uc3m | Universidad Carlos III de Madrid

University Degree in Biomedical Engineering
2018-2019

Bachelor Thesis

"Integrating omics data from phenotypically-related genodermatoses. A Cytoscape approach using biological networks"

Hermes Piedra Zayas

Tutor: Carlos León Canseco

Leganés, February 2019

ABSTRACT

The ongoing advance of high-throughput sequencing technologies is bringing to the biomedical research community the opportunity to disclose relatively uncharted and poorly addressed domains in genetic disorders. Specifically, this project aims to shed new light on the molecular mechanisms of three rare skin diseases: Recessive Dystrophic Epidermolysis Bullosa (RDEB), Kindler Syndrome (KS) and Xeroderma pigmentosum type C (XPC). To accomplish this, biological network construction is leveraged herein, by providing a convenient approach to integrate and downstream analyze molecular omics data obtained from the comparison of these three genodermatoses (RDEB, KS & XPC) against healthy control samples. Concretely, microRNAs, RNAs and protein datasets are conjointly combined in the form of graphs whose structure and arrangement can be analyzed. On this basis, and upon computational procedures, the representation of high-throughput omics data across networks serves for both a topological and functional characterization of the molecular entities embedded within the graphs. Cytoscape software harbors the toolkits needed to exploit the massive omics information presented in this work, closely operating with online ontologies containing crucial annotations on the molecular entities under the network conglomerates. Cytoscape platform is going to carry out the bioinformatics computational endeavours, conducting then to new insights where common mechanisms and candidate biomarkers shared by the three genodermatoses will be highlighted. In this manner, STRING, BiNGO and ClueGO (Cytoscape plug-ins) will assist in the finding of enriched functions (such as “cell adhesions” and “epidermal growth factor signaling”), whereas the topological analysis will rely on STRING and NetworkAnalyzer, following the principles of graph theory to identify candidate molecules like TFAP2A and L1CAM. With the aid of manual curations, these two approaches will stand for a narrowing-down strategy from which biological interpretations are obtained.

Keywords: Bioinformatics, Recessive Dystrophic Epidermolysis Bullosa, Kindler Syndrome, Xeroderma pigmentosum type C, biological networks, Cytoscape, network medicine.

ACKNOWLEDGMENTS

A mi familia y amigos por las veces que no supe explicarme. Por las veces que no estuve.

Por lo que he desgastado.

A Carlos por enseñarme el rigor científico y empujarme hacia lo desconocido.

Table of Contents

List of Figures

List of Tables

1. INTRODUCTION

a. Systems Biology and Network Medicine

Systems Biology is an emerging inter-disciplinary field which makes use of advanced technologies to carry out cutting-edge computational analyses with the aim of unraveling deep biological knowledge [1]. System Biology offers an approach complementary to traditional molecular biology by considering the interrelationships between the components of the cell, the potential emerging properties and treating the cell as an integrated system. Next-generation sequencing technologies (NGS) allow a rapid and effective breakthrough to the field of genomic research. Genomic sequencing projects have brought to the scientific community a massive amount of high-throughput (HT) data over the last ten years since the completion of the Human Genome Project [2]. It has become a well-established and standard technique, which generates enormous amounts of biological information. An ample range of instrumentation can be used to quantify and characterize different molecules within cells and tissues. NGS platforms stand as a crucial procedure for the comprehensive understanding of the mechanisms that comprise the living organism phenomena and are therefore essential tools in the field of System Biology in order to obtain the necessary data that can be used for modelling the cell response to different stimulus.

Research laboratories are taking advantage of the high potential of these technologies. Thanks to the technological improvement, clinical investigations are catching up with desktop sequencers. However, this exponential escalation is generating a vast number of sequences from different organisms, tissues, conditions etc, leading to tons of gigabytes which need to be successfully analysed. In order to give an appropriate biological interpretation to the different mechanisms and pathways that conform the functioning of a given organism, the sequenced data must be thoroughly analysed, solving the bioinformatics bottle-neck that is impeding to study the applications of NGS [3].

Even though NGS data is promising and can generate crucial information for the implementation of Systems Biology, the current data portals are extremely fragmented. Data integration is giving its first steps and there is still a long way to go in terms of workflow unification, tool and software development. This issue is causing the researchers to struggle when it comes to going through the generated information and employing it for downstream research. Several online platforms offer material retrieved from different experiments in a disordered and difficult to understand way, causing ambiguity and inefficiency. To prevent this, several repositories are collecting not only the data but also the experimental design. However, there is no consensus yet in terms of data sharing.

The solution lies in passing through the “analysis bottle-neck”, building up a unified platform where the content is easily accessed and managed [4]. Here is where the role of bioinformatics comes in. The *in silico* interplay between NGS and informatics databases and repositories will definitely assist in handling high-throughput (HT) data.

An efficient storage, sharing and post-processing of NGS data will enable the study of human diseases and model organisms through a variety of frameworks, being the so-called biological networks the most notorious and suitable one [5]. Essentially, these networks are biological

models represented by graphs where the connected entities stand for some phenomenon. Namely, the physical entities (in our case, the molecules of interest are represented by nodes in the network) are connected by some type of interaction between them (edges), and among the rest of the network.

Even though biological networks are recognized as a disruptive method that will shed some light on the biomedical interpretation of data, systems medicine encounters difficulties to explain some pathogenesis that have not been addressed in previous investigations yet. This presents a shifting of the classical molecular biology reductionist approach towards the integrated holism proposed by system biology [6]. So far, many health conditions have been successfully diagnosed and treated using the methodological reductionism bases. It basically defends that, the small parts of a system -its isolated molecules- can be pieced together by the sum of all its physics and chemistry, starting from the simplest level and working upwards the whole organization. Reductionism, epitomized by molecular biology in the 20th century [7], has thrivingly served to give a strong and valid scientific explanation to numerous diseases and mechanisms, especially those based on single gene mutations.

The debate escalates with the Systems Biology apogee, increasingly cited and referenced in many review articles regarding disease pathology. Likewise, complex traits need a complementary insight, because reductionism cannot explain the system mechanisms as a whole [8]. In this way, holism appears as an integrative approach where the global properties of the system -emerging from their interactions- are studied rather than each constituent part. Systems are deterministic, but cannot be forecasted – *the whole is greater than the sum of its parts.* Just to take a clear example, the behaviour of a water molecule cannot be deduced simply from the behaviour of oxygen and hydrogen. What is more, the behaviour of water cannot be deduced from a single water molecule. Properties like flow and expansion cannot be attributed to single molecules.

Despite their conflicting nature, the evolution of these approaches cannot take place unless holism nourishes from reductionism, having a mutual dependence and complementing each other. Following this premise, clinical treatments will shift from reactive and generalized to predictive, preventive and personalized medicine [9].

Pathologies can be perceived as alterations in any biological network, where an error in the system gives rise to a re-wiring event. Therefore, large-scale networks contribute as convenient toolkits for the prototypic study of modern Systems Biology throughout quantitative modelling. Gene expression provides information on how functionally-related molecules are ubiquitously regulated within the cells and stroma that conforms tissues. Having this in mind, one can consider that the interaction among the genes expressed in each tissue tend to aggregate in the same connected components. These connected components resemble a biological network conformation. In addition, only 2% of human diseases result from a single gene defect (monogenic). Even on that count, monogenic disorders are rather heterogeneous in nature [10]. Although monogenic disorders are caused by a single gene defect, there is a wide phenotypical variability in patients with the same causal mutation, and not all the observable disease traits can be explained by the gene defect but rather by a combination of several dysregulations in gene expression profiles. For that reason, networks can be considered better markers of disease than single genes or combinations of them, putting thus the spotlight on overall arrangements that can be treated as a single piece using global parameters.

There are two elementary objects involved in any network: nodes and edges [11]. Depending on the type of biological network represented, nodes and edges will take different aspect and meaning. Nevertheless, nodes tend to represent discrete entities -such as genes, proteins or diseases-, whereas the edges are the linking elements which symbolize a generic relationship, that can be for example an interaction or a transcriptional control.

b. Biological networks. Types and examination

Prior to analysing the biological networks, it is convenient to understand the main topological measurements and parameters of any network formed by nodes and edges (graph theory) [12]. Figure 1 collects the main metrics associated with graph topology:

The **degree** (K) of a node defines how many edges (connections) a node has to other nodes. Likewise, the degree distribution (P(k)) shows how many nodes have a particular degree k (Figure 2).

The **clustering coefficient** (C) gives an idea of the interconnectivity in the neighbourhood of a node. In other words, the cohesiveness of the neighbourhood of a node. The average clustering coefficient (c) can be computed as well and recognizes the overall tendency of nodes to form clusters. Clustering, or network transitivity, is a fundamental property of any type of network.

The **path length** counts the number of edges that separates two particular nodes within the graph. Likewise, the **shortest path** between two nodes finds the minimum number of edges (distance) needed to connect those nodes. Another interesting parameter, the **betweenness**, reflects node's centrality and its "influence" in the network by measuring the number of shortest paths that go through the node.

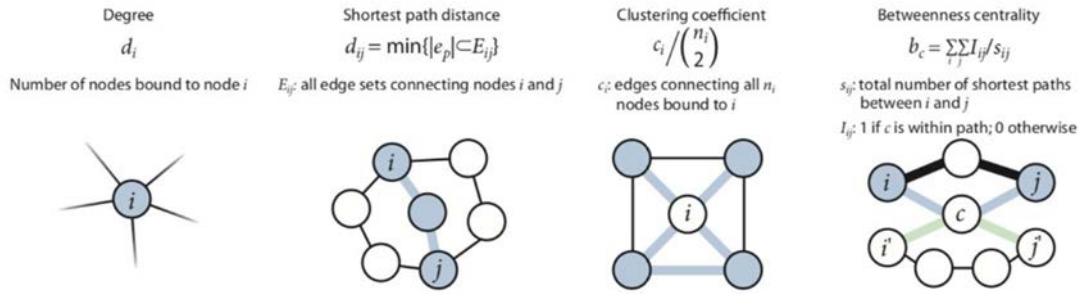


FIGURE 1. PICTORIAL GLOSSARY OF COMMON NETWORK METRICS

According to the degree distribution pattern, networks can be characterized in three different classes of graph models by means of a power-law mathematical correlation (Figure 2) [13]:

Random Networks: Their nodes share a random number of links. The degree distribution plot is a Gaussian curve, where nodes that have a significant deviation from the average are extremely rare.

Scale-free Networks: The majority of networks associated to cells and tissues adopt this structure. In them, there is a majority of nodes with very few connections, and a small number of nodes that have a high degree. The main graph properties will be determined by the special nodes that highly exceeds the average links, also known as hubs. The interactome –showcasing protein-protein interactions (PPIs)-, reactome and diseasome and many other complex graphs have a scale-free arrangement, where the feasibility to go from almost any node to any part of the graph in a minor number of steps confers them the so-called *small world* characteristic [14]. As its name indicates, they are growing networks marked by a preferential attachment where the hubs tend to get "richer" in terms of connectivity.

Hierarchical Network: These networks are a type of scale free networks, represented by a highly ordered structure, where the clusters combine in an iterative fashion, creating thus a ruled network. Consequently, their average clustering coefficient is markedly high. A few number of hubs are in charge of maintaining the conformation and connections between different clusters.

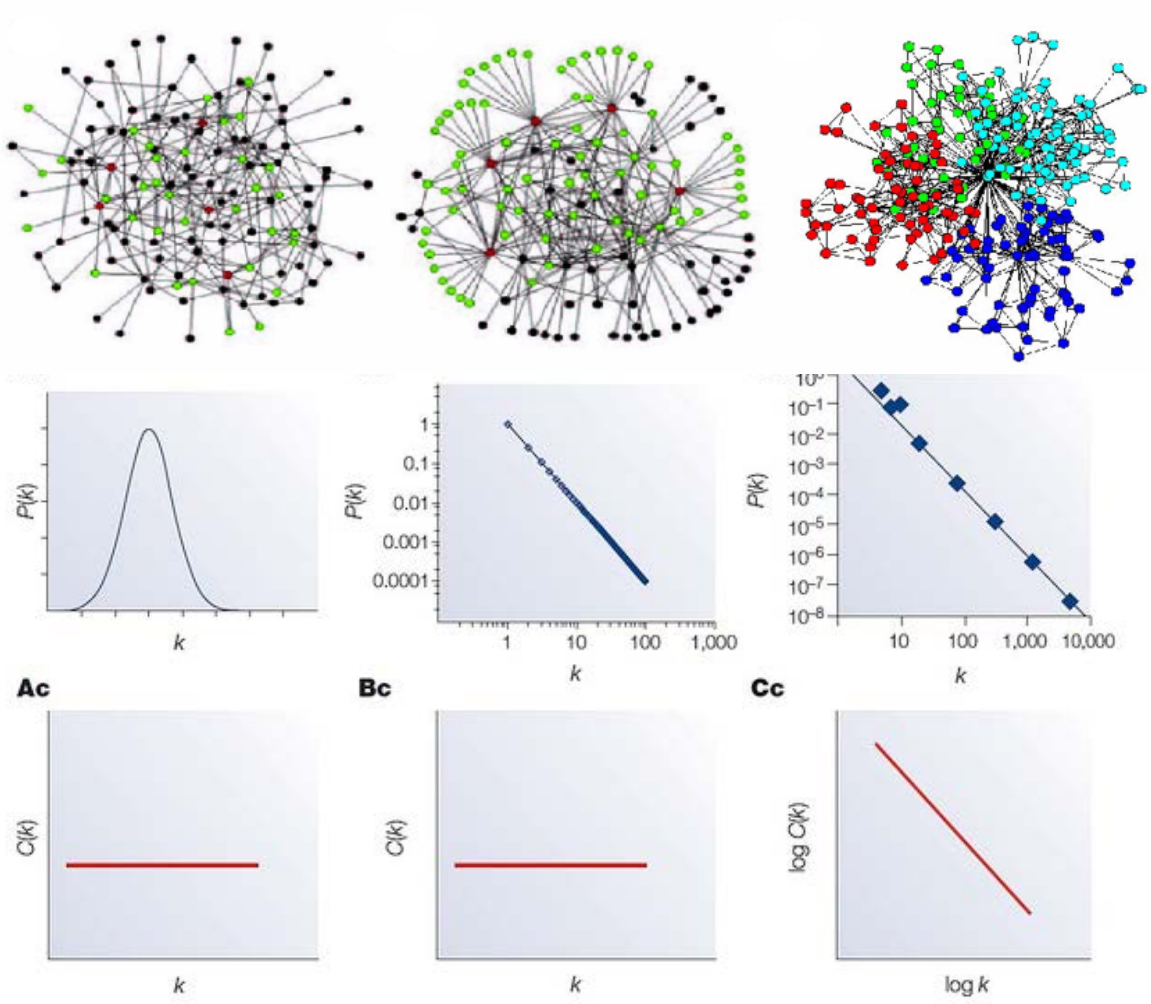


FIGURE 2. BARABÁSI-ALBERT MODEL MODIFIED FROM [13] & [15]

In order to analyse the data represented in networks, and eventually give a biomolecular interpretation of the phenomena under study, these graphs possess, in an inherent fashion, a topological structure and connectivity parameters. Upon computational analysis, the network showcases evidences of certain patterns and global values, which in turn enables the detection and identification of the distinct features of the cellular machinery, providing then solid indications underlying complex disease mechanisms [15].

c. Introduction to the diseases phenotypes

Genes are segments of DNA molecules located in each chromosome, both autosomal and sexual, and they are responsible for the inheritance patterns in living organisms. A gene is a functional unit whose expression controls the cell fate, and so the possible outcome of any functional tissue. Structure of genes can be seen in Figure 3. The term phenotype refers then to the interaction of the genotype –the complete heritable genome- with the environment that gives rise to the observable characteristics on each human being [16].

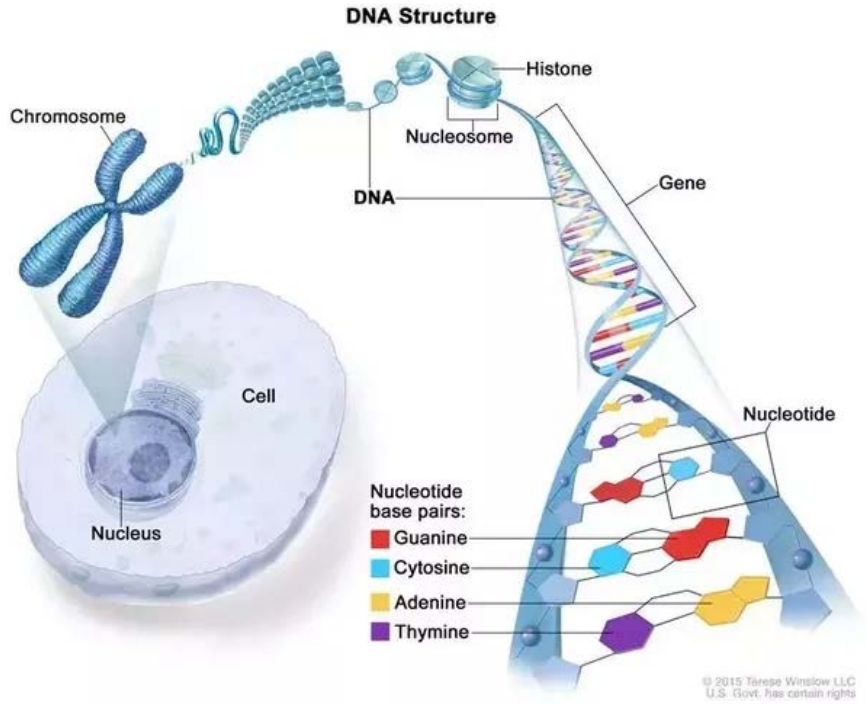


FIGURE 3. CHROMOSOME TO DNA. UPDATED FIGURE FROM [17]

Some gene alterations might occur during DNA replication or due to environmental factors, giving rise to mutations if they are located in sexual cells that are passed to the offspring [18].

If the mutation is carried only by one copy of the pair of chromosomes, it is called recessive. On the other hand, if it is present in both copies, it is a dominant mutation. The repercussions will vary in scope and gravity depending on the alleles (variant forms of a gene) and the function of that particular gene carrying the mutation. In addition, if both alleles of a gene are identical, the individual is called homozygous. Otherwise, the individual is heterozygous [17].

Genodermatoses are inherited genetic skin conditions. The three genodermatoses studied in this project are rare autosomal recessive diseases. Despite of their unalike genetic background [19], they have been grouped together because of the similarities they share at the phenotypic level (such as skin fragility, inflammation, cancer proneness), which will be studied in detail hereafter.

Two out of three are under the epidermolysis bullosa (EB) classification: EB is a rare genetic condition, which easily causes fragility and blisters along the skin in response to some minor injury or friction. According to its etiology, it is caused by mutations in genes expressing proteins related to the adhesion between dermis and epidermis [20].

The recessive dystrophic epidermolysis bullosa (RDEB) is one subtype of the four broad categories of EB. With a prevalence of 30.000 individuals in Europe [21], this fatal disorder is the most severe, even affecting the mucous membranes of the gastrointestinal (GI) tract and moist lining. After a blister heals, it leads to progressive scarring that might include digit fusion and even joint abnormalities. A mutation in the COL7A1 gene happens to be the cause of the two recessive and one dominant types of DEB. COL7A1 carries the instructions to assemble the collagen VII protein, the main constituent of anchoring fibrils, located at the dermal-epidermal basement membrane (Figure 4) [22].

Regarding the prognosis, life expectancy is significantly reduced due to the risk of squamous cell carcinoma (SCC) development. Its metastasis incurs an 87% of mortality by age 45 [23].

Kindler Syndrome (KS) is also a rare disease which belongs to the EB family, the most unusual one though, with only about 250 cases reported worldwide [24]. By the same token, skin blistering is its major clinical description, driving to fusion of fingers and toes. Moreover, it incurs in other skin abnormalities such as poikilodermia, characterized by pigmentation irregularities and small dilated blood vessels just under the skin. In addition, it might cause patients to be highly sensitive to ultraviolet (UV) light. It is caused by a mutation in the FERMT1 gene, which takes over the expression of kindlin-1 protein, a phosphoprotein that has a crucial role in the polarity, motility and proliferation of epidermal keratinocytes [25].

In the same fashion, KS also increases the risk of cutaneous squamous cell carcinoma (SCC) development.

Cell biological properties of kindlin-1 in basal keratinocytes are ubiquitous, since it is an essential player in the cutaneous epithelial stem cell homeostasis (Figure 4). Forming the backbone of the basement membrane, collagen VII also appears as an accountable component for signalling processes.

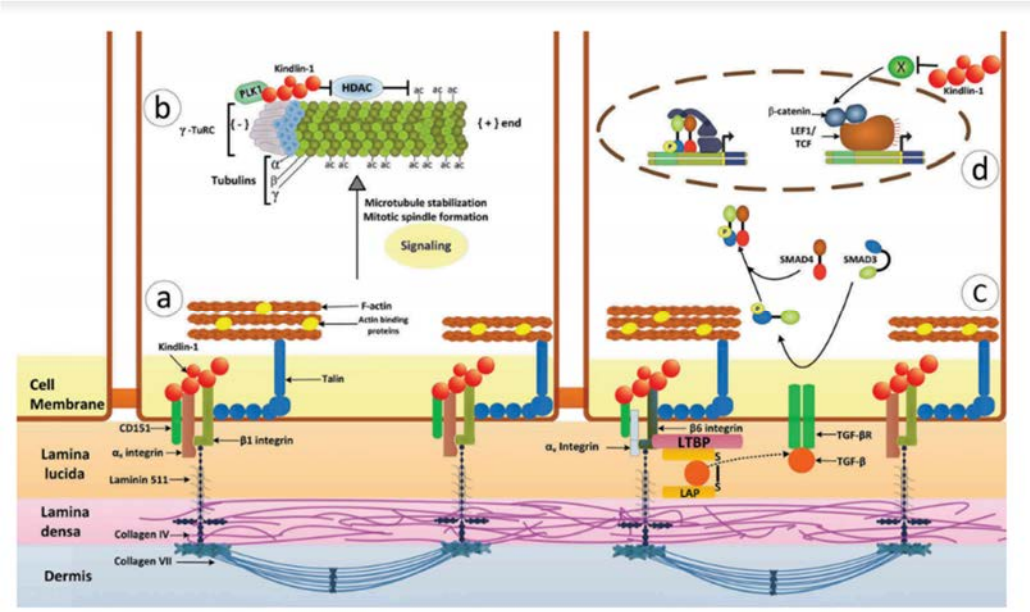


FIGURE 4. SCHEMATIC ILLUSTRATION OF THE FOUR MAIN ROLES OF KINDLIN-1. A) CELL ADHESION AND INTEGRIN SIGNALLING, B) MITOTIC SPINDLES AND CELL SURVIVAL, C) RELEASE OF TGFB AND D) SUPRESSION OF WNT SIGNALLING. SOURCE [26]

Xeroderma pigmentosum complementation group C (XPC). Aside from the EB subtypes, the other inherited skin condition studied along this project is XPC. With a frequency of 1 in 1 million in the United States and Europe, XPC is caused by a defect in the XPC gene, which forms part of the DNA repair mechanism. In the skin, its phenotype is characterized by severe sunburns present upon individual exposure to sunlight. In other words, patients with XPC have an extreme sensitivity to UV rays due to their incapacity to correct mutations [27]. Sunburns then transform into freckling in affected young children. Dry skin and changes in skin pigmentation ultimately make the name of xeroderma pigmentosum.

It is a quite heterogeneous disease, taking into account that at least eight different forms of XP have been found [28]. The mutated genes are involved in the nucleotide excision repair (NER) pathway [29] (Figure 5). As a result, DNA is not repaired properly and the accumulated abnormalities eventually make the XPC patients prone to malignant melanoma.

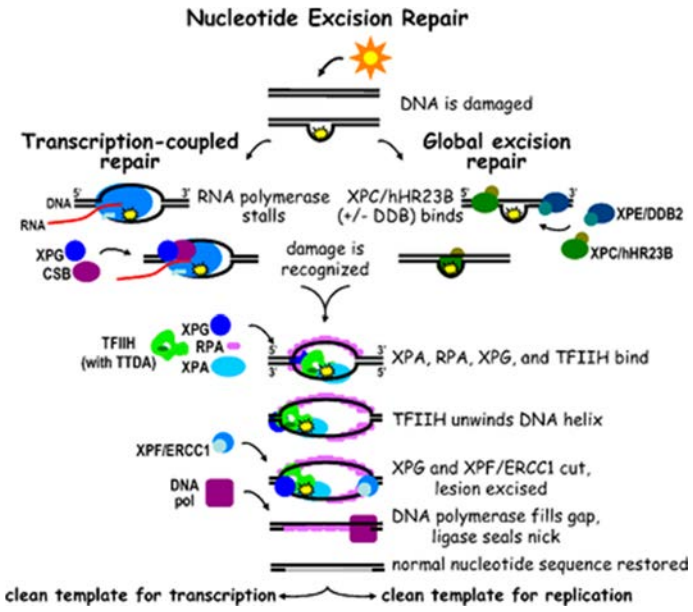


FIGURE 5. NER PATHWAY DIAGRAM. UPDATED FIGURE FROM [29]

Furthermore, these three genodermatoses lead to chronic irritation and inflammation, which is indeed a critical component in tumor progression. Inflammatory cells orchestrate the neoplastic environment, where the signalling molecules of the innate immune response combine with receptors for survival, proliferation and migration during metastasis [30]. In short, they are all cancer-prone. Besides, all these three genodermatoses imply vision impairment [23, 24, 27].

In spite of their matching related phenotype, the mutated genes of each disease have different chromosomal locations and a very low nucleotide base sequence similarity according to the National Center for Biotechnology Information (NCBI). Simply put, they seem to have a poor correlation at the genomic level.

Therefore, in order to obtain molecular mechanisms that can explain the similarities in the phenotype, other levels of data apart from genomics, must be analysed and integrated.

d. Integration of omics data

In order to build a bridge between System Biology and network medicine and, in turn, infer causal associations between gene expression and diseases, mathematical and relational models need to be introduced [31]. In this manner, the collected and shared data from bioinformatics repositories ought to be exploited after its discovering. Consequently, due to the massive amount of heterogeneous high-throughput data that NGS technologies can make available into these databases, computational systems biology ought to improve omics data integration [32].

Omics integration promises to be capable of disclosing hidden biological knowledge, addressing thus biomedical problems. Essentially, it combines multiple data types to compensate unreliable information in any independent data set and to obtain a further system biology insight into the functioning of cells as a whole.

There are two main sources for omics data [33]: the so-called knowledge databases, namely ontologies, which collect a detailed framework of the majority of mechanisms and functions of each individual component in a species, and the clinical data, that corresponds to experimental designs and patient sequencing data. The latter tends to have annotated expression values for every single sequenced entry.

To fully characterize the cellular mechanisms and thus the organ functioning (or malfunctioning), researchers need a methodology capable of increasing the value of both the previously deposited knowledge and the continuous rise of clinical annotated data. The major challenge that the omics technologies confront are the redundancy-free information –no futile overlaps among shared data- and the efficiency at ruling out the false positives [34].

Talking about omics integration means talking about the different projects that have previously appeared willing to investigate the complexity of the biological systems. The most relevant paradigms are the 1000 Genomes Project [35,36], Encyclopaedia of DNA Elements Project [37], Cancer Genome Atlas Project [38] and the Immunological Genome Project [39], among others. During the course of these ventures, novel methodologies to analyse and integrate data have evolved, providing a wide variety of algorithms for concrete purposes: alignment of markers with molecular activities, self-organizing maps, inference of regulatory networks, etc.

From all the integration approaches carried out, the most prominent algorithm established, able to work with multiple variables in a computationally feasible manner, is the meta-dimensional analysis [40]. It consists of a simultaneous integration of non-linear interactions with high-throughput data to get a global picture of a trait. The multivariable prediction model generated comes with a particular outcome. This method avoids the common limitations that can present the rest of the approaches, such as the multi-staged analysis, which assumes that genetic variations are hierarchical. Multi-staged might be very useful by the time of associating in a linear manner a change in gene expression with a change in the phenotype (Figure 6). However, it has been stated that disease patterns are non-linear and quite interactive in reality, having a more complex nature that cannot be fully fitted in a stepwise, linear model such as the multi-staged analysis. In our work, an approach to omics integration using biological networks will be committed, since graph constructions enable to undergo meta-dimensional analysis of omics data.

