A biocomputational platform for the automated construction of large-scale mathematical models of miRNA-transcription factor networks for studies on gene dosage compensation

Acón Man-Sai
Master Program on Bioinformatics and Systems Biology
University of Costa Rica (UCR)
San José, Costa Rica
m.acon@tecapro.com

Siles-Canales Francisco
PRIS Lab, Faculty of Engineering, University of Costa Rica (UCR)
San José, Costa Rica
francisco.siles@ucr.ac.cr

Mora-Rodriguez RA
LabQT, Research Center on Tropical Diseases (CIET)
Faculty of Microbiology, University of Costa Rica (UCR)
San José, Costa Rica
rodrigo.morarodriguez@ucr.ac.cr

Abstract— Cancer complexity and resistance is mediated by cell-to-cell heterogeneity, which is the consequence of the enormous instability of its genetic material. It is unknown how cancer cells are able to withstand the effects of these alterations, while normal cells are typically very sensitive. We hypothesize that cancer requires specific type of stability to survive the enormous chromosomal alterations. This stability may be mediated by a group of genes, whose expression is tightly regulated to maintain viability through a process called gene dosage compensation. This mechanism could be mediated by systems-level properties of complex networks of microRNAs (miRNA) and transcription factors (TF), regulating gene expression despite changes in copy number. Therefore, we designed a biocomputational platform to automatically construct large-scale mathematical models regulating the expression of several candidate genes under dosage compensation. This platform has a broader potential application to other scientific questions involving miRNA and TF networks.

Keywords—miRNAs, gene dosage compensation, cancer, systems biology

I. Introduction

Cancer robustness is enabled at the tumor cell population level by heterogeneity in therapy responses, which is driven by genomic instability[1] , specially by aneuploidy: gains and losses of whole or partial chromosomes. It is unknown how cancer cells deal with so much aneuploidy whereas normal cells are very sensitive. A possible explanation is given by the hypothesis of gene dosage compensation, a mechanism that has been described for other organisms to compensate the negative effects of aneuploidy [2]. It has been shown for aneuploid cancers that messenger RNA (mRNA) levels generally correlate well with an increased DNA copy number (gene dosage) but these changes are not reflected at the protein levels for several genes [3]. Several lines of evidence suggest the existence of a gene dosage compensation mechanism that provides stability to cancer despite its genomic instability. However, this mechanism must be able to regulate the expression of a handful of critical genes simultaneously. We hypothesize the existence of a complex regulatory network mediated by microRNAs (miRNAs) to compensate for gene dosage changes in aneuploid cancer cells. miRNAs are small endogenous RNA molecules that bind mRNAs and repress gene expression [4]. Currently, 1500 miRNAs have been described within the human genome [5] regulating the expression of nearly 30% of all genes [6]. Additionally, miRNAs can regulate hundreds of genes and their target genes can be regulated by several miRNAs [7]. Furthermore, they form regulatory interactions with transcription factors including feedback and feedforward loops leading to non-linear, systems-level properties such as bistability, ultrasensitivity and oscillations [8]. We suggest that the manipulation of specific nodes of this miRNA-based regulatory network could block gene dosage compensation, representing a specific target against cancer. Due to the complexity of this network, identification of optimal targets requires an advanced computational platform.

II. Materials and Methods

A. *Data sources and biocomputational platform*

For the biocomputational platform we gathered data from several sources. The primary sources were from experiments on the NCI60 panel: gene copy number [17], RNA gene expression [18] and protein expression [11]. MicroRNA related data was downloaded from Mirtarbase [19] and MiRBase [20]. For gene regulation data we relied on several sources: Transmir [21], Pazar [22], TRED (Transcriptional Regulatory Element Database) [23], CircuitsDB [24].

B. *Gene classification using Gaussian Mixture Models*

In order to classify genes according to their behavior, we developed a computer algorithm based on the Gaussian Mixture Model functions in MATLAB. An increasing number of components (ki) of the GMM model is added sequentially and the GMM training is performed for several iterations searching for the best fit to the experimental data. The resulting GMM is used to classify the cells of the original data set. A MANOVA test is applied to the resulting clusters to evaluate statistically the biological significance of adding another component to the GMM until the new component adds no further significance, finding thereby the optimal number of components with biological meaning.

