NATIONAL CENTER FOR THE MULTISCALE ANALYSIS OF GENOMIC AND CELLULAR NETWORKS

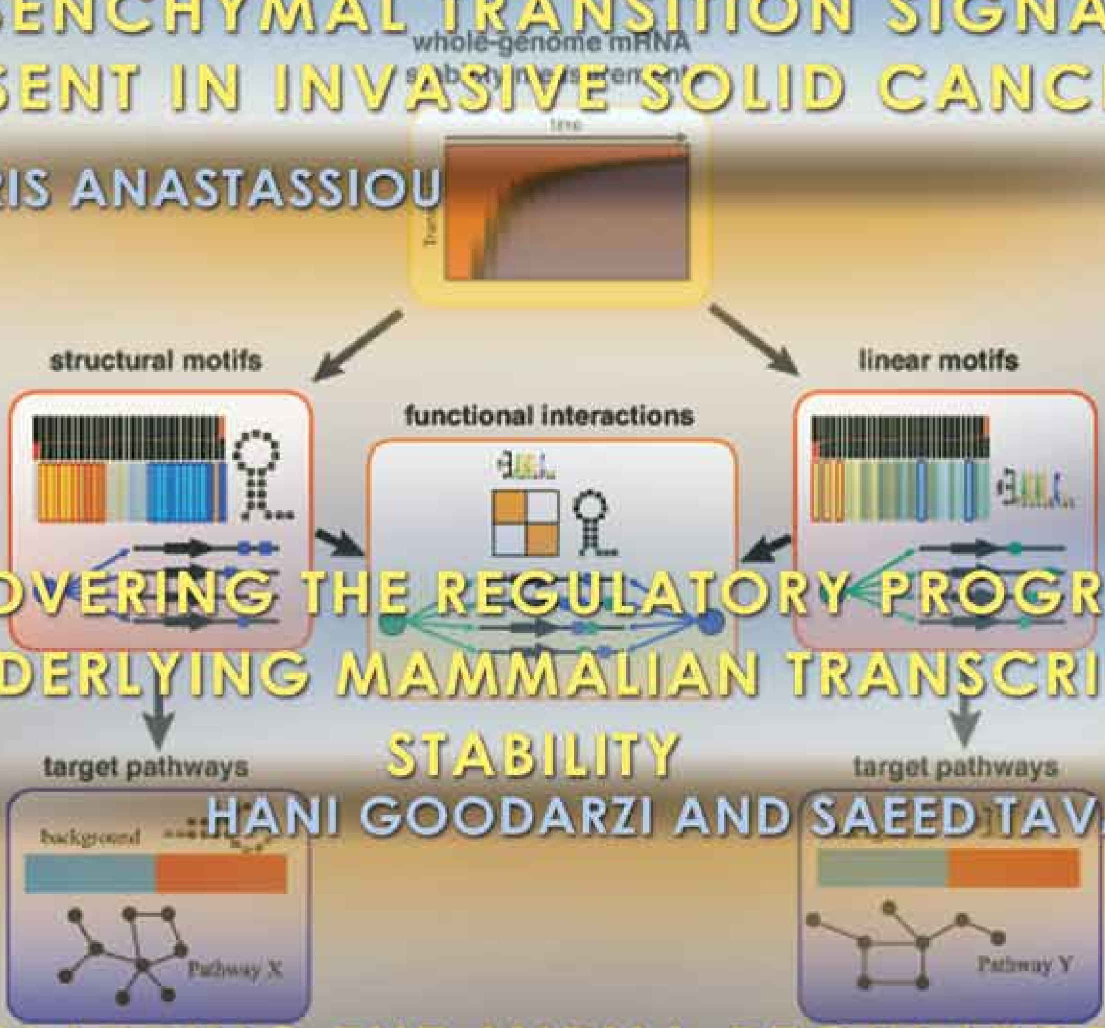
SPRING 2012

ISSUE NO.5

MAGNet NEWSLETTER

A MESENCHYMAL TRANSITION SIGNATURE PRESENT IN INVASIVE SOLID CANCERS

DIMITRIS ANASTASSIOU



DISCOVERING THE REGULATORY PROGRAMS UNDERLYING MAMMALIAN TRANSCRIPT STABILITY

HANI GOODARZI AND SAEED TAVAZOIE

MAPPING THE MIRNA REGULATORY NETWORK IN GLIOBLASTOMA

PAVEL SUMAZIN AND ANDREA CALIFANO

SPRING 2012
ISSUE NO.5

MAGNet Newsletter
http://magnet.c2b2.columbia.edu

MAGNet NEWSLETTER

FEATURES

SECTIONS

INTRODUCTION

2012 is an exciting year for the national Center for Multiscale Analysis of Genomic and Cellular Networks (MAGNet) at Columbia University. It is our seventh anniversary of successful growth and marks the first step to create a new Department of Systems Biology (DSB) at Columbia, which is the direct result of the success of MAGNet center related activities. The new Department will provide a collaborative environment across disciplines to advance our study of human disease. We already welcome four new faculty members who will be appointed in the DSB upon its creation, including Dr. Saeed Tavazoie, who joins us from Princeton at the full professor level, and Drs. Sagi Shapira, Yufeng Shen, and Peter Sims, who join us at the Assistant Professor level. MAGNet plays a central role in providing computational tools, methodologies, and information databases for the research community to investigate biological processes and disease states. Additionally, through our expanding infrastructure, we will continue to encourage scientific collaborations that benefit research centers and organization all over the world.

In an effort to keep you abreast of the accomplishments of MAGNet researchers, this issue highlights the results from three out of many articles recently published by MAGNet investigators. Consistent with the mission of MAGNet, each of these studies have contributed to the creation of novel computational tools that dissect and interrogate cellular and genomic network models to produce novel, biologically relevant hypotheses that were experimentally validated.

Our first feature article, published in Nature Precedings by Dr. Dimitris Anastassiou and collaborators, introduces a unique molecular signature common to many types of highly invasive solid tumors. Characterizing the genetic profiles of such tumors is important, since most cancer deaths result from progression to metastatic disease. While the biological mechanisms underlying metastatic progression remain largely elusive, there is increasing evidence suggesting a connection with a cancer specific epithelial-mesenchymal transition (EMT) that increases invasion potential. The gene expression signature proposed by Dr. Anastassiou and colleagues provides a unifying framework to study EMT transition in most invasive tumor types. This signature was discovered by computational analysis of gene expression patterns and was validated experimentally, in vivo, using a mouse xenograft model. Specifically, human cancer cells were implanted in immunocompromised mice to observe the complete tumor life cycle from proliferative disease to metastatic progression. In brief, the new signature includes many EMT markers, including transcription factor Slug, fibronectin, and a-SMA. Its presence has been confirmed in different cancer types by studying a variety of publicly available datasets and can be potentially used to identify diagnostic biomarkers, as well as candidate targets to abrogate metastatic progression.

The second feature article, published in Nature by Dr. Saeed Tavazoie's group, examines post-transcriptional regulation of mammalian RNA by developing a computational regulatory model. Specifically, the article introduces the TEISER algorithm (Tool for Eliciting Informative Structural Elements in RNA), to predict structural motifs in RNA that are important for its stability and regulation and that are universally informative across all RNA transcripts. These predictions were then experimentally validated to support their functional and regulatory roles. For instance, they identified HNRPA2B1, a human gene whose binding to one of their predicted structural RNA elements, sRSM1 (structural RNA Stability Motif-1), is crucial for the regulation of other target genes. The TEISER algorithm constitutes a significant addition to our repertoire of computational tools that can be used to study post-transcriptional regulatory networks.

Utilizing a similar approach, the third and final Feature article in this newsletter, published in Cell by a team of postdocs and graduate students in Dr. Califano's lab led by Dr. Sumazin, combines computational tools with biological research methods to characterize a novel microRNA (miR) mediated regulatory network. mRNA and miR expression data obtained from The Cancer Genome Atlas was analyzed using a newly developed HERMES algorithm. This revealed an extensive network in which pairs of competing endogenous RNA species (ceRNA) can regulate each other by sequestering miRs that target both species depending on their expression. These findings were experimentally confirmed and were shown to explain a significant component of PTEN expression variability in glioblastoma as a result of deletions of 13 ceRNA competing with PTEN. They propose an important role for genes never before associated with the disease and provide mechanistic insight into the depth gene regulation. This computational approach can also be applied to other biological abnormalities that result from RNA-RNA interactions and cause disease.

Collectively, the research investigations summarized above support MAGNet's broader goal of creating computational models of processes aimed at preventing disease, improving diagnostics and providing more therapeutic options. The integration of experimental biology with mathematical modeling results in fresh insights and new approaches to the management of diseases such as cancer. Our new department, which uniquely benefits from the success of MAGNet, will allow for the collaboration among clinicians and researchers from a variety of fields including oncology, mathematics, physics, information technology, imaging sciences, and computer science.

We look forward to another successful and productive MAGNet year.

--Andrea Califano

A MESENCHYMAL TRANSITION SIGNATURE PRESENT IN INVASIVE SOLID CANCERS

DIMITRIS ANASTASSIOU
DEPARTMENT OF ELECTRICAL ENGINEERING, COLUMBIA UNIVERSITY

INTRODUCTION

About two years ago we realized that a particular precise set of genes appeared to be coordinately and strongly overexpressed in some samples from all types of solid cancers [1]. These genes were significantly overexpressed only in samples that had exceeded a particular stage of invasiveness, specific to each cancer type. For example, this phenomenon occurred when ovarian cancer progressed to Stage III, colon cancer progressed to Stage II, and ductal carcinoma in situ (DCIS) progressed to invasive ductal carcinoma (IDC). Prominent among these genes were collagen COL11A1 and thrombospondin THBS2. Collagen COL11A1 is the best proxy of the whole signature in the following sense: Identifying the genes whose expression was most associated with that of COL11A1 consistently included all the other key genes of the signature somewhere at the top of the list, better than any other choice. And furthermore this association was true in all solid cancer types that we tried (glioblastoma being the only exception – more on that later), but never in data sets of normal samples. So, we analyzed many data sets from many cancer types and ended up with the list of the top genes of this "universal" cancer signature shown in Table 1 as those being most associated with COL11A1 in all datasets.

It is easy to confirm all of this. For example, we can just go to Supplementary Data 3 of a paper [2] comparing the gene expression of DCIS vs. IDC and then "sort largest to smallest" the genes in the "up in IDC" sheet of the Excel file in the column showing the fold change. The resulting list of top genes starts from COL11A1, COL10A1, MFAP5, LRRC15, INHBA, FBN1, SULF1, GREM1, COL5A2, LOX, COL5A1, THBS2. All 12 of them are in the list of Table 1. We can do the same in ovarian or colon cancer datasets using the staging thresholds mentioned above, and we will find similar results. Or, we can take the list of genes in Table 1 and use it as input for Gene Set Enrichment Analysis (GSEA) provided

Rank	Gene	Rank	Gene	Rank	Gene	Rank	Gene
1	COL11A1	17	FN1	33	LOXL2	49	COPZ2
2	THBS2	18	AEBP1	34	COL6A3	50	NOX4
3	COL10A1	19	SULF1	35	MXRA5	51	EDNRA
4	COL5A2	20	FBN1	36	MFAP5	52	ACTA2
5	INHBA	21	ASPN	37	NUAK1	53	PDGFRB
6	LRRC15	22	SPARC	38	RAB31	54	RCN3
7	COL5A1	23	CTSK	39	TIMP3	55	SNAI2
8	VCAN	24	TNFAIP6	40	CRISPLD2	56	C1QTNF3
9	FAP	25	HNT	41	ITGBL1	57	COMP
10	COL1A1	26	EPYC	42	CDH11	58	LGALS1
11	MMP11	27	MMP2	43	TMEM158	59	THY1
12	POSTN	28	PLAU	44	SPOCK1	60	PCOLCE
13	COL1A2	29	GREM1	45	SFRP4	61	COL6A2
14	ADAM12	30	BGN	46	SERPINF1	62	GLT8D2
15	COL3A1	31	OLFML2B	47	DCN	63	NID2
16	LOX	32	LUM	48	C7orf10	64	PRRX1

Table 1: COL11A1 associated gene signature

by the Broad Institute against the Molecular Signatures Database (MSigDB) (www.broadinstitute.org/gsea/msigdb). The results will include many "hits" with P value exactly equal to "zero." Among those (with "$P = 0e^{0}$"), there will be many occurrences of sets of genes expressed in higher-stage samples from many cancer types, such as nasopharyngeal, head and neck, urothelial, lymphomas, etc. Following are some examples:

MSigDB Gene Set Name	**Description**
SENGUPTA_NASOPHARYNGEAL_CARCINOMA_UP [286]	Genes up-regulated in nasopharyngeal carcinoma relative to the normal tissue.
GRUETZMANN_PANCREATIC_CANCER_UP [346]	Genes up-regulated in pancreatic ductal adenocarcinoma (PDAC) identified in a meta-analysis across four independent studies.
LINDGREN_BLADDER_CANCER_CLUSTER_2B [389]	Genes specifically up-regulated in Cluster IIb of urothelial cell carcinoma (UCC) tumors.
PICCALUGA_ANGIOIMMUNOBLASTIC_LYMPHOMA_MA_UP [207]	Up-regulated genes in angioimmunoblastic lymphoma (AILT) compared to normal T lymphocytes.
VECCHI_GASTRIC_CANCER_ADVANCED_VS_EARLARLY_UP [167]	Up-regulated genes distinguishing between two subtypes of gastric cancer: advanced (AGC) and early (EGC).
CROMER_TUMORIGENESIS_UP [44]	Tumorigenesis markers of head and neck squamous cell carcinoma (HNSCC): up-regulated in the 'early' tumors vs. normal samples.

Table 2: Cancer related MSigDB gene sets highly enriched in genes from Table 1.