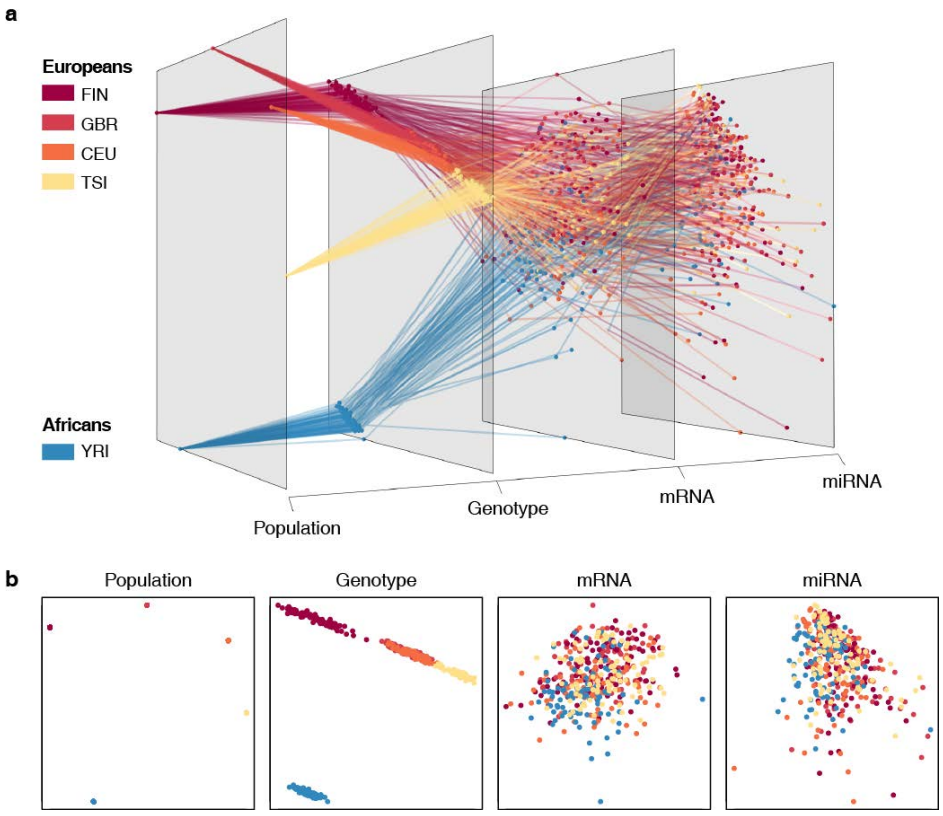


FIGURE 6.THE GEUVADIS PROJECT. COMBINATION OF MRNA AND MICRORNA SEQUENCED DATA TO CHARACTERIZE REGULATORY VARIATION IN HUMAN POPULATIONS. SOURCE [41]

All the massive amount of biological data is then grouped on different levels of regulation, resultant from each particular omic-technology employed to target the distinct molecular entities. The accepted classification of these omics levels is the following (Figure 7):

Genomics

Genomics is a catch-all term that collects studies of genes and genomes at the DNA level (variants, copy number, regulatory sequences, etc). It is the broadest and most mature of the omics disciplines. The retrieved genomic data has made sense to several gene-regulatory functions and procured the biogenesis of the species whose genomes have been completely sequenced today [42]. Therefore, genome-wide studies have provided evidences and new discoveries regarding the evolutionary tree of life.

Transcriptomics

Transcriptomics field yields information about RNA transcripts abundance. Since the transcripts level determines the gene expression state within the cells, measuring the RNA amount accounts for a snapshot about the active and dormant cellular processes that are taking place in a cell. Contrary to genomics, it has to deal with quantitative results. First attempts are dated to the 1990s and they have reached a point –as Microarray and RNA-Seq technologies flourished- where the volume of RNA collected can catch up with the recorded DNA entries [43]. Unlike Microarray technology, RNA-Seq is not dependent on any prior genomic knowledge: it examines the quantity and sequences of RNA in a sample by means of NGS (for this to occur, RNA

fragments are converted into a cDNA library). Essentially, RNA-Seq uses shorts reads of messenger RNA (mRNA) where intronic content has been discarded, and then aligned back to a reference genome or assembled, providing eventually a comprehensive map view of the whole transcriptome [44]. In this way, transcriptomics assists on capturing gene dysregulation patterns. It is true that alterations can occur out of the traditional genetic basis for inheritance, that is, without mutating the DNA codifying sequence (epigenetic). In this way, post transcriptional modifications such as mRNA silencing, orchestrated by microRNAs, will lay on this subcategory called epigenomics [45].

Proteomics

This discipline quantifies the protein amounts within cells at a certain time point. Coined in 1997, the proteome information explains the functions encoded by any gene, throughout the analysis of the overall composition, structure and activity of the proteins. On the basis of their mass and charge, proteins are characterized by using mass spectrometry strategies [46]. Proteomics data can appear in biological networks in the form of physical connections (at the molecular level) among proteins, which are called protein-protein interactions (PPIs).

Metabolomics

This unique biochemical approach seeks to identify the set of metabolites –any small substance involved in metabolism- found within tissues, biofluids and microbiome. It is closely related to the phenotype and apparently there exists a crosstalk between epigenetics and metabolomics [47]. For that reason, it is an emerging tool able to elicit powerful pathogenesis information. Mass spectrometry offers as well a sensitive technique to capture metabolomics data.

Considering this set of layers where the omics data can be batched, either vertical or horizontal integrative analysis can be performed [48]. In terms of suitability, each one can offer some benefits depending on the ultimate purpose. However, these different layers are not so descriptive on their own and molecular mechanisms cannot be fully explained with the information from just one. Their complementary interplay hence becomes more meaningful and efficient over the bioinformatics advancement. What is more, the published work has traditionally been focused on genomics and transcriptomics [49], that is, DNA and RNA sequenced molecules, lacking thus substantial meta-data required for the cross-analysis and integration along the rest of the layers. Besides, dimensional overfitting –the excessive use of the experimental space and the variables allowed in a specific study- is impeding as well the proper attainment of omics compendia [50].

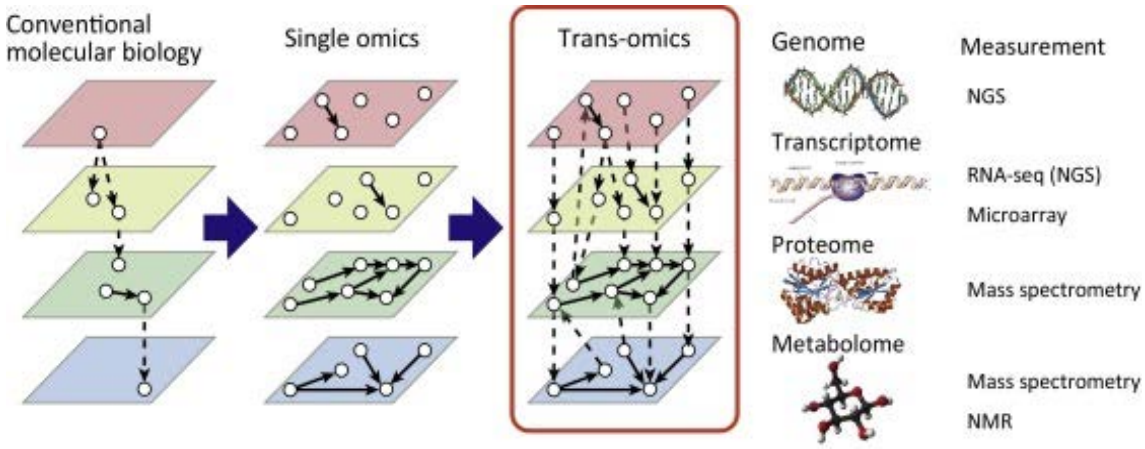


FIGURE 7. TRANS-OMICS APPROACH. SOURCE [51]

The omics-driven predictive modelling allows to view as a whole, through integration of the biological data, the complex crosstalk among all the molecular entities that are responsible of pathways, pathologies and diseases.

On top of that, the integration over multiple omics layers seems to be a reasonable approach in order to develop hypothesis that lead to the determination of molecular causes of disease. Nevertheless, as long as the peculiarities of any given experimental setting and the deprivation of standards are not tackled, data quality control is a major concern though [52].

e. A regulatory overview

As aforementioned, transcripts are studied by the transcriptomic field. Their regulation is a critical step in the central dogma process (DNA→RNA→Protein), since it determines the gene expression into proteins. On this basis, there exists a set of small molecules consisting of ~22 nucleotides whose performance can disrupt the translation process. These non-coding and single-chain molecules are named microRNA (miRNA), and belong to an RNA interference pathway which neutralizes mRNA molecules [53], controlling the activity of genes and inhibiting thus gene expression in cells.

These “knockdown” events can be used in loss-of-function studies, so microRNAs are considered relevant players in the human interactome. In this way, microRNA & non-coding RNA analysis is a valid approach to understand the protein role and abundance in a particular condition.

The mechanism of action consists in a hybridization where the base pair complementation between the microRNA and the target mRNA leads to the degradation of the latter [54] (Figure 8). However, the complete knowledge regarding its functioning and involvement in physiological and pathological processes is still scarce.

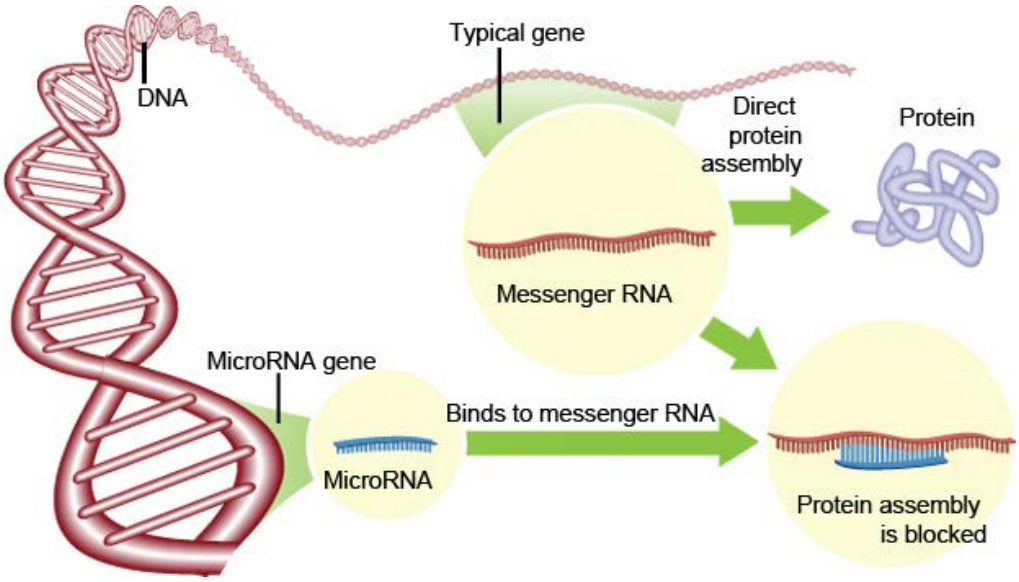


FIGURE 8. POST-TRANSCRIPTIONAL GENE SILENCING UNDER MICRORNA ACTION. SOURCE [55]

2. HYPOTHESES AND GOALS

Recessive Dystrophic Epidermolysis Bullosa (RDEB), Kindler Syndrome (KS) and Xeroderma Pigmentosum type C (XPC) are three phenotypically related genodermatoses. Even though these three inherited skin disorders hold a different genomic origin (mutations in COL7A1, FERMT1 and XPC respectively), some of their intra and extracellular life mechanisms could be somehow shared, which would explain eventually their common phenotypic traits.

By using omics technologies (RNA-Sequencing, microRNA Sequencing and Protein-protein interactions), both transcriptomics and epigenomics data from the three genodermatoses can be obtained and analysed. The hypothesis underpinned herein is that the construction and interpretation of biological networks gives rise to a traceable framework of interactions among entities that might serve as a convenient approach for the study of the aforementioned common hallmarks (cell-cell adhesion, inflammation processes and cancer proneness).

This blueprint proposes a novel procedure for narrowing down high-throughput clinical data to end up with candidate genes and their regulatory microRNAs which might govern the RDEB/KS/XPC transcriptional profile.

The aim of acquiring a global transcriptional profile for the three genodermatoses is endorsed by the idea that a common-injury event causes certain epigenetic changes which in turn lead to a stable cancer associated fibroblasts (CAF)-like phenotype [56]. This phenotype would be shared by the three diseases, regardless of their different genetic origin. As expected, in this paper, the omics profiling revealed a high resemblance among the genodermatoses in contrast to healthy individuals. Put another way, under this similar transcriptional signature, it can be speculated that a triggering event makes fibroblasts detect adverse cues and forces them to self-activate and start secreting aberrant extra-cellular matrix (ECM) molecules, driving into a CAF-like phenotype. Understanding the causal genes that are in charge of the convergent themes among genodermatoses is what gives a meaning to this transcriptome and epigenome sequencing.

Upon the construction of the biological networks, and performing system-level observations (both topological and functional), the initial RNA entries will be short-listed, leading to candidate biomarkers, whose regulation by microRNAs would hopefully serve as a working basis for prospective pharmacological research.

The **main goal** of this project is to develop hypothesis that can explain, through an integrative network approach, the mechanisms that govern the expression of genes and that regulate their cellular level (through microRNAs) in three different, but related, genodermatoses.

To accomplish this, some other goals are approached, such as the validation of the previous results from RNA Sequencing, the topological analysis of integrated networks and the biological interpretation of the results.

3. MATERIALS & METHODS

a. Data Origin & Statistical Analysis

The transcriptomic and epigenomic data herein utilized come from a study carried out by Chacón-Solano et al. [56], where skin biopsies of four different patient cohorts (a total of 9 RDEB, 3 KS, 3 XPC patients and 3 healthy individuals) were obtained following the World Medical Association Declaration of Helsinki Principles [57]. As a first step, patients were screened for the mutation responsible of the disease. Both keratinocytes and fibroblasts from the skin biopsies were isolated, cultured and grown until confluency was reached. Next, RNA from fibroblast samples was extracted and submitted to RNA-Sequencing analysis in order to obtain short sequence reads that were posteriorly aligned and assembled together. This RNA sequencing will determine which genes and microRNAs are induced/repressed in each of the aforementioned pathologies compared with the healthy volunteers. To discard any sort of bias, artifacts or batch effects in the linear model, a principal component analysis (PCA) was carried out: PCA confers to all the components independence of one another. The normalized RNA-Seq counts yielded 22970 transcripts, and sample processing ensured the quality control of the fastq sequences using FastQC [58]. Each of these transcripts was subjected to statistical analysis, comparing disease vs healthy samples, where different quantitative annotations were obtained:

LogFC: It measures the logarithm of fold change that exists in a specific variable for two distinct samples. In the Chacón-Solano et al. [56] study, it is referred as the gene expression of the patients reads when compared to the healthy individuals. In this way, a positive log fold change value will mean an overexpression of that specific gene when compared to the controls, and underexpression otherwise.

P-value: The probability, under the null hypothesis, that the differential expression between diseased and healthy entries is statistically significant. By scientific convention, the decision about the significance of a result is cut off at 0.05.

FDR: When dealing with large sequenced data, it is important to reduce the likelihood of false positives that can be misread. The false discovery rate stands as a corrected p-value when conducting multiple comparisons.

This whole RNA-Seq approach was also employed to sequence and render the microRNA reads of each sample.

Finally, data group comparison upon Venn's diagram visualization (Venny 2.1.0) [59] gave a result of 227 genes and 18 microRNAs that were commonly dysregulated in the three genodermatoses.

Undertaking network construction and analyses is the selected approach to downstream study these omics data in the present project.

A schematic overview of the data origin and previous statistical analysis can be seen in Figure 9. Now that all the entries are properly annotated, they have got a statistical meaning, representing hence the take-off point for the subsequent testing and plan of action.

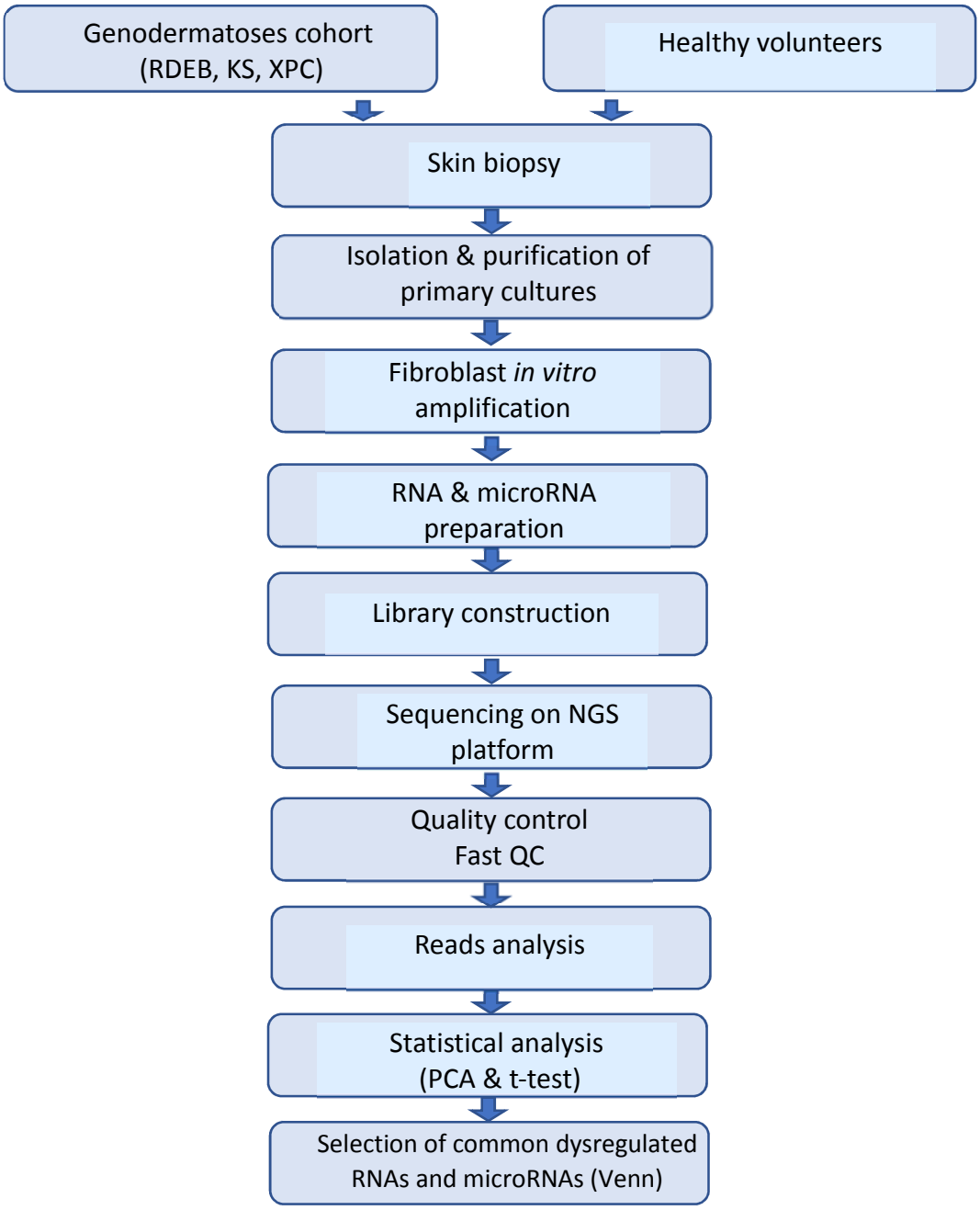


FIGURE 9. DATA ORIGIN THROUGH RNA-SEQ

b. Data Validation

Clinical data requires a reliable validation to ensure findings are not just random findings. Trustworthiness and veracity of microRNA sequencing data was assessed in this case. Specially in omics data, where the massive amount obtained is hampered by the appearance of possible false positives and negatives, due to artifacts for accuracy and precision of the technologies employed.

Among the 18 common microRNAs, several candidates were selected for biological validation, based on a topological analysis of microRNA networks from a previous study [60] and also their biological interest in the diseases. In order to study the biological relevance of these selected microRNAs, a bioinformatics analysis was carried out, with the aid of mirPath v.3 [61], a web-server employing a DIANA TOOLS algorithm. By means of this method, microRNA targets can be either predicted (microT-CDS) or empirically matched (TarBase). Following a leave-one-out cross-validation strategy, that is, disabling one microRNA at a time to evaluate how the rest of entries interact in their own right, and entering four microRNA every time, it is possible to subtly track their functioning in different KEGG pathways [62]. KEGG pathways are understood as a convenient compilation of databases which gather ubiquitous biological interactions and reactions in the form of diagrams. From this analysis, the most interesting (commonly dysregulated) miRNAs from a biological point of view were selected for qPCR validation.

Besides, another identified microRNA has proved to be a relevant signature in what RDEB concerns [63]. Although hsa-miR-29 was not among the common dysregulated ones (it was only statistically significant in RDEB and KS vs healthy comparisons), it apparently mediates the aberrant ECM synthesis through TGF-β1 induction, which at the same time is associated with inflammatory processes [63] and was therefore included in the study.

Proceeding with these microRNAs functional characterization, it was eventually decided to validate hsa-miR-10a-5p, hsa-miR-10a-3p, hsa-miR-29c-5p, hsa-miR-29c-3p, hsa-miR-129-5p and hsa-miR-195-5p, using RT-qPCR.

MicroRNA validation by RT-qPCR was carried out at CIEMAT, where fibroblasts samples from the RDEB/KS/XPC patients of the previous studies were grown at the cell culture lab. Isolated fibroblasts were subjected then to lysis processes by which all the microRNA content was released. The same procedure was applied to different healthy samples (different from the ones used in the microRNA-Seq) to further sustain the validation process. The extraction kit employed was mirVana^{TM} miRNA Kit (Applied Biosystems/ Ambion, USA), and the concentration of microRNA extracted was evaluated by Nanodrop 1000 (Agilent Technologies, USA) as a first instance, and then converted into cDNA using a reverse transcription polymerase chain reaction machine for their downstream quantification. RNU6 housekeeping microRNAs primers was used as control. Afterwards, quantitative PCR was undergone using Taq-Man technology and kit (Thermo Fisher thermocycler). Results were obtained in the form of CT, or minimal cycles, so they would need eventually a conversion (with algebraic formulas) in the form of logFC, which will indeed tell us the differential expression of the microRNAs of interest. Final statistical analysis (PRISM 6.0) was configured with Mann-Whitney (non-parametric) test.

c. Transcriptomics and epigenomics integration. Cytoscape

The maxim of biomedicine network analysis is to 1) begin with a reliable initial list of biological omics data, 2) display in a network their interactions across every molecular level and 3) gain biological insights from the integration analyses of the network.

In our case, after the screening, transcriptomics and epigenomics data was firstly obtained and validated. From that point, an integration analysis was conducted using Cytoscape [64]: a user-friendly, open source platform that is considered today as de-facto standard software in the bioinformatics state-of-the-art. Its range of possibilities is outstanding, with an architecture that

is continuously being upgraded by the scientific community. The core structure, designed at the Institute of Systems Biology in Seattle, is fairly compatible and interoperates with other existing bioinformatics tools. Cytoscape orchestrates every single step involved in the understanding of biomolecular interaction networks (Figure 10):

-Integration of biological networks with gene expression profiles and its management.

-Navigation and customization of network data displays.

-Import and export of all the constructed graphs and annotated tables.

-Work alongside biological literature found at biobanks and ontologies.

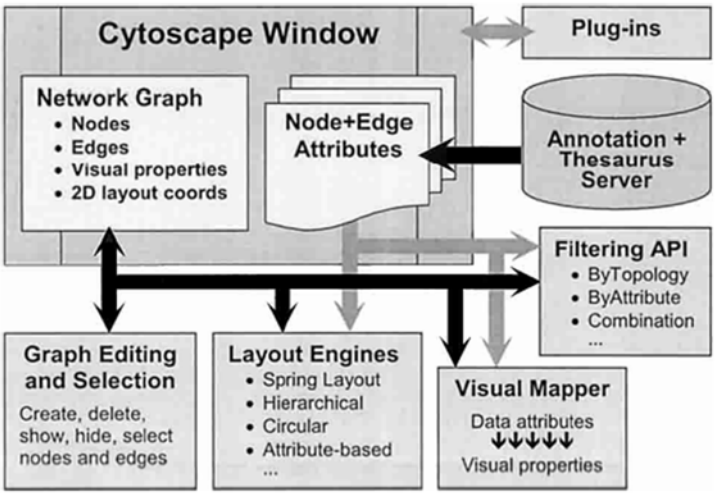


FIGURE 10. SCHEMATIC VIEW OF CYTOSCAPE CORE ARCHITECTURE. SOURCE [63]

Cytoscape plug-ins obtain the biological information from a number of databases or ontologies. The most common one is the Gene Ontology (GO), which classifies all genes and gene functions in different hierarchical terms (GO numbers) that define specific functions in three categories (Biological Processes, Molecular Functions and Cellular Components). Other databases commonly used for enrichment analysis are KEGG and Reactome. Furthermore, it is so multitasking that the plug-ins allow for the three steps for network analysis followed in this study:

1) Network generation, in order to generate the integrated omic data network, Cytoscape makes use of certain plug-ins that can be subdivided in:

Layout and visual properties manipulation. Cytoscape makes use of a wide array of instruments and options to tailor any biological network layout. These changes of graph appearance allow for a clearer visualization of the pursued molecular entities within the network without altering the natural graph structure, that is, nodes and edges distribution.

Importation of pairwise connections among classes of entities. Apps like STRING [65], which apply quality-controlled interactome data to incorporate, in our case, protein-protein interactions (PPIs) along already created graphs. In this way, data from different omics levels can be associated on the same network backbone, which entails a crucial step in Integrative Bioinformatics.

2) Topological analysis of molecular elements. In order to analyse the structure of the constructed graphs, a global topological analysis was used, which in turn draws the most relevant parameters and measures of each network component. Topology certainly highlights graph elements, such as nodes and their connections. However, these entities are not necessarily biomarkers for our diseases, they might be just related to the graph conformation, so their essentiality requires to be downstream studied.

Cluster and Motif Discoveries. Depending on how nodes and edges are arranged/distributed within the graph, different Apps can support the deep examination of collective structures of interest within large-scale networks [66]

3) Functional enrichment analyses. Since the ultimate goal has always been to obtain phenotype-genotype information regarding pathologies, a powerful tool is required to dissect the biological meaning out of the network constructs. To help on this pursuit, the functional enrichment stands as an exploratory procedure which uses statistical approaches to identify molecular mechanisms that are over-represented within a large set of genes/proteins (in our case the common differentially expressed genes). These relevant mechanisms may have an association with disease phenotypes. The method employed herein is the singular functional enrichment (SEA), which iteratively test the enrichment of each annotated gene by making use of an enrichment p-value. This enrichment probability tells the number of genes in a list that hit a given biological function as compared to random chance [67]. These genes are considered determinant for the specific disease so they can be figured as candidates in terms of the biological functions they are involved in. Two interfaces are going to assist within Cytoscape environment to undertake the functional enrichment analysis.

STRING. In addition to its employment in the detection of protein-protein interactions, its Cytoscape plugin can retrieve as well functional enrichments mainly from Gene Ontology (GO) [68] and KEGG Pathways terms after setting a confidence cutoff. GO belongs to the so-called online knowledge databases [33] whose information is subdivided in three categories: Biological Processes (BP), Molecular Function (MF) and Cellular Component (CC), all of them organized in a nested hierarchical fashion where putative functions are directly related to hierarchically successive entities. Yet, this approach is incomplete as not all the existing human proteins are annotated in the three GO ontologies. Figure 11 presents an extent of annotation of proteins in model species.

BINGO. By the same token, BINGO provides an exhaustive perusal of the GO categories that are statistically overrepresented in a set of genes present in a graph. By configuring a multi-parameter kit, the user can obtain the enriched GO annotations and propagate them upwards through the GO hierarchy. By these means, every gene annotated to a certain GO category will be fitted at once in all the corresponding parental categories [69].

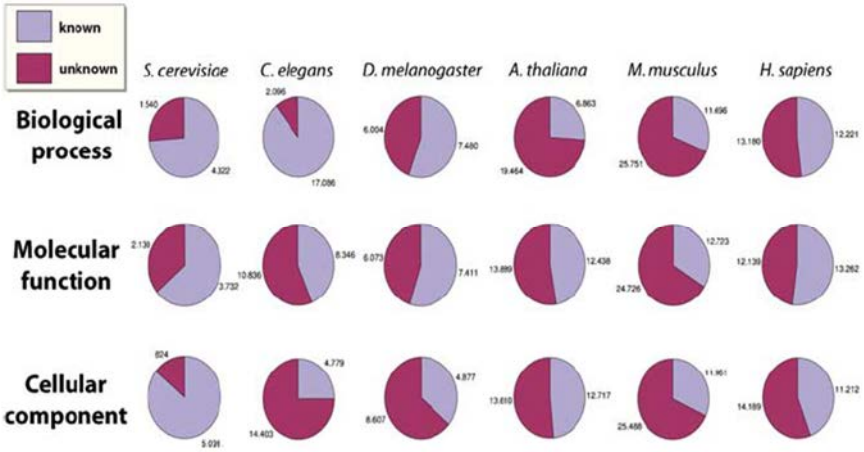


FIGURE 11. EXTENT OF ANNOTATION OF PROTEINS IN MODEL SPECIES. PIE CHARTS SHOWCASE YET A SUBSTANTIAL FRACTION OF UNANNOTATED PROTEINS, SINCE NOT ALL THE DATABASES IN HUMANS ARE COMPLETE. SOURCE [70]

Curation pipeline. Due to the non-directed nature of the SEAs, overly a huge number of pathways and functions tend to be retrieved in the analyses. Some of them are unrelated or redundant, hence certain strategies are employed to refine the obtained results.

