Bioinformatics Program

Technical University of Munich

Ludwig-Maximilians-Universität München

Master's Thesis in Bioinformatics

Patient-specific dysregulation analysis based on gene regulatory networks in cancer

Sebastian Dötsch

Bioinformatics Program

Technical University of Munich

Ludwig-Maximilians-Universität München

Master's Thesis in Bioinformatics

Patient-specific dysregulation analysis based on gene regulatory networks in cancer

Patientenspezifische Dysregulationsanalyse bei Krebs basierend auf genregulatorischen Netzwerken

Author: Sebastian Dötsch
Supervisor: Dr. Markus List Chair of Experimental
Bioinformatics,
TUM School of Life Sciences Technische Universität
München
Dr. Josch Pauling LipiTUM, Chair of Experimental
Bioinformatics,
TUM School of Life Sciences Technische Universität
München
Advisor: Alexander Dietrich, Technische Universität München
Submitted: 15.09.2023

Declaration of Originality

I confirm that this master's thesis is my own work and I have documented all sources and material used.

________________ ________________

Ort, Datum Unterschrift

Contents

Abstract

Introduction

1.1 Epigenetics

Epigenetics, an emerging field in molecular biology, gives insights into heritable gene regulation. It involves dynamic changes in DNA, chromatin, and noncoding RNAs. Histone tail modifications, including acetylation and DNA methylation are examples of mechanisms that cause these changes. They control chromatin structure and influence gene accessibility[1]. Noncoding RNAs such as microRNAs on the other hand fine-tune gene expression post-transcriptionally. Epigenetics thus offers insights into development, disease, and therapeutic opportunities[2]. The dynamic interplay between genetic and epigenetic mechanisms opens new approaches for our understanding of biological complexity and is therefore investigated.

1.1.1 DNA Methylation

DNA methylation, an important epigenetic modification, involves the addition of methyl groups to cytosine or adenine bases in DNA molecules. It typically occurs on the first mentioned by addition of the methyl group to the carbon-5 atom of cytosin. This reaction is catalyzed by the three DNA methyltranferases DNMT3A, DNMT3B and DNMT1. DNMT3A and DNMT3B are essential for de novo methylation, for example during embryogenesis and germ cell development [3]. DNMT1 on the other hand is responsible for maintaining methylation patterns and does this by catalyzing the addition of methyl groups to cytosine nucleotides [4]. Therefore this chemical alteration is involved in cellular development and maintenance of genome integrity. It regulates gene expression by inhibiting the binding of transcription factors (TFs) to DNA or recruiting repressor genes [5]. It additionally silences a large number of parasitic transposable, retroviral, repeat elements, which the mammalian genome gathered over time [6].

DNA methylation occurs very frequently at CpG sites (CpGs), 70% to 80% of CpG cytosines are methylated in mammals[7]. CpGs are regions in the DNA where a cytosine nucleotide is followed by a guanine nucleotide, if they occur with a atypically high frequency in a sequence it is called a CpG island[8]. These sites hold a central role in gene expression regulation, acting as epigenetic switches when located in promoter regions of genes[9]. Beta values, a critical metric in DNA methylation studies, quantify the degree of DNA methylation at CpG sites. By taking the methylated ($M{>}0$) and unmethylated ($U{>}0$) signal intensities, measured by the Illumina 450k array for instance, the beta value $b{=}M/(M{+}U{+}a)$ is calculated. To stabilize beta values when M and U are small, the offset $a \geq 0$ usually equal to 100 is added to $M+U$. Ranging from 0, completely unmethylated, to 1, completely methylated, beta values provide a quantitative measure of DNA methylation levels. They are instrumental in understanding epigenetic modifications associated with diverse biological processes, from normal development to diseases like cancer[10].

Analyzing beta values enables researchers to identify differentially methylated CpG sites, crucial for biomarker discovery and understanding disease mechanisms. This quantitative approach to epigenetics grants insights into the dynamic nature of DNA methylation, understanding their role in health and disease.

1.1.2 Regulatory Elements

Gene regulation is a finely-tuned process in which several molecular components play crucial roles. Cis and trans gene regulatory elements (REMs) are pivotal components of this process. REMs act as transcription factor binding sites (TFBs), by binding to TFs. Cis elements, non-coding DNA regions, are located proximal to specific genes and encompass promoters and enhancers. Promoters initiate gene transcription, while enhancers enhance gene expression. Cis elements often represent binding motifs for trans acting factors[11]. Trans elements are represented by TFs, regulating the expression of distant genes by precisely binding to cis elements. In most cases such complex interactions between cis regulatory elements and trans acting factors regulate gene expression. TFs are responsible for regulating the timing and location of gene transcription[12]. They serve as molecular switches, exerting control over gene activation or repression. TFs play indispensable roles in cellular processes, including differentiation and response to external stimuli[11].

1.2 EpiRegioDB

The EpiRegio[13] project includes a database and web server that aim to advance epigenomic research by providing a centralized and accessible platform for analyzing and exploring epigenetic data.

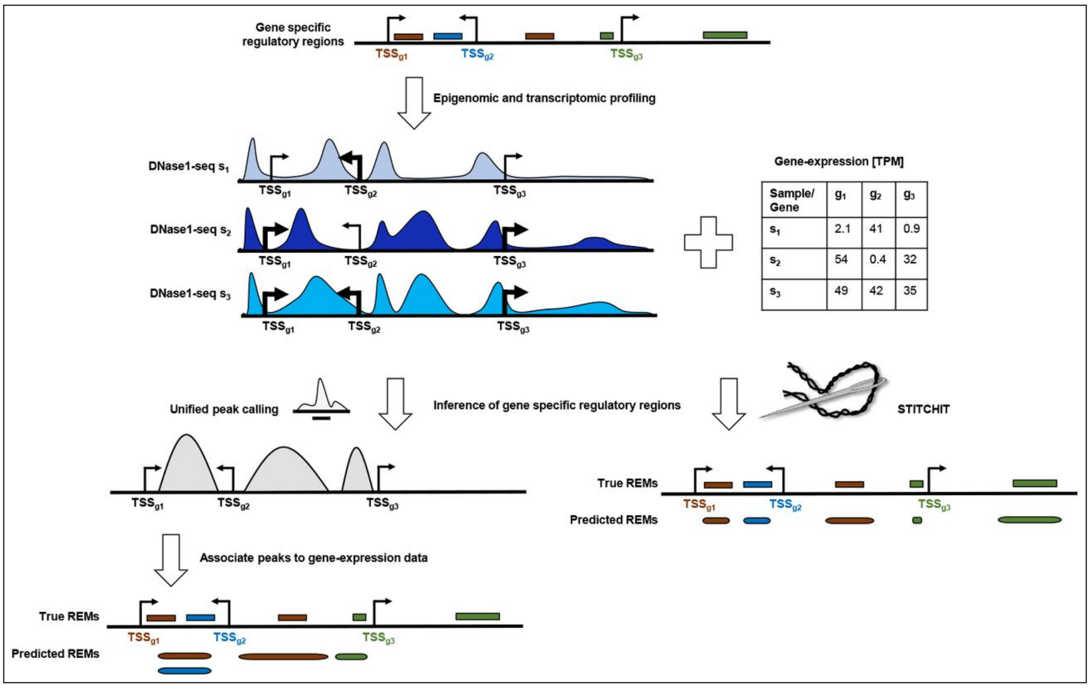
EpiRegio serves as a comprehensive resource for exploring epigenetic data including gene expression, REMs and their interactions. It integrates data from Roadmap[14] and Blueprint[15] and provides a user-friendly interface for data retrieval and analysis. The project aims to enable researchers to investigate epigenetic patterns across diverse cell types, tissues, and organisms by providing data about regulatory elements, their assigned genes and their regulation as previously explained. By consolidating data from multiple experiments and studies, EpiRegio enables researchers to gain insights into the role of epigenetic modifications in gene regulation and disease processes.

The project offers various features and tools to support epigenomic research. It provides data visualization capabilities, statistical analysis tools, and a search and query function to locate specific epigenetic marks or genomic regions of interest. The web server offers several visualization options, such as genomic tracks, heatmaps, and scatter plots, to facilitate the exploration and interpretation of epigenomic data. Additionally, users can perform statistical analyses to identify differentially methylated regions and other epigenetic features.

Overall, the EpiRegio project aims to accelerate scientific discoveries and advance the understanding of epigenetic mechanisms in health and disease.

1.2.1 Regulatory Elements in EpiRegioDB

EpiRegio contains data on REMs and information about their activity in different tissues, cell types, their target gene and its expression[16].


Gene specific regulatory regions
TSS_{g1}
TSS_{g2}
TSS_{g3}
Epigenomic and transcriptomic profiling
DNase1-seq s_1
DNase1-seq s_2
DNase1-seq s_3
Gene-expression [TPM]
Unified peak calling
Inference of gene specific regulatory regions
STITCHIT
True REMs
Predicted REMs
Associate peaks to gene-expression data

Sample/ Gene	g_1	g_2	g_3
s_1	2.1	41	0.9
s_2	54	0.4	32
s_3	49	42	35

Figure 1.1: An Illustration of the STITCHIT method and its performance compared to peak-calling methods from the STITCHIT paper[17].

REMs in EpiRegio are learned using STITCHIT[17], which is a method to identify gene-specific REMs based on the analysis of epigenetic signal of diverse human cell types with respect to gene expression. STITCHIT does this by solving a classification problem in which it segments a large genomic area around the target gene (Fig. 1.1). A linear regression using elastic-net penalty for feature selection and an Ordinary Least Squares regression to get a final regression coefficient and P-value are used. Regions exhibiting the epigenetic signal variance, which is linked to the expression of the gene, are highlighted in the resulting segmentation. Because STITCHIT is a peak-calling free approach it is able to identify REMs with high resolution and accuracy. Since genomic locations are not exclusive to REMs, REMs associated to different genes can overlap. Therefore Cluster of Regulatory EleMents (CREMs) are introduced, containing all, at least two, REMs overlapping with each other by at least one base pair without any break in between. In EpiRegio mentioned REMs, CREMs and their target gene are stored with multiple attributes. REMs and CREMs are stored with unique IDs, chromosome, start position and end position. Genes are stored with their Ensemble ID, also chromosome, start position, end position and their expression value in the respective cell type and sample. The linkage between CREMs and REMs and genes has the following activity values:

I) The `Model score` is a indicator on how important a REM is for the target gene expression compared to all other REMs assicated with that gene. The score is the absolute binary logarithm of the p-value of the regression coefficient for the connection between a REM and its associated gene. It is normalized with the maximal value to obtain a model score in the range [0, 1]. The larger the score, the more impact the REM has on the gene and vice versa. It is not cell type specific, thus can be used for comparison between REMs but not in between cell types.

