Original Article

A systems biological approach to identify key transcription factors and their genomic neighborhoods in human sarcomas

Antti Ylipää[1], Olli Yli-Harja[1], Wei Zhang[1,2], and Matti Nykter[1]

Abstract

Identification of genetic signatures is the main objective for many computational oncology studies. The signature usually consists of numerous genes that are differentially expressed between two clinically distinct groups of samples, such as tumor subtypes. Prospectively, many signatures have been found to generalize poorly to other datasets and, thus, have rarely been accepted into clinical use. Recognizing the limited success of traditionally generated signatures, we developed a systems biology-based framework for robust identification of key transcription factors and their genomic regulatory neighborhoods. Application of the framework to study the differences between gastrointestinal stromal tumor (GIST) and leiomyosarcoma (LMS) resulted in the identification of nine transcription factors (SRF, NKX2-5, CCDC6, LEF1, VDR, ZNF250, TRIM63, MAF, and MYC). Functional annotations of the obtained neighborhoods identified the biological processes which the key transcription factors regulate differently between the tumor types. Analyzing the differences in the expression patterns using our approach resulted in a more robust genetic signature and more biological insight into the diseases compared to a traditional genetic signature.

Key words Systems biology, transcription factor, gene regulation, binding motif, sarcoma

Introduction

Genetic diseases are attributed to changes in the expression pattern of one or more key genes. In cancer, abnormal expression of oncogenes and tumor suppressor genes is often caused by focal genetic events, including gene mutations and copy number changes, as well as epigenetic changes like altered promoter methylation[1-3]. Focal events are then followed by tumorigenic alterations in transcription programs, which further affect downstream gene expression patterns. A well known example of this trickle-down effect is the MDM2-p53 pathway. When MDM2, an inhibitor of the tumor suppressor gene p53, is deactivated by a single nucleotide polymorphism, p53 is overexpressed, resulting in aberrant expression of hundreds of downstream genes[4]. Thus, as gene expression is often controlled by hierarchical regulatory networks, experiments that measure expression patterns at steady-state levels, such as microarray, provide little insight into how gene expression is regulated.

Given the limitations of steady-state experiments, many genomic studies are conducted solely to identify genes that either are differentially expressed between two clinical groups, such as tumors with different clinical characteristics [5-19], or are correlated with clinical parameters like patient survival. The signature genes identified in these studies can be used, for example, as features in computational classifiers for discriminating patients as poor or long survivors, or as responders or non-responders to treatment. Nevertheless, although signature genes for some tumor subclasses have been successfully identified using computational methods, most signature gene lists have been found highly unstable and unreliable in prospective studies. Ill-conceived experiments, lack of clinical validation, biased selection of signature molecules, and overly positive error estimates for classifiers have been deemed downfalls of this approach in practice[20-25]. Further, global

Authors' Affiliations: [1]Department of Signal Processing, Tampere University of Technology, Tampere 33101, Finland; [2]Department of Pathology, the University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA.

Corresponding Author: Matti Nykter, Department of Signal Processing, Tampere University of Technology, Tampere 33101, Finland. Email: matti.nykter@tut.fi.

gene lists fail to shed light on molecular relationships and the biological difference that gene dysregulation conveys to the phenotype. An additional significant problem is that signature genes are often passive bystanders ("passenger genes") that do not have a major driving function ("driver genes") in the disease, such as promoting faster relapse or conveying resistance to treatment.

As a result of the shortcomings of gene-level analyses, there has been a transition towards module-level analysis of gene regulation. Several investigators propose that it is imperative to analyze smaller, coherent sets of genes instead of global signatures [26-32]. The coherence of a gene set can be defined without knowing the network topology by such parameters as shared biological function, molecule type or localization, such as in Gene Ontology [33]; or topologically, by genetic regulatory connections or other physical interactions. Gene regulatory networks include signaling pathways, metabolic pathways, and disease pathways and are made publicly available in several databases [34,35]. In oncology, pathways have replaced genes as primary building blocks after the observations that critical pathways, such as the Rb, WNT, PI3K/AKT, and p53 pathways[36], can be activated or deactivated by varied and mutually exclusive single gene mutations. Although we emphasize the gene set approach, information on just one or few dominant genes is sometimes sufficient to provide meaningful clinical information. There are a number of one-gene oncogenic events which define certain expression patterns and, thus, cancer subtypes [16,19,37,38]. Indeed, using a mouse model, a single oncogenic activation, such as elevated expression of PDGFB, suffices to induce brain tumor development[39].

Gastrointestinal stromal tumor (GIST) and leiomyosarcoma (LMS), two soft tissue sarcomas of the abdominal cavity, are morphologically remarkably similar, yet clinically and biologically very diverse. Clinicians primarily achieve correct diagnosis of GIST or LMS using histological examination and a number of immunohistochemical markers, including KIT, CD34, desmin, and smooth muscle actin [40]. Although some characteristic genetic markers of these tumors, such as gene mutations, have been known for a few years, the upstream causes and downstream effects of these signature mutations are not well understood. Of the greatest clinical relevance are gain-of-function mutations in *KIT* or *PDGFRA* genes in GIST. These mutations allow treatment with targeted therapy in patients with GIST but not LMS. Conversely, LMS is effectively treated with chemotherapy, whereas the objective response rate for GIST is negligible with this regimen[41-44]. To complement immunohistochemistry, computational classifiers, such as the *OBSCN/PRUNE2* gene classifier[45], have been previously devised for these tumors. However, like the immunohistochemical markers, the features in this genetic classifier have not been interpreted biologically, mainly due to the very limited knowledge of the marker genes. Although these malignancies are now recognized as two different tumor types as supported by our recent genomic characterization studies [46,47], they also share many common characteristics, like morphology and anatomical sites.

The unique historical and clinical relationship between GIST and LMS makes it interesting to compare the differences in their expression patterns. In this study, we developed a systems biological approach for identifying the genomic difference between tumors in more detail than can be attained with a straightforward list of signature genes. We applied this approach to investigate the key transcriptional regulators and their genomic neighborhoods that may cause the clinical differences between the two tumor types. The distinctive gene regulatory information that the results describe should prove more robust and biologically relevant than a signature gene list.

Materials and Methods

Gene expression measurements of GIST and LMS

We acquired 68 surgical specimens of primary tumors at the University of Texas MD Anderson Cancer Center under an Institutional Review Board-approved protocol with patient consent. Of these tumors, 37 were classified as GIST and 31 as LMS based on both clinicopathologic evaluation and molecular marker studies. Specifically, clinicopathologic observations included the site of the primary tumor, the pattern of metastatic spreading, and the efficacy of systemic therapy; molecular marker studies included immunostaining for KIT, CD34, desmin, and smooth muscle actin. All specimens were snap-frozen within 20 min of surgical resection and were verified by histopathologic examination to be composed of a minimum of 90% neoplastic cells. The gene expression profiles of these samples were measured with whole human genome oligo arrays with 44 000 60-mer probes (Agilent Technologies, Palo Alto, CA, USA) according to manufacturer's protocol. Arrays were scanned with Agilent's dual laser-based scanner and intensity values were read and processed with Agilent's Feature Extraction software version 8.0 with default parameters. The intensity values were quantile normalized in Matlab version R2009b (The MathWorks, Natick, MA, USA). All the data analysis was implemented in Matlab. The gene expression data are publicly available at http://www.cs.tut.fi/sgn/csb/GISTLMS/.

Finding master regulators

To gain confidence towards the generalization properties of the identified master regulators, we created 100 resampled[48] sets of data, each excluding a randomly chosen 15% of the total samples. We computed the differentially expressed genes for each of the sample sets by first applying the two-sided Wilcoxon rank sum test to compute a *P* value for differential expression for each gene. To correct for multiple comparisons, we computed false discovery rates (FDR) based on the *P* value distribution[49]. We used a *q* value of 0.005 as the threshold of significance.

Promoter analysis was applied to predict which transcription factors can regulate each gene by binding to its promoter region. To identify those promoters with binding sites corresponding to binding motifs obtained from the Biobase Transfac database release 2009.3 (BIOBASE GmbH, Wolfenbuettel, Germany), we used the MotifLocator algorithm[50] with a first order background model [51]. Significant *P* value for binding was estimated from randomly permuted promoter sequences. For promoter scanning, the DNA sequence 1 kb upstream from the transcription start site of each differentially expressed gene were downloaded from UCSC Genome Browser hg19 genome build.