As a method to meet our data with biobanks and ontologies, two different approaches were followed:

- ReVIGO. As its acronym indicates, it serves for reducing and visualizing Gene Ontology terms by semantic similarity. For this study, it was fine-tuned using a stringent clearance (0.7) and a semantic similarity measure=SimRel. Functional enrichment results will be grouped by ReVIGO [71].
- ClueGO. Genes of interest from enriched functions will be downstream studied in here as candidates for explaining the phenotypic correlations among the three genodermatoses. ClueGO comprises one of the most robust Cytoscape plug-ins, where functionally organized term networks are elaborated using GO, Reactome and KEGG assistance [72]. Since six microRNAs has been experimentally validated at the lab, they will be used as a way to reinforce this venture as well.

A detailed flowchart of the tools used, the workflow and the data interpretation are shown in Figure 12. In short, microRNA-omics data was imported into miRNet [73], generating a microRNA-RNA interactions table. After nomenclature conversion (miRCarta [74]), interactions were loaded into Cytoscape for networks construction, adding the protein-protein interactions (PPIs) within the network. Topological and enrichment analyses were then carried out, and a biological interpretation is subsequently performed.

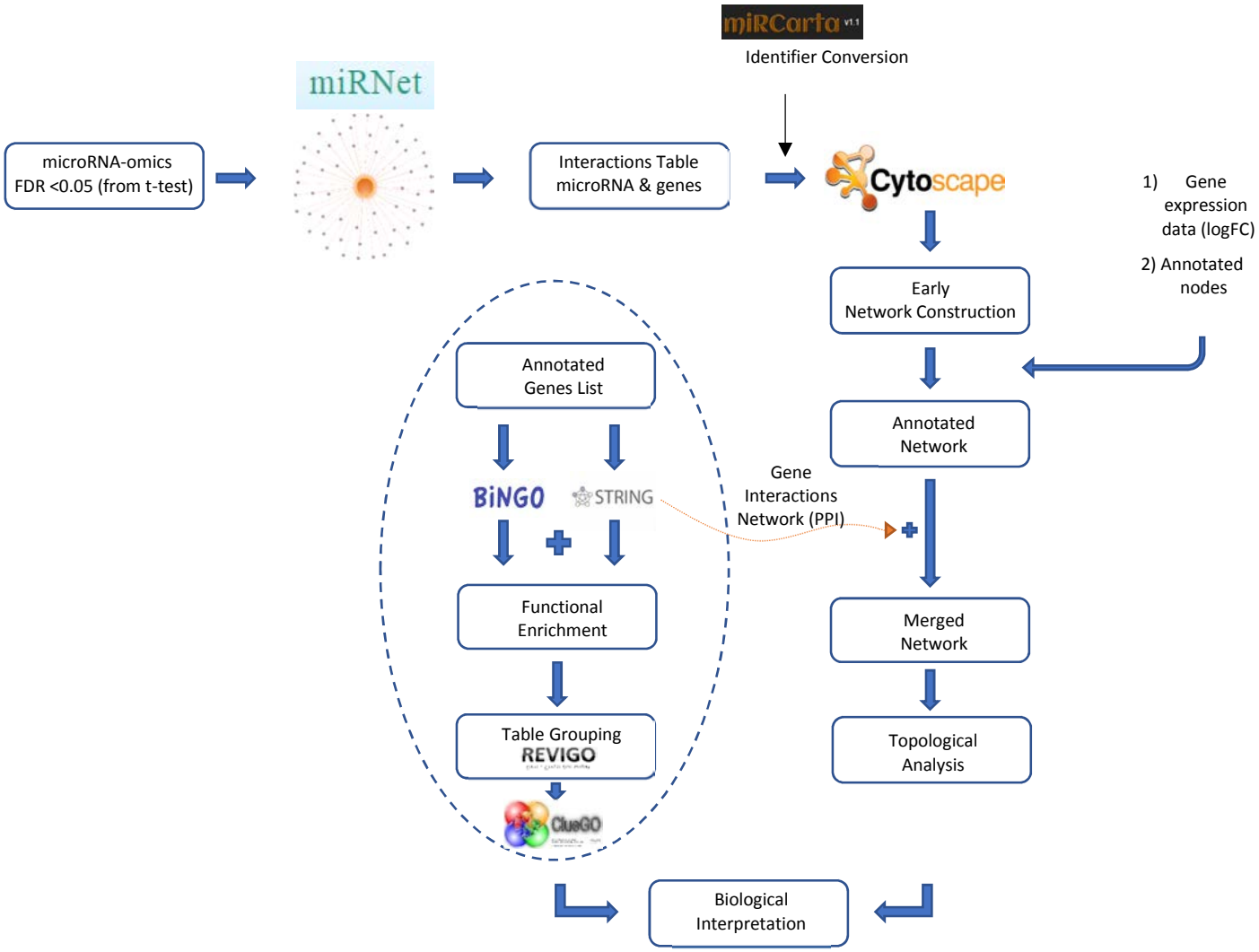


FIGURE 12. DETAILED INTEGRATION FLOWCHART

4. RESULTS

a. Validation of microRNAs

The first stage consisted in selecting and validating a small set of dysregulated microRNAs by RT-qPCR from a pool of 18 commonly dysregulated among the three genodermatoses, which was mainly performed assisted by miRPath [61]. MiR-10a and miR-10b were excluded of the *in silico* mirPath algorithm because miR-10 has already been confirmed in a wide array of species and its precursor family is highlighted prominently in the literature due to their association with numerous cancers [75] and notable protein synthesis, so there was no doubt its validation was mandatory. Same happened, as mentioned before, with miR-29c [63].

In this way, after running the online software, microRNAs with few hits and biological impact were discarded for validation. MiRPath interface view can be observed in Figure 13, where microRNAs were subjected to leave-one-out strategy. Heatmaps show the arbitrary log (p-value) for each analyzed microRNAs across different functions and pathways.

The first leave-one-out analysis (Heatmap 1) stated that, either leaving the miR-195-5p or the miR-129-5p out shortened the pathways list and impact, indicating thus their prominent involvement in the mechanisms of interest. Moreover, the miRPath web-server pointed that these particular microRNAs are involved (as the top-ranking list indicates) in the “Adherens junctions”, “TGF-β signaling” or “Melanoma” functions, among others. This is strengthened by the fact that they have proved to empirically regulate 1422 and 545 genes, respectively.

The second analysis (Heatmap 2) followed the same procedures, this time with the 3p versions of miR-195 and miR-129. In spite of the highly significant correlation of miR-129-2-3p to “ECM-receptor interactions”, its presence in other pathways of interest was slight.

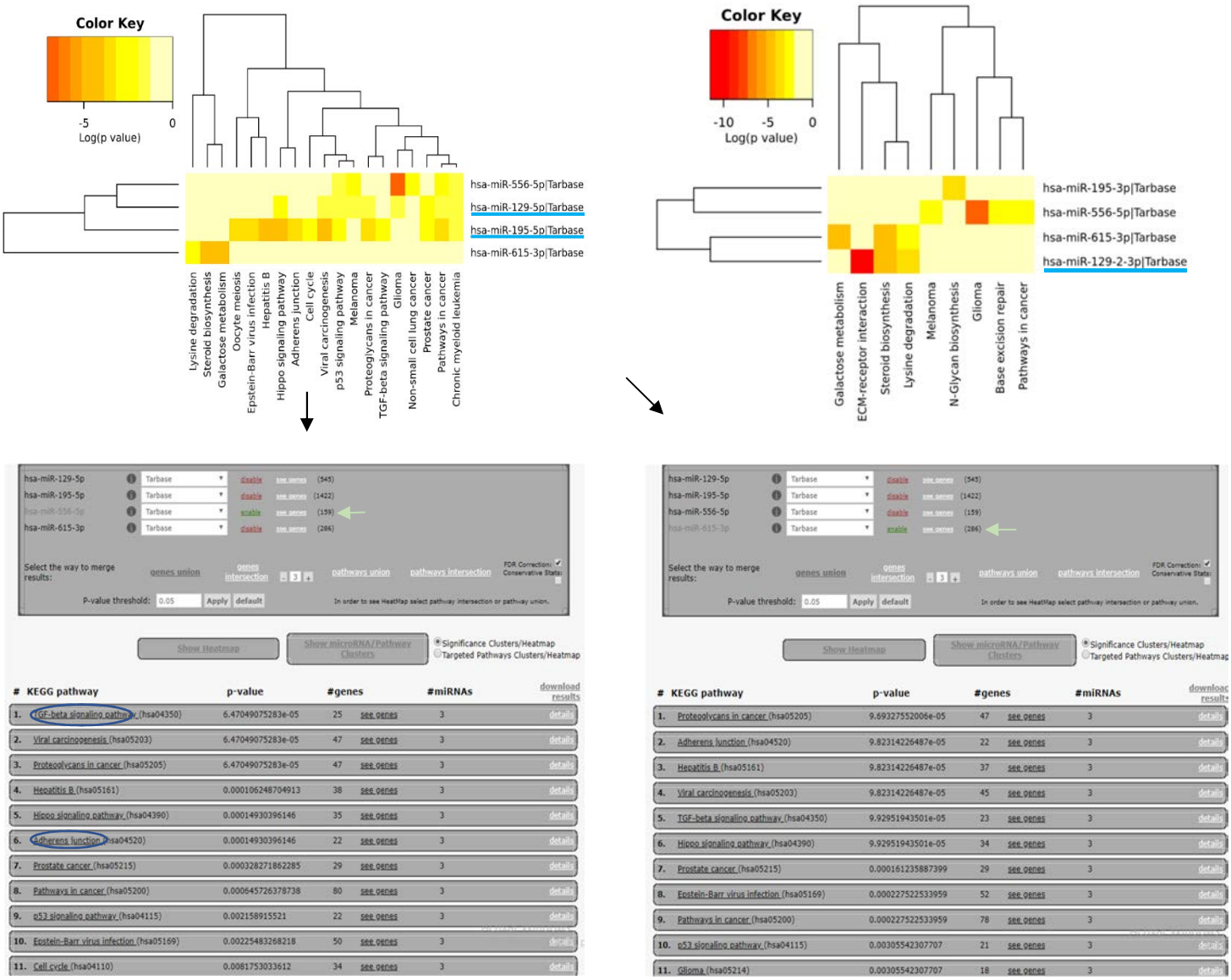


FIGURE 13. HEATMAPS AND TABLE ANALYSIS OUTPUTS FROM THE MIRPATH INTERFACE

After studying the impact of all 18 microRNAs, miR-195-5p, miR-10a-3p, miR-10a-5p, miR-29c-5p, miR-29c-3p and miR-129-5p were finally validated by RT-qPCR at CIEMAT, following the protocol described in Materials & Methods. The final statistical analysis displayed the results shown in Figure 14. Red arrows represent the expression status of each microRNA obtained in the previous RNA-Seq analysis (namely, upregulation or downregulation).

All the qPCR analyses (Figure 14) rendered a resembling tendency on microRNAs expression when compared to results previously obtained in RNA-Seq (with the only exception of miR-195-5p in Kindler Syndrome), thus validating these microRNAs for the subsequent data analysis.

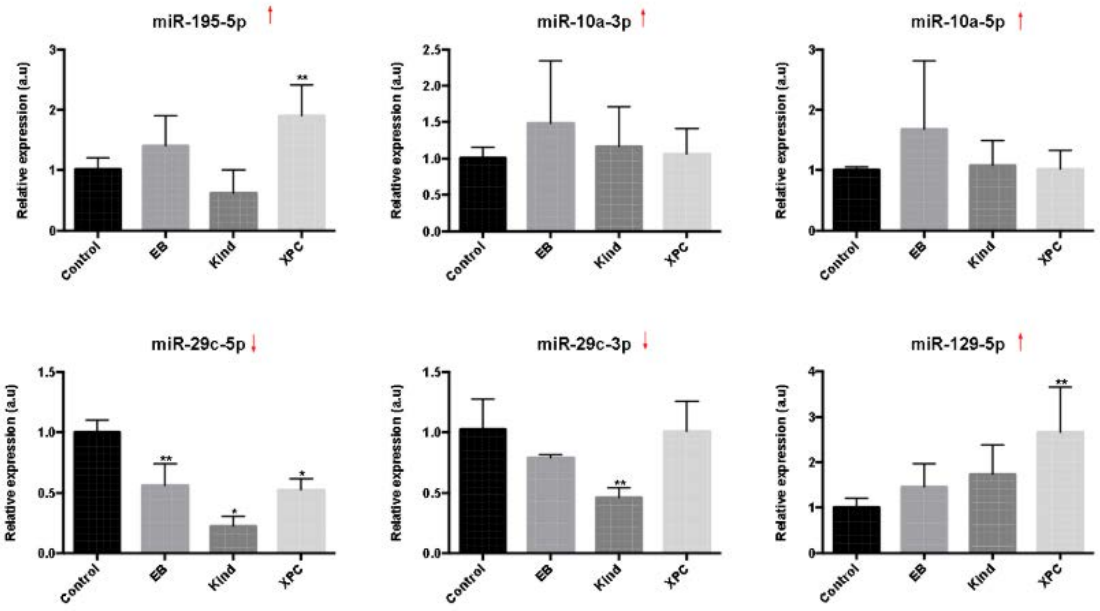


FIGURE 14. COLUMN CHARTS WITH ERROR BARS FOR THE SIX VALIDATED MICRORNAS. "*" MEANS P-VALUE < 0.05 AND "" MEANS P-VALUE<0.01 (N=3). ARROW IN RED SHOWS THE EXPRESSION TREND OBTAINED FROM RNA-SEQ ANALYSIS**

b. Network construction

Once validated the RNA-Seq and microRNA-Seq data, the next stage consisted in constructing the regulatory networks for each disease. In the following lines, the term "genes" is going to be used for convenience when referring to mRNAs or transcripts.

To build each specific network, links between the differentially expressed microRNAs from the RNA-Seq study and all their possible targets were established. This was done using miRNet [73], an online platform where the 27 differentially expressed microRNAs from the RDEB vs. healthy comparison, the 99 from KS vs. healthy and the 148 from XPC vs. healthy were independently uploaded as input files. These numbers came from selecting solely the entries corresponding to a FDR lower than 0.05 (Annex). In other words, in study [60], a list of 27/99/148 microRNAs were found to be differentially expressed in RDEB, KS and XPC respectively. This analysis of dysregulation was accomplished taking healthy controls as reference. After choosing the pertinent parameters to run the program (Organism=Homo sapiens, ID Type=miRBase ID and Target Type=Genes), a microRNA-mRNA interaction table is rendered consisting of meaningful columns regarding tissue specificity, validation method (both predictive and empirical) and links to literature, among others. The legitimacy of the predicted interactions is guaranteed by the miRanda confident scores. The number of detected microRNA target genes for each disease is shown in Table 1.

In spite of miRNet's capability to downstream display the interactions table as graphs, miRNet usage was no longer continued here since the author's purpose was to carry these tables to the

Cytoscape domain [64], as this tool enables a much more thorough study of complex networks by integrating expression data profiles (mRNAs) and protein-protein interaction networks. In this way, three tables were initially imported to Cytoscape, containing the 27 microRNAs and their 3781 targeted genes for RDEB, the 99 microRNAs and their 7532 genes for KS and the 148 microRNAs and their 8973 targeted genes for XPC, in compliance with miRNet results. To upload the microRNAs in Cytoscape, the identifier had to be changed for compatibility reasons. Hence, the microRNA identifier was converted into relative IDs (MIMAT) using miRCarta [74]. The next step was to import the logFC corresponding to both microRNAs and genes (again, filtering out at FDR=0.05), from the [60] study. Once that the expression data was imported, only a certain number of the genes targeted by miRNet have got a value for the logFC. This is due to the fact that obviously not all the predicted nor validated interactions between microRNAs and genes retrieved by miRNet are actually appearing in the [60] study, since we are only accounting for those genes being differentially expressed when compared to controls (healthy individuals). In this sense, biologically non-relevant genes were excluded from the network, ending up then with the desired bipartite networks where nodes correspond both to microRNAs and genes. The number of microRNAs, edges and differentially expressed (targeted) genes that arrange each network can be seen in Table 1. These nodes are consequently attributed with logFC values and information about the types of regulatory interactions among them.

TABLE 1. NUMBER OF DE NODES FOR EACH GENODERMATOSES NETWORK AND THEIR CONNECTING EDGES

	microRNAs	miRNet Genes	DE targeted Genes	Edges
RDEB	27	3781	152	229
KS	99	7532	135	290
XPC	148	8973	419	1210
Shared	18	3277	36 (out of 227)	47

As depicted in the above table, a commonly shared network was also constructed with the 18 microRNAs common to RDEB, KS and XPC. This simple check showcases that our initial hypothesis appears to be working satisfactorily, since 36 out of the 227 DE genes from the [60] study are actually being targeted by the shared differentially expressed microRNAs and, as proved further down in this section, they have reasons to be highlighted from a topological and functional point of view.

Apart from that, the “Common Network” is going to serve as a control for our analyses with the RDEB/KS/XPC networks, as its structure is indeed a subgraph of the later.

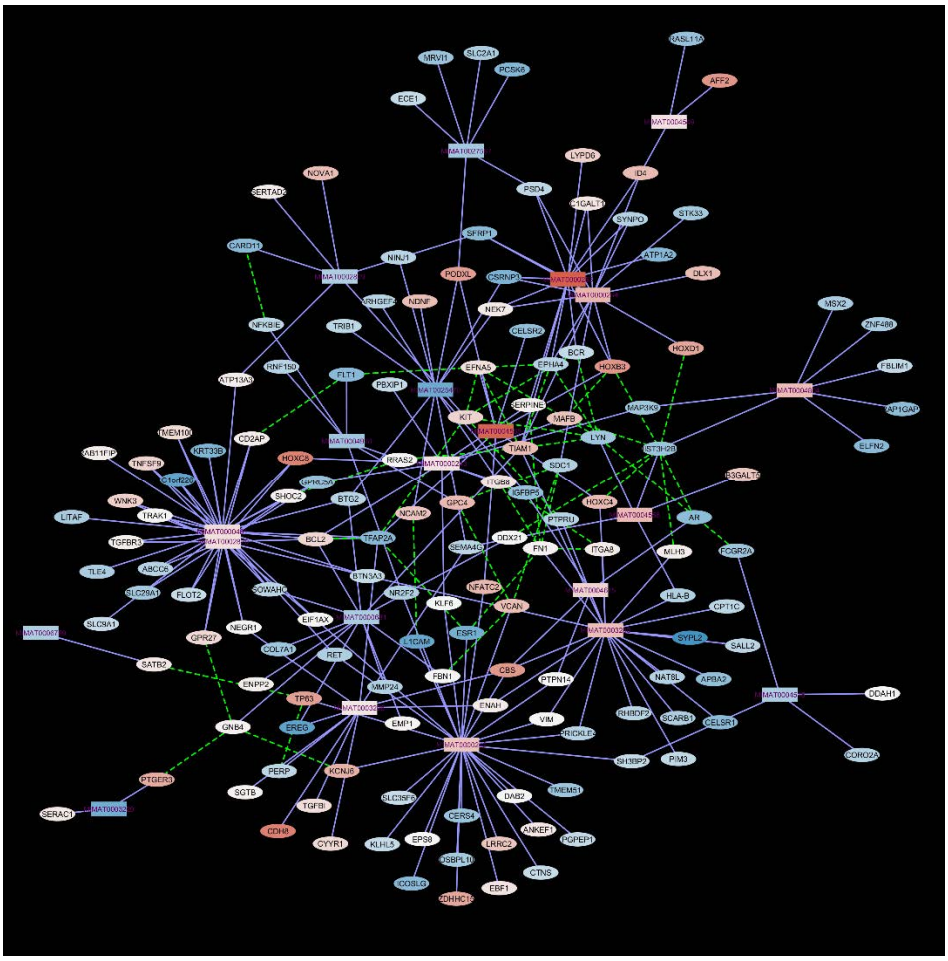
Once genes appearing in the [60] study are correctly matched to their regulatory microRNAs, protein-protein interations (PPIs) were incorporated into our networks (from the interactome). PPIs can be understood as physical contacts of high specificity that genes will have once they are translated into proteins. Therefore, a sound approach will be to investigate both the predicted and already experimental PPIs using STRING database [65], a functional protein association consortium. For integration purposes, there already exists a plugin within Cytoscape able to make use of STRING. In this manner, gene lists were uploaded as input files, confidence cutoff was set at 0.80 (a very stringent value to decrease the fraction of false positives) and the maximum number of proteins was left as default. The number of PPIs and interacting genes can be seen in Table 2.

TABLE 2. NUMBER OF PPIS PRESENT ON EACH DISEASE GRAPH

	Genes	Interacting Genes	Number of PPI
RDEB	152	42 (28%)	44
KS	135	33 (24%)	35
XPC	419	149 (36%)	307
Shared	36	18 (50%)	34

According to the above STRING outputs based on the Interactome content, a significant amount of the targeted genes (24-36%) also experiences other molecular associations among their translated protein chains apart from regulation by microRNAs. These PPIs secure the fact that a portion of the DE proteins from RDEB, KS and XPC in truth act conjointly in some cellular mechanisms. Moreover, Shared Network accounts for a 50% of gene associations by reason of topological and functional relations among them.

The identified PPIs are therefore integrated in the microRNAs → genes networks. At this point, a successful integration of epigenomics (microRNAs), transcriptomics (RNAs) and proteomics (PPIs) has been committed. For simplicity, from now on, these resulting networks will be called "Merged Networks", and are shown in Figure 15 for each disease.



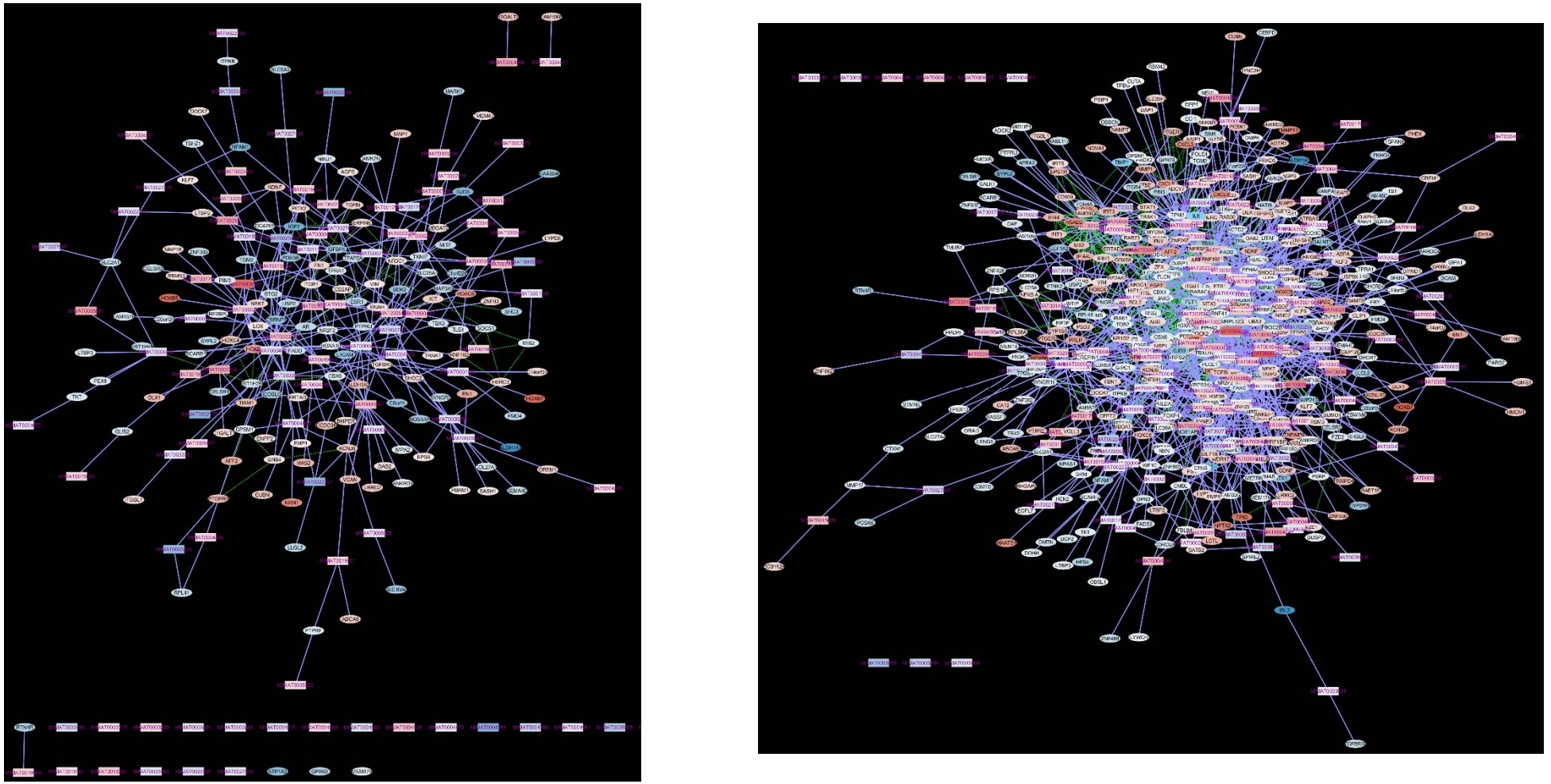


FIGURE 15. MERGED NETWORKS FOR A) RDEB, B) KS AND C) XPC. NODE FILL COLOR CORRESPONDS TO A CONTINUOUS MAPPING WHERE RED MEANS UP-REGULATION WHILE BLUE MEANS DOWN-REGULATION (LOGFC). PURPLE EDGES DENOTE MICRORNA → RNA INTERACTIONS, WHILE GREEN DOT LINES ARE PPIS BETWEEN GENES

c. RDEB, KS and XPC graphs under topological study

Cytoscape hosts a widget for the topological analysis of networks (NetworkAnalyzer), providing information about clusters, hubs, minimal paths or degree distribution. This widget is applied to the analysis of each Merged Network. A new window appears with diverse outputs regarding graph statistics. It is important to comment that PPI interactions are not directed, whereas microRNA→ genes are directed. The bipartite graph can therefore be considered one way or another. The topological analysis was done in both directed and undirected ways but since the results were very similar, from now on and for simplicity, the networks are considered non-directional. As said, all three bipartite graphs rendered scale-free distribution results, and the microRNAs →genes directionality does not affect to the overall topological computation (Figure 24). The most interesting topological parameters for each Merged Network can be seen in Table 3.

The RDEB Merged Network (Table 3) only presents one connected component, also known as clique: every node is somehow connected to every other node; there are no isolated elements. Conversely, KS and XPC present a little number of isolated microRNAs: they do not regulate any of the differentially expressed genes at all.

TABLE 3. SOME OF THE GLOBAL TOPOLOGICAL MEASURES FOR EACH GRAPH

	RDEB	**KS**	**XPC**
Node Degree Distribution	Scale-free	Scale-free	Scale-free
Clustering Coefficient	0.024	0.014	0.053
Isolated nodes	0	24	9
Characteristic path length	4.024	4.395	3.863
Avg. no. of neighbours	3.198	2.778	5.332
Density	0.019	0.012	0.009

A power-law line can be fitted to the **degree distribution** charts. It adopts the shape corresponding to a scale-free distribution: a small number of nodes (microRNAs or genes) are going to be highly connected (hubs), being responsible in turn of the network outcome. Due to the inherent architecture of our networks, microRNAs are mostly going to act as the hubs, having especially high degrees. Their role is to regulate the transcripts outcome. On the other hand, some interesting genes might also behave as hubs (if they are regulated by many microRNAs and hold protein-protein interactions with many other genes). Its disposition may help to identify active hotspots within the network which in turn could serve as potential biomarkers for the genodermatoses under study.

The values obtained for the **clustering coefficients** (Table 3), were as expected for a scale-free network: The XPC network presents the higher clustering coefficient, since a core-cluster is embedded within its giant component formed by over 400 nodes. That is the reason why XPC

global clustering coefficient is the biggest one. This can be explained due to the higher number of genes and microRNAs that were dysregulated in XPC, maybe as a result of its mutational nature (inability to repair damaged DNA leads to even more mutations).

In respect of the **characteristic path length**, every graph showcases a value typical of *small-world* networks: roughly 4 steps have to be taken in order to reach any point of the network, which is quite short indeed. There is a short distance between any pair of nodes. Hubs obviously govern this property, acting as end-to-end bridges. This is a classic result in biomolecular networks, manifesting that small failures or dysregulations in one of its nodes can rapidly diffuse and in a short number of steps, transmitting thus to other different parts of the network.

The average **number of neighbours** is fairly related to the clustering coefficient values: the larger the number of neighbours, the greater the clustering. However, there is a tendency in our networks by which nodes do not extensively share neighbours with others (that is, shared network distribution decays exponentially). This can be explained due to the bipartite nature of the networks: microRNAs regulate a high number of genes, but these genes in turn are not extensively associated with more entities. Of course, there are exceptions (genes with PPIs and genes co-regulated by more than one microRNAs), which actually entail important topological information, and they will be likely acting as hubs.

Density is understood as the portion of all potential connections that are real connections. In this manner, RDEB network showcases the greater density value, supporting thus its strong connectivity: from all the possible interactions that can arise from the nodes, the 1.9% are established indeed. It is large enough so it cannot go unnoticed on downstream analysis.

As a convenient way to locate hubs and identify relevant topological properties of each individual node, the topological parameters can be mapped to the network view. In our case, degree was mapped as the node size and the clustering coefficient as the node color. By using this configuration, an individual examination of each network's most interesting components can be undergone (Figures 16 to 24).

