Briefings in Bioinformatics, 2021, pp. 1–7

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

Deepening the knowledge of rare diseases dependent on angiogenesis through semantic similarity clustering and network analysis

Raquel Pagano-Márquez,[1] José Córdoba-Caballero,[1] Beatriz Martínez-Poveda,[1,2,3] Juan AG. Ranea,[1,3,4] Ana R. Quesada,[1,3,4] Elena Rojano,[1,3*] Pedro Seoane[1,3,4*] and Miguel Ángel Medina[1,3,4]

[1]Department of Molecular Biology and Biochemistry, University of Malaga, Andalucia Tech, Bulevar Louis Pasteur 31, E-29071, Malaga, Spain, [2]CIBER de Enfermedades Cardiovasculares, CIBERCV, Av. Monforte de Lemos, 3-5, Pabellon 11, Planta 0, 28029, Madrid, Spain, [3]Biomedical Research Institute of Malaga, IBIMA, Calle Doctor Miguel Diaz Recio 28, 29010, Malaga, Spain and [4]CIBER de Enfermedades Raras, CIBERER, Av. Monforte de Lemos, 3-5, Pabellon 11, Planta 0, 28029, Madrid, Spain
*Co-corresponding authors. email:seoanezonjic@uma.es; email:elenarojano@uma.es
FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Background: Angiogenesis is regulated by multiple genes whose variants can lead to different disorders, including rare diseases. Despite their low frequency in population, there are thousands of rare diseases and most are thought to be genetic. The large variety of genes associated with rare diseases make them a valuable source of genetic information to explore the mechanisms involved in different biological processes. For this reason, we use them to deepen the knowledge of angiogenesis.
Results: In this work, we propose a systemic approach supported by the use of pathological phenotypes to group diseases by semantic similarity. Applying this methodology allowed us to identify 158 angiogenesis related rare diseases and to group them in 18 clusters based on their phenotypes. Of them, 14 clusters had traceable gene connections in a high quality interaction network. When the genes of these clusters were compared with ClinVar genes and variant data, six of seven were identified as angiogenesis related genes by our methodology. Furthermore, it allowed us to identify common affected functions among these disease clusters. psz
Availability: https://github.com/ElenaRojano/angio_cluster.
Contact: seoanezonjic@uma.es and elenarojano@uma.es
Supplementary information: Supplementary data are available at *Briefings in Bioinformatics* online.

Key words: rare diseases, systems biology, semantic similarity, disease clustering, angiogenesis.

Introduction

Angiogenesis deregulation is associated with a large number of diseases, including different types of cancer, autoimmune and rare diseases [3, 49, 40]. This biological process is complex in molecular terms and its regulation is susceptible to changes in the genome [51]. Despite being an essential process for the maintenance of the organism and the knowledge of genes involved in its functioning, there is much more to investigate about the genes and the regulation of this process.

In this work, we rely on the analysis of rare angiogenesis-dependent diseases (A-RDs) to analyse the genetic factors involved in the angiogenesis deregulation. This heterogeneous group of diseases include approximately 7,000 different pathologies and around 80% are due to genetic causes [33, 34]. Each rare disease is unique and they are very heterogeneous at the genetic level, allowing us to explore the mechanisms of multiple biological processes [23]. Thanks to the large number of genes

described for rare diseases, in this study we selected A-RDs to deepen the knowledge of angiogenesis.

In 2012, our research group performed a systemic review of A-RDs, with a manual search of A-RDs used to identify disease-associated genes and available drugs for their treatment using Orphanet resources [40]. In the work presented here we update the list of A-RDs and use it to get a better understanding of the A-RDs and their associated genes. To do so, we use a semantic similarity measure and clustering analysis of these A-RDs. There are plenty of studies that describe several similarity measures [35] applied for disease analysis at the phenotypic level as the rare disease map (RDmap) [57] or for the differential diagnostics for common diseases [47]. Furthermore, the use of semantic similarity for data clustering is frequent for the stratification of patients in cohorts [48] or disease groups [2]. In addition, this methodology is also used for the identification of genes that could be involved in disease development. For example, there are essential resources such as the Monarch Initiative [30] and DisGeNET [36], and tools like Priori-T [38].

In this way, we apply this knowledge to A-RDs designing a full analysis protocol to analyse a pool of diseases. In the first stage, we searched for the publications from which to get the list of A-RDs and retrieved their pathological phenotypes. Then, in the second stage, we stratified the list of A-RDs using semantic similarity with their pathological phenotypes and hierarchical clustering. For each disease group, their associated genes were used to analyse their closeness in the human interaction network and functions involved.

Our motivation is to give a reliable and straightforward insight of the rare diseases related to angiogenesis, grouping them at a phenotypic level. With this stratification, we explored each A-RD group at a genetic level to analyse the biological functions of the associated genes and which other genes could be related. Finally, we identified close A-RD groups at both phenotypic and genetic levels, giving a reference point to A-RD researchers to elucidate the disease mechanisms. [psz]

Material and methods

Angiogenesis-related rare diseases selection

Information concerning angiogenesis-related rare diseases (A-RDs) was compiled with the criteria described in the work of Rodríguez-Caso and collaborators [40]. Following their procedure, we performed an advanced search for specific terms that emerged in publications of any year. This search was performed in the Web of Science (WOS) and PubMed databases. We personalized our search in the following way, according to the database consulted: in the case of WOS, we searched for terms related to rare diseases and angiogenesis "(TS = (rare disease AND angiogen*))", whereas in PubMed were used "((rare diseases [MeSH Terms]) OR (rare AND diseases) OR (rare diseases) OR (rare AND disease) OR (rare disease) AND angiogen*)". Both searches were made in August 2021. We exported the results of the articles corresponding to this search and eliminated the repeated records.

In the same line as Rodríguez-Caso and collaborators followed in their study, [40], to perform the A-RDs articles selection we made a search for terms in the title and abstract of all articles, and specifically in the abstract keywords and MESH terms of PubMed articles and in the author keyword and keyword plus of WOS articles. We searched for two groups of keywords: "angiogen" and "VEGF" to verify that the article had content about this biological process, or "rare" and "disease" to confirm that there were rare diseases mentioned in the article. All the articles that did not meet these search requirements were removed from the study.

We calculated a content score to prioritize articles according to where these terms were included in the publications. This score is calculated in the following way: if the searched terms are in the title, we add 3 points to the score, 2 if they are included in the keywords or MeSH terms and 1 if it is in the abstract. Using the content score, we focused on articles whose score was equal to or greater than 4 as we considered them as the most relevant.

We manually inspected these articles to verify that they were describing A-RDs. All diseases resulting from manual curation were searched in Orphanet to get an official ORPHA code. These ORPHA codes will be used to find additional information about A-RDs in different databases.

For each A-RD ORPHA code we retrieved both their phenotype and gene annotations. In the case of phenotypes, we selected all the HPOs related to each A-RD ORPHA codes from the HPO annotation website [20] (`http://purl.obolibrary.org/obo/hp/hpoa/phenotype.hpoa`). The genes associated with A-RDs were retrieved from the Monarch Initiative resources [30] (`https://data.monarchinitiative.org/tsv/all_associations/gene_disease.all.tsv.gz`). More information about the list of A-RDs, their HPOs and associated genes is available in Supplementary Table 1.

Disease workflow analysis overview[psz]

[psz 1] MT2

We developed a workflow to group diseases with similar phenotypes in order to determine the genetics and molecular processes common to these diseases. The workflow uses a list of disease codes, in this case A-RDs ORPHA codes, to retrieve their associated HPO terms and genes from the Monarch Initiative. Then, the following steps are performed: 1) grouping diseases in clusters by phenotypic similarity; 2) calculation of the average shortest path (ASP) between known genes associated with diseases in each cluster using STRING data; 3) for clusters that have all gene pairs with computable paths in the interaction network, we do an expansion of the cluster with available genes in all shortest paths, 4) enrichment analysis in the Gene Ontology (GO) for both raw and expanded gene clusters; and 5) gene cluster prioritization in the interaction network using the CRank algorithm.[psz] The overview of the methodology is given in Figure 1. The workflow was developed in AutoFlow [45] and is available at `https://github.com/ElenaRojano/angio_cluster`.

Establishing A-RDs groups by phenotypes[psz]

[psz 2] MT3

We used the Cohort Analyzer tool [41] to group diseases using semantic similarity. This tool, included in the Patient Exploration Tools Suite [42], calculates different statistics in a cohort of patients or a pool of diseases and uses different semantic similarity methods to group them according to their HPO profiles [41].

The steps to obtain the disease clusters are performed as follows. First, the Lin similarity measure [35] is used to calculate a semantic similarity matrix among the A-RD HPO terms. This matrix is transformed to a dissimilarity matrix (the Lin measure ranges between 0-1, it is transformed with 1 – similarity). This dissimilarity matrix is used to perform a hierarchical clustering with the R core function hclust using the Ward criterion

[31]. The resulting dendrogram is split with the cutreeDynamic function included in the dynamicTreeCut R package [22] to get the final disease clusters. This algorithm is a hard clustering procedure that iterates the dendrogram analysing the tips of the branches to identify possible tightly connected clusters. It is used with default settings except for the minClusterSize and deepSplit parameters. The minClusterSize is the minimum items that can contain a cluster and we calculated it as the 1% of the diseases that have HPO terms. And the deepSplit parameter configures several internal parameters that controls how the branch partitioning and the clustering merging is performed. We set it to 2 following the cutreeDynamic authors recommendations. [psz]

Exploring the molecular mechanisms involved in the disease clusters

Once we obtained the disease clusters grouped by their phenotypic similarity, we analysed their associated genes to explore the underlying molecular mechanisms. These clusters may have some variability in terms of the number of diseases they have. However, it is expected they share associated genes or at least having genes with similar functions. We eliminated clusters with a single disease as they were meaningless in this study.

Then, we selected the union of all disease-associated genes for each A-RD cluster and analysed how they were connected in the protein interaction network. We downloaded all human interactions from the STRING database [50] (version 11.0b) and selected those with a combined score higher or equal to 900, which indicates a high confidence interaction between proteins [53]. In an additional step, we computed the degree for each node in the network. All the degree values were converted into Z-scores and we removed nodes with a Z-score greater or equal than 2.5 as they were considered as hubs. [psz] Then, we used the final interaction network to calculate the ASP to measure the closeness of the disease associated genes in each cluster. For this, we used the NetworkX Python package [14]. In this case, we used the Dijkstra algorithm [5] to obtain the shortest paths in the interaction network and we transformed the STRING weights into distances performing the operation $1000 - STRING_{\text{weight}}$ [26]. Disease clusters with a pair of associated genes without computable shortest path between them were removed from this study.[psz]

A gene expansion of the genes associated with the disease clusters was performed using the interaction network. For this, we associated with each cluster all the genes that contributed to the ASP calculation from the interaction network. This expansion allows us to find genes very close to disease associated genes but that have not been initially described for the diseases for each cluster, and which could be considered as possible genes involved in disease development[psz]. Additionally, the CRank [59] algorithm was applied to the gene lists (raw or expanded) to rank the A-RD clusters in accordance with their network connectivity features at gene level. Furthermore, it allows us to evaluate the improvement of the expanded gene lists. This algorithm measures the magnitude of structural features and the robustness against noise for the clusters in the network using four different connectivity metrics: Likelihood, Density, Boundary and Allegiance. All these metrics are summarized in the CRank value, which ranges between 0 (the evaluated list is the most dispersed cluster from the clusters set in the network) and 1 (the evaluated list is the most connected and coherent cluster in the clusters set).