II) The `StandDnase1Log2` score is a measurement for the accessibility of the chromatin in the REMs, indicating the activity of a REM to its target gene. It is the Deoxyribonuclease 1 (DNase1) signal for a REM retrieved from the Roadmap[14] and Blueprint[15] data sets. DNase1 is an endonuclease coded by the human DNASE1 gene, which functions by cleaving DNA in an endonucleolytic manner[18]. The signal is retrieved by the DNase 1-seq method[19] and has been used as indicator for regulatory regions, which have been shown to map many types of cis regulatory elements. The DNase1 signal is log-transformed and standardized over all cell types[20]. The value inticates how active a REM is in a cell type for a sample and therefore allows the for the comparison between samples, since it is normalized for sequencing depth.

1.3 Gene regulation

Gene regulation is a fundamental biological process necessary for the proper functioning of cells and organisms. Genes contain the instructions for the synthesis of proteins, which perform essential tasks within cells, such as providing structural support, acting as transporters or catalyzing chemical reactions. However, not all genes should be active all the time. Gene regulation ensures that genes are activated or suppressed as needed to adapt to changing conditions and maintain cellular order.

Therefore the regulation process is not static, they exhibit dynamic behaviors, responding to developmental cues and environmental stimuli[22]. This leads to a precise control over the timing and amount of the protein synthesis, which is crucial for the normal functioning of cells and the overall development of an organism.[23]. Gene regulation relies on a complex system involving TFs, enhancers, and epigenetic modifications and often build a larger regulatory network. In this system TFs bind to specific DNA sequences in gene promoters and enhancers. By binding, they modulate the initiation and level of transcription, either activating or repressing gene expression. Activators facilitate the recruitment of RNA polymerase, essential for transcribing the gene into mRNA, while repressors inhibit this process (Fig 1.2). These factors often operate in combinations, enabling precise and context-dependent control of gene activity. Epigenetic modifications, like DNA methylation and histone modifications, further modulate gene accessibility[24] and together complex regulatory systems.

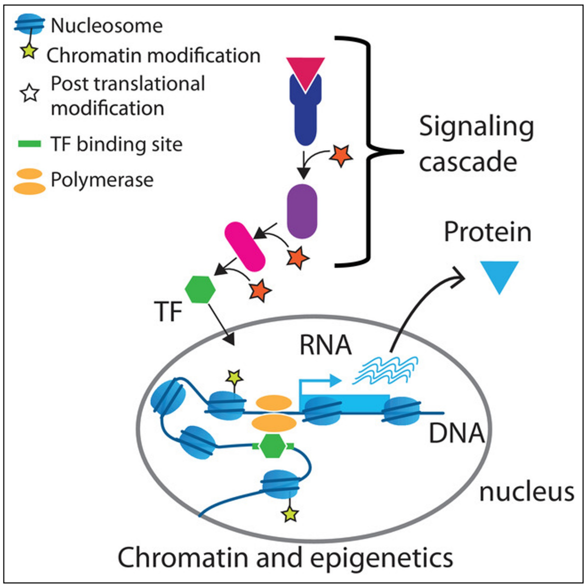


Figure 1.2: A simplified view of the TF activity in gene expression regulation[21].

1.3.1 Gene Regulatory Network

Gene Regulatory Networks (GRNs) computationally model the mentioned regulation system of gene expression in living organisms[22]. In GRNs the key players, TFs and genes, are represented as nodes connected through edges (Fig. 3.1). Edges contain a value describing the type and manner in which the TF regulates the expression of the gene. The edge weights can be either experimental, as already mentioned DNase 1 signals, or statistical values, as the model score of the STITCHIT method. Values from individual cell types can be used to create cell type specific GRNs to investigate in cell type specific regulations. The topology of GRNs varies, including hierarchical, modular, and feedback-driven structures. As gene regulatory systems are very complex, GRNs can become large and complicated. With centrality, clustering and other algorithms designed for large GRNs, these regulatory systems can be analysed efficiently. The results provide information about the pathways, structure and regulatory relationships between actors of the network[25].

Computational models and bioinformatics tools are essential for deciphering relationships and the complex dynamics of GRNs.

1.4 Dysregulation of DNA methylation in cancer

Dysregulation of DNA methylation, the abnormal or altered methylation, is associated with various diseases, including cancer. Aberrant hypermethylation, the excessive methylation, of tumor suppressor gene promoter results in an inactivation of the gene. It contributes to oncogenesis and uncontrolled cell growth. On the other hand global hypomethylation, the reduced or absent DNA methylation, can activate oncogenes, fostering genomic instability[26].

Therefore these alterations and dysregulations in GRNs and DNA methylation are considered promising in cancer research and efforts are being made to find hypermethylated promoters as biomarkers for cancer. Understanding the mechanisms underlying the dysregulated systems is crucial for epigenetic-based interventions in cancer treatment. Since DNA methylation is additionally reversible, it makes it vastly interesting in therapy research and is investigated in this work[6].

Methods

2.1 Packages

The web application and all further analyses were created with the help of several libraries and all in Python, R and Jupyter Notebook. The fundamental ones include dash[27], graph-tool[28], DysRegNet[29] and Locus Overlap analysis (LOLA)[30].

2.1.1 Dash

Dash is an open source Python framework that can be used to efficiently create analytical web apps. Dash was used in this work to create the web app as a user interface for user interaction with the graphs. Therefore dash callbacks were used to integrate interactive elements such as filters, selections and file uploads. Dash and dash-bootstrap-components were used for drop-down menus, checklists, sliders, text inputs and buttons. For the GRN display and interactive usage dash-cytoscape was used[27].

2.1.2 Graph-tool

Graph-tool is a python module that can be used to create, adapt, filter and analyse large networks. Since the library is implemented in C++, it has advantages over other Python libraries. In terms of memory usage and computation time, it is comparable to a normal C++ library, which is significantly faster than a normal Python library. In graph-tool, graphs with their contained nodes, edges and corresponding attributes can be saved and reloaded in `.gt-files` (graph files). The `.gt-file` format provides a simple binary format as an alternative to the text based `graphml` format for large graphs. Where `graphml` files can be time and memory consuming for input and output for large graphs, the `.gt-file` format handles these tasks in a compact and fast manner. In addition, many algorithms and other analysis tools are already integrated in the library[28].

This makes it a perfect tool for the analysis of large regulatory networks and was used to integrate the graphs in this work. The graph files are integrated in the dash web app and can be reloaded when selected. They serve as base for the visualization with dash-cytoscape and can be filtered quickly.

2.1.3 DysRegNet

DysRegNet[29] is a python package to detect patient-specific dysregulations within gene expression profiles. As input it requires a meta data table containing the meta data for all available samples, a GRN and an expression table containing the expression values for all nodes contained in the GRN. Additionally a number of filters and specified inputs can be given. As a result, it returns a table with a z-score for all predicted dysregulated edges for each case sample that can be integrated into a network with one or more connected components. The z-score indicates whether the edge in the disease sample is dysregulated compared to control samples. Positive values indicate activation and negative values indicate repression, the value 0 indicates that this edge is not significantly dysregulated.

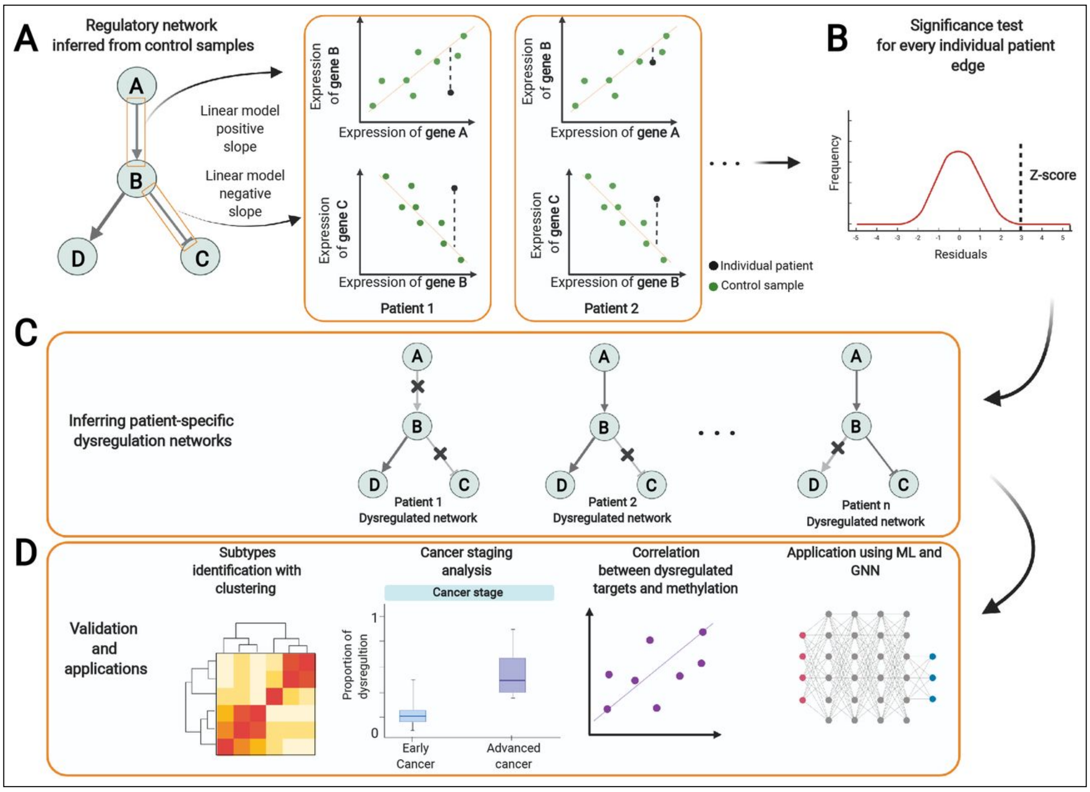
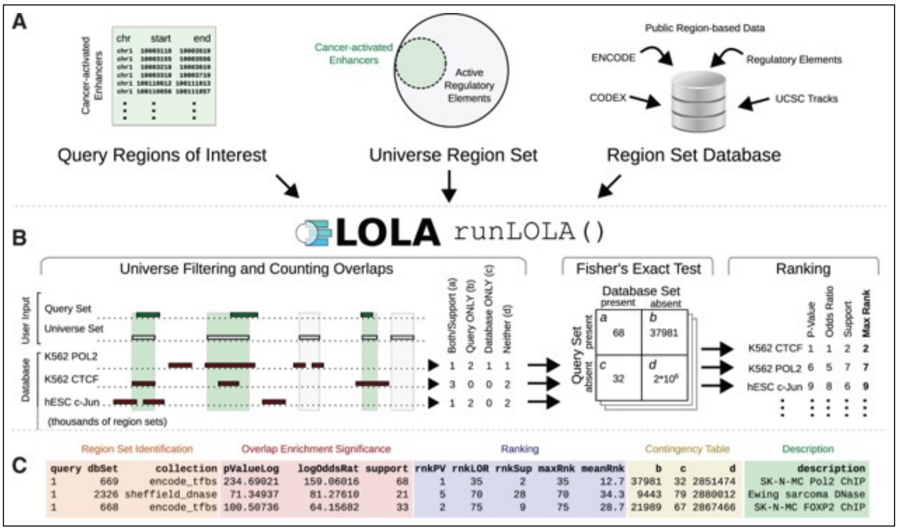