None of these cancer types had participated in any way whatsoever in the derivation of the signature. This validation of the signature by pointing to all kinds of cancer types in MSigDB suggests that the signature may reflect a universal biological mechanism present in the invasive stage of all solid cancers.

SO, WHAT DOES THIS SIGNATURE REPRESENT?

It so happens that many among the genes in Table 1 are known markers of a cell transdifferentiation process known as epithelial mesenchymal transition (EMT). It is believed that EMT is a key mechanism by which cancer cells lose cell adhesion and become migratory and invasive [3, 4] as a result of obtaining mesenchymal traits. Similar mechanisms are employed during early embryonic development, so it is also believed that EMT-based cancer cell invasiveness is achieved by reactivating such preexisting programs. So, it appeared that this signature is due, at least in part, to some cells having undergone an EMT.

It is tempting to hypothesize that the cancer cells themselves undergo an EMT at a particular stage of invasiveness, expressing the mesenchymal genes of the signature, such as fibroblast activation protein (FAP) and alpha-SMA (ACTA2). But these are genes typically expressed by fibroblasts, which are mesenchymal cells known to be part of the stroma, the connective tissue in the microenvironment adjacent to the tumor. Could it be that the cancer cells transform themselves into motile fibroblasts, and the signature is due to them? This is a controversial hypothesis, but the cancer research pioneer Bob Weinberg himself had made the following statement in his follow-up of the classic "Hallmarks of Cancer" paper [5], coauthored by Douglas Hanahan: "An EMT can convert epithelial carcinoma cells into mesenchymal, fibroblast-like cancer cells that may well assume the duties of cancer-associated fibroblasts (CAFs) in some tumors." Weinberg also writes in his book [6] that "In order to invade adjacent cell layers, carcinoma cells are required to remodel the nearby tissue environment by excavating passageways through the extracellular matrix (ECM) and pushing aside any cells that stand in their path." The genes of Table 1 fit this scenario quite well: they contain proteases, fibronectin, collagens, proteoglycans, a good recipe for remodeling the adjacent connective tissue and allowing the cancer cells to go through.

A more reasonable hypothesis would be that the signature is produced by CAFs that originate from other sources, maybe the bone marrow, or just the local stroma. Indeed, similar signatures have been labeled “stromal” because they were actually found in the stroma, as opposed to the tumor, after using laser capture microdissection to separate the tissues. Any presence of a mesenchymal signature inside the tumor could then be explained by “stromal infiltration” in the tumor.

But could it be that the truth is precisely the opposite? In other words, could it be that the presence of a mesenchymal signature inside the tumor is genuine as cancer cells start undergoing an EMT, but the strong presence of the signature of Table 1 in the stroma is due to the cancer cells themselves having invaded the stroma after undergoing a full fibroblastic transition?

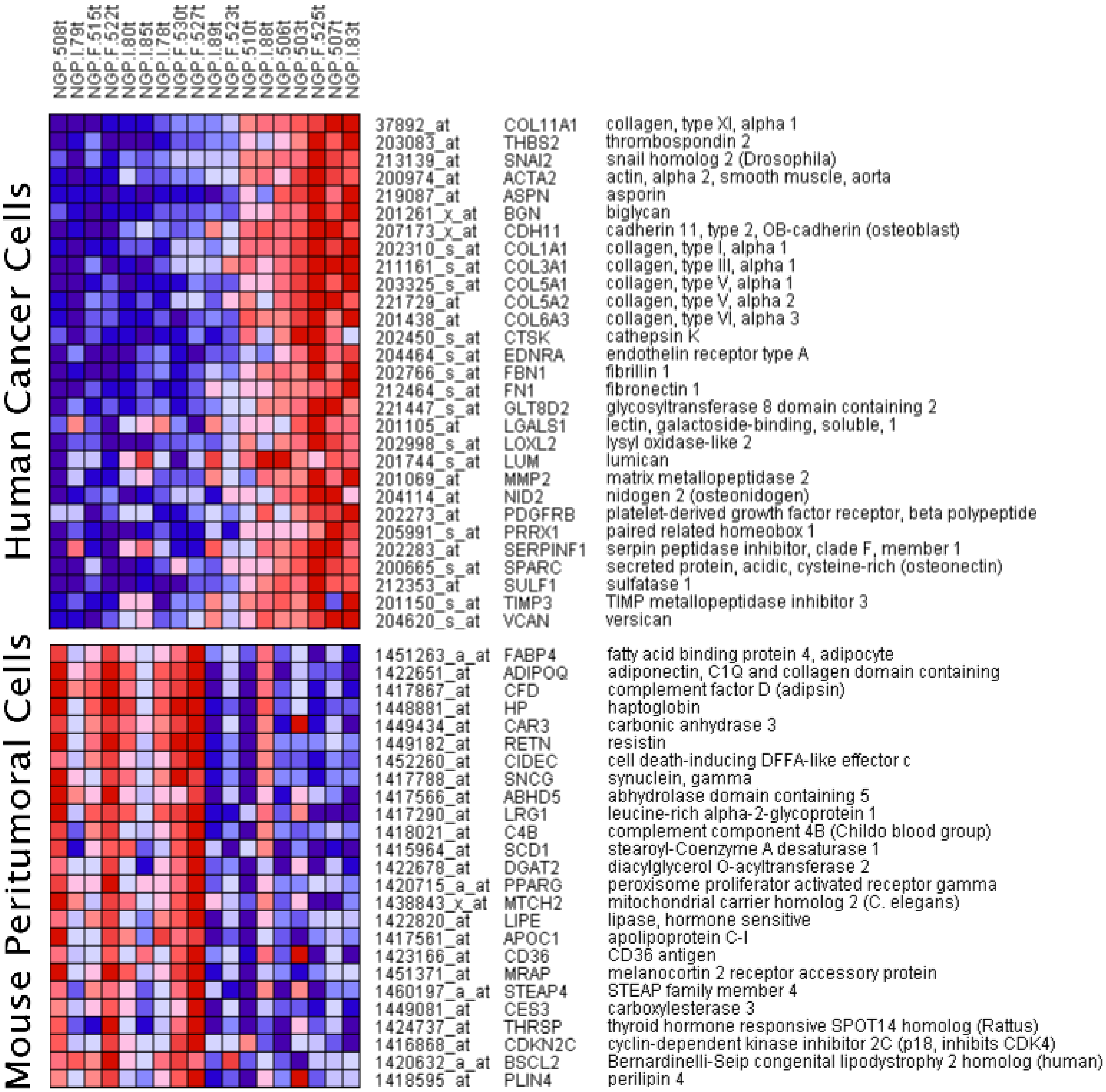


Figure 1: 29 of the genes in Table 1 are found to be strongly co-expressed in human cells (top panel). Further, the expression of these genes is anti-correlated with the expression of adipocyte markers in the peritumoral mouse tissue.

To try to find out, we collaborated with a team of medical researchers at Columbia including Dr. Jessica Kandel and Dr. Darrell Yamashiro, who had extensive experience in performing xenograft experiments, i.e., implanting human cancer cells in immunocompromised mice, so that the human tumors can be observed and scrutinized as they grow, invade the mouse stroma, and eventually metastasize.

HUMAN CANCER CELLS EXPRESS THE MESENCHYMAL TRANSITION SIGNATURE IN VIVO

The signature is clearly present in publicly available datasets of nonepithelial cancers such as neuroblastoma, and since the team had particular experience in these tumors, we decided to use human neuroblastoma cell lines in a total of 18 mice. The tumors were harvested at the proper time and were then profiled twice separately, using microarrays with species-specific probes for either human or mouse. Luckily, there was minimal or zero cross-species hybridization for the genes of the signature, as evidenced by the fact that their pairwise correlation across species was almost always negative.

Remarkably, most of the genes of the signature were found [7] to have been clearly overexpressed in some samples, but only in the human cells, and never in the mouse cells. And COL11A was also a good proxy for the other genes, including the key gene THBS2, as well as mesenchymal markers such as fibronectin and alpha-SMA, as well as the EMT inducing transcription factor SNAI2 (aka Slug). In fact, none of the other EMT-inducing transcription factors was upregulated, only Slug. But none of this had occurred in the mouse cells. Clearly, the human cancer cells had undergone a mesenchymal transition, which represents a more general process than what EMT is assumed to be, because the cells were not epithelial.

Figure 1 shows this coexpression in a heat map of some of these genes in the human cells of the 18 samples. It also shows that there is a strong correlation between the up-regulation of the signature genes in the human cells with a down-regulation of adipocyte (fat cell) markers, such as FABP4 and ADIPOQ, in the peritumoral mouse tissue. This correlation leads to the hypothesis that contextual microenvironmental interactions between cancer cells and peritumoral adipocytes contribute to a full transdifferentiation of the cancer cells into alpha SMA producing fibroblast-like cells (just as Hanahan and Weinberg wrote), which also secrete the key marker, collagen COL11A1. This contextual interaction has already been proposed [8], and it involves the expression of gene MMP11 (prominently included in the signature of Table 1) from the adipocytes. Consistently, we did not find MMP11 upregulated in the human cells.

This also explains why the mesenchymal transition signature in glioblastoma does not include the coexpression of COL11A1: there is no significant presence of adipocytes in the brain. However, the signature is still clearly present in glioblastoma, and in fact, as we recently found [9] by analyzing the dataset from The Cancer Genome Atlas (TCGA), it is associated with time to recurrence: All patients who had exceptionally long time to recurrence following successful treatment, also had exceptionally low levels of the signature. This is consistent with the concept [10] that EMT induces cell "stemness" (ability to self-renew as well as differentiate), so that the cells that did not have the signature had the least stemness, thus making tumor recurrence more unlikely. There are several other related results that we have found, which have been deposited in a preprint [11].

So, it appears that this precise cancer mesenchymal transition signature of Table 1 is present in solid cancers of all types, with the necessary condition that it can be found in significant amounts only if the tumor has exceeded a particular invasive stage. The signature may or may not be detected in samples that have reached or exceeded this invasive stage. However, the absence of the signature in a particular high-stage sample does not necessarily imply that the signature had not been present earlier in time or at some other neighboring location in the heterogeneous [12] tumor (in fact we saw evidence of this heterogeneity in our own xenografts). It is also unclear to what extent the underlying mechanism of mesenchymal transition is causal for invasion and metastasis. It is conceivable, however, that, at least in some cases, it plays a causal role, leading to the exciting possibility that its inhibition may lead to reduction of recurrence and metastasis applicable to multiple cancer types.

REFERENCES:

1. Kim, H., Watkinson, J., Varadan, V. & Anastassiou, D. Multi-cancer computational analysis reveals invasion-associated variant of desmoplastic reaction involving INHBA, THBS2 and COL11A1. BMC Med Genomics 3, 51 (2010).

2. Schuetz, C.S. et al. Progression-specific genes identified by expression profiling of matched ductal carcinomas in situ and invasive breast tumors, combining laser capture microdissection and oligonucleotide microarray analysis. Cancer Res 66, 5278-5286 (2006).

3. Hay, E.D. An overview of epithelio-mesenchymal transformation. Acta Anat (Basel) 154, 8-20 (1995).

4. Kalluri, R. & Weinberg, R.A. The basics of epithelial-mesenchymal transition. J Clin Invest 119, 1420-1428 (2009).

5. Hanahan, D. & Weinberg, R.A. Hallmarks of cancer: the next generation. Cell 144, 646-674 (2011).

6. Weinberg, R.A. The biology of cancer. (Garland Science, New York; 2007).

7. Anastassiou, D. et al. Human cancer cells express Slug-based epithelial-mesenchymal transition gene expression signature obtained in vivo. BMC Cancer 11, 529 (2011).

8. Motrescu, E.R. & Rio, M.C. Cancer cells, adipocytes and matrix metalloproteinase 11: a vicious tumor progression cycle. Biol Chem 389, 1037-1041 (2008).

9. Cheng, W.Y., Kandel, J.J., Yamashiro, D.J., Canoll, P. & Anastassiou, D. A multi-cancer mesenchymal transition gene expression signature is associated with prolonged time to recurrence in glioblastoma. PLoS One 7, e34705 (2012).

10. Mani, S.A. et al. The epithelial-mesenchymal transition generates cells with properties of stem cells. Cell 133, 704-715 (2008).

11. Anastassiou, D. Universality of a mesenchymal transition signature in invasive solid cancers. Available from Nature Precedings <http://hdl.handle.net/10101/npre.2012.6862.1> (2012).

12. Gerlinger, M. et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. The New England journal of medicine 366, 883-892 (2012).

DISCOVERING THE REGULATORY PROGRAMS UNDERLYING MAMMALIAN TRANSCRIPT STABILITY

HANI GOODARZI[1,2] AND SAEED TAVAZOIE[2]

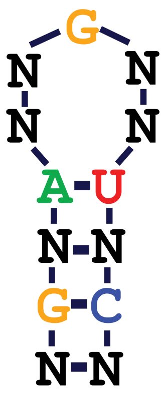
[1]LABORATORY OF SYSTEMS CANCER BIOLOGY, ROCKEFELLER UNIVERSITY

[2]DEPARTMENT OF BIOCHEMISTRY AND MOLECULAR BIOPHYSICS, AND SYSTEMS BIOLOGY INITIATIVE, COLUMBIA UNIVERSITY

In many aspects, RNA is a unique biological molecule. In addition to their role as information carriers, RNA molecules are capable of forming complex folds, where base-pairing between nucleotides create simple secondary structures (e.g. stem-loops) and higher order interactions between distal sequences form more complex tertiary structures. RNA structures affect a variety of cellular processes, such as splicing, localization, translation and RNA stability [1, 2]. Thus, developing real-world dynamical models of cellular behavior in large part relies on decoding post-transcriptional regulatory programs in RNA. Despite recent efforts [2-4], the vast landscape of RNA regulatory elements remains poorly characterized, mainly due to shortcomings in our ability to systematically explore the RNA secondary structures with important roles in regulatory interactions. As such, a full characterization of post-transcriptional regulatory programs relies on effective capturing of both local secondary structures as well as the underlying primary sequences that define post-transcriptional regulatory elements and their interactions [2, 5].

