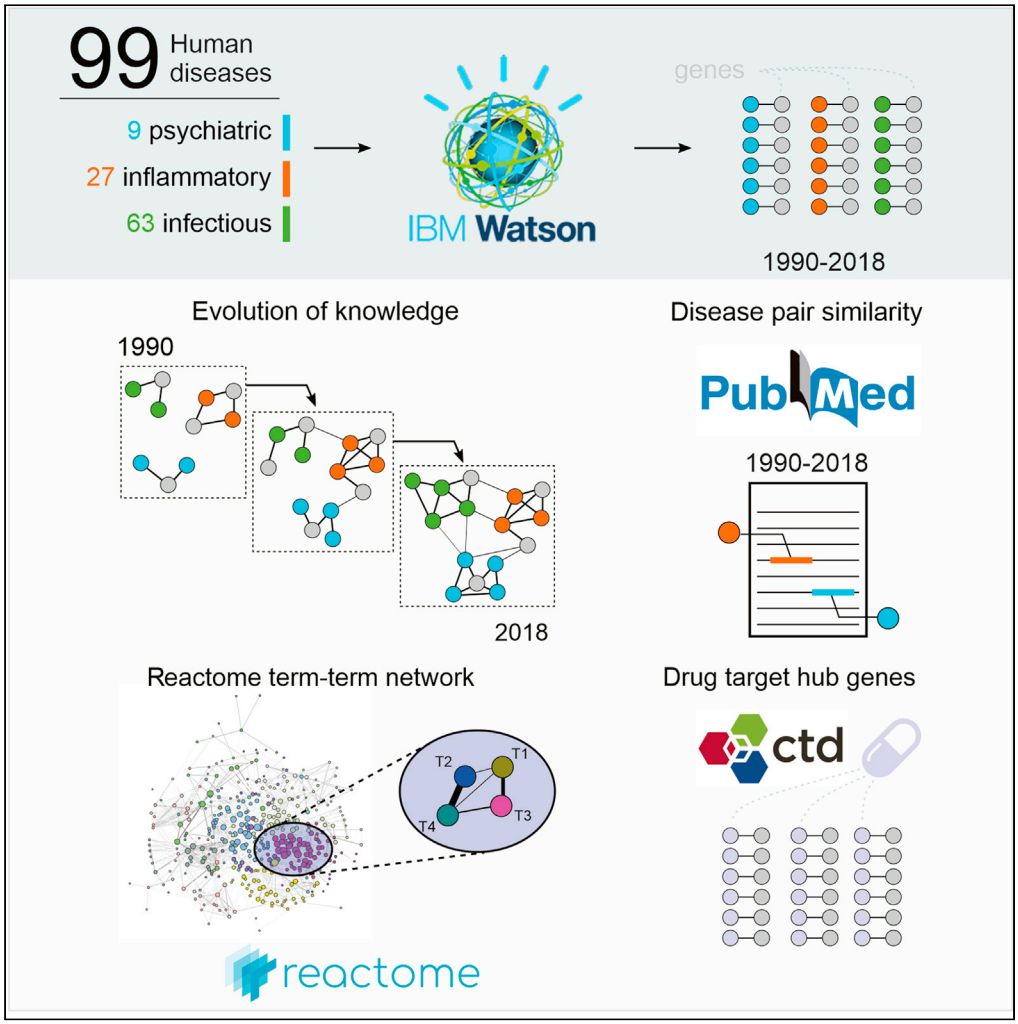
iScience

CellPress
OPEN ACCESS

Article

The evolution of knowledge on genes associated with human diseases



Thomaz Lüscher-Dias, Rodrigo Juliani Siqueira Dalmolin, Paulo de Paiva Amaral, Tiago Lubiana Alves, Viviane Schuch, Glória Regina Franco, Helder I. Nakaya

helder.nakaya@einstein.br

Highlights

Over 3,700 genes were associated with 99 human diseases in the scientific literature

The knowledge on human disease genes increased exponentially in the past 30 years

Knowledge networks revealed genes shared by very different diseases

Hub genes are known drug targets with drug repositioning potential

Lüscher-Dias et al., iScience 25, 103610
January 21, 2022
https://doi.org/10.1016/j.isci.2021.103610

iScience

CellPress
OPEN ACCESS

Article

The evolution of knowledge on genes associated with human diseases

Thomaz Lüscher-Dias,[1] Rodrigo Juliani Siqueira Dalmolin,[2,3] Paulo de Paiva Amaral,[4] Tiago Lubiana Alves,[5] Viviane Schuch,[5] Glória Regina Franco,[1] and Helder I. Nakaya[5,6,7,8,*]

SUMMARY

Thousands of biomedical scientific articles, including those describing genes associated with human diseases, are published every week. Computational methods such as text mining and machine learning algorithms are now able to automatically detect these associations. In this study, we used a cognitive computing text-mining application to construct a knowledge network comprising 3,723 genes and 99 diseases. We then tracked the yearly changes on these networks to analyze how our knowledge has evolved in the past 30 years. Our systems approach helped to unravel the molecular bases of diseases and detect shared mechanisms between clinically distinct diseases. It also revealed that multi-purpose therapeutic drugs target genes that are commonly associated with several psychiatric, inflammatory, or infectious disorders. By navigating this knowledge tsunami, we were able to extract relevant biological information and insights about human diseases.

INTRODUCTION

Thousands of biomedical scientific articles are published every day, piling up with millions of already published papers (Fortunato et al., 2018). The task of keeping up-to-date with this "knowledge tsunami" has become overwhelming for researchers in all areas of science. In this scenario, computational methods such as text mining, machine learning, and cognitive computing are helping scientists to summarize published scientific literature. Recently, machine learning text-mining bibliometric approaches have been used to analyze and integrate a variety of biological, medical, and environmental science data (Littmann et al., 2020; Tan et al., 2021; Zitnik et al., 2019). These include methods that integrate electronic health records (Rajkomar et al., 2018), capture latent knowledge from the material science literature (Tshitoyan et al., 2019), investigate the evolution of research in environmental sciences (Tan et al., 2021), and discover potential novel drugs to treat psychiatric and neurological disorders using cognitive computing and network medicine analysis of the medical literature (Lüscher Dias et al., 2020).

Particularly, the field of molecular biology has seen a remarkable increase in the number of new studies in recent decades. This has resulted in a large number of genes associated with diseases. As a positive consequence of this efflux of genetic knowledge, diseases that were previously not known to have common etiologies are now being connected through their shared alterations in gene expression and interaction patterns, which has opened many potential new roads for clinical advances (Brooks et al., 2014; Carson et al., 2017; Lees et al., 2011; Postma et al., 2011). One significant example of this trend is the association between psychiatric disorders and immune-related diseases (Gibney and Drexhage, 2013; Marrie et al., 2017; Wang et al., 2015).

Network medicine (Barabási et al., 2011), a contemporary approach to studying relationships between genes and diseases, has also been made possible because of the large amounts of data on genes and diseases available online. Moreover, knowledge networks, that is, complex graphs that connect concepts according to the established knowledge, can be analyzed under the network medicine framework to produce novel insights from medical knowledge (Bai et al., 2016; Lüscher Dias et al., 2020).

In this study, we used IBM Watson for Drug Discovery (WDD [Chen et al., 2016a]), a cognitive computing text-mining application, to extract known relationships between genes and psychiatric, inflammatory, and infectious diseases from the peer-reviewed literature published between 1990 and 2018. We

[1]Department of Biochemistry and Immunology, Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil

[2]Bioinformatics Multidisciplinary Environment—BioME, IMD, Federal University of Rio Grande do Norte, Natal, RN, Brazil

[3]Department of Biochemistry, CB, Federal University of Rio Grande do Norte, Natal, RN, Brazil

[4]Instituto de Ensino e Pesquisa, Insper, São Paulo, Brazil

[5]Department of Clinical and Toxicological Analyses, School of Pharmaceutical Sciences, University of São Paulo, São Paulo, Brazil

[6]Scientific Platform Pasteur-University of São Paulo, São Paulo, Brazil

[7]Hospital Israelita Albert Einstein, São Paulo, Brazil

[8]Lead contact

*Correspondence: helder.nakaya@einstein.br

https://doi.org/10.1016/j.isci.2021.103610

developed knowledge networks of genes and diseases and monitored the evolution of these relationships yearly. We then quantified and described how genes were connected to each category of disease over this period and how key biological functions unraveled as new genes were added to the network. We also found pairs of diseases from different categories that significantly share genes with each other, indicating underlying clinical proximity between diseases that have not been historically related. Lastly, we explored the genes that were common to all psychiatric, inflammatory, or infectious diseases and investigated which drugs target them. By using a network medicine approach, we were able to extract relevant biological information and new insights of genes, pathways, and therapeutic drugs associated with complex human disorders.

Methods

Watson for drug discovery

We built knowledge networks containing interactions between diseases and genes using the WDD (Chen et al., 2016a). The WDD database contains a corpus of data extracted from the biomedical scientific literature using a cognitive computing text-mining approach (Chen et al., 2016a). WDD has access to millions of abstracts in the MEDLINE platform and full texts in the PMC (PubMed Central) Open Access platform (Chen et al., 2016a). The MEDLINE database is controlled by the National Institutes of Health (NIH) and started in 1960 (NIH, 2021). The PubMed database, for instance, includes all MEDLINE references, which represents over 28 million of the 32 million references in the PubMed database. These references are published in over 5,200 journals (NIH, 2021). The PMC Open Access is a free archive for full-text biomedical and life sciences journal articles. Some PMC journals are also MEDLINE journals, and there are also reciprocal links between the full texts in PMC and corresponding citations in PubMed (NIH, 2021). In each document, the WDD algorithm detects relevant biomedical concepts, namely genes, chemicals, drugs, and diseases. This is performed using a machine learning annotator approach (WDD has a Hatz et al., 2019; High and Bakshi, 2019) dictionary of terms built using the annotator approach based on the thousands of terms extracted from its corpus that reconciles multiple synonyms of a term into a single, unambiguous concept. Next, WDD searches for meaningful associations between the detected unambiguous terms using a set of semantic annotators, i.e. nouns, verbs, and prepositions that convey a semantic relation between the terms. WDD attributes a confidence score (0%–100%) to each association based on the number of documents in which the relation is found and also on the semantic relevance of each link, determined by the machine learning annotator approach (Hatz et al., 2019; High and Bakshi, 2019). The detected terms and relationships that compose the WDD corpus then become available for online or API-mediated searches. The final user performs individual or grouped searches using terms of interest in the form of keywords (e.g. "Alzheimer's disease"). WDD automatically converts the queried keyword into the consensus term present in its dictionary. Terms that are not present in the WDD dictionary will not yield results. The search returns tables containing the connections of the queried term with other terms of interest, the score of each connection, the class to which the connected terms belong (drug, gene, or disease), the number of documents in which the connection could be detected, and the PMIDs or PMCIDs of the documents.

WDD queries

We performed independent searches on WDD using the names of 27 inflammatory diseases, 63 infectious diseases, and 9 psychiatric and neurological disorders (Table S1) as query keywords. All searches were performed in July 2018. WDD allows users to specify a year interval from which relationships will be extracted, so that only documents published in the defined time period are accessed. We defined 29 time intervals beginning in 1964 (the first year of records in WDD) and ending in one year from 1990 to 2018. WDD returned 29 lists of genes related to the human diseases extracted from the scientific literature in each interval. These associations are cumulative, that is, the genes associated with the diseases in 2018 include all the associations present in the previous years. The lists of genes related to human diseases were downloaded in table format and processed using custom R code. From the extracted relationships, we only kept connections between genes and diseases supported by a WDD confidence score of at least 50% and 2 documents of evidence, to reduce false-positive associations, as previously demonstrated by our group (Lüscher Dias et al., 2020). The tables containing the associations retrieved by WDD and the custom R code used to process, filter, and analyze data and to plot figures are available at https://doi.org/10.5281/zenodo.5217544 (Luscher Dias et al., 2021). Figure S1 summarizes all the methodological steps performed in this study.

Evolution of knowledge

We measured the similarity between all pairs of human diseases by calculating a Fisher's exact test for the gene overlap between each pair in each year from 1990 to 2018 (Figure S1). We used the total number of unique genes connected to the diseases in each year's as the Fisher's exact test universe. For each year, a disease-disease knowledge network was developed. In these networks, the nodes are the diseases, and the edges connect diseases that significantly share genes with each other. The edge weights are proportional to the significance of the gene sharing between each pair of diseases according to the Fisher's exact test. We used the $-\log_{10}$p value of the Fisher's exact test (also termed "disease-disease similarity score" here) for each disease pair. We removed edges with a Fisher's exact test p value > 0.05. The networks were constructed using the R package *igraph* (Csardi and Nepusz, 2006) and plotted using the package *ggraph with the kk layout*. We detected new genes each year by comparing the list of genes of the diseases in one year with the list of genes of the same disease in the previous year. Thus, we obtained a list of new genes that were added to the network in each year from 1991 to 2018. The total number of genes associated with each disease was also calculated for each year. Line, violin, and ridge plots were created to illustrate the results using *ggplot2* (Wickham, 2016).

