Grammar and Language of CRISPR/Cas-Targetable Sites in *Escherichia coli*, *Shigella*, *Pseudomonas*, and *Salmonella*: A Comprehensive Survey

Carla M. Mann[1*] (cmann@anl.gov), Rebecca Weinberg,[2] Gyorgy Babnigg,[2] Peter E. Larsen,[2] Marie-Francoise Gros,[2,3] Philippe Noirot,[2,3] **Dionysios A. Antonopoulos**,[2] and Arvind Ramanathan[1]

[1]Data Science and Learning Division, Argonne National Laboratory, Lemont, IL; [2]Biosciences Division, Argonne National Laboratory, Lemont, IL; [3]National Research Institute for Agriculture, Food and Environment (INRAE), France

Project Goals: The long-term goal for this Project is to realize secure biodesign strategies for microbial systems that operate in the dynamic abiotic and biotic conditions of natural environments, thus enabling systems-level and rational biological design for field use. There are several key challenges to incorporating safeguard systems at the design stage including: (1) lack of knowledge for how well safeguards operate across the broad set of environmental and physiological conditions that an organism experiences; (2) a need to integrate the safeguard with other cellular components so that it can sense and recognize specific signals from the intracellular or extracellular environment, and mediate a response; and (3) a need for rapid and reliable methods to engineer and optimize the biological components for safeguard construction and functional integration. To address these challenges, we propose to utilize recent advances in the fields of synthetic biology, artificial intelligence (AI), and automation, which are creating the conditions for a paradigm shift in our understanding of the ways that cellular function can be designed at the level of bacterial communities.

The advent of CRISPR/Cas revolutionized precision genome editing and engineering. However, the technology has introduced its own set of challenges, particularly in choosing a target site within a genome that matches potentially stringent experimental parameters, such as efficiency of double-strand break (DSB) induction, DSB location, and number of off-target sites. Even when a suitable target site has been identified in a model organism, there is no guarantee that that particular target will be present in closely related species. These requirements demonstrate the importance of understanding the genomic landscape of CRISPR/Cas-targetable sites to better understand the rules governing the behavior of gRNAs. To that end, we surveyed the genomes and CRISPR/Cas-targetable sites within 16,171 *Escherichia coli*, 446 *Shigella*, 7,867 *Pseudomonas*, and 14,604 *Salmonella* genomes from the Pathosystems Resource Integration Center (PATRIC) [1,2] database.

We analyzed the quality of the genomes present in the PATRIC database for each species using the scoring system devised by Land et al. [3] and demonstrate a species quality bias – nearly 97% of *E. coli* genomes meet our quality threshold, while only 15% of *Shigella*, 75% of *Pseudomonas*, and 65% of *Salmonella* genomes meet that threshold.

Using the high-quality genome assemblies, we calculated genomic GC content, identified rare codons within each species, and identified the species' core genes. We further identified all SpCas9, St1Cas9, SaCas9, NmCas9, CjCas9, and AsCas12a-targetable sites in each genome. Using MASH [4], an implementation of the MinHash algorithm for estimating genomic distance, we found that these Cas-targetable sites are as well conserved among each species as the overall

genomes themselves, although those sites located within highly conserved core genes and operons are themselves highly conserved.

References

1. Davis J.J. et al. *The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities.* Nucleic Acids Res. 2020 Jan 8;48(D1):D606-D612. DOI: 10.1093/nar/gkz943.
2. Wattam A.R. et al. *Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center*. *Nucleic Acids Res*. 2017 Jan 4;45(D1):D535-D542. DOI: 10.1093/nar/gkw1017
3. Land, M.L. et al. *Quality scores for 32,000 genomes*. Stand Genomic Sci. 2014; 9:20 DOI: 10.1186/1944-3277-9-20
4. Ondov B.D., et al. *Mash: fast genome and metagenome distance estimation using MinHash.* Genome Biol 2016 17:132 DOI: 10.1186/s13059-016-0997-x

This Project is funded by the Biological Systems Science Division's Genomic Science Program, within the U.S Department of Energy, Office of Science, Biological and Environmental Research.