Poster B-34

A computational system to select candidate genes for complex human traits

Authors:
Kyle J Gaulton (*Department of Genetics, University of North Carolina*)
Karen L Mohlke (*Department of Genetics, University of North Carolina*)
Todd J Vision (*Department of Biology, University of North Carolina*)

Short Abstract: A wealth of genomic data exists that can be mined to uncover functionally relevant links between genes and traits. We developed a computational system, named CAESAR, which exploits these resources by combining text mining, data mining and data integration to prioritize potential candidate genes for complex human traits.

Long Abstract:
The genes underlying complex human traits remain largely unknown. A wealth of genomic information in the form of publicly available databases and ontologies is underutilized as a potential resource for uncovering functionally relevant, but not directly obvious, links between genes and traits. We have developed a computational system, named CAESAR (CAndidatE Search And Rank), which exploits these resources by combining text mining, data mining and data integration to prioritize potential candidate genes for complex human traits. CAESAR represents a novel system in that candidate selection is not restricted to one or several genomic regions, although it can be used for this purpose, and is designed for candidate gene studies for which a modest to large number of genes are selected. The underlying principles can be applied to mono- or multi-factorial traits in any organism for which an annotated gene set exists.

CAESAR uses 11 publicly available databases and four well-characterized ontologies to perform searches, ultimately ranking up to one third of the known genes in the human genome. The databases were selected based on areas of biological knowledge researchers typically use to manually select candidate genes, including aspects of genomics, proteomics, metabolomics, expression, and mutant phenotype studies. CAESAR uses four possible methods of integration to combine the results of database searches into a prioritized candidate gene list. In effect, CAESAR mimics the steps a researcher would undertake in selecting candidate genes, albeit faster, potentially more thoroughly, and in a more quantitative manner.

The CAESAR algorithm is comprised of three main steps. First, previously implicated genes are identified and ontology terms are ranked based on similarity to an input text, referred to here as a corpus. If the user supplies an Online Mendelian Inheritance in Man (OMIM) identifier, CAESAR will use the OMIM record as the corpus. Alternately, the user can provide any other body of text, for instance a review article. The corpus is then used for information retrieval in two ways. First, human gene symbols are identified within the corpus. The extracted genes need not be true effectors of the disorder, and are only considered previously implicated as prior candidate genes. The master gene list, consisting of gene names, standard symbols, database identifiers, and corresponding mouse homologs, is

compiled from the National Center for Biotechnology Information (NCBI) Entrez Gene database and Ensembl. Second, the corpus is used as the query document to quantify the relevance of terms within several different standard biomedical ontologies using a vector space similarity-based method.

In the second step, genes associated with the above ontology terms are ranked for each data source independently by querying eight databases. The phenotype ontology results are used to query the mouse genome database (MGD) for genes producing a given phenotype when mutated and query the genetic association database (GAD) for positive evidence of association with a given phenotype. The anatomical ontology results are used to query the UniProt database for genes expressed in a given tissue. The gene ontology results are used to query the Gene Ontology Annotation (GOA) database for genes annotated with a given Gene Ontology biological process or molecular function term. Finally, the extracted genes and their associated protein products are used to query interaction databases BIND and HPRD for proteins interacting with extracted protein products, query the KEGG pathway database for proteins in pathways with extracted protein products, and query the InterPro protein domain database for proteins sharing protein domains with extracted protein products. The outputs of the eight database searches are eight ranked gene lists.

Finally, these gene lists are merged to produce a single ranked gene list using one of four methods, each of which weights the evidence across the lists to provide a final score for each gene. The first three methods calculate the maximum, average, and sum of the scores for a given gene across all gene lists, respectively. The fourth method calculates scores by considering both the length and total sum of each gene list.

To test the ability of CAESAR to select valid candidates, we selected 17 susceptibility genes from recently published reports providing strong evidence of statistical association with 11 common complex human disorders. On average, 10,950 genes were ranked per trait. 8 of the 17 tested genes (47%) ranked in the top 2% of all ranked genes for their tested trait. 4 of the 17 genes (24%) ranked in the top 1% of all ranked genes. Over all ranked genes, the mean rank was in the 85th quantile, significantly higher than expected by chance.

Previous attempts at candidate gene prioritization have each defined success by the presence of tested disease genes in an enriched candidate set. Here, we define success as a tested gene's presence in a subset of the top 2% of genes ranked for a trait. This subset can be considered the practical constraints on the selection of a medium to large candidate gene set for study. While the evaluation methods used in CAESAR and these other systems are somewhat arbitrary, the percentage of successful tests was comparable to slightly larger than in other systems.

The CAESAR system has been designed for candidate gene studies for which a modest to large number of genes are selected. It can also be used to select candidates from a linkage region and to prioritize SNPs showing significant association in a genome-wide association (GWA) study. CAESAR can function as a fully automated system but can also be guided where greater user control is desired.