Math. Model. Nat. Phenom.
Vol. 7, No. 1, 2012, pp. 337-368
DOI: 10.1051/mmnp/20127115

Analysis of the Growth Control Network Specific for Human Lung Adenocarcinoma Cells

G. Pinna[1], **A. Zinovyev** [2,3,4], **N. Araujo** [1], **N. Morozova** [1 *] **and A. Harel-Bellan**[1]

[1] CNRS FRE 3377, CEA Saclay, Gif-sur-Yvette, F-91191
and Universite Paris-Sud, Gif-sur-Yvette, F-91191, France
[2] Institut Curie, 26 rue d'Ulm, Paris, France
[3] INSERM, U900, F-75248 Paris, France
[4] Mines ParisTech, Centre for Computational Biology, F-77300 Fontainebleau, France

Abstract. Many cancer-associated genes and pathways remain to be identified in order to clarify the molecular mechanisms underlying cancer progression. In this area, genome-wide loss-of-function screens appear to be powerful biological tools, allowing the accumulation of large amounts of data. However, this approach currently lacks analytical tools to exploit the data with maximum efficiency, for which systems biology methods analyzing complex cellular networks may be extremely helpful. In this article we report such a systems biology strategy based on the construction of a Network for a biological process and specific for a given cell system (cell type). The networks are created from genome-wide loss-of-function screen datasets. We also propose tools to analyze network properties. As one of the tools, we suggest a mathematical model for discrimination between two distinct cell processes that may be affected by knocking down the activity of a gene, i. e., a decreased cell number may be caused by arrested cell proliferation or enhanced cell death. Next we show how this discrimination between the two cell processes helps to construct two corresponding subnetworks. Finally, we demonstrate an application of the proposed strategy to the identification and characterization of putative novel genes and pathways significant for the control of lung cancer cell growth, based on the results of a genome-wide proliferation/viability loss-of-function screen of human lung adenocarcinoma cells.

Key words: systems biology, network analysis, lung adenocarcinoma, genome-wide screen
AMS subject classification: 92B15, 92B05

*Corresponding author. E-mail: morozova@vjf.cnrs.fr

1. Introduction

Genome-wide loss-of-function (LOF) screens are powerful experimental techniques using RNA interference (RNAi) to obtain a comprehensive view of gene function in a given biological system. In these approaches, the genes from the whole genome are individually and systematically knocked down in a cell-based assay, and the resulting phenotypical modifications are quantified with aid of a relevant reporter. Usually this knockdown is performed by application of synthetic siRNAs (short interfering RNAs), which are short non-coding RNAs, each specifically designed for the knock-down of a given gene in the given type of genome (e.g. human, mice, fly, etc.). Genome-wide LOF screens are currently widely exploited biological tools, which generate a huge amount of data, but the analytical tools applied for processing these data and arriving at conclusions are generally not commensurate with the biological capability of this strong and expensive method. In practice, the results of such screens are either finally narrowed down to the discovery of 1-2 specific genes only, or, alternatively, undergo too-general systems biology analysis with conclusions barely (or not at all) meaningful with respect to real biological needs [5]-[14]. Here, we try to find a way between these Scylla and Charybdis, and to propose a powerful analytical tool, based on the concept of Specific Network, meaning a Network corresponding to a certain process and specific for a given cell system (e.g., given cell line, given cancer type, etc.). We suggest rules for construction of such a network based on datasets of genome-wide LOF screens performed in such a system, and provide tools for analysis of its specificity (non-randomness). The analysis of such a Specific Network helps to identify a set of genes and pathways in the given biological system that is significant for the given process under investigation (e.g. proliferation, differentiation, etc.).

It is very important to point out that network analysis, first of all, helps to ferret out those genes which are in fact very important for the process in the given system, but which were not identified in the screen due to technical/experimental reasons. Inclusion of additional genes-connectors to the genes found in the screen for creating a Specific Network, and next analysis of their characteristics in the corresponding network, helps to prove the importance of such genes using the ”guilty by association” principle, despite the fact that these genes were not justified by the initial LOF screen.

Next we show an application of proposed systems biology approaches for the identification of novel genes and pathways potentially associated with lung cancer, by analysis of the Specific Network, based on the results of a genome-wide LOF proliferation/viability screen of human lung adenocarcinoma cells (cell line A549). The A549 cell line is a human lung adenocarcinoma cell line, which belongs to Non-Small-Cell Lung cancers (NSCLC) and bears the K-Ras oncogene. K-Ras is a G-protein acting directly downstream of the epidermal growth factor receptor (EGFR), and an essential component of the EGFR pathway, which is one of the major signaling pathways deregulated in lung cancer [20]. Indeed, K-Ras mutations are found in 20-40% of lung adenocarcinomas [21]. An important feature of the A549 cell line is that it expresses normal (wild type) tumor suppressor protein p53. Non-small cell lung cancers, and particularly the NSCLC adenocrcinoma subtype, are one of the main causes of human lethality [15, 16], and the molecular mechanisms underlying its progression are still poorly understood. The therapies targeting well known pathways, such as the EGFR pathway [1, 2, 3, 17], were shown to be effective for only a small fraction of NSCLC tumors. A very few reports show a role for specific genes, such as EML4-ALK [4] or CRK

[18], in lung adenocarcinoma progression, and they do not clarify the corresponding pathways.

A genome-wide RNAi-based LOF screen performed in our lab on the human lung adenocarcinoma A549 cell line resulted in the discovery of 203 genes negatively regulating cell proliferation, with or without associated effects on cell viability (assessed by a metabolic assay and a cytotoxicity reporter assay, respectively). In this article, we report the construction of A549 Specific Network using the results of this genome-wide RNAi screen, and suggest tools for its analysis.

The systems biology analysis of this A549 Specific Network revealed around 30 genes as being significant ones for human lung adenocarcinoma cell growth, i.e., influencing cell proliferation activity, growth, or viability. Discrimination between participation of these genes in the processes of proliferation or the maintenance of cell survival was arrived at using mathematical modeling of these processes. This involved retrieving the corresponding proliferation and cell survival subnetworks from the A549 Specific Network, thereby elucidating several pathways and modules in this system.

2. Results

2.1. Biological Background - hit genes found in genome-wide loss-of-function screen of A549 cells

siRNAs (short interfering RNAs) are short non-coding RNAs that can induce sequence-specific gene knockdown via a silencing complex. In the course of a genome-wide LOF screen, a library containing siRNAs targeting each gene from the genome of interest is assembled, and next applied as separate samples to the cells growing in culture. This allows evaluation of the effect on the given biological process of knocking down each individual gene in the given cells, through quantification of a phenotypical readout corresponding to the process being investigated. For identification of the genes controlling cell proliferation, growth and viability in human lung adenocarcinoma cells (A549 cells), the WST-1 Cell Proliferation Assay was chosen as appropriate readout. A colorimetric assay, this semi-quantitative test estimates mitochondrial metabolic activity, thus giving an indirect indication of the number of live cells in a sample relative to an appropriate control population. In parallel, using separate aliquots from the same cultures, cell death was evaluated by a Lactate Dehydrogenase (LDH) release colorimetric assay, which measures the relative proportion of dead cells by quantifying the release of LDH enzymes into the cell culture supernatants upon cell membrane rupture (i.e., cell death).

For the technical details of genome-wide screen of A549 cells (Primary genome-wide screen, Candidate genes selection, Secondary (confirmation) screen and Hit genes selection), see Materials and Methods. Briefly, biostatistics analysis of the screening data identified 203 genes, mostly influencing A549 cell growth. As expected, well-characterized regulators of mitosis and cytokinesis (KIF11, PLK1, CDC2L1-2, E2F1) were identified as top hits, highlighting the reliability of the screen. There was also a characteristic pattern of RNAi screens with a high hit number involved in general gene expression (POLR2 subunits, elongation initiation factors, ribosomal subunits, spliceosome components). 16% of the total identified genes were not yet functionally annotated,

and 12% of the siRNAs targeted predicted genes, i.e., genes that correspond to open reading frames (ORFs) from a database and are not yet confirmed as real genes.

2.2. Specific Network construction and analysis of its specificity

2.2.1. General strategy for Specific Network construction

Generally speaking, a Specific Network for a biological system may be defined as a network built on the hit genes selected in the genome-wide screen performed for this system, and proved to be significantly different from a random network built using the same random number of genes and the same rules for network construction. In reality, depending upon the chosen readout for a given genome-wide screen, its Specific Network does not reflect the whole biological system, but more precisely, the particular process investigated in this biological system (e.g., cell proliferation, differentiation, oxidative stress, etc.). Here, we addressed the question of growth control in A549 human lung adenocarcinoma cells, and used a cell proliferation/viability assay as readout; thus in our case we will construct a Specific A549 Growth Control Network.

Another important remark is that a Specific Network for a given biological system explicitly depends on the global ineraction databases(s) used for its construction (e.g., protein-protein interaction database(s), protein-gene interaction database(s), metabolic database(s), etc.). Of course, the best results for Specific Network construction will be obtained by using the "merger" of protein-protein interaction, protein-gene interaction and metabolic databases. However, using only one global database for specific network constraction may be enough for addressing some important questions about particular molecular mechanisms. For example, for investigation of the signalling pathways underlying cell growth control in lung cancer, it is reasonable to construct a Specific Network using only protein-protein interaction (PPI) database(s), as most of these pathways consists of protein-protein interactions. Of course, information about other types of regulation will not be included in such an analysis, yet the results obtained with such a Specific Network can still be considered to be important ones, as long as the specificity (non randomness) of the network is well proven. To be accurate on this point, the Specific Network should be called, for example, the "A549 Specific PPI Network", if only protein-protein interaction database(s) were used for its construction.

Thus, strictly speaking, the "A549 Specific Network" that we analyse in this study and use as an example of an application of several new system biology tools for network analysis, should have the "full name" of "Specific PPI Network for Growth Control of A549 human lung adenocarcinoma cells".

We suggest the following rules for constructing a Specific Network using the results of a genome-wide screen:

1. The Network should be built on the hit genes (set of genes found in the genome-wide screen and statistically proven to be the most significant for the effect considered), using global ineraction databases (e.g., protein-protein interaction database(s), protein-gene interaction database(s), metabolic database(s), etc.). Normally, the proportion of hit genes should be around 1-5% of all the genes tested in the genome-wide screen.

2. If the connectivity of the obtained network, which we call the "Direct Network" (DN), is sufficient, the DN can be used as a Specific Network for further analysis. We may suggest that the criteria should be the size (in nodes) of the maximal connected component in the network, and this size for accepting the Direct Network as a Specific Network should be not less than 50% of the total number of genes in the hit list, proven to be statistically significant in this system; the best situation for accepting the Direct Network as a Specific Network is when the size of the maximal connected component is equal to or greater than 66-75% of the total number of genes in the hit list.

3. If the size of the maximal connected component of a Direct Network is smaller then 50% the number of hit genes, then the Specific Network should be built using hit genes together with Connectors (Network with Connectors, NC). This means that one should add to the nodes corresponding to hit genes all their first neighbors (FN) from an interaction database (FN are nodes with a 1-node distance from a given gene), and then choose for the Specific Network only those among the FN of the hit genes that each have more than one hit gene as their own FN (connectors). The advanced tool for construction of a "Network with Connectors" is BiNoM software [19].

4. In the case where the Direct Network has a size (in number of nodes) around 50% of the hit list, or if it is composed of several medium-size connected components (=subnetworks) that are not connected between each other (i.e., not producing the common network), the following strategy seems to be the best: to built the Network with Connectors, then check the specificity of both networks (Direct Network and Network with Connectors), and choose as Specific Network the one which has more specificity according to the tests.

In all cases the specificity of the obtained network must be checked.

2.2.2. Analysis of network specificity

We suggest here two ways for checking the specificity of the obtained network:

a) by evaluating *the global compactness of the network* and comparing it with that obtained for networks built on sets of randomly chosen genes;

b) by comparing its *degree of connectivity* with that of networks built on sets of randomly chosen genes.

a) *Global compactness of the network (Specificity as average internode distance)*

We choose a given list of genes for creating the Specific Network, and characterize their distribution in the global PPI network by a single number:

$$G(genes, network) = \frac{1}{N_{genes}(N_{genes}-1)/2} \sum_{i=1}^{N_{genes}} \sum_{j=i+1}^{N_{genes}} dist_{network}(gene_i, gene_j) \qquad (2.1)$$

where $dist$ is the number of edges in the shortest path connecting hit i and hit j in the global network and N_{genes} is the number of genes in the list. This can be called the *global compactness* measure G.