A
Regulatory network inferred from control samples
Linear model positive slope
Linear model negative slope
Expression of gene B
Expression of gene A
Expression of gene C
Expression of gene B
Patient 1
Patient 2
Individual patient
Control sample
B
Significance test for every individual patient edge
Frequency
Z-score
Residuals
C
Inferring patient-specific dysregulation networks
Patient 1 Dysregulated network
Patient 2 Dysregulated network
Patient n Dysregulated network
D
Validation and applications
Subtypes identification with clustering
Cancer staging analysis
Cancer stage
Proportion of dysregultion
Early Cancer
Advanced cancer
Correlation between dysregulated targets and methylation
Application using ML and GNN

Figure 2.1: A Illustration of the DysRegNet tool and workflow. Showing the fitting of the linear model for each edge, the patient individual significance test, the resulting networks and the subsequent result validation[29].

DysRegNet does this by fitting a linear model for each edge given by the input GRN using all control samples (Fig. 2.1). The model parameters are estimated in advance using Ordinary Least Squares. Then each patient sample is tested one after the other and the observed value is compared with the expected value from the linear model, which is comparable with an outlier detection task in regression analysis. A test sample specific z-score is calculated using a standardized residual. After analysing all patients the z-scores are transformed into p-values and corrected for multiple testing. It is important to mention that dysregnet can also account for covariates, which gives the tool an advantage over comparable tools that do not have this feature integrated.

2.1.4 LOLA

LOLA is an R package tool integrated in Bioconductor for enrichment analysis for genomic regions[30]. The analysis is based on testing for overlaps of genomic regions of interest and a database of regions from previous studies. Three components are needed for the analysis: (I) The query set - several or a list of genomic regions to be tested for enrichment; (II) A region universe - the background set of genomic regions contained in all query subsets; (III) The genomic region reference database to be tested for overlap with the query set. LOLA contains several core databases for the GRCh38 genome, among others and (I) and (II) can be read in as BED files. LOLA identifies all regions from the query set that overlap with each region set in the reference database, which is performed against the user specified region universe (Fig. 2.2). A single common base pair is sufficient for both regions to be considered overlapping, but can be adjusted by the user. Finally, LOLA uses Fisher's exact test with false discovery rate correction to evaluate the significance of the overlap. The resulting rank score is the worst (max) rank of P-value, log-odds ratio and number of overlapping regions. The resulting enriched regions are returned as a `data.table` object. This provides a powerful interface for sorting, exploring, visualising and further processing the results.



```
query dbSet       collection pValueLog    logOddsRat support rnkPV rnkLOR rnkSup maxRnk meanRnk     b  c       d          description
1       669      encode_tfbs 234.69021     159.06016      68     1     35      2     35    12.7 37981 32 2851474   SK-N-MC Pol2 ChIP
1      2326 sheffield_dnase   71.34937      81.27610      21     5     70     28     70    34.3  9443 79 2880012 Ewing sarcoma DNase
1       668      encode_tfbs 100.50736      64.15682      33     2     75      9     75    28.7 21989 67 2867466  SK-N-MC FOXP2 ChIP
```

Figure 2.2: A Illustration of the LOLA enrichment workflow[30].

2.2 Network generation and data preparation

2.2.1 Graph-files

For the creation of the graph files, all 10 files of the EpiRegioDB were viewed and all samples from the `sampleInfo_Roadmap_1.csv` and `sampleInfo_Blueprint_1.csv` project files were assigned to their cell types with help of the `CellTypeInfo.csv` file. The samples from both projects were matched with those of the `REMActivity_1.csv` file. This resulted in 165 samples in 46 cell types which are stored in the `sampleToCelltype.csv` file.

For each cell type, a directed graph file was created containing the edge, the REM and the gene as source and target node from the `REMAnnotationModelScore_1.csv` file. As an edge attribute the cell type specific edge weights contained in the cell type were stored as an internal `edge_property` from graph-tool under the sample name. As edge weight the value of the standDnase1Log2 column from the `REMActivity_1.csv` file was used. REMs associated with a CREM from the `clusterREMs_1.csv` file were merged into one CREM node. CREMs have a single edge to the corresponding gene for each contained REM. In addition, the all cell type graph file with the identical edges, genes and REMs as nodes was created using the `REMAnnotationModelScore_1.csv` file. As edge weight the value of the normModelScore column from the same file was used, which contains the model score across all cell types. As REM and gene internal `vertex_property` node attributes, type, name, chromosome, start, end and for the cell type specific graph files the gene expression for gene nodes for the respective sample were added from `GeneExpressionBlueprint_1.csv` and `GeneExpressionBlueprint_1.csv`. As additional edge attributes the rem name of the edge, for the all cell type graph file the p-value and for the cell type specific graph files the cell type and cell type ID were added.

As already mentioned, depending on whether it is a cell type specific or the all cell type graph, the edge and node attributes differ. To show the attributes values an example the graph for `CTID_0000006`, cd14-positive monocyte and the all cell type graph is taken.

Node attributes:

- name: name of the gene/REM
- type: node type, either gene oder rem
- chr: chromosome on which the gene/REM in located
- start: start coordinate of gene/REM
- end: end coordinate of gene/REM

Only for gene nodes:

- B_C0010KB1: expression value in sample B_C0010KB1
- B_C0011IB1: expression value in sample B_C0011IB1
- B_C001UYB4: expression value in sample B_C001UYB4

The all cell type graph contains these node attributes except for the expression values. The reason for this is that this graph contains model score values for all cell types and therefore has no expression value for all cell types at once. The difference with the CTID_0000006 graph concerning the edge attributes is that it has a separate edge attribute for each sample with its name.

Edge attributes:

- rem: REM name the edge is assigned to, only important if REM is part of a CREM
- celltype: CTID_0000006

- celltypeID: cd14-positive monocyte
- B_C0010KB1: standDnase1Log2 value for edge in sample B_C0010KB1
- B_C0011IB1: standDnase1Log2 value for edge in sample B_C0011IB1
- B_C001UYB4: standDnase1Log2 value for edge in sample B_C001UYB4

In the all cell type graph, the same attributes are present except for the standDnase1Log2 values, these are replaced by the following:

- score: model score over all cell types
- p-value: p-value for the model score

This produces 47 graph files in total with identical nodes and edges, each representing a cell type specific or a GRN for all cell types.

2.2.2 CpG-mapping and beta values

To use the GRN as input to DysRegNet an expression value for each node is needed. So far, the GRNs only contain expression values for gene nodes. To obtain expression values for the REMs as well, the beta values of all CpGs lying within the REM coordinates by taking the CPG coordinates from Illumina Methylation450K[31].

Since the coordinates of the CpGs are based on GRCh37 and all EpiRegio coordinates are based on GRCh38, the CpG coordinates were mapped to GRCh38 using UCSC LiftOver with default settings[32]. LiftOver takes a bed file containing the input coordinates from the selected assembly and convertes them to the coordinates of the desired output assembly. Coordinates that can't be converted result in a list of mismatches, the successfully converted coordinates are printed to a file. After manually filtering 71 mismatches from LiftOver and REMs to which no CpG site could be assigned, all remaining REMs have their assigned CpGs by mapping the CpG site coordinates to the REM coordinate ranges.

The beta values of the CpGs from the data being analysed were converted to m-values, as these show significantly better performance in identifying differentially methylated CpG sites[10] by the following equation:

$$m - value = \log_2 (\frac{beta}{1 - beta}) \tag{2.1}$$

To obtain a single expression value for a REM, the mean was calculated over all m-values of the CpGs contained in a REM.

This provides all input data, the GRN, expression values for all nodes, and the remaining data of respective data set for DysRegNet.

Results

3.1 Network

The resulting 47 GRNs serve as the basis for all analyses and visualizations. These very large networks contain several million nodes and edges. In detail every graph consists of:

Number of nodes	1.541.001
Number of edges	2.404.861
Number of genes	35.379
Number of REMs	1.140.336
Number of CREMs	365.286

Table 3.1: Element numbers contained in the network graph.

The basic structure of the graphs is composed of small clusters consisting of one gene, which has a large number of incoming edges from REMs and CREMs (Fig. 3.2). Genes have exclusively named incoming edges and no edges to other genes. REMs have exclusively a single outgoing edge to their assigned gene and no edge to other REMs. Only CREMs have multiple outgoing edges representing the edges of the contained REMs. CREMs can therefore also have edges to several different genes or several edges to the same gene, each representing a single REM edge. Thus, the mentioned small clusters are connected, if then only by edges from CREMs. Since not every gene has an edge to a CREM and only through it can the individual gene clusters be connected, the network is a disjoint network.

