Faculty of Science, Technology and Communication

Master in Integrated Systems Biology

MASTER thesis

by

Ganna Androsova
Born on 14 February 1991 in Ukraine

Network-based Approach Enabling Drug Repositioning for the Treatment of Myocardial Infarction

UNIVERSITÉ DU LUXEMBOURG

Faculty of Science, Technology and Communication

Master in Integrated Systems Biology

MASTER thesis

by

Ganna Androsova
Born on 14 February 1991 in Ukraine

Network-based Approach Enabling Drug Repositioning for the Treatment of Myocardial Infarction

Defense: 17 July 2014 in Luxembourg

Supervisor(s): Dr. Francisco Azuaje, CRP-Santé (Luxembourg)
Dr. Reinhard Schneider, LCSB (Luxembourg)

Jury members: Dr. Petr Nazarov, CRP-Santé (Luxembourg)
Dr. Roland Krause, LCSB (Luxembourg)

Acknowledgements

This research has been carried out as part of the INFUSED project at the NORLUX Neuro-Oncology Laboratory, Oncology Department, Centre de Recherche Public de la Santé (CRP-Santé). The Master degree was supported by the University of Luxembourg during 2012-2014 academic years.

I would like to thank my supervisor, Francisco Azuaje, for his guidance, inspiration and patience. The greatest benefit of this Master internship was to obtain skills that should fundamentally help my future research career. I would like to thank Francisco for trusting me in implementing methods and conducting research. I felt that each result might serve as a basis for other people to continue research, which made me strive for the greatest level of attention.

I would also like to thank our biologists, Sophie Rodius and Céline Jeanty who were kind enough to clarify all the unknown biological terms and processes. They have greatly contributed to the direction of my investigations.

Further thanks should be addressed to Petr Nazarov, whom I had a common language of Statistics with. I am grateful for his questions and ideas that gave rise to more motivation in my daily work and implementation of more advanced methods of analysis. I would like to thank as well the other members of the Genomics Research Unit of CRP-Santé.

I would also like to thank my internal supervisor, Reinhard Schneider, for giving his opinion about the results and suggesting additional ideas for analysis.

I would like to thank my course director who gave rise to this Master program that combined my knowledge in Biology and my passion for programming. He assisted me in all the organizational questions and had only the best interest for our future as graduates of the program.

Table of Contents

Abstract

Despite a notable reduction in incidence of acute myocardial infarction (MI), patients who experience it remain at risk for premature death and cardiac malfunction. The human cardiomyocytes are not able to achieve extensive regeneration upon MI. Remarkably, the adult zebrafish is able to achieve complete heart regeneration following amputation, cryoinjury or genetic ablation. This raises new potential opportunities on how to boost the heart healing capacity in humans. The objective of our research is to characterize the transcriptional network of the zebrafish heart regeneration, to describe underlying regulatory mechanisms, and to identify potential drugs capable to boost heart regeneration capacity.

Having identified the gene co-expression patterns in the data from a zebrafish cryoinjury model, we constructed a weighted gene co-expression network. To detect candidate functional sub-networks (modules), we used two different network clustering approaches: a density-based (ClusterONE) and a topological overlap-based (Dynamic Hybrid) algorithms.

We identified eighteen distinct modules associated with heart recovery upon cryoinjury. Functional enrichment analysis displayed that the modules are involved in different cellular processes crucial for heart regeneration, including: cell fate specification (p-value < 0.006) and migration (p-value < 0.047), cardiac cell differentiation (p-value < 3E-04), and various signaling events (p-value < 0.037). The visualization of the modules' expression profiles confirmed the relevance of these functional enrichments. Among the candidate hub genes detected in the network, there are genes relevant to atherosclerosis treatment and inflammation during cardiac arrest. Among the top candidate drugs, there were drugs already reported to play therapeutic roles in heart disease, though the majority of the drugs have not been considered yet for myocardial infarction treatment.

In conclusion, our findings provide insights into the complex regulatory mechanisms involved during heart regeneration in the zebrafish. These data will be useful for modelling specific network-based responses to heart injury, and for finding sensitive network points that may trigger or boost heart regeneration in the zebrafish, and possibly in mammals.

List of Abbreviations

BH	Benjamini-Hochberg
CD	candidate drug
CMap	Connectivity Map
CRP-Santé	Centre de Recherche Public de la Santé
CSRM	cubic spline regression model
CVD	cardiovascular disease
DEG	differentially expressed gene
dpi	day post-injury
FC	fold change
FDR	false discovery rate
GC	guanine-cytosine
hpi	hour post-injury
IMP	Integrative multi-species prediction
IQR	inter-quartile range
MI	myocardial infarction
MInfo	mutual information
ML	maximum likelihood
PCA	principle component analysis
STITCH	search tool for interacting chemicals
TO	topological overlap
TOM	topological overlap measure
WGCNA	weighted gene co-expression network analysis
WNC	weighted node connectivity

I. Introduction

1.1. Background

Cardiovascular disease (CVDs) is the main cause of mortality in Europe (Council of the European Union, 2006). Despite major advances in prevention, diagnosis and treatment measures for CVDs, it accounts for over 4 million deaths yearly, and about one-third of deaths in Luxembourg (Direction de la Santé, 2009). This outcome places cardiovascular health as a top priority of Europe and Luxembourg.

Often the result of coronary heart disease is myocardial infarction (MI). It causes irreparable damage to the hypoxic muscle, when blood flow in a segment of human heart becomes interrupted. Upon myocardial infarction, ischemic damage leads to apoptosis of cardiomyocytes. The heart replaces the infarcted myocardium with a non-contractile scar tissue instead of new muscle (Jennings et al., 1990). Such scar is beneficial over short term by preventing ventricular wall rapture. However, accumulation of the fibrotic scar tissue leads to adverse heart ventricle remodeling, cardiac hypertrophy, arrhythmias, and ultimately heart failure. Although a notable reduction in outcome and incidence of acute MI was achieved, patients that experience myocardial infarction remain at the risk of premature death and cardiac malfunction (Sidney et al., 2013). Therefore, novel therapeutic strategies are required for stimulating cardiac regeneration in humans.

1.2. Zebrafish as a model for heart regeneration

Although human cardiomyocytes have limited capacity for self-renewal, they are not able to achieve a complete regeneration after MI (Laflamme and Murry, 2011; Sedmera and Thompson, 2011). Remarkably, adult zebrafish achieves complete regeneration of heart following amputation, cryoinjury, or genetic ablation (Gonzalez-Rosa et al., 2011; Poss et al., 2002; Wang et al., 2011). Regarding the differences between human and zebrafish, the later has a two-chambered heart with one atrium and one ventricle, in contrast to the four-chambered human heart. Nonetheless, mammals and zebrafish have many striking similarities in embryonic morphogenesis and heart structure. A direct comparison of human and zebrafish protein-coding genes reveals that 71.4% of human genes have at least one zebrafish orthologue. Reciprocally, 69% of zebrafish genes have at least one human orthologue (Howe et al., 2013). This fact makes zebrafish a suitable organism for modelling myocardial infarction and observation of heart regeneration.

Experiments had shown that after amputation of ~20% of a cardiac ventricle, zebrafish heart is able to regenerate (Poss et al., 2002). The complete regrowth of the amputated region can be achieved at the 60th day post-amputation (Poss et al., 2002). However, ventricular resection does not fully resembles the mammalian myocardial infarction. The models based on tissue damage, rather than tissue removal, are more clinically meaningful. Cryoinjury damages around 25% of the ventricle and causes necrotic and apoptotic death, which resembles human cell death induced by ischemic damage (Gonzalez-Rosa and Mercader, 2012). Detailed steps of cardiac cryoinjury are shown on the Figure 1. The cryoinjured tissue is gradually substituted by fibrotic tissue, which is similar to human post-MI process. Unlike to post-MI process in mammals, zebrafish achieves the complete regeneration by replacement of the fibrotic tissue with newly formed cardiac tissue (Gonzalez-Rosa et al., 2011).

We applied a network-based approach to the data measured at different time points of the zebrafish heart regeneration (Gonzalez-Rosa et al., 2011). At 4hours after injury, one can observe massive cell death. This time point represents the early post-injury response, when thicken epicardium reactivates the gene expression involved in heart development (Gonzalez-Rosa et al., 2011). From the 1st and up to 4th, day cells extensively proliferate particularly in endocardium and myocardium regions. This step is required for regeneration of the damaged tissue. On the 7th day, one can observe the maximum extent of the formed transient scar. It gradually regresses until complete disappearance, which can be observed on the 130th day of post-injury (Gonzalez-Rosa et al., 2011).

1.3. The INFUSED project

The INFUSED project aims to improve understanding of the molecular systems underlying heart regeneration in the zebrafish and to discover new therapeutic strategies that, in the long term, could boost the heart regeneration capacity upon injury in humans. It is hosted at the Norlux Neuro-Oncology Laboratory at CRP-Santé in Luxembourg, in partnership with the Swiss Institute of Bioinformatics and Vital IT, the Genomics Research Unit of CRP-Santé.

To achieve this, a first step is to identify the key drivers of the heart regeneration in the zebrafish and their corresponding biological interaction networks. Then, to match the findings to chemical compounds that can accelerate/enhance the heart regeneration process in an *in vivo* model.

The top candidate targets and compounds will be validated on an *in vivo* model of myocardial infarction in zebrafish.

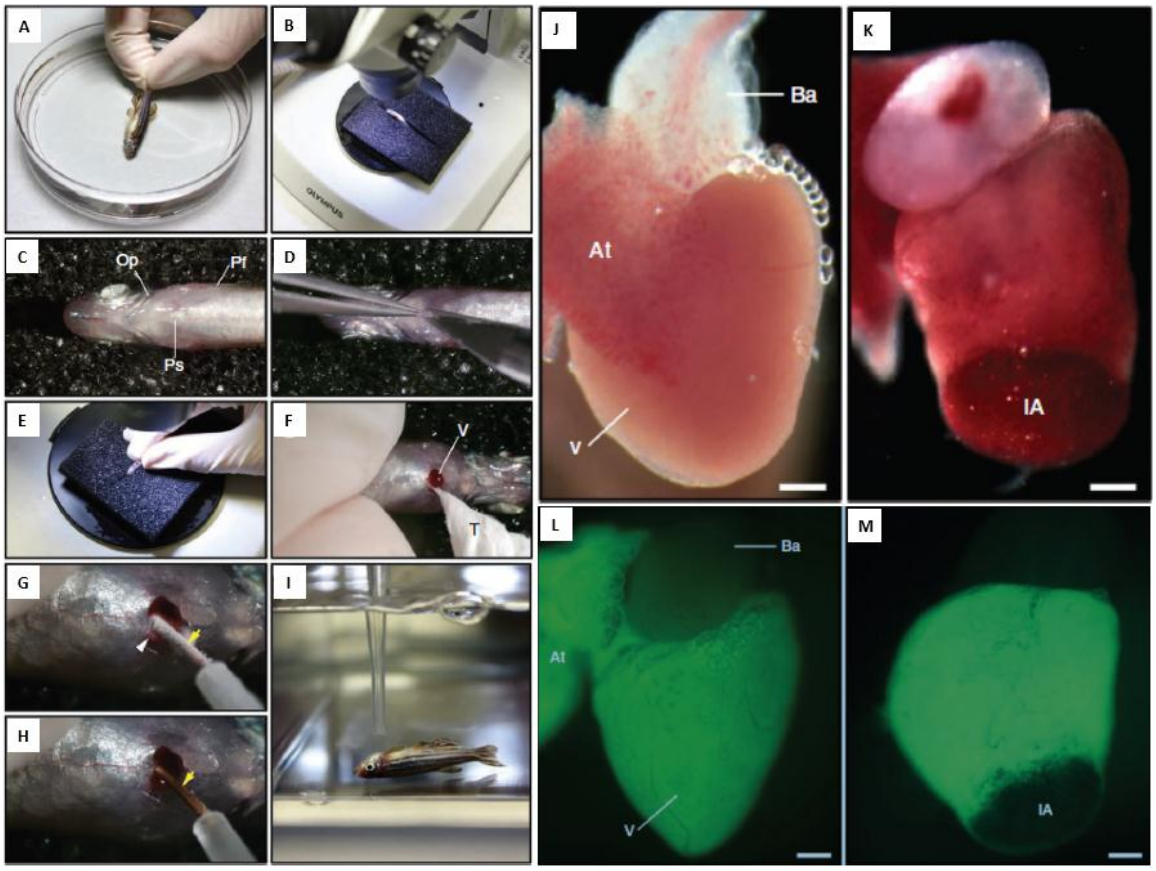


Figure 1. Cardiac cryoinjury in adult zebrafish. A) Anesthetized zebrafish. B) Zebrafish is placed ventral side up in holder under dissecting microscope. C) Zoomed view of the zebrafish in B. D) Making incision with forceps and microdissection scissors. E) After incision position thumb and index finger to compress abdomen. F) Drying heart ventricle with tissue. G) Cryoinjury of ventricle (white arrow) by frozen cryoprobe (yellow arrow). H) Cryoprobe is thawed and removed. I) Reanimation of zebrafish by gentle pipetting water onto the gills. (J-M) Effect of cryoinjury on the zebrafish heart ventricle. Whole-mount view of dissected control (J) and freshly cryoinjured (K) heart. Whole-mount view of dissected control (L) and freshly cryoinjured (M) heart with staining Tg(*cmlc2:GFP*) for visualization of GFP expression under a myocardial-specific promoter. L) Absence of myocardial staining in the injured area. Op, operculum; Pf, pectoral fins; Ps, pericardial sac; T, tissue; V, ventricle; At, atrium; Ba, bulbus arteriosus; IA, injury area. Adapted from (Gonzalez-Rosa and Mercader, 2012).

1.4. Data filtering approach

Filtering methods can be used to reduce amount of non-informative genes, which results in increased power to identify truly differentially expressed genes (DEGs). Common filtering methods include filtering by average signal intensity, variance, or detection of the "Present/Absent" calls (Bourgon et al., 2010; Hackstadt and Hess, 2009).

Filtering by mean signal intensity removes genes that have a signal close to background noise. The goal of filtering by overall variance is to remove all the genes with low variance across the array ignoring treatment conditions. The principle of this filtering approach is that the expression for equally expressed genes should not vary greatly between the control/treatment conditions, thus small overall variance is expected. Consequently, low variability indicates little information for any classification problem. Therefore filtering by variance removes genes with a low variance. MAS 5.0 detection call method is based on the use of the Wilcoxon Signed Rank test, which compares Perfect Match (PM) and Mismatch (MM) probes within the probe pair. Such comparison makes a "call" of Present or Absent for each probe. The rationale of this method is that if a transcript is not present in any of the samples, then there is no differential expression (Hackstadt and Hess, 2009). Thus, "Absent" transcripts are filtered out by MAS detection.

Each method has its advantages and limitations, which dependent on the type of dataset to be processed. One of our aims is to detect differentially expressed genes (DEGs) taking into account statistical significance. To achieve this, we used filtering approach proposed by Hackstadt and Hess (2009). They proved that filtering by variance on "original" scale, coupled with false discovery rate (FDR) control has many advantages: i) it increases power to detect DEGs, ii) it often decreases false discovery rate, iii) its performance does not depend on the assumption of the dependence or independence among genes or the proportion of differentially expressed genes in the dataset, iv) it proved to be more effective over a wide range of adjusted p-value cutoffs than filtering by mean values, v) variance filtering on the "original" scale data results in an increase of identified DEGs than filtering on the log2-transformed scale. Besides that, filtering by overall variance has been widely applied in other studies for co-expression network reconstruction (Kadarmideen and Watson-Haigh, 2012; Presson et al., 2008).

In order to filter by variance, a filtering threshold is required. Different papers state the variance threshold depending on overall variance or specific variance requirements. For example, Presson et al. (2008) filtered genes with variance above 66%,Calza et al. (2007) excluded 60% of the least varying genes. However, Hackstadt and Hess (2009) reported

that FDR power to identify differentially expressed genes was improved if the cutoff of the variance filtering was approximately close to the proportion of the truly null hypotheses in the dataset. The proportion of truly null hypotheses is regarded as the amount of genes in the dataset that are expected to be non-differentially expressed. Thus, a first step would be to determine the proportion of truly null hypotheses in the data.

Among the procedures to control/estimate FDR, the most common are Benjamini-Hochberg (BH) (Benjamini and Hochberg, 1995) and q-value method (Storey and Tibshirani, 2003). Both methods are based on the assumption that distribution of p-values follows a uniform distribution on [0, 1] interval. Q-value method adjusts raw p-values, resulting in q-values (details in Materials & Methods section).

Interestingly, filtering by variance coupled with q-value method results in bigger amount of identified DEGs, than filtering by variance coupled with BH (Hackstadt and Hess, 2009). Based on this reason, we used q-value approach to control false discovery rate (see Materials & Methods section).

1.5. Gene expression similarity measures

Gene co-expression methods are able to identify concordance of expression patterns between genes with biological and statistical value. It was proven that gene expression data are intrinsically noisy (Ingram et al., 2008). That explains why such data demand robust methods for biological patterns recognition. Hereafter we discuss the most commonly used methods and their possible alternatives.

Gene co-expression measures can be classified into two categories. The first category is able to detect linear and non-linear monotonic relationships. To this category belong correlation coefficients such as Pearson's (Pearson, 1920), Spearman's (Spearman, 1904) and Kendall's (Kendall, 1938). The second category is capable of detecting non-linear monotonic and non-monotonic associations. Mutual information (MInfo) (Qiu et al., 2009) and regression based models (Song et al., 2012) belong to this category. For construction of the global co-expression networks, both mutual information- and correlation-based methods performed equally well (Allen et al., 2012). Below we discuss specific differences as well as advantages and disadvantages of each of the above-mentioned correlation methods.

Unlike the second category, all correlation coefficients include the sign of correlation, which allows indicating positive and negative relationships between the variables.

Correlation coefficients result in accurate estimations with relatively few observations and retain biologically relevant hub genes and their connections within sub-networks (Kadarmideen and Watson-Haigh, 2012).

Pearson's correlation measures the degree of the linear dependence between two continuous random variables regardless of whether bivariate distribution is normal. This correlation measure provides complete and accurate description of the concordance but it is susceptible to outliers (observations that significantly deviates from the rest of the data) (Devlin et al., 1975; Hardin et al., 2007). Thus, for small sample sizes ($n < 20$) it is recommended to use Spearman's correlation (Song et al., 2012). With the increase of sample size, Spearman's performance is very similar to Pearson's (Song et al., 2012).

Spearman's and Kendall's co-expression measures are nonparametric rank-based methods, which use ranks of expression values instead of original values for analysis. They provide robust measure of a monotonic relationship between two continuous random variables. They are not susceptible to outliers (Hardin et al., 2007) and can be applied to ordinal data. For these reasons they are particularly suitable to identify genes that increase or decrease in a monotonic way. It was argued (Newson, 2002) that Kendall's parameters are more reliable and interpretable than Spearman's *rho* test. Though, Spearman's correlation successfully finds patterns from noisy gene expression data due to its robustness. Both Spearman's and Kendall's measure have the following advantages (Kumari et al., 2012): i) high power in identification of linear and non-linear monotonic relationships; ii) higher sensitivity, specificity, and predictive accuracy than Pearson's correlation coefficient; iii) identified network modules are significantly enriched with Gene Ontology terms. Their power depends on sample size, pre-assigned significance level, magnitude of trend and variation within the time course. This means the increase in sample size and absolute magnitude of trend improves performance of Spearman's and Kendall's method but the increase in variation with time series acts the opposite way (Newson, 2002). Similarly to all correlation coefficients, Spearman's and Kendall's measures are not capable of identifying non-linear, non-monotonic relationships in data (de Siqueira Santos et al., 2013). However, such type of problem is handled well by the mutual information and regression models, which are discussed below.