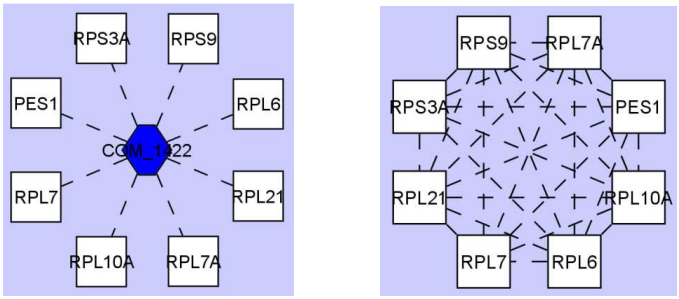
b) *Degree of connectivity of a network (Specificity by Connectivity test)*

Another test of non-randomness, inspired by percolation theory, can be proposed. On the figures showing the extracted networks, one can see that these networks contain a number of connected components of a certain size. It is important to estimate the probability of obtaining such connected components by chance.

2.2.3. Including information on protein complexes

Discussing the general strategy of analysis of network specificity, it is necessary to note that information on existing protein complexes is of the utmost importance for evaluation the specificity of the given network in comparison with the network built on a randomly chosen gene set. In other words, we would like to distinguish between those connections between proteins in the network belonging to well-known protein complexes, on the one hand, and those which were not previously investigated, on the other. This is especially important for the LOF screen data analysis, where the presence of the main housekeeping complexes is expected (e.g., elongation initiation complex, complex of ribosomal subunits, etc.), and thus high level of connectivity inside these complexes must not influence the evaluation of network specificity.

Two different approaches may be used to include the information about complexes (Figure 1):
1) representation of protein complexes as separate nodes, and
2) representation of protein complexes as clicks;
and the final result of network specificity evaluation can be strongly influenced by this choice.



1) Complex as a separate node 2) Complex as a click

Figure 1: Two possible representations of complexes in the PPI network.

Next we will illustrate the application of the two specificity tests to the Direct A459 Network and compare the specificity results obtained for this network when one or the other of the above-mentioned strategies for adding information about protein complexes is applied.

2.2.4. Analysis of A549 network specificity

The A549 Growth Control Network was created on 203 hit genes found in our genome-wide proliferation/viability LOF screen, using information on Protein-Protein Interaction (PPI) from the Human Protein Reference Database (HPRD, http://www.hprd.org) and visualizing it by Cytoscape Software [22]. In the network constructed using the Protein-Protein Interaction Database, all nodes

represent proteins produced from the corresponding genes, and edges represent interactions between these proteins. The Direct PPI A549 Network contains only 134 hits out of the 203 used to create the network, due to the fact that predicted genes (ORFs) and non-characterized genes are not included in the HPRD database (Figure 2).

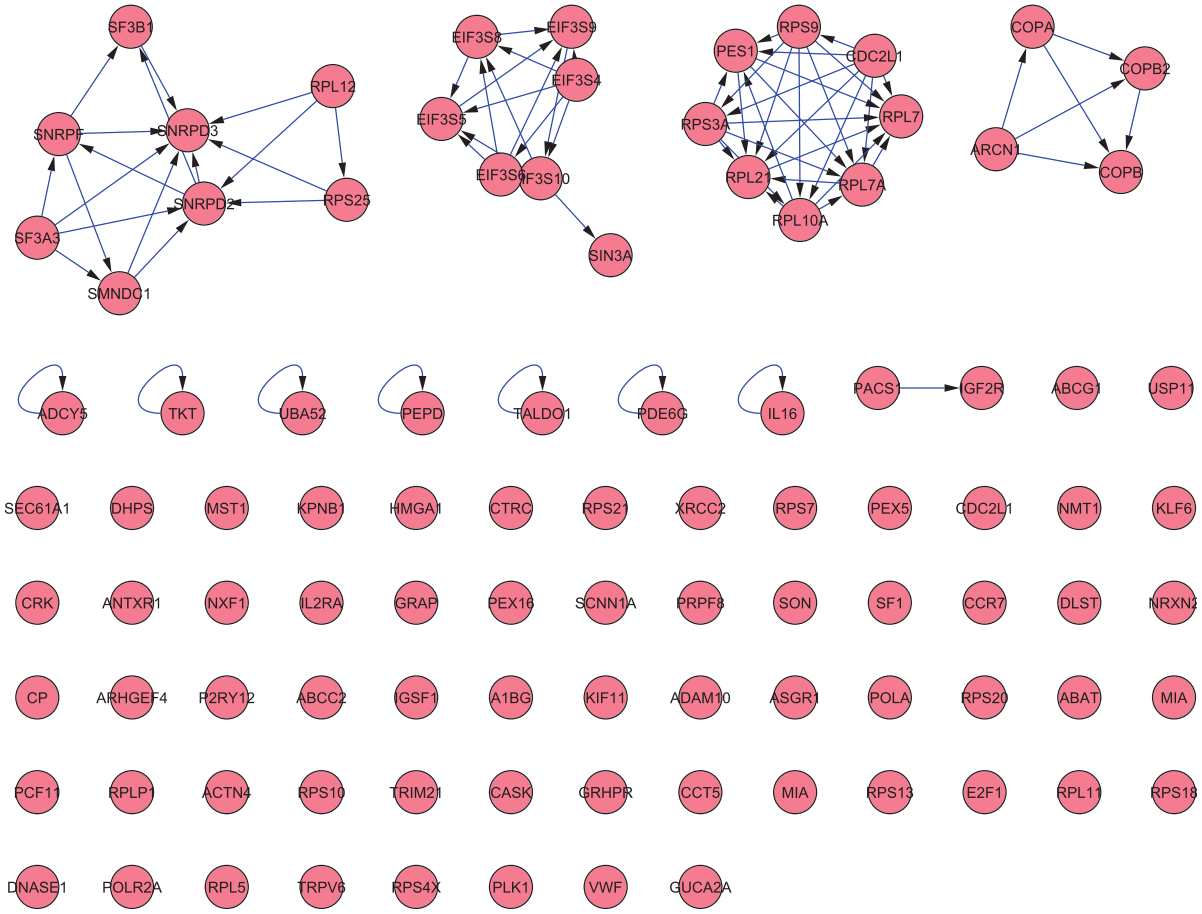


Figure 2: Network built on the hit list of 203 genes found in the genome-wide screen (Cytoscape Software, HPRD database, 134 genes found).

We will check the specificity of the A549 Direct Network and will compare the obtained results for 3 possible ways of including information about protein complexes:

1) No complexes are taken into account. Direct interactions only are added (Figure 2).

2) Complexes are added as clicks (Figure 3).

3) Complexes are added as additional nodes (Figure 4).

a) *Global compactness of the A459 Direct Network*

We will calculate the p-value of the global compactness measure G for the A549 Direct Network in several ways (Table 1):

1) we take 10,000 randomly sampled sets of 134 hits (purely random);

2) we take 10,000 randomly sampled sets of 134 hits but such that each set has nodes with the

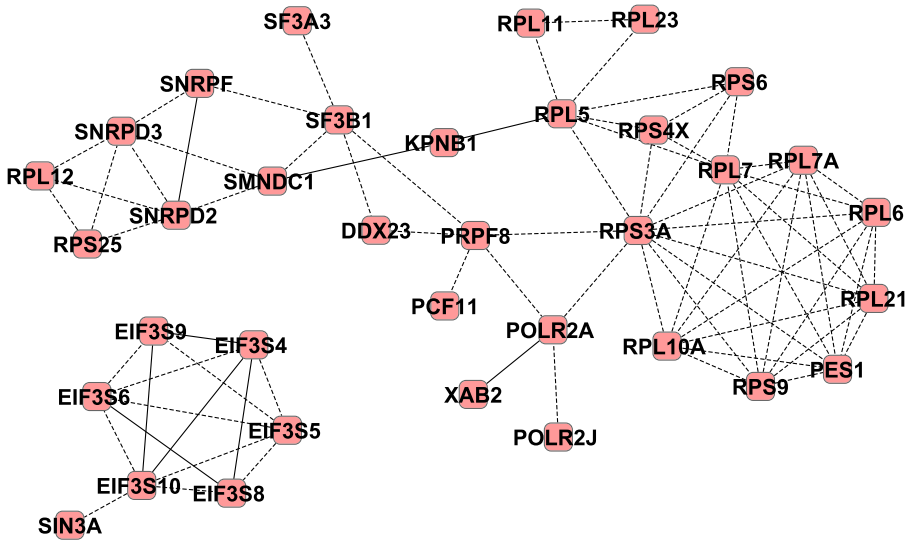


Figure 3: Complexes are added as clicks. Interactions inside complexes are shown as dashed lines. There are 70 ”inside-complex” additional interactions (plus direct ones), which creates two big connected components.

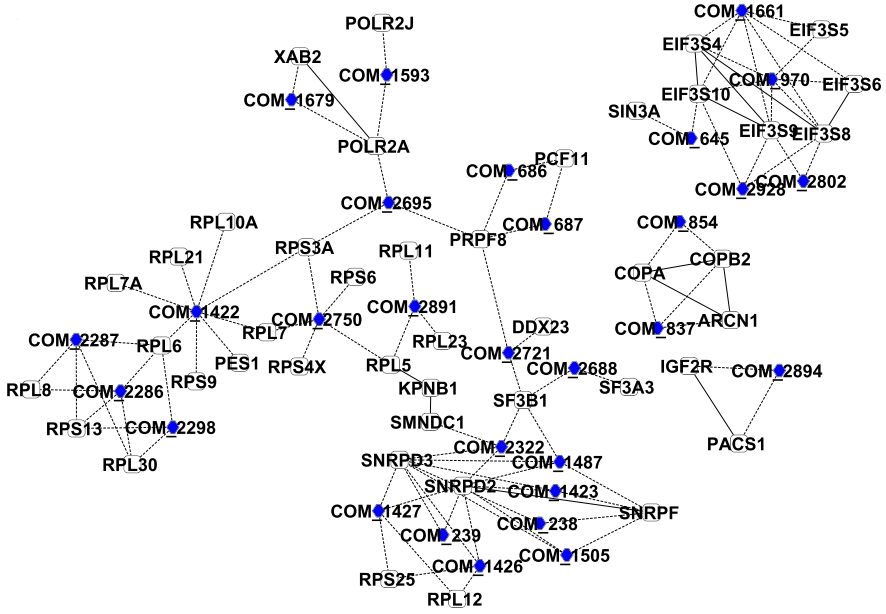


Figure 4: Complexes are added as additional nodes. Interactions inside complexes are shown as dashed lines. There are 29 ”complex” nodes (shown as blue hexagons) and 134 hit nodes on the graph, 16 direct and 95 ”inside-complex” interactions (dashed lines). Two large and 2 small connected components are created. Note: the color code is readable only in the electronic version.

same number of neighbors as in the hit set (conserving connectivity); in the case of the network where complexes are represented as nodes, the random nodes are chosen from 'non-complex' nodes.

3) we take 10,000 randomly sampled sets of 134 hits. We sample complexes presented in the hit set accordingly to the following rule: for each complex present in the initial hit list with k components (where k should be greater than or equal to M, the complex representation threshold), one randomly adds k components of the same complex to the sampling. Those proteins that do not belong to the represented complexes are sampled from the global network, with preservation of connectivity (as in case 2 above). Thus, each sampling generates 134 hits such that all complexes represented in the hit list are also represented by the same number of complex members in the sampling. Changing M, one can take into account only the most represented complexes, or else all complexes present in the hit list (conserving connectivity and complexes).

Analysis of the results in Table 1 leads to the following conclusions:

1) The distribution of hits is significantly more compact than random in the network where inside-complex interactions are represented as clicks. This significance remains valid for the purely random choice of nodes, for the choice conserving the connectivity distribution and for the sampling choice conserving representation of complexes represented in the initial hit list.

2) This compactness is explained mostly by inside-complex interactions, since, after removal of these interactions, the results become border-line (p-value = 0.1) if one conserves the connectivity distribution, and completely non-significant if one conserves the representation of complexes that are represented in the initial hit list.