Figure 3.1, a section of the webapp visualization, illustrates the individual cases all at once. Three genes can be seen as light blue nodes, REMs and CREMs are shown as black nodes and are all connected by their edges. `ENSG00000142627` in the upper right corner shows the case where a gene is connected only to REMs and not CREMs, creating an isolated clique with no connection to the rest of the network because REMs can exclusively have a single edge to a gene. The large cluster in the lower left center of the figure with `ENSG00000236908` and `ENSG00000275850` shows how two genes are connected via `CREM0068412`. Here, `ENSG00000236908` has only one incoming edge through the mentioned CREM and `ENSG00000275850` forms a cluster with several neighboring REMs and CREMs, while it is also connected to other genes via the latter, but these are not shown here. Also, one can see the case where a CREM has multiple edges to the same gene, each representing a separate REM edge.

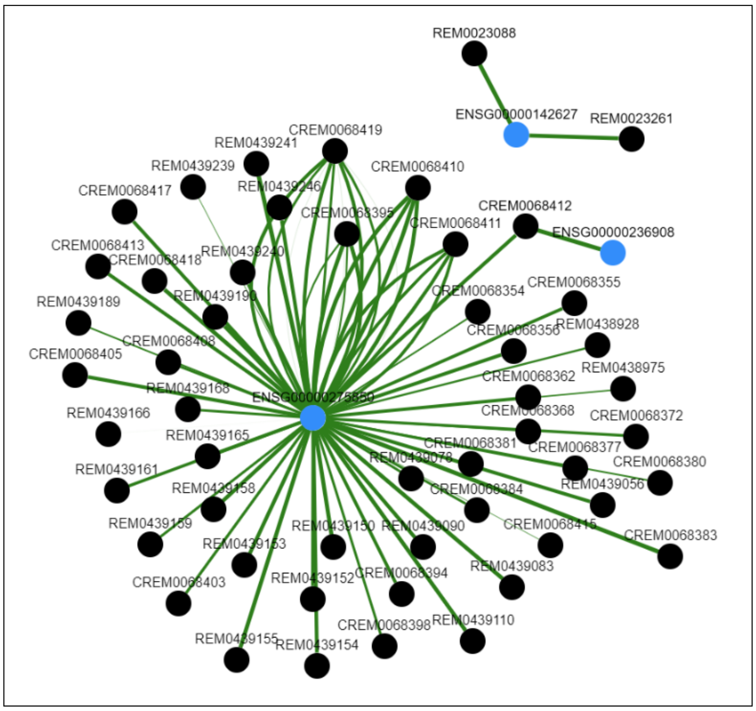


Figure 3.1: Visualization of three genes and their connected REMs and CREMs in the Webapp showing the clusters and disjoint structure of the network.

Looking at the topology, the graph consists of 2.188 individual components, which are shielded from each other and cannot be reached by each other. The sizes of the components range from 2 to the largest component with 16.500 contained nodes and form a total of 1.055 different sizes. Smaller components with up to 1.000 contained nodes occur significantly more often than extremely large ones. The most frequent components are those with 55 nodes, of which a total of 134 can be found in the entire network, and the least frequent components with 10.000 nodes or more are usually represented only once. That illustrates that the network consists of many individual components, which differ strongly by their size and frequency.

The average degree over all nodes within the graph is 3.12, but it differs enormously if you look at the type of nodes. The average degree for gene nodes is 67.97 and for REM and CREM nodes only 1.60. This value is small for REM and CREM nodes, since REMs always have only one edge to a certain gene and thus drive the average value down. Looking at the average degree for CREMs only, it is almost three times higher at 3.46. These numbers again describe the network structure with clusters of genes with many incoming edges from regulatory elements. The value of the maximum degree does not differ that much and is 108 for gene nodes and 122 for CREMs, for REMs the value is self-explanatory 1.0. However, 8.096 genes have a degree of 108, only `CREM0109128` has a degree of 122, which is enriched for the nuclear respiratory factor 1 gene (NRF1) using LOLA[30]. NRF1 functions as a TF, activating the expression of genes required

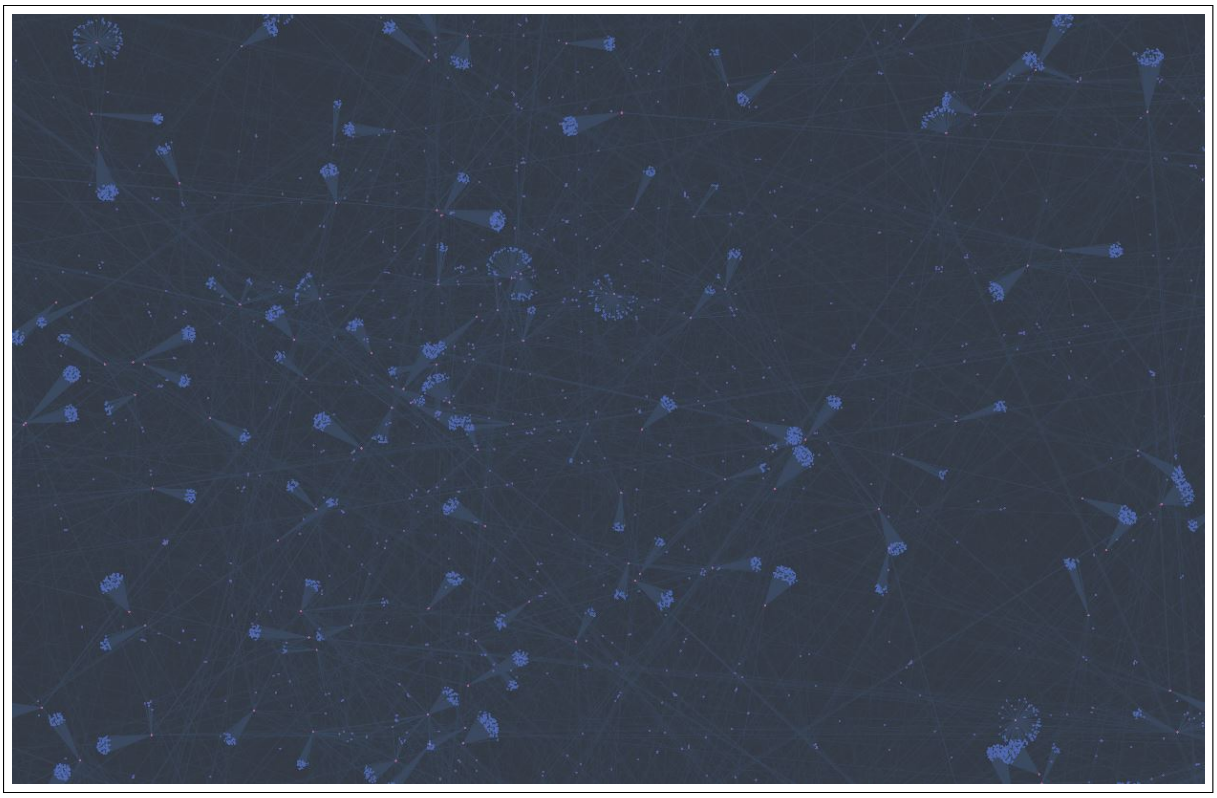


Figure 3.2: Part of the model score network graph visualized with Cosmograph[33] showing the clusters with one gene and multiple connected REMs and CREMs.

for mitochondrial DNA transciption and regulating cellular growth[34]. The TF is thus involved in many processes and therefore has a high number of interactions. The degree distribution of the CREMs (Fig. A.1) is very one-sided and most of them have degrees of 2 and from fifteen on there are only a few with higher degrees. The distribution of gene degrees (Fig. A.2), on the other hand, looks very different with two maxima at 54 and 108, showing that a maximum degree of 108 is a common case.

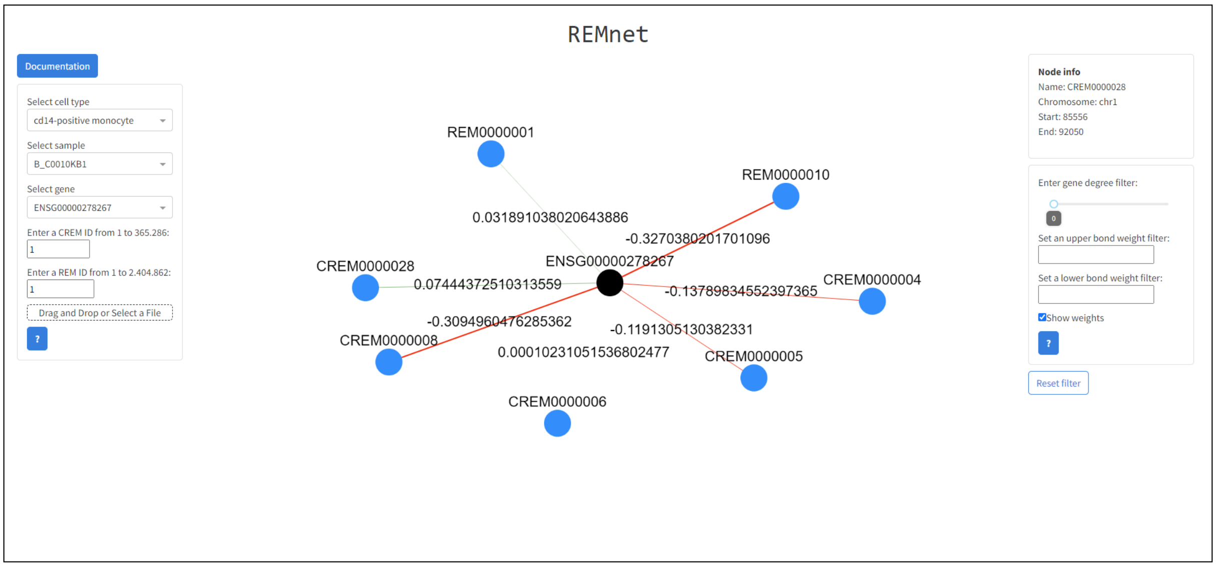
Results that support these findings can also be seen by looking at the clustering coefficients, which give 0 for both local and global clustering coefficients. For this purpose the methods `local_clustering` and `global_clustering integrated` with default settings from graph-tool were used. These minimal coefficients show that all nodes in this network have no edges between their neighboring nodes, and thus only nodes are connected to their immediate neighbors. This confirms the description of the basic network structure as described earlier.

3.2 Web application

To be able to interact with the created graph files, a user friendly web application (app) was designed. The app provides visualization, interaction and filtering of the networks and their properties for user specific usage. It was created as described in 2.1.1 using dash and dash-cytoscape.

3.2.1 Structure

The basic structure of the app is divided into three parts, which are arranged vertically next to each other. In the left column, the data can be selected for display and there is the option for a file upload. In the middle column is the display area of the elements selected in the left column. In the right column you can filter the elements to be visualized and display information about selected nodes.