C. *Ordinary differential equation modeling of miRNA-TF networks*

From the list of genes selected by the GMM we identified the miRNAs an TFs regulating each gene. We build the network topology using these three types of species as nodes and the regulation relations as edges. In the SBML model we defined each node as specie. For each species we defined two parameters: synthesis and degradation rates. For miRNAs we also define the repression rate and for TFs we defined activation and repression rates. Finally, we added synthesis and degradation reactions for each species in the SBML. We also created experimental data files using CGH, RNA and microRNA expression. We imported the model into COPASI. Once in COPASI we included the experimental data files, fitted the model with the Parameter Estimation Task.

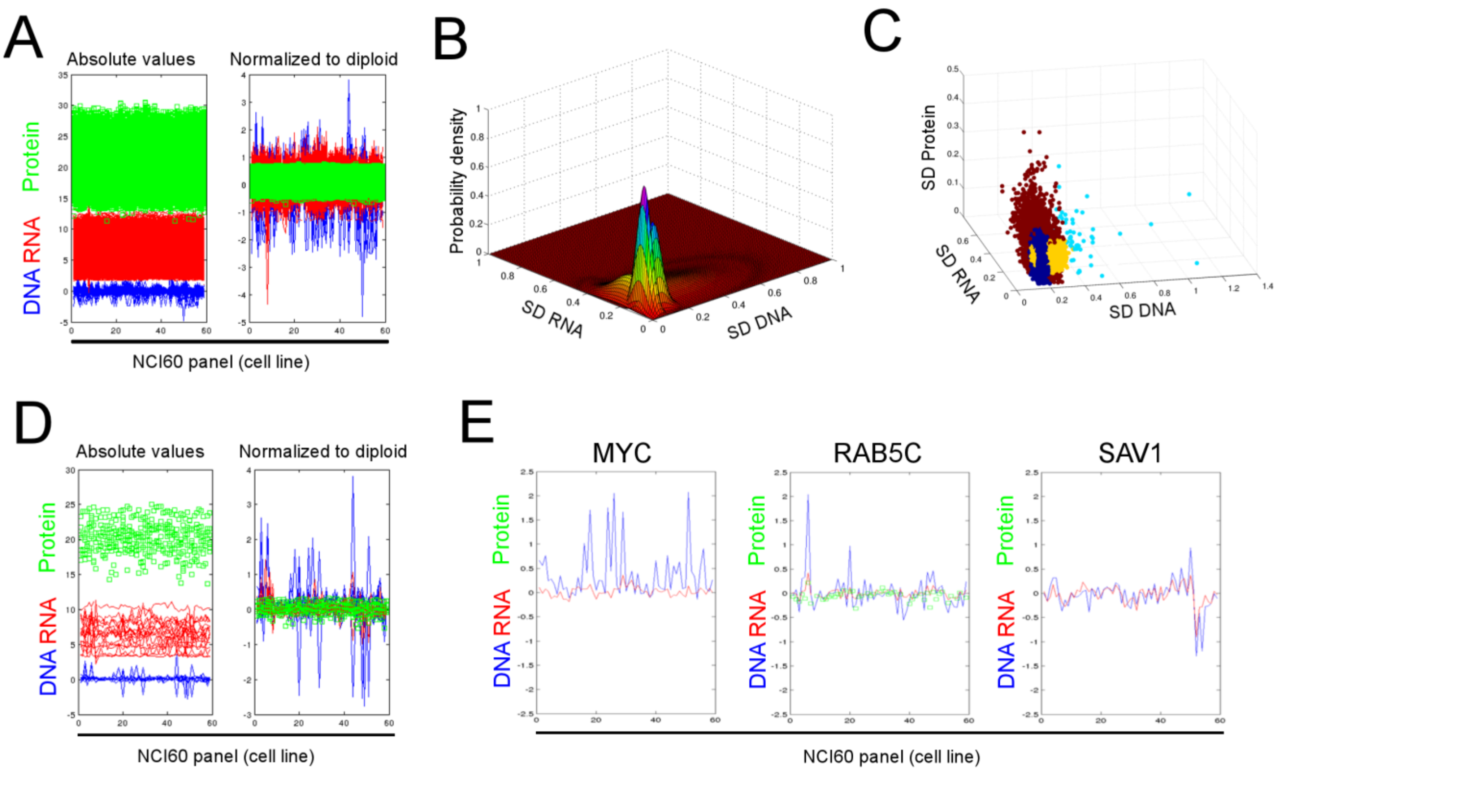

Figure 1. Identification of candidate genes under dosage compensation. A. Input data of gene copy number (DNA), gene expression (RNA) and protein levels (protein) of the NCI60 panel. The absolute values are shown on the left panel and the right panel corresponds to the log2 values normalized to the averaged RNA and protein of the diploid cell lines for the respective gene (Normalized to diploid). B. Gaussian Mixture Model to identify a cluster of subpopulation of genes with high SD DNA and low SD RNA and/or low SD Protein (white arrow). C. Gene Clustering according to the model in B showing the standard deviations (SD) of the DNA, RNA and protein levels for each gene across the 59 cell lines of the NCI60 panel. The cyan cluster contains candidate genes under dosage compensation, characterized by high SD DNA, low SD Protein (middle) and low SD RNA (right). D. Absolute and Normalized values of selected candidate genes under dosage compensation. E. Examples of candidate genes under dosage compensation (MYC and RAB5C) compared to a non-candidate gene (SAV1).

III. Results

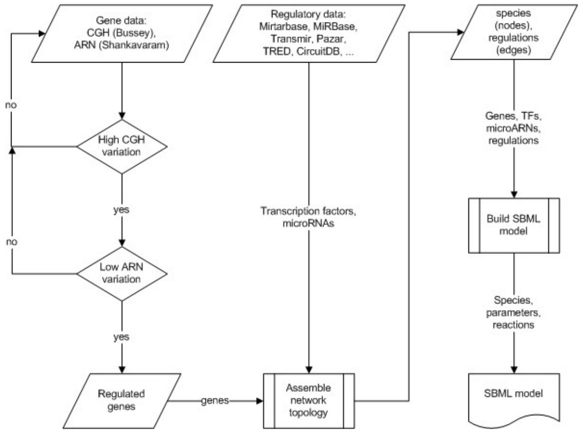

Figure 2. The workflow starts by analizing CGH and ARN variation of the genes. Those genes with high CGH variation and low ARN variation are selected. These genes, along with regulation relations between transcription factors/microARNs and genes as well as to each other are assembled together to conform the topology of the gene dosage compensation network. Finally, the SBML model of the network is constructed using the list of nodes and edges of the network, converting each node into a species and using the edges for the reactions of synthesis and degradation in the SBML model.

A. *Candidate genes under gene dosage compensation are present across the NCI-60 panel*