Evolution of disease relationships between categories

We selected the top 9 diseases of each category (psychiatric, inflammatory, infectious) that were connected to the most genes in 2018 ("top 9 diseases"). Then, we detected the top diseases from the other two categories that had the highest disease-disease similarity score with the top 9 diseases (Figure S1). We analyzed how the relationship between these similar pairs of distinct categories evolved from 1990 to 2018. We used the *MeSH.db* R package (Tsuyuzaki et al., 2015) to obtain the MeSH IDs and MeSH terms of all 99 diseases. Using the MeSH terms of the diseases in each pair, we used the *easyPubMed* R package to search in PubMed for papers in which both disease MeSHes of each pair were found together. We then used an adapted version of the *fetch_pubmed_data* function (see code in [Luscher Dias et al., 2021]) of the *easyPubMed* package to retrieve the number of papers that contained the searched MeSH pairs in each year from 1990 to 2018. We used the disease-disease similarity score and the number of papers in 2018 that contained MeSH terms from both diseases to calculate a similarity-to-paper ratio for each disease pair as follows:

$$similarity.paperratio = \frac{dis-dis-similarity}{number\ of\ papers}$$

Low similarity-to-paper ratios were considered as cases of low knowledge gap between the gene sharing and the general scientific interest in the disease pairs. Pairs with low ratios included those in which the diseases did not share a significant amount of genes or pairs of similar diseases for which there is also a proportional number of papers that cite the two diseases together. Intermediated ratio values were considered as cases of intermediate knowledge gap, that is, the diseases in the pair are similar in the genes they share, but the number of papers on the two diseases together is not proportionally high. High similarity-to-paper ratios were interpreted as cases of a large knowledge gap. The pairs that had high ratios include diseases that share a significant proportion of their genes but that have almost never been studied together, evidenced by the very low number of papers including the two MeSH terms.

Evolution of biological pathways

We used the *enricher* function of the R package *clusterProfiler* (Yu et al., 2012) to perform an ORA against Reactome pathways of the genes associated with the top 9 diseases of each category in each year. We selected the significant Reactome pathways (p.adjust <0.01) of the top 9 diseases in 2018 and calculated the significance of the gene overlap between these pathways with Fisher's exact test (Figure S1). We considered only the genes of each significant pathway that were also present in the 2018 gene-disease network. By doing this, we limited pathways to cluster according to the genes shared from our dataset, not all the genes in the pathways. We then built a pathway network connecting the significant Reactome terms using the $-\log_{10}$p value of the Fisher's exact tests as edge weights, similar to what was done for the disease-disease network in Figure 1A. We detected clusters of pathways in this network using the *cluster_louvain* function (Blondel et al., 2008) of the *igraph* R package (Csardi and Nepusz, 2006). Edge weights were considered for the cluster detection. We calculated the weighted degree of each pathway in the network using the *strength* function of the *igraph* package (Csardi and Nepusz, 2006). We manually annotated the detected clusters for their major biological function using the pathways with the highest weighted degree in each cluster as reference. The significance values ($-\log_{10}$pval) of ORA for the pathways

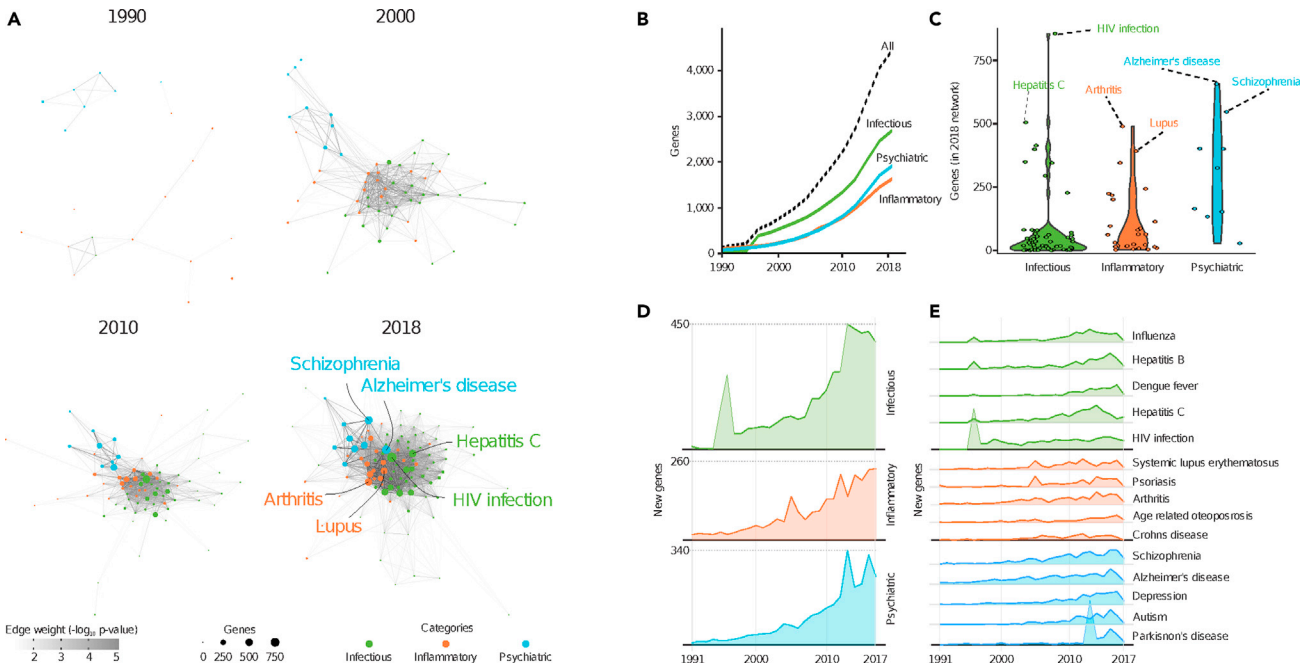


Figure 1. Evolution of knowledge on the molecular bases of human diseases

(A) Representative disease-disease knowledge networks on infectious, inflammatory, and psychiatric disorders in 1990, 2000, 2010, and 2018. All quantitative analyses were performed using the 29 yearly networks from 1990 to 2018. Nodes represent diseases and are proportional to the number of genes associated with each disease in each year. Edge weights are proportional to the significance of gene-sharing between each pair of diseases. Only edges with a p value < 0.01 are depicted.

(B) Cumulative number of genes associated with each disease category and with all diseases from 1990 to 2018.

(C) Distribution of the number of genes associated with each disease and category in 2018.

(D) Number of new genes added to the network in each category per year.

(E) Number of new genes added to the network in selected diseases each year. Color code: green—infectious diseases, orange—inflammatory diseases, and blue—psychiatric disorders.

in each cluster were used to make box and ridge plots to illustrate the results for each disease in 2018 and how these results changed from 1990 to 2018.

Evolution of drug target hub genes

Using the 2018 gene-disease network, we detected the genes common to all three categories of diseases ("hub genes") (Figure S1). We used the R package *UpsetR* to visualize the number of genes shared and exclusive to the disease categories. We downloaded the drug-gene and the drug-disease interaction databases from the CTD (http://ctdbase.org/ [Davis et al., 2021]). We used the MeSH terms of the 99 diseases to filter the drug-disease database and kept only interactions between drugs and diseases that were listed as "therapeutic" by CTD. These are cases of a "chemical that has a known or potential therapeutic role in a disease (e.g., chemical X is used to treat leukemia)," according to the CTD glossary (Davis et al., 2021). We filtered the drug-gene database and kept only the interactions between the therapeutic drugs and the hub genes of our analysis. This final drug-gene list was used to detect the top 20 drugs that target the most hub genes and the top 20 hub genes most targeted by the therapeutic drugs. We visualized these drug-gene interactions in a network built with the R packages *igraph* and plotted with *ggplot2* and *ggraph*. We used the yearly gene-disease networks to detect when the top 20 drug target hub genes were first connected to diseases in each category to build a timeline.

Quantification and statistical analysis

All analyzes were performed in the R environment (Version 4.1.0). R packages used to perform the analyses and plot figures were igraph (v.1.2.6), ggraph (v.2.0.5), ggplot2 (v.3.3.5), clusterProfiler (v.4.0.2), MeSH.db (v.1.15.1), easyPubMed (v.2.13), and UpSetR (v.1.4.0). Specific details of each analysis are described in the Method details section and in the Results section, as well as figure legends. The significance of the

difference between the number of published papers between diseases connected to more or less than 100 genes in the 2018 network was determined with a t test (p value < 0.05). Fisher's exact test result significance was established at p values < 0.01 (network edge filter) or p.adjust <0.01 (Reactome functional enrichment).

RESULTS

Evolution of knowledge on the molecular bases of human diseases

We used WDD, a cognitive computing text-mining application, to identify connections between genes and diseases in millions of peer-reviewed studies (Chen et al., 2016a). For each year from 1990 to 2018, we queried WDD to obtain gene sets related to 99 inflammatory, psychiatric, and infectious diseases (Table S1). WDD detects terms of interest, such as genes and diseases, in scientific texts (e.g., PubMed abstracts and full text journal articles) and finds contextual elements connecting them (e.g., prepositions and verbs). These connections can be extracted from many distinct sources of evidence such as gene expression alterations, genome-wide association studies, or protein expression experiments. A confidence score is established for each relationship based on the strength of the detected semantic association and also the number of documents in which the connection is found. However, the type of study from which the association is obtained is not considered for the calculation of the evidence score. Here, we kept only gene-disease relationships with a confidence score equal or higher than 50% and that were supported by at least 2 studies.

Next, we built yearly disease-disease networks connecting inflammatory, infectious, and psychiatric diseases according to the significance of the genes shared by each pair of diseases (Figure 1A). These networks were cumulative: the 2018 network (Figure 1A, bottom right network) displays all connections found in the entire period, whereas the network of the year 2000 (Figure 1A, top right network), for instance, contains all connections from 1990 up to that year. The 1990 network (Figure 1A, top left network) depicts the relationships between diseases from the beginning of the literature registries up to 1990.

We then assessed how these relationships evolved over the past three decades (1990–2018) and explored the historical trends of the new genes connected to the network during the period (Figures 1B–1E and Table S1). In 1990, only 95 genes were connected in the network (Figure 1B), and no association between psychiatric disorders and inflammatory or infectious diseases could be established through shared genes (Figure 1A). Accordingly, the overall similarity between diseases (between or within categories) was low in 1990 (Figure S2). From 1990 to 2010, with the constant increase in the number of genes associated with diseases in all categories, a preliminary approximation between inflammatory and infectious diseases was observed (Figure 1A, second panel, and Figure S2A). During the next 9 years (2010–2018), the new genes added to the network (Figure 1B) resulted in a strengthening of the connections between infectious and inflammatory diseases and a fast approximation between psychiatric disorders and the other two categories (Figure 1A, fourth panel, and Figure S2A). Meanwhile, the proximity of diseases within the same categories also increased (Figure S2B). Inflammatory diseases occupy a central position in the 2018 network (Figure 1A, fourth panel), which reflects their high between- and within-category similarities sustained throughout the 30-year period (Figure S2). Psychiatric and infectious diseases presented the lowest similarity between each other (Figures 1A and S2).

In 2018, a total of 3,723 genes were present in the network (Figure 1B). The number of genes associated with each disease in the three different categories in 2018 also varied (Figure 1C). The infectious diseases with the highest number of connected genes in 2018 were hepatitis B (414 genes), hepatitis C (506 genes), and HIV infection (856 genes; Figure 1C). However, 55 of 63 infectious diseases were connected to less than 100 genes in 2018 (Figure 1C). The most connected inflammatory diseases were psoriasis (346 genes), systemic lupus erythematosus (393 genes), and arthritis (490 genes; Figure 1C). In the category of psychiatric disorders, Alzheimer disease was the most connected (657 genes), followed by schizophrenia (547 genes) and depression (402 genes; Figure 1B). The imbalance in the distribution of genes connected to infectious diseases likely reflects a bias in the research interest toward the discovery of genes related to diseases already connected to more genes. In fact, we detected a positive correlation between the number of papers published on human diseases and the number of genes connected to the diseases in the 2018 network (Figure S2C).2.

Distinct historical trends of discovery were seen for each disease category (Figures 1D and Table S1). Prominent peaks of gene-association discovery occurred in 1996 for infectious diseases, in 2005 for inflammatory

diseases, and in 2013 for psychiatric disorders (Figure 1C). From 2010 to 2017, the rate of gene discovery in all three categories increased (Figure 1C). The significant increase in the number of genes associated with infectious diseases observed in 1996 was mostly driven by 154 new genes associated with HIV infection (Figure 1D), which corresponded to 50% of the new genes added to the network in that year (Table S1). The triple therapy for HIV using nucleoside reverse-transcriptase inhibitors and protease inhibitors was established in 1996 (Hammer et al., 1996), which likely influenced this outburst of genetic discovery. The 2005 increase in the number of genes associated with inflammatory diseases was mostly related to the new genes connected to psoriasis (41 genes) and systemic lupus erythematosus (33 genes; Figure 1D), which together corresponded to 20% of the new genes associated with all of the diseases in 2005 (Table S1). The Th_{17} cell lineage was discovered in 2005 (Langrish et al., 2005), a cell type that has since been strongly associated with autoimmune and infectious diseases (Zambrano-Zaragoza et al., 2014). In 2013, a large number of new genes were associated with Parkinson disease (165 genes Figure 1D), which corresponded to 17% of the new genes in the network in that year (Table S1). We could not detect any specific scientific landmark in 2013 that could explain this peak. Nevertheless, important genes related to the innate immune response to pathogens and inflammation are among the new genes associated with Parkinson disease in 2013, such as interleukin 1 beta (IL1B) and the p105 subunit of the nuclear factor kappa B (NFKB1).