Mutual information is an information-theoretic approach that measures non-linear monotonic/non-monotonic associations between discrete or categorical variables (Daub et al., 2004). The disadvantages of the mutual information include: susceptibility to outliers; dependence on the sample size (decrease of power is proportional to the

decrease of the observations number); potential computationally intensive estimation of MInfo between quantitative variables (Song et al., 2012).

Song et al. (2012) had shown that polynomial and spline regression model based statistical approaches can be superior alternatives to mutual information. They are computationally simpler and faster than MI, easier in calculation and able to include additional covariates for analysis (Song et al., 2012). Both polynomial and cubic spline regression models correctly capture non-linear trends. The obtained results include model fitting indices and statistical tests that can be used for further statistical procedures. Despite the similarities between the two regression models, a clear advantage of spline regression model has been stated. In fact polynomial regression fit can be adversely affected by outliers, which bend the fitting curve in the "wrong" direction. The consequence of such influence is affected estimation of β coefficients. Spline regression overcomes mentioned above problem by fitting model on sub-intervals of the variables range (Song et al., 2012).

1.6. Construction of the weighted gene co-expression networks

Co-expression measurements are further used for construction of the co-expression networks. Here, the co-expression network is referred as undirected, weighted gene network. The nodes of such network correspond to genes, and edges between nodes are the pairwise correlations between the genes. Calculated gene co-expressions are organized into a correlation matrix. Such matrix is further transformed into a weighted co-expression network by applying soft threshold β. Soft thresholding is achieved through raising each value of the correlation matrix to power β. It helps to retain the continuous nature of similarity information. Soft thresholding leads to highly robust results and allows a simple geometric interpretation of network topological properties (Horvath and Dong, 2008; Langfelder and Horvath, 2008).

Since we do not neglect negative correlations, the constructed network type is called "unsigned". The unsigned network keeps both positive and negative correlations as edges between the nodes, while signed network would neglect them and nodes would be considered unconnected (Song et al., 2012). The details of gene co-expression network reconstruction are explained in Materials & Methods section.

1.7. Fundamental network properties

We further evaluate constructed gene co-expression network by fundamental network properties (topological parameters). Such network properties include connectivity, density, centralization, heterogeneity, and clustering coefficient (Horvath and Dong, 2008). These network properties enable us to access topology and functional parameters of the network. Below, we discuss each fundamental concept.

Connectivity is a synonym of node degree distribution. It denotes the sum of edges' weights between a specific gene and other genes. Thus, a higher connectivity defines a stronger correlation connectivity of the gene. This allows to distinguish nodes in the network and to identify hub genes. Hub genes are potential key regulators of the network. They are connected to a significant number of nodes in the network, thus they are of biological interest (Horvath and Dong, 2008).

Another network parameter is *density*. It is based on the sum of the connectivity with respect to size of the network. Network density represents the overall influence among the nodes. High density value is an indication that nodes are strongly connected in the network (Horvath and Dong, 2008).

A complementary term to connectivity is network *centralization*. It is able to reveal if highly connected hub gene forms a center in the network. For example, centralization value of 1 denotes that node A is connected to all other nodes that are not connected to each other, thus node A is center of the network (Horvath and Dong, 2008).

The network *heterogeneity* measures the variation of connectivity patterns between the nodes (Horvath and Dong, 2008). Barabasi and Oltvai (2004) had shown that networks, exhibiting an approximate scale-free topology, are very heterogeneous.

Clustering coefficient measures the cohesiveness of a node's neighborhood. Cohesiveness is a tendency of any node to connect to each other. Thus, clustering coefficient is useful determinant of the node "affection" among its neighbors. It also describes the structural (hierarchical) properties of the network (Horvath and Dong, 2008).

1.8. Module detection approaches

A module (also called a “biological cluster”) is a group of densely interconnected nodes of the network. Genes that participate in a specific pathway or biological function usually form dense regions in the network. Thus, detection of such regions might reveal the modules that regulate specific or multiple biological processes.

There are two types of the clustering algorithms for module identification: network clustering algorithms and attribute clustering algorithms (Morris et al., 2011). The attribute clustering groups the nodes based on similarity of nodes’ or edges’ attributes. The network clustering algorithms find densely connected regions by a local approach, starting with a node neighborhood, or a global approach, starting with the entire graph, and iteratively, partitioning network into the clusters (Morris et al., 2011).

Network hierarchical clustering is an attribute clustering algorithm for detecting modules of closely related genes. It was shown that unsupervised hierarchical clustering based on topological overlap measure (TOM) followed by Dynamic Branch Cut method performs well in module detection (Li and Horvath, 2007; Yip and Horvath, 2007; Zhang and Horvath, 2005). The identified modules by this method were highly significantly enriched with known gene ontology terms (Dong and Horvath, 2007). For the module identification, we first measure interconnectedness between genes with a topological overlap measure. Based on the TOMs values, hierarchical clustering organizes genes into a dendrogram with distinct branches corresponding to modules. The drawback of the hierarchical clustering is the difficulty in determination how many modules are present in dataset. Identification of the modules from the hierarchical dendrogram is a challenging problem for most of the fixed cutting threshold methods. Thus, a Dynamic Branch Cut method, based on the dendrogram branches shape, was proposed that is called Dynamic Hybrid algorithm (Langfelder and Horvath, 2008). This method offers such advantages compared to fixed cut height methods: i) capability of identifying nested clusters, ii) flexibility due to the tuning cluster shape parameters, iii) suitability for automation, iv) combination of the advantages of hierarchical clustering and partitioning around medoids that leads to better detection of the outliers (Langfelder and Horvath, 2008). Although, Dynamic Hybrid algorithm provides improved flexibility for module identification, choosing an optimal cutting parameters remains an open question.

ClusterONE (Clustering with Overlapping Neighborhood Expansion) was reported to exhibit better performance than other network clustering approaches, matching more modules with a higher accuracy (Nepusz et al., 2012). It was shown that proteins can

participate in more than one complex (Pu et al., 2009), and the same rule applies to the genes. ClusterONE is capable to detect overlapping modules as well as handle weighted connections between the nodes. Identified modules by ClusterONE tended to consist of proteins in the same cellular component and likely to have similar functions or participate in the same biological process (Nepusz et al., 2012).

Due to the different strong points between hierarchical clustering combined with Dynamic Hybrid algorithm and ClusterONE, we were interested to compare the identified modules by both methods. The comparison includes the statistical intramodular properties, Jaccard similarity index, and predicted functional enrichment of the modules. We describe the details of comparison approaches in the Materials & Methods section.

1.9. Drug repositioning for the treatment of myocardial infarction

Investigation of molecular basis of disease provide opportunity to translate the findings into medical treatments. However, introduction of a new drug to the market typically takes 10-15 year of enormous efforts and more than 1 billion US dollars (Sleigh and Barton, 2010). A new drug must be shown to be effective, safe and potent.

A cost-effective solution is drug repositioning. Drug repositioning refers to investigation of the existing (approved) drugs/chemical compounds for novel indications/applications (Sleigh and Barton, 2010). Since the approved drugs have been optimized for safety and efficacy, and tested in humans, there is available detailed information on their formulation, pharmacology and potential toxicity. As drug repositioning is based on previous research and development efforts, thus new candidate compounds may be rapidly ready for clinical trials and integration into health care.

This project provides results that will enable drug repositioning for the treatment of myocardial infarction. Drug repositioning in cardiovascular diseases was not widely investigated, giving more opportunities for new discoveries. It starts with identification of the key targets that can boost heart regeneration process upon injury. Application of the network-based approach might reveal not only candidate drug targets, but also possible effects on down-stream processes, targeted pathways, and network perturbation effects. Also the findings from network-based drug repositioning may detect undesired drug-target interactions (Azuaje et al., 2011).

Having identified the hub genes and important modules in zebrafish heart regeneration, we can proceed to detection of chemical compounds/drugs, capable to target hub genes

and/or induce the module-specific signatures. A module-specific signature is a ranked list of significantly DEGs. The highest up-regulated genes are in the top of the signature and the greatest down-regulated genes are in the bottom of the signature. Such module-specific signature can be queried through the Connectivity Map (CMap), a comprehensive and constantly updating database of the genomic profiles of many drugs vs. controls (Lamb et al., 2006).

Another database of chemical compound-target associations is STITCH ('search tool for interacting chemicals'). It combines the sources of chemical compound-target interactions from experimental databases, drug–target databases, text mining results, and pathway databases and predictions (Kuhn et al., 2012). Both, CMap and STITCH, might indicate potential drugs/chemical compounds that are capable to boost or perturn the heart regeneration process in zebrafish.

1.10. Aims of the study

The main purpose of this study is to characterize the transcriptional network of the zebrafish heart regeneration and underlying regulatory mechanisms enabling drug repositioning.

The specific objectives of our research are:

1. to compare the performance of different gene expression similarity measures for the network re-construction;
2. to identify the regulatory sub-structures of the network (modules) by two different network clustering approaches: a density-based (ClusterONE) and a topological overlap-based (Hybrid Dynamic) algorithms;
3. to compare the topological parameters and similarities between the detected modules by different approaches, and to characterize the functional enrichment of the modules;
4. to identify the modules and regulatory hub genes responsible for the zebrafish heart regeneration process;
5. to visualize the dynamics of the regeneration process in the zebrafish heart by different approaches;
6. to identify relevant biological pathways at each post-cryoinjury time point;
7. to map the most promising modules to the drug/chemical compounds association databases in order to detect possible enhancers or effectors of the heart regeneration process.

II. Materials and Methods

2.1. Microarray data source and pre-processing

Experiments in the zebrafish were conducted at the Norlux Laboratory of CRP-Santé, Luxembourg. The microarray data were generated and pre-processed at the Genomics Research Unit of CRP-Santé. The resulting dataset contains 27 samples of biological triplicates for 3 control conditions (Sham1, Sham2, Sham3) and 6 post-cryoinjury time points. In the case of cryoinjury samples, hearts were dissected and ventricles were recovered 4h, 1 day, 3 days, 7 days, 14 days or 90 days post-injury. The cryoinjury procedure was performed according to (Gonzalez-Rosa and Mercader, 2012). Sham1 is the control of repetition of the whole cryoinjury procedure except that cryoprobe was not frozen. Sham2 is the control of incision through body wall and pericardial sac. For both Sham1 and Sham2 the hearts were dissected at 4 hours post-injury (hpi). Sham3 is the control of dissection of zebrafish that aged 90 days: healthy wild-type fish, untouched. Five ventricles were pooled for each sample.

The original dataset with 27 samples (75,212 genes) was normalized by Robust Multi-Array Average method with GC (guanine-cytosine base pairs) correction in Partek software (http://www.partek.com). The dataset was filtered by intensity levels exceeding background noise (intensity threshold above 6) (Nazarov et al., 2013).

2.2. Estimating the proportion of truly null hypotheses and q-values

For determination of the differential expression, we have to estimate p-values by two-sample t-statistics with an equal variance. For this we have two treat data as two groups: “reference” and “damaged” states. As reference state we considered Sham3 and 90 days post-injury (dpi) samples. Damaged state included Sham1, Sham2 and post-cryoinjury time points from 4 hours to 14 days. The approach of such data division into two states resulted from the principle component analysis (Appendix Figure 1) and differential expression assessment (Appendix Figure 2) provided by the Genomics Research Unit of CRP-Santé, Luxembourg.

P-values were calculated using *multtest* R package (Pollard et al., 2005). Raw p-values were adjusted by q-value statistics, resulting in adjusted p-values. For a specific feature (gene), q-value is the expected proportion of false positives among all the features that are as extreme as or more extreme than the one observed. For m, number of tested hypotheses, the obtained p-values are: $p_1, p_2, \ldots, p_m$. Assuming that p-values are uniformly

distributed under the null hypotheses, the proportion of truly null features ($\hat{\pi}_0$) can be quantified by

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda; i = 1, \ldots m\}}{m(1-\lambda)},$$

where λ is tuning parameter. p-values of truly alternative features will tend to be close to zero.

2.3. Filtering by variance

Filtering by overall variance removes the genes with low variability across all the arrays ignoring treatment assignments. To assess the variance of transcript cluster across all the samples, we used inter-quartile range (IQR). First, the variance of expression values are calculated for each transcript cluster and then the transcript clusters are ranked by their means. The transcript clusters falling below the established threshold are filtered out. Using (Hackstadt and Hess, 2009) approach, we approximated the variance threshold to the estimated proportion of the truly null hypotheses.

We apply filtering procedure to 20,104 gene set by variance exceeding the proportion of the truly null hypotheses on the original scale with with *genefilter* package (Gentelman et al., 2009).

2.4. FDR estimation by Storey and Tibshirani approach

Having filtered by variance on “original” scale, we obtained dataset containing 12,263 genes. Our next step was to estimate the amount of differentially expressed genes in this dataset. Detection of DEGs usually is accompanied by the indication of the statistical relevance of the findings. False discovery rate is one of such statistical measures. For *m* number of tested hypotheses and calculated p-value (p_i) for feature *i*, with respect to the proportion of truly null hypotheses, q-value considering threshold $t \geq p_i$ is calculated as

$$\hat{q}_i(p_i) = \min_{t \geq p_i} \frac{\hat{\pi}_0 \cdot m \cdot t}{\#\{p_i \leq t\}},$$

where *t* is the threshold at which FDR is controlled (Storey and Tibshirani, 2003). q-values were calculated by *qvalue* package (Dabney and Storey, 2014).

We considered several q-value threshold (FDR control levels): 0.05, 0.01, 0.005, and 0.001, which denote 5%, 1%, 0.5% and 0.1% of false positives in the data respectively.

2.5. Detection of the gene expression similarity patterns

2.5.1. Pearson's product-moment correlation coefficient

Pearson's correlation is a measure of association that detects linear dependence between two variables (Pearson, 1920). Being divided by the product of standard deviations, it results in the slope obtained by the linear regression of X and Y. For n observations from two random variables, let the mean of X be determined as $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ and the mean of Y would be $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$. Following our previous assumption, Pearson's correlation is defined as $$\rho_{Pearson}(X,Y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}.$$

Pearson's correlation assumes values between -1 and +1. Positive values describe increasing (positive) linear relationship and negative values describe decreasing linear relationship. The weak point of the Pearson's measure is that the interpretation of $-1 < \rho_{Pearson}(X,Y) < +1$ can assume non-correlated X and Y. In fact, these two variables may be associated in non-linear fashion, which Pearson's coefficient was unable to detect (de Siqueira Santos et al., 2013).

Pearson's correlation was calculated using function *cor* with parameter *method* = *"pearson"* using package *stats* (http://www.r-project.org).

2.5.2. Spearman's correlation coefficient

Spearman's rank correlation is a nonparametric measure of association that identifies non-linear monotonic relationships between two variables. It does not require assumptions of linearity, nor data measurements at the interval or ratio scales and can be used for ordinal variables. For calculation of $\rho_{Spearman}$ between two variables X and Y, raw values of x_i and y_i should be converted to ranks:

$$\rho_{Spearman}(X,Y) = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)},$$

where d_i is the difference between the ranks of the i^{th} observations of the two variables, and n is the number of pairs of values (Spearman, 1904).

Same as Pearson's correlation coefficient, Spearman's measure results in values between -1 and +1. A negative sign implies that with an increase of the rank by one variable – the rank of the other variable will decrease. A positive sign indicates positive correlation, e.g. with increase of rank by one variable – the rank of another variable will also increase (Spearman, 1904).

Spearman's correlation was calculated using function *cor* with parameter *method* = *"spearman"* from the *stats* package (http://www.r-project.org).

2.5.3. Kendall τ rank correlation coefficient

Kendall's correlation provides a statistical test to test the independence of two variables and is capable in identifying linear and non-linear monotonic relationships (Kendall, 1938). For a pair of randomly taken objects, τ is interpreted as the difference between the probability for these objects to be in the same order and the probability of these objects being in different order. For n observations from two random variables, the estimation of Kendall's correlation is the following:

$$\tau(X,Y) = \frac{(No\ of\ concordant\ pairs) - (No\ of\ discondant\ pairs)}{0.5n(n-1)}.$$

Concordant means that ranks of both elements agree in case both $x_i > x_j$ and $y_i > y_j$ or in case both $x_i < x_j$ and $y_i < y_j$. For the case when $x_i > x_j$ and $y_i < y_j$, or when $x_i < x_j$ and $y_i > y_j$, such pairs are called *discordant*. If $x_i = x_j$ and $y_i = y_j$, the pair is considered neither *concordant* nor *discordant* (Kendall, 1938).

Similarly to Spearman's correlation coefficient, Kendall's rank coefficient takes the values between [-1, 1] following the same interpretation (Kendall, 1938).

Kendall τ rank correlation coefficients were calculated using function *cor* with parameter *method* = *"kendall"* from the package *stats* (http://www.r-project.org).

2.5.4. Cubic spline regression model

Local spline regression model applies fitting on the subintervals of the range of x (Song et al., 2012). The boundaries of the subintervals are called knots. Let m be the number of observations, if m > 100 use 5 knots, if m < 30 use 3 knots, otherwise 4 knots. Taking into account the above mentioned determinants, cubic spline regression model with fitting polynomial of degree 3 to sub-intervals with n knots can be described as

$$E(y) = \beta_0 1 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{i=1,\dots,n} \beta_{3+i}(x - knot_i)_+^3\ ,$$

where $(x - knot1)_+$ is a vector i^{th} component of which is equal to $x[i] - knot1$ if $x[i] - knot1 \geq 0$ and 0 otherwise.

Next step is the calculation of the model fitting indices. For a set of m variables $x_1, x_2, \dots x_m$, model fitting indices, as elements of an m x m association matrix, are calculated as $R_{ij}^2 = R^2\ (x_i, x_j) = cor(x, M\hat{\beta})^2$, where R_{ij}^2 is an overall model fitting index for

pair of x_i and x_j. Parameter vector $\hat{\beta}$ is calculated by the least squares estimate ($\hat{\beta} = (M^{\tau}M)^{-}M^{\tau}y$), where $^{-}$ is the (pseudo) inverse, and $^{\tau}$ denotes transposed matrix. This method guarantees the smoothness of the regression line (Song et al., 2012).

Spline regression adjacencies were calculated by function *adjacency.splineReg* from the *WGCNA* package (Langfelder and Horvath, 2008).

2.5.5. Mutual information

As entropy estimation method we used the maximum likelihood (ML) estimator (also known as empirical estimator). The data was discretized by the default estimation of number of bins: $numberBins = sqrt(m)$, where m is the number of samples.
The connection between observed counts y_k and frequencies $\theta_k$is calculated by multinomial distribution (Hausser and Strimmer, 2009):

$$Prob(y_1, \dots, y_p;\ \theta_1, \dots, \theta_p) = \frac{n!}{\prod_{k=1}^{p} y_k!} \prod_{k=1}^{p} \theta_k^{y_k},\ \text{with } \theta_k > 0.$$

Mutual information was calculated by function *mutualInfoAdjacency* from *WGCNA* package (Langfelder and Horvath, 2008) with parameter *entropyEstimationMethod* = *"ML"*.

2.6. Construction of the weighted gene co-expression networks

Having determined the co-expression similarity between gene pairs by the above mentioned co-expression measures, we are able to define weighted network adjacency. Calculation of adjacent values preserves the continuous nature of the co-expression information (Zhang and Horvath, 2005). Similar to other publications (Ghazalpour et al., 2006; Horvath et al., 2006), our interest was to keep both strong positive and stong negative relationships as edges in the network. In order to keep such associations connected, we choose the unsigned type of the weighted adjacency matrix:

$$a_{ij} = \left|cor(x_i, x_j)\right|^{\beta},$$

where $\beta \geq 1$ is a soft threshold ($\beta = 6$ by default for unsigned adjacency). Such soft thresholding leads to a weighted gene co-expression network, emphasizing strong correlations and neglecting weak ones (Zhang and Horvath, 2005).

We used *adjacency* function with parameter *power = 6* implemented in *WGCNA* package (Langfelder and Horvath, 2008).