In order to identify possible genes under gene dosage compensation, we compared copy number, gene expression and proteomic data of the NCI60 panel. We considered input data including high resolution Copy Number Variation data (DNA) of the NCI-60 Cancer Cell lines from 4 different platforms [9], the Gene Transcript (RNA) Average Intensities of 5 Platforms [10], and the protein levels (Protein) of a global proteome analysis of the NCI-60 cell line panel [11]. Figure 1A left shows the relative variation of these absolute DNA, RNA and Protein levels. However, once we normalized the data in the same way (log2 values normalized to the averaged RNA and protein of the diploid cell lines for the respective gene), it can be observed in Figure 1A right that the DNA values have a higher amplitude, followed by RNA levels and protein levels.

Indeed, we are interested in those genes with high variation in DNA levels and low variation in RNA or protein levels, as candidates to study possible mechanisms of gene dosage compensation. Therefore, we considered the Standard Deviation (SD) of the DNA, RNA and Protein values across the 59 cell lines and classify those genes using a Gaussian mixture model (GMM) from the data (Figure 1B). One cluster was identified having high SD of DNA values and proportionally low SD of RNA and SD of Protein values (Figure 1C). The genes contained within this cluster presented the behaviour of interest (Figure 1D), where the corresponding SD of RNA and/or Protein is proportionally low compared to the high SD of DNA values. Moreover, we discarded those genes with orthologues in X/Y chromosomes since they cannot be differentiated using with microarray techniques. In addition, we discarded genes without any reported interactions with microRNA or Transcription Factor forming regulatory loops (see below). Furthermore, we discarded genes with high DNA variation due mostly to deletion (DNA values lower than -0.25) and obtained a list of 19 gene candidates under dosage compensation including ANKFY1, ATP1B2. PGR, DCUN1D5, MMP12, BIRC2, ATM, NPAT, CUL5, STAT3, KCNH4, RAB5C, TRIM37, ZNF217 and MTSS1, MYC, SEMAD3D, BIRC2, ZNF217, FOXC1 y PDCD10. Among those, we can highlight the presence of the oncogene MYC, which presents high frequency of amplifications in the NCI60 panel without the corresponding increase in RNA levels (Figure 1E right). For many genes there is also protein data, which confirms this behaviour as is the case for RAB5C, which RNA and protein levels are maintained quite constant despite DNA variations (Figure 1E middle), compared to a non-candidate gene such as SAV1 (Figure 1E right).

