGATGATTTGGCCACCTTGAGTCCAC

TCAGGAATCACAGGGTCGTCACCATCAATTTGACATCGATAACGTTGTAGA

ATGATGCTGCGGTATCACCCACATCCCATACCGCAAGGATGACATGGTACC

TTCGCGTTCAGGCCACGT

Reverse sequencing product

ATTTTTAGGTGACACTATAGAATACTCAAGCTATGCATCCAACGCGTTGGGA

GCTCTCCCATATGGTCGACCTGCAGGCGGCCGCGAATTCACTAGTGATTATG

AAATCATTCCCTAATAAATCGCTCGTTGCCCTTGCTATCGCAAGCATGAGCT

CTGGTGTCTTAGCACACGGTTACGTATCGGAATCCAACGATGGCGTCGCAG

CAAGTCGTGCAGCGCTGTGTAAGTACCCAACATCGGATACCAATGAAAGAA

ACACTAACTGTGGCGCAATTCAATATGAACCACAAAGCGTGGAAGGCCCAG

ATGGCTTCCCAGAGACAGGCCCTCGCGACGGGAAAATTGCCAGCGCAGAAA

CAGCTCGAGCAGCTGCACTAGACGAACAAACCGCCGACCGCTGGGTAAAAC

GTCCGATAAAATCTGGTACACAAACGTTTGAATGGACTTTTACAGCAAACC

ACGTGACTCGTGATTGGAAATACTACATCACAAAACCGAATTGGAACCCGA

ACGCTTCATTGTCACGCGACTCGTTCGCTCTCAACCCTTTCTGTGTGGTTGAC

GGCAATATGGTTCAACCACCAAAACAGATGAGCCACCAATGTAACGTGCCT

GAACGCGAAGGGTACCATGTCATCCTTGCGGTATGGGATGTGGGTGATACC

GCAGCATCATTCTACAACGTTATCGATGTCAAATTTGATGGTGACGACCCTG

TGATTCCTGAGTGGACTCAAGGTGGCCAAATCATCCCTACTATGAATCTAAA

GGTTGGCGATTCGGTTTACACTCGCGTGTTTGATCAATCTGGAGAGAACGTT

GCATACCGCACTGAACTAGCAATTAGTAACGACGCACTCACTCAAGCGAAA

AACTGGTCTTACGCCTTGGCGAGTAAAGTCAACCAAGAGCAAACTAAGCTT

CAAGCTGGTCAATATTCGGAAGACAAGTTCACGCCAGTGTACGGCACGAAC

CCAATTTACCTGCAATCAAACAGTGCCTTGAGCGTGTTGAAATTGGCTACAT

ATTGAACGCCAGTACCTGAGT

Colony 2

Forward sequencing product

CGGCCCGTGATTGTTAATACGACTCACTATAGGGCGAATTGGGCCCGACGT

CGCATGCTCCCGGCCGCCATGGCGGCCGCGGGAATTCGATTTTAGTTCAGTT

TATCCCAAGCCATTTCCCAGTGTGAGCCTGTCGCTGGCTCAAATTCTGTTGC

ACTCTCACTCCACTGTACGCAGTAGCCAGAGTAAGGGAATGGTTTACATTG

ATAAATCGCGCCGTCGCTTGCTAACACTTTTGTACCTGCTGTGTAGCTAGAA

AGCCCCTCTGGGAATACAAAGTCGTATTCGCCTGATGGTGGTGGGGTGACT

TCCTCACTTTTTAGATGGAAATCCAGAGTTTGCTGATCGACAAGGTTGCCAT

CGGTATCTTTAATGCGGGTTACTAACATGTGGTGCCCCGGTTCTGATTTGCT

CAGCGTGAGTTCTACTTGTTCGCTTGCACCATCTTGAATTGAACCCGTCCAA

CTTGCTAGTGGTTCGCGATGATGATTGTAAACCGTTAACTCGGCGGTTAAGT

CGCCCTCTGCGGTAAGAGTGAGGTCTAAAGAGGTTGGTTCTGTAGCAATTA

TGTACTCACTGGCAAGACCATCAACCGTGAGTGAATACTCAGGTACTGGCG

TTTCAATATTGTAGCCAATTTCAACACGCTCAAGGCCACTGTTTGATTGCAG

GTAAATTGGGTTCGTGCCGTACACTGGCGTGAACTTGTCTTCCGAATATTGA

CCAGCTTGAAGCTTAGTTTGCTCTTGGTTGACTTTACTCGCCAAGGCGTAAG

ACCAGTTTTTCGCTTGAGTGAGTGCGTCGTTACTAATTGCTAGTTCAGTGCG

GTATGCAACGTTCTCTCCAGATTGATCAAACACGCGAGTGTAAACCGAATC

GCCAACCTTTAGATTCATAGTAGGGATGATTTGGCCACCTTTGAGTCCACTC

AGGAATCACAGGGTCGTCACCATCAAATTTGACATCGATAACGTTGTAGAA

TGATGCTGCGGTATCCCCCCACATCCCATACCGCAAGGATGACATGGTACC

CTTTCGCGTTCAGGCACG

Reverse sequencing product

CCACTGATTACGCCAAGCTATTTAGGTGACACTATAGAATACTCAAGCTATG

CATCCAACGCGTTGGGAGCTCTCCCATATGGTCGACCTGCAGGCGGCCGCG

AATTCACTAGTGATTATGAAATCATTCCCTAATAAATCGCTCGTTGCCCTTG

CTATCGCAAGCATGAGCTCTGGTGTCTTAGCACACGGTTACGTATCGGAATC

CAACGATGGCGTCGCAGCAAGTCGTGCAGCGCTGTGTAAGTACCCAACATC

GGATACCAATGAAAGAAACACTAACTGTGGCGCAATTCAATATGAACCACA

AAGCGTGGAAGGCCCAGATGGCTTCCCAGAGACAGGCCCTCGCGACGGGA

AAATTGCCAGCGCAGAAACAGCTCGAGCAGCTGCACTAGACGAACAAACC

GCCGACCGCTGGGTAAAACGTCCGATAAAATCTGGTACACAAACGTTTGAA

TGGACTTTTACAGCAAACCACGTGACTCGTGATTGGAAATACTACATCACA

AAACCGAATTGGAACCCGAACGCTTCATTGTCACGCGACTCGTTCGCTCTCA

ACCCTTTCTGTGTGGTTGACGGCAATATGGTTCAACCACCAAAACAGATGA

GCCACCAATGTAACGTGCCTGAACGCGAAGGGTACCATGTCATCCTTGCGG

TATGGGATGTGGGTGATACCGCAGCATCATTCTACAACGTTATCGATGTCAA

ATTTGATGGTGACGACCCTGTGATTCCTGAGTGGACTCAAGGTGGCCAAAT

CATCCCTACTATGAATCTAAAGGTTGGCGATTCGGTTTACACTCGCGTGTTT

GATCAATCTGGAGAGAACGTTGCATACCGCACTGAACTAGCAATTAGTAAC

GACGCACTCACTCAAGCGAAAAACTGGTCTTACGCCTTGGCGAGTAAAGTC

AACCAAGAGCAAACTAAGCTTCAAGCTGGTCAATATTCGGAAGACAAGTTC

ACGCCAGTGTACGGCACGAACCCAATTTACCTGCAATCAAACAGTGGCCTT

GAGCGTGTGAATTTGGCTACAATATT

Colony 3

Forward sequencing product

TGGCCGTGATTTTGTAATACGACTCACTATAGGGCGAATTGGGCCCGACGTC

GCATGCTCCCGGCCGCCATGGCGGCCGCGGGAATTCGATTTTAGTTCAGTTT

ATCCCAAGCCATTTCCCAGTGTGAGCCTGTCGCTGGCTCAAATTGTGTTGCA

CTCTCACTCCACTGTACGCAGTAGCCAGAGTAAGGGAATGGTTTACATTGAT

AAATCGCGCCGTCGCTTGCTAACACTTTTGTACCTGCTGTGTAGCTAGAAAG

CCCCTCTGGGAATACAAAGTCGTATTCGCCTGATGGTGGTGGGGTGACTTCC

TCACTTTTTAGATGGAAATCCAGAGTTTGCTGATCGACAAGGTTGCCATCGG

TATCTTTAATGCGGGTTACTAACATGTGGTGCCCCGGTTCTGATTTGCTCAG

CGTGAGTTCTACTTGTTCGCTTGCACCATCTTGAATTGAACCCGTCCAACTT

GCTAGTGGTTCGCGATGATGATTGTAAACCGTTAACTCGGCGGTTAAGTCGC

CCTCTGCGGTAAGAGTGAGGTCTAAAGAGGTTGGTTCTGTAGCAATTATGT

ACTCACTGGCAAGACCATCAACCGTGAGTGAATACTCAGGTACTGGCGTTT

CAATATTGTAGCCAATTTCAACACGCTCAAGGCCACTGTTTGATTGCAGGTA

AATTGGGTTCGTGCCGTACACTGGCGTGAACTTGTCTTCCGAATATTGACCA

GCTTGAAGCTTAGTTTGCTCTTGGTTGACTTTTACTCGCCAAGGCGTAAGAC

CAGTTTTTCGCTTGAGTGAGTGCGTCGTTACTAATTGCTAGTTCAGTGCGGT

ATGCAACGTTCTCTCCAGATTGATCAAACACGCGAGTGTAAACCGAATCGC

CAACCTTTAGATTCATAGTAGGGATGATTTGGGCCACCTTGAGTCCACTCAG

GAATCACAGGGTCGTCACCATCAATTTGACATCGATAACGTTGTAGAATTG

ATGCTGCGGTATCCACCCAACAATCCCATACCGCAAGGAATGACATTGGTA

ACCCTTTCGCGTTCA

Reverse sequencing product

TTGACTTGATTCGCCAAGCTATTTAGGTGACACTATAGAATACTCAAGCTAT

GCATCCAACGCGTTGGGAGCTCTCCCATATGGTCGACCTGCAGGCGGCCGC

GAATTCACTAGTGATTATGAAATCATTCCCTAATAAATCGCTCGTTGCCCTT

GCTATCGCAAGCATGAGCTCTGGTGTCTTAGCACACGGTTACGTATCGGAAT

CCAACGATGGCGTCGCAGCAAGTCGTGCAGCGCTGTGTAAGTACCCAACAT

CGGATACCAATGAAAGAAACACTAACTGTGGCGCAATTCAATATGAACCAC

AAAGCGTGGAAGGCCCAGATGGCTTCCCAGAGACAGGCCCTCGCGACGGG

AAAATTGCCAGCGCAGAAACAGCTCGAGCAGCTGCACTAGACGAACAAAC

CGCCGACCGCTGGGTAAAACGTCCGATAAAATCTGGTACACAAACGTTTGA

ATGGACTTTTACAGCAAACCACGTGACTCGTGATTGGAAATACTACATCAC

AAAACCGAATTGGAACCCGAACGCTTCATTGTCACGCGACTCGTTCGCTCTC

AACCCTTTCTGTGTGGTTGACGGCAATATGGTTCAACCACCAAAACAGATG

AGCCACCAATGTAACGTGCCTGAACGCGAAGGGTACCATGTCATCCTTGCG

GTATGGGATGTGGGTGATACCGCAGCATCATTCTACAACGTTATCGATGTCA

AATTTGATGGTGACGACCCTGTGATTCCTGAGTGGACTCAAGGTGGCCAAA

TCATCCCTACTATGAATCTAAAGGTTGGCGATTCGGTTTACACTCGCGTGTT

TGATCAATCTGGAGAGAACGTTGCATACCGCACTGAACTAGCAATTAGTAA

CGACGCACTCACTCAAGCGAAAAACTGGTCTTACGCCTTGCCGAGTAAAGT

CAACCAAGAGCAAACTAGCTTCAAGCTGGTCAATATTCGGAAGACAAGTTC

ACGCCAGTGTACGGCACGAACCCAATTTACCTGCAATCAAACAGTGGGGTT

GAAGCGTGTTGAAATTGGCTAACAAT

Colony 4

Forward sequencing product

CGTTCCGGTGGGAATTGTAATACGACTCACTATAGGGCGAATTGGGCCCGA

CGTCGCATGCTCCCGGCCGCCATGGCGGCCGCGGGAATTCGATTTTAGTTCA

GTTTATCCCAAGCCATTTCCCAGTGTGAGCCTGTCGCTGGCTCAAATTTTGT

TGCACTCTCACTCCACTGTACGCAGTAGCCAGAGTAAGGGAATGGTTTACA

TTGATAAATCGCGCCGTCGCTTGCTAACACTTTTGTACCTGCTGTGTAGCTA

GAAAGCCCCTCTGGGAATACAAAGTCGTATTCGCCTGATGGTGGTGGGGTG

ACTTCCTCACTTTTTAGATGGAAATCCAGAGTTTGCTGATCGACAAGGTTGC

CATCGGTATCTTTAATGCGGGTTACTAACATGTGGTGCCCCGGTTCTGATTT

GCTCAGCGTGAGTTCTACTTGTTCGCTTGCACCATCTTGAATTGAACCCGTC

CAACTTGCTAGTGGTTCGCGATGATGATTGTAAACCGTTAACTCGGCGGTTA

AGTCGCCCTCTGCGGTAAGAGTGAGGTCTAAAGAGGTTGGTTCTGTAGCAA