2.7. Estimating network topology properties

We estimated the fundamental network properties to compare the topology of the constructed networks (Horvath and Dong, 2008).

Following the definitions of weighted gene co-expression network construction, connectivity of the i^{th} gene is $k_i = \sum_{j \neq i} a_{ij}$. For comparison of connectivity between the networks, it should be scaled $K_i = \frac{k_i}{\max_j(k_j)}$, where K_i is the scaled connectivity between 0 and 1, $\max_j(k_j)$ is the maximum connectivity. The high values of the connectivity indicate that node i is well connected to the other nodes. To obtain a network/module connectivity, we calculate mean connectivity for all nodes in the network/module.

Network density is calculated by off-diagonal adjacency values as: $Density = \frac{\sum_i \sum_{j \neq i} a_{ij}}{n(n-1)}$. Density indicates if network/module is tight or cohesive.

Network centralization is also called a degree centralization (Freeman, 1978). It is calculated as $Centralization = \frac{n}{n-2}(\frac{k_{max}}{n-1} - Density)$, where k_{max} is the maximum connectivity. Centralization of 1 indicates that network has star topology (only one node is connected to all the rest nodes). In contrast, centralization of 0 represents that all nodes are equally connected to each other (Dong and Horvath, 2007).

Heterogeneity is a coefficient of variation of the connectivity distribution: $Heterogeneity = \frac{\sqrt{var(k)}}{mean(k)}$. High heterogeneity indicates that network has free-scale topology, whenever randomly constructed network would have heterogeneity of 0 (Dong and Horvath, 2007).

The local density measure is defined as clustering coefficient. It indicates the affection among the nodes. It is calculated as $ClusterCoefficient_i = \frac{\sum_{l \neq i} \sum_{m \neq i,l} a_{il} a_{lm} a_{mi}}{\{(\sum_{l \neq i} a_{il})^2 - \sum_{l \neq i} (a_{il})^2\}}$. Cluster coefficient takes a range of values between 0 and 1 (Dong and Horvath, 2007).

The all above mentioned properties were calculated by function *fundamentalNetworkConcepts* in *WGCNA* package (Langfelder and Horvath, 2008). The network connectivity was calculated as mean of all scaled connectivities of the network. The network clustering coefficient was calculated as mean of all clustering coefficients of the network.

2.8. Module detection

2.8.1. Hierarchical clustering combined with Dynamic Hybrid algorithm

The obtained adjacency matrix is further mapped to the topological overlap matrix (TOM). TOM serves as a filter that diminishes the effect of the weak and spurious connections,

which leads to more robust networks (Li and Horvath, 2007; Yip and Horvath, 2007). TOM based adjacency function (A_{TOM}) replaces the original adjacencies ($A^{original}$) by a measure of interconnectness based on shared neighbors:

$$A_{TOM}(A^{original})_{ij} = \frac{\sum_{l \neq i,j} A_{il}^{original} A_{jl}^{original} + A_{ij}^{original}}{\min\left(\sum_{l \neq i} A_{il}^{original}, \sum_{l \neq j} A_{jl}^{original}\right) - A_{ij}^{original} + 1}.$$

TOM based adjacency function is implemented in WGCNA R package as *TOMsimilarity* function (Langfelder and Horvath, 2008)

Next, hierarchical clustering iteratively merges two closest genes (determined by topological overlap dissimilarity: $TOMdissimilarity(i) = 1 - A_{TOM}(i)$) into a new composite object. Hierarchical clustering produces a dendrogram containing information on which objects were merged at each step and on their merging height (dissimilarity between the merged objects) (Langfelder and Horvath, 2008).

Dynamic Hybrid algorithm is applied for module detection from the gene dendrogram. It identifies modules as branches if: i) they contain a certain minimum number of objects; ii) they have excluded greatly distant genes (even belonging to the same branch); iii) they are distinct from their surroundings; iv) the core (tip of the branch) of the module is tightly connected (Langfelder and Horvath, 2008).

As mentioned above, there are no optimal parameters known that can be used in this approach. We followed the tutorials provided for *WGCNA* package (http://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/Tutorials) and set minimum module size = 30, *deepSplit = 2.* Such settings favor specificity, but does not neglect sensitivity of the module detection (Langfelder and Horvath, 2008).

2.8.2. ClusterONE

ClusterONE is clustering with overlapping neighborhood expansion method based on cohesiveness measure. According to Nepusz et al. (2012), a module should satisfy two structural properties: i) subunits of the module should have many reliable interactions; ii) module should be well separated from the rest of the network. These two properties form a quality measure called cohesiveness. Cohesiveness is calculated as $f(V) = \frac{w^{in}(V)}{w^{in}(V) + w^{bound}(V) + p|V|}$, where $w^{in}(V)$ is total weight of the edges contained entirely by a group of genes V, $w^{bound}(V)$ is the total weight of the edges connecting that group with the rest of the network, $p|V|$ is a penalty term that models uncertainty in the data, assuming the existence of yet undiscovered interactions in gene interaction network (Nepusz et al., 2012).

All identified modules have an associated p-value, which is calculated by one-sided Mann-Whitney U test on in- and out- weights of the module nodes. Low p-value indicates that in-weights are significantly larger than out-weights. P-value below 0.05 denote that module is statistically valid and is not a result of random fluctuations.

As it was suggested by authors (Nepusz et al., 2012), we used the default parameters of ClusterONE. For weighted networks, density threshold is 0.3 and merging threshold is 0.8. ClusterONE was implemented using its *Cytoscape* plugin (Shannon et al., 2003).

2.9. Jaccard similarity index

Jaccard index estimates the similarity between two sets of objects (Jaccard, 1901). Our aim was to evaluate the similarity between the modules identified by two approaches: Dynamic Hybrid algorithm and ClusterONE. Jaccard similarity index is calculated as follows:

$$Jaccard\ similarity\ index\ = \frac{size\ of\ overlap}{size\ of\ union\ of\ two\ modules}$$

For calculation of Jaccard index we used function *comparelists* (gives the intersection between the modules) and function *union* from the R package *made4* (Culhane et al., 2005). Each pair of modules (from Dynamic Hybrid algorithm and ClusterONE) is given as input for calculation of Jaccard index. Size of overlap and union were calculated as length of the resulting intersection and union correspondingly.

The resulting similarity indices we presented as color gradient matrix, created by function *labeledHeatmap* from *WGCNA* package (Langfelder and Horvath, 2008).

2.10. Prediction of functional enrichment

Due to the peculiarity of our data (significant proportion of non-annotated zebrafish genome in the network) we needed to widen functional enrichment analysis by accounting statistically significant predictions. The expansion of the gene sets is possible by mining functional relationships predicted from integrated high-throughput data. This principle is implemented in the Integrative Multi-species Prediction (IMP) database (Wong et al., 2012). It overlays the gene set of the module on functional networks, and expands gene set by mining functional relationships within the network. IMP identifies overrepresented pathways among genes of the module and predicted genes using annotations from Gene Ontology Consortium (Ashburner et al., 2000), MetaCyc (Caspi et al., 2012), KEGG database (Kotera M et al., 2012), Reactome (Matthews et al., 2009). Statistical

significance is calculated using hypergeometric distribution. Obtained p-values are multiple test corrected by controlling false discovery rate (Wong et al., 2012).

In IMP parameters we choose "*Organism – Danio Rerio*", inserted gene list of the module, checked the relevance of recognized input gene list and set a maximum number of predicted associated genes (50).

2.11. Visualization of the gene expression patterns during regeneration

To visualize network and module structure, network is transformed into a Cytoscape readable format (Shannon et al., 2003). For this we used function *exportNetworkToCytoscape* from *WGCNA* package (Langfelder and Horvath, 2008)). In order to make network visually comprehensive and not overloaded with weak connections, we filter the correlations below 0.799. Similar threshold was stated by (Borate et al., 2009), observing sharp increase of biological relationships above 0.8 correlation value. Since we are dealing with adjacency values, we recalculate a threshold as $a_{ij} = \left|cor_{ij}\right|^6 = 0.26$. For function *exportNetworkToCytoscape,* we set parameters *threshold = 0.26*, *weighted = TRUE* (because our network is weighted).

Next step was the calculation of the average values of gene expression per each sample. Then we calculated a fold change (FC) of gene expression per each time point relatively to Sham3. These values were passed to Cytoscape network as node attributes. We denoted the gene expression FC by gradation of colors: gene down-regulation (values below 1) in blue, no differential expression relatively to Sham3 (values close to 1) in white, and gene up-regulation (values above 1) in red. At each post-cryoinjury time point, we took the snapshot of gene expression in the network/module and created a GIF animation in Adobe Photoshop® CS6 software.

Another visualization approach also uses gene expression FC. The calculated FCs are further transformed into log2(FC) and plotted as a graph.

2.12. wiPER: a hub gene detection approach

wiPER analyses the observed connectivity values with statistical significance inference at the level of individual genes (nodes of the network) (Azuaje, 2014). The weighted node connectivity for gene i is calculated as

$$WNC_i = \sum_j^N w_{i,j},$$

where node i is connected to node j, and $w_{i,j}$ reflects the strength of the connection (correlation) between node i and node j. For each estimated WNC_i score p-value is

calculated with a permutation-based test that calculates the proportion of $WNC_i \geq WNC_{i,random}$, where the $WNC_{i,random}$ is an (empirical null) distribution of random WNC_i values obtained from thousands of permuted samples (Azuaje, 2014). Each p-value is adjusted with a Bonferroni correction to account for multiple testing. Here p-values were estimated with 10^5 random permutations.

2.13. Matching chemical compounds/drugs to the module profiles

2.13.1. Chemical compound association by STITCH

We analyzed all module profiles in STITCH (database state on May 19, 2014) (Kuhn et al., 2012). A set of zebrafish genes belonging to analyzed module served as an input to STITCH. A confidence score was set to 0.700 and up to 400 interactors were retrieved. Since many of the compounds were associated to several modules, we calculated the combined score per compound:

$$Comb\ Score_i = median\ (all\ STICH\ scores_i) * freq_i / max_{j \in \tau} freq_j,$$

where $Comb\ Score_i$ is calculated for compound i in list τ of all association results, $STICH\ score$ is the confidence score of compound i association with a module, $freq_i$ is the number of modules that compound i was associated to, $max_{j \in \tau} freq_j$ is maximum frequency of compound association to the modules in the list τ of all association results (for our results $max_{j \in \tau} freq_j = 27$). The combined scores were further normalized (Renda and Straccia, 2003) as $Norm\ Comb\ Score_i = \frac{Comb\ Score_i - min_{j \in \tau} Comb\ Score_j}{max_{j \in \tau} Comb\ Score_j - min_{j \in \tau} Comb\ Score_j}$. At the end of this procedure, we retrieve a list of chemical compounds associated with investigated network modules.

2.13.2. Identification of the candidate drugs by CMap

Since the first 4 hours upon cryoinjury are associated with major inflammatory processes, we consider that 1day post-injury is a starting point of heart regeneration. We hypothesize that inducing similar differential expression of genes as observed at 1 day post-injury, may boost heart regeneration capacity soon after damage.

Having calculated gene differential expression of 1 dpi/control condition (Sham3), genes were ordered from the highest up-regulation to the largest down-regulation and submitted to CMap query. If the drug was identified for several module-specific signatures, we calculated for each drug the mean values of enrichment, p-value, and specificity. The top drugs were identified by criteria: positive enrichment > 0.9 or negative enrichment < -0.9, ranking > 20, percent non-null = 100. Drugs showing positive enrichment are those drugs likely to induce the expression patter observed in the module.

III. Results

3.1. Filtering of the pre-processed dataset

We analyzed a dataset of 20,104 genes that contained annotated and non-annotated genes.

In order to estimate the proportion of null hypotheses, we calculated and adjusted p-values as described in the Materials & Methods section. The estimated proportion of the null hypotheses in the data is $\hat{\pi}_0$ = 0.399 (Appendix Figure 3B). Thus, we filtered out by variance ~39% of gene from the initial dataset.

The resulting dataset contained 12,263 genes and we calculated the adjusted p-values of the differentially expressed genes with FDR control. Chosen q-value threshold below 0.005 that results in 3,581 genes. Such number of genes for network construction reduces the computational burden and satisfies optimal visualization requirements (Langfelder and Horvath, 2008). In the chosen dataset, we checked as well the presence of genes reported to be involved in zebrafish heart regeneration (Gemberling et al., 2013; Smart and Riley, 2012). The dataset of 3,581 genes contains *gata4, tgf-β, fgf-1, fgf-17, wnt, notch, raldh2, wt1, igf2b*. Other potentially relevant genes, such as: *vegf, pdgf, tbx18, fgf-2, bmp-1, -4, and hand2,* were not found in this dataset. These genes have no significant differential expression as defined by our chosen q-value threshold.

The last filtering of the data was removal of the duplicated probes that denote same gene. We retained the single probe for each gene with the highest variance. Resulting dataset contained 3,477 genes.

Table 1. Cumulative number of significant features for different q-values thresholds.

q-value threshold	< 0.001	< 0.005	< 0.01	< 0.05
Number of retained genes	2,534	3,581	4,159	6,210

3.2. Construction of the weighted gene co-expression networks

For the dataset with 3,477 genes we estimated the gene expression similarity by five measures: Pearson's, Spearman's, and Kendall's correlation coefficients, cubic spline regression model and mutual information. The five correlation matrices further served for construction of the weighted gene co-expression networks with 3,477 nodes. For each

constructed network we also calculated the fundamental network properties (Table 2). Each network is expected to contain 6,043,026 potential edges. Upon conversion of the network into the Cytoscape format, the edges that are close to zero up to the size of mantissa are removed. This explains the difference of the initial number of edges before filtering. Due to the computational reasons we have to filter the weak edges (correlation < 0.8) (Table 2). The number of edges after filtering greatly vary between the networks. This indicates that fixed filtering threshold is not suitable for different topologies of the networks, thus threshold adjusting possibilities could be further investigated.

The network constructed by cubic spline regression model has the highest density, centralization and mean clustering coefficient. The highest mean scaled connectivity was obtained by mutual information network. This network as well has the second highest density and the second highest mean clustering coefficient.

Pearson's and Spearman's correlations results in the networks with quite similar topology. We can see that all the network properties have approximately the same values. Kendall's correlation based network resulted in the smallest density, centralization, mean clustering coefficient and mean scaled connectivity. Thus, this network is the least preferred to work with.

Since we are interested in the network with the topology that can result in a significant amount of functionally enriched modules, all five networks were further processed for the module detection with the Dynamic Hybrid algorithm.

Table 2. Fundamental network properties for the weighted gene co-expression networks.

Network concept	No of edges before filtering	No of edges after filtering	Density	Centralization	Heterogeneity	Mean Clustering Coefficient	Mean scaled connectivity
Pearson's network	6,043,026	439,381	0.074	0.105	0.498	0.167	0.412
Spearman's network	6,042,265	412,773	0.070	0.100	0.529	0.163	0.412
Kendall's network	6,042,916	4,830	0.014	0.030	0.619	0.044	0.313
CSRM network	6,043,026	5,049,141	0.439	0.181	0.171	0.471	0.708
MInfo network	6,039,695	2,829,679	0.332	0.098	0.106	0.340	0.773

3.3. Module detection with the Dynamic Hybrid algorithm

After construction of the network the next step is the module detection. Results of the network clustering by Dynamic Hybrid algorithm are presented on the Figure 2 and Table 3. The genes of all the networks, except Kendall's network, were assigned to modules. In the Kendall's network one gene was not clustered to any of the modules. On Figure 2 each dendrogram shows hierarchically clustered genes and their assignment to the modules (indicated by respective color underneath). It is important to mention that the same colors of the modules on different dendrograms do not indicate that modules are the same.

The major deviation of the network topology and module detection results has been shown by CSRM network and MInfo network. Dynamic Hybrid algorithm could detect only 3 and 2 large modules in these networks (Table 3). Such large modules are not reliable for precise analysis or specific conclusions.

Despite the significant differences in network properties of Pearson's and Kendall's correlation-based networks, Dynamic Hybrid algorithm determined the same number of modules in each of them (Table 3). The opposite situation is observed for Pearson's and Spearman's correlation-based networks. Being closely topologically similar, it was identified 16 modules in Pearson's network and only 14 modules in Spearman's.

Since Pearson's correlation-based network resulted in relatively high general network properties and majority of the publications with WGCNA framework used Pearson's correlation coefficient, we choose to focus on this network for further analysis.

We checked the Pearson's network module preservation statistics relatively to the Spearman's and Kendall's network modules (method described (Langfelder et al., 2011)). The preservation analysis revealed that module connectivity patterns, density, and separability are highly preserved (Bonferroni corrected p-value = 1.48×10^{-8} and smaller) between the Pearson and Spearman/Kendall networks (Appendix Figure 4). The exception is Module 13A with moderate preservation of density between the networks (Bonferroni corrected p-value = 0.468).

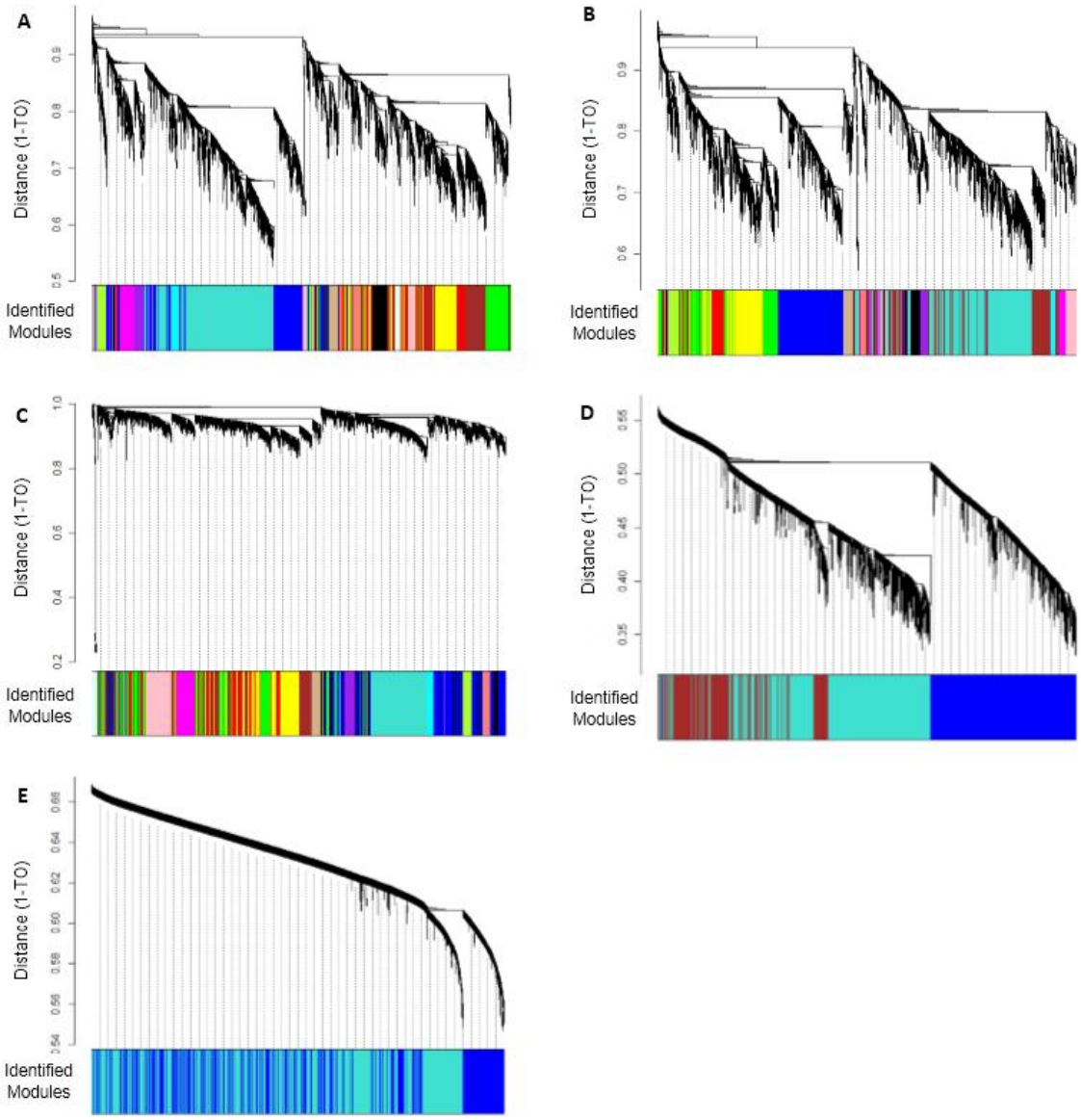


Figure 2. Gene dendrograms obtained by average linkage hierarchical clustering. (A-D) Vertical "leaves" of the dendrogram represent genes. The y-axis represents network distance, which is determined by 1 – topological overlap (TO). Values closer to 1 indicate greater dissimilarity of probe expression profiles across the samples. Color blocks below denote the module assignment determined by Dynamic Hybrid algorithm. The height cut off for dendrogram dissection into modules was determined by Dynamic Hybrid algorithm. A) Dendrogram constructed for the Pearson's correlation-based network with height cut off 0.978. B) Gene dendrogram constructed for the Spearman's correlation-based network with height cut off 0.967. C) Dendrogram constructed for the Kendall's correlation-based network with height cut off 0.996. D) Dendrogram constructed for the regression model correlation-based network with height cut off 0.562. E) Dendrogram constructed for the mutual information-based network with height cut off 0.741.

