



RESEARCH ARTICLE

Architecture of the superintegron in *Vibrio cholerae*: identification of core and unique genes [v1; ref status: indexed, <http://f1000r.es/w6>]

Michel A Marin, Ana Carolina P Vicente

Laboratory of Molecular Genetics of Microorganisms, Oswaldo Cruz Institute (IOC), Rio de Janeiro, 4365, PO Box 926 CEP 21045-900, Brazil

v1 First published: 27 Feb 2013, 2:63 (doi: [10.12688/f1000research.2-63.v1](https://doi.org/10.12688/f1000research.2-63.v1))
 Latest published: 27 Feb 2013, 2:63 (doi: [10.12688/f1000research.2-63.v1](https://doi.org/10.12688/f1000research.2-63.v1))

Abstract

Background: *Vibrio cholerae*, the etiologic agent of cholera, is indigenous to aquatic environments. The *V. cholerae* genome consists of two chromosomes; the smallest of these harbors a large gene capture and excision system called the superintegron (SI), of ~120 kbp. The flexible nature of the SI that results from gene cassette capture, deletion and rearrangement is thought to make it a hotspot of *V. cholerae* diversity, but beyond the basic structure it is not clear if there is a core genome in the SI and if so how it is structured. The aim of this study was to explore the core genome structure and the differences in gene content among strains of *V. cholerae*.

Methods: From the complete genomes of seven *V. cholerae* and one *Vibrio mimicus* representative strains, we recovered the SI sequences based on the locations of the structural gene *Int14* and the *V. cholerae* repeats. Analysis of the pangenome, including cluster analysis of functional genes, pangenome profile analysis, genetic variation analysis of functional genes, strain evolution analysis and function enrichment analysis of gene clusters, was performed using a pangenome analysis pipeline in addition to the R scripts, splitsTree4 and genoPlotR.

Results and conclusions: Here, we reveal the genetic architecture of the *V. cholerae* SI. It contains eight core genes when *V. mimicus* is included and 21 core genes when only *V. cholerae* strains are considered; many of them are present in several copies. The *V. cholerae* SI has an open pangenome, which means that *V. cholerae* may be able to import new gene cassettes to SI. The set of dispensable SI genes is influenced by the niche and type species. The core genes are distributed along the SI, apparently without a position effect.

Article Status Summary

Referee Responses

Referees	1	2	3
v1 published 27 Feb 2013	report	report	report

- 1 Thandavarayan Ramamurthy**, National Institute of Cholera and Enteric Diseases India
- 2 Yan Boucher**, University of Alberta Canada
- 3 Nur Hasan**, University of Maryland USA

Latest Comments

No Comments Yet

Corresponding author: Michel A Marin (mfabanto@ioc.fiocruz.br)

How to cite this article: Marin MA and Vicente ACP (2013) **Architecture of the superintegron in *Vibrio cholerae*: identification of core and unique genes** [v1; ref status: indexed, <http://f1000r.es/w6>] *F1000Research* 2013, 2:63 (doi: [10.12688/f1000research.2-63.v1](https://doi.org/10.12688/f1000research.2-63.v1))

Copyright: © 2013 Marin MA and Vicente ACP. This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the [Creative Commons Zero "No rights reserved" data waiver](#) (CC0 1.0 Public domain dedication).

Grant information: This work was supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) doctoral fellowship to MAM. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: No relevant competing interests were disclosed.

First published: 27 Feb 2013, 2:63 (doi: [10.12688/f1000research.2-63.v1](https://doi.org/10.12688/f1000research.2-63.v1))

First indexed: 14 Apr 2014, 2:63 (doi: [10.12688/f1000research.2-63.v1](https://doi.org/10.12688/f1000research.2-63.v1))

Introduction

Vibrio cholerae is a diverse, environmental, gram-negative bacterial species that can be pathogenic and can cause cholera, a severe diarrheal disease that occurs most frequently in epidemic form^{1,2}. The *V. cholerae* genome consists of two chromosomes. The largest chromosome of 2.96 Mbp encodes most essential genes. The 1.07 Mbp small chromosome contains few essential genes and the superintegron (SI), a large gene capture and excision system of ~120 kbp² (Figure 1). The SI is characterized by a site-specific integrase gene (*IntI4*) closely associated with a cognate recombination site *attI* and a promoter *Pc* followed by a large array of gene cassettes. Within the SI, the gene cassettes generally consist of a promoterless open reading frame (ORF) flanked by two recombination sites termed *V. cholerae* repeats (VCRs)³. Cassettes can be excised from any position in the array through VCR × VCR recombination mediated by the integrase. The resulting circular intermediate can then be integrated, preferentially through *attI* × VCR recombination by the integrase, bringing the cassette under control of *Pc*^{4,5}. Since gene cassettes are usually promoterless, only the first few cassettes are expressed by *Pc* and the rest of the array can be seen as a reservoir of standing genetic variation⁵.

