frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 20 August 2021
doi: 10.3389/fcell.2021.703489

Reconstruction of the Cytokine Signaling in Lysosomal Storage Diseases by Literature Mining and Network Analysis

Silvia Parolo[1], Danilo Tomasoni[1], Pranami Bora[1], Alan Ramponi[1], Chanchala Kaddi[2], Karim Azer[2†], Enrico Domenici[1,3], Susana Neves-Zaph[2] and Rosario Lombardo[1*]*

[1] Fondazione the Microsoft Research-University of Trento Centre for Computational and Systems Biology, Rovereto, Italy, [2] Data and Data Science – Translational Disease Modeling, Sanofi, Bridgewater, NJ, United States, [3] Department of Cellular, Computational and Integrative Biology (CIBIO), University of Trento, Trento, Italy

OPEN ACCESS

Edited by:
Adelaide Fernandes,
University of Lisbon, Portugal

Reviewed by:
Gregory M. Pastores,
Mater Misericordiae University Hospital, Ireland
Einat Vitner,
Israel Institute for Biological Research (IIBR), Israel

***Correspondence:**
Susana Neves-Zaph
Susana.Zaph@sanofi.com
Rosario Lombardo
lombardo@cosbi.eu

†Present address:
Karim Azer,
Axcella Health, Cambridge, MA, United States

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: 30 April 2021
Accepted: 30 July 2021
Published: 20 August 2021

Citation:
Parolo S, Tomasoni D, Bora P, Ramponi A, Kaddi C, Azer K, Domenici E, Neves-Zaph S and Lombardo R (2021) Reconstruction of the Cytokine Signaling in Lysosomal Storage Diseases by Literature Mining and Network Analysis. Front. Cell Dev. Biol. 9:703489. doi: 10.3389/fcell.2021.703489

Lysosomal storage diseases (LSDs) are characterized by the abnormal accumulation of substrates in tissues due to the deficiency of lysosomal proteins. Among the numerous clinical manifestations, chronic inflammation has been consistently reported for several LSDs. However, the molecular mechanisms involved in the inflammatory response are still not completely understood. In this study, we performed text-mining and systems biology analyses to investigate the inflammatory signals in three LSDs characterized by sphingolipid accumulation: Gaucher disease, Acid Sphingomyelinase Deficiency (ASMD), and Fabry Disease. We first identified the cytokines linked to the LSDs, and then built on the extracted knowledge to investigate the inflammatory signals. We found numerous transcription factors that are putative regulators of cytokine expression in a cell-specific context, such as the signaling axes controlled by STAT2, JUN, and NR4A2 as candidate regulators of the monocyte Gaucher disease cytokine network. Overall, our results suggest the presence of a complex inflammatory signaling in LSDs involving many cellular and molecular players that could be further investigated as putative targets of anti-inflammatory therapies.

Keywords: cytokine, text-mining, natural language processing, systems biology, lysosomal storage diseases, Gaucher, Fabry, ASMD

INTRODUCTION

Lysosomal storage diseases (LSDs) are a group of rare metabolic disorders in which a defect in the gene encoding a lysosomal protein causes the accumulation of substrates inside the lysosome (Platt et al., 2018). This ultimately leads to numerous clinical manifestations. Among them, several LSDs have been associated with abnormalities of the immune system and chronic inflammation (Bosch and Kielian, 2015; Pandey et al., 2017, 2018; Rigante et al., 2017).

Despite the growing body of evidence supporting a connection between LSDs and inflammation, the existing knowledge is distributed across numerous scientific publications, and a unifying knowledgebase is still missing. Text-mining is increasingly used to extract knowledge from unstructured text of scientific articles (Huang and Lu, 2016; Przybyla et al., 2016; Westergaard et al., 2018; Lee et al., 2020). Successful examples of text-mining in the biomedical field include extractions

of gene-disease associations (Piñero et al., 2015; Zhou and Fu, 2018), protein–protein interactions (Saik et al., 2016; Szklarczyk et al., 2019), drug discovery (Azer et al., 2019; Zheng et al., 2019; Hansson et al., 2020) and clinical trial design (Michelini et al., 2018). Text-mining has also been applied to the study of the immune system. Recently, a cell-cytokine network was built by PubMed mining (Kveler et al., 2018), and specific conditions such as the immune response to psychological stress (Priyadarshini and Aich, 2012) or the immune-related adverse events of immuno-oncology drugs (Yu et al., 2020) were investigated by mining of the literature. The application of text-mining to rare diseases is a less explored area; however, it could provide an important contribution to better understand the underlying biological processes that drive clinical manifestations and to develop new treatments (Boycott and Ardigó, 2018; Sakate et al., 2018; Shen et al., 2018; Roessler et al., 2021).

In this study, we developed an integrative analysis based on text-mining and network analysis to study the cytokines involved in three related LSDs: Gaucher Disease (GD), Acid Sphingomyelinase Deficiency (ASMD), and Fabry Disease (FD).

Gaucher Disease is caused by a deficient lysosomal β-glucosidase activity, which results in the accumulation of glucosylceramide (GL1) in macrophages, leading to hepatosplenomegaly, anemia, skeletal lesions, and, in some cases, neurological manifestations. Three main forms of GD are usually distinguished: type 1 (GD1) is the chronic non-neurological form, type 2 (GD2) is the acute neurological form that leads to premature death in early childhood and type 3 (GD3) is the chronic neurological form also associated with damage to peripheral tissues (Sidransky, 2004; Stirnemann et al., 2017). Currently, pharmacological treatments are available for the non-neuronopathic forms of GD and include both enzyme replacement therapies (ERT) and substrate reduction therapies (SRT) (Weinreb et al., 2013; Peterschmitt et al., 2018). ASMD is caused by mutations in the gene encoding the lysosomal enzyme acid sphingomyelinase, that converts sphingomyelin to ceramide in lysosomes. Historically, ASMD has been referred to as Niemann-Pick Disease. While an infantile neurovisceral type of ASMD (ASMD type A previously known as NPD A) has an extremely severe course involving the central nervous system (CNS), and usually a premature death in early childhood, ASMD type B (NPD B) typically displays visceral and pulmonary involvement without CNS involvement and a more heterogeneous time course (Schuchman and Desnick, 2017). Olipudase alfa, a recombinant human acid sphingomyelinase is currently in clinical trials for AMSD treatment (Thurberg et al., 2020). FD is caused by mutations in the gene encoding the enzyme alpha galactosidase A that leads to globotriaosylceramide accumulation mainly in endothelial cells, kidney cells, and cardiomyocytes. In agreement with this, the most relevant clinical manifestations of FD are quite different from those of GD and ASMD and they include renal failure, cardiac and cerebrovascular disease (Wanner et al., 2018). ERT therapy is available for FD as well (Azevedo et al., 2021).

Independently of the disease-specific therapies, the treatment of chronic inflammation is emerging as a potential adjuvant therapy for LSD patients (Platt, 2018), and a comprehensive analysis of the inflammatory signaling involved in LSDs can help to understand the dysregulated pathways. With that goal, computational mining of literature performed in this study resulted with the identification of a list of cytokines associated with GD, FD, and ASMD, which were then used as seed for a systems biology workflow providing insight into the inflammatory processes associated to LSDs.

MATERIALS AND METHODS

Text-Mining Pipeline

Over 31 million abstracts from PubMed and 6.1 million full texts from PubMed Central were harmonized and indexed on a Solr[1] instance along with the clinical trial descriptions from ClinicalTrials.org. Documents underwent an automatic annotation of genes, proteins, diseases, species and chemicals leveraging state-of-the-art Machine Learning (ML) methods DNorm (Leaman et al., 2013), GNorm (Wei et al., 2015), Huner (Weber et al., 2020), TaggerOne (Leaman and Lu, 2016), MutationFinder (Caporaso et al., 2007). The tagged entities were organized and curated along with paper's keywords and MeSH terms to identify the relevant search terms. Word2Vec (Mikolov et al., 2013; Pyysalo et al., 2013) was also used to identify lists of mentions appearing in contexts statistically similar to the ones of the input keywords. A variety of data-driven search terms were identified in this way, including some frequent mistypings (e.g., "neimann-pick") and other non-standard mentions. Starting from the simple "Gaucher disease," "Fabry disease," and "ASMD" queries, we picked relevant search terms from the unbiased information coming from ML data-driven suggestions of annotations. Some exclusions were also easily identified from the structured results, leading to the following three queries used to identify the relevant literature corpora (on the 31st of December 2020):

> Title, abstract, keyword, mesh: ("acid beta-glucosidase deficiency" OR gaucher OR "gba deficiency" OR "glucocerebrosidase deficiency" OR "glucosylceramide beta-glucosidase deficiency").
> Title, abstract, keyword, mesh: (fabry OR "Alpha-galactosidase deficiency" OR "alpha-galactosidase A deficiency" OR "GLA deficiency" OR "angiokeratoma corporis diffusum").
> AND NOT title, abstract, keyword, mesh: ("Fabry-Pérot" OR "Fabry-Perot" OR "Pérot-Fabry" OR "Perot-Fabry").
> [Title, abstract, keyword, mesh: (asmd OR "ASM deficiency" OR "Acid Sphingomyelinase deficiency" OR "neimann pick" OR "smpd1 deficiency" OR "smpd1 mutation" OR "neimann-pick" OR "neimann-pick")].
> AND NOT keyword: ["asmd, absolute standardized mean difference" OR "absolute standardized mean difference" OR "adaptive steered molecular dynamics (asmd)" OR "adaptive steered molecular dynamics"].