TTATGTACTCACTGGCAAGACCATCAACCGTGAGTGAATACTCAGGTACTG

GCGTTTCAATATTGTAGCCAATTTCAACACGCTCAAGGCCACTGTTTGATTG

CAGGTAAATTGGGTTCGTGCCGTACACTGGCGTGAACTTGTCTTCCGAATAT

TGACCAGCTTGAAGCTTAGTTTGCTCTTGGTTGACTTTACTCGCCAAGGCGT

AAGACCAGTTTTTCGCTTGAGTGAGTGCGTCGTTACTAATTGCTAGTTCAGT

GCGGTATGCAACGTTCTCTCCAGATTGATCAAACACGCGAGTGTAAACCGA

ATCGCCAACCTTTAGATTCATAGTAGGGATGATTTGGCCAACTTGAGTCCAC

TCAGGAATCACAGGGTCGTCACCATCAAATTGACATCCGATAACGTTGTAG

AATGATGCTGCGGTATCACCCACATCCCATACGCAGGGATGACATGGTAAC

CTTCGCGTTCAGGCACGT

Reverse sequencing product

CGCCCAAGGCTATTTTAGGTGACACTATAGAATACTCAAGCTATGCATCCA

ACGCGTTGGGAGCTCTCCCATATGGTCGACCTGCAGGCGGCCGCGAATTCA

CTAGTGATTATGAAATCATTCCCTAATAAATCGCTCGTTGCCCTTGCTATCG

CAAGCATGAGCTCTGGTGTCTTAGCACACGGTTACGTATCGGAATCCAACG

ATGGCGTCGCAGCAAGTCGTGCAGCGCTGTGTAAGTACCCAACATCGGATA

CCAATGAAAGAAACACTAACTGTGGCGCAATTCAATATGAACCACAAAGCG

TGGAAGGCCCAGATGGCTTCCCAGAGACAGGCCCTCGCGACGGGAAAATTG

CCAGCGCAGAAACAGCTCGAGCAGCTGCACTAGACGAACAAACCGCCGAC

CGCTGGGTAAAACGTCCGATAAAATCTGGTACACAAACGTTTGAATGGACT

TTTACAGCAAACCACGTGACTCGTGATTGGAAATACTACATCACAAAACCG

AATTGGAACCCGAACGCTTCATTGTCACGCGACTCGTTCGCTCTCAACCCTT

TCTGTGTGGTTGACGGCAATATGGTTCAACCACCAAAACAGATGAGCCACC

AATGTAACGTGCCTGAACGCGAAGGGTACCATGTCATCCTTGCGGTATGGG

ATGTGGGTGATACCGCAGCATCATTCTACAACGTTATCGATGTCAAATTTGA

TGGTGACGACCCTGTGATTCCTGAGTGGACTCAAGGTGGCCAAATCATCCCT

ACTATGAATCTAAAGGTTGGCGATTCGGTTTACACTCGCGTGTTTGATCAAT

CTGGAGAGAACGTTTGCATACCGCACTGAACTAGCAATTAGTAACGACGCA

CTCACTCAAGCGAAAAACTGGTCTTACGCCTTGGCGAGTAAAGTCAACCAA

GAGCAAACTAAGCTTCAAGCTGGGTCAATATTTCGGAAGACAAGTTCACGC

CAGTGTACGGCACGTACCCCAATTTTACCTGCAATCAAACCAGTTGCCCTTG

AGCCGTGTTGAAATTTGGCTTACAA

Colony 5

Forward sequencing product

ACGGCCAGGATTGTATACGACTCACTATAGGGCGAATTGGGCCCGACGTCG

CATGCTCCCGGCCGCCATGGCGGCCGCGGGAATTCGATTTAGTTCAGTTTAT

CCCAAGCCATTTCCCAGTGTGAGCCTGTCGCTGGCTCAAATTATGTTGCACT

CTCACTCCACTGTACGCAGTAGCCAGAGTAAGGGAATGGTTTACATTGATA

AATCGCGCCGTCGCTTGCTAACACTTTTGTACCTGCTGTGTAGCTAGAAAGC

CCCTCTGGGAATACAAAGTCGTATTCGCCTGATGGTGGTGGGGTGACTTCCT

CACTTTTTAGATGGAAATCCAGAGTTTGCTGATCGACAAGGTTGCCATCGGT

ATCTTTAATGCGGGTTACTAACATGTGGTGCCCCGGTTCTGATTTGCTCAGC

GTGAGTTCTACTTGTTCGCTTGCACCATCTTGAATTGAACCCGTCCAACTTG

CTAGTGGTTCGCGATGATGATTGTAAACCGTTAACTCGGCGGTTAAGTCGCC

CTCTGCGGTAAGAGTGAGGTCTAAAGAGGTTGGTTCTGTAGCAATTATGTA

CTCACTGGCAAGACCATCAACCGTGAGTGAATACTCAGGTACTGGCGTTTC

AATATTGTAGCCAATTTCAACACGCTCAAGGCCACTGTTTGATTGCAGGTAA

ATTGGGTTCGTGCCGTACACTGGCGTGAACTTGTCTTCCGAATATTGACCAG

CTTGAAGCTTAGTTTGCTCTTGGTTGACTTTACTCGCCAAGGCGTAAGACCA

GTTTTTCGCTTGAGTGAGTGCGTCGTTACTAATTGCTAGTTCAGTGCGGTAT

GCAACGTTCTCTCCAGATTGATCAAACACGCGAGTGTAAACCGAATCGCCA

ACCTTTAGATTCATAGTAGGGATGATTTGGGCCACCTTGAGTCCACTCAGGA

ATCACAGGGTCGTCACCATCAAATTTGACATCGATAACGTTGTAGATGATG

CTGCGGTATCACCCACATCCCATACCGCAAGGATGACAATGGTAACCCTTTC

GCGTTCAGCACGTTAC

Reverse sequencing product

ATTTTTAAGGTTGACACTATAGAATACTCAAGCTATGCATCCAACGCGTTGG

GAGCTCTCCCATATGGTCGACCTGCAGGCGGCCGCGAATTCACTAGTGATT

ATGAAATCATTCCCTAATAAATCGCTCGTTGCCCTTGCTATCGCAAGCATGA

GCTCTGGTGTCTTAGCACACGGTTACGTATCGGAATCCAACGATGGCGTCGC

AGCAAGTCGTGCAGCGCTGTGTAAGTACCCAACATCGGATACCAATGAAAG

AAACACTAACTGTGGCGCAATTCAATATGAACCACAAAGCGTGGAAGGCCC

AGATGGCTTCCCAGAGACAGGCCCTCGCGACGGGAAAATTGCCAGCGCAGA

AACAGCTCGAGCAGCTGCACTAGACGAACAAACCGCCGACCGCTGGGTAA

AACGTCCGATAAAATCTGGTACACAAACGTTTGAATGGACTTTTACAGCAA

ACCACGTGACTCGTGATTGGAAATACTACATCACAAAACCAAATTGGAACC

CGAACGCTTCATTGTCACGCGACTCGTTCGCTCTCAACCCTTTCTGTGTGGTT

GACGGCAATATGGTTCAACCACCAAAACAGATGAGCCACCAATGTAACGTG

CCTGAACGCGAAGGGTACCATGTCATCCTTGCGGTATGGGATGTGGGTGAT

ACCGCAGCATCATTCTACAACGTTATCGATGTCAAATTTGATGGTGACGACC

CTGTGATTCCTGAGTGGACTCAAGGTGGCCAAATCATCCCTACTATGAATCT

AAAGGTTGGCGATTCGGTTTACACTCGCGTGTTTGATCAATCTGGAGAGAA

CGTTGCATACCGCACTGAACTAGCAATTAGTAACGACGCACTCACTCAAGC

GAAAACTGGTCTTACGCCTTGGCGAGTAAAGTCAACCAAGAGCAAACTAAG

CTTCAAGCTGGTCAATATTCGGAAGACAAGTTCCACGCCAGTGTACGGCAC

GAACCCAAATTTACCCTGCAATCAAACAGTGGCCTTGAGCGTGTTGGAAAT

TGGGCTACAATATTGAAACGCCAG

Colony 6

Forward sequencing product

ATGCCCGTGGATTTGTTATACGACTCACTATAGGGCGAATTGGGCCCGACGT

CGCATGCTCCCGGCCGCCATGGCGGCCGCGGGAATTCGATTATGAAATCAT

TCCCTAATAAATCGCTCGTTGCCCTTGCTATCGCAAGCATGAGCTCTGGTGT

CTTAGCACACGGTTACGTATCGGAATCCAACGATGGCGTCGCAGCAAGTCG

TGCAGCGCTGTGTAAGTACCCAACATCGGATACCAATGAAAGAAACACTAA

CTGTGGCGCAATTCAATATGAACCACAAAGCGTGGAAGGCCCAGATGGCTT

CCCAGAGACAGGCCCTCGCGACGGGAAAATTGCCAGCGCAGAAACAGCTC

GAGCAGCTGCACTAGACGAACAAACCGCCGACCGCTGGGTAAAACGTCCG

ATAAAATCTGGTACACAAACGTTTGAATGGACTTTTACAGCAAACCACGTG

ACTCGTGATTGGAAATACTACATCACAAAACCGAATTGGAACCCGAACGCT

TCATTGTCACGCGACTCGTTCGCTCTCAACCCTTTCTGTGTGGTTGACGGCA

ATATGGTTCAACCACCAAAACAGATGAGCCACCAATGTAACGTGCCTGAAC

GCGAAGGGTACCATGTCATCCTTGCGGTATGGGATGTGGGTGATACCGCAG

CATCATTCTACAACGTTATCGATGTCAAATTTGATGGTGACGACCCTGTGAT

TCCTGAGTGGACTCAAGGTGGCCAAATCATCCCTACTATGAATCTAAAGGTT

AGCGATTCGGTTTACACTCGCGTGTTTGATCAATCTGGAGAGAACGTTGCAT

ACCGCACTGAACTAGCAATTAGTAACGACGCACTCACTCAAGCGAAAAACT

GGTCTTACGCCTTGGCGAGTAAAGTCAACCAAGAGCAAACTAAGCTTCAAG

CTGGTCAATATTCGGAAGACAGTTCACGCCAGTGTACGGCACGAAACCCAA

TTTACCTGCAATCAAACAGTGGCCCTGAGCGTGTTGAAATTGGCTACAATTA

TTGAAACGCCAGTACCTGAGTATTC

Reverse sequencing product

AGGCCCCAAGGCTATTTTAGGTGACACTATAGAATACTCAAGCTATGCATC

CAACGCGTTGGGAGCTCTCCCATATGGTCGACCTGCAGGCGGCCGCGAATT

CACTAGTGATTTAGTTCAGTTTATCCCAAGCCATTTCCCAGTGTGAGCCTGT
CGCTGGCTCAAATTATGTTGCACTCTCACTCCACTGTACGCAGTAGCCAGAG
TAAGGGAATGGTTTACATTGATAAATCGCGCCGTCGCTTGCTAACACTTTTG
TACCTGCTGTGTAGCTAGAAAGCCCCTCTGGGAATACAAAGTCGTATTCGCC
TGATGGTGGTGGGGTGACTTCCTCACTTTTTAGATGGAAATCCAGAGTTTGC
TGATCGACAAGGTTGCCATCGGTATCTTTAATGCGGGTTACTAACATGTGGT
GCCCCGGTTCTGATTTGCTCAGCGTGAGTTCTACTTGTTCGCTTGCACCATCT
TGAATTGAACCCGTCCAACTTGCTAGTGGTTCGCGATGATGATTGTAAACCG
TTAACTCGGCGGTTAAGTCGCCCTCTGCGGTAAGAGTGAGGTCTAAAGAGG
TTGGTTCTGTAGCAATTATGTACTCACTGGCAAGACCATCAACCGTGAGTGA
ATACTCAGGTACTGGCGTTTCAATATTGTAGCCAATTTCAACACGCTCAAGG
CCACTGTTTGATTGCAGGTAAATTGGGTTCGTGCCGTACACTGGCGTGAACT
TGTCTTCCGAATATTGACCAGCTTGAAGCTTAGTTTGCTCTTGGTTGACTTTA
CTCGCCAAGGCGTAAGACCAGTTTTTCGCTTGAGTGAGTGCGTCGTTACTAA
TTGCTAGTTCAGTGCGGGTATGCAACGTTCTCTCCAGATTGATCAAACACGC
GAGTGTAAACCGGATCGCTAACCTTTAGATTCATAGTAGGGATGATTTGGG
CCACCTTGAGTCCACTCAGGAATCACAGGGTCGTCACCATCAATTTGACATC
GATAACGGTTGTAGATTGATGCTGCGGTATCCCCCCACATCCCATACCGCAG
GGATGACATTGC

Colony 7

Forward sequencing product

GGGCCCGTGCATTGTATACGACTCACTATAGGGCGAATTGGGCCCGACGTC
GCATGCTCCCGGCCGCCATGGCGGCCGCGGGAATTCGATTATGAAATCATT
CCCTAATAAATCGCTCGTTGCCCTTGCTATCGCAAGCATGAGCTCTGGTGTC
TTAGCACACGGTTACGTATCGGAATCCAACGATGGCGTCGCAGCAAGTCGT

