MicroRNA Expression Regulation in the Context of Cell Growth Stimulation
**Background**

Despite years of intensive cancer research, malignancies have been the second most common cause of death in the U.S. in 2007. In the quest for new drug targets, one class of molecules has lately risen to high prominence: microRNAs (miRNAs).

MicroRNAs are small, non-coding, 18-22 bp long RNA single stranded regulatory molecules that were first described in 1993 in *C. elegans* and play a key role in cell development, proliferation and differentiation through down-regulation of target mRNA-translation. Due to their role as developmental regulators, several authors have highlighted the importance of miRNAs in carcinogenesis, identifying miRNA expression signatures unique to the respective tumors. Even more importantly, certain expression patterns seem to correlate with prognosis, including therapeutic outcome. However, very little is known about miRNA regulation itself. Cellular growth and proliferation are extrinsically controlled by so-called growth factors. Some of these are potent mitogens, like epithelial growth factor (EGF), platelet-derived growth factor (PDGF) or transforming growth factor (TGF-β), whereas others enhance angiogenesis, such as vascular endothelial growth factor (VEGF) or fibroblast growth factor (FGF). The fact that all of these factors have been associated with human cancer development emphasizes the importance of the aforementioned signaling pathways.

Recently, a transcription factor of the Smad family has been identified as a bridge linking TGF-β signaling to the miRNA “miR-24”-controlled cell differentiation. However, current understanding of miRNA regulation through transcription factors is limited to small scale analyses and very little is known about how external growth signals influence miRNA expression patterns.

Thus, we propose to systematically study the effects of transcription factor and external growth factor stimulation on miRNA expression by integrating data from biological and computational experiments.

**Aim 1: Identification of growth factor specific microRNA profiles**

There is rich evidence in the literature demonstrating a role for miRNAs in the regulation of cell growth, differentiation and proliferation. However, very little is known about how miRNAs are themselves regulated. There is some preliminary evidence that external growth signaling may affect miRNA regulatory activity, but further experiments are required.

Theoretically, the cell could modulate microRNA activity through multiple mechanisms. First, miRNA expression levels could be altered, for example through regulatory elements...
like transcription factors. Sun and his colleagues showed, for example, how TGF-β activates proteins of the Smad family that in turn down-regulate miR-24 expression\textsuperscript{12}. Second, miRNA biogenesis could be modified. Biologically active miRNAs are ordinarily processed from so-called primary microRNAs (pri-miRNAs) containing a hairpin structure, which itself is known as precursor miRNA (pre-miRNA). After excision of the hairpin element and export out of the cell’s nucleus, further processing removes the hairpin section to produce the mature miRNA. Third, miRNA target recognition and degradation could be modified through post-transcriptional modification of the miRNA itself, its targeted mRNAs, or post-translational modifications of proteins (RNA-induced silencing complex, RISC) involved in the degradation or translational repression of targeted mRNA. 

Our approach will focus on alteration of miRNA expression. We hypothesize that it is possible to extract signatures specific to the growth signaling pathways of EGF, VEGF, PDGF, FGF and TGF-β. All of these have been shown to have pro-proliferative activity and their dysregulation has been associated with cancer pathogenesis\textsuperscript{8,11}.

We suggest using HeLa cell lines that will be exposed to these stimulus conditions. Additionally, we will include a non-stimulated negative control. In order to determine expression levels, we will use the newly evolved ‘Deep Sequencing’ methods, such as the Illumina/Solexa or Roche/454 system. Although to the best of our knowledge there is no comprehensive study showing superiority over microArrays in terms of sensitivity and specificity in expression analysis of miRNAs, vendors promise robust and high quality detection of low-abundance RNA molecules (www.illumina.com).

Different techniques will allow us to identify clusters of miRNAs specific to particular pathways. First, supervised learning methods like logistic regression, used in conjunction with variable selection methods, have the potential to identify significant predictors of a pathway. The performance of the model can be evaluated using the area under a Receiver Operator Characteristics (ROC) curve\textsuperscript{14}. Furthermore, we can determine statistical significance of each predictor included in our model. Second, unsupervised learning methods like hierarchical clustering allow us to group miRNAs into clusters that have similar variation. Caveats of these latter methods involve the determination of the optimal number of clusters, which can be assessed using Gap statistics\textsuperscript{15}, and the need to label the clusters, which is usually done by experts based on gene ontology pathway annotation of corresponding cluster members.
We will compare both techniques and try to combine the results to extract a robust miRNA expression signature. Thus, we will use both statistical and biological methods to determine the degree of miRNA expression alteration under the influence of growth factor stimulation.

**Aim 2: Determining the exact position of miRNA start of transcription**

Whereas in the past eight years attention was mainly drawn to target recognition and miRNA biogenesis\(^{16}\), focus starts shifting towards miRNA regulation itself\(^{17,18}\). Out of the ≈ 500 known human miRNAs, about 50% are located in intronic regions and are therefore believed to be regulated by the promoter regions of the respective host gene\(^{19}\). The remaining 50% map to intergenic regions and there is strong evidence that most of them are transcribed as independent transcriptional units with their own promoters. RNA polymerase II is believed to produce most of the primary miRNA transcripts with a unique 5’-methylguanosine cap and a 3’ polyadenylated tail\(^{20}\). However, very few promoter regions so far have been biologically validated\(^{12,21,22}\) and prediction of transcription factor binding sites (TFBSs) is limited by incomplete knowledge about biochemical processes that drive protein DNA interaction. Saini et al., for example, used a software package called Eponine\(^{23}\) to predict miRNA transcription start sites (TSSs) at a 1 kb resolution and compared the predictions to existing EST, ditag, CpG-island and TFBS information, identifying putative promoter regions and potential transcription factor binding motifs\(^{18}\). Fujita et al., in contrast, searched a region 100kb upstream of known miRNA hairpin structures of cross-species conserved miRNAs to identify regions of high conservation and enrichment for TFBS motifs. Out of the 59 regions they identified, they were able to validate 6\(^{17}\).