[1]https://lucene.apache.org/solr/

The same methods described above were used also for the concepts to be mined in the text. The list of concepts corresponding to disease synonyms was defined by querying the following databases and ontologies: OMIM[2], orphanet[3], ICD-10[4], MeSH[5], disease ontology[6]. Alternative names of the mutated gene/protein were identified from NCBI gene[7], UniProt[8], MedlinePlus[9]. Synonyms of the accumulated metabolites and the deacylated forms (lyso-species) were instead retrieved from PubChem[10] and ChEBI[11] databases. Specifically, for GD we searched synonyms of glucosylceramide and glucosylsphingosine (lyso-GL1), for FD we searched synonyms of globotriaosylceramide and globotriaosylsphingosine (lyso-Gb3), and for ASMD we searched synonyms of sphingomyelin and lyso-sphingomyelin (lyso-SM). Moreover, we included additional keywords from scientific articles identified in PubMed, from machine-learning annotation of texts and by the expanding the seed terms in the target corpora. The dictionary of cytokines was defined by leveraging the cytokine registry from the ImmPort (Bhattacharya et al., 2018) and the CytReg (Santoso et al., 2020) databases.

Disease synonyms, metabolites, cytokines and genes were identified by the entity recognition task powered by *ad hoc* linguistic variant-detection (Ramponi et al., 2020). Paragraphs where such mentions occurred were not analyzed for co-mention, given that we were not looking at summarization statistics which are known to suffer from poor precision associations in the absence of intensive human screening. The text was rather analyzed using a hi-precision method that proved highly reliable on five different benchmark corpora (Ramponi et al., 2020). The method identified syntactic relational structures among the entities and extracted linguistic associations between cytokines and disease-related concepts. The result is a structured format defining effector, affected and a verbal association among the two, therefore allowing for further systems analysis. The results were filtered to retain only sentences mentioning at least one disease-cytokine association. Sentences from clinical trial records and the method section of the full text articles were excluded due to their low information content. We also excluded from subsequent analyses all cytokine names that could not be mapped unambiguously to a gene symbol.

Evaluation of Cell-Type Expression Specificity of the Cytokine

Gene expression information for eight blood cell types: eosinophils, basophils, neutrophils, T-cells, B-cells, monocytes, NK-cells, and dendritic cells was downloaded from Human Protein Atlas (HPA) (Uhlén et al., 2015)[12]. We analyzed the data from the consensus dataset in Blood Atlas, and we considered as expressed all genes with an HPA normalized expression value >1, and "cell-specific" all genes classified by HPA as cell-type enriched or group enriched.

Construction of the Cytokine Gene Regulatory Networks

The transcription factor (TF)–cytokine gene regulatory network (GRN) was taken from CytReg database (Santoso et al., 2020). Disease specific, immune cell TF-cytokine networks were created by first selecting human interactions and filtering the original network to retain only the cytokines identified by text-mining, and then filtering the resulting network to keep only the genes expressed in the cell type of interest (HPA normalized expression value >1). The co-expression between the TF-cytokine pairs of the GD monocyte GRN network was tested using the blood monocyte expression data downloaded from HPA Blood Atlas. For each gene, we computed the average monocyte expression at sample level (average of classical monocyte, intermediate monocyte, and non-classical monocyte) and we performed the correlation test using the Pearson method. The cytokine GRN network of GD monocyte was visualized using Cytoscape, version 3.7.1[13].

Ligand-Receptor Analysis

Ligand-receptor (LR) pairs were downloaded from CellTalkDB (Shao et al., 2020). Pathway enrichment analysis was performed using the enrichPathway function from the r package ReactomePA, which uses the hypergeometric test to evaluate whether specific Reactome pathways are enriched in a gene list (Yu and He, 2016). As background genes we used all the cytokines present in the dictionary used for text-mining and their receptors, as reported in CellTalkDB. Cytokines not present in CellTalkDB were not included in the background gene list. The minGSSize was set equal to 5. To control for multiple testing, we used the Benjamini-Hochberg correction. To build the disease-immune cell networks, we identified the cytokine receptors from CellTalkDB. We then assessed the expression of the genes encoding cytokine and receptors produced by the immune cells using the HPA data described above, and we removed non-expressed genes from the networks. The cell-cell interactions were scored based on the number of LR pairs connecting them. By analyzing the frequency distribution of the number of LR pairs connecting the cells, we selected the cell-cell interactions with a number of LR pairs above the 75th percentile and we created the network. The igraph package for R was used to create and plot the networks[14].

Enrichment Analysis of Cell-Specific Cytokines

We defined the lists cell-specific cytokine-sets by selecting for each cell the genes encoding cytokines that HPA Blood Atlas

[2]https://www.omim.org/
[3]https://www.orpha.net/
[4]https://icd.who.int/
[5]https://www.ncbi.nlm.nih.gov/mesh/
[6]https://disease-ontology.org/
[7]https://www.ncbi.nlm.nih.gov/gene
[8]https://www.uniprot.org/
[9]https://medlineplus.gov/
[10]https://pubchem.ncbi.nlm.nih.gov/
[11]https://www.ebi.ac.uk/chebi/
[12]http://www.proteinatlas.org
[13]https://cytoscape.org/
[14]https://igraph.org

reports as "cell-type enriched" or "group enriched." The gene-set collection of cell-specific cytokines was built using the loadGSC function of the Bioconductor package piano. The enrichment of GD cytokines in monocyte-specific cytokines was tested using the Fisher's exact test implemented in the runGSAhyper function of the Bioconductor package piano. The analysis was performed using all the cytokines present in the cytokine dictionary used for the text-mining analysis corresponding to a human gene symbol as universe.

Search of Transcriptomic and Proteomic Datasets

To identify relevant transcriptomic datasets, NCBI Gene Expression Omnibus (GEO)[15] and EBI ArrayExpress[16] were queried on 12th July 2021. The searches were performed using the following keywords: "gaucher" and "fabry." GD proteomic datasets were identified by searching Omics DI[17] using the keywork "gaucher" and NCBI PubMed using the query "proteomic[TIAB] AND gaucher[TIAB]" (search performed on 20th July 2021). The results were manually curated to exclude non-relevant datasets.

RESULTS

Text-Mining Analysis Identifies a List of Cytokines Associated With LSDs in the Literature

To identify the cytokines associated with GD, FD, and ASMD, we devised a text-mining pipeline that uses Natural Language Processing (NLP) techniques to process all PubMed and PubMed Central (PMC) documents to identify sentences describing a relationship between disease-related concepts and at least one cytokine (**Figure 1A**). First, we collected a corpus of scientific publications related to the three diseases. The available documents are both PubMed citations and full-text articles. Indeed, due to copyright restrictions, in some cases the complete article was not available for text-mining, and only the citation (including title, abstract, and the MeSH terms) could be automatically parsed. For GD, we identified 6,068 articles published from 1912 to 2020, 917 of which were review articles. Among all the available documents, 693 were full-text articles, and the others were citations. For FD, we retrieved a set of 4,982 documents published between 1947 and 2020, out of which 809 were review articles. The full text was available for 730 documents. For ASMD, we identified 4,301 articles mentioning ASMD or Niemann-Pick (NP) disease in the text having a range of publication dates from 1940 to 2020. In this case, the number of identified review articles was 648 and the number of full-text documents available for the text-mining analysis was 853 (**Figure 1B**). The NP disease mentions were retrieved as described in the methods and further classified in type A, B, C, D, E, F, and generic NP, allowing to filter only for NP type A and B in the relations.

In addition to the search of scientific articles, we also compiled a catalog of publicly available transcriptomic datasets related to the three diseases. By querying NCBI Gene Expression Omnibus and ArrayExpress, we identified 17 GD transcriptomic studies. Twelve studies are derived from GD mouse models, four from GD human tissues/cells and one from an experiment in fruit flies. For FD, four studies were identified, three derived from mouse models and one from human organoids. For ASMD, no publicly available transcriptomic study was identified (**Supplementary Table 1**). We also searched for published GD proteomic studies by querying Omics DI (see text footnote 17) and NCBI PubMed and we identified four published proteomics studies (**Supplementary Table 1**). For FD, a recent review already compiled a comprehensive list of proteomic studies (Rossi et al., 2021) and we could not identify any additional study. We did not find any ASMD-related proteomic study.

The retrieved scientific articles were analyzed to identify relevant sentences following the approach detailed in section "Materials and Methods" and summarized in **Figure 1A**. Briefly, the entire set of documents related to the disease of interest was computationally processed to identify sentences with a linguistic relationship between a disease-related concept (disease, mutated enzyme, accumulated metabolites) and at least one cytokine. In this work, we focused on a list of cytokines identified starting from the cytokine registry made available by the ImmPort project (Bhattacharya et al., 2018) and from the CytReg database (Santoso et al., 2020). Growth factors and hormones without a primary role in the immune system and the cytokine receptors were not included in the text-mining search. For GD, we identified 280 relevant sentences from 102 articles (**Supplementary Table 2**). These sentences mention 44 GD-related cytokines, out of which 34 could be assigned unambiguously to a human gene symbol. A visual summary of the disease-cytokine associations present in the GD literature is shown in **Figure 2A**. This chart shows the directed relations between the cytokines and the disease-related concepts automatically identified by the text-mining pipeline. The chemokine CCL18 is the most cited cytokine in GD literature, being mentioned together with a disease term in 114 sentences from 52 articles, followed by TNF with 59 sentences from 28 articles (**Figure 2B**). For FD, the list of cytokines was obtained from 163 sentences in 47 articles (**Supplementary Table 3** and **Supplementary Figure 1**). These sentences report 16 cytokines, and for 12 of them we could find the corresponding gene symbol (**Figure 2A**). For ASMD, we found 15 cytokines in 72 sentences from 24 articles (**Supplementary Table 4** and **Supplementary Figure 2**), and we identified the corresponding gene symbol for 12 cytokines (**Figure 2A**). For both FD and ASMD, the most cited cytokine is TNF. It is also worth noting that six cytokines, namely CCL5, CXCL8, IL1B, IL4, IL6, and TNF, are shared among the three lists, suggesting a shared inflammatory signal among the diseases (**Figure 2C**).