The functions of the majority of the SI genes are unknown; however, a few genes have been characterized and it has been suggested that they are involved in adaptive functions such as toxin-antitoxin (TA) loci. TA loci consist of two genes in an operon encoding a ‘toxin’ and an ‘antitoxin’. The expression of the toxins reduces cell growth and prevents colony formation, thus exerting a bacteriostatic rather than bacteriocidal condition. However, cell viability can be rescued by later overproduction of the cognate antitoxins⁶.

The pangenome describes the complete repertoire of genes in a bacterial species, which includes the “core genome” containing genes present in all strains, a “dispensable genome” containing genes present in two or more strains, and “unique genes” specific to single strains⁷. Previous phylogeographic analysis, considering *V. cholerae* strains and its sister species *Vibrio metecus*⁸, showed that, in contrast to the core genome, the SI displays strong geographical differentiation, and cassettes from the *V. cholerae* group cluster with those of *V. metecus* from the same place rather than with cassettes from geographically distinct *V. cholerae*. It suggested

that SI structure is influenced by geographic boundaries and in response to environmental conditions. The flexible nature of the SI that results from gene cassette capture, deletion and rearrangement is thought to make it a hotspot of *V. cholerae* diversity, but beyond the basic structure it is not clear if there is a core genome in the SI and if so how it is structured. The aim of this work was to explore the core genome structure and the differential gene content among strains of *V. cholerae*.

Methodology

Based on the complete genomes of seven *V. cholerae* and one *V. mimicus* representative strains (Table 1), we searched repeats above 10 nucleotides and used one VCR sequence (AAC AAA CGC CTC AAG AGG GAC TGT CAA CGC GTG GCG TTT CCA GTC CCA TTG AGC CGT GGT GGT TTC GGT TGT TGT GTT TGA GTT TAG TGT TAT GCG TTG TCA GCC CCT TAG GCG GGC G) to search for sequences with more of 45% nucleotide identity. The SI sequences were recovered using the locations of the structural gene *IntI4* and VCRs identified with the UGENE software⁹. Cluster analysis of functional genes was performed using the pangenome analysis pipeline¹⁰, which searches for homologs or orthologs among multiple genomes using the MultiParanoid (MP) method (based on a 90% nucleotide identity threshold). For each pair of genes in the same cluster, the local matched region is no less than 25% of the longer protein coding sequence and the global matched region is no less than 50% of the longer protein coding sequence. The minimum score value and E-value in BLAST are 50 and 1e-8¹⁰. The gene content was converted to a presence/absence (0/1) matrix and then the core, dispensable and unique genes were identified by in-house R scripts. The phylogenetic tree based in gene content and split network for gene content were constructed with SplitsTree4¹¹ using the GeneContentDistance method¹². The SI structure and comparison of seven *V. cholerae* and, their sister species, *V. mimicus* were performed using genoPlotR¹³.

Results and discussion

SI regions were extracted from the seven *V. cholerae* and one *V. mimicus* genomes (Table 1). The 1285 genes recovered were clustered and a total of 408 clusters were detected (Figure 2A; Table S1). The pangenome of the SI of *Vibrio* strains evaluated was 408 genes, of which eight correspond to core genes, 196 are