Evolution of disease relationships between categories

Next, we investigated the evolution of the similarity between diseases from different categories according to their shared genes (see STAR Methods section). For the top 9 most connected diseases of each category in 2018 (i.e., diseases connected to more genes), we detected the diseases from the other two categories with the most significant gene sharing between them and analyzed how these relationships evolved from 1990 to 2018 (Figures 2, S2, S3, and S4). Alzheimer disease was the psychiatric disorder with the highest similarity to inflammatory diseases in 2018, including arthritis and systemic lupus erythematosus (Figure 2A). The relationships between Alzheimer disease and these disorders grew steadily in significance from 1990 to 2018 (Figure S3A), which captures the now well-established relevance of inflammatory processes in the pathophysiology of Alzheimer disease (Newcombe et al., 2018). Surprisingly, fibromyalgia was similar to several psychiatric diseases: depression, anxiety, bipolar disorder, schizophrenia, and Huntington disease (Figures 2A and S3). The total number of genes associated with fibromyalgia in 2018 was low (25 genes), but 72% of these (17 genes) are also associated with depression. These are genes related to nervous system development, such as brain derived neurotrophic factor (BDNF), nerve growth factor (NGF), and neuropeptide Y (NPY), and inflammatory response, including interleukin-6 (IL-6), C-X-C motif chemokine ligand 8 (CXCL8), and tumor necrosis factor (TNF). In fact, fibromyalgia patients often present psychiatric comorbidities such as depression and anxiety (Galvez-Sánchez et al., 2020).

Among infectious diseases, herpes was the most similar disease to autism, schizophrenia, and Huntington disease and was also among the top 5 most similar infectious diseases to depression, Parkinson disease, and Alzheimer disease (Figures 2B and S4). Herpes infection might be associated with the development of Alzheimer disease (Harris and Harris, 2015); the typical amyloid-β deposition that occurs in the brain of Alzheimer disease patients could be an innate immunity mechanism to fight herpes virus infections (Eimer et al., 2018). Our results indicate that there has been latent evidence of that association since the early 2000s in the scientific literature (Figure S4A). In the 2005 network, Alzheimer disease and herpes virus infection shared 14 genes, which represented 58% of the known genes associated with herpes infection at that time.

Autoimmune inflammatory diseases, such as systemic lupus erythematosus, arthritis, and psoriasis, also showed strong gene sharing with viral infections such as hepatitis B and C, respiratory syncytial virus (RSV) infection, influenza, and HIV (Figures 2C and S5). The association between viral infections and autoimmune diseases is well documented (Getts et al., 2013). For instance, the SARS-CoV-2 virus can trigger Guillain–Barré syndrome, a neurological autoimmune disease, in COVID-19 patients (Dalakas, 2020). Dengue patients also present a higher risk of developing autoimmune diseases, such as systemic lupus erythematosus and vasculitis (Li et al., 2018), an association that was also captured in our analysis of the scientific literature since the late 1990s (Figure S5I).

We then examined the number of publications retrieved from PubMed using the topmost similar pairs of diseases from distinct categories as queries (see STAR Methods section; Figure 3). The goal was to find out whether the gene-sharing similarities between diseases from different categories detected in our networks

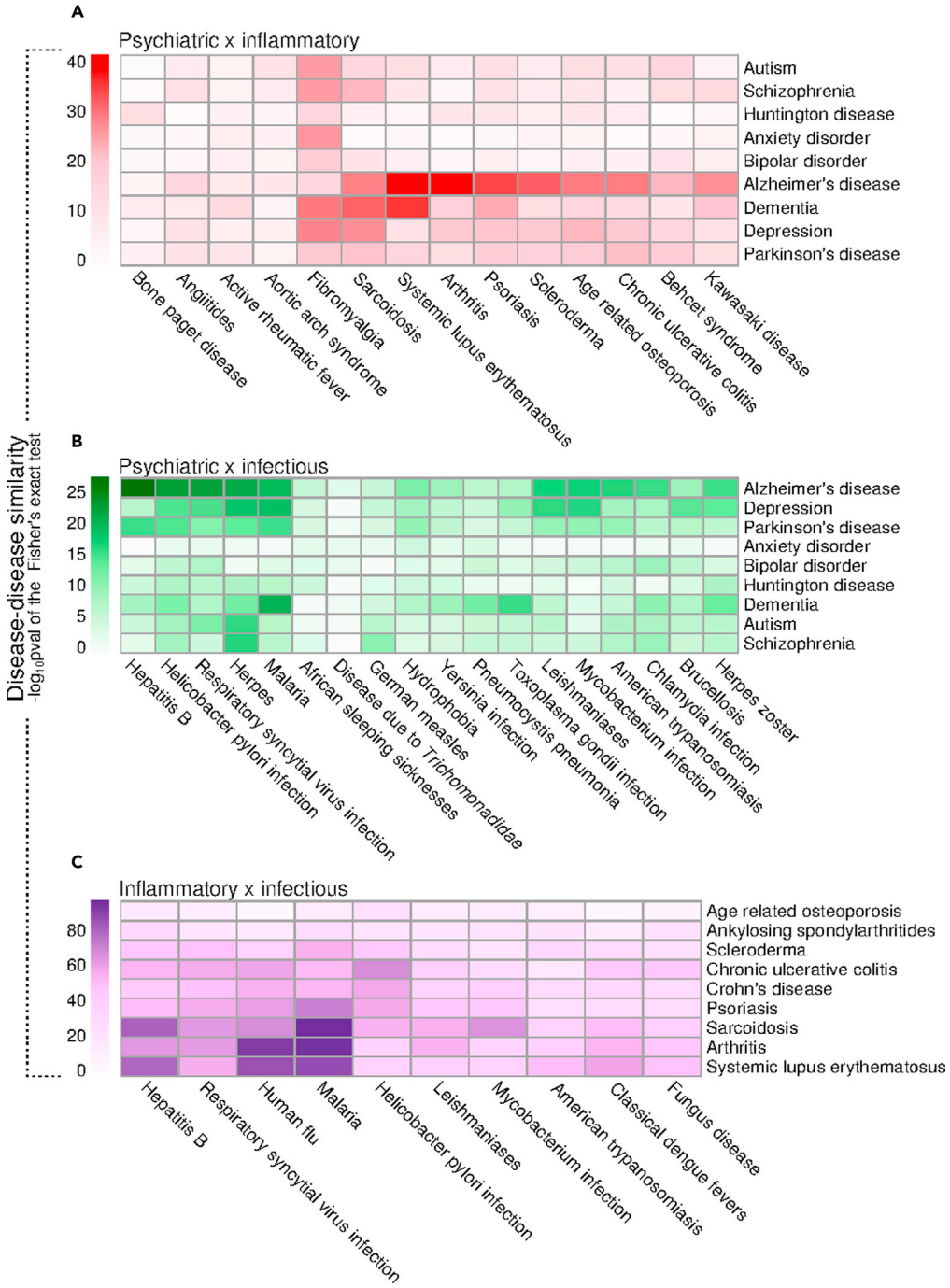


Figure 2. Evolution of disease relationships between categories

Disease-disease similarity between diseases of different categories in the 2018 network according to their shared genes. The similarity score was defined as the $-\log_{10}$pval of the Fisher's exact test result of the gene overlap between each disease pair. Each heatmap represents the similarity score between diseases of two different categories: (A) psychiatric versus inflammatory diseases.

(B) psychiatric versus infectious diseases.

(C) inflammatory versus infectious diseases.

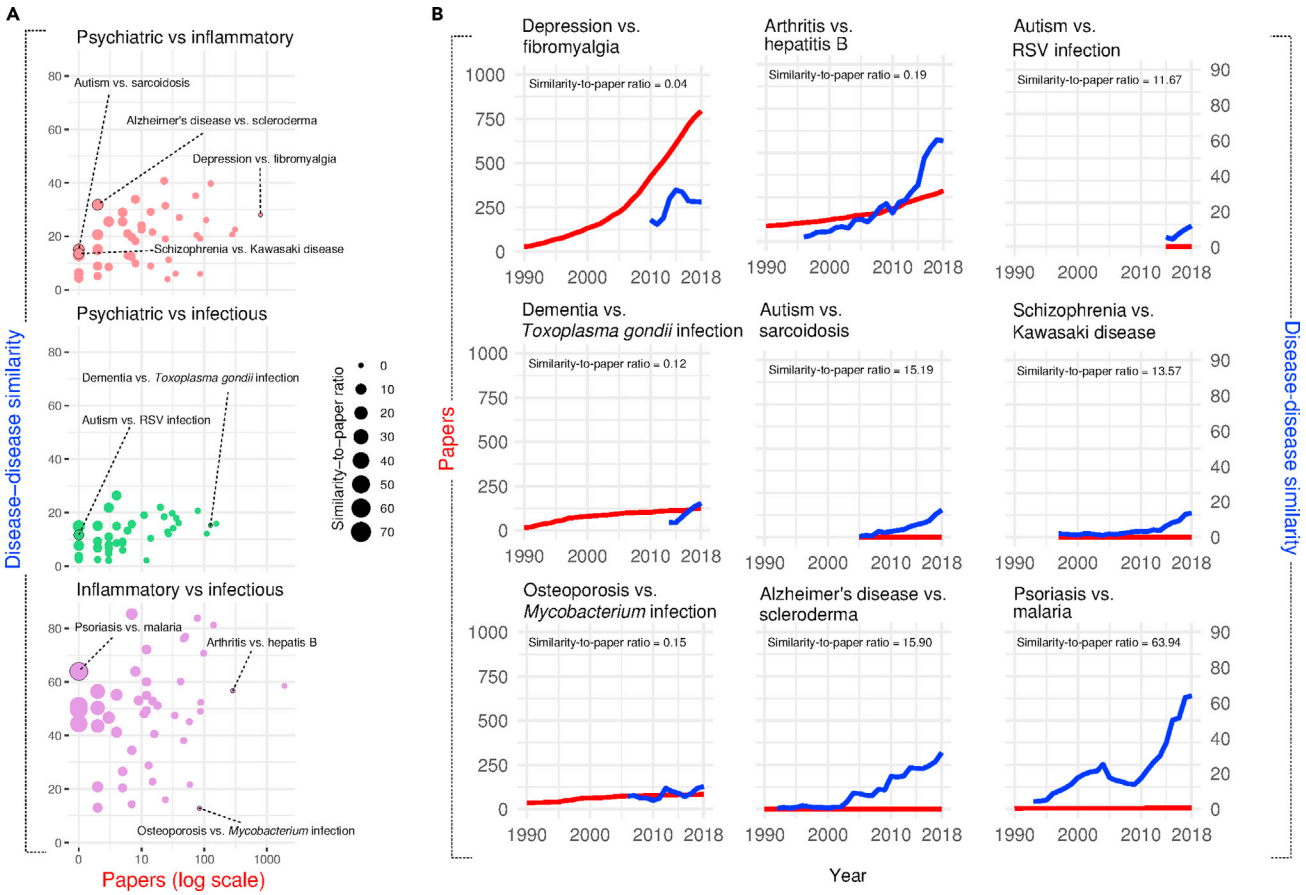


Figure 3. Evolution of the knowledge gap between diseases of different categories

(A) Number of papers versus disease-disease similarity for all disease pairs from distinct categories. Each point represents a disease pair, and the size of the point is proportional to the similarity-to-paper ratio for that pair. This index was obtained as a ratio of the similarity score to the total number of papers published for each disease pair in 2018.

(B) Selected cases of disease pairs with low to high similarity-to-paper ratios depicting the evolution in the number of papers on each pair and the evolution of the similarity between them.

could also be captured from direct co-occurrence in the general peer-reviewed literature over the 30-year period. For each disease pair, we obtained a ratio between the similarity score of the diseases (i.e., the significance of the gene sharing between them) and the total number of studies retrieved from PubMed that mention both diseases of the pairs together (Table S2). This similarity-to-paper ratio was used to detect potentially understudied pairs of diseases that significantly share genes. Low similarity-to-paper ratio values (Figures 3A and 3B, and Table S2) represent similar diseases with many papers already published about them or dissimilar disease pairs. An example of such a pair is fibromyalgia and depression. These diseases have significant gene sharing and also hundreds of scientific papers that explore their relationship in the literature (Figure 3B). Conversely, the genetic association between osteoporosis and mycobacterial infection is low and so is the number of papers that investigate these diseases together (Figure 3B). These cases were considered as examples of a low knowledge gap between the genetic similarity obtained from our network analysis and the established literature coverage of the disease pairs.