For each gene, we determined a set of transcription factors whose binding sites lie up to 1 kb upstream of the gene's transcription start site ($P < 0.001$). Using these data, we tested which transcription factor binding motifs occur more often than expected at random at the upstream regions of differentially expressed genes. We used a hypergeometric distribution with a *P* value threshold of 0.05 to test for statistical significance. We repeated the process of finding the enriched transcription factor binding motifs for each of the 100 resampled sets to obtain the most frequently identified binding motifs and the corresponding transcription factors, which we termed the "master regulators".

Constructing genomic neighborhoods

Construction of a genomic neighborhood around a master regulator began downstream with the inclusion of the predicted targets using Ingenuity Pathway Analysis (IPA) software (Ingenuity Systems, http://www.ingenuity.com). Regulatory relationships were then established between the master regulator and its predicted targets in IPA. A *q* value cutoff of 0.005 was set to identify molecules whose expression was differently regulated between the sample sets. The connections to molecules that were not differentially expressed were consequentially removed from the network. Taking advantage of the global molecular network developed from information contained in Ingenuity's Knowledge Base, known targets of the master regulator were included in the network, whereas the non-differentially expressed molecules were excluded.

Genomic neighborhoods were then algorithmically generated based on the connectivity of the master regulator and its targets. Further downstream, we added the molecules that are established targets for genes regulated by the master regulator excluding those that are not differentially expressed. We continued adding differentially expressed target genes hierarchically until there were no more target genes to add or we reached a major hub, such as an extensively studied transcription factor that plays a role in many different biological processes and regulates numerous genes under varying conditions. The rationale for excluding the downstream targets of central transcription factors was to limit the size of the neighborhood. We then proceeded similarly upstream from the master regulator by adding the differentially expressed known regulators of the master regulator until we reached the desired depth of the regulatory pathway. A cartoon presentation of neighborhood generation is shown in Figure 1.

Rather than analyzing the biological functions of all the differentially expressed genes simultaneously, we applied enrichment analysis to identify the most significant cancer-critical functions for each genomic neighborhood and master regulator. Molecules that were associated with pathways in Ingenuity's Knowledge Base were considered for the analysis. Right-tailed Fisher's exact test was used to calculate a *P* value to determine the statistical significance for each biological function. A *P* value threshold of 0.05 was chosen as significance level for functional enrichments. All steps in the analysis are schematically summarized in Figure 2.

ChIP-sequencing validation

To show the regulatory potential of the most prominent master regulator, SRF, in inducing expression differences in the established markers as well as its neighborhood molecules, we validated some of the SRF protein-DNA interactions using publicly available ChIP-seq data [52]. In addition to a library from ChIPs against SRF, we used a negative control library of reversed cross-links and no immunoprecipitation (RX-noIP) as a reference. Both data were measured from a human Jurkat cell line and sequenced using the Solexa platform. We re-aligned the sequenced reads against the most recent human genome build (hg19) using Bowtie algorithm [53]. The sequence alignment of SRF and RX-noIP libraries yielded 8.6 and 17.3 million mapped tags, respectively. Regions of significantly high read density ("peaks") were identified by MACS[54] and QuEST[52] software and smoothed read density profiles were illustrated in Integrative Genomic Browser (http://www.bioviz.org/igb/).

Results

Promoter enrichment analysis revealed nine transcription factors that distinguish GIST from LMS

Promoter analysis for 261 transcription factors identified 433 binding motifs that were located up to 1 kb from the transcription start sites of genes throughout the genome ($P < 0.001$). The total number of genes which harbor a particular binding motif in their promoter sequence ranged from 3 to 5951 genes per motif. Iteratively leveraging the promoter analysis data and the GIST and LMS gene expression data, we found 74 different transcription factor binding motifs that were enriched in at least 1 of the 100 resampled sets of differentially expressed genes. The motifs corresponded to 58 differentially expressed (Wilcoxon rank sum; $P < 0.05$) transcription factors. The use of resampling

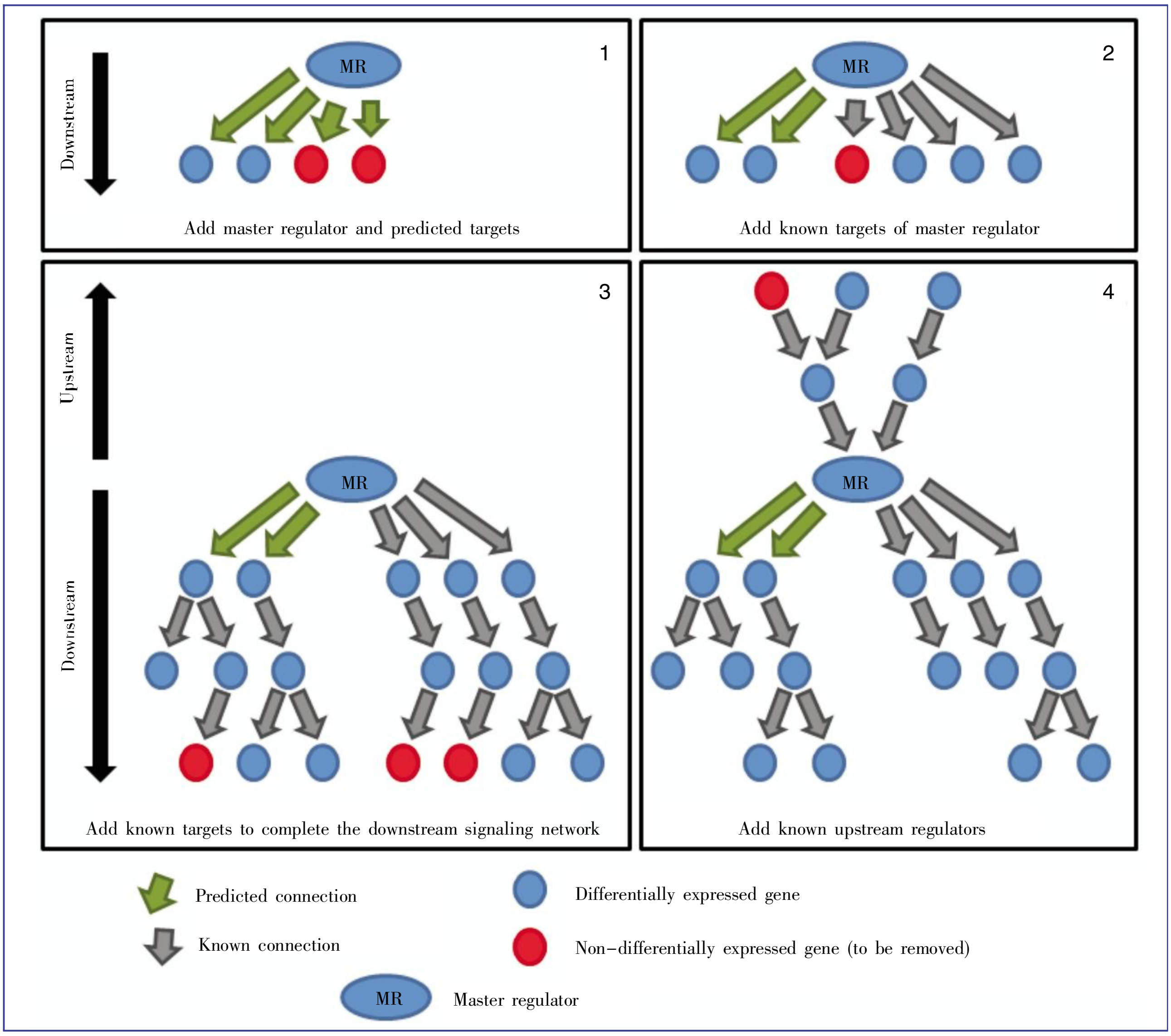
Downstream
MR
1
Add master regulator and predicted targets
MR
2
Add known targets of master regulator
Upstream
Downstream
MR
3
Add known targets to complete the downstream signaling network
MR
4
Add known upstream regulators
Predicted connection
Known connection
Differentially expressed gene
Non-differentially expressed gene (to be removed)
MR
Master regulator

Figure 1. Cartoon illustration of creating genomic neighborhoods. First, the mAdd known targets to complete the downstream signaling networkaster regulator (MR) is added along with its predicted target genes (green arrows). Second, known targets of the master regulator are added. Third, to complete the downstream signaling network, genes are hierarchically added according to regulatory relationships. Fourth, direct and indirect regulators of the master regulator are added. Non-differentially expressed genes (red) are removed after each step.