Select cell type: cd14-positive monocyte
Select sample: B_C0010KB1
Select gene: ENSG00000278267
Enter a CREM ID from 1 to 365.286: 1
Enter a REM ID from 1 to 2.404.862: 1

Node info
Name: CREM0000028
Chromosome: chr1
Start: 85556
End: 92050

Enter gene degree filter: 0
Set an upper bond weight filter:
Set a lower bond weight filter:
☑ Show weights

Figure 3.3: App view on start with the selection options on the left, the visualization in the middle and the filtering on the right.

3.2.2 Selection

The left column is the selection column. It has two help buttons for the user, which can be opened and closed like all help button by simple clicks. The `Documentation` button gives an overview of which actions can be performed in which parts of the app. The `?` help button gives instructions to the user on how to use the selection and upload options.

Below that is the window for all display choices. In the first `Select cell type` drop-down menu, all cell types contained in Roadmap[14] and Blueprint[15] can be selected. Selecting one of the cell types will reload the respective graph file as described in 2.1.2. This results in a additional short loading time, but allows to speed up the subsequent work in the graph file. For the loading time a spinning loading circle is displayed in the upper center of the page. Below that, it is possible to choose from all the samples available for this cell type in the `Select sample` drop-down. If the all cell type graph is selected, the model score and the p-value are selectable instead of the sample names as explained in 2.2.1. In the next three input fields, genes, REMs or CREMs can be selected. Only one of the three can be shown in the visualization, the last selection is always shown. In the gene `Select gene` drop-down menu, genes can be selected by their ensemble ID. For the selection of REMs and CREMs, the desired ID must be entered. All drop-down menus can be searched for the desired selections.

In addition to the above options, the user can upload a CSV file with individualized nodes and edges. This function is useful if only selected or several disconnected elements are to be displayed. The file must be in the format `'REM,GENE'` as header with two columns where each row describes an edge between the respective REM and Gen. Other file formats and formattings are not accepted.

3.2.3 Visualisation

In the app center, individual genes, REMs, CREMs and their neighboring nodes and edges or a user-defined input of nodes by the file upload can be displayed.

When a gene, REM or CREM is selected, it is displayed as seed node in black. All neighboring nodes are displayed in light blue and the edges leading to them. Above each node the corresponding name is displayed. Depending on the previously selected sample, the adjusted edge weights are shown on the edges. For a simplified illustration, the edges are colored green for positive edge weights and red for negative edge weights. In addition, the edge width adapts to the weights and becomes wider the further away the value is from 0. If the file upload function is used, all edges specified in the file are displayed. Genes are always displayed as black nodes and REMs and CREMs as light blue nodes (Fig. 3.1).

These display options are available, since the display of the entire GRN is not useful for several reasons. On the one hand, it would be extremely inefficient, creating long loading times and thus a poor user experience. On the other hand, the display of the entire GRN would be confusing and the user would have difficulties to orientate himself, as in Fig. A.4.

3.2.4 Filtering

The right column of the app contains information about nodes and filtering options for the GRN.

In the upper bin, information about nodes for which more details are requested can be displayed. The information appears by simply clicking on a node. The name of the node, on which chromosome it is located and the start and end position in the genome are given. The display is available for genes as well as REMs and CREMs.

The box below contains the filter options for the network visualization. With the help of the first `Enter gene degree filter` slider the gene degree can be filtered. By moving the slider, genes with a higher gene degree than the filter value are filtered. This affects the `Select gene` dropdown menu where the selectable genes are adjusted according to the filter. In the `Upper bond weight filter` and `Lower bond weight filter` input fields the edge weights can be filtered. In the former, edge weights smaller than the filter are filtered and larger ones are left out. With the latter it is the opposite, edge weights larger than the filter are filtered and smaller ones are left out. Thus, the edge weights can be narrowed down from both sides. The filters are transferred to the visualization after selection and all edges excluded by the filters and nodes connected only by the excluded edge are removed. Using the `Show weights` select box, the edge weights can be shown or hidden. The `?` help button gives instructions to the user on how to use filters correctly. The `Reset filter` button resets all selected filters to their default values.

All these functions and filters together form the app, which allows the user to view and filter desired parts of the GRN.

3.3 Use case

The GRN can be applied in combination with DysRegNet and LOLA to identify and investigate patient-specific dysregulations in a disease data set. A corresponding use case is described below for breast cancer (BRCA).

3.3.1 Breast Cancer

With 2.3 million new cases of BRCA, it is the most common type of cancer in women[35]. Breast cancer also affects men, who account for less than 1% of breast cancer cases. The risk of developing breast cancer also depends on several factors, among which the mutations in the tumour suppressor genes BRCA1/BRCA2 and older age contribute negatively. Numbers of cases and the death rate increased between 1990 and 2016. Death rates have doubled in 60 of 102 countries such as Saudi Arabia and Paraguay, and even tripled in ten countries such as Iran and Algeria. The number of cases also more than doubled in 43 of 102 countries, such as Afghanistan and Brazil[36]. Survival depends on both stage and molecular subtype. Looking at the histology, BRCA can be divided into subtypes Luminal A, Luminal B, HER2 enriched, Basal-like and Normal Breast-like. The subdivision is based on different gene expression levels such as those of estrogen receptors (ER), progesterone receptors (PR) or human epidermal growth factor receptor 2 (HER2). Luminal breast cancers account for almost 70% of all breast cancers in the western population and are characterised by the presence of ER. Luminal A and Luminal B breast cancers are differentiated by the absence of PR and/or the presence of HER2 in Luminal B tumours. The HER2-enriched subtype, which is characterised by a high expression of HER2 and the absence of ER and PR, accounts for 10-15% of breast cancer cases. Basal-like, often used interchangeably with triple-negative breast cancer, is characterised by the absence

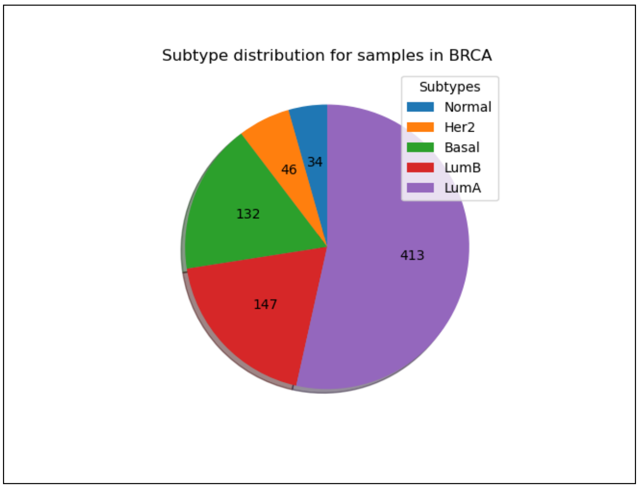


Figure 3.4: Distribution of subtypes in samples of the DysRegNet result from TCGA BRCA data set.

of both ER and PR as well as HER2[35]. This subtype accounts for about 20% of breast cancers. The Normal Breast-Like subtype accounts for between 5-10% of breast cancer cases, but is poorly characterised as its gene expression levels are intermediate between Luminal and Basal-like subtypes[37].

The Xena platform[38] contains the `GDC TCGA Breast Cancer (BRCA)` data set with meta data, expression data and methylation data for breast cancer patients from The Cancer Genome Atlas (TCGA)[39]. The meta data contains 1.284 samples with 140 identifiers each, distributed across both sexes, several ages, with a control group and others. The expression data contain values for 1.217 samples over 60.484 genes. The expression values are specified FPKM (Fragments Per Kilobase Million) and log2(x+1) transformed. The methylation data consist of 890 samples and 485.578 CpGs containing the respective beta values for Methylation450K[31].

3.3.2 DysRegNet

Given a disease dataset, there is interest in finding dysregulations that differ between case and control samples, as these can be used to infer the changes in affected patients. With the help of DysRegNet, patient-specific dysregulations of gene expression profiles can be inferred. For the DysRegNet inputs, the data was adapted accordingly. The GRN was extracted from the all cell type graph file in the form `REM - GENE` for all edges contained. In the meta table, the control samples were encoded with 0 and the case samples with 1. Missing values for in the `days_to_birth.demographic` column were completed with the mean column value. With the help of the methylation data, the CpGs were matched to their respective REMs as described in 2.2.2, the beta values were converted to m-values and the mean value of all contained CpGs was assigned to each REM as expression value.

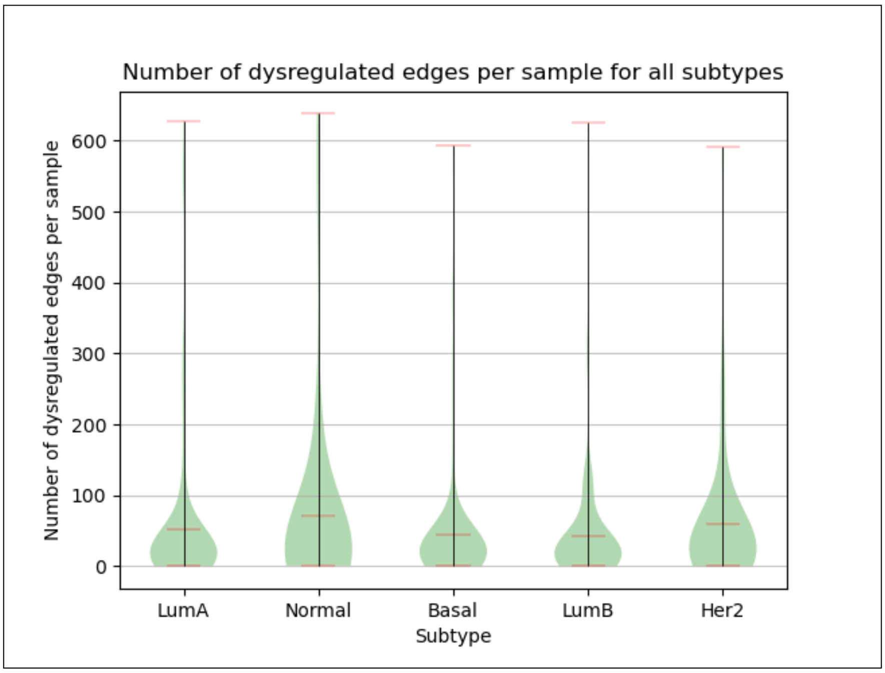


Figure 3.5: Number of dysregulated edges per sample in the DysRegNet result. Grouped by BRCA subtypes.