On the microRNAs side, as expected, they mostly stand as distinctive big nodes with a reddish color. This means that, they have a large degree value due to their targeted gene regulation and that their clustering coefficient tends to be low: their neighbours (RNAs) does not share many other edges with other nodes. These genes belong thus to the periphery (leafs in the three structure of the network). This consolidates the assumption that the most relevant hubs will have reversely acquired centrality properties within the network.

RDEB Merged Network

Among the most connected microRNAs (hubs), it is important to highlight the MIMAT0000242 (hsa-miR-129-5p, Degree=29), MIMAT0000461 (hsa-miR-195-5p, Degree=28) and MIMAT0000253 (hsa-miR-10a-5p, Degree=14). Moreover, they are all traversed by a high number of shortest paths (Av. Shortest paths=2.95, 3.03 and 3.43 respectively) accounting thus for a high betweenness centrality (0.287, 0.189 and 0.088 respectively).

As can be seen in Figure 16, a cluster in the Merged Network, the gene with the highest betweenness centrality is TIAM1 (0.121), which is linked to 8 elements in total: MIMAT0000253, MIMAT0000254 (hsa-miR-10b-5p), EPHA4, LYN, EFNA5, MIMAT0003283, RRAS2 and MIMAT0003298. That is, four different microRNAs have showed to dysregulate the translation of the TIAM1 mRNA into protein and, on top of that, it somehow experiences physical

associations with four DE genes. It is closely followed by TFAP2A, presenting 7 different edges that link to: MIMAT0004555, ESR1, MIMAT0000461 (has-miR-195-5p), BCL2, MIMAT0002820, MIMAT0003298, MIMAT00045555 and KIT. This time, 5 different microRNAs silence the same gene which is in turn associated with three more proteins.

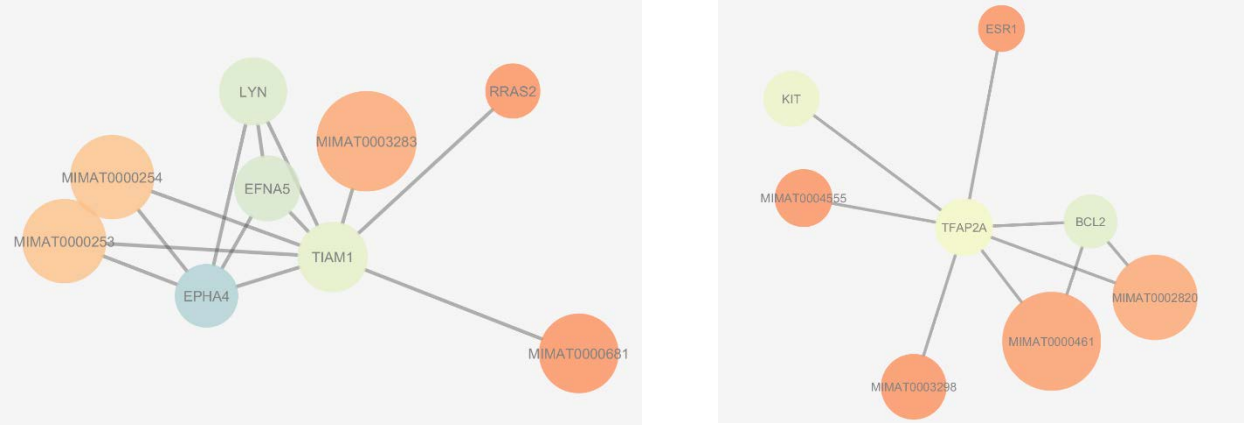


FIGURE 16. TIAM1 & TFAP2A CLUSTERS

Having a look to the clustering coefficient, ITGA8 has got by far the largest value (0.67). It is easily located in the network due to its bluish mapping (Figure 17). ITGA8 possess a strategic location in which it is connected to FN1, ITGB8 and MIMAT0003283. ITGB8 in addition exhibits PPIs with FN1, which is also regulated by MIMAT0003283.

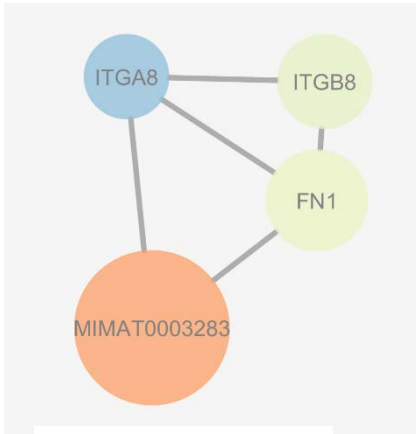


FIGURE 17. ITGA8 CLUSTER

KS Merged Network

The topological mapping immediately displays a more homogeneous graph where the microRNAs do not regulate so many genes and the peripheral nodes are particularly small in size compared to the RDEB network. In fact, the most connected microRNAs are MIMAT0000424, MIMAT0000252, MIMAT0000242 (hsa-miR-129-5p), MIMAT0000461 (hsa-miR-195-5p) and MIMAT0003283 (Degrees=16 and 14 for the first two and 13 for the last three). Moreover, their neighbours tend to share edges with other elements in the graph (either genes or microRNAs), showcasing then the large centrality of these 6 microRNAs: they are greatly traversed by a high number of nodes.

BTG2 is postulated as one of the most relevant genes herein (Figure 18). With a degree of 11 and an average shortest path length of 3.29, it is impressively regulated by 11 different microRNAs among whom one should point out MIMAT0000252, MIMAT0000461 (hsa-miR-195-5p) and MIMAT0000681 (hsa-miR-29c-3p).

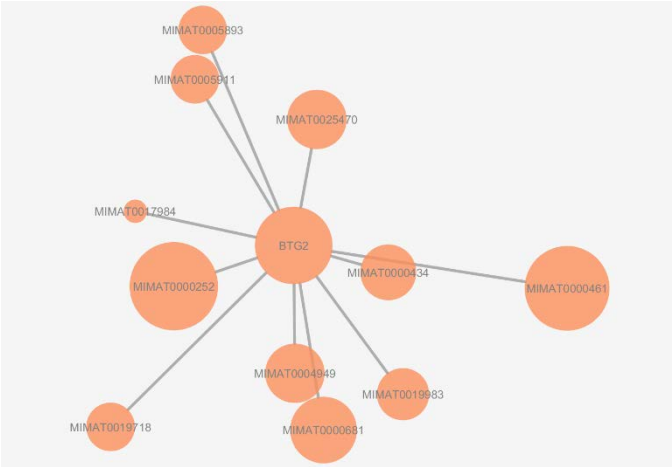


FIGURE 18. BTG2 AND ITS NEIGHBORS

It is important to mention as well HOXA10 (Figure 19). With a degree of 8 and an average shortest path length of 3.27, it resembles to BTG2's structure. Among the 13 different microRNAs, the most significant are MIMAT000424 and MIMAT0000461 (hsa-miR-195-5p). No PPIs have been recognized either for HOXA10.

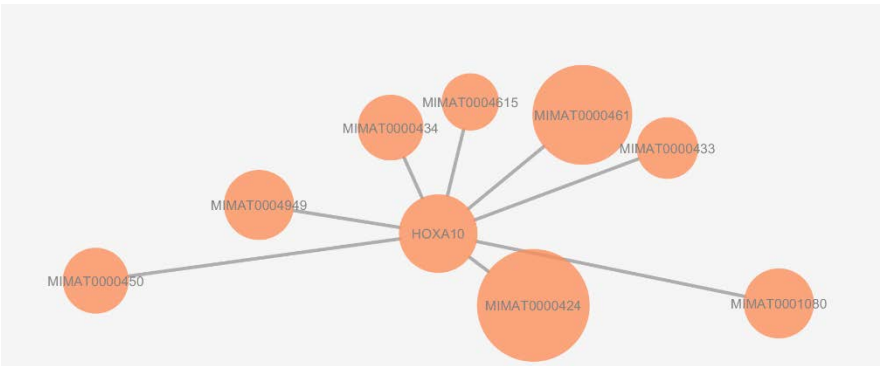


FIGURE 19. HOXA10 AND ITS NEIGHBORS

Regarding the clustering coefficients, there are two community structures that can be easily perceived at a glance by their non-reddish mapping (Figure 20): The first one is represented by SERPINE1, IGF2 and FN1, which seem to work together. SERPINE1 is regulated by two interesting microRNAs that are MIMAT0000243 and MIMAT0000437. Moreover, FN1 interacts with ITGB1 and ESR1. The other interesting clustered structure is formed by the genes SOCS2, RNF182, HERC3 and MIB2. SOCS2 also connects to KIT and MIMAT00004333, which regulates HOXA10.

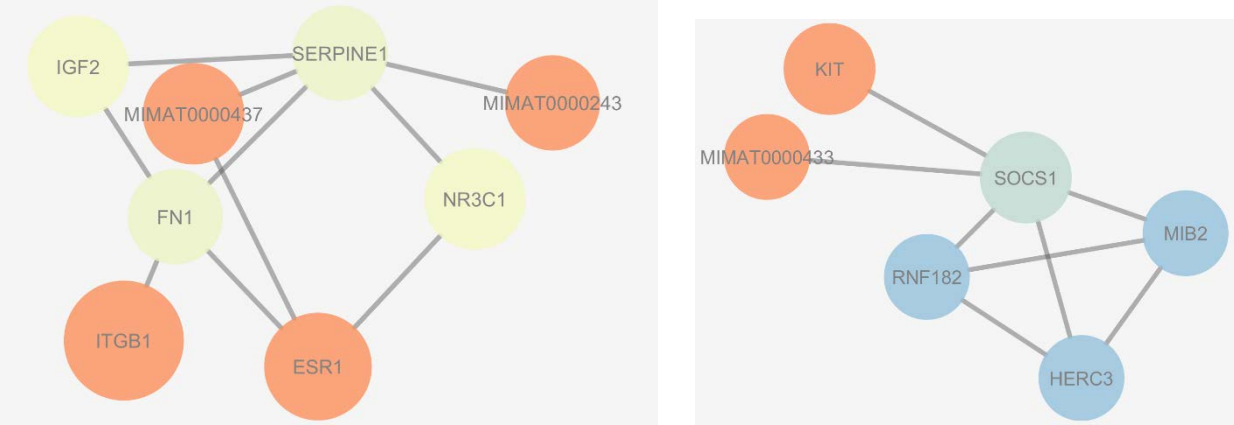


FIGURE 20. SUBGRAPHS WITH HIGHER CLUSTERING COEFFICIENTS

XPC Merged Network

A giant core-cluster can be appreciated the moment we map the topological parameters, with small-size genes in the periphery as well. Since it stands as the largest-scale graph (568 nodes and 1517 edges, from which 307 are PPIs), its topological analysis is not very intuitive.

Regarding the degree, MIMAT0000646, MIMAT000075 and MIMAT0000242 (hsa-miR-129-5p) are pinpointed as the most relevant (Degree= 55, 42 and 37 respectively). However, a bunch of other microRNAs also have important degree values (in the order of three and two dozens). For instance, MIMAT0000461 (hsa-miR-195-5p) has got 32 edges and adopts a very interesting location inside the giant core-cluster where it shows a highly-ranked betweenness centrality (0.037).

Among the genes with larger degree values, STAT1 and HERC5 (Figure 21) look to have an important influence in the core-cluster too (Degree= 21 and 19 respectively). On top of that, STAT1 is also regulated by the major microRNA of this graph: MIMAT0000646. Looking deeply, in fact, a PPI is held between STAT1 and HERC5, which conjointly give rise to an impressive cluster where every node is a gene. Simply put, a highly interconnected PPIs cluster is found embedded in the XPC Merged Network, where two quite relevant genes (STAT1 and HERC5) appear to have a notorious role. The bluish mapping of these cluster nodes confirms the aforementioned.

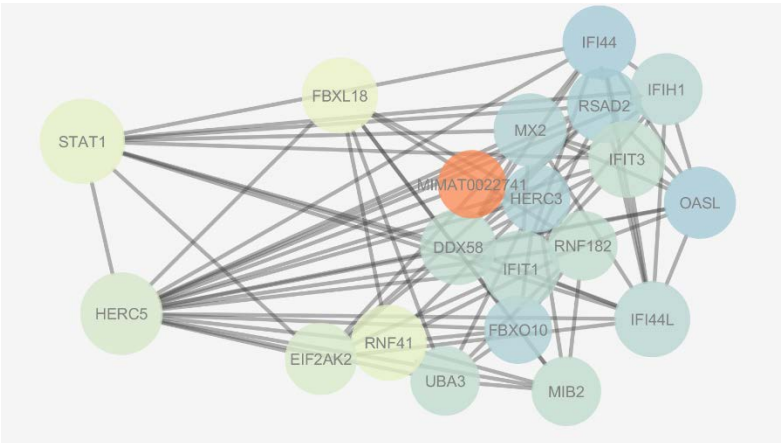


FIGURE 21. HERC5 CLUSTER

These two genes are followed in the degree ranking by ITGB1 (Figure 22), with 16 edges that connects to a wide array of nodes: 6 microRNAs of different node-size regulate ITGB1 and 10 genes with significant clustering coefficients (pale node-color), being: ITGB4, TGM2, L1CAM, FN1, KDR, ITGB3, FYN, TGFB1, FBN1 and ITGB8. Several of them have been previously commented and will be discussed later on.

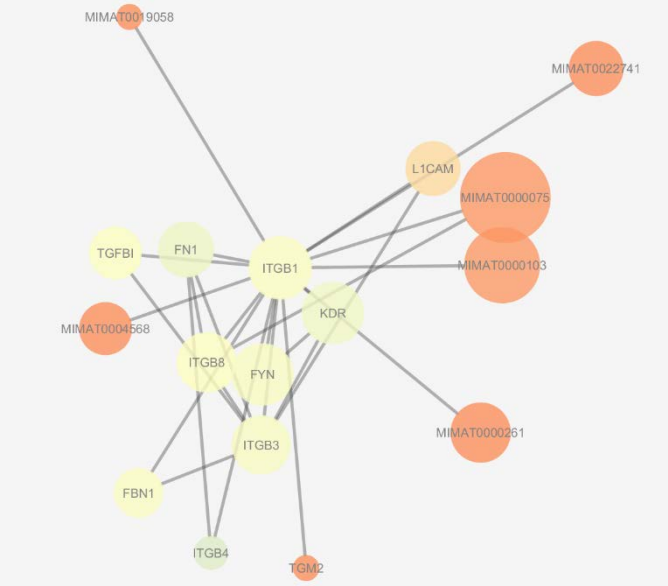


FIGURE 22. ITGB1 CLUSTER

Some of the top-ranked gene nodes with higher betweenness centrality values are included in the previous clusters. Others like ESR1, TFAP2A and BTG2 (Figure 23) deserve as well being mentioned (betweenness centrality= 0.014, 0.008 and 0.014 and degree=12,10,11 respectively). Into the bargain, TFAP2A presents PPI with ESR1, which is in turn connected to FN1 and regulated (among 6 others) by MIMAT0000242 (hsa-miR-129-5p, degree=37).

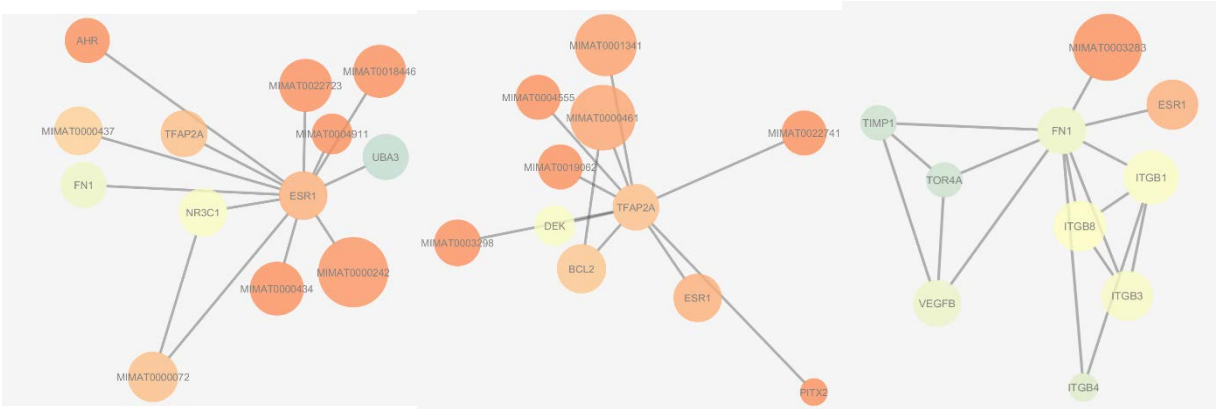
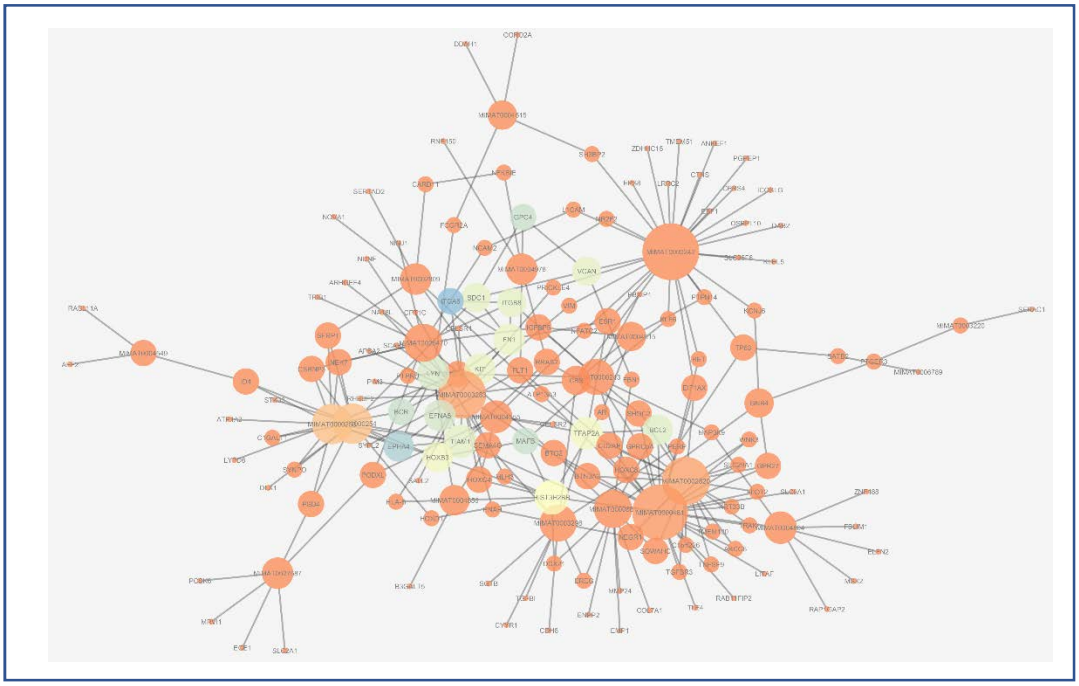
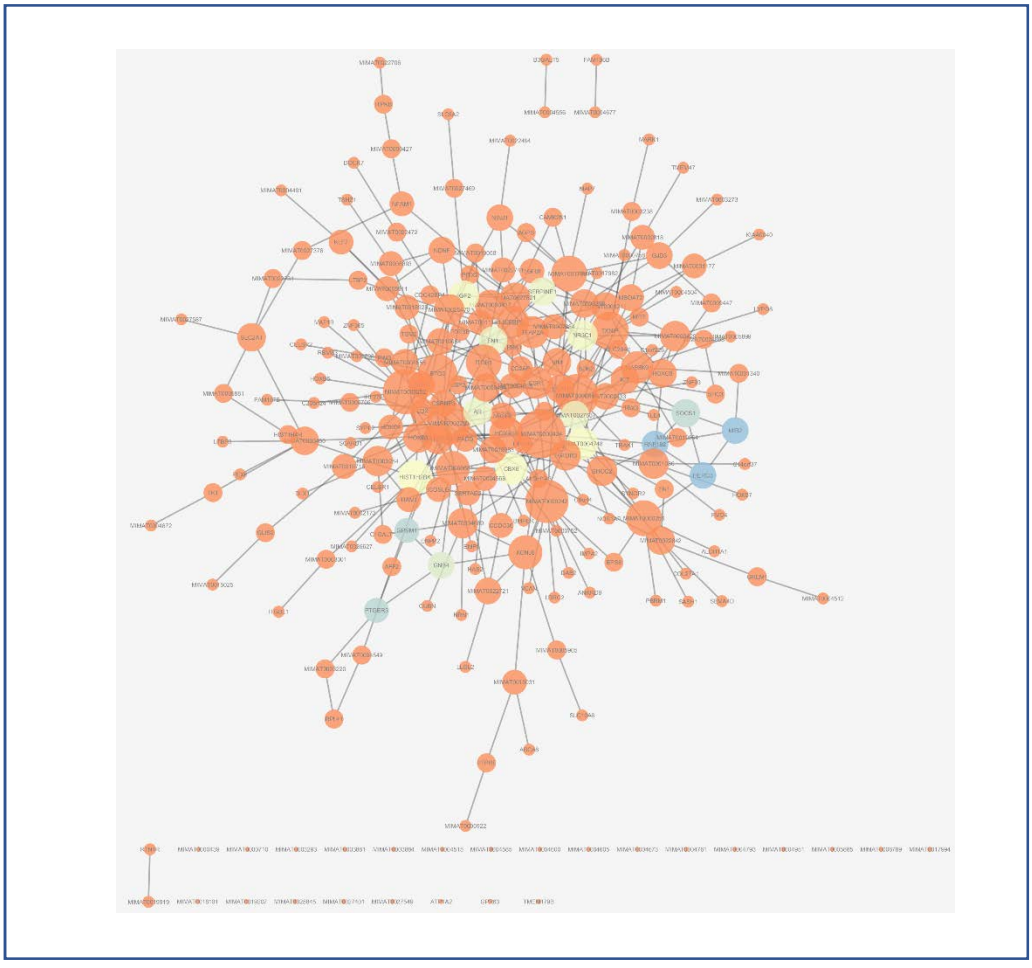


FIGURE 23. ESR1, TFAP2A AND FN1 CLUSTERS





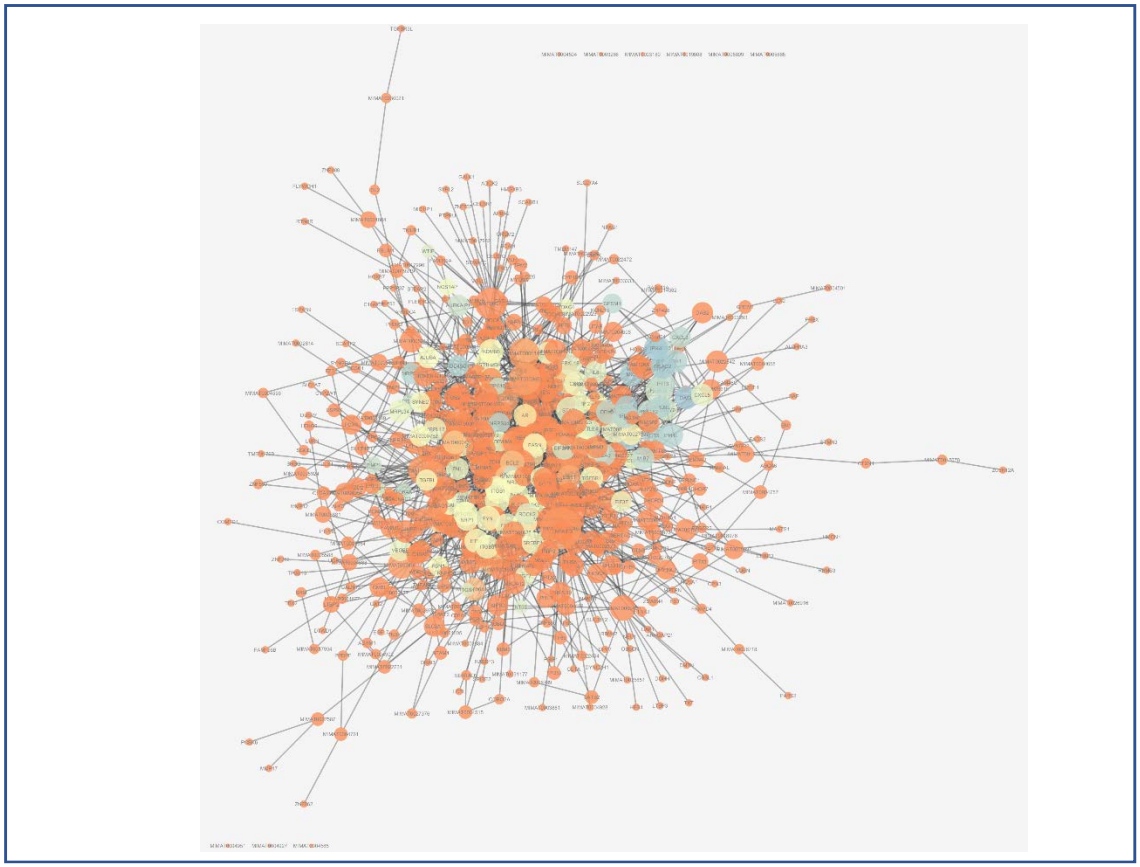
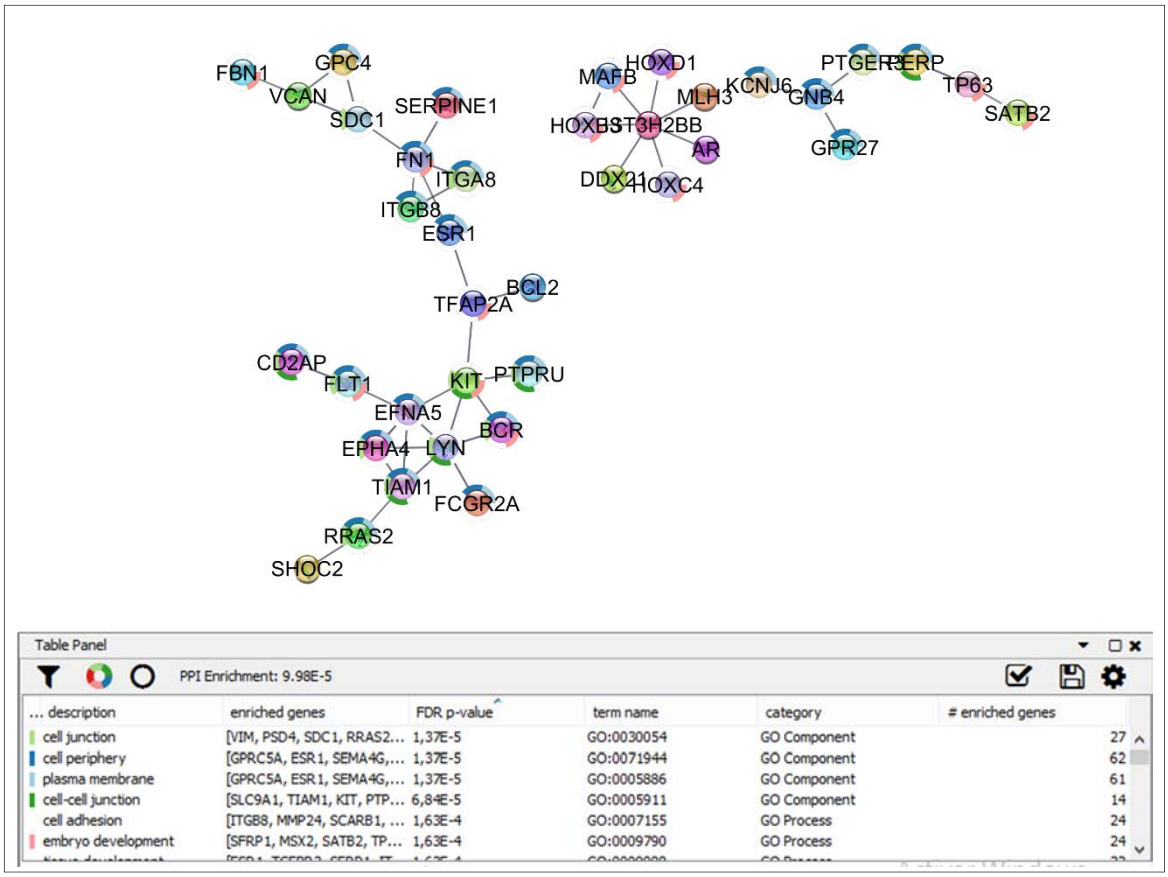


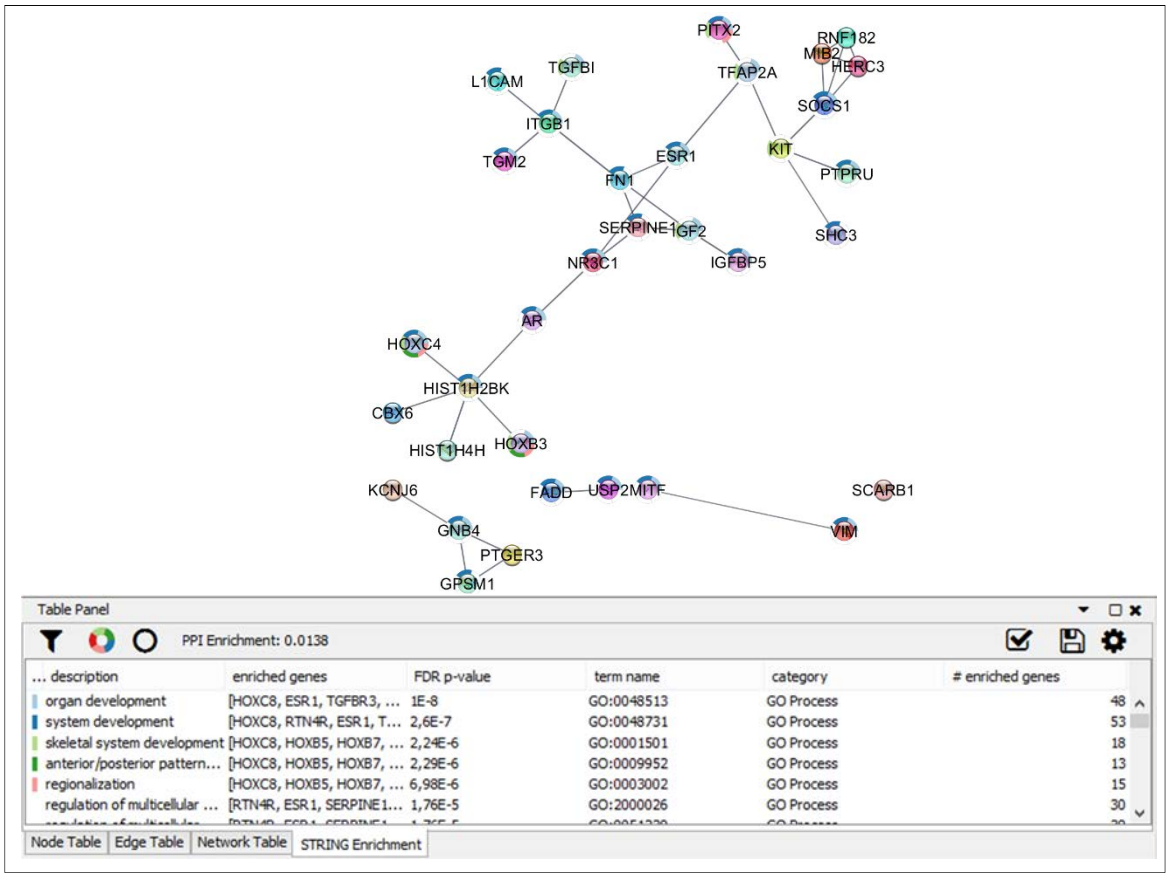
FIGURE 24. DEGREE AND CLUSTERING COEFFICIENT MAPPED FOR EACH NODE ON THE GENODERMATOSES GRAPHS A) RDEB, B) KS AND C) XPC

d. Functional enrichment analysis

Functional enrichment analysis was performed using two different plugins from Cytoscape: BiNGO and STRING. By loading the DE gene lists (those genes that are differentially expressed and are regulated by the microRNAs), a table is rendered where the DE genes are grouped on different enriched GO and KEGG functional categories. Basically, if a significant number of genes are identified to presumably mediate in any of the biological categories recorded in the ontologies, it is marked then in the correspondent enriched function. By these means, every DE gene is computationally tested, retrieving enriched functions in which they are involved (Figure 25). In the STRING plug-in, an overall automatically computed PPI enrichment score will tell whether the network PPIs are more significant among themselves than a pure random set of PPIs of similar size picked from the genome. In other words, if PPI p-value enrichment is lower than 0.05, it can be stated that the enriched genes are at least partially biologically connected, as a group.

