Analysis of 19 Highly Conserved *Vibrio cholerae* Bacteriophages Isolated from Environmental and Patient Sources Over a Twelve-Year Period

Angus Angermeyer [1], Moon Moon Das [2], Durg Vijai Singh [2] and Kimberley D. Seed [1,3,*]

[1] Department of Plant and Microbial Biology, University of California Berkeley, Berkeley, CA, USA; angermeyer@berkeley.edu (A.A.); kseed@berkeley.edu (K.D.S.)
[2] Department of Infectious Disease Biology, Institute of Life Sciences, Nalco Square, Bhubaneswar 751023, India; moon.scial@gmail.com (M.M.D.); durg.singh@gmail.com (D.V.S.)
[3] Chan Zuckerberg Biohub, San Francisco, CA, USA; kseed@berkeley.edu
* Correspondence: kseed@berkeley.edu; Tel.: +1-510-664-7711

Abstract: The *Vibrio cholerae* biotype 'El Tor' is responsible for all current epidemic and endemic cholera outbreaks worldwide. These outbreaks are clonal and are hypothesized to originate from the coastal areas near the Bay of Bengal where the lytic bacteriophage ICP1 specifically preys upon these pathogenic outbreak strains. ICP1 has also been the dominant bacteriophage found in cholera patient stool since 2001. However, little is known about its genomic differences between ICP1 strains collected over time. Here we elucidate the pan-genome and phylogeny of ICP1 strains by aligning, annotating and analyzing the genomes of 19 distinct isolates collected between 2001 and 2012. Our results reveal that ICP1 isolates are highly conserved and possess a large core-genome as well as a smaller, somewhat flexible accessory-genome. Despite its overall conservation, ICP1 strains have managed to acquire a number of unknown genes as well as a CRISPR-Cas system, which is known to be critical for its ongoing struggle for co-evolutionary dominance over its host. This study describes a foundation on which to construct future molecular and bioinformatic studies of this *V. cholerae*-associated bacteriophages.

Keywords: Cholera; Bacteriophage; *Vibrio cholerae*; Pan-genome; Myovirus; Phylogenetics; Vibrio phage

1. Introduction

Vibrio cholerae is a globally-distributed bacterium and the causal agent of the disease cholera, a potentially severe intestinal illness that affects ~1-5 million people resulting in up to ~140,000 deaths annually [1]. The current (seventh) pandemic is comprised of *V. cholerae* biotype 'El Tor' (predominantly serotype O1) and is responsible for current endemic and epidemic disease [2]. Epidemic disease outbreaks sweep the globe periodically and have been traced back to a single lineage that has emerged from the Bay of Bengal region in multiple waves over the last half-century [3]. Despite the overall genetic heterogeneity of this lineage, in which individual outbreaks are nearly always clonal [4], there is an abundance of subtle variation and horizontal transfer that has been observed between outbreaks over time [5,6]. It is hypothesized that the Bay of Bengal serves as a reservoir where El Tor strains circulate throughout the year exchanging genetic material and undergoing ecological selection before infiltrating coastal communities. They are subsequently transported by infected individuals to larger cities where they can be transmitted globally [5,7]. This mechanism is thought to create a bottleneck for strains and result in the clonality of outbreaks.

Bacteriophage, viruses that uniquely infect bacteria, are extremely abundant in the environment where they can outnumber their prokaryotic hosts by several orders of magnitude [8]. As such, bacteriophage play a key role in the evolution of their hosts through both selection and phage-mediated lateral gene transfer [9]. These processes are likely to be very important to *V. cholerae* strain evolution in the Bay of Bengal as well. Previous work has identified a *V. cholerae*

O1-specific [10] lytic *myoviridae* bacteriophage (ICP1) to be of particular interest in this system [11]. In Bangladesh, ICP1 has been found in water samples [12,13] and it has been identified as the dominant phage in cholera patient stool samples since 2001 [11]. The persistence of this phage over time indicates that *V. cholerae* has strategies to limit ICP1 predation, and that ICP1 can evolve to overcome such defenses. Indeed, from this natural genetic laboratory, several complex and surprising adaptations/acquisitions have occurred in the race for survival between *V. cholerae* and ICP1. These include self-mobilizing chromosomal islands that can provide a rapid and efficient response to ICP1 infection [14] and the first known example of a bacteriophage-encoded CRISPR-Cas system [15]. Initial characterization of eight ICP1 isolates collected between 2001-2011 noted the relative low level of diversity and lack of major genomic rearrangements, deletions or insertions [11] (with the exception of its remarkable CRISPR-Cas acquisition [15]). Here we build upon the initial characterization of ICP1 to perform a comparative genomic analysis on 19 individual ICP1 isolates to reconstruct their phylogenetic relationships over time, identify the core and accessory genomes, and infer possible gene function where possible.

V. cholerae is an organism that affects millions and appropriately, it is well-studied with modern bioinformatic and sequencing tools. It is important that their concomitant bacteriophage are similarly studied to help us better elucidate the important role that bacteriophage likely play in the evolution of *V. cholerae* and the epidemiology of the ongoing cholera pandemic.

2. Materials and Methods

Nineteen ICP1 bacteriophage genomes were acquired from various sources (Table 1). Those genomes not specifically described below were downloaded from the NCBI GenBank database. Their metadata (if available) were used to inform our reporting of isolation source and isolation year. In cases where this information differed from what was reported in a genome's original publication we deferred to the GenBank database. Isolation year was used to standardize the ICP1 bacteriophage naming convention: ICP1_YEAR_X where "X" is sequentially assigned letter (A-Z) based on the order in which genomes were named. The only exception is the original, 'ancestral' ICP1 isolate, which is simply referred to as 'ICP1' [11].

Genomes not already available on GenBank include: ICP1_2006_E and ICP1_2011_A, which were isolated and sequenced as described previously [15]. ICP1_2011_A was assembled using CLC Genomics Workbench v10 (Qiagen, Redwood City, CA) and ICP1_2006_E was assembled with IDBA-UD v1.1.3 [16] (default settings) after fastq read filtering with USEARCH v10.0.240 [17] fastq_filter (-fastq_maxee_rate 0.001 -fastq_maxns 1 -fastq_truncqual 15 -fastq_maxee 0.25). ICP1_2011_B was assembled from an existing diarrheal stool metagenomic sample [18] (SRA: PRJEB9150; Run: ERR866578) using USEARCH filtering and IDBA-UD *de novo* assembly as above (same parameters) to generate an incomplete ICP1 contig. That contig was then used as a reference sequence to reassemble the same filtered reads using IDBA-Hybrid [16] with default settings. Finally, ICP1_2012_A was isolated from a cholera patient stool sample collected in Silvassa, India [19]. Genomic DNA libraries were sequenced using an Illumina Mi-Seq (Genotypic Technology, Bangalore, India). Genome assembly was performed with CLC Genomics Workbench v10.

Whole-genome alignment was performed on all 19 genomes using progressiveMauve [20] (build: Feb 25 2015) with default settings. The Mauve xmfa alignment output file was converted to Phylip format using BioPython v1.71 [21] and a maximum likelihood, unrooted, phylogenetic tree was constructed using PhyML v20120412 [22] while calculating bootstrap support (-s BEST --rand_start --n_rand_starts 10 -b 100). A companion phylogeny based on the concatenated core-genome (described below) was constructed using the same methods. Dendroscope3 v3.5.9 [23] was used to generate a tanglegram joining both trees with ICP1 ancestral set as the outgroup. The whole-genome alignment was also used to determine a consensus ICP1 sequence using CLC Genomics Workbench v10, which all ICP1 genomes were visually mapped back onto using BRIG v0.95 [24].

All extant CDS annotations, hereafter referred to as 'Open Reading Frames' (ORFs), in the ICP1 ancestral GenBank file were blasted (BLASTn v2.6.0 [25]) against every other ICP1 genomic sequence to find homologous ORFs. Putative hits were considered homologs, and annotated as such, if the following conditions were met: the subject hit was a complete ORF (start and stop codons; codon table=11), e-value≤ 1e-10, %identity≥85, and the two matched ORFs were within 10% sequence length of each other. After identifying existing homologs, *de novo* ORF prediction was performed on all genomes to find additional possible coding sequences using Prodigal v2.6.3 [26] with default settings and a confidence cutoff ≥95%. All newly identified putative-ORFs from each genome were then blasted against each other with the same conditions as above, grouped by homology and given an iterative numerical identifier based on their locations in the genome relative to the extant ORFs (i.e. ORF1, ORF2, ORF2.1, ORF2.2, ORF3, etc.). ORFs were then categorized into groups based on how many of the 19 genomes they were found it. If an ORF was present in all 19 genomes, it was considered part of the core-genome, otherwise, it was considered part of the accessory-genome. Core and accessory ORFs were also mapped to the BRIG alignment. The core-genome ORF protein sequences were concatenated by strain to create a core-genomic, pseudo-genome for each ICP1 strain and subjected to the phylogenetic analysis as above. The core and accessory ORFs were also queried against the NCBI's Conserved Domain Database (https://www.ncbi.nlm.nih.gov/cdd) to determine if any possessed interesting domain homology that wasn't detected by BLAST alone.

Table 1. ICP1 bacteriophage strains.

