Next Generation Sequencing: Signature Sequence Detection for *In Silico* Primer Design

Shane Mitchell, Mychal Ivancich, Danielle Montoya, Jane Tang, Mitchell Holland, Nate Dellinger, Sterling Thomas

Noblis Inc, Reston, VA

Abstract

Identification of bacterial organisms and viruses plays an important role in a multitude of areas such as healthcare, biodefense, research, and food safety. Continually increasing databases of sequence data has made it possible to identify genomic regions specific to species, even serotypes of bacteria, as well as various classes of virus. As a part of an internal research project, Noblis examined the possibility of identifying regions of specific genomes that could be used to differentiate them from all other known genomes, including those of closely related organisms.

Background

Primer Design

- Rational design of highly discriminatory primers is essential for diagnostic assays
- Primers must also have good sensitivity, high selectivity, low likelihood to develop secondary structures, and function at a reasonable temperature.
- Many existing assays are designed from limited sequence information or are gene specific

Multiple Reference Analysis

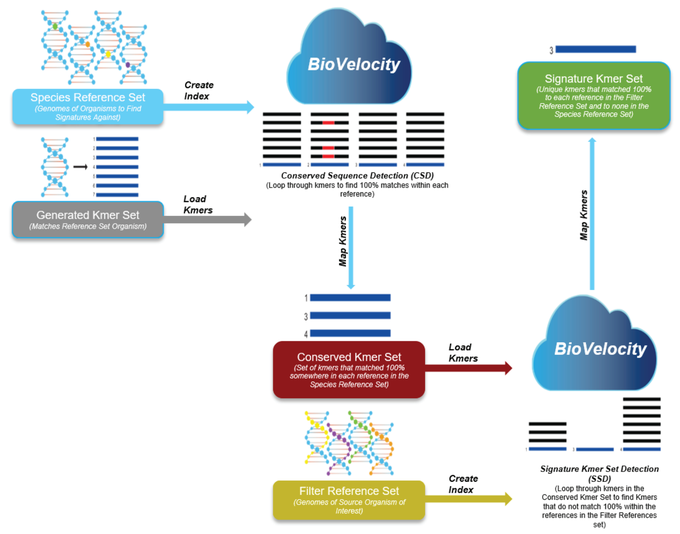
- Discovering Signature Sequence Domains (SSDs) by an exhaustive comparison of one reference to many can help identify regions for primer design
- SSDs are defined as a set of unique sequences that are only found within a specified read-set. If they are detected against a large reference set, such as all bacteria, they are highly specific and are likely good targets for primer design

Objectives

- Determine genomic regions which are unique at different phylogenetic levels, such as genus, species, and strain
- Evaluate the viability of these regions for primer design *in silico*

Methods

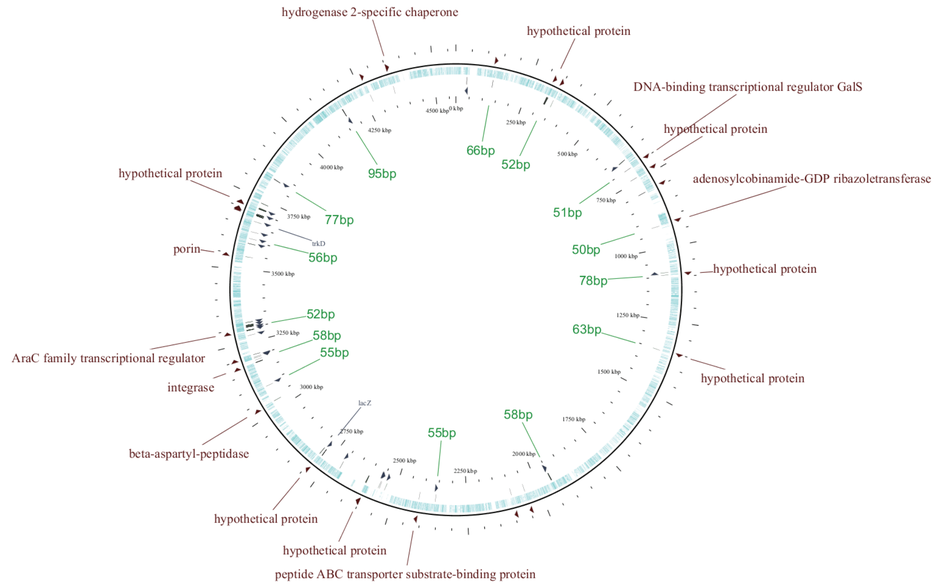
- Select and kmerize the reference genome(s) for target organism(s)
- Select and kmerize the reference genomes for organisms from which it would be desirable to generate distinct primers
- Find kmers that are present only in the target set and are not present in the larger set
- Map signature kmers back to reference and evaluate potential for primer design.



Flow graph of reference data to produce Signature Sequence Domains (SSDs)

Results

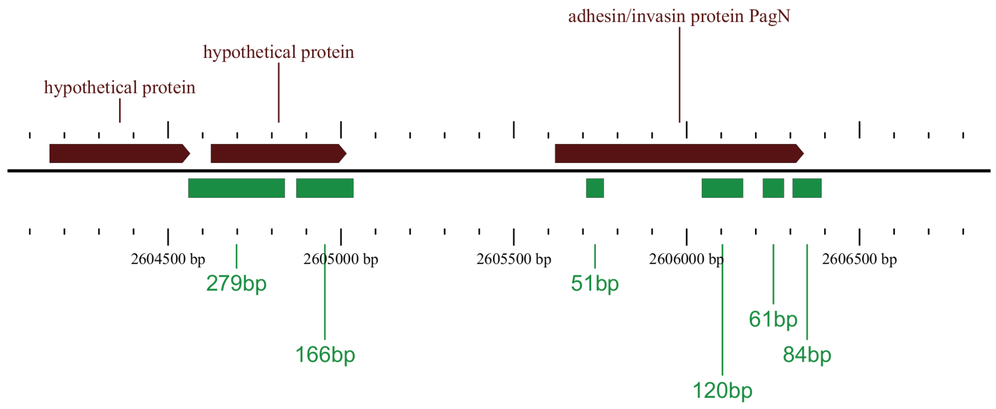
The identified signature regions were evaluated as potential primers using the open source tool Primer 3 and were confirmed for specificity using BLAST. This pipeline supports continual generation of primers, particularly to combat assay degradation occurring over time.



Multiple signature regions were observed (shown in green). This image show the regions detected in the *Salmonella enterica* subsp. *arizonae serovar*

Conclusion

- Signature Sequence Domains can be detected using many different organisms such as Salmonella and E. coli as well as viruses such as Ebola.
- Sequences were found both inside and outside coding regions
- Sufficient sequences were found to generate viable primer targets
- Larger reference sets produce more specific signatures
- Future direction would include verifying potential primers in a wet lab



Zoomed in view of a region of the above genome. Signature regions occurring in genes coding for protein production were observed (shown in green)

Acknowledgement

Thank you to the Bioinformatics Lab at Noblis and SFAF 2017. Special thanks to Mychal Ivancich for developing the SSD technique.