A number of approaches have been developed to tackle this challenging problem. These methods rely on free energy minimization, local sequence alignments or a combination of both alignments and secondary structure predictions to identify putative structural elements (also called structural motifs) [2, 5, 6]. However, *in silico* predictions ignore the conducive role of RNA binding proteins and complexes in facilitating the formation of certain secondary structures in vivo. We therefore sought to bypass the need for predicting structures by efficiently enumerating a large space of potential structural motifs. Based on this strategy, we developed TEISER (Tool for Eliciting Informative Structural Elements in RNA) that systematically explores the space of possible small stem-loops and reveals structural motifs that explain different aspects of experimentally observed RNA behavior [7]. In this approach, stem-loops are represented as context-free grammars that provide a principled data structure for efficient handling of both structure and sequence components (Figure 1) For example, in a whole-genome transcript stability dataset, TEISER discovers structural motif in 5' and 3' untranslated regions (UTRs) that are strongly correlated with the stability of target mRNAs.

In order to evaluate TEISER, we chose to focus on a single factor in RNA metabolism, namely mRNA



$S_0=\varnothing$

$S_1=S_0\mathbf{N}$

$S_2=S_1\mathbf{N}$

$S_3=S_2\mathbf{G}$

$S_4=S_3\mathbf{N}$

$S_5=S_4\mathbf{N}$

$S_6=\mathbf{A}S_5\mathbf{U}$

$S_7=\mathbf{N}S_6\mathbf{N}$

$S_8=\mathbf{G}S_7\mathbf{C}$

$S_9=\mathbf{N}S_8\mathbf{N}$

Figure 1. *Efficient representation of structural elements*. Each structural motif is defined as a series of context- free statements that define the structure and sequence of the motif. A context-free grammar is a set of production rules that describes how phrases are made from their building blocks. Considering a structured RNA molecule as a phrase, its potential building blocks are the different base pairs and bulges (or loops). Internal loops can be represented as a combination of left and right bulges in the middle of phrases. The context-free grammar that we have used to represent structural motifs contains the following production rules: S→S[AUCGN], S→ [AUCGN]S, S→ [AUCGN]S[AUCGN]; wherein the first production rule depicts a right bulge, the second production rule results in a left bulge, and the third production rule creates a base-pairing. In this figure, we have provided the production rules for an exemplary motif.

decay, and insulate it from other aspects (e.g. transcription rate). For transcript stability measurements, we used a non-invasive and powerful pulse-chase experiment based on the seamless incorporation of 4-thiouridine (4sU) into cellular RNA. In this method, cells are incubated with 4sU for a short period, which labels the transcripts produced during this period. Removing 4sU from the media marks the beginning of a time-series experiment in which, at every time point, the fraction of the mRNA population that remains labeled is identified. Using this information, a relative decay rate can be calculated for each transcript based on the rate at which labeled transcripts are replaced by newly synthesized unlabeled mRNAs in the population.

Analyzing this dataset of mRNA decay rates using TEISER, we successfully identified and characterized eight highly significant structural elements that are strongly correlated and most likely causally involved in mRNA stability see(Figure 2). These putative regulatory elements show a variety of other characteristics that support their functionality. For example, some of these motifs are correlated with transcript stability in mouse [8]. They are also highly conserved between human and mouse genomes [7].

To biologically validate our findings, we chose sRSM1 (structural RNA Stability Motif-1) for further analysis and thorough functional characterization. We used a state-of-the-art approach based

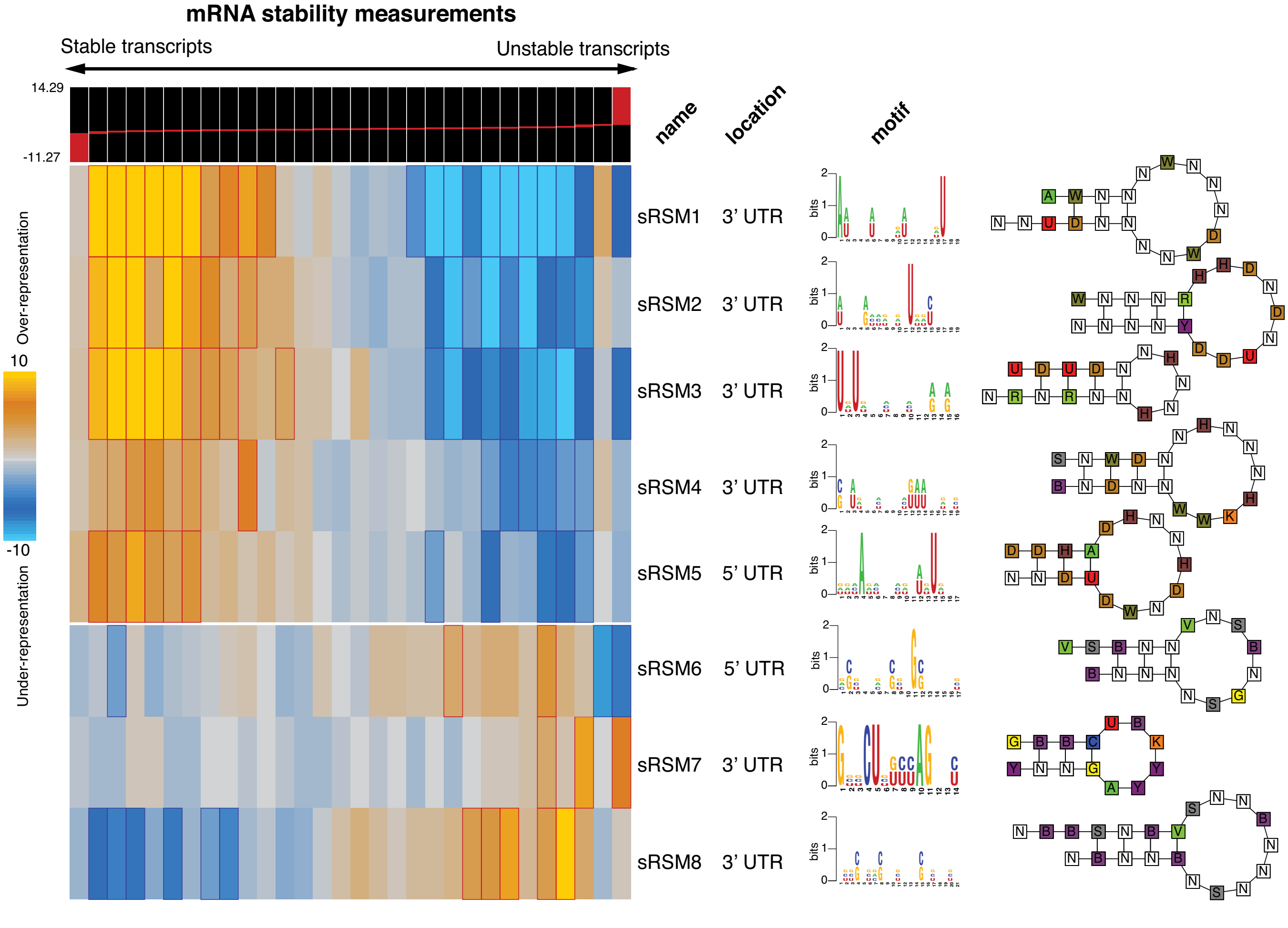


Figure 2. *Discovery of RNA structural motifs informative of genome wide transcript stability*. Each RNA structural motif is shown along with its pattern of enrichment/depletion across the range of mRNA stability measurements throughout the genome. The transcripts are partitioned into equally populated bins based on their stability measures, going from left (highly stable) to right (unstable). In the heatmap representation, a gold entry marks the enrichment of the given motif in its corresponding stability bin, while a light-blue entry indicates motif depletion in the bin. Red and blue borders mark highly significant motif enrichments and depletions, respectively. Included are the motif names, their location (5'UTR or 3'UTR), their sequence information and a structural illustration of each motif generated using the following single letter nucleotide code: Y=[UC], R=[AG], K=[UG], M=[AC], S=[GC], W=[AU], B=[GUC], D=[GAU], H=[ACU], V=[GCA] and N=any nucleotide.

on mass spectrometry [9] to discover candidates that bind, *in vitro*, to the chemically synthesized double-stranded oligonucleotides carrying instances of sRSM1, but not to oligonucleotides with randomly shuffled versions of the motif. Through this approach, we identified HNRPA2B1 as a promising trans-acting factor that binds sRSM1 and regulates the stability of its targets *in vivo*. We subsequently showed that knocking down HNRNPA2B1 results in a significant decrease in the expression level of sRSM1 transcripts and confirmed that this down-regulation is due to changes in stability where sRSM1 targets show a marked increase in their decay rates [7].

In order to show that HNRNPA2B1 directly interacts with sRSM1 target genes in vivo, we took advantage of a method based on cross-linking and immunoprecipitation which, through local UV photoreactivity of bases and amino-acids, enables the detection of direct physical interactions [10]. We expressed a tagged clone of HNRPA2B1, and after crosslinking, immunoprecipitated this protein and the target mRNA molecules that were bound to it (a method called RIP-chip [11]). We observed that sRSM1 targets are significantly overrepresented in the immunoprecipitated population, which indicates a direct interaction between mRNAs carrying sRSM1 in their 3' UTR sequence (Figure 3). A modified version of this approach, followed by high-throughput sequencing (HITS-CLIP [12]), enabled us to footprint the sequences that are directly bound by HNRPA2B1 in vivo and show that sRSM1 elements are significantly overrepresented among the binding sites. These observations clearly demonstrate that HNRPA2B1 directly interacts with sRSM1 in vivo and functions to stabilize its target transcripts through this regulatory element [7]. A modified version of this approach, followed by high-throughput sequencing (HITS-CLIP [12]), enabled us to footprint the sequences that are directly bound by HNRPA2B1 in vivo and show that sRSM1 elements are significantly overrepresented among the binding sites. These observations clearly demonstrate that HNRPA2B1 directly interacts with sRSM1 in vivo and functions to stabilize its target transcripts through this regulatory element [7].

Early on we realized that sRSM1 is a very abundant motif, likely functional in about 4,000 transcripts. HNRNPA2B1, which is also highly expressed in the cell, is a suitable binding partner. However, such a large regulon implies that modulation of HNRNPA2B1 expression and/or activity could result in far-reaching effects on a variety of biological processes. Indeed, we observed that knocking down this protein results in a slight (~10%) but significant increase in growth rate.

The addition of TEISER to our arsenal of computational tools enabled us to create a powerful pipeline for studying post-transcriptional regulatory networks. This approach, outlined in Figure 4 , relies on the identification of regulatory elements, their functional interactions and downstream targets to portray a comprehensive picture of regulatory networks. In this case, in parallel to TEISER, we used a computational platform called FIRE [4] to discover linear regulatory elements (i.e. sequence motifs based on nucleotide preferences at each position) whose presence or absence is strongly correlated with mRNA stability. We identified a large set of lRSMs (linear RNA Stability Motifs), including six known microRNA recognition sites that, combined with the structural motifs identified through TEISER, comprise a large set of RSMs with potentially key roles in post-transcriptional regulation. Identifying post-transcriptional regulatory elements, however, is only the first step in deciphering the post-

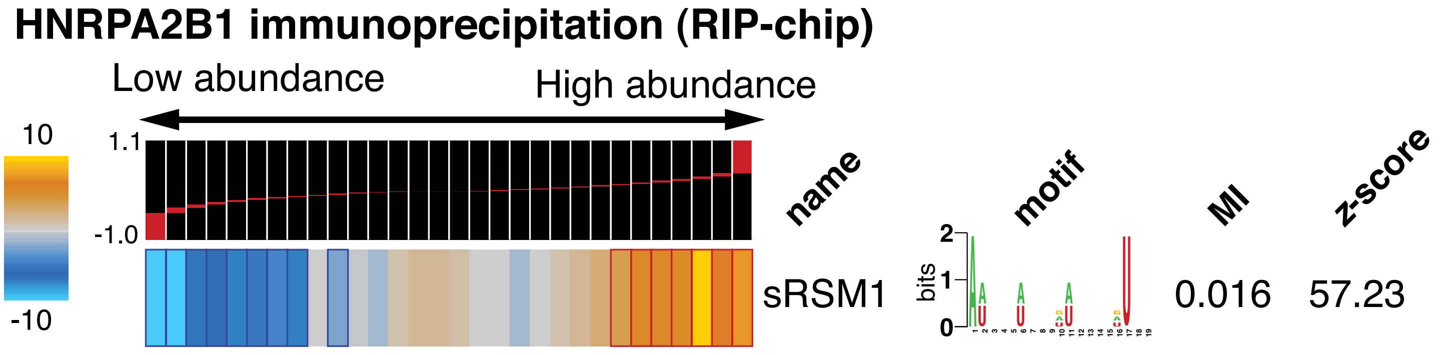


Figure 3. *HNRPA2B1 directly binds sRSM1 targets in vivo*. Using UV-crosslinking followed by immunoprecipitation, mRNAs that bind HNRPA2B1 were extracted and compared against the input mRNA population (RIP-chip). The value assigned to each mRNA denotes its abundance in the immunoprecipitated sample relative to the input control. Bins to the right contain the mRNAs that were captured as interacting partners with HNRPA2B1. Similar to the prior examples, TEISER was used to show the enrichment/depletion pattern of transcripts carrying the sRSM1 structural motif in their 3' UTRs. The values associated with each transcript were calculated as the average of log ratios from biological replicates.

transcriptional regulatory program underlying transcript stability. We then used iPAGE [13] to discover the cellular pathways and processes that are likely targeted and modulated by each element [7]. For example, we observed that sRSM1 is significantly overrepresented in the 3' UTRs of the genes involved in "Notch signaling", while avoiding the UTRs of other pathways such as "nucleosome assembly".