We finally performed the functional enrichment analysis for the genes associated with the disease clusters in their expanded form or not, using the clusterProfiler R package [58] . This enrichment analysis was performed in molecular function and biological process GO sub-ontologies. The p-value associated with each functional category was calculated using the Over Representation Analysis (ORA) algorithm and corrected by multiple testing with the Benjamini-Hochberg method. Functional categories with adjusted p-value equal or less than 0.01 are reported. As the functional categories belong to the GO, when a functional category and its parent are significant for the same clusters, the parental terms are removed to simplify the interpretation of the results. The visualization of the enrichment results are generated with the dotplot function of the clusterProfiler R package. [psz]

When the enrichment analysis displays a large amount of functional terms, we use a summary representation. For it, the terms for each disease cluster are sorted by their adjusted p-value and the top N categories (custom threshold) with the lowest p-value are selected for each cluster. Then, using specific functions of clusterProfiler, we calculate a Wang semantic similarity matrix between the selected functional terms [54]. This matrix is hierarchically clustered using the hclust R function with the average method, and the resulting dendrogram is split with the treecut function setting h to $1-S$, where S is a custom similarity threshold used to get the GO clusters. For each similarity cluster, the common ancestor in all GO terms is searched and used as a representative term of the cluster. All child terms are replaced by this representative term and it gets the A-RD cluster relations available in their children. A parental cleaning process is applied as previously described. The results of this analysis are plotted with the heatmaply R package [9], building a heat map that groups rows and columns by their similarity vector. We show only a dendrogram for columns, corresponding to the disease clusters. [psz]

Results and Discussion

Retrieving angiogenesis-related rare diseases

For this work, we performed a bibliographic search on angiogenesis-related rare diseases (A-RDs) in PubMed and WOS databases. We performed an automated scoring of the found articles depending on where the search terms for these diseases appeared and to select those with the largest amount of information regarding A-RDs. From an original list of 1107 articles, 242 were related to A-RDs. We selected and inspected them manually, resulting in 158 A-RDs that were extracted from the Orphanet database. Of these diseases, 107[psz] were characterized with HPO terms and 109[psz] have associated genes in the Monarch Initiative database (Supplementary Table 1).

Characterization and clustering of angiogenesis-related rare diseases

We calculated with the Cohort Analyzer some statistics of our A-RD list. The full report is available in the GitHub repository at `https://github.com/ElenaRojano/angio_cluster`. The A-RD list includes a large number of different pathological phenotypes: 1,476[psz]. Likewise, the average of HPOs used to describe each disease is high: 28.36[psz]. This detailed description of the diseases will allow us to cluster the diseases in a more precise and informative way.

Cohort Analyzer computes the frequency for each phenotype in the disease list. In Table 1 we show the top 10

[psz 3] MA-07

[psz 4] MT5

[psz 5] MT6

[psz 6] MT7

[psz 7] R3

[psz 8] UD

[psz 9] UD

most frequent HPOs. We also check in the current bibliography its relationship with angiogenesis. For example, the terms HP:"Seizure", HP:"Headache" and HP:"Hypertension" are quite related to endothelial dysfunctions in patients with preeclampsia and hypertensive encephalopathy, and have been associated with dysregulations of vascular endothelial growth factor (VEGF) in endothelial cells [21, 27, 24].[psz] In fact, there are studies that relates VEGF-induced angiogenesis and the phenotypes HP:"Hepatomegaly", HP:"Splenomegaly" and HP:"Thrombocytopenia" [56, 55]. Dysregulations affecting VEGF levels lead to blood vessel anomalies and consequently produce all these symptoms. The HP:"Fatigue" term is mostly related to patients with cancer [15], and it also has been reported along with HP:"Weight loss" and HP:"Fever" in a patient with hemophagocytic lymphohistiocytosis, a rare immune disease [25]. Other phenotypes observed in patients with this rare syndrome include the top terms HP:"Splenomegaly", HP:"Hepatomegaly" and HP:"Abdominal pain" [7]. Taken altogether, this information shows that top 10 most frequent HPOs are related to alterations of the angiogenic process.

Cohort Analyzer tool computes the semantic similarity of the HPO profiles associated with each A-RD by the Lin method. Then, these A-RDs are clustered using these similarity values. In this case, the tool generated 18 different clusters (Supplementary Table 1, "ClusterID" column) with an average of diseases per cluster of 5.88[psz].

It is worth mentioning that from the initial list of 158 A-RDs, 47[psz] do not have HPOs described. It also draws attention that most of them are different types of cancer, including retinoblastoma, various sarcomas such as liposarcoma and rhabdomyosarcoma, and carcinomas including pancreatic and renal cell carcinomas, among others. However, the HPO has not yet included all the pathological phenotypes used to describe the different types of cancer available in Orphanet. It should be mentioned that the medical focus of the HPO in its early years was the phenotypic characterisation of Mendelian diseases [12], and many types of cancer are produced by somatic mutations in individual cells that do not follow a pattern of inheritance [37]. This would explain why for many types of A-RDs there are no annotations found for this ontology and suggest that they are enriched in oncologic diseases.

The angiogenesis map of genes and diseases

[psz 10] UD

In Supplementary Table 1 we show in which clusters the A-RDs have been grouped and the genes they have. As can be seen, from the 107[psz] diseases with pathological phenotypes available for this study, 23 have no genes described. Diseases with HPO description were used to perform the clustering. This does not mean that this information is not valuable, but quite the opposite: it is possible to determine whether diseases within the same cluster participate in the same biological processes to extrapolate the information to diseases whose genetics are still unknown. In Supplementary Figure 1 can be observed the robustness of the cluster procedure and the semantic similarity selection.[psz]

We generated a network representation with Cytoscape [46] to create the angiogenesis map of genes and diseases (Figure 2), related by the computed A-RDs clusters. The frequency of occurrence for each gene can be observed in Supplementary Figure 2A. As can be seen, most genes are connected to a single disease.[psz] We observe in Figure 2[psz] that most of the clusters (green circles) are connected between them by at least one gene (lilac circles). The connected clusters present at least one disease (salmon circles) that has been described with the same gene. Some isolated clusters are also observed, such as clusters 4, 10, 12, 14, 16 and 18. For example, cluster 4 is a tight gene-disease cluster with genes such as *TET2*, which plays a key role in erythropoiesis and its mutations are associated with anemia [10]. This gene has been described in five diseases, three of them are different types of anemia, and the other two are polycythemia vera (a blood cancer characterized by excessive production of red blood cells) and primary myelofibrosis (a bone marrow disease that affects the correct production of blood cells). As can be seen, these A-RDs share not only characteristics at the phenotypic level but also at the genetic level. [psz]

It is also remarkable that cluster 1 has twelve different diseases, each one described with at least one gene except the systemic sclerosis syndrome. This cluster overlaps with clusters 2, 3, 5 and 8 due mainly to Cowden syndrome whose genes connect with A-RDs characterized by the development of tissue tumors (malign or benign), a characteristic phenotype of Cowden syndrome [11]. [psz]

In the case of overlapping clusters, cluster 5 is an interesting example of both similar phenotypic features and genetic basis. Intrahepatic cholestasis of pregnancy disease has two associated genes: the ATP binding cassette subfamily B member 4 (*ABCB4*) and the ATPase Phospholipid Transporting 8B1 (*ATP8B1*). The latter gene belongs to the same family as the gene associated with Wilson disease, the *ATP7B* gene [43]. This suggests a very similar genetic basis for both diseases, supported by a high phenotypic similarity. For this reason, this approach could be used to identify some of these genes as involved in diseases that have not genes associated yet. [psz]

Is interesting to mention the connection of clusters 1, 3 and 8[psz] due to the *SDHB* and *SDHC* genes, shared between three diseases: Cowden syndrome, gastrointestinal stromal tumor and hereditary pheochromocytoma-paraganglioma. Furthermore, from the same gene family, the genes *SDHD* and *SDHA* are shared for some pairs of these diseases. All these diseases are characterized for generating benign overgrowths in different tissues and following an inheritance pattern [32].

In addition, we can find some diseases with a large number of genes, such as amyotrophic lateral sclerosis (ALS) in cluster 17[psz]. In fact, these genes are only connected to ALS. It is known that ALS is produced by mutations in a single or several genes at the same time [28] and this explains the large number of associated genes. Among them, we found angiogenin (*ANG*), a gene that stimulates angiogenesis in healthy and tumor tissue.

Altogether, the results shown and discussed in this section clearly show that the angiogenesis map of genes and diseases is very useful to extract new relations between genes and diseases. For this reason, we perform further analysis at gene interaction and functional levels.

Mapping the disease clusters onto the human interactions network

[psz 11] UD

[psz 12] MA-07

Once we have the A-RD clusters and the genes associated with the diseases, we can explore how these genes are related between them. To measure the proximity of the genes for each cluster, we mapped them to a high quality STRING human interaction network that includes interactions with a combined score higher or equal to 900 and removes hub nodes as described in Material and Methods. The degree of distribution of this network is shown in the Supplementary Figure 2B.[psz] This proximity measure is performed through the average shortest path

(ASP) calculation, which gives the number of nodes between two genes. If only a single disease-associated gene appeared in the interaction network, as it happens with cluster 14[psz], the cluster is discarded. Genes that were not found in the interactions network are available in the[psz] Supplementary Table 3.

For each cluster gene list, we calculated the ASP values using the interaction network considering[psz] the interaction weight itself (Supplementary Table 2, column ASP_value). We removed clusters 2, 5, 11 and 17[psz] because no direct path could be established between all its associated genes. Thus, we considered these clusters as unconnected to the gene level. Figure 3 shows for the remaining clusters the ASP, the number of disease-associated genes in the cluster and the proportion of the diseases in the[psz] cluster that have at least one associated gene by Monarch. First, we do not appreciate any relationship between the ASP values and the number of genes associated with each cluster. Interestingly, the A-RD clusters have ASP values equal or below 8[psz] and most of them are between 3.8-5.6[psz]. This means that most of the A-RD clusters have very similar diseases at the phenotype level and their affected genes are close in the interaction network. This could suggest that the A-RDs within a cluster may have alterations in the same biological processes in which different genes are involved. It may be discussed that the gene closeness is due to the association of genes with a specific A-RD; however, in Figure 3 it is shown that from 60% to 80% of the A-RDs in a cluster have associated at least one gene. If we focus on specific A-RD clusters, to the right of the figure we observe three clusters with the highest number of genes per cluster that ranges from 20 to 25. It is interesting that their ASP values range from 4.23 to 5.58, whereas cluster 16, with a lower number of associated genes, has an ASP value of 8. There are also six clusters with ASP values in a range from 3.84 to 5.21 and with ten or fewer genes per cluster, pointing to coherent clusters at both phenotype and interaction levels. Finally, clusters 13, 15 and 18 have both the lowest number of associated genes (4 or less) and the lowest ASP values, from 1,69 to 1. Consequently, these three clusters have associated genes very close in the interaction network.[psz] This evidence supports that the disease clusters are coherent at both phenotypic and interaction levels. The genes associated with each cluster are very close in the interactome and this suggests that they are involved in the same biological mechanism.

