biomolecules

MDPI

Article

Darling: A Web Application for Detecting Disease-Related Biomedical Entity Associations with Literature Mining

Evangelos Karatzas [1,*,†], Fotis A. Baltoumas [1,*,†], Ioannis Kasionis [1,†], Despina Sanoudou [2,3,4], Aristides G. Eliopoulos [3,4,5], Theodosios Theodosiou [6], Ioannis Iliopoulos [6] and Georgios A. Pavlopoulos [1,3,*]

1 Institute for Fundamental Biomedical Research, Biomedical Sciences Research Center "Alexander Fleming", 16672 Vari, Greece; gkasionis2@gmail.com
2 Clinical Genomics and Pharmacogenomics Unit, 4th Department of Internal Medicine, School of Medicine, National and Kapodistrian University of Athens, 11527 Athens, Greece; dsanoudou@bioacademy.gr
3 Center for New Biotechnologies and Precision Medicine, School of Medicine, National and Kapodistrian University of Athens, 11527 Athens, Greece; eliopag@med.uoa.gr
4 Biomedical Research Foundation of the Academy of Athens, 4 Soranou Ephessiou Street, 11527 Athens, Greece
5 Department of Biology, School of Medicine, National and Kapodistrian University of Athens, Mikras Asias 75, 11527 Athens, Greece
6 Department of Basic Sciences, School of Medicine, University of Crete, 71003 Heraklion, Greece; theodosios.theodosiou@gmail.com (T.T.); iliopj@med.uoc.gr (I.I.)
* Correspondence: karatzas@fleming.gr (E.K.); baltoumas@fleming.gr (F.A.B.); pavlopoulos@fleming.gr (G.A.P.)
† These authors contributed equally to this work.

Abstract: Finding, exploring and filtering frequent sentence-based associations between a disease and a biomedical entity, co-mentioned in disease-related PubMed literature, is a challenge, as the volume of publications increases. Darling is a web application, which utilizes Name Entity Recognition to identify human-related biomedical terms in PubMed articles, mentioned in OMIM, DisGeNET and Human Phenotype Ontology (HPO) disease records, and generates an interactive biomedical entity association network. Nodes in this network represent genes, proteins, chemicals, functions, tissues, diseases, environments and phenotypes. Users can search by identifiers, terms/entities or free text and explore the relevant abstracts in an annotated format.

Keywords: text-mining; data integration; bioinformatics; named-entity recognition; literature-derived associations

check for updates

Citation: Karatzas, E.; Baltoumas, F.A.; Kasionis, I.; Sanoudou, D.; Eliopoulos, A.G.; Theodosiou, T.; Iliopoulos, I.; Pavlopoulos, G.A. Darling: A Web Application for Detecting Disease-Related Biomedical Entity Associations with Literature Mining. *Biomolecules* **2022**, *12*, 520. https://doi.org/10.3390/biom12040520

Academic Editor: Lukasz Kurgan

Received: 1 March 2022
Accepted: 28 March 2022
Published: 30 March 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

PubMed® today (02/2022) hosts more than 33 million biomedical abstracts, whereas PubMed Central® Open Access Subset (PMC OA Subset) [1] contains more than 7 Million full-text articles. The ever-increasing amount of literature is posing numerous challenges for bioscientists, as parsing these texts and extracting associations among biomedical entities is neither easy nor trivial. This is particularly true for disease-related research, where a wealth of knowledge on the relations between bioentities (genes, proteins, chemicals, etc.) and pathological conditions is available, especially since the rise of high-throughput experimental methods [2]. There is, therefore, a great need for the development of effective and user-friendly methods for the automated recognition, visualization and analysis of disease-related bioentity associations.

Towards this end, several text-mining approaches have been implemented [3–7]. BioTextQuest [8], for example, retrieves PubMed articles and clusters them based on their biomedical terms. DrugQuest [9] applies text mining on the DrugBank database [10], in order to explore drug associations. DISEASES [11] is a system for extracting disease–gene associations from biomedical abstracts. PREGO [12] uses text mining to link microorganisms with environmental processes and functions. Reflect [13] and EXTRACT [14]

perform Named Entity Recognition (NER) on web pages on the fly. FACTA [15] is a text search engine for identifying associated biomedical concepts. OnTheFly [16] parses Office documents, images and PDF files to identify biomedical terms in their text and perform functional enrichment and biological network analysis. CoPub [17] uses Medline abstracts to calculate robust statistics for keyword co-occurrences. NETME [18] offers a knowledge network construction, with term associations in biomedical literature. PubAnnotation [19] is an open, Agile text mining framework to aid researchers throughout the entire annotation process. PubTator [20] provides automated annotations from state-of-the-art text mining systems for genes/proteins, genetic variants, diseases, chemicals, species and cell lines. MetaMap [21] provides access to concepts in the unified medical language system (UMLS) Metathesaurus, from biomedical text. Medline Ranker [22] scores abstracts from Medline, according to a training set of abstracts or a MeSH term. LipiDisease [23] performs disease enrichment analysis on lipids using biomedical literature data. Finally, PESCADOR [24] extracts and analyzes a network of gene and protein interactions from a set of Medline abstracts.

Despite the increasing number of text-mining solutions, effective text mining and analysis of disease-related literature remains challenging. For one thing, the majority of currently available approaches, such as those referenced above, are specialized towards specific bioentity types (e.g., genes, proteins, chemicals, etc.). However, diseases are often complex phenotypes, depending on a multitude of different factors, from gene expression, protein function and chemical substances to cell tissues and even environmental factors. Furthermore, most of these services often offer limited options in the visualization and analysis of their components. To address these challenges, in this article, we present Darling, a novel web application to query scientific publications associated with diseases, identify and visualize bioentities of various types and construct knowledge-based biological interaction networks. Out of a plethora of articles and available databases (reviewed in [25]), we focus on disease-centric repositories and generate a non-redundant set of publications, associated with entries in the OMIM [26], Human Phenotype Ontology (HPO) [27] and DisGeNET [28] databases. The abstracts of the publications are parsed through Named Entity Recognition (NER) to identify a wide range of biomedical terms (genes, chemicals, organisms, ontology terms, diseases, phenotypes and environments). Sentence-based associations among the various biomedical entities are presented in an interactive network [29,30], as well as in searchable and sortable tables, while abstracts are shown in annotated format. Statistics regarding the frequencies of the queried entity types are also presented. Darling is available at http://darling.pavlopouloslab.info or http://bib.fleming.gr:8084/app/darling (accessed on 28 February 2022).

2. Materials and Methods

2.1. Data Collection

The database records (October 2021 data) of OMIM (25,767 entries), HPO (4645 entries) and the human subset of DisGeNET v. 7.0 (30,170 entries) were parsed and their associated publications were isolated, resulting in a non-redundant set of 881,185 articles. The article abstracts were retrieved from PubMed using the Entrez Direct API [31] and were analyzed through NER to isolate bioentities, using the EXTRACT tagger [14,32]. The EXTRACT tagger uses a dictionary-based approach, through which biological and biomedical terms, both canonical and synonyms (e.g., gene name aliases), are assigned to their unique identifiers; thus producing concept-normalized results. The extracted bioentities were assigned to their proper database identifiers, resulting in a non-redundant set of 78,938 terms including genes (protein-coding and other gene types, e.g., micro-RNAs), chemical compounds, Gene Ontology terms, tissues, diseases, organisms, phenotypes and environments. In the dataset, each term is represented by its unique identifier to the relevant database (Table 1), its canonical name, and a number of alternative names/synonyms, as found through the mining of the publications. A knowledge-based interaction network was constructed from these terms (nodes), using their co-occurrence to define interactions (edges). Specifically, two

terms were defined as interaction partners if they were mentioned in the same sentence in the text, with their edge weight defined as the sum of the two terms' co-mentions in the analyzed abstracts. The aforementioned approach resulted in knowledge network consisting of 78,938 nodes and 5,235,076 edges. Table 1 summarizes the number of biomedical terms identified for each category. A flowchart demonstrating the data retrieval and analysis procedure is shown in Figure 1.