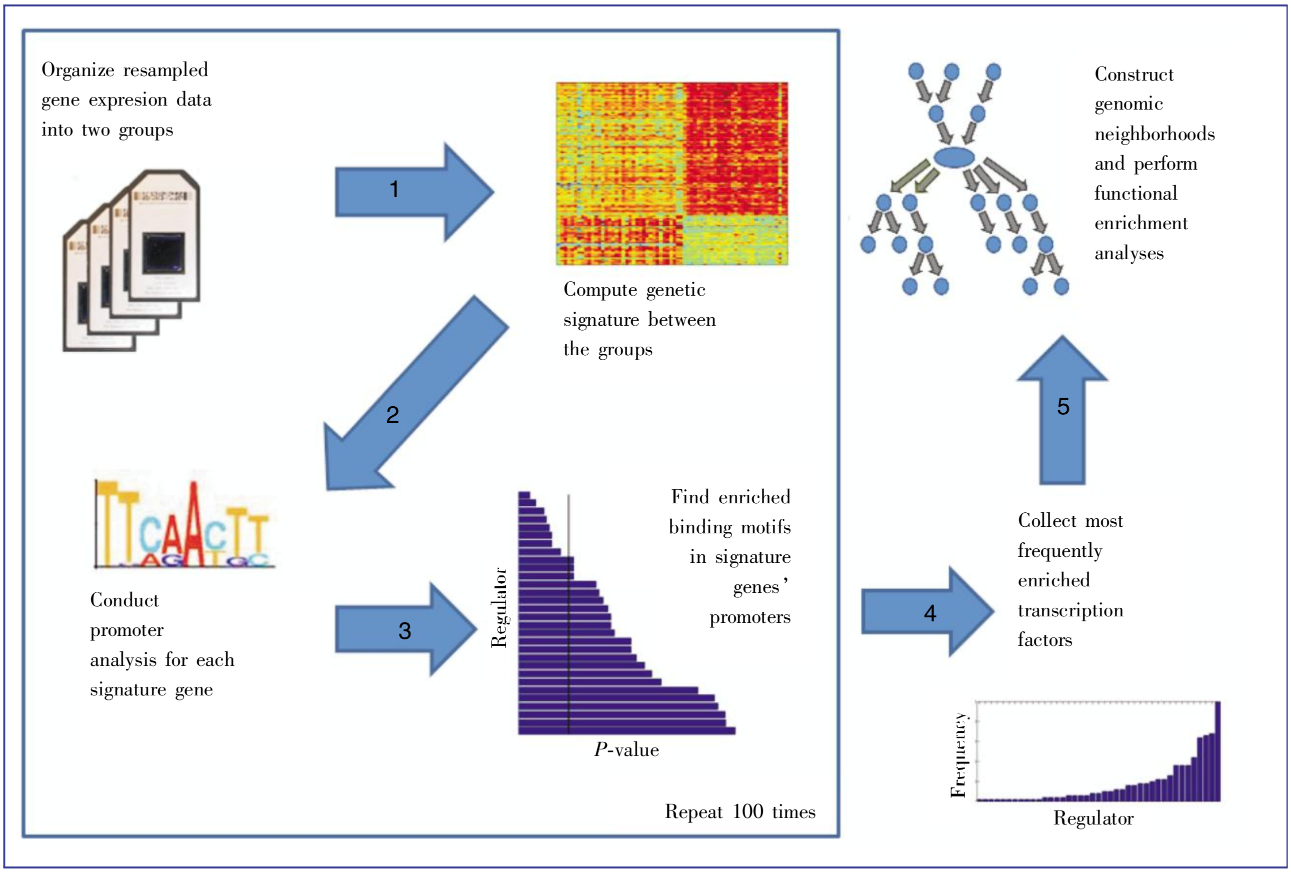
Organize resampled gene expresion data into two groups
1
Compute genetic signature between the groups
2
Conduct promoter analysis for each signature gene
3
Regulator
P-value
Find enriched binding motifs in signature genes' promoters
Repeat 100 times
4
Collect most frequently enriched transcription factors
Frequency
Regulator
5
Construct genomic neighborhoods and perform functional enrichment analyses

Figure 2. Schematic of constructing genomic neighborhoods. First, resampled gene expression data are organized into two clinically relevant groups. Second, a set of differentially expressed genes between the two groups is computed. Third, promoters of the signature genes are scanned for transcription factor binding motifs. Fourth, significantly enriched motifs for the signature gene set are computed. This procedure is repeated 100 times for resampled data. Fifth, the most frequently enriched transcription factors are used further in the pathway analysis. Sixth, genomic neighborhoods are constructed around the most important transcription factors and the neighborhoods are subsequently used for computing functional enrichments in them.

methodology in uncovering the enriched binding motifs revealed that most of the 58 transcription factors would probably not generalize well in other data sets because their binding motifs were enriched in only small portion of the resampled sample sets. Only 9 transcription factors had an enriched binding motif in more than 30% of the resampled sets (Table 1). We held the enrichment frequency (a measure of robustness) in greater importance than the *P* value for differential expression and other statistical or biological parameters.

Serum response factor (SRF) was the only factor with an enriched motif in every signature (100/100 enrichment frequency). In fact, SRF had 6 different robust SRF-binding motifs, whereas the other transcription factors only had one robust motif with the exception of MYC (Figure 3). The next most robust binding motifs corresponded to 2 transcription factors, NK2 transcription factor-related locus 5 (NKX2-5) and coiled-coil domain containing 6 (CCDC6), which had an enriched motif in over 90% of the resampled gene sets (99/100 and 91/100 enrichment frequency, respectively). Six transcription factors [lymphoid enhancer-binding factor 1 (LEF1), vitamin D receptor (VDR), zinc finger protein 350 (ZNF350), tripartite motif-containing 63 (TRIM63), v-maf musculoaponeurotic fibrosarcoma oncogene homolog (MAF), and v-myc myelocytomatosis viral oncogene homolog (MYC)] fell between 30% and 50% in motif enrichment frequency. We termed these 9 genes the master regulators. The remaining 49 transcription factors with enrichmen t frequency below 30% were regarded as not robust enough and, thus, were excluded from subsequent analyses. The correlation between the expressions of 4 master regulators, SRF, CCDC6, MAF, and NKX2-5, and their predicted targets is shown in Figure 4. We clearly observed an expression-prom oting effect for the 3 first master regulators and an inhibitory effect for NKX2-5, particularly in the samples where NKX2-5 was highly expressed (far right in Figure 4D).

Table 1. Summary of top master regulators and their genomic neighborhoods

Symbol	Gene name	Frequency[a]	*P* value[b]	Size[c]	Major biological functions
SRF	Serum response factor	100/100	< 0.001	159	Cell movement; cell death; cell growth and proliferation; cell development; cell morphology
NKX2-5	NK2 transcription factor related, locus 5	99/100	< 0.001	24	Cell morphology; cell growth and proliferation; cell death; cell development; carbohydrate metabolism
CCDC6	Coiled-coil domain containing 6	91/100	< 0.001	38	Cell movement; cell-to-cell signaling and interaction; cell morphology; cell assembly and organization; cell development
LEF1	Lymphoid enhancer-binding factor 1	44/100	0.002	85	Cell movement; cell cycle; cell death; cell growth and proliferation; gene expression
VDR	Vitamin D receptor	41/100	0.016	10	Cell growth and proliferation; cell development; cell cycle; protein synthesis; cellular compromise
ZNF350	Zinc finger protein 350	38/100	0.032	50	Cell death; cell morphology; cell development; cell-to-cell signaling and interaction; cell assembly and organization
TRIM63	Tripartite motif-containing 63	36/100	0.003	72	Cell cycle; cell growth and proliferation; cell death; gene expression; cell development
MAF	V-maf musculo-aponeurotic fibrosarcoma oncogene	33/100	0.001	64	Cell cycle; cell death; cell growth and proliferation; cell development; DNA replication, recombination, and repair
MYC	V-myc myelo-cytomatosis viral oncogene	32/100	0.020	115	Cell death; cell cycle; cell growth and proliferation; cell development; cell morphology

[a]Frequency of significant enrichments of the top binding motif in 100 resampled sample sets; [b]*P* value for differential expression; [c]number of genes in the genomic neighborhood.

