Document S1

Classification of O-antigen biosynthetic gene

The proteins on O-antigen biosynthetic gene cluster were categorized using the Pfam databases (https://pfam.xfam.org) to identify glycosyltransferases and aligned against O-antigen processing-related genes of *wzx*, *wzy*, *wzm* and *wzt* from *E. coli* and *Shigella* sp. by BLAST+ (v2.6.0) to classify the O-antigen processing system (*wzx*/*wzy* or *wzm*/*wzt*) [1–3]. Hypothetical proteins were manually curated using the NCBI BLASTp search.

Additionally, the pairwise alignments of 206 O-antigen gene clusters were performed by Chromosomal Comparison of BioNumerics software with default parameters (Applied Maths). Reference strain pairs showing identity scores >90% were examined to ascertain whether these O-serogroups were distinguishable according to the cross-reactivity against O antisera. When reference strains seemed to possess an O-antigen identical to another, we eliminated the strain with a higher reference number from the O-serogroup list.

Pan-genome analysis

Pan-genome analysis of 190 *V. cholerae* genomes was performed using the Roary (v3.11.2) pipeline with default parameters [4]. The extracted gene clusters were then classified into "core" and "non-core" (accessory) genomes based on whether the clusters consisted of genes from all strains; only clusters with genes from all genomes were included in the "core" set; otherwise, they were classified into the "accessory" set. The curve fitting of the pan-genome growth was performed using a power-law regression, which was based on Heaps' Law described in previous reports [5, 6]. The fitting was conducted using panGP (v1.0) to fit the equation (y = Apan x Bpan + Cpan), in which y and x were the pan-genome size and the number of genomes, respectively [7]. Bpan was equivalent to the γ parameter for estimating whether a pan-genome is open or closed. Alignment sequences of identified core genes were used for reconstruction of the phylogenetic tree. The profile of pan-genome gene clusters was used for downstream analysis. Pan-genome profile and the relevant statistics were shown in circular phylogram or bar plots visualizing by using anvi'o (v5) [8]. Orthologous gene clusters in the circular phylogram were organized by Euclidean distance and the Ward linkage algorithm in the anvio5 platform.

Phylogenetic analysis

In the large datasets of sequence alignments with *V. cholerae* core genes, we used FastTree (v2.1.1) to estimate an approximate maximum-likelihood phylogenetic tree using the General Time Reversible (GTR) + G + I nucleotide substitution model [9]. The reliability of the phylogenetic tree was determined with local support values based on the Shimodaira-Hasegawa test. The *V. cholerae* population structure was analyzed via the Bayesian-based clustering algorithm hierBAPS to assign lineages, implemented in R [10]. Average nucleotide identity values were calculated on the whole-genome level using FastANI (v1.3) with default parameters and the pairwise comparison was visualized in a heatmap matrix which was implemented in R using heatmap.2 from the gplots package [11]. The other ML trees were reconstructed using RaXML-NG v0.90 with the GTR + G + I substitution model. Support for the branches was estimated from 100 bootstrapping replicates [12]. To investigate the genetic differentiation between phylogenetic clusters, the Weir and Cockerham's fixation index (Fst) value for SNPs on core genome sequences between each pair of clusters was calculated by VCFtools (v0.1.17) with haploid mode and 1 kb sliding window (-- haploid -fst-window-size 1000) (https://github.com/jydu/vcftools). The Fst value between phylogenetic clusters was visualized using Manhattan plots and box plots using R with the ggplot2 package.

COG and KO analyses

COGs were predicted by searching the amino acid sequences of the genes against COG hidden Markov models downloaded from the eggNOG 4.5 database using HMMER3 with an e-value cut-off of 1e-5 [13, 14]. For further core and non-core genome annotations, protein functions were assigned using the BlastKOALA tool via the KEGG database [15]. KEGG Orthology (KO) terms were divided into 18 supercategories including "Genetic Information Processing", "Metabolism", "Cellular Processes", and "Environmental Information Processing", based on the KEGG hierarchy classification.

Ortholog analysis of O-antigen or SI structural genes

For ortholog analysis within O-antigen biosynthesis gene clusters or SI structural genes, predicted proteins were clustered into orthologous groups using OrthoFinder (v2.3.7) with a cut-off of 1e-50, as described in previous reports [16, 17].