Table 3. Description of the identified modules in the networks. Module colors are matching the ones presented on Figure 2.

Module No.	Number of genes in the detected module					Color name of the module	Color of the module
	Pearson's network	Spearman's network	Kendall's network	CSRM network	MInfo network		
1	907	855	540	1432	1619	turquoise	
2	464	538	508	1209	1858	blue	
3	391	417	397	836	-	brown	
4	318	386	344	-	-	yellow	
5	265	311	336	-	-	green	
6	183	167	262	-	-	red	
7	162	162	219	-	-	black	
8	156	149	219	-	-	pink	
9	142	123	156	-	-	magenta	
10	94	103	91	-	-	purple	
11	82	88	84	-	-	green-yellow	
12	79	84	77	-	-	tan	
13	74	52	72	-	-	salmon	
14	62	42	68	-	-	cyan	
15	52		67	-	-	midnight-blue	
16	46	-	36	-	-	light-cyan	

3.4. Analysis of networks representing different biological states

We were also interested in separating the data into two heart regeneration states and construct corresponding networks. "Reference state" network includes only data points from Sham3 and 90 days post injury sample. "Damaged state" network includes all the remaining samples. We constructed two Pearson correlation-based networks and calculated their topological parameters (Appendix Table 1). Reference state network is more dense, centralized and has higher mean clustering coefficient then the Damaged state network.

Further we identified 41 modules by Dynamic Hybrid method in Reference state network, whenever 21 modules in the Damaged state network (Appendix Figure 5). We estimated Jaccard similarity between the two networks (Appendix Figure 6). The greatest Jaccard similarity index was detected between Module 15_dam and Module 9_ref (Jaccard similarity index = 0.267). The rest of the modules do not share clear similarity between

two states. These results indicate several issues with the network separation into two states:

1. Reference state network was built on the 6 data points of the samples (triplicates of the Sham3 and triplicate of the 90 days). This might be too few points for reliable detection of concordance pattern between the genes. Also it is known that Pearson's correlation coefficient is susceptible to outliers for small sample sizes.
2. The estimated general network properties are greatly dissimilar. This corresponds to the difference in the constructed networks' topology, which is also influenced by the estimation of the concordance between the genes.
3. The last issue is the dissimilarity between the modules in the Reference and Damaged state. One would expect a partial overlap between the two states. The Jaccard similarity indicated very poor similarity results, suggesting that two networks don't share similarity patterns.

Generalizing the findings, we conclude that separation of the network into Reference and Damaged states doesn't bring the desired basis for analysis of the two states separately. Thus, we proceed with the analysis of the Global network constructed with the Pearson's correlation.

3.5. Module detection with ClusterONE

The Pearson's correlation-based network was converted into Cytoscape edge and node attribute files as discussed in Materials & Methods section. To identify modules by density-based approach, we used ClusterONE. It identified 15 significant modules in the given network. Further we compared the detected modules by their topological parameters and Jaccard similarity.

3.6. Comparison of the detected modules

3.6.1. Intramodular topological properties

We calculated the intramodular properties to compare modules detected by approach A - Dynamic Hybrid algorithm, and approach B - ClusterONE. Detailed results are presented in the Appendix Table 2.

From the topological parameters of the detected modules we observe:

1. the intramodular centralization is approximately the same (~0.16) for all detected modules (exception is Module 8B), meaning that all modules have the central elements connected with other nodes that are not connected to each other;

2. Module 8B has very different topology than the rest of the modules. The intramodular density, mean clustering coefficient and mean connectivity of this module have extremely high values. Since such strong connectivity pattern was found between 13 genes the respective centralization and heterogeneity are diminished;
3. the intramodular heterogeneity is much higher in Dynamic Hybrid modules, indicating greater variation of the connectivity and module tendency to scale-free topology;
4. the higher values of the mean clustering coefficient in ClusterONE modules indicate the tendency of the nodes to connect to each other, which increases the influence among the intramodular genes;
5. ClusterONE results in modules with density higher than 0,3, which is determined by the default parameters, whenever Dynamic Hybrid algorithm gives the variability of the density values and intramodular connectivity.

We further compare the modules obtained by different methods by calculation of Jaccard similarity between them.

3.6.2. Jaccard similarity

Since the modules were detected by different approaches we are interested to see extend of similarity and overlap between the modules. As we can see from Figure 3, the greatest gene membership overlap is between Module 11A and Module 7B (0.693). From topological point of view, Module 7B is only similar based on centralization value to the Module 11A (Appendix Table 2). Though, Module 11A has higher heterogeneity, and Module 7B has higher density, clustering coefficient and connectivity.

The Jaccard similarity also shows that the two module detection methods can identify unique modules that have insignificant overlap with other modules (for example, Module 5A, 10A, 12A-15A, 15B).

From Jaccard similarity observation, indeed we can see that ClusterONE is capable of identification of overlapping modules. ClusterONE can assign the same genes to different modules, whenever Dynamic Hybrid algorithm doesn't support this option. Significant overlap between the modules (Jaccard similarity index > 0.1) is observed for:

- Module 1A <-> Modules 1 to 5B;
- Module 2A <-> Modules 2B, 4B, 5B;
- Module 3A <-> Modules 6B, 9 to 14B;

- Module 4A <-> Modules 6B, 9B, 11B, 12B;
- Module 6A <-> Modules 6B, 10B, 12B, 13B;

Jaccard similarity enables us to see the extend of an overlap between the modules detected by two different approaches. We expect the similar functional enrichment between the modules that have high Jaccard similarity index (section 3.7).

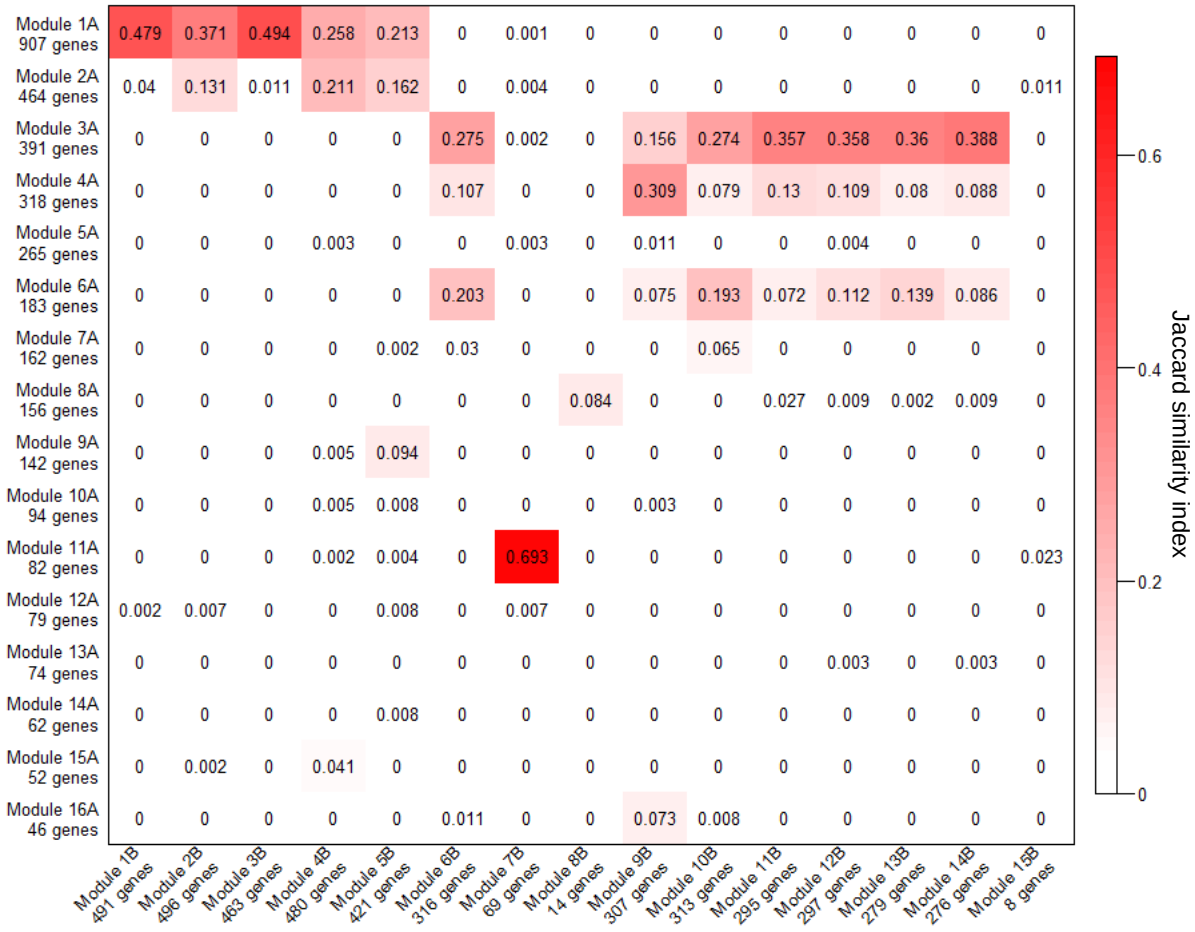


Figure 3. Jaccard similarity between the modules detected with the Dynamic Hybrid algorithm and ClusterONE. The modules detected with the Dynamic Hybrid algorithm are represented on the left axis and denoted by letter “A”. The modules detected with ClusterONE are on the bottom axis and denoted by letter “B”. The greater saturated red color stands for the greater similarity.

3.7. IMP Functional Enrichment Analysis

Functional enrichment analysis revealed the variety of the processes that are directly and indirectly related to the heart regeneration events (Figure 4): embryonic development processes and morphogenesis, cell fate specification, cardiac cell differentiation, cell migration, ribosome biogenesis and various signaling events.

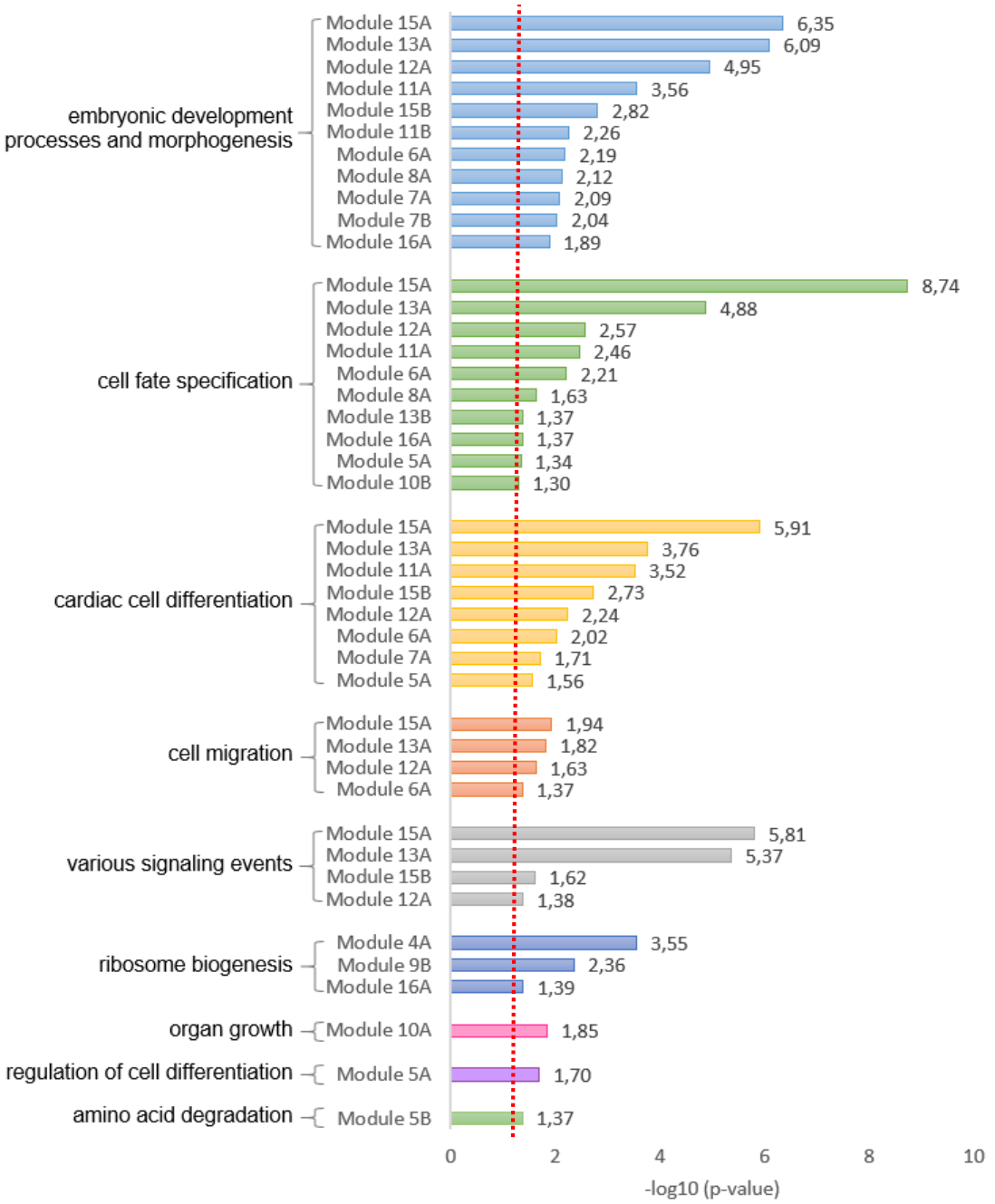


Figure 4. IMP functional enrichment of the modules. The horizontal axes of the figure denote the $-\log_{10}$(adjusted p-value). The p-values were multiple test corrected by controlling FDR rate. The significance threshold for the results is presented by the red dashed line.

3.8. Visualization of the gene expression patterns in the global network and modules

We visualized the expression changes of the Pearson's correlation-based network in the Cytoscape (details in the Materials & Methods section). In Figure 5, the nodes correspond to the genes, and the edges correspond to the correlation/connection between the genes. The strength of the regulatory interactions between the genes are denoted by the thickness of the edge and color gradation: white and thin edge corresponds to very weak or no correlation; black and thick edge corresponds to the very strong (>0.95) correlation between genes.

At 4 hours (Figure 5) we observe massive down-regulation and partial up-regulation of a considerable number of genes. Picture changes at 1 day post-injury, when amount of down-regulated genes decreases, and we see a strong up-regulation in the left part of the network. Similar differential expression can be observed at 3 dpi (Figure 5). From 7 to 90 dpi there is a gradual decrease in differential expression, and network state becomes closer to the Control condition (Sham3).

It is very informative to visualize functional sub-structures of the network, thus similar approach of visualization was applied to the functionally enriched modules (GIF figures in attached folder). The module expression patterns vary from mixed up- and down-regulation of the genes to the solely down-regulation at all post-injury time points.

The bright example of strong up- and down-regulation can be observed in the Module 12A from 4 hpi to 3 dpi (Appendix Figure 7). It is interesting to note that this Module is involved in embryonic development (smallest p-value = 2.37×10^{-6}), cardiac cell differentiation (smallest p-value = 0.0058) and cell migration (smallest p-value = 0.0233) – functions which are indeed important at early stages of heart regeneration.

Another example is Module 7B which is down-regulated at all post-cryoinjury time points (Appendix Figure 8). This module is involved in embryonic development processes and morphogenesis (smallest p-value = 0.0092).

Another approach of visualization of the gene expression patterns was implement in Excel (described in Materials & Methods section). The modules were grouped by similarity of their gene differential expression giving rise to 4 groups. First group is characterized by strong differential expression at 4 hours and its sharp decrease at other time points. This group includes Modules 3A, 4A, 6A, 6B, 9B, 10B, 11B, 12B, 13B, 14B, 16A (Figure 6A). In the second group of modules (1A, 1B, 2A, 2B, 3B, 4B, 5B, 15A) we observe the

strongest differential expression at 1-3 days post-cryoinjury and its decrease at 7dpi (Figure 6B). Third group (5A, 7A, 7B, 9A, 10A, 11A, 14A, 15B) is characterized by a peak DE at 4 hpi - 3dpi, then a decrease at 7dpi, and a final increase at 14 dpi (Figure 6C). The last group of modules (Modules 8A, 8B, 12A, 13A) has significant up- and down-regulation of the genes that starts at 4 hpi and remains up to 90 dpi (Figure 6D).