Cases with an intermediate similarity-to-paper ratio (Figures 3A and 3B and Table S2) were considered as cases of moderate knowledge gap (Figure 3A), which was the case for arthritis and hepatitis B (Figure 3B). As previously mentioned, several recent studies have explored the association between viral infections and autoimmune diseases (Dalakas, 2020; Getts et al., 2013; Li et al., 2018). In 2018, there were over 250 published papers in which arthritis and hepatitis B were mentioned together (Figure 3B). Virally mediated arthritis represents ~1% of all arthritis cases, including cases related to hepatitis B infection (Marks and Marks, 2016). Scientists have detected the hepatitis B virus in the synovial fluid of rheumatological patients,

which could contribute to the pathogenesis of arthritis (Chen et al., 2018a). Although these diseases are known to be clinically associated at least since the 1970s (Mirise and Kitridou, 1979), our results show that the knowledge on the gene sharing between them increased rapidly after 2015, which was not followed at the same rate by the number of papers published on the two diseases together. This represents a potential gap to be explored by novel research on the genetic bases of the relationship between arthritis and hepatitis B.

Lastly, we considered the disease pairs with strong gene sharing and few studies supporting a direct association as cases of a high knowledge gap (Figures 3A and 3B, and Table S2). We suggest that these cases might represent potentially underexplored fields of research that deserve further investigation. Surprisingly, the number of papers published until 2018 that mentioned psoriasis and malaria together was neglectable (Figure 3B). These diseases share 31 genes, one-third of the genes associated with psoriasis, and over 10% of the genes associated with malaria in the 2018 network. Hydroxychloroquine, a drug used to treat malaria (Ben-Zvi et al., 2012) and rheumatic diseases, such as arthritis and lupus (Ben-Zvi et al., 2012), can trigger psoriatic lesions (Balak and Hajdarbegovic, 2017). Among a few papers in which malaria and psoriasis are mentioned together, there is a report from 2014 that describes cases of hydroxychloroquine-induced psoriasis in patients undergoing malaria treatment (Gravani et al., 2014). The authors of this study suggest that there should be guidelines for the management of psoriasis patients who are also at risk of malaria (Gravani et al., 2014). Our findings corroborate the need for future studies to investigate the association between these diseases.

Evolution of biological pathways

We performed a gene overrepresentation analysis (ORA) against Reactome pathways with the genes associated with the top 9 most connected diseases in each year from 1990 to 2018 (Figures 4–6 and Table S3). We detected 433 Reactome pathways that presented significant enrichment (p.adjust <0.01) among the genes of at least one disease (Table S3). Functional enrichment analysis, such as ORA, often yields too many significant pathways, making these results difficult to interpret at the individual pathway level. For this reason, we used a network approach to reduce the complexity of the obtained set of enriched pathways (see STAR Methods section). Briefly, we built a pathway network (Figure 4) with the significant Reactome pathways obtained from the ORA. We connected these pathways to each other according to the gene sharing between them, similar to what was done in Figure 1A. We then identified 11 clusters of closely connected pathways in the network and annotated these clusters according to the main biological functions of the pathways within them (Figure 4 and Table S3). One of the detected clusters grouped several pathways associated with interferon-stimulated genes, interleukins, and antigen presentation (Figure 4 and Table S3). The pathways in this cluster were significantly enriched among the genes of diseases in all categories, including malaria, HIV infection, arthritis, lupus, depression, and Alzheimer disease (Figure 5). The pathways related to interleukin signaling (e.g., "interleukin 10 signaling"), for instance, were among the top enriched pathways associated with depression genes in the 2018 network (Figure 5 and Table S3). Another cluster of pathways that showed consistent enrichment across all disease categories was NFκB-mediated inflammation induced by Toll-like receptors (TLRs), T cell receptors (TCRs), and B cell receptors (BCRs; Figure 4). These results illustrate the most recurring theme detected in our study: psychiatric, inflammatory, and infectious diseases share common immunological mechanisms that are mostly related to innate immunity and inflammation.

Conversely, we found a cluster of closely connected pathways related to neurotransmission that were enriched mostly among the genes of psychiatric disorders (Figure 4 and Table S3). However, three inflammatory and infectious diseases (hepatitis B, arthritis, and HIV infection) presented enrichment for pathways in this cluster (Figure 5 and Figure S6). The genes related to these diseases presented enrichment for the pathway "transcriptional regulation MECP2," a member of the neurotransmission cluster. Methyl CpG binding protein 2 (MECP2) is located in the X chromosome, and mutations in this gene are the primary cause of Rett syndrome (Liyanage and Rastegar, 2014). There is no evidence in the scientific literature that there is a link between HIV infection or hepatitis B and Rett syndrome, but recent studies indicate a link between this neurodevelopmental disorder and autoimmune diseases, including arthritis (De Felice et al., 2016). Moreover, AIDS patients can develop neurological manifestations similar to those observed in Rett patients, such as cognitive dysfunction and movement disorders (Brew and Garber, 2018). Our results suggest that the similarity between Rett syndrome and autoimmune diseases might also occur for infectious diseases of viral etiology.

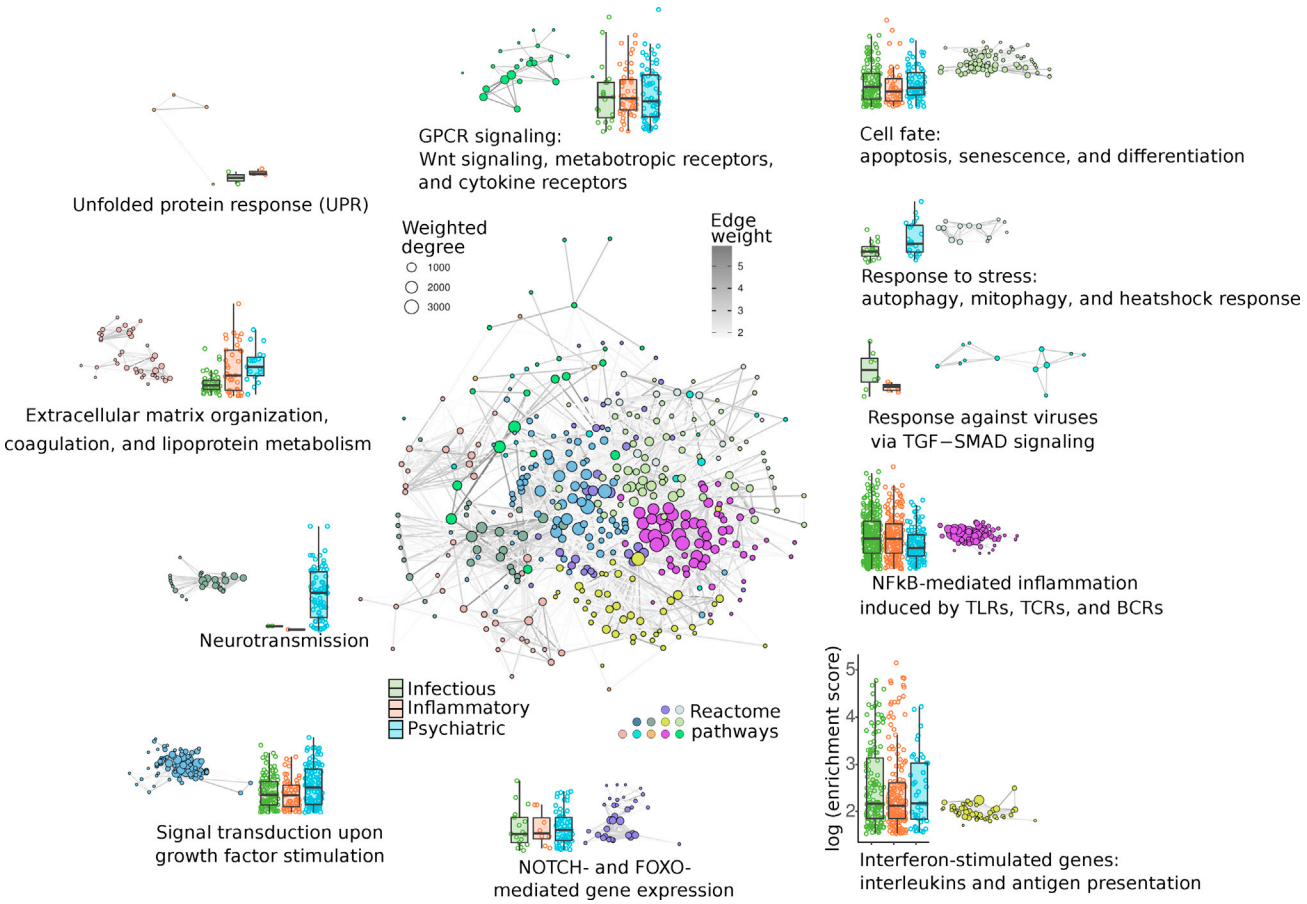


Figure 4. Reactome term network built from the ORA results of the genes associated with human diseases in 2018

Significant Reactome ORA terms (p.adjust <0.01) obtained from the genes of the top 9 diseases in the 2018 network were connected to each other according to the significance of the gene sharing between them (edge weight). Only terms with a gene sharing with a p.adjust <0.01 were connected. We detected 11 clusters (node colors) of closely related terms using the Louvain clustering algorithm in the R package *igraph* (Csardi and Nepusz, 2006) and compared the enrichment score distribution of the terms in these clusters in each disease category (boxplots). Boxplots are colored according to the disease categories: green—infectious diseases, orange—inflammatory diseases, and light blue—psychiatric disorders. Dots in the boxplots represent individual enriched Reactome pathways that belong to each network cluster.

We also detected other clusters of pathways with similar enrichment results between diseases of different categories (Figure 4). The genes related to arthritis and those related to Alzheimer disease presented enrichment for pathways related to the extracellular matrix organization, coagulation, and lipoprotein metabolism (Figure 5). In arthritis, fibroblast-like synoviocytes become hyper-inflammatory and disrupt the extracellular matrix integrity, which leads to the degradation of synovial joint collagen (Nygaard and Firestein, 2020). In Alzheimer disease, some extracellular matrix macromolecules seem to promote the production and stabilization of amyloid β, whereas others act to protect neurons from amyloidosis (Sethi and Zaia, 2017). The pathways in the signal transduction on growth factor stimulation and GPCR-mediated signaling clusters were also enriched among the genes of diseases in all categories (Figures 4 and 5, and S6). This result was expected because the genes involved in signal transduction and intracellular signaling are usually shared between cellular pathways and are involved in virtually all biological functions relevant to diseases (Figure 5 and S6).

After determining the major biological functions related to the genes connected to infectious, inflammatory, and psychiatric diseases in the 2018 network, we investigated how this knowledge evolved from 1990 to 2018 (Figure 6). The pathways related to interferon-stimulated genes, interleukins, and antigen presentation became enriched for the genes associated with inflammatory and infectious diseases already since the early 1990s (Figure 6). Surprisingly, this enrichment appeared earlier for inflammatory

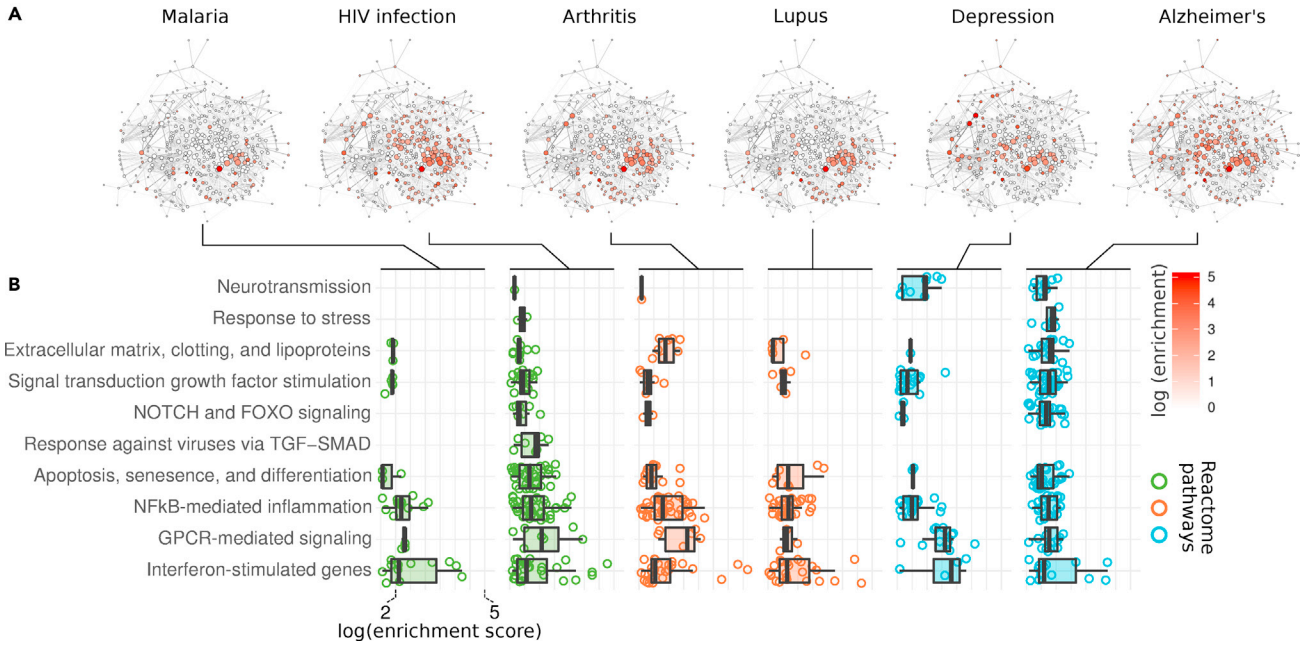


Figure 5. Key biological pathways are enriched among the genes associated with human diseases in 2018

(A) ORA networks depicting the enrichment score of Reactome pathways in selected infectious, inflammatory, and psychiatric disorders. The networks in A have the same topology of the network in Figure 04. The nodes are colored according to the logarithm of enrichment score ($-\log_{10}$pval) of the terms represented by each node.