The REMs and their newly generated expression values were added to the expression table. Since the meta data, the expression table and the methylation data do not contain the same samples, they were filtered to the largest intersection for all three tables. In the GRN, the REMs and genes were adjusted according to those contained in the expression table, since not all REMs could be assigned CpGs and thus expression values. This results in 861 samples in the meta table as well as the Expression table. The input GRN and expression table values contain entries for 75,543 REMs and 34,934 genes. To run DysRegNet, the default input values as in the example of the DysRegNet paper[29] were taken. Thus the inputs are: (I) expression, meta and methylation tables described above; (II) covariates race, gender and days to birth; (III) R2 threshold = 0.2 and normaltest = True; (IV) sample type as condition column. The result is a table with 778 case samples and 169,498 predicted dysregulated edges connecting REM and gene. If only the edges

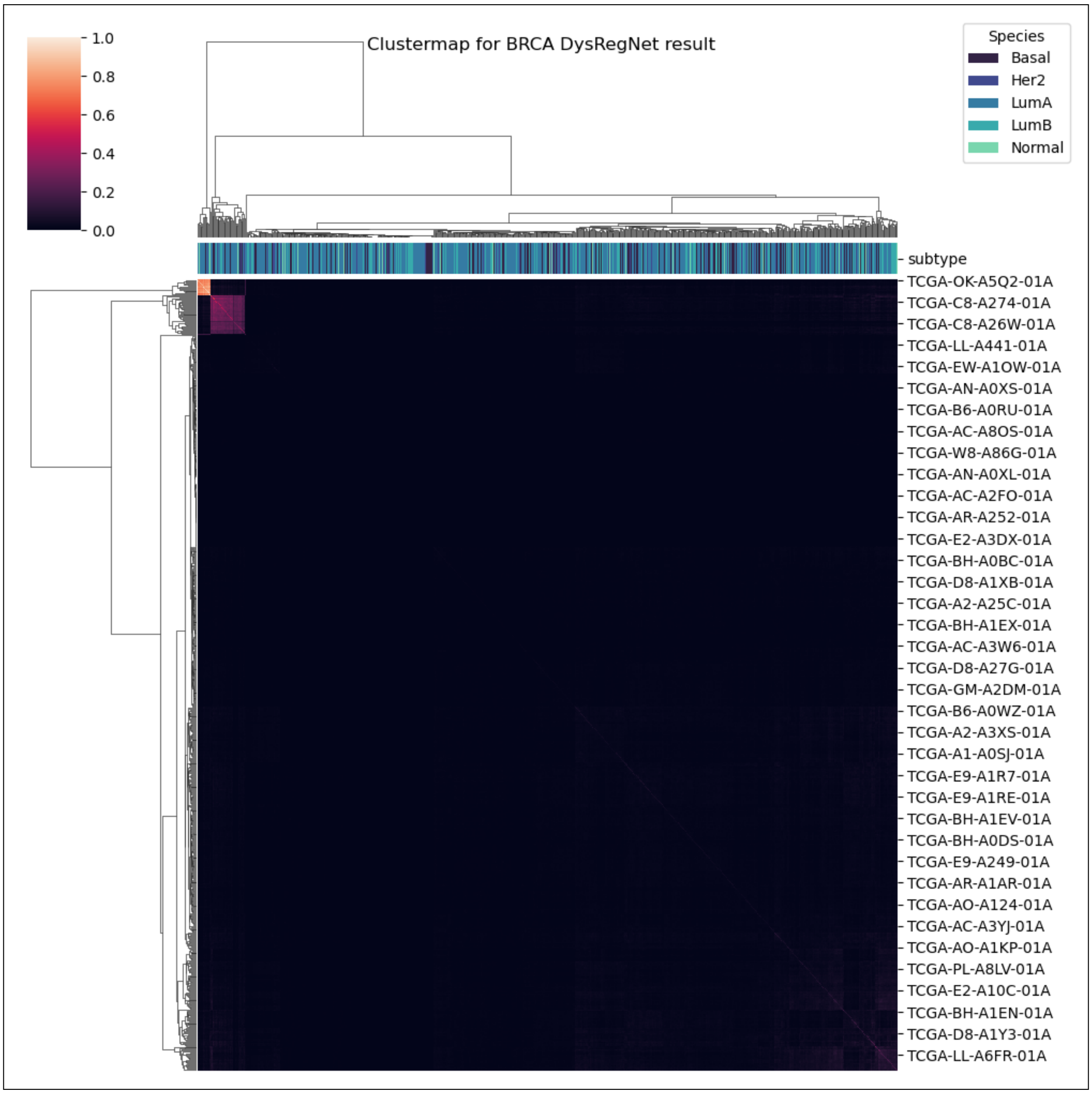


Figure 3.6: Clustering Heatmap of the number of equal dysregulated edges for each pair of samples.

that have a z-score unequal 0 in one of the samples are considered and are thus predicted dysregulated, 1.890 edges remain. Only named dysregulated edges are used for further analysis.

In order to be able to cluster the edges and thus obtain subtype-specific groups of dysregulated edges, the samples were assigned their subtypes. For the assignment, the table of PanCancerAtlas subtypes from TCGAbiolinks[40] was used. It was filtered according to the cancer type BRCA and the samples for which a subtype is given. This resulted in a table of 772 samples with their associated subtypes. In the BCRA subtype distribution (Fig. 3.4), it is noticeable that the luminal subtypes make up almost two-thirds of all samples, as they account for the largest proportion of breast cancer cases, described in 3.3.1. The HER2 and basal subtypes represent the smallest groups. Looking at the number of dysregulated edges per sample (Fig. 3.5), it becomes apparent that the distributions for each subtype are similar. The distributions differ minimally in their maximum values around 600 and their mean values do not diverge much. This provides a good basis for differentiating the subtypes according to their dysregulated edges, as they are comparable. In contrast, the subtype distributions for the number of dysregulated samples per edge differ significantly (Fig. A.5). They are proportionally similar to the sizes of the subtypes in Figure 3.4. This is due to the fact that samples from the larger subtypes are more likely to be dysregulated in more edges.

For the clustering of the edges, all edges were compared with each other in pairs. For each pair, the number of edges that are dysregulated in both samples, i.e. have a z-score unequal 0, was determined. This results in a symmetrical matrix with all samples on both the x and y axis, with the number of equal dysregulated edges in each cell. The matrix was normalised in order to obtain values between 0 and 1. Clustering this table using a clustering heatmap and displaying the corresponding subtypes of samples yields the result in Figure 3.6. The x axis shows the respective subtypes of the samples and lighter values in the matrix indicate a higher number of equal dysregulated edges per sample pair. The samples do not cluster into their subtypes as expected. Two clusters can be seen in the upper left, however, these are divided across all subtypes and therefore are not subtype-specific. Apart from that, no other clusters are visible.

3.3.3 LOLA enrichment

The DysRegNet results provide dysregulated edges between REM and gene per case sample. However, only genomic coordinates are given for the REMs. Therefore, it is interesting to investigate which TFBs represent the dysregulated REMs. For this purpose, the LOLA enrichment tool was used, which provides an enrichment analysis for genomic regions. From the DysRegNet results table, a .bed file was created for each sample containing the coordinates of the REMs in the dysregulated edges. Samples without dysregulated edges were filtered out, leaving 746 samples and .bed files. These were taken as query sets and the hg38 genome from the LOLA core data base as the data base for input. All edges from the DysRegNet result table that were dysregulated in at least one sample formed the region universe for input. The LOLA results provide the enriched TFBs for the REMs from each sample file. For each TFBs, the results of the statistical test are given by p-value, log-odds ratio, number of overlaps and ranks. The meanRank and maxRank rank the results using a combination of named values and often provide the most interpretable results[30]. The results for each sample were ordered by their ascending maxRank and

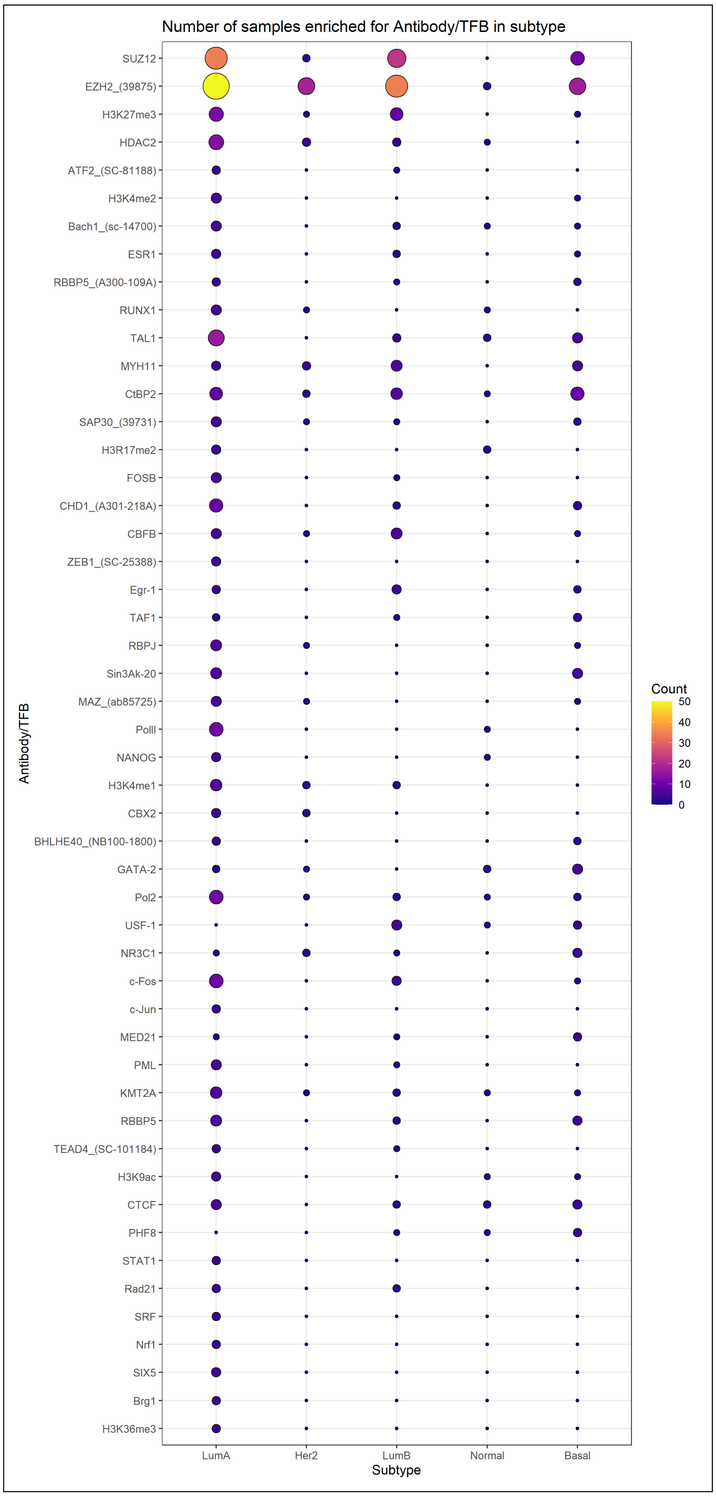


Figure 3.7: Illustration for the number of samples in which the TFBs is enriched, shown for each subtype.