These data suggest the existence of 19 candidate genes under dosage compensation across the NCI60 panel, partially related by common chromosomal locations or other functional interactions.

B. *A putative network of miRNA-transcription factor regulatory loops links all candidate genes under potential dosage compensation*

miRNAs and transcription factor networks have been implicated in regulation of gene expression [12], including gene dosage compensation by regulatory feed-forward loops [13]. Since there is expression data available of miRNA [14] and transcription factors (TF)[10] for the NCI60 panel, we asked whether there is a connection between our 19 candidate genes and miRNA/transcription factor regulation. In order to explore this connection, we calculated the correlation coefficients between the Z-scores of copy number variations of the candidate genes and the Z-scores of miRNA/TF expression data across the NCI60 panel. As depicted in figure 2A, there are both miRNAs and TF with high positive or negative correlation. This result suggests the existence of miRNAs and TFs that potentially regulate the expression of clusters of candidate genes simultaneously. To evaluate whether there are any reported or predicted connections between these candidate genes and miRNAs/TF, we constructed a network of putative regulatory interactions based on the information available at the databases Mirtarbase, MiRBase, Pazar, TRED and the study of Neph et al [15]. In addition, we included the interactions between miRNAs and TFs using the information available at Transmir and CircuitsDB 2. This generated a network with 540 nodes and