(B) ORA enrichment score distribution of the terms in the clusters and diseases from panel (A). Boxplots are colored according to the category of each disease: green—infectious, orange—inflammatory, and light blue—psychiatric. Dots in the boxplots represent individual Reactome pathways that belong to the clusters listed in the y axis and that were enriched in each disease.

diseases, despite the highly relevant role of interferon-stimulated genes and antigen presentation in infectious diseases. Conversely, there was a significant increase in the enrichment of these pathways for the genes related to depression, autism, and schizophrenia since 2010 (Figure 6). Recently, the specific roles of the immune system in psychiatric diseases began to be revealed (Chen et al., 2016b; de Baumont et al., 2015; Dong et al., 2018b; Madore et al., 2016; Yuan et al., 2019). Particularly, neuroglial cells have gained importance as key neuroimmune players in the development of autism (microglia and oligodendrocytes [Scuderi and Verkhratsky, 2020], Alzheimer disease (microglia [Clayton et al., 2017], and schizophrenia (astrocytes [Gandal et al., 2018]. The association of pathways related to apoptosis, senescence, and cell differentiation with psychiatric disorders has also occurred recently, except with Alzheimer disease, which began early in the period (Figure 6). Alzheimer, Parkinson, and Huntington diseases are neurodegenerative conditions in which chronic neuronal death happens in distinct parts of the brain (Dugger and Dickson, 2017). We also found an increasing association in recent years of genes related to autism and depression to cell fate pathways (Figure 6), showing that these disorders might also have a neurodegenerative component. In fact, apoptosis and cell death in response to stress and inflammation are relevant factors in the pathogenesis of autism (D. Dong et al., 2018a) and depression (Leonard, 2018).

Evolution of drug target hub genes

Lastly, we examined how drugs that are used to treat inflammatory, infectious, and psychiatric diseases target the genes that are shared between the three categories. We found that 345 genes were common to all disease categories (Figure 7A). Ninety-nine genes were shared only between inflammatory and psychiatric diseases; 259 were common only between psychiatric and infectious diseases; and a total of 409 genes were related exclusively to inflammatory and infectious diseases (Figure 7A). The remaining genes were unique to inflammatory (493 genes), psychiatric (869 genes), and infectious diseases (1,209 genes; Figure 7A).

We used the comparative toxicogenomics database (CTD [Davis et al., 2021]) to find drugs that have a therapeutic relationship with the top 9 diseases and the list of genes that these drugs affect (see STAR Methods

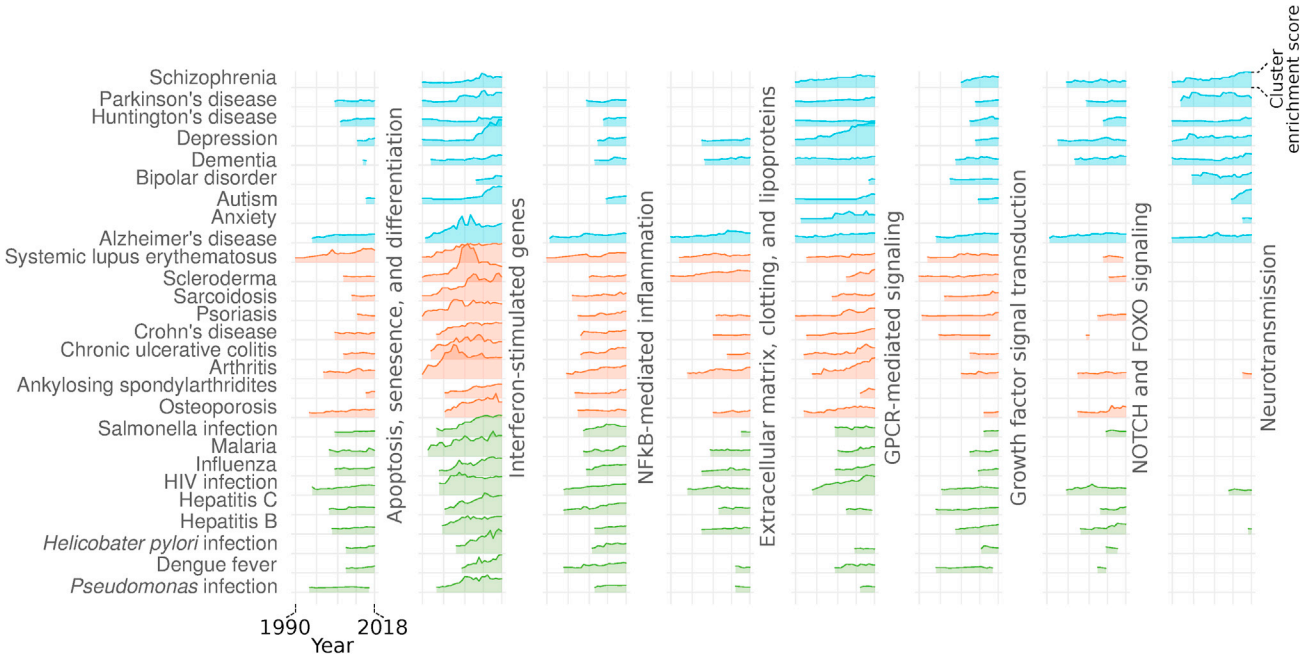


Figure 6. Evolution of knowledge on biological pathways

Ridge plots of the enrichment score of selected clusters from the network in Figure 04 for the top 9 diseases in each category from 1990 to 2018. The height of the ridges are proportional to the mean enrichment score (mean $\log_{10}$pval) of the Reactome pathways in each cluster listed in the y axis.

section). From these lists, we highlight the top 20 most common target genes of the therapeutic drugs listed by CTD (Figure 7B). Among these genes, IL-6, TNF, and interferon gamma (IFNG) were already connected to inflammatory diseases in the 1990 network and were gradually related to diseases in the other two categories until 2002 (Figure 7C). Interleukin-1 beta (IL-1B), B cell lymphoma 2 (BCL2), tumor protein P53 (TP53), and CXCL8 also appeared in our networks in the early 1990s and were first connected to inflammatory diseases (Figure 7C). Eight drug target genes were first connected to psychiatric disorders (Figure 7C): caspase 3 (CASP3; 1996), prostaglandin-endoperoxide synthase 2 (PTGS2; 1997), heme oxygenase 1 (HMOX1; 2000), BCL-2-associated X (BAX), mitogen-activated protein kinase 1 (MAPK1; 2001), RAC-alpha serine/threonine-protein kinase (AKT1; 2003), nuclear factor erythroid 2-related factor 2 (NFE2L2; 2007), and mitogen-activated protein kinase 1 (MAPK3; 2008). The other 5 genes were first connected to infectious diseases (Figure 7C): NFkB P65 subunit (RELA; 1996), poly (ADP-Ribose) polymerase 1 (PARP1), ATP binding cassette subfamily B member 1 (ABCB1; 1999), cyclin-dependent kinase inhibitor 1A (CDKN1A; 2001), and caspase 8 (CASP8; 2010). All top 20 drug target genes were first connected to one of the categories until 2010, with the majority of new connections happening in the 1990s (Figure 7C). These are very well-known genes involved in inflammation (e.g., IL6 and IL1B), innate immunity (e.g., IFNG), apoptosis (e.g., CASP3 and CASP8), cell cycle (e.g., TP53), and other key biological functions that are altered in several diseases.

Next, we found the top 20 therapeutic drugs that affect the most hub genes of inflammatory, psychiatric, and infectious diseases (Figure 7D). Valproic acid, a class I histone deacetylase (HDAC) inhibitor (Göttlicher et al., 2001), was the drug that affected the most hub genes, 259 (Figure 7D). According to CTD, among the diseases we analyzed in this study, valproic acid is a therapeutic drug for anxiety, autism, bipolar disorder, and schizophrenia (Figure 7E). This drug is also an efficient anti-convulsant used to treat epilepsy (Tomson et al., 2016) because it facilitates gamma-aminobutyric acid (GABAergic) neurotransmission (Chateauvieux et al., 2010). There is extensive evidence in the literature of the anti-inflammatory effects of valproic acid and its potential use to treat conditions such as spinal cord injury (Chen et al., 2018b), renal ischemia (Costalonga et al., 2016), and sepsis-induced heart failure (Shi et al., 2019). Valproate was also speculated as a potential repurposing candidate to treat diseases caused by infectious agents, such as COVID-19 (Pitt et al., 2021) and toxoplasmosis (Goodwin et al., 2008). HDAC inhibitors promote epigenetic modifications in the genome that induce the expression of genes in many biological functions and cell types (Hull et al., 2016). This could explain valproic acid's versatility and why it ranked first in our analysis.

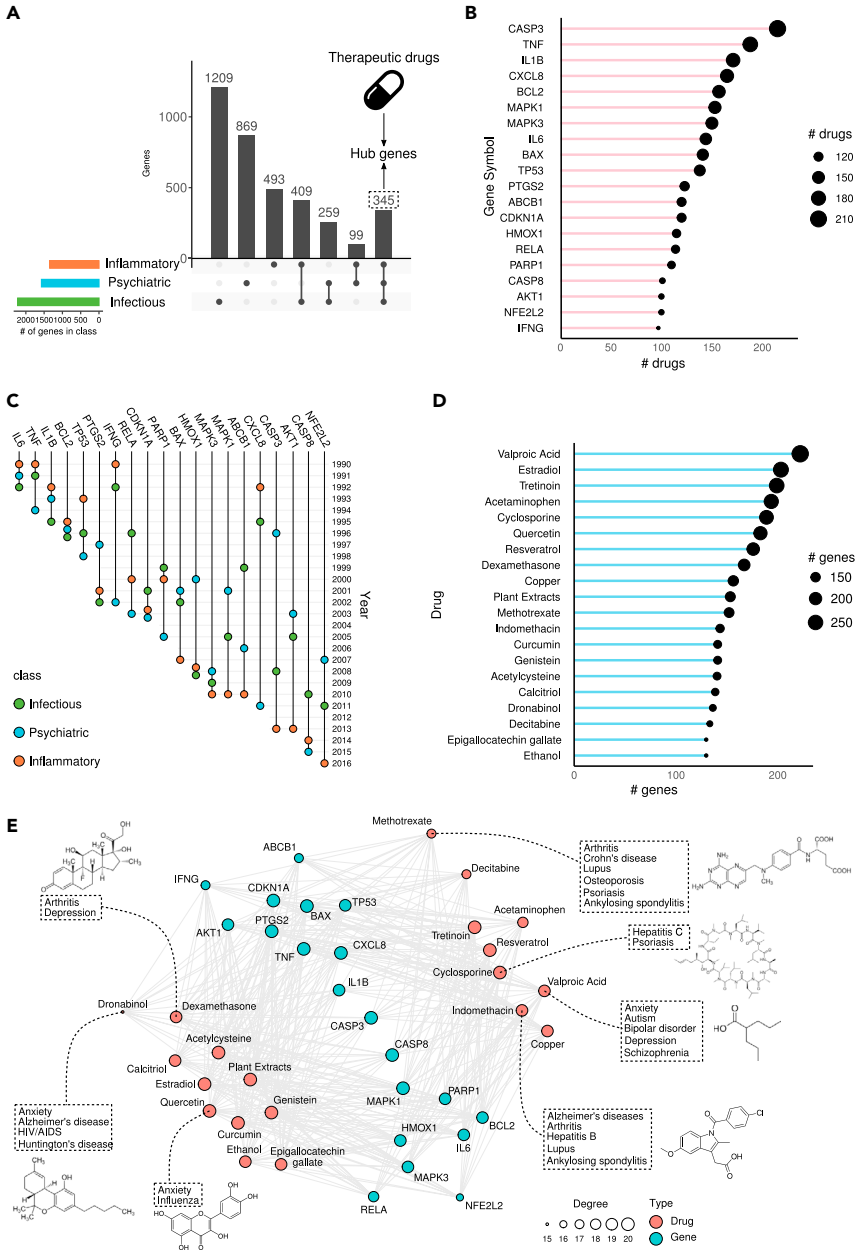


Figure 7. Evolution of drug target hub genes

(A) Upset plot showing the common genes between all categories (hub genes), between two categories exclusively and genes that are unique to each category.

(B) Number of therapeutic drugs of inflammatory, infectious, and psychiatric diseases that target the top 20 target hub genes according to the comparative toxicogenomics database (CTD).

(C) Timeline of the association of the top 20 target hub genes to the gene-disease network. The year in which each gene was associated with the first disease of each category is depicted by the circles with distinct colors for each category.

(D) Number of hub genes targeted by the top 20 drugs that target more hubs according to CTD. (E) Drug-gene network depicting the top 20 drugs and that target hub genes. We selected a few drugs and illustrated their molecular structure and diseases for which they are listed as therapeutic according to CTD.