In addition, and for visual purposes, charts by colour can be plotted along the PPIs networks, representing the participation of each gene in the five most relevant enriched functions (according to the adjusted p-value). However, these most relevant enriched functions might correspond to generalist processes or pathways, not saying too much about specific molecular mechanisms.


FBN1
GPC4
VCAN
SDC1
SERPINE1
FN1
ITGA8
ITGB8
ESR1
TFAP2A
BCL2
CD2AP
FLT1
KIT
PTPRU
EFNA5
BCR
EPHA4
LYN
TIAM1
FCGR2A
RRAS2
SHOC2
MAFB
HOXD1
MLH3
HOXB3
HIST3H2BB
AR
DDX21
HOXC4
KCNJ6
GNB4
PTGER3
ERP
TP63
SATB2
GPR27
Table Panel
PPI Enrichment: 9.98E-5
... description
enriched genes
FDR p-value
term name
category
enriched genes
cell junction [VIM, PSD4, SDC1, RRAS2... 1,37E-5 GO:0030054 GO Component 27
cell periphery [GPRC5A, ESR1, SEMA4G,... 1,37E-5 GO:0071944 GO Component 62
plasma membrane [GPRC5A, ESR1, SEMA4G,... 1,37E-5 GO:0005886 GO Component 61
cell-cell junction [SLC9A1, TIAM1, KIT, PTP... 6,84E-5 GO:0005911 GO Component 14
cell adhesion [ITGB8, MMP24, SCARB1, ... 1,63E-4 GO:0007155 GO Process 24
embryo development [SFRP1, MSX2, SATB2, TP... 1,63E-4 GO:0009790 GO Process 24


PITX2
RNF182
MIB2
HERC3
L1CAM
TGFBI
TFAP2A
SOCS1
ITGB1
KIT
TGM2
ESR1
PTPRU
FN1
SERPINE1
IGF2
SHC3
NR3C1
IGFBP5
AR
HOXC4
HIST1H2BK
CBX6
HIST1H4H
HOXB3
KCNJ6
FADD
USP2
MITF
SCARB1
VIM
GNB4
PTGER3
GPSM1
Table Panel
PPI Enrichment: 0.0138
... description
enriched genes
FDR p-value
term name
category
enriched genes
organ development [HOXC8, ESR1, TGFBR3, ... 1E-8 GO:0048513 GO Process 48
system development [HOXC8, RTN4R, ESR1, T... 2,6E-7 GO:0048731 GO Process 53
skeletal system development [HOXC8, HOXB5, HOXB7, ... 2,24E-6 GO:0001501 GO Process 18
anterior/posterior pattern... [HOXC8, HOXB5, HOXB7, ... 2,29E-6 GO:0009952 GO Process 13
regionalization [HOXC8, HOXB5, HOXB7, ... 6,98E-6 GO:0003002 GO Process 15
regulation of multicellular ... [RTN4R, ESR1, SERPINE1... 1,76E-5 GO:2000026 GO Process 30
Node Table
Edge Table
Network Table
STRING Enrichment

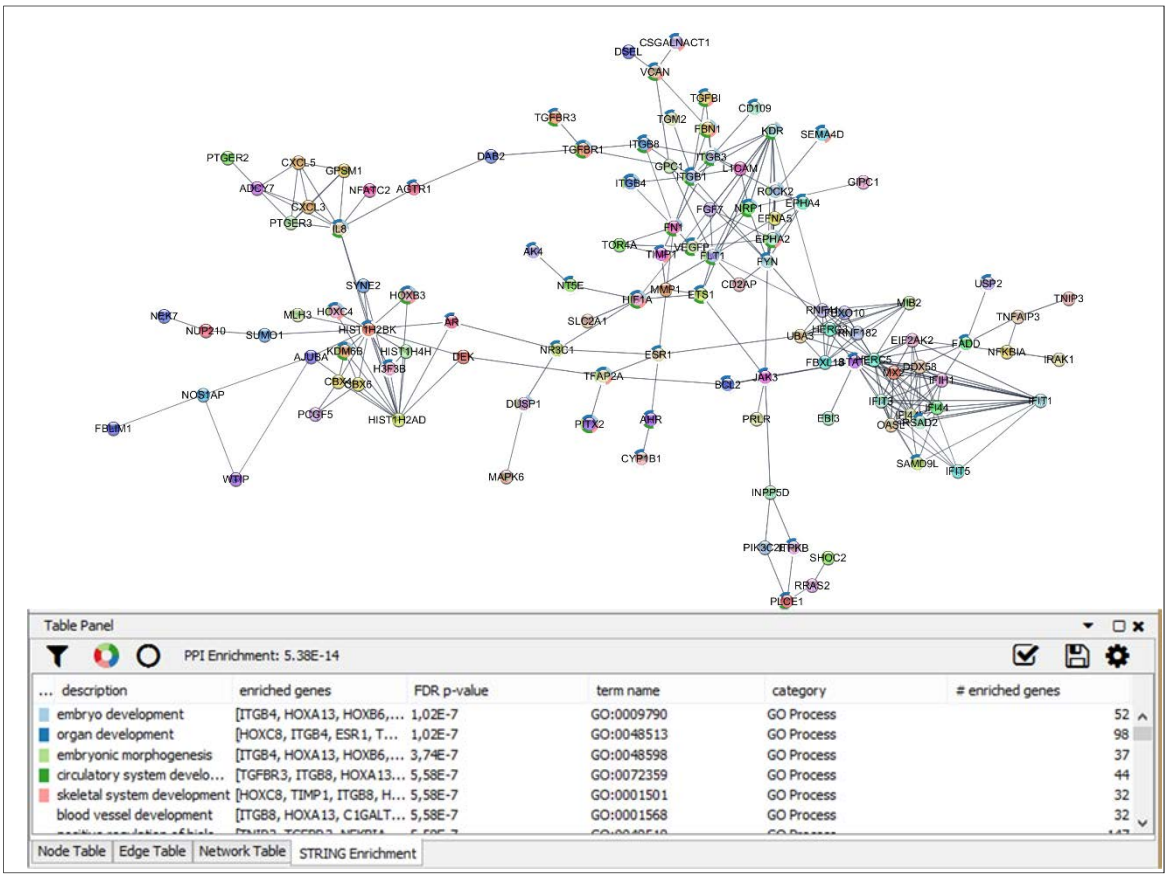


FIGURE 25. STRING PPIS AND FUNCTIONAL ENRICHMENT ANALYSIS NETWORKS FOR THE DE GENES CORRESPONDING TO A) RDEB, B) KS AND C) XPC.

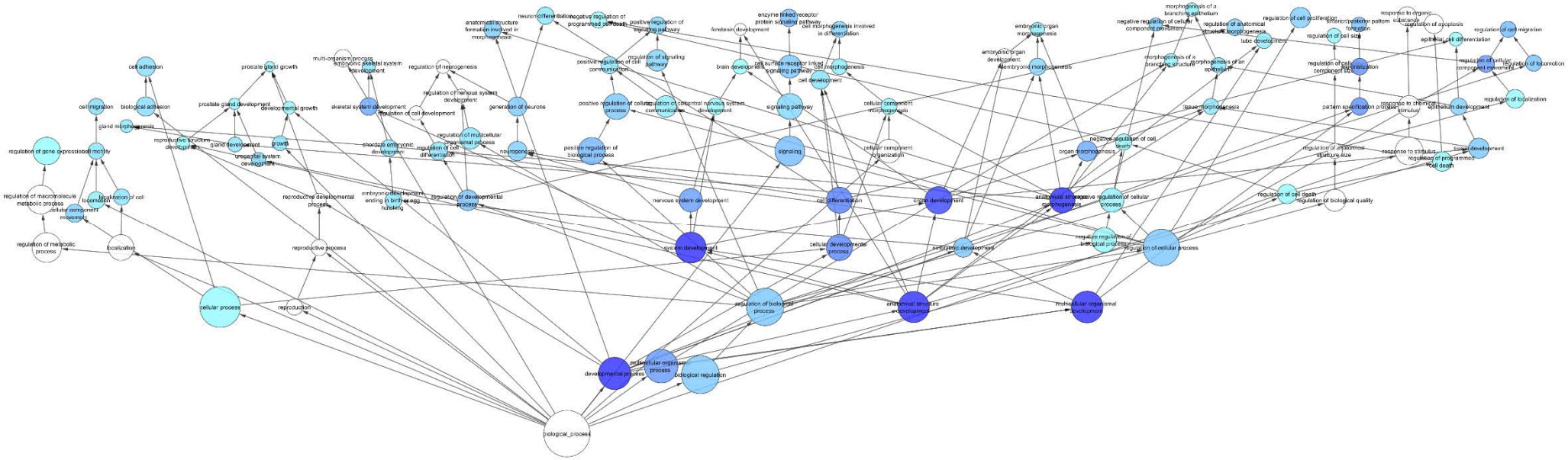
The other plug-in, BINGO generates hierarchical models where the plotted elements are naturally enriched functions. Setting up a proper Hierarchical Layout after its creation, the graph adopts a directed tree fashion, also known as dendogram. This graph is nothing but an enriched subnetwork out of all the GO terms that have been recorded over the years in the Gene Ontology Consortium [68]. Due to its inherent structure, they exhibit manifest family relations, where the higher-level "descendants" come from certain "ancestors" and these branches, in turn, from a root node.

In this way, these hierarchical graphs are constituted by nodes (representing each ontology term) whose parameters are computed by BINGO, yielding 1) an adjusted p-value which corresponds to the gradual node fill color scale (the lower the p-value, the darker the fill color) and 2) a "node size mapping" provided by the amount of functional overrepresented genes integrated in every enriched function.

Some nodes act as interconnectors: they are not enriched nor contain any functional genes at all, but are used as a convenient bridge between enriched functions that are not directly linked by themselves. These interconnectors are depicted as white (or empty) nodes and, by algorithm default, sometimes they appear as end-nodes as well.

The Gene Ontology is composed of three different functional categories 1) Biological Process (BP), 2) Molecular Function (MF) and 3) Cellular Component (CC), and therefore three hierarchical graphs are retrieved for each genodermatoses DE gene list, thus ending up eventually with 9 different BiNGO enriched networks. As our purpose is to study the common

resemblances among the functional enrichment analyses for the three diseases and in order to simplify the analysis, Cytoscape was used to generate networks intersections between the three diseases for each functional category. In this manner, three final intersected networks are obtained where the enriched elements are jointly shared by RDEB, KS and XPC (Figure 26).


biological_process
cellular process
reproduction
reproductive process
regulation of biological process
multicellular organism process
developmental process
biological regulation
anatomical structure development
multicellular organismal development
system development
cellular developmental process
nervous system development
cell differentiation
organ development
anatomical structure morphogenesis
regulation of cellular process
signaling
signaling pathway
cell development
cell morphogenesis
generation of neurons
neurogenesis
regulation of developmental process
regulation of multicellular organismal process
regulation of cell differentiation
positive regulation of biological process
regulation of nervous system development
regulation of neurogenesis
cell adhesion
biological adhesion
cell migration
gland morphogenesis
prostate gland development
prostate gland growth
gland development
growth
developmental growth
reproductive structure development
urogenital system development
skeletal system development
multi-organism process
embryonic skeletal system development
chordate embryonic development
embryonic development ending in birth or egg hatching
reproductive developmental process
regulation of gene expression, epigenetic
regulation of macromolecule metabolic process
cellular component movement
localization of cell
localization
regulation of metabolic process
neuron differentiation
anatomical structure formation involved in morphogenesis
negative regulation of programmed cell death
positive regulation of signaling pathway
forebrain development
enzyme linked receptor protein signaling pathway
cell morphogenesis involved in differentiation
regulation of signaling pathway
positive regulation of cell communication
brain development
cell surface receptor linked signaling pathway
positive regulation of cellular process
regulation of cell communication
regulation of central nervous system development
cellular component morphogenesis
cellular component organization
positive regulation of biological process
embryonic organ development
embryonic organ morphogenesis
morphogenesis of a branching epithelium
negative regulation of cellular component movement
regulation of anatomical structure morphogenesis
lung development
regulation of cell proliferation
branching morphogenesis of an epithelial tube
regulation of cell size
morphogenesis of a branching structure
morphogenesis of an epithelium
tissue morphogenesis
organ morphogenesis
negative regulation of cell development
regulation of cell development
response to inorganic substance
regulation of apoptosis
regulation of cell migration
epithelial cell differentiation
regulation of cell motility
regulation of locomotion
pattern specification process
response to chemical stimulus
epithelium development
regulation of localization
regulation of anatomical structure size
response to stimulus
regulation of programmed cell death
regulation of cell death
regulation of biological quality
negative regulation of biological process
positive regulation of cellular process

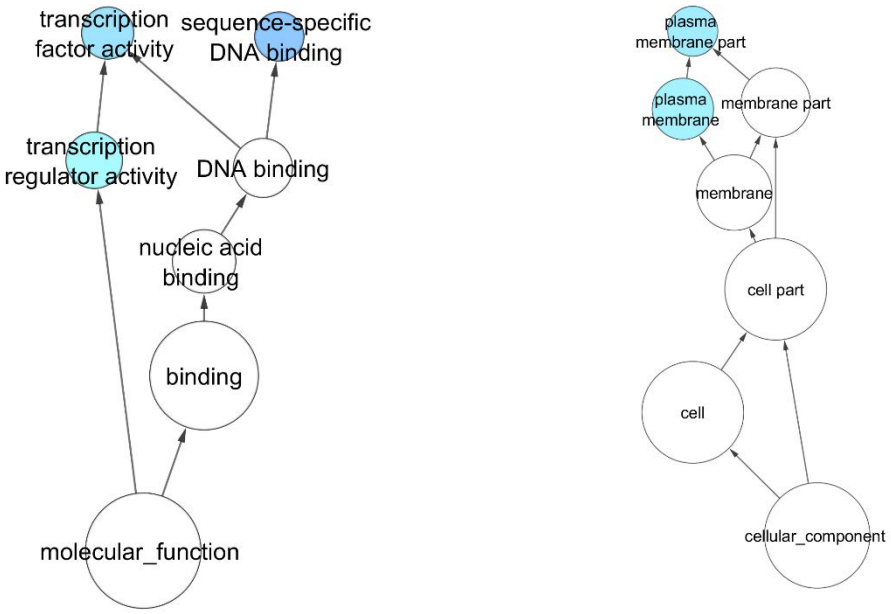


FIGURE 26. BINGO FUNCTIONAL ENRICHMENT NETWORKS FOR THE INTERSECTION OF A) BP, B) MF AND C) CC. NODE SIZE IS MAPPED WITH THE NUMBER OF GENES ENCOMPASSED, AND NODE COLOR IS MAPPED WITH THE ENRICHMENT P-VALUE

The leafs, or end-nodes, are going to be mainly the enriched functions of interest, since they retain statistically significant biological diversity. This arborescence can be studied to determine the disease molecular development that caused the shared phenotype signature by the three genodermatoses.

The enriched functions both from STRING (KEGG database) and BINGO (Gene Ontology) are collected all together in a table according to the enrichment p-values they rendered in both executions. Some of these enriched functions were to appear in both analyses, while other appeared just in one. An insightful scheme is depicted in Figure 27. By performing both single enrichment analyses (SEAs), we are increasing the likelihood to cover the majority of the enriched elements.

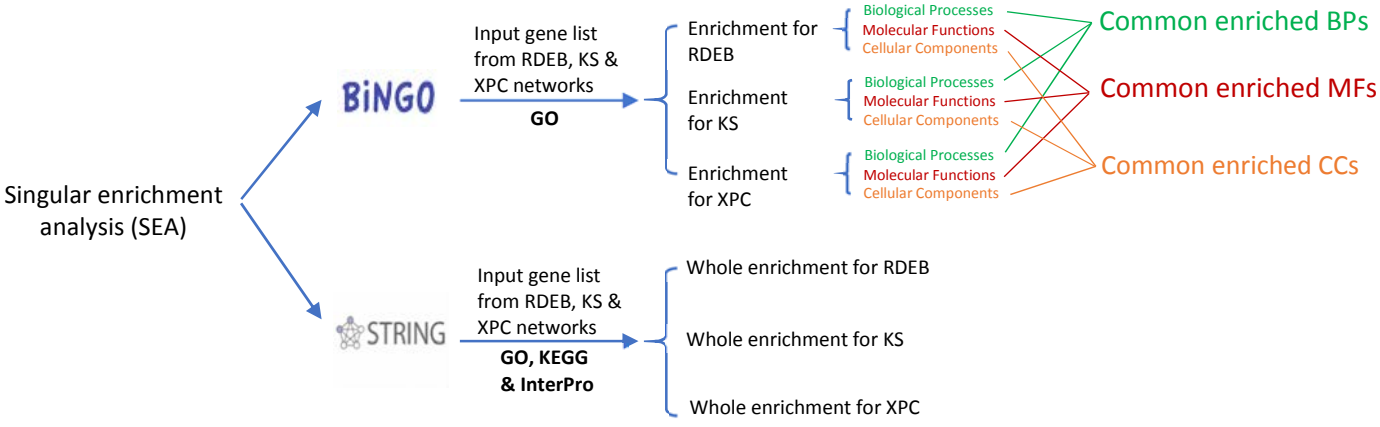


FIGURE 27. SEAS APPROACHES. IT IS IMPORTANT TO NOTE THAT STRING DOES NOT RETRIEVE NETWORKS WITH ENRICHED FUNCTIONS, ONLY PPIS GRAPHS

e. Grouping and operating with the enriched functions. ReVIGO & ClueGO

Merging parallel SEA datasets is a convenient and necessary point to further explore the enriched functions. The generated data was compiled in order to make it properly accessible for its succeeding analysis. By these means, system-level intepretations can be performed as a conclusion.

A priori one might argue that BiNGO SEAs are more reliable since they already include an intersection of the enriched functions shared by the three diseases. However, STRING SEAs potentially enclose important evidences too as they include a broadly used database (KEGG) and would therefore stand for a wider functional coverage.

Since 804 out of the 831 functions gathered from all the SEA datasets were detected at the GO server, the grouping was carried out using ReVIGO [71], a bioinformatics tool that clusters gene ontology terms according to similarities in their functions. The rest of the entries correspond to strikes at either KEGG or InterPro (protein domain) databases, and as the number was not too large they did not necessarily need to be grouped.

It is important to note that each single enriched function could have been rendered for one, two or even for the three diseases. In other words, not all the enriched functions have the same distribution across RDEB, KS and XPC. For instance, “sequence-specific DNA binding”, (which corresponds to the GO term GO:0043565) is an enriched function only for KS and XPC, while “cellular response to lipid” (or GO:0071396) is enriched only in RDEB and XPC. Moreover, as two analogous SEA approaches were followed (BINGO & STRING), some enriched functions even have two parallel (non-identical) p-values resulting from both algorithms.

By the time of grouping the enriched functions by their GO semantic similarity using ReVIGO, those entries that had got more than one p-value (either due to their manifestation in at least two genodermatoses or because they were part of both STRING & BiNGO results) are considered only once at the ReVIGO input interface and have been assigned with the highest p-value found for them in the compiled dataset. For instance, “central nervous system development”, or GO:0007417, has gotten an enrichment p-value of 7.60E-03 in BINGO for RDEB, while 4.57E-04, 2.50E-04 and 7.55E-06 in STRING for RDEB, KS and XPC respectively. This GO term accounts then for 4 entries with the same description name but with different false discovery rates. Therefore, we entered 7.60E-03 (the highest one, securing the statistical certainty) as the input for GO:00007417. Yet, the process is completely consistent since ReVIGO only uses p-values to guide the grouping selection, if possible. They do no mediate in the choice of the clusters’ representatives. Other than that, p-values have no further use in biological interpretation of results, since all the enriched functions are statistically significant ($p<0.05$) due to the nature of the functional enrichment analysis (only considers functions that are statistically significant enriched in the DE list of genes).

After setting up the input data, i.e, every GO Term ID and the enrichment p-value, they can be loaded in ReVIGO (Figure 28), where the allowed semantic similarity was 0.7 (Figure 28). In this way, 521 enriched GO terms were uploaded (283 repeated terms were excluded).

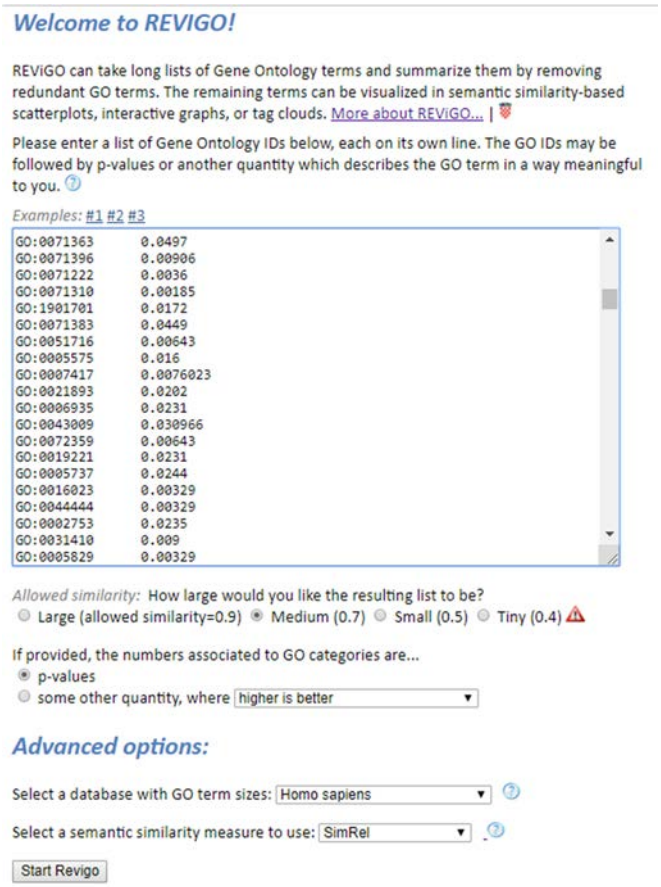


FIGURE 28. REVIGO INPUT INTERFACE

The output of ReVIGO provided several clusters of functions. The retrieved table (Figure 29) includes columns regarding frequency (the greater the frequency, the more general the GO term is), log10 p-value, uniqueness (to check for atypical values, also known as outliers) and dispensability (a measure to group GO terms within the clusters).

GO:0009605	response to external stimulus	12.043 %	-4.9281	0.96	0.14
GO:0071396	cellular response to lipid	3.076 %	-2.0429	0.92	0.14
GO:0009755	*hormone-mediated signaling pathway*	*1.281 %*	*-1.5287*	*0.87*	*0.72*
GO:0043401	*steroid hormone mediated signaling pathway*	*1.050 %*	*-2.0031*	*0.87*	*0.71*
GO:0048857	neural nucleus development	0.364 %	-1.9245	0.87	0.15
GO:0060021	palate development	0.508 %	-2.4698	0.88	0.15
GO:0019222	regulation of metabolic process	35.730 %	-1.4572	0.90	0.15
GO:0006952	defense response	8.904 %	-1.5768	0.95	0.16
GO:0009719	response to endogenous stimulus	9.175 %	-1.4841	0.96	0.16
GO:1904018	positive regulation of vasculature development	0.814 %	-3.2565	0.70	0.16
GO:0048514	*blood vessel morphogenesis*	*2.856 %*	*-2.7190*	*0.72*	*0.93*
GO:0072358	*cardiovascular system development*	*3.531 %*	*-2.1918*	*0.75*	*0.82*
GO:0001568	*blood vessel development*	*3.347 %*	*-3.2027*	*0.74*	*0.79*
GO:0001525	*angiogenesis*	*2.389 %*	*-2.3036*	*0.73*	*0.91*
GO:0045766	*positive regulation of angiogenesis*	*0.716 %*	*-2.8794*	*0.68*	*0.86*
GO:0045765	*regulation of angiogenesis*	*1.264 %*	*-2.5482*	*0.69*	*0.92*
GO:0007507	*heart development*	*2.891 %*	*-3.1186*	*0.74*	*0.80*
GO:1901342	*regulation of vasculature development*	*1.408 %*	*-2.7570*	*0.71*	*0.85*
GO:0021559	trigeminal nerve development	0.069 %	-1.5498	0.86	0.17
GO:0090092	regulation of transmembrane receptor protein serine/threonine kinase signaling pathway	1.223 %	-2.8069	0.88	0.17
GO:0030510	*regulation of BMP signaling pathway*	*0.479 %*	*-1.4498*	*0.80*	*0.80*
GO:0001763	morphogenesis of a branching structure	1.114 %	-3.2565	0.81	0.17
GO:0044260	cellular macromolecule metabolic process	48.748 %	-2.5784	0.92	0.18
GO:0031344	regulation of cell projection organization	3.306 %	-1.8386	0.91	0.18
GO:0060839	endothelial cell fate commitment	0.035 %	-3.5200	0.84	0.19
GO:0097101	*blood vessel endothelial cell fate specification*	*0.017 %*	*-1.6946*	*0.80*	*0.92*
GO:0060846	*blood vessel endothelial cell fate commitment*	*0.017 %*	*-1.6946*	*0.80*	*0.92*
GO:0060847	*endothelial cell fate specification*	*0.017 %*	*-3.0721*	*0.85*	*0.85*
GO:0008610	lipid biosynthetic process	3.866 %	-1.4306	0.92	0.20
GO:0003002	regionalization	1.922 %	-2.7144	0.82	0.20
GO:0009952	*anterior/posterior pattern specification*	*1.171 %*	*-2.4056*	*0.83*	*0.86*
GO:0009954	*proximal/distal pattern formation*	*0.179 %*	*-1.9957*	*0.86*	*0.70*
GO:0044707	single-multicellular organism process	35.118 %	-2.5272	0.86	0.21
GO:0060322	head development	4.091 %	-1.8386	0.83	0.22

FIGURE 29. SNAPSHOT OF THE GROUPED GO TERMS

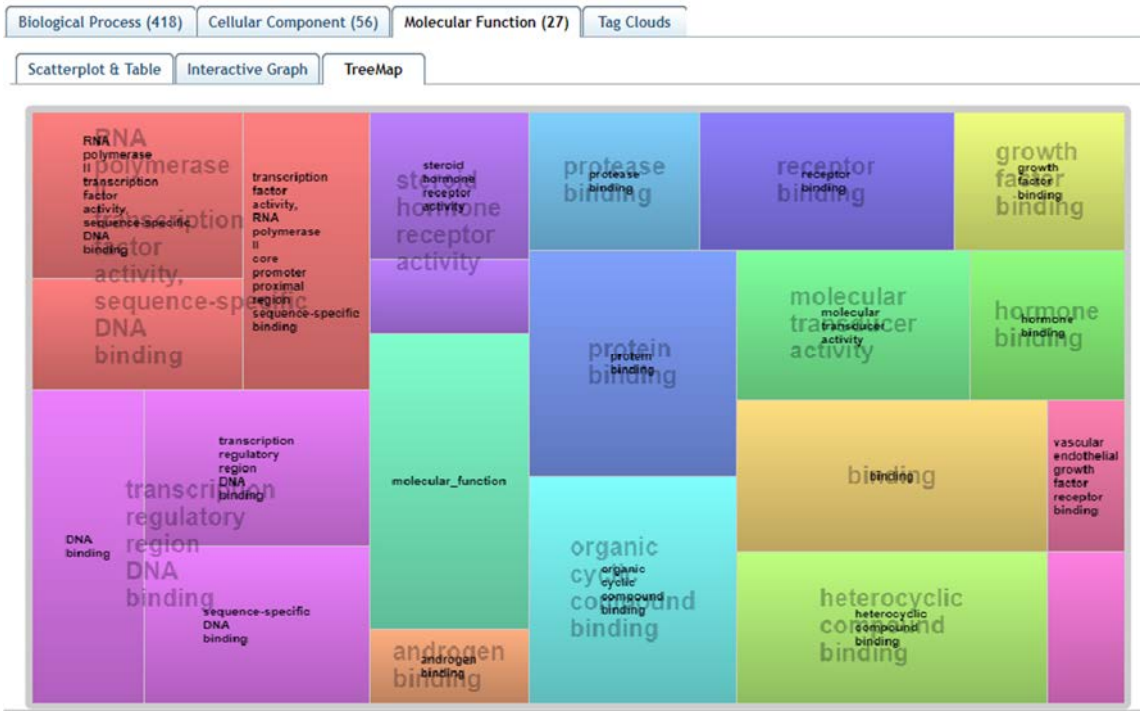


FIGURE 30. ENRICHED MF TREEMAP

Upon visualizing the results in Figure 30, it is appropriate to highlight that the Molecular Function category (MF) displays a set of 27 enriched elements, where the majority relates to binding processes and functions, whether for intra or extracellular environments.

