MOSES: A New Approach to Integrate Interactome Topology and Functional Features for Disease Gene Prediction

Manuela Petti, Lorenzo Farina, Federico Francone, Stefano Lucidi, Amalia Macali, Laura Palagi, Marianna de Santis

genes

MDPI

Article

MOSES: A New Approach to Integrate Interactome Topology and Functional Features for Disease Gene Prediction

Manuela Petti *, Lorenzo Farina, Federico Francone, Stefano Lucidi, Amalia Macali, Laura Palagi and Marianna De Santis

Department of Computer, Control and Management Engineering, Sapienza University of Rome, 00185 Rome, Italy; lorenzo.farina@uniroma1.it (L.F.); federicofrancone94@gmail.com (F.F.); lucidi@diag.uniroma1.it (S.L.); macali.1656011@studenti.uniroma1.it (A.M.); laura.palagi@uniroma1.it (L.P.); marianna.desantis@uniroma1.it (M.D.S.)

* Correspondence: manuela.petti@uniroma1.it

Abstract: Disease gene prediction is to date one of the main computational challenges of precision medicine. It is still uncertain if disease genes have unique functional properties that distinguish them from other non-disease genes or, from a network perspective, if they are located randomly in the interactome or show specific patterns in the network topology. In this study, we propose a new method for disease gene prediction based on the use of biological knowledge-bases (gene-disease associations, genes functional annotations, etc.) and interactome network topology. The proposed algorithm called MOSES is based on the definition of two somewhat opposing sets of genes both disease-specific from different perspectives: warm seeds (i.e., disease genes obtained from databases) and cold seeds (genes far from the disease genes on the interactome and not involved in their biological functions). The application of MOSES to a set of 40 diseases showed that the suggested putative disease genes are significantly enriched in their reference disease. Reassuringly, known and predicted disease genes together, tend to form a connected network module on the human interactome, mitigating the scattered distribution of disease genes which is probably due to both the paucity of disease-gene associations and the incompleteness of the interactome.

Keywords: disease gene prediction; data integration; precision medicine; computational biology

Citation: Petti, M.; Farina, L.; Francone, F.; Lucidi, S.; Macali, A.; Palagi, L.; De Santis, M. MOSES: A New Approach to Integrate Interactome Topology and Functional Features for Disease Gene Prediction. *Genes* **2021**, *12*, 1713. https://doi.org/10.3390/genes12111713

Academic Editors: Stefano Lonardi and Sven Rahmann

Received: 24 September 2021
Accepted: 25 October 2021
Published: 27 October 2021

1. Introduction

Precision medicine has been defined as "an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person." [1]. This definition is mainly related to the experimental, methodological, and technological developments of the last decades (e.g., next generation sequencing) that gave birth to new possibilities in the practice of healthcare based on individually tailored therapies. Disease genes identification is an important goal of biomedical research, and one of the main challenges aimed at the development of personalized treatments. In fact, a disease is rarely a consequence of an abnormality on a single gene, but it is usually the result of perturbations involving sets of genes and their relationships (e.g., alteration in molecular interactions, pathways). Disease genes (or seed genes) are those genes whose mutations are involved in diseases, and it is still uncertain whether such genes have unique properties that distinguish them from non-disease genes. In the last decades, numerous databases for gene annotations have been proposed providing information about genes and related diseases. Online Mendelian Inheritance in Man (OMIM) [2], curated by the NCBI and Johns Hopkins University, is one of the most widely used source of information for disease-gene associations, other examples are: PheGenI [3], DisGeNET [4], eDGAR [5]. Despite the several available resources offer different levels of information about the genetic basis of human diseases, knowledge about associations between disease-causing genes and diseases is still incomplete. Moreover, the identification

of specific disease genes is often impaired by gene pleiotropy, by the polygenic nature of many diseases, by the influence of a plethora of environmental factors, and by genome variability [6]. Various experimental techniques such as genome-wide association studies (GWAS) and linkage analysis are used to identify new seed genes, but the disadvantage of these high-throughput techniques is that often, they provide long lists of candidate genes and thus require validation procedures that make these methods time-consuming and expensive.

The described open problems combined with the importance of exploiting disease-gene associations to determine personalized treatments paved the way for the development of computational methods. In this context, algorithms for disease gene prediction have been proposed to use and/or integrate the large amount of available omics data and knowledge-based resources (gene annotations, disease-gene associations, etc.). Typical inputs of these algorithms are a set of seed genes (gathered from knowledgebases such as OMIM) and at least a second source of information (protein-protein interactions, functional ontologies, gene expression data, etc.). Instead, the output of these gene prioritization methods are typically subsets of candidate seed genes or genes rankings where top positions are related to high likelihood of involvement in generating a disease phenotype. Several reviews provide a description and a classification of the available algorithms for disease gene prioritization [7–9]. Here we briefly describe the three main categories: filtering-based techniques, similarity-based techniques, and network-based techniques. Filtering methods require the definition of filters based on the available knowledge of the molecular basis of the disease under investigation. Similarity-based techniques provide a gene prioritization based on a similarity measure between candidate genes and seed genes: the calculation of the similarity can exploit text-mining approaches [10] and can be based on functional profiles of genes [11]. Finally, network-based methods represent biological data as networks and apply graph mining techniques to rank genes. This last class of algorithms is also the last one developed in time in the wake of the introduction and success of network science in biomedical research [12–17]. Several methods have been proposed based on different strategies (network propagation [18,19], module-based [20,21]). Alongside these approaches, recently new network-based methods have been developed that use other genes besides the seed genes to help in the prediction of new disease genes [22–24]: in detail, these algorithms exploit genes associated with related diseases.

In this work, we introduce a new method for disease gene prediction based on the use of knowledge-bases, network topological features, and on the k-means algorithm applied to binary data. The proposed algorithm called MOSES (warM and cOld Seeds for disEase geneS) is based on the definition of two different and opposing sets of genes. In fact, for a specific disease, we define the known disease genes as warm seeds, and we identify as cold seeds the genes far from the warm seeds in the interactome (high path length between warm seeds and cold seeds) and not involved in the functions characterizing the disease genes (e.g., molecular pathways). In detail, given the two sets of genes, the key point of the proposed procedure is to distinguish between warm seeds and cold seeds exploiting the topology of the human interactome and the set of functionalities disrupted in the diseases. Regarding the second key point, it is not a priori known how to recognize similarities in functional profile, so that, in practice, we cannot use these similarities to decide if a generic gene is a disease gene or not. To overcome this issue, MOSES is based on the well-known data mining technique *k*-means clustering [25] and exploits an adaptive strategy to guide the clustering procedure to group the majority of known disease genes in a specific cluster. The hypothesis is that this specific cluster contains unknown disease genes (putative disease genes), besides containing known ones. The new approach can exploit and integrate different sources of information and here, we propose a first use based on known disease-gene associations, protein-protein human interactome and two types of functional gene annotations (Gene Ontology terms and *KEGG* pathways).

In the present study, we applied MOSES to a set of 40 diseases. To test the predictive power of MOSES, we performed a computational validation (10-fold cross-validation).

Furthermore, we used the enrichment analysis tool Enrichr [26] for checking if the putative genes are enriched in the disease to which the disease genes belong, and we studied the topological features of the predicted disease module (network module composed of known and putative disease genes).

2. Methods

The algorithm MOSES (warM and cOld Seeds for disEase geneS) is based on the definition and the characterization of two different and opposing sets of genes: warm seeds (WSs) and cold seeds (CSs). A warm seed is a disease gene, while a cold seed is a gene satisfying two constraints: (i) network-based distance and (ii) functional distance from the warm seeds. The first constraint imposes high path length between WSs and CSs in the interactome (see Figure 1), while the second distance requires that WSs and CSs are involved in totally different biological functions.