Standardized Name	Previous Name	Isolation Year	Isolation Source	Genome Size (bp)	GenBank Accession	Genome Citation
ICP1	-	2001	Stool	125,956	HQ641347	Seed *et al.* 2011
ICP1_2001_A	-	2001	Stool	124,826	HQ641353	Seed *et al.* 2011
ICP1_2001_B	JSF1	2001	Water	126,082	KY883636	Naser *et al.* 2017
ICP1_2001_C	JSF2	2001	Water	126,082	KY883637	Naser *et al.* 2017
ICP1_2001_D	JSF4	2001*	Water*	124,261	KY065147	Naser *et al.* 2017
ICP1_2001_E	JSF5	2001*	Water	132,142	KY883634	Naser *et al.* 2017
ICP1_2001_F	JSF6	2001*	Water	133,685	KY883635	Naser *et al.* 2017
ICP1_2004_A	-	2004	Stool	128,083	HQ641354	Seed *et al.* 2011
ICP1_2005_A	-	2005	Stool	129,373	HQ641352	Seed *et al.* 2011
ICP1_2006_A	-	2006	Stool	123,104	HQ641351	Seed *et al.* 2011
ICP1_2006_B	-	2006	Stool	123,097	HQ641350	Seed *et al.* 2011
ICP1_2006_C	-	2006	Stool	124,497	HQ641349	Seed *et al.* 2011
ICP1_2006_D	-	2006	Stool	124,497	HQ641348	Seed *et al.* 2011
ICP1_2006_E	-	2006	Stool	128,298	TBD	This study
ICP1_2009_A	JSF13	2009	Water	128,814	KY883638	Naser *et al.* 2017
ICP1_2011_A	-	2011	Stool	126,861	TBD	This study
ICP1_2011_B	-	2011	Stool	125,128	TBD	This study
ICP1_2011_C	JSF14	2011	Water	125,096	KY883639	Naser *et al.* 2017
ICP1_2012_A	-	2012	Stool	121,418	TBD	This study

* GenBank metadata was reported in cases where it differed from the original publication.

3. Results

3.1. Genome Characteristics and Phylogeny

ICP1 genomes were analyzed from isolates collected over a 12-year period, from 2001 to 2012 (Table 1). The isolates were derived from both environmental water samples and patient stool samples collected in Bangladesh (n=18) and India (ICP1_2012_A). Genome length was slightly variable with an average of 126,384bp (stdev: 3008bp). The maximum likelihood phylogenetic analysis grouped both the whole-genome and core-genome alignments into several general clusters (Figure 1). Cluster A contains the five most recently isolated strains as well as ICP1_2004_A. This cluster also contains five of the seven total CRISPR-positive strains in the dataset. Cluster B contains four of the six 2001 isolates, ICP1_2006_E and two of the CRISPR-positive strains. Clusters C and D contain isolates from intermediate years 2005 and 2006, with one CRISPR-positive represented in cluster C. The CRISPR-positive strain ICP1_2001_F did not cluster with other isolates. ICP1 ancestral and ICP1_2001C also cluster closely together but are not specifically highlighted.

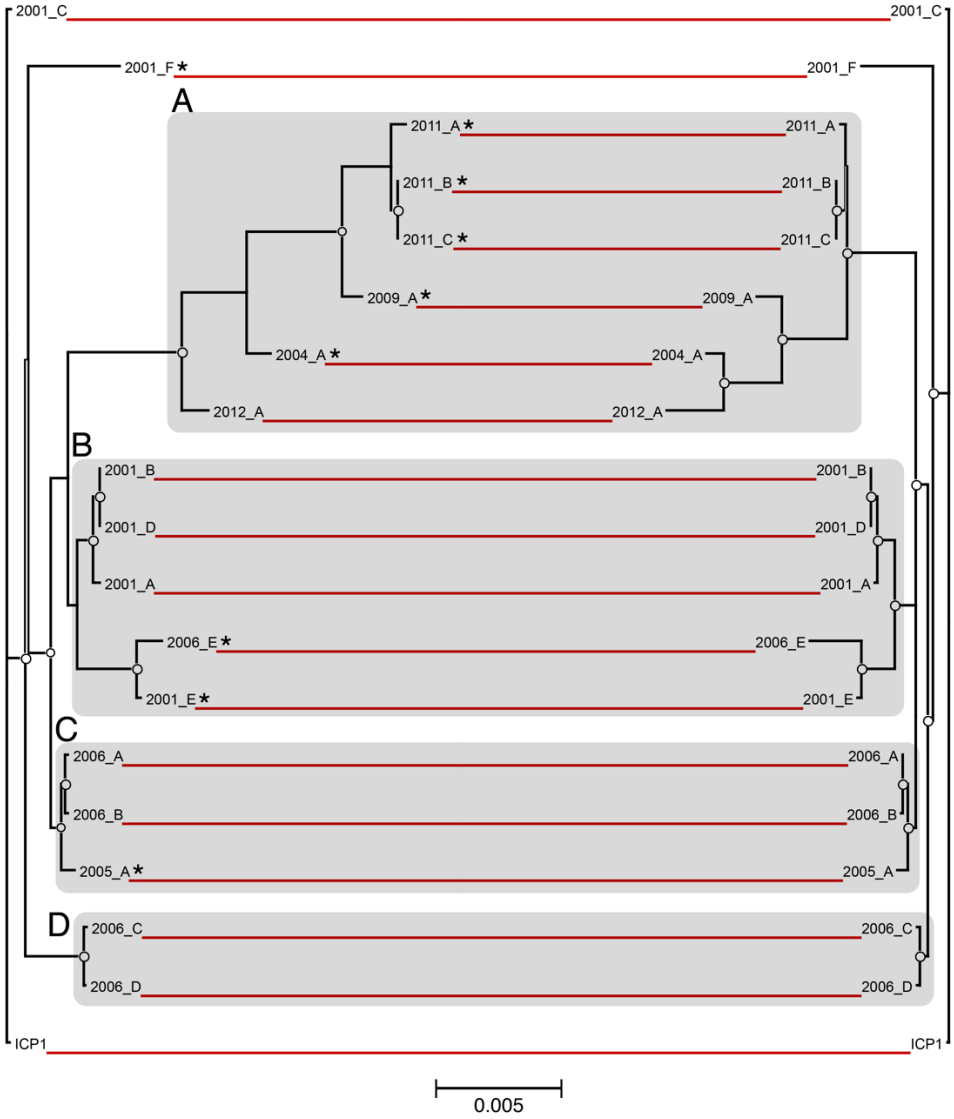


Figure 1. Phylogenetic comparison of ICP1 whole and core-genomes. Maximum likelihood phylogenetic trees (unrooted) were constructed based on a multiple whole-genome alignment of all 19 ICP1 phage sequences (left) as well as an alignment of the concatenated ORFs for each phage that comprise the core genome (right). The red lines connect identical leaves between trees to indicate relative phylogeny. The circles represent nodes with ≥90% bootstrap support (n=100). The scale bar measures nucleotide substitutions per base pair. Distinct phylogenetic clusters are shaded grey and CRISPR strains are marked with '*'.

Overall, the topologies between whole-genome and core-genome trees were highly similar with a few minor exceptions. Both share identical clustering and have almost no leaf-level differences within clusters B, C and D. Cluster A showed an inversion of the phylogenetic differences of the leaves between the two alignments. For instance, ICP1_2012_A has the most divergent core-genome (from all other strains) but is the least divergent within its cluster when the whole-genome was considered.

3.2. Genome Alignment Visualization

The consensus genomic sequence constructed from the whole-genome alignment was 151,832bp long and contained all of the coding and non-coding regions from each genome. It was used as a reference to map the genomes and visualize the overall multiple-genome alignment (Figure 2). Variable regions of insertions and deletions were visible as gaps in the circular alignment. Similar to the whole-genome phylogeny, there was not a clear progression of sequence divergence based on isolation chronology. No large regions of GC content difference were observed.