ReVIGO also displays interactive graphs to examine the clusters formed (Figure 31).

As an example, the interactive graph for the Cellular Components (CC) from ReVIGO was opened with Cytoscape, and includes three well-established clusters where every single enriched function (now represented like nodes) has to do with membrane and binding structures. First cluster contains tight entities of the plasma membrane, the nodes of the second one bears relation with the extracellular space, while the third cluster is about cell-cell junctions.

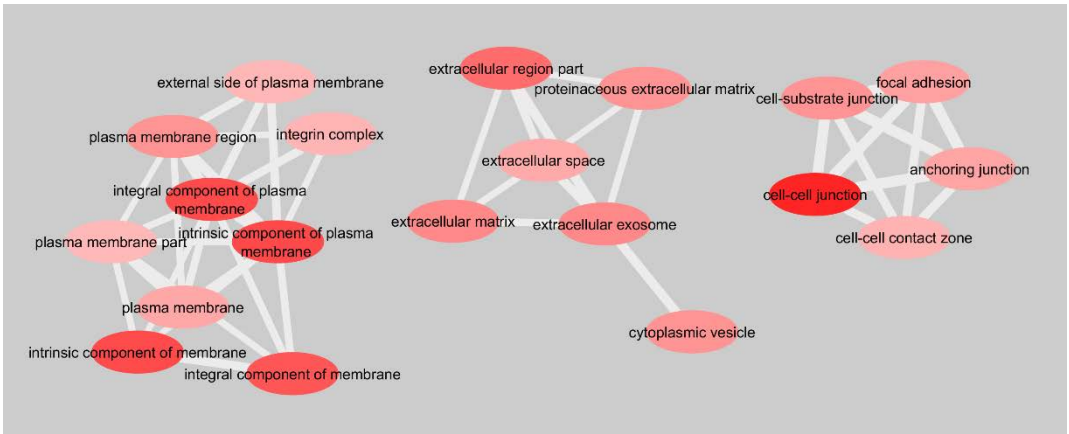


FIGURE 31. CELLULAR COMPONENTS (CC) CLUSTERS

This ReVIGO validation provides a confirmation that the candidate BP, CC and MF of interest for our study are, as expected, sufficiently represented in the overall picture: “cell-cell adhesion”, “extracellular matrix structures”, “regulation of inflammation processes”, “development”, “chemotaxis” and “cell cycle” are, among others, the most recurrent enriched functions in the three genodermatoses, which makes sense according to the phenotype of the diseases.

However, at this stage, the results from ReVIGO only provide unrefined aggrupation of functions, so a more sophisticated approach is used hereafter with a different Cytoscape tool, named ClueGO. Up to now, a systemic analysis has taken place where all the enriched functions have contributed equally (yet depending on their enrichment p-value). ReVIGO grouping allowed to keep track of the most relevant enriched hits. The grouped dataset was curated afterwards, to only consider those enriched functions somehow conjointly shared by all RDEB, KS and XPC (Figure 32). As can be seen in this figure, some functions, processes or pathways can be commonly dysregulated among the three diseases, although the individual DE genes that trigger the dysregulation may be different. Manual curation allows us to specifically select those enriched functions that are indeed shared by RDEB, KS and XPC. With that in mind, we can perform another downstream analysis, i.e. ClueGO, in step with ReVIGO outcomes.

ClueGO is another Cytoscape plugin where gene lists can also be systemically analysed, though encompassing more options and features. Multiple annotation and ontology resources cover the running of the program, which allows to display the results in graph form too. Basically, genes are mapped on the selected ontologies and if the predefined criteria are met, the algorithm sets the interrelationships among genes.

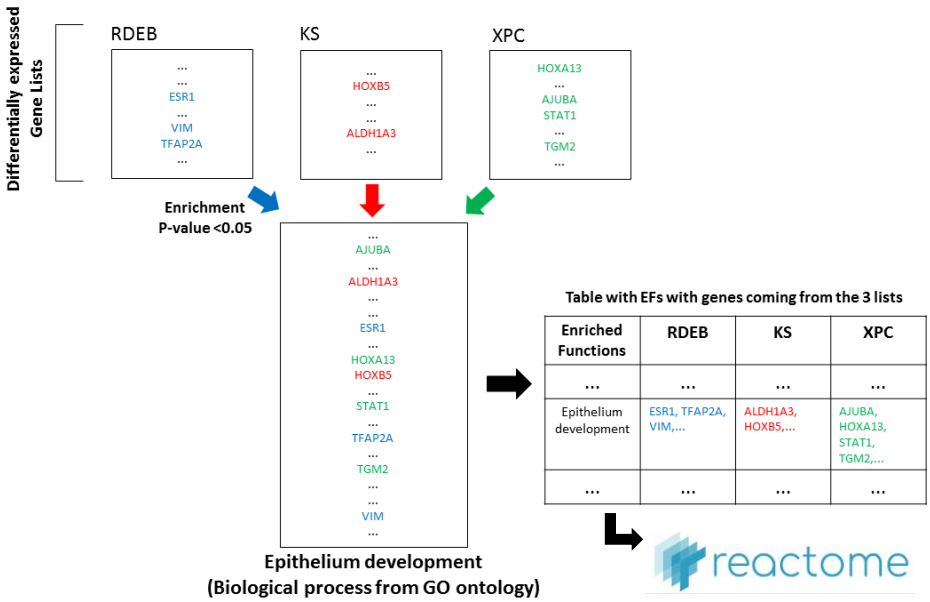


FIGURE 32. AN ILLUSTRATIVE EXAMPLE IS PROPOSED TO EVIDENCE THE TABLE CONSTRUCTION BASED ON ENRICHMENT P-VALUES, TAKING SIMULTANEOUSLY THE THREE GENE LISTS AS INPUTS

To do so, the three gene lists (RDEB, KS and XPC) for each enriched function were imported in the ClueGO input panel. The database employed in this approach was REACTOME, for both Reactions and Pathways. Overall, "REACTOME Pathways" gathers 2227 recognized pathways and 10629 available unique genes and "REACTOME Reactions", for its part, 11635 terms with 10894 genes. In a general picture, REACTOME is the most compatible ontology with ClueGO, so its running fits properly with the results rendered by both STRING & BiNGO. In Figure 33, the ClueGO input panel is showcased, where the RDEB/KS/XPC DE genes for the "epithelium development" enriched function has been introduced.

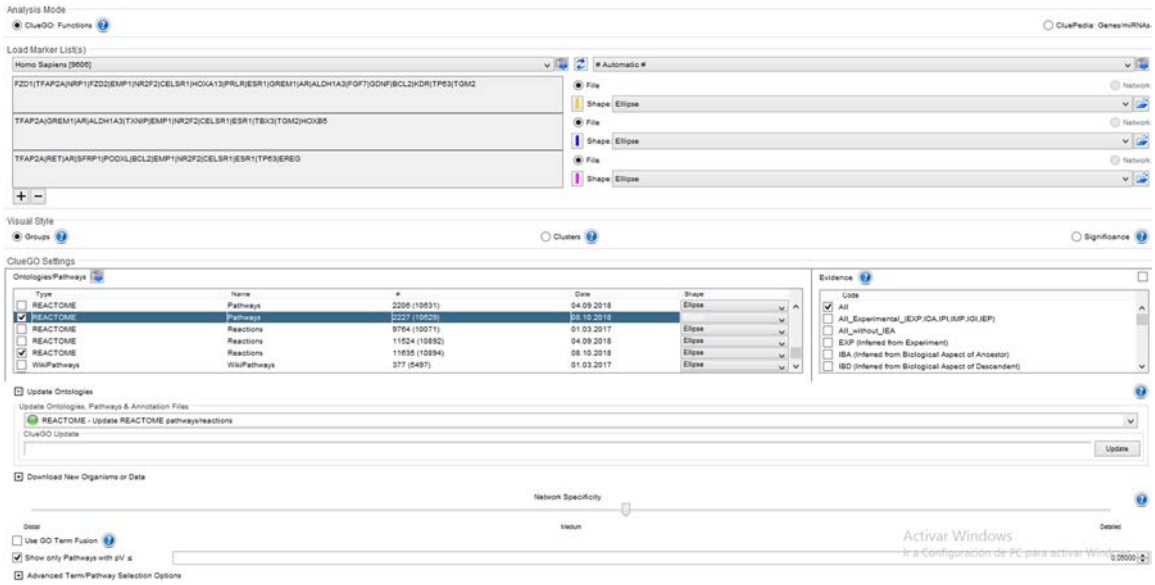


FIGURE 33. INPUT PANEL FOR CLUEGO. DE GENES FROM RDEB/KS/XPC HAVE BEEN IMPORTED IN THE UPPER-LEFT CORNER

One by one, all the functions that have been found to be enriched with DE genes from the three genodermatoses (in RDEB/KS/XPC) are subjected to a final functional analysis with REACTOME. A total of 39 functions from BiNGO enrichment study fulfilled the aforementioned requisite, whereas 35 enriched functions from STRING met the needs. A thorough scrutiny of the results led to the construction of a worksheet (Table 4 for BiNGO and Table 5 for STRING) where all the retrieved elements by REACTOME were compiled, for every enriched function. In these tables, columns exhibit recurrent hits retrieved by ClueGO and marked for each enriched function. The total number of recurrent hits is shown at the last column and row of each worksheet. These hits might explain some of the hallmarks from the genodermatoses phenotype.

TABLE 4. WORKSHEET FOR BINGO EFS

BiNGO EFs \ Recurrent hits	Elastic fibre formation	TFAP2A regulation of GFs and their receptors	Integrins/ Syndecans	Semaphorin	Neutropilin interacts with VEGF and VEGFR	Signal transduction by L1	EPHA binds	Chemokines	NR-MED1 coactivator complex	NOTCH	PI3K	Interleukin signaling	Exocytosis of platelets	SCF-KIT	
Anatomical structure development	x	x	Both	x	x	x	x TIAM1		x			x	x	x	11
Anatomical structure formation. Morphogenesis					x										1
Anatomical structure morphogenesis		x			x		x EFNAs								3
Biological regulation		x			x		x TIAM1					x		x	5
Cell differentiation				x	x	x	x TIAM1		x		x			x	7
Cell migration			Both	x		x									3
Cell motility			Both	x		x									3
Cell surface receptor linked signaling pathway	x		Both			x	x TIAM1			x				x	6
Cellular component movement			Both	x		x	X NGEF								4
Cellular developmental process					x	x	x TIAM1		x		x			x	6
Developmental process	x	x	Both	x	x	x	x TIAM1		x			x	x	x	11
Generation of neurons				x			x TIAM1								2
Localization of cell								x							1
Locomotion			Both	x		x		x							4
Multicellular organismal development	x	x	Both	x	x	x	x TIAM1		x		x	x	x	x	12
Multicellular organismal process	x	x	Both	x	x	x	x TIAM1		x		x	x	x	x	12
Nervous system development				x			x TIAM1								2
Neurogenesis				x			x TIAM1								2
Neuron differentiation							x EFNAs								1
Organ development	x	x	Both		x	x	x NGEF				x	x		x	9
Organ morphogenesis		x													1
Plasma membrane part	x		Both			x	x EFNAs				x				5
Positive regulation of biological process		x			x	x			x	x (3)					5
Positive regulation of cellular process				x	x	x			x	x (1)	x				6
Positive regulation of signaling pathway											x				1
Regulation of anatomical structure morphogenesis				x							x				2
Regulation of biological process		x			x		x TIAM1		x			x		x	6
Regulation of cell migration					x										1
Regulation of cell proliferation					x						x			x	3
Regulation of cellular component movement					x									x	2
Regulation of cellular process		x			x		x TIAM1		x			x		x	6
Regulation of developmental process		x		x	x				x		x				5
Regulation of locomotion					x										1
Regulation of multicellular organismal process		x									x				2
Regulation of signaling pathway											x				1
Signaling	x		Both	x	x	x	x TIAM1		x	x	x	x		x	11
Signaling pathway	x		Both	x	x	x	x TIAM1			x	x			x	9
Skeletal system development	x														1
System development	x	x					x TIAM1				x		x	x	6
	11	14	13	17	21	17	21	2	12	5	16	9	5	16	

TABLE 5. WORKSHEET FOR STRING EFS

STRING EFs / Recurrent hits	Elastic fibre formation	TFAP2A regulation of GFs and their receptor	Integrins/ Syndecans	Semaphorin	Neutropilin interacts with VEGF and VEGFR	Signal transduction by L1	EPHA binds	Chemokines	NR-MED1 coactivator complex	NOTCH	PI3K	Interleukin signaling	Exocytosis of platelets	SCF-KIT	
Anatomical structure development	x		Both		x	x			x			x			6
Anatomical structure formation. Morphogenesis			Syndecans		x										2
Anatomical structure morphogenesis	x	x	Both		x	x							x		6
Brain development				x											1
Cardiovascular system development	x		Both		x	x									4
Cell communication	x	x	Both		x	x		x	x						7
Cell differentiation		x	Both	x		x									4
Cell growth						x									1
Cell migration					x	x	x TIAM1	x				x			5
Cell surface receptor signaling pathway	x		Both	x	x	x	x TIAM1	x		x (1)		x			9
Central nervous system development				x											1
Circulatory system development	x		Both		x	x									4
Developmental process	x		Both		x	x			x			x			6
Embryonic morphogenesis	x		Both	x		x									4
Embryonic organ morphogenesis		x													1
Enzyme linked receptor protein signaling pathway				x	x		x TIAM1							x	4
Epithelial cell differentiation		x									x				2
Epithelium development		x									x				2
Extracellular matrix organisation	x		Both												2
Gland development									x						1
Head development				x											1
Localization of cell			Syndecans		x	x	x TIAM1	x				x			6
Locomotion			Syndecans	x	x	x		x				x			6
Morphogenesis of an epithelium		x													1
Movement of cell or subcellular component			Syndecans	x	x	x		x				x			6
Nervous system development						x									1
Protein binding		x	Both		x		x NGEF	x			x	x	x	x	9
Regulation of anatomical structure morphogenesis		x		x	x		x								4
Regulation of apoptotic process				x					x			x			3
Regulation of cell adhesion				x			x TIAM1						x		3
Regulation of cell differentiation	x		Both	x		x						x			5
Regulation of cellular component movement				x	x			x				x	x		5
Regulation of signal transduction			Both	x	x	x	x TIAM1		x	x (1)		x	x		9
Skeletal system development		x													1
Vasculature development	x		Both		x	x									4
	11	10	18	15	18	18	8	8	6	2	3	12	5	2	

Thanks to the traceability of the enrichment analysis, the dysregulation distribution for each enriched function was surveyed (Figure 34). Basically, gene dysregulation was assessed within each enriched function: according to the logFC values taken from the RNA-Seq, genes were initially mapped in accordance with their up- or down-regulation (Figure 15. logFC>0, overexpressed in blue and logFC<0, underexpressed in red). The idea was then to generate a distribution of the dysregulation pattern for each enriched function. For example, if "Generation of neurons" included 59 dysregulated genes, count how many of them have been presumably identified as over- and under-expressed. The plotted distribution evidenced that no abrupt changes are present for any enriched function (Figure 34). That is, the total number of dysregulated genes encountered for any enriched function is proportionally halved (up vs. down) for most of the cases. This scenario approximates to the idea that there exists a genetic expression balance, where the cellular functions potentially involved in a genodermatoses experience a state of equilibrium thanks to the cell's ability to level over- and under- transcripts expression.

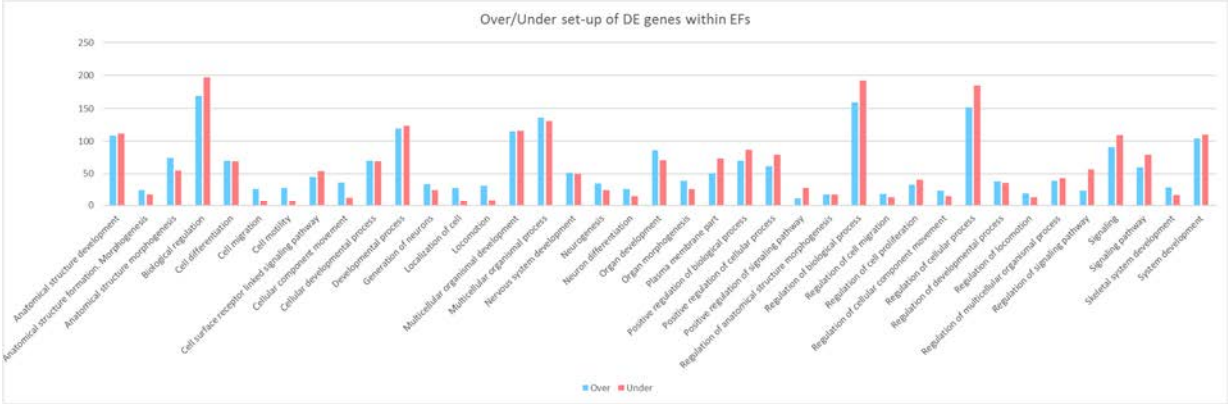


FIGURE 34. COLUMN CHART EXHIBITING THE DYSREGULATION DISTRIBUTION FOR EACH ENRICHED FUNCTION FROM BINGO

5. DISCUSSION

In a large extent, this project puts the spotlight on bioinformatics analysis to drive omics data integration from three phenotypically-related genordermatoses. The necessity for a convenient approach to work through massive omics data has been brought to the fore herein. To accomplish these endeavours, biological network construction has been proposed as a baseline. In this way, the generation of RNA graphs regulated by microRNAs has contributed to frame a global picture were mRNA expression and diversification for RDEB, KS and XPC are exposed. Networks allowed to perform system-level observations, ensuing in turn a narrowing-down strategy to finally identify certain candidate agents. These agents could eventually bring to light possible relations between lesser known genes and gear prospective blueprints towards them. To go over these agents, two approaches were used herein: the topological and the functional analyses.

a. Topological characterization

Even though there is not a clear consensus on how hubs can be identified or defined, network statistics definitely assist on the attempt to find and examine them. Topology provides bona fide results where novel candidate hub genes can be inspected. Once the hubs are recognized, one can make hypothesis regarding their essentiality on the graph structure. Understanding how the network behave when perturbations are introduced and targeted to the hubs is crucial to get phenotypic insights.

With that said, one should note that the constructed graphs for RDEB/KS/XPC contain nodes that can represent either mRNAs or microRNAs, that is, they are bipartite graphs. This conditions the topological results, since the majority of the hubs turn out to be microRNAs: by removing them from the networks, the structure of the web is deeply compromised. This makes sense, since the main structure of the network is based on microRNAs and their mRNA targets. Clear examples of microRNA hubs are 1) MIMAT0000242, MIMAT0000461 and MIMAT0000253 for RDEB (which have been indeed validated with the RT-qPCR), 2) MIMAT0000242, MIMAT0000252 (validated), MIMAT0000461 and MIMAT0000681 (validated) for KS and 3) MIMAT0000646, MIMAT000075 and MIMAT0000242 for XPC.

On the other side, certain genes have also been located as hubs for the studied networks. However, it should be noted that the Cytoscape algorithm which renders network properties (NetworkAnalyzer) is purely graph theoretical and uses no biological criteria, so the conclusions for gene hubs might not apply for other different biological networks, but they do for the RDEB/KS/XPC dysregulation networks constructed in this project. In this way, one can profile the candidate hub genes by the network they have been found in:

RDEB:

TIAM1 accounts for the gene with highest betweenness centrality, being connected to hub microRNAs and genes with high clustering coefficients (such as EPHA4, EFNA4 and LYN), so it should be remarked. Its value for betweenness centrality is closely followed by TFAP2A, or transcription factor AP-2, which exhibits as well a key interconnection with microRNAs and

relevant genes (Figure 16). TIAM1, EPHAs and EFNAs are molecules already studied which have showed to be entangled in skin inflammation processes and carcinogenesis [76].

By its part, ITGA8 presents the largest clustering coefficient, having tight connections with corresponding elements (Figure 17). ITGA8 is an integrin precursor protein, which shows PPIs with ITGB8 and FN1 (fibronectin). FN1, adds to the bargain a PPI with SDC1 gene, which translates into syndecans: transmembrane proteins that ligate with fibroblasts growth factors, among others [77]. In this way, Figure 17 manifests a conformation where two out of the three hallmarks of the genodermatoses common phenotype appear [19]: membrane and cell binding structures (integrins and syndecans) and inflammation (fibronectin).

KS:

BTG2 gene manifests silencing by two of the validated microRNAs (among others not validated): MIMAT0000461 and MIMAT0000681 (Figure 18). It is well-studied and have been recognised as a fundamental tumour suppressor [78]. With a shortest path length of 3.29, BTG2 presents an interesting silencing by 11 microRNAs, representing a strong communicative point within the KS graph.

For the same reason, HOXA10 shows a similar arrangement (Figure 19): highly silenced by 13 microRNAs and possessing a shortest path length of 3.27. HOXA10 gene downregulation has been observed in early stages of the human developmental process, being a critical regulator of megakaryocyte development (related again to inflammation responses) [79].

Regarding clustering coefficients, there is a conformation where SERPINE1, IGF2 and FN1 appear to work together. SERPINE1 is an inhibitor of fibrinolysis and IGF2 stands for the insulin-like growth factor 2, crucial for cellular growth and proliferation during embryogenesis [80]. These three molecules might mediate tightly a specific cellular function. Furthermore, the subgraphs exhibit connections with ITGB1 and ESR1. These connections of dysregulated transcripts might hold important information about KS loss-of-function events.

Another subgraph is shown in Figure 20, this time with the largest clustering coefficient values for the KS network. Two elements should be highlighted: 1) SOCS1 is a suppressor of cytokine signalling through the JAK/STAT3 pathway [81], and appears to be a negative regulator in IGF1R signalling pathway. 2) Mutations in HERC3 have been associated with carcinomas [82]. Again, inflammation- and cancer proneness-related structures.

XPC:

STAT1 and HERC5 are ranked among the genes with larger degree values by the Cytoscape algorithm in the XPC microRNA $\rightarrow$ RNA network (degree=21 and 19 respectively). JAK/STAT signalling pathway has been previously associated with increased inflammation and oxidative stress (brain and skin aging). During cell signalling, it mediates integrin and actin filaments dynamics, affecting adhesion and cell movement [83]. By its part, HERC5 is required for interferon functioning during immune responses [84]). STAT1 and HERC5 are deeply embedded in a giant cluster found within the XPC graph (Figure 21). A remarkable number of PPIs has been found across it, and the majority of the genes are mapped with large clustering coefficients, as expected. Important information regarding this cluster might be downstream retrieved by refining the clustering tools. Agglomerative methods (based on degree) find only the core community, leaving out the periphery, which tends to be neglected [85]. A shift toward iterative algorithms where the edges with highest betweenness centralities are successively removed is

required for a better performance in finding cluster communities. That is why in this study only obvious-to-the-eye clusters have been interpreted. As a reflection for this cluster, it is important to point that large clusters do not necessarily group genes functionally, because it contains genes from many unrelated diseases. Alternatively, it indicates that complex phenotypes are entangled together genetically [85], which again proves the necessity of addressing the optimization of the clustering algorithms.

ITGB1 shows a large degree in the XPC network (Figure 22): 16 edges with relevant PPIs such as FN1, more integrins (ITGB8 and ITGB3), TGFB1 (also known as Transforming Growth Factor Beta 1), L1CAM and several inhibitory microRNAs. Special attention deserves L1CAM gene, which provides instructions for producing the L1 cell adhesion molecule. L1 has shown adhesive roles in cell-cell interactions and associates with β1 integrins on the cell surface to induce cell surface signalling, stimulating thus cell migration and outgrowth [86].

Finally, other top-ranked gene nodes in what betweenness centrality means deserve emphasis as well (Figure 23): 1) ESR1, or estrogen receptor 1, is tightly linked to FN1 and TFAP2A (among others). In addition, it is regulated by the validated MIMAT0000242. Besides, TFAP2A shows silencing by validated MIMAT0000461. 2) FN1 cluster manifest relationships with four different integrin family members (ITGB1, ITGB3, ITGB4, ITGB8) and with VEGFB, among others. Vascular Endothelial Growth Factor B (VEGFB) has been widely studied: it is a ligand for neutropilin-1 and is associated with platelets. In addition, diseases like macular degeneration, aging and focal adhesions relate to VEGFB, which prove phenotypically relevance on the genodermatoses disorders [87].

TABLE 6. TOPOLOGICAL RESULTS (HUB GENES)

microRNA-RNA network	Topological results (hub genes)
RDEB	TIAM1 TFAP2A ITGA8 ITGB8 FN1 SDC1
KS	BTG2 HOXA10 SERPINE1 IGF2 FN1 ESR1 SOCS1 HERC3
XPC	STAT1 HERC5 ITGB1 L1CAM ESR1 FN1 TFAP2A VEGF

Aforementioned RNAs stand for relevant candidates for further research. Their individual topological parameters (computed at Cytoscape) allow to describe them as hubs for the RDEB/KS/XPC constructed networks. To do so, values for degree, betweenness centrality and shortest path distance have been considered regardless of previous biases. Network properties influence the likelihood and phenotypic consequences of disease mutations. For that reason, if a random perturbation were introduced in the microRNAs-RNAs networks, graph tolerance would be higher if hubs were not compromised. That is, a perturbation that affected the hubs would quickly spread through the network. These bipartite graphs are widely governed by the hubs, where in the majority of the cases, microRNAs control the interconnections, and only a few RNAs partake in the topological properties. The ratio of microRNA/RNA nodes is notably small, so the bipartition is not equally distributed: many RNAs link to few microRNAs, and some transcripts interconnections turn out to be PPIs. According to this system, the underexpression of microRNA is related to an overexpression of its targeted genes. The absence of a microRNA would enable the transcripts (RNAs) to express and codify the final protein [55], and viceversa: if a microRNA is up-regulated, it will largely silence its targeted transcripts, impeding the translation into proteins. This would bring huge decompensations in the normal functioning of the analysed fibroblasts for RDEB/KS/XPC patients.

Furthermore, global topological results provide with more valuable insights as well: the clustering coefficient is significantly higher than expected from random chance, the small average shortest path length and the appropriate density values allows us to conclude that our merged networks are actually real-world. Under this premise, they might experience preferential attachment, a process where “rich nodes get richer” [14], that is, microRNAs would target even more genes and hub genes, by their part, would present new PPIs. In other words, the presumable power-law distribution (Figure 2) would endorse the point made herein. High degree nodes tend to attach to low degree nodes, and this disassortativity event has been proved typical for biological networks [14]. In this way, the aforementioned genes identified as hubs would be able to reveal “guilty by association” connections to other genes. In other words, they might predict the cellular function of up-to-now unknown proteins.