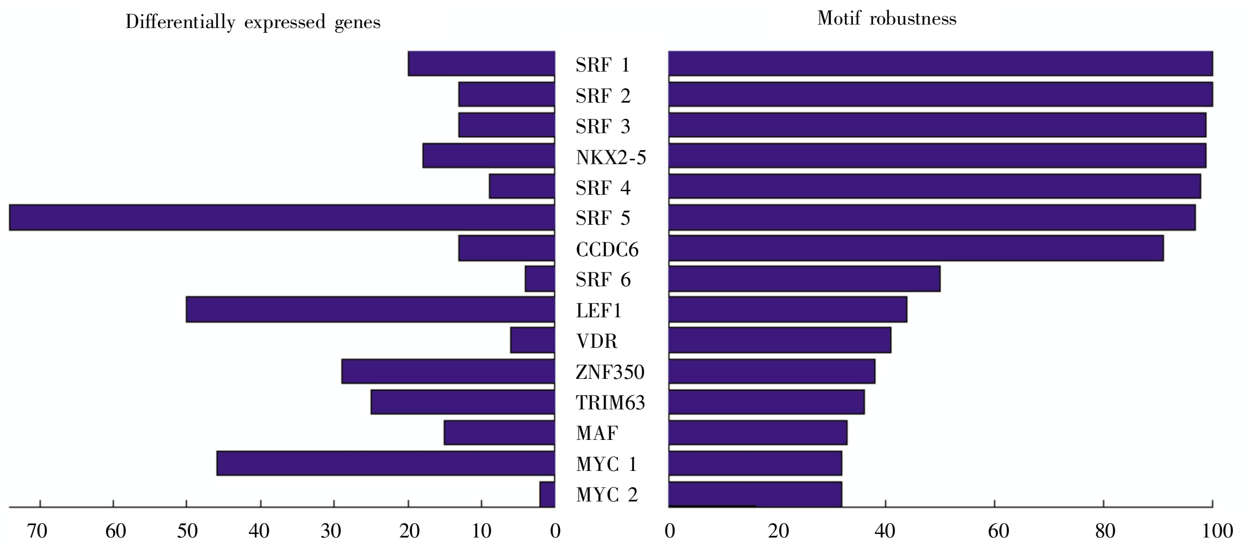

Figure 3. Robust transcription factor binding motifs. The motifs are indicated by the name of the corresponding transcription factor with a running number in the case that has more than one motif per factor. The left side indicates the number of differentially expressed genes that have the binding motif in their promoter region. The right side shows the percentage of resampled signatures where the motif was enriched as a measure of robustness.

Key transcription factors control genes with distinct biological functions

To further investigate the roles of the 9 master regulators in creating the clinical differences between GIST and LMS, their genomic neighborhoods were constructed as described in Methods. The sizes of the resulting gene regulatory networks varied from 10 genes in the VDR neighborhood to 159 genes in the SRF neighborhood. We went on to associate biological processes to the neighborhoods using IPA gene annotations and enrichment analysis to the respective genes. The top 5 processes for each genomic neighborhood are listed in Table 1. Invariably, common

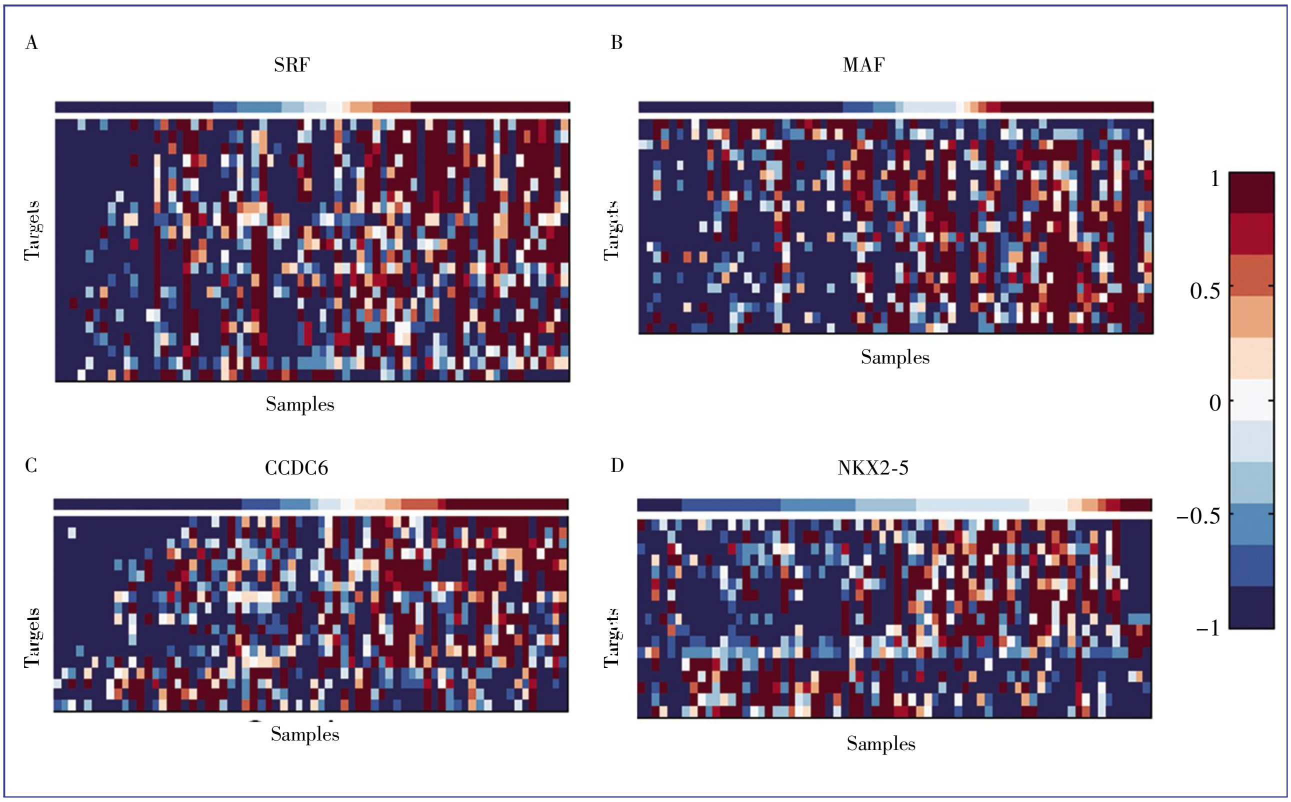

Figure 4. Correlation between master regulator expression and predicted target gene expression. The most strongly expressed predicted target genes of 4 master regulators (SRF, MAF, CCDC6 and NKX2-5) are sorted based on their normalized expression. Red color implies high expression and blue color low expression values. A general enhancing effect is clearly seen for the first 3 regulators whereas NKX2-5 seems to inhibit most of its targets.

cancer-related processes, such as cell cycle, cell death and cell proliferation, appeared on the top of each list. To further establish the functional differences and similarities in neighborhoods of the 9 master regulators, and to relate those to the set of 2330 differentially expressed genes between GIST and LMS (q < 0.005) in the entire cohort, we systematically compared their functional enrichment scores in several categories. An illustration in Figure 5 shows negative log-transformed enrichment P values in 18 broad biological processes that were relevant for all the gene sets.