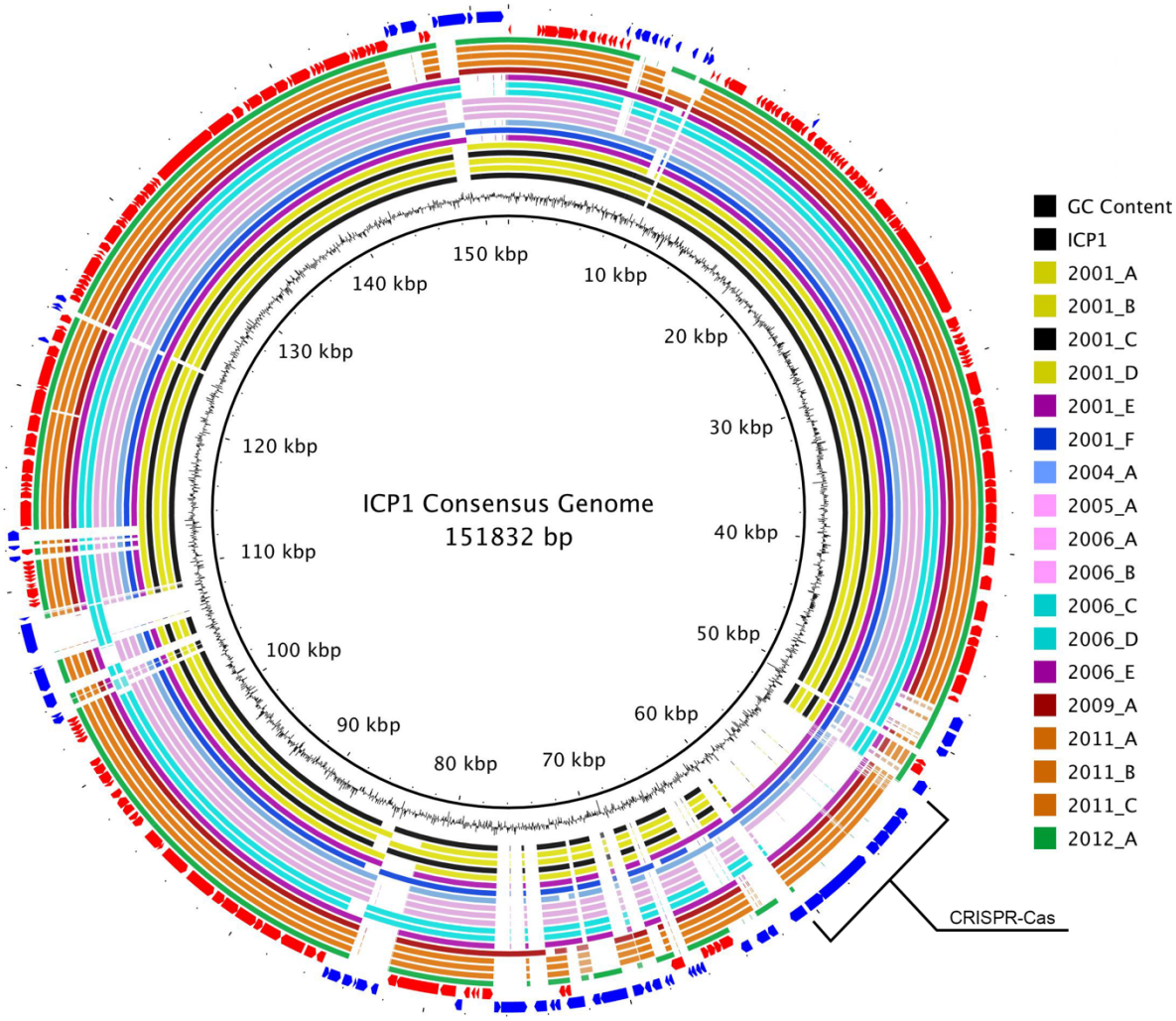


Figure 2. ICP1 pan-genome consensus alignment. A BLASTn-based whole genome alignment of all 19 ICP1 phage genomes using the MAUVE alignment consensus sequence as reference. The innermost ring is the consensus sequence. The next ring represents the GC content for that region. The following 19 rings display the alignment for each genome and are colored by phylogenetic similarity as determined by analysis of whole genomes (Figure 1 left). The second to last ring (red) represents the core-genomic ORFs while the outermost ring (blue) is the accessory-genome ORFs. The CRISPR-Cas insertion region is labeled.

3.3. ORF Annotation

Among all 19 genomes, a total of 269 distinct ORFs (based on homology cutoffs) were identified. Of these, 230 were originally annotated in the ICP1 ancestral genome and 39 were called by Prodigal. The number of ORFs per genome varied from 215 to 232 and demonstrated a slight but significant inverse relationship with year of isolation, i.e. more recent strains had fewer ORFs (Figure S1A). A weaker significant positive trend was observed between number of ORFs and genome length (Figure S1B) which is likely due to the CRISPR-Cas insertions, however there was not a statistically significant correlation between genome length and isolation year.

3.3. Core and Accessory Genome Analysis

To estimate gene diversity among the genomes we divided the total pan-genomic ORF complement into core and accessory groups. The core ORFs, those that were found in all 19 genomes, comprised ~70% of the total number of ORFs (185 out of 269) while the other 84 ORFs were considered to be accessory (Figure 3).

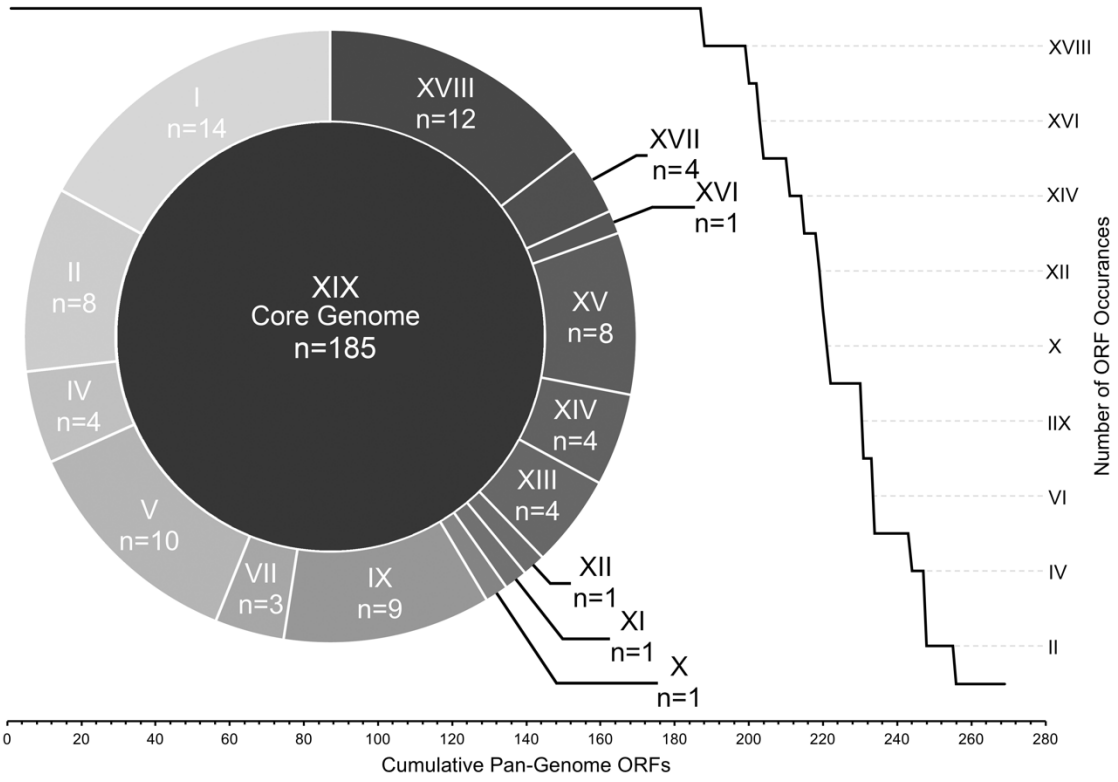


Figure 3. ICP1 pan-genome ORF allocation. All ICP1 ORFs arranged by number of genomes in which they were detected. ORF bins are represented by roman numerals, i.e. 'XIX' ORFs were found in all 19 genomes (core), 'X' ORFs were found in exactly 10 genomes, etc. The line graph demonstrates the overall cumulative curve of core and accessory ORF prevalence. The doughnut chart provides exact number of ORFs within each accessory bin.

Despite their occurrence across all genomes in this study, the core-genome ORFs were not all equally conserved at the sequence level. By comparing average pairwise nucleotide and amino acid sequence identity, we were able to resolve the core ORFs into three distinct groups: conserved-core, synonymous-core and divergent-core (Figure 4). The conserved-core was comprised of 49 ORFs that shared perfect sequence identity among all the genomic sequences. Correspondingly, they shared identical amino acid sequence identity. The synonymous-core included 26 ORFs that

possessed a small amount of nucleotide diversity between genomes, but all mutations were silent, and the amino acid primary structure was therefore identical. Finally, the divergent-core contained the remaining 110 ORFs which were diverse in both nucleotide and amino acid pairwise identity. These divergent ORFs encompassed a range pairwise identity similarities from 99.97% to 91.57% at the nucleotide level and 99.98% to 91.83% for amino acid sequences (Table S1). The degree of pairwise identity difference was not correlated with an ORF's sequence length (Figure 4).

