Systems Biology Research Symposium
Oral Presentation Session

Grand Ballroom
Tuesday, June 5th
7:00-8:30pm

A Text-Based Strategy to Identify Disease Variants: Looking for Gene Relationships Across Implicated Loci (GRAIL)

Soumya Raychaudhuri[1-3], David Altshuler[2,3], and Mark Daly[2,3]

[1]Divisionof Rheumatology, Immunology, and Allergy, Brigham and Women's Hospital, Boston, Massachusetts, 02115, USA. [2]Broad Institute, Cambridge, Massachusetts, 02142 USA [3]Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, 02114, USA.
Presenter’s email address: soumya@broad.mit.edu

Translating a set of disease regions into insight about pathogenic mechanisms requires not only the ability to identify the key disease genes within them, but also the biological relationships between those key genes. Here we describe a statistical method, *Gene Relationships Among Implicated Loci* (*GRAIL*) that takes a list of disease regions and automatically assesses the degree of relatedness of implicated genes using 250,000 PubMed abstracts. We tested GRAIL, by assessing its ability to separate true disease regions from many false positives disease regions in two separate practical applications in human genetics. First, we took 75 nominally associated Crohn’s disease SNPs, and applied GRAIL to identify a subset of 13 SNPs with highly related genes. Ten of these convincingly validated in follow-up genotyping. Next, we applied GRAIL to 170 rare deletion events seen in schizophrenia cases (less than one-third of which are contributing to disease risk). We demonstrate that GRAIL is able to identify a subset of 16 deletions containing significantly related genes; intriguingly many of these genes are expressed in the central nervous system and play a role in neuronal synapses. These genes constitute a very exciting set for additional investigation. Finally, we demonstrate novel applications to identifying novel oncogenes from a large collection of somatic copy number alterations in cancer, and in identifying novel rheumatoid arthritis genetic risk loci by prioritizing SNPs for replication.

Key words: genetics, systems biology, text mining, genomics, genome-wide association studies (GWAS)