The edge-permutation procedure retains the number of edges and the degree distribution of the network.[43] Two edges, A-B and X-Y, are chosen at random and reshuffled to create the edges A-Y and X-B. Reshuffling is skipped if the edges A-Y and X-B already exist. Reshuffling is performed 10,000 times, resulting is an edge-randomized version of CDN-o, which we call CDN-er and for which we can again compute the GEF. We constructed 1,000 versions of CDN-er and plotted the distribution of the resulting GEF values in Figure S4A. As one can see, the p value of the CDN is less than 0.001 because none of the edge-randomized CDNs achieved the same or a smaller GEF than the original CDN.

We additionally performed a test in which we randomized the HPO terms associated with each disease (ar). For this, we randomly selected 50% of the terms associated with each disease and replaced them with randomly selected HPO terms. We computed the randomized CDN (called CDN-ar) by using the above procedures used to construct the CDN-o. We repeated this procedure 100 times and computed the GEF for each CDN-ar. Note that each CDN-ar might not have the same amount of nodes and edges as the CDN-o. When using the same simcut (2.0) used for constructing the CDN-o, we obtained much smaller networks (fewer than 100 nodes). The distribution of GEF values of CDN-ar with simcut 2.0 is shown in Figure S4B. No CDN-ar achieved a GEF less than or equal to the CDN-o GEF, which corresponds to a p value of less than 0.01. We modified the simcut to 1.4 because it leads to CDN-ar versions with approximately the same amount of nodes as CDN-o. The distribution of the resulting GEF values is shown in Figure S4C. Again, not a single CDN-ar constructed with a simcut of 1.4 achieved a GEF less than or equal to the CDN-o GEF, which corresponds to a p value of less than 0.01.

GWAS Data

GWAS Central provides a comprehensive collection of summary-level genetic-association data and advanced visualization tools to allow comparison and discovery of datasets from the perspective of genes, genome regions, phenotypes, or traits.[33] The project collates association data and study metadata from many disparate sources, including the National Human Genome Research Institute GWAS Catalog,[35] and receives frequent data submissions from researchers who wish to make their research findings publicly available. All gathered and submitted data are extensively curated by a team of post-doctoral genetics researchers who manually evaluate each study for its range of phenotype content and apply appropriately chosen MeSH terms. As of December 2014, the resource contained 69 million p values for over 1,800 studies.

Data and metadata for up to 1,000 associations can be freely downloaded from the BioMart-based system (GWAS Mart), and larger custom data dumps (up to and including the complete database) are available via contacting the GWAS Central development team and agreeing with a data-sharing statement. Thus, to provide data for the present study, we generated a tab-separated file representing 1,574 studies and 34,252 unique SNPs (annotated to 675 unique MeSH terms) and containing the GWAS Central study identifier, PubMed identifier, dbSNP "rs" identifier, p value, and MeSH identifier for all associations with $p < 1 \times 10^{-5}$. We compiled the list of genes considered for our experiments by retrieving the "mapped genes" column from the database SCAN and identifying those genes corresponding to the GWAS Central SNPs. Where no mapped genes were reported, we used the upstream, as well as downstream, genes listed by SCAN.[44]

Results

Generation of Phenotype Annotations for Common Disease by CR

We applied a phenotype-aware CR system (the Bio-LarK Concept Recognizer[40]) to all available abstracts in PubMed in order to extract phenotypic annotations for common diseases. We first retrieved the MeSH terms associated with PubMed abstracts and used them to retain only those abstracts focused on diseases. 5,136,645 of 22,376,811 articles listed in PubMed had an abstract and could be assigned to such a MeSH disease term (see Material and Methods for a description of our inclusion criteria for MeSH disease entries; a total of 3,145 diseases were included). Second, we applied CR on the resulting set, after which a total of 930,805 HPO annotations were assigned to 3,145 common diseases. Finally, we filtered this initial set of HPO terms, by using a ranking-and-clustering method with the aim of maximizing the F-score computed on a manually curated gold-standard set of 41 common diseases (see Material and Methods). This approach aims to maximize the text-mining accuracy, defined as the harmonic mean of the precision and recall of the derived annotations. This final set comprised 132,006 HPO annotations covering 4,459 unique HPO terms. The mean number of annotations per disease was 41.97 (range, 1–271; median, 32) and consisted of terms belonging to all of the top-level HPO categories (Figure S5). Figure 2 provides an overview of the analysis procedures used to generate and validate the common-disease annotations.

As an example, Table S1 lists the annotations produced for "giant cell arteritis" (MeSH: D013700), which includes terms such as "vasculitis" (HP: 0002633), "granulomatosis" (HP: 0002955), and "amaurosis fugax" (HP: 0100576). The annotations are highly accurate, although some nuances are not detected by the CR process. For instance, "facial palsy" (HP: 0010628) and "renal amyloidosis" (HP: 0001917) are classic manifestations of giant cell arteritis. The list of phenotypic manifestations is by no means complete, given that it failed to identify manifestations such as "dysphagia" (HP: 0002015), "trismus" (HP: 0000211), and "encephalopathy" (HP: 0001298). Nonetheless, the CR process was able to capture a largely accurate subset of phenotypic abnormalities for giant cell arteritis, such that 64% of the annotations were true positives.

We estimated the overall quality of the HPO annotations by inspecting the automatically extracted annotations for a set of 41 common diseases randomly chosen from 13 upper-level DO[45] categories that had a MeSH disease identifier and thus could be analyzed analogously to the common MeSH diseases. The process involved manually validating of all HPO annotations extracted by the CR process and comparing them to the results of detailed manual curation