Table 1. Identified biomedical terms in a set of 881,185 articles mentioned in OMIM, HPO and DisGeNET databases.

Entity Type	Resource	#Terms
Chemicals	PubChem [33]	23,593
Genes/Proteins	ENSEMBL [34], miRBase [35], Gene Cards [36]	19,731
GO—Biological Process	Gene Ontology [37]	6002
GO—Molecular Function	Gene Ontology [37]	3176
GO—Cellular Component	Gene Ontology [37]	1842
Tissues	BRENDA Tissue Ontology (BTO) [38]	4229
Diseases	Disease Ontology [39], AmyCo [40]	6172
Organisms	NCBI Taxonomy [41]	11,212
Environments	Environmental Ontology (ENVO) [42]	363
Phenotypes	Mammalian Phenotype Ontology [43], Cell Line Data Base (CLDB) [44]	2618

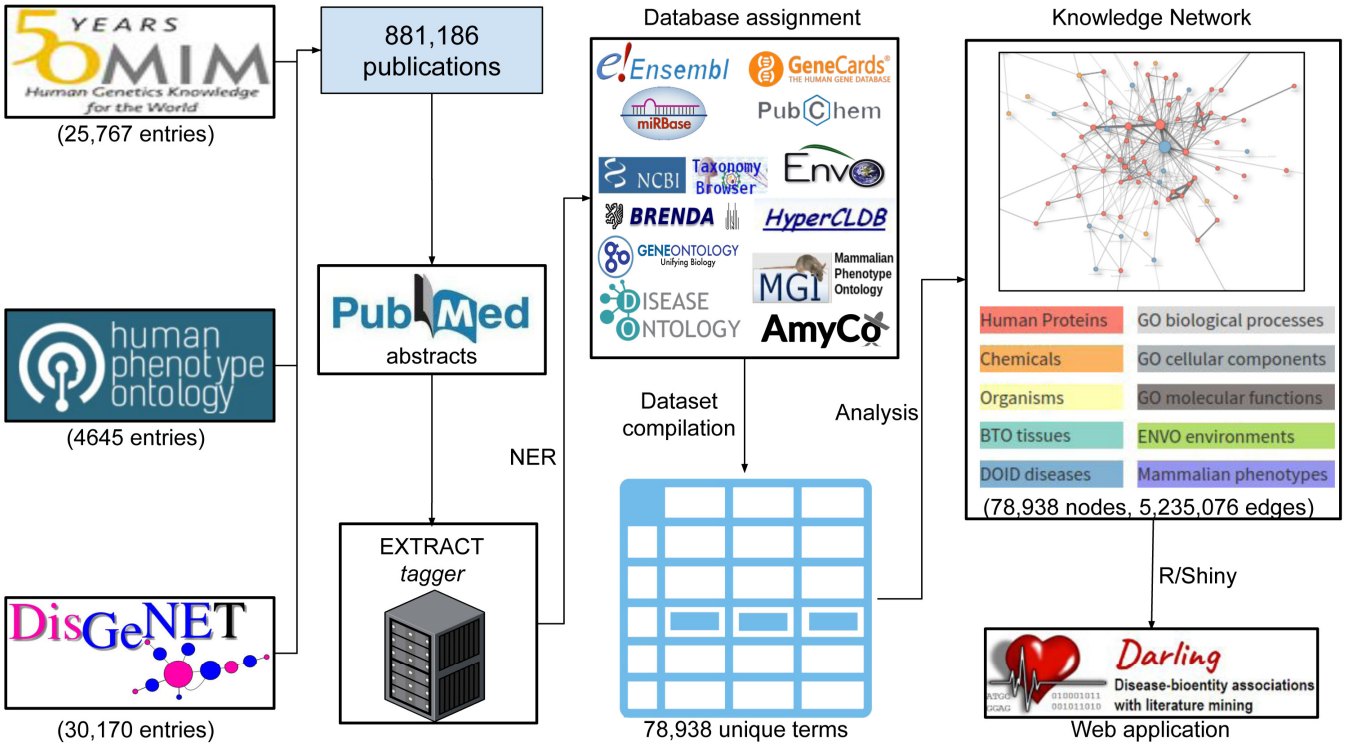

Figure 1. Flowchart of the data retrieval procedure implemented in Darling.

2.2. Darling Application and Analysis

Query: Darling's GUI, offers different query options through three tabs. These are: *(i) Disease Search, (ii) Bioentity Search* and *(iii) Literature Search.*

In the first case, one can directly query any of the OMIM, HPO and DisGeNET databases using their original identifiers or disease names. Users can only query one of

the databases each time but, with each query, one can append the article result list for further analysis. Duplicated retrieved terms are discarded in the next step of the analysis. In the case of free text querying (e.g., disease name), users can force Darling to look for exact matches or substrings in the database's record names. In the case of using database identifiers, users can use lists of IDs separated by spaces or commas to retrieve the results of multiple disease entries.

In the second tab, one can search for a bioentity term using free text and exact or partial string matches. In this case, the user can search for chemicals, proteins or tissues stored in Darling's database and perform a non-disease-centric analysis from a different starting point (e.g., a chemical). Notably, exact matches refer to the bioentity terms identified by the EXTRACT tagging service.

The third option is the most flexible as one can use a list of PubMed identifiers or free text to look for exact or partial matches in article titles. In this case, Darling will search for terms (e.g., "CRISPR-Cas9" or "mir-19") that may not appear in its dictionary or any of the OMIM, HPO and DisGeNET record names.

In every case, after submitting a search query, Darling will fetch all matched articles (Figure 2). The collected articles are then summarized in an interactive and sortable table for review prior to further analysis; thus, users may either keep all retrieved articles or focus on a subset. When multiple search queries are executed, the results of each query can also be filtered to include the intersection (only the common results) or union (all results) of the queries. Users may also choose to filter the NER results and subsequently the retrieved associations by selecting one or more bioentity types (genes, proteins, chemicals, functions, tissues, diseases, environments and phenotypes). Notably, all of these actions can be applied on the set of 881,185 articles mentioned in OMIM, HPO and DisGeNET databases. OMIM's body text is not processed due to license restrictions.

Tables and statistics: Upon selecting articles and applying entity-type filters, Darling will mine all articles of interest and retrieve the corresponding NER results from its database (pre-calculated with the use of EXTRACT [14]). Identified terms are reported in searchable and sortable tables along with their synonyms, official symbols, database identifiers and links to the original source. Identified terms can be reported altogether or separately in corresponding tabs—one per category (genes/proteins, chemicals, functions, tissues, diseases, environments and phenotypes). Extra columns indicate how many times a term was found in the retrieved abstracts as well as in how many articles this term was detected. Interactive ordered bar blots are generated to show such frequencies while interactive pie charts show the overall coverage of terms and articles retrieved for every bioentity category. Finally, word clouds show the most common terms, scaled by their frequency.

Network: In addition to the tables, Darling generates an interactive association network of the identified bioentities. Network nodes may fall into any of the identified bioentity types (genes/proteins, chemicals, organisms, GO terms, tissues, diseases, environments and phenotypes) and are assigned a certain color (distinct per category). Node sizes can be adjusted according to how many times they were identified in a selected set of abstracts (total frequency) whereas network edges can be interactively filtered according to the total times two adjacent entities were located in the same sentence (edge weight). At any stage of analysis, users may limit the visible nodes to certain bioentity types. For aesthetical convenience, users may adjust the network view using various offered layouts [45]. Characteristic examples are the force-directed ones such as Fruchterman–Reingold [46] and Kamada–Kawai [47] or the plain ones such as grid, random and circular layout. The network is fully interactive and comes with control buttons for positioning, zooming and recentering. Nodes can be dragged and positioned anywhere on the plane.