	Network without complexes	Network with complexes as clicks	Network with complexes as nodes
Purely random	**0.020**	**<0.0001**	**0.005**
Conserving connectivity	**0.100**	**0.005**	0.128
Conserving connectivity and complexes $\boldsymbol{k} >= 5$	0.270	**0.013**	0.229
Conserving connectivity and complexes $\boldsymbol{k} >= 3$	0.451	**0.030**	0.331
Conserving connectivity and complexes (all)	0.563	**0.080**	0.453

Table 1: P-values of G calculated using different settings. Significant p-values are underlined and in bold. Borderline significant p-values are only in bold

An additional methodological conclusion is that the representation of complexes as nodes, while useful for visualization, does not allow us to make conclusions about compactness (the results are not significantly different from random sampling when preserving at least the connectivity distribution).

b) *Degree of connectivity of the A549 Direct Network*

We will estimate the probability of obtaining such connected components as we obtained in the A549 Direct Network (Figure 2) by chance. We implement this idea using the same networks and sampling strategies as for the previous test. The results are presented in Table 2. We can see that the appearance of a big connected component, of size 28 in the network with complex information included, is significant only if we use sampling preserving the connectivity, but is expected to happen if we also consider the representation of (at least the largest) complexes.The other connected component, of size 7, can be arrived at by chance with the sampling conserving the connectivity.

Thus, the main conclusion from the analysis of specificity of the Direct Network by the connectivity test is that we cannot consider this network to be specific.

Specificity of the A549 Network with Connectors (NC)

As the connectivity of the A549 Direct Network appears to be very small (the size of the maximal connected component is 8 out of a possible 134 nodes), then according to the rules we have proposed for Specific Network construction, the Specific Network for A549 lung adenocarcinoma cells should be built on the hit genes together with connectors (Figure 5). In this case, 103 of the 134 nodes in the HPRD-curated hit list are present in the maximal connected component, which now consists of 245 nodes in total. This gives 50.7% of hit genes involvement in the network (77.4% of the HPRD-curated hit list), and if we prove that this network is specific, we will be able to derive meaningful conclusions from its analysis.

It is very important to note that, according to its definition, the specificity measured as average internode distance (G) is applicable to the list of genes (hit list in our case), and thus it will be the same for both the Direct Network and for the Network with Connectors that are built using the same hit list. And, as we already showed, this specificity is well proven (for the network with complexes as clicks, with conserving connectivity, and considering all complexes, the p-value for G is less than 0.1).

The evaluation of A549 NC specificity by the Degree of connectivity test also shows statistically significant results: the probability of obtaining 103 of the 134 random nodes in one connected component, by applying the strategy with connectors, using conserving connectivity, and considering all complexes, is around 0.05 (Table 3). This means that the A549 Specific Network, built as a Network with Connectors on the hit genes of our genome-wide screen using the HPRD database, is indeed found to be specific by both specificity tests.

2.3. Analysis of A549 Specific Network

2.3.1. Defining the most important nodes in the A549 LOF Specific Network

There are two main parameters which determine the relative importance of a node in the graph/ network: its connectivity and its centrality (betweenness) in the given network.

	Network without complexes	Network with complexes as clicks
Distribution of sizes of connected components	2:5 3:2 5:1	2:4 3:1 7:1 28:1
Purely random	2:8, p-value=0.085 3:3, p-value=0.074 5:1, p-value=0.235	2:7, p-value=0.363 3:3, p-value=0.165 **7:2, p-value=0.001** **28:1, p-value=0.000**
Conserving connectivity	2:8, p-value=0.175 3:3, p-value=0.115 5:1, p-value=0.184	2:7, p-value=0.904 3:3, p-value=0.856 7:2, p-value=0.221 **28:1, p-value=0.005**
Conserving connectivity and complexes $k >= 5$	2:8, p-value=0.415 3:3, p-value=0.465 5:1, p-value=0.753	2:7, p-value=0.536 3:3, p-value=0.602 7:2, p-value=0.309 28:1, p-value=0.372
Conserving connectivity and complexes (all)	2:8, p-value=0.762 3:3, p-value=0.834 5:1, p-value=0.844	2:7, p-value=0.272 3:3, p-value=0.549 7:2, p-value=0.264 28:1, p-value=0.928

Table 2: Expected number of connected components of certain sizes in the extracted networks (only direct connections - inside-complex or others - are considered). In the first row, the records X:Y show the distribution of connected components' sizes, X=size of the connected component, Y=number of connected components of size X. In the rows beneath, X:Y means Y=number of connected components of size X or larger. The p-value is estimated for the appearance of Y connected components of size X or bigger (for example, for appearance of one large connected component of size 28 or more, which is the case for the hit list).

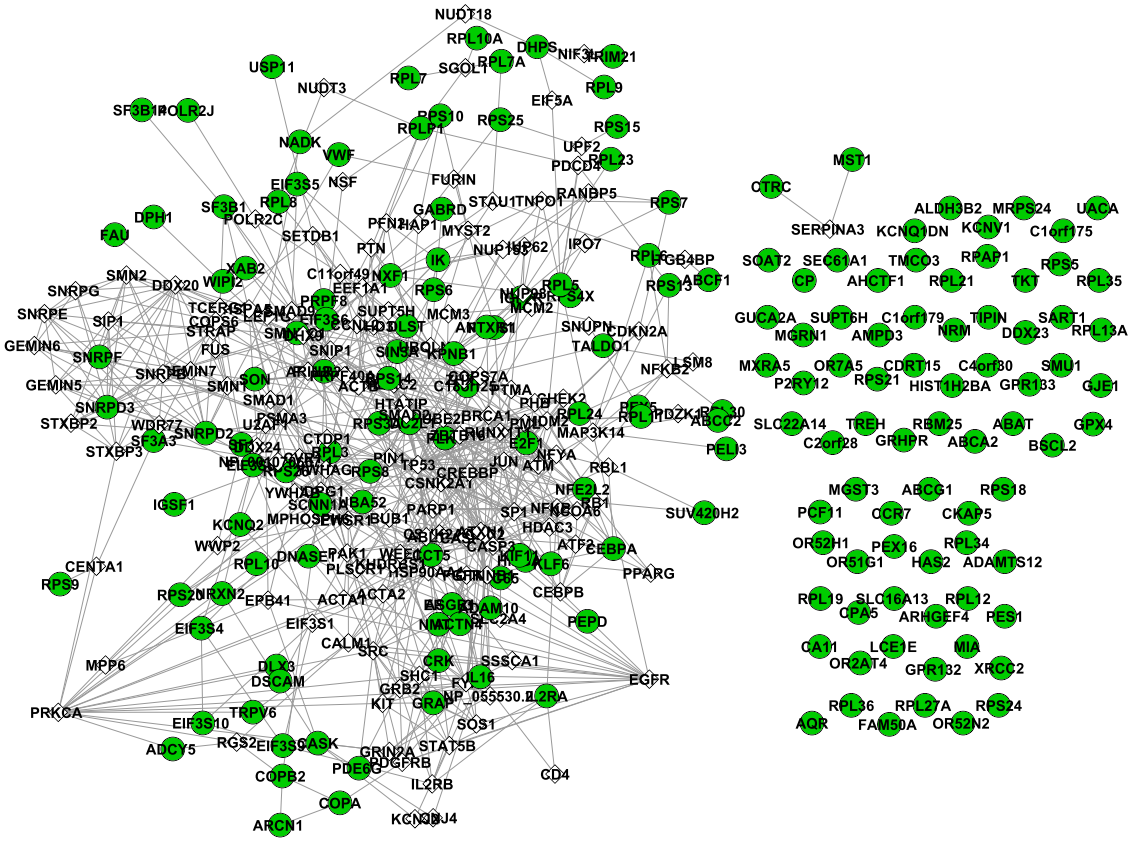


Figure 5: Network including hit genes (green, circle nodes) and connectors (genes connected two at least two hit genes: white, rhomb nodes) (Cytoscape Software, HPRD database). Note: the color code is readable only in the electronic version.

	Network without complexes	Network with complexes as clicks
Distribution of sizes of connected components	103:1	110:1
Purely random	**103:1, p-value=0.000**	**110:1, p-value=0.000**
Conserving connectivity	**103:1, p-value=0.001**	110:1, p-value=0.102
Conserving connectivity and complexes $k >= 5$	**101:1, p-value=0.013**	110:1, p-value=0.067
Conserving connectivity and complexes (all)	101:1, p-value=0.154	**110:1, p-value=0.047**

Table 3: Expected number of connected hits in one connected component for the Network with Connectors.

The connectivity of a node is determined by its number of connections (edges) with its first neighbors, and the nodes with the highest level of connectivity (named hubs) are thought to be of the most importance for the given system.

Betweenness is one of the most frequently used measures of the centrality of a node within a network, showing the significance of a node for the pathways within a network.

Namely, nodes with higher levels of betweenness occur on more shortest pathways between other nodes in the network, than those that have lower levels of betweenness:

$$B(n) = \sum \left(\frac{P_{ab}(n)}{P_{ab}} \right), a \neq b \neq n, \tag{2.2}$$

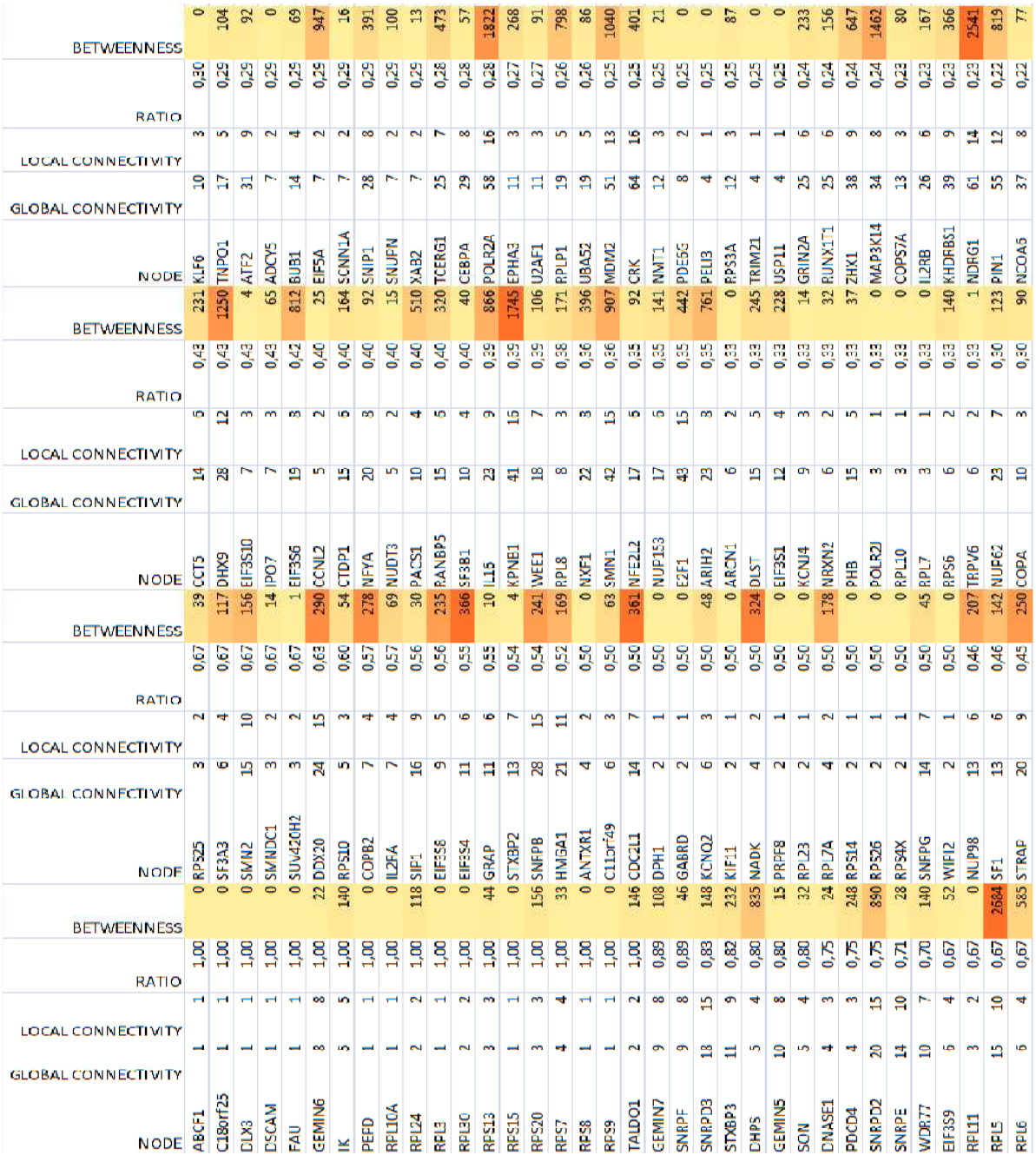
where $B(n)$ is the betweenness of the node n, P_{ab} is the number of shortest paths from node a to node b, and $P_{ab}(n)$ is the number of shortest paths from a to b that pass through a node n.

