AN AUTOMATED APPROACH TO IDENTIFYING DISEASE-GENE ASSOCIATIONS FROM THE MEDICAL LITERATURE TO INFORM GENE PANEL DESIGN

Mark J Kiel MD PhD1, Matthew Schu PhD2, Steve A Schwartz1, Victor Weigman PhD2
1GENOMENON, Ann Arbor MI; 2I’Q Solutions - EA Genomics, Morrisville NC

Rational approaches to the design of gene panels for next-generation molecular diagnostic assays are made challenging by the amount and complexity of information that must be manually collected and assessed from the medical literature. We have developed a novel text-mining infrastructure that automatically identifies disease-gene-treatment associations from the titles and abstracts of more than 264M scientific publications in PubMed and disease-gene-variant associations from the full-text of 3.9M prioritized articles. For this work, curated lists of diseases and genes comprising 11.7K and 50.9K total entries, respectively, were used as initial search parameters to identify high-yield content. Next, custom-designed algorithms were used to scan full-text articles using a comprehensive variant listing comprising 620M total entries for every possible cDNA or protein variant in all gene transcripts sorted by biological outcome as second-tier search parameters. In total, we identified 909K putative disease-gene associations [24 articles per association on average] and 543K total variants distributed across all articles. This information was organized according to the strength of the association based on the total number and quality of individual citations and the position of disease-gene and disease-gene-variant key terms within the text.

We decided to demonstrate this approach in Acute Myeloid Leukemia (AML) to create an NGS panel, where outcome is very poor and lists of well-known mutations and biomarkers for treatment are not well characterized outside the context of cytogenetics. In total, 11K unique variants in 151 genes were found to be associated with AML and ranked according to the number of citations for each. Each variant had been cited at least once from 3,865 individual scientific publications. These variants were further classified according to the journals in which they appeared and the resulting data was manually inspected for accuracy. Additionally, custom search algorithms identified recurrent gene amplifications, gene deletions and gene-fusions resulting from translocations and were also catalogued and incorporated in the final panel design. The final in silico panel design was compared to commercially-available panels and included several genes with a clear association to AML with potential clinical significance that are otherwise not present on any such panel. In conclusion, we have developed and tested a tool that rapidly and comprehensively interrogates, organizes and displays genomic targets for diseases and has promising applications for clinical panel construction with significantly reduced curation time and research effort.

INTRODUCTION

Translating genomic knowledge from the medical literature into clinical insight to inform gene panel design is a challenging and subjective process.

METHODS

First, meaningful content was identified based on scans of the titles and abstracts for disease terms and gene names. Next, automated full-text data processing identified disease-gene-variant associations from prioritized content. The resultant database was queried by disease to identify recurrently associated genes and variant names. Gene amplifications, deletions and gene fusion events were also identified.

Acute Myeloid Leukemia was selected as a test case due to the complexity of the pathogenetic mechanisms contributing to this disease and the substantial interest in effectively targeting this disease.

RESULTS + CONCLUSION

Our novel approach to automated, evidence-based gene panel design eliminates hours of searching and allows for more focused attention on assessing actionable and clinical significance of candidate biomarkers.

Interested in learning more? Contact us at info@genomenon.com www.genomenon.com

Contact us at info@genomenon.com