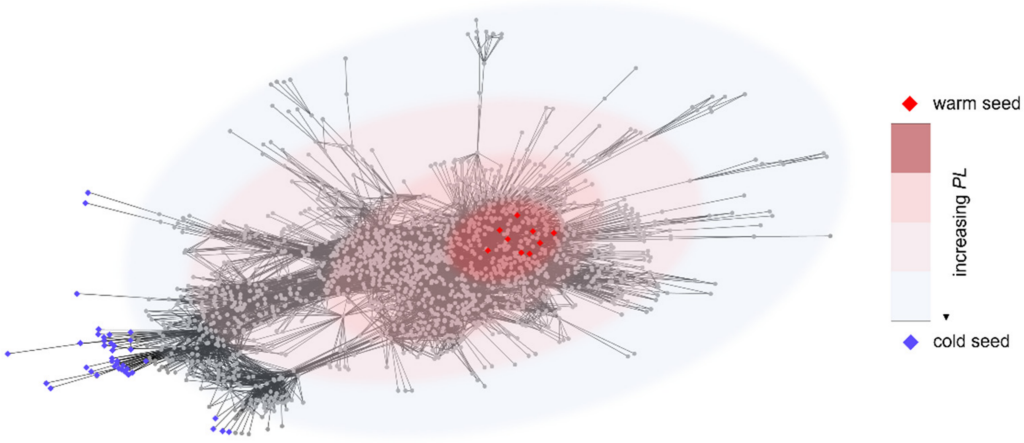

Figure 1. Example of network-based distance between warm seeds (red diamonds) and cold seeds (blue diamonds). The color of background ovals codes for the path length (PL) between WSs and CSs.

Its functioning requires three sequential phases described in detail in the following sections.

2.1. Functional Characterization of the Warm Seeds

The first step of the algorithm is to functionally characterize the WSs (i.e., the known disease genes of the disease under investigation) by means of the enrichment analysis (hypergeometric distribution with FDR correction). Different databases can be exploited, and thus integrated, such as Gene Ontology database, *KEGG* pathways, miRTarBase, TRRUST, etc. Fixed a significance threshold, for each of the considered databases, MOSES identifies M significant annotations: only databases for which $M \geq 2$ are considered for the next steps.

2.2. Identification and Enrichment Analysis of the Cold Seeds

To identify the cold seeds, the algorithm first applies the constraint of network-based distance in the selection of a set of genes in the interactome far from the disease genes. Given a specific disease characterized by P disease genes (or warm seeds, set S_0) and the interactome I composed of N genes, the iterative procedure to identify these peripheral genes is described below:

1. identification of the non-seeds set NS_i. At the first iteration, NS_1 is the difference set between the interactome and the disease genes: $NS_1 = I - S_0$, $\#(NS_1) = N - P$
2. identification of the first neighbors of genes in S_i (set FN_i)
3. update of the sets S_i and NS_i:

$$S_i = S_{i-1} \cup FN_{i-1}$$

$$NS_i = NS_{i-1} - FN_{i-1}$$

The procedure stops when the ratio between the cardinalities of sets S_0 and NS_i is equal to or greater than 10^{-1}. Once the set of peripheral genes has been identified, the MOSES algorithm extracts from this set, the cold seeds selecting only genes not involved in the WSs significant annotations (GO terms, *KEGG* pathways, MicroRNA-Target interactions, transcription factor-target regulatory relationships, etc.).

To be consistent with the WSs characterization, the algorithm selects also for the CSs the first M annotations of the considered databases with smaller *p*-values according to the hypergeometric distribution with FDR correction. Now, for each type of annotation, a set denoted by *J*, made of 2*M* annotations, can be built: the first *M* terms functionally describe the WSs, while the second half is related to the CSs characterization.

2.3. Optimized Clustering Phase and Selection of Putative Disease Genes

The set *J* is the input of the clustering phase: it is used to identify the subset of G_J genes, namely the genes in the interactome involved in the selected annotations, and allows to build the G_J-by-2*M* matrix to be subjected to the clustering procedure. This step is based on the use of the popular *k*-means clustering algorithm [25]: let *k* be a fixed integer number, the *k*-means separates the input set of genes into *k* clusters. MOSES algorithm proceeds as follows. Starting from $k = 2$, it iteratively applies the *k*-means clustering algorithm to the G_J-by-2*M* matrix, until it identifies a reasonable value for *k*. Note that, in the first iteration of the algorithm, namely when $k = 2$, it is likely that the known disease genes belonging to G_J are grouped within the same cluster, due to the specific choice of the set *J*. Then, the process goes on increasing the number *k* of clusters incrementally by one. The algorithm ends up at the first iteration for which the number of clusters k_{max} is the maximum number of clusters so that a given percentage *q*% of disease genes within G_J is in the same cluster C^*. This percentage belongs to the range (60%, 90%): the threshold of 60% is set to obtain more than half of the disease genes in one out of *k* clusters. In the case, at the first iteration ($k = 2$), the disease genes within G_J are divided into two clusters containing each less than 90% of them, MOSES sets *q*% equal to the higher percentage only if $q\% \geq 60\%$. Note that WSs e CSs do not share any annotations by construction, hence the optimal cluster cannot contain both the type of seeds.

The procedure outlined above is repeated considering each database (Gene Ontology database, *KEGG* pathways, *miRTarBase*) and allows to obtain the sets of genes C_i^* (i = GO; *KEGG*; miRTarBase). The new algorithm performs the intersection among the sets C_i^* returning a batch of known and putative disease genes.

The above-described procedure is synthesized in Figure 2.

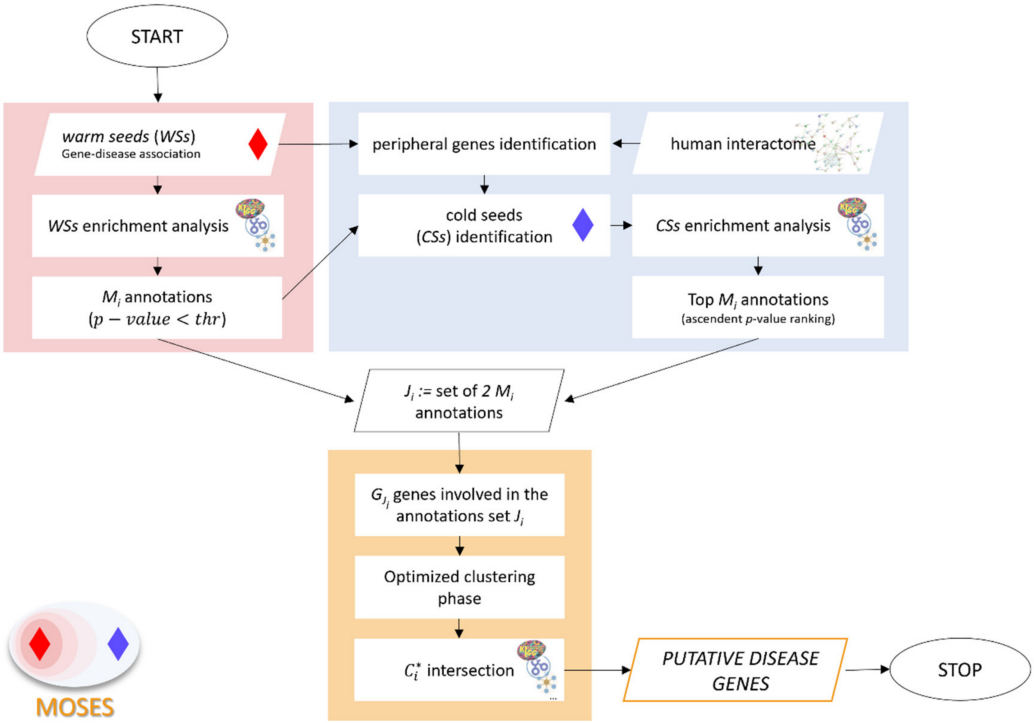

Figure 2. Flowchart of MOSES algorithm. The background rectangles identify the three sequential phases: red, blue and orange respectively for: (1) WSs functional characterization, (2) CSs identification and characterization, and (3) optimized clustering phase. The enrichment analysis can be performed considering different types of gene annotations (Gene Ontology database, KEGG pathways, miRTarBase) and obtaining for each of them M_i annotations, with i = GO; KEGG; miRTarBase, etc.