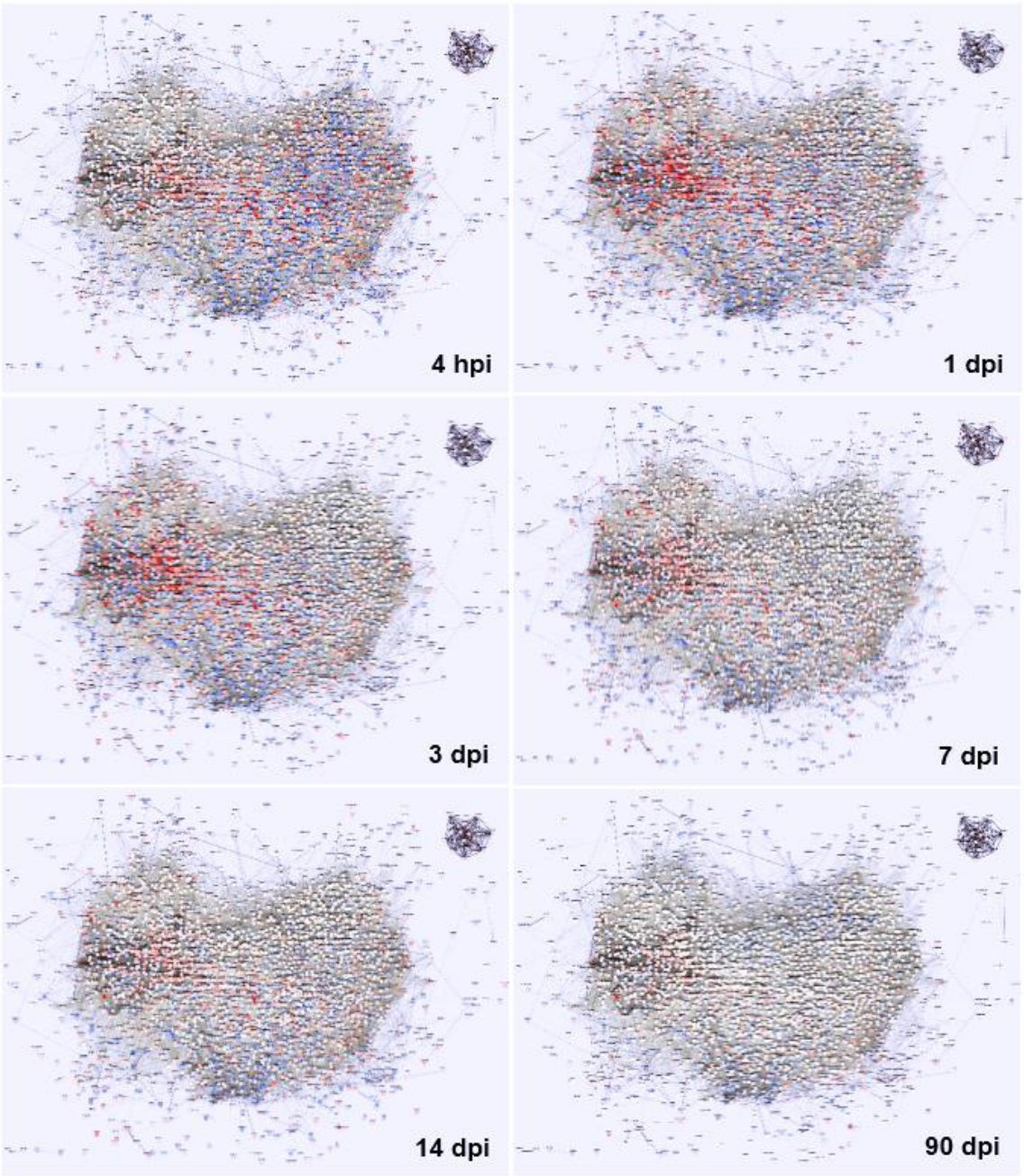


Figure 5. Dynamic visualization of the gene expression patterns in the network. Nodes correspond to genes with colors indicating differential expression. Up-regulation is shown in gradation of red, down-regulation – in gradation of blue color. The edges denote correlation between the genes. Strong connections are visualized with darker and thicker

edges, whereas weak connections appear thinner and translucent. Hpi indicates hour post injury; dpi, day post injury.

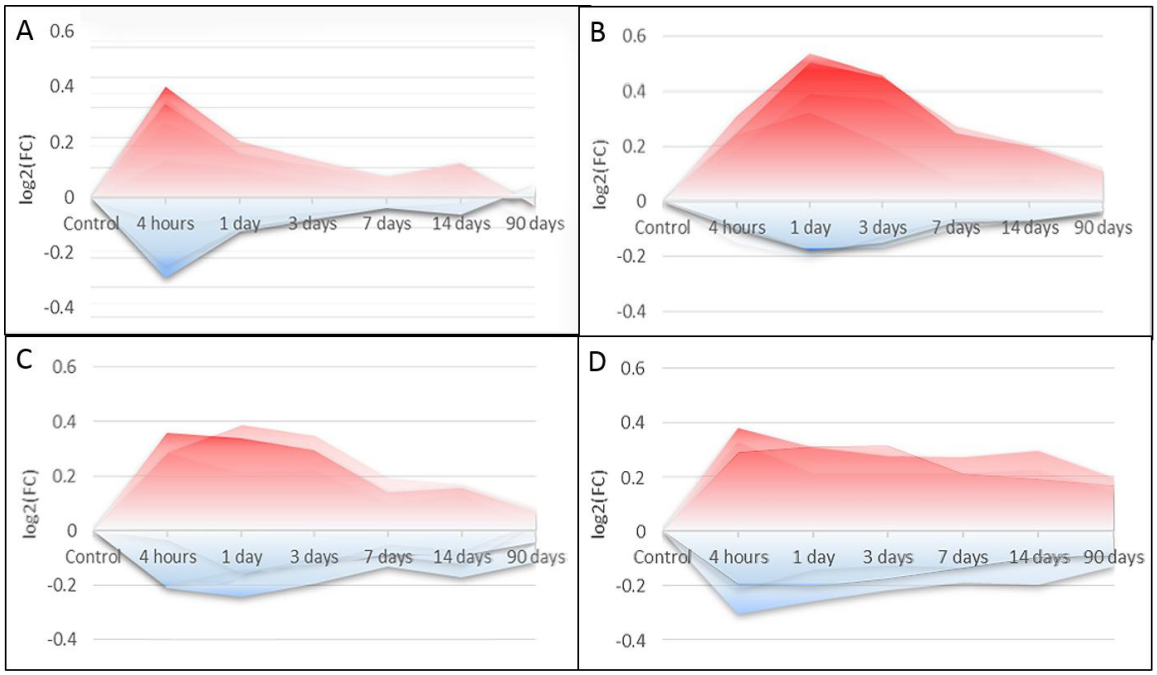


Figure 6. Visualization of the grouped differential expression patterns of the modules. (A-D) Horizontal x-axis denotes the time points upon cryoinjury and starting from the control state (Sham3). Vertical y-axis indicate the logarithmic fold change of the average up- and down-regulation of the genes in corresponding module. Up-regulation is shown by gradation of red and down-regulation – by gradation of blue color. A) The plot is constructed by overlapping the expression profiles of the group 1: Modules 3A, 4A, 6A, 6B, 9B, 10B, 11B, 12B, 13B, 14B, 16A. B) The plot is constructed by overlapping the expression profiles of the group 2: Modules 1A, 1B, 2A, 2B, 3B, 4B, 5B, 15A. C) The plot is constructed by overlapping the expression profiles of the group 3: Modules 5A, 7A, 7B, 9A, 10A, 11A, 14A, 15B. D) The plot is constructed by overlapping the expression profiles of the group 4: Modules 8A, 8B, 12A, 13A.

3.9. Targeted pathways during the heart regeneration process

A graphical summary of targeted pathways at each time point post-cryoinjury (Figure 7) was synthesized from the results of the module functional/pathway enrichment analysis and patterns of the gene expression. For this, first, we overlapped the plots of all four groups (section 3.8), obtaining the generalized plot of the network's differential expression (Figure 7). Second, we filtered the IMP results to retain only pathway enrichment from verified source databases (PID (Schaefer et al., 2009), KEGG (Kotera et al., 2012), and BioCarta (Nishimura, 2001)). Also we group the pathway enrichment of the modules that have similar differential expression.

First group, with the strongest differential expression at 4 hours (Figure 6A), is enriched various signaling pathways (transcription factor mediated signaling pathways, cell-cell signaling pathways, endothelial cell-specific factor signaling pathways, signaling involved in the regulation of cellular differentiation, proliferation, and apoptosis, various transcription pathways).

Second group, with the strongest differential expression at 1-3 days post-cryoinjury and its decrease at 7dpi (Figure 6B), is enriched in biosynthetic and degradation events, pathways regulating cell cycle, cell-cell adhesion, and motility.

Tough third group (peak DE at 4 hpi - 3dpi, a decrease at 7dpi, and increase at 14 dpi (Figure 6C) combines enrichment of the first and second groups, it has also enrichment in Notch signaling pathway, hormone signaling pathways, regulation of transcription, signaling involved in cell growth, cell differentiation, apoptosis, cellular homeostasis. It might be explained due to additional differential expression at 14 dpi.

Forth group, with significant up- and down-regulation of the genes at 4 hpi - 90 dpi (Figure 6D), is enriched in pathways that contribute to the progression of the heart over time, induce apoptosis, regulate transcription and cytoskeleton, control pro-inflammatory reactions in response to tissue injury. This group is also enriched in growth factor signaling pathways and cell-cell signaling pathways.

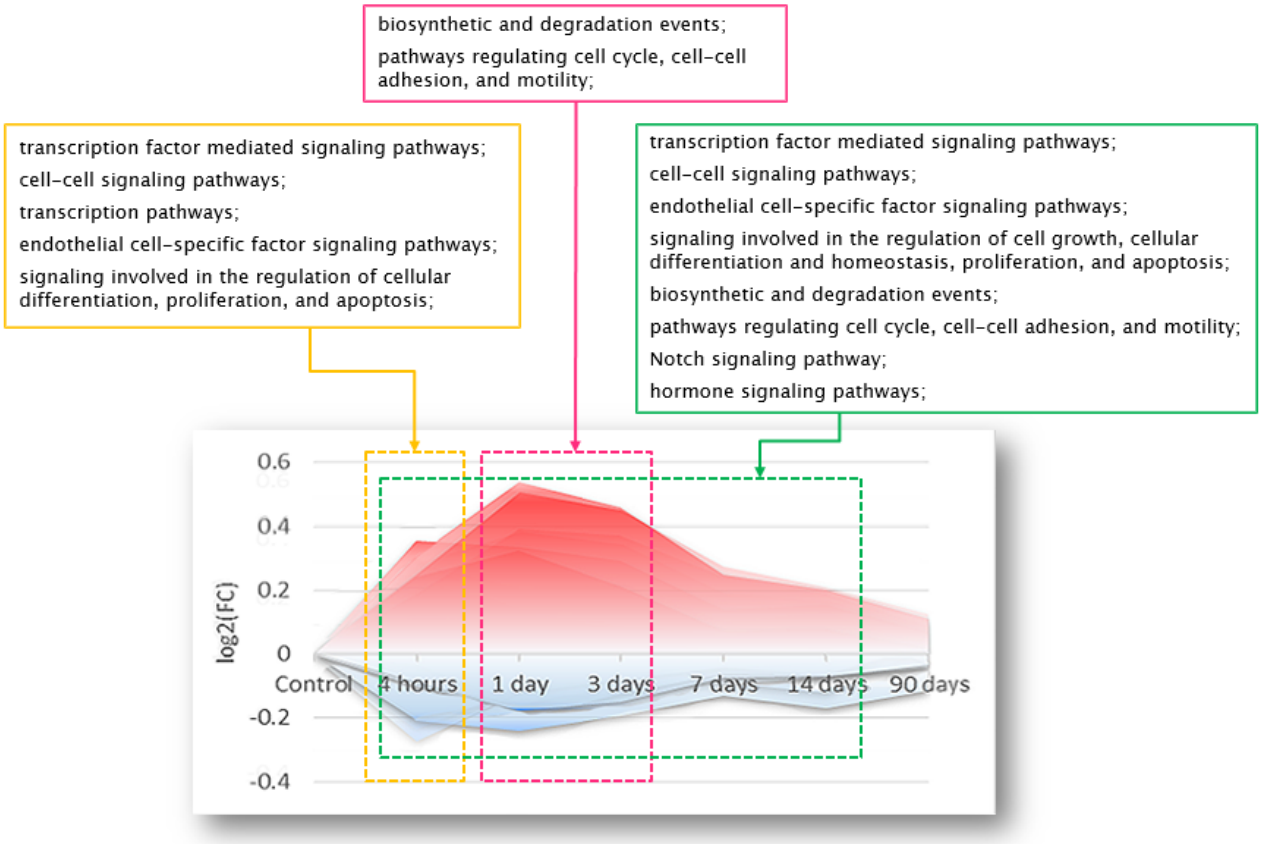


Figure 7. Targeted pathways at post-cryoinjury intervals. This plot was created by overlapping the expression patterns of the functionally enriched modules (4A to 8A, 10A to 13A, 15A, 16A, 5B, 7B, 9B to 11B, 13B, 15B). Horizontal x-axis denotes the time points upon cryoinjury and starting from the control state (Sham3). Vertical y-axis indicate the logarithmic fold change of the average up- and down-regulation of the genes in corresponding module. The up-regulation of the genes is shown by intensity of the red color, down-regulation of the genes – in gradation of blue color. Pathway enrichments are provided according to grouped expression pattern of the modules.

3.10. Hub gene detection with wiPER

We identified candidate hub genes in the global network with wiPER (Azuaje, 2014). The majority of top hub genes are members of the Modules 1A, 1B, 2A, 2B, 3B, 4B, 5B (Table 4). We also observe a variety of module memberships for lower-ranked hub genes (results not shown).

Previously reported genes that play role in heart regeneration (*gata4, tgf-β, fgf-1, fgf-17, wnt, notch, raldh2, wt1, igf2b)* were not identified as significant hub genes indicating low connectivity of these genes in our network.

Table 4. Top-10 annotated hub genes detected by wiPER. The names of the genes are not revealed due to ongoing discussions on publication and intellectual property issues. The WNC score stands for a weighted node connectivity score. p-values were adjusted by Bonferroni correction.

Hub Genes	Module Membership	Observed WNC Score	Adjusted p-value	Expression/disease association
gene A	1A, 1B, 2B, 3B, 4B, 5B	384.71	0	small cell lung cancer; rheumatoid arthritis; autoimmune disorders
gene B	1A, 1B, 2B, 4B, 5B	366.34	0	end of fetal development; hemophagocytic lymphohistiocytosis
gene C	1A, 1B, 2B, 3B, 4B, 5B	364.89	0	pulmonary tuberculosis
gene D	1A, 1B, 2B, 3B, 4B, 5B	362.85	0	prostate cancer; intravascular atherosclerosis
gene E	1A, 1B, 2B, 3B, 4B, 5B	360.68	0	Alzheimer's disease; schizophrenia
gene F	3A, 6B, 9B, 10B, 11B, 12B, 13B, 14B	357.15	0	at stage of brain development; colorectal cancer; basal and squamous cell cancers
gene G	2A, 4B, 5B	351.19	0	heart regeneration; heart response to chronic constant hypoxia
gene H	1A, 1B, 2B, 3B, 4B, 5B	341.01	0	heart response to chronic constant hypoxia; head, neck, lung squamous cell carcinoma
gene I	1A, 1B, 2B, 3B, 4B, 5B	339.10	0	-
gene J	1A, 1B, 2B, 3B, 4B, 5B	335.90	0	-

3.11. Identification of the candidate chemical compounds by STITCH

Each functionally enriched module was analyzed in the STITCH database. The results contain a prediction of gene/protein/enzyme/chemical compound association for the input set of genes. We were interested in chemical compound predictions with the high confidence score (higher than 0.7) that are based on the experimental, databases and text-mining results. Further, we calculated a combined score per each chemical compound (explained in Materials & Methods) and highlighted the hub genes among the putative targets of the compound.

The top-4 chemical compounds had high confidence scores of association with all functionally enriched modules. Due to potential intellectual property protection measures, we cannot reveal their identities here. However, these compounds are known to be involved in energy metabolism and reduction of the infarct size and platelet aggregation

in myocardial infarction. They are also capable to improve heart function (in an animal model) after myocardial infarction.

Table 5. Top-10 associated chemical compounds among the analyzed modules by STITCH.

Chemical compound	Concentration(g/mol)	Normalized Combined Score
carbohydrate	~ 400	1.00
carbohydrate	~ 500	1.00
carbohydrate	~ 500	0.96
carbohydrate	~ 450	0.75
amino acid	~ 100	0.65
enzyme	~ 750	0.62
amino acid	~ 100	0.62
amino acid	~ 150	0.61
enzyme	~ 750	0.61
carbohydrate	~ 550	0.57

3.12. CMap drug associations with the module signatures

Also we were interested to find drugs that are potentially capable to induce the observed module-specific gene expression responses during heart regeneration upon cryoinjury.

Several drugs showed potential to induce different module-specific signatures, indicating the putative targeting capacity of the identified drugs. (Table 6). The results included drugs that are capable to induce the observed module-specific gene expression response or opposite one (up-regulated genes will be down-regulated and *vice versa*). Among the top drugs, there were drugs already reported to play therapeutic role in heart disease, though the majority of the drugs have not been considered yet for myocardial infarction treatment.

Table 6. Drugs with the highest positive/negative enrichment associated to the modules signatures. We calculated the mean values of the enrichment score, permutation p-value, and specificity of drugs associated with several modules. Positive enrichment scores indicate the drugs capable to induce the gene expression as observed in module-specific signature. Negative enrichment scores indicate that drugs are capable to induce the opposite gene expression than observed in the module-specific signature. The permutation p-value estimates the probability to observe by random chance gene expression changes induced by a given drug, relatively to all available drugs in CMap database. CD(number) stands for candidate drug.

Drug	Mean Enrichment	Mean p-value (x 10^{-3})	Mean specificity	Module membership	Current clinical use
CD1	0.916	1.60	0.077	12A, 6A, 13A, 11A, 8A	anti-manic therapeutic treatment for bipolar disorder
CD2	0.961	2.64	0	12A	possible role in wound healing
CD3	0.935	0.04	0.076	12A, 13A, 11A, 8A	potential therapeutic agent for colorectal adenoma and neurodegenerative diseases
CD4	0.914	0.06	0.058	8A, 11A	potential anti-spastic agent after subarachnoid hemorrhage
CD5	0.943	6.53	0.109	6A, 8A, 13A, 13B	PI3K inhibitor
CD6	-0.958	0	0.014	5B, 8A, 11A, 12A, 13A, 13B	investigated in cardio-protection, reduces myocardial infarction
CD7	-0.991	0	0.000	8A, 11A, 12A, 13A	emetic and ameobicide agent
CD8	-0.981	0	0.010	8A, 11A, 12A, 13A	anti-protozoal drug, potential therapeutic treatment of type 1 diabetes
CD9	-0.975	1.37	0.012	13B	potential treatment of advanced gastrointestinal cancers
CD10	-0.983	0.66	0.086	13B	potential treatment of myelodysplastic syndromes, chronic myelomonocytic leukemia, acute myeloid leukemia, advanced solid tumors or lymphoma
CD11	-0.901	3.00	0.039	5B, 6A, 12A, 13A, 13B	antihistamine drug

IV. Discussion

4.1. Comparison of the biological sample controls for the recovery process upon cryoinjury

Our dataset contained three different controls for cryoinjury procedure. Sham1 is the control of cryoinjury procedure with unfrozen cryoprobe. Sham2 is the control of incision through body wall and pericardial sac. Sham3 is the control of dissection of zebrafish that aged 90 days.

Since the hearts were dissected at 4 hours post-injury for Sham1 and Sham2 samples, we can investigate the possible improvements. For example, performance of the control procedure for each respective time point of cryoinjury. Comparison of cryoinjured and control samples of respective time points could indicate more specific differential expression of genes or reveal specific modules implicated at each time point.

PCA results indicate that gene expression of Sham3 is mostly similar to the gene expression at 90 dpi sample, while Sham2 is more similar to 4 hours post-injury (Appendix Figure 1). The differential expression analysis confirms the results of the PCA, indicating that Sham1 differential expression is closest to sample of 4 hours post-injury (Appendix Figure 2A). In addition, we can see that Sham3 is the best control condition (Appendix Figure 2C), whose number of DEGs gradually decreases with the further post-injury time points. Ideally, we would expect that the 90 dpi sample would have a very no DEGs in relation to Sham3, but Figure 2C indicates the presence of differentially expressed genes. Such results might be due to batch effects in Sham3 and 90 days sample, or due to not completed regeneration of the heart at the 90 dpi.

The above-mentioned observations indicate that Sham3 is the most suitable control sample for the recovery process upon cryoinjury.

4.2. Performance of the filtering approach

We applied different approaches for the filtering of the pre-processed dataset. First, we did not know which of the performed control samples (Sham1, Sham2 or Sham3) is the most reliable control for the cryoinjury procedure. Therefore, in order to estimate the proportion of the null hypotheses, we treated at first data as two groups: all Shams vs. all cryoinjured samples. The estimated proportion of the null hypotheses $\hat{\pi}_0$ = 0.489 for this approach (Appendix Figure 3A).

This insight for another approach was given by the principle component analysis (Appendix Figure 1) and differential expression estimation (Appendix Figure 2), provided by the Genomics Research Unit of CRP-Santé, Luxembourg.

The data was treated as two groups: Sham3 and 90 days vs. the rest of the samples. The estimated proportion of the null hypotheses was $\hat{\pi}_0$ = 0.399 (Appendix Figure 3B). This indicates that treating data by the latter approach can identify a greater amount of DEGs. Thus filtering by variance should remove 39% of the genes with the lowest variability across all the arrays.