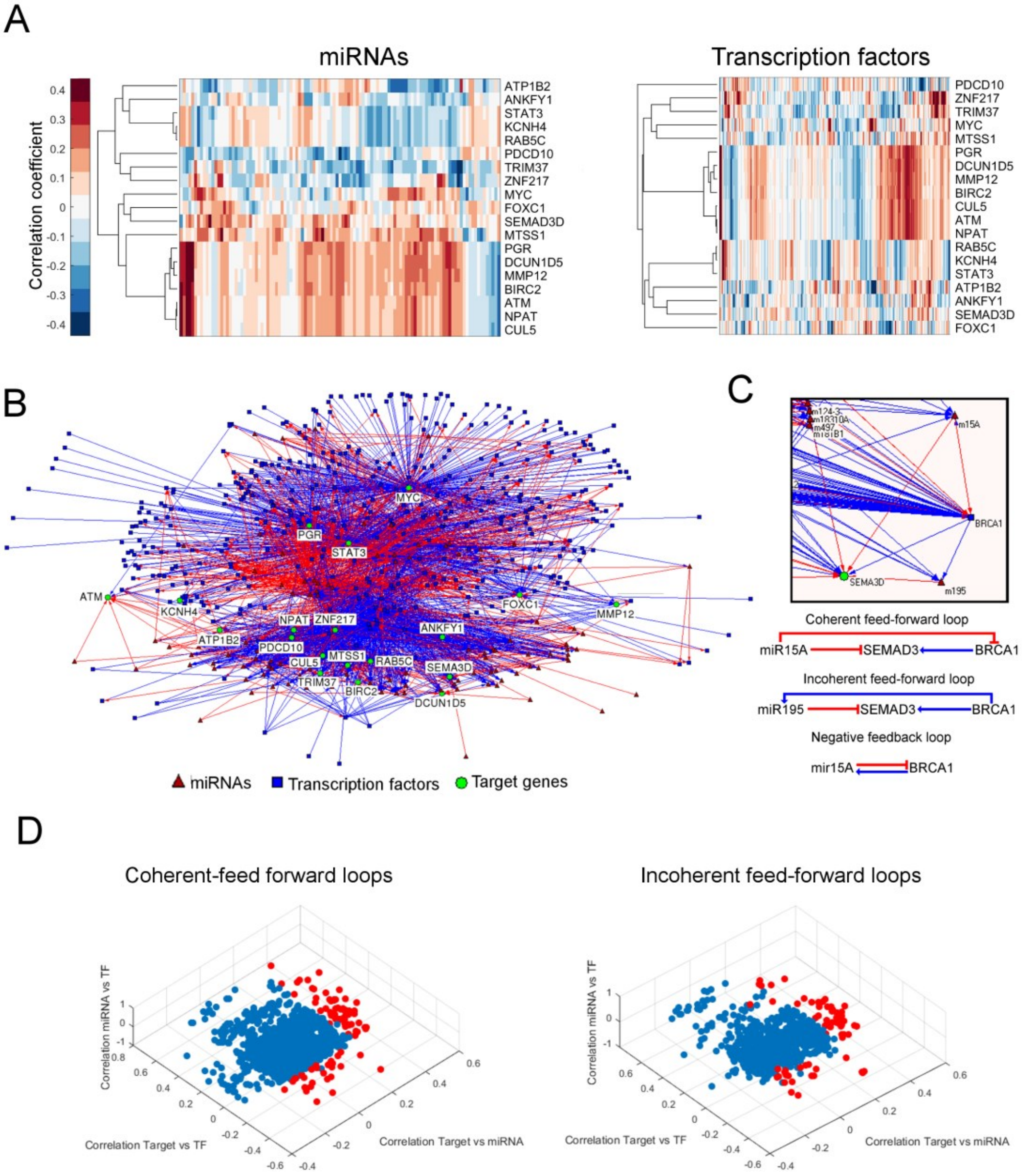

Figure 3. A putative regulatory network is automatically constructed from available TF and miRNA interaction data. Correlation coefficients of candidate target genes with miRNAs and transcription factors (A). Putative network topology of target-miRNA-TF interactions (B) including examples of coherent feedforward, incoherent feed-forward and negative feedback loops (C). The correlation coefficients from A are employed to choose those forward loops with the highest likelihoods to be real interactions (D).

2848 interactions, which connect all 19 candidate genes (targets) (Figure 3B). In order to examine the target/miRNA/TF network for the presence of regulatory interactions with potential activity on gene dosage compensation, we searched for motifs with systems-level properties including positive and negative feedback loops (between miRNAs and TFs), coherent feed-forward loops and incoherent feed-forward loops [8]. We identified a total number of 2500 putative regulatory motifs. For example, miRN15A participates in a coherent feed-forward loop inhibiting both SEMAD3 and BRCA1, a positive regulator of BRCA1. Also, BRCA1 forms an incoherent feed-forward loop, activating the transcription of SEMAD3 but also miR195, a negative regulator of SEMAD3. In addition, miR15A forms a negative feedback loop with BRCA1 (Figure 2C). In total, this network includes 78 negative feedback loops but no positive feedback loops between TFs and miRNAs, 1422 coherent feed-forward loops and 1000 incoherent feed-forward loops formed by the interaction of TFs and miRNAs with the target genes. MYC, ZNF217 and STAT3 are the genes involved in the majority of regulatory motifs.

To gain insight into these target/miRNA/TF interactions, we calculated the correlation coefficients among the values of the elements for each regulatory loop including miRNA levels (miR), target copy number (CN) and transcription factor expression (TF). The correlations between miR/CN, TF/CN and TF/miR displayed a high density around zero for both coherent and incoherent feed-forward loops. However, several interactions are separated from that central core, suggesting that they could participate in real regulatory loops (Figure 2D). Therefore, we selected the regulatory loops that could play a role in gene dosage compensation, having at least one of its correlations separated from the central core. For coherent feed-forward loops (where a miRNA inhibits a TF) we included loops with a positive miR/CN correlation (higher than 0.25), a negative TF/CN correlation (lower than -0.25), or a negative miR/TF correlation (lower than -0.25). For the incoherent feed-forward loops (where a TF activates a miRNA) we selected those with positive miR/CN correlations (higher than 0.25), negative CN/TF correlations (lower than -0.25) or positive TF/miR correlations (higher than 0.35). These putative regulatory loops generated a simplified regulatory network of the interactions with the highest correlations expected for gene dosage compensation (Figure 3A).