3. Data and Preprocessing

In the present work, to avoid selection bias, we applied the MOSES algorithm to 40 diseases selected from those provided in [21]. The selection criterion is related to the number of disease genes (warm seeds set, S_0): #(S_0) = P, P∈(25,150). As described in detail in [21], the disease-gene associations were retrieved from OMIM (Online Mendelian Inheritance in Man; http://www.ncbi.nlm.nih.gov/omim [2], accessed on 26 April 2019) and from the PheGenI database (Phenotype-Genotype Integrator; http://www.ncbi.nlm.nih.gov/gap/PheGenI [3], accessed on 26 April 2019). We used the human protein–protein interactome provided in [27] (243,603 protein-protein interactions connecting 16,677 unique proteins) and we considered two kinds of annotations: GO terms (Gene Ontology database, biological process, downloaded 26 April 2019) and pathways (*KEGG* gene set from the Molecular Signatures Database, version 6.2). The available GO terms (biological process) were not propagated upwards on the GO tree and were prefiltered as follows [20]:

i. annotations labeled with evidence code IPI (Inferred from Physical Interaction) were excluded to avoid circularity;
ii. annotations not associated with the gene products (evidence code "NOT") were excluded.

4. Results and Discussion

MOSES algorithm is based on the definition of two different and opposing sets of genes (warm seeds and cold seeds) and its functioning required the above described sequential phases. The putative disease genes returned by the algorithm are characterized by two important properties: the network-based proximity and the functional similarity with the original disease genes (here defined warm seeds). This is possible thanks to the new definition of the cold seeds: genes far from the disease genes in the interactome and

not involved in their functions. Furthermore, it is worth noting that MOSES can exploit, and thus integrate, different sources of information.

As described in the previous section, we applied MOSES to 40 diseases and for the first step of the algorithm (functional characterization of the WSs by means of the enrichment analysis), we set the significant threshold equal to 5×10^{-2}: for 27 out of the 40 diseases, MOSES selected $M \geq 2$ significant annotations in both cases (GO-BP terms, *KEGG* pathways). As MOSES has been thought to exploit data integration in the prediction of new disease genes, we considered only the subset of 27 diseases and in Table 1, we show for all of them: the number of WSs, the number of genes identified by MOSES applying the first constraint of network-based distance (peripheral genes) and the number of CSs (i.e., peripheral genes functionally distant from the WSs). It is worth noting that the application of the functional distance constraint further filters the set of peripheral genes proving that the integration between protein-protein interactome topology and gene functional annotations databases allows to appropriately identify the two opposing sets of gens.

Table 1. Cardinalities of the 3 genes sets: WSs, peripheral genes (application of network-based distance constraint) and CSs (application of both network-based and functional distance constraints).

Disease	Warm Seeds	Peripheral Genes	Cold Seeds
Amino acid metabolism inborn errors	52	119	113
Anemia, hemolytic	29	155	143
Arrhythmias, cardiac	30	171	163
Arthritis, rheumatoid	42	87	77
Asthma	37	91	85
Bile duct diseases	31	109	103
Blood coagulation disorders	40	142	129
Blood platelet disorders	26	193	170
Carbohydrate metabolism inborn errors	77	81	79
Cardiomyopathies	50	70	63
Celiac disease	36	137	120
Colitis, ulcerative	56	90	72
Colorectal neoplasms	42	79	68
Crohn disease	72	65	56
Diabetes mellitus, type 2	73	77	75
Head and neck neoplasms	35	87	80
Leukemia, myeloid	43	97	93
Lipid metabolism disorders	50	93	83
Lung diseases, obstructive	40	88	82
Lupus erythematosus	75	51	48
Lysosomal storage diseases	45	152	150
Multiple sclerosis	69	71	62
Muscular dystrophies	36	113	107
Psoriasis	54	86	76
Renal tubular transport inborn errors	34	229	211
Spinocerebellar ataxias	28	147	132
Spinocerebellar degenerations	30	147	137

For the optimized clustering phase, we used the *k*-means algorithm implementation in Matlab (*k*-means++ algorithm), setting as input parameters hamming distance and 50 replicates (number of times the *k*-means algorithm is run with different centroids).

In Figure 3 we show the clustering phase application to amino acid metabolism inborn errors, characterized by 52 WSs and 113 CSs. For the warm seeds, MOSES selected $M = 25$ significant GO-BP terms (p-value $< 5 \times 10^{-2}$ according to the hypergeometric test with FDR correction) leading to the set J composed of $2M = 50$ annotations (the second half of them is related to the cold seeds functional characterization). 1594 genes (32 of which are disease genes) of the interactome are involved in the 50 selected annotations. At the first iteration, the *k*-means algorithm produces two clusters with 1345 and 249 genes, respectively. One of the clusters contains 100% of disease genes, thus the process goes on increasing the number of clusters k and stops with $k_{max} = 3$, as one cluster contains the largest percentage of seeds equal to 0.84% (27 out of 32 disease genes, see Figure 3).

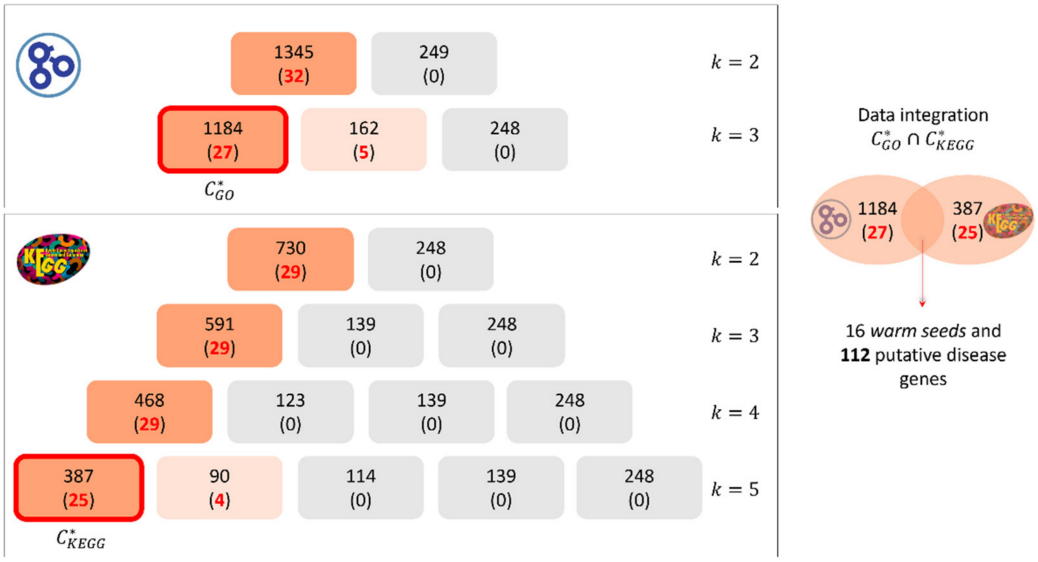

Figure 3. Clustering process applied to the disease amino acid metabolism inborn errors using GO-BP annotations (top panel) and KEGG pathways (bottom panel). On the right, the procedure of data integration and the identification of putative disease genes are shown.