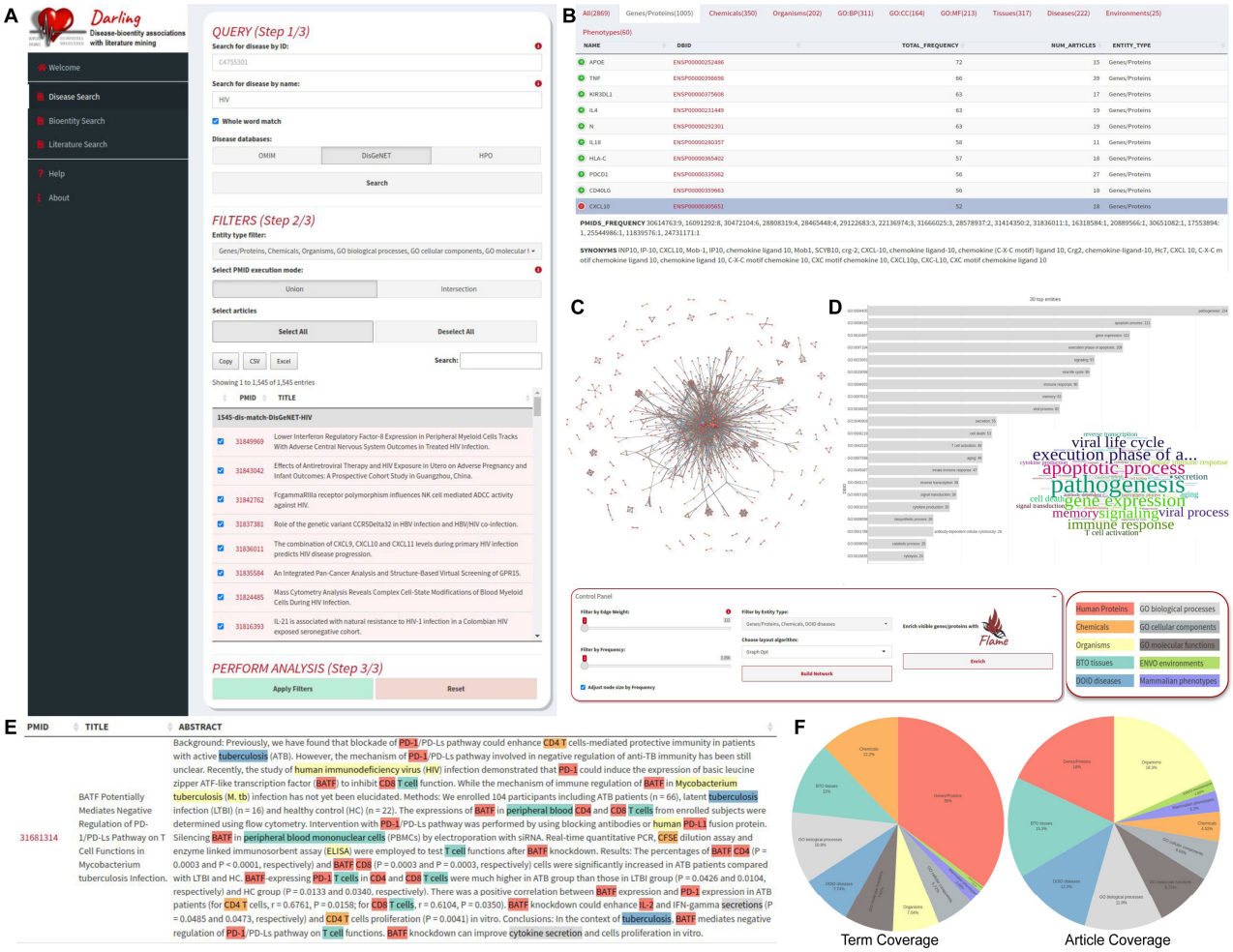

Figure 2. The Darling Graphical User Interface of Darling. (**A**) The input form of the *Disease Search* query. Users can perform searches using a disease's name or database identifier, against the data retrieved from OMIM, HPO or DisGeNET. In the example, the term "HIV" is searched against DisGeNET. The search form initially returns the publications associated with the disease. Users can then choose the publications and entity types of their interest and perform an analysis, using the form elements at the bottom of the page. (**B**) Excerpts of the results retrieved for the search. A total of 2869 entities have been retrieved and organized in distinct categories. For each term, its database identifier, canonical name, synonym terms and associated PubMed identifiers (PMIDs) are shown; in addition, the frequency of the terms, both total and for each distinct publication, is calculated. (**C**) A knowledge-based association network generated by the search results. Users can adjust the elements of the network and select from a list of different visualization layouts. (**D**) Bar-plot (left) and word cloud (right) representations of the most frequent GO biological processes associated with HIV. (**E**) A graphical representation of one of the publication abstracts, with extracted bioentities highlighted in color. (**F**) Pie charts of the overall coverage of terms and articles retrieved for every bioentity category.

Besides visualization, similarly to the NAP application [48,49] or Cytoscape's Network analyzer [50], Darling offers basic network topological analysis where users can see numerical values for the numbers of nodes and edges, density, modularity, radius, average path length, average connectivity, average clustering coefficient, betweenness and eccentricity centrality.

For more comprehensive visualization and analysis, at any stage, the network's edge list can be exported in a tab-delimited file format and visualized with external viewers [51–53] (e.g., Cytoscape [54], Gephi [55], NORMA [56], Arena3D^{web} [57]). Bidirectional edges (e.g., AB-BA) are kept only once.

Annotated text: At any stage of the analysis, the relevant PubMed article abstracts are reported in an annotated format in a separate table. Users can read these abstracts with the identified terms highlighted in different colors according to the tagged entity category. On mouse-hovering or clicking over a term, a popup window with relevant links to the corresponding databases is generated on-the-fly.

Functional Enrichment: After the network generation and the application of any filtering options, all of the visible identified genes and proteins can be sent to the Flame application [58] for functional and literature enrichment analysis. Genes and Proteins will be first converted to ENSEMBL identifiers and can then be analyzed for KEGG [59,60], Reactome [61,62] and Wiki Pathways [63] or for the biological functions [37] they are involved in. Flame utilizes g:Profiler [64] and aGOTool [65] at its backend for functional and literature enrichment and offers appealing visualizations for easier interpretation of the reported results. In addition, Flame can construct protein–protein interaction networks, by retrieving evidence from the STRING database [66].

2.3. Implementation

Darling is organized in a *MySQL* database which is periodically updated. The GUI and backend are mainly written in *R/Shiny*. The interactive network is visualized with the *R/visNetwork* library and network topological analysis is performed using the *R/igraph* library [67]. Plots are generated with the use of *R/Plotly* [68], while wordclouds with the *R/wordcloud2* library. The *EXTRACT* API [14] is utilized to display popup windows for bioentity terms in the annotated abstracts.

3. Results

3.1. Investigating the Link between Obesity and Cardiovascular Diseases

To demonstrate Darling's capacity for the extraction of biological information and knowledge discovery, we investigated genes and pathways that may link cardiovascular disease (CVD) to obesity. We queried DisGeNET, using the disease term "cardiovascular", and obtained the 5000 most recent articles, 100 of which also contained the term "obesity". This group entailed 317 entities (Figure 3A) that included 109 unique genes/proteins associated with "insulin receptor signaling", "metabolic disease", "energy homeostasis" and "cytolysis" gene ontology (GO) biological processes (Figure 3B). By using Darling, we constructed a co-occurrence network of genes, phenotypes and tissues predicted to link CVD to obesity (Figure 3C). A major neighborhood in this network (subnetwork 1) is associated with "insulin resistance", "abnormal inflammatory response" and "cardiac hypertrophy" and linked to the adipose tissue, liver and blood. This group mostly entails components of the adiponectin pathway, including adiponectin (ADIPOQ), its receptors ADIPOR1 and ADIPOR2, and their downstream adaptors APPL1/2, which transduce the anti-atherogenic and anti-inflammatory effects of adiponectin. The group also includes the inflammation marker CRP, which is elevated in both obesity and CVD, and GAS6, which has been implicated in atherosclerosis, thrombosis and innate immune reactions [69].

