Abstract

Analysis of genotypic and phenotypic diversity of *Aggregatibacter actinomycetemcomitans* using high-throughput sequencing and comparative genomics

Weerayuth Kittichotirat

Bumgarner Laboratory
Department of Microbiology
University of Washington, USA

A bacterial species can be described by a pan-genome consisting of a core gene pool shared by all strains, plus a variable gene pool consisting of partially shared and strain-specific genes. The elucidation of the pan-genome for a species from sequence analysis of related strains can help in understanding of how genetic variability drives pathogenesis within a bacterial species. In this study, we have developed a genome sequencing, annotation and comparison pipeline for studying the genotype-phenotypic correlations between closely related bacterial strains. We have primarily applied our pipeline to analyze a collection of strains of an oral bacterium known as *Aggregatibacter actinomycetemcomitans*. *A. actinomycetemcomitans* is a human pathogen that is heavily associated with aggressive periodontitis and other systemic infections. Not all strains of this bacterium exhibit virulent phenotypes and by using comparative genomics approach we have characterized significant genomic variations found within this bacterial species at both structural as well as gene content level. At the structural level, we have shown that strains from the same serotype appear to have fairly conserved genomic arrangement whereas strains from serotype a can have significantly different arrangement from strains belonging to serotype b or c. These results suggest a significant evolutionary divergence between serotype a strains and serotypes b/c strains of *A. actinomycetemcomitans*. Additionally, the distinct patterns of genome arrangement may suggest phenotypic differences between serotype a and serotypes b/c strains. At the gene content level, we have shown that the pan-genome of *A.*

actinomycetemcomitans is large and only about 60% of all gene content constitute the core gene pool shared by all strains. Our mathematical extrapolation of the gene content data also suggests that the *A. actinomycetemcomitans* pan-genome is opened ended and that new genes will continue to be identified when genomes of additional strains are sequenced. Addition, by performing pairwise comparisons, we show that the gene content variation within each serotype can be greater than between serotypes, suggesting that this classical way of bacteria typing does not reflect the genetic diversity of *A. actinomycetemcomitans*. Together, these findings have implications for pathogenesis of *A. actinomycetemcomitans* and could be used to elucidate the molecular basis of variable virulence among *A. actinomycetemcomitans* strains. They also raise questions on the working definition of a bacterial species and suggest that the research strategies for bacterial genomes, which are typically based on the concept of limited genomic variability within a species, may need to be reconsidered. Besides *A. actinomycetemcomitans*, the computational infrastructures developed during the course of this dissertation have also been used to analyze closely related strains of other bacterial species, which has resulted in several publications.