TFBs with the smallest maxRank were taken. By calculating the number of taken enriched TFBs of all samples per BRCA subtype, a table with subtypes on the X axis and all uniquely enriched TFBs on the Y axis was generated. The table was filtered for all TFBs that occur at least twice in a subtype. This leaves 50 TFBs and their number for the five subtypes in Figure 3.7.

Discussion

- discuss results
- how to use web app
- how to interpret genes and rems
- select multiple genes
- select network components
- more filter, gene filter both directions

Bibliography

[1] Esteller M. Portela, A. Epigenetic modifications and human disease. *Nat Biotechnol*, 2010.

[2] Jones PA. Sharma S, Kelly TK. Epigenetics in cancer. *Carcinogenesis*, 2010.

[3] M Okano et al. Dna methyltransferases dnmt3a and dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 1999.

[4] Xu J. Zhang, W. Na methyltransferases and their roles in tumorigenesis. *Biomark Res*, 2017.

[5] Le T. Moore, L. and G. Fan. Dna methylation and its basic function. *Neuropsychopharmacol*, 2013.

[6] Lakshminarasimhan R and Liang G. The role of dna methylation in cancer. *Adv Exp Med Biol.*, 2016.

[7] Giorgio Bernardi Kamel Jabbari. Cytosine methylation and cpg, tpg (cpa) and tpa frequencies. *Gene*, 2004.

[8] Illingworth RS and Bird AP. Cpg islands–'a rough guide'. *FEBS Lett.*, 2009.

[9] Daiya Takai and Peter A. Jones. Comprehensive analysis of cpg islands in human chromosomes 21 and 22. *Proceedings of the National Academy of Sciences*, 99(6):3740–3745, 2002.

[10] Zhang X. Huang CC. Du, P. et al. Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics 11*, 2010.

[11] Stampfel G. Shlyueva, D. and A. Stark. Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet*, 2014.

[12] Jia Y. Wang Y. Wang, Q. et al. Evolution of cis- and trans-regulatory divergence in the chicken genome between two contrasting breeds analyzed using three tissue types at one-day-old. *BMC Genomics*, 2019.

[13] Nina Baumgarten, Dennis Hecker, Sivarajan Karunanithi, Florian Schmidt, Markus List, and Marcel H Schulz. Epiregio: analysis and retrieval of regulatory elements linked to genes. *Nucleic Acids Research*, 2020.

[14] Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 2015.

[15] Hendrik G Stunnenberg et al. The international human epigenome consortium: A blueprint for scientific collaboration and discovery. *Cell*, 2016.

[16] Elemento O. Doane AS. Regulatory elements in molecular networks. *Wiley Interdiscip Rev Syst Biol Med*, 2017.

[17] Florian Schmidt, Alexander Marx, Marie Hebel, Martin Wegner, Nina Baumgarten, Manuel Kaulich, Jonathan Göke, Jilles Vreeken, and Marcel H. Schulz. Integrative analysis of epigenetics data identifies gene-specific regulatory elements. *bioRxiv*, 2019.

[18] National Library of Medicine. Dnase1 deoxyribonuclease 1. `https://www.ncbi.nlm.nih.gov/gene/1773`, 2023. (accessed: 15.09.2023).

[19] John Sam et al. Genome-scale mapping of dnase i hypersensitivity. *Current protocols in molecular biology*, 2013.

[20] Shulha HP Meltzer P Margulies EH Weng Z Furey TS Crawford GE. Boyle AP, Davis S. High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 2008.

[21] Evangelia Petsalaki Judith B. Zaugg Paula Weidemüller, Maksim Kholmatov. Transcription factors: Bridge between cell signaling and gene regulation. 2021.

[22] Enze Liu, Lang Li, and Lijun Cheng. Gene regulatory network review. In Shoba Ranganathan, Michael Gribskov, Kenta Nakai, and Christian Schönbach, editors, *Encyclopedia of Bioinformatics and Computational Biology*, pages 155–164. Academic Press, Oxford, 2019.

[23] Lewis J Alberts B, Johnson A et al. Molecular biology of the cell. 4th edition. *New York: Garland Science*, 2002.

[24] Bacopoulou F. Vlachakis D. Chrouso G.P. Mitsis T., Efthimiadou A. and Eliopoulos E. Transcription factors and evolution: An integral part of gene expression (review). 2020.

[25] Isabelle S. Peter. Chapter nine - the function of architecture and logic in developmental gene regulatory networks. In Isabelle S. Peter, editor, *Gene Regulatory Networks*, volume 139 of *Current Topics in Developmental Biology*, pages 267–295. Academic Press, 2020.

[26] Marta Kulis and Manel Esteller. 2 - dna methylation and cancer. In Zdenko Herceg and Toshikazu Ushijima, editors, *Epigenetics and Cancer, Part A*, volume 70 of *Advances in Genetics*, pages 27–56. Academic Press, 2010.

[27] Plotly Technologies Inc. Dash plotly. `https://plot.ly`. (accessed: 15.09.2023).

[28] Tiago P. Peixoto. The graph-tool python library. *figshare*, 2014.

[29] Olga Lazareva, Zakaria Louadi, Johannes Kersting, Jan Baumbach, David B. Blumenthal, and Markus List. Dysregnet: Patient-specific and confounder-aware dysregulated network inference. *bioRxiv*, 2022.

[30] Christoph Bock Nathan C. Sheffield. Lola: enrichment analysis for genomic region sets and regulatory elements in r and bioconductor. 2016.

[31] Inc. Illumina. Illumina methylation450k. `https://support.illumina.com/array/array_kits/infinium_humanmethylation450_beadchip_kit/downloads.html`. (accessed: 15.09.2023).

[32] UCSC Genome Browser Group. Ucsc liftover. `https://genome.ucsc.edu/cgi-bin/hgLiftOver`. (accessed: 15.09.2023).

[33] Nitika Rokotyan, Olga Stukova, and Denis Ovysyannikov. Cosmograph: Gpu-accelerated force graph layout and rendering. `https://cosmograph.app`, 2022.

[34] National Library of Medicine. Nrf1 nuclear respiratory factor 1. `https://www.ncbi.nlm.nih.gov/gene/4899`, 2023. (accessed: 15.09.2023).

[35] Forma A Baj J Sitarz R Stanisławek A. Łukasiewicz S, Czeczelewski M. Breast cancer-epidemiology, risk factors, classification, prognostic markers, and current treatment strategies-an updated review. 2021.

[36] R. Sharma. Breast cancer incidence, mortality and mortality-to-incidence ratio (mir) are associated with human development, 1990–2016: evidence from global burden of disease study 2016. 2019.

[37] Barutca S. Yersal O. Biological subtypes of breast cancer: Prognostic and therapeutic implications. 2014.

[38] Hastie M. Goldman M.J., Craft B. et al. Visualizing and interpreting cancer genomics data via the xena platform. 2020.

[39] Mills GB Shaw KR Ozenberger BA Ellrott K Shmulevich I Sander C Stuart JM Weinstein JN, Collisson EA. The cancer genome atlas pan-cancer analysis project. 2013.

[40] Catharina Olsen Luciano Garofano Claudia Cava Davide Garolini Thais S. Sabedot Tathiane M. Malta Stefano M. Pagnotta Isabella Castiglioni Michele Ceccarelli Gianluca Bontempi Houtan Noushmehr Antonio Colaprico, Tiago C. Silva. Tcgabiolinks: an r/bioconductor package for integrative analysis of tcga data. 2016.

Appendix

A.1 Figures

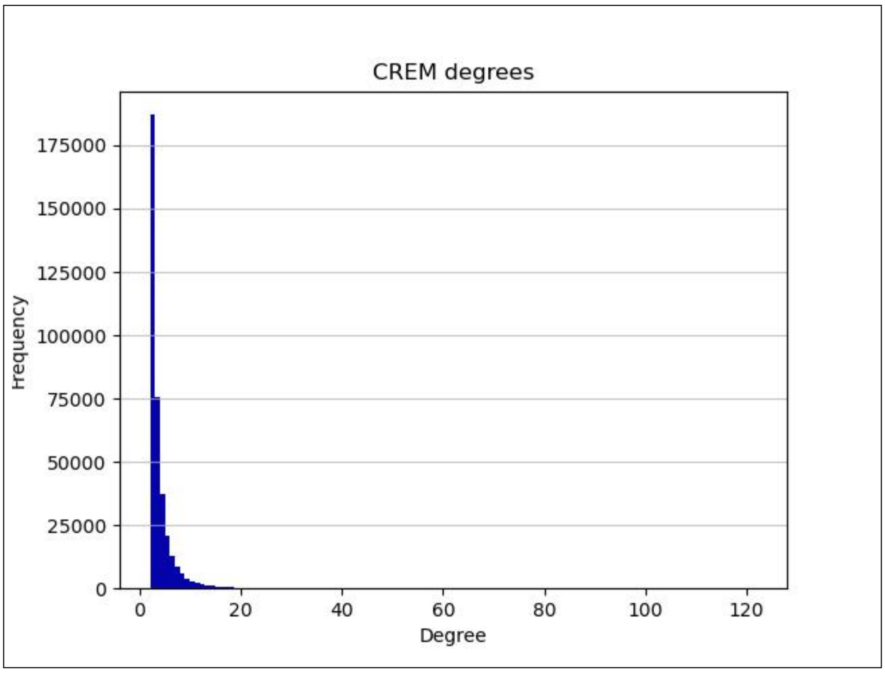


Figure A.1: Distribution of CREM degrees in GRN.