We observed that the broadest terms, genetic disorder and cancer, were the most enriched for the set of all differentially expressed genes. This was expected because the global signature is a composition of all the cancer-related processes that are different between GIST and LMS, and therefore, the only unifying functional themes must be very broad. Among the more specialized terms, there was a large variation of enrichment scores between the neighborhoods. Some neighborhoods greatly exceeded the score of the global gene signature in many categories. Since the network genes are functionally coherent subsets of all differentially expressed genes, we expected them to gain a much higher score in specialized categories and possibly a lower score in broad categories. Overall, the highest significance scores in most categories were given to the SRF, LEF1, and CCDC6 neighborhoods. The diversity of SRF's cellular functions was also seen as a significant enrichment in most functional categories (Figure 5).

We provided schematics of the genomic neighborhoods of SRF (Figure 6A) and ZNF350 (Figure 6B). In Figure 6A, we highlighted three partially overlapping biological processes to which the SRF neighborhood is very significantly associated: cancer, muscle development, and cellular movement. To show the link from SRF to three generally cancer-related signaling pathways, we included PTEN, VEGF and ATM signaling, which are also more strictly defined than most processes and, therefore, contain fewer genes in general. For ZNF350 (Figure 6B), we highlighted the genes that are associated with cancer, cell death, and integrin signaling.

Experimental identification of binding sites by ChIP-seq experiments validated 28 differentially expressed predicted targets of SRF, such as a computational marker PRUNE2. In addition to validating

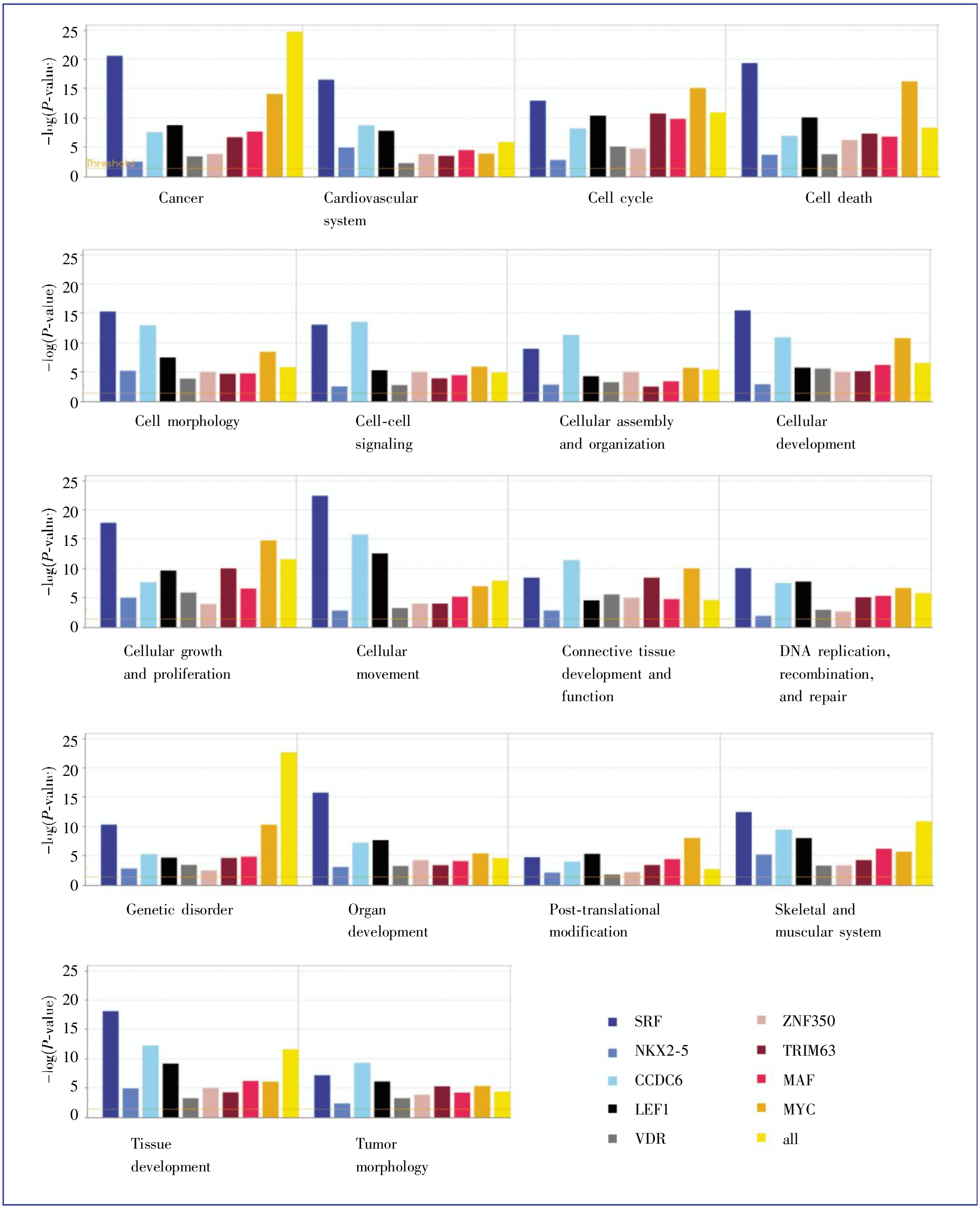

Figure 5. Functional enrichment analysis for genomic neighborhoods. Enrichment scores of genes in 9 neighborhoods are compared against the global signature (all differentially expressed genes) in 18 categories. Interestingly, the global signature reaches the highest score in the broadest categories: cancer and genetic disorder. The neighborhoods score the highest in more specialized categories.