The sentences were manually annotated for mentions of clinical studies, animal models, tissues, and cells (**Supplementary Tables 2–4**). Overall, 66 sentences from

[15]http://ncbi.nlm.nih.gov/geo/

[16]https://www.ebi.ac.uk/arrayexpress/

[17]https://www.omicsdi.org/

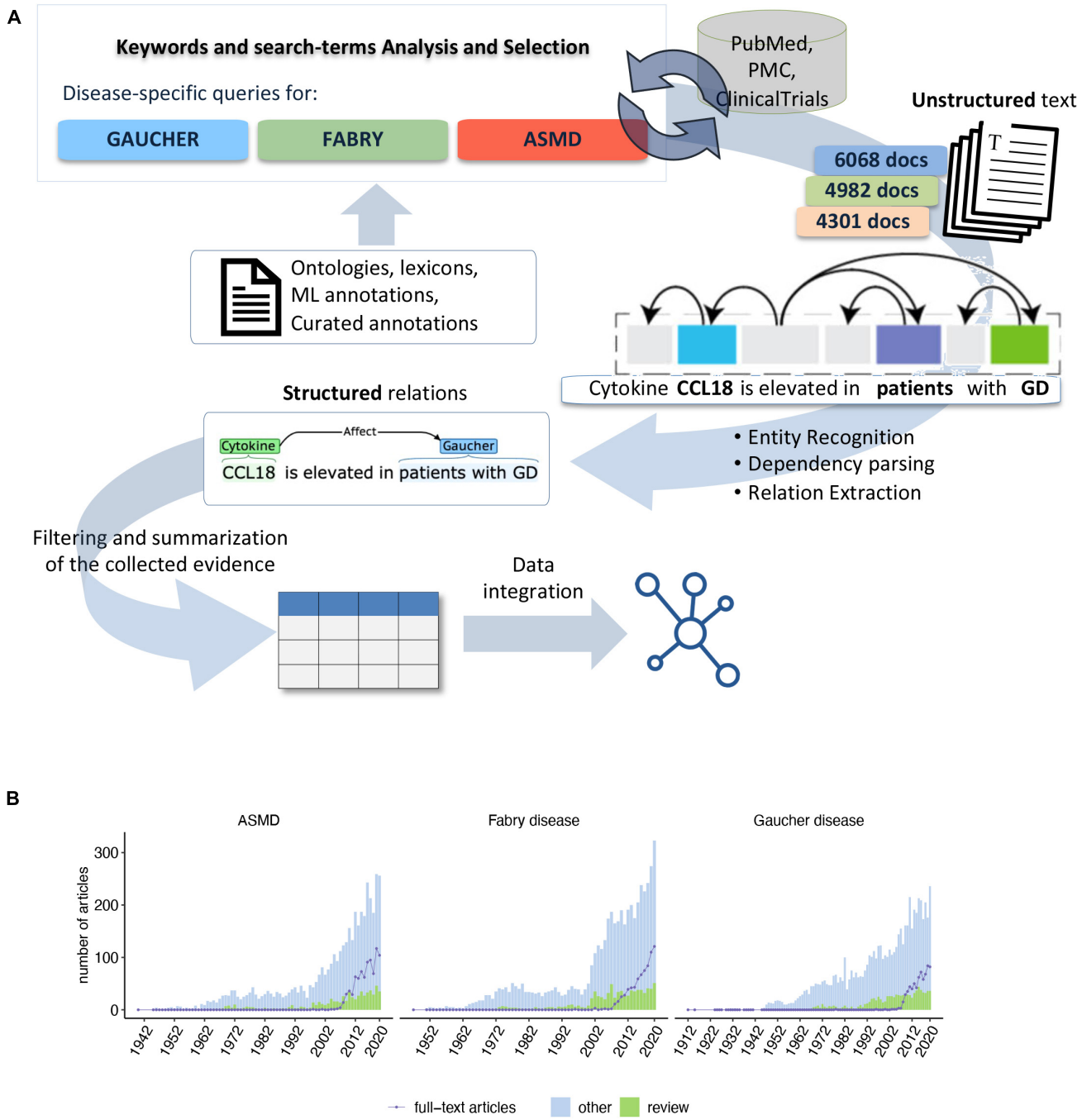

FIGURE 1 | (A) Overview of the text-mining pipeline: Ontologies, DBs, and ML annotations were used to select specific queries and search terms on PubMed, PMC for the three diseases. Over 15,000 documents were retrieved and processed with our NLP-based analytics platform. The structured knowledge derived from the linguistic relationship extraction was normalized, analyzed, and integrated with biological databases to derive a systems outlook of the inflammatory signal. **(B)** Summary of the disease-specific scientific literature we analyzed to identify the cytokines. For each disease is shown the number of articles per year of publication. The number of review articles is highlighted in green. The violet line shows the number of full-text articles available for text-mining.

GD results report findings from clinical studies, 17 sentences findings from studies performed on animal models, 38 sentences findings from *in vitro* studies, 6 sentences mixed findings, and 153 sentences do not specify the type of study. Most of the sentences indicating the tissue where the cytokine has been measured report blood findings, with 55 sentences mentioning serum, plasma, or a generic mention to blood/circulation. When we checked the type of disease, we identified 33 GD sentences that are specific for one type of disease, while the others do not specify any GD type.

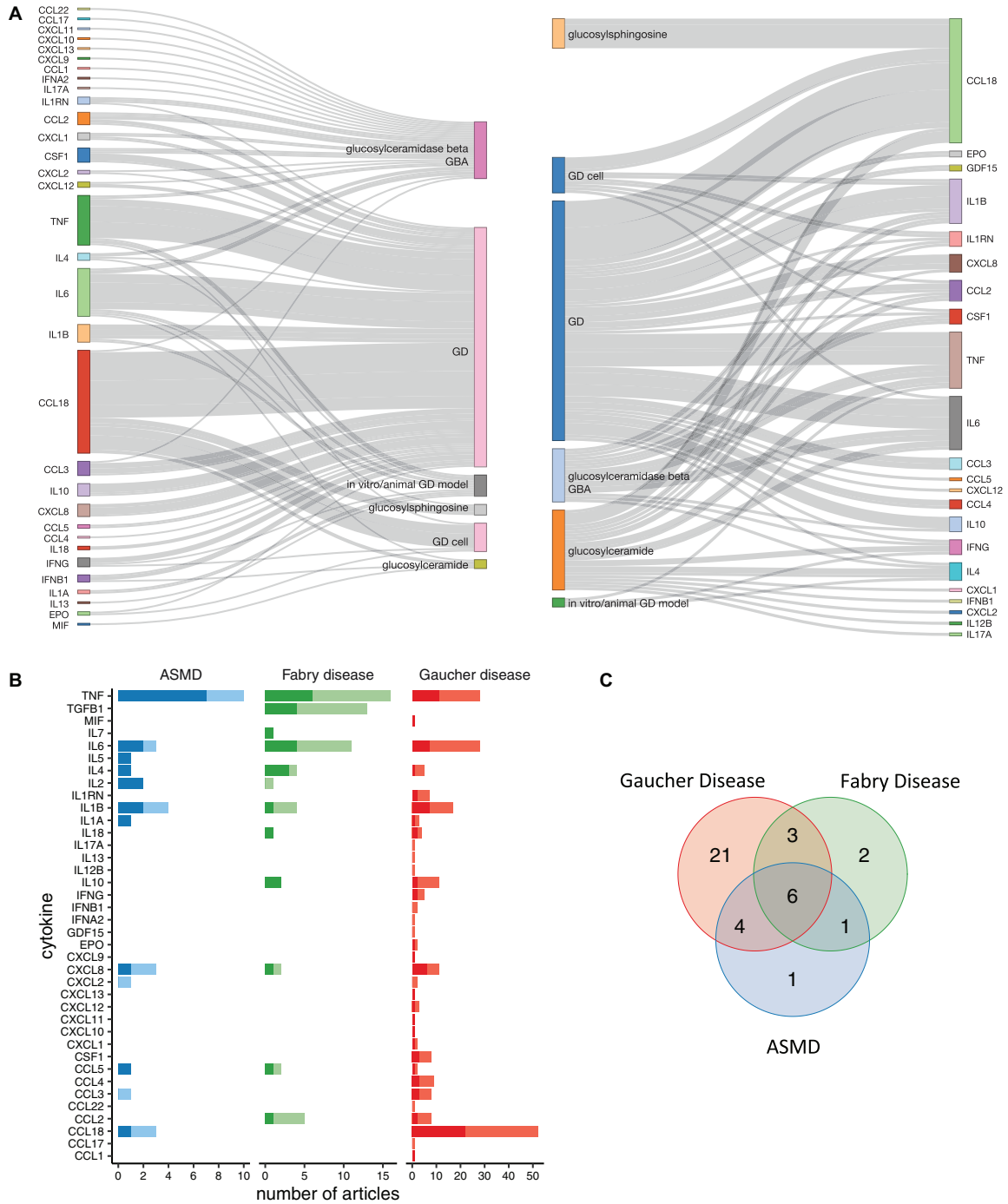

FIGURE 2 | Cytokines identified by the text-mining pipeline. **(A)** Linguistic structure of the mined text, shown as a bi-partite information flow, in the GD literature. The chart summarizes the directional associations between cytokines and the disease-related terms. The relations derived from sentences in which the cytokine is the subject are shown on the left, instead those in which the cytokine is more of an object are shown on the right. **(B)** The chart shows the number of papers mentioning each cytokine in association with the disease. Darker colors indicate that at least one sentence mentioning cytokine-disease association is present in the title or abstract of the article, suggesting that the identified association is probably a key finding. Lighter colors indicate the cytokine-disease associations derived from other sections of the article. **(C)** Venn diagram of the lists of cytokines obtained for the three diseases.