Functional analysis of the A-RD clusters

In the[psz] previous sections, we determined the disease similarity and the gene closeness in the human interaction network for each disease cluster. In this section, we focus on[psz] the functional perspective of these A-RD clusters. In this way, we used the gene lists with a computable average ASP value and performed a Gene Ontology (GO) enrichment analysis. In Figure 4 we found significant categories in GO molecular function for 10 of the 13 clusters. It is shown a great specialization in the functions for each disease cluster. Clusters 1 and 3 have the highest number of significant categories and the largest gene lists (25 and 20, respectively). The highest functional overlap is observed in clusters 1, 3 and 8, with the two categories "succinate dehydrogenase activity" and "electron transfer activity", both directly related to the electron transport chain in mitochondria. Looking closely at the genes for each cluster, we verified that all three have genes that code for different subunits of the succinate dehydrogenase complex (*SDH* gene, Figure 2). Regarding the specific functions of these clusters, we found cluster 15 with "galactosidase activity" related terms, which are known to be associated with angiogenesis [8]. In the same way, cluster 6 have the category "Hsp90 protein binding" and this complex is known to be involved in angiogenesis as well [17]. [psz]

Regarding GO biological process, in Figure 5, we found significant terms for 10 of the 13 clusters. We also observed the specialization in the functions for each disease cluster mentioned for annotations in GO molecular function. Clusters 1 and 3 have the largest number of functions, likely due to their substantial and diverse gene lists. In fact, cluster 1 presents a high variability of functional annotations but also includes specific processes related to the *VEGF* signalling pathway. It is remarkable that cluster 6, which merely includes three genes: *TSC1*, *TSC2* and *GLA* (Figure 2), has a noticeable number of functions related to mitophagy and changes in the glucose metabolism, a common feature in many oncological diseases or angiogenesis deregulation [44]. Is worth to mention that related to angiogenesis, when detailed results of biological process are observed, Supplementary Figure 3, cluster 3 have associated the terms "negative regulation of epithelial cell proliferation" and "endothelial cell differentiation". [psz]

This functional analysis supports the relationship between the selected diseases and the angiogenesis mechanism in which they relay. Furthermore, it allows to inspect the functional specialization for each disease cluster and which functions are shared by different groups of diseases reflecting the interconnection of the different angiogenesis related mechanisms.

A-RD clusters gene expansion to find unknown disease associated genes

We explored the phenotypic, interaction and functional levels of the clustered A-RDs and the evidence shown in this work avails the relationships between the A-RDs, as well as those between them and their associated genes. To deepen the results, we can identify new putative candidates and members of molecular mechanisms. To do this, we used the ASP calculation to take the genes in these short paths and expand the gene list for each disease cluster (Supplementary Table 2, ASP_expanded_genes column).

In Supplementary Figure 4 we can see how the number of genes associated with clusters 1, 3 and 8 range from 250 to > 500. This is clearly due to the number of associated genes in the Monarch Initiative to the cluster diseases, from 20 to 25 genes (Figure 3). With this number of genes, these clusters likely will be uninformative. [psz] In addition, the expanded gene lists were explored to identify new functions and connections between the disease clusters, repeating the functional analysis.

When we explored the summary results for GO molecular function (Figure 6), we observed that a larger number of clusters (12) have significant functional categories than without gene expansion (10). Additionally, the gene expansion increased the number of significant functions and the functional overlap between disease clusters, specially highlighted when full results are inspected (Supplementary Figure 5). We also observed a functional specialization for all disease clusters in Figure 6. However, we found two disease cluster pairs, 1-9 and 3-8, which present a noticeable number of functions with several categories shared across the pairs. Clusters 1 and 9 have DNA repair and histone regulation functions in common, both related to angiogenesis involved in neovascularization in several pathologies [4]. Regarding clusters 3 and 8, they both share terms related to cascade signaling regulation, including "insuline-like

[psz 13] UD

[psz 14] R3

[psz 15] UD

[psz 16] MA-07

[psz 17] R3

growth factor binding" and "AMP activated protein kinase activity", coupled to "oxidoreductase activity" and "ubiquitin protein ligase binding". All these activities are used as angiogenesis inhibition targets [18, 19]. Furthermore, cluster 12 has a large amount of specific functions and it is very different from the others. In fact, this cluster is focused around the terms "mitogen-activated protein kinase binding" and "MAP kinase activity". The first term corresponds with an essential function in endothelial cells angiogenesis performed by the regulatory gene mitogen-activated protein kinase phosphatase-1 (*MKP-1*), related to the second term. In addition, clusters 4 and 16 are worth mentioning due to their number of functional terms. Cluster 4 has several signaling related terms along with the "platelet-derived growth factor receptor binding" and "growth factor receptor activity", which are essential in vasculature formation, acting in a similar way to *VEGF* pathway [13]. In the case of cluster 16, it has a pair of terms related to transcription factor activity and the "nitric-oxide synthase binding" term, along with signaling or chromatin regulation involved functions. It is also known that nitric-oxide triggers angiogenesis through the *VEGF* pathway [29].

In the case of GO biological process (Supplementary Figure 6), due to the large number of functions associated with the genes for each cluster, we show only the summary results. As in the case of GO molecular function, here we can see again how the gene expansion finds functional terms for two clusters that were not available with the original genes. [psz]

In any case, the gene expansion approach allows us to identify the participation of diseases or molecular mechanisms of previously not related genes, and to contribute to reveal the biological basis of these diseases.

Compare ClinVar data with angiogenesis related genes[psz]

[psz 18] R1

To illustrate the value of the relationships found between A-RDs, their associated genes and the inferred genes through interaction data, we compared our results with genomic known data. For this, we explored known pathogenic variants related to angiogenesis. We searched for the keyword "*angiogen**" in the ClinVar database and selected variants with clinical significance defined as pathogenic and whose length was less or equal to 50 nt to ensure that they only affected a single gene. We got a list of 16 pathogenic variants included in the Supplementary Table 4. From this pool of variants, we used the gene identifiers associated by ClinVar and checked if they were involved in angiogenesis by performing a bibliographic search (Supplementary Table 5). In Table 2 we show the comparison between the angiogenesis related genes from ClinVar and the gene lists obtained in this study. The genes *ANG*, *SDHA*, *SDHD*, *TP53* and *VEGFC* were found between the genes associated with the A-RDs by the Monarch Initiative. The *F7* gene was found when the gene clusters were expanded with the interaction data. This *F7* gene is a coagulation factor related to angiogenesis, but the study [1] was overlooked by our bibliographic search. This work has all the features described in Material and methods to be included; however, it has no mention to the *VEGF* gene although it relates *F7* with angiogenesis. There are three genes that our study did not relate to angiogenesis: *MT-TE*, *RNASE4* and *AIMP*. In the first case, *MT-TE* is a mitochondrial gene that encodes a tRNA for glutamic acid. Mutations in this gene are known that produces myopathies [16] or diabetes mellitus [39], but there are no studies relating this gene to angiogenesis at the disease level. This gene does not encode a protein, therefore the protein interaction data is useless. In the case of the *RNASE4* gene, we found its genomic coordinates overlapping the *ANG* gene coordinates and its eight associated variants are also affecting the *ANG* gene. In fact, when the identifiers of these variants were inspected in Supplementary Table 4 for these two genes, they referenced the *ANG* gene but not *RNASE4*. Consequently, we could consider that the pathogenic variants affect the *ANG* gene but not *RNASE4*, and *ANG* was identified as an angiogenesis related gene. In the case of the *AIMP* gene, it was not identified at all due to the two following reasons: this gene causes the Pelizaeus-Merzbacher-like disease encoded as ORPHA:280293 but it has not HPO terms described in the Monarch Initiative, and its related publication ([6]) does not include the "rare" keyword. Consequently, our criteria ignores it although it lists five different OMIM entries. This highlights that our methodology can identify six of seven angiogenesis related genes with known pathogenic variants. [psz]

Prioritization of gene groups associated with A-RD clusters in the interaction network[psz]

Finally, we applied a prioritization approach to the A-RD clusters using their associated genes. In this way, we could rank the A-RD clusters in accordance with the network connectivity features of the associated genes in the interaction network, rewarding the clusters with high interconnectivity. This ranking allows us to select which clusters (raw or expanded) are promising candidates for downstream experiments. For this reason, we used the CRank algorithm [59] that uses Likelihood, Density, Boundary and Allegiance metrics to characterize network connectivity features for each cluster in an integrated way. These metrics measure the structural features and the robustness against noise of the clusters in the network structure. The integrated CRank value was computed for both raw and expanded clusters, as shown in 7. Regarding the A-RD clusters with the raw gene lists, clusters 7, 4 and 13 were in the top three with 1, 1 and 0.86 CRank values, respectively. When the gene lists were expanded with the ASP computation, three clusters highly increased their CRank measures, whereas six clusters decreased their values to a lesser extent. Noteworthy, with the expanded gene list, clusters 7 and 4 decreased their CRank to 0.58, but clusters 15, 18 and 6 increased this value to 0.92, 1 and 0.84, respectively. As can be seen, the top of prioritization changes if the gene clusters are expanded. [psz] In this way, the gene association for the ASP calculation can be measured, giving the opportunity to choose which clusters need further investigation.

[psz 19] UD

Concluding remarks

The approach presented in this work has allowed us to deepen our knowledge of angiogenesis-related rare diseases (A-RDs) at the genetic, phenotypic and molecular levels. Starting from the work of Rodríguez-Caso and collaborators, we have characterized the phenotypes of A-RDs to be able to group them according to their semantic similarity. This allowed us to analyse disease clusters at the genetic level, exploring the molecular mechanisms involved in the development of different diseases. Likewise, we have demonstrated the coherence of the diseases within each cluster at the genetic level with the use of network characteristics such as the average shortest path. This approach allows the identification of clusters whose genes are

very close in the network of interactions and that could be involved in related molecular mechanisms. Besides, we propose the CRank measure to prioritize the A-RD clusters and to select candidates for downstream experiments in wet laboratories to validate the new gene associations.

In addition, this strategy has been essential to determine the common molecular mechanisms of these diseases. It also allowed us to explore the putative genes that could be associated with the A-RDs and whose function is not well characterized, considering them as possible genes involved in the disease development. Furthermore, we have compared the results in this work with ClinVar angiogenesis-related data and we have achieved to list six of seven genes, one of them through the interaction data. This demonstrates that further experimental validation of the genes involved in angiogenesis, it would be necessary to verify their function for our study. Finally, our protocol can be extrapolated to the analysis of other diseases or biological processes. [psz]

Key Points

- We propose a systemic methodology for the study of a set of rare diseases, grouping them according to their phenotypic similarity and analysing them at a functional level using their disease-associated genes.
- This methodology is used to identify possible genes involved in angiogenesis-related rare diseases for those cases in which the genetic cause or functional impact is not known.
- We applied our methodology to the study of angiogenesis-related rare diseases, but it can be used to analyse other human genetic diseases.