The Darling co-occurrence network also indicated an interaction between fat mass and obesity-associated protein (FTO) and apolipoprotein E (APOE), linked to inflammatory processes (Figure 3C). Several FTO polymorphisms are associated with increased risk for weight gain [70] and the APOE ε4 variant is a genetic risk factor for atherosclerosis and CVD in humans [71]. Experimental evidence suggests that expression of APOE ε4 leads to elevated intracellular and circulating cholesterol levels and heightened inflammatory reactions compared to other variants [71]. A putative mechanistic link between APOE and FTO is underscored by studies showing that overexpression of FTO in APOE-deficient mice reduces cholesterol and inflammatory cytokine synthesis by macrophages and alleviates atherosclerosis associated with the absence of APOE [72].

Another neighborhood of interest identified by Darling (subnetwork 2; Figure 3C) entails the growth differentiation factor 15 (GDF15) and its receptor, GFRAL. Circulating GDF15 crosses the blood brain axis to bind GFRAL in neurons of the hindbrain, leading

to reduced appetite and food intake. The serum levels of GDF15 dramatically increase in cancer-associated cachexia but are also found elevated in obesity, presumably acting as a compensatory mechanism to reduce appetite [73]. GDF15 has been reported as a prospective biomarker of CVD and independent predictor of all-cause mortality [74,75]. Interestingly, CRP has been found to induce the expression of GDF15 [76].

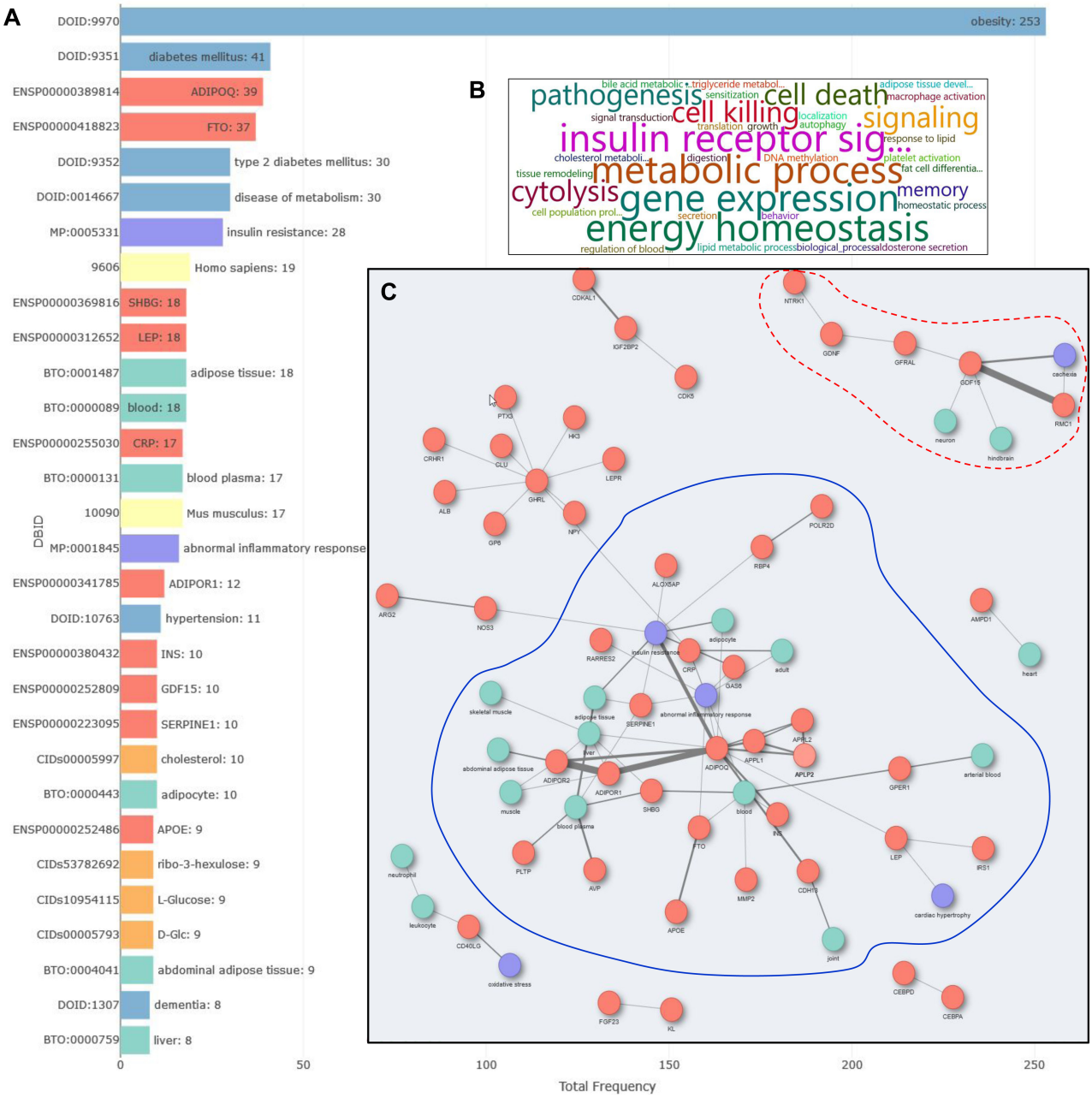

Figure 3. Assessment of Darling's capacity for extraction of biological information and knowledge discovery. (**A**) Top entities (chemicals, phenotypes, proteins/genes, organisms, GO:BP, GO:CC and GO:MF, tissues, environments and diseases) plotted against frequency of occurrence in 100 DisGeNET-related publications. (**B**) GO Biological processes predicted to be mostly associated with cardiovascular disease. (**C**) Co-occurrence network depicting genes (orange circles), phenotypes (purple circles) and tissues (green circles) predicted to link cardiovascular disease and obesity. Subnetwork 1, discussed in the Results section, is demarcated by the blue line and subnetwork 2 by the red line.

Overall, the aforementioned observations demonstrate the capacity of Darling for the extraction of biological information and knowledge discovery.

3.2. Querying Multiple Disease Databases Simultaneously with Darling

In a second case study, we asked whether Darling could facilitate the extraction of biological information on a disease by combining several disease libraries. To this end, we interrogated OMIM, HPO and DisGeNET for Cornelia de Lange, a rare genetic syndrome characterized by slow growth rates, leading to short stature, intellectual disability that ranges from moderate to severe, congenital heart defects and bone abnormalities, among others. Through Darling, we queried OMIM for "Cornelia de Lange" and obtained 127 entries that generated 264 entities. By exploring the same query against the OMIM, HPO and DisGeNET compendiums together, Darling retrieved 318 entries that generated 292 unique articles and 712 entities. The co-occurrence network of genes/proteins, GO Biological Processes and DOID diseases derived by these 712 entities yielded superior information compared to the respective network derived from OMIM only (Figure 4).