Most of the FD-related sentences refer to *in vitro* studies (52 out of 163 sentences) and in particular they mention cell cultures of podocytes (11 sentences), a kidney epithelial cell-type particularly affected by globotriaosylceramide accumulation (Waldek and Feriozzi, 2014). *In vitro* studies are also the most common ones in ASMD sentences, with 26 sentences. In this

case, most of the studies mentioned by the sentences refer to experiments performed on fibroblasts (22 sentences).

Immune Cell-Type Specificity of the Identified Cytokines

Since cytokines are molecular mediators of cell communication, mainly among immune cells, we set out to investigate the immune-cell specificity of the cytokines identified by the text-mining pipeline. According to the consensus dataset from the HPA Blood Atlas (Uhlén et al., 2015), 15 literature-derived GD cytokines are cell-specific (cell type enriched or cell group enriched) and six of them, namely CCL1, CCL2, CXCL10, CXCL12, IL1RN, and erythropoietin (EPO), are monocyte-specific (**Figures 3A,B**). When tested in a overrepresentation analysis, the monocyte-specific cytokine-set resulted significantly enriched in GD cytokines (Fisher's exact test p-value = 0.04). The other cell-specific cytokine-sets were not tested for their enrichment in GD cytokines due to the small number of overlapping genes (less than five genes). It is worth noting that CCL18, the cytokine with the highest number of citations in GD literature, was "not detected" in the HPA Blood Atlas dataset, thus we could not evaluate its cell specificity. We identified cell-specific cytokines also for ASMD and FD. However, in these cases we only found one or two cytokines overlapping the cell-specific cytokine-sets (**Figures 3A,B**) and thus we could not test the significance of the overrepresentation for any cell-type.

Construction of a GD-Cytokine Gene Regulatory Network

Next, we decided to investigate the gene regulatory network (GRN) behind the expression of cytokine genes identified by text-mining. We focused on the regulation of GD cytokine gene expression in monocytes since this is the cell type with the highest number of cell-specific cytokines (**Figure 3B**). We built the cytokine GRNs by identifying the TFs regulating the expression of the text-mining derived GD cytokines leveraging the information provided by CytReg, a recently published database of TFs-cytokine interactions (Santoso et al., 2020) which we filtered according to the monocyte gene expression reported in the Blood Atlas data (**Figure 4**).

Overall, we can observe that numerous TFs contribute to the regulation of cytokine gene expression. Among the identified

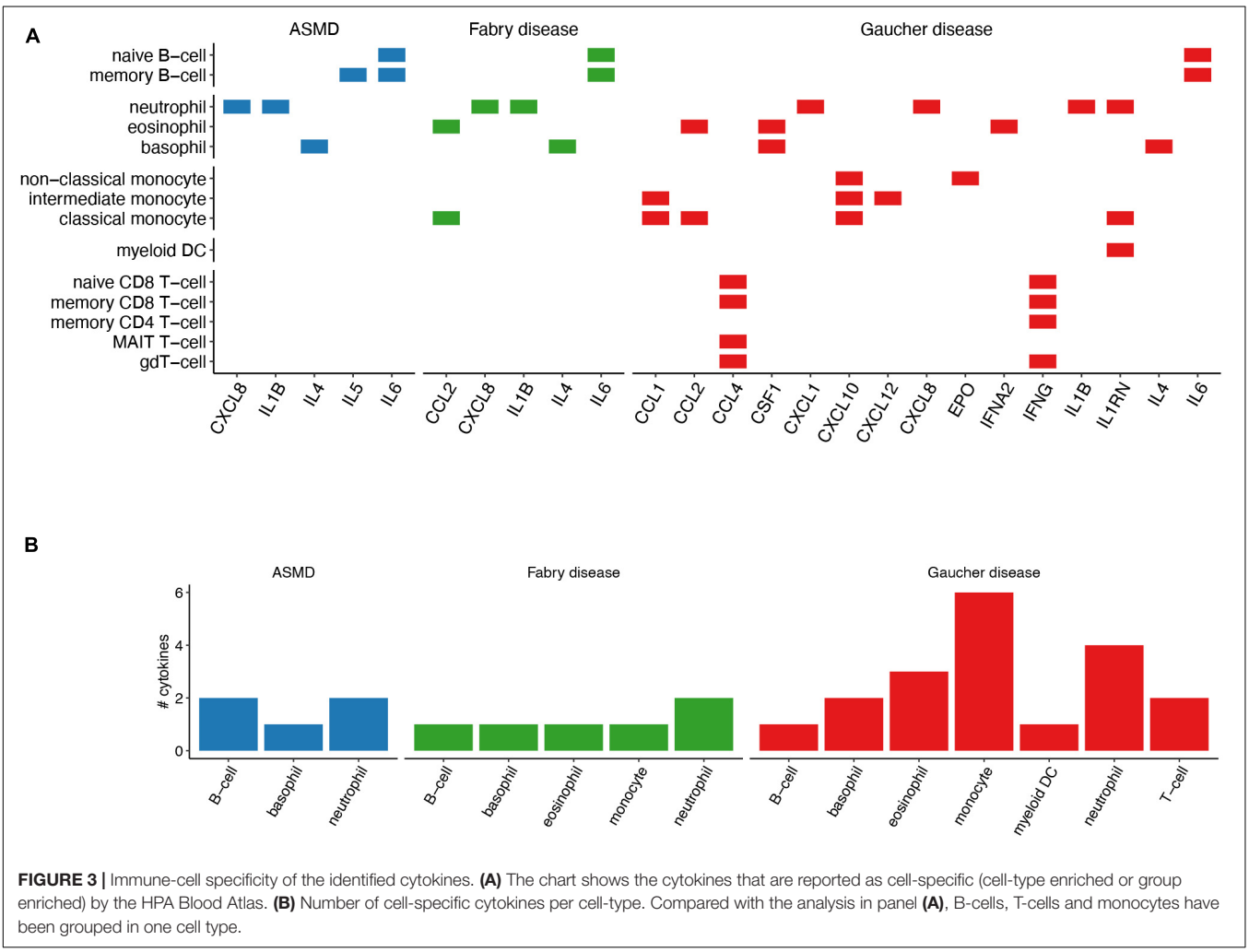

FIGURE 3 | Immune-cell specificity of the identified cytokines. **(A)** The chart shows the cytokines that are reported as cell-specific (cell-type enriched or group enriched) by the HPA Blood Atlas. **(B)** Number of cell-specific cytokines per cell-type. Compared with the analysis in panel **(A)**, B-cells, T-cells and monocytes have been grouped in one cell type.

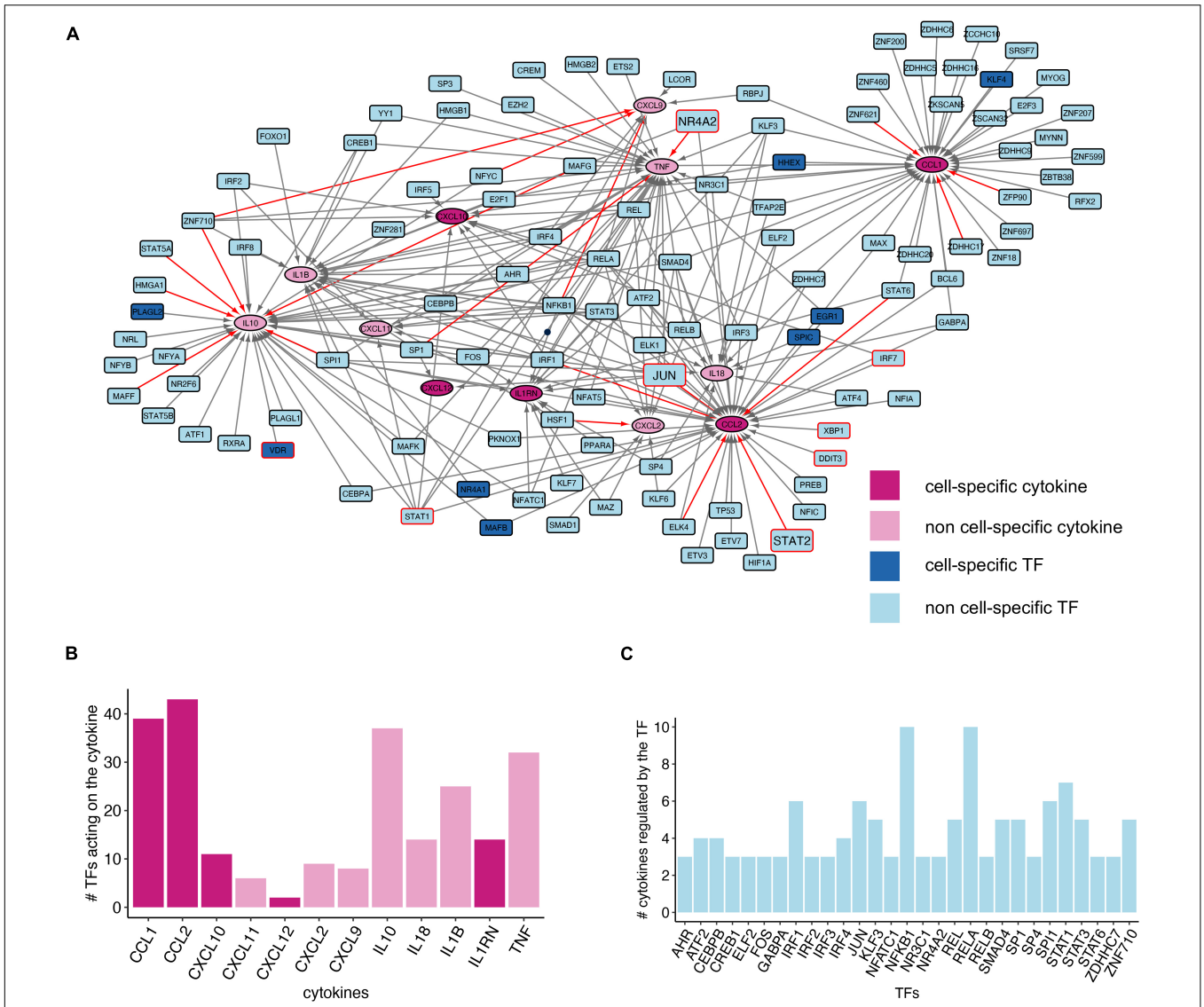

FIGURE 4 | Cytokine gene regulatory network of GD monocytes. **(A)** Network plot showing the TF-cytokine interactions. Only genes expressed in monocytes according to the HPA Blood Atlas data are reported. The nodes with a red border were identified in GD literature and the red edges indicate a significant correlation between TFs and cytokines. The big nodes (NR4A2, JUN, and STAT2) correspond to TFs identified in the GD literature with monocyte gene expression significantly correlated with a cytokine in the network. **(B)** Bar chart showing the number of incoming edges for all the cytokines in the chart, corresponding to the number of putative TFs regulating cytokine gene expression. **(C)** Bar chart showing the number of outgoing edges for the TF nodes with more than two degrees which corresponds to the number of cytokine genes regulated by each TF.