Competing interests

There is NO Competing Interest.

Author contributions statement

E.R. and P.S.Z. conceived the methodology. R.P.M., J.C.C., E.R. and P.S.Z., developed the software that implements the protocol. R.P.M., J.C.C., E.R., P.S.Z. and M.A.M. analysed the results and provided interpretation. R.P.M., E.R. and P.S.Z. wrote the manuscript. B.M.P., J.A.G., A.R.Q. and M.A.M. were involved in planning of the study, contributed to the acquisition of funding for research and headed the project. All authors read and approved the final version of the manuscript.

Acknowledgments

This work was supported by the Spanish Ministry of Science, Innovation and Universities (grant PID2019-105010RB-I00, grant PID2019-108096RB-C21), the Andalusian Government and FEDER (grants UMA18-FEDERJA-102, UMA18-FEDERJA-220, PY20_00257, PY20_00372 and funds from the group PAIDI BIO 267); the Ramón Areces foundation, which funds project for the investigation of rare disease (National call for research on life and material sciences, XIX edition) and the University of Malaga (Ayudas del I Plan Propio). The "CIBER de Enfermedades Raras" and "CIBER de Enfermedades Cardiovasculares" are initiatives from the ISCIII (Spain). The funders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript. The authors thank the Supercomputing and Bioinnovation Center (SCBI) of the University of Malaga for their provision of computational resources and technical support (http://www.scbi.uma.es/site).

References

1. F. Bernardi and G. Mariani. Biochemical, molecular and clinical aspects of coagulation factor VII and its role in hemostasis and thrombosis. *Haematologica*, 106(2):351–362, jan 2021.
2. P. Buphamalai, T. Kokotovic, V. Nagy, and J. Menche. Network analysis reveals rare disease signatures across multiple levels of biological organization. *Nature Communications*, 12(1):6306, 2021.
3. P. Carmeliet and R. K. Jain. Angiogenesis in cancer and other diseases. *Nature*, 407(6801):249–257, 2000.
4. T. Chavakis, V. V. Orlova, and H. F. Langer. A possible crosstalk between DNA repair pathways and angiogenesis, 2009.
5. E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik 1959 1:1*, 1(1):269–271, dec 1959.
6. M. Feinstein, B. Markus, I. Noyman, H. Shalev, H. Flusser, I. Shelef, K. Liani-Leibson, Z. Shorer, I. Cohen, S. Khateeb, S. Sivan, and O. S. Birk. Pelizaeus-merzbacher-like disease caused by AIMP1/p43 homozygous mutation. *American Journal of Human Genetics*, 2010.
7. D. N. Fisman. Hemophagocytic syndromes and infection. *Emerging Infectious Diseases*, 6(6):601, 2000.
8. G. Fontemaggi, S. Dell'Orso, D. Trisciuoglio, T. Shay, E. Melucci, F. Fazi, I. Terrenato, M. Mottolese, P. Muti, E. Domany, D. Del Bufalo, S. Strano, and G. Blandino. The execution of the transcriptional axis mutant p53, E2F1 and ID4 promotes tumor neo-angiogenesis. *Nature Structural and Molecular Biology*, 16(10):1086–1093, sep 2009.
9. T. Galili, A. O'Callaghan, J. Sidi, and C. Sievert. Heatmaply: An R package for creating interactive cluster heatmaps for online publishing. *Bioinformatics*, 34(9), 2018.
10. L. Ge, R.-p. Zhang, F. Wan, D.-y. Guo, P. Wang, L.-x. Xiang, and J.-z. Shao. TET2 Plays an Essential Role in Erythropoiesis by Regulating Lineage-Specific Genes via DNA Oxidative Demethylation in a Zebrafish Model. *Molecular and Cellular Biology*, 34(6):989, mar 2014.
11. M. A. Gosein, D. Narinesingh, C. A. A. C. Nixon, S. R. Goli, P. Maharaj, and A. Sinanan. Multi-organ benign and malignant tumors: Recognizing Cowden syndrome: A case report and review of the literature, 2016.
12. T. Groza, S. Köhler, D. Moldenhauer, N. Vasilevsky, G. Baynam, T. Zemojtel, L. Schriml, W. Kibbe, P. Schofield, T. Beck, D. Vasant, A. Brookes, A. Zankl, N. Washington, C. Mungall, S. Lewis, M. A. Haendel, H. Parkinson, and P. Robinson. The Human Phenotype Ontology: Semantic Unification of Common and Rare Disease. *American Journal of Human Genetics*, 97(1):111, 2015.
13. R. P. Gude, P. Patil, M. Z. Kamran, and P. N. Goel. Development of Novel Anti-Cancer Strategies Based on Angiogenesis Inhibition. *Anti-Angiogenesis Drug Discovery and Development*, 2:147–190, jan 2014.
14. A. A. Hagberg, D. A. Schult, and P. J. Swart. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy)*, 2008.
15. C. Himbert, J. Ose, T. Lin, C. Warby, B. Gigic, K. Steindorf, P. Schrotz-King, C. Abbenhardt-Martin, L. Zielske,

J. Boehm, and C. Ulrich. Inflammation- and angiogenesis-related biomarkers are correlated with cancer-related fatigue in colorectal cancer patients: Results from the ColoCare Study. *European journal of cancer care*, 28(4), jul 2019.

16. R. Horvath, J. P. Kemp, H. A. Tuppen, G. Hudson, A. Oldfors, S. K. Marie, A. R. Moslemi, S. Servidei, E. Holme, S. Shanske, G. Kollberg, P. Jayakar, A. Pyle, H. M. Marks, E. Holinski-Feder, M. Scavina, M. C. Walter, J. Oku, A. Günther-Scholz, P. M. Smith, R. McFarland, Z. M. Chrzanowska-Lightowlers, R. N. Lightowlers, M. Hirano, H. Lochmüller, R. W. Taylor, P. F. Chinnery, M. Tulinius, and S. Dimauro. Molecular basis of infantile reversible cytochrome c oxidase deficiency myopathy. *Brain*, 2009.
17. M. Iwabayashi, Y. Taniyama, F. Sanada, J. Azuma, K. Iekushi, H. Kusunoki, A. Chatterjee, K. Okayama, H. Rakugi, and R. Morishita. Role of serotonin in angiogenesis: Induction of angiogenesis by sarpogrelate via endothelial 5-HT1B/Akt/eNOS pathway in diabetic mice. *Atherosclerosis*, 220(2):337–342, feb 2012.
18. H. J. Jung, K. H. Kim, N. D. Kim, G. Han, and H. J. Kwon. Identification of a novel small molecule targeting UQCRB of mitochondrial complex III and its anti-angiogenic activity. *Bioorganic and Medicinal Chemistry Letters*, 21(3):1052–1056, feb 2011.
19. H. J. Jung and H. J. Kwon. Exploring the role of mitochondrial UQCRB in angiogenesis using small molecules. *Molecular BioSystems*, 9(5):930–939, 2013.
20. S. Köhler, N. A. Vasilevsky, M. Engelstad, E. Foster, J. McMurry, S. Aymé, G. Baynam, S. M. Bello, C. F. Boerkoel, K. M. Boycott, M. Brudno, O. J. Buske, P. F. Chinnery, V. Cipriani, L. E. Connell, H. J. Dawkins, L. E. DeMare, A. D. Devereau, B. de Vries, H. V. Firth, K. Freson, D. Greene, A. Hamosh, I. Helbig, C. Hum, J. A. Jähn, R. James, R. Krause, S. J. F. Laulederkind, H. Lochmüller, G. J. Lyon, S. Ogishima, A. Olry, W. H. Ouwehand, N. Pontikos, A. Rath, F. Schaefer, R. H. Scott, M. Segal, P. I. Sergouniotis, R. Sever, C. L. Smith, V. Straub, R. Thompson, C. Turner, E. Turro, M. W. Veltman, T. Vulliamy, J. Yu, J. von Ziegenweidt, A. Zankl, S. Züchner, T. Zemojtel, J. O. Jacobsen, T. Groza, D. Smedley, C. J. Mungall, M. Haendel, and P. N. Robinson. The Human Phenotype Ontology in 2017. *Nucleic Acids Research*, 45(D1):D865–D876, jan 2017.
21. C. Lamy and J. L. Mas. Hypertensive Encephalopathy. *Stroke*, pages 734–740, 2011.
22. P. Langfelder, B. Zhang, and S. Horvath. Defining clusters from a hierarchical cluster tree: The Dynamic Tree Cut package for R. *Bioinformatics*, 24(5):719–720, 2008.
23. C. E. Lee, K. S. Singleton, M. Wallin, and V. Faundez. Rare Genetic Diseases: Nature's Experiments on Human Development. *iScience*, 23(5), may 2020.
24. L. Lenke, G. Martínez de la Escalera, C. Clapp, T. Bertsch, and J. Triebel. A Dysregulation of the Prolactin/Vasoinhibin Axis Appears to Contribute to Preeclampsia. *Frontiers in Endocrinology*, 10:893, jan 2020.
25. L. Levy, A. Nasereddin, M. Rav-Acha, M. Kedmi, D. Rund, and M. E. Gatt. Prolonged Fever, Hepatosplenomegaly, and Pancytopenia in a 46-Year-Old Woman. *PLoS Medicine*, 6(4), apr 2009.
26. B.-Q. Li, T. Huang, L. Liu, Y.-D. Cai, and K.-C. Chou. Identification of Colorectal Cancer Related Genes with mRMR and Shortest Path in Protein-Protein Interaction Network. *PLOS ONE*, 7(4):e33393, apr 2012.
27. S. E. Maynard and S. A. Karumanchi. Angiogenic Factors and Preeclampsia. *Seminars in nephrology*, 31(1):33, jan 2011.
28. R. Mejzini, L. L. Flynn, I. L. Pitout, S. Fletcher, S. D. Wilton, and P. A. Akkari. ALS Genetics, Mechanisms, and Therapeutics: Where Are We Now? *Frontiers in Neuroscience*, 13(1310), dec 2019.
29. L. Morbidelli, S. Donnini, and M. Ziche. Role of Nitric Oxide in the Modulation of Angiogenesis. *Current Pharmaceutical Design*, 105(18):2133–2135, 2005.
30. C. J. Mungall, J. A. McMurry, S. Kohler, J. P. Balhoff, C. Borromeo, M. Brush, S. Carbon, T. Conlin, N. Dunn, M. Engelstad, E. Foster, J. P. Gourdine, J. O. Jacobsen, D. Keith, B. Laraway, S. E. Lewis, J. N. Xuan, K. Shefchek, N. Vasilevsky, Z. Yuan, N. Washington, H. Hochheiser, T. Groza, D. Smedley, P. N. Robinson, and M. A. Haendel. The Monarch Initiative: An integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Research*, 45(D1):D712–D722, jan 2017.
31. F. Murtagh and P. Legendre. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification*, 2014.
32. E. Nazar, F. Khatami, H. Saffar, and S. M. Tavangar. The Emerging Role of Succinate Dehyrogenase Genes (SDHx) in Tumorigenesis. *International Journal of Hematology-Oncology and Stem Cell Research*, 13(2):72, 2019.
33. S. Nguengang Wakap, D. M. Lambert, A. Olry, C. Rodwell, C. Gueydan, V. Lanneau, D. Murphy, Y. Le Cam, and A. Rath. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *European Journal of Human Genetics*, 28(2):165–173, feb 2020.
34. S. Pavan, K. Rommel, M. E. M. Marquina, S. Höhn, V. Lanneau, and A. Rath. Clinical practice guidelines for rare diseases: The orphanet database. *PLoS ONE*, 12(1), jan 2017.
35. C. Pesquita, D. Faria, H. Bastos, A. E. Ferreira, A. O. Falcão, and F. M. Couto. Metrics for GO based protein semantic similarity: A systematic evaluation. *BMC Bioinformatics*, 9(SUPPL. 5):1–16, apr 2008.
36. J. Piñero, J. M. Ramírez-Anguita, J. Saüch-Pitarch, F. Ronzano, E. Centeno, F. Sanz, and L. I. Furlong. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research*, 48(D1):D845–D855, 2020.
37. A. Poduri, G. D. Evrony, X. Cai, and C. A. Walsh. Somatic Mutation, Genomic Variation, and Neurological Disease. *Science (New York, N.Y.)*, 341(6141):1237758, 2013.
38. A. Rao, T. Joseph, V. G. Saipradeep, S. Kotte, N. Sivadasan, and R. Srinivasan. Priori-T: A tool for rare disease gene prioritization using MEDLINE. *PLoS ONE*, 2020.
39. L. Rigoli, F. Prisco, R. A. Caruso, D. Iafusco, G. Ursomanno, D. Zuccarello, N. Ingenito, M. Rigoli, and I. Barberi. Association of the T14709C mutation of mitochondrial DNA with maternally inherited diabetes mellitus and/or deafness in an Italian family [2], 2001.
40. L. Rodríguez-Caso, A. Reyes-Palomares, F. Sánchez-Jiménez, A. R. Quesada, and M. Á. Medina. What is known on angiogenesis-related rare diseases? A systematic review of literature. *Journal of Cellular and Molecular Medicine*, 16(12):2872–2893, dec 2012.
41. E. Rojano, J. Córdoba-Caballero, F. M. Jabato, D. Gallego, M. Serrano, B. Pérez, Á. Parés-Aguilar, J. R. Perkins, J. A. G. Ranea, and P. Seoane-Zonjic. Evaluating, Filtering and Clustering Genetic Disease Cohorts Based on Human