GCAGCGCTGTGTAAGTACCCAACATCGGATACCAATGAAAGAAACACTAAC

TGTGGCGCAATTCAATATGAACCACAAAGCGTGGAAGGCCCAGATGGCTTC

CCAGAGACAGGCCCTCGCGACGGGAAAATTGCCAGCGCAGAAACAGCTCG

AGCAGCTGCACTAGACGAACAAACCGCCGACCGCTGGGTAAAACGTCCGAT

AAAATCTGGTACACAAACGTTTGAATGGACTTTTACAGCAAACCACGTGAC

TCGTGATTGGAAATACTACATCACAAAACCGAATTGGAACCCGAACGCTTC

ATTGTCACGCGACTCGTTCGCTCTCAACCCTTTCTGTGTGGTTGACGGCAAT

ATGGTTCAACCACCAAAACAGATGAGCCACCAATGTAACGTGCCTGAACGC

GAAGGGTACCATGTCATCCTTGCGGTATGGGATGTGGGTGATACCGCAGCA

TCATTCTACAACGTTATCGATGTCAAATTTGATGGTGACGACCCTGTGATTC

CTGAGTGGACTCAAGGTGGCCAAATCATCCCTACTATGAATCTAAAGGTTG

GCGATTCGGTTTACACTCGCGTGTTTGATCAATCTGGAGAGAACGTTGCATA

CCGCACTGAACTAGCAATTAGTAACGACGCACTCACTCAAGCGAAAAACTG

GTCTTACGCCTTGGCGAGTAAAGTCAACCAAGAGCAAACTAAGCTTCAAGC

TGGTCAATATTCGGAAGACAGTTCACGCCAGTGTACGGCACGAACCCAATT

TACCTGCAATCAAACAGTGTCTTGAGCGTGTTGAAATTGGCTACATATTGAA

ACGCCAGTACCTGAGTATTCACCCCA

Reverse sequencing product

TAGGACATGATTTCCGCCAAGCTATTTAGGTGACACTATAGAATACTCAAG

CTATGCATCCAACGCGTTGGGAGCTCTCCCATATGGTCGACCTGCAGGCGG

CCGCGAATTCACTAGTGATTTAGTTCAGTTTATCCCAAGCCATTTCCCAGTG

TGAGCCTGTCGCTGGCTCAAATTGTGTTGCACTCTCACTCCACTGTACGCAG

TAGCCAGAGTAAGGGAATGGTTTACATTGATAAATCGCGCCGTCGCTTGCT

AACACTTTTGTACCTGCTGTGTAGCTAGAAAGCCCCTCTGGGAATACAAAGT

CGTATTCGCCTGATGGTGGTGGGGTGACTTCCTCACTTTTTAGATGGAAATC

CAGAGTTTGCTGATCGACAAGGTTGCCATCGGTATCTTTAATGCGGGTTACT

AACATGTGGTGCCCCGGTTCTGATTTGCTCAGCGTGAGTTCTACTTGTTCGC

TTGCACCATCTTGAATTGAACCCGTCCAACTTGCTAGTGGTTCGCGATGATG

ATTGTAAACCGTTAACTCGGCGGTTAAGTCGCCCTCTGCGGTAAGAGTGAG

GTCTAAAGAGGTTGGTTCTGTAGCAATTATGTACTCACTGGCAAGACCATCA

ACCGTGAGTGAATACTCAGGTACTGGCGTTTCAATATTGTAGCCAATTTCAA

CACGCTCAAGGCCACTGTTTGATTGCAGGTAAATTGGGTTCGTGCCGTACAC

TGGCGTGAACTTGTCTTCCGAATATTGACCAGCTTGAAGCTTAGTTTGCTCT

TGGTTGACTTTACTCGCCAAGGCGTAAGACCAGTTTTTCGCTTGAGTGAGTG

CGTCGTTACTAATTGCTAGTTCAGTGCGGTATGCAACGTTCTCTCCAGATTG

ATCAAACACGCGAGTGTAAACCGAATCGCCAACCTTTAGATTCATAGTAGG

GATGATTTGGCCACCTTGAGTCCACTCAAGAATCACAGGGTCGTCACCATC

AAATTTGACATCGATAACGTTGTAGAATGATGCTGCGGTATCACCAACATC

CCATACCGCAGATGACA

Colony 8

Forward sequencing product

NNNNN

Reverse sequencing product

NNNNN

Colony 9

Forward sequencing product

AGGCCCGGTGTTTTGGTAATACGACTCACTATAGGGGCGATTGGGCCCGAC

GTCGCATGCTCCCGGCCGCCATGGCGGCCGCGGGAATTCGATTATCCACAG

AATCAGGGGATAACGCAGGAAAGAACATGTGAGCAAAAGGCCAGCAAAAG

GCCAGGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGC

CCCCCTGACGAGCATCACAAAAATCGACGCTCAAGTCAGAGGTGGCGAAAC

CCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTG

CGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCC

TTCGGGAAGCGTGGCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCG

GTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAACCCCCCGTTCAGC

CCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGTAAG

ACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGC

GAGGTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGG

CTACACTAGAAGAACAGTATTTGGTATCTGCGCTCTGCTGAAGCCAGTTACC

TTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAAGCACCGCTG

Reverse sequencing product

CACTCAAAAACACCCATACCCCATTTAGAATAAGGCCCCATTGTTAATATTG

AAGCCAATGCCCTTCAGGTTCATCTTGTTCCCAATGCCCATCTTGCTCATATT

GCACCCAATCCCCTTCTTGCTCATCTTGCACCCAATCCCCATCATGCTCTCCA

TGCACCCAATGCCTTTCTTCTTCATCATGCACCCAATGCCCATCATGCTCTCC

ATGCCCTCAATGCCTATCATGCCCTTCATGCTCTCCATGCCCATCATGTTCAT

CATGCAATTAATCCCATTTTGCTCTTCATGCTTTTATTTCCTTCATGCTTATTT

TCTTCCTTTCCTCATCATGCTCATCATTTTTCTTATTCTTATTATGCTCATGAT

GCTTTCTTTGCACACGATGG

Colony 10

Forward sequencing product

CGGGCCGGTGAATTGTAATACGACTCACTATAGGGCGAATTGGGCCCGACG

TCGCATGCTCCCGGCCGCCATGGCGGCCGCGGGAATTCGATTATGAAATCA

TTCCCTAATAAATCGCTCGTTGCCCTTGCTATCGCAAGCATGAGCTCTGGTG

TCTTAGCACACGGTTACGTATCGGAATCCAACGATGGCGTCGCAGCAAGTC

GTGCAGCGCTGTGTAAGTACCCAACATCGGATACCAATGAAAGAAACACTA

ACTGTGGCGCAATTCAATATGAACCACAAAGCGTGGAAGGCCCAGATGGCT

TCCCAGAGACAGGCCCTCGCGACGGGAAAATTGCCAGCGCAGAAACAGCTC

GAGCAGCTGCACTAGACGAACAAACCGCCGACCGCTGGGTAAAACGTCCG

ATAAAATCTGGTACACAAACGTTTGAATGGACTTTTACAGCAAACCACGTG

ACTCGTGATTGGAAATACTACATCACAAAACCGAATTGGAACCCGAACGCT

TCATTGTCACGCGACTCGTTCGCTCTCAACCCTTTCTGTGTGGTTGACGGCA

ATATGGTTCAACCACCAAAACAGATGAGCCACCAATGTAACGTGCCTGAAC

GCGAAGGGTACCATGTCATCCTTGCGGTATGGGATGTGGGTGATACCGCAG

CATCATTCTACAACGTTATCGATGTCAAATTTGATGGTGACGACCCTGTGAT

TCCTGAGTGGACTCAAGGTGGCCAAATCATCCCTACTATGAATCTAAAGGTT

GGCGATTCGGTTTACACTCGCGTGTTTGATCAATCTGGAGAGAACGTTGCAT

ACCGCACTGAACTAGCAATTAGTAACGACGCACTCACTCAAGCGAAAAACT

GGTCTTACGCCTTGGCGAGTAAAGTCAACCAAGAGCAAACTAAGCTTCAAG

CTGGTCAATATTCGGAAGACAAGTTCACGCCAGTGTACGGCACGAACCCAA

TTTACCTGCAATCAAACAGTGGCCTTGAGCGTGTTGAATTGGCTACATATTG

AACGCCAGTACCTGAGTATTCACTC

Reverse sequencing product

AGGATTCCGCCCAAGCTATTTAGGTGACACTATAGAATACTCAAGCTATGC

ATCCAACGCGTTGGGAGCTCTCCCATATGGTCGACCTGCAGGCGGCCGCGA

ATTCACTAGTGATTTTAGTTCAGTTTATCCCAAGCCATTTCCCAGTGTGAGC

CTGTCGCTGGCTCAAATTATGTTGCACTCTCACTCCACTGTACGCAGTAGCC

AGAGTAAGGGAATGGTTTACATTGATAAATCGCGCCGTCGCTTGCTAACAC

TTTTGTACCTGCTGTGTAGCTAGAAAGCCCCTCTGGGAATACAAAGTCGTAT

TCGCCTGATGGTGGTGGGGTGACTTCCTCACTTTTTAGATGGAAATCCAGAG

TTTGCTGATCGACAAGGTTGCCATCGGTATCTTTAATGCGGGTTACTAACAT

GTGGTGCCCCGGTTCTGATTTGCTCAGCGTGAGTTCTACTTGTTCGCTTGCA
CCATCTTGAATTGAACCCGTCCAACTTGCTAGTGGTTCGCGATGATGATTGT
AAACCGTTAACTCGGCGGTTAAGTCGCCCTCTGCGGTAAGAGTGAGGTCTA
AAGAGGTTGGTTCTGTAGCAATTATGTACTCACTGGCAAGACCATCAACCG
TGAGTGAATACTCAGGTACTGGCGTTTCAATATTGTAGCCAATTTCAACACG
CTCAAGGCCACTGTTTGATTGCAGGTAAATTGGGTTCGTGCCGTACACTGGC
GTGAACTTGTCTTCCGAATATTGACCAGCTTGAAGCTTAGTTTGCTCTTGGT
TGACTTTACTCGCCAAGGCGTAAGACCAGTTTTTCGCTTGAGTGAGTGCGTC
GTTACTAATTGCTAGTTCAGTGCGGTATGCAACGTTCTCTCCAGATTGATCA
AACACGCGAGTGTAAACCGAATCGCCAACCTTTAGATTCATAGTAGGTATG
ATTTGGCCACCTTGAGTCCACTCAGGAATCACCAGGGTCGTCACCATCAAAT
TTGACATCGATAACGTGTAGATGATGCTGCGGTATCCCCCACCATTCCCATA
CGCAAGGATGACATGG

Colony 11

Forward sequencing product

GGGGCCAGTGAATTTGTAATACGACTCACTATAGGGCGAATTGGGCCCGAC
GTCGCATGCTCCCGGCCGCCATGGCGGCCGCGGGAATTCGATTATGAAATC
ATTCCCTAATAAATCGCTCGTTGCCCTTGCTATCGCAAGCATGAGCTCTGGT
GTCTTAGCACACGGTTACGTATCGGAATCCAACGATGGCGTCGCAGCAAGT
CGTGCAGCGCTGTGTAAGTACCCAACATCGGATACCAATGAAAGAAACACT
AACTGTGGCGCAATTCAATATGAACCACAAAGCGTGGAAGGCCCAGATGGC
TTCCCAGAGACAGGCCCTCGCGACGGGAAAATTGCCAGCGCAGAAACAGCT
CGAGCAGCTGCACTAGACGAACAAACCGCCGACCGCTGGGTAAAACGTCCG
ATAAAATCTGGTACACAAACGTTTGAATGGACTTTTACAGCAAACCACGTG
ACTCGTGATTGGAAATACTACATCACAAAACCGAATTGGAACCCGAACGCT

TCATTGTCACGCGACTCGTTCGCTCTCGACCCTTTCTGTGTGGTTGACGGCA

ATATGGTTCAACCACCAAAACAGATGAGCCACCAATGTAACGTGCCTGAAC

GCGAAGGGTACCATGTCATCCTTGCGGTATGGGATGTGGGTGATACCGCAG

CATCATTCTACAACGTTATCGATGTCAAATTTGATGGTGACGACCCTGTGAT

TCCTGAGTGGACTCAAGGTGGCCAAATCATCCCTACTATGAATCTAAAGGTT

GGCGATTCGGTTTACACTCGCGTGTTTGATCAATCTGGAGAGAACGTTGCAT

ACCGCACTGAACTAGCAATTAGTAACGACGCACTCACTCAAGCGAAAAACT

GGTCTTACGCCTTTGGCGAGTAAAGTCAAACAAAGAGCAAACTAAGCTTCA

AGCTGGTCAATATTCGGGAAGAACAAGTTCACGCCAGTGTACGGCACGAAC

CCAATTTTACCTGCAATCAAACAGTTGTCCTTGAGCGTGTTGAAATTGGGCT

TACAATATTGAAAACGCCAGTACC

Reverse sequencing product

AGACTGGATTACGCCAAGCTATTTAGGTGACACTATAGAATACTCAAGCTA

TGCATCCAACGCGTTGGGAGCTCTCCCATATGGTCGACCTGCAGGCGGCCG

CGAATTCACTAGTGATTTTAGTTCAGTTTATCCCAAGCCATTTCCCAGTGTG

AGCCTGTCGCTGGCTCAAATTTTGTTGCACTCTCACTCCACTGTACGCAGTA

GCCAGAGTAAGGGAATGGTTTACATTGATAAATCGCGCCGTCGCTTGCTAA

CACTTTTGTACCTGCTGTGTAGCTAGAAAGCCCCTCTGGGAATACAAAGTCG

TATTCGCCTGACGGTGGTGGGGTGACTTCCTCACTTTTTAGATGGAAATCCA

GAGTTTGCTGATCGACAAGGTTGCCATCGGTATCTTTAATGCGGGTTACTAA

CATGTGGTGCCCCGGTTCTGATTTGCTCAGCGTGAGTTCTACTTGTTCGCTTG

CACCATCTTGAATTGAACCCGTCCAACTTGCTAGTGGTTCGCGATGATGATT

GTAAACCGTTAACTCGGCGGTTAAGTCGCCCTCTGCGGTAAGAGTGAGGTC

TAAAGAGGTTGGTTCTGTAGCAATTATGTACTCACTGGCAAGACCATCAAC

CGTGAGTGAATACTCAGGTACTGGCGTTTCAATATTGTAGCCAATTTCAACA

CGCTCAAGGCCACTGTTTGATTGCAGGTAAATTGGGTTCGTGCCGTACACTG

GCGTGAACTTGTCTTCCGAATATTGACCAGCTTGAAGCTTAGTTTGCTCTTG
GTTGACTTTACTCGCCAAGGCGTAAGACCAGTTTTTCGCTTGAGTGAGTGCG
TCGTTACTAATTGCTAGTTCAGTGCGGTATGCAACGTTCTCTCAGATTGATC
AAACACGCGAGTGTAAACCGAATCGCCAACCTTTTAGATTCATAGTAGGGA
TGATTTGGCCACCTTGAGTCCACTCAAGAATCACAGGGTCGTCACCATCAA
ATTTTGACATCGATAACGTTGTAGAATGATGCTGCGGTATCACCCACATCCC
ATACCGCAAGGATGAC

Colony 12

Forward sequencing product

NNNNN

Reverse sequencing product

NNNNN

Colony 13

Forward sequencing product

CACGGGCCCGTTGTATTTGTTATACGACTCACTATAGGGCGAATTGGGCCCG
ACGTCGCATGCTCCCGGCCGCCATGGCGGCCGCGGGAATTCGATTTTAGTTC
AGTTTATCCCAAGCCATTTCCCAGTGTGAGCCTGTCGCTGGCTCAAATTTTG
TTGCACTCTCACTCCACTGTACGCAGTAGCCAGAGTAAGGGAATGGTTTAC
ATTGATAAATCGCGCCGTCGCTTGCTAACACTTTTGTACCTGCTGTGTAGCT
AGAAAGCCCCTCTGGGAATACAAAGTCGTATTCGCCTGATGGTGGTGGGGT
GACTTCCTCACTTTTTAGATGGAAATCCAGAGTTTGCTGATCGACAAGGTTG
CCATCGGTATCTTTAATGCGGGTTACTAACATGTGGTGCCCCGGTTCTGATT
TGCTCAGCGTGAGTTCTACTTGTTCGCTTGCACCATCTTGAATTGAACCCGT
CCAACTTGCTAGTGGTTCGCGATGATGATTGTAAACCGTTAACTCGGCGGTT