Considering *KEGG* database, for the warm seeds, the algorithm selected $M = 10$ significant pathways (p-value $< 5 \times 10^{-2}$ according to the hypergeometric test with FDR correction) leading to the set J composed of $2M = 20$ terms. The genes involved in the 20 selected pathways are 978, 29 of them are disease genes. In this case, the process goes on increasing the number of clusters k until the selection of $k_{max} = 5$, as it is the first iteration for which we obtain a cluster containing the 86.2% of the disease genes (see Figure 3).

Thus, for this disease, we obtain the cluster C^*_{GO} made of 1184 genes (27 out of them being disease genes) and the cluster C^*_{KEGG} made of 387 genes (25 out of them being disease genes) The intersection $C^*_{GO} \cap C^*_{KEGG}$, returns 138 genes: 16 disease genes and 112 putative disease genes (set PG).

In 5 diseases among the 27 studied, the clustering phase failed in the identification of the cluster C^* using at least one type of annotations (GO-BP terms, K_{EGG} pathways). For example, in the case of asthma, at the first iteration ($k = 2$) the k-means algorithm applied to GO-BP data, returns 2 clusters containing 57% and 43% of disease genes: the warm seeds are therefore divided into two halves. In these cases, the use of a third (or more) database(s) could help to overcome the limitation encountered with a specific type of gene annotations. For the other 22 diseases, the MOSES algorithm identified the set of putative disease genes (PG).

4.1. Computational Cross-Validation and Comparison with Random Walk with Restart

To test the predictive power of MOSES, we performed the 10-fold cross-validation. For each disease, we randomly split the disease genes set S_0 into 10 subsets. Each time, we selected one subset as probe set S_P and the rest nine subsets as training warm seeds set S_T. Then we measured MOSES ability to recover genes in S_P. Furthermore, to evaluate the relative performance of MOSES, we considered as a reference another algorithm for candidate gene prioritization. We selected the random walk with restart algorithm (RWR) [18]: it is a ranking algorithm exploiting global network topology and it was shown to outperform other methods [19].

For each disease, we applied RWR (restart probability $r = 0.7$) to the same training sets used with MOSES, and to make comparable the outputs of the two algorithms (a finite set of *PG* putative disease genes for MOSES and a genes ranking for RWR), we considered the top *PG* positions in the case of RWR. As a measure of performance, we considered the percentage of recovered warm seeds in the test set *SP*. In Figure 4, we show in detail the difference between the two algorithms: we can appreciate the tendency of MOSES to outperform RWR in most cases, however, in few cases, RWR works better than MOSES. As expected indeed, there is not a universal best algorithm, but in general, the selection of the algorithm should be taken considering different factors of the available data. However, results of the statistical comparison (paired *t*-test) between the two algorithms show that overall MOSES performances are significantly higher than overall RWR performances (p-value = 9.6×10^{-03}).

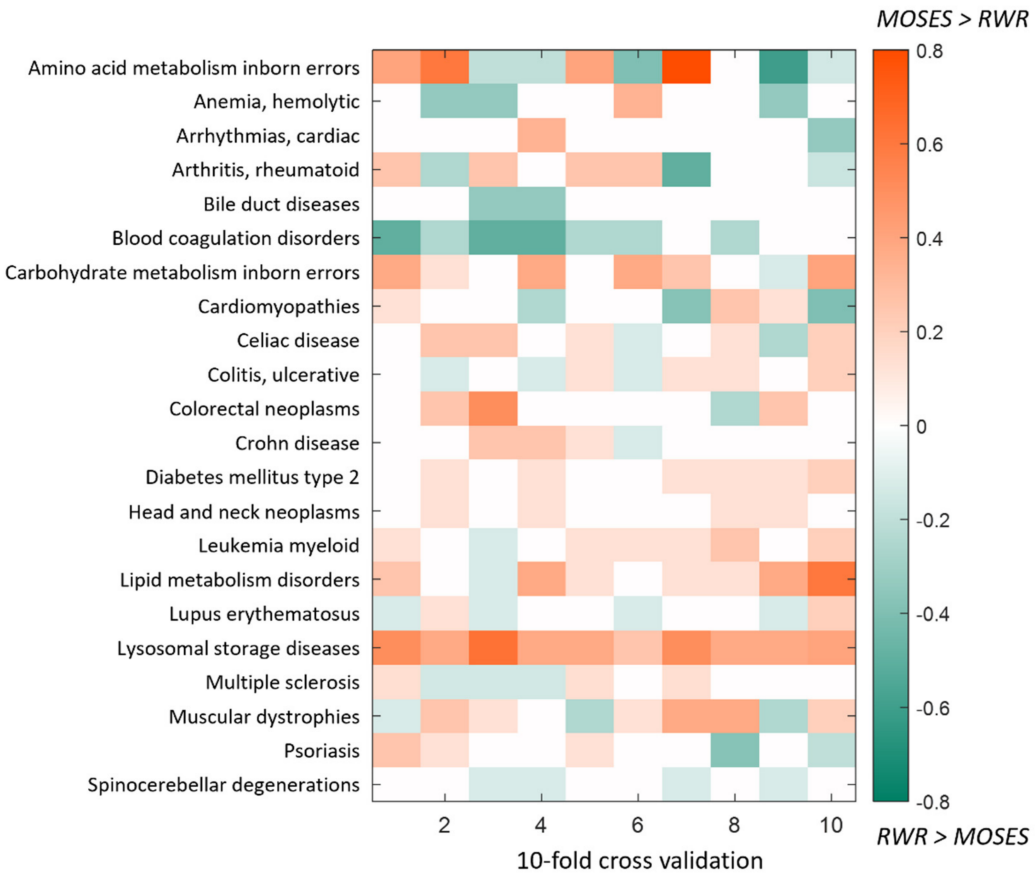

Figure 4. 10-fold cross-validation. Difference between MOSES and RWR performances. The performances are computed as the percentage of recovered warm seeds in the test set S_P. Rows and columns represent respectively the diseases and the cross-validation iterations. In the case of positive values (orange pixels), MOSES outperforms RWR, while negative values (green pixels) refer to the opposite situation.

4.2. Enrichment Analysis of Putative Disease Genes

We used the enrichment analysis tool Enrichr [23] to check if the putative genes are enriched in the disease to which the disease genes belong (category: Diseases/Drugs, section: *DisGeNET*). To find the corresponding disease in Enrichr, we referred to the International Statistical Classification of Diseases (ICD-11 for Mortality and Morbidity Statistics, version: 09/2020). Results are shown in Table 2. For 19 out of 22 diseases, the adjusted *p*-value is below the threshold of 0.05. Only in one case (*hemolytic anemia*), the *p*-value is above the significance threshold, while for two diseases (carbohydrate metabolism inborn errors, lipid metabolism disorders) we did not find the corresponding disease in Enrichr.

Table 2. Enrichment analysis of putative disease genes performed with Enrichr (Diseases/Drugs category, DisGeNET section). For each disease, we show the number of putative disease genes (PG), the corresponding DisGeNET disease name, the number of validated PG and the adjusted *p*-value retrieved from Enrichr; *p*-values below the significance threshold are highlighted in red.