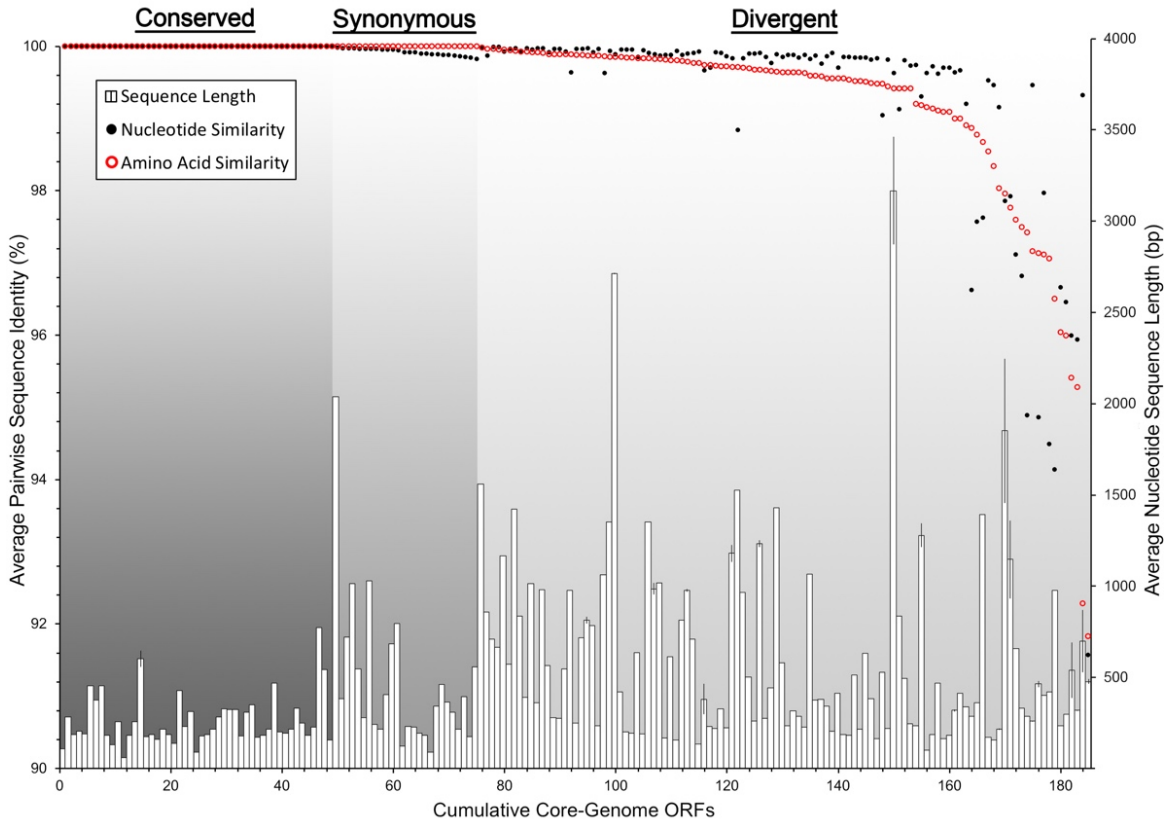


Figure 4. ICP1 core-genome ORF divergence. All ICP1 core-genome ORFs arranged by average pairwise similarity (171 pairwise comparisons) of both DNA nucleotide sequence alignments (black dots) and amino acid residue alignments (red circles). The histogram shows average nucleotide sequence length and standard deviation among the 19 sequences per ORF. The graph is divided into three sections by types of ORF similarity: Conserved (identical nucleotide and amino acid sequences), Synonymous (identical amino acid sequences, but silent nucleotide mutations) and Divergent (dissimilarity in both nucleotide and amino acid sequences).

The accessory ORFs were distributed among all 19 genomes in a complex manner (Table S2) and grouped into 15 levels of occurrence. These ranged from occurrences in 18 genomes (n=12) to singletons (n=14) (Figure 3) with several patterns of accessory ORF co-occurrence (Table S2). The most obvious example was the CRISPR-Cas locus which was previously known to reside in 8 of these genomes [13,15] and confirmed through our annotation methods. Other examples of sequential ORF co-occurrence included a locus containing ORFs 115, 116, 117, 117.1 and 118 which were found in the same five genomes, ORFs 160, 162, 163 in a different set of five overlapping genomes and ORFs 222, 223, 225 in 15 genomes.

3.3. Conserved Functional Domains

The majority of ORFs in both the core and accessory genomes were classified as hypothetical due to a lack of an informative BLAST identification. Only 18 of the core ORFs (9.7%) currently had a predicted function, two in the conserved-core, three in the

synonymous-core and 13 in the divergent-core (Table S1). The accessory core contains 11 ORFs with a putative or known function (Table S2). The NCBI conserved domain search identified 18 additional core-genome ORFs and 11 accessory-genome ORFs that shared at least partial amino acid sequence homology to known functional domains.

4. Discussion

ICP1 is a *V. cholerae*-infecting bacteriophage that appears to be prevalent throughout the Bay of Bengal's coastal areas and is transferred readily alongside *V. cholerae* into humans during cholera outbreaks in the region [11]. It is becoming increasingly theorized that this region is the source of most if not all global cholera outbreaks [3,4,27], and therefore a better understanding of this co-evolving predator is an important area of ongoing cholera research. In this study we have performed a comprehensive phylogenetic analysis on all available, well-sequenced ICP1 isolates to elucidate their genetic divergence over time and provide a platform on which to develop future ICP1-related bioinformatic analyses.

We have found that the genomes of ICP1 are surprisingly well-conserved between all isolates over the twelve-year period in which they were isolated. This is demonstrated in the whole-genome phylogeny which, while resolvable into distinct phylogenetic clusters, still only represents a maximum variation of approximately 1 nucleotide substitution per 100 base pairs between the most divergent isolates (Figure 1), many of which are likely silent or non-coding. A high degree of genomic conservation is also indicated by the relatively large core-genome shared between all isolates (Figure 3). This conservation is not only surprising due to the amount of time the core-genome has remained stable, but also due to the complex conditions that likely exist in the coastal and ocean environments where ICP1 is competing against a host population that is almost certainly more diverse than clonal outbreak strains. In at least one other study that examined multiple strains of a single marine bacteriophage, Far-T4, it was shown that sequence variability among strains was at least an order of magnitude greater than for ICP1 [28]. In contrast, a study from a less variable environment found that strains of a *Y. enterocolitica*-infecting podovirus were more highly conserved than ICP1 [29]. However, it must be considered that these environmental pressures may actually be what drives this bacteriophage to maintain a large core-genome, of which the vast majority of ORFs are hypothetical. This large gene complement may be providing the flexibility needed to compete in the Bay of Bengal. Currently we can only speculate on the exact reason for the conservation of ICP1's genome, but expanding the repertoire of well-sequenced genomes will help to put a finer constraint on the core-genome and perhaps identify genetic targets for future study.

Despite the stable conservation of the core-genome, ICP1 also possesses a diverse collection of accessory genes that have been integrated into several locations around the pan-genome (Figure 2). These accessory ORFs appear to be both single acquisitions as well as integrations of larger loci (Table S2). The most notable of the latter is the previously described acquisition of a CRIPSR-Cas system which is used as a weapon in the arms race between ICP1 and *V. cholerae* [14,15]. This may suggest that other mechanisms of molecular warfare would be likely targets for acquisition, but if so, the conserved domain analyses failed to reveal mechanisms already known. Interestingly, though, the trend has been for genomic ORF counts to diminish over time culminating in the latest isolate from India in 2012 which possessed the fewest overall number of ORFs, i.e. the smallest accessory-genome (Figure S1). This could be due to shedding of outdated or detrimental genes, possibly because the acquisition of a CRISPR-Cas system provided enough flexibility to obviate the need for other mechanisms. And although ICP1_2012_A is CRISPR-Cas negative, the other five most recent isolates are positive. It is also possible that this is a regional difference though, and more sequenced genomes will help to determine if this trend is chronological or geographical or spurious.

Querying the ORFs against the NCBIs conserved domain database returned several hits having to do with replication, nucleotide metabolism, recombination, endonuclease activity and a few other basic functions (Table S1, S2). However, what may be most telling is that 80% of all ORFs do

not possess homology to any conserved domain indicating that there is a great deal left to learn about the interaction between ICP1 and *V. cholerae*. It should also be noted that the annotation of ORFs necessarily requires certain assumptions to be made about similarity cutoffs and thresholds and as such these ORF calls are best viewed as estimates. However, we were conservative in our methods and are confident that the a very large proportion of the calls are accurate.

As we advance our understanding of how cholera spreads globally, it will be important to also continue tracking ICP1's phylogeny and genetic composition so that we may develop a better understanding of its co-evolution with *V. cholerae* and attempt to disentangle the complex molecular and ecological interactions that may play an important role in defining cholera outbreaks.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Figure S1: ORF occurrence trends by year and genome length, Table S1: Core-genome ORFs pairwise similarity and CDD hits, Table S2: Accessory-genome ORFs ICP1 strain matrix.

Author Contributions: Angus Angermeyer and Kimberley Seed conceived of the study; Angus Angermeyer analyzed the data; Moon Moon Das and Durg Vijai Singh contributed sequencing data and reagents. Angus Angermeyer and Kimberley Seed wrote the paper.

Funding: This research was funded by the National Institute of Allergy and Infectious Diseases grant number R01AI127652 and the Chan Zuckerberg Biohub.

Acknowledgments: We wish to thank Zach Barth for his helpful assistance and **Genotypic Technology, Bangalore, India** for the sequencing of ICP1_2012_A.

Conflicts of Interest: The authors declare no conflict of interest. The funding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