AAGTCGCCCTCTGCGGTAAGAGTGAGGTCTAAAGAGGTTGGTTCTGTAGCA

ATTATGTACTCACTGGCAAGACCATCAACCGTGAGTGAATACTCAGGTACT

GGCGTTTCAATATTGTAGCCAATTTCAACACGCTCAAGGCCACTGTTTGATT

GCAGGTAAATTGGGTTCGTGCCGTACACTGGCGTGAACTTGTCTTCCGAATA

TTGACCAGCTTGAAGCTTAGTTTGCTCTTGGTTGACTTTACTCGCCAAGGCG

TAAGACCAGTTTTTCGCTTGAGTGAGTGCGTCGTTACTAATTGCTAGTTCAG

TGCGGTATGCAACGTTCTCTCCAGATTGATCAAACACGCGAGTGTAAACCG

AATCGCCAACCTTTTAGATTCATAGTAAGGGATGATTTGGCCACCTTGAGTC

CACTCAGGAATCACAGGGTCGTCACCATCAAATTTGACATCGATTAACGTT

GTAGATTGATGCTGCGGTATCACCCACATCCCATACCGCAAGGATGACAAT

GTACCCTTCGCGTTCAG

Reverse sequencing product

TCCGGCCCAAGGCTATTTAGGGTGACACTATAGAATACTCAAGCTATGCAT

CCAACGCGTTGGGAGCTCTCCCATATGGTCGACCTGCAGGCGGCCGCGAAT

TCACTAGTGATTATGAAATCATTCCCTAATAAATCGCTCGTTGCCCTTGCTA

TCGCAGGCATGAGCTCTGGTGTCTTAGCACACGGTTACGTATCGGAATCCA

ACGATGGCGTCGCAGCAAGTCGTGCAGCGCTGTGTAAGTACCCAACATCGG

ATACCAATGAAAGAAACACTAACTGTGGCGCAATTCAATATGAACCACAAA

GCGTGGAAGGCCCAGATGGCTTCCCAGAGACAGGCCCTCGCGACGGGAAA

ATTGCCAGCGCAGAAACAGCTCGAGCAGCTGCACTAGACGAACAAACCGCC

GACCGCTGGGTAAAACGTCCGATAAAATCTGGTACACAAACGTTTGAATGG

ACTTTTACAGCAAACCACGTGACTCGTGATTGGAAATACTACATCACAAAA

CCGAATTGGAACCCGAACGCTTCATTGTCACGCGACTCGTTCGCTCTCAACC

CTTTCTGTGTGGTTGACGGCAATATGGTTCAACCACCAAAACAGATGAGCC

ACCAATGTAACGTGCCTGAACGCGAAGGGTACCATGTCATCCTTGCGGTAT

GGGATGTGGGTGATACCGCAGCATCATTCTACAACGTTATCGATGTCAAATT

TGATGGTGACGACCCTGTGATTCCTGAGTGGACTCAAGGTGGCCAAATCAT

CCCTACTATGAATCTAAAGGTTGGCGATTCGGTTTACACTCGCGTGTTTGAT

CAATCTGGAGAGAACGTTGCATACCGCACTGAACTAGCAATTAGTAACGAC

GCACTCACTCAAGCGAAAAACTGGTCTTACGCCTTGACGAGTAAAGTCAAC

CAAGAGCAAACTAAGCTTCAAGCTGGTCAATATTCGGGAAGAACAAGTTCA

CGCCAGTGTACGGCCACGAACCCCAATTTACCTGCAATCAAACAGTGCCTT

GAGCGTGTGAAATTGGCTAACAATTAT

Colony 14

Forward sequencing product

CGCGGGCCGTTGAATTGTATACGACTCACTATAGGGCGAATTGGGCCCGAC

GTCGCATGCTCCCGGCCGCCATGGCGGCCGCGGGAATTCGATTTTAGTTCAG

TTTATCCCAAGCCATTTCCCAGTGTGAGCCTGTCGCTGGCTCAAATTGTGTT

GCACTCTCACTCCACTGTACGCAGTAGCCAGAGTAAGGGAATGGTTTACAT

TGATAAATCGCGCCGTCGCTTGCTAACACTTTTGTACCTGCTGTGTAGCTAG

AAAGCCCCTCTGGGAATACAAAGTCGTATTCGCCTGATGGTGGTGGGGTGA

CTTCCTCACTTTTTAGATGGAAATCCAGAGTTTGCTGATCGACAAGGTTGCC

ATCGGTATCTTTAATGCGGGTTACTAACATGTGGTGCCCCGGTTCTGATTTG

CTCAGCGTGAGTTCTACTTGTTCGCTTGCACCATCTTGAATTGAACCCGTCC

AACTTGCTAGTGGTTCGCGATGATGATTGTAAACCGTTAACTCGGCGGTTAA

GTCGCCCTCTGCGGTAAGAGTGAGGTCTAAAGAGGTTGGTTCTGTAGCAAT

TATGTACTCACTGGCAAGACCATCAACCGTGAGTGAATACTCAGGTACTGG

CGTTTCAATATTGTAGCCAATTTCAACACGCTCAAGGCCACTGTTTGATTGC

AGGTAAATTGGGTTCGTGCCGTACACTGGCGTGAACTTGTCTTCCGAATATT

GACCAGCTTGAAGCTTAGTTTGCTCTTGGTTGACTTTACTCGCCAAGGCGTA

AGACCAGTTTTTCGCTTGAGTGAGTGCGTCGTTACTAATTGCTAGTTCAGTG

CGGTATGCAACGTTCTCTCCAGATTGATCAAACACGCGAGTGTAAACCGAA

TCGCCAACCTTTAGATTCATAGTAGGGATGATTTGGGCCACCTTGAGTCCAC

TCAGGAATCACAGGGTCGTCACATCAAATTTGACATCGATACGTTGTAGAA

TGATGCTGCGGTATCACCCACATCCCATACCGCAAGGATGACATGGTACCC

TTCGCGTTCAGGCACGTT

Reverse sequencing product

ACCGCCCCAAGGCTTATTTTAGGTGACACTATAGAATACTCAAGCTATGCAT

CCAACGCGTTGGGAGCTCTCCCATATGGTCGACCTGCAGGCGGCCGCGAAT

TCACTAGTGATTATGAAATCATTCCCTAATAAATCGCTCGTTGCCCTTGCTA

TCGCAAGCATGAGCTCTGGTGTCTTAGCACACGGTTACGTATCGGAATCCA

ACGATGGCGTCGCAGCAAGTCGTGCAGCGCTGTGTAAGTACCCAACATCGG

ATACCAATGAAAGAAACACTAACTGTGGCGCAATTCAATATGAACCACAAA

GCGTGGAAGGCCCAGATGGCTTCCCAGAGACAGGCCCTCGCGACGGGAAA

ATTGCCAGCGCAGAAACAGCTCGAGCAGCTGCACTAGACGAACAAACCGCC

GACCGCTGGGTAAAACGTCCGATAAAATCTGGTACACAAACGTTTGAATGG

ACTTTTACAGCAAACCACGTGACTCGTGATTGGAAATACTACATCACAAAA

CCGAATTGGAACCCGAACGCTTCATTGTCACGCGACTCGTTCGCTCTCAACC

CTTTCTGTGTGGTTGACGGCAATATGGTTCAACCACCAAAACAGATGAGCC

ACCAATGTAACGTGCCTGAACGCGAAGGGTACCATGTCATCCTTGCGGTAT

GGGATGTGGGTGATACCGCAGCATCATTCTACAACGTTATCGATGTCAAATT

TGATGGTGACGACCCTGTGATTCCTGAGTGGACTCAAGGTGGCCAAATCAT

CCCTACTATGAATCTAAAGGTTGGCGATTCGGTTTACACTCGCGTGTTTGAT

CAATCTGGAGAGAACGTTGCATACCGCACTGAACTAGCAAATTAGTAACGA

CGCACTCACTCAAGCGAAAAACTGGTCTTACGCCTTGGCGAGTAAAGTCAA

CCAAGAGCAAACTAAGCTTCAAGCTGGTCAATATTCGGAAGACAAGTTCAC

GCCAGTGTACGGCACGAACCCAATTTACCTGCAATCAAAACAGTGGCCTTG

AGCGTGTTGAAATTGGCTACAATATT

Colony 15

Forward sequencing product

CGGCCGTGGATTTGTAATACGACTCACTATAGGGCGAATTGGGCCCGACGT

CGCATGCTCCCGGCCGCCATGGCGGCCGCGGGAATTCGATTATGAAATCAT

TCCCTAATAAATCGCTCGTTGCCCTTGCTATCGCAAGCATGAGCTCTGGTGT

CTTAGCACACGGTTACGTATCGGAATCCAACGATGGCGTCGCAGCAAGTCG

TGCAGCGCTGTGTAAGTACCCAACATCGGATACCAATGAAAGAAACACTAA

CTGTGGCGCAATTCAATATGAACCACAAAGCGTGGAAGGCCCAGATGGCTT

CCCAGAGACAGGCCCTCGCGACGGGAAAATTGCCAGCGCAGAAACAGCTC

GAGCAGCTGCACTAGACGAACAAACCGCCGACCGCTGGGTAAAACGTCCG

ATAAAATCTGGTACACAAACGTTTGAATGGACTTTTACAGCAAACCACGTG

ACTCGTGATTGGAAATACTACATCACAAAACCGAATTGGAACCCGAACGCT

TCATTGTCACGCGACTCGTTCGCTCTCAACCCTTTCTGTGTGGTTGACGGCA

ATATGGTTCAACCACCAAAACAGATGAGCCACCAATGTAACGTGCCTGAAC

GCGAAGGGTACCATGTCATCCTTGCGGTATGGGATGTGGGTGATACCGCAG

CATCATTCTACAACGTTATCGATGTCAAATTTGATGGTGACGACCCTGTGAT

TCCTGAGTGGACTCAAGGTGGCCAAATCATCCCTACTATGAATCTAAAGGTT

GGCGATTCGGTTTACACTCGCGTGTTTGATCAATCTGGAGAGAACGTTGCAT

ACCGCACTGAACTAGCAATTAGTAACGACGCACTCACTCAAGCGAAAAACT

GGTCTTACGCCTTGGCGAGTAAAGTCAACCAAGAGCAAACTAAGCTTCAAG

CTGGTCAATATTCGGAAGACAAGTTCACGCCAGTGTACGGCACGAACCCAA

TTTACCTGCAATCAAACAGTGCCTTGAGCGTGTTTGAAATTGGCCTACAATA

TTGAAACCCCAGTACCTGAGTATTCA

Reverse sequencing product

CGGCCCAAAGGCTTATTTTAGGTGACACTATAGAATACTCAAGCTATGCATC

CAACGCGTTGGGAGCTCTCCCATATGGTCGACCTGCAGGCGGCCGCGAATT

CACTAGTGATTTTAGTTCAGTTTATCCCAAGCCATTTCCCAGTGTGAGCCTG

TCGCTGGCTCAAATTTTGTTGCACTCTCACTCCACTGTACGCAGTAGCCAGA

GTAAGGGAATGGTTTACATTGATAAATCGCGCCGTCGCTTGCTAACACTTTT

GTACCTGCTGTGTAGCTAGAAAGCCCCTCTGGGAATACAAAGTCGTATTCG

CCTGATGGTGGTGGGGTGACTTCCTCACTTTTTAGATGGAAATCCAGAGTTT

GCTGATCGACAAGGTTGCCATCGGTATCTTTAATGCGGGTTACTAACATGTG

GTGCCCCGGTTCTGATTTGCTCAGCGTGAGTTCTACTTGTTCGCTTGCACCAT

CTTGAATTGAACCCGTCCAACTTGCTAGTGGTTCGCGATGATGATTGTAAAC

CGTTAACTCGGCGGTTAAGTCGCCCTCTGCGGTAAGAGTGAGGTCTAAAGA

GGTTGGTTCTGTAGCAATTATGTACTCACTGGCAAGACCATCAACCGTGAGT

GAATACTCAGGTACTGGCGTTTCAATATTGTAGCCAATTTCAACACGCTCAA

GGCCACTGTTTGATTGCAGGTAAATTGGGTTCGTGCCGTACACTGGCGTGAA

CTTGTCTTCCGAATATTGACCAGCTTGAAGCTTAGTTTGCTCTTGGTTGACTT

TACTCGCCAAGGCGTAAGACCAGTTTTTCGCTTGAGTGAGTGCGTCGTTACT

AATTGCTAGTTCAGTGCGGTATGCAACGTTCTCTCCAGATTGATCAAACACG

CGAGTGTAAACCGAATCGCCAACCTTTAGATTCATAGTAGGGATGATTTGG

CCACCTTGAGTCCACTCAGGATCACAGGGTCGTCACCATCAAATTTGACATC

GATAACGTTGTAGAATGATGCTGCGGTATCACCCACATCCCATACCGCAGA

TGACATGTACCCT

Colony 16

Forward sequencing product

GCGGGCCCGGTTGAATTGTAATACGACTCACTATAGGGCGAATTGGGCCCG

ACGTCGCATGCTCCCGGCCGCCATGGCGGCCGCGGGAATTCGATTTTAGTTC

AGTTTATCCCAAGCCATTTCCCAGTGTGAGCCTGTCGCTGGCTCAAATTGTG

TTGCACTCTCACTCCACTGTACGCAGTAGCCAGAGTAAGGGAATGGTTTAC

ATTGATAAATCGCGCCGTCGCTTGCTAACACTTTTGTACCTGCTGTGTAGCT

AGAAAGCCCCTCTGGGAATACAAAGTCGTATTCGCCTGATGGTGGTGGGGT

GACTTCCTCACTTTTTAGATGGAAATCCAGAGTTTGCTGATCGACAAGGTTG

CCATCGGTATCTTTAATGCGGGTTACTAACATGTGGTGCCCCGGTTCTGATT

TGCTCAGCGTGAGTTCTACTTGTTCGCTTGCACCATCTTGAATTGAACCCGT

CCAACTTGCTAGTGGTTCGCGATGATGATTGTAAACCGTTAACTCGGCGGTT

AAGTCGCCCTCTGCGGTAAGAGTGAGGTCTAAAGAGGTTGGTTCTGTAGCA

ATTATGTACTCACTGGCAAGACCATCAACCGTGAGTGAATACTCAGGTACT

GGCGTTTCAATATTGTAGCCAATTTCAACACGCTCAAGGCCACTGTTTGATT

GCAGGTAAATTGGGTTCGTGCCGTACACTGGCGTGAACTTGTCTTCCGAATA

TTGACCAGCTTGAAGCTTAGTTTGCTCTTGGTTGACTTTACTCGCCAAGGCG

TAAGACCAGTTTTTCGCTTGAGTGAGTGCGTCGTTACTAATTGCTAGTTCAG

TGCGGTATGCAACGTTCTCTCCAGATTGATCAAACACGCGAGTGTAAACCG

AATCGCCAACCTTTAGATTCATAGTAGGGATGATTTGGGCCACCTTGAGTCC

ACTCAGGAATCACAGGGTCGTCACCATCAAATTTGACATCGATAACGTTGT

AGAATGATGCTGCGGTATCACCCCACATCCCATACCGCAAGGATGACATGG

TACCCTTCGCGTTCAGGC

Reverse sequencing product

TTAACTGGATTACGCCAAGCTATTTAGGTGACACTATAGAATACTCAAGCTA

TGCATCCAACGCGTTGGGAGCTCTCCCATATGGTCGACCTGCAGGCGGCCG

CGAATTCACTAGTGATTGTTCGCTCTCAACCCTTTCTGTGTGGTTGACGGCA

ATATGGTTCAACCACCAAAACAGATGAGCCACCAATGTAACGTGCCTGAAC

GCGAAGGGTACCATGTCATCCTTGCGGTATGGGATGTGGGTGATACCGCAG

CATCATTCTACAACGTTATCGATGTCAAATTTGATGGTGACGACCCTGTGAT

TCCTGAGTGGACTCAAGGTGGCCAAATCATCCCTACTATGAATCTAAAGGTT

GGCGATTCGGTTTACACTCGCGTGTTTGATCAATCTGGAGAGAACGTTGCAT

ACCGCACTGAACTAGCAATTAGTAACGACGCACTCACTCAAGCGAAAAACT

GGTCTTACGCCTTGGCGAGTAAAGTCAACCAAGAGCAAACTAAGCTTCAAG

CTGGTCAATATTCGGAAGACAAGTTCACGCCAGTGTACGGCACGAACCCAA

TTTACCTGCAATCAAACAGTGGCCTTGAGCGTGTTGAAATTGGCTACAATAT

TGAAACGCCAGTACCTGAGTATTCACTCACGGTTGATGGTCTTGCCAGTGAG

TACATAATTGCTACAGAACCAACCTCTTTAGACCTCACTCTTACCGCAGAGG

GCGACTTAACCGCCGAGTTAACGGTTTACAATCATCATCGCGAACCACTAG

CAAGTTGGACGGGTTCAATTCAAGATGGTGCAAGCGAACAAGTAGAACTCA

CGCTGAGCAAATCAGAACCGGGGCACCACATGTTAGTAACCCGCATTAAAG

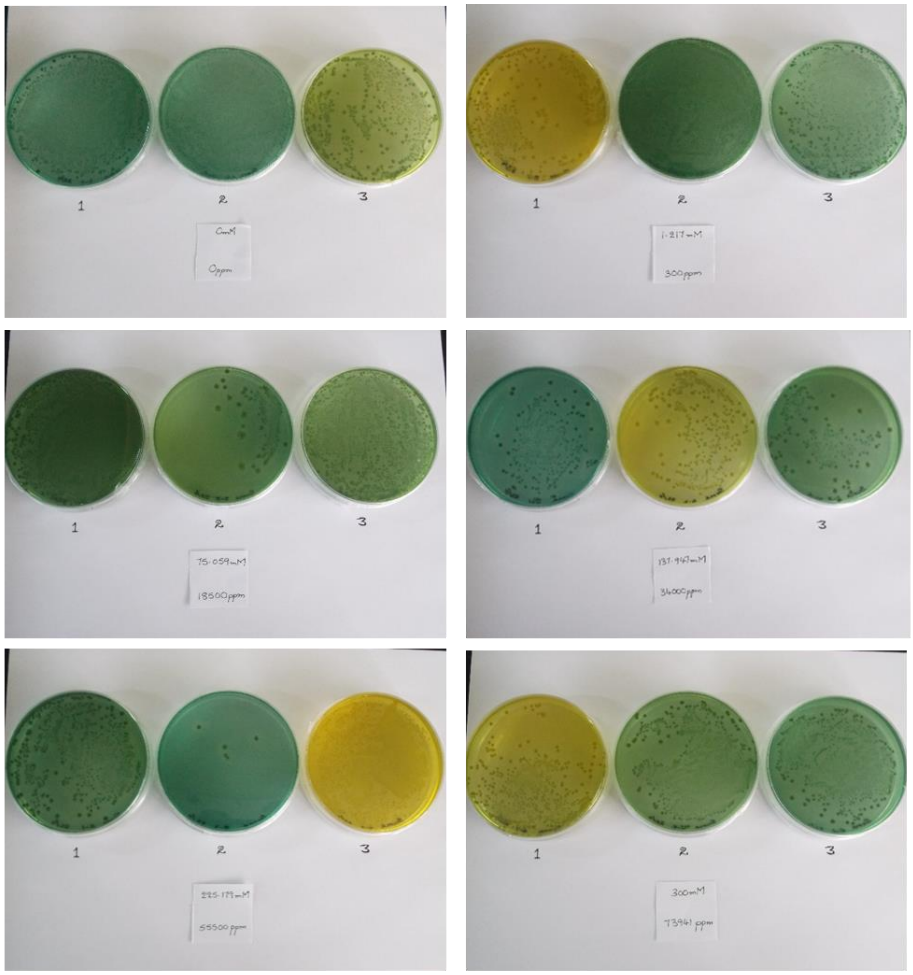
ATACCGATGGCAACCTTGTCGATCAGCAAACTCTGGATTTCCATCTAAAGTG

AGGAAGTCACCCCACCACCATCAGGCGAATACGACTTTGTATCCCAGAGGG

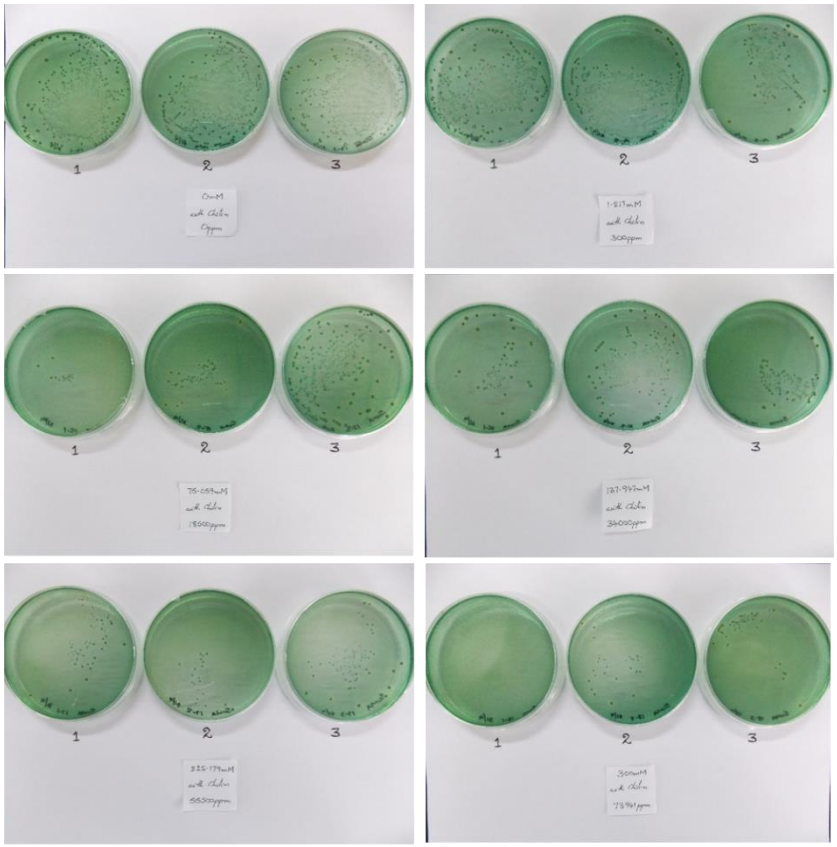
GCTTTCTAGCTACACAGCAGTACAAAAGTGTTAGCAAGCGACGCGCGATTT

ATCATGTAACCATTCCTACTC

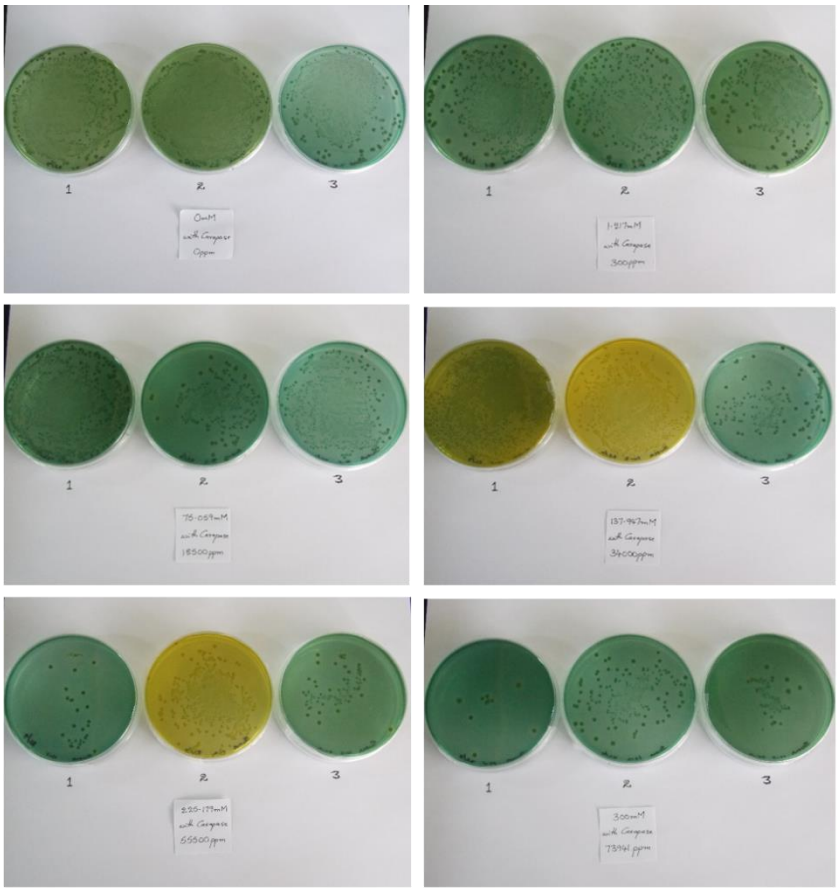
(i) *Vibrio parahaemolyticus* PCV08-7 bacterial plate count



AppFigure1. Figure representing petri plates which have been spread plate with different concentrations of $MgSO_4.7H_20$ (0ppm, 300ppm, 18500ppm, 34000ppm, 55500ppm ,73941ppm)



AppFigure2. Figure representing petri plates which have been spread plate with different concentrations of $MgSO_4.7H_20$ (0ppm, 300ppm, 18500ppm, 34000ppm , 55500ppm ,73941ppm) with added Chitin (Sigma) by weight 0.05gm/3ml.