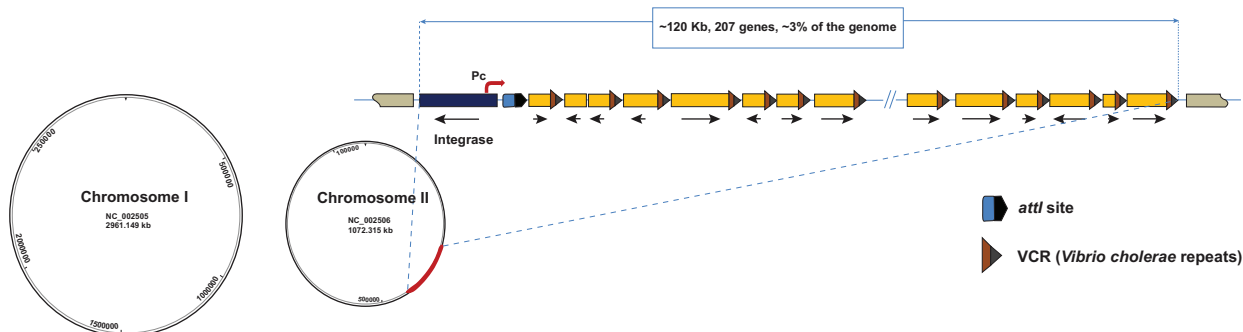


Figure 1. Schematic organization of the *Vibrio cholerae* genome and the superintegron (SI). The functional platform of the SI consists of an integrase gene, a cassette promoter (*Pc*), and a primary recombination site (*attI*). The system maintains an array of several cassettes, which generally consist of a promoterless ORF flanked by two recombination sites termed VCR (*V. cholerae* repeats).

Table 1. Superintegron regions extracted from *V. cholerae* and *V. mimicus* genomes.

Organism	Serogroup/ Biotype	Geographical origin	Source of isolation	Year of isolation	Start	End*	Size (bp)	G+C (%)	ORFs	Locus Int/4	Accession in NCBI
<i>V. cholerae</i> N16961	O1 El Tor	Bangladesh	Clinical	1975	309750	435418	125669	42.20	166	VCA0291	NC_002506
<i>V. cholerae</i> 2010EL1786	O1 El Tor	Haiti	Clinical	2010	36195	135658	99464	42.08	138	Vch1786_I10037	NC_016446
<i>V. cholerae</i> MJ-1236	O1 El Tor	Mattlab, Bangladesh	Clinical	1994	931735	1050596	118862	41.46	135	VCD_000984	NC_012667
<i>V. cholerae</i> O395	O1 Classical	India	Clinical	1965	799827	916350	116524	41.35	175	VCO395_0938	NC_009456
<i>V. cholerae</i> LMA3984	O1	Para, Brazil	Environmental	2007	294428	332847	38420	42.70	47	VCLMA_B0259	NC_0117269
<i>V. cholerae</i> M66-2	O1	Indonesia	Clinical	1937	310949	409433	98485	42.15	133	VCM66_A0290	NC_012580
<i>V. cholerae</i> IEC224	O1	Para, Brazil	Clinical	1990s	309717	435237	125521	42.21	167	O3Y_14823	NC_016945
<i>V. mimicus</i> MB-451	ND	Mattlab, Bangladesh	Clinical	ND	744870	872905	128036	41.39	115	VII_000636	NZ_ADAAF01000002

*Nucleotide position on the chromosome. ND, not determined.