Among the other top 20 drugs, we found molecules that are currently under investigation for repositioning from one disease category to another. Methotrexate (Figure 7D), which affects 141 genes among the 345 hubs, is used to treat several inflammatory diseases, including psoriasis, lupus, and arthritis (Figure 7E). Recently, a randomized clinical trial revealed a potential for methotrexate to treat positive symptoms in

schizophrenia patients (Chaudhry et al., 2020). The authors of the trial argue that this effect of methotrexate might be achieved through resetting of systemic regulatory T cell control of immune signaling, which is also the way this drug is thought to act in autoimmune diseases (Chaudhry et al., 2020). The use of anti-inflammatory drugs for the treatment of neuropsychiatric diseases gained traction in recent years (Kohler et al., 2016; Ozben and Ozben, 2019; Pandurangi and Buckley, 2020; Rosenblat et al., 2016) influenced by the increasing evidence that these disorders have underlying immune causes, which we have extensively demonstrated in this study. Dexamethasone (Figure 7D) is a glucocorticoid anti-inflammatory drug listed in CTD as a therapy for arthritis and depression (Figure 7E), but it is also used to treat several other inflammatory disorders. Indeed, dexamethasone was one of the few drugs submitted to randomized clinical trials that reduced mortality in COVID-19 patients subjected to invasive ventilation (RECOVERY Collaborative Group et al., 2021). Several of the other top 20 drugs were also listed in CTD to be used as therapy for diseases of different categories, such as cyclosporine (hepatitis C and psoriasis), indomethacin (Alzheimer and autoimmune diseases), dronabinol (neuropsychiatric diseases and HIV infection), and quercetin (anxiety and influenza; Figure 7E).

DISCUSSION

Similar to the exponential increase in the number of published papers seen in the past decades (Fortunato et al., 2018), the number of genes associated with psychiatric, inflammatory, and infectious diseases have also increased significantly in the past 30 years. This rapid growth in knowledge about the genetic underpinnings of these diseases can be directly attributed to at least two historical landmarks: the publication of the human genome in 2001 (International Human Genome Sequencing Consortium et al., 2001; Venter et al., 2001) and the advent of high-throughput DNA-sequencing technologies (Margulies et al., 2005). Discrete advances in genes associated with specific diseases could also be spotted throughout the period analyzed here. In 1996, the triple therapy for HIV was developed using nucleoside reverse-transcriptase inhibitors and protease inhibitors (Hammer et al., 1996). In the same year, 50% of the new genes added to the knowledge network were connected to HIV infection. In 2005, a peak of novel genes associated with psoriasis and systemic lupus erythematosus was detected. This year also saw the discovery of the Th_{17} cell lineage (Langrish et al., 2005). The central role of these pro-inflammatory cells in the pathogenesis of autoimmune and infectious diseases was later identified (Zambrano-Zaragoza et al., 2014). Indeed, the key genes of the differentiation and maintenance of the Th_{17} phenotype in $CD4^+$ T lymphocytes, such as interleukin 17F (IL17F), interleukin-21 (IL21), the peroxisome proliferator-activated receptor gamma (PPARG), and the fatty acid-binding protein 5 (FABP5), were connected to psoriasis and systemic lupus erythematosus in the network in 2005 (Hwang, 2010; Nalbant and Eskier, 2016).

One of the advantages of using text mining and network medicine to study the relationships between genes and diseases is the possibility of detecting novel connections from established scientific knowledge. When two diseases share a genetic mechanism, they can also present common clinical or epidemiological characteristics, despite having distinct etiological backgrounds (Barabási et al., 2011). These similarities can inform researchers of potential treatment options (Lüscher Dias et al., 2020). Here, we showed that diseases from inflammatory, psychiatric, and infectious etiologies significantly share genes with each other. This sharing was strong between disease pairs that were well studied together, such as depression and fibromyalgia. Conversely, the gene sharing between psoriasis and malaria could be perceived in our knowledge networks since the 2000s, but the number of papers featuring the two conditions together in PubMed is virtually null. We detected a few such cases, mostly involving neglected infectious diseases, which could explain the knowledge gap. We also found cases of diseases that just recently began to share genes that also lack many publications directly connecting them in the literature. A case in point is autism and RSV. We also found disease pairs, such as dementia and *Toxoplasma gondii* infection, for which there have been direct associations in the literature since 1990 but that just recently started to share genes in the network. Our results reveal potentially underexplored pathways for future research on the association between diseases of distinct categories and also for the discovery of new genes related to well-studied disease pairs.

The sharing of genes between diseases from distinct categories also reflects in the overlap of biological functions, particularly those related to immunological processes. The genes of several diseases in all categories presented enrichment for Reactome pathways related to the interferon response, cytokines, and NFkB-mediated inflammation. This pattern was detectable in our networks since the early 1990s for inflammatory diseases and gradually appeared for infectious and psychiatric diseases as well. Pathways associated with neurotransmission were almost exclusively enriched among the genes of psychiatric diseases.

Nevertheless, we found enrichment for a neurotransmission-related pathway, "transcriptional regulation by MECP2," among the genes of HIV infection and hepatitis B that could point to a connection between these disorders and Rett syndrome, a neurological condition. Our functional enrichment results also highlighted the relevance of core cellular functions in diseases of all categories, such as signal transduction and the regulation of gene expression by transcription factors.

Our network medicine text mining approach also revealed how shared genes between disease categories can signal toward common therapeutic solutions. The findings presented in the last section of our study emphasize the relevance of drugs that target shared genes for the treatment of distinct diseases. Our results show that the genes targeted by therapeutic drugs shared by inflammatory, psychiatric, and infectious diseases have been associated with these disorders early in the past 30 years of scientific research. These genes are associated with inflammation, the cell cycle, apoptosis, and central pathways of cellular function. We also demonstrated that well-established and promising cases of repositioning involve drugs that target shared genes between diseases. Future studies should aim to reveal more common molecular mechanisms between these categories of diseases as well as to harness that knowledge for novel drug discovery and repurposing.

In summary, we could apply a machine learning and cognitive computing text-mining strategy using WDD to extract knowledge about genes related to inflammatory, infectious, and psychiatric diseases from the scientific literature and depict how this knowledge evolved during the past 30 years.

Limitations of the study

Previous work from our group (Lüscher Dias et al., 2020) revealed that WDD occasionally includes false gene-disease associations due to misleading or ambiguous sentences in the source documents. We made an effort to prevent those mistakes here by restricting associations supported by at least two documents and with a WDD confidence score higher than 50%. However, because the number of associations detected in this work was too high, we were not able to manually curate them to guarantee that every connection was supported by the documents used by WDD. Moreover, the gene-disease associations obtained with WDD were retrieved from diverse types of scientific documents, including low-throughput single-gene studies, omics, and genome-wide association studies. Therefore, the nature of the association of a gene with a given disease in our results might differ from the nature of the association of the same gene with other diseases. We treated all associations between genes and diseases equally, regardless of their nature, so the interpretation of the results reported here must take this into consideration. Lastly, as has been discussed previously in this manuscript, the number of genes associated with each disease is likely biased toward more studied diseases. Therefore, the results presented here might significantly change, as new genes are associated with the analyzed diseases, especially those that are less well studied.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Watson for drug discovery
  - WDD queries
  - Evolution of knowledge
  - Evolution of disease relationships between categories
  - Evolution of biological pathways
  - Evolution of drug target hub genes
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.103610.

ACKNOWLEDGMENTS

This work was supported by Brazilian National Council for Scientific and Technological Development (grant numbers 313662/2017-7); the São Paulo Research Foundation (grant numbers 2018/14933-2 and 2018/21934-5).

AUTHOR CONTRIBUTIONS

Conceptualization, Investigation: TLD, RJSD, VS, GRF, and HIN. Software Programming, Formal analysis, and Data Curation: TLD, VS, and TLA. Visualization: TLD. Resources: TLD and TLA. Writing—Original Manuscript: TLD, HIN. Writing—Review & Editing: TLD, RJSD, PPA, VS, GRF, and HIN; Supervision and Funding acquisition: GRF and HIN.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 20, 2021
Revised: November 5, 2021
Accepted: December 8, 2021
Published: January 21, 2022

REFERENCES

Bai, T., Gong, L., Wang, Y., Wang, Y., Kulikowski, C.A., and Huang, L. (2016). A method for exploring implicit concept relatedness in biomedical knowledge network. BMC Bioinformatics *17*, 265. https://doi.org/10.1186/s12859-016-1131-5.

Balak, D.M., and Hajdarbegovic, E. (2017). Drug-induced psoriasis: Clinical perspectives. Psoriasis (Auckl) *7*, 87–94. https://doi.org/10.2147/PTT.S126727.

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: A network-based approach to human disease. Nat. Rev. Genet. *12*, 56–68. https://doi.org/10.1038/nrg2918.

de Baumont, A., Maschietto, M., Lima, L., Carraro, D.M., Olivieri, E.H., Fiorini, A., Barreta, L.A.N., Palha, J.A., Belmonte-de-Abreu, P., Moreira Filho, C.A., and Brentani, H. (2015). Innate immune response is differentially dysregulated between bipolar disease and schizophrenia. Schizophr. Res. *161*, 215–221. https://doi.org/10.1016/j.schres.2014.10.055.

Ben-Zvi, I., Kivity, S., Langevitz, P., and Shoenfeld, Y. (2012). Hydroxychloroquine: From malaria to autoimmunity. Clin. Rev. Allergy Immunol. *42*, 145–153. https://doi.org/10.1007/s12016-010-8243-x.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. J. Stat. Mech. *2008*, P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008.

Brew, B.J., and Garber, J.Y. (2018). Neurologic sequelae of primary HIV infection. Handb. Clin. Neurol. *152*, 65–74. https://doi.org/10.1016/B978-0-444-63849-6.00006-2.

Brooks, P.J., Tagle, D.A., and Groft, S. (2014). Expanding rare disease drug trials based on shared molecular etiology. Nat. Biotechnol. *32*, 515–518. https://doi.org/10.1038/nbt.2924.

Carson, M.B., Liu, C., Lu, Y., Jia, C., and Lu, H. (2017). A disease similarity matrix based on the uniqueness of shared genes. BMC Med. Genomics *10*, 26. https://doi.org/10.1186/s12920-017-0265-2.

Chateauvieux, S., Morceau, F., Dicato, M., and Diederich, M. (2010). Molecular and therapeutic potential and toxicity of valproic acid. J. Biomed. Biotechnol. *2010*. https://doi.org/10.1155/2010/479364.

Chaudhry, I.B., Husain, M.O., Khoso, A.B., Husain, M.I., Buch, M.H., Kiran, T., Fu, B., Bassett, P., Qurashi, I., Ur Rahman, R., et al. (2020). A randomised clinical trial of methotrexate points to possible efficacy and adaptive immune dysfunction in psychosis. Transl. Psychiatry *10*, 415. https://doi.org/10.1038/s41398-020-01095-8.

Chen, Y., Elenee Argentinis, J.D., and Weber, G. (2016a). IBM watson: How cognitive computing can be applied to big data challenges in life sciences research. Clin. Ther. *38*, 688–701. https://doi.org/10.1016/j.clinthera.2015.12.001.

Chen, H., Liu, S., Ji, L., Wu, T., Ji, Y., Zhou, Y., Zheng, M., Zhang, M., Xu, W., and Huang, G. (2016b). Folic acid supplementation mitigates alzheimer's disease by reducing inflammation: A randomized controlled trial. Mediators Inflamm. *2016*, 5912146. https://doi.org/10.1155/2016/5912146.

Chen, Y.-L., Jing, J., Mo, Y.-Q., Ma, J.-D., Yang, L.-J., Chen, L.-F., Zhang, X., Yan, T., Zheng, D.-H., Pessler, F., and Dai, L. (2018a). Presence of hepatitis B virus in synovium and its clinical significance in rheumatoid arthritis. Arthritis Res. Ther. *20*, 130. https://doi.org/10.1186/s13075-018-1623-y.

Chen, S., Ye, J., Chen, X., Shi, J., Wu, W., Lin, W., Lin, W., Li, Y., Fu, H., and Li, S. (2018b). Valproic acid attenuates traumatic spinal cord injury-induced inflammation via STAT1 and NF-κB pathway dependent of HDAC3. J. Neuroinflammation *15*, 150. https://doi.org/10.1186/s12974-018-1193-6.

Clayton, K.A., Van Enoo, A.A., and Ikezu, T. (2017). Alzheimer's disease: The role of microglia in brain homeostasis and proteopathy. Front. Neurosci. *11*, 680. https://doi.org/10.3389/fnins.2017.00680.

Costalonga, E.C., Silva, F.M.O., and Noronha, I.L. (2016). Valproic acid prevents renal dysfunction and inflammation in the ischemia-reperfusion injury model. Biomed. Res. Int. *2016*, 5985903. https://doi.org/10.1155/2016/5985903.

Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. Int. J. Complex Syst. *1695*, 1–9.

Dalakas, M.C. (2020). Guillain-Barré syndrome: The first documented COVID-19-triggered autoimmune neurologic disease: More to come with myositis in the offing. Neurol. Neuroimmunol. Neuroinflamm. *7*. https://doi.org/10.1212/NXI.0000000000000781.

Davis, A.P., Grondin, C.J., Johnson, R.J., Sciaky, D., Wiegers, J., Wiegers, T.C., and Mattingly, C.J. (2021). Comparative toxicogenomics database (CTD): update 2021. Nucleic Acids Res. *49*, D1138–D1143. https://doi.org/10.1093/nar/gkaa891.