Disease	#PG	DisGeNET Disease	#Validated	Adjusted *p*-Value
Amino acid metabolism, inborn errors	122	Amino Acid Metabolism, Inborn Errors	2	1.68×10^{-02}
Anemia, hemolytic	50	Anemia, Hemolytic	2	7.52×10^{-02}
Arrhythmias, cardiac	59	Cardiac Arrhythmia	5	7.08×10^{-04}
Arthritis, rheumatoid	447	Rheumatoid Arthritis	156	7.92×10^{-49}
Bile duct diseases	55	Bile Duct Diseases	1	3.71×10^{-02}
Blood coagulation disorders	104	Blood Coagulation Disorders	13	9.73×10^{-10}
Carbohydrate metabolism inborn errors	256	-	-	-
Cardiomyopathies	32	Cardiomyopathies	22	1.04×10^{-04}
Celiac disease	112	Celiac Disease	16	4.16×10^{-10}
Colitis, ulcerative	165	Ulcerative Colitis	68	7.88×10^{-45}
Colorectal neoplasms	1160	Colorectal Carcinoma	433	2.13×10^{-84}
Crohn disease	162	Crohn Disease	58	3.50×10^{-34}
Diabetes mellitus, type 2	52	Diabetes Mellitus, Non-Insulin-Dependent	29	4.68×10^{-15}
Head and neck neoplasms	412	Malignant Head and Neck Neoplasm	52	1.21×10^{-21}
Leukemia, myeloid	184	Myeloid Leukemia	22	3.32×10^{-08}
Lipid metabolism disorders	43	-	-	-
Lupus erythematosus	248	Lupus Erythematosus, Systemic	103	1.19×10^{-59}
Lysosomal storage diseases	112	Lysosomal Storage Diseases	5	3.18×10^{-03}
Multiple sclerosis	396	Multiple Sclerosis	101	1.31×10^{-37}
Muscular dystrophies	122	Muscular Dystrophy, Duchenne	6	6.63×10^{-03}
Psoriasis	421	Psoriasis	77	3.51×10^{-27}
Spinocerebellar degenerations	38	Ataxia, Spinocerebellar	2	3.52×10^{-02}

4.3. Study of the Predicted Disease Module

Putative disease genes are identified by MOSES exploiting the protein-protein interactome topology and the set of functionalities disrupted in the diseases. The integration of these different types of information agrees with the hypothesis of overlap among disease module, topological module (locally dense network neighborhood) and functional module (aggregation of nodes with similar or related functions in the same network neighborhood) [13]. While the topological and functional modules are concepts widely applied in different fields and suitable also in the case of biological networks, the network disease module is a recent key concept of network medicine [13,28,29]. This concept was raised from some broadly accepted hypotheses and organizational principles of disease genes [30]. In particular, genes (or gene products) involved in the same disease tend to interact (local hypothesis) and to cluster in connected subnetworks (*disease module hypothesis*). Moreover, genes in a disease module are often involved in the same biological functions (*functional coherence hypothesis*).

In the light of these considerations, for each disease, we studied the topology of the network module composed of the known disease genes (the warm seeds, set S_0) and the candidate genes (set PG) suggested by MOSES. We focused on the largest connected component (LCC) of the disease module investigating the size of the LCC consisting of warm seeds only ($|LCC_{WS}|$), the size of the LCC considering the set $S_0 \cup PG$ ($|LCC_{WS+PG}|$) and the number of warm seeds in LCC_{WS+PG}. In Table 3, we show the above describe measures for each disease.

Table 3. Study of the largest connected component (LCC) of the disease module. For each disease, we show the number of warm seeds (WSs), the number of putative disease genes (PG), the size of the LCC consisting of warm seeds only ($|LCC_{WS}|$), the size of the LCC considering the set $S_0 \cup PG$ ($|LCC_{WS+PG}|$) , the number of warm seeds in LCC_{WS+PG}, the 95th percentile threshold of the distribution of the 1000 LCCs of the random disease module (known disease genes and random genes). In the column $|LCC_{WS+PG}|$, bold text highlights values above $|LCC_{WS+RG}|$ threshold.

Disease	#WSs	#PGs	$\|LCC_{WS}\|$	$\|LCC_{WS+PG}\|$	#WSs in LCC_{WE+PG}	$\|LCC_{WS+RG}\|$ Threshold
Amino acid metabolism inborn errors	52	122	11	**42**	14	27
Anemia, hemolytic	29	50	11	**55**	12	16
Arrhythmias, cardiac	30	59	2	**36**	6	16
Arthritis, rheumatoid	42	447	6	**306**	31	201
Bile duct diseases	31	55	3	**35**	7	12
Blood coagulation disorders	40	104	22	**98**	34	37
Carbohydrate metabolism inborn errors	77	256	9	**168**	39	96
Cardiomyopathies	50	32	27	**42**	32	33
Celiac disease	36	112	2	**57**	7	15
Colitis, ulcerative	56	165	5	**140**	22	44
Colorectal neoplasms	42	1160	18	**992**	35	771
crohn disease	72	162	10	**150**	27	57
Diabetes mellitus type 2	73	52	7	**19**	9	16
Head and neck neoplasms	35	412	6	**320**	25	172
Leukemia myeloid	43	184	16	**136**	32	69

Table 3. *Cont.*

Disease	#WSs	#PGs	$\|LCC_{WS}\|$	$\|LCC_{WS+PG}\|$	#WSs in LCC_{WE+PG}	$\|LCC_{WS+RG}\|$ Threshold
Lipid metabolism disorders	50	43	11	**37**	19	17
Lupus erythematosus	75	248	5	**180**	39	92
Lysosomal storage diseases	45	112	8	13	5	20
Multiple sclerosis	69	396	11	**287**	40	185
Muscular dystrophies	36	122	12	**84**	24	31
Psoriasis	54	421	6	**309**	36	194
Spinocerebellar degenerations	30	38	2	**37**	9	12

In all the cases except for lysosomal storage diseases, the largest connected component of the predicted disease module (known and putative disease genes) contains a higher number of WSs with respect to the size of the *LCC* composed of warm seeds only. This result suggests that the extension of the disease module with the identified candidate genes, mitigates the WSs scattered distribution in the human interactome. The obtained disease modules are thus in accordance with the strategy of the recent Seed Connector Algorithm (SCA) [31]. Indeed, SCA proposes to add few additional linking genes (seed connectors) to the disease genes set, on the basis of the hypothesis that such seed connectors are hidden disease module elements that are critical for interpreting the functional context of disease genes [31]. However, the main and substantial difference between the two algorithms is that SCA builds the network module forcing the presence of the disease genes connected component, while with the application of MOSES, this topological property of the disease module is the result of the algorithm procedure. Furthermore, MOSES does not impose the presence of a single connected component.

Figure 5 shows the case of *ulcerative colitis* (see supplementary material for the other diseases). For this disease, the largest connected component of *warm seeds* only is composed of 5 nodes: thus only 5 out of 56 known disease genes are directly connected in the human interactome. Adding the 165 putative genes suggested by the MOSES algorithm to the disease module, we obtain the largest connected component composed of 140 nodes, 22 of which are WSs.

To verify that the obtained size of the LCC_{WS+PG} has not been obtained by chance, for each disease, we generated 1000 times a random disease module composed of the known disease genes and genes randomly selected from the interactome (*WSs* were excluded during random picks). In particular, the number of random genes (*RG*) is equal to the number of putative disease genes. For all the diseases except for *lysosomal storage diseases*, the size of the LCC_{WS+PG} is above the threshold of 95th percentile of the distribution of the largest connected components composed of the known disease genes and the random genes ($|LCC_{WS+RG}|$, see Table 3).