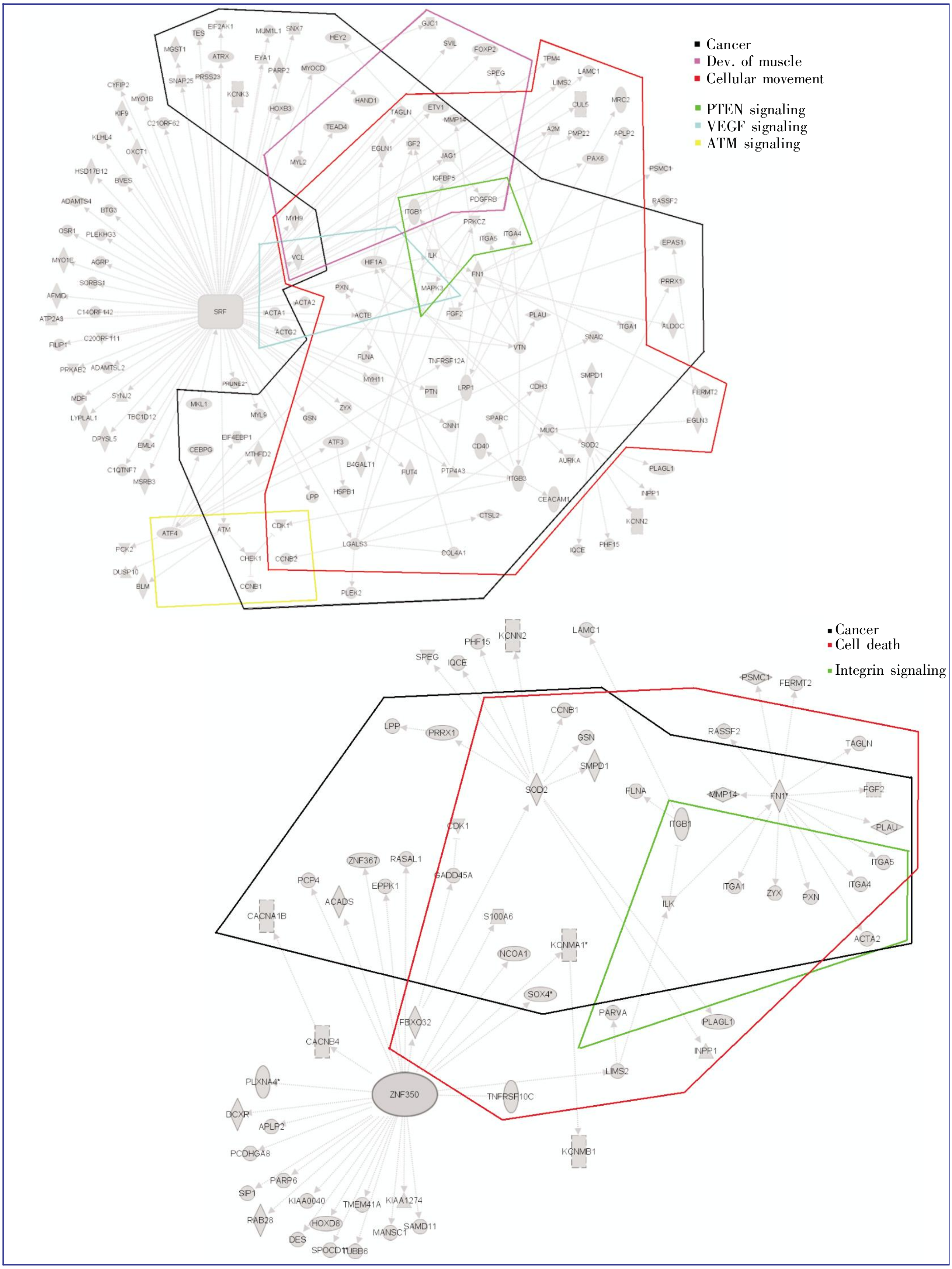
■ Cancer
■ Dev. of muscle
■ Cellular movement
■ PTEN signaling
■ VEGF signaling
■ ATM signaling
SRF
TES
EIF2AK1
MUM1L1
SNX7
HEY2
GJC1
MGST1
ATRX
EYA1
PARP2
MYOCD
SVIL
FOXP2
SPEG
TPM4
LAMC1
LIMS2
CYFIP2
SNAP25
PRSS23
KCNK3
HAND1
HOXB3
MYO1B
KIF9
C21ORF62
TAGLN
ETV1
MMP14
CUL5
MRC2
KLHL4
TEAD4
A2M
PMP22
APLP2
OXCT1
EGLN1
IGF2
JAG1
HSD17B12
BVES
MYL2
PAX6
PSMC1
IGFBP5
ADAMTS4
BTG3
PDGFRB
RASSF2
ITGB1
MYH9
PRKCZ
OSR1
PLEKHG3
ITGA5
ITGA4
EPAS1
MYO1E
AGRP
VCL
ILK
HIF1A
SORBS1
FN1
PRRX1
AFMID
PXN
MAPK3
ACTA2
ATP2A8
C14ORF142
ACTA1
ACTB
FGF2
PLAU
ITGA1
ALDOC
ACTG2
FILIP1
C20ORF111
SNAI2
VTN
FLNA
PRKAB2
ADAMTSL2
TNFRSF12A
SMPD1
MYH11
PRUNE2*
PTN
LRP1
CDH3
FERMT2
MDFI
SYNJ2
MKL1
MYL9
ZYX
LYPLAL1
TBC1D12
GSN
CNN1
SPARC
EGLN3
MUC1
DPYSL5
EML4
EIF4EBP1
SOD2
CEBPG
ATF3
CD40
MTHFD2
AURKA
C1QTNF7
B4GALT1
FUT4
PTP4A3
PLAGL1
MSRB3
ITGB3
HSPB1
INPP1
LPP
CEACAM1
CTSL2
KCNN2
CDK1
ATF4
ATM
PHF15
PCK2
LGALS3
IQCE
COL4A1
CHEK1
CCNB2
DUSP10
BLM
CCNB1
PLEK2
▪ Cancer
▪ Cell death
▪ Integrin signaling
PHF15
KCNN2
LAMC1
SPEG
IQCE
PSMC1
FERMT2
CCNB1
LPP
PRRX1
RASSF2
GSN
TAGLN
SMPD1
SOD2
FLNA
MMP14
FN1*
FGF2
CDK1
ITGB1
PLAU
ZNF367
RASAL1
ITGA5
PCP4
GADD45A
EPPK1
ACADS
ITGA1
ZYX
PXN
ITGA4
ILK
CACNA1B
S100A6
ACTA2
KCNMA1*
NCOA1
SOX4*
PARVA
PLAGL1
FBXO32
CACNB4
INPP1
LIMS2
PLXNA4*
ZNF350
TNFRSF10C
DCXR
APLP2
KCNMB1
PCDHGA8
PARP6
SIP1
KIAA0040
TMEM41A
KIAA1274
RAB28
HOXD8
SAMD11
MANSC1
DES
SPOCD1
TUBB6

Figure 6. Graphical representation of the genomic neighborhoods of SRF and ZNF350. Genes are represented as nodes, and the biological relationship between two nodes is represented as an edge. All edges are either between the master regulator and its predicted targets or supported by at least one reference from the literature, a textbook, or canonical information stored in the Ingenuity Pathways Knowledge Base. Nodes are displayed using various shapes that represent the functional class of the gene product. In panel A for SRF, colored boxes encircle genes that are annotated to biological processes (cancer, muscle development, and cell movement) or signaling pathway (PTEN signaling, VEGF signaling, and ATM signaling). All genes shown here are differentially expressed between GIST and LMS (q < 0.001). Regulatory connections are shown by dotted arrows pointing the direction of regulation. MKL1 is the only upstream molecule known to regulate SRF, which, in turn, regulates the expression of several marker genes, including the computational marker PRUNE2 and MYOCD, which regulates immunohistochemical marker smooth muscle actin. In panel B for ZNF350, colored boxes encircle genes that are annotated to cancer, cell death, or integrin signaling. Notably, ZNF350 is predicted to bind to the promoter of muscle-specific immunohistochemical marker of LMS, desmin (DES).

some of the predictions, several canonical SRF targets, such as actin genes and FHL2, also showed pronounced peaks in ChIP-seq data. SRF autoregulation by binding to its own promoter and binding to the co-activator MKL1 were clearly observed. However, there were no SRF binding events detected in the promoter of another co-activator, myocardin (Figure 7).

Discussion

In this study, we developed a methodology for studying the global gene expression differences between two groups of samples in a biologically relevant context. Our method is a resampling-based computational approach that couples gene expression data with promoter scanning, database information, and gene set enrichment analysis to find the most robust and biologically relevant transcription factors, the master regulators. A difference between the master regulators and common marker genes is that markers are usually designed to be optimal for classification purposes, whereas the master regulators are likely more biologically relevant. Nevertheless, the master regulators may also be used as a more robust genetic signature, although tumor classification was not included in the scope of this study. The paramount importance of the master regulators is that they can control most of the difference in gene regulation patterns that are observed in the disease phenotypes through their extensive regulatory connections. Despite their biological relevance, the regulators themselves do not always appear on top of the list of the most differentially expressed genes, which makes it hard to identify them with conventional methods.

In addition to identifying the key transcription factors,

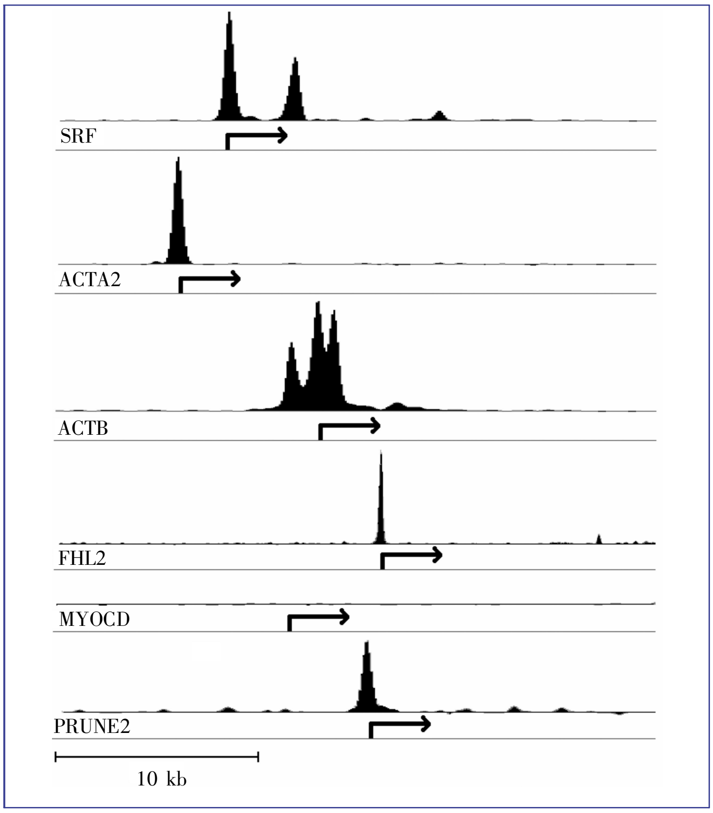

Figure 7. ChIP-seq validation of SRF neighborhood. The vertical lines represent 30 kb sequence around transcription start sites (arrows point to the direction of transcription) of 9 predicted or known target genes of SRF. Read density profiles from a ChIP-seq experiment are shown for each gene. There is a pronounced region of high read density ("peak") within the promoter showing autoregulation of SRF, regulation of actin genes (ACTA2, ACTB) and FHL2. SRF has not bound to the promoter of myocardin (MYOCD), one of the established targets of SRF. Predicted target PRUNE2 also shows a peak at their promoter region.