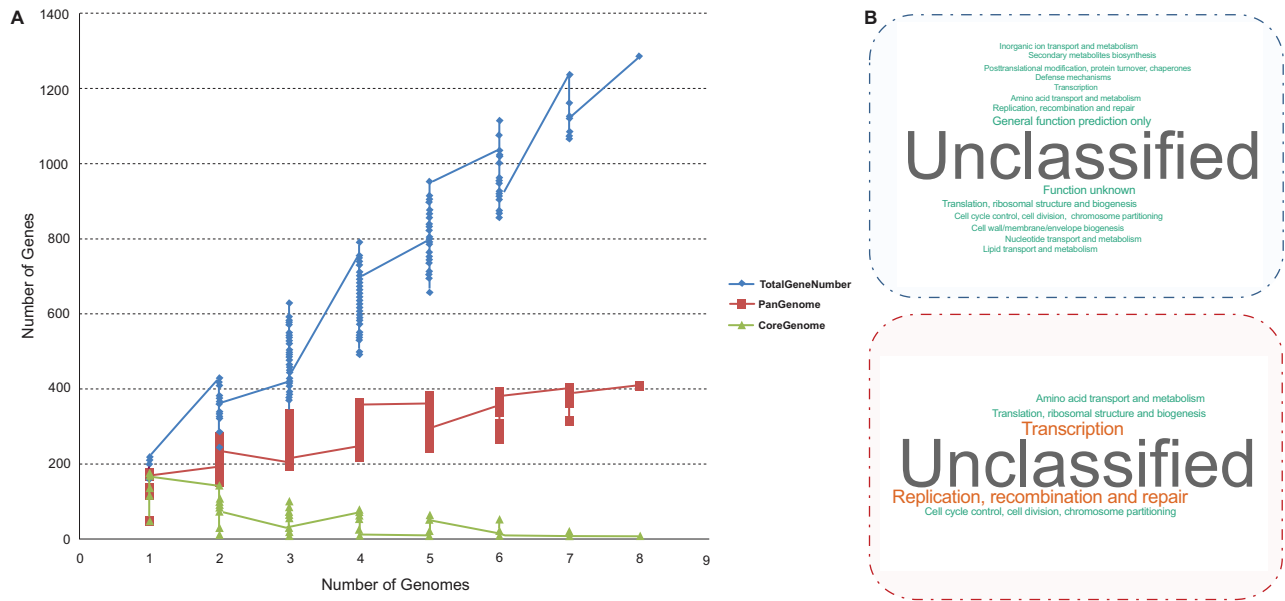


Figure 2. (A) Pangenome plot of the SI region considering seven *V. cholerae* and one *V. mimicus* genomes. 1285 total genes, 408 pangenome clusters and eight core clusters were identified. (B) Word clouds of cluster function enrichment comparison according to clusters of orthologous groups (COG) for whole and core clusters identified are shown at the top and bottom, respectively. Clusters that are not assigned in the COG classification were excluded from the figure.

distributed or dispensable genes and 204 are unique genes. Six of the eight core genes are present in many copies (Table 2). The pangenome profile analysis shows that the cluster numbers of core genome are almost the same, when the SI considered reaches nine, while the pangenome is still increasing (Figure 2A). We infer that the *V. cholerae* SI has an open pangenome, which means that *V. cholerae* may have the ability to import new SI gene cassettes, which affect its plasticity and diversity. On the other hand, the set of SIs, from clinical and environmental lineages, used in this study are apparently representative of this species because allowed to establish that the core genome is close to being completed.

Function enrichment analysis of gene clusters were performed according to description of gene annotation (File S1) supplied to the pangenome analysis pipeline¹⁰. From the 408 clusters, 329 were unclassified by the function enrichment analysis. Following the categorization of Cluster of Orthologous Groups (COG), the characterized clusters were rich in the following categories: translation, ribosomal structure and biogenesis, transcription, replication, recombination and repair, cell cycle control, cell division, chromosome partitioning, defense mechanisms, cell wall/membrane/envelope biogenesis and posttranslational modification, protein turnover, chaperones, amino acid transport and metabolism, nucleotide transport and metabolism, lipid transport and metabolism, secondary metabolites biosynthesis, transport and catabolism (Figure 2B).

In the SI, random excisions occur throughout the cassette array to form nonreplicative circular intermediates containing one or several cassettes; integration events preferentially occur at the *attI* site⁵ and are subjected to selection. It is expected that SI core genes would be

arranged and stay together; however, we found the core genes are distributed along the SI (Figure 3), apparently without any position effect.

We identified 204 unique genes, 94 belonging to *V. mimicus* MB451, nine to LMA3984, 45 to O395, nine to 2010EL1786, 14 to MJ1236, seven to IEC224, 20 to M66, and six to N16961 (Figure 3; Table S1). Considering only the *V. cholerae* SI, there are 21 core genes, most of them present in many copies and rich in the transcription, replication, recombination and repair, translation, ribosomal structure and biogenesis categories.

Pandey and Gerdes¹⁴ identified 13 TA loci within the SI of the N16961 strain. Here we identified six TA genes as part of core SI genes (Table 2), of which the *relB* genes (VCA0349 and VCA0504) were present in all *V. cholerae* strains (including *V. mimicus*) SIs. The *parE* (VCA0359), *relB* (VCA0477) and *relE* (VCA0489) genes were present in all *V. cholerae* SIs. Moreover, we also identified two *higBA* loci (VCA0469 and VCA468), which encode mRNA cleaving enzymes and can stabilize plasmids⁶, as well as SI genes. The previous authors¹⁴ also identified *higBA-I* TA loci (VCA0392 and VCA0391); in our results, these two TA loci are present in all clinical *V. cholerae* strains (Table S1). These results suggest that *V. cholerae* TA loci function as essential stress response elements that help cells survive⁶, as well as act to stabilize the massive arrays of SI cassettes, as reported previously¹⁵.

A previous study suggested that SI structure is influenced by geographic boundaries in response to environmental conditions⁸. Here, we found that the clinical nature of the *V. cholerae* and *V. mimicus* strains evaluated were not grouped together by the analyses

Table 2. Core genes of the *V. cholerae* SI. The table shows the clusters, conservation level between genomes, the functional categories, gene description and the corresponding locus tag in the reference N16961 genome.