Biological networks function, in large part, through direct and indirect interactions between regulatory modules. The last step in our analytical pipeline aims at discovering putative interactions between the identified regulators. In this context, we defined interaction as a higher than expected chance of finding two given elements co-occurring in the same UTRs, or on the negative side, as lower than expected chance (i.e. avoidance). For example, in case of sRSM1, we observed significant positive and negative interactions with a number of structural and linear motifs, including sRSM8 and sRSM3. These interactions reflect cross talk, or insulation, between the underlying regulatory modules that interact with these elements.

These results demonstrate that while post-transcriptional regulatory mechanisms are poorly characterized, they have potentially far-reaching impact on specific cellular processes. In Figure 5, we have included the totality of the predicted interactions discovered in our mRNA stability data. It is important to note that we have focused on a single cell-line under static conditions and the complete network is most likely significantly more complex and carries more nodes. Nevertheless, the majority of these elements and interactions in this study fall outside of what is currently known. This regulatory map, and those from other types of biological data, set the stage for molecular dissection and predictive modeling of post-transcriptional regulation from sequence. We should also emphasize that our computational framework can be easily applied to the discovery of regulatory

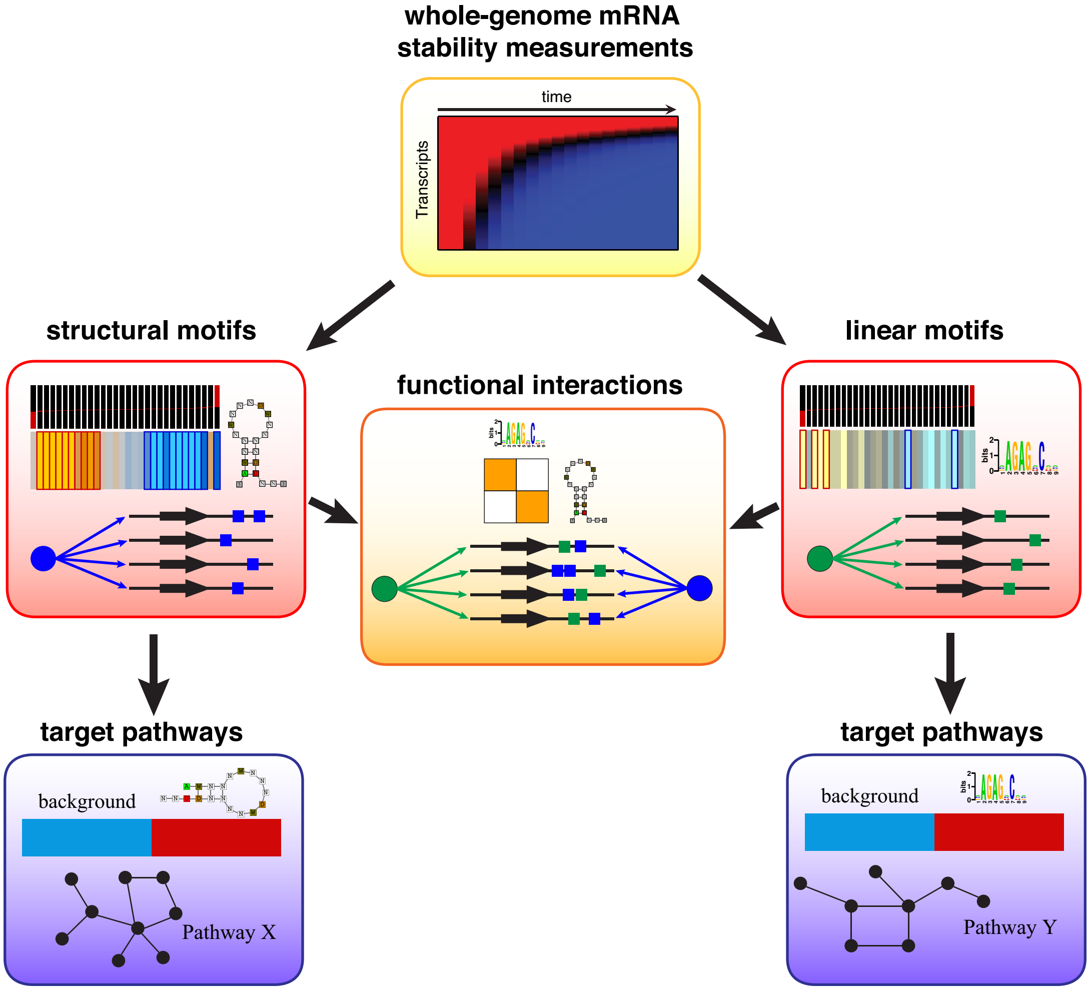


Figure 4. *Schematic of our analysis pipeline*. In the first step, we identify the putative structural and linear motifs that are informative of a given whole-genome dataset. The resulting putative motifs then serve as building blocks for the regulatory modules formed based on their potential interactions and the pathways they most likely target.

elements and interactions underlying other aspects of RNA behavior, including splicing, localization and translation.

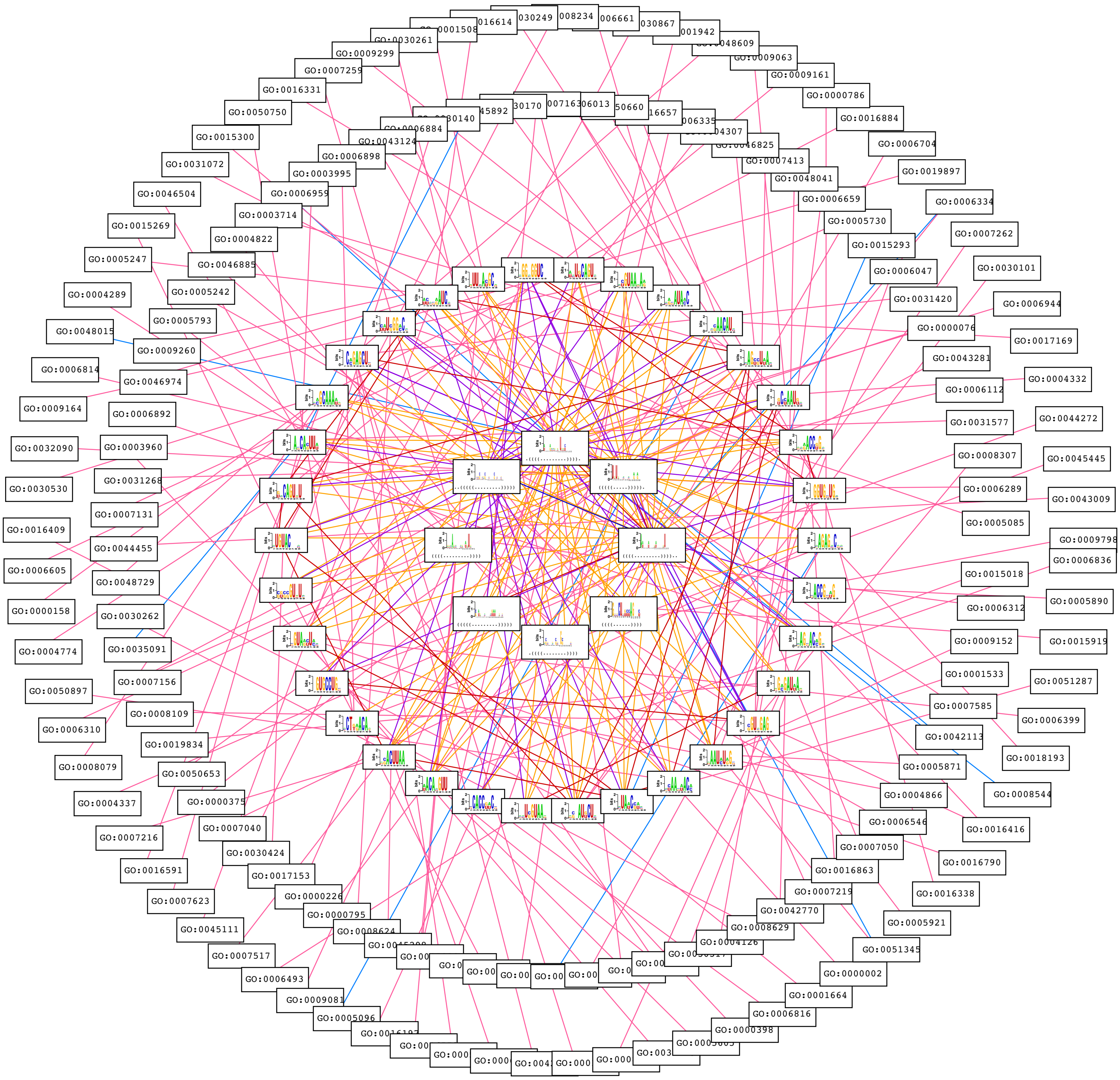


Figure 5. *Predicted post-transcriptional regulatory network underlying transcript stability.* Combining results from TEISER, FIRE and the regulatory interaction maps, we created a network of dependencies that comprises an inferred post-transcriptional regulatory network based on our observations. We also used iPAGE to identify the pathways that are likely targeted by the identified elements. In this figure, the red and blue edges show positive and negative interactions between structural and linear motifs, respectively. The orange and purple edges show the interactions between structural motifs and linear ones, respectively. The pink and light blue edges connect each element with its target pathways. A pink edge means that the genes in a given pathway carry the specified motif more than expected by chance, whereas, light blue edges show that the motif is under-represented in the genes of the corresponding pathway.

REFERENCES

1. Barash, Y., J.A. Calarco, W. Gao, Q. Pan, X. Wang, O. Shai, B.J. Blencowe, and B.J. Frey, Deciphering the splicing code. Nature, 2010. 465(7294): p. 53-9.

2. Rabani, M., M. Kertesz, and E. Segal, Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes. Proceedings of the National Academy of Sciences of the United States of America, 2008. 105(39): p. 14885-14890.

3. Dolken, L., Z. Ruzsics, B. Radle, C.C. Friedel, R. Zimmer, J. Mages, R. Hoffmann, P. Dickinson, T. Forster, P. Ghazal, and U.H. Koszinowski, High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. Rna-a Publication of the Rna Society, 2008. 14(9): p. 1959-1972.

4. Elemento, O., N. Slonim, and S. Tavazoie, A universal framework for regulatory element discovery across all Genomes and data types. Molecular Cell, 2007. 28(2): p. 337-350.

5. Pavesi, G., G. Mauri, M. Stefani, and G. Pesole, RNAProfile: an algorithm for finding conserved secondary structure motifs in unaligned RNA sequences. Nucleic Acids Research, 2004. 32(10): p. 3258-69.

6. Hofacker, I.L., M. Fekete, and P.F. Stadler, Secondary structure prediction for aligned RNA sequences. Journal of molecular biology, 2002. 319(5): p. 1059-66.

7. Goodarzi, H., H.S. Najafabadi, P. Oikonomou, T.M. Greco, L. Fish, R. Salavati, I.M. Cristea, and S. Tavazoie, Systematic discovery of structural elements governing stability of mammalian messenger RNAs. Nature, 2012. advance online publication. doi:10.1038/nature11013

8. Schwanhäusser, B., D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach, Global quantification of mammalian gene expression control. Nature, 2011. 473: p. 337–342.

9. Windbichler, N. and R. Schroeder, Isolation of specific RNA-binding proteins using the streptomycin-binding RNA aptamer. Nature Protocols, 2006. 1(2): p. 638-U4.

10. Jensen, K.B. and R.B. Darnell, CLIP: crosslinking and immunoprecipitation of in vivo RNA targets of RNA-binding proteins. Methods in molecular biology, 2008. 488: p. 85-98.

11. Keene, J.D., J.M. Komisarow, and M.B. Friedersdorf, RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. Nature Protocols, 2006. 1(1): p. 302-7.

12. Licatalosi, D.D., A. Mele, J.J. Fak, J. Ule, M. Kayikci, S.W. Chi, T.A. Clark, A.C. Schweitzer, J.E. Blume, X.N. Wang, J.C. Darnell, and R.B. Darnell, HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature, 2008. 456(7221): p. 464-U22.

13. Goodarzi, H., O. Elemento, and S. Tavazoie, Revealing global regulatory perturbations across human cancers. Molecular Cell, 2009. 36(5): p. 900-11.

MAPPING THE MIRNA REGULATORY NETWORK IN GLIOBLASTOMA

PAVEL SUMAZIN[1] AND ANDREA CALIFANO[1,2]
[1]SYSTEMS BIOLOGY INITIATIVE, COLUMBIA UNIVERSITY
[2]DEPARTMENT OF BIOMEDICAL INFORMATICS, COLUMBIA UNIVERSITY

INTRODUCTION

Deciphering the molecular processes and networks that drive cellular pathophysiology in complex diseases such as cancer has been greatly facilitated by the use of *in silico* analysis. Glioblastoma Multiforme (GBM), a therapy resistant brain cancer is a prime example, where a molecular network controlled by three genes, C/EBPβ, C/EBPδ, and STAT3, appears to be involved in driving the most aggressive form of the malignancy, characterized by a mesenchymal phenotype [1]. Despite this and other mechanistic studies addressing the etiology of high-grade gliomas, it is increasingly evident that some if not most of the genetic mechanisms and variability associated with this disease is still elusive and likely mediated by complex regulatory networks.

MAGNet scientists working on glioma research recently identified a large and previously uncharacterized layer of regulation that allows RNA molecules to regulate each other via a hidden layer of microRNAs (miRs), through a "sponge" mechanisms first demonstrated in plants [2] and later shown to allow interaction between genes and pseudogenes [3]. The fundamental result of the Columbia team is that this regulatory mechanism rather than being relegated to pseudogenes is ubiquitous and mediates pathogenic events, including downregulation of key tumor suppressors as well as upregulation of oncogenes as a result of deletion events by previously unrelated transcribed loci.

