Data Integration for Genome-wide Association Studies of Human Diseases

Qi Sun, Lalit Ponnala, Cornell University

Managing data for high throughput genomics studies requires researchers to deal with issues including the integration of heterogeneous data sets, the use of tools for data access, and the issues arising around confidentiality. We have been working together with researchers of Dr. Ron Crystal's group of Weill Cornell Medical School to develop a data processing pipeline for their COPD project (Chronic Obstructive Pulmonary Disease, which is the 4th leading cause of death in the United States). Microsoft SQL Server 2005 was used as the database engine for this project. We developed a schema that can accommodate the heterogeneous multi-media clinical data, and the high-throughput genotyping data from multiple platforms. The built-in SQL Server encryption functions were used for storing sensitive patient information. On the client side, users can enter and retrieve data through VSTO add-ins for Excel, as most researchers are already familiar with using Excel. Our experience showed that the SOAP web service and the Excel based client applications can be a versatile solution for data integration of high throughput genomics and proteomics projects.