TFs, there are well-known regulators of genes involved in the immune expression, such as members of the NF-κB (REL, RELA, RELB, and NFKB1) and different members of the interferon-regulatory factor (IRF) proteins family (**Figure 4C**). Moreover, we noticed the presence of the nuclear receptor NR4A2, a TF encoded by a gene harboring genetic variants that have been associated with familial Parkinson's disease susceptibility (Le et al., 2003). Similarly, NR3C1 has been associated with epigenetic deregulations in Parkinson's disease (Fernández-Santiago et al., 2015). Since GD patients are at higher risk of developing Parkinson's disease (Behl et al., 2021), this finding deserves further investigation.

To investigate the TFs present in the cytokine GRN of GD monocytes, we used TFs as seeds of the text-mining analysis aiming at the identification of TFs reported within the context of GD. This analysis identified seven TFs, namely DDIT3, JUN, IRF7, NR4A2, STAT1, STAT2, VDR, and XBP1 (**Supplementary Table 5**). DDIT3 and XBP1 are TFs regulating *CCL2* expression in the GRN network and are both involved in the Unfolded Protein Response (UPR), a mechanism activated by the endoplasmic reticulum (ER) to cope with stress conditions. The text-mining analysis identified contradictory findings related to the UPR induction in GD. Indeed, we found both articles reporting UPR activation in GD and PD patients with mutations in *GBA* gene

and articles showing lack of evidence (Farfel-Becker et al., 2009; Gegg et al., 2012; Maor et al., 2013; Braunstein et al., 2018; Do et al., 2019; Ivanova et al., 2019). The master regulator of type I interferon signaling IRF7, which in our network regulates *CCL2* and *CXCL10* gene expression (**Figure 4**), was reported by two studies as elevated in the neurological forms of GD (Vitner et al., 2016; Melamed et al., 2020). The phosphorylated form of STAT2, a regulator of *CCL2* expression in the network in **Figure 4**, was also up-regulated in the brain of a GD mouse model (Vitner et al., 2016). On the other hand, INFG-induced STAT1 activation was shown to be inhibited in Gaucher cells (Batta et al., 2018). STAT1 in our network regulates several cytokine genes: *IL10*, *IL1B*, *CCL2*, *TNF*, *CXCL10*, *CXCL11*, and *CXCL9*. Our analysis also pointed out a study that investigated *NR4A2* expression in dopaminergic neurons obtained from GD iPSC. This study showed a decrease, albeit not significant, of *NR4A2* expression in these cells (Awad et al., 2017). Moreover, text-mining identified a study indicating that the expression of glucocerebrosidase gene is affected by JUN (Moran et al., 1997), a TF that in our network regulates *IL10*, *IL1B*, *CCL2*, *TNF*, and *CXCL12*. Finally, we identified several studies reporting polymorphisms in the VDR gene possibly associated with GD phenotypes (Vlieger et al., 2002; Greenwood et al., 2010a,b; Lieblich et al., 2011; Zhang et al., 2012; Mistry et al., 2013; Gervas-Arruga et al., 2015; Zimmermann et al., 2018; Kałużna et al., 2019).

We also investigated the correlation pattern of the identified TF-cytokine pairs by leveraging the monocyte gene expression data from the HPA Blood Atlas. In total, we identified 20 TF-cytokine gene pairs with a significant correlation (**Supplementary Table 6**). Among them, STAT2-CCL2, JUN-CCL2, and NR4A2-TNF involve TFs already described in the GD literature.

Construction of a Cytokine-Driven Immune Cell-Communication Network

Cytokines are molecular effectors that mediate the communication between cells, mainly of the immune system. In this study, we explored the cytokine signaling by leveraging publicly available data on cytokine-receptor interactions. We first identified the receptors of the text-mining derived cytokines from CellTalkDB, a recently published, curated database of ligand-receptor LR pairs (Shao et al., 2020). We identified 180 cytokine-receptor (CR) pairs for GD, 80 for ASMD and 95 for FD. Pathway analysis of the GD cytokines and their receptors identified "interleukin 10 signaling" as the most enriched pathway, with 26 genes. This pathway is also among the top significantly enriched pathways when considering the genes derived from the list of FD cytokines and their corresponding receptors. For ASMD, instead, the most enriched pathways are related to TNF signaling, with five cytokines and receptors (**Figure 5A**). Having identified the cytokine receptors, we assessed the immune cell expression of the genes encoding the cytokines and their receptors using the gene expression data from the HPA Blood Atlas, and we reconstructed a putative cell-cell interaction network based on the number of CR pairs linking two cells (**Figure 5B**). In the GD network, the intercellular interaction supported by the highest number of CR pairs is the one going from monocytes (producing the cytokine) to T cells (expressing the receptor), with 58 CR interactions.

In the FD network, the interaction dendritic cells - > T cells is the one with the highest number of CR pairs with 44 CR interactions, while in the ASMD network the cell-cell communication supported by the highest number of CR pairs is between neutrophils and T cells, with 34 interactions.

DISCUSSION

In this study, we investigated the inflammatory signaling characterizing GD, FD, and ASMD. Our integrative approach starts from literature computational mining to identify the cytokines associated with the three diseases, and then combines the literature findings with other data sources through a systems biology framework. Compared with previous efforts to identify immune cells and molecular mediators by text-mining (Kveler et al., 2018), in this study we focused on three specific LSDs. This allowed us to set up a tailored approach that reflects the characteristics of the LSDs of interest. For example, we could evaluate the association between the specific disease genes (mutated genes) or accumulated metabolites and cytokines. Indeed, the primary driver of the LSD phenotype is the accumulation of lipids that cause cellular, tissue and organ dysfunctions that are frequently coupled with chronic inflammation. The causes of the observed chronic inflammation are still not fully understood. For example, in GD and ASMD, these undegraded lipids mainly accumulate in macrophages (Pandey and Grabowski, 2013), instead in FD the vasculature is particularly affected (Bodary et al., 2007). The accumulation of glucosylceramide in GD macrophages causes their activation, disrupts autophagy and starts a cascade of inflammatory events that worsen the disease itself (Simonaro, 2016). To take into account these aspects in the text-mining analysis, we set up a search for cytokines described in association with accumulating lipids and their deacylated forms. These results were merged with those obtained by searching disease-cytokine and mutated enzyme-cytokine mentions to obtain a more comprehensive characterization of the LSD cytokines. On the other hand, cytokine mentions not linked to any disease-related concept were not considered to avoid false-positive results. For example, if a sentence mentioned a cytokine related to an immune cell but did not specify any disease-related concept, this sentence was not included among the results because it could be potentially referred to another condition. We are aware that this approach can lower the recall rate and hamper the identification of some cytokines truly associated with LSDs but at the same time it increases the precision. Future extensions of this analysis could consider the extraction of relations within paragraphs to increase the number of cytokine-disease relations diving into "discourse parsing" and more research into coreference and anaphora resolution (Soricut and Marcu, 2003; Sukthanker et al., 2020). Another factor that can limit the power of the text-mining analysis is the unavailability of the entire document for many disease-relevant articles (**Figure 1**).