ClusterID	Conservation level	COG*	Description	Locus_tag in N16961
1	8	-	hypothetical protein	VCA0407,VCA0353,VCA0336, VCA0297,VCA0302
2	8	COG0456R	acetyltransferase	VCA0470
3	8	-	lipoprotein	VCA0425,VCA0414
4	8	-	hypothetical protein	VCA0381,VCA0435,VCA0357, VCA0306
5	8	-	hypothetical protein	VCA0434,VCA0411
7	8	COG4974L	site-specific recombinase IntI4	VCA0291
8	8	-	relB protein	VCA0349,VCA0504
9	8	COG1670J	acetyltransferase	VCA0505,VCA0436,VCA0417, VCA0316
24	7	COG0110R	acetyltransferase	VCA0473
25	7	COG3668R	plasmid stabilization element ParE	VCA0359
27	7	COG2944K	virulence gene repressor RsaL	VCA0469
31	7	-	hypothetical protein	VCA0497
32	7	COG1694R	mazG-related protein	VCA0485
33	7	-	cytotoxic translational repressor of toxin-antitoxin stability system	VCA0468
34	7	COG0346E	glyoxalase/bleomycin resistance protein	VCA0506,VCA0347
35	7	-	hypothetical protein	VCA0486
37	7	COG2161D	antitoxin of toxin-antitoxin stability system	VCA0477
40	7	COG0456R	GCN5-related N-acetyltransferase	VCA0382
41	7	COG1943L	IS1004 transposase	VCA0493
43	7	COG3668R	plasmid stabilization system protein	VCA0489
44	7	COG3636K	hypothetical protein	VCA0498

*COG: Cluster of Orthologous Groups; "-" depicts no COG assignment.

performed. Therefore, the ability of *V. cholerae* to cause disease must be explained by other virulence factors found outside the SI region.

There are 199 clusters involved with indel or mutation events (Table S2). As for the non-synonymous/synonymous substitution (dN/dS) ratio, we found that 30 clusters were suffering positive selection pressure (dN/dS > 1). At the same time, we could also select those variable clusters as the markers for different strains. Based on pangenome profiles and single nucleotide polymorphism (SNP) information, gene content and phylogenetic trees were

constructed (Figure 4). The SNP information from SI was useful for separating *V. cholerae* from *V. mimicus*, but nevertheless lacked the resolution to distinguish between the different lineages of *V. cholerae*. However, using gene content information (Figure 4), a good resolution was reached that was coherent with the evolution of the species and the environmental or clinical nature of the strains. These results indicate that the evolution of *V. cholerae* into different lineages is reflected in the diversity of the SI, which would be also influenced by horizontal gene transfer in these region, as proposed elsewhere^{8,16,17}.