These results indicate that several putative regulatory loops link all the candidate target genes. These regulatory motifs with potential systems-level properties are widely present within this putative network. The high correlation/anticorrelation for the copy number variations of some target genes and the expression levels of some miRNA/TFs suggests that some of these regulatory motifs may be involved in gene dosage compensation. However, the assessment of the complexity of this regulatory network requires a systems-level approach.

C. A large scale mathematical model of miRNA-transcription factor interactions with the candidate genes

The simplified network includes 434 nodes and 2745 arcs (Figure 4A). Due to the high complexity of miRNA/TF regulatory netwoks, we proceeded to a systems-level approach in order to gain insight into the complexity of the gene dosage compensation mediated by miRNAs and TFs. Therefore, we constructed a mathematical model using COPASI using our automated biocomputational platform considering basic network motifs for the target gene, miRNA and TF (Figure 4B). The biochemical model includes 182 species, 364 reactions and 823 parameters and was fitted using the Parameter Estimation function of COPASI. The resulting model presents a good fitting of the data. This result indicates that we have now the first large-scale mathematical model to perform future studies on gene dosage compensation.

Taken together, these results indicate that several genes with high copy number variations have very low changes in expression, suggesting that they are under the influence of gene dosage compensation. Those candidate genes are highly interconnected with miRNAs and transcription factors leading to the formation of different types of regulatory loops that could contribute to the mechanism of gene dosage compensation. The high complexity of the resulting network required a dedicated computational platform for the automatic construction of a large-scale mathematical model of gene regulation that can be used to perform studies on gene dosage compensation.

IV. Discussion

We developed a computational platform to automatically construct large scale models of miRNAs and TF interactions with a novel approach. The classical approach starts with those disregulated miRNAs followed by the identification of gene targets, which is very inefficient because each miRNA can alter the expression of hundreds of genes by only 1.5 to 4 fold [8] and it is the cooperative effect of miRNA networks that makes them robust regulators [16]. It is therefore very hard to identify single miRNA-target interactions with relevant biological function, requiring extensive molecular biology work for validation. Therefore, we propose the first systems-level approach in the opposite direction, identifying first targets under gene dosage compensation and second, identifying their regulating miRNAs.

To our knowledge, previous bioinformatic work focused on miRNA networks based on differential expression data between tumoral and normal tissues. Those differences are a consequence of genetic instability and as such, highly heterogeneous among cancer types, because they arise from different evolutionary trajectories of cancer. In contrast, our work is the first to explore the core stability of gene expression in cancer, which mediates its survival despite its genetic instability. This stability core presumably would be homogeneous among cells and cancer types; and it might be represented by a set of genes tightly regulated by stable miRNA networks to ensure gene dosage compensation. miRNA networks are robust regulators of gene expression upon environmental changes [16] and they show adaptation to gene dosage through the formation of regulatory circuits with transcription factors [13]. Thus, we hypothesize that cancer has a robust Achilles-Heel due to an increased sensitivity to perturbations in these circuits, which is not necessarily reflected as differences in miRNA expression levels but at systems-level properties.

In conclusion, the present work led to the construction of a complex mathematical model to study gene dosage compensation and formulated model-driven hypothesis for the identification of novel targets against aneuploid cancer.