Phenotype Ontology Data with Cohort Analyzer. *Journal of Personalized Medicine 2021, Vol. 11, Page 730*, 11(8):730, jul 2021.

42. E. Rojano, P. Seoane-Zonjic, F. M. Jabato, J. R. Perkins, and J. A. G. Ranea. Comprehensive Analysis of Patients with Undiagnosed Genetic Diseases Using the Patient Exploration Tools Suite (PETS). In *In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 12108, pages 775–786, 2020.
43. S. Roy, C. J. McCann, M. Ralle, K. Ray, J. Ray, S. Lutsenko, and S. Jayakanthan. Analysis of Wilson disease mutations revealed that interactions between different ATP7B mutants modify their properties. *Scientific Reports 2020 10:1*, 10(1):1–15, aug 2020.
44. N. Sawada and Z. Arany. Metabolic Regulation of Angiogenesis in Diabetes and Aging. *Physiology*, 32(4):290, jun 2017.
45. P. Seoane, S. Ocaña, R. Carmona, R. Bautista, E. Madrid, A. Torrres, and G. Claros. AutoFlow, a Versatile Workflow Engine Illustrated by Assembling an Optimised de novo Transcriptome for a Non-Model Species, such as Faba Bean (Vicia faba). *Current Bioinformatics*, 11(4):440–450, 2016.
46. P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, 13(11):2498, nov 2003.
47. L. T. Slater, A. Karwath, J. A. Williams, S. Russell, S. Makepeace, A. Carberry, R. Hoehndorf, and G. V. Gkoutos. Towards similarity-based differential diagnostics for common diseases. *Computers in Biology and Medicine*, 133(104360), 2021.
48. L. T. Slater, J. A. Williams, A. Karwath, H. Fanning, S. Ball, P. N. Schofield, R. Hoehndorf, and G. V. Gkoutos. Multi-faceted semantic clustering with text-derived phenotypes. *Computers in Biology and Medicine*, 138, 2021.
49. Z. Szekanecz and A. E. Koch. Mechanisms of Disease: angiogenesis in inflammatory diseases. *Nature Clinical Practice Rheumatology 2007 3:11*, 3(11):635–643, nov 2007.
50. D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork, L. J. Jensen, and C. Mering. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47(D1):D607–D613, jan 2019.
51. E. A. Trifonova, M. G. Swarovskaya, O. A. Ganzha, O. V. Voronkova, T. V. Gabidulina, and V. A. Stepanov. The interaction effect of angiogenesis and endothelial dysfunction-related gene variants increases the susceptibility of recurrent pregnancy loss. *Journal of Assisted Reproduction and Genetics*, 36(4):717–726, 2019.
52. N. X. Vinh, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 2010.
53. C. von Mering, L. J. Jensen, B. Snel, S. D. Hooper, M. Krupp, M. Foglierini, N. Jouffre, M. A. Huynen, and P. Bork. STRING: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic Acids Research*, 33(Database Issue):D433, jan 2005.
54. J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C. F. Chen. A new method to measure the semantic similarity of GO terms. *Bioinformatics*, 23(10):1274–1281, 2007.
55. Y. Xu, Y.-Y. Xiao, M. Simon, T. T. Aye, K. Khiani, Y. Liu, Y. Huang, and J. Burton. Plasma Vascular Endothelial Growth Factor (VEGF) Levels Correlate with Thrombocytopenia of Various Etiology. *Blood*, 124(21):4991–4991, dec 2014.
56. Y. Xue, F. Chen, D. Zhang, S. Lim, and Y. Cao. Tumor-derived VEGF modulates hematopoiesis. *Journal of Angiogenesis Research*, 1(1):9, 2009.
57. J. Yang, C. Dong, H. Duan, Q. Shu, and H. Li. RDmap: a map for exploring rare diseases. *Orphanet Journal of Rare Diseases*, 16(101), 2021.
58. G. Yu, L. G. Wang, Y. Han, and Q. Y. He. ClusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS A Journal of Integrative Biology*, 16(5):284–287, may 2012.
59. M. Zitnik, R. Sosič, and J. Leskovec. Prioritizing network communities. *Nature Communications*, 2018.

Raquel Pagano-Márquez is a PhD student at the Department of Molecular Biology and Biochemistry, University of Malaga, Spain. Her research focuses on the analysis and interpretation of rare diseases dependent on angiogenesis. **José Córdoba-Caballero** is a PhD student at the Department of Molecular Biology and Biochemistry, University of Malaga, Spain. His research focuses on the development of bioinformatics tools for the analysis of patients with rare diseases. **Beatriz Martínez-Poveda** is an Associate Professor at the Department of Molecular Biology and Biochemistry, University of Malaga, Spain. Her research focuses on analysing the molecular factors involved in angiogenesis-related diseases, including cancer and atherosclerosis. **Juan AG. Ranea** is a professor at the Department of Molecular Biology and Biochemistry, University of Malaga, Spain. He leads a research group focused on bioinformatics and systems biology for the analysis of rare diseases. **Ana R. Quesada** is a professor at the Department of Molecular Biology and Biochemistry, University of Malaga, Spain. She leads a research group focuses on the analysis of the molecular signaling pathways associated with angiogenesis and the search for modulating compounds of angiogenic activity. **Elena Rojano** is a post-doctoral researcher in bioinformatics at the Department of Molecular Biology and Biochemistry, University of Malaga, Spain. Her research focuses on systems biology methods for associating pathological phenotypes to genomic variants from patients with rare diseases. **Pedro Seoane-Zonjic** is a post-doctoral researcher in bioinformatics at the CIBER of Rare Diseases (CIBERER), University of Malaga, Spain. His current research focuses on developing software for the analysis of rare diseases from high-throughput genomic data. **Miguel Ángel Medina** is a professor at the Department of Molecular Biology and Biochemistry, University of Malaga, Spain. He leads a research group focused on systems biology, angiogenesis and cancer research, as well as rare diseases.

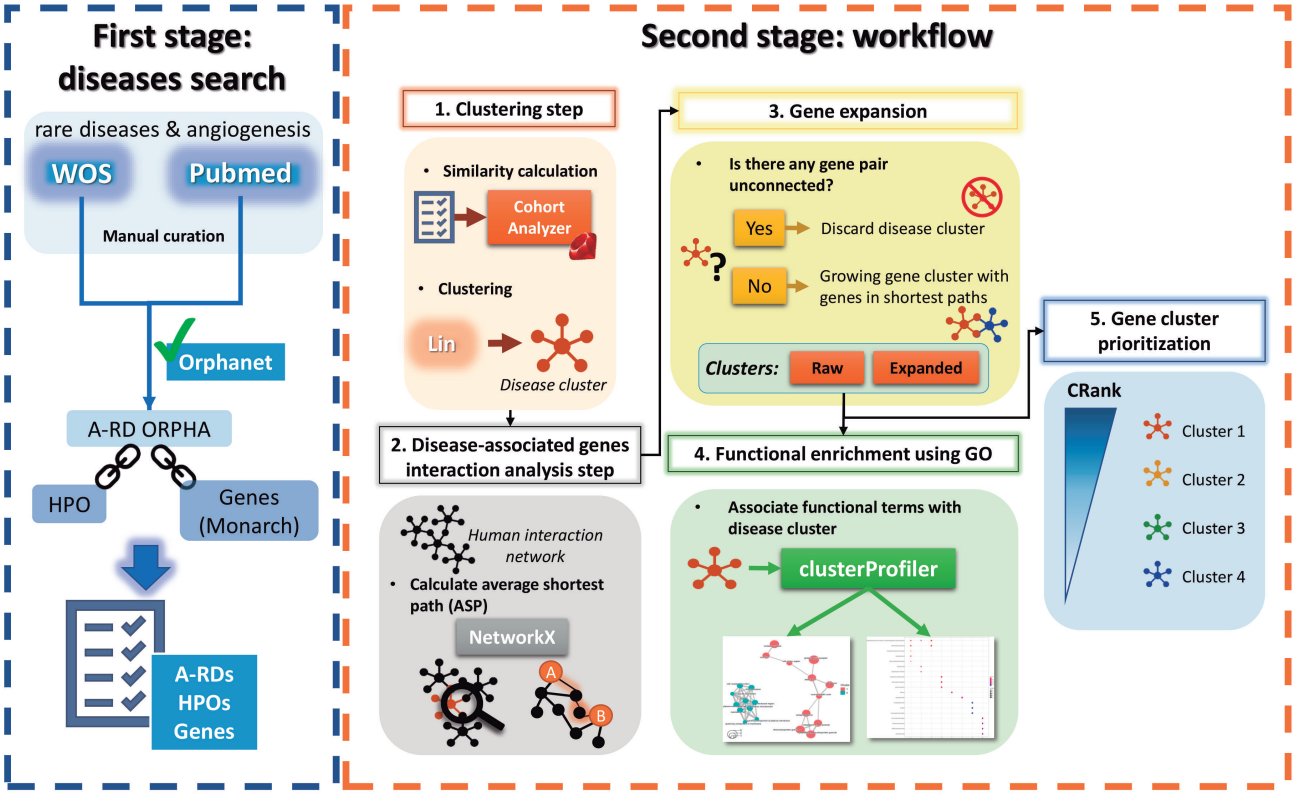


Fig. 1. Representation of the main stages and steps followed in this work. The first stage consisted on the search for angiogenesis-related rare diseases (A-RDs) through a systematic review of the literature. HPO terms associated with each A-RD were searched in the HPO annotation website and genes from the Monarch Initiative. The second stage describes the workflow developed in this work. The first step of this stage consists in retrieving the disease clusters. Then, in the second step, the average shortest path (ASP) is calculated among genes for each disease cluster in the interaction network. For clusters with computable paths between its associated genes in the interaction network, they are expanded with the genes that are presented in the computed shortest paths. Finally, raw and expanded gene clusters are used in GO enrichment analysis and cluster prioritization in the interaction network.[psz]