The connectivity and betweenness of all 245 genes (proteins) in the A549 Specific Network (including hit genes and connectors) was analysed (Table 4). This analysis identified several proteins (equivalent to the corresponding genes) with very high levels of connectivity in the given network, i.e., equal to or greater than 10 (Table 4, first column). However, it is also very important to know the level of connectivity for the given protein in the HPRD global network, and then to evaluate the ratio of the connectivity of a protein within the given network to its connectivity in the HPRD global network (Table 4, second and third columns). This ratio is an especially important criterion in the case of proteins with a high level of global connectivity, because these proteins may use only a small percentage of their connectivity potential within the given network, while still having very high level of "absolute connectivity". The fourth column presents the value of betweenness for each protein.

From this analysis we may derive the following main conclusions about the most important genes in this system:

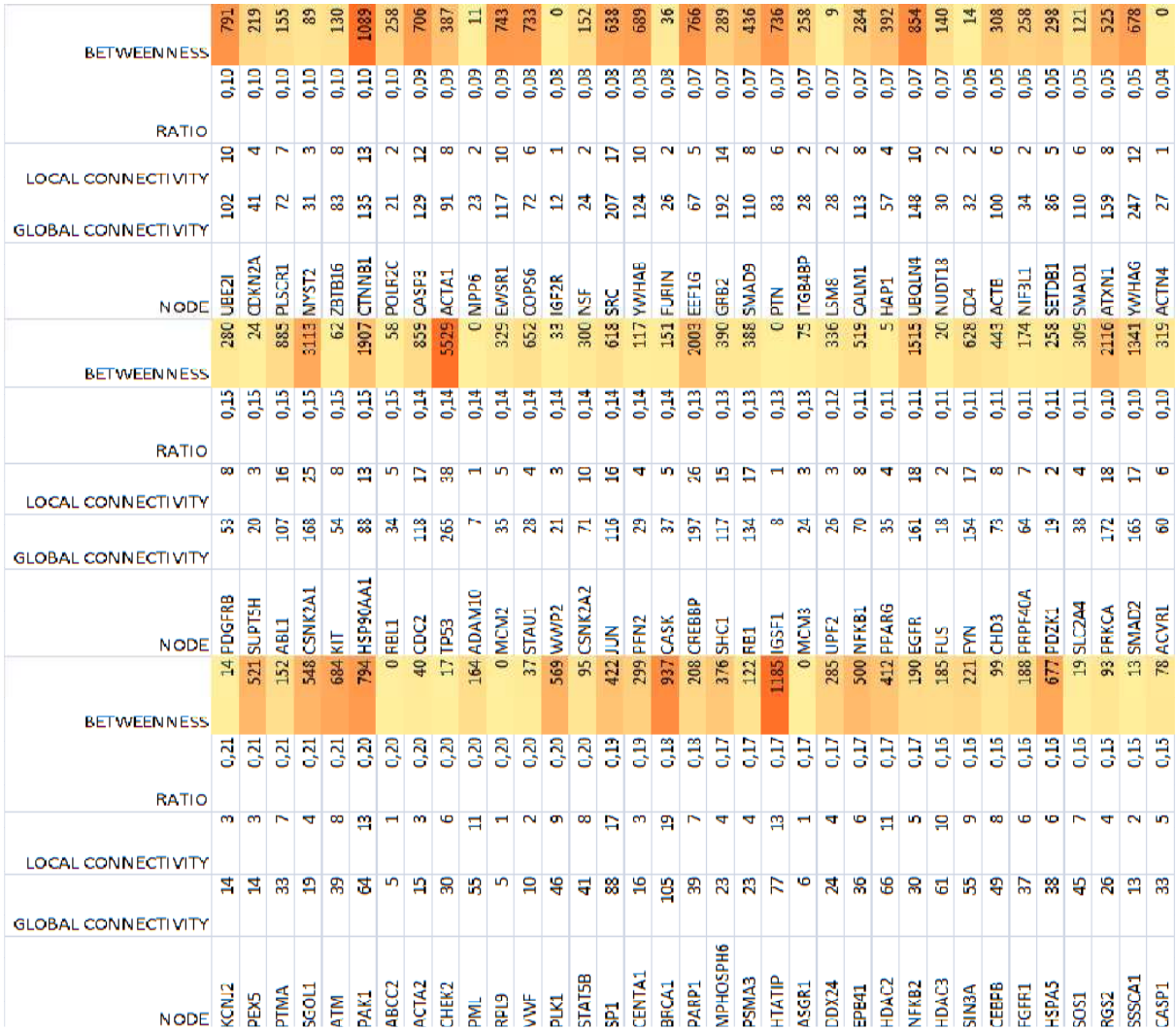
1. There are several genes (proteins) with significantly high levels of connectivity (> 10) in the Specific Network (two times or more greater than the average connectivity of a node in the global HPRD Network). Besides the expected pivotal proteins in cell proliferation and cell survival (such as TP53, RB1, SMAD2, JUN, E2F1, CDC2, EGFR, POLR2A, CASP3, spliceosome components (SNRPs), casein kinase, histones deacetylases, protein kinase C), and those already mentioned to be important in lung cancers (EGFR, HMGA1, GRB2), there are some that are less obvious, the presence of which in the list of hubs may point to the importance of these genes in A549 cells. The list of these genes includes well-known cancer genes that have not previously been shown to be important for lung cancer: BRCA1, CREBBP, SRC, MDM2, CRK and other proteins from different functional groups, NDRG1, PML, PAKI, SP1, PIN1, KPNB1, HTATIP, ABL1, SHC1, SMN1, FYN, DDX20, YWHAG, DHX9, CTNNB1, HSP90AA1. The three proteins from this list bearing the highest levels of connectivity are TP53, CREBBP, and CSNK2A1 (connectivity values of 38, 26, and 25, respectively).

2. Proteins that show the highest values for the ratio of connectivity-in-the-specific-network to connectivity-in-the-HPRD-global-network and also have a connectivity in the specific network greater or equal to 4 belong mostly to housekeeping-protein complexes: EIF3S8, EIF3S4, EIF3S9 - proteins of the elongation initiation complex RPL5, RPL6,RPS7 - ribosomal subunit proteins



NODE	GLOBAL CONNECTIVITY	LOCAL CONNECTIVITY	RATIO	BETWEENNESS	NODE	GLOBAL CONNECTIVITY	LOCAL CONNECTIVITY	RATIO	BETWEENNESS	NODE	GLOBAL CONNECTIVITY	LOCAL CONNECTIVITY	RATIO	BETWEENNESS	NODE	GLOBAL CONNECTIVITY	LOCAL CONNECTIVITY	RATIO	BETWEENNESS
ABCF1	1	1	1,00	0	RPS25	3	2	0,67	39	CCT5	14	5	0,43	231	KLF6	10	3	0,30	0
C18orf25	1	1	1,00	0	SF3A3	6	4	0,67	117	DHX9	28	12	0,43	1250	TNPO1	17	5	0,29	104
CLIC3	1	1	1,00	0	SMIN2	15	10	0,67	156	EIF3S10	7	3	0,43	4	ATF2	31	9	0,29	92
DSCAM	1	1	1,00	0	SMNDC1	3	2	0,67	14	IPO7	7	3	0,43	65	ADCY5	7	2	0,29	0
FAU	1	1	1,00	0	SUV420H2	3	2	0,67	1	EIF3S6	19	3	0,42	812	BUB1	14	4	0,29	69
GEMIN6	8	8	1,00	22	DDX20	24	15	0,63	290	CCNL2	5	2	0,40	25	EIF5A	7	2	0,29	947
IK	5	5	1,00	140	RPS10	5	3	0,60	54	CTDP1	15	5	0,40	164	SCNN1A	7	2	0,29	16
PEPD	1	1	1,00	0	COPB2	7	4	0,57	278	NFYA	20	8	0,40	92	SNIP1	28	8	0,29	391
RPL10A	1	1	1,00	0	IL2RA	7	4	0,57	69	NUDT3	5	2	0,40	15	SNUPN	7	2	0,29	100
RPL24	2	2	1,00	118	SIP1	16	9	0,56	30	PACS1	10	4	0,40	510	XAB2	7	2	0,29	13
RPL3	1	1	1,00	0	EIF5B	9	5	0,56	235	RANBP5	15	5	0,40	320	TCERG1	25	7	0,28	473
RPL30	2	2	1,00	0	EIF5A	11	6	0,55	366	SF3B1	10	4	0,40	40	CEBPA	29	8	0,28	57
RPS13	3	3	1,00	44	GRAP	11	6	0,55	10	IL15	23	9	0,39	866	POLR2A	58	16	0,28	1822
RPS15	1	1	1,00	0	STXBP2	13	7	0,54	4	KPNB1	41	15	0,39	1745	EPHA3	11	3	0,27	268
RPS20	3	3	1,00	156	SNRPB	28	15	0,54	241	WEE1	18	7	0,39	106	U2AF1	11	3	0,27	91
RPS7	4	4	1,00	33	HMGA1	21	11	0,52	169	RPL8	8	3	0,38	171	RPLP1	19	5	0,26	798
RPS8	1	1	1,00	0	ANTXR1	4	2	0,50	0	NXF1	22	8	0,36	396	UBA52	19	5	0,26	86
RPS9	1	1	1,00	0	C11orf49	6	3	0,50	63	SMN1	42	15	0,36	907	MDM2	51	13	0,25	1040
TALDO1	2	2	1,00	146	CDC2L1	14	7	0,50	361	NFE2L2	17	5	0,35	92	CRK	64	16	0,25	401
GEMIN7	9	8	0,89	108	DPH1	2	1	0,50	0	NUP153	17	5	0,35	141	NMT1	12	3	0,25	21
SNRPF	9	8	0,89	46	GABRD	2	1	0,50	0	E2F1	43	15	0,35	442	PDE5C	8	2	0,25	0
SNRPD3	18	15	0,83	148	KCNQ2	6	3	0,50	48	ARIH2	23	8	0,35	761	PELI3	4	1	0,25	0
STXBP3	11	9	0,82	232	KIF11	2	1	0,50	0	ARCN1	6	2	0,33	0	RPS3A	12	3	0,25	87
DHPS	5	4	0,80	835	NADK	4	2	0,50	324	DLST	15	5	0,33	245	TRIM21	4	1	0,25	0
GEMIN5	10	8	0,80	15	PRPF8	2	1	0,50	0	EIF3S1	12	4	0,33	228	USP11	4	1	0,25	0
SON	5	4	0,80	32	RPL23	2	1	0,50	0	KCNJ4	9	3	0,33	14	GRIN2A	25	6	0,24	233
DNASE1	4	3	0,75	24	RPL7A	4	2	0,50	178	NRXN2	6	2	0,33	32	RUNX1T1	25	6	0,24	156
PDCD4	4	3	0,75	248	RPS14	2	1	0,50	0	PHB	15	5	0,33	37	ZHX1	38	9	0,24	647
SNRPD2	20	15	0,75	890	RPS26	2	1	0,50	0	POLR2J	3	1	0,33	0	MAP3K14	34	8	0,24	1462
SNRPE	14	10	0,71	28	RPS4X	2	1	0,50	0	RPL10	3	1	0,33	0	COPS7A	13	3	0,23	80
WDR77	10	7	0,70	140	SNRPG	14	7	0,50	45	RPL7	3	1	0,33	0	IL2RB	26	6	0,23	167
EIF3S9	6	4	0,67	52	WIPI2	2	1	0,50	0	RPS6	6	2	0,33	140	KHDRBS1	39	9	0,23	366
RPL11	3	2	0,67	0	NUP98	13	6	0,46	207	TRPV6	6	2	0,33	1	NDRG1	61	14	0,23	2541
RPL5	15	10	0,67	2684	SF1	13	6	0,46	142	NUP62	23	7	0,30	123	PIN1	55	12	0,22	819
RPL6	6	4	0,67	585	STRAP	20	9	0,45	250	COPA	10	3	0,30	90	NCOA5	37	8	0,22	77