Filtering by variance can be performed on the “original” scale or on the log2-transformed scale of the data (Hackstadt and Hess, 2009). Comparison of both approaches revealed that filtering by variance on the “original” scale results in the increase of power to identify DEGs (Hackstadt and Hess, 2009). We explored both approaches and confirmed the earlier findings. Filtering on the “original” scale indeed results in greater amount of significant DEGs (6,210 genes are below q-value threshold of 0.05) than the filtering on log2-transformed scale (5,917 genes are below q-value threshold of 0.05).

In the dataset filtered by variance, we estimated the amount of differentially expressed genes that have statistical relevance. Different FDR control levels were checked: 0.05, 0.01, 0.005, and 0.001. The choice of the threshold should result in the optimal size of the network to work with. To reduce the computational burden and obtain optimal visualization, it was recommended to retain not more than 3,600 genes (Langfelder and Horvath, 2008). The q-value threshold of 0.005 was chosen for network construction from 3,581 genes.

4.3. Comparison of the reconstructed networks by different gene expression similarity measures

For our analysis we have implemented five different measures of the gene expression similarity. The widely recognized advantage of mutual information is its ability to detect non-linear relationships (Song et al., 2012). The same ability to detect non-linear relationship has been shown by the cubic spline regression model (Song et al., 2012). The estimated topological parameters of the networks, reconstructed by these two measures, resulted in the highest values of network density, mean clustering coefficient, and mean scaled connectivity (Table 2). Such results indicate that MInfo and CSRM found tightly connected patterns in the data. Though, it might also indicate that the determined relationships in the data were over-fitted and led to many false positive edges.

The module detection with standard settings proves that genes are highly connected and it is difficult to distinguish modules in the network. Dynamic Hybrid algorithm resulted in poor network clustering and detected 3 modules in CSRM network and 2 modules in MInfo network (Table 3). The smallest module of these networks consists of 836 genes, which is not reliable for precise analysis and specificity of the findings. We assume that adjustment of the clustering parameters according to the network topology could improve the results of the module detection. For example, increase of the module search depth (setting paremeter “DeepSplit = 4”) and decrease of the minimum module size to 10 (due to setting of “DeepSplit = 4”, module detection results in bigger amount of the modules with smaller size). If we apply such parameters to the regression model and mutual information networks, we are able to detect 13 and 6 modules respectively. We will discuss the results estimated by the same settings of the module detectionmethods.

The correlation coefficient measures are an attractive alternative to MInfo and CSRM due to the reasons discussed in the Introduction part. From the results obtained (Table 2), we can see that Pearson’s and Spearman’s correlation-based networks have quite similar topologies. Spearman’s correlation is robust to outliers, which is particularly important for small sample sizes ($n < 20$). However, with the increase of the sample number, Spearman’s performance is similar to Pearson’s, which is reflected in the results (Song et al., 2012).

Despite the similarity of the topological parameters of Pearson’s and Spearman’s networks, the Dynamic Hybrid algorithm detected 16 (Pearson’s network) and 14 (Spearman’s network) modules (Table 3). The size of the detected modules is more suitable for analysis than the large modules from CSRM and MInfo networks.

The Kendall’s network has shown the weakest topological parameters among the five estimated networks (Table 2). The higher heterogeneity value of the Spearman’s and Kendall’s networks can be explained by the decreased values of the other network properties, which are reciprocally related to heterogeneity. Dynamic Hybrid algorithm determined the same amount of modules in Kendall’s network as in Pearson’s one, but modules have different sizes (Table 3).

At this stage, it is not possible to offer conclusive evidence of the performance superiority of any of these techniques. For this, for example, we would need to perform functional enrichment analysis of the modules in each constructed network. Due to time constraints, we conclude that CSRM and MInfo constructed the networks with highly dense topologies. Kendall’s correlation-based network had the weakest topology among all networks.

Despite the topological similarity of the Pearson's and Spearman's correlation-based networks, we opt for the network with the larger amount of detected and relatively high topological parameters. From the abovementioned comparison, we conclude that Pearson's network has an optimal topology and a suitable amount of modules for further investigations.

To determine if the detected modules in the Pearson's network can be found in the Spearman's and Kendall's networks, we performed a module preservation analysis (Langfelder et al., 2011). The results had shown that the majority of the Pearson's modules preserve their connectivity patterns, density, and separability (Appendix Figure 4).

4.4. Comparison of modules detected by Dynamic Hybrid algorithm and ClusterONE

4.4.1. Module topology

As we expected, the peculiarities of each method are reflected in the topological parameters of the detected modules. For example, all modules detected by ClusterONE have a density above 0.3, and an overall higher intramodular clustering coefficient and connectivity than modules detected by Dynamic Hybrid algorithm (Appendix Table 2). This indicates that genes in ClusterONE modules are more tightly connected to each other and have higher intramodular influence. The smaller heterogeneity of the ClusterONE modules (~0.2 and smaller) can be explained by the fact that modules are more interconnected. On the other hand, Dynamic Hybrid modules have higher heterogeneity values indicating the variability of the connectivity patterns in the modules.

Since intramodular centralization values are approximately the same between the majority of the detected modules (~0.16), we conclude that modules follow approximately the star-topology (having central nodes, which might be regulatory hub genes).

The module detection by Dynamic Hybrid algorithm and ClusterONE results in a great variability of modules' topological parameters. The implementation of the two different network clustering approaches gives us an opportunity to look at the same network under different lenses.

4.4.2. Module overlapping

Jaccard similarity indicates the overlap between the modules detected by different approaches. Its results show that two network clustering methods can identify on the one hand the unique modules that have no overlap with other modules (for example, Module 5A, 10A, 12A-15A, 15B) and on the other hand highly overlapping modules (for example, Module 11A and 7B).

The general observation from the Jaccard similarity analysis is that ClusterONE is indeed capable to identify overlapping modules, while Dynamic Hybrid algorithm assigns each gene to only one module (for example, Dynamic Hybrid Module 1A has significant overlap with ClusterONE Modules 1-5B).

Jaccard similarity results are useful for further analysis of modules. In the case of the analysis of very large modules (for example, Module 1A containing 907 genes), the statistically significant functional enrichment might be potentially undetected. Thereon, we were able to focus our attention on other modules that have high Jaccard similarity with this large module. Functional enrichment of these modules would partially describe the possible functional enrichment of the large module. For example, Module 1A is not statistically significantly functionally enriched, but overlapping Module 5B is enriched in amino acid degradation. It remains an open question if we should proceed with large modules or try to dissect them into smaller sub-structures.

The application of two different network clustering methods enables us to retrieve more biologically relevant information from the same network.

4.4.3. Module functional enrichment

The functional enrichment analysis showed that Dynamic Hybrid algorithm results in more functionally enriched modules than ClusterONE. Among the Dynamic Hybrid modules there were 11 out of 16 functionally enriched modules. Among ClusterONE modules there were 7 out 15 functionally enriched modules. Additionally, Dynamic Hybrid modules had a wider variety of the significantly enriched processes (for example, Module 12A and 13A are functionally enriched in 5 different processes) whenever ClusterONE detected only one module that is moderately functionally enriched (Module 15B is functionally enriched in 3 different processes).

As observed on Figure 4, a majority of the modules are functionally enriched in embryonic developmental processes and embryonic morphogenesis. Such modules can be

significantly involved at 4 hours post-injury response, when reactivation of the genes involved in heart development occurs (Gonzalez-Rosa et al., 2011). To translate the mRNA into protein, ribosomes are required. Therefore modules functionally enriched in ribosome biogenesis might play part at this and later time points. At the early post-injury response, one can observe inflammation process, which leads us to think that modules involved in various signaling events might be responsible for driving inflammation.

The modules involved in cell fate specification (such as Module 5A: p-value = 0.02; Module 10B: p-value = 0.0499; Module 13B: p-value = 0.0426; Module 6A: p-value = 0.0061; Module 8A: p-value = 0.014; Module 11A: p-value = 3.03×10^{-4}; Module 13A: p-value = 1.68×10^{-5}; Module 15A: p-value = 1.82×10^{-9}) might play a role at 1-7 dpi when cells extensively proliferate to form the fibrotic scar. Since gradual substitution of the fibrotic scar by new cardiomyocytes was previously observed, we can consider that modules, functionally enriched in cell migration and cardiac cell differentiation (Figure 4), might take part in scar substitution.

Our hypotheses on module implication at different stages of heart regeneration might be "validated" by visualization of the gene expression dynamics. Such visualization shows the time frames of the particular module gene expression and functional enrichment in these time frames.

4.5. Dynamic visualization of the zebrafish heart recovery

The dynamic visualization of the global network provided us with the insight that the strongest gene differential expression takes place from 4 hours up to 3 days post-cryoinjury. We believe that, during this time interval, the major contributing events to heart regeneration occur.

The fact that we can still see some differential expression at 90 days in comparison to Control condition (Sham3) might be due to the incomplete regeneration of the heart, or by a Control sample that does not perfectly represent the healthy zebrafish heart.

Different publications state different time intervals required for the complete regeneration of the heart (Gonzalez-Rosa et al., 2011; Poss et al., 2002). It was demonstrated that new cardiomyocytes replace most of the lost ventricular tissue by the 30th day post-amputation and the heart achieves complete regrowth of the amputated region at the 60th day post-amputation (Poss et al., 2002).Though it was not stated that the heart has reached a healthy state identical to the initial one. Another publication demonstrated that the fibrotic

scar completely disappears at 130th day post-cryoinjury and heart regeneration is completed in cryoinjured model (Gonzalez-Rosa et al., 2011).

We also visualized the gene expression patterns of each detected module. This gave us the possibility to see at which time point(s) a functionally enriched module plays a role. Some modules had an early post-injury response (the strongest differential expression at 4 hours and its sharp decrease at other time points; Figure 6A). The functional enrichment of these modules (4A, 6A, 9B, 10B, 11B, 13B, 16A) indicate that they are involved in embryonic developmental processes, ribosome biogenesis and cell fate specification (Figure 4). This functional enrichment is relevant to the expected events at early post-injury response.

Visualization of another group of modules (1A, 1B, 2A, 2B, 3B, 4B, 5B, 15A) indicated the strongest differential expression at 1-3 days post-cryoinjury and its decrease starting 7dpi (Figure 6B). Module 15A is of particular interest, because it is functionally enriched in a variety of processes: embryonic development, cell fate specification, various signaling events, cell migration and cardiac cell differentiation (Figure 4). Another functionally enriched module is Module 5B responsible for amino acid degradation. These modules might be of particular interest since they are involved at the post-inflammatory stage. At this time interval, all cell types proliferate extensively. The high levels of proliferation in a cryoinjury can be observed at the border zones close to the injured area of endocardium and myocardium, leading the regeneration of the damaged tissue (Gonzalez-Rosa and Mercader, 2012). Thus, an enhance of the differential expression pattern is observed in these modules could either increase the speed of the heart recovery process or provide ability of regeneration in a human heart. The latter, though, is a hypothesis that will require both computational and experimental validation.

The visualization of Modules 5A, 7A, 7B, 9A, 10A, 11A, 14A, 15B showed the strongest differential expression at 4 hpi - 3dpi, then its decrease at 7dpi, and an increase at 14 dpi (Figure 6C). Functional enrichment of these modules (5A, 7A, 7B, 9A, 10A, 11A, 14A, 15B) indicates the variety of the processes: from regulation of cell differentiation up to the organ growth (Figure 4). The regulation of the genes belonging to these modules could be promising but there is a great chance of increasing the inflammatory response since these modules are differentially expressed starting 4hpi.

The expression pattern visualization of the Modules 8A, 8B, 12A, and 13A showed significant DE of the genes starting 4 hpi and remaining up to 90 dpi (Figure 6D). Functional enrichment of Modules 12A and 13A involves a great variety of the processes:

embryonic development, cell fate specification and migration, cardiac cell differentiation and various signaling events (Figure 4).

4.6. Pathway targeting during heart regeneration process in zebrafish

We analyzed the regulation of the regeneration process at the pathway level. The implicated pathways in regeneration were obtained by the combination of the module functional/pathway enrichment results and the modules' gene differential expression at different time points.

The relevance of obtained results is proved by the known events that take place during heart regeneration. The early post-injury response is inflammation. The modules involved in various signaling events have significant gene DE at 4 hours post-injury. These modules are functionally enriched in signaling pathways regulated by transcription factors, endothelial cell-specific factors, cell-cell signaling pathways and signaling involved in the regulation of cellular differentiation, proliferation and apoptosis. The observed functional enrichment may indeed initiate the cascade of inflammatory response.

Results of functional/pathway enrichment of modules, with highest gene DE at 1-3 dpi, suggest that targeted pathways are responsible for cell cycle, cell motility and cell-cell adhesion. The biosynthetic and degradation event also takes place at 1-3 days post-injury. These results confirm the earlier findings of cell proliferation in endocardium and myocardium regions, resulting in heart tissue regeneration (Gonzalez-Rosa et al., 2011). From literature review, we know that the major contributing events to heart regeneration take place at 1-4 dpi (Gonzalez-Rosa et al., 2011). All further time points contribute to regeneration by following the obtained changes and gradual regression of the transient scar until its complete disappearance.

We were unable to obtain the results of pathway targeting specific to 7, 14, and 90 days post-injury though the modules, with significant gene differential expression at these time points, among mentioned above pathways are also involved in Notch and hormone signaling.

4.7. Detected hub genes

Among the results of hub gene detection, we presented only the top-10 annotated genes. We found as well non-annotated genes among the top hubs, but data are not presented, as no characterization of these genes is available. Some hub genes are particularly of interest since they have not been studied yet as potential molecular targets for myocardial infarction treatment or relevant to regeneration in zebrafish.

Among the candidate hub genes detected in the network, there are genes known to be relevant to intravascular atherosclerosis treatment, and as indicators of inflammation upon cardiac arrest. Some hub genes are characteristic to different cancer types: small cell lung cancer, prostate cancer, colorectal cancer, neck, head squamous cell carcinoma. The implication of the genes in cancer might indicate their possible role in the cell cycle, cell proliferation, motility and migration, which is required for the heart regeneration process.

As future outlook it would be interesting to independently validate the role of the top hub genes in heart regeneration. The related findings would be interesting in both directions: if they are capable to boost heart regeneration or slow it down, indicating new targets for further studies.

4.8. The candidate chemical compounds and drugs

Integrative approach for drug repositioning implemented in our study takes advantage of multiple independent databases and provides predictions of novel indications for known drugs.

For the identification of the chemical compound association with the modules of interest we used STITCH database. The advantage of this approach is the prediction of the associations based on different sources: databases, experimental and text-mining results. STITCH provides an indication of the compound targets (genes or other chemicals) and sources of the identified interaction. We checked the presence of hub genes among the targets of the candidate compounds. Consequently, the choice of the chemical compound can be based also on the confidence of targeting predicted interactors.

The majority of the predicted compounds are carbohydrates. A possible explanation to this might be the requirement of large energy supply for all metabolic processes. It is not surprising that all reconstruction changes in the cell require energetic basis, which can be provided by carbohydrates.

Among the identified candidate compounds, some are involved in the reduction of the infarct size and platelet aggregation in myocardial infarction. Other candidate compounds are known to improve heart function upon myocardial infarction. An experimental validation is therefore required to further estimate their potential in treatment of myocardial infarction.

Another approach for drug repositioning includes the drug association to specific module signatures in CMap. The positive enrichment scores (Table 6) indicate that a drug is capable of inducing gene expression observed in a module during heart regeneration upon cryoinjury. The negative enrichment indicates that drug induces the opposite gene expression than the one observed in the module-specific signature. The top drugs were selected based on their mean enrichment score and specificity calculated from association to module signatures.

Drugs associated with several module signatures are particularly interesting. Since such drugs have an increased effect during treatment because they regulate several sub-structures of the network at the same time.

None of the predicted candidate drugs have been reported to play role as treatment of MI. This fact provides room for detailed investigation and experimental validation to estimate their potential value as MI treatment.

4.9. Conclusions and outlook

This thesis provides a comprehensive integrative analysis of the transcriptional network underlying zebrafish heart regeneration.

1. From comparison of the performance of different gene expression similarity measures for the network re-construction, Pearson correlation coefficient resulted in the most optimal network topology for our study. The majority of the identified modules in the Pearson's network were statistically functionally enriched.
2. Detected modules by ClusterONE have more dense topological structure, though Dynamic Hybrid algorithm identified more modules that are functionally enriched.
3. From functional enrichment analysis and dynamical visualization of the gene expression patterns, we identified modules that are potentially critical to the zebrafish heart regeneration process.
4. We identified candidate regulatory hub genes of the network. Among the top hub genes, there are genes relevant to: different cancer types, intravascular atherosclerosis treatment, and indication of inflammation upon cardiac arrest.
5. We identified pathways that are significantly implicated at different time intervals of the regeneration process upon cryoinjury. These findings resemble the processes that take place at each step of regeneration.
6. The identified candidate drugs/chemical compounds are potential treatments for myocardial infarction. Among the top drugs, we detected putative enhancers and triggers of gene expression during heart regeneration process in zebrafish.

These and other findings are currently undergoing deeper computational analyses. The top, most promising candidate drugs and targets will be independently validated using our zebrafish *in vivo* model.

In conclusion, our findings provide insights into the complex regulatory mechanisms involved during heart regeneration of the zebrafish. These data will be useful for modelling specific network-based responses to heart injury, and for finding sensitive network points that may trigger or boost heart regeneration.

V. References

Allen, J.D., Xie, Y., Chen, M., Girard, L., and Xiao, G. (2012). Comparing statistical methods for constructing large scale gene networks. PloS one *7*, e29348.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T.*, et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics *25*, 25-29.

Azuaje, F.J. (2014). Selecting biologically informative genes in co-expression networks with a centrality score. Biology Direct *9*, 12.

Azuaje, F.J., Zhang, L., Devaux, Y., Wagner, D.R. (2011). Drug-target network in myocardial infarction reveals multiple side effects of unrelated drugs. Scientific Reports *1,* article 54.

Barabasi, A.L., Altvai, Z.N. (2004). Network biology: understanding the cell's functional organization. Nature Reviews Genetics *5*, 101-113.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological) *57*, 289-300.

Borate, B.R., Chesler, E.J., Langston, M.A., Saxton, A.M., and Voy, B.H. (2009). Comparison of threshold selection methods for microarray gene co-expression matrices. BMC research notes *2*, 240.

Bourgon, R., Gentleman, R., and Huber, W. (2010). Independent filtering increases detection power for high-throughput experiments. Proceedings of the National Academy of Sciences of the United States of America *107*, 9546-9551.

Calza, S., Raffelsberger, W., Ploner, A., Sahel, J., Leveillard, T., and Pawitan, Y. (2007). Filtering genes to improve sensitivity in oligonucleotide microarray data analysis. Nucleic acids research *35*, e102.

Caspi, R., Altman, T., Dreher, K., Fulcher, C.A., Subhraveti, P., Keseler, I.M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L.A.*, et al.* (2012). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic acids research *40*, D742-753.

Council of the European Union (2006) Council Conclusions on Common Values and Principles in EU Health Systems, 2733rd Employment, Social Policy, Health and Consumer Affairs Council Meeting, Luxembourg, 1-2 June.

Culhane, A.C., Thioulouse, J., Perriere, G., and Higgins, D.G. (2005). MADE4: an R package for multivariate analysis of gene expression data. Bioinformatics (Oxford, England) *21*, 2789-2790.

Dabney, A. and Storey, J.D. (2014). qvalue: Q-value estimation for false discovery rate control. R package version 1.38.0.