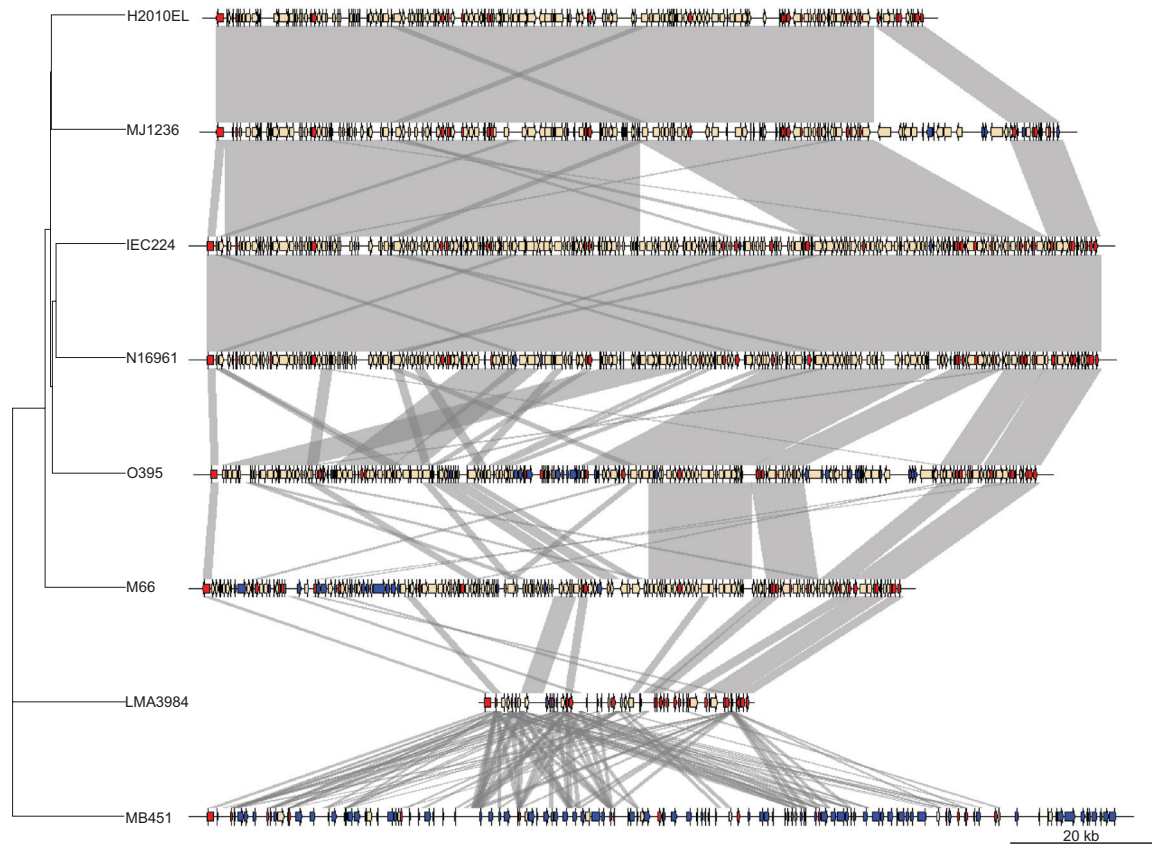


Figure 3. Superintegron (SI) structure and comparison of seven strains of *V. cholerae* and a strain of *V. mimicus*. The core, dispensable and unique genes are indicated by red, cream and blue arrows, respectively. Vertical blocks between sequences indicate regions with more than 1 kb of shared similarity shaded according to BLASTn. A phylogenetic tree based on gene content of the SI is shown on the left.

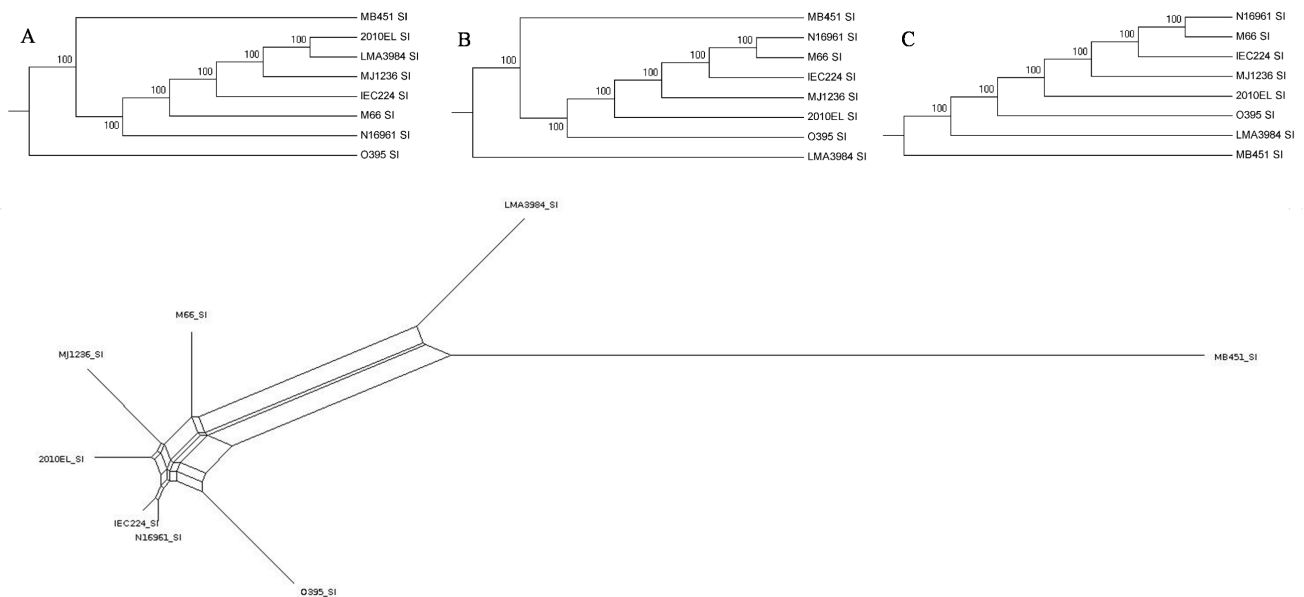


Figure 4. Top: Phylogenetic trees for the *V. cholerae* SI based on SNPs constructed by the Maximum Likelihood (A), Neighbor-Joining (B) and UPGMA (C) methods. The numbers indicate the bootstrap values. **Bottom:** Split network for gene content based on the 408 genes in seven *V. cholerae* and one *V. mimicus* genomes. The network was constructed with SplitsTree4 using the GeneContentDistance method¹².

