Automated Identification, Prioritization and Visualization of Large-Scale Genomic Data

M. J. Kiel [1,2] N. T. P. Patel [1,2] R. W. Peng [3] S. A. Schwartz [2,3] M. S. Lim [1,2,4] K. S. J. Elenitoba-Johnson [1,2,5]

1) Department of Pathology, University of Michigan, Ann Arbor MI; 2) GENOMENON, Ann Arbor MI; 3) AlfaJango, Ann Arbor MI; 4) Department of Hematopathology, University of Pennsylvania, Philadelphia PA; 5) Center for Personalized Diagnostics, University of Pennsylvania, Philadelphia PA.

MASTERMIND

INTRODUCTION

The tertiary component of genome sequence analysis requires identification of informative associations between diseases or drugs and genes and geneic mutations for extraction of clinicobiological meaning from patient data. The accuracy and efficiency of tertiary analysis is limited by inaccessibility and non-uniformity of this information in both public databases and primary scientific articles. To overcome these limitations, we have developed MASTERMIND - a suite of novel analytic and visualization tools that dramatically reduces the time and effort required to organize and integrate genomic information from any data source including millions of full-text scientific articles and dozens of heterogeneous variant databases.

METHODS

To comprehensively interrogate genomic data from both structured and unstructured databases, an automated querying architecture was designed using customized open-source analytics engines and a combination of publicly available and custom developed APIs. Curated lists of diseases and gene transcripts with synonyms comprising 11.7K and 50.9K total entries were used as initial query parameters. Custom-designed algorithms were used to generate comprehensive mutation query lists comprising 602M total entries sorted by biological outcome and used as second-tier queries. Using titles and abstracts of 24M primary articles, we identified 909K putative disease-gene associations which we then confirmed by automated scanning of 5.8M full-text articles to comprehensively identify all disease - mutation citations within each article. Integrated metadata for each finding was used to prioritize disease - gene - mutation associations in accordance with the abundance and quality of supporting evidence. To facilitate rapid comprehension, these associations were then organized within a singular graphical UI with display of all relevant information from the primary source material used to drive data prioritization including interactive access to annotated full-text articles. This method of rapidly assimilating disease - gene - mutation associations for display was reproduced for several additional structured databases including ClinVar.

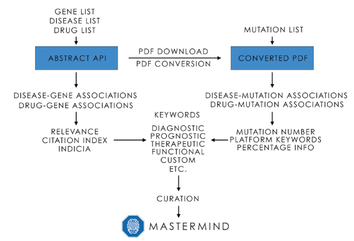

Figure 1. Schematic of Information Flow for MASTERMIND Database Assembly. Query lists for genes, diseases and drugs are used to prioritize articles containing genetic information through the eutils API. PMIDs and doi values corresponding to each of these articles are used to download full-text PDFs followed by conversion to searchable text. A comprehensive mutation list is then used to identify and categorize genetic variants within the text, figures and tables. Additional keywords are used to categorize article content. Meta-information about each article is also captured followed by curation of identified genetic variants for inclusion in the MASTERMIND database.

Figure 2. Screenshot Illustrating the Ability to Query MASTERMIND by Disease and/or Gene and Mutation. The resulting database can be queried by searching either for specific disease entities to identify all associated genes or otherwise by searching for a specific gene to identify all associated diseases. Synonyms for diseases and genes are recognized within the query window. Specific mutations within any given gene can also be queried to identify all articles or databases describing that specific mutation.

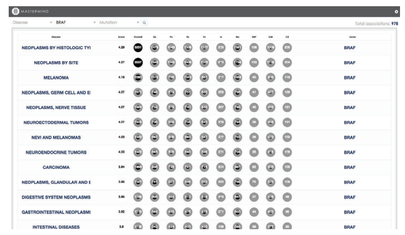

Figure 3. MASTERMIND Screenshot Illustrating the Identification of Disease Associations for BRAF. Every possible combination of disease - gene association is automatically screened for during assembly of the MASTERMIND database and the results are prioritized by the number of articles containing any specific disease - gene association. The number of articles is further fractionated by identifying within each article keywords with relevance to the following categories of significance: diagnostic, prognostic, therapeutic, functional, etc. For the BRAF example shown, association with melanoma, neuroectodermal tumors, gastrointestinal neoplasm as well as thyroid malignancy (not shown) were readily identified.