Non-coding RNAs (ncRNAs) have been recognized as regulating gene expression at the mRNA transcriptional, stability and translational levels with microRNAs being the most widely studied of the small ncRNA class. Additionally microRNAs themselves are subject to sophisticated regulation and control[4], and altered expression can lead to pathologies and cancer [5]. The term "microRNA (miR) sponge" was coined by Phillip Sharp in 2007 to refer to an engineered molecular tool by which whole miRNA families could be functionally blocked by effectively sequestering miRNA from their endogenous targets[6]. At the time, little was know about naturally occurring miRNA sponges and their role in normal or pathological states. Since then, a few studies have shown roles for endogenous miRNA sponges in plants [2] prokaryotes [4] and metazoans [7]. Notably these initial studies introduced a fundamentally new dimension in cellular regulation; namely that, between transcription and translation lies the domain of protein-coding messenger RNA (mRNA) and non-coding RNA *cross* regulation. We are only beginning to understand the extent to which this cross talk exists in biological systems and how dysregulation of miRNA function affect homeostasis of the cell. This is yet another realm that is benefiting from network biology approaches.

IDENTIFYING A GENOME WIDE HUMAN GLIOMA MIRNA NETWORK USING HERMES

We developed a new algorithm, Hermes, which systematically infers candidate modulators of miR activity from large collections of genome-wide expression profiles of both genes and miRs from the same tumor samples. Hermes employs multivariate analysis and is an extension of the modulator inference by network dynamics (MINDy) algorithm we previously developed and which uses measurements from information theory to identify genes that modulate transcription factor activity via posttranslational modifications [8]. In essence, MINDy and Hermes make inferences by estimating two quantities from information theory: the mutual information (MI) and conditional mutual information (CMI). The MI quantifies how much one variable informs about another variable (i.e., high MI between two variables implies that knowledge about the first variable is predictive of state of the second variable). The CMI calculates the expected value of MI of two variables given the third variable. Using Hermes [9] to analyze 262 glioblastoma samples that were patient matched for

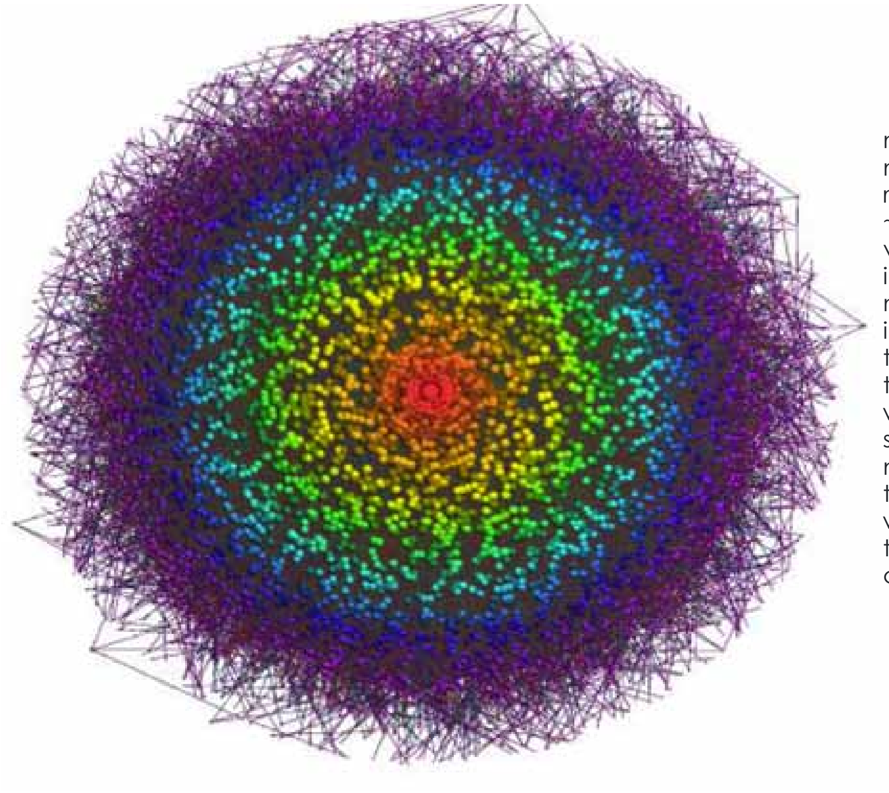


Figure 1. *The mPR network*.

Genome-wide inference of sponge modulators identified a miR program-mediated post transcriptional regulatory (mPR) network including ~248,000 interactions. Its graphic visualization uses nodes to represent individual RNAs and edges to represent miR program-mediated RNA-RNA interactions. Nodes near the center of the graph are contained within more tightly regulated, dense subgraphs, with the densest 564 node subgraph shown in red at the center of the network. The network is scale free, and the color bands, which include nodes with similar connectivity, have a size that increases exponentially with the distance from the center.

gene and miRNA expression profiles from *The Cancer Genome Atlas (TCGA)* [10], we defined a vast and previously unknown post-transcriptional miRNA mediated regulatory network in GBM comprising more than 248,000 miRNA mediated interactions involving miRNA/mRNA cross talk (Figure 1).

MIRNA ACTIVITY MODULATORS

Previous research has identified individual miRs which have key roles in gliomagenesis and progression [11] but very little is know regarding the molecular regulators of miRNA activity on their targets. The data obtained from our Hermes analysis strongly indicates that there are a large number of genes, which are affected by miR activity modulators that have a prominent significance in GBM. We looked at two specific mechanisms by which miR activity is modulated (Figure 2, A and B). The first mechanism, known as the sponge effect, involves the interaction of RNAs of different genes through a common miR program. Any changes in the RNA of gene 1 affects the function of the miRs that control its interactions with the RNA of gene 2, thus affecting the expression levels/functioning of gene 2 as well. The second mechanism, nonsponge modulation consists of protein-protein and miR-protein regulation of the miR mediated posttranscriptional apparatus. Using the miR program-mediated regulatory network (mPR) obtained from Hermes (Figure 2C), we found that the increase in expression of the modulator gene is linked with an increase in the expression of several miRs as well as their respective target genes. Hermes analysis unveiled an astonishing 7,000 gene transcripts (i.e. miR "sponges") and 148 genes involved in nonsponge interactions.

THE GBM MIRNA NETWORK REGULATES ESTABLISHED ONCOGENIC PATHWAYS

The mPR regulatory network we have identified facilitates communication between mRNAs and microRNAs. We further used this methodology to examine specific genes that have been associated with GBM pathophysiology. For this analysis, we chose the phosphatase and tensin homolog gene (PTEN) tumor suppressor, down regulation of which is a hallmark of gliomagenesis. A comparable range of expression between intact and heterozygous deleted PTEN loci is highly suggestive that its expression may be tightly regulated and that a variety of additional mechanisms are capable of

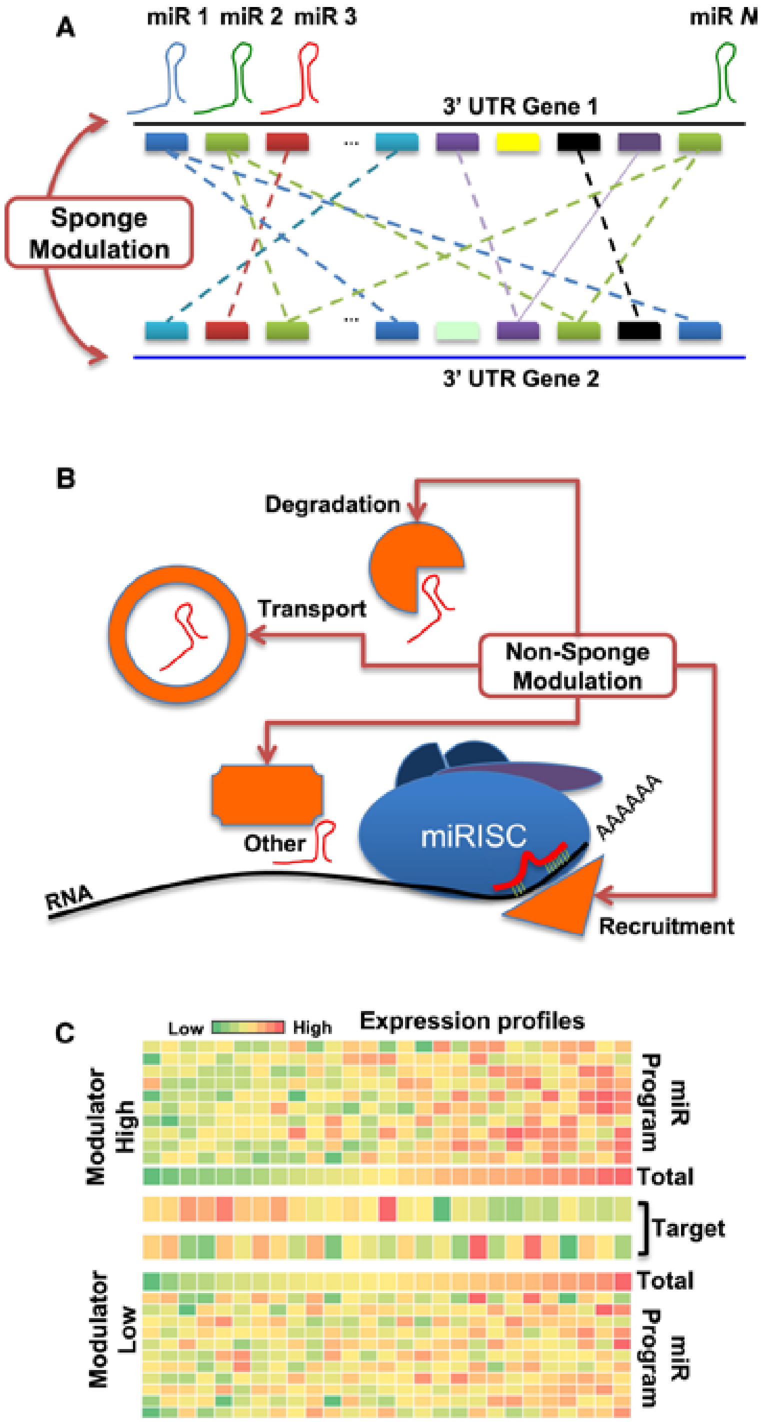


Figure 2. *miR Activity Modulators*: miR activity modulation may be implemented by several distinct mechanisms. We consider competition by RNAs for common miR programs (sponge effect) separately from other mechanisms, such as those driven by protein-protein or protein-miR interactions. (A) RNAs modulate each other through their common miR regulatory program. Up/down changes to the expression of one RNA perturb the relative abundance of functioning miRs that target both RNAs, leading to a corresponding up/downregulation of the second RNA. (B) Nonsponge modulators regulate miR activity by assisting or inhibiting components of the miR-mediated posttranscriptional regulatory apparatus. These regulators may help or prevent recruitment of miRISC to the target RNA or affect target degradation and transport. (C) To identify candidate modulators, we sought out instances in which the correlation between the total expression of a miR program and its target is dependent on the expression of a candidate modulator. This image visualizes a simplification of the process. The top heatmap shows expression of miRs in a program (rows) across all samples (columns) in which the modulator expression is high, with the bottom line showing the total expression of the miR program in the sample. Samples are sorted low to high based on miR program expression. Below that is the expression of the target of the miR program. The top heatmap shows strong inverse correlation between miR-program expression and target expression, consistent with an active miR program. The bottom heatmap shows the same data but this time for samples in which modulator expression is low. Here, the negative correlation between miR program expression and target expression is reduced, which is indicative of a suppressed miR program.

down regulating this gene in tumors.

Studies show that PTEN is controlled by miRs [12], [13]. Interestingly, our computational analysis indicates that PTEN participates in a total of 534 interactions in the mPR network and that its expression is regulated by the sponge effect, potentially mediated by a large number of regulators.

We experimentally manipulated a subset of these putative PTEN mPR regulators, whose loci are often deleted in patient tumors with an intact PTEN locus (Figure 3). Decreasing their mRNA levels by siRNA in cell lines that serve as models for glioblastoma showed a striking decrease in expression of a PTEN 3'UTR luciferase reporter. Additionally this regulatory mechanism was symmetrical; expression of the PTEN 3'UTR alone induced an increase in the expression of the majority of regulators. Furthermore these expression and silencing experiments had effects on the proliferation of the glioblastoma cell lines. Transfection of PTEN 3'UTR upregulated the expression of its mPR neighbors, increased PTEN (protein) concentration, and reduced tumor cell growth rates. Conversely, siRNA-mediated silencing of mPR regulators reduced PTEN 3'UTR-luciferase expression and significantly accelerated SNB19 and SF188 cell growth, respectively.

Thus in analyzing other genes that steer the progression of GBM tumors, 13 are commonly deleted in GBM and collaborate through miRs to disturb PTEN expression and functionality. This proves to be as deleterious as tumors that have DNA mutations in the PTEN gene. Moreover, this also partly accounts for the genetic variability among GBM patients where a large percent exhibit PTEN negative tumors, while others show an intact PTEN but without expression of the gene. The 13 genes that we found in the mpR regulatory network of PTEN could be, in part, the cause of PTEN suppression (Figure 3).