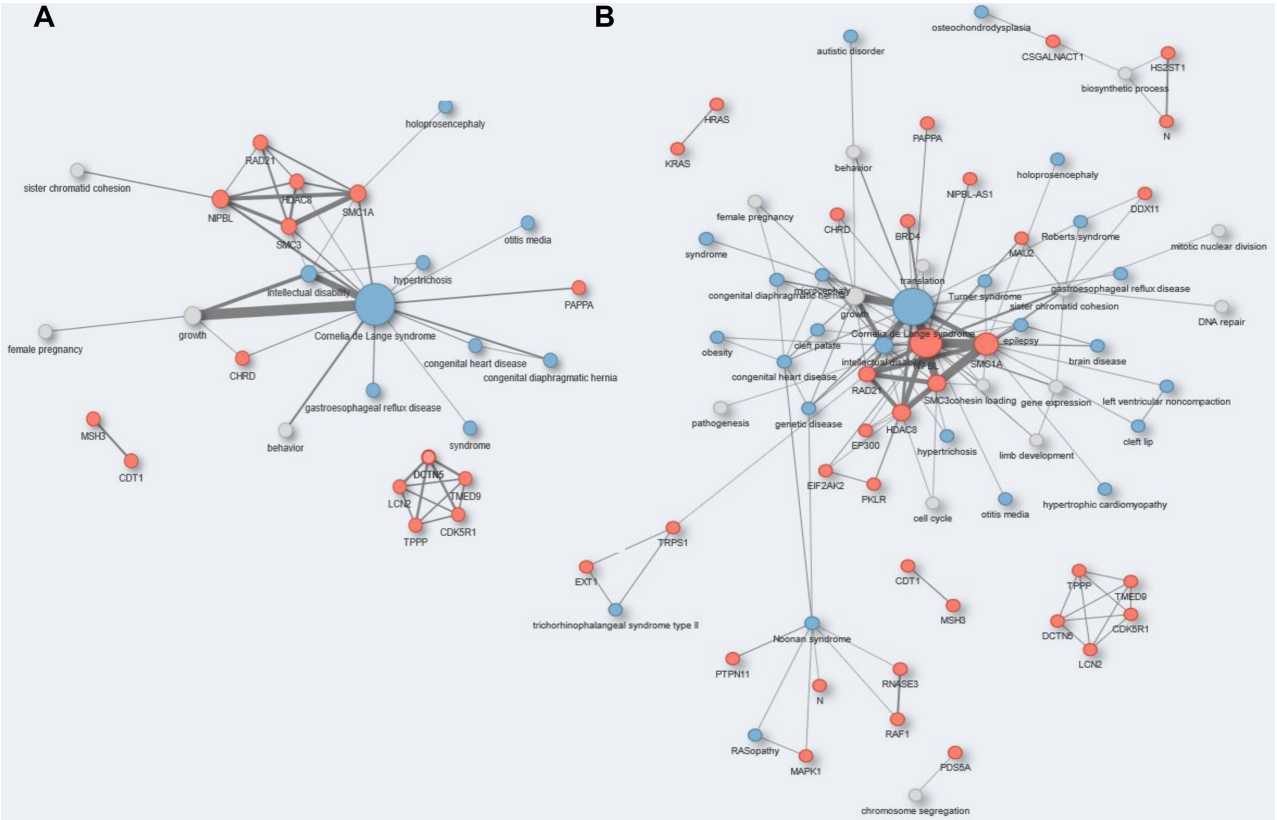

Figure 4. Networks of genes/proteins, GO Biological Processes and DOID diseases generated by Darling query of OMIM (**A**) or OMIM, HPO and DisGeNET compendiums together (**B**) for "Cornelia de Lange". Entities (264 and 712, respectively) were used to build co-occurrence networks (filter by frequency = 5).

A major neighborhood in both networks (Figure 4A,B) contained the NPBL, SMC1A, SMC3, HDAC8 and RAD21 genes, which are found mutated in >80% of Cornelia de Lange patients. These genes encode for regulators of the cohesin complex and are involved in chromosome condensation, chromosome segregation and DNA repair. Additional genes found exclusively in the network generated from OMIM, HPO and DisGeNET include BRD4 and MAU2. Mutations in both genes have recently been detected in Cornelia de Lange patients and have been functionally implicated in disease pathogenesis through their interaction with NPBL [77,78].

Interactions with Wilson–Turner and Roberts syndromes were also indicated in this network (Figure 4B). Roberts syndrome bears developmental abnormalities, similar to Cornelia de Lange, such as limb abnormalities, retarded growth and intellectual impairment. Mechanistically, Roberts syndrome has been linked to mutations in ESCO2 gene, which encodes a cohesin acetyltransferase and modulator of double strand break repair [79]. Wilson–Turner syndrome is a rare X-linked multisystem genetic disease that also manifests with intellectual disability, dysmorphic facial features and short stature and has been linked to a mutation in the HDAC8 gene [80]. Overall, the aforementioned examples demonstrate the capacity of Darling for the extraction of biological information and acceleration of knowledge discovery.

4. Discussion

Darling is a text-mining application, aiming to aid researchers in associating different biomedical entities in a knowledge network, generated by literature mining. A great advantage of Darling is its high quality back-end NER tagger, which makes it more competitive compared to other similar applications, both in terms of annotation and data integration. In addition, Darling only focuses on a subset of disease-centric articles, which have been manually curated in the OMIM, HPO and DisGeNET databases, rather than the whole PubMed space. Taking into account that PubMed currently contains many review articles and has also recently started to support preprints [81], we believe that this is the safest approach, in order to eliminate possible false-positive term associations. Nevertheless, we plan to extend Darling's functionality in the future and cover literature coming from more databases, as well as support full text articles.

In its core, Darling contains a relational database, consisting of all relevant bioentity information and associations. Term frequencies per article, their respective canonical names and the relative tagged documents are all pre-calculated, further speeding up the execution of the application, and are served via an interactive GUI. Therefore, Darling does not depend on external web services, as opposed to other similar applications (e.g., NETME), which query the various databases (e.g., PubMed) on the fly, resulting in time-consuming requests. This may secure an always up-to-date information schema but comes at the cost of speed, performance and web-service dependencies. To keep up to date, Darling's database will be annually updated, including new OMIM, HPO and DisGeNET entries, as well as their associated publications and extracted bioentities. Furthermore, in future versions, Darling will implement additional databases, support more model organisms and enable the detection of abstract-based associations (currently only offers sentenced-based), something which may increase the network's complexity.

Overall, we believe that Darling outperforms most of the currently available tools, in terms of performance, variety of identified entity terms and quality of results. It is a powerful tool, which can simplify the way researchers query and explore existing knowledge, while also identifying novel indirect associations among biomedical entities, which may be the pivot elements for new hypotheses and discoveries.

Author Contributions: Conceptualization, G.A.P. and I.I.; methodology, I.I., G.A.P. and T.T.; software, E.K., F.A.B. and I.K.; validation, A.G.E. and D.S. formal analysis, A.G.E. and T.T.; investigation, I.K.; data curation, E.K. and F.A.B.; writing—original draft preparation, G.A.P., E.K., F.A.B., A.G.E., D.S. and I.I.; visualization, I.K.; supervision, G.A.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I) under the "First Call for H.F.R.I Research Projects to support faculty members and researchers and the procurement of high-cost research equipment grant", Grant ID: 1855-BOLOGNA. GAP was also supported by the project 'The Greek Research Infrastructure for Personalized Medicine (pMedGR)' (MIS 5002802), which is implemented under the Action 'Reinforcement of the Research and Innovation Infrastructure', funded by the Operational Program 'Competitiveness, Entrepreneurship and Innovation' (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Darling is available online at http://darling.pavlopouloslab.info (accessed on 28 February 2022).