Characterization of genomic islands

Identified GIs were characterized using BLAST+ (v2.6.0) with sequences of known GIs as query sequences, which showed >80% query coverage and >80% query identity [3]. We used entire sequences of VPI-1, VPI-2, VSP-I and VSP-II in N16961 for characterization. A GI possessing *intI4* with *V. cholerae* repeats and *int*^{SXT} was categorized as an SI and an integrative conjugative element (ICE), respectively [18–20]. Type 3 secretion systems (T3SSs) were detected by nine genes (*vcsS2J2U2V2Q2R2T2C2N2*) encoding the structural components of T3SS and genes encoding the DUF4123 domain protein and Zn-binding Pro-Ala-Ala-Arg domain protein, *hcp* and *vgrG* were used as queries to detect the Type 6 secretion system [21–23]. Further characteristics were determined as follows. The CRISPR-Cas module and prophage region were identified using CRISPRFinder (https://crispr.i2bc.paris-saclay.fr/) and PHASTER (https://phaster.ca/), respectively [24, 25]. The results of the O-antigen biosynthetic gene cluster assembly were also incorporated in the GI features.

Uncharacterized GIs were manually characterized using the NCBI BLASTp search.

Supplementary References

1. El-Gebali S, Mistry J, Bateman A, et al. The Pfam protein families database in 2019. *Nucleic Acids Res*. 2019; 47: D427-32.

2. Iguchi A, Iyoda S, Kikuchi T, et al. A complete view of the genetic diversity of the *Escherichia coli* O-antigen biosynthesis gene cluster. *DNA Res*. 2015; 22: 101-7.

3. Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009; 10: 421.

4. Page AJ, Cummins CA, Hunt M, et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*. 2015; 31: 3691-93.

5. Tettelin H, Masignani V, Cieslewicz MJ, et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A*. 2005; 102: 13950-55.

6. Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. *Curr Opin Microbiol*. 2008; 11: 472-77.

7. Zhao Y, Jia X, Yang J, et al. PanGP: a tool for quickly analyzing bacterial pan-genome profile. *Bioinformatics*. 2014; 30: 1297-99.

8. Eren AM, Esen ÖC, Quince C, et al. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*. 2015; 3: e1319.

9. Price MN, Dehal PS, Arkin AP. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One*. 2010; 5: e9490.

10. Tonkin-Hill G, Lees JA, Bentley SD, et al. RhierBAPS: An R implementation of the population clustering algorithm hierBAPS. *Wellcome Open Res*. 2018; 3: 93.

11. Jain C, Rodriguez-R LM, Phillippy AM, et al. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun*. 2018; 9: 5114.

12. Kozlov AM, Darriba D, Flouri T, et al. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*. 2019; 35: 4453-55.

13. Her HL, Wu YW. A pan-genome-based machine learning approach for predicting antimicrobial resistance activities of the *Escherichia coli* strains. *Bioinformatics*. 2018; 34: i89-i95.

14. Huerta-Cepas J, Szklarczyk D, Forslund K, et al. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res*. 2016; 44: D286-93.

15. Kanehisa M, Sato Y, Morishima K. BlastKOALA and GhostKOALA: KEGG Tools for functional characterization of genome and metagenome sequences. *J Mol Biol*. 2016; 428: 726-31.

16. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics.

Genome Biol. 2019; 20: 238.

17. Bentley SD, Aanensen DM, Mavroidi A, et al. Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. *PLoS Genet*. 2006; 2: e31.

18. Karaolis DK, Johnson JA, Bailey CC, et al. A *Vibrio cholerae* pathogenicity island associated with epidemic and pandemic strains. *Proc Natl Acad Sci U S A*. 1998; 95: 3134-9.

19. Dziejman M, Balon E, Boyd D, et al. Comparative genomic analysis of *Vibrio cholerae*: genes that correlate with cholera endemic and pandemic disease. *Proc Natl Acad Sci U S A*. 2002; 99: 1556-61.

20. Jermyn WS, Boyd EF. Characterization of a novel *Vibrio* pathogenicity island (VPI-2) encoding neuraminidase (*nanH*) among toxigenic *Vibrio cholerae* isolates. *Microbiology (Reading)*. 2002; 148: 3681-3693.

21. Alam A, Miller KA, Chaand M, et al. Identification of *Vibrio cholerae* type III secretion system effector proteins. *Infect Immun*. 2011; 79: 1728-40.

22. Dziejman M, Serruto D, Tam VC, et al. Genomic characterization of non-O1, non-O139 *Vibrio cholerae* reveals genes for a type III secretion system. *Proc Natl Acad Sci U S A*. 2005; 102: 3465-70.

23. Labbate M, Orata FD, Petty NK, et al. A genomic island in *Vibrio cholerae* with VPI-1 site-specific recombination characteristics contains CRISPR-Cas and type VI secretion modules. *Sci Rep*. 2016; 6: 36891.

24. Grissa I, Vergnaud G, Pourcel C. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res*. 2007; 35: W52-57.

25. Arndt D, Grant JR, Marcu A, et al. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res*. 2016; 44: W16-21.