Daub, C.O., Steuer, R., Selbig, J., and Kloska, S. (2004). Estimating mutual information using B-spline functions--an improved similarity measure for analysing gene expression data. BMC bioinformatics *5*, 118.

de Siqueira Santos, S., Takahashi, D.Y., Nakata, A., and Fujita, A. (2013). A comparative study of statistical methods used to identify dependencies between gene expression signals. Brief Bioinform.

Devlin, S.J., Gnanadesikan, R., and Kettenring, J.R. (1975). Robust estimation and outlier detection with correlation coefficients. Biometrika *62*, 531-545.

Direction de la Santé (2009). National Statistics of all causes of death, Statistiques des causes de décès. Ministry of Health edition, Luxembourg.

Dong, J., and Horvath, S. (2007). Understanding network concepts in modules. BMC systems biology *1*, 1-20.

Freeman, L.C. (1978). Centrality in social networks conceptual clarification. Social Networks *1*, 215-239.

Gemberling, M., Bailey, T.J., Hyde, D.R., and Poss, K.D. (2013). The zebrafish as a model for complex tissue regeneration. Trends in genetics : TIG *29*, 611-620.

Gentleman, R.C., Carey, V.J., Huber, W., Hahne, F. (2009).Genefilter: methods for filtering genes from microarray experiments. R Package Version 1.24.2., Bioconductor 2.6.

Ghazalpour, A., Doss, S., Zhang, B., Wang, S., Plaisier, C., Castellanos, R., Brozell, A., Schadt, E.E., Drake, T.A., Lusis, A.J., *et al.* (2006). Integrating genetic and network analysis to characterize genes related to mouse weight. PLoS genetics *2*, e130.

Gonzalez-Rosa, J.M., Martin, V., Peralta, M., Torres, M., and Mercader, N. (2011). Extensive scar formation and regression during heart regeneration after cryoinjury in zebrafish. Development (Cambridge, England) *138*, 1663-1674.

Gonzalez-Rosa, J.M., and Mercader, N. (2012). Cryoinjury as a myocardial infarction model for the study of cardiac regeneration in the zebrafish. Nature protocols *7*, 782-788.

Hackstadt, A.J., and Hess, A.M. (2009). Filtering for increased power for microarray data analysis. BMC bioinformatics *10*, 11.

Hardin, J., Mitani, A., Hicks, L., and VanKoten, B. (2007). A robust measure of correlation between two genes on a microarray. BMC bioinformatics *8*, 220.

Hausser, J., and Strimmer, K. (2009). Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. J Mach Learn Res Journal of Machine Learning Research *10*, 1469-1484.

Horvath, S., and Dong, J. (2008). Geometric interpretation of gene coexpression network analysis. PLoS Comput Biol *4*, e1000117.

Horvath, S., Zhang, B., Carlson, M., Lu, K.V., Zhu, S., Felciano, R.M., Laurance, M.F., Zhao, W., Qi, S., Chen, Z.*, et al.* (2006). Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. Proceedings of the National Academy of Sciences of the United States of America *103*, 17402-17407.

Howe, K., Clark, M.D., Torroja, C.F., Torrance, J., Berthelot, C., Muffato, M., Collins, J.E., Humphray, S., McLaren, K., Matthews, L.*, et al.* (2013). The zebrafish reference genome sequence and its relationship to the human genome. Nature *496*, 498-503.

Ingram, P.J., Stumpf, M.P.H., and Stark, J. (2008). Nonidentifiability of the Source of Intrinsic Noise in Gene Expression from Single-Burst Data. PLoS Comput Biol *4*, e1000192.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin del la Société Vaudoise des Sciences Naturelles *37*, 547-579.

Jennings, R.B., Murry, C.E., Steenbergen, C., Jr., and Reimer, K.A. (1990). Development of cell injury in sustained acute ischemia. Circulation *82*, II2-12.

Kadarmideen, H.N., and Watson-Haigh, N.S. (2012). Building gene co-expression networks using transcriptomics data for systems biology investigations: Comparison of methods using microarray data. Bioinformation *8*, 855-861.

Kendall, M.G. (1938). A new measure of rank correlation. Biometrika *30*, 81-93.

Kotera, M., Hirakawa, M., Tokimatsu, T., Goto, S., and Kanehisa, M. (2012). The KEGG databases and tools facilitating omics analysis: latest developments involving human diseases and pharmaceuticals. Methods in molecular biology (Clifton, NJ) *802*, 19-39.

Kuhn, M., Szklarczyk, D., Franceschini, A., von Mering, C., Jensen, L.J., and Bork, P. (2012). STITCH 3: zooming in on protein-chemical interactions. Nucleic acids research *40*, D876-880.

Kumari, S., Nie, J., Chen, H.-S., Ma, H., Stewart, R., Li, X., Lu, M.-Z., Taylor, W.M., and Wei, H. (2012). Evaluation of Gene Association Methods for Coexpression Network Construction and Biological Knowledge Discovery. PloS one *7*, e50411.

Laflamme, M.A., and Murry, C.E. (2011). Heart regeneration. Nature *473*, 326-335.

Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., Blat, I.C., Wrobel, M.J., Lerner, J., Brunet, J.P., Subramanian, A., Ross, K.N.*, et al.* (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science (New York, NY) *313*, 1929-1935.

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics *9*, 559.

Langfelder, P., Luo, R., Oldham, M.C., and Horvath, S. (2011). Is my network module preserved and reproducible? PLoS Comput Biol *7*, e1001057.

Li, A., and Horvath, S. (2007). Network neighborhood analysis with the multi-node topological overlap measure. Bioinformatics (Oxford, England) *23*, 222-231.

Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B.*, et al.* (2009). Reactome knowledgebase of human biological pathways and processes. Nucleic acids research *37*, D619-622.

Morris, J.H., Apeltsin, L., Newman, A.M., Baumbach, J., Wittkop, T., Su, G., Bader, G.D., and Ferrin, T.E. (2011). clusterMaker: a multi-algorithm clustering plugin for Cytoscape. BMC bioinformatics *12*, 436.

Nazarov, P.V., Reinsbach, S.E., Muller, A., Nicot, N., Philippidou, D., Vallar, L., and Kreis, S. (2013). Interplay of microRNAs, transcription factors and target genes: linking dynamic expression changes to function. Nucleic acids research *41*, 2817-2831.

Nepusz, T., Yu, H., and Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. Nature methods *9*, 471-472.

Newson, R. (2002). Parameters behind "nonparametric" statistics: Kendall's tau, Somers' D and median differences. Stata Journal *2*, 45-64.

Nishimura, D. (2001). BioCarta. Biotech Software & Internet Report *2*, 117-120.

Pearson, K. (1920). Notes on the history of correlation. Biometrika *13*, 25-45.

Pollard, K.S., Dudoit, S., and van der Laan, M.J. (2005). Multiple Testing Procedures: the multtest Package and Applications to Genomics. In Bioinformatics and Computational Biology Solutions Using R and Bioconductor, R. Gentleman, V. Carey, W. Huber, R. Irizarry, and S. Dudoit, eds. (Springer New York), pp. 249-271.

Poss, K.D., Wilson, L.G., and Keating, M.T. (2002). Heart regeneration in zebrafish. Science (New York, NY) *298*, 2188-2190.

Presson, A.P., Sobel, E.M., Papp, J.C., Suarez, C.J., Whistler, T., Rajeevan, M.S., Vernon, S.D., and Horvath, S. (2008). Integrated weighted gene co-expression network analysis with an application to chronic fatigue syndrome. BMC systems biology *2*, 95.

Pu, Z.E., Hou, Y.C., Xu, X.X., Yan, Z.H., Wei, Y.M., Lan, X.J., and Zheng, Y.L. (2009). Genetic Diversity among Barley Populations from West China Based on RAMP and RAPD Markers.

Qiu, P., Gentles, A.J., and Plevritis, S.K. (2009). Fast calculation of pairwise mutual information for gene regulatory network reconstruction. Computer methods and programs in biomedicine *94*, 177-180.

Renda, M.E., and Straccia, U. (2003). Web metasearch: rank vs. score based rank aggregation methods. In Proceedings of the 2003 ACM symposium on Applied computing (Melbourne, Florida: ACM), pp. 841-846.

Schaefer, C.F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., and Buetow, K.H. (2009). PID: the Pathway Interaction Database. Nucleic acids research *37*, D674-679.

Sedmera, D., and Thompson, R.P. (2011). Myocyte proliferation in the developing heart. Developmental dynamics : an official publication of the American Association of Anatomists *240*, 1322-1334.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research *13*, 2498-2504.

Sidney, S., Rosamond, W.D., Howard, V.J., and Luepker, R.V. (2013). The "heart disease and stroke statistics--2013 update" and the need for a national cardiovascular surveillance system. Circulation *127*, 21-23.

Sleigh, S.H., and Barton, C.L. (2010). Repurposing Strategies for Therapeutics. Pharm Med *24*, 151-159.

Smart, N., and Riley, P.R. (2012). The epicardium as a candidate for heart regeneration. Future cardiology *8*, 53-69.

Song, L., Langfelder, P., and Horvath, S. (2012). Comparison of co-expression measures: mutual information, correlation, and model based indices. BMC bioinformatics *13*, 328.

Spearman, C. (1904). The proof and measurement of association between two things. The American journal of psychology *15*, 72-101.

Storey, J.D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences of the United States of America *100*, 9440-9445.

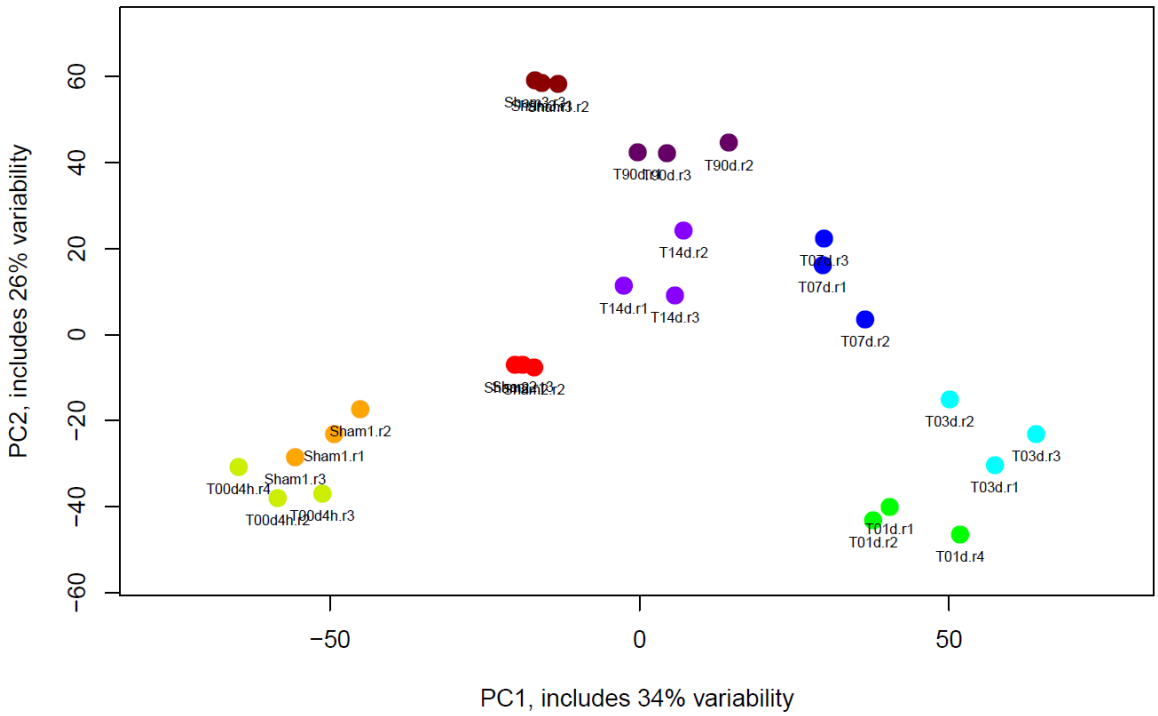
Wang, J., Panakova, D., Kikuchi, K., Holdway, J.E., Gemberling, M., Burris, J.S., Singh, S.P., Dickson, A.L., Lin, Y.F., Sabeh, M.K.*, et al.* (2011). The regenerative capacity of zebrafish reverses cardiac failure caused by genetic cardiomyocyte depletion. Development (Cambridge, England) *138*, 3421-3430.

Wong, A.K., Park, C.Y., Greene, C.S., Bongo, L.A., Guan, Y., and Troyanskaya, O.G. (2012). IMP: a multi-species functional genomics portal for integration, visualization and prediction of protein functions and networks. Nucleic acids research *40*, W484-490.

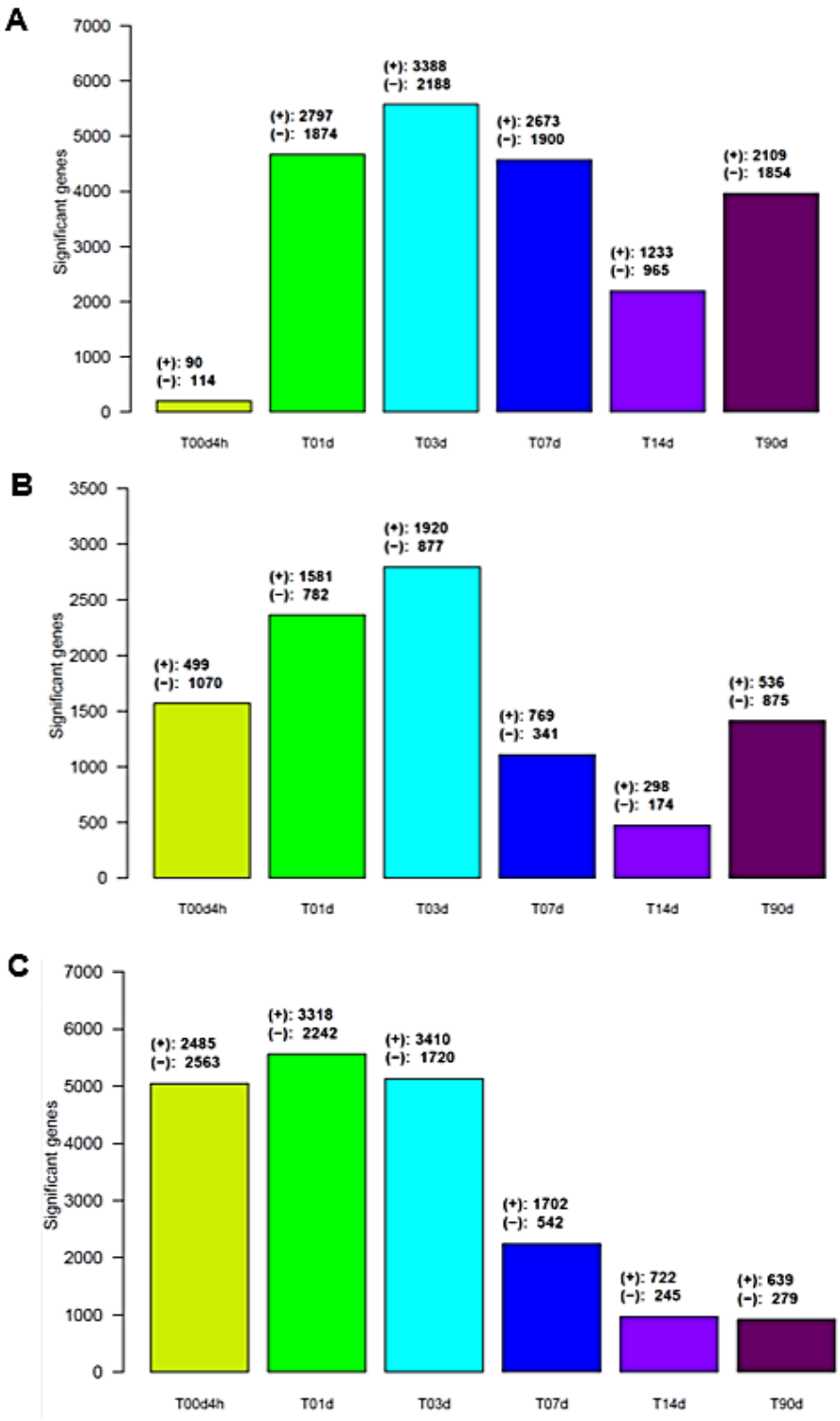
Yip, A.M., and Horvath, S. (2007). Gene network interconnectedness and the generalized topological overlap measure. BMC bioinformatics *8*, 22.

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. Statistical applications in genetics and molecular biology *4*, Article17.

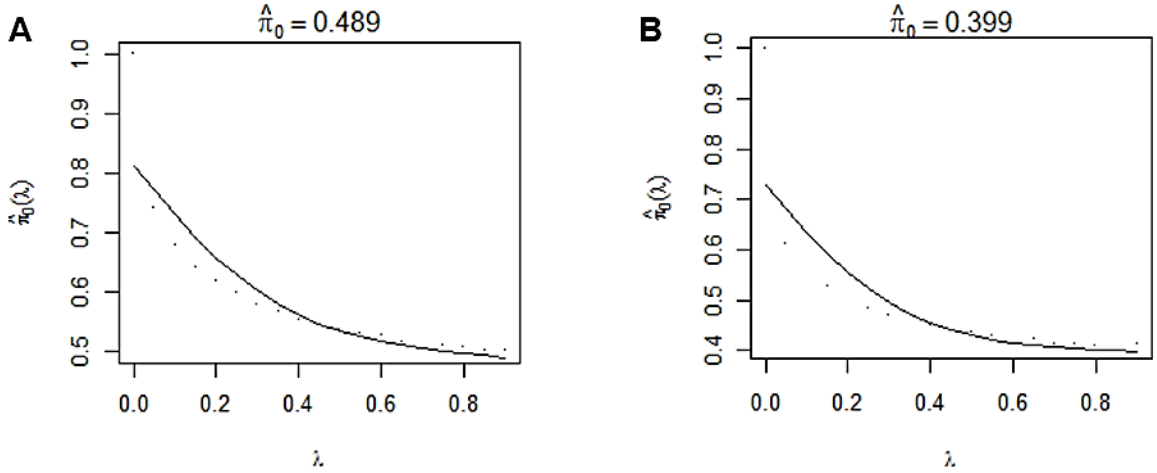
VI. Appendices



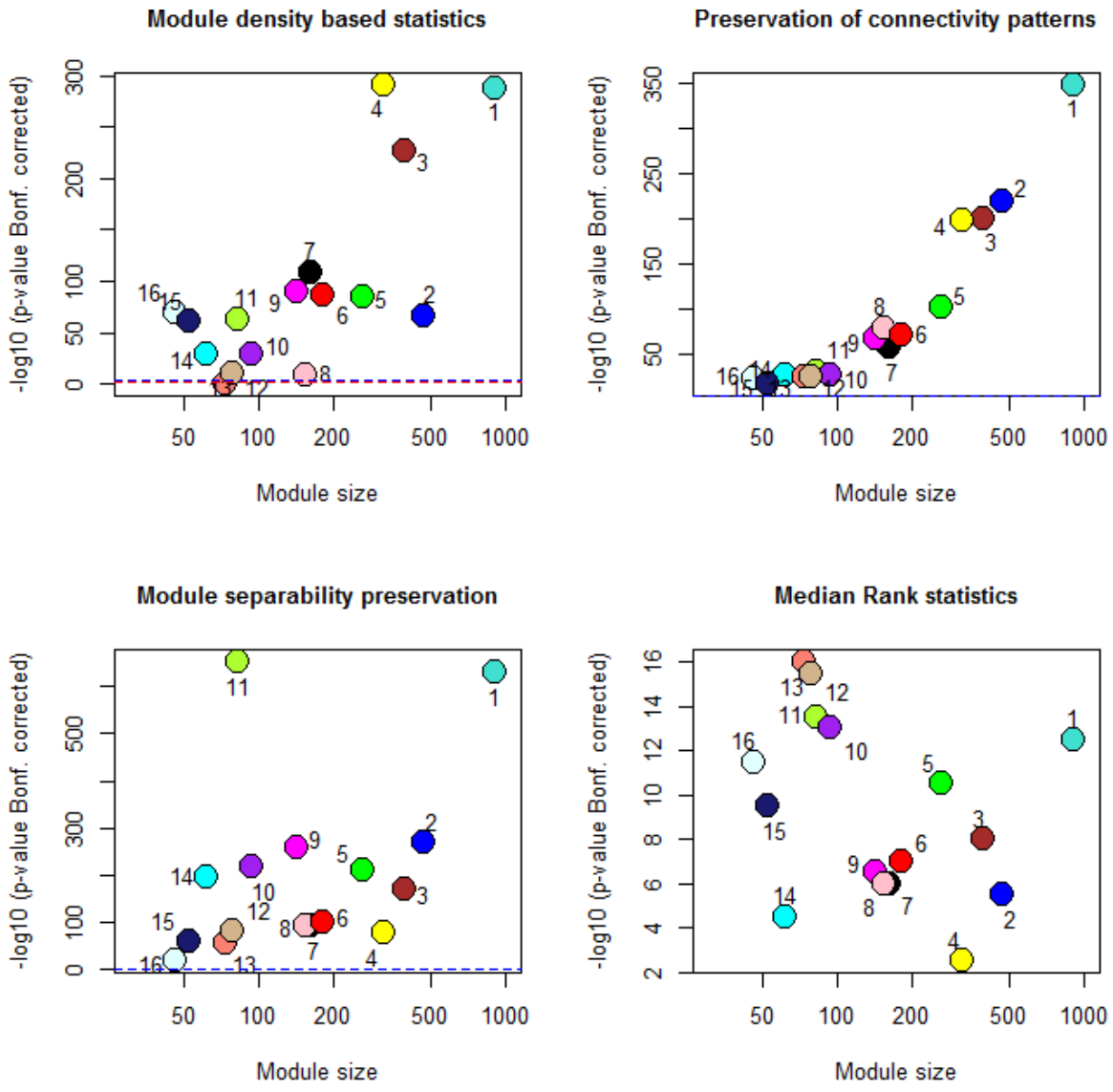
Appendix Figure 1. Principle component analysis of the pre-processed dataset. are denoted as PC1 (first principal component) and PC2 (second principal component) varibility. Sham1, Sham2 and Sham3 are the different controls explained in Materials & Methods section. T00d4h is 4 hours post-injury. Time points: T01d, T03d, T07d, T14d and T90d denote post-injury samples at day 1, 3, 7, 14, and 90 respectively. r1, r2, r3 or r4 are the respective biological replicates of the sample. This figure was provided by the Genomics Research Unit of CRP-Santé.