Acknowledgments: We would like to thank Yorgos Sofianatos, supported by the Marie Skłodowska-Curie Individual Fellowships—MSCA-IF-EF-CAR (Grant ID: 838018—H2020-MSCA-IF-2018).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Roberts, R.J. PubMed Central: The GenBank of the published literature. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 381–382. [CrossRef] [PubMed]
2. Lightbody, G.; Haberland, V.; Browne, F.; Taggart, L.; Zheng, H.; Parkes, E.; Blayney, J.K. Review of applications of high-throughput sequencing in personalized medicine: Barriers and facilitators of future progress in research and clinical application. *Brief. Bioinform.* **2019**, *20*, 1795–1811. [CrossRef] [PubMed]
3. Cheerkoot-Jalim, S.; Khedo, K.K. A systematic review of text mining approaches applied to various application areas in the biomedical domain. *J. Knowl. Manag.* **2020**, *25*, 642–668. [CrossRef]
4. Przybyła, P.; Shardlow, M.; Aubin, S.; Bossy, R.; Eckart de Castilho, R.; Piperidis, S.; McNaught, J.; Ananiadou, S. Text mining resources for the life sciences. *Database* **2016**, *2016*, baw145. [CrossRef] [PubMed]
5. Rebholz-Schuhmann, D.; Oellrich, A.; Hoehndorf, R. Text-mining solutions for biomedical research: Enabling integrative biology. *Nat. Rev. Genet.* **2012**, *13*, 829–839. [CrossRef] [PubMed]
6. Wang, L.L.; Lo, K. Text mining approaches for dealing with the rapidly expanding literature on COVID-19. *Brief. Bioinform.* **2021**, *22*, 781–799. [CrossRef] [PubMed]
7. Papanikolaou, N.; Pavlopoulos, G.A.; Theodosiou, T.; Iliopoulos, I. Protein-protein interaction predictions using text mining methods. *Methods S. Diego Calif.* **2015**, *74*, 47–53. [CrossRef]
8. Papanikolaou, N.; Pavlopoulos, G.A.; Pafilis, E.; Theodosiou, T.; Schneider, R.; Satagopam, V.P.; Ouzounis, C.A.; Eliopoulos, A.G.; Promponas, V.J.; Iliopoulos, I. BioTextQuest(+): A knowledge integration platform for literature mining and concept discovery. *Bioinforma. Oxf. Engl.* **2014**, *30*, 3249–3256. [CrossRef]
9. Papanikolaou, N.; Pavlopoulos, G.A.; Theodosiou, T.; Vizirianakis, I.S.; Iliopoulos, I. DrugQuest—A text mining workflow for drug association discovery. *BMC Bioinform.* **2016**, *17*, 182. [CrossRef]
10. Wishart, D.S.; Feunang, Y.D.; Guo, A.C.; Lo, E.J.; Marcu, A.; Grant, J.R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074–D1082. [CrossRef]
11. Pletscher-Frankild, S.; Pallejà, A.; Tsafou, K.; Binder, J.X.; Jensen, L.J. DISEASES: Text mining and data integration of disease-gene associations. *Methods S. Diego Calif.* **2015**, *74*, 83–89. [CrossRef]
12. Zafeiropoulos, H.; Paragkamian, S.; Ninidakis, S.; Pavlopoulos, G.A.; Jensen, L.J.; Pafilis, E. PREGO: A Literature and Data-Mining Resource to Associate Microorganisms, Biological Processes, and Environment Types. *Microorganisms* **2022**, *10*, 293. [CrossRef] [PubMed]
13. Pafilis, E.; O'Donoghue, S.I.; Jensen, L.J.; Horn, H.; Kuhn, M.; Brown, N.P.; Schneider, R. Reflect: Augmented browsing for the life scientist. *Nat. Biotechnol.* **2009**, *27*, 508–510. [CrossRef] [PubMed]
14. Pafilis, E.; Buttigieg, P.L.; Ferrell, B.; Pereira, E.; Schnetzer, J.; Arvanitidis, C.; Jensen, L.J. EXTRACT: Interactive extraction of environment metadata and term suggestion for metagenomic sample annotation. *Database J. Biol. Databases Curation* **2016**, *2016*, baw005. [CrossRef]
15. Tsuruoka, Y.; Tsujii, J.; Ananiadou, S. FACTA: A text search engine for finding associated biomedical concepts. *Bioinformatics* **2008**, *24*, 2559–2560. [CrossRef] [PubMed]
16. Baltoumas, F.A.; Zafeiropoulou, S.; Karatzas, E.; Paragkamian, S.; Thanati, F.; Iliopoulos, I.; Eliopoulos, A.G.; Schneider, R.; Jensen, L.J.; Pafilis, E.; et al. OnTheFly2.0: A text-mining web application for automated biomedical entity recognition, document annotation, network and functional enrichment analysis. *NAR Genom. Bioinform.* **2021**, *3*, lqab090. [CrossRef]
17. Fleuren, W.W.M.; Verhoeven, S.; Frijters, R.; Heupers, B.; Polman, J.; van Schaik, R.; de Vlieg, J.; Alkema, W. CoPub update: CoPub 5.0 a text mining system to answer biological questions. *Nucleic Acids Res.* **2011**, *39*, W450–W454. [CrossRef]
18. Muscolino, A.; Di Maria, A.; Rapicavoli, R.V.; Alaimo, S.; Bellomo, L.; Billeci, F.; Borzì, S.; Ferragina, P.; Ferro, A.; Pulvirenti, A. NETME: On-the-fly knowledge network construction from biomedical literature. *Appl. Netw. Sci.* **2022**, *7*, 1–24. [CrossRef]
19. Kim, J.-D.; Wang, Y.; Fujiwara, T.; Okuda, S.; Callahan, T.J.; Cohen, K.B. Open Agile text mining for bioinformatics: The PubAnnotation ecosystem. *Bioinformatics* **2019**, *35*, 4372–4380. [CrossRef]
20. Wei, C.-H.; Kao, H.-Y.; Lu, Z. PubTator: A web-based text mining tool for assisting biocuration. *Nucleic Acids Res.* **2013**, *41*, W518–W522. [CrossRef]
21. Aronson, A.R.; Lang, F.-M. An overview of MetaMap: Historical perspective and recent advances. *J. Am. Med. Inform. Assoc.* **2010**, *17*, 229–236. [CrossRef]