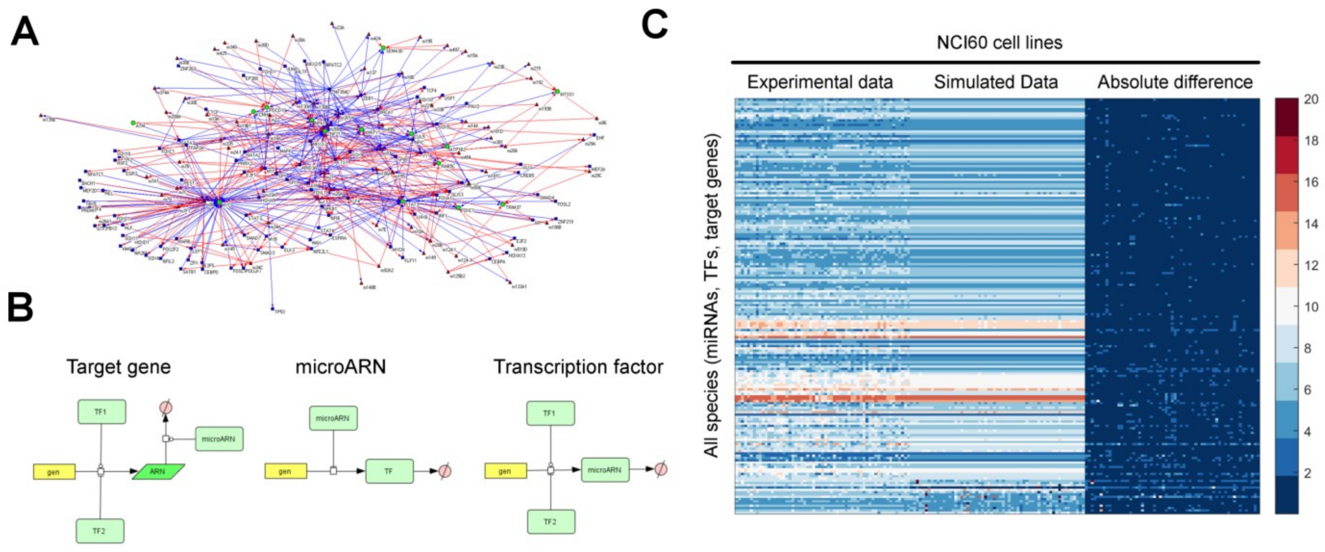

Figure 4. A mathematical model of miRNA-TF factor regulation related to genes under dosage compensation. Simplified network topology (A). Basic structural motifs for mathematical modeling (B). Model fitting to experimental data (C).

References

[1] H. Kitano, "Cancer as a robust system: implications for anticancer therapy.," *Nat. Rev. Cancer*, vol. 4, no. 3, pp. 227–35, Mar. 2004.

[2] R. H. Devlin, D. G. Holm, and T. a Grigliatti, "Autosomal dosage compensation Drosophila melanogaster strains trisomic for the left arm of chromosome 2.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 79, no. 4, pp. 1200–4, Feb. 1982.

[3] S. Stingele, G. Stoehr, K. Peplowska, J. Cox, M. Mann, and Z. Storchova, "Global analysis of genome, transcriptome and proteome reveals the response to aneuploidy in human cells.," *Mol. Syst. Biol.*, vol. 8, no. 608, p. 608, Jan. 2012.

[4] M. R. Fabian, N. Sonenberg, and W. Filipowicz, "Regulation of mRNA translation and stability by microRNAs.," *Annu. Rev. Biochem.*, vol. 79, pp. 351–79, Jan. 2010.

5] M. Malumbres, "miRNAs versus oncogenes: the power of social networking," *Mol. Syst. Biol.*, vol. 8, no. 569, pp. 1–2, Feb. 2012.

[6] L. P. Lim, N. C. Lau, P. Garrett-Engele, A. Grimson, J. M. Schelter, J. Castle, D. P. Bartel, P. S. Linsley, and J. M. Johnson, "Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs.," *Nature*, vol. 433, no. 7027, pp. 769–73, Feb. 2005.

[7] W. Ritchie, J. E. J. Rasko, and S. Flamant, "MicroRNA target prediction and validation.," *Adv. Exp. Med. Biol.*, vol. 774, pp. 39–53, Jan. 2013.

[8] J. Vera, X. Lai, U. Schmitz, and O. Wolkenhauer, "MicroRNA-regulated networks: the perfect storm for classical molecular biology, the ideal scenario for systems biology.," *Adv. Exp. Med. Biol.*, vol. 774, pp. 55–76, Jan. 2013.

[9] J. Yao, S. Weremowicz, B. Feng, R. C. Gentleman, J. R. Marks, R. Gelman, C. Brennan, and K. Polyak, "Combined cDNA array comparative genomic hybridization and serial analysis of gene expression analysis of breast tumor progression.," *Cancer Res.*, vol. 66, no. 8, pp. 4065–78, Apr. 2006.

[10] W. H. Gmeiner, W. C. Reinhold, and Y. Pommier, "Genome-wide mRNA and microRNA profiling of the NCI 60 cell-line screen and comparison of FdUMP[10] with fluorouracil, floxuridine, and topoisomerase 1 poisons.," *Mol. Cancer Ther.*, vol. 9, no. 12, pp. 3105–14, 2010.