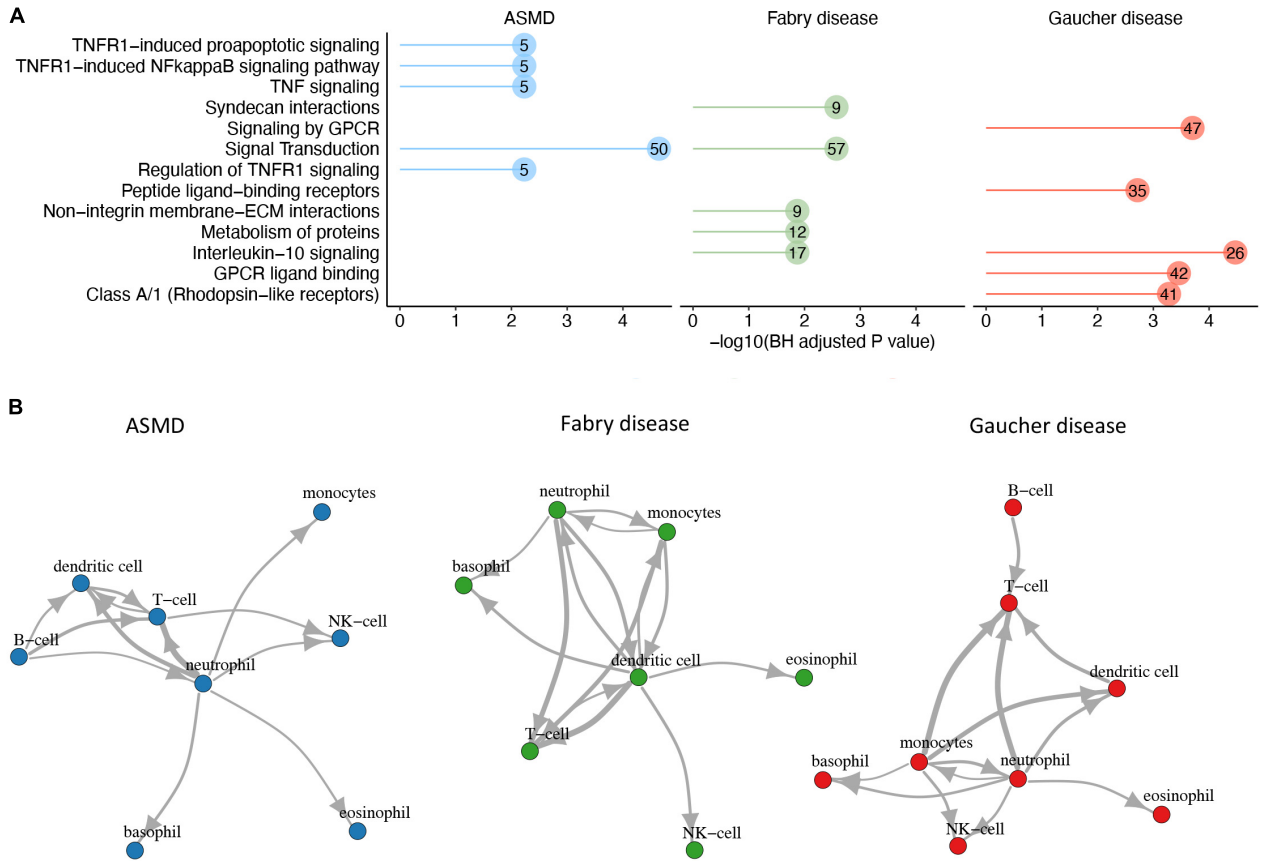

FIGURE 5 | Cytokine-receptor analysis to build cell-cell networks. **(A)** Functional enrichment analysis. For each disease, the top 5 Reactome pathways enriched in disease cytokines and their receptors are shown. The numbers within the dots indicate the number of cytokines/receptors belonging to the pathway. **(B)** For each disease, the top cell-cell interactions involving the literature-derived cytokines and their receptors are shown. The cell-cell interactions were scored based on the number of CR pairs involved, and those above the 75th percentile are shown in the network chart. The thickness of the edge is proportional to the number of interacting CR pairs that in ASMD network ranges from 20 to 34, in FD network from 28 to 44 and in GD from 33 to 56.

Indeed, a recent study showed that text mining of full-text articles to identify protein–protein, disease-gene, and protein subcellular associations outperforms the analysis using abstracts only (Westergaard et al., 2018).

To extend the analysis we performed, other players of the inflammatory response, such as immune cells and molecules belonging to the complement system, could be included in the analysis. Indeed, a dysregulation of the complement pathway has been described in GD (Pandey et al., 2017, 2018) and its investigation in a systems biology framework could provide hints on the interplay with other inflammatory players. The integration of external data sources, such as gene expression data of immune cells, TF-cytokine gene interactions, and LR interactions allowed us to gain insights into the inflammatory signaling network. Indeed, cytokines are molecular effectors involved in cell-cell signaling (Armingol et al., 2020), and their production is regulated at the transcriptional level by combinations of TFs (Pro et al., 2018; Santoso et al., 2020). In this study, to evaluate the cell specificity and build the GRN, we relied on gene expression data of blood immune cells from the blood of healthy donors. Our literature-derived list of GD cytokines is significantly enriched in monocyte-specific cytokines. This finding is in agreement with the hallmark of GD pathophysiology, i.e., the accumulation of GL1 in the cells of the macrophage-monocyte system. Indeed, macrophages are the mediators of the removal of erythrocytes and leukocytes, which contain large amount of GL1 whose accumulation leads to clinical manifestations such as splenomegaly and hepatomegaly (Stirnemann et al., 2017). The significance of the TF-cytokine interactions in the GD monocyte network was tested by computing the gene expression correlation. This analysis allowed us to identify three TFs, namely STAT2, JUN, and NR4A2, that could be further investigated in the context of GD. The HPA Blood Atlas dataset used to perform the correlation analysis, however, includes only six samples and thus the analysis had a limited statistical power. The availability of disease-specific immune cell transcriptomics datasets, for example derived from single-cell sequencing experiments, would allow investigating more precisely the inflammatory signaling characterizing these diseases and to consider additional disease-specific immune cell types.

Our results can be used as a basis to further investigate the interplay between lipid accumulation and inflammation.

To support drug discovery and development for LSDs, we recently developed quantitative systems pharmacology (QSP) models for GD type I and for ASMD, and a QSP platform that also includes FD is under development (Kaddi C. et al., 2018; Kaddi C. D. et al., 2018; Abrams et al., 2020). QSP models are mathematical tools that allow studying *in silico* the perturbations exerted by drugs on a biological system and test hypotheses on their mechanism of action. Literature mining can be effectively incorporated in the multistep process that leads to model development, becoming particularly useful for the definition of the model scheme and facilitating the identification of key biological processed to represent, along with data sources and parameter constraints needed for the model development (Azer et al., 2021).

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. The data can be found here: http://tcm.zju.edu.cn/celltalkdb/, http://www.proteinatlas.org, https://cytreg.bu.edu/search.html. Scientific articles processed with NLP methods were retrieved from PubMed and PubMed Central.

AUTHOR CONTRIBUTIONS

SP, CK, KA, and RL conceived the study. SP and RL designed the analyses, curated the data, and wrote the original draft. SP, RL, PB, and DT performed the analyses. DT, AR, and RL developed the text-mining pipeline. ED and SN-Z supervised the project. ED managed the project and provided resources. SP, DT, PB, AR, CK, KA, ED, SN-Z, and RL discussed the results and reviewed the draft. All authors contributed to the article and approved the submitted version.

FUNDING

This study was partially funded by Sanofi.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.703489/full#supplementary-material

Supplementary Figure 1 | Linguistic structure of the sentences with cytokine-disease associations in the FD literature.

Supplementary Figure 2 | Linguistic structure of the sentences with cytokine-disease associations in the ASMD literature.

Supplementary Table 1 | Selected GD and FD transcriptomic and proteomic datasets.

Supplementary Table 2 | Selected sentences from GD literature.

Supplementary Table 3 | Selected sentences from FD literature.

Supplementary Table 4 | Selected sentences from ASMD literature.

Supplementary Table 5 | Results of text-mining analysis of TF search in the GD literature.

Supplementary Table 6 | Significant results of TF-cytokine pairs correlation analysis.

REFERENCES

Abrams, R., Kaddi, C. D., Tao, M., Leiser, R. J., Simoni, G., Reali, F., et al. (2020). A quantitative systems pharmacology model of gaucher disease type 1 provides mechanistic insight into the response to substrate reduction therapy with eliglustat. *CPT Pharmacometr. Syst. Pharmacol.* 9, 374–383. doi: 10.1002/psp4.12506

Armingol, E., Officer, A., Harismendy, O., and Lewis, N. E. (2020). Deciphering cell–cell interactions and communication from gene expression. *Nat. Rev. Genet.* 22, 71–88. doi: 10.1038/s41576-020-00292-x

Awad, O., Panicker, L. M., Deranieh, R. M., Srikanth, M. P., Brown, R. A., Voit, A., et al. (2017). Altered differentiation potential of gaucher's disease iPSC neuronal progenitors due to Wnt/β-catenin downregulation. *Stem Cell Rep.* 9, 1853–1867. doi: 10.1016/j.stemcr.2017.10.029

Azer, K., Kaddi, C. D., Barrett, J. S., Bai, J. P. F., McQuade, S. T., Merrill, N. J., et al. (2021). History and future perspectives on the discipline of quantitative systems pharmacology modeling and its applications. *Front. Physiol.* 12:637999. doi: 10.3389/fphys.2021.637999

Azer, K., Michelini, S., Giampiccolo, S., Parolo, S., Leonardelli, L., Lombardo, R., et al. (2019). TB knowledgebase: interactive application for extracting knowledge from the TB literature to inform TB drug and vaccine development. *Int. J. Tubercul. Lung Dis.* 22:S592.

Azevedo, O., Gago, M. F., Miltenberger-Miltenyi, G., Sousa, N., and Cunha, D. (2021). Review fabry disease therapy: state-of-the-art and current challenges. *Int. J. Mol. Sci.* 22, 1–16. doi: 10.3390/ijms22010206

Batta, G., Soltész, L., Kovács, T., Bozó, T., Mészár, Z., Kellermayer, M., et al. (2018). Alterations in the properties of the cell membrane due to glycosphingolipid accumulation in a model of Gaucher disease. *Sci. Rep.* 8:157. doi: 10.1038/s41598-017-18405-8

Behl, T., Kaur, G., Fratila, O., Buhas, C., Judea-Pusta, C. T., Negrut, N., et al. (2021). Cross-talks among GBA mutations, glucocerebrosidase, and α-synuclein in GBA-associated Parkinson's disease and their targeted therapeutic approaches: a comprehensive review. *Transl. Neurodegener.* 10:4. doi: 10.1186/s40035-020-00226-x

Bhattacharya, S., Dunn, P., Thomas, C. G., Smith, B., Schaefer, H., Chen, J., et al. (2018). ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. *Sci. Data* 5:180015. doi: 10.1038/sdata.2018.15

Bodary, P. F., Shayman, J. A., and Eitzman, D. T. (2007). α-Galactosidase A in vascular disease. *Trends Cardiovasc. Med.* 17, 129–133. doi: 10.1016/j.tcm.2007.02.006

Bosch, M. E., and Kielian, T. (2015). Neuroinflammatory paradigms in lysosomal storage diseases. *Front. Neurosci.* 9:417. doi: 10.3389/fnins.2015.00417