22. Fontaine, J.-F.; Barbosa-Silva, A.; Schaefer, M.; Huska, M.R.; Muro, E.M.; Andrade-Navarro, M.A. MedlineRanker: Flexible ranking of biomedical literature. *Nucleic Acids Res.* **2009**, *37*, W141–W146. [CrossRef] [PubMed]
23. More, P.; Bindila, L.; Wild, P.; Andrade-Navarro, M.; Fontaine, J.-F. LipiDisease: Associate lipids to diseases using literature mining. *Bioinformatics* **2021**, *37*, 3981–3982. [CrossRef] [PubMed]
24. Barbosa-Silva, A.; Fontaine, J.-F.; Donnard, E.R.; Stussi, F.; Ortega, J.M.; Andrade-Navarro, M.A. PESCADOR, a web-based tool to assist text-mining of biointeractions extracted from PubMed queries. *BMC Bioinform.* **2011**, *12*, 435. [CrossRef] [PubMed]
25. Baltoumas, F.A.; Zafeiropoulou, S.; Karatzas, E.; Koutrouli, M.; Thanati, F.; Voutsadaki, K.; Gkonta, M.; Hotova, J.; Kasionis, I.; Hatzis, P.; et al. Biomolecule and Bioentity Interaction Databases in Systems Biology: A Comprehensive Review. *Biomolecules* **2021**, *11*, 1245. [CrossRef]
26. Amberger, J.S.; Bocchini, C.A.; Schiettecatte, F.; Scott, A.F.; Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* **2015**, *43*, D789. [CrossRef]
27. Köhler, S.; Gargano, M.; Matentzoglu, N.; Carmody, L.C.; Lewis-Smith, D.; Vasilevsky, N.A.; Danis, D.; Balagura, G.; Baynam, G.; Brower, A.M.; et al. The Human Phenotype Ontology in 2021. *Nucleic Acids Res.* **2021**, *49*, D1207–D1217. [CrossRef]
28. Piñero, J.; Ramírez-Anguita, J.M.; Saüch-Pitarch, J.; Ronzano, F.; Centeno, E.; Sanz, F.; Furlong, L.I. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* **2019**, *48*, gkz1021. [CrossRef]
29. Koutrouli, M.; Karatzas, E.; Paez-Espino, D.; Pavlopoulos, G.A. A Guide to Conquer the Biological Network Era Using Graph Theory. *Front. Bioeng. Biotechnol.* **2020**, *8*, 34. [CrossRef]
30. Pavlopoulos, G.A.; Secrier, M.; Moschopoulos, C.N.; Soldatos, T.G.; Kossida, S.; Aerts, J.; Schneider, R.; Bagos, P.G. Using graph theory to analyze biological networks. *BioData Min.* **2011**, *4*, 10. [CrossRef] [PubMed]
31. Kans, J. *Entrez Direct: E-Utilities on the Unix Command Line*; National Center for Biotechnology Information (US): Rockville, MD, USA, 2022.
32. Pafilis, E.; Jensen, L.J. Real-time tagging of biomedical entities. *BioRxiv* **2016**, 078469. [CrossRef]
33. Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B.A.; Thiessen, P.A.; Yu, B.; et al. PubChem in 2021: New data content and improved web interfaces. *Nucleic Acids Res.* **2021**, *49*, D1388–D1395. [CrossRef] [PubMed]
34. Howe, K.L.; Achuthan, P.; Allen, J.; Allen, J.; Alvarez-Jarreta, J.; Amode, M.R.; Armean, I.M.; Azov, A.G.; Bennett, R.; Bhai, J.; et al. Ensembl 2021. *Nucleic Acids Res.* **2021**, *49*, D884–D891. [CrossRef] [PubMed]
35. Kozomara, A.; Birgaoanu, M.; Griffiths-Jones, S. miRBase: From microRNA sequences to function. *Nucleic Acids Res.* **2019**, *47*, D155–D162. [CrossRef] [PubMed]
36. Stelzer, G.; Rosen, N.; Plaschkes, I.; Zimmerman, S.; Twik, M.; Fishilevich, S.; Stein, T.I.; Nudel, R.; Lieder, I.; Mazor, Y.; et al. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr. Protoc. Bioinform.* **2016**, *54*, 1–30. [CrossRef]
37. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **2004**, *32*, D258–D261. [CrossRef] [PubMed]
38. Chang, A.; Schomburg, I.; Placzek, S.; Jeske, L.; Ulbrich, M.; Xiao, M.; Sensen, C.W.; Schomburg, D. BRENDA in 2015: Exciting developments in its 25th year of existence. *Nucleic Acids Res.* **2015**, *43*, D439–D446. [CrossRef]
39. Schriml, L.M.; Mitraka, E.; Munro, J.; Tauber, B.; Schor, M.; Nickle, L.; Felix, V.; Jeng, L.; Bearer, C.; Lichenstein, R.; et al. Human Disease Ontology 2018 update: Classification, content and workflow expansion. *Nucleic Acids Res.* **2019**, *47*, D955–D962. [CrossRef]
40. Nastou, K.C.; Nasi, G.I.; Tsiolaki, P.L.; Litou, Z.I.; Iconomidou, V.A. AmyCo: The amyloidoses collection. *Amyloid* **2019**, *26*, 112–117. [CrossRef]
41. Schoch, C.L.; Ciufo, S.; Domrachev, M.; Hotton, C.L.; Kannan, S.; Khovanskaya, R.; Leipe, D.; Mcveigh, R.; O'Neill, K.; Robbertse, B.; et al. NCBI Taxonomy: A comprehensive update on curation, resources and tools. *Database* **2020**, *2020*, baaa062. [CrossRef]
42. Buttigieg, P.L.; Morrison, N.; Smith, B.; Mungall, C.J.; Lewis, S.E. ENVO Consortium The environment ontology: Contextualising biological and biomedical entities. *J. Biomed. Semant.* **2013**, *4*, 43. [CrossRef] [PubMed]
43. Smith, C.L.; Eppig, J.T. The mammalian phenotype ontology: Enabling robust annotation and comparative analysis. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **2009**, *1*, 390–399. [CrossRef] [PubMed]
44. Romano, P.; Manniello, A.; Aresu, O.; Armento, M.; Cesaro, M.; Parodi, B. Cell Line Data Base: Structure and recent improvements towards molecular authentication of human cell lines. *Nucleic Acids Res.* **2009**, *37*, D925–D932. [CrossRef] [PubMed]
45. Pavlopoulos, G.A.; Paez-Espino, D.; Kyrpides, N.C.; Iliopoulos, I. Empirical Comparison of Visualization Tools for Larger-Scale Network Analysis. *Adv. Bioinform.* **2017**, *2017*, 1278932. [CrossRef]
46. Fruchterman, T.M.J.; Reingold, E.M. Graph drawing by force-directed placement. *Softw. Pract. Exp.* **1991**, *21*, 1129–1164. [CrossRef]
47. Kamada, T.; Kawai, S. An algorithm for drawing general undirected graphs. *Inf. Process. Lett.* **1989**, *31*, 7–15. [CrossRef]
48. Theodosiou, T.; Efstathiou, G.; Papanikolaou, N.; Kyrpides, N.C.; Bagos, P.G.; Iliopoulos, I.; Pavlopoulos, G.A. NAP: The Network Analysis Profiler, a web tool for easier topological analysis and comparison of medium-scale biological networks. *BMC Res. Notes* **2017**, *10*, 278. [CrossRef]
49. Koutrouli, M.; Theodosiou, T.; Iliopoulos, I.; Pavlopoulos, G.A. The Network Analysis Profiler (NAP v2.0): A web tool for visual topological comparison between multiple networks. *EMBnet. J.* **2021**, *26*, e943. [CrossRef]