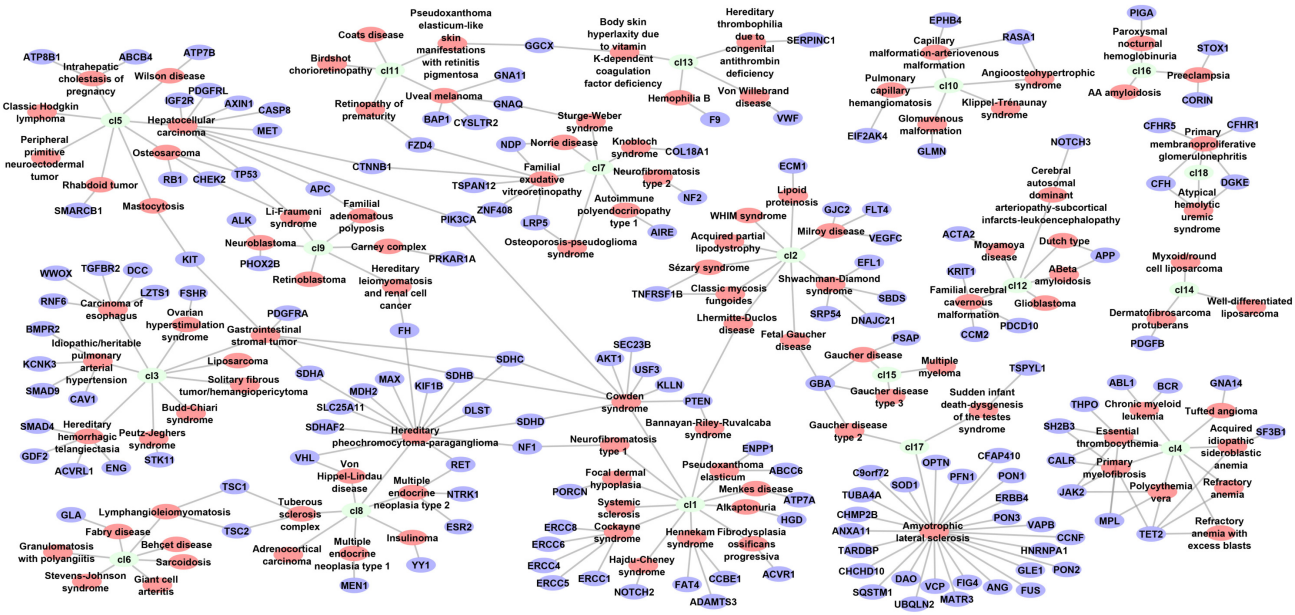


Fig. 2. Angiogenesis map of genes and diseases representation. Salmon circles represent A-RDs and green circles in which cluster they belong. Lilac circles are the genes associated with each disease.

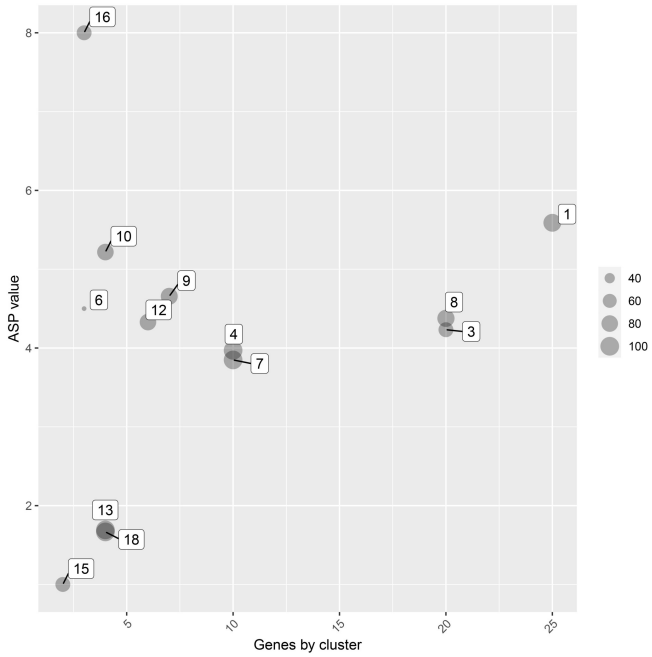


Fig. 3. Scatter plot representing the average shortest path (ASP) calculated among genes within each disease cluster. X-axis represents the number of genes by cluster and the Y-axis the ASP value. Dot size represents the number of diseases by cluster. Dots numbers are the identifier for each disease cluster.

Fig. 4. Dot plot for results obtained with the clusterProfiler R package in GO molecular function. X-axis includes the A-RD cluster identifiers and the number of genes by cluster between brackets. Y-axis represents each GO molecular function term associated with the genes for these clusters. Colour scale represents the adjusted p-value (red: lower, blue: higher) and dot size indicates the proportion of genes in the functional category that are annotated in the cluster.

Fig. 5. Heat map for results obtained with clusterProfiler R package in GO biological process. X-axis shows the A-RD cluster identifiers and Y-axis shows the summarized terms for the enrichment results as described in Material and methods, using 50 terms per cluster and a similarity threshold of 0.7.

Fig. 6. Heat map for results obtained with clusterProfiler R package in GO molecular function for expanded clusters. X-axis shows the A-RD cluster identifiers and Y-axis show the summarized terms for the enrichment results as described in Material and methods, using 35 terms per cluster and a similarity threshold of 0.6.

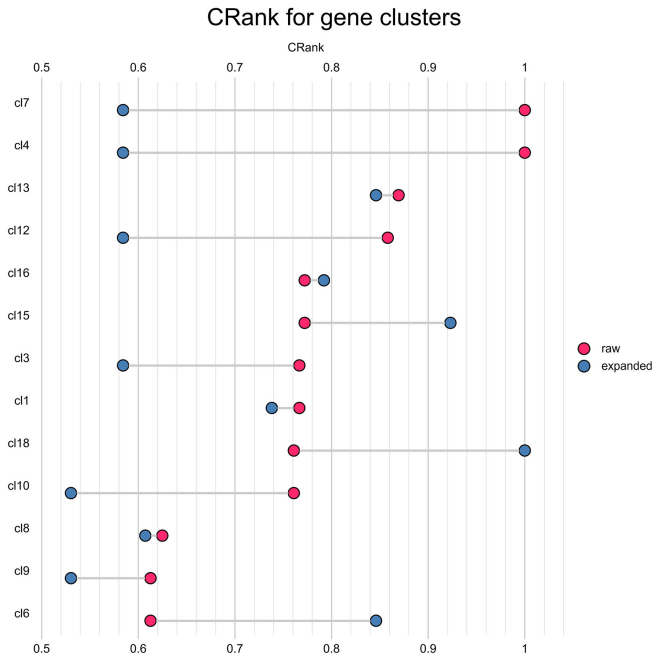


Fig. 7. CRank values for the genes associated with each A-RD cluster, using the interaction network from STRING filtered by a combined score of 900. Series show the CRank for raw clusters (pink) and gene expanded clusters (blue).

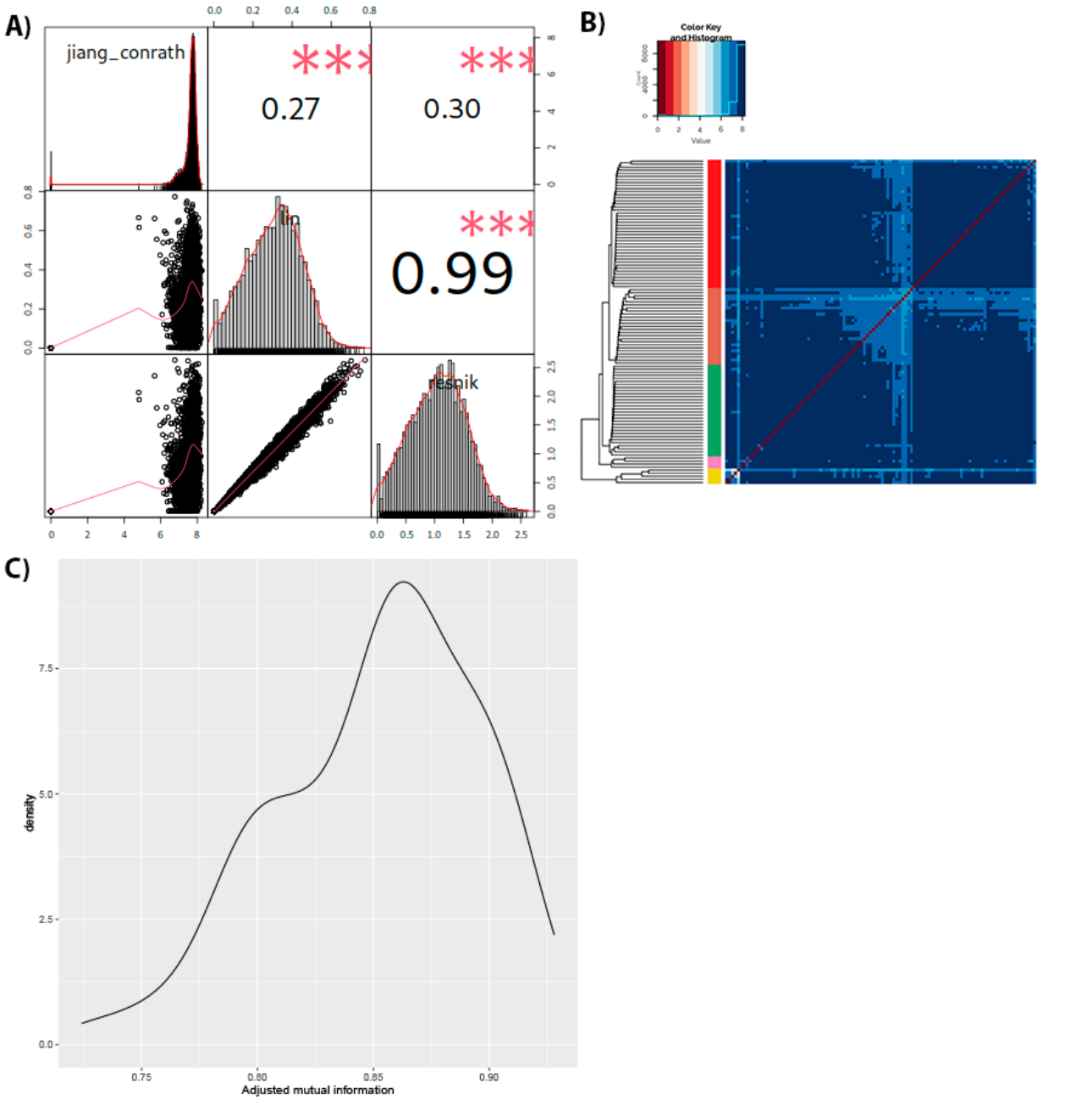
Table 1. Top ten most frequent HPOs in the A-RD cohort.