AppFigure3. Figure representing petri plates which have been spread plate with different concentrations of $MgSO_4.7H_20$ (0ppm, 300ppm, 18500ppm,34000ppm , 55500ppm ,73941ppm) with added carapace of prawn by its dry weight at 0.05g/3ml

(j) *Vibrio parahaemolyticus* PCV08-7 RNA isolation and cDNA synthesis with Nanodrop quality readings

AppTable1. Table representing concentration of RNA and cDNA isolations for different Magnesium concentrations in triplicates.

Sample	260/280	260/230	Conc (ng/ul) RNA	260/280	260/230	Conc (ng/ul) cDNA
1 0ppm	2.15	1.71	513.3	1.74	2.03	1191.4
2 0ppm	2.11	2.13	606.6	1.74	2.08	1217.1
3 0ppm	2.19	2.12	463.4	1.72	2.06	1172.2
4 300ppm	2.06	1.05	53.5	1.62	1.94	1195
5 300ppm	2.04	1.25	38.2	1.60	1.95	1242.5
6 300ppm	2.07	1.64	57.9	1.61	1.95	1248.7
7 18500ppm	2.22	1.73	31.3	1.61	1.96	1092.4
8 18500ppm	2.17	1.59	48.7	1.61	1.96	1245.1
9 18500ppm	2.16	1.92	69.6	1.62	1.96	1254.2
10 34000ppm	2.14	1.90	51.8	1.60	1.84	1239
11 34000ppm	2.35	1.64	29.9	1.61	1.96	1233.2
12 34000ppm	2.20	1.71	62.8	1.61	1.95	1230.5
13 55500ppm	2.16	1.83	58.2	1.61	1.93	1183.8
14 55500ppm	2.12	2.02	69.1	1.62	1.97	1232.8
15 55500ppm	2.18	1.42	61.8	1.61	1.96	1265.9
16 73941ppm	2.23	2.09	52.6	1.61	1.97	1239.8
17 73941ppm	2.25	1.68	37.5	1.60	1.97	1224.4
18 73941ppm	2.12	2.04	100.3	1.62	1.96	1268.3

AppTable2. Table representing concentration of RNA and cDNA isolations for different Magnesium concentrations in triplicates in the presence of Chitin.

Sample	260/280	260/230	Conc (ng/ul) RNA	260/280	260/230	Conc (ng/ul) cDNA
1 0ppm	2.26	2.34	104.2	1.56	1.82	883.1
2 0ppm	2.24	2.76	168.7	1.58	1.83	879.6
3 0ppm	2.25	2.78	180.7	1.57	1.84	885.3
4 300ppm	2.34	2.77	89.5	1.57	1.84	865.8
5 300ppm	2.24	2.5	210.2	1.57	1.83	891.6
6 300ppm	2.35	2.18	56.4	1.56	1.82	870.8
7 18500ppm	2.27	2.76	145.2	1.56	1.81	882.2
8 18500ppm	2.28	2.79	142.0	1.55	1.81	872.7
9 18500ppm	2.26	2.76	168.8	1.56	1.84	874.9
10 34000ppm	2.27	2.87	148.7	1.57	1.83	883.5
11 34000ppm	2.25	2.56	176.8	1.57	1.81	880.7
12 34000ppm	2.26	2.31	138.8	1.57	1.82	882.3
13 55500ppm	2.33	2.94	94.8	1.56	1.82	873.5
14 55500ppm	2.3	2.79	118.8	1.57	1.82	878.2
15 55500ppm	2.28	2.85	153.2	1.58	1.83	882.8
16 73941ppm	2.34	3.16	91.2	1.57	1.85	864.2
17 73941ppm	2.56	3.32	103.0	1.57	1.83	866.7
18 73941ppm	2.27	2.39	118.4	1.56	1.81	884.6

AppTable3. Table representing concentration of RNA and cDNA isolations for different Magnesium concentrations in triplicates in the presence of carapace.

Sample	260/280	260/230	Conc (ng/ul) RNA	260/280	260/230	Conc (ng/ul) cDNA
1 0ppm	2.21	1.42	29.3	1.61	1.97	1239.1
2 0ppm	2.17	1.98	75	1.61	1.96	1247.2
3 0ppm	2.18	2.05	88.5	1.62	1.97	1250.2
4 300ppm	2.18	1.46	32.1	1.61	1.97	1220.9
5 300ppm	2.17	0.85	39.9	1.61	1.95	1232.9
6 300ppm	2.17	1.30	38.8	1.6	1.96	1261.6
7 18500ppm	2.2	1.82	44.9	1.61	1.95	1239.9
8 18500ppm	2.18	1.89	53.7	1.61	1.96	1199.7
9 18500ppm	2.15	1.78	74.7	1.61	1.95	1232.8
10 34000ppm	2.1	2.00	64.4	1.61	1.95	1242.8
11 34000ppm	2.14	1.56	65.7	1.61	1.95	1221.2
12 34000ppm	2.12	1.91	69.2	1.61	1.94	1248.6
13 55500ppm	2.16	2.06	49.4	1.61	1.95	1237.9
14 55500ppm	2.19	1.93	51.9	1.61	1.95	1233.0
15 55500ppm	2.17	1.75	57	1.61	1.92	1212.9
16 73941ppm	2.1	1.81	52.1	1.61	1.96	1226.5
17 73941ppm	2.18	1.57	39	1.61	1.95	1240.2
18 73941ppm	2.15	1.81	68.7	1.62	1.95	1231.2

(k) Absolute quantification

Absolute quantification Real Time PCR values

Well	Sample Name	Detector	Task	Ct	StdDev Ct	Average
E6	0.15_0.005	gbpA	Unknown	34.82	1.133	
F6	0.15_0.005	gbpA	Unknown	33.24	1.133	33.56
G6	0.15_0.005	gbpA	Unknown	32.62	1.133	
B6	0.15_0.05	gbpA	Unknown	33.12	1.005	
C6	0.15_0.05	gbpA	Unknown	33.02	1.005	32.49
D6	0.15_0.05	gbpA	Unknown	31.33	1.005	
A6	0.15_0.5	gbpA	Unknown	30.4	0.329	
G5	0.15_0.5	gbpA	Unknown	29.74	0.329	30.0967
H5	0.15_0.5	gbpA	Unknown	30.15	0.329	
D5	0.15_5	gbpA	Unknown	27.31	0.072	
E5	0.15_5	gbpA	Unknown	27.19	0.072	27.2267
F5	0.15_5	gbpA	Unknown	27.18	0.072	
A5	0.15_50	gbpA	Unknown	23.17	0.023	
B5	0.15_50	gbpA	Unknown	23.13	0.023	23.1567
C5	0.15_50	gbpA	Unknown	23.17	0.023	
H6	0.15_BLANK	gbpA	Unknown	30.92		
E4	0.3_0.005	gbpA	Unknown	34.32	1.647	
F4	0.3_0.005	gbpA	Unknown	33.44	1.647	32.9633
G4	0.3_0.005	gbpA	Unknown	31.13	1.647	
B4	0.3_0.05	gbpA	Unknown	34.61	0.781	
C4	0.3_0.05	gbpA	Unknown	34.42	0.781	34.0667
D4	0.3_0.05	gbpA	Unknown	33.17	0.781	
A4	0.3_0.5	gbpA	Unknown	30.39	0.6	
G3	0.3_0.5	gbpA	Unknown	31.39	0.6	30.7
H3	0.3_0.5	gbpA	Unknown	30.32	0.6	
D3	0.3_5	gbpA	Unknown	28.15	0.215	
E3	0.3_5	gbpA	Unknown	28.22	0.215	28.3067
F3	0.3_5	gbpA	Unknown	28.55	0.215	
A3	0.3_50	gbpA	Unknown	23.93	0.204	
B3	0.3_50	gbpA	Unknown	23.79	0.204	23.97
C3	0.3_50	gbpA	Unknown	24.19	0.204	
H4	0.3_BLANK	gbpA	Unknown	31.21		
E2	0.6_0.005	gbpA	Unknown	33.14	1.88	
F2	0.6_0.005	gbpA	Unknown	31.88	1.88	32.65
G2	0.6_0.005	gbpA	Unknown	32.93	1.88	
B2	0.6_0.05	gbpA	Unknown	31.5	1.231	
C2	0.6_0.05	gbpA	Unknown	32.92	1.231	32.79
D2	0.6_0.05	gbpA	Unknown	33.95	1.231	
A2	0.6_0.5	gbpA	Unknown	30.23	0.112	
G1	0.6_0.5	gbpA	Unknown	30.4	0.112	30.2733
H1	0.6_0.5	gbpA	Unknown	30.19	0.112	
D1	0.6_5	gbpA	Unknown	26.82	0.319	
E1	0.6_5	gbpA	Unknown	27.02	0.319	27.0967
F1	0.6_5	gbpA	Unknown	27.45	0.319	
A1	0.6_50	gbpA	Unknown	22.68	0.742	
B1	0.6_50	gbpA	Unknown	24.13	0.742	23.3067
C1	0.6_50	gbpA	Unknown	23.11	0.742	
H2	0.6_BLANK	gbpA	Unknown	32.83		

(l) Relative quantification

Real time values in the presence of only Magnesium sulfate hepta hydrate

Plate	Well ID	Well	Sample	Detector	Task	Ct	Ct std err	Avg Ct
RELATIVE_10thSept14	1	A1	gbpA_1	Crustin	ENDO	24.639	0.437	25.289
RELATIVE_10thSept14	13	B1	gbpA_1	Crustin	ENDO	26.12	0.437	25.289
RELATIVE_10thSept14	25	C1	gbpA_1	Crustin	ENDO	25.107	0.437	25.289
RELATIVE_10thSept14	37	D1	gbpA_2	Crustin	ENDO	22.186	0.135	22.413
RELATIVE_10thSept14	49	E1	gbpA_2	Crustin	ENDO	22.401	0.135	22.413
RELATIVE_10thSept14	61	F1	gbpA_2	Crustin	ENDO	22.653	0.135	22.413
RELATIVE_10thSept14	2	A2	gbpA_3	Crustin	ENDO	24.702	0.077	24.824
RELATIVE_10thSept14	14	B2	gbpA_3	Crustin	ENDO	24.805	0.077	24.824
RELATIVE_10thSept14	26	C2	gbpA_3	Crustin	ENDO	24.966	0.077	24.824
RELATIVE_10thSept14	38	D2	gbpA_4	Crustin	ENDO	19.399	0.136	19.658
RELATIVE_10thSept14	50	E2	gbpA_4	Crustin	ENDO	19.72	0.136	19.658
RELATIVE_10thSept14	62	F2	gbpA_4	Crustin	ENDO	19.856	0.136	19.658
RELATIVE_10thSept14	3	A3	gbpA_5	Crustin	ENDO	19.457	0.108	19.586
RELATIVE_10thSept14	15	B3	gbpA_5	Crustin	ENDO	19.502	0.108	19.586
RELATIVE_10thSept14	27	C3	gbpA_5	Crustin	ENDO	19.8	0.108	19.586
RELATIVE_10thSept14	39	D3	gbpA_6	Crustin	ENDO	18.771	0.102	18.96
RELATIVE_10thSept14	51	E3	gbpA_6	Crustin	ENDO	18.989	0.102	18.96
RELATIVE_10thSept14	63	F3	gbpA_6	Crustin	ENDO	19.12	0.102	18.96
Relative2_10thSept14	1	A1	gbpA_7	Crustin	ENDO	20.752	0.403	21.275
Relative2_10thSept14	13	B1	gbpA_7	Crustin	ENDO	22.067	0.403	21.275
Relative2_10thSept14	25	C1	gbpA_7	Crustin	ENDO	21.007	0.403	21.275
Relative2_10thSept14	37	D1	gbpA_8	Crustin	ENDO	21.122	0.025	21.123
Relative2_10thSept14	49	E1	gbpA_8	Crustin	ENDO	21.168	0.025	21.123
Relative2_10thSept14	61	F1	gbpA_8	Crustin	ENDO	21.08	0.025	21.123
Relative2_10thSept14	2	A2	gbpA_9	Crustin	ENDO	19.574	0.126	19.824
Relative2_10thSept14	14	B2	gbpA_9	Crustin	ENDO	19.935	0.126	19.824
Relative2_10thSept14	26	C2	gbpA_9	Crustin	ENDO	19.965	0.126	19.824
Relative_3rd30thSept14	1	A1	gbpA_10	gbpA	ENDO	19.724	0.416	20.256
Relative_3rd30thSept14	13	B1	gbpA_10	gbpA	ENDO	21.076	0.416	20.256
Relative_3rd30thSept14	25	C1	gbpA_10	gbpA	ENDO	19.967	0.416	20.256
Relative_3rd30thSept14	37	D1	gbpA_11	gbpA	ENDO	22.539	0.17	22.843
Relative_3rd30thSept14	49	E1	gbpA_11	gbpA	ENDO	22.86	0.17	22.843
Relative_3rd30thSept14	61	F1	gbpA_11	gbpA	ENDO	23.129	0.17	22.843
Relative_3rd30thSept14	2	A2	gbpA_12	gbpA	ENDO	20.437	0.119	20.597
Relative_3rd30thSept14	14	B2	gbpA_12	gbpA	ENDO	20.526	0.119	20.597
Relative_3rd30thSept14	26	C2	gbpA_12	gbpA	ENDO	20.83	0.119	20.597
Relative_3rd30thSept14	38	D2	gbpA_13	gbpA	ENDO	19.952	0.152	20.223
Relative_3rd30thSept14	50	E2	gbpA_13	gbpA	ENDO	20.239	0.152	20.223
RELATIVE_10thSept14	7	A7	IC_deaD_	Crustin	ENDO	23.69	0.119	23.785
RELATIVE_10thSept14	19	B7	IC_deaD_	Crustin	ENDO	23.643	0.119	23.785
RELATIVE_10thSept14	31	C7	IC_deaD_	Crustin	ENDO	24.021	0.119	23.785
RELATIVE_10thSept14	43	D7	IC_deaD_	Crustin	ENDO	21.456	0.043	21.422
RELATIVE_10thSept14	55	E7	IC_deaD_	Crustin	ENDO	21.337	0.043	21.422
RELATIVE_10thSept14	67	F7	IC_deaD_	Crustin	ENDO	21.472	0.043	21.422
RELATIVE_10thSept14	8	A8	IC_deaD_	Crustin	ENDO	21.446	0.149	21.67
RELATIVE_10thSept14	20	B8	IC_deaD_	Crustin	ENDO	21.612	0.149	21.67
RELATIVE_10thSept14	32	C8	IC_deaD_	Crustin	ENDO	21.951	0.149	21.67
RELATIVE_10thSept14	44	D8	IC_deaD_	Crustin	ENDO	18.794	0.033	18.847
RELATIVE_10thSept14	56	E8	IC_deaD_	Crustin	ENDO	18.838	0.033	18.847
RELATIVE_10thSept14	68	F8	IC_deaD_	Crustin	ENDO	18.908	0.033	18.847
RELATIVE_10thSept14	9	A9	IC_deaD_	Crustin	ENDO	18.822	0.058	18.805
RELATIVE_10thSept14	21	B9	IC_deaD_	Crustin	ENDO	18.697	0.058	18.805
RELATIVE_10thSept14	33	C9	IC_deaD_	Crustin	ENDO	18.895	0.058	18.805
RELATIVE_10thSept14	45	D9	IC_deaD_	Crustin	ENDO	18.1	0.042	18.147
RELATIVE_10thSept14	57	E9	IC_deaD_	Crustin	ENDO	18.112	0.042	18.147
RELATIVE_10thSept14	69	F9	IC_deaD_	Crustin	ENDO	18.23	0.042	18.147
Relative2_10thSept14	4	A4	IC_deaD_	Crustin	ENDO	20.421	0.085	20.581
Relative2_10thSept14	16	B4	IC_deaD_	Crustin	ENDO	20.609	0.085	20.581
Relative2_10thSept14	28	C4	IC_deaD_	Crustin	ENDO	20.712	0.085	20.581
Relative2_10thSept14	40	D4	IC_deaD_	Crustin	ENDO	20.473	0.132	20.735
Relative2_10thSept14	52	E4	IC_deaD_	Crustin	ENDO	20.888	0.132	20.735
Relative2_10thSept14	64	F4	IC_deaD_	Crustin	ENDO	20.845	0.132	20.735
Relative2_10thSept14	5	A5	IC_deaD_	Crustin	ENDO	19.333	0.07	19.446
Relative2_10thSept14	17	B5	IC_deaD_	Crustin	ENDO	19.429	0.07	19.446
Relative2_10thSept14	29	C5	IC_deaD_	Crustin	ENDO	19.575	0.07	19.446
Relative_3rd30thSept14	44	D8	gbpA_IC_	gbpA	ENDO	19.702	0.113	19.928
Relative_3rd30thSept14	56	E8	gbpA_IC_	gbpA	ENDO	20.031	0.113	19.928
Relative_3rd30thSept14	68	F8	gbpA_IC_	gbpA	ENDO	20.05	0.113	19.928
Relative_3rd30thSept14	9	A9	gbpA_IC_	gbpA	ENDO	21.869	0.103	22.067
Relative_3rd30thSept14	21	B9	gbpA_IC_	gbpA	ENDO	22.114	0.103	22.067
Relative_3rd30thSept14	33	C9	gbpA_IC_	gbpA	ENDO	22.217	0.103	22.067
Relative_3rd30thSept14	45	D9	gbpA_IC_	gbpA	ENDO	20.27	0.112	20.486
Relative_3rd30thSept14	57	E9	gbpA_IC_	gbpA	ENDO	20.54	0.112	20.486
Relative_3rd30thSept14	69	F9	gbpA_IC_	gbpA	ENDO	20.647	0.112	20.486
Relative_3rd30thSept14	10	A10	gbpA_IC_	gbpA	ENDO	19.459	0.13	19.655
Relative_3rd30thSept14	22	B10	gbpA_IC_	gbpA	ENDO	19.604	0.13	19.655