50. Assenov, Y.; Ramírez, F.; Schelhorn, S.-E.; Lengauer, T.; Albrecht, M. Computing topological parameters of biological networks. *Bioinformatics* **2008**, *24*, 282–284. [CrossRef]
51. Gehlenborg, N.; O'Donoghue, S.I.; Baliga, N.S.; Goesmann, A.; Hibbs, M.A.; Kitano, H.; Kohlbacher, O.; Neuweger, H.; Schneider, R.; Tenenbaum, D.; et al. Visualization of omics data for systems biology. *Nat. Methods* **2010**, *7*, S56–S68. [CrossRef]
52. Pavlopoulos, G.A.; Wegener, A.-L.; Schneider, R. A survey of visualization tools for biological network analysis. *BioData Min.* **2008**, *1*, 12. [CrossRef]
53. Pavlopoulos, G.A.; Malliarakis, D.; Papanikolaou, N.; Theodosiou, T.; Enright, A.J.; Iliopoulos, I. Visualizing genome and systems biology: Technologies, tools, implementation techniques and trends, past, present and future. *GigaScience* **2015**, *4*, 38. [CrossRef] [PubMed]
54. Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N.S.; Wang, J.T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **2003**, *13*, 2498–2504. [CrossRef] [PubMed]
55. Bastian, M.; Heymann, S.; Jacomy, M. Gephi: An Open Source Software for Exploring and Manipulating Networks. *Proc. Int. AAAI Conf. Web Soc. Media* **2009**, *3*, 361–362.
56. Koutrouli, M.; Karatzas, E.; Papanikolopoulou, K.; Pavlopoulos, G.A. NORMA: The Network Makeup Artist—A Web Tool for Network Annotation Visualization. *Genom. Proteom. Bioinform.* **2021**, S1672022921001303. [CrossRef]
57. Karatzas, E.; Baltoumas, F.A.; Panayiotou, N.A.; Schneider, R.; Pavlopoulos, G.A. Arena3Dweb: Interactive 3D visualization of multilayered networks. *Nucleic Acids Res.* **2021**, *49*, W36–W45. [CrossRef] [PubMed]
58. Thanati, F.; Karatzas, E.; Baltoumas, F.A.; Stravopodis, D.J.; Eliopoulos, A.G.; Pavlopoulos, G.A. FLAME: A Web Tool for Functional and Literature Enrichment Analysis of Multiple Gene Lists. *Biology* **2021**, *10*, 665. [CrossRef] [PubMed]
59. Kanehisa, M.; Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **2000**, *28*, 27–30. [CrossRef] [PubMed]
60. Okuda, S.; Yamada, T.; Hamajima, M.; Itoh, M.; Katayama, T.; Bork, P.; Goto, S.; Kanehisa, M. KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Res.* **2008**, *36*, W423–W426. [CrossRef]
61. Fabregat, A.; Jupe, S.; Matthews, L.; Sidiropoulos, K.; Gillespie, M.; Garapati, P.; Haw, R.; Jassal, B.; Korninger, F.; May, B.; et al. The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* **2018**, *46*, D649–D655. [CrossRef]
62. Koutrouli, M.; Hatzis, P.; Pavlopoulos, G.A. Exploring Networks in the STRING and Reactome Database. In *Systems Medicine*; Wolkenhauer, O., Ed.; Academic Press: Oxford, UK, 2021; pp. 507–520, ISBN 978-0-12-816078-7.
63. Martens, M.; Ammar, A.; Riutta, A.; Waagmeester, A.; Slenter, D.N.; Hanspers, K.; A Miller, R.; Digles, D.; Lopes, E.N.; Ehrhart, F.; et al. WikiPathways: Connecting communities. *Nucleic Acids Res.* **2021**, *49*, D613–D621. [CrossRef] [PubMed]
64. Raudvere, U.; Kolberg, L.; Kuzmin, I.; Arak, T.; Adler, P.; Peterson, H.; Vilo, J. g: Profiler: A web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **2019**, *47*, W191–W198. [CrossRef] [PubMed]
65. Schölz, C.; Lyon, D.; Refsgaard, J.C.; Jensen, L.J.; Choudhary, C.; Weinert, B.T. Avoiding abundance bias in the functional annotation of post-translationally modified proteins. *Nat. Methods* **2015**, *12*, 1003–1004. [CrossRef] [PubMed]
66. Szklarczyk, D.; Gable, A.L.; Nastou, K.C.; Lyon, D.; Kirsch, R.; Pyysalo, S.; Doncheva, N.T.; Legeay, M.; Fang, T.; Bork, P.; et al. The STRING database in 2021: Customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* **2021**, *49*, D605–D612. [CrossRef] [PubMed]
67. Csardi, G.; Nepusz, T. The igraph software package for complex network research. *InterJ. Complex Syst.* **2006**, *1695*, 1–9.
68. Sievert, C. *Interactive Web-Based Data Visualization with R, Plotly, and Shiny*; CRC Press, Taylor and Francis Group: Boca Raton, FL, USA, 2020; ISBN 978-1-138-33149-5.
69. Laurance, S.; Lemarié, C.A.; Blostein, M.D. Growth Arrest-Specific Gene 6 (gas6) and Vascular Hemostasis. *Adv. Nutr.* **2012**, *3*, 196–203. [CrossRef]
70. Gkouskou, K.; Vlastos, I.; Karkalousos, P.; Chaniotis, D.; Sanoudou, D.; Eliopoulos, A.G. The "Virtual Digital Twins" Concept in Precision Nutrition. *Adv. Nutr.* **2020**, *11*, 1405–1413. [CrossRef]
71. Gkouskou, K.; Vasilogiannakopoulou, T.; Andreakos, E.; Davanos, N.; Gazouli, M.; Sanoudou, D.; Eliopoulos, A.G. COVID-19 enters the expanding network of apolipoprotein E4-related pathologies. *Redox Biol.* **2021**, *41*, 101938. [CrossRef]
72. Mo, C.; Yang, M.; Han, X.; Li, J.; Gao, G.; Tai, H.; Huang, N.; Xiao, H. Fat mass and obesity-associated protein attenuates lipid accumulation in macrophage foam cells and alleviates atherosclerosis in apolipoprotein E-deficient mice. *J. Hypertens.* **2017**, *35*, 810–821. [CrossRef]
73. Breit, S.N.; Brown, D.A.; Tsai, V.W.-W. The GDF15-GFRAL Pathway in Health and Metabolic Disease: Friend or Foe? *Annu. Rev. Physiol.* **2021**, *83*, 127–151. [CrossRef]
74. Hagström, E.; Held, C.; Stewart, R.A.H.; Aylward, P.E.; Budaj, A.; Cannon, C.P.; Koenig, W.; Krug-Gourley, S.; Mohler, E.R., III; Steg, P.G.; et al. Growth Differentiation Factor 15 Predicts All-Cause Morbidity and Mortality in Stable Coronary Heart Disease. *Clin. Chem.* **2017**, *63*, 325–333. [CrossRef] [PubMed]
75. Wiklund, F.E.; Bennet, A.M.; Magnusson, P.K.E.; Eriksson, U.K.; Lindmark, F.; Wu, L.; Yaghoutyfam, N.; Marquis, C.P.; Stattin, P.; Pedersen, N.L.; et al. Macrophage inhibitory cytokine-1 (MIC-1/GDF15): A new marker of all-cause mortality. *Aging Cell* **2010**, *9*, 1057–1064. [CrossRef] [PubMed]
76. Kim, Y.; Noren Hooten, N.; Evans, M.K. CRP Stimulates GDF15 Expression in Endothelial Cells through p53. *Mediat. Inflamm.* **2018**, *2018*, e8278039. [CrossRef]

77. Olley, G.; Ansari, M.; Bengani, H.; Grimes, G.R.; Rhodes, J.; von Kriegsheim, A.; Blatnik, A.; Stewart, F.J.; Wakeling, E.; Carroll, N.; et al. BRD4 interacts with NIPBL and BRD4 is mutated in a Cornelia de Lange-like syndrome. *Nat. Genet.* **2018**, *50*, 329–332. [CrossRef] [PubMed]
78. Parenti, I.; Diab, F.; Gil, S.R.; Mulugeta, E.; Casa, V.; Berutti, R.; Brouwer, R.W.W.; Dupé, V.; Eckhold, J.; Graf, E.; et al. MAU2 and NIPBL Variants Impair the Heterodimerization of the Cohesin Loader Subunits and Cause Cornelia de Lange Syndrome. *Cell Rep.* **2020**, *31*, 107647. [CrossRef] [PubMed]
79. Whelan, G.; Kreidl, E.; Peters, J.-M.; Eichele, G. The non-redundant function of cohesin acetyltransferase Esco2: Some answers and new questions. *Nucl. Austin Tex* **2012**, *3*, 330–334. [CrossRef] [PubMed]
80. Harakalova, M.; van den Boogaard, M.-J.; Sinke, R.; van Lieshout, S.; van Tuil, M.C.; Duran, K.; Renkens, I.; Terhal, P.A.; de Kovel, C.; Nijman, I.J.; et al. X-exome sequencing identifies a HDAC8 variant in a large pedigree with X-linked intellectual disability, truncal obesity, gynaecomastia, hypogonadism and unusual face. *J. Med. Genet.* **2012**, *49*, 539–543. [CrossRef]
81. NIH Preprint Pilot. Available online: https://www.ncbi.nlm.nih.gov/pmc/about/nihpreprints/ (accessed on 10 February 2022).