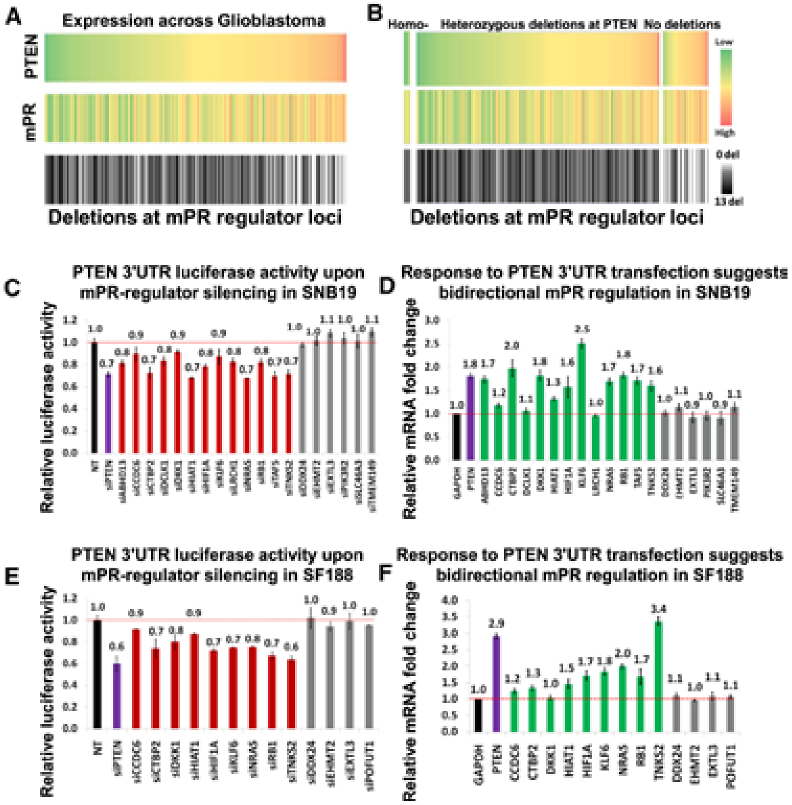


Figure 3. *PTEN Expression Is Correlated with the Expression of Its mPR Regulators*.

(A) PTEN is targeted by > 500 mPR regulators, and its expression is correlated with both their total gene expression and with deletions at their loci; in aggregate, 97% of the TCGA glioma tumors have at least one deletion in a PTEN mPR regulator locus. We selected 13 mPR regulators of PTEN with enriched locus deletions in PTEN intact tumors. As shown, their collective deletions and total expression are both significantly correlated with PTEN expression (pD < 2 3 10_10 and pE < 5 3 10_23, respectively).

(B) Surprisingly, the correlation between PTEN and the aggregate expression across the 13 genes is significant in both samples with an intact PTEN locus and samples with heterozygous deletions (rD = 0.40, pD < 10_9 and rWT = 0.46, pWT < 4 3 10_4 by Pearson correlation, respectively). The range of PTEN expression in PTEN heterozygously deleted samples and in samples with an intact PTEN locus was virtually the same.

(C) Individual siRNA-mediated silencing of 13 PTEN mPR regulators reduced PTEN 3'UTR luciferase activity in SNB19 cells at 24 hr. Negative control targets (in gray) were unaffected.

(D) Ectopic expression of PTEN 3'UTR increased expression of 13 PTEN mPR regulators in SNB19 cells at 24 hr, compared to empty vector. Negative control targets (in gray) were unaffected.

(E and F) Results in SNB19 were replicated in SNF188 cells for genes that are expressed in this cell line. Fold change was measured by qRT-PCR. Data are represented as mean ± SEM.

The strongly indicates that there are a large number of genes, which are affected by miR activity modulators that have a prominent significance in GBM.

Other established genes promoting tumorigenesis and glioma subtype that were tested included PDGFRA, RB1, VEGFA, STAT3, and RUNX1, which together form a dense subgraph of mutually mPR interacting genes. These genes have numerous miRNAs in common and in particular, loss of either PTEN or RB1 suggests a dramatic cross talk involving a shared miRNA/mRNA subnetwork. Hermes was also able to identify 148 nonsponge miRNA modulators. Amongst these we experimentally confirmed that several predicted regulators mediated effects on PTEN and RUNX1.

Taken together, these experimental results confirmed the validity of the software-model-predicted mPR network and are highly suggestive of a miR network mediating interactions between established oncogenic pathways. Moreover, they provide a mechanistic explanation for the loss of PTEN expression.

REFERENCES

1. Carro, M.S., et al., The transcriptional network for mesenchymal transformation of brain tumours. Nature, 2010. 463(7279): p. 318-25.

2. Franco-Zorrilla, J.M., et al., Target mimicry provides a new mechanism for regulation of microRNA activity. Nat Genet, 2007. 39(8): p. 1033-7.

3. Poliseno, L., et al., A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature, 2010. 465(7301): p. 1033-8.

4. Krol, J., I. Loedige, and W. Filipowicz, The widespread regulation of microRNA biogenesis, function and decay. Nat Rev Genet, 2010. 11(9): p. 597-610.

5. Garzon, R., G.A. Calin, and C.M. Croce, MicroRNAs in Cancer. Annu Rev Med, 2009. 60: p. 167-79.

6. Ebert, M.S., J.R. Neilson, and P.A. Sharp, MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells. Nat Methods, 2007. 4(9): p. 721-6.

7. Ebert, M.S. and P.A. Sharp, Emerging roles for natural microRNA sponges. Curr Biol, 2010. 20(19): p. R858-61.

8. Wang, K., et al., Genome-wide identification of post-translational modulators of transcription factor activity in human B cells. Nat Biotechnol, 2009. 27(9): p. 829-39.

9. Sumazin, P., et al., An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma. Cell, 2011. 147(2): p. 370-81.

10. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature, 2008. 455(7216): p. 1061-8.

11. Sana, J., et al., MicroRNAs and glioblastoma: roles in core signalling pathways and potential clinical implications. J Cell Mol Med, 2011. 15(8): p. 1636-44.

12. Verhaak, R.G., et al., Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell, 2010. 17(1): p. 98-110.

13. Kim, H., et al., Integrative genome analysis reveals an oncomir/oncogene cluster regulating glioblastoma survivorship. Proc Natl Acad Sci U S A, 2010. 107(5): p. 2183-8.

HOX PROTEINS REVEAL THEIR SECRETS

RICHARD MANN, HARMEN BUSSEMAKER, AND BARRY HONIG LABS

Hox proteins have very similar sequence preferences when binding DNA as a monomer. Mutations in the corresponding HOX genes however confer dramatically different phenotypes. MAGNet investigators Richard Mann, Harmen Bussemaker, and Barry Honig set out to address this paradox using high-throughput sequencing. They developed an integrated experimental and computational approach – named SELEX-seq – that can be used to determine binding affinities for all possible DNA sequences for any transcription factor complex. Applying this method to each of the eight *Drosophila* Hox proteins in complex with another homeodomain protein named Exd, they showed that this co-factor evokes striking differences in DNA binding preference between the eight Hox-Exd heterodimers (Figure 1). Differences in minor groove shape in the center of the binding seem to contribute to this "latent specificity" of Hox proteins.

REFERENCES

M. Slattery, T. Riley, P. Liu, N. Abe, P. Gomez-Alcala, I. Dror, T. Zhou, R. Rohs, B. Honig, H.J. Bussemaker, R.S. Mann (2011), Cofactor Binding Evokes Latent Differences in DNA Binding Specificity between Hox Proteins, Cell 147:6: 1270-1282

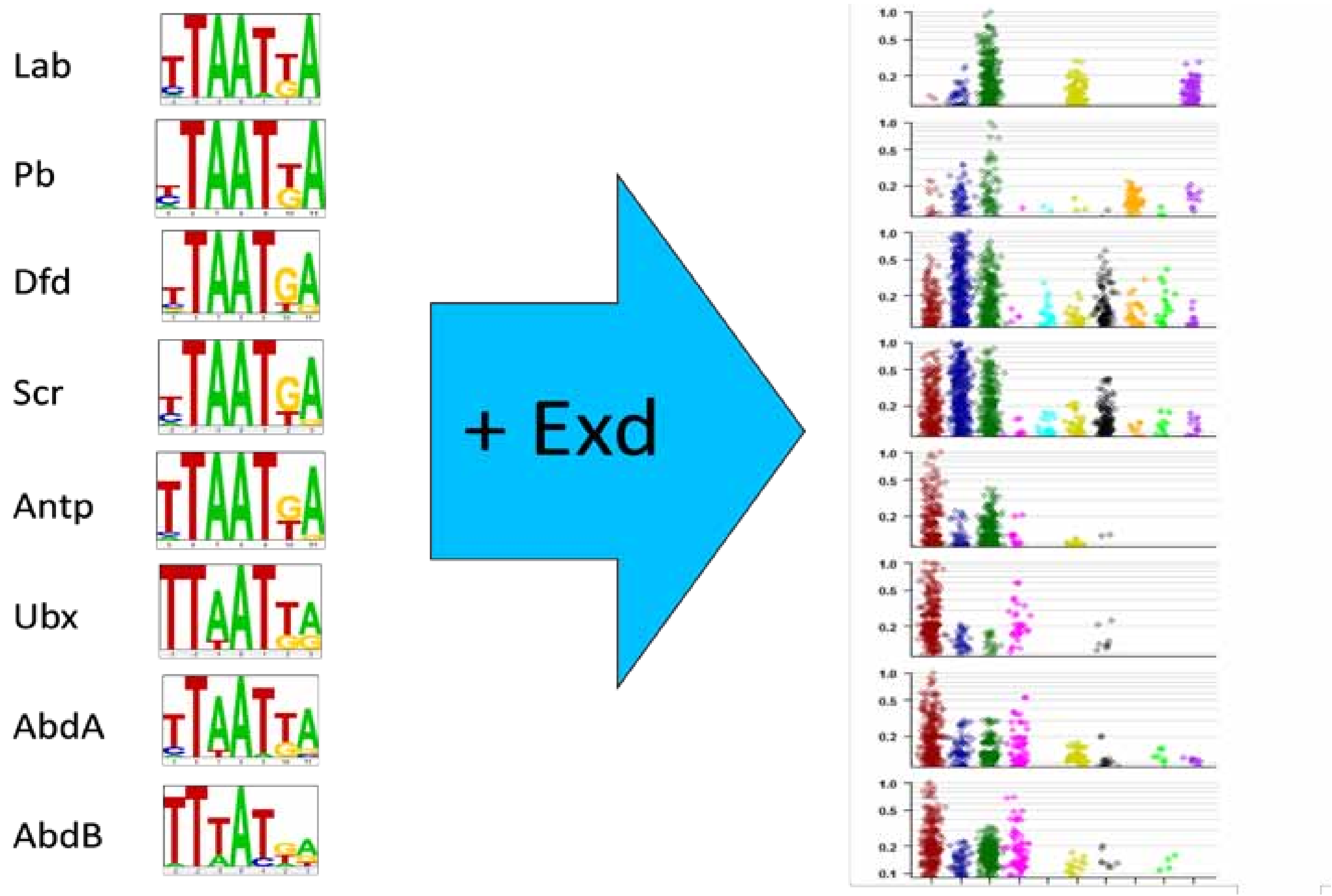


Figure 1: Dimerization with Extradentical (Exd) reveals latent DNA binding specificities across the conserved family of Hox transcription factors in Drosophila. The sequence logos on the left are from a study by Noyes et al. (2008); the "specificity fingerprints" on the right are from Slattery, Riley, Liu et al. (2011).

INFERENCE OF MODULES REGULATED BY EQTLS

ITSIK PE'ER AND DANA PE'ER LABS

Cataloging the association of transcripts to genetic variants can offer significant insights to the regulatory structure of human transcription. To find such relationships we have developed a novel approach, which entails detection and analysis of modules of transcripts, each co-associated with a single genetic variant. First, we search pair-wise connections between transcripts whose levels are co-associated with the same SNP. Second, we combine these pairs into modules that share an associated main SNP. We then assign a confidence score to each module. Finally, we find secondary SNPs whose association to transcript levels in a module is conditioned on the main SNP. We applied our method to existing data on genetics of gene expression in the liver. The modules we discover are significantly more, larger and denser than those found in permuted data (Figure 2). We quantify the confidence in a module as a likelihood score, and prune a subset of 95 distinct modules with FDR<0.02. We systematically look for cis effects that can explain multiple reported modules. We observe similar annotations of modules from two sources of information: the enrichment of a module in gene subsets and locus annotation of the genetic variants. This and further phenotypic analysis provide a validation for our methodology.

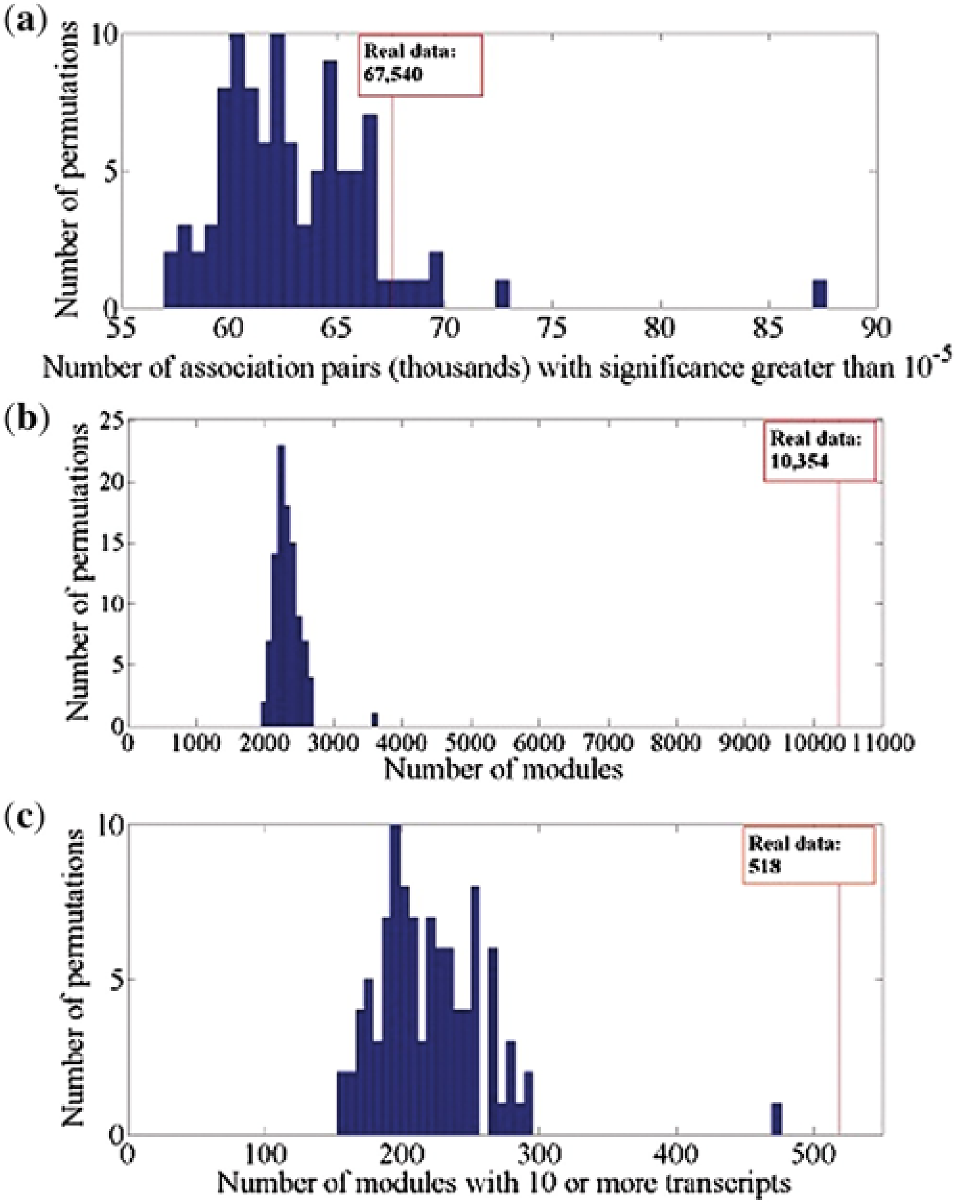


Figure 2: The number of (a) association pairs (b) modules and (c) large modules in real data compared with 100 permuted data sets. Although only 93 out of the 100 permuted data sets have fewer association pairs than in the real data, all of them have fewer (large) modules.

