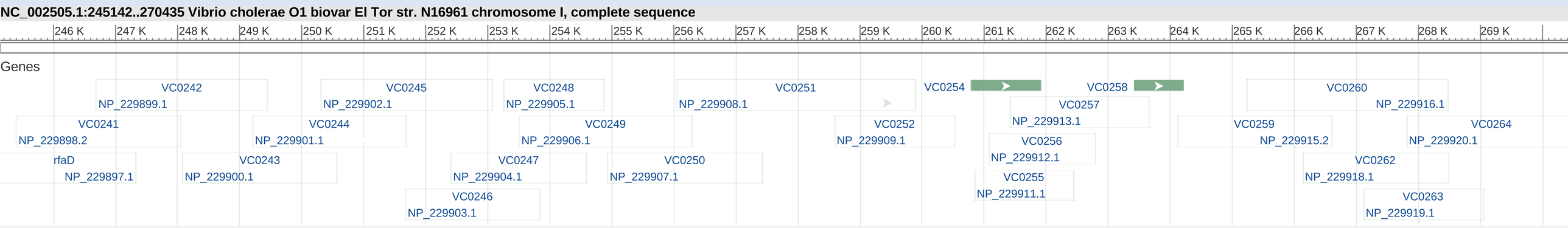
Phylogenetic Analysis of the O-antigen Biosynthesis Genes in *Vibrio cholerae*

Daniel Antonio Negrón[1], Bruce Goodwin[2], Michael Smith, Ph.D.[2], Shanmuga Sozhamannan, Ph.D.[2,3]

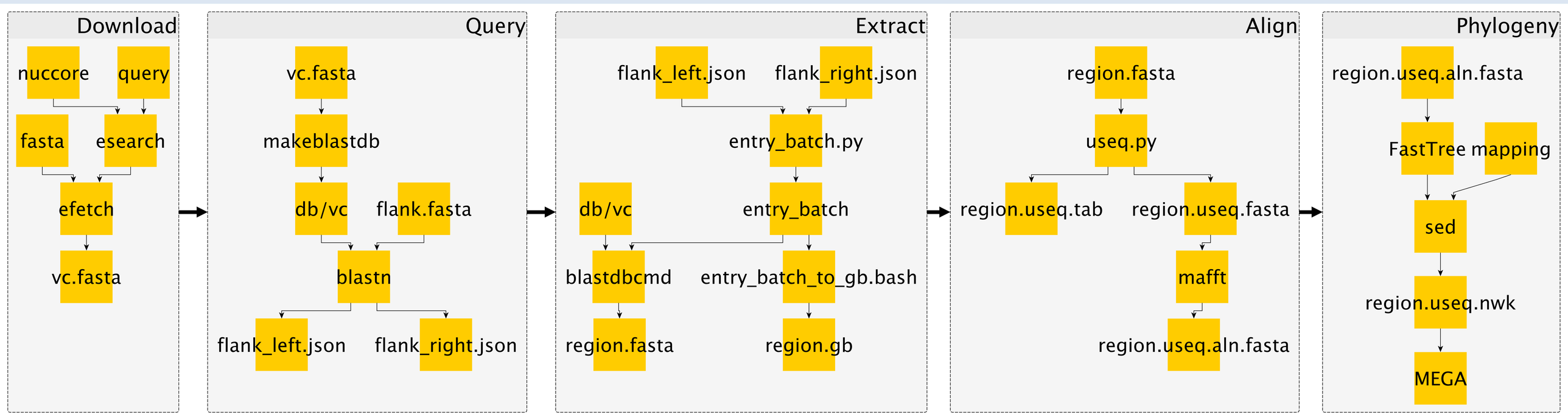
[1]Noblis, Inc., [2]Defense Biological Product Assurance Office, [3]The Tauri Group, LLC