In conclusion, topology helps to identify biomarkers which correspond in turn to active hotspots for further analysis.

b. Functional characterization

It is important to highlight that both initial Enrichment widgets (BiNGO & STRING) sort every term by a p-value obtained from a Hypergeometric test, that is corrected by the Benjamini and Hochberg FDR method [67]. That is, they follow a similar statistical approach. Despite following analogous algorithms, one can consider STRING being more complete since it provides more orthology information (by invoking more knowledge databases such as KEGG and InterPro).

Enriched functions from both plug-ins have been studied in tandem. Grouping by semantic similarity (ReVIGO) has sustained the project was on the right track. Enriched functions for the RDEB/KS/XPC gene lists were identified. By curating only those functions which had been enriched with genes from the three genodermatoses simultaneously (Figure 33), a definite record was reached out, including exclusively functions presumably shared by the three skin disorders together.

A final observation of the dysregulated genes in the enrichment analysis (Figure 34) was also performed. Each gene associated to an enriched function was assessed in terms of its logFC, which can hold either a positive (over-expressed) or a negative value (under-expressed). Overall, the functional profile manifests a paired fashion, where the number of up-regulated genes appear to compensate the amount of down regulation. For the majority of the enriched functions, there is a balance on gene expression. On this basis, one can argue that cells carry out compensatory mechanisms by which their resilience to outweigh abrupt dysregulatory challenges is effective. This steady state of internal conditions is called homeostasis [30], and can be hypothesized that the number of PPIs in each system help in these compensatory mechanisms by transmitting the information and regulatory processes across the network.

This constituted a turning point, from which final Cytoscape analyses could be undergone. In fact, ClueGO plug-in showed a perfect match for the desired exercise, since it is able to make use of REACTOME to concentrate on very specific terms, especially those from transcriptional control. In this way, and employing downstream enrichment analysis, REACTOME provided with hits that were successively allocated in a worksheet (Table 4 and Table 5). The hits provided by REACTOME confirmed several of the genes highlighted in the topological analysis, as well as

pointed out certain mechanisms related to the three phenotypical hallmarks of our genodermatoses.

Cell adhesions:

Elements crucial for tissue architecture are highlighted as hits in the REACTOME results. Concretely, “elastic fiber formation”, TFAP2A, Integrins & Syndecans, Neutropilin, VEGF, L1, EPHAs and TIAM1.

-Elastic fibers are bundles of proteins produced by fibroblasts and secreted in the extracellular matrix of connective tissues, conferring biomechanical elasticity and resilience to them [88]. COL7A1 and kindlin-1 fall within this class of molecules. Additionally, they are a major contributor in the degenerative changes in sun-damaged skin [89], which relates to XPC disorder.

-TFAP2A, mentioned above, exerts influence on proper body wall development and coordinates keratinocyte proliferation and differentiation. Plus, it has been related already to melanoma progression [90]

- Integrins and syndecans experience a synergistic control of cell adhesion [77], regulating cell behavior in response to the external environment too. These large complexes of cell surface receptors allow for growth factor activation and tumour suppression and progression, among others.

-Semaphorins are versatile proteins associated with axon guidance and neural system development. Besides, some classes can regulate integrins and have specific roles in immune function [91]

-Neutropilin is a protein receptor active in neurons. It is a co-receptor for semaphorins, having a combined role in the nervous system development in vertebrates [92]. As stated in the worksheet (Tables 4 & 5), neutropilin interacts with VEGF.

-VEGF, or vascular endothelial growth factor, is an important signaling protein involved in angiogenesis and immune response. VEGF has been seen to stimulate SDC1 to capture epidermal growth factor receptors [93]. Platelets has been observed to accumulate and transport VEGF in cancer patients [87]

-L1, mentioned in topological results as well, attains adhesive roles in cell-cell interactions. It provides neurons with cues in order to drive axonal growth and guidance. It is also able to associate with integrins and binds the SEMA3A receptor neutropilin-1 (partially reversing the effects of VEGF, [93])

-EPHA is closely related to TIAM1 by mediating neurite outgrowth. The best documented function of its signaling is the cell adhesion, positioning and migration. It has previously shown skin inflammation [76]

-TIAM1 deficiency protects again Ras-1 induced skin carcinogenesis. Its dysregulation causes tumor onset and progression. TIAM1 has been associated to hemidesmosomes and focal adhesions on epidermolysis bullosa and kindler syndrome [94].

Inflammatory processes:

-Chemokines are a family of small cytokines that induce directed chemotaxis, recruiting cells of the immune system and conducting them to a site of infection [30]. All of them interact with

transmembrane receptors and have different roles: from controlling lymphocytes response to promoting angiogenesis, although they are functionally divided into homeostatic and inflammatory.

-NR-MED1 coactivator complex has been related to integrin and PI3K signaling events. MED1 functions as a nuclear receptor coactivator, and its overexpression regulates the expression of proinflammatory chemokines [95].

Cancer proneness:

-NOTCH is a well-studied epidermal growth factor, having broad roles in neurogenesis, angiogenesis, cardiac functioning and bone regeneration. NOTCH signaling is known to occur inside ciliated, differentiating cells found in the first epidermal layers during early skin development [96].

-PI3Ks are a family of enzymes involved in a wide range of cellular functions, all of them resulting in cancer. It has got pathways intertwined with NOTCH. As signal transducer, it is implicated in a tumor insensitivity pathway. Its tight relation to the epidermal growth factor receptor (EGFR) has appeared in cancer studies [97].

-Interleukin, as well as chemokines, are a group of cytokines that mediate immune responses (T cell signaling). Intracellular networks describe connections where different types of interleukins activate PI3K and relates to tumor cell survival [30].

-The exocytosis of platelets plays a critical role in different aspects of the immune response (hemostasis, thrombosis, vascular remodeling and healing). In fact, platelets are related to Ras oncogenes, besides being used in skin wound healing [87].

-SCF-KIT stands for a pathway that plays a crucial role in the human melanocyte homeostasis [97]. SCF is expressed by fibroblasts, promoting proliferation, survival, migration and differentiation of melanocyte progenitors. KIT stimulation activates PI3K pathway, among others.

Some of the genes involved in the aforementioned complexes and structures can be actually ubicated within the constructed transcriptional regulatory networks (Table 7). Provided that they are contained in the graphs, their expression values are statistically different from those observed in the healthy samples. In the light of this, it can be argued that the three skin disorders might register certain phenotypic convergences under those transcriptional alterations.

Highlighted genes in Table 7 refer to nodes common to the three networks, understood as differentially expressed transcripts.

TABLE 7. FUNCTIONAL RESULTS FROM REACTOME

microRNA-RNA network	Functional results from REACTOME (corresponding genes)
RDEB	ITGA8 ITGB8 EPHA4 KIT **L1CAM** SDC1 **SEMA4G TFAP2A** TIAM1
KS	ITGB1 KIT **L1CAM SEMA4D TFAP2A** TIAM1
XPC	EPHA2 EPHA4 ITGB1 ITGB3 ITGB8 **L1CAM SEMA4C SEMA4D TFAP2A** VEGFB

c. Unifying topological & functional results

A remark needs to be taken in what topological and functional outcomes concerns. Their kind is different, so is its origin: topology comes from the purest graphical point of view, whereas functional results make use of enrichment strategies to invoke ontologies and databases from which matched biological processes, molecular functions and cellular components are retrieved. Among the Computational Biology community, there is a postulation that defends *essential* and *disease* genes show qualitative different behaviour. *Essentiality* were to be a topology-related aspect, whereas function-related anomalies within the networks were to be attributed to *disease* genes. According to this perception, disease genes would avoid dense-clustering neighborhoods unlike the essential genes, because the formers would tend to escape most vital cellular components while affecting lesser physiological processes [98]. In this manner, disease mutations could preferentially occur in non-essential genes for the network. Anyway, the approach followed in this work contemplates topology and functional analyses regardless of distinctions, where genes are considered as equally relevant whatever they were retrieved either from topological or functional analysis. Both of them deserve attention and discussion, since no previous bias was upheld.

Notwithstanding, a meeting point is identified for this particular genodermatoses study. According to both algorithms, some differentially expressed entities, here represented as specific nodes, are acknowledged as relevant within the gene lists and biological networks constructed. Concretely, L1CAM, TFAP2A are highlighted both from topology and functional enrichment analyses. By these means, and having a look at their connections at the constructed biological networks, final insights are gained (Table 8):

In RDEB network, L1CAM is underexpressed (logFC=-5.078), and regulated by the validated microRNA MIMAT0000242 (hsa-miR-129-5p) which, in turn, is overexpressed (logFC=3.604).

In KS network, L1CAM is underexpressed (logFC=-5.724), and regulated by the validated microRNA MIMAT0000242 (hsa-miR-129-5p) which, in turn, is overexpressed (logFC=3.582)

In XPC network, L1CAM is underexpressed (logFC=-7.364), and regulated by the validated microRNA MIMAT0000242 (hsa-miR-129-5p) which, in turn, is overexpressed (logFC=3.999). Other microRNAs seem to be silencing L1CAM, all of them showing overexpression (MIMAT0000449 logFC=3.763, MIMAT0000727 logFC=1.135, MIMAT0022925 logFC= 1.662, MIMAT0004494 logFC=1.419).

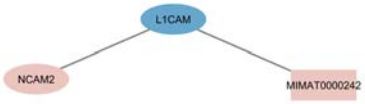
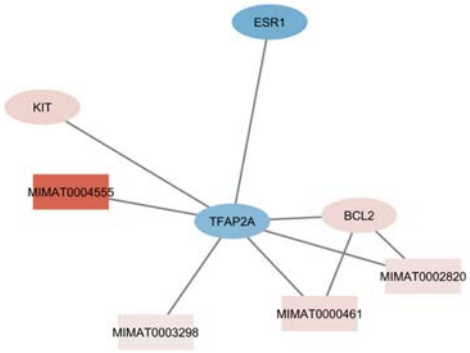
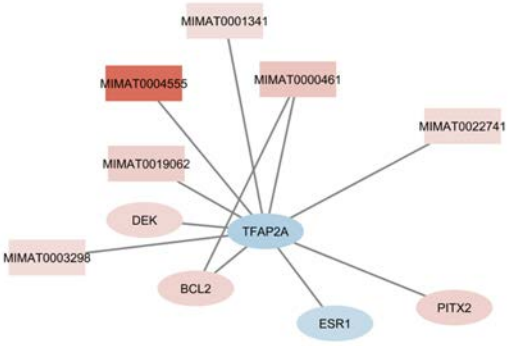
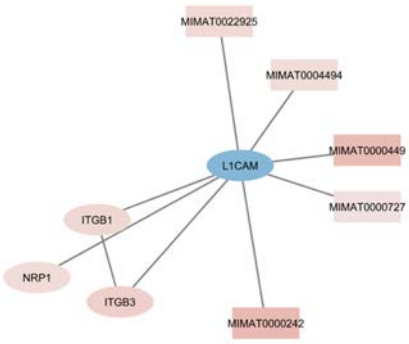
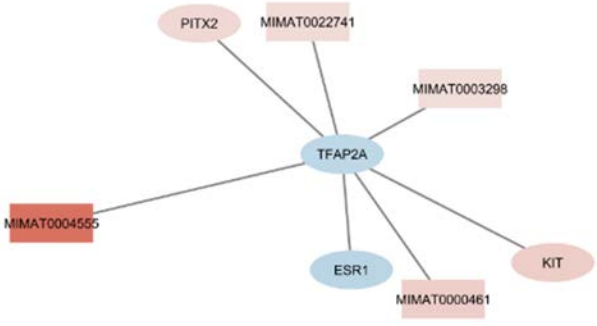
In RDEB network, TFAP2A is underexpressed (logFC=-3.635), and regulated by the validated microRNAs MIMAT0004555 (hsa-miR-10a-3p) and MIMAT000461 (hsa-miR-195-5p) (among others) which, in turn, are overexpressed (logFC=8.590 and logFC=2.555)

In KS network, TFAP2A is underexpressed (logFC=-3.41), and regulated by the validated microRNAs MIMAT0004555 (hsa-miR-10a-3p) and MIMAT000461 (hsa-miR-195-5p) (among others) which, in turn, are overexpressed (logFC=9.535 and logFC=2.936)

In XPC network, TFAP2A is underexpressed (logFC=-4.636), and regulated by the validated microRNAs MIMAT0004555 (hsa-miR-10a-3p) and MIMAT000461 (hsa-miR-195-5p) (among others) which, in turn, are overexpressed (logFC=9.483 and logFC=3.149)

It is important to recall that this whole project takes off from microRNAs. They have been the bedrock for all the actions and analyses carried out. That is why a look back was needed. Being able to come to an end by highlighting relevant microRNAs (Table 8) for further studies, is a matter of good sense.

TABLE 8. L1CAM & TFAP2A BEHAVIOUR IN THE BIOLOGICAL NETWORKS

	L1CAM	TFAP2A
RDEB		
KS		
XPC		

6. SOCIO-ECONOMIC IMPACT

Rare Disorders are considered as such due to their low prevalence in the population (less than 5 in 10.000 inhabitants). However, The World Health Organization (WHO) calculates that about 7.000 Rare Disorders affect 7% of individuals, globally. FEDER (Federación Española de Enfermedades Raras) estimates that over 3 million people currently suffer from Rare Disorders in Spain, impairing their physical, mental and behavioural faculties with different degrees of disability. These pathologies are, for the most part, chronic and degenerative. They deprive self-sufficiency and autonomy, compromising quality of life and even its expectancy in some cases (35% of Rare Disorder patient deaths occur in the first year of life [99]).

Specifically, RDEB, KS and XPC correspond to inherited skin disorders where respective mutations in proteins COL7A1, FERMT1 and XPC impede to reach a curative treatment for the moment. Gene therapy can be provided in some cases today (throughout Recombinant DNA Technologies [100]) where the genomic mutation is corrected in stem cells and infused afterwards in the patient. Yet, this experimental technique is not extended to the clinics, and therefore, discovery of biomarkers and possible drug targets for symptom alleviation is needed.

Studies like this one, represent a blueprint where system-level observations lead to new insights about the molecular mechanisms that are entangled in the shared signature of the genodermatoses. By doing so, and employing a bioinformatics approach, the scientific community can take advantage of omics data, generating information and eventually gaining perspectives on the cellular functioning.

The identification of potential candidate genes or microRNAs whose mechanisms and interactions might be involved in the common hallmarks of RDEB/KS/XPC (skin fragility, inflammation and cancer proneness) gives a prospective opportunity for the development of new drugs that could target specific structures and processes within tissues, serving as a palliative treatment for the genodermatoses and improving thus the quality of life and the patients.

On the other hand, by carrying out bioinformatic studies to approach Rare Disorders, economic costs are being reduced substantially. Narrowing-down strategies allow to rule out a huge number of molecules, validating thus only the promising ones in the lab. Bioinformatic analyses stand as an efficient tool where time-consuming research procedures can be shortened, scientific knowledge can be expanded and palliative treatments developed without directly compromising the patient's state. Besides, bioinformatic studies are not dramatically affected by the lack of funding in research, since resources like laboratory equipment and devices are not necessarily needed throughout the whole line of research, and NGS technologies are already benefiting from a standardized use and reduction in prices.

The positive bioinformatic advantages seem to be quite obvious, having a hopeful influence in health and clinical needs. However, computational algorithms require capital and upgrading investment in order to provide more accurate and refined bioinformatic results, decreasing thus the false positives rate and making the effort worthwhile. In the same way, public bioinformation infrastructure seems inadequate, so massive data storage and accessibility demands higher equity injections.

The estimated budget to carry out the thesis project is indicated in Table 9.

TABLE 9. ESTIMATED EXPENDITURE ON THE RESEARCH PROJECT

FUNGIBLES	microRNA Seq	Illumina sequencing	Sample preparation	4,758.36 €
			Sequencing	4,261.22 €
		Bioinformatics	Data handling and delivery	416.72 €
			Data analysis	2,496.94 €
	RT-qPCR		microRNA Kits	1,895.00 €
			TaqMan microRNA primers	1,652.00 €
EQUIPMENT	Bioinformatics analysis	Hardware	Computer	700.00 €
		Software	Databases, programs, computing tools, papers	100.00 €
PERSONNEL	Research staff	Research coordinator	250 hours	3,750.00 €
		Research assistant	550 hours	6,600.00 €
			TOTAL:	26,630.24 €

7. CONCLUSIONS AND FUTURE DIRECTIONS

MicroRNA-RNA networks are poorly understood today. Not only this type of graphs, but in general the analysis of integrated biological networks. Their bipartite nature entails a challenge when facing analytical approaches to extract useful information and to convey, eventually, a biological interpretation out of the omics data integration. The differences between epigenomics and transcriptomics are latent, but not exclusionary. Both records can be combined and assessed conjointly in an attempt to offer a bioinformatics solution for high-throughput data orchestration. Engaging with NGS technologies implies continuous improvements and discoveries on cellular mechanisms and medical conditions. It is all about data and its consistency. Merging strategies by repository curation and data cross-referencing ensures high standards, leading ultimately to a unified method through which scaling-up approaches could be applied to other systems and lines of research, extending its reproducibility. In this particular case, a potential way to engender trust in data and guarantee the proper value of integrative bioinformatics, metabolomics data could be incorporated: metabolic pathways described in biological literature still preserves hidden cellular knowledge. Metabolites suffer a magnified amplification of the effects occurring at the transcriptome and proteome level, so their information is very reliable, due to its closeness to the phenotype. It represents a crosstalk between Epigenetics and Metabolomics. Upon its integration, biological networks can be satisfactorily sharpened and biomarkers can be sufficiently refined to allow for drug repurposing approaches. By meeting these necessities, bioinformatic investigation might position itself in a cutting-edge Information Technology strategy to complement, and even strengthen, biomedical sciences.

In this project, a global approach to the analysis of networks was performed, providing valuable information regarding tentative biomarkers and functional mechanisms that can partially explain some of the common phenotypic hallmarks shared by three genodermatoses. This approach allows to further generate new hypothesis (that must be eventually validated) and opens new fields of research in terms of development of drugs (or drug repurposing) that could target these highlighted genes or miRNAs and therefore mitigate some of the symptoms of these diseases.

8. BIBLIOGRAPHY

[1] “What is Systems Biology?” Institute for Systems Biology. [Online] Available at: https://systemsbiology.org/about/what-is-systems-biology/

[2] Collins, F. S., Morgan, M., & Patrinos, A. (2003). The Human Genome Project: lessons from large-scale biology. Science, 300(5617), 286-290.

[3] Morozova, O., & Marra, M. A. (2008). Applications of next-generation sequencing technologies in functional genomics. Genomics, 92(5), 255-264.

[4] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., ... & DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research.

[5] Zhu, X., Gerstein, M., & Snyder, M. (2007). Getting connected: analysis and principles of biological networks. Genes & development, 21(9), 1010-1024.

[6] Mazzocchi, F. (2008). Complexity in biology: Exceeding the limits of reductionism and determinism using complexity theory. EMBO reports, 9(1), 10-14.

[7] Fang, F. C., & Casadevall, A. (2011). Reductionistic and holistic science. Infection and immunity.

[8] Engel, G. L. (1977). The need for a new medical model: a challenge for biomedicine. Science, 196(4286), 129-136.

[9] Hood, L., Heath, J. R., Phelps, M. E., & Lin, B. (2004). Systems biology and new technologies enable predictive and preventative medicine. Science, 306(5696), 640-643.

[10] Loscalzo, J., Kohane, I., & Barabasi, A. L. (2007). Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. Molecular systems biology, 3(1), 124.

[11] Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. Proceedings of the national academy of sciences, 99(12), 7821-7826.

[12] Doncheva, N. T., Assenov, Y., Domingues, F. S., & Albrecht, M. (2012). Topological analysis and interactive visualization of biological networks and protein structures. Nature protocols, 7(4), 670.

[13] Albert, R. (2005). Scale-free networks in cell biology. Journal of cell science, 118(21), 4947-4957.

[14] Wang, X. F., & Chen, G. (2003). Complex networks: small-world, scale-free and beyond. IEEE circuits and systems magazine, 3(1), 6-20.

[15] Albert, R., & Barabási, A. L. (2002). Statistical mechanics of complex networks. Reviews of modern physics, 74(1), 47.

[16] Taylor, Peter and Lewontin, Richard, "The Genotype/Phenotype Distinction", The Stanford Encyclopedia of Philosophy (Summer 2017 Edition), Edward N. Zalta (ed.), [Online] Available at: <https://plato.stanford.edu/archives/sum2017/entries/genotype-phenotype/>.

[17] Lodish H, Berk A, Zipursky SL, et al. Molecular Cell Biology. 4th edition. New York: W. H. Freeman; 2000. Section 8.1, Mutations: Types and Causes. [Online] Available at: https://www.ncbi.nlm.nih.gov/books/NBK21578/

[18] Stenson, P. D., Mort, M., Ball, E. V., Howells, K., Phillips, A. D., Thomas, N. S., & Cooper, D. N. (2009). The human gene mutation database: 2008 update. Genome medicine, 1(1), 13.

[19] Harper, J. I., & Trembath, R. C. (2004). Genetics and genodermatoses. Rook's Textbook of dermatology, 383-468.

[20] Epidermolysis Bullosa. (2017, June 6). Physiopedia, . Retrieved 10:55, November 20, 2018 [Online]Available at: https://www.physiopedia.com/index.php?title=Epidermolysis_Bullosa&oldid=174400.

[21] "Recessive dystrophic epidermolysis bullosa". Orphanet. [Online] Available at: https://www.orpha.net/consor/cgi-bin/OC_Exp.php?lng=EN&Expert=89842_as of November 2018

[22] Yan, Y., Meng, Z., Hao, S., Wang, F., Jin, X., Sun, D., ... & Ma, X. (2018). Five Novel COL7A1 Gene Mutations in Three Chinese Patients with Recessive Dystrophic Epidermolysis Bullosa. Annals of Clinical & Laboratory Science, 48(1), 100-105.

[23] "Dystrophic Epidermolysis Bullosa" U.S. National Library of Medicine. Genetics Home Reference. [Online] Available at: https://ghr.nlm.nih.gov/condition/dystrophic-epidermolysis-bullosa as of November 2018

[24] "Kindler Syndrome" U.S National Library of Medicine. Genetics Home Reference [Online] Available at: https://ghr.nlm.nih.gov/condition/kindler-syndrome#statistics as of November 2018

[25] Wang, P., Zhan, J., Song, J., Wang, Y., Fang, W., Liu, Z., & Zhang, H. (2017). Differential expression of Kindlin-1 and Kindlin-2 correlates with esophageal cancer progression and epidemiology. Science China Life Sciences, 60(11), 1214-1222.

[26]. Youssefian, L., Vahidnezhad, H., Saeidian, A. H., Ahmadizadeh, K., Has, C., & Uitto, J. (2016). Kindler syndrome, an orphan disease of cell/matrix adhesion in the skin–molecular genetics and therapeutic opportunities. Expert Opinion on Orphan Drugs, 4(8), 845-854.

[27] "Xeroderma Pigmentosum" U.S National Library of Medicine. Genetics Home Reference. [Online] Available at: https://ghr.nlm.nih.gov/condition/xeroderma-pigmentosum

[28] Kraemer, K. H., Lee, M. M., & Scotto, J. (1987). Xeroderma pigmentosum: cutaneous, ocular, and neurologic abnormalities in 830 published cases. Archives of dermatology, 123(2), 241-250.

[29] Cleaver, J. E. (1968). Defective repair replication of DNA in xeroderma pigmentosum. nature, 218(5142), 652.

[30] Coussens, L. M., & Werb, Z. (2002). Inflammation and cancer. Nature, 420(6917), 860.

[31] Schadt, E. E., Lamb, J., Yang, X., Zhu, J., Edwards, S., GuhaThakurta, D., ... & Lum, P. Y. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. Nature genetics, 37(7), 710.

[32] Ge, H., Walhout, A. J., & Vidal, M. (2003). Integrating 'omic'information: a bridge between genomics and systems biology. TRENDS in Genetics, 19(10), 551-560.

[33] Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., ... & Tegnér, J. (2014). Data integration in the era of omics: current and future challenges.

[34] Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. Genome biology, 18(1), 83.

[35] Gonçalo R, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs Ra, McVean Ga: A map of human genome variation from population-scale sequencing. Nature 2010, 467(7319):1061-73, doi:10.1038/nature09534

[36] Goncalo R, Auton A, Brooks LD, M. a, Durbin RM, Handsaker RE, McVean Ga: An integrated map of genetic variation from 1,092 human genomes. Nature 2012, 491(7422):56-65, doi:10.1038/nature11632

[37] ENCODE Project Consortium. (2004). The ENCODE (ENCyclopedia of DNA elements) project. Science, 306(5696), 636-640.

[38] Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., ... & Cancer Genome Atlas Research Network. (2013). The cancer genome atlas pan-cancer analysis project. Nature genetics, 45(10), 1113.

[39] Heng, T. S., Painter, M. W., Elpek, K., Lukacs-Kornek, V., Mauermann, N., Turley, S. J., ... & Davis, S. (2008). The Immunological Genome Project: networks of gene expression in immune cells. *Nature immunology*, *9*(10), 1091.

[40] Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., & Kim, D. (2015). Methods of integrating data to uncover genotype–phenotype interactions. Nature Reviews Genetics, 16(2), 85.

[41] Lappalainen, T., Sammeth, M., Friedländer, M. R., AC't Hoen, P., Monlong, J., Rivas, M. A., ... & Barann, M. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature, 501(7468), 506.

[42] Liolios, K., Tavernarakis, N., Hugenholtz, P. & Kyrpides, N. C. The genomes on line database (GOLD) v.2: a monitor of genome projects worldwide. Nucleic Acids Res. 34, D332–D334 (2006)

[43] Joyce, A. R., & Palsson, B. Ø. (2006). The model organism as a system: integrating'omics' data sets. Nature reviews Molecular cell biology, 7(3), 198.

[44]. J Mackenzie, R. (2018). RNA-Seq: Basics, Applications and Protocol. Technology Networks.

[45] Holliday, R. (2006). Epigenetics: a historical overview. Epigenetics, 1(2), 76-80.

[46] Aebersold, R., & Mann, M. (2003). Mass spectrometry-based proteomics. Nature, 422(6928), 198.

[47] Petersen, A. K., Zeilinger, S., Kastenmüller, G., Römisch-Margl, W., Brugger, M., Peters, A., ... & Huber, F. (2013). Epigenetics meets metabolomics: an epigenome-wide association study with blood serum metabolic traits. *Human molecular genetics*, *23*(2), 534-545.

[48] Toyoda, T., & Wada, A. (2004). Omic space: coordinate-based integration and analysis of genomic phenomic interactions. Bioinformatics, 20(11), 1759-1765.

[49] Kiechle, F. L., & Holland-Staley, C. A. (2003). Genomics, transcriptomics, proteomics, and numbers. Archives of pathology & laboratory medicine, 127(9), 1089-1097.

[50] Meng, C., Kuster, B., Culhane, A. C., & Gholami, A. M. (2014). A multivariate approach to the integration of multi-omics datasets. BMC bioinformatics, 15(1), 162.

[51] Yugi, K., Kubota, H., Hatano, A., & Kuroda, S. (2016). Trans-omics: how to reconstruct biochemical networks across multiple 'omic'layers. Trends in biotechnology, 34(4), 276-290.

[52] Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G., & Milanesi, L. (2016). Methods for the integration of multi-omics data: mathematical aspects. BMC bioinformatics, 17(2), S15.

[53] Lewis, B. P., Burge, C. B., & Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. cell, 120(1), 15-20.

[54] Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116, 281–297 (2004)

[55] "Paleoaerie". Tag Archives: MicroRNA. Image by Steve Karp [Online] Available at https://paleoaerie.org/tag/microrna/

[56] Chacón-Solano et al. Fibroblasts activation and abnormal extracellular matrix remodelling as common hallmarks in three cancer-prone genodermatoses, British Journal of Dermatology, accepted https://doi.org/10.1111/bjd.17698.

[57] World Medical Association. (2001). World Medical Association Declaration of Helsinki. Ethical principles for medical research involving human subjects. Bulletin of the World Health Organization, 79(4), 373.

[58]. Bioinformatics, B. (2011). FastQC: a quality control tool for high throughput sequence data. Cambridge, UK: Babraham Institute.

[59] Oliveros, J. C. (2016). Venny. An interactive tool for comparing lists with Venn's diagrams. 2007–2015.

[60] Santos, A. (2018). Bioinformatic interpretation of microRNA role in three phenotypically related genodermatoses. Final Degree Project

[61] Vlachos, I. S., Zagganas, K., Paraskevopoulou, M. D., Georgakilas, G., Karagkouni, D., Vergoulis, T., ... & Hatzigeorgiou, A. G. (2015). DIANA-miRPath v3. 0: deciphering microRNA function with experimental support. Nucleic acids research, 43(W1), W460-W466.

[62] Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research, 28(1), 27-30.

[63] Oever, M. V., Muldoon, D., Mathews, W., McElmurry, R., & Tolar, J. (2016). miR-29 Regulates Type VII Collagen in Recessive Dystrophic Epidermolysis Bullosa. The Journal of investigative dermatology, 136(10), 2013.

[64] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., ... & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research, 13(11), 2498-2504.