Boycott, K. M., and Ardigó, D. (2018). Addressing challenges in the diagnosis and treatment of rare genetic diseases. *Nat. Rev. Drug Discov.* 17, 151–152. doi: 10.1038/nrd.2017.246

Braunstein, H., Maor, G., Chicco, G., Filocamo, M., Zimran, A., and Horowitz, M. (2018). UPR activation and CHOP mediated induction of GBA1 transcription in Gaucher disease. *Blood Cells Mol. Dis.* 68, 21–29. doi: 10.1016/j.bcmd.2016.10.025

Caporaso, J. G., Baumgartner, W. A., Randolph, D. A., Cohen, K. B., and Hunter, L. (2007). MutationFinder: a high-performance system for extracting point mutation mentions from text. *Bioinformatics* 23, 1862–1865. doi: 10.1093/bioinformatics/btm235

Do, J., McKinney, C., Sharma, P., and Sidransky, E. (2019). Glucocerebrosidase and its relevance to Parkinson disease. *Mol. Neurodegener.* 14:36. doi: 10.1186/s13024-019-0336-2

Farfel-Becker, T., Vitner, E., Dekel, H., Leshem, N., Enquist, I. B., Karlsson, S., et al. (2009). No evidence for activation of the unfolded protein response in neuronopathic models of Gaucher disease. *Hum. Mol. Genet.* 18, 1482–1488. doi: 10.1093/hmg/ddp061

Fernández-Santiago, R., Carballo-Carbajal, I., Castellano, G., Torrent, R., Richaud, Y., Sánchez-Danés, A., et al. (2015). Aberrant epigenome in iPSC -derived dopaminergic neurons from Parkinson's disease patients. *EMBO Mol. Med.* 7, 1529–1546. doi: 10.15252/emmm.201505439

Gegg, M. E., Burke, D., Heales, S. J. R., Cooper, J. M., Hardy, J., Wood, N. W., et al. (2012). Glucocerebrosidase deficiency in substantia nigra of parkinson disease brains. *Ann. Neurol.* 72, 455–463. doi: 10.1002/ana.23614

Gervas-Arruga, J., Cebolla, J. J., De Blas, I., Roca, M., Pocovi, M., and Giraldo, P. (2015). The influence of genetic variability and proinflammatory status on the development of bone disease in patients with Gaucher disease. *PLoS One* 10:e0126153. doi: 10.1371/journal.pone.0126153

Greenwood, A., Altarescu, G., Zimran, A., and Elstein, D. (2010a). Vitamin D Receptor (VDR) polymorphisms in the cardiac variant of gaucher disease. *Pediatr. Cardiol.* 31, 30–32. doi: 10.1007/s00246-009-9538-7

Greenwood, A., Elstein, D., Zimran, A., and Altarescu, G. (2010b). Effect of vitamin D receptor (VDR) genotypes on the risk for osteoporosis in type 1 Gaucher disease. *Clin. Rheumatol.* 29, 1037–1041. doi: 10.1007/s10067-010-1464-9

Hansson, L. K., Hansen, R. B., Pletscher-Frankild, S., Berzins, R., Hansen, D. H., Madseni, D., et al. (2020). Semantic text mining in early drug discovery for type 2 diabetes. *PLoS One* 15:e0233956. doi: 10.1371/journal.pone.0233956

Huang, C. C., and Lu, Z. (2016). Community challenges in biomedical text mining over 10 years: success, failure and the future. *Brief. Bioinform.* 17, 132–144. doi: 10.1093/bib/bbv024

Ivanova, M. M., Changsila, E., Iaonou, C., and Goker-Alpan, O. (2019). Impaired autophagic and mitochondrial functions are partially restored by ERT in Gaucher and Fabry diseases. *PLoS One* 14:e0210617. doi: 10.1371/journal.pone.0210617

Kaddi, C., Reali, F., Marchetti, L., Niesner, B., Parolo, S., Simoni, G., et al. (2018). Integrated quantitative systems pharmacology (QSP) model of lysosomal diseases provides an innovative computational platform to support research and therapeutic development for the sphingolipidoses. *Mol. Genet. Metab.* 123, S73–S74. doi: 10.1016/j.ymgme.2017.12.183

Kaddi, C. D., Niesner, B., Baek, R., Jasper, P., Pappas, J., Tolsma, J., et al. (2018). Quantitative systems pharmacology modeling of acid sphingomyelinase deficiency and the enzyme replacement therapy olipudase alfa is an innovative tool for linking pathophysiology and pharmacology. *CPT Pharmacometr. Syst. Pharmacol.* 7, 442–452. doi: 10.1002/psp4.12304

Kałużna, M., Trzeciak, I., Ziemnicka, K., Machaczka, M., and Ruchała, M. (2019). Endocrine and metabolic disorders in patients with Gaucher disease type 1: a review. *Orphanet J. Rare Dis.* 14:275. doi: 10.1186/s13023-019-1211-5

Kveler, K., Starosvetsky, E., Ziv-Kenet, A., Kalugny, Y., Gorelik, Y., Shalev-Malul, G., et al. (2018). Immune-centric network of cytokines and cells in disease context identified by computational mining of PubMed. *Nat. Biotechnol.* 36, 651–659. doi: 10.1038/nbt.4152

Le, W., dong, Xu, P., Jankovic, J., Jiang, H., Appel, S. H., et al. (2003). Mutations in NR4A2 associated with familial Parkinson disease. *Nat. Genet.* 33, 85–89. doi: 10.1038/ng1066

Leaman, R., Doğan, R. I., and Lu, Z. (2013). DNorm: Disease name normalization with pairwise learning to rank. *Bioinformatics* 29, 2909–2917. doi: 10.1093/bioinformatics/btt474

Leaman, R., and Lu, Z. (2016). TaggerOne: joint named entity recognition and normalization with semi-Markov models. *Bioinformatics* 32, 2839–2846. doi: 10.1093/bioinformatics/btw343

Lee, J., Lee, D., and Lee, K. H. (2020). Literature mining for context-specific molecular relations using multimodal representations (COMMODAR). *BMC Bioinform.* 21:250. doi: 10.1186/s12859-020-3396-y

Lieblich, M., Altarescu, G., Zimran, A., and Elstein, D. (2011). Vitamin D Receptor (VDR) polymorphic variants in patients with cancer and Gaucher disease. *Blood Cells Mol. Dis.* 46, 92–94. doi: 10.1016/j.bcmd.2010.09.002

Maor, G., Rencus-Lazar, S., Filocamo, M., Steller, H., Segal, D., and Horowitz, M. (2013). Unfolded protein response in Gaucher disease: from human to Drosophila. *Orphanet J. Rare Dis.* 8:140. doi: 10.1186/1750-1172-8-140

Melamed, S., Avraham, R., Rothbard, D. E., Erez, N., Israely, T., Klausner, Z., et al. (2020). Innate immune response in neuronopathic forms of Gaucher disease confers resistance against viral-induced encephalitis*. *Acta Neuropathol. Commun.* 8:144. doi: 10.1186/s40478-020-01020-6

Michelini, S., Balakrishnan, B., Parolo, S., Matone, A., Mullaney, J. A., Young, W., et al. (2018). A reverse metabolic approach to weaning: in silico identification of immune-beneficial infant gut bacteria, mining their metabolism for prebiotic feeds and sourcing these feeds in the natural product space. *Microbiome* 6:171. doi: 10.1186/s40168-018-0545-x

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). "Efficient estimation of word representations in vector space," in *Proceedings of the 1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, Scottsdale, AR.

Mistry, P. K., Taddei, T., vom Dahl, S., and Rosenbloom, B. E. (2013). Gaucher disease and malignancy: a model for cancer pathogenesis in an inborn error of metabolism. *Crit. Rev. Oncog.* 18, 235–246. doi: 10.1615/CritRevOncog.2013006145

Moran, D., Galperin, E., and Horowitz, M. (1997). Identification of factors regulating the expression of the human glucocerebrosidase gene. *Gene* 194, 201–213. doi: 10.1016/S0378-1119(97)00148-0

Pandey, M. K., Burrow, T. A., Rani, R., Martin, L. J., Witte, D., Setchell, K. D., et al. (2017). Complement drives glucosylceramide accumulation and tissue inflammation in Gaucher disease. *Nature* 543, 108–112. doi: 10.1038/nature21368

Pandey, M. K., and Grabowski, G. A. (2013). Immunological cells and functions in Gaucher disease. *Crit. Rev. Oncog.* 18, 197–220. doi: 10.1615/CritRevOncog.2013004503

Pandey, M. K., Grabowski, G. A., and Köhl, J. (2018). An unexpected player in Gaucher disease: the multiple roles of complement in disease development. *Semin. Immunol.* 37, 30–42. doi: 10.1016/j.smim.2018.02.006

Peterschmitt, M. J., Cox, G. F., Ibrahim, J., MacDougall, J., Underhill, L. H., Patel, P., et al. (2018). A pooled analysis of adverse events in 393 adults with Gaucher disease type 1 from four clinical trials of oral eliglustat: evaluation of frequency, timing, and duration. *Blood Cells Mol. Dis.* 68, 185–191. doi: 10.1016/j.bcmd.2017.01.006

Piñero, J., Queralt-Rosinach, N., Bravo, À, Deu-Pons, J., Bauer-Mehren, A., Baron, M., et al. (2015). DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database* 2015:bav028. doi: 10.1093/database/bav028

Platt, F. M. (2018). Emptying the stores: lysosomal diseases and therapeutic strategies. *Nat. Rev. Drug Discov.* 17, 133–150. doi: 10.1038/nrd.2017.214

Platt, F. M., d'Azzo, A., Davidson, B. L., Neufeld, E. F., and Tifft, C. J. (2018). Lysosomal storage diseases. *Nat. Rev. Dis. Prim.* 4:27. doi: 10.1038/s41572-018-0025-4

Priyadarshini, S., and Aich, P. (2012). Effects of psychological stress on innate immunity and metabolism in humans: a systematic analysis. *PLoS One* 7:e43232. doi: 10.1371/journal.pone.0043232

Pro, S. C., Imedio, A. D., Santoso, C. S., Gan, K. A., Sewell, J. A., Martinez, M., et al. (2018). Global landscape of mouse and human cytokine transcriptional regulation. *Nucleic Acids Res.* 46, 9321–9337. doi: 10.1093/nar/gky787

Przybyla, P., Shardlow, M., Aubin, S., Bossy, R., De Castilho, R. E., Piperidis, S., et al. (2016). Text mining resources for the life sciences. *Database* 2016, 1–30. doi: 10.1093/database/baw145

Pyysalo, S., Ginter, F., Moen, H., Salakoski, T., and Ananiadou, S. (2013). Distributional semantics resources for biomedical text processing. *Proc. LBM* 2013, 39–44.