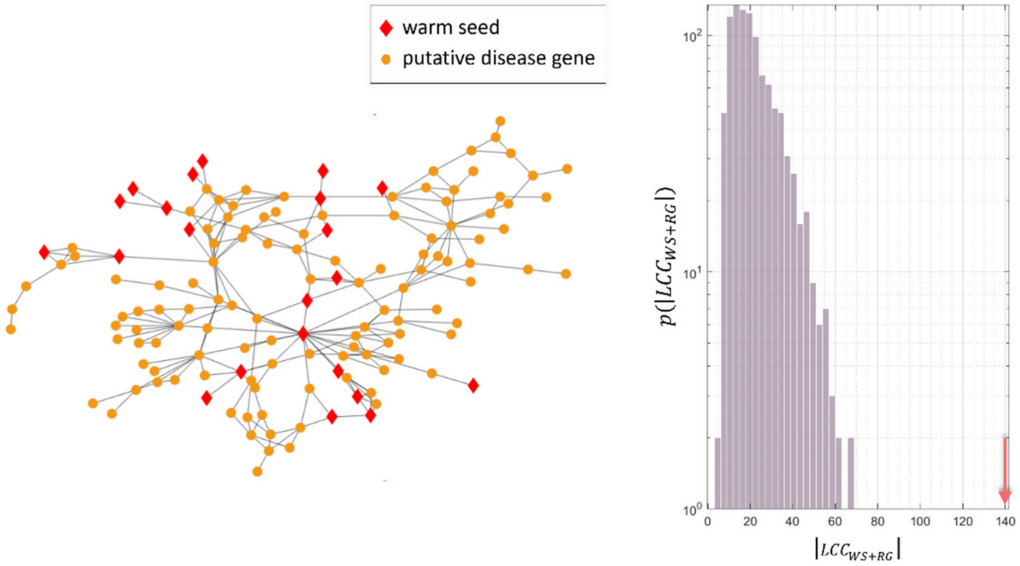

Figure 5. Largest connected component (*LCC*) of the predicted disease module (*warm seeds* and putative disease genes) for *ulcerative colitis*. Node shape codes for the type of genes: red diamonds represent the *warm seeds* (22 nodes), while orange dots represent the putative disease genes (118 nodes). On the right, distribution of the size of the 1000 *LCCs* of the random disease modules ($|LCC_{WS+RG}|$) obtained adding to the *warm seeds*, a set of randomly selected genes with cardinality equal to the set of putative genes; the orange arrow indicates the size of the LCC_{WS+PG} shown in the left panel.

4.4. Case Studies on Colorectal Neoplasms and Rheumatoid Arthritis

Among the studied diseases, we present in this section the results obtained for rheumatoid arthritis and colorectal neoplasms.

4.4.1. Rheumatoid Arthritis

Rheumatoid arthritis (RA) is an autoimmune and inflammatory disease, which means that the immune system attacks healthy cells by mistake, causing inflammation (painful swelling) in the affected parts of the body.

The disease genes for RA used in this work (Supplementary Table S1), are enriched in 18 KEGG pathways (hypergeometric test, FDR less than 0.05): among them, it is worth noting the presence of notch signaling pathway [32,33], cell adhesion molecules CAMs [34] and Jak-STAT pathway [35].

Starting from the 42 original disease genes, the MOSES algorithm identified 447 putative genes (Supplementary Table S1). The functional enrichment analysis showed they are enriched in 41 KEGG pathways (hypergeometric test, FDR less than 0.05), including 16 of the 18 characterizing the disease genes. However, none of them are in the top 3 (FDR ascending order). Indeed, the putative genes resulted mainly associated with neuroactive ligand receptor interaction (FDR = 1.44×10^{-43}), olfactory transduction (FDR = 7.82×10^{-20}) and metabolism of xenobiotics by cytochrome P450 (FDR = 4.28×10^{-19}). Interestingly, in relation to the olfactory transduction pathway, disturbances in the olfactory function have been investigated mainly in neurological/neurodegenerative disorders and only recently in autoimmune diseases [36–38]. In particular, in [39], Li and colleagues carried out a whole-exome sequencing study in a Han (Chinese ethnic group) patient cohort and identified genes enriched in the olfactory transduction pathway, suggesting the potential involvement of this pathway in RA disease progression.

Furthermore, performing the enrichment analysis with Enrichr (category: transcription, section: TRRUST Transcription Factors 2019) and focusing on the top 3 positions (ascending order based on adjusted *p*-value), we found that the putative genes are significantly enriched in RelA (adjusted *p*-value = 1.104×10^{-14}), NF-κB1 (adjusted *p*-value = 4.502×10^{-14}) and CIITA (adjusted *p*-value = 4.734×10^{-12}). NF-κB is a collective name for dimeric tran-

scription factors comprised of the Rel family of proteins that include RelA (p65), c-Rel, RelB, NF-κB1 (p50), and NF-κB2 (p52): NF-κB has been well recognized as a pivotal regulator of inflammation in rheumatoid arthritis [40].

4.4.2. Colorectal Neoplasms

Colorectal cancer is one of the most common cancers in the world and also one of the leading causes of cancer-related death worldwide [41].

The disease genes here used (Supplementary Table S2), are enriched in 19 K_{EGG} pathways (hypergeometric test, FDR less than 0.05): among them, as expected, there are colorectal cancer pathway, pathways in cancer, and mismatch repair. For this disease, starting from the 42 original disease genes, MOSES suggested 1160 putative genes (Supplementary Table S2). The functional enrichment analysis showed they are enriched in 65 KEGG pathways (hypergeometric test, FDR less than 0.05), including almost all those characterizing the disease genes (16 out of 19 terms). It is worth noting that only for the putative genes and at the first position in their pathways ranking based on FDR ascending order, we found cytokine-cytokine receptor interaction (FDR = 4.18×10^{-63}). Indeed, cytokine and cytokine receptor interaction networks are crucial aspects of inflammation and tumor immunology particularly for colorectal cancer [42,43]. Moreover, using the Enrichr platform (category: transcription, section: miRTarBase 2017), we found that putative disease genes are enriched in mir-145-5p (adjusted p-value = 4.97×10^{-9}; top position in the ascending order). miR-145 has frequently been investigated in colorectal cancer [44]: this miRNA acts as a tumor suppressor [45,46] and has been reported to be down-regulated in colon carcinomas [47].

5. Conclusions

In this work, we introduce the algorithm MOSES based on the new definition of warm seeds and cold seeds. In particular, the identification of the cold seeds requires the application of two constraints of distance from the known disease genes (here defined warm seeds): network-based distance and functional distance. MOSES exploits thus the advantages of network-based approaches and the use of disease genes functional features: indeed, it suggests a finite set of putative disease genes characterized by the two important properties of network-based proximity and functional similarity with the original disease genes. The use of these two seeds sets is innovative in the fact that we consider genes far away from each other to identify putative genes, whereas most disease gene prediction algorithms are based on the idea that putative genes are "near" in some sense to the known disease genes. Future analysis will be aimed at the integration of more types of gene annotations, to overcome the limitation encountered in the present study.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/genes12111713/s1, Figure S1: Largest connected component (LCC) of the predicted disease module (*warm seeds* and putative disease genes) for *amino acid metabolism inborn errors*, Figure S2: largest connected component (LCC) of the predicted disease module (*warm seeds* and putative disease genes) for *anemia, hemolytic*, Figure S3: Largest connected component (LCC) of the predicted disease module (*warm seeds* and putative disease genes) for *arrhythmias, cardiac*, Figure S4: Largest connected component (LCC) of the predicted disease module (*warm seeds* and putative disease genes) for *arthritis, rheumatoid*, Figure S5: Largest connected component (LCC) of the predicted disease module (*warm seeds* and putative disease genes) for *bile duct diseases*, Figure S6: Largest connected component (LCC) of the predicted disease module (*warm seeds* and putative disease genes) for *blood coagulation disorders*, Figure S7: Largest connected component (LCC) of the predicted disease module (*warm seeds* and putative disease genes) for *carbohydrate metabolism inborn errors*, Figure S8: Largest connected component (LCC) of the predicted disease module (*warm seeds* and putative disease genes) for *cardiomyopathies*, Figure S9: Largest connected component (LCC) of the predicted disease module (*warm seeds* and putative disease genes) for *celiac disease*, Figure S10: Largest connected component (LCC) of the predicted disease module (*warm seeds* and putative disease genes) for *colorectal neoplasms*, Figure S11: Largest connected component (LCC) of the predicted disease module