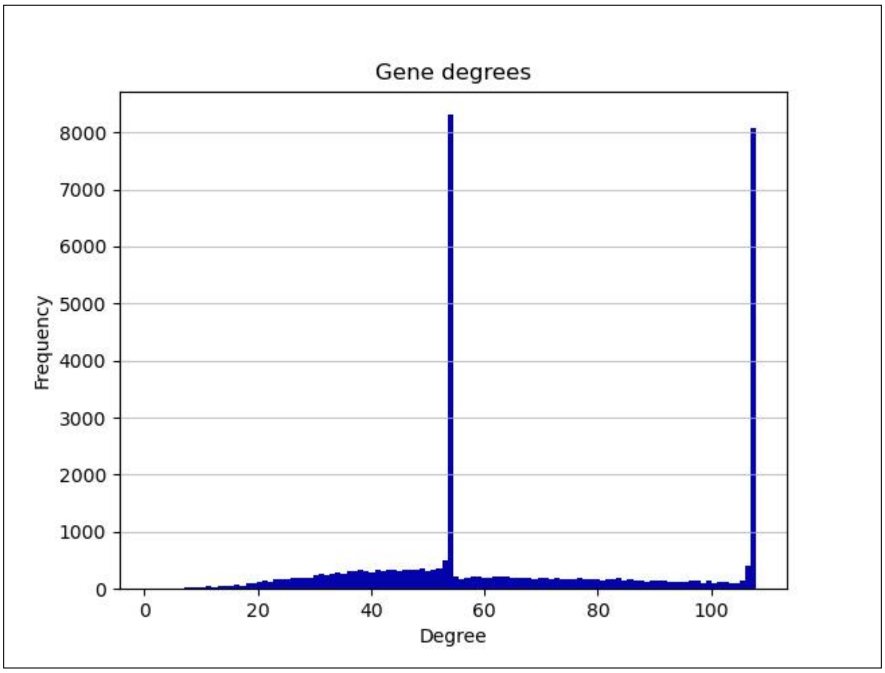


Figure A.2: Distribution of gene degrees in GRN.

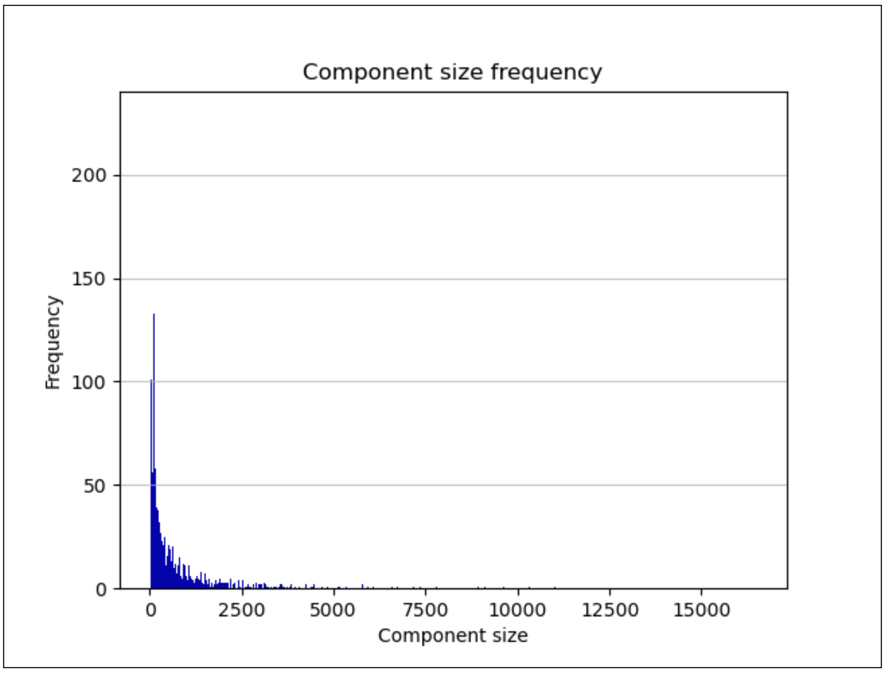


Figure A.3: Frequency of all component sizes in the network.

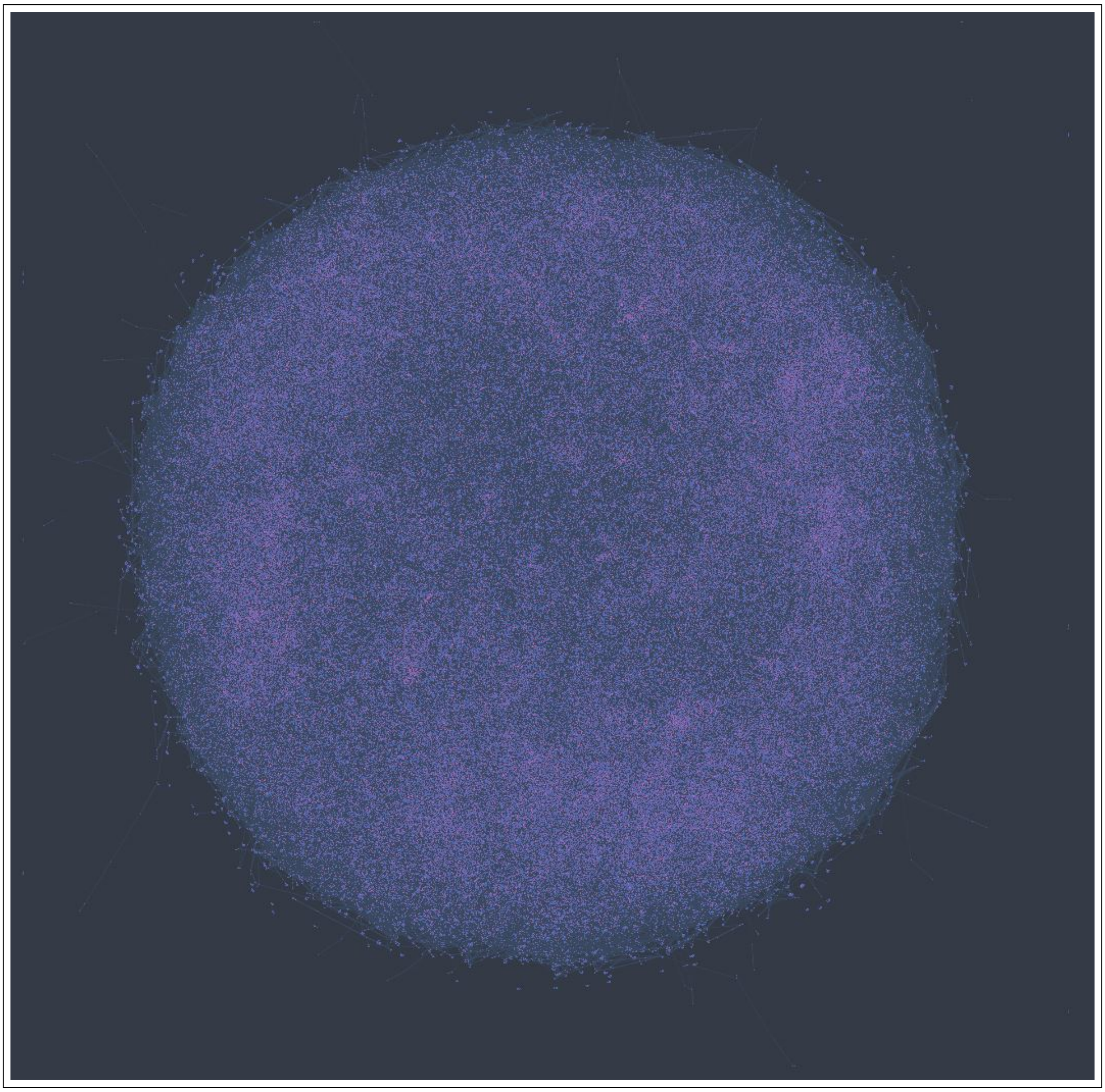


Figure A.4: Visualization of the entire with all nodes and edges with cosmograph[33].

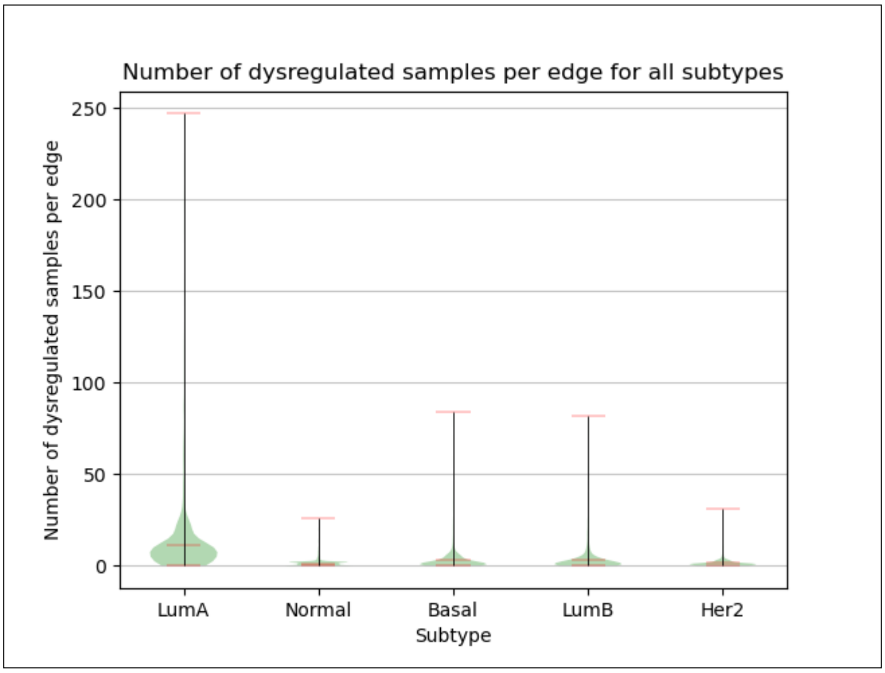


Figure A.5: Number of dysregulated samples per edge in the DysRegNet result. Grouped by BRCA subtypes.