Introduction



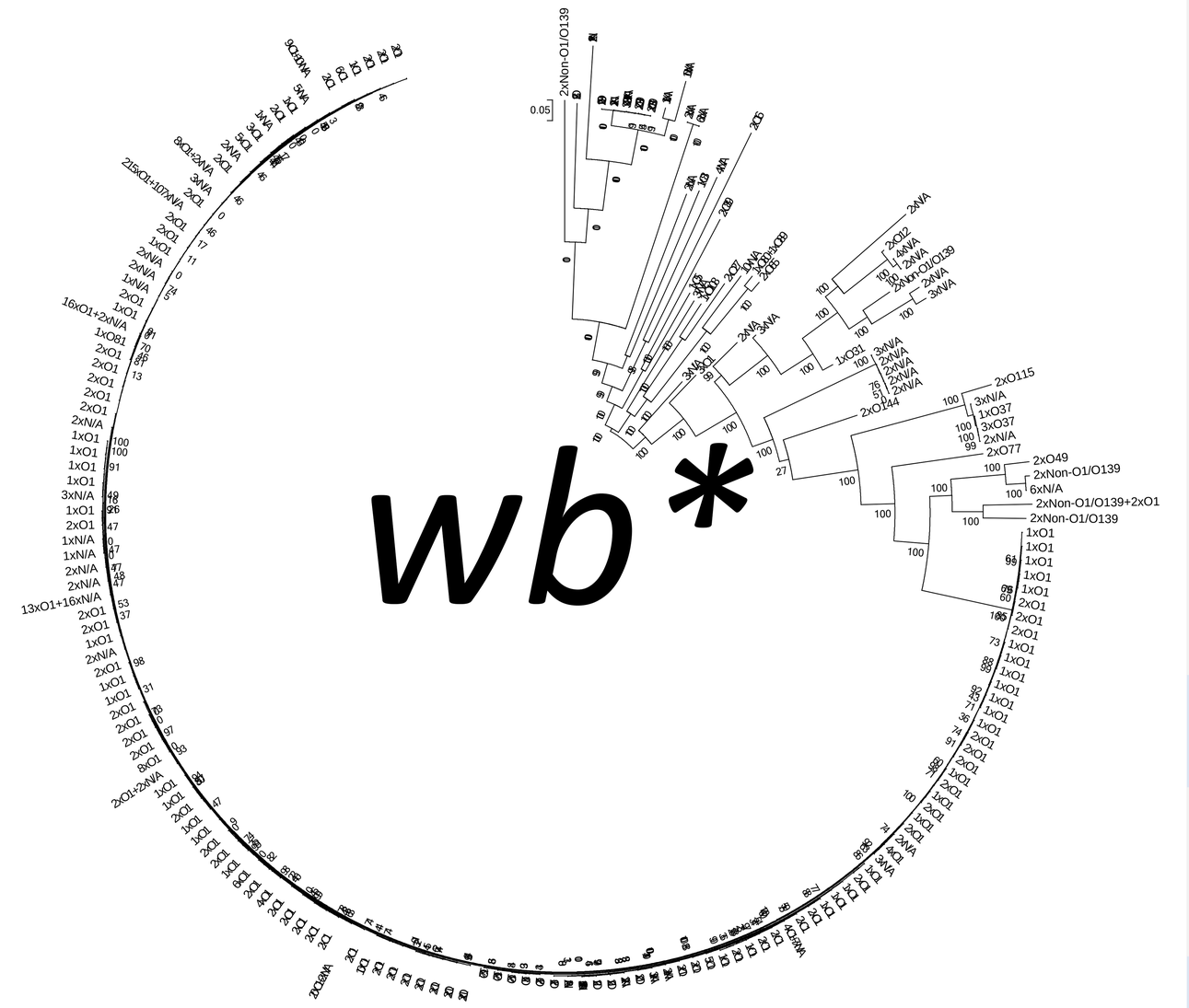
The lipopolysaccharide (LPS) of *Vibrio cholerae* is a virulence factor involved in host-pathogen interactions. In particular, the O-antigen constituent of the LPS exhibits diverse genetic organization and is useful for classifying *Vibrio* strains and serogroups. Consequently, this provides valuable information for research into the ongoing pandemic. Our previous study developed a simple and effective bioinformatics pipeline to analyze the *wb** gene cluster involved in O-antigen biosynthesis. The pipeline successfully extracted these regions from publicly available, whole genome sequencing data and generated a bootstrapped, approximately-maximum-likelihood phylogenetic tree. This follow-up study compares the *wb** region against the genomic backbone using phylogenic methods. Results from this study facilitate the identification and analysis of horizontal gene transfer (HGT) events, particularly those involving epidemic and non-epidemic strains.

Methods



The flowcharts illustrate the generalized pipeline for the extraction of flanked regions. The download phase issues a query to the NCBI and downloads the results as FASTA files. The query phase makes a BLAST database to search for hits to the two flank sequences. The extraction phase relies on two in-house scripts that produce an entry_batch file and download a corresponding GenBank file representative of the extracted regions. The entry_batch lists the header, range, and strand of each region. The GenBank is useful for extracting genes and metadata, particularly serogroup information. The alignment phase removes duplicate sequences, producing a non-redundant FASTA file and tab-delimited file matching the representative sequence to duplicates. Next, the MAFFT tool generates a multiple alignment file. The FastTree program performs approximate, maximum-likelihood estimation with bootstrapping. The sed command re-labels the taxon identifiers and MEGA renders the tree.

Results



Approximately-maximum-likelihood phylogenetic tree of the *wb** region. Each label groups the regions of serogroups with 100% identity: label = N_kS_k, N = number of taxa, S = serogroup, k = serogroup index.

Conclusion

The pipeline successfully extracted the *wb** regions. The corresponding tree consists of a large clade of O1 taxa and a small clade of O139 taxa. The rest of the tree consists of other serogroup taxa. Two O1 serogroup taxa occur within this portion.

Multiple Non-O1/O139 taxa were observed in the tree. Horizontal gene transfer may be responsible for the occurrence of conserved flanking genes, *gmhD* and *rjg*, across the serogroups. Additional analysis is required to investigate the genes that constitute the region.

Metadata is critical for conducting additional analysis. A semi-automated process extracted serogroup data. Considerable manual effort was required to obtain as much information as possible. This is due to user error and lack of standards when submitting sequence feature qualifiers to GenBank. Thus, further literature review is required to obtain missing serogroup information.

Further investigation aims to compare the evolution of the genomic backbone with the *wb** region. This includes phylogenetic analysis of the *wb** regions for each serogroup separately. The generalized conserved region pipeline may be applied to the study of other gene clusters.

References

- Cock, P. J. A. *et al*. Biopython: freely available Python tools for computational molecular biology and. *Bioinformatics* **25,** 1422–1423 (2009).
- Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10,** 421 (2009).
- Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE* **5,** e9490 (2010).
- Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Molecular Biology and Evolution* **26,** 1641–1650 (2009).
- Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* **30,** 772–780 (2013).
- Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution* msw054 (2016). doi:10.1093/molbev/msw054
- Van Rossum, G. *Python tutorial, Technical Report CS-R9526*. (Centrum voor Wiskunde en Informatica (CWI), 1995).
- R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2016).

Acknowledgements

Special thanks to the Big Data Laboratory at Noblis, Inc., especially Mitch Holland for reviewing this poster. Thank you, SFAF 2016!