HPO	%
Seizure	28.03
Fatigue	25.23
Abdominal pain	20.56
Hepatomegaly	19.62
Splenomegaly	18.69
Weight loss	18.69
Thrombocytopenia	16.82
Fever	14.95
Headache	14.95
Hypertension	14.95

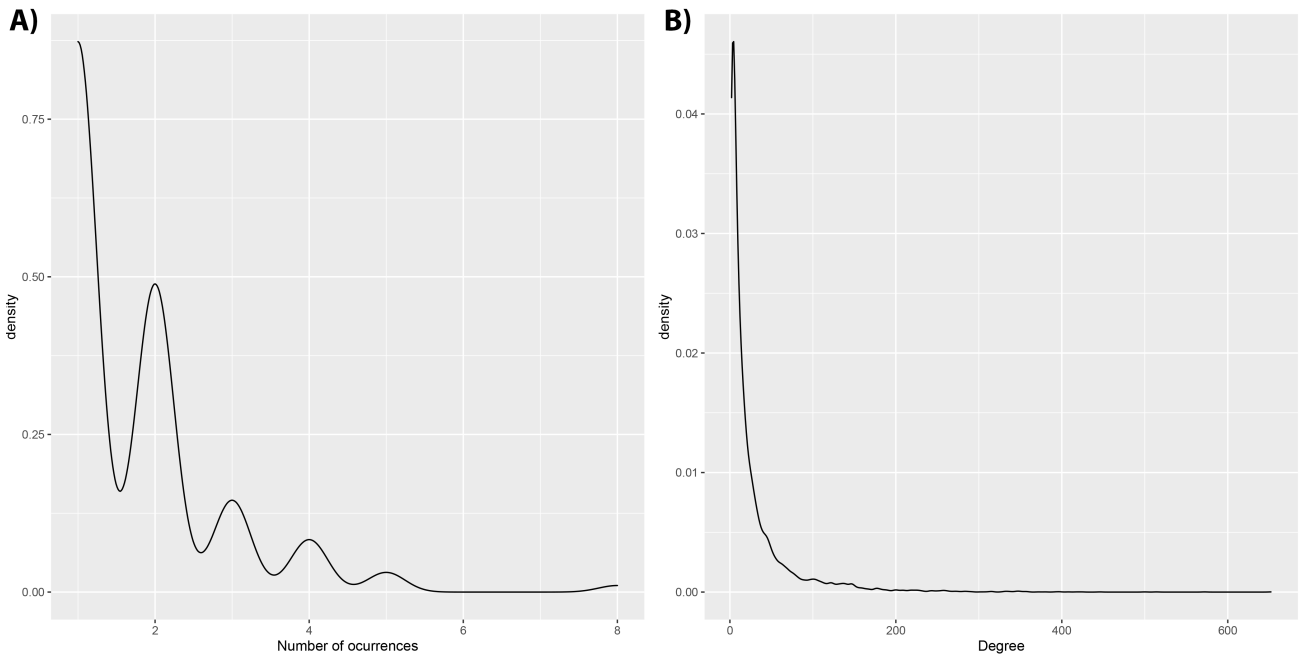
Table 2. List of genes affected by pathogenic variants associated with angiogenesis in ClinVar database. The Match column shows if the gene is identified in this study. *Associated genes* means that the gene is found in the list of genes retrieved from the Monarch Initiative, whereas *Expanded genes* means that the gene was found in the gene lists obtained with the STRING interaction data.

Gene	Associated variants	Match
AIMP1	1	No
ANG	8	Associated genes
RNASE4	8	No
F7	1	Expanded genes
MT-TE	1	No
SDHA	1	Associated genes
SDHD	1	Associated genes
TP53	1	Associated genes
VEGFC	1	Associated genes

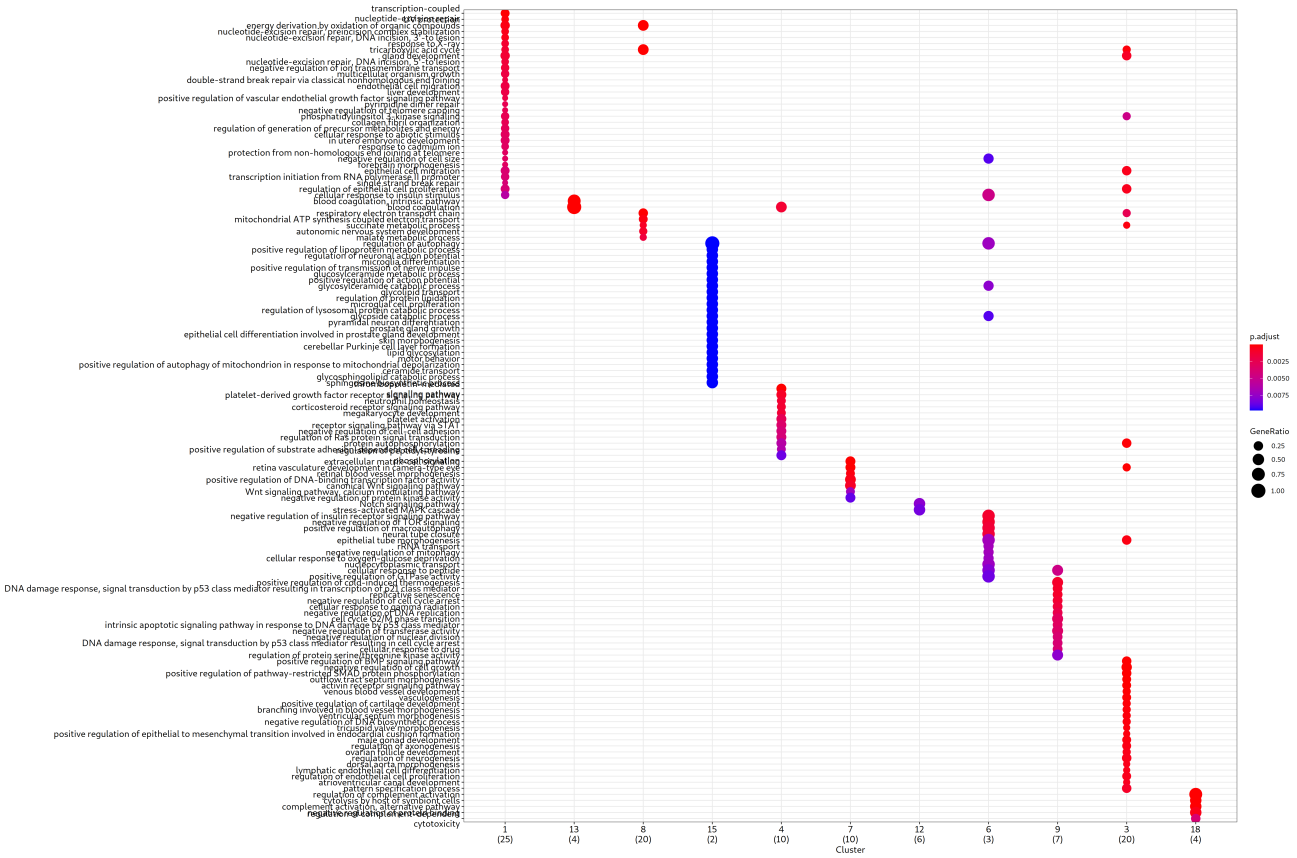
SUPPLEMENTARY DATA



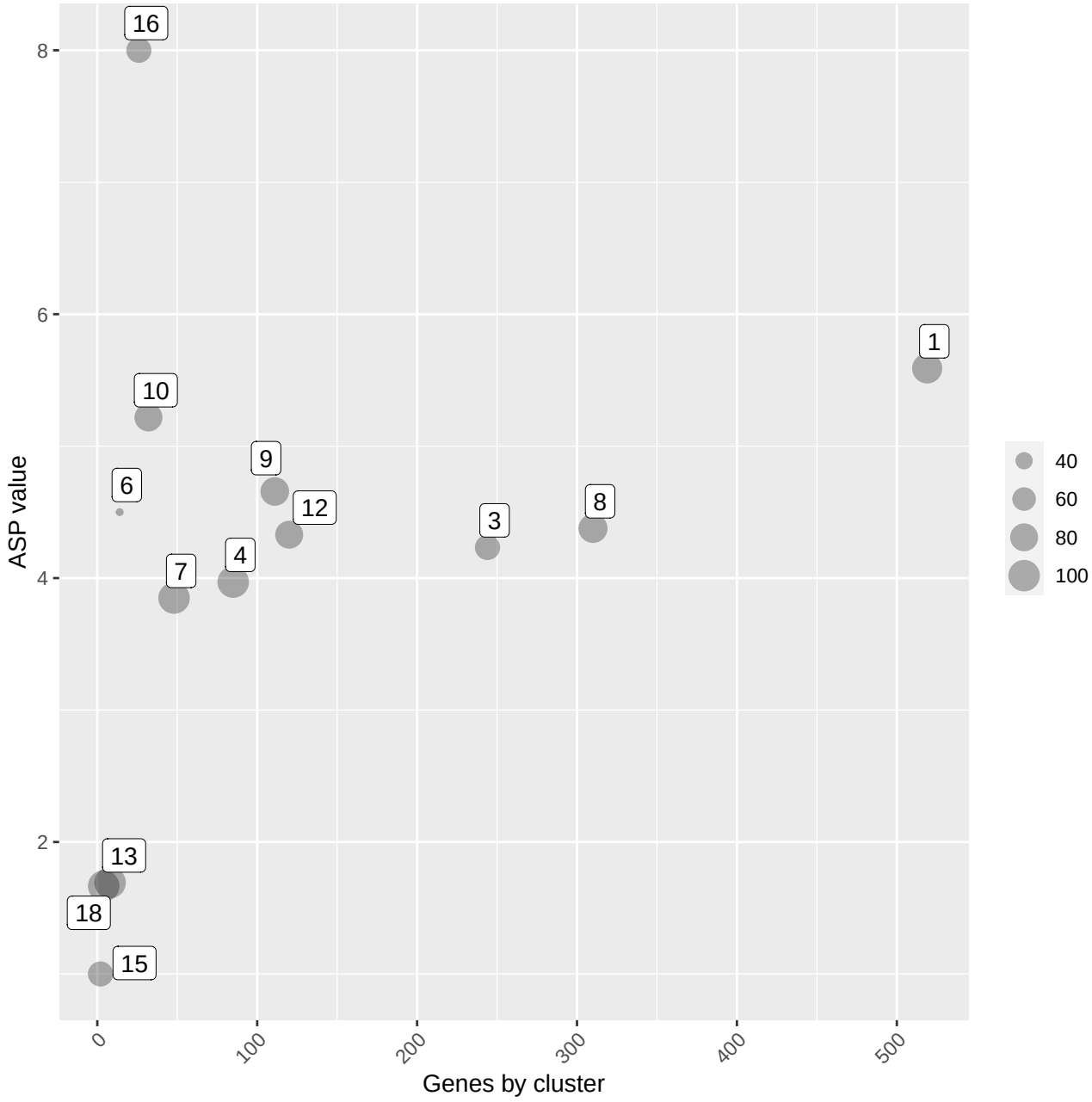
Supplementary Figure 1. Robustness of the clustering according to the list of the diseases and the semantic similarity measure. A) Correlation coefficient between Resnik, Lin and Jiang-Conrath similarity measures that Cohort Analyzer can compute. B) Jiang-Conrath similarity matrix and its clustering. Only can extract five clusters (coloured segments in vertical bar) so this semantic similarity measure is discarded. C) Distribution of the adjusted mutual information [52] computed between the clustering of the full set of diseases and 100 samples with the 99% of the initial elements. A value of 1 means that the sample clustering is identical to the full clustering whereas a value of 0 means that both partitions are completely different. The clustering uses a Lin similarity measure as its values range from 0 to 1 and has a correlation coefficient of 0.99 with Resnik, meaning that both are equivalent.