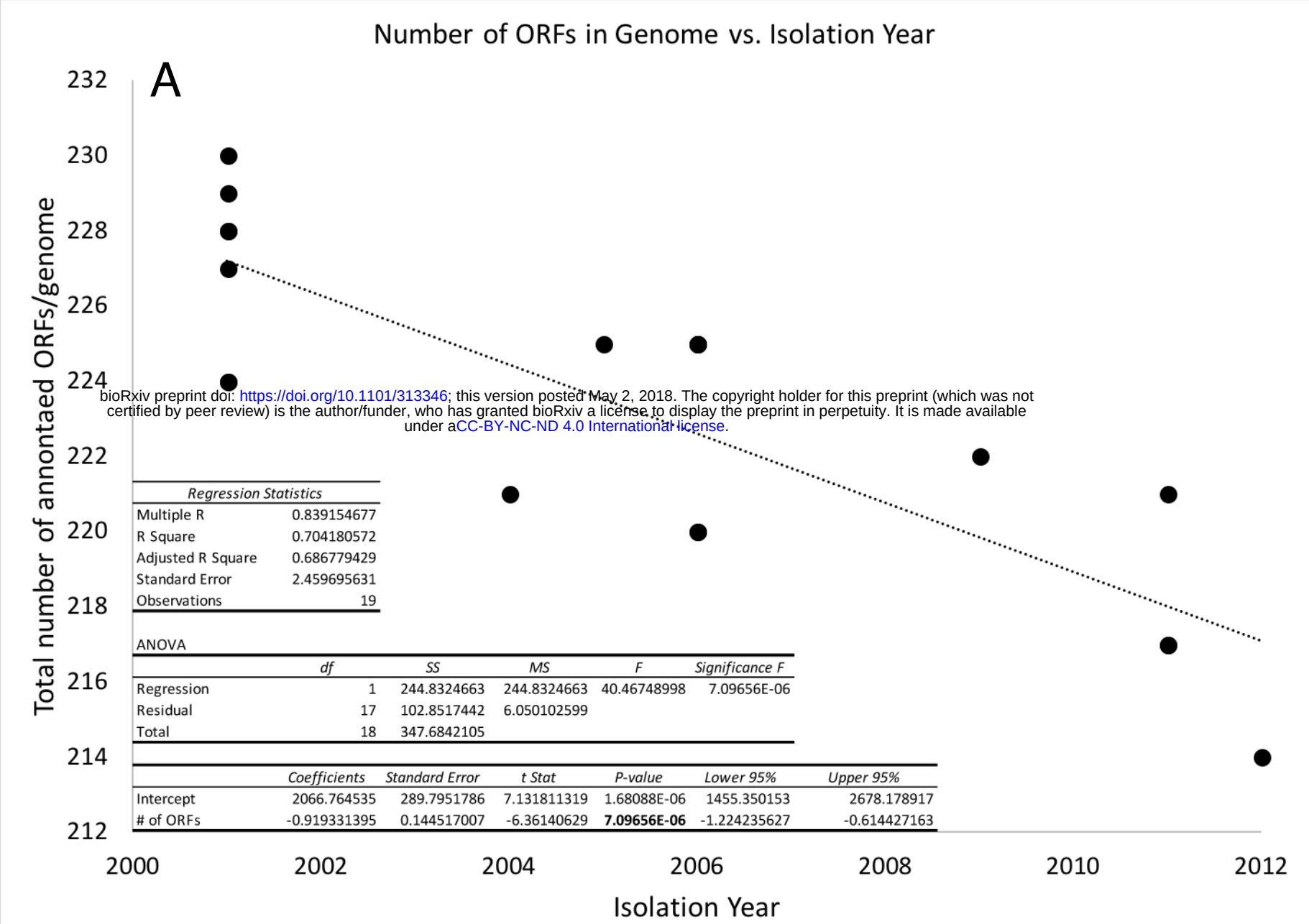
References

1. Ali, M.; Nelson, A. R.; Lopez, A. L.; Sack, D. A. Updated global burden of cholera in endemic countries. *PLoS Negl Trop Dis* **2015**, *9*, e0003832.
2. Safa, A.; Nair, G. B.; Kong, R. Y. C. Evolution of new variants of Vibrio cholerae O1. *Trends Microbiol.* **2010**, *18*, 46–54.
3. Mutreja, A.; Kim, D. W.; Thomson, N. R.; Connor, T. R.; Lee, J. H.; Kariuki, S.; Croucher, N. J.; Choi, S. Y.; Harris, S. R.; Lebens, M.; Niyogi, S. K.; Kim, E. J.; Ramamurthy, T.; Chun, J.; Wood, J. L. N.; Clemens, J. D.; Czerkinsky, C.; Nair, G. B.; Holmgren, J.; Parkhill, J.; Dougan, G. Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* **2011**, *477*, 462–465.
4. Moore, S.; Thomson, N.; Mutreja, A.; Piarroux, R. Widespread epidemic cholera caused by a restricted subset of Vibrio cholerae clones. *Clin. Microbiol. Infect.* **2014**, *20*, 373–379.
5. Weill, F.-X.; Domman, D.; Njamkepo, E.; Tarr, C.; Rauzier, J.; Fawal, N.; Keddy, K. H.; Salje, H.; Moore, S.; Mukhopadhyay, A. K.; Bercion, R.; Luquero, F. J.; Ngandjio, A.; Dosso, M.; Monakhova, E.; Garin, B.; Bouchier, C.; Pazzani, C.; Mutreja, A.; Grunow, R.; Sidikou, F.; Bonte, L.; Breurec, S.; Damian, M.; Njanpop-Lafourcade, B.-M.; Sapriel, G.; Page, A.-L.; Hamze, M.; Henkens, M.; Chowdhury, G.; Mengel, M.; Koeck, J.-L.; Fournier, J.-M.; Dougan, G.; Grimont, P. A. D.; Parkhill, J.; Holt, K. E.; Piarroux, R.; Ramamurthy, T.; Quilici, M.-L.; Thomson, N. R. Genomic history of the seventh pandemic of cholera in Africa. *Science* **2017**, *358*, 785–789.
6. Cho, Y.-J.; Yi, H.; Lee, J. H.; Kim, D. W.; Chun, J. Genomic evolution of Vibrio cholerae. *Curr Opin Microbiol* **2010**, *13*, 646–651.
7. Domman, D.; Quilici, M.-L.; Dorman, M. J.; Njamkepo, E.; Mutreja, A.; Mather, A. E.; Delgado, G.; Morales-Espinosa, R.; Grimont, P. A. D.; Lizárraga-Partida, M. L.; Bouchier, C.; Aanensen, D. M.;

Kuri-Morales, P.; Tarr, C. L.; Dougan, G.; Parkhill, J.; Campos, J.; Cravioto, A.; Weill, F.-X.; Thomson, N. R. Integrated view of Vibrio cholerae in the Americas. *Science* **2017**, *358*, 789–793.

8. Parikka, K. J.; Le Romancer, M.; Wauters, N.; Jacquet, S. Deciphering the virus-to-prokaryote ratio (VPR): insights into virus-host relationships in a variety of ecosystems. *Biol Rev Camb Philos Soc* **2017**, *92*, 1081–1100.
9. Ofir, G.; Sorek, R. Contemporary Phage Biology: From Classic Models to New Insights. *Cell* **2018**, *172*, 1260–1270.
10. Seed, K. D.; Faruque, S. M.; Mekalanos, J. J.; Calderwood, S. B.; Qadri, F.; Camilli, A. Phase Variable O Antigen Biosynthetic Genes Control Expression of the Major Protective Antigen and Bacteriophage Receptor in Vibrio cholerae O1. *PLoS Pathog* **2012**, *8*, e1002917–13.
11. Seed, K. D.; Bodi, K. L.; Kropinski, A. M.; Ackermann, H.-W.; Calderwood, S. B.; Qadri, F.; Camilli, A. Evidence of a dominant lineage of Vibrio cholerae-specific lytic bacteriophages shed by cholera patients over a 10-year period in Dhaka, Bangladesh. *MBio* **2011**, *2*, e00334–10.
12. Faruque, S. M.; Naser, I. B.; Islam, M. J.; Faruque, A. S. G.; Ghosh, A. N.; Nair, G. B.; Sack, D. A.; Mekalanos, J. J. Seasonal epidemics of cholera inversely correlate with the prevalence of environmental cholera phages. *Proc Natl Acad Sci USA* **2005**, *102*, 1702–1707.
13. Naser, I. B.; Hoque, M. M.; Nahid, M. A.; Tareq, T. M.; Rocky, M. K.; Faruque, S. M. Analysis of the CRISPR-Cas system in bacteriophages active on epidemic strains of Vibrio cholerae in Bangladesh. *Scientific Reports* **2017**, *7*, 14880.
14. O'Hara, B. J.; Barth, Z. K.; McKitterick, A. C.; Seed, K. D. A highly specific phage defense system is a conserved feature of the Vibrio cholerae mobilome. *PLoS Genet.* **2017**, *13*, e1006838.
15. Seed, K. D.; Lazinski, D. W.; Calderwood, S. B.; Camilli, A. A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. *Nature* **2013**, *494*, 489–491.
16. Peng, Y.; Leung, H. C. M.; Yiu, S. M.; Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **2012**, *28*, 1420–1428.
17. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **2010**, *26*, 2460–2461.
18. David, L. A.; Weil, A.; Ryan, E. T.; Calderwood, S. B.; Harris, J. B.; Chowdhury, F.; Begum, Y.; Qadri, F.; LaRocque, R. C.; Turnbaugh, P. J. Gut microbial succession follows acute secretory diarrhea in humans. *MBio* **2015**, *6*, e00381–15.
19. Das, M. M.; Bhotra, T.; Zala, D.; Singh, D. V. Phenotypic and genetic characteristics of Vibrio cholerae O1 carrying Haitian ctxB and attributes of classical and El Tor biotypes isolated from Silvassa, India. *Journal of Medical Microbiology* **2016**, *65*, 720–728.
20. Darling, A. E.; Mau, B.; Perna, N. T. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* **2010**, *5*, e11147.
21. Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J. L. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009**, *25*, 1422–1423.
22. Guindon, S.; Dufayard, J.-F.; Lefort, V.; Anisimova, M.; Hordijk, W.; Gascuel, O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **2010**, *59*, 307–321.
23. Huson, D. H.; Scornavacca, C. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Syst. Biol.* **2012**, *61*, 1061–1067.

24. Alikhan, N.-F.; Petty, N. K.; Ben Zakour, N. L.; Beatson, S. A. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* **2011**, *12*, 402.
25. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T. L. BLAST+: architecture and applications. *BMC Bioinformatics* **2009**, *10*, 421.
26. Hyatt, D.; Chen, G.-L.; LoCascio, P. F.; Land, M. L.; Larimer, F. W.; Hauser, L. J. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **2010**, *11*, 119.
27. Hu, D.; Liu, B.; Feng, L.; Ding, P.; Guo, X.; Wang, M.; Cao, B.; Reeves, P. R.; Wang, L. Origins of the current seventh cholera pandemic. *Proc Natl Acad Sci USA* **2016**, *113*, E7730–E7739.
28. Roux, S.; Enault, F.; Ravet, V.; Pereira, O.; Sullivan, M. B. Genomic characteristics and environmental distributions of the uncultivated Far-T4 phages. *Front Microbiol* **2015**, *6*, 199.
29. Salem, M.; Skurnik, M. Genomic Characterization of Sixteen Yersinia enterocolitica-Infecting Podoviruses of Pig Origin. *Viruses* **2018**, *10*.