The above approaches show a diversity that reflects our limited knowledge of the exact location of where miRNA transcription starts. In theory, it should be possible to identify pri-miRNA transcripts in nuclear cell lysates, map them to the genome, and identify the exact TSS. However, due to limitations of large-scale detection of low-abundance RNA sequences, very few pri-miRNA sequences are known\(^{20,21,24}\). We propose to combine RLM-RACE PCR with deep-sequencing technology to detect low abundance pri-miRNA structures and determine the exact position of the TSS in the human genome.

We will use a mixture of nuclear fractions of all differently stimulated cells described in aim 1. Different stimulus conditions should lead to different expression patterns and hence to increased total amount of miRNA biogenesis products. We will use RNA ligase-mediated rapid amplification of 5’ cDNA ends (RLM-RACE) as described in Michalosky et al.\(^{25}\) to amplify 5’-methylguanosine capped RNA molecules. In short, an initial PCR reaction with a
universal 5’ primer (e.g. GeneRacer) is performed with primers matching to as many pre-miRNA genes as possible. To design the primers, we will use the software Prime 3, described in Rozen et al.\textsuperscript{26} Next, we will sequence the products, using a deep-sequencing platform. These systems are able to produce a high number of reads and should therefore be able to detect low-abundance RNA molecules. By having complete sequence information, we can also detect miRNAs at different stages of their biogenesis and hence convert findings into a probabilistic model\textsuperscript{27}. We will then use BLAST\textsuperscript{28} to map the identified sequences to the human genome and localize the exact position of the TSS to help defining miRNA promoter regions.

It will be important to check our results for consistency with previously published studies\textsuperscript{20,21,24}. Furthermore, TSSs for intronic miRNAs should match to annotated TFBSs of the corresponding host genes, which we will check using the UCSC genome browser\textsuperscript{29}. We will also calculate the odds ratio of CG dinucleotide content in the upstream regions of the promoter regions\textsuperscript{30} over the CG content of a randomly selected cohort of regions in the human genome of similar size. We furthermore will use USCS’s BLASTZ alignment-tool\textsuperscript{31} to assess conservation of these regions in other species.

In sum we propose to design a biotechnological experiment followed by a computational method to determine the exact promoter regions of all known human miRNAs.

**Aim 3: Identification of transcription factor binding sites**

Several reports have shown examples of vertebrate miRNA gene regulation by known transcription factors\textsuperscript{32}. However, genome-wide screens for miRNA regulating transcription factor binding sites have been limited by lack of well-annotated promoter regions and hence leading to decreased predictive accuracy. Fujita and colleagues for example considered a region of 100kb upstream of known miRNA hairpin structures to identify promoter regions, using phylogenetic footprinting and position weighted matrices (PWMs) for RNA polymerase II binding motifs\textsuperscript{17}. Saini et al. used a TSS prediction algorithm with 1kb accuracy to limit potential promoter regions to a 2kb window\textsuperscript{18}. We hypothesize that we might increase specificity of TFBS identification by employing annotation data from aim 2 and therefore we expect to be able to reduce our region of interest. We furthermore propose to integrate these findings with expression data derived from aim 1 to link transcription factor regulation to growth signaling.

First we will define the region that we will search as a 1 kb window upstream\textsuperscript{33} of the TSSs identified in aim 2. Then, we will scan these regions for transcription factor binding motifs
using two different methods. First, we will try to find TFBSs based on existing knowledge, using so-called position weight matrices (PWMs). A software interface like MatInspector\textsuperscript{34} calculates PWM probability scores based on TRANSFAC\textsuperscript{35}, the largest and most commonly used database. Second, we will make use of methods that extract commonly shared sequences in promoter regions of eukaryotes. Melina II\textsuperscript{36} combines multiple algorithms, like MEME\textsuperscript{37} and Gibbs recursive sampler\textsuperscript{38} and is therefore likely to outperform single-method approaches. We would expect to find co-regulated miRNAs to be members of the same expression clusters defined in aim 1.

A major limitation of most in silico TFBS predictions is the high false positive rate\textsuperscript{39}. The assessment of functional significance is an especially complex problem, which usually is counteracted by adding phylogenetic and biochemical layers of information. However, stringent cut-off scores will greatly reduce sensitivity. As our intention is to determine possible links to signaling pathways, we prefer to privilege sensitivity. We hope to be able to identify transcription factor binding motifs in the promoter regions of miRNAs specific to the respective growth factor pathways that resemble known pathways in the KEGG, BIOCARTA, and AMBION databases.

\textbf{Future Directions}

We are proposing a combination of computational and biological approaches to provide groundwork for high-level investigations of the linkages between growth factor regulation and miRNA expression. However, our approach targets only a subset of possible regulatory mechanisms, omitting important mechanisms such as post-translational modification or splicing. Further studies using labeled amino-acid mass spectrometry to determine phosphorylation status could yield insights to further advance our knowledge about the integration of external signals and miRNA-based translational regulation. It will also be important to assess biological relevance of our findings through experimental validation, such as gel-shift assays to capture proteins binding to our identified TFBSs, followed by mass spectrometry to determine the identity of these proteins. Despite these limitations, our work will hopefully help clarify the relationship between processes in cellular signaling. This could serve as the basis for finding new drug targets in dysregulative states such as cancer, which continues to be one of the most common causes of death.