Dong, D., Zielke, H.R., Yeh, D., and Yang, P. (2018a). Cellular stress and apoptosis contribute to the pathogenesis of autism spectrum disorder. Autism Res. *11*, 1076–1090. https://doi.org/10.1002/aur.1966.

Dong, Y., Lagarde, J., Xicota, L., Corne, H., Chantran, Y., Chaigneau, T., Crestani, B., Bottlaender, M., Potier, M.-C., Aucouturier, P., et al. (2018b). Neutrophil hyperactivation correlates with Alzheimer's disease progression. Ann. Neurol. *83*, 387–405. https://doi.org/10.1002/ana.25159.

Dugger, B.N., and Dickson, D.W. (2017). Pathology of neurodegenerative diseases. Cold Spring Harb. Perspect. Biol. *9*. https://doi.org/10.1101/cshperspect.a028035.

Eimer, W.A., Vijaya Kumar, D.K., Navalpur Shanmugam, N.K., Rodriguez, A.S., Mitchell, T., Washicosky, K.J., György, B., Breakefield, X.O., Tanzi, R.E., and Moir, R.D. (2018). Alzheimer's disease-associated β-amyloid is rapidly seeded by herpesviridae to protect against brain infection. Neuron *100*, 1527–1532. https://doi.org/10.1016/j.neuron.2018.11.043.

De Felice, C., Leoncini, S., Signorini, C., Cortelazzo, A., Rovero, P., Durand, T., Ciccoli, L., Papini, A.M., and Hayek, J. (2016). Rett syndrome: An autoimmune disease? Autoimmun. Rev. *15*, 411–416. https://doi.org/10.1016/j.autrev.2016.01.011.

Fortunato, S., Bergstrom, C.T., Börner, K., Evans, J.A., Helbing, D., Milojević, S., Petersen, A.M., Radicchi, F., Sinatra, R., Uzzi, B., et al. (2018). Science of science. Science *359*, 6397. https://doi.org/10.1126/science.aao0185.

Galvez-Sánchez, C.M., Montoro, C.I., Duschek, S., and Reyes Del Paso, G.A. (2020). Depression and trait-anxiety mediate the influence of clinical pain on health-related quality of life in fibromyalgia. J. Affect. Disord. *265*, 486–495. https://doi.org/10.1016/j.jad.2020.01.129.

Gandal, M.J., Zhang, P., Hadjimichael, E., Walker, R.L., Chen, C., Liu, S., Won, H., van Bakel, H., Varghese, M., Wang, Y., et al. (2018). Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. Science *362*. https://doi.org/10.1126/science.aat8127.

Getts, D.R., Chastain, E.M.L., Terry, R.L., and Miller, S.D. (2013). Virus infection, antiviral immunity, and autoimmunity. Immunol. Rev. *255*, 197–209. https://doi.org/10.1111/imr.12091.

Gibney, S.M., and Drexhage, H.A. (2013). Evidence for a dysregulated immune system in the etiology of psychiatric disorders. J. Neuroimmune Pharmacol. *8*, 900–920. https://doi.org/10.1007/s11481-013-9462-8.

Goodwin, D.G., Strobl, J., Mitchell, S.M., Zajac, A.M., and Lindsay, D.S. (2008). Evaluation of the mood-stabilizing agent valproic acid as a preventative for toxoplasmosis in mice and activity against tissue cysts in mice. J. Parasitol. *94*, 555–557. https://doi.org/10.1645/GE-1331.1.

Göttlicher, M., Minucci, S., Zhu, P., Krämer, O.H., Schimpf, A., Giavara, S., Sleeman, J.P., Lo Coco, F., Nervi, C., Pelicci, P.G., and Heinzel, T. (2001). Valproic acid defines a novel class of HDAC inhibitors inducing differentiation of transformed cells. EMBO J. *20*, 6969–6978. https://doi.org/10.1093/emboj/20.24.6969.

Gravani, A., Gaitanis, G., Zioga, A., and Bassukas, I.D. (2014). Synthetic antimalarial drugs and the triggering of psoriasis-do we need disease-specific guidelines for the management of patients with psoriasis at risk of malaria? Int. J. Dermatol. *53*, 327–330. https://doi.org/10.1111/ijd.12231.

Hammer, S.M., Katzenstein, D.A., Hughes, M.D., Gundacker, H., Schooley, R.T., Haubrich, R.H., Henry, W.K., Lederman, M.M., Phair, J.P., Niu, M., et al. (1996). A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. AIDS clinical trials group study 175 study team. N. Engl. J. Med. *335*, 1081–1090. https://doi.org/10.1056/NEJM199610103351501.

Harris, S.A., and Harris, E.A. (2015). Herpes simplex virus type 1 and other pathogens are key causative factors in sporadic alzheimer's disease. J. Alzheimers Dis. *48*, 319–353. https://doi.org/10.3233/JAD-142853.

Hatz, S., Spangler, S., Bender, A., Studham, M., Haselmayer, P., Lacoste, A.M.B., Willis, V.C., Martin, R.L., Gurulingappa, H., and Betz, U. (2019). Identification of pharmacodynamic biomarker hypotheses through literature analysis with IBM Watson. PLoS One *14*, e0214619. https://doi.org/10.1371/journal.pone.0214619.

High, R., and Bakshi, T. (2019). Cognitive Computing with IBM Watson: Build Smart Applications Using Artificial Intelligence as a Service (Packt Publishing), pp. 1–256.

RECOVERY Collaborative Group, Horby, P., Lim, W.S., Emberson, J.R., Mafham, M., Bell, J.L., Linsell, L., Staplin, N., Brightling, C., Ustianowski, A., Elmahi, E., et al. (2021). Dexamethasone in hospitalized patients with Covid-19. N. Engl. J. Med. *384*, 693–704. https://doi.org/10.1056/NEJMoa2021436.

Hull, E.E., Montgomery, M.R., and Leyva, K.J. (2016). HDAC inhibitors as epigenetic regulators of the immune system: Impacts on cancer therapy and inflammatory diseases. Biomed. Res. Int. *2016*, 8797206. https://doi.org/10.1155/2016/8797206.

Hwang, E.S. (2010). Transcriptional regulation of T helper 17 cell differentiation. Yonsei Med. J. *51*, 484–491. https://doi.org/10.3349/ymj.2010.51.4.484.

Kohler, O., Krogh, J., Mors, O., and Benros, M.E. (2016). Inflammation in depression and the potential for anti-inflammatory treatment. Curr. Neuropharmacol. *14*, 732–742. https://doi.org/10.2174/1570159x14666151208113700.

International Human Genome Sequencing Consortium, Whitehead Institute for Biomedical Research, Center for Genome Research, Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921. https://doi.org/10.1038/35057062.

Langrish, C.L., Chen, Y., Blumenschein, W.M., Mattson, J., Basham, B., Sedgwick, J.D., McClanahan, T., Kastelein, R.A., and Cua, D.J. (2005). IL-23 drives a pathogenic T cell population that induces autoimmune inflammation. J. Exp. Med. *201*, 233–240. https://doi.org/10.1084/jem.20041257.

Lees, C.W., Barrett, J.C., Parkes, M., and Satsangi, J. (2011). New IBD genetics: Common pathways with other diseases. Gut *60*, 1739–1753. https://doi.org/10.1136/gut.2009.199679.

Leonard, B.E. (2018). Inflammation and depression: A causal or coincidental link to the pathophysiology? Acta Neuropsychiatr. *30*, 1–16. https://doi.org/10.1017/neu.2016.69.

Li, H.-M., Huang, Y.-K., Su, Y.-C., and Kao, C.-H. (2018). Increased risk of autoimmune diseases in dengue patients: A population-based cohort study. J. Infect. *77*, 212–219. https://doi.org/10.1016/j.jinf.2018.03.014.

Littmann, M., Selig, K., Cohen-Lavi, L., Frank, Y., Hönigschmid, P., Kataka, E., Mösch, A., Qian, K., Ron, A., Schmid, S., et al. (2020). Validity of machine learning in biology and medicine increased through collaborations across fields of expertise. Nat. Mach. Intell. https://doi.org/10.1038/s42256-019-0139-8.

Liyanage, V.R.B., and Rastegar, M. (2014). Rett syndrome and MeCP2. Neuromolecular Med. *16*, 231–264. https://doi.org/10.1007/s12017-014-8295-9.

Lüscher Dias, T., Schuch, V., Beltrão-Braga, P.C.B., Martins-de-Souza, D., Brentani, H.P., Franco, G.R., and Nakaya, H.I. (2020). Drug repositioning for psychiatric and neurological disorders through a network medicine approach. Transl. Psychiatry *10*, 141. https://doi.org/10.1038/s41398-020-0827-5.

Luscher Dias, T., Juliani Siqueira Dalmolin, R., de Paiva Amaral, P., Lubiana Alves, T., Schuch, V., Franco, G.R., and Imoto Nakaya, H. (2021). csbl-usp/evolution_of_knowledge: First release of the code for the paper "The evolution of knowledge on genes associated with human diseases". Zenodo. https://doi.org/10.5281/zenodo.5217544.

Madore, C., Leyrolle, Q., Lacabanne, C., Benmamar-Badel, A., Joffre, C., Nadjar, A., and Layé, S. (2016). Neuroinflammation in autism: Plausible role of maternal inflammation, dietary omega 3, and microbiota. Neural Plast. *2016*, 3597209. https://doi.org/10.1155/2016/3597209.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature *437*, 376–380. https://doi.org/10.1038/nature03959.

Marks, M., and Marks, J.L. (2016). Viral arthritis. Clin. Med. *16*, 129–134. https://doi.org/10.7861/clinmedicine.16-2-129.

Marrie, R.A., Walld, R., Bolton, J.M., Sareen, J., Walker, J.R., Patten, S.B., Singer, A., Lix, L.M., Hitchon, C.A., El-Gabalawy, R., et al.; CIHR team in defining the burden and managing the effects of psychiatric comorbidity in chronic immunoinflammatory disease (2017). Increased incidence of psychiatric disorders in immune-mediated inflammatory disease. J. Psychosom. Res. *101*, 17–23. https://doi.org/10.1016/j.jpsychores.2017.07.015.

Mirise, R.T., and Kitridou, R.C. (1979). Arthritis and hepatitis. West. J. Med. *130*, 12–17.

Nalbant, A., and Eskier, D. (2016). Genes associated with T helper 17 cell differentiation and function. Front Biosci. (Elite Ed.) *8*, 427–435. https://doi.org/10.2741/e777.

Newcombe, E.A., Camats-Perna, J., Silva, M.L., Valmas, N., Huat, T.J., and Medeiros, R. (2018). Inflammation: The link between comorbidities, genetics, and Alzheimer's disease. J. Neuroinflammation *15*, 276. https://doi.org/10.1186/s12974-018-1313-3.

Nygaard, G., and Firestein, G.S. (2020). Restoring synovial homeostasis in rheumatoid arthritis by targeting fibroblast-like synoviocytes. Nat. Rev. Rheumatol. *16*, 316–333. https://doi.org/10.1038/s41584-020-0413-5.

Ozben, T., and Ozben, S. (2019). Neuro-inflammation and anti-inflammatory treatment options for Alzheimer's disease. Clin. Biochem. *72*, 87–89. https://doi.org/10.1016/j.clinbiochem.2019.04.001.

Pandurangi, A.K., and Buckley, P.F. (2020). Inflammation, antipsychotic drugs, and evidence for effectiveness of anti-inflammatory agents in schizophrenia. Curr. Top. Behav. Neurosci. *44*, 227–244. https://doi.org/10.1007/7854_2019_91.

Pitt, B., Sutton, N.R., Wang, Z., Goonewardena, S.N., and Holinstat, M. (2021). Potential repurposing of the HDAC inhibitor valproic acid for patients with COVID-19. Eur. J. Pharmacol. *898*, 173988. https://doi.org/10.1016/j.ejphar.2021.173988.

Postma, D.S., Kerkhof, M., Boezen, H.M., and Koppelman, G.H. (2011). Asthma and chronic obstructive pulmonary disease: common genes, common environments? Am. J. Respir. Crit. Care Med. *183*, 1588–1594. https://doi.org/10.1164/rccm.201011-1796PP.

Rajkomar, A., Oren, E., Chen, K., Dai, A.M., Hajaj, N., Hardt, M., Liu, P.J., Liu, X., Marcus, J., Sun, M., et al. (2018). Scalable and accurate deep learning with electronic health records. npj Digital Med. *1*, 18. https://doi.org/10.1038/s41746-018-0029-1.

Rosenblat, J.D., Kakar, R., Berk, M., Kessing, L.V., Vinberg, M., Baune, B.T., Mansur, R.B., Brietzke, E., Goldstein, B.I., and McIntyre, R.S. (2016). Anti-inflammatory agents in the treatment of bipolar depression: A systematic review and meta-analysis. Bipolar Disord. *18*, 89–101. https://doi.org/10.1111/bdi.12373.

Scuderi, C., and Verkhratsky, A. (2020). The role of neuroglia in autism spectrum disorders. Prog. Mol. Biol. Transl. Sci. *173*, 301–330. https://doi.org/10.1016/bs.pmbts.2020.04.011.