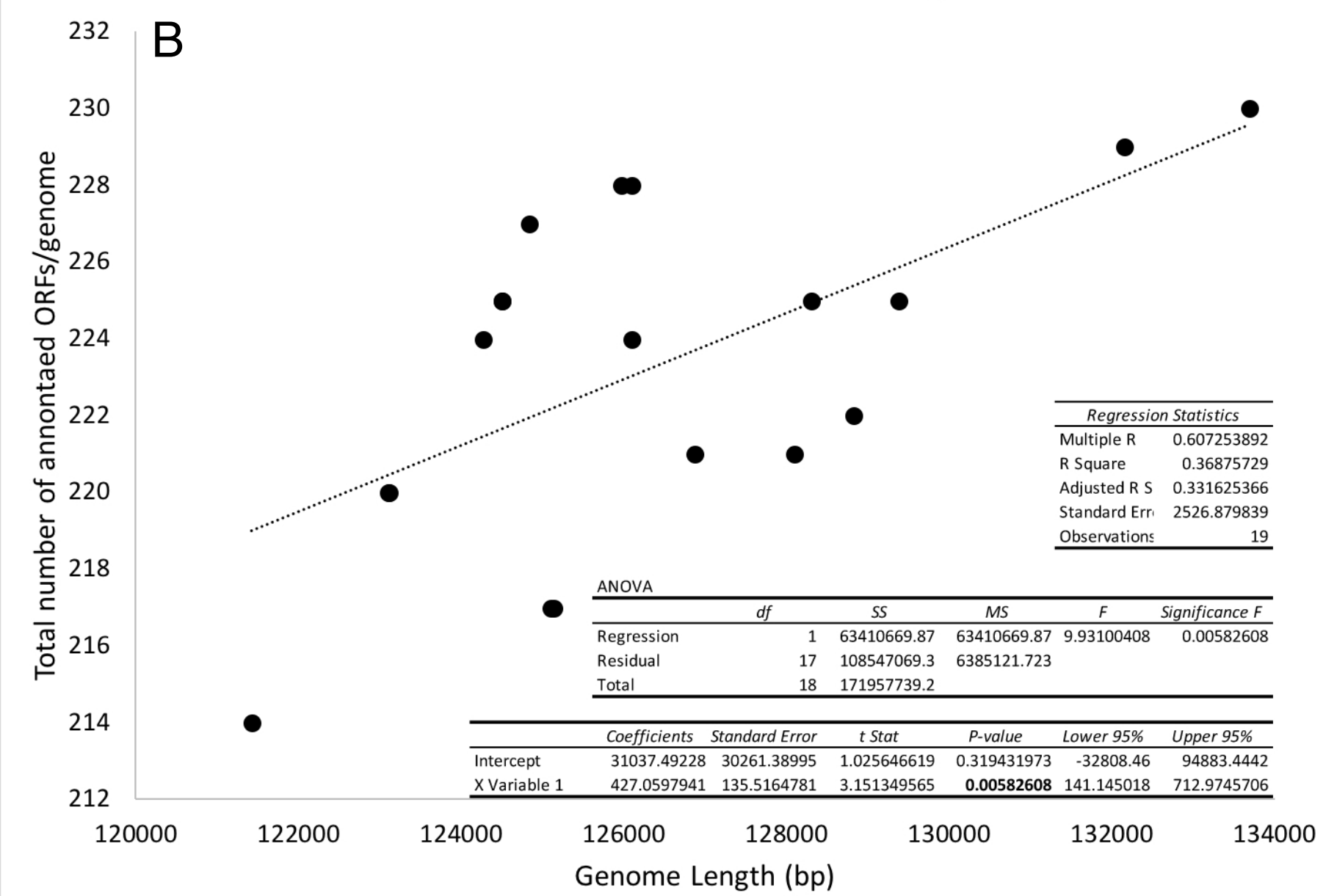


Regression Statistics	
Multiple R	0.839154677
R Square	0.704180572
Adjusted R Square	0.686779429
Standard Error	2.459695631
Observations	19

ANOVA

	df	SS	MS	F	Significance F
Regression	1	244.8324663	244.8324663	40.46748998	7.09656E-06
Residual	17	102.8517442	6.050102599		
Total	18	347.6842105			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	2066.764535	289.7951786	7.131811319	1.68088E-06	1455.350153	2678.178917
# of ORFs	-0.919331395	0.144517007	-6.36140629	**7.09656E-06**	-1.224235627	-0.614427163



Regression Statistics	
Multiple R	0.607253892
R Square	0.36875729
Adjusted R S	0.331625366
Standard Err	2526.879839
Observations	19

ANOVA

	df	SS	MS	F	Significance F
Regression	1	63410669.87	63410669.87	9.93100408	0.00582608
Residual	17	108547069.3	6385121.723		
Total	18	171957739.2			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	31037.49228	30261.38995	1.025646619	0.319431973	-32808.46	94883.4442
X Variable 1	427.0597941	135.5164781	3.151349565	**0.00582608**	141.145018	712.9745706

Figure S1: ORF linear regression statistics. ANOVA analyses of the linear regressions between (A) number of ORFs/genome vs genome isolation year and (B) ORFs/genome vs. genome length.

Table S1: Core-genome Information

	ORF Name	Nucleotide Similarity	Amino Acid Similarity	nuc_stdev	aa_stdev	Known Function	CDD Domain name	E-Value	PSSM-ID
Conserved Core	ORF1	100.00	100.00	0.00	0.00	-	-	-	-
	ORF10	100.00	100.00	0.00	0.00	-	-	-	-
	ORF109	100.00	100.00	0.00	0.00	-	-	-	-
	ORF12	100.00	100.00	0.00	0.00	-	-	-	-
	ORF13	100.00	100.00	0.00	0.00	-	-	-	-
	ORF134	100.00	100.00	0.00	0.00	-	-	-	-
	ORF135	100.00	100.00	0.00	0.00	-	-	-	-
	ORF138	100.00	100.00	0.00	0.00	-	-	-	-
	ORF142	100.00	100.00	0.00	0.00	-	-	-	-
	ORF154	100.00	100.00	0.00	0.00	-	-	-	-
	ORF158	100.00	100.00	0.00	0.00	-	-	-	-
	ORF164	100.00	100.00	0.00	0.00	-	-	-	-
	ORF167	100.00	100.00	0.00	0.00	-	-	-	-
	ORF168	100.00	100.00	0.00	0.00	-	-	-	-
	ORF171	100.00	100.00	0.00	0.00	-	HAD_like superfamily	1.87E-04	328728
	ORF175	100.00	100.00	0.00	0.00	-	-	-	-
	ORF180	100.00	100.00	0.00	0.00	-	-	-	-
	ORF195	100.00	100.00	0.00	0.00	-	-	-	-
	ORF206	100.00	100.00	0.00	0.00	-	-	-	-
	ORF214	100.00	100.00	0.00	0.00	-	-	-	-
	ORF216	100.00	100.00	0.00	0.00	-	-	-	-
	ORF218	100.00	100.00	0.00	0.00	-	-	-	-
	ORF226	100.00	100.00	0.00	0.00	-	-	-	-
	ORF227	100.00	100.00	0.00	0.00	-	-	-	-
	ORF26	100.00	100.00	0.00	0.00	-	-	-	-
	ORF29	100.00	100.00	0.00	0.00	-	-	-	-
	ORF3	100.00	100.00	0.00	0.00	-	-	-	-
	ORF32	100.00	100.00	0.00	0.00	-	-	-	-
	ORF33	100.00	100.00	0.00	0.00	-	-	-	-
	ORF34	100.00	100.00	0.00	0.00	-	-	-	-
	ORF39	100.00	100.00	0.00	0.00	-	DUF3696 superfamily	6.88E-03	331172
	ORF41	100.00	100.00	0.00	0.00	-	-	-	-
	ORF42	100.00	100.00	0.00	0.00	-	-	-	-
	ORF47	100.00	100.00	0.00	0.00	-	-	-	-
	ORF48	100.00	100.00	0.00	0.00	-	-	-	-
	ORF50	100.00	100.00	0.00	0.00	-	DUF1778 superfamily	5.16E-06	321696
	ORF52	100.00	100.00	0.00	0.00	-	-	-	-
	ORF54	100.00	100.00	0.00	0.00	-	-	-	-
	ORF55	100.00	100.00	0.00	0.00	ribonuclease H	RNase_HI_prokaryote_like	1.19E-56	260010
	ORF56	100.00	100.00	0.00	0.00	-	RNase_H_like superfamily	2.12E-14	326352
	ORF60	100.00	100.00	0.00	0.00	-	-	-	-
	ORF61	100.00	100.00	0.00	0.00	-	-	-	-
	ORF62	100.00	100.00	0.00	0.00	-	-	-	-
	ORF64	100.00	100.00	0.00	0.00	-	-	-	-
	ORF65	100.00	100.00	0.00	0.00	-	-	-	-
	ORF66	100.00	100.00	0.00	0.00	-	Peptidases_S8_S53 superfamily	5.78E-03	324584
	ORF75	100.00	100.00	0.00	0.00	putative baseplate assembly protein	Phage_base_V superfamily	9.89E-03	327437
	ORF82	100.00	100.00	0.00	0.00	-	-	-	-
	ORF9	100.00	100.00	0.00	0.00	-	-	-	-
Synonymous Core	ORF57	99.98	100.00	0.02	0.00	putative primase/helicase	RecA-like_NTPases superfamily	7.80E-28	333705
	ORF166	99.97	100.00	0.08	0.00	-	-	-	-
	ORF80	99.97	100.00	0.06	0.00	HNH homing endonuclease	HNHc superfamily	7.94E-05	320750
	ORF189	99.97	100.00	0.05	0.00	-	DnaQ_like_exo superfamily	1.39E-05	324557
	ORF137	99.96	100.00	0.07	0.00	-	-	-	-
	ORF213	99.96	100.00	0.11	0.00	-	-	-	-
	ORF81	99.96	100.00	0.05	0.00	-	Macoilin superfamily	3.48E-04	313022
	ORF43	99.96	100.00	0.13	0.00	-	-	-	-
	ORF45	99.95	100.00	0.14	0.00	-	-	-	-
	ORF74	99.95	100.00	0.10	0.00	-	-	-	-
	ORF4	99.95	100.00	0.07	0.00	-	-	-	-
	ORF112	99.94	100.00	0.06	0.00	putative DNA-binding protein Roi	Phage_pRha superfamily	2.26E-08	324635
	ORF191	99.91	100.00	0.25	0.00	-	-	-	-
	ORF46	99.91	100.00	0.17	0.00	-	-	-	-
	ORF30	99.91	100.00	0.18	0.00	-	-	-	-
	ORF44	99.90	100.00	0.20	0.00	-	-	-	-
	ORF210	99.89	100.00	0.22	0.00	-	-	-	-
	ORF153	99.88	100.00	0.34	0.00	-	-	-	-
	ORF139	99.88	100.00	0.14	0.00	-	-	-	-
	ORF83	99.88	100.00	0.12	0.00	-	-	-	-
	ORF7	99.88	100.00	0.14	0.00	-	-	-	-
	ORF194	99.87	100.00	0.16	0.00	-	-	-	-
	ORF2	99.86	100.00	0.22	0.00	-	-	-	-
	ORF181	99.85	100.00	0.22	0.00	-	-	-	-
	ORF15	99.84	100.00	0.25	0.00	-	-	-	-
	ORF126	99.82	100.00	0.36	0.00	-	-	-	-
	ORF176	99.97	99.98	0.03	0.06	putative exodeoxyribonuclease	-	-	-
	ORF173	99.87	99.96	0.09	0.11	recombination-associated protein RdgC	RdgC superfamily	5.57E-21	321354
	ORF170	99.99	99.96	0.04	0.13	-	-	-	-
	ORF179	99.98	99.95	0.05	0.14	-	-	-	-
	ORF68	99.92	99.95	0.11	0.11	-	-	-	-
	ORF49	99.96	99.94	0.07	0.16	-	-	-	-
	ORF73	99.96	99.94	0.04	0.10	putative baseplate component	Baseplate_J superfamily	1.29E-05	321435
	ORF196	99.91	99.92	0.12	0.22	putative adenine methyltransferase	Dam	2.70E-32	223415
	ORF131	99.97	99.92	0.08	0.24	-	-	-	-
	ORF76	99.95	99.92	0.07	0.13	-	-	-	-
	ORF71	99.97	99.91	0.08	0.26	-	-	-	-
	ORF211	99.97	99.91	0.05	0.15	ClpP ATP-dependent protease subunit	crotonase-like superfamily	6.26E-19	329030
	ORF120	99.91	99.89	0.10	0.22	-	-	-	-
	ORF63	99.96	99.89	0.11	0.33	-	-	-	-
	ORF132	99.96	99.88	0.11	0.34	-	-	-	-
	ORF177	99.92	99.88	0.13	0.23	-	-	-	-
	ORF79	99.63	99.88	0.55	0.16	-	DUF2130 superfamily	5.56E-03	331406
	ORF140	99.96	99.87	0.12	0.37	-	-	-	-
	ORF207	99.96	99.87	0.07	0.20	putative Gp5 baseplate hub subunit and tail lysozyme	NLPC_P60 superfamily	3.15E-35	328779
	ORF59	99.97	99.87	0.08	0.38	-	-	-	-
	ORF72	99.91	99.87	0.08	0.18	-	-	-	-