Relative_3rd30thSept14	62	F2	gbpA_13	gbpA	ENDO	20.478	0.152	20.223
Relative_3rd30thSept14	3	A3	gbpA_14	gbpA	ENDO	19.269	0.073	19.385
Relative_3rd30thSept14	15	B3	gbpA_14	gbpA	ENDO	19.364	0.073	19.385
Relative_3rd30thSept14	27	C3	gbpA_14	gbpA	ENDO	19.521	0.073	19.385
Relative_3rd30thSept14	39	D3	gbpA_15	gbpA	ENDO	20.221	0.567	20.986
Relative_3rd30thSept14	51	E3	gbpA_15	gbpA	ENDO	22.094	0.567	20.986
Relative_3rd30thSept14	63	F3	gbpA_15	gbpA	ENDO	20.642	0.567	20.986
Relative_3rd30thSept14	4	A4	gbpA_16	gbpA	ENDO	21.702	0.072	21.845
Relative_3rd30thSept14	16	B4	gbpA_16	gbpA	ENDO	21.909	0.072	21.845
Relative_3rd30thSept14	28	C4	gbpA_16	gbpA	ENDO	21.925	0.072	21.845
Relative_3rd30thSept14	40	D4	gbpA_17	gbpA	ENDO	21.819	0.093	21.991
Relative_3rd30thSept14	52	E4	gbpA_17	gbpA	ENDO	22.012	0.093	21.991
Relative_3rd30thSept14	64	F4	gbpA_17	gbpA	ENDO	22.141	0.093	21.991
Relative_3rd30thSept14	5	A5	gbpA_18	gbpA	ENDO	18.974	0.042	18.945
Relative_3rd30thSept14	17	B5	gbpA_18	gbpA	ENDO	18.998	0.042	18.945
Relative_3rd30thSept14	29	C5	gbpA_18	gbpA	ENDO	18.863	0.042	18.945

Relative_3rd30thSept14	34	C10	gbpA_IC_	gbpA	ENDO	19.902	0.13	19.655
Relative_3rd30thSept14	46	D10	gbpA_IC_	gbpA	ENDO	19.252	0.119	19.464
Relative_3rd30thSept14	58	E10	gbpA_IC_	gbpA	ENDO	19.663	0.119	19.464
Relative_3rd30thSept14	70	F10	gbpA_IC_	gbpA	ENDO	19.477	0.119	19.464
Relative_3rd30thSept14	11	A11	gbpA_IC_	gbpA	ENDO	19.608	0.099	19.805
Relative_3rd30thSept14	23	B11	gbpA_IC_	gbpA	ENDO	19.883	0.099	19.805
Relative_3rd30thSept14	35	C11	gbpA_IC_	gbpA	ENDO	19.923	0.099	19.805
Relative_3rd30thSept14	47	D11	gbpA_IC_	gbpA	ENDO	21.391	0.164	21.487
Relative_3rd30thSept14	59	E11	gbpA_IC_	gbpA	ENDO	21.263	0.164	21.487
Relative_3rd30thSept14	71	F11	gbpA_IC_	gbpA	ENDO	21.806	0.164	21.487
Relative_3rd30thSept14	12	A12	gbpA_IC_	gbpA	ENDO	21.465	0.026	21.517
Relative_3rd30thSept14	24	B12	gbpA_IC_	gbpA	ENDO	21.54	0.026	21.517
Relative_3rd30thSept14	36	C12	gbpA_IC_	gbpA	ENDO	21.547	0.026	21.517
Relative_3rd30thSept14	48	D12	gbpA_IC_	gbpA	ENDO	18.7	0.141	18.976
Relative_3rd30thSept14	60	E12	gbpA_IC_	gbpA	ENDO	19.067	0.141	18.976
Relative_3rd30thSept14	72	F12	gbpA_IC_	gbpA	ENDO	19.161	0.141	18.976

Real time values in the presence of Magnesium sulfate and chitin

Plate	Well ID	Well	Sample	Detector	Task	Ct	Ct std err	Avg Ct
Relative_4th1stOct14	14	B2	gbpA_37	gbpA	Target	26.846	0.293	27.23
Relative_4th1stOct14	26	C2	gbpA_37	gbpA	Target	27.037	0.293	27.23
Relative_4th1stOct14	38	D2	gbpA_37	gbpA	Target	27.805	0.293	27.23
Relative_4th1stOct14	50	E2	gbpA_38	gbpA	Target	27.198	0.249	27.575
Relative_4th1stOct14	62	F2	gbpA_38	gbpA	Target	27.482	0.249	27.575
Relative_4th1stOct14	74	G2	gbpA_38	gbpA	Target	28.045	0.249	27.575
Relative_4th1stOct14	3	A3	gbpA_39	gbpA	Target	26.035	0.752	27.121
Relative_4th1stOct14	15	B3	gbpA_39	gbpA	Target	26.761	0.752	27.121
Relative_4th1stOct14	86	H2	gbpA_39	gbpA	Target	28.566	0.752	27.121
Relative_4th1stOct14	27	C3	gbpA_40	gbpA	Target	27.961	0.319	28.565
Relative_4th1stOct14	39	D3	gbpA_40	gbpA	Target	28.69	0.319	28.565
Relative_4th1stOct14	51	E3	gbpA_40	gbpA	Target	29.044	0.319	28.565
Relative_4th1stOct14	63	F3	gbpA_41	gbpA	Target	27.554	0.434	27.875
Relative_4th1stOct14	75	G3	gbpA_41	gbpA	Target	28.734	0.434	27.875
Relative_4th1stOct14	87	H3	gbpA_41	gbpA	Target	27.339	0.434	27.875
Relative_4th1stOct14	4	A4	gbpA_42	gbpA	Target	26.481	0.317	27.004
Relative_4th1stOct14	16	B4	gbpA_42	gbpA	Target	27.576	0.317	27.004
Relative_4th1stOct14	28	C4	gbpA_42	gbpA	Target	26.957	0.317	27.004
Relative_4th1stOct14	40	D4	gbpA_43	gbpA	Target	28.087	0.002	28.572
Relative_4th1stOct14	52	E4	gbpA_43	gbpA	Target	28.574	0.002	28.572
Relative_4th1stOct14	64	F4	gbpA_43	gbpA	Target	28.57	0.002	28.572
Relative_4th1stOct14	5	A5	gbpA_44	gbpA	Target	27.423	0.583	28.582

Plate	Well ID	Well	Sample	Detector	Task	Ct	Ct std err
Relative_4th1stOct14	20	B8	IC_deaD_	deaD	ENDO	21.495	0.002
Relative_4th1stOct14	32	C8	IC_deaD_	deaD	ENDO	22.218	0.002
Relative_4th1stOct14	44	D8	IC_deaD_	deaD	ENDO	22.214	0.002
Relative_4th1stOct14	56	E8	IC_deaD_	deaD	ENDO	21.859	0.113
Relative_4th1stOct14	68	F8	IC_deaD_	deaD	ENDO	22.169	0.113
Relative_4th1stOct14	80	G8	IC_deaD_	deaD	ENDO	22.224	0.113
Relative_4th1stOct14	9	A9	IC_deaD_	deaD	ENDO	21.266	0.005
Relative_4th1stOct14	21	B9	IC_deaD_	deaD	ENDO	21.256	0.005
Relative_4th1stOct14	92	H8	IC_deaD_	deaD	ENDO	21.663	0.005
Relative_4th1stOct14	33	C9	IC_deaD_	deaD	ENDO	20.735	0.217
Relative_4th1stOct14	45	D9	IC_deaD_	deaD	ENDO	21.308	0.217
Relative_4th1stOct14	57	E9	IC_deaD_	deaD	ENDO	21.442	0.217
Relative_4th1stOct14	69	F9	IC_deaD_	deaD	ENDO	21.504	0.103
Relative_4th1stOct14	81	G9	IC_deaD_	deaD	ENDO	21.417	0.103
Relative_4th1stOct14	93	H9	IC_deaD_	deaD	ENDO	21.761	0.103
Relative_4th1stOct14	10	A10	IC_deaD_	deaD	ENDO	20.297	0.253
Relative_4th1stOct14	22	B10	IC_deaD_	deaD	ENDO	20.877	0.253
Relative_4th1stOct14	34	C10	IC_deaD_	deaD	ENDO	21.154	0.253
Relative_4th1stOct14	46	D10	IC_deaD_	deaD	ENDO	21.719	0.095
Relative_4th1stOct14	58	E10	IC_deaD_	deaD	ENDO	22.047	0.095
Relative_4th1stOct14	70	F10	IC_deaD_	deaD	ENDO	21.872	0.095
Relative_4th1stOct14	11	A11	IC_deaD_	deaD	ENDO	21.328	0.26

Relative_4th1stOct14	76	G4	gbpA_44	gbpA	Target	29.272	0.583	28.582
Relative_4th1stOct14	88	H4	gbpA_44	gbpA	Target	29.049	0.583	28.582
Relative_4th1stOct14	17	B5	gbpA_45	gbpA	Target	27.976	0.215	27.654
Relative_4th1stOct14	29	C5	gbpA_45	gbpA	Target	27.74	0.215	27.654
Relative_4th1stOct14	41	D5	gbpA_45	gbpA	Target	27.245	0.215	27.654
Relative_4th1stOct14	53	E5	gbpA_46	gbpA	Target	27.069	0.113	27.294
Relative_4th1stOct14	65	F5	gbpA_46	gbpA	Target	27.427	0.113	27.294
Relative_4th1stOct14	77	G5	gbpA_46	gbpA	Target	27.386	0.113	27.294
Relative_4th1stOct14	6	A6	gbpA_47	gbpA	Target	27.555	0.283	27.6
Relative_4th1stOct14	18	B6	gbpA_47	gbpA	Target	27.134	0.283	27.6
Relative_4th1stOct14	89	H5	gbpA_47	gbpA	Target	28.11	0.283	27.6
Relative_4th1stOct14	30	C6	gbpA_48	gbpA	Target	27.282	0.188	27.655
Relative_4th1stOct14	42	D6	gbpA_48	gbpA	Target	27.806	0.188	27.655
Relative_4th1stOct14	54	E6	gbpA_48	gbpA	Target	27.878	0.188	27.655
Relative Q	1	A1	gbpA_49	gbpA	Target	28.52	0.124	28.767
Relative Q	2	A2	gbpA_49	gbpA	Target	28.911	0.124	28.767
Relative Q	3	A3	gbpA_49	gbpA	Target	28.869	0.124	28.767
Relative Q	4	A4	gbpA_50	gbpA	Target	26.89	0.478	27.813
Relative Q	5	A5	gbpA_50	gbpA	Target	28.491	0.478	27.813
Relative Q	6	A6	gbpA_50	gbpA	Target	28.057	0.478	27.813
Relative Q	13	B1	gbpA_51	gbpA	Target	30.444	0.528	29.49
Relative Q	14	B2	gbpA_51	gbpA	Target	29.405	0.528	29.49
Relative Q	15	B3	gbpA_51	gbpA	Target	28.621	0.528	29.49
Relative Q	16	B4	gbpA_52	gbpA	Target	28.192	0.018	28.201
Relative Q	17	B5	gbpA_52	gbpA	Target	28.175	0.018	28.201
Relative Q	18	B6	gbpA_52	gbpA	Target	28.236	0.018	28.201
Relative Q	25	C1	gbpA_53	gbpA	Target	28.141	0.076	28.267
Relative Q	26	C2	gbpA_53	gbpA	Target	28.403	0.076	28.267
Relative Q	27	C3	gbpA_53	gbpA	Target	28.256	0.076	28.267
Relative Q	28	C4	gbpA_54	gbpA	Target	28.42	0.216	28.818
Relative Q	29	C5	gbpA_54	gbpA	Target	29.162	0.216	28.818
Relative Q	30	C6	gbpA_54	gbpA	Target	28.873	0.216	28.818

Relative_4th1stOct14	82	G10	IC_deaD_	deaD	ENDO	21.409	0.26
Relative_4th1stOct14	94	H10	IC_deaD_	deaD	ENDO	22.145	0.26
Relative_4th1stOct14	23	B11	IC_deaD_	deaD	ENDO	20.775	0.209
Relative_4th1stOct14	35	C11	IC_deaD_	deaD	ENDO	21.127	0.209
Relative_4th1stOct14	47	D11	IC_deaD_	deaD	ENDO	21.498	0.209
Relative_4th1stOct14	59	E11	IC_deaD_	deaD	ENDO	20.895	0.133
Relative_4th1stOct14	71	F11	IC_deaD_	deaD	ENDO	21.251	0.133
Relative_4th1stOct14	83	G11	IC_deaD_	deaD	ENDO	21.328	0.133
Relative_4th1stOct14	12	A12	IC_deaD_	deaD	ENDO	21.224	0.203
Relative_4th1stOct14	24	B12	IC_deaD_	deaD	ENDO	20.567	0.203
Relative_4th1stOct14	95	H11	IC_deaD_	deaD	ENDO	21.112	0.203
Relative_4th1stOct14	36	C12	IC_deaD_	deaD	ENDO	20.639	0.186
Relative_4th1stOct14	48	D12	IC_deaD_	deaD	ENDO	20.716	0.186
Relative_4th1stOct14	60	E12	IC_deaD_	deaD	ENDO	21.23	0.186
Relative Q	49	E1	IC_deaD_	deaD	mixed	22.476	0.016
Relative Q	50	E2	IC_deaD_	deaD	mixed	22.424	0.016
Relative Q	51	E3	IC_deaD_	deaD	mixed	22.436	0.016
Relative Q	52	E4	IC_deaD_	deaD	mixed	22.096	0.042
Relative Q	53	E5	IC_deaD_	deaD	mixed	22.241	0.042
Relative Q	54	E6	IC_deaD_	deaD	mixed	22.165	0.042
Relative Q	61	F1	IC_deaD_	deaD	mixed	22.652	0.056
Relative Q	62	F2	IC_deaD_	deaD	mixed	22.844	0.056
Relative Q	63	F3	IC_deaD_	deaD	mixed	22.718	0.056
Relative Q	64	F4	IC_deaD_	deaD	mixed	23.264	0.062
Relative Q	65	F5	IC_deaD_	deaD	mixed	23.476	0.062
Relative Q	66	F6	IC_deaD_	deaD	mixed	23.331	0.062
Relative Q	73	G1	IC_deaD_	deaD	mixed	23.006	0.02
Relative Q	74	G2	IC_deaD_	deaD	mixed	23.053	0.02
Relative Q	75	G3	IC_deaD_	deaD	mixed	23.075	0.02
Relative Q	76	G4	IC_deaD_	deaD	mixed	23.399	0.042
Relative Q	77	G5	IC_deaD_	deaD	mixed	23.453	0.042
Relative Q	78	G6	IC_deaD_	deaD	mixed	23.544	0.042