Table 4: Connectivity and Betweennes of proteins in the A549 Specific Network



NODE	GLOBAL CONNECTIVITY	LOCAL CONNECTIVITY	RATIO	BETWEENNESS	NODE	GLOBAL CONNECTIVITY	LOCAL CONNECTIVITY	RATIO	BETWEENNESS	NODE	GLOBAL CONNECTIVITY	LOCAL CONNECTIVITY	RATIO	BETWEENNESS
KCNJ2	14	3	0,21	14	FDGFRB	53	8	0,15	280	UBE2I	102	10	0,10	791
PEX5	14	3	0,21	521	SLPTSH	20	3	0,15	24	CDKN2A	41	4	0,10	219
PTMA	33	7	0,21	152	ABL1	107	16	0,15	885	PLSCR1	72	7	0,10	155
SGOL1	19	4	0,21	548	CSNK2A1	168	25	0,15	3113	NYST2	31	3	0,10	89
ATM	39	8	0,21	684	KIT	54	8	0,15	62	ZBTB16	83	8	0,10	130
PAK1	64	13	0,20	794	HSP90AA1	88	13	0,15	1907	CTNNB1	135	13	0,10	1089
ABCC2	5	1	0,20	0	REL1	34	5	0,15	58	POLR2C	21	2	0,10	258
ACTA2	15	3	0,20	40	CDC2	118	17	0,14	859	CASP3	129	12	0,09	706
CHEK2	30	6	0,20	17	TP53	265	38	0,14	5529	ACTA1	91	8	0,09	387
PML	55	11	0,20	164	ADAM10	7	1	0,14	0	NPP6	23	2	0,09	11
RPL9	5	1	0,20	0	MCM2	35	5	0,14	325	EWSR1	117	10	0,09	743
VWF	10	2	0,20	37	STAU1	28	4	0,14	652	COPS6	72	6	0,08	733
PLK1	46	9	0,20	569	WWP2	21	3	0,14	33	IGF2R	12	1	0,08	0
STAT5B	41	8	0,20	95	CSNK2A2	71	10	0,14	300	NSF	24	2	0,08	152
SP1	88	17	0,19	422	JUN	116	16	0,14	618	SRC	207	17	0,08	638
CENTA1	16	3	0,19	299	PFN2	29	4	0,14	117	YWHAE	124	10	0,08	689
BRCA1	105	19	0,18	937	CASK	37	5	0,14	151	FUBIN	26	2	0,08	36
PARP1	39	7	0,18	208	CREBBP	197	26	0,13	2003	EEF1G	67	5	0,07	766
NPHOSPH6	23	4	0,17	376	SHC1	117	15	0,13	390	GRB2	192	14	0,07	289
PSMA3	23	4	0,17	122	RB1	134	17	0,13	388	SMAD9	110	8	0,07	436
HTATIP	77	13	0,17	1185	IGSF1	8	1	0,13	0	PTN	83	6	0,07	736
ASGR1	6	1	0,17	0	MCM3	24	3	0,13	75	ITGB4BP	28	2	0,07	258
DDX24	24	4	0,17	285	UPF2	26	3	0,12	336	LSM8	28	2	0,07	9
EPB41	36	6	0,17	500	NFKB1	70	8	0,11	515	CALM1	113	8	0,07	284
HDAC2	66	11	0,17	412	PPARG	35	4	0,11	5	HAP1	57	4	0,07	392
NFKB2	30	5	0,17	190	EGFR	161	18	0,11	1515	UBQLN4	148	10	0,07	854
HDAC3	61	10	0,16	185	FUS	18	2	0,11	20	NUDT18	30	2	0,07	140
SIN3A	55	9	0,16	221	FYN	154	17	0,11	628	CD4	32	2	0,06	14
CEBPB	49	8	0,16	99	CHD3	73	8	0,11	443	ACTB	100	6	0,06	308
FGFR1	37	6	0,16	188	PRPF40A	64	7	0,11	174	NIF3L1	34	2	0,06	258
HSPA5	38	6	0,16	677	PDZK1	19	2	0,11	258	SETDB1	86	5	0,06	298
SOS1	45	7	0,16	19	SLC2A4	38	4	0,11	309	SMAD1	110	6	0,06	121
RGS2	26	4	0,15	93	PRKCA	172	18	0,10	2116	ATXN1	159	8	0,06	525
SSSCA1	13	2	0,15	13	SMAD2	165	17	0,10	1341	YWHAG	247	12	0,06	678
CASP1	33	5	0,15	78	ACVR1	60	6	0,10	319	ACTN4	27	1	0,04	0

Table 4: Connectivity and Betweennes of proteins in the A549 Specific Network (continuation)

SNRPD3, SNRPD2, SNRPF, SNRPE, SNRPG - spliceosome components GEMIN6, GEMIN7, GEMIN5 - GEM-associated proteins (GEM is a GTP-binding protein). But there are also some proteins in the list that are not associated with protein complexes: STXBP3, SON, SF3A3, DHPS, COPB2, WDR77, DDX20, SMN2, IK, GRAP, IL2RA, SIP1, HMGA1.

3. The third, and possibly, the best value to assign to a protein in the Specific Network for evaluating its significance by connectivity level would be the weighted connectivity (WC), which is the number of edges for the given protein in the network, weighted by its ratio to connectivity in the corresponding global network. Table 5, first column, represents proteins from the A549 Specific Network with the highest score (equal or more than 5) of weighted connectivity. We can see that proteins of the spliceosome complex, ribonucleoproteins (GEMIN), as well as SMN proteins are found to be pivotal in the given Specific Network. There are two proteins whose highest level of significance for the given Specific Network can be elucidated only by checking their weighted connectivity score: STXBP3 and SIP1 (marked in red in the Table).

4. Table 5, second column, shows the most important proteins according to their betweenness, and finally, the third column represents those genes which are the most important for the A549 Specific Network by both weighted connectivity and betweenness. Figure 6a shows the distribution of values of betweenness for proteins in the A549 Specific Network; the threshold for considerably high values corresponds to the inflection of the curve.

5. According to the evaluation of the importance of a protein by two parameters, B and WC, (Table 5, third column), RPL5 has an extremely high level of significance, definitely unexplainable by its main known function (ribosomal subunit protein).

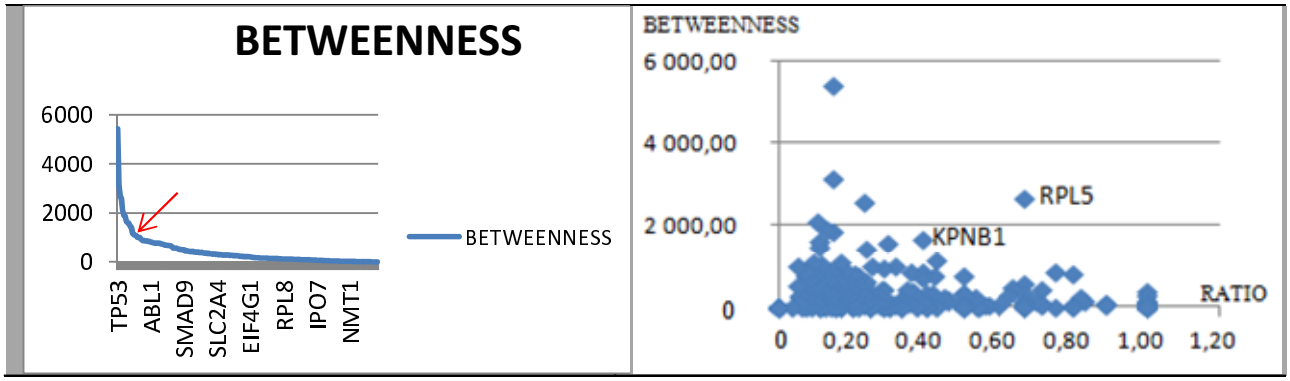


Figure 6: Betweenness. a) Values for genes in the A549 Specific Network. The inflection point of the curve corresponding to the threshold for considerably high values, is marked by arrow. b) Distribution of the values of betweenness together with ratio of connectivity in the A549 Specific Network.

Finally, we can see that the importance of two proteins in the A549 proliferation/viability Specific Network - RPL5 (apart from all other RPL and RPS proteins) and KPNB1 - was suggested by all three approaches considered in Table 5. Moreover, on the plot showing the distribution of values of betweenness together with the ratio of connectivity in the A549 Specific Network (Figure 6b), RPL5 protein appears to be the most significant one, whereas KPNB1 is the second in

NODE	**WEIGHTED CONNEC-TIVITY (WC)**		**NODE**	**BETWEEN NESS (B)**		**NODE**	**Both WC and B**
SNRPD3	12,5		TP53	5529		TP53	30126
SNRPD2	11,3		CSNK2A1	3113		**RPL5**	17894
DDX20	9,4		**RPL5**	2684		CSNK2A1	11580
SNRPB	8,0		NDRG1	2541		**KPNB1**	10897
GEMIN6	8,0		PRKCA	2116		**SNRPD2**	10013
STXBP3	7,4		CREBBP	2003		NDRG1	8164
SNRPE	7,1		HSP90AA1	1907		**POLR2A**	8041
GEMIN7	7,1		**POLR2A**	1822		CREBBP	6872
SNRPF	7,1		**KPNB1**	1745		DHX9	6430
RPL5	6,7		EGFR	1515		SMN1	4859
SMN2	6,7		MAP3K14	1462		PRKCA	3985
GEMIN5	6,4		SMAD2	1341		HSP90AA1	3663
KPNB1	6,2		DHX9	1250		MDM2	3447
HMGA1	5,8		HTATIP	1185		BRCA1	3223
TP53	5,4		CTNNB1	1089		**IL16**	3051
SMN1	5,4		MDM2	1040		EGFR	3049
E2F1	5,2					MAP3K14	2751
DHX9	5,1					**EIF3S6**	2735
SIP1	5,1					DDX20	2715
IK	5,0					**DHPS**	2673

Table 5: The most important hubs for the A549 Specific Network. Genes from the initial hit list are marked in bold

significance.

Additional information known about the RPL5 protein is that it specifically interacts with the beta subunit of casein kinase II [25]. Interestingly, it was shown that, in colorectal cancers, expression of the RPL5 gene in tumour tissue differs from that in adjacent normal tissues [24]. KPNB1 protein is a member of the importin beta family, involved in nuclear-cytoplasmic transport [23]. Thus, the significance of these two proteins for a lung cancer cell line is a completely unexpected result which may be of considerable importance.

In addition to these two proteins, our bioinformatics analysis suggests that CSNK2A1, SNRPD2, NDRG1, CREBBP, DHX9, and IL16 proteins may be the most important ones for A549 cell growth/viability. It is important to mention that IL16 was found to be very significant by the two-parameters test, although it was not identified by any of the single parameter tests.

2.3.2. Creating subnetworks specific to individual biological processes in the system

As a first step, simple mathematical modeling, was undertaken to elucidate the interplay between proliferation and viability processes in the phenotypical effect of knocking down each individual gene.

According to the method applied for reading the screening data, the effect of an siRNA-induced knockdown of mRNA expression for a given gene is quantified 3 days post siRNA treatment, by simultaneous measurement of (1) the final quantity of cells in the corresponding samples and (2) the release into the supernatant of a cytoplasmic enzyme by dying/dead cells. This final effect of the siRNAs on cell quantities could essentially be due to cell cycle arrest, cell death, or to various combinations of these two processes. Additionally, both of these processes could occur with different kinetics over the time-course of the experiment (for example, a late cell death combined with an early proliferation inhibition).

To elucidate the process in which the significant genes and pathways we found are mainly involved, namely to discriminate between their influence on cell proliferation and cell survival, a mathematical modeling, taking advantage of the availability of data about these two processes, was performed and next applied to the results of our proliferation and toxicity assays (see Material and Methods sections for experimental details).

As a result, we have obtained a list of genes with significant effects on cell survival and a list of genes with significant effects on proliferation, which allow us to retrieve the corresponding Proliferation and Cell survival subnetworks from the A549 Specific Network.

a) *Mathematical model*

For creating the simple mathematical model, we assume that the quantity of cells growing in the culture increases at a normal rate of proliferation in the normal (control) case, and increases at an unknown rate of proliferation and decreases by cell death in the experimental case.

Thus for live cells we can write:

$$\frac{da}{dt} = pa - \mu a \tag{2.3}$$

for dead cells:

$$\frac{db}{dt} = \mu a \tag{2.4}$$

where a is the amount of alive cells, b is the amount of dead cells, p is the coefficient of proliferation, and μ is the coefficient of cytotoxicity (cell death).

The solution for a will be:

$$a(t) = a_0 e^{(p-\mu)t} \tag{2.5}$$

Let a_c be the final amount of live cells in the control, a_0 is the initial amount of cells in the tube (well), then

$$a_c = a_0 e^{p_c t} = 8a_0 \tag{2.6}$$

where p_c is the rate of proliferation in control case, coefficient 8 came from the experimental data for the control case; we will consider $t = 1$, as the end of experiment is the first point of measurement.

Two biological measurements, namely, NPI (Normalized Percent of Inhibition) as the result of proliferation screen, and T (percent of dead cells) as the result of toxicity screen, were made at the end of the experiment for each tube, corresponding to each gene.

