REPEATANALYZER: A TOOL FOR ANALYZING AND MANAGING SHORT-SEQUENCE REPEAT DATA

Abstract

by Helen Noel Catanese, M.S.
Washington State University
May 2017

Chair: Assefaw H. Gebremedhin

Short-sequence repeats (SSRs) are repetitive DNA elements, which occur in both coding and non-coding regions and may be exact or inexact copies. When heterogeneous SSRs are present at a given locus, it is possible to take advantage of the pattern of different repeat sequences to genotype microbial strains. This Master's thesis presents a new software tool, called RepeatAnalyzer, for analysis and management of SSR data, along with case studies that illustrate the tool's various functionalities. In particular, RepeatAnalyzer is a suite of algorithms for tracking, managing, analyzing, and cataloging repeating genetic elements like SSRs for use in genotyping and genetic characterization of a wide variety of microbial species. We demonstrate the capabilities of RepeatAnalyzer using *Anaplasma marginale* as a model species and *msp1a* as a model gene. *A. marginale* is a tick-borne bacterial pathogen that infects cattle for which there is currently no reliable vaccine. Characterization of this microbe is a necessary step toward developing such a vaccine. RepeatAnalyzer's genotyping functionality is validated for all SSRs and genotypes reported in 21 publications, using 380 *A. marginale* isolates gathered from the five publications within that list which provided access to their data. The tool produces accurate results

in every case. The analysis and visualization functionalities of the tool are demonstrated using several examples, including a comparison of the diversity between Mexican and U.S. *A. marginale* strains, and producing geographic plots based on sample queries. RepeatAnalyzer's applicability to a variety of species is shown on SSR data from *Anaplasma centrale*, a species closely related to *A. marginale* and *Streptococcus pneumoniae*, a species that commonly infects humans and is widely studied.

We also conduct hierarchical and graph based clustering analysis of *A. marginale* SSRs. This shows that the majority of clusters are distributed widely, rather than being geographically isolated. Additionally, the clustering analysis shows that the number of SSR groups in the Northern hemisphere was significantly lower than in the Southern hemisphere.