File S1

1 Data File

<http://dx.doi.org/10.6084/m9.figshare.177719>**File S2**

1 Data File

<http://dx.doi.org/10.6084/m9.figshare.177727>**Table S1**

1 Data File

<http://dx.doi.org/10.6084/m9.figshare.177734>**Table S2**

2 Data Files

<http://dx.doi.org/10.6084/m9.figshare.177736>**Supplementary tables**

Table S1. Orthologs clusters identification among SIs from *V. cholerae* and *V. mimicus* genomes. These clusters were identified using the pangenome analysis pipeline¹⁰, strains without genes in the cluster are marked with “-”.

Table S2. Clusters involved with indel or mutation events. The 1st column is the Cluster ID, which is consistent with the ID in **Table S1**. The 2nd column is the cluster conservation of current cluster. The 3rd column is the variation position, which counts according to the alignment result of protein sequences in this cluster. For indel events, the position is an integer. For synonymous mutation and non synonymous mutation, the position is a floating number, in which the integer part marks the position of the amino acid in the

Conclusions

In this study, we have revealed the genetic architecture of the *V. cholerae* SI, which contains eight core genes, many of them present in many copies. The *V. cholerae* SI has an open pangenome, which means that *V. cholerae* may have the ability to import new gene cassettes into the SI. The set of the dispensable SI gene cassettes is influenced by the niche and type species. The core genes are distributed along the SI, apparently without a position effect.

Author contributions

MAM and ACPV designed the study. MAM and ACPV analyzed the data. MAM and ACPV wrote the article. All authors have approved the final manuscript for publication.

Competing interests

No relevant competing interests were disclosed.

Grant information

This work was supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) doctoral fellowship to MAM.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

alignment result of protein sequences, while the decimal part mark the position of codon. The 4th column shows the amino acid types on current position. The 5th column shows the nucleotide types on current position, indel is marked with “-”. The 6th column shows all gene nucleotide profile in current position (for indel, amino acid will be listed). The 7th column shows the variation type (indel, synonymous and non synonymous). The CDS.variation.analysis spreadsheet shows the summary result for CDS.variation.

Supplementary files

File S1. Gene data of SI region from seven *V. cholerae* and one *V. mimicus* genomes used in this study.

File S2. R scripts used in this study.

References

- Bloesch M, Schoolnik GK: **Serogroup conversion of *Vibrio cholerae* in aquatic reservoirs.** *PLoS Pathog.* 2007; **3**(6): e81.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Heidelberg JF, Eisen JA, Nelson WC, *et al.*: **DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*.** *Nature.* 2000; **406**(6795): 477–83.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Mazel D: **Integrans: agents of bacterial evolution.** *Nat Rev Microbiol.* 2006; **4**(8): 608–20.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Rowe-Magnus DA, Guérout AM, Mazel D: **Super-integrans.** *Res Microbiol.* 1999; **150**(9–10): 641–51.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Cambray G, Guérout AM, Mazel D: **Integrans.** *Annu Rev Genet.* 2010; **44**: 141–66.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Christensen-Dalsgaard M, Gerdes K: **Two *higBA* loci in the *Vibrio cholerae* superintegron encode mRNA cleaving enzymes and can stabilize plasmids.** *Mol Microbiol.* 2006; **62**(2): 397–411.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Medini D, Donati C, Tettelin H, *et al.*: **The microbial pan-genome.** *Curr Opin Genet Dev.* 2005; **15**(6): 589–94.
[PubMed Abstract](#) | [Publisher Full Text](#)