Real time values in the presence of Magnesium and carapase

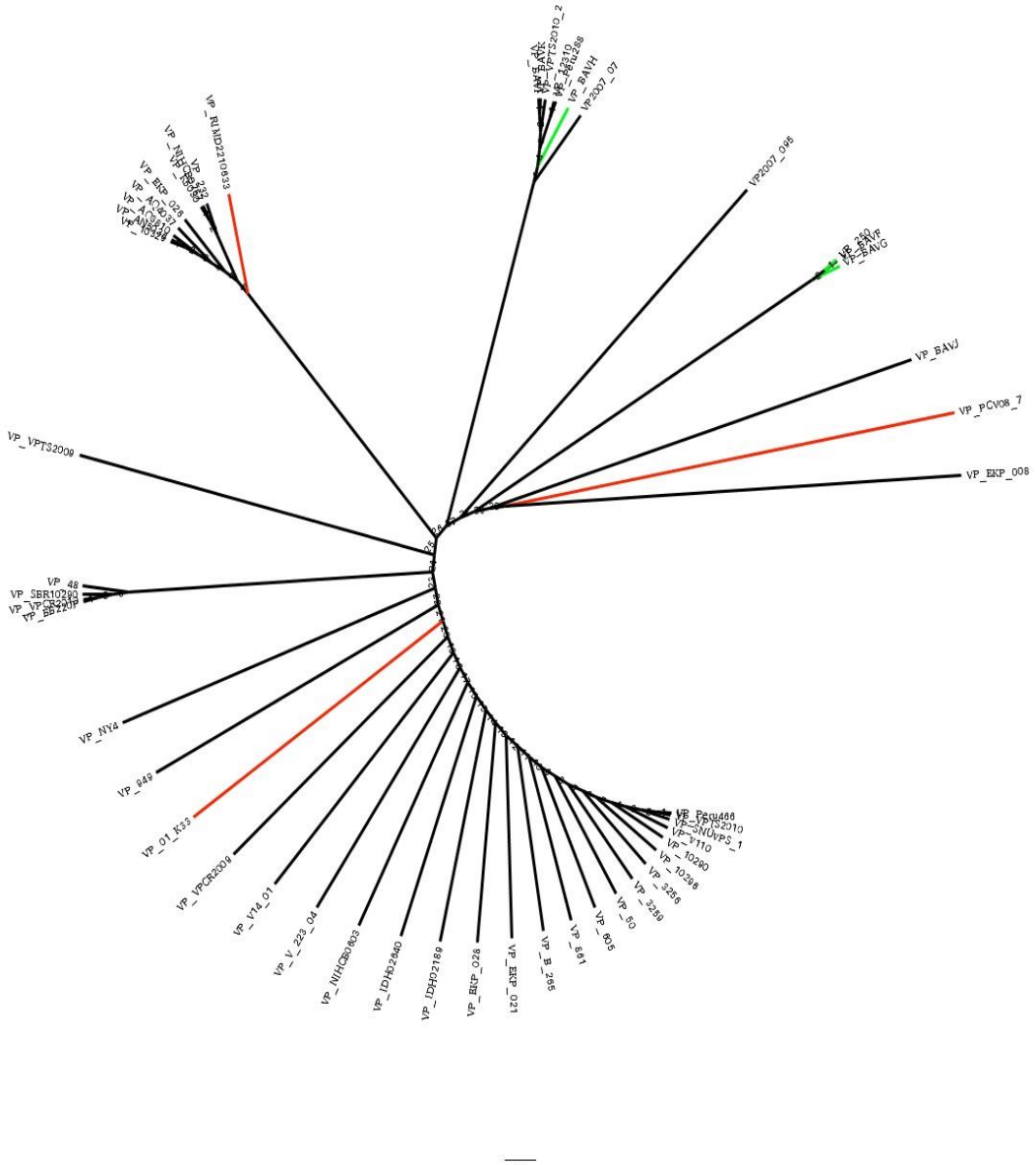
Plate	Well ID	Well	Sample	Detector	Task	Ct	Ct std err	Avg Ct
RELATIVE_10thSept14	4	A4	gbpA_19	Crustin	ENDO	21.259	0.125	21.009
RELATIVE_10thSept14	16	B4	gbpA_19	Crustin	ENDO	20.862	0.125	21.009
RELATIVE_10thSept14	28	C4	gbpA_19	Crustin	ENDO	20.907	0.125	21.009
RELATIVE_10thSept14	40	D4	gbpA_20	Crustin	ENDO	19.455	0.174	19.794
RELATIVE_10thSept14	52	E4	gbpA_20	Crustin	ENDO	20.028	0.174	19.794
RELATIVE_10thSept14	64	F4	gbpA_20	Crustin	ENDO	19.899	0.174	19.794

Plate	Well ID	Well	Sample	Detector	Task	Ct	Ct std err
RELATIVE_10thSept14	10	A10	IC_deaD_	Crustin	ENDO	21.152	0.124
RELATIVE_10thSept14	22	B10	IC_deaD_	Crustin	ENDO	20.75	0.124
RELATIVE_10thSept14	34	C10	IC_deaD_	Crustin	ENDO	20.818	0.124
RELATIVE_10thSept14	46	D10	IC_deaD_	Crustin	ENDO	19.561	0.246
RELATIVE_10thSept14	58	E10	IC_deaD_	Crustin	ENDO	20.408	0.246
RELATIVE_10thSept14	70	F10	IC_deaD_	Crustin	ENDO	20.062	0.246

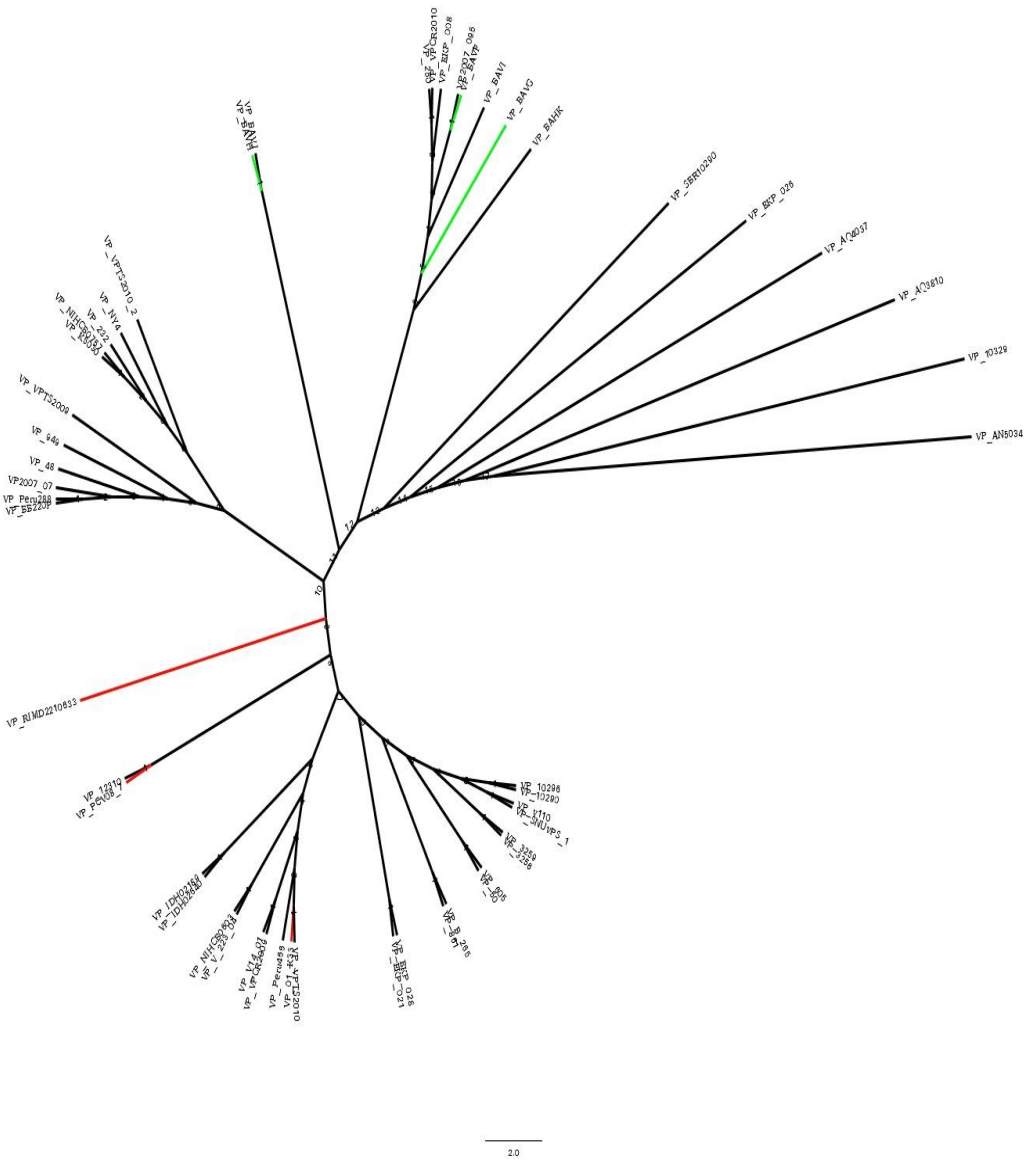
RELATIVE_10thSept14	5	A5	gbpA_21	Crustin	ENDO	19.656	0.08	19.814
RELATIVE_10thSept14	17	B5	gbpA_21	Crustin	ENDO	19.871	0.08	19.814
RELATIVE_10thSept14	29	C5	gbpA_21	Crustin	ENDO	19.915	0.08	19.814
RELATIVE_10thSept14	41	D5	gbpA_22	Crustin	ENDO	20.844	0.124	21.07
RELATIVE_10thSept14	53	E5	gbpA_22	Crustin	ENDO	21.093	0.124	21.07
RELATIVE_10thSept14	65	F5	gbpA_22	Crustin	ENDO	21.272	0.124	21.07
RELATIVE_10thSept14	6	A6	gbpA_23	Crustin	ENDO	20.856	0.07	20.984
RELATIVE_10thSept14	18	B6	gbpA_23	Crustin	ENDO	20.997	0.07	20.984
RELATIVE_10thSept14	30	C6	gbpA_23	Crustin	ENDO	21.099	0.07	20.984
RELATIVE_10thSept14	42	D6	gbpA_24	Crustin	ENDO	20.217	0.193	20.444
RELATIVE_10thSept14	54	E6	gbpA_24	Crustin	ENDO	20.287	0.193	20.444
RELATIVE_10thSept14	66	F6	gbpA_24	Crustin	ENDO	20.829	0.193	20.444
Relative2_10thSept14	38	D2	gbpA_25	Crustin	ENDO	21.133	0.201	21.467
Relative2_10thSept14	50	E2	gbpA_25	Crustin	ENDO	21.441	0.201	21.467
Relative2_10thSept14	62	F2	gbpA_25	Crustin	ENDO	21.828	0.201	21.467
Relative2_10thSept14	3	A3	gbpA_26	Crustin	ENDO	20.177	0.075	20.324
Relative2_10thSept14	15	B3	gbpA_26	Crustin	ENDO	20.428	0.075	20.324
Relative2_10thSept14	27	C3	gbpA_26	Crustin	ENDO	20.367	0.075	20.324
Relative2_10thSept14	39	D3	gbpA_27	Crustin	ENDO	20.142	0.073	20.283
Relative2_10thSept14	51	E3	gbpA_27	Crustin	ENDO	20.384	0.073	20.283
Relative2_10thSept14	63	F3	gbpA_27	Crustin	ENDO	20.323	0.073	20.283
Relative_3rd30thSept14	41	D5	gbpA_28	gbpA	ENDO	19.438	0.15	19.736
Relative_3rd30thSept14	53	E5	gbpA_28	gbpA	ENDO	19.858	0.15	19.736
Relative_3rd30thSept14	65	F5	gbpA_28	gbpA	ENDO	19.911	0.15	19.736
Relative_3rd30thSept14	6	A6	gbpA_29	gbpA	ENDO	20.272	0.18	20.629
Relative_3rd30thSept14	18	B6	gbpA_29	gbpA	ENDO	20.844	0.18	20.629
Relative_3rd30thSept14	30	C6	gbpA_29	gbpA	ENDO	20.771	0.18	20.629
Relative_3rd30thSept14	42	D6	gbpA_30	gbpA	ENDO	19.742	0.1	19.94
Relative_3rd30thSept14	54	E6	gbpA_30	gbpA	ENDO	20.053	0.1	19.94
Relative_3rd30thSept14	66	F6	gbpA_30	gbpA	ENDO	20.026	0.1	19.94
Relative_3rd30thSept14	7	A7	gbpA_31	gbpA	ENDO	19.705	0.197	20.027
Relative_3rd30thSept14	19	B7	gbpA_31	gbpA	ENDO	19.991	0.197	20.027
Relative_3rd30thSept14	31	C7	gbpA_31	gbpA	ENDO	20.384	0.197	20.027
Relative_3rd30thSept14	43	D7	gbpA_32	gbpA	ENDO	21.351	0.176	21.665
Relative_3rd30thSept14	55	E7	gbpA_32	gbpA	ENDO	21.683	0.176	21.665
Relative_3rd30thSept14	67	F7	gbpA_32	gbpA	ENDO	21.959	0.176	21.665
Relative_3rd30thSept14	8	A8	gbpA_33	gbpA	ENDO	20.236	0.176	20.554
Relative_3rd30thSept14	20	B8	gbpA_33	gbpA	ENDO	20.584	0.176	20.554
Relative_3rd30thSept14	32	C8	gbpA_33	gbpA	ENDO	20.843	0.176	20.554
Relative_4th1stOct14	1	A1	gbpA_34	gbpA	Target	19.215	0.609	20.071
Relative_4th1stOct14	13	B1	gbpA_34	gbpA	Target	21.249	0.609	20.071
Relative_4th1stOct14	25	C1	gbpA_34	gbpA	Target	19.749	0.609	20.071
RELATIVE_10thSept14	73	G1	IC_deaD_	Crustin	ENDO	22.111	0.651	
RELATIVE_10thSept14	74	G2	IC_deaD_	Crustin	ENDO	20.359	0.651	
RELATIVE_10thSept14	75	G3	IC_deaD_	Crustin	ENDO	20.006	0.651	
RELATIVE_10thSept14	85	H1	IC_deaD_	Crustin	ENDO	21.138	0.749	
RELATIVE_10thSept14	86	H2	IC_deaD_	Crustin	ENDO	23.495	0.749	
RELATIVE_10thSept14	87	H3	IC_deaD_	Crustin	ENDO	21.379	0.749	
RELATIVE_10thSept14	76	G4	IC_deaD_	Crustin	ENDO	21.958	0.427	
RELATIVE_10thSept14	77	G5	IC_deaD_	Crustin	ENDO	22.633	0.427	
RELATIVE_10thSept14	78	G6	IC_deaD_	Crustin	ENDO	21.155	0.427	
RELATIVE_10thSept14	88	H4	IC_deaD_	Crustin	ENDO	21.714	0.48	
RELATIVE_10thSept14	89	H5	IC_deaD_	Crustin	ENDO	20.089	0.48	
RELATIVE_10thSept14	90	H6	IC_deaD_	Crustin	ENDO	20.597	0.48	
Relative2_10thSept14	41	D5	IC_deaD_	Crustin	ENDO	20.813	0.091	
Relative2_10thSept14	53	E5	IC_deaD_	Crustin	ENDO	21.125	0.091	
Relative2_10thSept14	65	F5	IC_deaD_	Crustin	ENDO	20.998	0.091	
Relative2_10thSept14	6	A6	IC_deaD_	Crustin	ENDO	19.992	0.054	
Relative2_10thSept14	18	B6	IC_deaD_	Crustin	ENDO	19.966	0.054	
Relative2_10thSept14	30	C6	IC_deaD_	Crustin	ENDO	20.14	0.054	
Relative2_10thSept14	42	D6	IC_deaD_	Crustin	ENDO	19.728	0.109	
Relative2_10thSept14	54	E6	IC_deaD_	Crustin	ENDO	20.074	0.109	
Relative2_10thSept14	66	F6	IC_deaD_	Crustin	ENDO	20.029	0.109	
Relative_3rd30thSept14	73	G1	gbpA_IC_	gbpA	ENDO	19.457	0.074	
Relative_3rd30thSept14	74	G2	gbpA_IC_	gbpA	ENDO	19.679	0.074	
Relative_3rd30thSept14	75	G3	gbpA_IC_	gbpA	ENDO	19.46	0.074	
Relative_3rd30thSept14	85	H1	gbpA_IC_	gbpA	ENDO	20.505	0.05	
Relative_3rd30thSept14	86	H2	gbpA_IC_	gbpA	ENDO	20.674	0.05	
Relative_3rd30thSept14	87	H3	gbpA_IC_	gbpA	ENDO	20.558	0.05	
Relative_3rd30thSept14	76	G4	gbpA_IC_	gbpA	ENDO	19.642	0.047	
Relative_3rd30thSept14	77	G5	gbpA_IC_	gbpA	ENDO	19.662	0.047	
Relative_3rd30thSept14	78	G6	gbpA_IC_	gbpA	ENDO	19.793	0.047	
Relative_3rd30thSept14	88	H4	gbpA_IC_	gbpA	ENDO	19.966	0.022	
Relative_3rd30thSept14	89	H5	gbpA_IC_	gbpA	ENDO	20.03	0.022	
Relative_3rd30thSept14	90	H6	gbpA_IC_	gbpA	ENDO	20.033	0.022	
Relative_3rd30thSept14	79	G7	gbpA_IC_	gbpA	ENDO	21.444	0.064	
Relative_3rd30thSept14	80	G8	gbpA_IC_	gbpA	ENDO	21.548	0.064	
Relative_3rd30thSept14	81	G9	gbpA_IC_	gbpA	ENDO	21.325	0.064	
Relative_3rd30thSept14	91	H7	gbpA_IC_	gbpA	ENDO	20.637	0.028	
Relative_3rd30thSept14	92	H8	gbpA_IC_	gbpA	ENDO	20.548	0.028	
Relative_3rd30thSept14	93	H9	gbpA_IC_	gbpA	ENDO	20.623	0.028	
Relative_4th1stOct14	7	A7	IC_deaD_	deaD	ENDO	18.61	0.217	
Relative_4th1stOct14	19	B7	IC_deaD_	deaD	ENDO	19.103	0.217	
Relative_4th1stOct14	31	C7	IC_deaD_	deaD	ENDO	19.347	0.217	