Appendix Figure 2. The differential expression of the post-injured samples in relation to Shams. A) Detected differentially expressed genes with regards to Sham1. B) Detected differentially expressed genes with regards to Sham2. C) Detected differentially expressed genes with regards to Sham1. T00d4h, T01d, T03d, T07d, T14d and T90d are post-injury samples at 4 hours, day 1, 3, 7, 14, and 90 respectively. This figure was provided by the Genomics Research Unit of CRP-Santé.



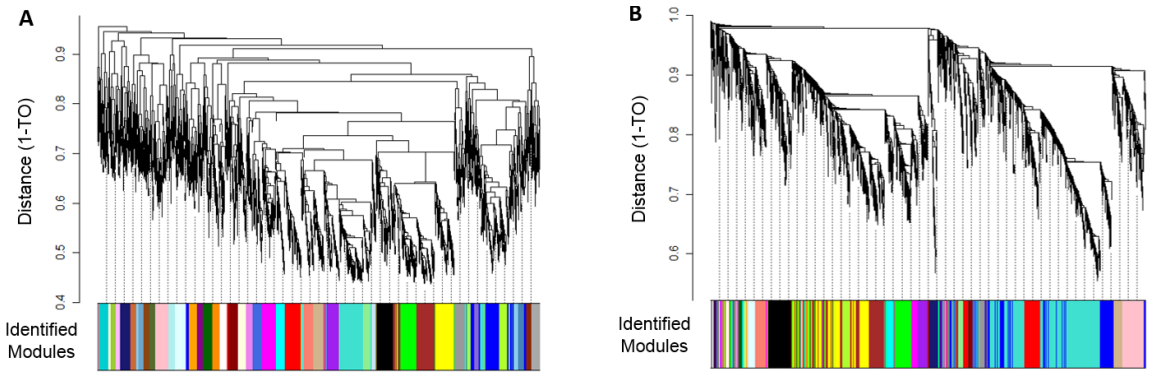
Appendix Figure 3. Estimation of the proportion of the null hypotheses in the dataset. The y-axis denote the $\hat{\pi}_0$(proportion of the null hypotheses) estimation dependent on the parameter λ, the x-axis denote the range of tuning parameter λ. A) Dataset is treated as two groups: all Shams vs. all post-injury samples; B) Dataset is treated as two groups: Sham3+90dpi vs. all the remaining samples



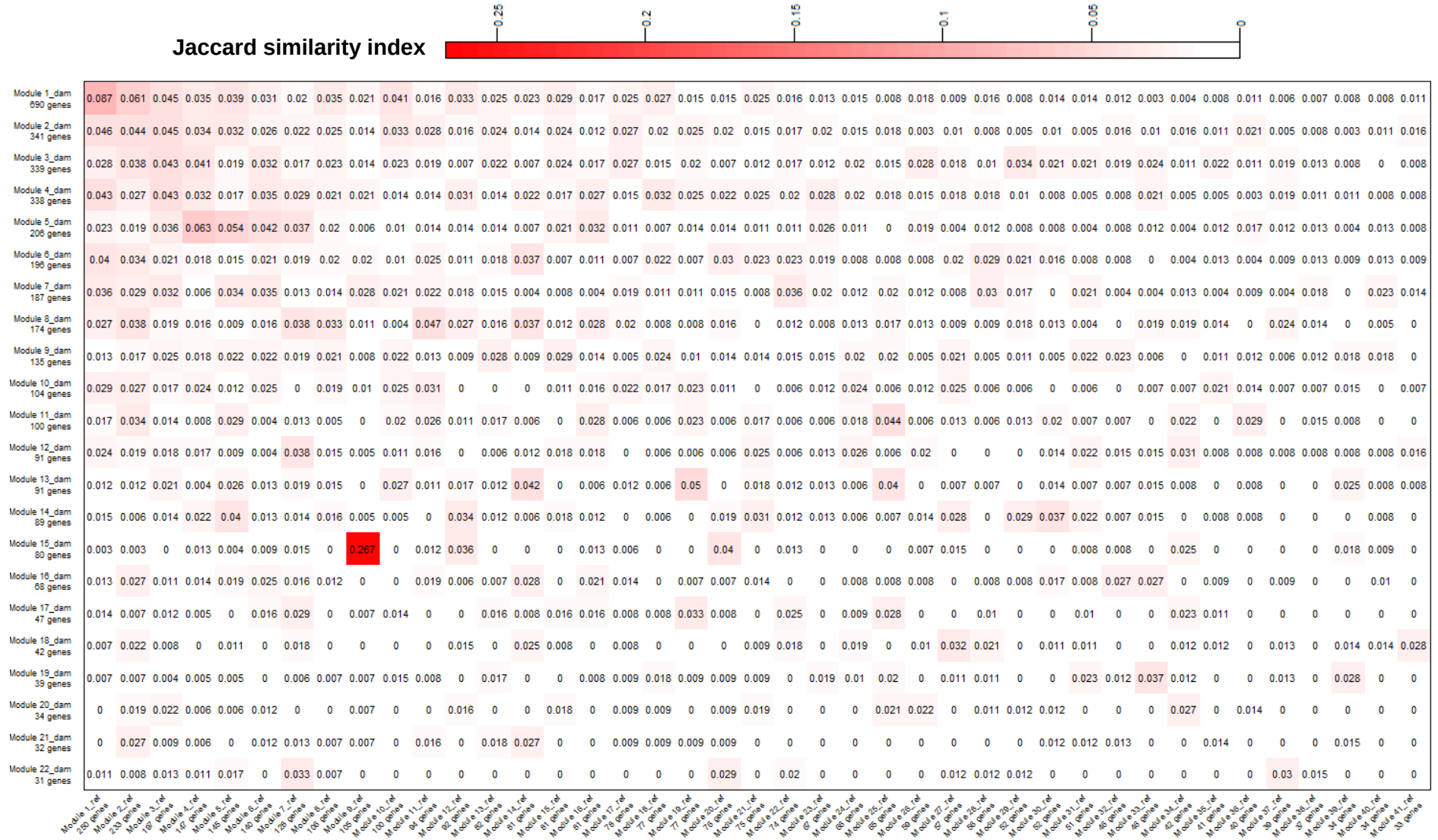
Appendix Figure 4. Module preservation statistics between Pearson's network modules and Spearman's and Kendall's network modules. Bonferroni corrected permutation p-value was calculated for each module in Pearson-based network as a preservation measure of module connectivity, density, or separability. Median Rank statistics is a composite measure of module preservation parameters (connectivity and density) independent of module size. Red dashed line indicates the lower limit of moderate preservation (Bonferroni corrected p-value < 0.05), blue dashed line indicates the lower limit of strong preservation (Bonferroni corrected p-value < 0.001). Details of the preservation analysis methods are described in (Langfelder et al., 2011).

Appendix Table 1. Comparison of the general network properties between global network, Reference state and Damaged state.

Network	Density	Centralization	Heterogeneity	Mean Clustering Coefficient	Mean Scaled Connectivity
Pearson's global	0.074	0.105	0.498	0.167	0.412
Reference state	0.100	0.133	0.611	0.257	0.429
Damaged state	0.032	0.059	0.672	0.127	0.353



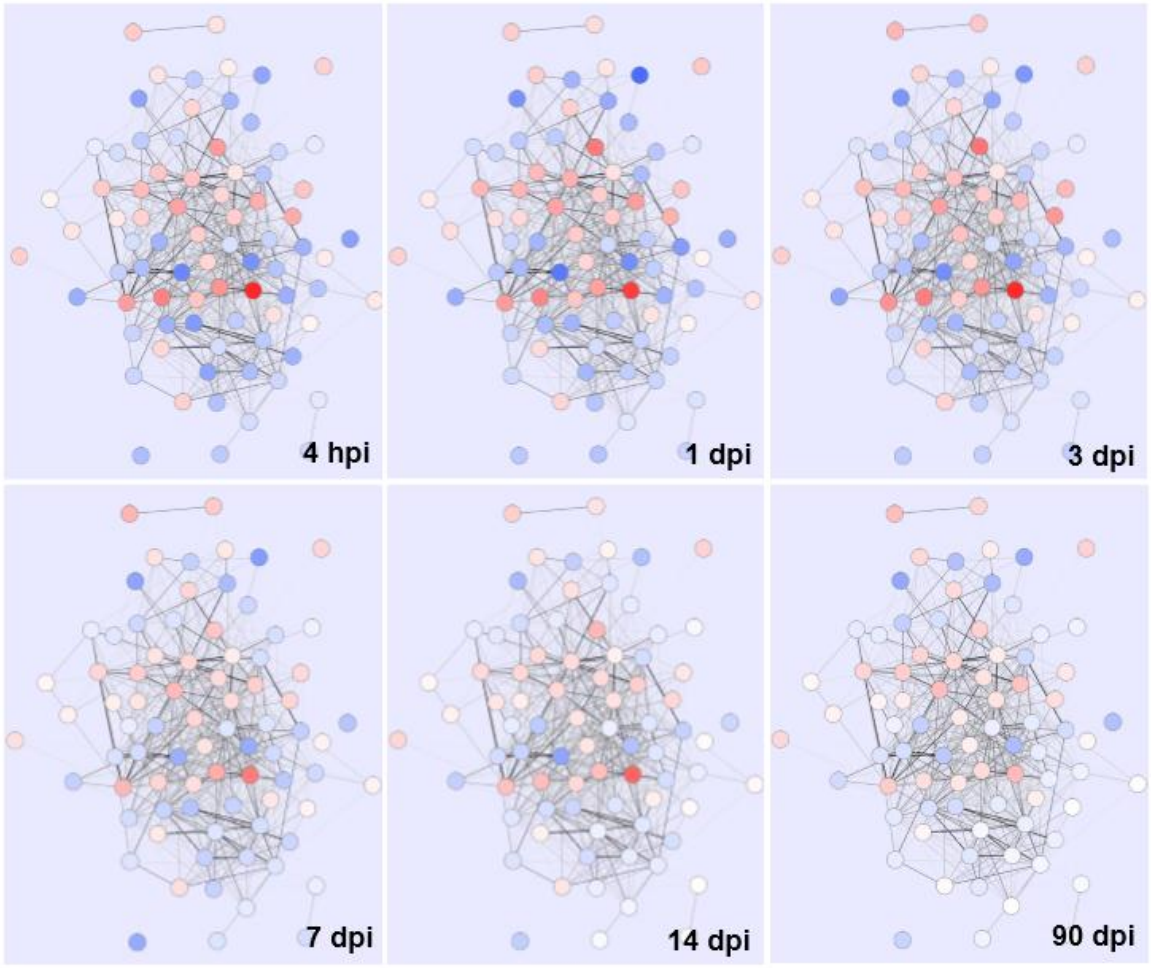
Appendix Figure 5. Gene dendrograms obtained by average linkage hierarchical clustering. (A-B) Vertical "leaves" of the dendrogram represent genes. The y-axis represents network distance, which is determined by 1 – topological overlap (TO). Values closer to 1 indicate greater dissimilarity of probe expression profiles across the samples. Color blocks below denote the module assignment determined by Dynamic Hybrid algorithm. A) Dendrogram constructed for the Reference state network with height cut off 0.951, containing 41 modules (determined by Dynamic Hybrid algorithm). B) Gene dendrogram constructed for the Damaged state network with height cut off 0.987, containing 22 modules (determined by Dynamic Hybrid algorithm).



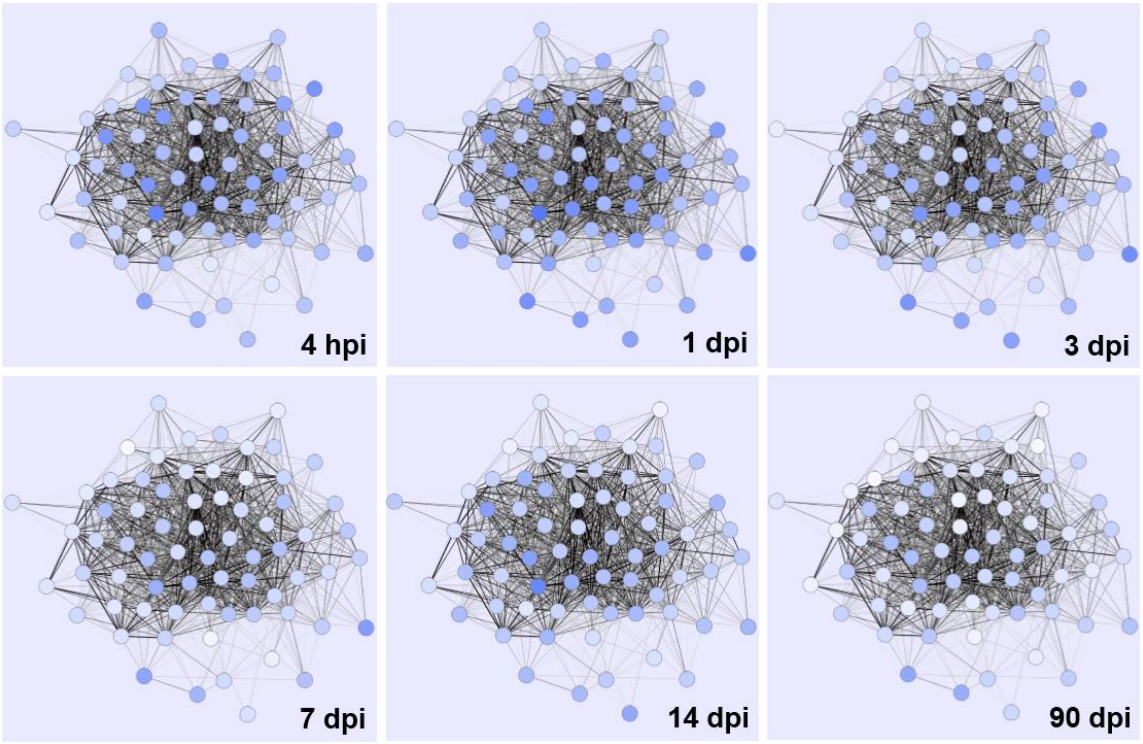
Appendix Figure 6. Jaccard similarity index between the modules detected by Dynamic Hybrid algorithm. On the left axis the modules denoted by "_dam" correspond to modules in Damaged state. On the bottom axis modules denoted by "_ref" correspond to modules in Reference state network. The greater saturated red color stands for the greater similarity.

Appendix Table 2. Intramodular topological properties. Modules detected by Dynamic Hybrid algorithm are denoted by A. Modules detected by ClusterONE are denoted by B.

Module	Number of genes	Intramodular Density	Intramodular Centralization	Intramodular Heterogeneity	Intramodular Clustering Coefficient	Intramodular Connectivity
1A	907	0,184	0,191	0,463	0,272	0,492
2A	464	0,134	0,185	0,554	0,228	0,421
3A	391	0,213	0,166	0,400	0,288	0,563
4A	318	0,210	0,189	0,440	0,299	0,528
5A	265	0,185	0,169	0,413	0,256	0,525
6A	183	0,199	0,194	0,461	0,288	0,509
7A	162	0,197	0,149	0,372	0,260	0,573
8A	156	0,124	0,137	0,507	0,234	0,478
9A	142	0,191	0,144	0,321	0,240	0,574
10A	94	0,232	0,166	0,366	0,302	0,588
11A	82	0,296	0,198	0,357	0,380	0,605
12A	79	0,192	0,145	0,336	0,245	0,576
13A	74	0,170	0,131	0,441	0,248	0,571
14A	62	0,207	0,128	0,406	0,281	0,626
15A	52	0,223	0,141	0,291	0,264	0,622
16A	46	0,325	0,181	0,250	0,370	0,653
1B	491	0,348	0,188	0,247	0,395	0,651
2B	496	0,354	0,188	0,231	0,399	0,654
3B	463	0,355	0,190	0,235	0,399	0,652
4B	480	0,340	0,197	0,243	0,388	0,633
5B	421	0,337	0,184	0,224	0,376	0,648
6B	316	0,351	0,149	0,169	0,375	0,704
7B	69	0,345	0,196	0,283	0,408	0,644
8B	14	0,993	0,003	0,006	0,993	0,997
9B	307	0,350	0,149	0,179	0,377	0,703
10B	313	0,328	0,157	0,223	0,365	0,678
11B	295	0,350	0,150	0,186	0,378	0,701
12B	297	0,351	0,152	0,187	0,378	0,699
13B	279	0,350	0,154	0,199	0,381	0,696
14B	276	0,350	0,157	0,198	0,381	0,692
15B	8	0,370	0,155	0,203	0,401	0,761



Appendix Figure 7. Visualization of the gene expression patterns in Module 12A. Nodes correspond to genes with colors indicating differential expression in relation to Sham3. Up-regulation is shown in gradation of red, down-regulation – in gradation of blue color. The edges denote correlation between the genes. Strong connections are visualized with darker and thicker edges, whereas weak connections appear thinner and translucent. Hpi indicates hour post injury; dpi, day post injury.



Appendix Figure 8. Visualization of the gene expression patterns in Module 7B. Nodes correspond to genes with colors indicating differential expression in relation to Sham3. Up-regulation is shown in gradation of red, down-regulation – in gradation of blue color. The edges denote correlation between the genes. Strong connections are visualized with darker and thicker edges, whereas weak connections appear thinner and translucent. Hpi indicates hour post injury; dpi, day post injury.