8. Boucher Y, Cordero OX, Takemura A, *et al.*: **Local mobile gene pools rapidly cross species boundaries to create endemicity within global *Vibrio cholerae* populations.** *MBio.* 2011; **2**(2): e00335–10.
[PubMed Abstract](#) | [Publisher Full Text](#)
9. Okonechnikov K, Golosova O, Fursov M: **Unipro UGENE: a unified bioinformatics toolkit.** *Bioinformatics.* 2012; **28**(8): 1166–7.
[PubMed Abstract](#) | [Publisher Full Text](#)
10. Zhao Y, Wu J, Yang J, *et al.*: **PGAP: pan-genomes analysis pipeline.** *Bioinformatics.* 2012; **28**(3): 416–8.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
11. Huson DH, Bryant D: **Application of phylogenetic networks in evolutionary studies.** *Mol Biol Evol.* 2006; **23**(2): 254–67.
[PubMed Abstract](#) | [Publisher Full Text](#)
12. Huson DH, Steel M: **Phylogenetic trees based on gene content.** *Bioinformatics.* 2004; **20**(13): 2044–9.
[PubMed Abstract](#) | [Publisher Full Text](#)
13. Guy L, Kultima JR, Andersson SG: **genoPlotR: comparative gene and genome visualization in R.** *Bioinformatics.* 2010; **26**(18): 2334–5.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
14. Pandey DP, Gerdes K: **Toxin-antitoxin loci are highly abundant in free-living but lost from host-associated prokaryotes.** *Nucleic Acids Res.* 2005; **33**(3): 966–76.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
15. Rowe-Magnus DA, Guerout A, Biskri L, *et al.*: **Comparative analysis of superintegrons: engineering extensive genetic diversity in the Vibrionaceae.** *Genome Res.* 2003; **13**(3): 428–42.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
16. Gao Y, Pang B, Wang HY, *et al.*: **Structural variation of the superintegron in the toxigenic *Vibrio cholerae* O1 El Tor.** *Biomed Environ Sci.* 2011; **24**(6): 579–92.
[PubMed Abstract](#) | [Publisher Full Text](#)
17. Feng L, Reeves PR, Lan R, *et al.*: **A recalibrated molecular clock and independent origins for the cholera pandemic clones.** *PLoS One.* 2008; **3**(12): e4053.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Current Referee Status:

Referee Responses for Version 1



Nur Hasan

Maryland Pathogen Research Institute, University of Maryland, College Park, MD, USA

Approved: 14 April 2014

Referee Report: 14 April 2014

In this manuscript Marin and Vicente investigated the genomic diversity of *V. cholerae* Super Integron (SI) with the aim to identify a set of orthologous genes that are conserved and unique among and in between *V. cholerae* and *V. mimicus* SI's. While one must appreciate the efforts that have gone into the analyses, unfortunately, given the known diversity of *V. cholerae* SI, the number of genomes analyzed was very limited, and was not a good representative of all major phyletic lineages of *V. cholerae* either. Yet, the manuscript provides some valuable information about the diversity and repository of SI genes and their biological functions.

The core genes estimates for SI (21 and 8, among *V. cholerae* and *V. cholerae-V. mimicus* respectively) reported in this study may not be very meaningful as the core genome might diminish or at least reduced further if additional genomes from distinct lineages are included. I strongly recommend inclusion of additional genomes at least from the major phylogenetic lineages of *V. cholerae* O1.

The function of TA loci as an essential stress response element needs to be supported by some experimental data.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.



Yan Boucher

Department of Biological Sciences, University of Alberta, Edmonton, AB, Canada

Approved with reservations: 14 March 2014

Referee Report: 14 March 2014

The article, although it has a very narrow focus, investigates an interesting question about integron regions in *V. cholerae* that scientists have so far mostly applied to whole genome sequences: what is their pan-genome and core genome? The scope of the question is quite narrow, as it focuses on a single species (*V. cholerae*), but is certainly novel.

The methods and approach used are sound and the results generally well presented. I have two major issues with the analysis:

1. Why use only seven genomes when >200 are available? This is quite puzzling to me, as no data should be excluded if it is available. Some of the >200 *V. cholerae* genomes available are relatively

redundant (they all belong to the 7th pandemic group) and present little gene content/SNP diversity, but the variations they present are crucial in understanding small scale variability. Many available environmental *V. cholerae* genomes were also not included and should have been.

2. The article is mostly descriptive and needs to make more informative statements about the results. What is the significance of such a small core genome? How does the pan-genome of the integron overlap with the rest of the genome? What is the meaning of the specific subset of functions found in the integron? How are sites under positive selection distributed in the integron? What genes are under positive/ negative selection? Much more can be said about integrons from the analyses performed.

Basically, the manuscript is interesting but could really benefit from a broader (more genomes) analysis and a more in-depth look at the results to infer hypotheses about integrons in *V. cholerae*.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.



Thandavarayan Ramamurthy

National Institute of Cholera and Enteric Diseases, Kolkata, India

Approved: 09 May 2013

Referee Report: 09 May 2013

In this manuscript the authors have explored the pattern of the superintegrons (SIs) in *V. cholerae* using the published DNA sequences of relevant strains. This study has shown the dynamic nature of *V. cholerae* O1 and genetic relatedness of SIs at the biotype level.

Comments:

- Abstract: Results and Conclusions: replace the word 'reveal' with "describe"
- Introduction: first paragraph: replace the word 'standing' with "standby"
- Page 6. The role of TA gene should be validated with non-toxigenic strains of *V. cholerae* (e.g., sequence comparison with non-O1, non-O139 strains of *V. cholerae*)

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.