(*warm seeds* and putative disease genes) for *crohn disease*, Figure S12: Largest connected component (LCC) of the predicted disease module (*warm seeds* and putative disease genes) for *diabetes mellitus type 2*, Figure S13: Largest connected component (LCC) of the predicted disease module (*warm seeds* and putative disease genes) for *head and neck neoplasms*, Figure S14: Largest connected component (LCC) of the predicted disease module (*warm seeds* and putative disease genes) for *leukemia, myeloid*, Figure S15: Largest connected component (LCC) of the predicted disease module (*warm seeds* and putative disease genes) for *lipid metabolism disorders*, Figure S16: Largest connected component (LCC) of the predicted disease module (*warm seeds* and putative disease genes) for *lupus erythematosus*, Figure S17: Largest connected component (LCC) of the predicted disease module (*warm seeds* and putative disease genes) for *lysosomal storage diseases*, Figure S18: Largest connected component (LCC) of the predicted disease module (*warm seeds* and putative disease genes) for *multiple sclerosis*, Figure S19: Largest connected component (LCC) of the predicted disease module (*warm seeds* and putative disease genes) for *muscular dystrophies*, Figure S20: Largest connected component (LCC) of the predicted disease module (*warm seeds* and putative disease genes) for *psoriasis*, Figure S21: Largest connected component (LCC) of the predicted disease module (*warm seeds* and putative disease genes) for *spinocerebellar degenerations*, Table S1: disease genes and putative genes for *rheumatoid arthritis*, Table S2: disease genes and putative genes for *colorectal cancer*.

Author Contributions: Conceptualization, M.P., L.F., S.L., L.P. and M.D.S.; Methodology, M.P., L.F. and S.L.; Software, M.P. and F.F.; Validation, M.P., F.F., A.M. and M.D.S.; Writing—Original Draft Preparation, M.P. and M.D.S.; Writing—Review & Editing, M.P., L.F., S.L., L.P. and M.D.S.; Supervision, L.F., S.L. and L.P.; Funding Acquisition, M.D.S. All authors have read and agreed to the published version of the manuscript.

Funding: M.D.S. was supported by a Sapienza University of Rome grant—n. RM120172A2970290.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study as well as the code used are available at https://github.com/ManuelaPetti/MOSES.git (accessed on 24 May 2021).

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

References

1. What Is Precision Medicine? MedlinePlus Genetics. Available online: https://medlineplus.gov/genetics/understanding/precisionmedicine/definition/ (accessed on 21 October 2021).
2. Hamosh, A.; Scott, A.F.; Amberger, J.S.; Bocchini, C.A.; Valle, D.; McKusick, V.A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **2002**, *30*, 52–55. [CrossRef]
3. Ramos, E.M.; Hoffman, D.; Junkins, H.A.; Maglott, D.; Phan, L.; Sherry, S.T.; Feolo, M.; Hindorff, L.A. Phenotype-Genotype Integrator (PheGenI): Synthesizing genome-wide association study (GWAS) data with existing genomic resources. *Eur. J. Hum. Genet.* **2014**, *22*, 144–147. [CrossRef] [PubMed]
4. Piñero, J.; Bravo, À.; Queralt-Rosinach, N.; Gutiérrez-Sacristán, A.; Deu-Pons, J.; Centeno, E.; García-García, J.; Sanz, F.; Furlong, L.I. DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* **2017**, *45*, D833–D839. [CrossRef]
5. Babbi, G.; Martelli, P.L.; Profiti, G.; Bovo, S.; Savojardo, C.; Casadio, R. eDGAR: A database of Disease-Gene Associations with annotated Relationships among genes. *BMC Genom.* **2017**, *18*, 554. [CrossRef]
6. Bromberg, Y. Chapter 15: Disease Gene Prioritization. *PLoS Comput. Biol.* **2013**, *9*, e1002902. [CrossRef] [PubMed]
7. Moreau, Y.; Tranchevent, L.-C. Computational tools for prioritizing candidate genes: Boosting disease gene discovery. *Nat. Rev. Genet.* **2012**, *13*, 523–536. [CrossRef] [PubMed]
8. Piro, R.M.; di Cunto, F. Computational approaches to disease-gene prediction: Rationale, classification and successes. *FEBS J.* **2012**, *279*, 678–696. [CrossRef] [PubMed]
9. Kaushal, P.; Singh, S. Network-based disease gene prioritization based on Protein–Protein Interaction Networks. *Netw. Modeling Anal. Health Inform. Bioinform.* **2020**, *9*, 55. [CrossRef]
10. van Driel, M.A.; Bruggeman, J.; Vriend, G.; Brunner, H.G.; Leunissen, J.A.M. A text-mining analysis of the human phenome. *Eur. J. Hum. Genet.* **2006**, *14*, 535–542. [CrossRef] [PubMed]
11. Freudenberg, J.; Propping, P. A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics* **2002**, *18*, S110–S115. [CrossRef] [PubMed]