our approach results in ma rked improvement in the interpretability of the global genetic signature. Only recently have enrichment analyses become the most common analysis for interpreting the biological theme of gene lists. In contrast to individually studying the top genes, enrichment analysis finds statistically significant over- or under-representations of gene categories. Although this is a marked improvement to single gene analysis, applying it directly to the gene list has drawbacks of its own. Mainly, the differentially expressed genes are usually a product of numerous aberrant pathways. Therefore, the signature genes rarely have any real unified biological themes. The presented work extends the standard enrichment analysis workflow towards systems biology by refining the signature and integrating additional information. Linking transcription factors to their differentially expressed up- and down-stream genes results in gene regulatory networks that are functionally coherent subsets of the global genetic signature.

Genomic neighborhoods have several advantages over the use of a global genetic signature. First, enrichment analysis now gives us information not only about the functional difference between the samples but also on the function of the neighborhood. Second, it extends the knowledge on the functional role of the master regulator, even if information on the gene itself is scarce or unavailable. Third, the network topology contains direct information on the biochemical mechanism, which might lead to a differential clinical property in the phenotype. On the other hand, drawbacks include the difficulty of predicting transcription factor binding in mammals, lack of knowledge of gene functionality and regulation, and the inability to measure post-transcriptional regulation. Additionally, the selected significance threshold has an effect on the final network and even the master regulator itself. While computing enrichments alleviates this problem, the method could benefit from inclusion of probabilistic features making strict thresholds unnecessary. Even accounting for these obstacles, we believe that other investigators will benefit from using our methodology in their research.

We demonstrated our approach by using it to investigate the biological differences in two soft tissue sarcomas, GIST and LMS. We uncovered 9 transcription factors that were robust and differentially expressed between the tumors. The most prominent regulator SRF is a widely expressed transcription factor that participates in many global and tissue-specific processes like muscle cell differentiation and growth factor-induced cell proliferation, but it has also been linked to human diseases like heart disease and cancer [55]. Previous studies have also shown that overexpression of SRF co-factor myocardin (MYOCD) leads to increased migration ability in LMS[56]. This is well in accordance with our observation that the SRF neighborhood is strongly associated with cellular movement. The second transcription factor, LEF1, is a mediator of Wnt signaling through which it plays a role in tissue specification and apoptosis [57,58]. This is reflected by high enrichment scores in the cell cycle, cell death, and tissue development categories in our study. The remarkably high enrichment of the LEF1 neighborhood in the cellular movement category is also a noteworthy finding. There is less evidence on the functional role of ubiquitously expressed CCDC6, but it has been linked to papillary thyroid carcinomas through fusion to RET proto-oncogene [59]. Also worth noting is the strong association of the MYC neighborhood with cell death, growth, and proliferation, as could be expected as per its well-known biological effects as an oncogene [60]. One concern is that many genes in human genome do not have any functional annotations. This fact affects particularly smaller gene sets which may contain only a few annotated genes and also likely induces a bias towards finding only the known functions for neighborhoods around extensively studied transcription factors such as MYC. The lack of knowledge on regulatory interactions is also one probable cause for the consistently lower significance scores of neighborhoods of NKX2-5, VDR, MAF, TRIM63, ZNF350.

Although SRF and ZNF350 have not been associated with processes like cellular movement, cell death, and integrin signaling before, they have the potential to regulate the processes and signaling pathways through their extensive regulatory connections. Therefore, we associated SRF and ZNF350 with these cellular functions through our analysis and proposed the relevant regulatory relationships in detail. Similarly, we proposed several new associations for all 9 master regulators through functional enrichment analysis of their genomic neighborhoods. The 9 master regulators may directly or indirectly explain the differential regulation of many computational and immunohistochemical markers used to differentiate the clinical groups. For example, an immunohistochemical marker for LMS, desmin, is predicted to be directly regulated by ZNF350. Another example of direct regulation is SRF's predicted target PRUNE2, which has been used as a computational marker in an earlier 2-gene classifi er differentiating between GIST and LMS[45]. In addition, SRF is known to regulate its own muscle-specific co-activator MYOCD, which, together with SRF, forms a complex that is a master regulator of smooth muscle gene expression pattern [61]. One of the genes in this pattern is smooth muscle actin, which is another common immunohistochemical marker used in diagnosing LMS. Cell type-specific regulation for SRF has been shown before[62] and probably accounts for many of the targets without visible peaks.

Uncovering SRF as an extremely robust master regulator differentiating between GIST and LMS is the most important finding of this work. Differential activation of SRF co-activators has been suggested as one way of directing SRF regulation to subsets of target genes in a tissue-specific manner. The only known co-factors that were differentially expressed in our data are NKX2-5, MYOCD, and MKL1. While MKL1 is ubiquitously expressed, MYOCD is required for the expression of smooth muscle specific genes[63]. Association to muscle development, on the other hand, may be explained by the cell lineage-specific expression rather than by the difference in tumor biology as leiomyosarcomas are tumors of the smooth muscle. Thus, they likely have a more smooth muscle cell-specific expression pattern than GISTs, which are thought to arise from interstitial cells of Cajal[64]. Having matching control samples would likely remove cell type-specific findings, but it is often very difficult to obtain such a sample for various practical reasons, such as the limited availability of healthy tissue. Even if some of the transcription factors were unrelated to these cancers, uncovering a fundamental regulatory program like smooth muscle-specific regulation supports our assertion that we can find meaningful gene regulatory networks and key factors behind them.

We showcased here our novel data analysis framework in a topical biological setting and provided solid hypotheses for further experimental work in this field. Particularly, the functional associations of the 9 master regulators warrant experimental validation in the context of GIST and LMS. An in depth investigation into these transcription factors may increase our knowledge of the functional differences between the tumors. Methodologically, genomic neighborhoods could be built similarly for studying different biological settings, such as tumor versus normal tissue, early stage versus late stage tumor, or primary tumor versus metastasis. By identifying key transcription factors, application into the aforementioned scenarios could help us understand the major mechanisms of tumorigenesis, disease progression, and metastasis.

Conclusions

In conclusion, we developed a framework that applies a systems approach to analyze the functional differences in gene expression patterns. The approach integrates several levels of data such as gene expression measurements, transcription factor binding site predictions, regulatory network topology, and information on gene function. Using a set of expression measurements from GIST and LMS, we demonstrated the applicability of our approach. Analyzing the differences in the expression patterns yielded many interesting starting points for future research that can potentially lead to better diagnostic methods and deeper understanding into biological differences of these tumors. Specifically, we uncovered 9 differentially expressed transcription factors around which we generated gene regulatory networks of differentially expressed genes using promoter analysis and literature-based database information. These master regulators and their genomic neighborhoods may be the epicenter of different clinical properties of GIST and LMS.

Acknowledgment

We thank Antti Larjo for participating in the analysis of the sequencing data. This work was supported by Project for the Biological Information and Information Processing Properties of Biological Systems from the Academy of Finland (No. 122973); Project for the Structure-dynamics Relationships in Biological Network from the Academy of Finland (No. 132877); Finnish Funding Agency for Technology and Innovation Finland Distinguished Professor program (No. 1480/31/09).

Received: 2010-11-17; revised: 2010-11-29; accepted: 2010-12-10.

References

[1] Knudson AG. Cancer genetics [J]. Am J Med Genet, 2002,111 (1):96-102.

[2] Baylin SB, Herman JG. DNA hypermethylation in tumorigenesis: epigenetics joins genetics [J]. Trends Genet, 2000,16(4):168-174.