VARIANTS IN EXONS AND IN TRANSCRIPTION FACTORS AFFECT GENE EXPRESSION IN TRANS

ITSIK PE'ER LAB

In recent years many genetic variants (eSNPs) have been found to be associated with gene expression. However, the causal variants and the regulatory mechanisms by which they act remain mostly unknown. Here we present a comprehensive analysis of trans-eSNPs, integrating SNPs that are fully ascertained from genomic sequencing data with transcriptional profiling (RNA-seq) in the same cohort. When considering interpretable genomic regions containing candidate eSNPs, we observe enrichment of such variants in exons. We thus focus on exonic eSNPs, and consider eSNPs within the span of Transcription Factors (TFs) for comparison. In both cases, these variants define the spanning source gene, along with the respective gene target of association. We map the source and target genes onto a Protein-Protein Interaction (PPI) network and study their topological properties.

When considering pairs of eSNP exon source with its corresponding target, the stronger their association, the closer they are within the PPI network (permutation p<9.9e-4) and the higher the degree of the target (permutation p<0.002). Expression analysis demonstrates that these source–target pairs are more likely to be co-expressed (p<5.4e-5) and the eSNP tends to have a cis effect, modulating the expression of the source gene (p<2.3e-13). In contrast, source-target pairs with a TF eSNP are not observed to have such properties. We do observe these latter pairs to reside within the same PPI cluster more than expected by chance (permutation p<0.0043), and to assemble functionally enriched units of a TF source along with its gene targets.

Our results suggest two modes of trans regulation: TF variation frequently acts via a modular regulation mechanism, with multiple targets that share a function with the TF source. Notwithstanding, exon variation often acts by a local cis effect, and propagates through shorter paths of interacting proteins across functional clusters of the PPI network.

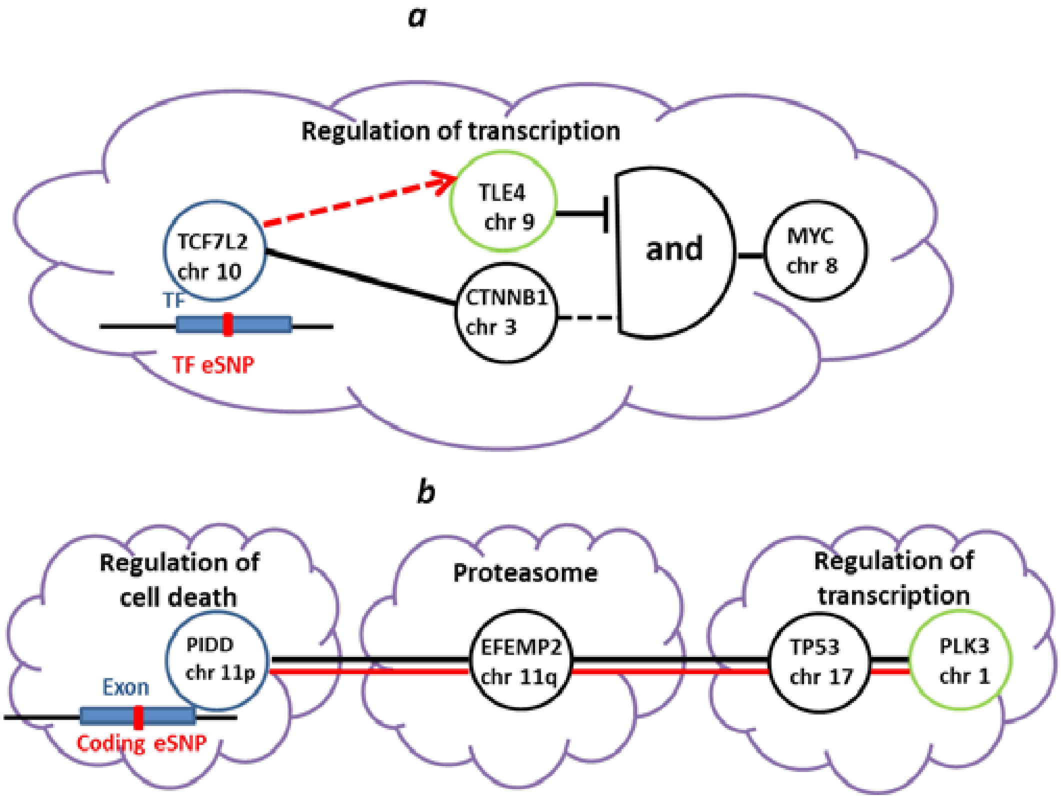


Figure 3: *Examples of TF and exon source-target pairs*. TF/exon source and target genes are denoted by blue and green circles respectively. The PPI cluster is denoted by a purple cloud. The genomic location of the TF/exon is denoted by the blue rectangle, and the eSNP associated to the target genes is marked in red. PPI edges are denoted by black solid lines and nodes in the PPI are denoted by black circles. Exonic eSNP interaction is denoted by a solid red line and TF eSNP interaction is denoted by a red dashed line. (a) Network motif I1-FFL: TCF7L2 activates MYC (in presence of CTNNB1) but also represses MYC by activating the repressor TLE4 (via an eSNP). (b) The shortest path on the PPI network between PIDD source and its gene target PLK3. There is a significant correlation between the expression of the source and target genes.

INTERROGATING MOUSE AND HUMAN PROSTATE CANCER INTERACTOMES TO UNDERSTAND THERAPEUTIC RESPONSE

CORY ABATE-SHEN, MICHAEL SHEN, AND ANDREA CALIFANO LABS

Last year we introduced in this newsletter a new project in which we have been developing mouse and human prostate cancer interactomes that have enabled the first cross-species integration of master regulators of therapeutic response for prostate cancer. The generation of the mouse prostate cancer interactome is particularly novel since it was developed in vivo from 13 different and distinct strains of mice, each treated with 14 distinct pharmacological perturbations using drugs that are relevant for prostate cancer. The human interactome was developed from a published data set with extensive clinical outcome data. To effectively link analyses of the mouse interactome, which was built on pharmacological perturbations, with that of the human interactome, which was built on clinical specimens, we are in the process of developing a human xenograft interactome that will be perturbed by the same pharmacological agents as we had used for the mouse interactome.

We have performed cross species interrogation of the mouse and human interactome to evaluate the therapeutic response to drug treatment for alleviating tumor and metastatic burden in a mouse model of metastatic prostate cancer. Importantly the cross-species verification has led to identification of two master regulators, namely FOXM1 and CENP-F, which synergize to promote progressive metastatic prostate cancer and are targeted for drugs that inhibit the metastatic phenotype. FOXM1 and CENP-F, which were identified exclusively based on these computational cross-species analyses, are expressed robustly in advanced prostate tumors and metastases and are predictive of clinical outcome. These findings establish a new paradigm for identification of master regulators of drug response based on cross-species interrogation of mouse and human regulatory networks with preclinical data from mouse models.

MICROWELL ARRAYS FOR SINGLE CELL SYSTEMS BIOLOGY

PETER SIMS LAB

Microfluidic devices fabricated in polydimethylsiloxane (PDMS) using soft lithography have enabled numerous biological applications over the last decade. Simple features like pumps, valves, chambers, and channels can be integrated into complex devices to facilitate high-throughput experiments in small volumes. We are developing a new tool for single cell transcriptomics that capitalizes on the unique properties of one such microfluidic feature - the PDMS microwell array. In our latest experiments, we are depositing individual cells in micrometer-scale wells arranged in a large array. PDMS is transparent to visible light, and so we can immediately apply conventional phenotypic analysis using immunofluorescence or reporter assays on a microscope. Furthermore, the microwell array can be reversibly sealed against a flat surface, isolating its contents in a few picoliters. By functionalizing the flat surface, we can capture RNA from the lysates of individual cells.

Coupling microfluidic RNA capture with high-sensitivity fluorescence microscopy will allow detection and quantification of the captured transcripts using fluorescent probes, sequencing, or PCR. Both digital PCR and multiplex sequencing have recently been demonstrated using the reversibly sealable PDMS microwell platform (Men, Fu, Chen, Sims, Greenleaf, Huang, Analytical Chemistry, 2012; and Sims, Greenleaf, Duan, Xie, Nature Methods, 2011). Going forward, our ultimate goal is a system that directly links phenotypic characterization by optical imaging with genome-wide expression profiling for thousands of individual cells in parallel. Not only will this highly scalable technology provide the statistical power needed to dissect and relate phenotypic and transcriptional subpopulations, it could eventually form the basis of an inexpensive diagnostic platform that is immune to compositional heterogeneity.

INFLUENZA HOST CLASSIFICATION

CHRIS WIGGINS LAB

MKBoost, an adaboost classifier with a mismatch k-mer feature space, was used to build predictive classifier of influenza host organism in order to identify sequence elements important in host adaptation. The classifier was built using publicly available sequence data from the NCBI influenza virus resource, trained on hemagglutinin (HA) sequences from all influenza subtypes, initially restricted to avian and human host isolates. The HA protein was chosen because of its known role in viral host specificity, primarily in binding to either alpha(2,3)-linked (avian) or alpha(2,6)-linked (human) sialic acids.

After training, the algorithm yielded high (>98% AUC) classification accuracy within 4 to 5 rounds of boosting. Following repeated cross-validation folds, selected k-mers were ranked based on their harmonic mean round of appearance, a test of selection robustness, under the assumption that k-mers robustly selected earlier in boosting carry more biological importance. When these selected k-mers were mapped onto a reference hemagglutinin sequence (Figure 4), known host-specific features in the receptor binding region were recovered, specifically those in the 190-helix and the 220-loop. Additionally, there is significant overlap with those mutations identified in (Imai 2012). This validates that the classifier is selecting regions biologically relevant for host adaptation. Those regions outside of the known domains will be selected for experimental validation studies.

Future work will use this method to identify motifs relevant for host adaptation in the internal genome segments, specifically those that have been implicated in host-specific interactions and virulence, including the polymerase proteins PB1 and PB2, and the nonstructural NS1 protein.

REFERENCES

Imai, M., Watanabe, T., Hatta, M., Das, S. C., Ozawa, M., Shinya, K., Zhong, G., et al. (2012). Experimental adaptation of an influenza H5 HA confers respiratory droplet transmission to a reassortant H5 HA/H1N1 virus in ferrets. Nature, Đ. doi:10.1038/nature10831

SUPERVISED LEARNING OF VIRAL ONCOGENICITY

CHRIS WIGGINS LAB

There is a great wealth of information contained in genomic sequence data, and advances in sequencing technology have only added to the massive troves of such data. It is now possible to tackle difficult questions relating genotype and phenotype of various organisms in a data driven fashion. When attempting to link genomic motifs to observed organism-level function, the ideal organism would be nothing more than a complex of its gene-products, essentially a virus. This is a key reason for targeting problems specifically within the sub-domain of viral sequence data, the relatively direct mapping of genotype to phenotype. There is already a short list of consensus oncogenic viruses in the medical literature, and the question of which genomic features predict tumor growth can and should be cast as a computational task leveraging the abundance of sequence data for these very oncoviruses.

Previous work within the Wiggins group has established an algorithm, MKBoost, which learns boosted decision tree classifiers for sequence data. The sequences are represented as collections of k-mers and the discriminative rules that build our classifiers are the presence/absence of specific k-mers.

Current work is focused on amino acid sequences from the proteins of human papilloma viruses (HPV). Protein sequences are collected from NCBI's viral genome resources and annotated with a binary label representing whether or not the particular HPV subtype of the isolate is known to be associated with human malignancy. Decision trees trained on this labeled data achieve ~90% AUC at 10 rounds of boosting. Beyond good classification performance though, it is of interest whether the k-mer rules selected for the trees are robust across different cross validation folds. Indeed the highest selected motifs consistently fell in either L1 (capsid protein) or E6 (canonical oncoprotein).

Interesting next steps include training only on E6 sequences to learn specific regions of the protein which contain predictive k-mers, and also applying these high performance classifiers to non-HPV sequences with unclear oncogenic potential, such as human adenoviruses.