By definition:

$$NPI = 1 - \frac{a_f}{a_c}, T = \frac{b_f}{a_c} \tag{2.7}$$

where a_f is the final amount of live cells and b_f is the final amount of dead cells for the siRNA sample being considered.

From initial system of equations we have for T:

$$T = \frac{\frac{\mu}{p-\mu}(e^{p-\mu} - 1)}{e^{p_c}} \tag{2.8}$$

Next we get:

$$\frac{a_f}{a_c} = e^{p-p_c-\mu}, \tag{2.9}$$

$$NPI = 1 - e^{p-p_c-\mu}, \tag{2.10}$$

$$p - p_c - \mu = \ln(1 - NPI). \tag{2.11}$$

Now we would like to find all cases in corresponding to the condition $p_c - p > \mu$ which means that the decrease in proliferation rate due to knockdown of the given gene is greater than the decrease which occurs due to the toxicity effect of its knockdown:

$$p_c - p - \mu > 0 \tag{2.12}$$

From (2.11) we have $p - p_c = \ln(1 - NPI) + \mu$, thus, our condition for such cases will be:

$$\mu < -\frac{1}{2}\ln(1 - NPI)$$

Now from (2.6) and (2.8) we can derive an expression

$$\mu = \frac{T(2 + \ln(1 - NPI))}{\frac{7}{8} - NPI}$$

from which we have our final criteria to distinguish those cases in which the decrease in proliferation rate due to knockdown of a gene is greater than the decrease which occurs due to the toxicity effect of its knockdown, based on the two corresponding experimental values NPI and T:

$$T < -\frac{1}{2}\ln(1 - NPI)(\frac{7}{8} - NPI)/(2 + \ln(1 - NPI))$$

.

Next, this criteria was applied to the NPI and T results obtained for A549 human lung adenocarcinoma cell line genes. As the experiments for all of the individual siRNAs were done in triplicate, both NPI and T values were included in (8), together with their standard deviations.

As a result of the analysis, the following values were assigned to the corresponding pair of T/NPI data for each siRNA:

"1"- if the total decrease in final cell quantity is due rather to a decrease in proliferation rate than to cell death because of cytotoxicity,

"2"- if both effects are equal for the given siRNA treatment.

"3"- if the total decrease in final cell quantity is due rather to the influence of the toxicity effect than to proliferation-rate slowing or cell-cycle arrest.

Finally, after summarizing results obtained for the 4 siRNAs tested per gene, we got values allowing us to distribute genes of interest into the three groups described above. Table 6 illustrates this technique applied to the list of hit genes of the A549 screen.

Next, it is possible, by applying thresholds defined from the details of the experiment/assay, to discriminate specific groups of genes of interest. For example, calculating the thresholds from reciprocal results of the controls in both assays and taking into consideration their standard deviations, it was possible to determine the specific groups of genes whose knockdown effect on cells is "toxic only" and "arrest of proliferation only", and to identify specific cases such as "overproliferation" (Table 7). Overproliferation means that knockdown of a gene results in more active proliferation than in normal cells (NPI is less than zero). If needed, these genes can be easily included in the mathematical model as an additional group. But for the A549 screen, where this case was excluded from the selection of hits, it is still important to note that for the cases where a considerable cytotoxic effect of the siRNA is coupled with no or a very small effect on NPI, we also have a case of overproliferation - as defined above - which compensates the loss of cells due to toxicity. Elucidation of these specific groups, and especially of the group of genes the knockdown of which causes proliferation arrest without affecting cell death is extremely important for possible cancer therapies.

Cell death effect is stronger than the Inhibition of Proliferation	**EQUAL effect of Cell death and Inhibition of Proliferation**	**Inhibition of Proliferation is stronger than Cell death effect**		
ABCA2	ABAT	ADAM10	GRAP	RPS10
ABCF1	ABCC2	ADAMTS12	GUCA2A	RPS13
ACTN4	ABCG1	ANTXR1	HAS2	RPS14
AMPD3	ADCY5	AQR	HMGA1	RPS18
DKFZp434G0625	ALDH3B2	ARCN1	IGF2R	RPS20
DKFZp434P0216	C18orf25	ARHGEF4	KCNQ1DN	RPS21
DPH2L1	CA11	ASGR1	KCNQ2	RPS24
E2F1	CCR7	BSCL2	LCE1E	RPS25
GPX4	CEBPA	TMCO3	MGC27169	RPS3A
GRHPR	COPB2	C2orf28	MGC3196	RPS4X
HOM-TES-103	CTRC	CASK	MGRN1	RPS5
IGSF1	DHPS	CCT5	MGST3	RPS6
IK	DLX3	CDC2L1	MRPS24	RPS8
KIF11	DNASE1	CDC2L2	NFE2L2	RPS9
KPNB1	DSCAM	CDRT15	NMT1	SART1
MGC35521	FLJ20643	ch-TOG	NRM	SCNN1A
NXF1	FLJ46354	COPA	NUT	SELV
OR51G1	GPR133	COPB	OR2AT4	SF3A3
OR52H1	HIST1H2BA	CP	OR7A5	SF3B14
OR52N2	IL16	CPA5	PACS1	SMNDC1
PLK1	IL2RA	CRK	PB1	SMU1
POLA	KCNV1	DDX23	PDE6G	SNRPF
RRP22	KLF6	DKFZp564I1922	PEPD	SOAT2
SNRPD3	MIA	DLST	PES1	SON
TREH	MST1	EIF3S10	PEX16	SUPT6H
TRPV6	NRXN2	EIF3S4	PEX5	TALDO1
XAB2	P2RY12	EIF3S5	RBM25	TKT
	PCF11	EIF3S6	RPL10A	TRIM21
	POLR2A	EIF3S8	RPL11	UACA
	PRPF8	EIF3S9	RPL12	UBA52
	RPAP1	ELYS	RPL19	VprBP
	RPS7	ET	RPL21	VWF
	SC65	FAM50A	RPL24	WIPI-2
	SEC61A1	FAU	RPL3	XRCC2
	SF1	FLJ13052	RPL30	
	SF3B1	FLJ20280	RPL35	
	SIN3A	FLJ20516	RPL36	
	SIT	FLJ44076	RPL5	
	SLC16A13	GA17	RPL7	
	SLC22A14	GABRD	RPL7A	
	SNRPD2	GD:PTPRD	RPL8	
	SUV420H2	GJE1	RPL9	
	USP11	GPR132	RPLP1	

Table 6: Effect of gene inhibition on cell fate

Cell survival only	Proliferation only			Toxic with overproliferation
IK	RPL10A	EIF3S10	PEX5	DKFZp434G0625
DKFZp434P0216	RPL19	EIF3S6	ARCN1	HIST1H2BA
NXF1	RPL11	EIF3S4	ANTXR1	SC65
TRPV6	RPL21	EIF3S8	ASGR1	AMPD3
	RPL24	EIF3S9	BSCL2	
	RPL3	ET	C2orf28	
	RPL30	FLJ13052	CDC2L2	
	RPL35	GD:PTPRD	CP	
	RPL36	GJE1	SCNN1A	
	RPL5	KCNQ2	SELV	
	RPL7	MGC27169	SUPT6H	
	RPL7A	MGRN1	TALDO1	
	RPL8	NFE2L2	UBA52	
	RPL9	NMT1	VWF	
	RPLP1	NRM		
	RPS10	NUT		
	RPS3A	PACS1		
	RPS5	PB1		
	RPS9	PEX16		

Table 7: Special group of genes: genes controlling only one cell process (cell survival or cell proliferation) and genes having toxic and overproliferation effects.

Following this analysis, it is of particular interest to notice the striking difference between three housekeeping complexes, obviously important for cell growth, namely: elongation initiation factors, ribosomal subunits, and spliceosome components. According to the mathematical modelling, the knockdown of all elongation initiation factors and all ribosomal subunits has exclusively affects proliferation arrest, whereas knockdown of spliceosome components also has partially or completely toxic effects.

b) *Extraction of Proliferation and Cell Survival Subnetworks from A549 Specific Network*

Next, based on the results of our mathematical modeling, Proliferation and Cell Survival subnetworks were derived from the A549 Specific Network (Figure 7 and Figure 8). This allows us to identify some of the specific pathways important for proliferation or for cell survival in our system. For subnetwork extraction, all genes from the Specific Network were analyzed using our mathematical model, and the specific processes affected by each of them were determined. Then, for the Proliferation Subnetwork, all genes with proliferation arrest and equal effect were extracted, and, correspondingly, for the Cell Survival Subnetwork, all genes with cell death and equal effects were extracted.

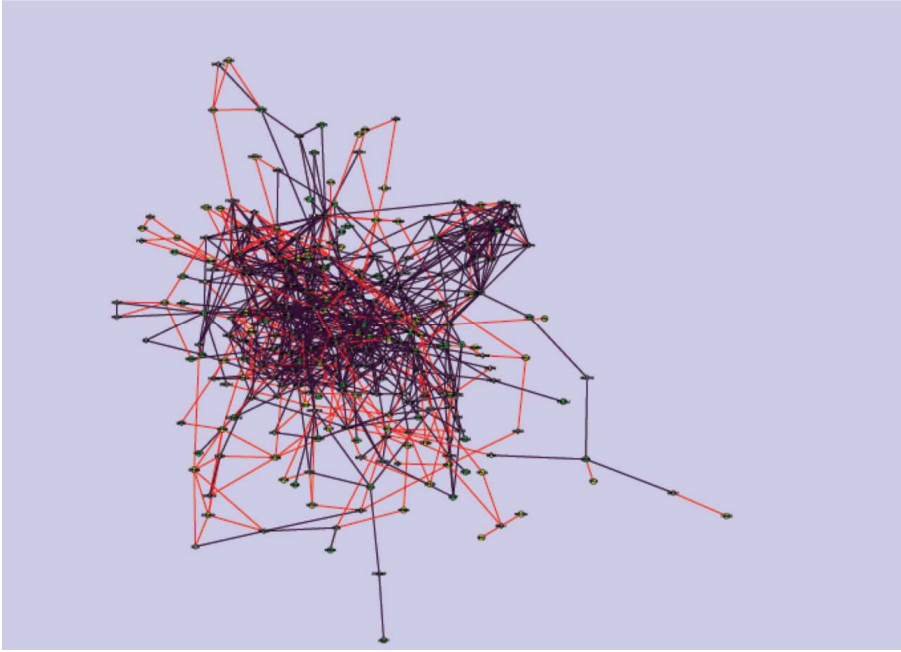


Figure 7: Network specific for A549, with subnetwork of genes the knockdown of which mainly causes *proliferation arrest* shown in red (color figure in the electronic version).

Figure 9 presents some examples of pathways and modules which can be highlighted in the Proliferation subnetwork after analysis in Cytoscape Software. Figure 9a (Prolif. 1) is interesting because of pinpointing the importance of COPS proteins, the functions of which are not yet very well characterized, in connecting the elongation initiation complex with FAU protein on the one hand side, and with the NFE2L2 - HMGA1 pathway on the other. Figure 9b (Prolif. 2) suggests the possibility of associating a pathway containing DLST, TALDO1 and KCNQ2 proteins with A549

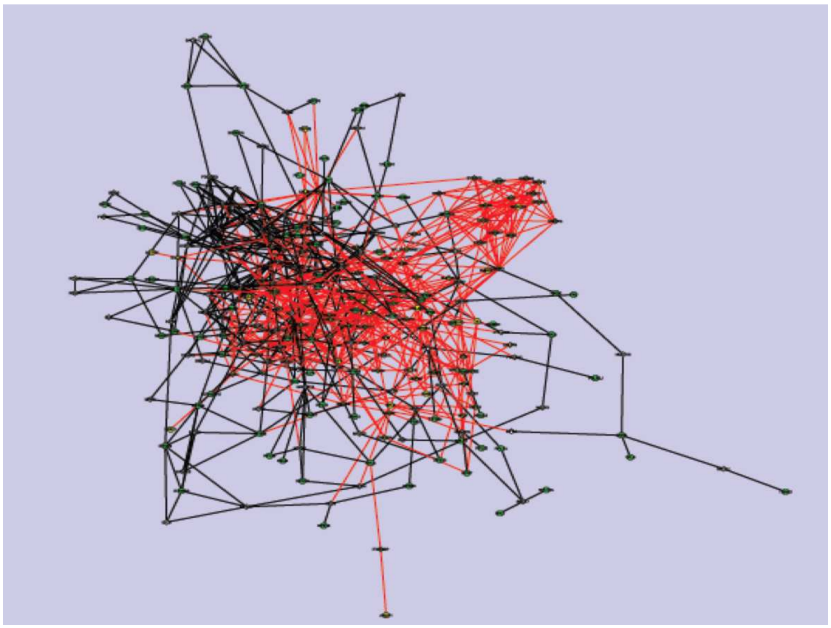


Figure 8: Network specific for A549, with subnetwork of genes the knockdown of which causes mostly *cell death* shown in red (color figure in the electronic version).

