Computational Tools for the Analysis of High-Throughput Immunoglobulin Sequencing

Yale University

PI: Steven H. Kleinstein

Grant Number: 3RO1A1104739

The ability of our immune system to respond effectively to pathogenic challenge or vaccination depends on a diverse repertoire of Immunoglobulin (Ig) receptors expressed by B lymphocytes. Each Ig receptor is unique, having been assembled during lymphocyte development by somatic recombination of gene segments. During the course of an immune response, B cell that initially bind antigen with low affinity through their Ig receptor are modified through cycles of somatic hypermutation (SHM) and affinity-dependent selection to produce high- affinity memory and plasma cells. This affinity maturation is a critical component of T cell dependent adaptive immune responses. It helps guard against rapidly mutating pathogens and underlies the basis for many vaccines. Large-scale characterization of B cell Ig repertoires is now feasible in humans. Driven by the dramatic improvements in high-throughput sequencing technologies, these data are opening up exciting avenues of inquiry. Features of the B cell repertoire, including polymorphisms, biased segment usage and diversity, can be correlated with clinically relevant outcomes, such as susceptibility to infection or vaccination response. These data can also contribute to basic understanding of B cells and adaptive immunity. In particular, the ability to estimate positive and negative selection from Ig mutation patterns has broad applications not only for understanding the immune response to pathogens, but is also critical to determining the role of somatic hypermutation in autoimmunity and B cell cancers. Although promising, repertoire-scale data also present fundamental challenges for analysis requiring the development of new techniques and the rethinking of existing methods that are not scalable to the millions of sequences being generated. This proposal describes novel approaches for the analysis of high-throughput Ig sequencing data sets enabled through a combination of bioinformatics and statistics method development, computational modeling and sequence data-mining. New ways to characterize repertoire properties will be developed that have the potential for use as biomarkers for disease risk, diagnosis and prognosis. Specifically, methods will be developed to: (Aim 1) group sequences into clones and improve V(D)J segment assignment, thus allowing identification of somatic mutations, (Aim 2) model SHM mutability and substitution patterns so they can be quantified and compared across groups, thus providing insights into underlying mutation mechanisms, and (Aim 3) quantify selection and characterize clonal diversity, providing information on affinity maturation and response dynamics. These methods will be validated through a combination of simulation-based studies, as well as testing on new experimental gold-standard data sets from both human and murine systems. All of the methods will be made widely available through web interfaces and distribution of open-source code. PUBLIC HEALTH RELEVANCE: This project will develop and validate computational methods to analyze large-scale immunoglobulin sequencing data sets that have become possible with the advent of next-generation sequencing technologies. Through quantitative characterization of the immunoglobulin repertoire, these methods will provide insights into the mechanisms underlying autoimmune disease, as well as biomarkers for susceptibility to infection or vaccination response. In addition, methods to identify and analyze mutation patterns will be used to investigate the role of somatic hypermutation in physiological and pathological immune responses.