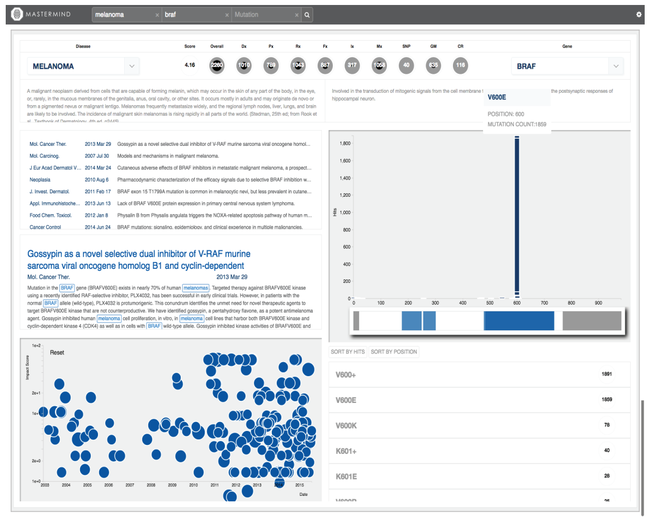

Figure 4. MASTERMIND Screenshot Illustrating the Results for the Melanoma-BRAF Disease-Gene Association. For any disease - gene association, all relevant articles are displayed by title in list-form (top left) prioritized according to the location of the keywords within the text (i.e. appearance in the title or abstract or full text or supplemental text). The title and abstract of selected article is displayed in full with all relevant keywords highlighted (middle left). The landscape of articles containing all selected keywords is displayed by publication date (x-axis) and citation index (y-axis; bottom left). Each mutation identified in any of the articles containing the disease - gene association are displayed (right) in list form for filtering (bottom right) and in graphical form (top right) according to the number of articles discussing this mutation displayed along the linear axis of the protein with functional domains depicted as colored boxes. This plot highlights the recurrently mutated V600 residue within BRAF.

Figure 5. Representative Article Illustrating MASTERMIND PDF Conversion with Illuminated Gene Symbol and Disease Keywords. Every mention of any disease or gene or its corresponding synonym is recognized and highlighted for easy identification of high-yield information within the text. Mutations are identified and highlighted whether described as nucleotide level changes or protein level changes. Information contained within tables and figures are also identified following conversion of PDF to searchable content. Within the MASTERMIND user-interface, sentence fragments containing mention of any specific mutation are extracted and displayed for regard content comprehension and localization of information in context (not shown). Shown above is identification of the *NOTCH2* p.V1667I mutation in association with Splenic Marginal Zone Lymphoma.

DISCUSSION

MASTERMIND autonomously and comprehensively interrogates, organizes and displays genome variant data from structured and unstructured data sources of genomic information. Here we demonstrate how we have engineered MASTERMIND to automatically identify disease - gene - mutation associations from multiple primary scientific articles in a hypothesis-neutral and highly parallelized process and then organize the results in a graphical user interface emphasizing the strength of the evidence supporting the likelihood of pathogenecity for individual mutations. MASTERMIND has promising applications in expediting tertiary analysis of human genome sequencing data in clinical assays of individual patients for large gene panels, whole exome and whole genome sequencing assays. The interface allows molecular diagnosticians to quickly assess whether a variant of uncertain significance (VUS) has ever been reported previously; and if so, how many times and in what journals and in association with which diseases and, finally, whether any evidence exists that the variant is useful for diagnostic, prognostic or therapeutic medical decision-making. The potential of integrating an automated annotation pipeline into routine diagnostic genome sequencing interpretation has the potential to further reduce the time and expense associated with tertiary analysis.

NEXT STEPS

While we have focused our initial efforts on extracting genomic knowledge from primary scientific articles, we have also engineered the back-end processing of MASTERMIND to interrogate other sources of unstructured text data (such as OMIM and clinicaltrials.gov) as well as tabular and relational databases (such as dbSNP, ClinVar and COSMIC). The results of querying these databases are automatically intercalated into the pre-existing entries of MASTERMIND where redundancies appear. Associated metadata comprising each entry from these external data sources is also extracted and used to further annotate MASTERMIND entries for ease of information retrieval.

GENOMENON