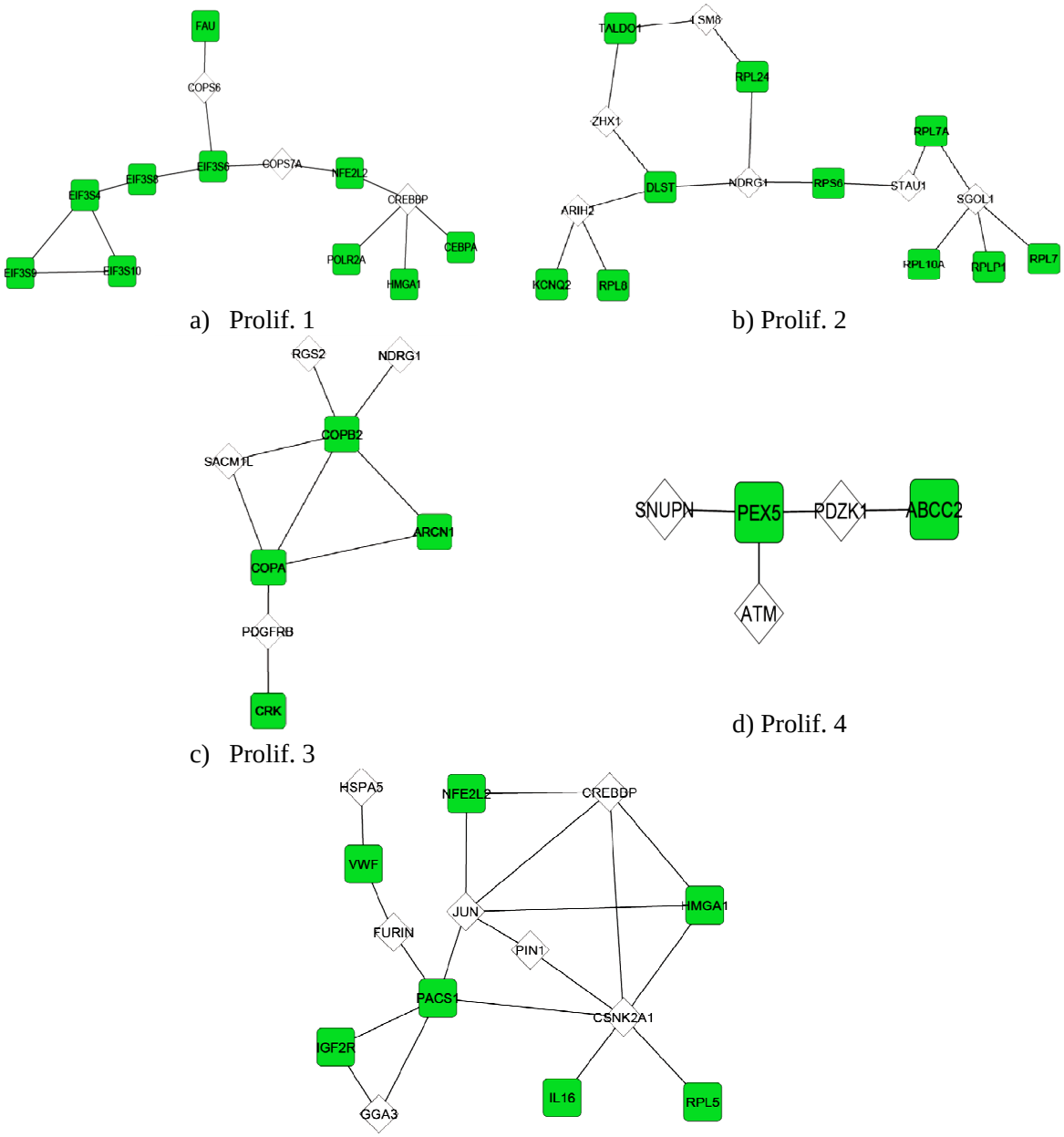
proliferation activity through its connection with RPL protein complexes. Figure 9c (Prolif. 3) emphasizes the importance of connection between the coatomer protein complex and CRK protein in this system. Figure 9d. (Prolif. 4) may suggest an additional previously unknown small pathway influencing A549 proliferation activity. Finally, Figure 9e (Prolif. 5) shows, in the module responsible for proliferation of A549 cells, detailed connections between proteins previously associated with cancer progression. Figure 10 presents some examples of pathways and modules which can be highlighted in the Cell Survival subnetwork.

Figure 10a (Cell Survival module 1) shows that the suppression of members of the spliceosome complex is much more crucial for A549 cell survival, than it is for their proliferation activity. Cell Survival module 2 (Figure 10b), which includes hit genes ADCY5, TRPV6, DLX3(B) and three small putative cell-survival pathways (Figure 10c) presents an opportunity to discover new pathways important for A549 viability and survival.

Figure 11 helps to elucidate the possible role of RPL5 in the A549 Specific Network, showing that it is an important hub in this system, located on the intersection of proliferation pathways (RPL and RPS proteins), cell survival pathways (SMNDC1, KPNB1 proteins), and, via casein kinase II, pathways involved in the regulation of cancerogenesis.

2.4. Discussion

In this work, we suggest several analytical tools for systems biology analysis of the results of genome-wide loss-of-function screens. First of all, we introduce the definition of Specific Network for the given system, and propose the rules for its construction based on the list of hit genes



a) Prolif. 1

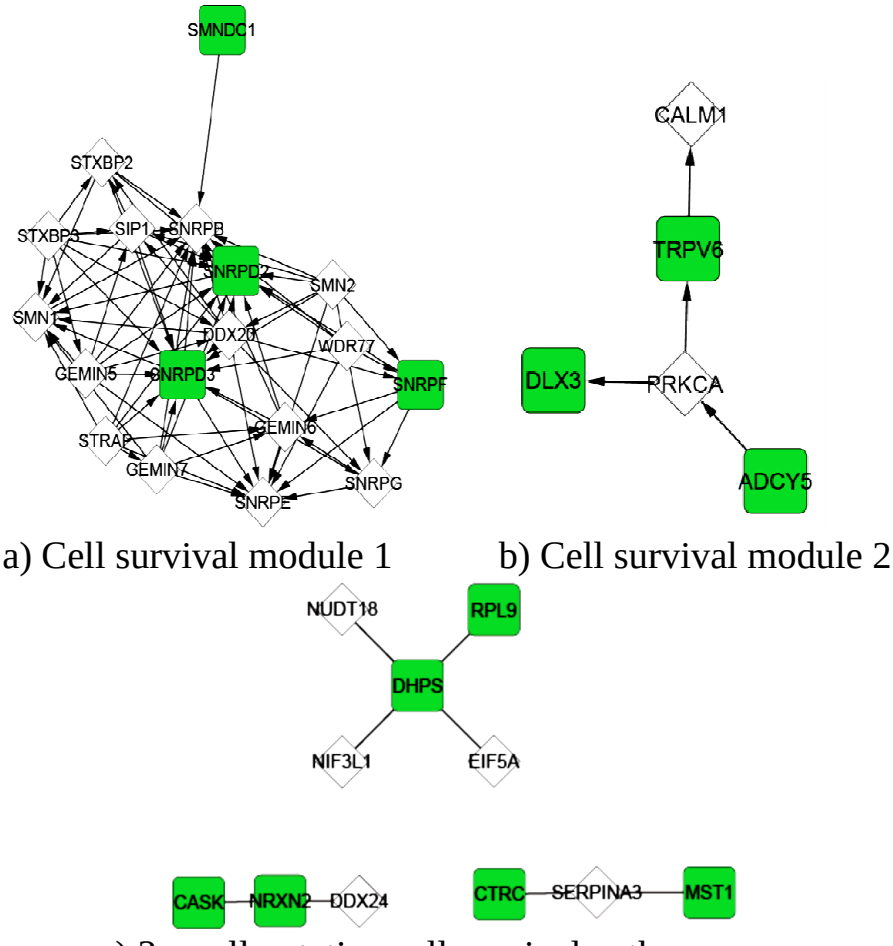
b) Prolif. 2

c) Prolif. 3

d) Prolif. 4

e) Prolif. 5

Figure 9: Pathways and modules from Proliferation subnetwork.



a) Cell survival module 1 b) Cell survival module 2

c) 3 small putative cell survival pathways

Figure 10: Pathways and modules from Cell survival subnetwork.

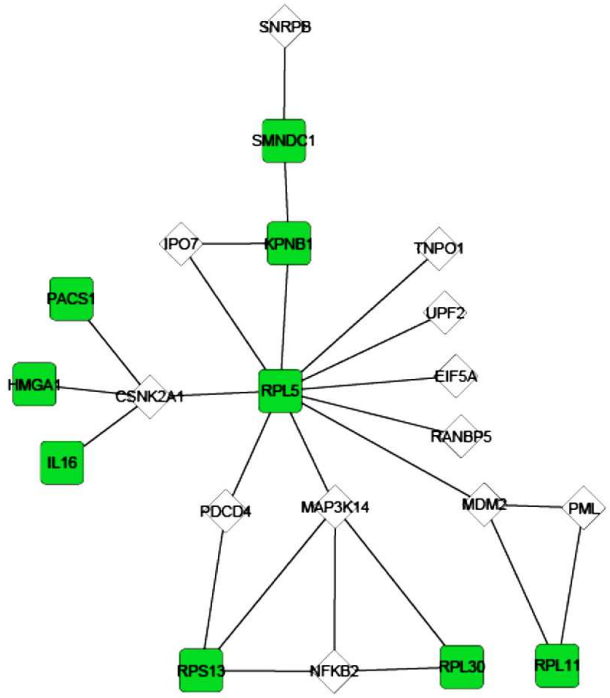


Figure 11: RPL5 is a hub on the intersection of prolifeation (RPL and RPS proteins) and cell survival pathways (SMNDC1, KPNB1 proteins).

identified in the course of a genome-wide LOF screen. Second, we worked out two different strategies for checking the specificity of the created network, which allows us to state that the initial hit list and Specific Networks driven from our analysis are relevant to the investigated biological system. Including information on protein complexes in the evaluation of network specificity is a crucial element for a genome-wide LOF screen network analysis, because the presence of housekeeping complexes, such as ribosomal subunit complexes, spliceosome complexes, etc., must not bias the network specificity results. Next, we present our mathematical model allowing the discrimination between two processes that may result simultaneously from the knockdown of a given gene. Namely, in the example discussed above, where the readout of the screen quantifies the final amount of cells after gene knockdown, the mathematical model we suggest permits discrimination between its effect on cell proliferation versus effect on cell survival processes. At the network level, this discrimination also allows the creation of corresponding subnetworks from the Specific Network, which in turn, opens avenues for elucidating pathways important for each of these processes in the system.

All the systems biology methods reported in this paper were illustrated by application to the results of a genome-wide LOF screen of A549 human lung adenocarcinoma cell proliferation/viability. The analysis of the A549 proliferation/viability Specific Network revealed 29 most important genes (hubs) and several important pathways and modules relevant to human lung adenocarcinoma proliferation activity and/or cell viability in this system. Understanding the involvement of these genes and pathways in cell proliferation or in cell survival/maintenance could be

particularly useful to orientate follow-up research on the given biological system, based on the importance of the identified subnetworks and their derived pathways. Genes from pathways found to be preferentially involved in cancer cell survival would be expected to be more relevant targets for further cancer therapy research than genes involved in cancer cell proliferation processes. Eventually, that hypothesis should be tested both from bioinformatics (by comparison of the relative importance of proliferation and survival subnetworks obtained from LOF screens in cancer cells and their non-tumoral counterparts) and experimental points of view.

Some of the results of our analysis of the A549 Specific Network are summarized below:

- some genes known to be involved in cancer progression, but not previously identified as being the most significant ones in lung adenocarcinoma, showed up as being highly important for this system, because they had the highest levels of connectivity and centrality (betweenness) in the A549 proliferation/viability Specific Network (BRCA1, CREBBP, SRC, MDM2, CRK);

- some new genes previously not linked to cancer progression, were found to be extremely important for A549 cell growth: RPL5, KPNB1, SNRPD2, DHPS, CSNK2A1,SNRPD2, NDRG1, CREBBP, DHX9 , IL16.

- several new pathways important for A549 cell proliferation, and several new pathways important for A549 cell survival were also suggested.