Ramponi, A., Giampiccolo, S., Tomasoni, D., Priami, C., and Lombardo, R. (2020). High-precision biomedical relation extraction for reducing human curation efforts in industrial applications. *IEEE Access* 8, 150999–151011. doi: 10.1109/ACCESS.2020.3014862

Rigante, D., Cipolla, C., Basile, U., Gulli, F., and Savastano, M. C. (2017). Overview of immune abnormalities in lysosomal storage disorders. *Immunol. Lett.* 188, 79–85. doi: 10.1016/j.imlet.2017.07.004

Roessler, H. I., Knoers, N. V. A. M., van Haelst, M. M., and van Haaften, G. (2021). Drug repurposing for rare diseases. *Trends Pharmacol. Sci.* 42, 255–267. doi: 10.1016/j.tips.2021.01.003

Rossi, F., L'Imperio, V., Marti, H. P., Svarstad, E., Smith, A., Bolognesi, M. M., et al. (2021). Proteomics for the study of new biomarkers in Fabry disease: state of the art. *Mol. Genet. Metab.* 132, 86–93. doi: 10.1016/J.YMGME.2020.10.006

Saik, O. V., Ivanisenko, T. V., Demenkov, P. S., and Ivanisenko, V. A. (2016). Interactome of the hepatitis C virus: literature mining with ANDSystem. *Virus Res.* 218, 40–48. doi: 10.1016/j.virusres.2015.12.003

Sakate, R., Fukagawa, A., Takagaki, Y., Okura, H., and Matsuyama, A. (2018). Trends of clinical trials for drug development in rare diseases. *Curr. Clin. Pharmacol.* 13, 199–208. doi: 10.2174/1574884713666180604081349

Santoso, C. S., Li, Z., Lal, S., Yuan, S., Gan, K. A., Agosto, L. M., et al. (2020). Comprehensive mapping of the human cytokine gene regulatory network. *Nucleic Acids Res.* 48, 12055–12073. doi: 10.1093/nar/gkaa1055

Schuchman, E. H., and Desnick, R. J. (2017). Types A and B Niemann-Pick disease. *Mol. Genet. Metab.* 120, 27–33. doi: 10.1016/j.ymgme.2016.12.008

Shao, X., Liao, J., Li, C., Lu, X., Cheng, J., and Fan, X. (2020). CellTalkDB: a manually curated database of ligand–receptor interactions in humans and mice. *Brief. Bioinform.* 22, bbaa269. doi: 10.1093/bib/bbaa269

Shen, F., Liu, S., Wang, Y., Wen, A., Wang, L., and Liu, H. (2018). Utilization of electronic medical records and biomedical literature to support the diagnosis of rare diseases using data fusion and collaborative filtering approaches. *J. Med. Internet Res.* 6:e11301. doi: 10.2196/11301

Sidransky, E. (2004). Gaucher disease: complexity in a "simple" disorder. *Mol. Genet. Metab.* 83, 6–15. doi: 10.1016/J.YMGME.2004.08.015

Simonaro, C. M. (2016). Lysosomes, lysosomal storage diseases, and inflammation*. *J. Inborn Errors Metab. Screen.* 4, doi: 10.1177/2326409816650465

Soricut, R., and Marcu, D. (2003). "Sentence level discourse parsing using syntactic and lexical information," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, NAACL '03*, Stroudsburg, PA, 149–156.

Stirnemann, J. Ô, Belmatoug, N., Camou, F., Serratrice, C., Froissart, R., Caillaud, C., et al. (2017). A review of gaucher disease pathophysiology, clinical presentation and treatments. *Int. J. Mol. Sci.* 18:441. doi: 10.3390/ijms18020441

Sukthanker, R., Poria, S., Cambria, E., and Thirunavukarasu, R. (2020). Anaphora and coreference resolution: a review. *Inf. Fusion* 59, 139–162. doi: 10.1016/j.inffus.2020.01.010

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47, D607–D613. doi: 10.1093/nar/gky1131

Thurberg, B. L., Diaz, G. A., Lachmann, R. H., Schiano, T., Wasserstein, M. P., Ji, A. J., et al. (2020). Long-term efficacy of olipudase alfa in adults with acid sphingomyelinase deficiency (ASMD): further clearance of hepatic sphingomyelin is associated with additional improvements in pro- and anti-atherogenic lipid profiles after 42 months of treatment. *Mol. Genet. Metab.* 131, 245–252. doi: 10.1016/j.ymgme.2020.06.010

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., et al. (2015). Tissue-based map of the human proteome. *Science* 347:1260419. doi: 10.1126/science.1260419

Vitner, E. B., Farfel-Becker, T., Ferreira, N. S., Leshkowitz, D., Sharma, P., Lang, K. S., et al. (2016). Induction of the type I interferon response in neurological forms of Gaucher disease. *J. Neuroinflamm.* 13:104. doi: 10.1186/s12974-016-0570-2

Vlieger, E. J. P., Maas, M., Akkerman, E. M., Hollak, C. E. M., and Den Heeten, G. J. (2002). Vertebra disc ratio as a parameter for bone marrow involvement and its application in Gaucher disease. *J. Comput. Assist. Tomogr.* 26, 843–848. doi: 10.1097/00004728-200209000-00031

Waldek, S., and Feriozzi, S. (2014). Fabry nephropathy: a review – how can we optimize the management of Fabry nephropathy? *BMC Nephrol.* 15:72. doi: 10.1186/1471-2369-15-72

Wanner, C., Arad, M., Baron, R., Burlina, A., Elliott, P. M., Feldt-Rasmussen, U., et al. (2018). European expert consensus statement on therapeutic goals in Fabry disease. *Mol. Genet. Metab.* 124, 189–203. doi: 10.1016/J.YMGME.2018.06.004

Weber, L., Münchmeyer, J., Rocktäschel, T., Habibi, M., and Leser, U. (2020). HUNER: Improving biomedical NER with pretraining. *Bioinformatics* 36, 295–302. doi: 10.1093/bioinformatics/btz528

Wei, C. H., Kao, H. Y., and Lu, Z. (2015). GNormPlus: an integrative approach for tagging genes, gene families, and protein domains. *Biomed Res. Int.* 2015:918710. doi: 10.1155/2015/918710

Weinreb, N. J., Goldblatt, J., Villalobos, J., Charrow, J., Cole, J. A., Kerstenetzky, M., et al. (2013). Long-term clinical outcomes in type 1 Gaucher disease following 10 years of imiglucerase treatment. *J. Inherit. Metab. Dis.* 36, 543–553. doi: 10.1007/s10545-012-9528-4

Westergaard, D., Stærfeldt, H. H., Tønsberg, C., Jensen, L. J., and Brunak, S. (2018). A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. *PLoS Comput. Biol.* 14:e1005962. doi: 10.1371/journal.pcbi.1005962

Yu, G., and He, Q.-Y. (2016). ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol. BioSyst.* 12, 477–479. doi: 10.1039/C5MB00663E

Yu, Y., Ruddy, K., Mansfield, A., Zong, N., Wen, A., Tsuji, S., et al. (2020). Detecting and filtering immune-related adverse events signal based on text mining and observational health data sciences and informatics common data model: Framework development study. *JMIR Med. Inform.* 8:e17353. doi: 10.2196/17353

Zhang, C. K., Stein, P. B., Liu, J., Wang, Z., Yang, R., Cho, J. H., et al. (2012). Genome-wide association study of N370S homozygous Gaucher disease reveals the candidacy of CLN8 gene as a genetic modifier contributing to extreme phenotypic variation. *Am. J. Hematol.* 87, 377–383. doi: 10.1002/ajh.23118

Zheng, S., Dharssi, S., Wu, M., Li, J., and Lu, Z. (2019). Text mining for drug discovery. *Methods Mol. Biol.* 1939, 231–252. doi: 10.1007/978-1-4939-9089-4_13

Zhou, J., and Fu, B. Q. (2018). The research on gene-disease association based on text-mining of PubMed. *BMC Bioinform.* 19:37. doi: 10.1186/s12859-018-2048-y

Zimmermann, A., Popp, R. A., Rossmann, H., Bucerzan, S., Nascu, I., Leucuta, D., et al. (2018). Gene variants of osteoprotegerin, estrogen-, calcitonin-and vitamin D-receptor genes and serum markers of bone metabolism in patients with gaucher disease type. *Ther. Clin. Risk Manag.* 14, 2069–2080. doi: 10.2147/TCRM.S177480

Conflict of Interest: CK and SN-Z are current employees of Sanofi and may hold shares and/or stock options in the company. KA was employee of Sanofi while the initial phase of this study was conducted. SP, RL, DT, PB, and ED were contracted by Sanofi while the initial phase of this study was conducted.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Parolo, Tomasoni, Bora, Ramponi, Kaddi, Azer, Domenici, Neves-Zaph and Lombardo. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.