Divergent Core

ORF188	99.96	99.86	0.13	0.40	-	-	-	-
ORF124	99.63	99.86	0.46	0.16	-	-	-	-
ORF204	99.93	99.85	0.06	0.14	ribonucleoside diphosphate reductase, beta chain	Ferritin_like superfamily	1.00E-52	320867
ORF58	99.89	99.85	0.09	0.12	DNA polymerase	DNA_pol_A superfamily	5.25E-33	322025
ORF169	99.95	99.85	0.10	0.30	-	-	-	-
ORF202	99.95	99.84	0.15	0.46	-	-	-	-
ORF144	99.95	99.84	0.16	0.48	-	-	-	-
ORF36	99.85	99.83	0.14	0.23	-	-	-	-
ORF11	99.95	99.83	0.16	0.49	-	-	-	-
ORF215	99.90	99.83	0.08	0.14	DNA ligase	CDC9 superfamily	1.57E-99	330238
ORF28	99.88	99.82	0.23	0.37	-	HNHc	1.08E-05	238038
ORF122	99.86	99.82	0.12	0.22	putative major head protein	Phage_cap_E superfamily	6.03E-24	309113
ORF155	99.87	99.81	0.37	0.56	-	-	-	-
ORF78	99.88	99.81	0.13	0.26	-	-	-	-
ORF151	99.93	99.80	0.19	0.59	-	-	-	-
ORF208	99.88	99.80	0.12	0.22	PhoH family protein	P-loop_NTPase superfamily	4.93E-31	328724
ORF53	99.89	99.78	0.16	0.31	-	-	-	-
ORF172	99.91	99.77	0.14	0.33	-	-	-	-
ORF25	99.92	99.76	0.23	0.70	-	-	-	-
ORF192	99.66	99.74	0.45	0.49	-	-	-	-
ORF145	99.70	99.74	0.46	0.53	-	-	-	-
ORF27	99.91	99.73	0.18	0.55	-	-	-	-
ORF8	99.91	99.72	0.15	0.44	-	-	-	-
ORF6	99.86	99.72	0.31	0.83	-	-	-	-
ORF130	99.83	99.71	0.12	0.26	-	-	-	-
ORF127	98.84	99.71	2.24	0.47	-	-	-	-
ORF193	99.83	99.70	0.13	0.25	putative thymidylate synthase	Thy1 superfamily	1.56E-04	332234
ORF152	99.90	99.69	0.10	0.30	-	-	-	-
ORF220	99.89	99.67	0.17	0.52	-	-	-	-
ORF121	99.90	99.67	0.07	0.21	-	T5orf172	8.07E-05	313714
ORF143	99.85	99.67	0.26	0.53	-	-	-	-
ORF133	99.77	99.66	0.18	0.40	-	-	-	-
ORF128	99.88	99.65	0.11	0.34	terminase large subunit	Terminase_6 superfamily	7.55E-28	321850
ORF40	99.85	99.64	0.14	0.34	-	-	-	-
ORF217	99.88	99.64	0.19	0.58	-	-	-	-
ORF184	99.88	99.63	0.16	0.50	-	-	-	-
ORF187	99.84	99.63	0.23	0.50	-	-	-	-
ORF150	99.88	99.63	0.20	0.60	-	-	-	-
ORF70	99.82	99.59	0.09	0.30	-	-	-	-
ORF31	99.86	99.59	0.17	0.51	-	-	-	-
ORF185	99.75	99.58	0.21	0.40	-	-	-	-
ORF51	99.85	99.55	0.17	0.52	-	-	-	-
ORF212	99.90	99.55	0.19	0.72	-	-	-	-
ORF77	99.70	99.55	0.63	0.64	-	-	-	-
ORF156	99.85	99.55	0.24	0.72	-	-	-	-
ORF136	99.85	99.53	0.25	0.75	-	-	-	-
ORF200	99.84	99.52	0.19	0.58	-	-	-	-
ORF186	99.84	99.51	0.22	0.66	-	-	-	-
ORF209	99.84	99.50	0.14	0.43	-	-	-	-
ORF123	99.81	99.48	0.24	0.57	-	-	-	-
ORF108	99.83	99.48	0.27	0.83	-	-	-	-
ORF107	99.04	99.48	1.41	0.64	-	-	-	-
ORF199	99.82	99.44	0.22	0.67	-	-	-	-
ORF203	99.63	99.42	0.83	1.27	-	Ribonuc_red_lgC superfamily	3.16E-87	332162
ORF125	99.13	99.42	0.94	0.46	-	Peptidase_S78_2 superfamily	4.65E-10	317012
ORF205	99.81	99.42	0.19	0.56	-	-	-	-
ORF219	99.72	99.41	0.26	0.70	-	-	-	-
ORF37	99.74	99.19	0.26	0.78	-	-	-	-
ORF165	99.30	99.18	1.32	1.43	-	-	-	-
ORF129	99.62	99.15	0.51	1.36	-	-	-	-
ORF141	99.72	99.14	0.27	0.82	-	-	-	-
ORF198	99.62	99.10	0.29	0.63	-	-	-	-
ORF201	99.70	99.09	0.30	0.93	-	-	-	-
ORF67	99.70	99.08	0.32	0.97	-	-	-	-
ORF110	99.63	99.00	0.63	1.60	-	-	-	-
ORF69	99.67	98.99	0.22	0.68	-	-	-	-
ORF94	99.20	98.90	1.98	2.19	-	-	-	-
ORF103	96.62	98.87	3.49	0.87	-	-	-	-
ORF93	97.56	98.77	3.12	1.64	-	-	-	-
ORF84	97.62	98.67	3.12	1.99	-	DUF3383 superfamily	4.95E-09	314693
ORF189.1	99.52	98.54	0.58	1.76	-	-	-	-
ORF197	99.46	98.34	0.42	1.30	-	-	-	-
ORF157	99.15	98.03	0.94	1.79	-	-	-	-
ORF113	97.85	97.96	2.06	2.60	-	NRDD superfamily	5.49E-149	330954
ORF174	97.92	97.76	8.27	8.40	-	T5orf172	1.47E-04	313714
ORF91	97.11	97.59	2.66	2.23	-	-	-	-
ORF92	96.81	97.49	3.76	2.96	-	-	-	-
ORF161	94.88	97.41	6.15	3.11	-	-	-	-
ORF35	99.45	97.16	1.10	5.70	-	-	-	-
ORF87	94.86	97.13	4.65	2.77	-	-	-	-
ORF119	97.97	97.12	2.17	2.95	-	-	-	-
ORF86	94.49	97.05	5.22	2.98	-	-	-	-
ORF85	94.13	96.50	7.18	4.47	-	BF2867_like_C	2.18E-03	240526
ORF159	96.66	96.04	3.20	3.89	-	-	-	-
ORF104	96.45	96.00	5.71	6.21	-	-	-	-
ORF221	95.99	95.40	5.24	5.88	-	-	-	-
ORF97	95.94	95.27	6.13	6.91	-	PP-binding superfamily	1.14E-07	324546
ORF5	99.32	92.28	0.85	10.28	-	GIY-YIG_SF superfamily	5.49E-03	326551
ORF114	91.57	91.83	6.46	6.32	-	Radical_SAM superfamily	1.93E-74	327492

