Spacer2PAM: an R Package for CRISPR Protospacer Adjacent Motif Prediction from Spacer Sequences

Grant A. Rybnicky[1]* (grantrybnicky2023@u.northwestern.edu), Michael Köpke[2], and **Michael C. Jewett**[3]

[1]Interdisciplinary Biological Sciences Graduate Program, Northwestern University, Evanston, IL; [2]LanzaTech, Skokie, IL; [3]Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL

http://jewettlab.northwestern.edu/; https://www.lanzatech.com/

Project Goals: We are addressing the challenge of designing, building, and optimizing biosynthetic pathways in cells in an interdisciplinary venture that establishes the clostridia Foundry for Biosystems Design (cBioFAB). Working both *in vitro* and *in vivo*, the goal is to interweave and advance state-of-the-art computational modeling, genome editing, omics measurements, systems-biology analyses, and cell-free technologies to expand the set of platform organisms that meet DOE bioenergy goals. cBioFAB will (i) reconceive how we engineer complex biological systems by linking pathway design, prospecting, validation, and production in an integrated framework, (ii) enable systems-level analysis of the David T. Jones collection, one of the largest collections of clostridia strains in the world, to uncover novel metabolic pathways, regulatory networks, and genome editing machinery, and (iii) open new paths for synthesis of next-generation biofuels and bioproducts from lignocellulosic biomass.

Advancement in CRISPR-based genome editing tools have greatly accelerated the ability to modify and domesticate a variety of microorganisms, but some bacteria remain recalcitrant to these technologies. Heterologous expression of CRISPR nucleases often cause cellular toxicity by unclear mechanisms in many bacteria, limiting the ability to edit and modify these organisms. However, endogenously encoded CRISPR-Cas systems are broadly present among bacteria and archaea. Harnessing these endogenous systems presents away to circumvent the drawbacks associated with heterologous nuclease expression and enables access organisms that were previously inaccessible. One hurdle to using endogenous systems is identification of functional protospacer adjacent motif (PAM) sequences, a requirement for nuclease targeting. Most current experimental methods determine PAM sequences via selection of a pooled randomized PAM library in the presence of a CRISPR system and use of next generation sequencing data as a readout of PAM frequencies. However, this technique is not feasible in organisms with very low transformation efficiencies and Cas proteins that are difficult to recombinantly express. Computational approaches attempt to address this same problem by reversing the process of spacer acquisition *in silico*. Since spacers in CRISPR arrays originated in organisms encoding a PAM next to the spacer sequence, sequence alignment can be used to link spacers to potential PAMs. A few tools have used this approach, but efforts thus far have not been continuous from spacer to PAM prediction or do not allow user input of curated data. Here we present Spacer2PAM, an R package that allows users to predict PAM sequences and design targeted PAM libraries from user-provided CRISPR array spacer sequences. The package includes functions to standardize and manipulate sequence and alignment data as well as predict and visualize PAM sequences from those data. We identify two regimes in which the tool is effective; a quick method to predict a PAM likely to be functional and a comprehensive method to design targeted PAM libraries. The quick method was able to identify functional PAM sequences for 8 out of 10 model CRISPR systems tested, with the other 2 systems yielding partial predictions. The comprehensive method was able to inform targeted library design that would achieve a functional PAM in 16 transformations or fewer for 9 out of the 10 model CRISPR systems. Using this tool, we apply the quick and comprehensive methods to organisms with unique carbon metabolism and suggest functional PAMs as well as targeted PAM libraries. We anticipate this consolidated and improved computational pipeline will enable faster domestication of endogenous and novel CRISPR systems, especially in organisms that have poor transformation efficiencies.

We acknowledge the Department of Energy grant DE-SC0018249 for funding of this project.