CLUES TO THE GENETICS ROOTS OF AUTISM

DENNIS VITKUP LAB

The genetic architecture of autism is turning out to be even more complex than the disease's diverse clinical manifestations. Large genetic studies have ruled out the hypothesis that autism is due to genetic malfunctions in a single gene or a small core set of genes. Instead, there is growing consensus that genetic mutations in many hundreds of genes contribute to autism spectrum disorders. To understand molecular networks underlying the autistic phenotype, Dennis Vitkup's lab has developed NETBAG, a novel method for network-based analysis of genetic associations. NETBAG was used to identify a large biological network of genes affected by rare de-novo copy number variants (CNVs) associated with autism (Gilman 2011). The genes forming the identified network are primarily related to synapse development, axon targeting, and neuron motility (Figure 5). The network

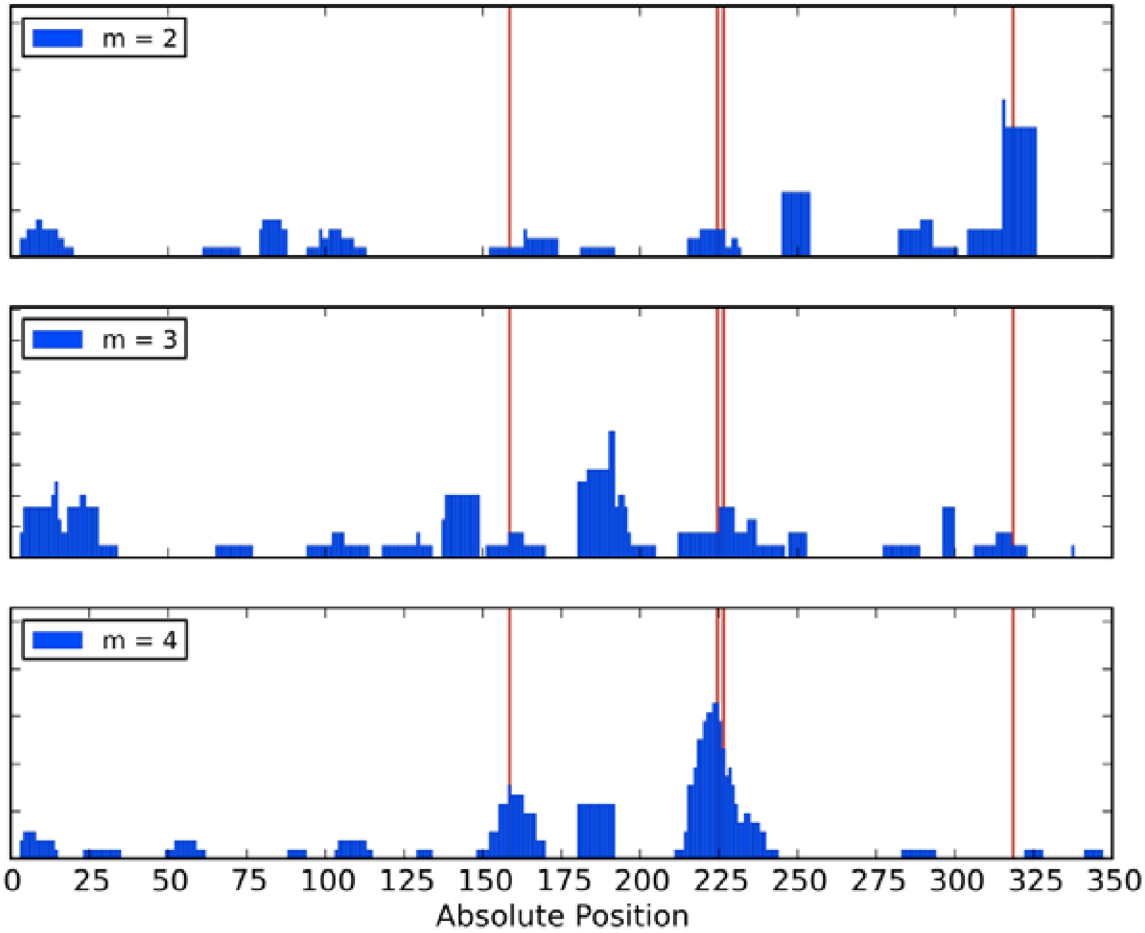


Figure 4: Histogram showing selected motifs mapped onto a reference H5 type hemagglutinin (PDBID: 2IBX) for mismatch m=2,3,4. The weight at each residue reflects the robustness of selection of motifs at that site under 5-fold cross validation. Host-adaptive mutations identified in (Imai 2012) are highlighted in red.

is also strongly related to genes previously implicated in autism and intellectual disability phenotypes. Overall, the analysis of de-novo variants supports the hypothesis that perturbed synaptogenesis lies at the heart of autism. The results of the study are also consistent with the hypothesis that significantly stronger functional perturbations are required to trigger the autistic phenotype in females compared to males. More generally, the study provides proof of the principle that networks underlying complex human phenotypes can be identified by a network-based analysis of relevant rare genetic variants..

REFERENCES

Gilman SR, Iossifov I, Levy D, Ronemus M, Wigler M, Vitkup D. (2011), Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses, Neuron, 70(5):898-907.

WEB-ENABLED ACCESS TO MAGNET TOOLS

ARIS FLORATOS LAB

geWorkbench (http://www.geworkbench.org/), the bioinformatics platform of the MAGNet Center, is an open source Java application that provides access to the computational tools and data resources of the MAGNet Center as well as to 3rd party analysis and visualization modules for a wide range of genomics domains (Floratos 2010). A key goal of geWorkbench is to make it easier for non-computational biologists to leverage the power of advanced software tools such as those developed in the MAGNet laboratories. To that end, geWorkbench offers a uniform graphical user interface that provides integrated access to more than 70 modules, thus eliminating the need to individually download and deploy a large number of tools.

In the past geWorkbench was available only as a thick Java client, delivered through a self-extracting installer. One of the benefits of the thick client model is that it allows leveraging the rich visualization and interactivity capabilities of the Swing framework in Java. We have now complemented the Java client with geWorkbench-Web, a web-enabled version of the application that allows accessing many of the available components through a browser. While the web interface lacks some of the more advanced interactive features of the thick client, it offers an attractive alternative for instant and installation-free access to the geWorkbench computational services, leaving the thick client as an option for power users that wish to take advantage of the more advanced Swing features. A pilot version of geWorkbench-Web with a small number of modules is currently under testing. A production release, comprising many of the components in the Java version, is planned for the end of 2012.

REFERENCES

Floratos A, Smith K, Ji Z, Watkinson J, Califano A. (2010), geWorkbench: an open source platform for integrative genomics, Bioinformatics, 26(14):1779-80. Epub 2010 May 28.

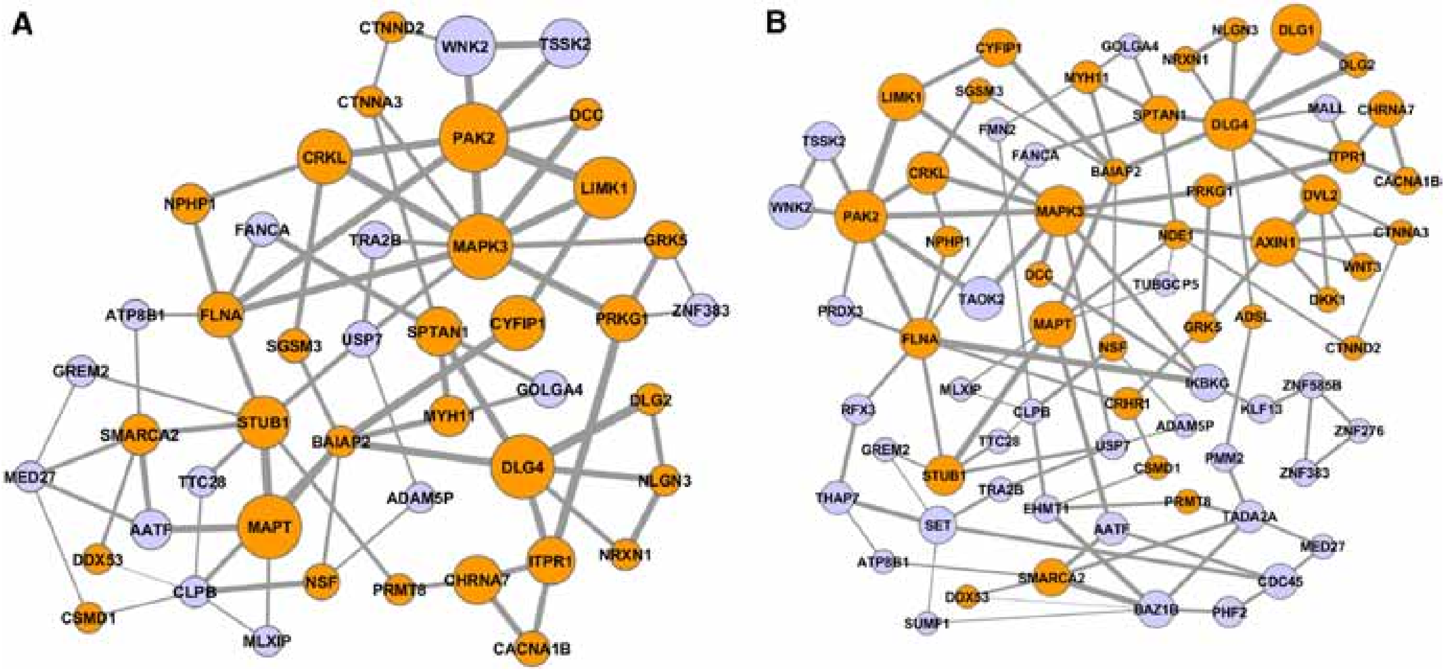
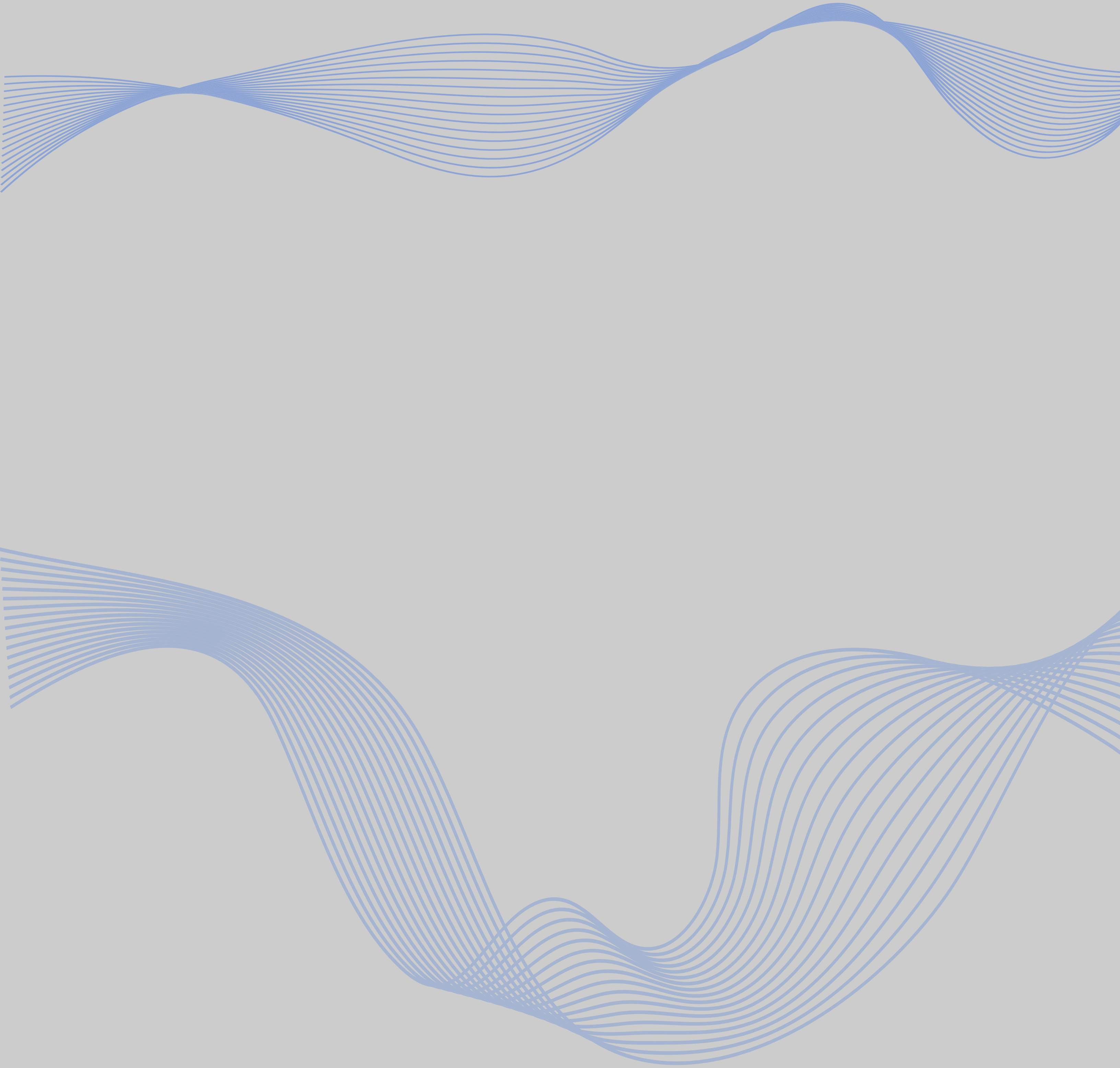


Figure 5: Gene Clusters Found using NETBAG analysis of De Novo CNV regions observed in autistic individuals(A) Highest scoring cluster obtained using the search procedure with up to one gene per each CNV region.(B) Cluster obtained using the search with up to two genes per region. Genes (nodes) with known functions in the brain and nervous systems are colored in orange. Node sizes represent the importance of each gene to the overall cluster score. Edge widths are proportional to the prior likelihood that the two corresponding genes contribute to a shared genetic phenotype.



Columbia University
Center for Computational Biology and Bioinformatics
1130 St. Nicholas Avenue
New York, NY, 10032

SPRING 2012
Issue No. 5