Table S1: Core-genome information. The data for each ORF represented in Figure 4 is listed in columns 2 and 3. Columns 4 and 5 contain the standard deviations for those pairwise similarity values. The other columns contain information about putative gene function and conserved domain homology.

# Occurances	ORF Name	ICP1	ICP1_2001_A	ICP1_2001_B	ICP1_2001_C	ICP1_2001_D	ICP1_2001_E	ICP1_2001_F	ICP1_2004_A	ICP1_2005_A	ICP1_2006_A	ICP1_2006_B	ICP1_2006_C	ICP1_2006_D	ICP1_2006_E	ICP1_2009_A	ICP1_2011_A	ICP1_2011_B	ICP1_2011_C	ICP1_2012_A	Putative Function	Domain Name	E-Value	PSSM_ID
XVIII	ORF19	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		-	-	-	-
	ORF20	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		-	-	-	-
	ORF38	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		-	-	-	-
	ORF95	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	-	-	-	-
	ORF96	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	-	-	-	-
	ORF105	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	-	-	-	-
	ORF106	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	-	-	-	-
	ORF111	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	-	-	-	-
	ORF178	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	-	-	-	-
	ORF180.1	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		-	-	-	-
	ORF182	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	-	-	-	-
	ORF183	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	-	-	-	-
XVII	ORF146	■	■	■	■	■	■	■	■	■	■	■			■	■	■	■	■	■	-	-	-	-
	ORF148	■	■	■	■	■	■	■	■	■	■	■			■	■	■	■	■	■	-	-	-	-
	ORF147	■	■	■	■	■	■	■	■	■	■	■			■	■	■	■	■	■	helicase	P-loop_NTPase super family	5.71e-12	328724
	ORF149	■	■	■	■	■	■	■	■	■	■	■			■	■	■	■	■	■	-	-	-	-
XVI	ORF21	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■				-	-	-	-
XV	ORF14	■	■	■	■	■	■	■	■	■	■	■	■	■	■					■	-	-	-	-
	ORF98	■	■	■	■	■	■	■	■	■	■	■	■	■	■					■	-	-	-	-
	ORF99	■	■	■	■	■	■	■	■	■	■	■	■	■	■					■	-	-	-	-
	ORF100	■	■	■	■	■	■	■	■	■	■	■	■	■	■					■	-	-	-	-
	ORF101	■	■	■	■	■	■	■	■	■	■	■	■	■	■					■	-	-	-	-
	ORF222	■	■	■	■	■	■	■	■	■	■	■	■	■	■					■	-	-	-	-
	ORF223	■	■	■	■	■	■	■	■	■	■	■	■	■	■					■	-	-	-	-
	ORF225	■	■	■	■	■	■	■	■	■	■	■	■	■	■					■	-	-	-	-
XIV	ORF16	■	■	■	■	■	■	■							■	■	■	■	■	■	-	-	-	-
	ORF114.1		■	■		■	■		■	■	■	■			■	■	■	■	■	■	-	-	-	-
	ORF228	■	■	■	■	■		■		■	■	■				■	■	■	■	■	-	Pribosyltran_N superfamily	1.78166E-22	330902
	ORF229	■	■	■	■	■		■		■	■	■				■	■	■	■	■	-	PRK09198	0	236407
XIII	ORF89	■	■	■	■	■	■	■		■	■	■	■	■						■	antirepressor protein	Bro-N superfamily	3.42643E-23	324611
	ORF90	■	■	■	■	■	■	■		■	■	■	■	■						■	-	-	-	-
	ORF102	■	■	■	■	■	■	■		■	■	■	■	■	■						-	-	-	-
	ORF114.2		■	■		■	■			■	■	■			■	■	■	■	■	■	-	-	-	-
XII	ORF88	■	■	■	■	■	■	■			■	■	■	■						■	-	T5orf172	2.97023E-06	313714
XI	ORF22	■	■		■					■	■	■	■	■		■	■			■	-	-	-	-
X	ORF24	■	■		■					■	■	■	■	■							-	-	-	-
IX	ORF15.1									■	■	■	■	■		■	■	■	■		-	-	-	-
	ORF17	■	■	■	■	■	■	■	■						■						-	-	-	-
	ORF18	■	■	■	■	■	■	■	■						■						-	-	-	-
	Cas1						■	■	■	■					■	■	■	■	■		Cas1	Cas1_I-II-III superfamily	8.59874E-89	321096
	Cas3						■	■	■	■					■	■	■	■	■		Cas3	P-loop_NTPase superfamily	3.8104E-118	328724
	Csy1						■	■	■	■					■	■	■	■	■		Csy1	Csy1_I-F superfamily	4.37493E-05	324501
	Csy2						■	■	■	■					■	■	■	■	■		Csy2	Csy2_I-F superfamily	1.24047E-14	325199
	Csy3						■	■	■	■					■	■	■	■	■		Csy3	Csy3_I-F superfamily	1.6744E-104	324502
	Csy4						■	■	■	■					■	■	■	■	■		Csy4	Cas6_I-F superfamily	2.50407E-28	324504
VII	ORF23	■	■							■	■	■	■	■							-	-	-	-
	ORF87.1						■		■						■	■	■	■	■		-	-	-	-
	putative_Rha_protein						■		■						■	■	■	■	■		Rha_protein	Phage_pRha	5.83871E-25	312982
V	ORF115	■			■			■					■	■							-	-	-	-
	ORF116	■			■			■					■	■							HNH endonuclease	-	-	-
	ORF117	■			■			■					■	■							-	-	-	-
	ORF117.1	■			■			■					■	■							-	-	-	-
	ORF118	■			■			■					■	■							-	-	-	-
	ORF160	■	■	■	■	■															-	-	-	-
	ORF162	■	■	■	■	■															-	-	-	-
	ORF163	■	■	■	■	■															-	-	-	-
	ORF113x		■	■		■	■								■						-	NRDD superfamily	8.63697E-73	330954
	putative_DNA-methyltransferase						■		■				■	■	■						DNA-methyltransferase	AdoMet_MTases superfamily	2.75054E-52	327401
IV	ORF97.1															■	■	■	■		-	-	-	-
	ORF181.1						■		■				■	■							-	-	-	-
	ORF5x															■	■	■	■		-	-	-	-
	HigA															■	■	■	■		-	antidote_HigA	4.07815E-27	274228
II	ORF15.2												■	■							-	-	-	-
	ORF15.3												■	■							-	PTH2_family superfamily	0.000800162	327521
	ORF145.1												■	■							-	-	-	-
	ORF147.1												■	■							-	-	-	-
	ORF165x		■																	■	-	-	-	-
	ORF174x												■	■							-	T5orf172	0.000295242	313714
	ORF192x			■		■															-	-	-	-
	ORF87p						■	■													-	-	-	-
I	ORF22.1																			■	-	-	-	-
	ORF94.1							■													-	-	-	-
	ORF94.2							■													-	-	-	-
	ORF96.1							■													-	PP-binding superfamily	1.18907E-09	324546
	ORF104.1															■					-	-	-	-
	ORF104.2															■					-	-	-	-
	ORF104.3															■					-	Lipase_GDSL_2	6.70426E-18	316033
	ORF104.4															■					-	-	-	-
	ORF179.1							■													-	-	-	-
	ORF228.1							■													-	-	-	-
	ORF203x																■				-	Ribonuc_red_lgC superfamily	5.56357E-84	332162
	ORF228x				■																-	PRTases_typeI superfamily	1.87368E-08	320892
	ORF86p							■													-	-	-	-
	putative_DNA-methyltransferase_x														■						DNA-methyltransferase_x	-	-	-

Table S2: Accessory-genome ORF occurrence matrix. Each ORF in the accessory genome is listed along with any putative function or conserved domain homology information. A black square indicates that the ORF occurs in a specific genome.