tranSMART Supports a Post-GWAS Data Coordinating Center

Dan Stuart, M.S. ¹; John Harju, MBA ¹; Wei Shi, MBA, M.S. ¹; David Hanauer, M.D., M.S. ¹; Dina Aronzon, M.S. ²; Jinlei Liu, M.S. ²; Raveen Sharma, M.S. ²; Frank J. Manion, M.S. ¹; Haiping Xia, PhD ²; Carolyn M. Hutter, PhD ³; Dr. Stephen B. Gruber, PhD, M.D. ⁴
¹ University of Michigan Comprehensive Cancer Center, Informatics Core, Ann Arbor, MI; ² Recombinant Data Corp., Newton, MA; ³ The Fred Hutchinson Cancer Research Center, Seattle, WA; ⁴ USC Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, CA

Abstract

The University of Michigan Comprehensive Cancer Center Informatics Core, in cooperation with Recombinant Data Corp, has implemented the open-source tranSMART application (http://www.transmartproject.org/) to support data management and analysis for the Colorectal Transdisciplinary Study (CORECT) Consortium coordinating center. CORECT is a post-GWAS consortium study of the genetic and epidemiologic risk factors of colorectal cancer, and collects data from multiple member institutions. tranSMART is used within CORECT to support data intake, distribution, and analysis.

Problem

The goals of the CORECT project are to thoroughly investigate and identify susceptibility loci for colorectal cancer, to characterize the biologic basis of inherited susceptibility, and to recognize how genetic variation may be quantified and modified by genetic and environmental risk factors. The requirements for the coordinating center include secure data storage, data distribution, data versioning, and management of data access between member institutions. These requirements are applied to both SNP data and epidemiological data from member sites.

Solution

To achieve these goals, the UM Cancer Center Informatics Core has collaborated with Recombinant Data Corp to implement the tranSMART application. Recombinant developed SNP data visualization and export capabilities for tranSMART to support the goals of the CORECT consortium, while consortium members at Fred Hutchinson Cancer Research Center developed a harmonized ontology to characterize epidemiological data. tranSMART’s i2b2-based Dataset Explorer facilitates exploration and characterization of data through a query-builder interface based on the epidemiological ontology. The Dataset Explorer also generates visualizations and statistical analytics with R to compare subsets of a clinical data set, allowing consortium members to perform initial analyses and explore the data for further research. Members can then export a complete data package (including epidemiological and genotype data in the PLINK format) for advanced analysis. Epidemiological and SNP data for approximately 4,200 subjects have been loaded into our tranSMART data warehouse, and additional data sets will be loaded as they become available throughout 2012.

The partnership between the UM Cancer Center Informatics Core and Recombinant Data Corp. will drive the evolution of tranSMART and the integration of additional tools. tranSMART can be extended with additional R libraries, by integrating new i2b2 modules (cells), or by adding new functionality to tranSMART’s open-source code base. We will use CORECT to identify scientific use cases that can be addressed in future versions of tranSMART, and we are looking at incorporating web services to allow integrating tranSMART with other data management applications.

1956