Relative_4th1stOct14	37	D1	gbpA_35	gbpA	Target	20.603	0.139	20.874	Relative_4th1stOct14	43	D7	IC_deaD_	deaD	ENDO	20.277	0.214
Relative_4th1stOct14	49	E1	gbpA_35	gbpA	Target	20.954	0.139	20.874	Relative_4th1stOct14	55	E7	IC_deaD_	deaD	ENDO	20.78	0.214
Relative_4th1stOct14	61	F1	gbpA_35	gbpA	Target	21.066	0.139	20.874	Relative_4th1stOct14	67	F7	IC_deaD_	deaD	ENDO	21.001	0.214
Relative_4th1stOct14	2	A2	gbpA_36	gbpA	Target	20.021	0.216	20.287	Relative_4th1stOct14	8	A8	IC_deaD_	deaD	ENDO	19.074	0.171
Relative_4th1stOct14	73	G1	gbpA_36	gbpA	Target	20.124	0.216	20.287	Relative_4th1stOct14	79	G7	IC_deaD_	deaD	ENDO	19.284	0.171
Relative_4th1stOct14	85	H1	gbpA_36	gbpA	Target	20.715	0.216	20.287	Relative_4th1stOct14	91	H7	IC_deaD_	deaD	ENDO	19.66	0.171

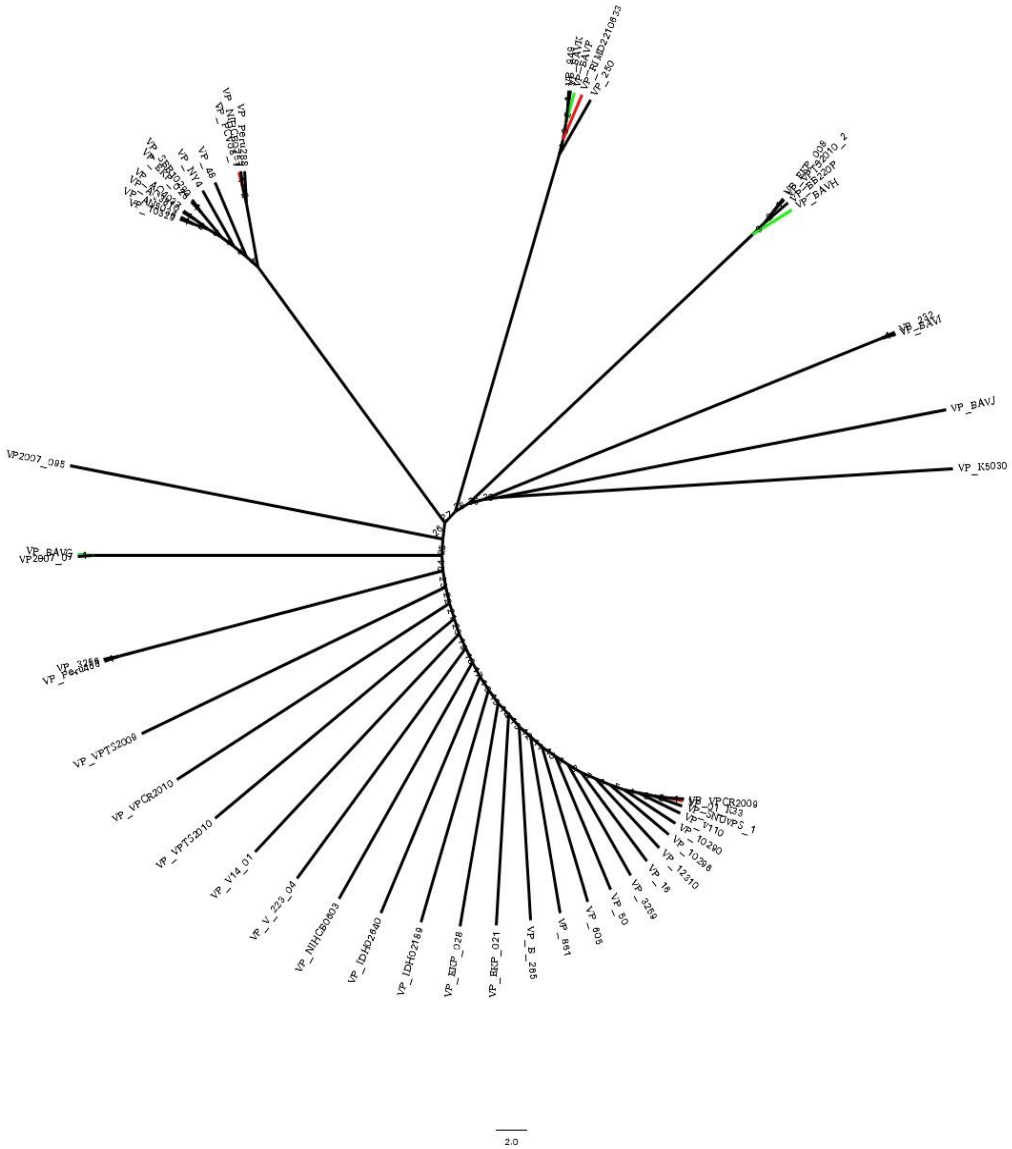
(m) Phylogenetic tree representing the selected 52 *V. parahaemolyticus* genomes in relation with (i) Chitinase 1 (ii) Chitinase 2 (iii) Chitinase 3 (iv) Chitinase 4 (v) Chitinase 5 genes. Marked in RED are the three genomes VPPCV08-7, VPRIMD2210633 and VP01:K33 genomes. In GREEN are the three EMS strains from Thailand VPBAVF (*Vibrio parahaemolyticus* TUMST_DE1_S1), VPBAVG (*Vibrio parahaemolyticus* TUMSAT_DE2_S2) and VPBAVH (*Vibrio parahaemolyticus* TUMSAT_D06_S3)



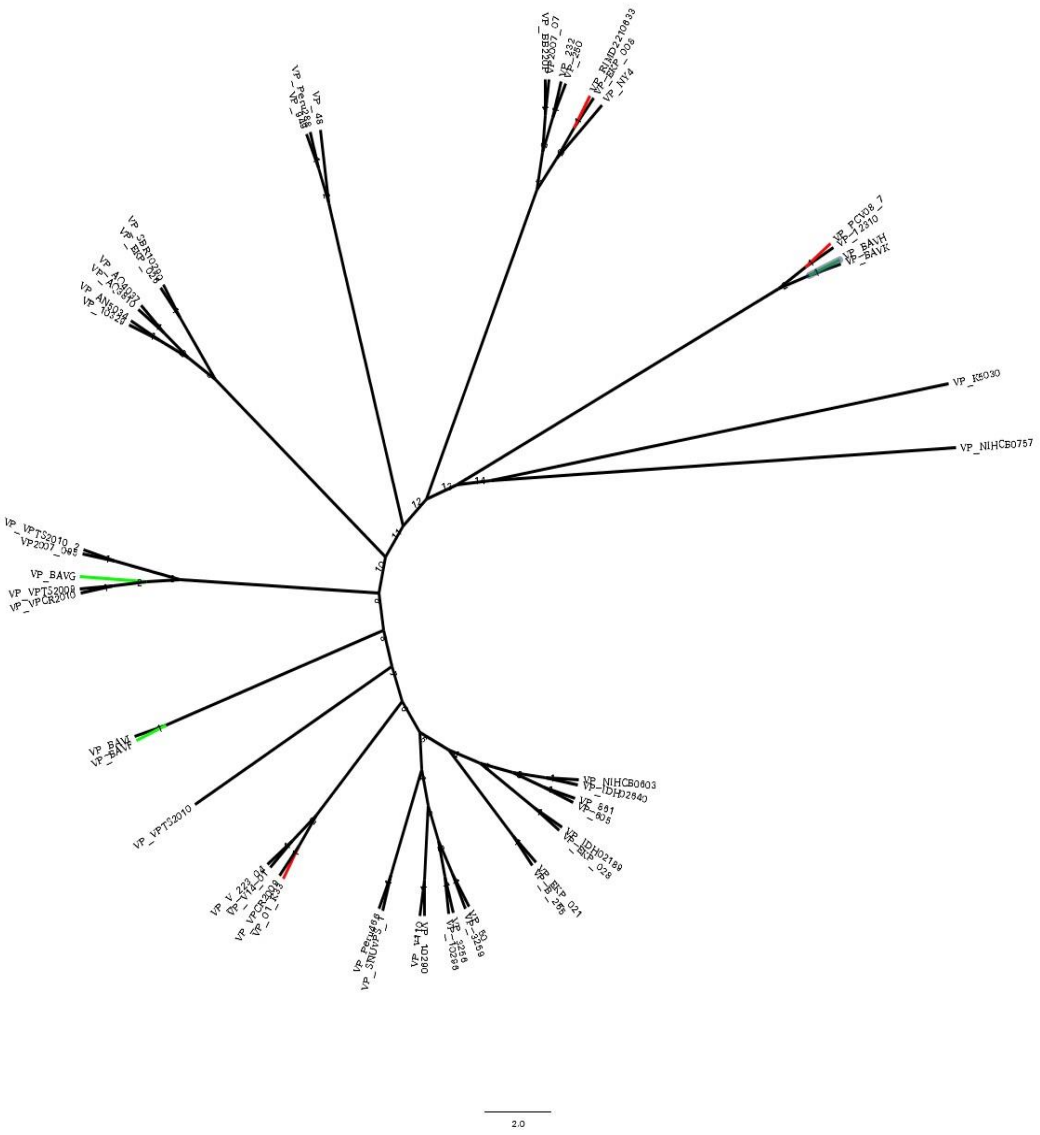
(i)


VP_BAN1
VP_BAVI
VP_BAVG
VP_BAHK
VP_SBR02290
VP_BKP_029
VP_AQ8057
VP_AQ9610
VP_ID329
VP_AN5034
VP_VPTS2010_2
VP_NV4
VP_946
VP_48
VP2007_07
VP_RLMD2210633
VP_10296
VP_10290
VP_1110
VP_1358
VP_BKP_021
VP_VPTS2010
2.0

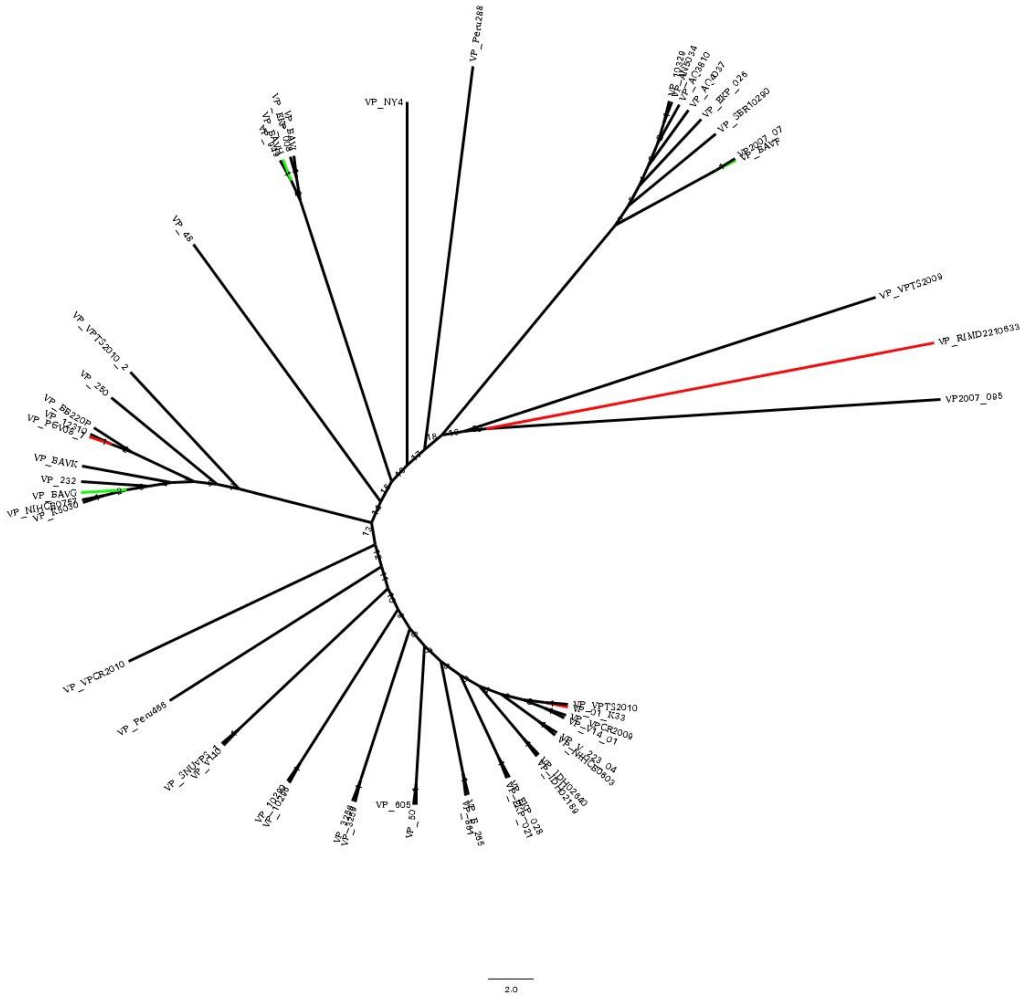
(ii)


VP_250
VP_BAVJ
VP_K5030
VP2007_095
VP_VPT32009
VP_VPCR2010
VP_VPT52010
VP_V14_01
VP_V_223_04
VP_NIHCB0603
VP_IDH02460
VP_IDH02189
VP_EKP_028
VP_EKP_021
VP_B_265
VP_851
VP_805
VP_50
VP_3259
VP_19
VP_12310
VP_10280
VP_10368
VP_7110
2.0

(iii)



(iv)


VP_Peru268
VP_NY4
VP_VPT52009
VP_RIMD2210633
VP2007_095
VP_48
VP_VPT52010_2
VP_250
VP_BAVK
VP_232
VP_BAVG
VP_VPCR2010
VP_Peru486
VP_SNUVP1.1
VP_805
VP_VPT32010
VP_VPCR2009
VP_IDH2149
2.0

(v)