[65] Snel, B., Lehmann, G., Bork, P., & Huynen, M. A. (2000). STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic acids research, 28(18), 3442-3444.

[66] Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. Proceedings of the national academy of sciences, 99(12), 7821-7826.

[67] Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2008). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic acids research, 37(1), 1-13.

[68] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., ... & Harris, M. A. (2000). Gene Ontology: tool for the unification of biology. Nature genetics, 25(1), 25.

[69] Maere, S., Heymans, K., & Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics, 21(16), 3448-3449.

[70] Sharan, R., Ulitsky, I., & Shamir, R. (2007). Network-based prediction of protein function. Molecular systems biology, 3(1), 88.

[71] Supek, F., Bošnjak, M., Škunca, N., & Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. PloS one, 6(7), e21800.

[72] Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., ... & Galon, J. (2009). ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics, 25(8), 1091-1093.

[73] Fan, Y., Siklenka, K., Arora, S. K., Ribeiro, P., Kimmins, S., & Xia, J. (2016). miRNet-dissecting miRNA-target interactions and functional associations through network-based visual analysis. Nucleic acids research, 44(W1), W135-W141.

[74] Backes, C., Fehlmann, T., Kern, F., Kehl, T., Lenhof, H. P., Meese, E., & Keller, A. (2017). miRCarta: a central repository for collecting miRNA candidates. Nucleic acids research, 46(D1), D160-D167.

[75] Lund, A. H. (2010). miR-10 in development and cancer. Cell death and differentiation, 17(2), 209.

[76] Habets, G. G., Scholtes, E. H., Zuydgeest, D., van der Kammen, R. A., Stam, J. C., Berns, A., & Collard, J. G. (1994). Identification of an invasion-inducing gene, Tiam-1, that encodes a protein with homology to GDP-GTP exchangers for Rho-like proteins. Cell, 77(4), 537-549.

[77] Woods, A., Longley, R. L., Tumova, S., & Couchman, J. R. (2000). Syndecan-4 binding to the high affinity heparin-binding domain of fibronectin drives focal adhesion formation in fibroblasts. Archives of biochemistry and biophysics, 374(1), 66-72.

[78] Duriez, C., Falette, N., Audoynaud, C., Moyret-Lalle, C., Bensaad, K., Courtois, S., ... & Puisieux, A. (2002). The human BTG2/TIS21/PC3 gene: genomic structure, transcriptional regulation and evaluation as a candidate tumor suppressor gene. *Gene*, *282*(1), 207-214.

[79] Magnusson, M., Brun, A. C., Miyake, N., Larsson, J., Ehinger, M., Bjornsson, J. M., ... & Karlsson, S. (2007). HOXA10 is a critical regulator for hematopoietic stem cells and erythroid/megakaryocyte development. *Blood*, *109*(9), 3687-3696.

[80] Martin-Trujillo, A., van Rietschoten, J. G., Timmer, T. C., Rodríguez, F. M., Huizinga, T. W., Tak, P. P., ... & Verweij, C. L. (2010). Loss of imprinting of IGF2 characterises high IGF2 mRNA-expressing type of fibroblast-like synoviocytes in rheumatoid arthritis. *Annals of the rheumatic diseases*, *69*(6), 1239-1242.

[81] Madonna, S., Scarponi, C., Doti, N., Carbone, T., Cavani, A., Scognamiglio, P. L., ... & Albanesi, C. (2013). Therapeutical potential of a peptide mimicking the SOCS 1 kinase inhibitory region in skin immune responses. *European journal of immunology*, *43*(7), 1883-1895.

[82] Yoo, N. J., Park, S. W., & Lee, S. H. (2011). Frameshift mutations of ubiquitination-related genes HERC2, HERC3, TRIP12, UBE2Q1 and UBE4B in gastric and colorectal carcinomas with microsatellite instability. *Pathology-Journal of the RCPA*, *43*(7), 753-755.

[83] Darnell, J. E., Kerr, I. M., & Stark, G. R. (1994). Jak-STAT pathways and transcriptional activation in response to IFNs and other extracellular signaling proteins. *Science*, *264*(5164), 1415-1421.

[84] Shi, H. X., Yang, K., Liu, X., Liu, X. Y., Wei, B., Shan, Y. F., ... & Wang, C. (2010). Positive regulation of interferon regulatory factor 3 activation by Herc5 via ISG15 modification. *Molecular and cellular biology*, *30*(10), 2424-2436.

[85] Wilkinson, D. M., & Huberman, B. A. (2004). A method for finding communities of related genes. *proceedings of the national Academy of sciences*, *101*(suppl 1), 5241-5248.

[86] Chan, Y. M., Yu, Q. C., LeBlanc-Straceski, J., Christiano, A., Pulkkinen, L., Kucherlapati, R. S., ... & Fuchs, E. (1994). Mutations in the non-helical linker segment L1-2 of keratin 5 in patients with Weber-Cockayne epidermolysis bullosa simplex. Journal of cell science, 107(4), 765-774.

[87] Golebiewska, E. M., & Poole, A. W. (2014). Secrets of platelet exocytosis–what do we really know about platelet secretion mechanisms?. *British journal of haematology*, *165*(2), 204-216.

[88] Greenlee, T. K., Ross, R., & Hartman, J. L. (1966). The fine structure of elastic fibers. *The Journal of cell biology*, *30*(1), 59-71.

[89] Chen, V. L., Fleischmajer, R., Schwartz, E., Palaia, M., & Timpl, R. (1986). Immunochemistry of elastotic material in sun-damaged skin. *Journal of investigative dermatology*, *87*(3), 334-337.

[90] Hallberg, A. R., Vorrink, S. U., Hudachek, D. R., Cramer-Morales, K., Milhem, M. M., Cornell, R. A., & Domann, F. E. (2014). Aberrant CpG methylation of the TFAP2A gene constitutes a mechanism for loss of TFAP2A expression in human metastatic melanoma. *Epigenetics*, *9*(12), 1641-1647.

[91] Mack, M., Wendelschafer-Crabb, G., McAdams, B., Hordinsky, M., Kennedy, W., & Tolar, J. (2015). Peripheral neuro-immune pathology in recessive dystrophic epidermolysis bullosa. *The Journal of investigative dermatology*, *135*(4), 1193.

[92] Gu, C., Rodriguez, E. R., Reimert, D. V., Shu, T., Fritzsch, B., Richards, L. J., ... & Ginty, D. D. (2003). Neuropilin-1 conveys semaphorin and VEGF signaling during neural and cardiovascular development. *Developmental cell*, *5*(1), 45-57.

[93] Rapraeger, A. C., Ell, B. J., Roy, M., Li, X., Morrison, O. R., Thomas, G. M., & Beauvais, D. M. (2013). Vascular endothelial-cadherin stimulates syndecan-1-coupled insulin-like growth factor-1 receptor and cross-talk between αVβ3 integrin and vascular endothelial growth factor receptor 2 at the onset of endothelial cell dissemination during angiogenesis. *The FEBS journal*, *280*(10), 2194-2206.

[94] Tsuruta, D., Hashimoto, T., Hamill, K. J., & Jones, J. C. (2011). Hemidesmosomes and focal contact proteins: functions and cross-talk in keratinocytes, bullous diseases and wound healing. *Journal of dermatological science*, *62*(1), 1-7.

[95] Stumpf, M., Waskow, C., Krötschel, M., van Essen, D., Rodriguez, P., Zhang, X., ... & Borggrefe, T. (2006). The mediator complex functions as a coactivator for GATA-1 in erythropoiesis via subunit Med1/TRAP220. *Proceedings of the National Academy of Sciences*, *103*(49), 18504-18509.

[96] Gutierrez, A., & Look, A. T. (2007). NOTCH and PI3K-AKT pathways intertwined. *Cancer cell*, *12*(5), 411-413.

[97] Grichnik, J. M., Burch, J. A., Burchette, J., & Shea, C. R. (1998). The SCF/KIT pathway plays a critical role in the control of normal human melanocyte homeostasis. *Journal of Investigative Dermatology*, *111*(2), 233-238.

[98] Feldman, I., Rzhetsky, A., & Vitkup, D. (2008). Network properties of genes harboring inherited disease mutations. *Proceedings of the National Academy of Sciences*, *105*(11), 4323-4328.

[99]. Información General Sobre Enfermedades Raras. Acta Sanitaria. [Online] Available at: https://www.actasanitaria.com/wp-content/uploads/2015/10/informacion-general-sobre-enfermedades-raras-.pdf

[100]. Micklos, D. A., Freyer, G. A., & Lauter, S. Z. (1990). DNA science: A first course in recombinant DNA technology (pp. 256-257). Carolina Biological Supply Company.

ANNEX

TABLE WITH THE 27 SIGNIFICANT MICRORNAS IN RDEB VS HEALTHY

Gene_id	logFC	PValue	FDR
hsa-miR-10a-5p	8.824	5.040E-13	6.310E-10
hsa-miR-10a-3p	8.590	1.107E-12	6.931E-10
hsa-miR-556-5p	-4.536	1.383E-11	5.772E-09
hsa-miR-6507-5p	-4.779	5.103E-08	1.597E-05
hsa-miR-6842-3p	-2.425	7.527E-08	1.885E-05
hsa-miR-195-3p	2.902	3.217E-07	6.712E-05
hsa-miR-556-3p	-3.859	7.267E-07	1.300E-04
hsa-miR-129-5p	3.604	1.863E-06	2.915E-04
hsa-miR-129-2-3p	2.975	4.151E-06	5.774E-04
hsa-miR-29b-2-5p	-2.150	7.285E-06	8.440E-04
hsa-miR-10b-5p	4.379	7.415E-06	8.440E-04
hsa-miR-29c-5p	-1.901	1.737E-05	1.813E-03
hsa-miR-29c-3p	-2.172	2.798E-05	2.694E-03
hsa-miR-10b-3p	4.185	3.675E-05	3.286E-03
hsa-miR-1295a	-3.015	6.489E-05	5.416E-03
hsa-miR-4488	-2.192	7.257E-05	5.679E-03
hsa-miR-615-3p	3.925	8.503E-05	6.262E-03
hsa-miR-148a-5p	2.112	1.078E-04	7.499E-03
hsa-miR-146b-5p	-1.976	1.146E-04	7.552E-03
hsa-miR-615-5p	4.028	2.479E-04	1.531E-02
hsa-miR-874-5p	-1.612	2.686E-04	1.531E-02
hsa-miR-497-5p	2.303	2.690E-04	1.531E-02
hsa-miR-195-5p	2.555	5.443E-04	2.931E-02
hsa-miR-935	-2.624	5.618E-04	2.931E-02
hsa-miR-629-3p	1.973	6.219E-04	3.011E-02
hsa-miR-148a-3p	2.069	6.253E-04	3.011E-02
hsa-miR-1468-5p	-2.008	6.669E-04	3.092E-02

TABLE WITH THE 99 SIGNIFICANT MICRORNAS IN KS VS HEALTHY

Gene_id	logFC	PValue	FDR
hsa-miR-1291	5.774	1.049E-17	1.011E-14
hsa-miR-10a-5p	10.496	2.217E-17	1.069E-14
hsa-miR-1260a	5.360	1.503E-15	4.830E-13
hsa-miR-10a-3p	9.535	3.761E-13	9.064E-11
hsa-miR-483-3p	-6.817	6.627E-13	1.278E-10
hsa-miR-483-5p	-7.343	1.383E-12	2.221E-10
hsa-miR-556-5p	-6.282	4.896E-12	6.742E-10
hsa-miR-3607-5p	4.149	8.124E-11	9.789E-09
hsa-miR-1295a	-4.676	2.420E-09	2.593E-07
hsa-miR-1468-5p	-3.589	3.161E-08	3.047E-06
hsa-miR-129-5p	3.582	4.865E-08	4.263E-06
hsa-miR-129-2-3p	3.237	1.004E-07	7.597E-06
hsa-miR-1247-5p	-5.513	1.025E-07	7.597E-06
hsa-miR-3605-3p	-2.427	1.662E-07	1.144E-05
hsa-miR-7974	2.918	3.096E-07	1.989E-05
hsa-miR-6842-3p	-2.770	6.803E-07	3.773E-05
hsa-miR-10b-5p	5.137	7.019E-07	3.773E-05
hsa-miR-29c-3p	-2.119	7.044E-07	3.773E-05
hsa-miR-148a-3p	2.276	1.562E-06	7.927E-05
hsa-miR-3615	-2.124	4.902E-06	2.363E-04
hsa-miR-137	2.503	5.442E-06	2.498E-04
hsa-miR-195-5p	2.936	6.427E-06	2.816E-04
hsa-miR-6507-5p	-5.372	7.726E-06	3.108E-04
hsa-miR-3184-3p	-1.852	7.739E-06	3.108E-04
hsa-miR-31-3p	1.948	1.090E-05	4.202E-04
hsa-miR-561-5p	-1.929	1.152E-05	4.271E-04
hsa-miR-3529-3p	2.033	1.671E-05	5.965E-04
hsa-miR-4767	2.362	1.929E-05	6.641E-04
hsa-miR-10b-3p	4.932	2.146E-05	7.088E-04
hsa-miR-4517	2.388	2.206E-05	7.088E-04
hsa-miR-4653-5p	3.374	2.941E-05	9.145E-04
hsa-miR-221-5p	1.722	4.374E-05	1.318E-03
hsa-miR-29b-2-5p	-2.095	5.826E-05	1.702E-03
hsa-miR-877-5p	-1.821	6.856E-05	1.944E-03
hsa-miR-142-5p	3.386	8.156E-05	2.162E-03
hsa-miR-149-5p	-1.864	8.172E-05	2.162E-03
hsa-miR-5701	-6.849	8.297E-05	2.162E-03
hsa-miR-155-3p	3.482	1.150E-04	2.918E-03
hsa-miR-142-3p	3.617	1.246E-04	3.081E-03
hsa-miR-181a-5p	-1.970	1.466E-04	3.533E-03
hsa-let-7i-3p	1.954	1.542E-04	3.625E-03
hsa-miR-548ab	-1.939	2.066E-04	4.735E-03

hsa-miR-133a-3p	-3.795	2.112E-04	4.735E-03
hsa-miR-328-3p	-1.558	2.410E-04	5.281E-03
hsa-miR-128-3p	-1.500	2.565E-04	5.495E-03
hsa-miR-4521	2.209	2.626E-04	5.504E-03
hsa-miR-624-5p	2.134	2.851E-04	5.847E-03
hsa-miR-1305	2.312	3.011E-04	6.047E-03
hsa-miR-100-3p	1.602	3.220E-04	6.336E-03
hsa-miR-629-3p	1.893	3.352E-04	6.463E-03
hsa-miR-98-3p	-1.439	3.613E-04	6.830E-03
hsa-miR-6784-3p	-2.060	4.059E-04	7.525E-03
hsa-miR-1247-3p	-6.422	4.445E-04	8.085E-03
hsa-miR-148a-5p	2.012	4.925E-04	8.792E-03
hsa-miR-887-3p	-1.571	5.145E-04	9.019E-03
hsa-miR-454-5p	-1.491	5.389E-04	9.276E-03
hsa-miR-326	-1.541	5.721E-04	9.564E-03
hsa-miR-556-3p	-3.428	5.754E-04	9.564E-03
hsa-miR-7-5p	2.166	6.960E-04	1.137E-02
hsa-miR-145-5p	1.770	7.213E-04	1.159E-02
hsa-miR-6868-3p	-2.724	7.424E-04	1.173E-02
hsa-miR-6737-3p	-1.711	7.964E-04	1.238E-02
hsa-miR-6769b-3p	-1.581	8.849E-04	1.354E-02
hsa-miR-19b-1-5p	1.875	9.072E-04	1.367E-02
hsa-miR-496	1.472	9.418E-04	1.397E-02
hsa-miR-195-3p	2.011	1.012E-03	1.478E-02
hsa-miR-1972	1.740	1.148E-03	1.652E-02
hsa-miR-135b-5p	2.387	1.179E-03	1.671E-02
hsa-miR-1268a	1.630	1.202E-03	1.679E-02
hsa-miR-6716-3p	-1.554	1.219E-03	1.679E-02
hsa-miR-144-5p	-2.106	1.414E-03	1.910E-02
hsa-miR-3677-3p	1.400	1.426E-03	1.910E-02
hsa-miR-615-3p	3.246	1.502E-03	1.983E-02
hsa-miR-423-3p	-1.340	1.523E-03	1.984E-02
hsa-miR-1910-5p	-1.637	1.549E-03	1.991E-02
hsa-miR-5683	3.181	1.585E-03	2.011E-02
hsa-miR-605-5p	1.655	1.665E-03	2.085E-02
hsa-miR-3152-5p	2.936	1.835E-03	2.242E-02
hsa-miR-196b-5p	-1.531	1.842E-03	2.242E-02
hsa-miR-573	2.556	1.861E-03	2.242E-02
hsa-miR-130b-5p	-1.310	2.000E-03	2.380E-02
hsa-miR-4803	2.149	2.079E-03	2.444E-02
hsa-miR-2116-3p	-1.470	2.124E-03	2.467E-02
hsa-miR-33b-5p	-1.445	2.184E-03	2.506E-02
hsa-miR-668-3p	-1.428	2.447E-03	2.775E-02
hsa-miR-6750-3p	-2.200	2.513E-03	2.817E-02
hsa-miR-3152-3p	2.117	2.664E-03	2.952E-02

hsa-miR-4731-3p	2.260	2.792E-03	3.059E-02
hsa-miR-423-5p	-1.194	3.026E-03	3.278E-02
hsa-miR-106b-3p	-1.241	3.327E-03	3.531E-02
hsa-miR-153-3p	-2.482	3.350E-03	3.531E-02
hsa-miR-34c-3p	1.376	3.369E-03	3.531E-02
hsa-miR-365a-3p	1.176	3.514E-03	3.642E-02
hsa-miR-6824-3p	-1.627	3.600E-03	3.691E-02
hsa-miR-29c-5p	-1.312	3.797E-03	3.853E-02
hsa-miR-3157-5p	2.022	3.860E-03	3.876E-02
hsa-miR-656-5p	1.817	4.191E-03	4.165E-02
hsa-miR-3613-3p	1.315	4.278E-03	4.208E-02
hsa-miR-1254	-1.344	4.758E-03	4.633E-02

TABLE WITH THE 148 SIGNIFICANT MICRORNAS IN XPC VS HEALTHY

Gene_id	logFC	PValue	FDR
hsa-miR-10a-5p	10.442	4.086E-74	3.641E-71
hsa-miR-10a-3p	9.483	1.155E-44	5.144E-42
hsa-miR-1260a	7.239	1.617E-23	4.803E-21
hsa-miR-6087	6.211	2.032E-14	4.526E-12
hsa-miR-129-5p	3.999	2.297E-13	4.093E-11
hsa-miR-129-2-3p	3.639	4.422E-13	6.567E-11
hsa-miR-195-5p	3.149	2.842E-10	3.618E-08
hsa-miR-1295a	-5.668	2.501E-09	2.785E-07
hsa-miR-1291	5.148	1.005E-08	9.951E-07
hsa-miR-6842-3p	-3.317	3.823E-08	3.186E-06
hsa-miR-10b-5p	5.152	3.934E-08	3.186E-06
hsa-miR-155-3p	4.731	5.253E-08	3.900E-06
hsa-miR-1468-5p	-3.251	1.197E-07	8.207E-06
hsa-miR-137	2.870	2.004E-07	1.266E-05
hsa-let-7i-3p	2.504	2.300E-07	1.266E-05
hsa-miR-5683	5.262	2.316E-07	1.266E-05
hsa-miR-556-5p	-4.168	2.493E-07	1.266E-05
hsa-miR-146a-5p	3.763	2.557E-07	1.266E-05
hsa-miR-3607-5p	3.859	6.783E-07	3.181E-05
hsa-miR-148a-3p	2.188	8.616E-07	3.838E-05
hsa-miR-147b	3.088	1.132E-06	4.802E-05
hsa-miR-548ab	-2.636	1.697E-06	6.873E-05
hsa-miR-877-5p	-2.057	3.409E-06	1.321E-04
hsa-miR-4454	2.278	4.725E-06	1.754E-04
hsa-miR-10b-3p	4.989	4.996E-06	1.781E-04
hsa-miR-100-3p	1.974	5.690E-06	1.950E-04
hsa-miR-195-3p	2.365	6.012E-06	1.984E-04
hsa-miR-3605-3p	-1.940	1.082E-05	3.442E-04
hsa-miR-3117-3p	2.637	1.236E-05	3.798E-04
hsa-miR-19b-1-5p	2.456	1.423E-05	4.221E-04
hsa-miR-142-5p	2.802	1.468E-05	4.221E-04
hsa-miR-483-5p	-4.169	2.110E-05	5.876E-04
hsa-miR-1247-5p	-4.264	2.401E-05	6.481E-04
hsa-miR-7974	2.170	2.798E-05	7.331E-04
hsa-miR-146a-3p	6.520	3.530E-05	8.987E-04
hsa-miR-15a-3p	2.503	3.795E-05	9.392E-04
hsa-miR-3615	-1.724	3.980E-05	9.585E-04
hsa-miR-4524a-5p	2.469	4.376E-05	1.026E-03
hsa-miR-3184-3p	-1.577	4.560E-05	1.042E-03
hsa-miR-129-1-3p	5.018	4.733E-05	1.054E-03
hsa-miR-3614-5p	4.423	5.032E-05	1.078E-03
hsa-miR-155-5p	1.753	5.080E-05	1.078E-03

hsa-miR-4488	-3.355	5.820E-05	1.206E-03
hsa-miR-4473	-2.102	6.801E-05	1.377E-03
hsa-miR-891a-5p	7.110	7.145E-05	1.415E-03
hsa-miR-183-5p	3.802	7.650E-05	1.482E-03
hsa-miR-6507-5p	-6.470	8.237E-05	1.561E-03
hsa-miR-624-5p	2.423	9.000E-05	1.671E-03
hsa-miR-561-5p	-1.840	9.521E-05	1.731E-03
hsa-miR-4517	2.201	1.018E-04	1.814E-03
hsa-miR-196b-5p	-1.569	1.115E-04	1.948E-03
hsa-miR-142-3p	2.619	1.261E-04	2.160E-03
hsa-miR-454-5p	-1.637	1.286E-04	2.162E-03
hsa-miR-1908-3p	-2.026	1.415E-04	2.334E-03
hsa-miR-668-3p	-1.885	1.965E-04	3.183E-03
hsa-miR-615-3p	3.474	2.180E-04	3.468E-03
hsa-miR-3529-3p	1.677	2.483E-04	3.882E-03
hsa-miR-615-5p	3.663	2.771E-04	4.257E-03
hsa-miR-21-3p	1.419	3.466E-04	5.234E-03
hsa-miR-4773	1.970	3.713E-04	5.513E-03
hsa-miR-6737-3p	-2.056	4.175E-04	6.098E-03
hsa-miR-98-3p	-1.477	4.295E-04	6.173E-03
hsa-miR-146b-3p	-2.345	4.685E-04	6.625E-03
hsa-miR-874-5p	-1.858	4.786E-04	6.654E-03
hsa-miR-326	-1.874	4.891E-04	6.654E-03
hsa-miR-31-3p	1.421	4.929E-04	6.654E-03
hsa-miR-34c-3p	1.744	5.020E-04	6.676E-03
hsa-miR-20a-5p	1.370	5.965E-04	7.816E-03
hsa-miR-487b-3p	-1.321	6.156E-04	7.933E-03
hsa-miR-486-3p	-1.808	6.257E-04	7.933E-03
hsa-miR-887-3p	-1.518	6.321E-04	7.933E-03
hsa-miR-3613-3p	1.528	7.211E-04	8.924E-03
hsa-miR-1268a	1.681	7.725E-04	9.428E-03
hsa-miR-149-5p	-1.478	8.263E-04	9.949E-03
hsa-miR-19a-5p	2.202	9.320E-04	1.107E-02
hsa-miR-549a	-1.617	1.017E-03	1.192E-02
hsa-miR-431-3p	-1.369	1.090E-03	1.261E-02
hsa-miR-4767	2.370	1.157E-03	1.322E-02
hsa-miR-6868-3p	-2.539	1.183E-03	1.334E-02
hsa-miR-29b-2-5p	-1.709	1.226E-03	1.358E-02
hsa-miR-3613-5p	1.427	1.235E-03	1.358E-02
hsa-miR-1299	2.329	1.286E-03	1.398E-02
hsa-miR-6747-3p	-2.775	1.305E-03	1.401E-02
hsa-miR-328-3p	-1.237	1.607E-03	1.704E-02
hsa-miR-16-1-3p	1.630	1.719E-03	1.802E-02
hsa-miR-424-5p	1.460	1.743E-03	1.806E-02
hsa-miR-1233-3p	-3.758	1.870E-03	1.915E-02

hsa-miR-2116-3p	-1.627	1.915E-03	1.939E-02
hsa-miR-128-3p	-1.219	1.999E-03	2.001E-02
hsa-miR-585-5p	1.921	2.038E-03	2.018E-02
hsa-miR-944	6.216	2.108E-03	2.062E-02
hsa-miR-483-3p	-2.997	2.129E-03	2.062E-02
hsa-miR-5701	-4.472	2.164E-03	2.074E-02
hsa-miR-548h-3p	1.768	2.236E-03	2.120E-02
hsa-miR-3194-5p	3.070	2.352E-03	2.195E-02
hsa-miR-503-3p	1.662	2.365E-03	2.195E-02
hsa-miR-584-5p	-1.178	2.488E-03	2.263E-02
hsa-miR-1270	1.620	2.498E-03	2.263E-02
hsa-miR-4524a-3p	2.686	2.515E-03	2.263E-02
hsa-miR-6787-3p	-2.832	2.566E-03	2.287E-02
hsa-miR-548t-3p	1.652	2.616E-03	2.308E-02
hsa-miR-181a-5p	-1.479	2.711E-03	2.368E-02
hsa-miR-486-5p	-1.578	2.776E-03	2.399E-02
hsa-miR-1296-3p	1.947	2.800E-03	2.399E-02
hsa-miR-744-5p	-1.171	2.866E-03	2.420E-02
hsa-miR-431-5p	-1.139	2.920E-03	2.420E-02
hsa-miR-133a-3p	-2.232	2.928E-03	2.420E-02
hsa-miR-423-3p	-1.118	2.933E-03	2.420E-02
hsa-miR-548z	1.907	3.125E-03	2.554E-02
hsa-miR-450a-1-3p	1.921	3.414E-03	2.765E-02
hsa-miR-629-3p	1.511	3.446E-03	2.766E-02
hsa-miR-485-5p	-1.145	3.573E-03	2.842E-02
hsa-miR-1273g-3p	1.661	3.631E-03	2.863E-02
hsa-miR-598-3p	-1.567	3.663E-03	2.863E-02
hsa-miR-7641	1.958	3.854E-03	2.986E-02
hsa-miR-543	-1.128	4.052E-03	3.112E-02
hsa-miR-181b-5p	-1.385	4.124E-03	3.140E-02
hsa-miR-6720-3p	-1.790	4.215E-03	3.183E-02
hsa-miR-7706	-1.129	4.336E-03	3.237E-02
hsa-miR-1252-5p	-3.811	4.379E-03	3.237E-02
hsa-miR-140-5p	1.180	4.396E-03	3.237E-02
hsa-miR-221-5p	1.091	4.487E-03	3.277E-02
hsa-miR-548am-5p	1.678	4.563E-03	3.306E-02
hsa-miR-34b-5p	1.315	4.635E-03	3.331E-02
hsa-miR-148a-5p	1.530	5.180E-03	3.692E-02
hsa-miR-27a-5p	1.102	5.315E-03	3.758E-02
hsa-miR-708-3p	1.873	5.375E-03	3.771E-02
hsa-miR-3607-3p	-1.284	5.485E-03	3.818E-02
hsa-miR-374a-5p	1.135	5.543E-03	3.829E-02
hsa-let-7d-5p	-1.154	5.708E-03	3.891E-02
hsa-miR-4707-3p	-2.406	5.743E-03	3.891E-02
hsa-miR-190a-5p	1.155	5.764E-03	3.891E-02

hsa-miR-1254	-1.364	6.095E-03	4.060E-02
hsa-miR-1972	1.512	6.106E-03	4.060E-02
hsa-miR-4521	-1.795	6.193E-03	4.087E-02
hsa-miR-432-5p	-1.086	6.298E-03	4.126E-02
hsa-miR-18a-5p	1.135	6.548E-03	4.259E-02
hsa-miR-30c-2-3p	-1.086	7.089E-03	4.554E-02
hsa-miR-106a-5p	1.507	7.105E-03	4.554E-02
hsa-miR-655-5p	-1.560	7.289E-03	4.639E-02
hsa-miR-145-5p	1.205	7.524E-03	4.752E-02
hsa-miR-1305	1.537	7.611E-03	4.752E-02
hsa-miR-874-3p	-1.231	7.627E-03	4.752E-02
hsa-miR-374a-3p	1.046	7.849E-03	4.799E-02
hsa-miR-3120-3p	1.233	7.853E-03	4.799E-02
hsa-miR-3157-5p	1.941	7.863E-03	4.799E-02
hsa-miR-675-5p	4.852	8.154E-03	4.942E-02
hsa-miR-1180-3p	-1.086	8.253E-03	4.968E-02