12. Silverman, E.K.; Schmidt, H.H.H.W.; Anastasiadou, E.; Altucci, L.; Angelini, M.; Badimon, L.; Balligand, J.; Benincasa, G.; Capasso, G.; Conte, F.; et al. Molecular networks in Network Medicine: Development and applications. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **2020**, *12*, e1489. [CrossRef] [PubMed]
13. Barabási, A.-L.; Gulbahce, N.; Loscalzo, J. Network Medicine: A Network-based Approach to Human Disease. *Nat. Rev. Genet.* **2011**, *12*, 56–68. [CrossRef] [PubMed]
14. Tieri, P.; Farina, L.; Petti, M.; Astolfi, L.; Paci, P.; Castiglione, F. Network Inference and Reconstruction in Bioinformatics. In *Encyclopedia of Bioinformatics and Computational Biology*; Ranganathan, S., Gribskov, M., Nakai, K., Schönbach, C., Eds.; Academic Press: Oxford, UK, 2019; pp. 805–813. Available online: http://www.sciencedirect.com/science/article/pii/B978012809633820290 2 (accessed on 26 April 2019).
15. Bassett, D.S.; Sporns, O. Network neuroscience. *Nat. Neurosci.* **2017**, *20*, 353–364. [CrossRef] [PubMed]
16. Toppi, J.; Petti, M.; Fallani, F.D.V.; Vecchiato, G.; Maglione, A.G.; Cincotti, F.; Salinari, S.; Mattia, D.; Babiloni, F.; Astolfi, L. Describing relevant indices from the resting state electrophysiological networks. In Proceedings of the 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, San Diego, CA, USA, 28 August–1 September 2012; pp. 2547–2550. Available online: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6346483 (accessed on 7 June 2015).
17. Barabási, A.-L.; Oltvai, Z.N. Network biology: Understanding the cell's functional organization. *Nat. Rev. Genet.* **2004**, *5*, 101–113. [CrossRef] [PubMed]
18. Köhler, S.; Bauer, S.; Horn, D.; Robinson, P.N. Walking the Interactome for Prioritization of Candidate Disease Genes. *Am. J. Hum. Genet.* **2008**, *82*, 949–958. [CrossRef] [PubMed]
19. Navlakha, S.; Kingsford, C. The power of protein interaction networks for associating genes with diseases. *Bioinformatics* **2010**, *26*, 1057–1063. [CrossRef] [PubMed]
20. Petti, M.; Bizzarri, D.; Verrienti, A.; Falcone, R.; Farina, L. Connectivity Significance for Disease Gene Prioritization in an Expanding Universe. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2020**, *17*, 2155–2161. [CrossRef] [PubMed]
21. Ghiassian, S.D.; Menche, J.; Barabási, A.-L. A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome. *PLoS Comput. Biol.* **2015**, *11*, e1004120. [CrossRef] [PubMed]
22. Selim, S.Z.; Ismail, M.A. K-Means-Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality. *IEEE Trans. Pattern Anal. Mach. Intell.* **1984**, *1*, 81–87. [CrossRef] [PubMed]
23. Garcia-Vaquero, M.L.; Gama-Carvalho, M.; Rivas, J.D.L.; Pinto, F.R. Searching the overlap between network modules with specific betweeness (S2B) and its application to cross-disease analysis. *Sci. Rep.* **2018**, *8*, 11555. [CrossRef]
24. Cáceres, J.J.; Paccanaro, A. Disease gene prediction for molecularly uncharacterized diseases. *PLoS Comput. Biol.* **2019**, *15*, e1007078. [CrossRef] [PubMed]
25. Maiorino, E.; Baek, S.H.; Guo, F.; Zhou, X.; Kothari, P.H.; Silverman, E.K.; Barabási, A.L.; Weiss, S.T.; Raby, B.A.; Sharma, A. Discovering the genes mediating the interactions between chronic respiratory diseases in the human interactome. *Nat. Commun.* **2020**, *11*, 811. [CrossRef]
26. Kuleshov, M.V.; Jones, M.R.; Rouillard, A.D.; Fernandez, N.F.; Duan, Q.; Wang, Z.; Koplev, S.; Jenkins, S.L.; Jagodnik, K.M.; Lachmann, A.; et al. Enrichr: A comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **2016**, *44*, W90–W97. [CrossRef] [PubMed]
27. Cheng, F.; Kovács, I.A.; Barabási, A.-L. Network-based prediction of drug combinations. *Nat. Commun.* **2019**, *10*, 1197. [CrossRef]
28. Goh, K.-I.; Cusick, M.E.; Valle, D.; Childs, B.; Vidal, M.; Barabási, A.-L. The human disease network. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 8685–8690. [CrossRef] [PubMed]
29. Caldera, M.; Buphamalai, P.; Müller, F.; Menche, J. Interactome-based approaches to human disease. *Curr. Opin. Syst. Biol.* **2017**, *3*, 88–94. [CrossRef]
30. Paci, P.; Fiscon, G.; Conte, F.; Wang, R.-S.; Farina, L.; Loscalzo, J. Gene co-expression in the interactome: Moving from correlation toward causation via an integrated approach to disease module discovery. *NPJ Syst. Biol. Appl.* **2021**, *7*, 1–11. [CrossRef]
31. Wang, R.-S.; Loscalzo, J. Network-Based Disease Module Discovery by a Novel Seed Connector Algorithm with Pathobiological Implications. *J. Mol. Biol.* **2018**, *430*, 2939–2950. [CrossRef] [PubMed]
32. Keewan, E.; Naser, S.A. The Role of Notch Signaling in Macrophages during Inflammation and Infection: Implication in Rheumatoid Arthritis? *Cells* **2020**, *9*, 111. [CrossRef] [PubMed]
33. Park, J.; Kim, S.; Kim, K.; Jin, C.; Choi, K.Y.; Jang, J.; Choi, Y.; Gwon, A.; Baik, S.; Yun, U.J.; et al. Inhibition of notch signalling ameliorates experimental inflammatory arthritis. *Ann. Rheum. Dis.* **2015**, *74*, 267–274. [CrossRef] [PubMed]
34. da Rosa Franchi Santos, L.F.; Costa, N.T.; Maes, M.; Simão, A.N.C.; Dichi, I. Influence of treatments on cell adhesion molecules in patients with systemic lupus erythematosus and rheumatoid arthritis: A review. *Inflammopharmacology* **2020**, *28*, 363–384. [CrossRef]
35. Walker, J.G.; Smith, M.D. The Jak-STAT pathway in rheumatoid arthritis. *J. Rheumatol.* **2005**, *32*, 1650–1653. [PubMed]
36. Vieira Borba, V.; Shoenfeld, N.; Perricone, C.; Shoenfeld, Y. Chapter 27-Smell and Autoimmunity—State of the Art. In *Mosaic of Autoimmunity*; Perricone, C., Shoenfeld, Y., Eds.; Academic Press: Oxford, UK, 2019; pp. 269–277. Available online: https://www.sciencedirect.com/science/article/pii/B978012814307000027X (accessed on 21 May 2021).

37. Perricone, C.; Shoenfeld, N.; Agmon-Levin, N.; de Carolis, C.; Perricone, R.; Shoenfeld, Y. Smell and Autoimmunity: A Comprehensive Review. *Clin. Rev. Allergy Immunol.* **2013**, *45*, 87–96. [CrossRef]
38. Moscavitch, S.-D.; Szyper-Kravitz, M.; Shoenfeld, Y. Autoimmune pathology accounts for common manifestations in a wide range of neuro-psychiatric disorders: The olfactory and immune system interrelationship. *Clin. Immunol.* **2009**, *130*, 235–243. [CrossRef]
39. Li, Y.; Leung, E.L.; Pan, H.; Yao, X.; Huang, Q.; Wu, M.; Xu, T.; Wang, Y.; Cai, J.; Li, R.; et al. Identification of potential genetic causal variants for rheumatoid arthritis by whole-exome sequencing. *Oncotarget* **2017**, *8*, 111119–111129. [CrossRef]
40. Makarov, S.S. NF-κB in rheumatoid arthritis: A pivotal regulator of inflammation, hyperplasia, and tissue destruction. *Arthritis Res. Ther.* **2001**, *3*, 200. [CrossRef] [PubMed]
41. Siegel, R.L.; Miller, K.D.; Fuchs, H.E.; Jemal, A. Cancer Statistics, 2021. *CA Cancer. J. Clin.* **2021**, *71*, 7–33. [CrossRef] [PubMed]
42. Lasry, A.; Zinger, A.; Ben-Neriah, A.L.A.Z.Y. Inflammatory networks underlying colorectal cancer. *Nat. Immunol.* **2016**, *17*, 230–240. [CrossRef] [PubMed]
43. West, N.R.; McCuaig, S.; Franchini, F.; Powrie, F. Emerging cytokine networks in colorectal cancer. *Nat. Rev. Immunol.* **2015**, *15*, 615–629. [CrossRef] [PubMed]
44. Akao, Y.; Nakagawa, Y.; Naoe, T. MicroRNA-143 and -145 in Colon Cancer. *DNA Cell Biol.* **2007**, *26*, 311–320. [CrossRef] [PubMed]
45. Qin, J.; Wang, F.; Jiang, H.; Xu, J.; Jiang, Y.; Wang, Z. MicroRNA-145 suppresses cell migration and invasion by targeting paxillin in human colorectal cancer cells. *Int. J. Clin. Exp. Pathol.* **2015**, *8*, 1328. [PubMed]
46. Wang, Z.; Zhang, X.; Yang, Z.; Du, H.; Wu, Z.; Gong, J.; Yan, J.; Zheng, Q. MiR-145 regulates PAK4 via the MAPK pathway and exhibits an antitumor effect in human colon cells. *Biochem. Biophys. Res. Commun.* **2012**, *427*, 444–449. [CrossRef] [PubMed]
47. Slaby, O.; Svoboda, M.; Fabian, P.; Smerdova, T.; Knoflickova, D.; Bednarikova, M.; Nenutil, R.; Vyzula, R. Altered Expression of miR-21, miR-31, miR-143 and miR-145 Is Related to Clinicopathologic Features of Colorectal Cancer. *Oncology* **2007**, *72*, 397–402. [CrossRef] [PubMed]