Sethi, M.K., and Zaia, J. (2017). Extracellular matrix proteomics in schizophrenia and Alzheimer's disease. Anal. Bioanal. Chem. *409*, 379–394. https://doi.org/10.1007/s00216-016-9900-6.

Shi, X., Liu, Y., Zhang, D., and Xiao, D. (2019). Valproic acid attenuates sepsis-induced myocardial dysfunction in rats by accelerating autophagy through the PTEN/AKT/mTOR pathway. Life Sci. *232*, 116613. https://doi.org/10.1016/j.lfs.2019.116613.

Tan, H., Li, J., He, M., Li, J., Zhi, D., Qin, F., and Zhang, C. (2021). Global evolution of research on green energy and environmental technologies:A bibliometric study. J. Environ. Manage. *297*, 113382. https://doi.org/10.1016/j.jenvman.2021.113382.

Tomson, T., Battino, D., and Perucca, E. (2016). The remarkable story of valproic acid. Lancet Neurol. *15*, 141. https://doi.org/10.1016/S1474-4422(15)00398-1.

Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K.A., Ceder, G., and Jain, A. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. Nature *571*, 95–98. https://doi.org/10.1038/s41586-019-1335-8.

Tsuyuzaki, K., Morota, G., Ishii, M., Nakazato, T., Miyazaki, S., and Nikaido, I. (2015). MeSH ORA framework: R/Bioconductor packages to support MeSH over-representation analysis. BMC Bioinformatics *16*, 45. https://doi.org/10.1186/s12859-015-0453-z.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The sequence of the human genome. Science *291*, 1304–1351. https://doi.org/10.1126/science.1058040.

Wang, Q., Yang, C., Gelernter, J., and Zhao, H. (2015). Pervasive pleiotropy between psychiatric disorders and immune disorders revealed by integrative analysis of multiple GWAS. Hum. Genet. *134*, 1195–1209. https://doi.org/10.1007/s00439-015-1596-8.

Wickham, H. (2016). ggplot2 - Elegant Graphics for Data Analysis, 2nd (Cham: Springer International Publishing). https://doi.org/10.1007/978-3-319-24277-4.

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: An R package for comparing biological themes among gene clusters. OMICS *16*, 284–287. https://doi.org/10.1089/omi.2011.0118.

Yuan, N., Chen, Y., Xia, Y., Dai, J., and Liu, C. (2019). Inflammation-related biomarkers in major psychiatric disorders: A cross-disorder assessment of reproducibility and specificity in 43 meta-analyses. Transl. Psychiatry *9*, 233. https://doi.org/10.1038/s41398-019-0570-y.

Zambrano-Zaragoza, J.F., Romo-Martínez, E.J., Durán-Avelar, M.de J., García-Magallanes, N., and Vibanco-Pérez, N. (2014). Th17 cells in autoimmune and infectious diseases. Int. J. Inflam. *2014*, 651503. https://doi.org/10.1155/2014/651503.

Zitnik, M., Nguyen, F., Wang, B., Leskovec, J., Goldenberg, A., and Hoffman, M.M. (2019). Machine learning for integrating data in biology and medicine: principles, practice, and opportunities. Inf. Fusion *50*, 71–91. https://doi.org/10.1016/j.inffus.2018.09.012.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Raw data	Watson for Drug Discovery	https://doi.org/10.1016/j.clinthera.2015.12.001
Processed data	This paper	https://doi.org/10.5281/zenodo.5217544
Drug gene interactions	Comparative Toxicogenomics Database	http://ctdbase.org/
Code for all analyses and figures	This paper	https://doi.org/10.5281/zenodo.5217544
Software and algorithms		
IBM Watson for Drug Discovery	IBM	https://www.ibm.com/cloud/watson-discovery
R version 4.1.0	The R Project for Statistical Computing	https://www.r-project.org/
Inkscape	Inkscape	https://inkscape.org/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Helder Nakaya (helder.nakaya@einstein.br).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- WDD data have been deposited at Zenodo and are publicly available as of the date of publication. DOIs are listed in the key resources table.
- All original code has been deposited at Zenodo and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Watson for drug discovery

We built knowledge networks containing interactions between diseases and genes using the WDD (Chen et al., 2016a). The WDD database contains a corpus of data extracted from the biomedical scientific literature using a cognitive computing text-mining approach (Chen et al., 2016a). WDD has access to millions of abstracts in the MEDLINE platform and full texts in the PMC (PubMed Central) Open Access platform (Chen et al., 2016a). The MEDLINE database is controlled by the National Institutes of Health (NIH) and started in 1960 (NIH, 2021). The PubMed database, for instance, includes all MEDLINE references, which represents over 28 million of the 32 millions references in the PubMed database. These references are published in over 5,200 journals (NIH, 2021). The PMC Open Access is a free archive for full-text biomedical and life sciences journal articles. Some PMC journals are also MEDLINE journals and there are also reciprocal links between the full texts in PMC and corresponding citations in PubMed (NIH, 2021). In each document, the WDD algorithm detects relevant biomedical concepts, namely genes, chemicals, drugs, and diseases. This is performed using a machine learning annotator approach (WDD has a Hatz et al., 2019; High and Bakshi, 2019) dictionary of terms built using the annotator approach based on the thousands of terms extracted from its corpus that reconciles multiple synonyms of a term into a single, unambiguous concept. Next, WDD searches for meaningful associations between the detected unambiguous terms using a set of semantic annotators, i.e. nouns, verbs, and prepositions that convey a semantic relation between the terms. WDD attributes a confidence score (0–100%) to each association based on the number of documents in which the relation is found and also on the semantic relevance of each link,

determined by the machine learning annotator approach (Hatz et al., 2019; High and Bakshi, 2019). The detected terms and relationships that compose the WDD corpus then become available for online or API-mediated searches. The final user performs individual or grouped searches using terms of interest in the form of keywords (e.g. "Alzheimer's disease"). WDD automatically converts the queried keyword into the consensus term present in it's dictionary. Terms which are not present in the WDD dictionary will not yield results. The search returns tables containing the connections of the queried term with other terms of interest, the score of each connection, the class to which the connected terms belong (drug, gene, or disease) the number of documents in which the connection could be detected, and the PMIDs or PMCIDs of the documents.

WDD queries

We performed independent searches on WDD using the names of 27 inflammatory diseases, 63 infectious diseases, and 9 psychiatric and neurological disorders (Table S1) as query keywords. All searches were performed in July 2018. WDD allows users to specify a year interval from which relationships will be extracted, so that only documents published in the defined time period are accessed. We defined 29 time intervals beginning in 1964 (the first year of records in WDD) and ending in one year from 1990 to 2018. WDD returned 29 lists of genes related to the human diseases extracted from the scientific literature in each interval. These associations are cumulative, that is, the genes associated with the diseases in 2018 include all the associations present in the previous years. The lists of genes related to human diseases were downloaded in table format and processed using custom R code. From the extracted relationships, we only kept connections between genes and diseases supported by a WDD confidence score of at least 50% and 2 documents of evidence, to reduce false positive associations, as previously demonstrated by our group (Lüscher Dias et al., 2020). The tables containing the associations retrieved by WDD and the custom R code used to process, filter, and analyze data and to plot figures are available at https://doi.org/10.5281/zenodo.5217544 (Luscher Dias et al., 2021). Figure S1 summarides all the methodological steps performed in this study.

Evolution of knowledge

We measured the similarity between all pairs of human diseases by calculating a Fisher's exact test for the gene overlap between each pair in each year from 1990 to 2018 (Figure S1). We used the total number of unique genes connected to the diseases in each year's as the Fisher's exact test universe. For each year, a disease-disease knowledge network was developed. In these networks, the nodes are the diseases and the edges connect diseases that significantly share genes with each other. The edge weights are proportional to the to the significance of the gene sharing between each pair of diseases according to the Fisher's exact test. We used the $-\log_{10}$p value of the Fisher's exact test (also termed "disease-disease similarity score" here) for each disease pair. We removed edges with a Fisher's exact test p value > 0.05. The networks were constructed using the R package *igraph* (Csardi and Nepusz, 2006) and plotted using the package *ggraph with the kk layout*. We detected new genes in each year by comparing the list of genes of the diseases in one year to the list of genes of the same disease in the previous year. Thus, we obtained a list of new genes that were added to the network in each year from 1991 to 2018. The total number of genes associated with each disease was also calculated for each year. Line, violin, and ridge plots were created to illustrate the results using *ggplot2* (Wickham, 2016).

Evolution of disease relationships between categories

We selected the top 9 diseases of each category (psychiatric, inflammatory, infectious) that were connected to the most genes in 2018 ("top 9 diseases"). Then, we detected the top diseases from the other two categories that had the highest disease-disease similarity score with the top 9 diseases (Figure S1). We analyzed how the relationship between these similar pairs of distinct categories evolved from 1990 to 2018. We used the *MeSH.db* R package (Tsuyuzaki et al., 2015) to obtain the MeSH IDs and MeSH terms of all 99 diseases. Using the MeSH terms of the diseases in each pair, we used the *easyPubMed* R package to search in PubMed for papers in which both disease MeSHes of each pair were found together. We then used an adapted version of the *fetch_pubmed_data* function (see code in [Luscher Dias et al., 2021]) of the *easyPubMed* package to retrieve the number of papers that contained the searched MeSH pairs in each year from 1990 to 2018. We used the disease-disease similarity score and the number of papers in 2018 that contained MeSH terms from both diseases to calculate a similarity-to-paper ratio for each disease pair as follows:

$$similarity.paperratio = \frac{dis - dis - similarity}{number\ of\ papers}$$

Low similarity-to-paper ratios were considered as cases of low knowledge gap between the gene sharing and the general scientific interest in the disease pairs. Pairs with low ratios included those in which the diseases did not share a significant amount of genes or pairs of similar diseases for which there is also a proportional number of papers that cite the two diseases together. Intermediated ratio values were considered as cases of intermediate knowledge gap, that is, the diseases in the pair are similar in the genes they share, but the number of papers on the two diseases together is not proportionally high. High similarity-to-paper ratios were interpreted as cases of a large knowledge gap. The pairs that had high ratios include diseases that share a significant proportion of their genes but that have almost never been studied together, evidenced by the very low number of papers including the two MeSH terms.

Evolution of biological pathways

We used the *enricher* function of the R package *clusterProfiler* (Yu et al., 2012) to perform an ORA against Reactome pathways of the genes associated with the top 9 diseases of each category in each year. We selected the significant Reactome pathways (p.adjust <0.01) of the top 9 diseases in 2018 and calculated the significance of the gene overlap between these pathways with Fisher's exact test (Figure S1). We considered only the genes of each significant pathway that were also present in the 2018 gene-disease network. By doing this, we limited pathways to cluster according to the genes shared from our dataset, not all the genes in the pathways. We then built a pathway network connecting the significant Reactome terms using the $-\log_{10}$p value of the Fisher's exact tests as edge weights, similar to what was done for the disease-disease network in Figure 1A. We detected clusters of pathways in this network using the *cluster_louvain* function (Blondel et al., 2008) of the *igraph* R package (Csardi and Nepusz, 2006). Edge weights were considered for the cluster detection. We calculated the weighted degree of each pathway in the network using the *strength* function of the *igraph* package (Csardi and Nepusz, 2006). We manually annotated the detected clusters for their major biological function using the pathways with the highest weighted degree in each cluster as reference. The significance values ($-\log_{10}$pval) of ORA for the pathways in each cluster were used to make box and ridge plots to illustrate the results for each disease in 2018 and how these results changed from 1990 to 2018.

Evolution of drug target hub genes

Using the 2018 gene-disease network, we detected the genes common to all three categories of diseases ("hub genes") (Figure S1). We used the R package *UpsetR* to visualize the number of genes shared and exclusive to the disease categories. We downloaded the drug-gene and the drug-disease interaction databases from the CTD (http://ctdbase.org/ [Davis et al., 2021]). We used the MeSH terms of the 99 diseases to filter the drug-disease database and kept only interactions between drugs and diseases that were listed as "therapeutic" by CTD. These are cases of a "chemical that has a known or potential therapeutic role in a disease (e.g., chemical X is used to treat leukemia)", according to the CTD glossary (Davis et al., 2021). We filtered the drug-gene database and kept only the interactions between the therapeutic drugs and the hub genes of our analysis. This final drug-gene list was used to detect the top 20 drugs that target the most hub genes and the top 20 hub genes most targeted by the therapeutic drugs. We visualized these drug-gene interactions in a network built with the R packages *igraph* and plotted with *ggplot2* and *ggraph*. We used the yearly gene-disease networks to detect when the top 20 drug target hub genes were first connected to diseases in each category to build a timeline.

QUANTIFICATION AND STATISTICAL ANALYSIS

All analyzes were performed in the R environment (Version 4.1.0). R packages used to perform the analyses and plot figures were: igraph (v.1.2.6), ggraph (v.2.0.5), ggplot2 (v.3.3.5), clusterProfiler (v.4.0.2), MeSH.db (v.1.15.1), easyPubMed (v.2.13), and UpSetR (v.1.4.0). Specific details of each analysis are described in the Method details section and in the Results section, as well as figure legends. The significance of the difference between the number of published papers between diseases connected to more or less than 100 genes in the 2018 network was determined with a t test (p value < 0.05). Fisher's exact test result significance was established at p values < 0.01 (network edge filter) or p. adjust <0.01 (Reactome functional enrichment).