[11] A. M. Gholami, H. Hahne, Z. Wu, F. Auer, C. Meng, M. Wilhelm, and B. Kuster, "Global proteome analysis of the NCI-60 cell line panel," *Cell Rep.*, vol. 4, no. 3, pp. 609–620, 2013.

[12] S. Arora, R. Rana, a. Chhabra, a. Jaiswal, and V. Rani, "MiRNA-transcription factor interactions: A combinatorial regulation of gene expression," *Mol. Genet. Genomics*, vol. 288, no. 3–4, pp. 77–87, 2013.

[13] L. Bleris, Z. Xie, D. Glass, A. Adadey, E. Sontag, and Y. Benenson, "Synthetic incoherent feedforward circuits show adaptation to the amount of their genetic template," *Mol. Syst. Biol.*, vol. 7, no. 519, pp. 1–12, 2011.

[14] P. E. Blower, J. S. Verducci, S. Lin, J. Zhou, J.-H. Chung, Z. Dai, C.-G. Liu, W. Reinhold, P. L. Lorenzi, E. P. Kaldjian, C. M. Croce, J. N. Weinstein, and W. Sadee, "MicroRNA expression profiles for the NCI-60 cancer cell panel.," *Mol. Cancer Ther.*, vol. 6, no. 5, pp. 1483–1491, 2007.

[15] S. Neph, A. B. Stergachis, A. Reynolds, R. Sandstrom, E. Borenstein, and J. A. Stamatoyannopoulos, "Circuitry and dynamics of human transcription factor regulatory networks.," *Cell*, vol. 150, no. 6, pp. 1274–1286, 2012.

[16] H. Herranz and S. M. Cohen, "MicroRNAs and gene regulatory networks: managing the impact of noise in biological systems.," *Genes Dev.*, vol. 24, no. 13, pp. 1339–44, Jul. 2010.

[17] K. J. Bussey, K. Chin, S. Lababidi, M. Reimers, W. C. Reinhold, W. L. Kuo, F. Gwadry, H. Kouros-Mehr, J. Fridlyand, A. Jain, C. Collins, S. Nishizuka, G. Tonon, A. Roschke, K. Gehlhaus, I. Kirsch, D. A. Scudiero, J. W. Gray and J. N. Weinstein, J. N, "Integrating data on DNA copy number with gene expression levels and drug sensitivities in the NCI-60 cell line panel", . Molecular Cancer Therapeutics, 5:853-867, 2006.

[18] U. T. Shankavaram, W. C. Reinhold, S. Nishizuka, S. Major, D. Morita, K. K. Chary, M. A. Reimers, U. Scherf, A. Kahn, D. Dolginow, J. Cossman, E. P. Kaldjian, D. A. Scudiero, E. Petricoin, L. Liotta, J. K. Lee and J. N. Weinstein, "Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study", Molecular Cancer Therapeutics, 6:820-832, 2007.

[19] S. Hsu, Y. Tseng, S. Shrestha, Y. Lin, A. Khaleel, C. Chou, C. Chu, H. Huang, C. Lin, S. Ho, T. Jian, F. Lin, T. Chang, S. Weng, K. Liao, I. Liao, C. Liu and H. Huang, "miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions", Nucleid Acids Research, pages 78-85, 2014.

[20] A. Kozomara and S. Griffiths-Jones, "miRBase: annotating high confidence microRNAs using deep sequencing data", Nucleic Acids Research, 42:D68-D73, 2014.

[21] J. Wang, M. Lu, C. Qiu, and Q. Cui, "TransmiR: a transcription factor-microRNA regulation database", Nucleic Acids Research, 38(1):D119-D122, 2008.

[22] E. Portales-Casamar, D. Arenillas, J. Lim, M. Swanson, S. Jiang, A. McCallum, S. Kirov and W. Wasserman, "The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences", Nucleic Acids Research, 37:D54-D60, 2009.

[23] F. Zhao, Z. Xuan, L. Liu and M. Zhang, "TRED: a Transcriptional Regulatory Element Database and a platform for in silico gene regulation studies", Nucleic Acids Research, vol. 33, 2005.

[24] O. Friard, A. Re, D. Taverna, M. De-Bortoli and D. Corá, "CircuitsDB: a database of mixed microRNA/transcription factor feed-forward regulatory circuits in human and mouse", BMC Bioinformatics, 11:435, 2010.