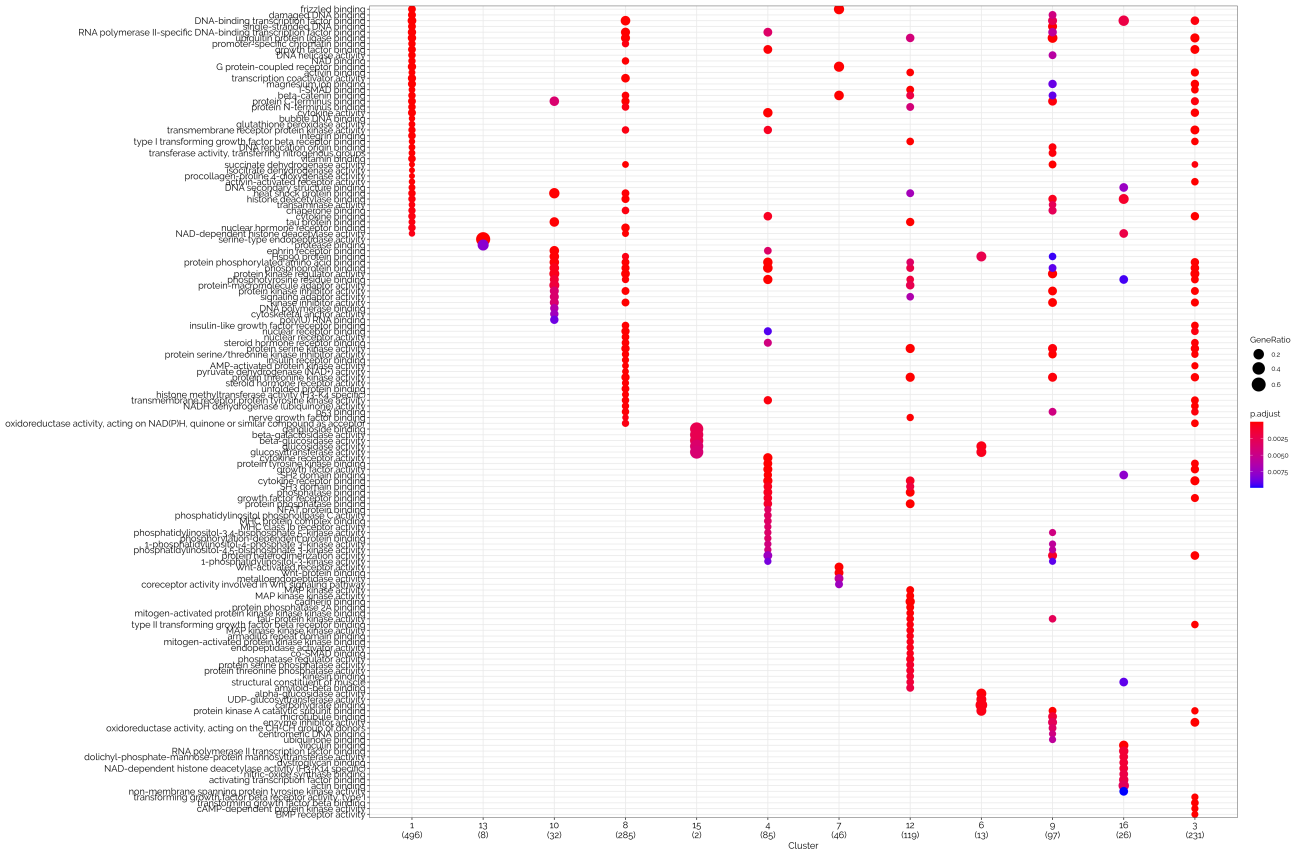
Supplementary Figure 2. A) Distribution of the frequency for each gene in the disease list retrieved in this study. B) Distribution of the gene degree in the STRING network when filtered by a combined score of 900. All genes with degree greater or equal than 200 are removed.



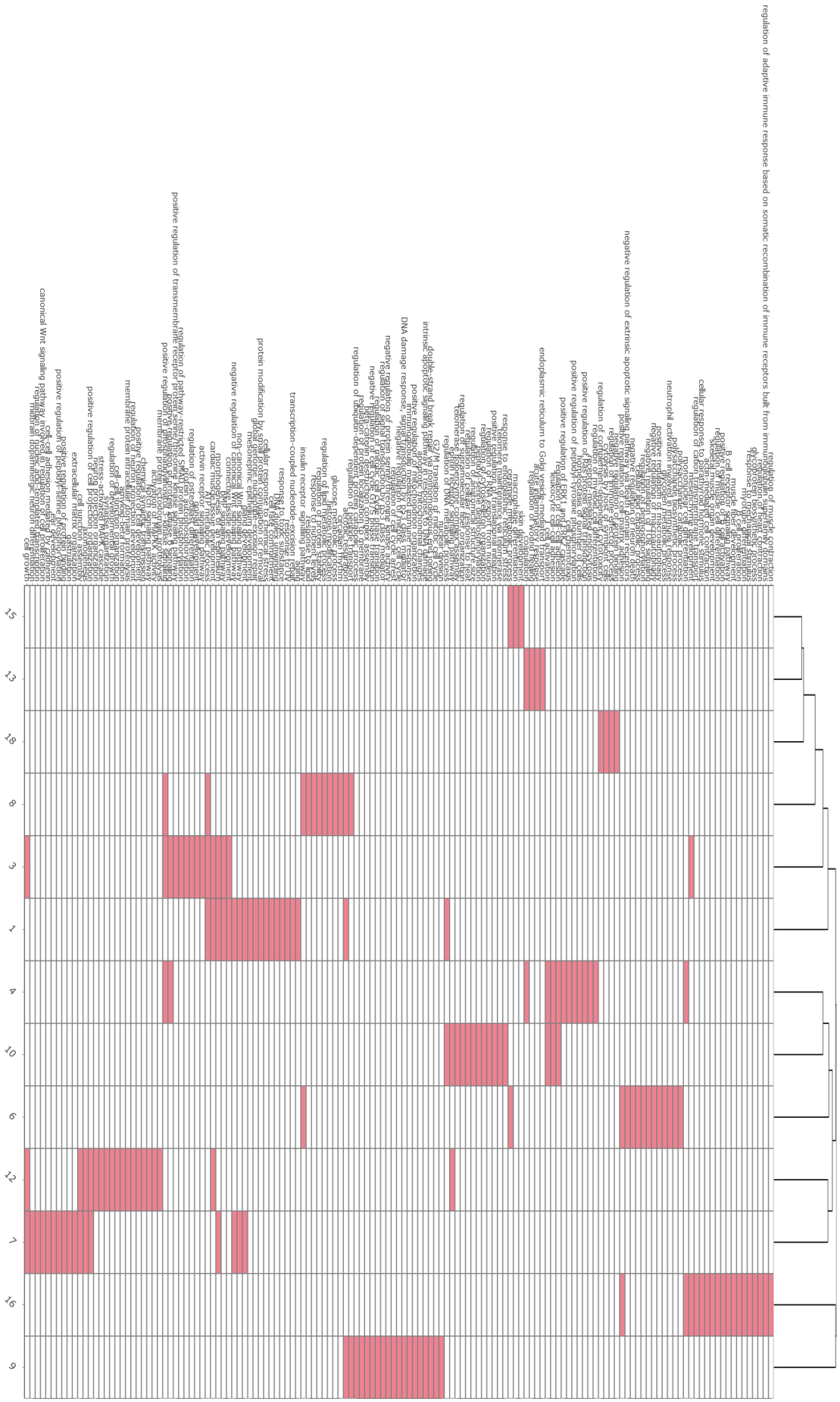
Supplementary Figure 3. Dot plot for results obtained with the clusterProfiler R package in GO biological process. X-axis includes the A-RD cluster identifiers and the number of genes by cluster between brackets. Y-axis represents each GO biological Process term associated with the genes for these clusters. Colour scale represents the adjusted p-value (red: lower, blue: higher) and dot size indicates the number of genes annotated with the same functional category.



Supplementary Figure 4. Average shortest path (ASP) calculated amongst genes within each disease cluster. Results shown correspond to the gene expanded network. X-axis represents the number of genes by cluster and the Y-axis the ASP value. Dots size represents the number of diseases by cluster.



Supplementary Figure 5. Dot plot for results obtained with the clusterProfiler R package in GO molecular function (gene expanded network). X-axis includes the A-RD cluster identifiers and the number of genes by cluster between brackets. Y-axis represents each GO molecular function term associated with the genes for these clusters. Colour scale represents the adjusted p-value (red: lower, blue: higher) and the dot size indicates the number of genes annotated with the same functional category.



Supplementary Figure 6. Heat map for results obtained with clusterProfiler R package in GO biological process for expanded clusters. X-axis shows the A-RD cluster identifiers and Y-axis shows the summarized terms for the enrichment results as described in methods using 35 terms per cluster and a similarity threshold of 0.6.

Table 1. For each A-RD identifier, it is included the full disease name, in which cluster it belongs, the HPO codes and descriptions and which genes has associated.

sup_tab1_disease_table.xlsx

Table 2. For each cluster, it is included the average shortest path (ASP) calculated amongst its genes, the gene identifiers according to the MONDO database and the genes obtained from the gene expansion.

sup_tab2_cluster_table.xlsx

Table 3. List of gene identifiers that have not been found in the STRING human interactions network.

sup_tab3_missing_genes.txt.xslx

Table 4. List of variants found in ClinVar when query for *"angiogen*"* keyword on 16th December 2021. Records were filtered by Clinical significance at least Likely pathogenic and variant length less or equal to 50 nt to ensure that only one gene is affected by the variant.

sup_tab4_clinvar_pathogenic_records.xlsx

Table 5. Genes associated with ClinVar variants and their relation with the angiogenesis process.

sup_tab5_clinvar_genes.xlsx