Statistically, the most pivotal genes/proteins obtained in the analysis are RPL5 and KPNB1 proteins. The importance of RPL5, apart from all other RPL and RPS proteins, was suggested by all the approaches we used in network analysis. Definitely, the extremely high rank of RPL5 significance for the A549 network cannot be explained by its main known function (ribosomal subunit protein) in the translational process, suggesting possibile additional specific functions of this protein in lung adenocarcinoma cells. It was shown that, in colorectal cancers, expression of RPL5 in tumours differs from that seen in adjacent normal tissues [24], and that RPL5 specifically interacts with the beta subunit of casein kinase II [25]. One possible explanation of RPL5's importance may be obtained from pathway analysis, showing that RPL5 is a hub on the intersection of proliferation (RPL and RPS proteins) and cell survival pathways (SMNDC1, KPNB1 proteins).

KPNB1 protein is a member of the importin beta family, involved in nuclear-cytoplasmic transport [23]. Mathematical modeling shows that its knockdown results mainly in cell death rather than in proliferation arrest, thus elucidating the importance of this protein for maintaining A549 viability. The significance of these two proteins for lung cancer cells was not previously shown, and thus may be of considerable importance.

Analysis of proliferation and cell-survival subnetworks highlights the striking difference in this biological system between three housekeeping complexes that are obviously important for cell growth, namely: elongation initiation factors, ribosomal subunits, and spliceosome components. Whereas elongation initiation and ribosomal subunits complexes belong to a proliferation subnetwork, the spliceosome complex belongs to a cell death subnetwork, indicating its high importance for cell survival.

Another interesting fact is that the analysis of the A549 Specific Network pinpointed the spliceosome component SNRPD2 as one the most important proteins in this system, though no information about its possible connection with cancer progression currently exists. Interestingly, there is some cancer-related data in the HPRD database for other members of SNRPD family,

especially SNRPD3, including one in lung cancer (J. of Proteome research).

In conclusion, this study outlines a systems biology strategy for the identification of genes and pathways that are important for a given biological system. The approach we suggest opens up new perspectives for the most important goal of such genome-wide techniques - i.e., suggestions for researching possible novel therapeutic targets - because it helps to determine the most important genes in the given biological system, which may be different from the most active genes in the list of hits identified by a genome-wide LOF screen. Additionally, this approach helps to bring to light other important genes of the system, which may potentially be found in the analysis of the corresponding Specific Network even if they were missed during the LOF screen for technical or procedural reasons. Consequently, we suggest that further biological analyses should be performed on those genes identified in our study as being very important ones, but that until now have not been functionally associated with oncogenesis and/or with development of human lung adenocarcinoma.

2.5. Materials and Methods

The A549 human lung adenocarcinoma cell line was provided by G. Kroemer, INSERM U848, Institut Gustave Roussy, Villejuif, France. The A549 cell proliferation and viability genome-wide screen was conducted in 96-well tissue culture plates, using a library assembled with individual synthetic siRNAs of predicted sequences targeting the whole human genome (22,950 targeted genes, 2 different siRNAs per gene, Qiagen). With the aid of automation (8-channel Star, Hamilton), siRNAs targeting each human gene and the control siRNAs were individually transfected in triplicate into A549 cells using Hiperfect (Qiagen); 72 hours after transfection, a toxicity assay and a proliferation/viability assay were performed in parallel on separate aliquots from each well. For the toxicity assay, an aliquot of supernatant from each well was used for quantification of lactate dehydrogenase (LDH), which is released by dead or dying cells (LDH Cytotoxicity Assay, Roche). Next, the remainder of the material from the well was tested for cell proliferation/viability using the WST-1 assay (Roche).

Candidate gene selection: Raw data were rescaled according to control-based (Normalized Percentage of Inhibition, NPI) and sample-based normalizations (robust Z-Score, B-score), and, when needed, corrected for positional effects using the sample well correction method (Kevorkov et al. 2005). A siRNA was defined as being active when its normalized value deviated by more than 3*MAD from the sample siRNAs median in at least one of the normalization procedures. The candidate list was constructed with genes for which at least one siRNA was defined as active.

Secondary screen: The candidate genes were then reassayed for proliferation and cytotoxicity in a confirmatory screen identical to the Primary screen but with 4 siRNAs tested per gene (the 2 that were previously tested + 2 additional siRNAs). To select the hit genes, the balanced activity of each set of 4 siRNAs/gene on cell proliferation was computed as a single score. An arbitrary threshold, corresponding to the score of 4 siRNAs with a theoritical activity of 20% inhibition of A549 cell growth, was set as hit selection threshold.

For the selection of Candidate and Hit Genes, mostly influencing A549 cell growth, biostatistics analysis of screen datasets was performed under the R environment with Bioconductor packages.

Acknowledgements

This work was supported by the APO-SYS EU FP7 project, the Institut National du Cancer (INCa) and the Agence National de la Recherche (ANR). AZ is a member of the Systems Biology of Cancer team, certified by the Ligue Nationale Contre le Cancer. The study was also funded by the Projet Incitatif Collaboratif "Bioinformatics and Biostatistics of Cancer" at the Institut Curie. We are grateful to Dr. Guido Kroemer (Institut Gustave Roussy, France) for providing the A549 human lung adenocarcinoma cell line, and to Dr. Linda L. Pritchard for critical reading of the manuscript.

References

[1] C.J. Creighton, J.L. Bromberg-White, D.E. Misek, D.J. Monsma, F. Brichory, R. Kuick, T.J. Giordano, W. Gao, G.S. Omenn, C.P. Webb, S.M. Hanash. *Analysis of tumor-host interactions by gene expression profiling of lung adenocarcinoma xenografts identifies genes involved in tumor formation*. Mol Cancer Res., 3 (2005), No. 3, 119–29.

[2] Y. Murakami. *Functional cloning of a tumor suppressor gene, TSLC1, in human non-small cell lung cancer*. Oncogene, 7 (2002), No. 21(45), 6936–48.

[3] Y. Jiang, L. Cui, T.A. Yie, W.N. Rom, H. Cheng, K.M. Tchou-Wong. *I*nhibition of anchorage-independent growth and lung metastasis of A549 lung carcinoma cells by IkappaBbeta. Oncogene, 26 (2001), No. 20(18), 2254–63.

[4] M. Soda, et al. *Identification of the transforming EML4-ALK fusion gene in non-small cell lung cancer*. Nature, 448 (2007), 561–566.

[5] R. Kittler , L. Pelletier, A.K. Heninger, M. Slabicki, M. Theis, L. Miroslaw, I. Poser, S. Lawo, H. Grabner, K. Kozak, J. Wagner, V. Surendranath, C. Richter, W. Bowen, A.L. Jackson, B. Habermann, A.A. Hyman, F. Buchholz *Genome-scale RNAi profiling of cell division in human tissue culture cells.* Nat Cell Biol., 9 (2007), No. 12, 1401–12.

[6] M.A. Pujana, J.D. Han, L.M. Starita, K.N. Stevens, M. Tewari, J.S. Ahn, G. Rennert, V. Moreno, T. Kirchhoff, B. Gold, V. Assmann, W.M. Elshamy , J.F. Rual, D. Levine, L.S. Rozek, R.S. Gelman, K.C. Gunsalus, R.A. Greenberg, B. Sobhian, N. Bertin, K. Venkatesan, N. Ayivi-Guedehoussou, X. Sole, P. Hernindez, C. Lazaro, K.L. Nathanson, B.L. Weber, M.E. Cusick, D.E. Hill, K. Offit, D.M. Livingston, S.B. Gruber, J.D. Parvin, M. Vidal. *Network modeling links breast cancer susceptibility and centrosome dysfunction.* Nat Genet., 39 (2007), No. 11, 1338–49.

[7] M. Vidal *A biological atlas of functional maps.* Cell, 9 (2001), No. 104(3), 333–339.

[8] E. Segal, N. Friedman, D. Koller, A. Regev *A module map showing conditional activity of expression modules in cancer.* Nat Genet., 36 (2004), No. 10, 1090–1098.

[9] A. Beyer, S. Bandyopadhyay, T. Ideker. *Integrating physical and genetic maps: from genomes to interaction networks.* Nat Rev Genet., 8 (2007), No.9, 699–710.

[10] T. Haberichter, B. Mädge, R.A. Christopher, N. Yoshioka, A. Dhiman, R. Miller, R. Gendelman, S.V. Aksenov, I.G. Khalil, S.F. Dowdy. *A systems biology dynamical model of mammalian G1 cell cycle progression.* Mol Syst Biol., 3 (2007), No. 84.

[11] O. Sahin, C. Löbke, U. Korf, H. Appelhans , Sültmann H, Poustka A, Wiemann S, Arlt D. *Combinatorial RNAi for quantitative protein network analysis.* Proc Natl Acad Sci U S A., 17 (2007) No. 104(16), 6579–84.

[12] A.3rd Bankhead, I. Sach, C. Ni, N. LeMeur, M. Kruger, M. Ferrer, R. Gentleman, C. Rohl *Knowledge based identification of essential signaling from genome-scale siRNA experiments.* BMC Syst Biol., 5 (2009), No. 3:80.

[13] B. Lehner, C. Crombie, J. Tischler, A. Fortunato, A.G. Fraser. *Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways.* Nat Genet., 38 (2006), No. 8, 896–903.

[14] M. Mukherji, R. Bell, L. Supekova, Y. Wang, A.P. Orth, S. Batalov, L. Miraglia, D. Huesken, J. Lange, C. Martin, S. Sahasrabudhe, M. Reinhardt, F. Natt, J. Hall, C. Mickanin, M. Labow, S.K. Chanda, C.Y. Cho, P.G. Schultz. *Genome-wide functional analysis of human cell-cycle regulators.* PNAS 103 (2006), No. 40, 14819–14824.

[15] M.H. Beers *Lung Carcinoma*. In *The Merck manual of diagnosis and therapy* (R.S. Porter, and T.V. Jones, editors). Rahway: Merck & Co., Inc. (2008), 2992.

[16] A. Jemal, et al. *Annual report to the nation on the status of cancer, 1975-2005, featuring trends in lung cancer, tobacco use, and tobacco control.* J Natl Cancer Inst, 100 (2008), No. 23, 1672–1694.

[17] R.K. Kancha, N. von Bubnoff, C. Peschel, J. Duyster *Functional analysis of epidermal growth factor receptor (EGFR) mutations and potential implications for EGFR targeted therapy.* Clin Cancer Res., 15 (2009), No. 2, 460–467.

[18] C.T. Miller, G. Chen, T.G. Gharib, H. Wang, D.G. Thomas, D.E. Misek, T.J. Giordano, J. Yee, M.B. Orringer, S.M. Hanash, D.G. Beer. *Increased C-CRK proto-oncogene expression is associated with an aggressive phenotype in lung adenocarcinomas.* Oncogene 22 (2003), No. 39, 7950–7957.

[19] A. Zinovyev, E. Viara, L. Calzone, E. Barillot. *BiNoM: a Cytoscape plugin for manipulating and analyzing biological networks.* Bioinformatics, 24 (2008), No.6, 876–877.

[20] H. Shigematsu, A.F. Gazdar *Somatic mutations of epidermal growth factor receptor signaling pathway in lung cancers.* Int J Cancer, 118 (2006), No. 2, 257–262.

[21] C. Mascaux, N. Iannino, B. Martin, M. Paesmans, T. Berghmans, M. Dusart, A. Haller, P. Lothaire, A.P. Meert, S. Noel, J.J. Lafitte, J.P. Sculier. *The role of RAS oncogene in survival of patients with lung cancer: a systematic review of the literature with meta-analysis.* Br J Cancer., 92 (2005), No. 1, 131–139.

[22] M. Smoot, K. Ono, J. Ruscheinski, P.-L. Wang, T. Ideker. *Cytoscape 2.8: new features for data integration and network visualization.* Bioinformatics, 27 (2011), No. 3, 431–432.

[23] Y.M. Chook, G. Blobel. *Karyopherins and nuclear import.* Curr Opin Struct Biol., 11 (2001), No. 6, 703–715.

[24] B. Lü, J. Xu, Y. Zhu, H. Zhang, M. Lai. *Systemic analysis of the differential gene expression profile in a colonic adenoma-normal SSH library.* Clin Chim Acta., 378 (2007), No.1-2, 42–47.

[25] J.W. Park, Y.S. Bae. *Phosphorylation of ribosomal protein L5 by protein kinase CKII decreases its 5S rRNA binding activity.* Biochem Biophys Res Commun. 263 (1999), No. 2, 475–481.