[3] Baylin SB, Esteller M, Rountree MR, et al. Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer [J]. Hum Mol Genet, 2001,10(7):687-692.

[4] Bond GL, Hu W, Levine A. A single nucleotide polymorphism in the MDM2 gene: from a molecular and cellular explanation to clinical effect [J]. Cancer Res, 2005,65(13):5481-5484.

[5] Beer DG, Kardia SL, Huang CC, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma [J]. Nat Med, 2002,8(8):816 - 824.

[6] Bhattacharjee A, Richards WG, Staunton J, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses [J]. Proc Natl Acad Sci U S A, 2001,98(24):13790-13795.

[7] Camós M, Esteve J, Jares P, et al. Gene expression profiling of acute myeloid leukemia with translocation t (8;16)(p11;p13) and MYST3-CREBBP rearrangement reveals a distinctive signature with a specific pattern of HOX gene expression [J]. Cancer Res, 2006,66(14):6947:6954.

[8] Chang HY, Nuyten DS, Sneddon JB, et al. Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival [J]. Proc Natl Acad Sci U S A, 2005,102(10):3738-3743.

[9] Huang F, Reeves K, Han X, et al. Identification of candidate molecular markers predicting sensitivity in solid tumors to dasatinib: rationale for patient selection [J]. Cancer Res, 2007,67(5):2226-2238.

[10] Kim S, Dougherty ER, Shmulevich I, et al. Finding strong gene sets for classification of gliomas [J]. Mol Cancer Ther, 2002,1:

1229-1236.

[11] Kobayashi T, Yamaguchi M, Kim S, et al. Microarray reveals differences in both tumors and vascular specific gene expression in de novo CD5$^+$ and CD5$^-$ diffuse large B-cell lymphomas [J]. Cancer Res, 2003,63(1):60-66.

[12] Iizuka N, Oka M, Yamada-Okabe H, et al. Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection [J]. Lancet, 2003,361(9361):923-929.

[13] Pomeroy SL, Tamayo P, Gaasenbeek M, et al. Prediction of central nervous system embryonal tumour outcome based on gene expression [J]. Nature, 2002,415(6870):436-442.

[14] Ramaswamy S, Ross KN, Lander ES, et al. A molecular signature of metastasis in primary solid tumors [J]. Nat Genet, 2003,33(1):49-54.

[15] Rosenwald A, Wright G, Chan WC, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma [J]. N Engl J Med, 2002,346(25):1937-1947.

[16] van't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer [J]. Nature, 2002,415(6871):530-536.

[17] van de Vijver MJ, He YD, van't Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer [J]. N Engl J Med, 2002,347(25):1999-2009.

[18] Wang Y, Klijn JG, Zhang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer [J]. Lancet, 2005,365(9460):671-679.

[19] Yeoh EJ, Ross ME, Shurtleff SA, et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling [J]. Cancer Cell, 2002,1(2):133-143.

[20] Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data [J]. Proc Natl Acad Sci U S A, 2002,99(10):6562-6566.

[21] Garber K. Genomic medicine. Gene expression tests foretell breast cancer's future [J]. Science, 2004,303 (5665):1754-1755.

[22] Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: a multiple random validation strategy [J]. Lancet, 2005,365(9458):488-492.

[23] Miller LD, Long PM, Wong L, et al. Optimal gene expression analysis by microarrays [J]. Cancer Cell, 2002,2(5):353-361.

[24] Simon R. Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data [J]. Br J Cancer, 2003,89(9):1599-1604.

[25] Wessels LFA, Reinders MJT, Hart AAM, et al. A protocol for building and evaluating predictors of disease state based on microarray data [J]. Bioinformatics, 2005,21(19):3755-3762.

[26] Carro MS, Lim WK, Alvarez MJ, et al. The transcriptional network for mesenchymal transformation of brain tumours [J]. Nature, 2010,464(7279):318-325.

[27] Huang E, Ishida S, Pittman J, et al. Gene expression phenotypic models that predict the activity of oncogenic pathways [J]. Nat Genet, 2003,34(2):226-230.

[28] Lamb J, Ramaswamy S, Ford HL, et al. A mechanism of cyclin D1 action encoded in the patterns of gene expression in human cancer [J]. Cell, 2003,114(3):323-324.

[29] Mootha VK, Lindgren CM, Eriksson KF, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulared in human diabetes [J]. Nat Genet, 2003,34(3):267-274.

[30] Rhodes DR, Yu J, Shanker K, et al. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression [J]. Proc Natl Acad Sci U S A, 2004,101(25):9309-9314.

[31] Segal E, Friedman N, Koller D, et al. A module map showing conditional activity of expression modules in cancer [J]. Nat Genet, 2004,36(10):1090-1098.

[32] Segal E, Friedman N, Kaminski N, et al. From signatures to models: understanding cancer using microarrays [J]. Nat Genet, 2005,37(Suppl):S38-S45.

[33] The Gene Ontology Consortium. Gene ontology: tool for the unification of biology [J]. Nat Genet, 2000,25(1):25-29.

[34] Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes [J]. Nucleic Acids Res, 2000,28(1):27-30.

[35] Pathway Commons database [http://www.pathwaycommons.org/pc/].

[36] Vogelstein B, Kinzler K. Cancer genes and the pathways they control [J]. Nat Med, 2004,10(8):789-799.

[37] Sørlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications [J]. Proc Natl Acad Sci U S A, 2001,98(19):10869-10874.

[38] Verhaak RG, Hoadley KA, Purdom E, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1 [J]. Cancer Cell, 2010,17(1):98-110.

[39] Dunlap SM, Celestino J, Wang H, et al. Insulin-like growth factor binding protein 2 promotes glioma development and progression [J]. Proc Natl Acad Sci U S A, 2007,104(28):11736-11741.

[40] Demetri G, Benjamin R, Blanke C, et al. NCCN Task Force report: management of patients with gastrointestinal stromal tumor (GIST) – update of the NCCN clinical practice guidelines [J]. J Natl Compr Canc Netw, 2007,5(Suppl 2):S1-S29.

[41] Joensuu H. Gastrointestinal stromal tumor (GIST) [J]. Ann Oncol, 2006,17(Suppl 10):280-286.

[42] Katz S, DeMatteo R. Gastrointestinal stromal tumors and leiomyosarcomas [J]. J Surg Oncol, 2008,97(4):350-359.

[43] Heinrich M, Corless C, Duensing A, et al. Pdgfra activating mutations in gastrointestinal stromal tumors [J]. Science, 2003,299(5607):708-710.

[44] Trent JC, Lazar AJ, Zhang W. Molecular approaches to resolve diagnostic dilemmas: the case of gastrointestinal stromal tumor and leiomyosarcoma [J]. Future Oncol, 2007,3(6):629-637.

[45] Price N, Trent J, El-Naggar A, et al. Highly accurate two-gene classifier for differentiating gastrointestinal stromal tumors and leiomyosarcomas [J]. Proc Natl Acad Sci U S A, 2007,104(9):3414-3419.

[46] Yang D, Ylipää A, Yang J, et al. An integrated study of aberrant gene copy number and gene expression in GIST and LMS [J]. Technol Cancer Res Treat, 2010,9(2):171-178.

[47] Ylip ää A, Hunt K, Yang J, et al. Integrative genomic characterization and a genomic staging system for gastrointestinal stromal tumors [J]. Cancer, 2010 Sep 3. [Epub ahead of print]

[48] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining inference, and prediction [M]. Second edition. Springer-Verlag, 2008:763.

[49] Storey J. A direct approach to false discovery rates [J]. Journal of the Royal Statistical Society, Series B (Methodological), 2002,64(3):479-498.

[50] Thijs G, Moreau Y, deSmet F, et al. INCLUSive: integrated clustering, upstream sequence retrieval and motif sampling [J]. Bioinformatics, 2002,18(2):331-332.

[51] Nykter M, L ähdesm äki H, Rust A, et al. A data integration framework for prediction of transcription factor targets: a BCL6 case study [J]. Ann N Y Acad Sci, 2009,1558:205-214.

[52] Valouev A, Johnson DS, Sundquist A, et al. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data [J]. Nat Methods, 2008,5(9):829-834.

[53] Langmead B, Trapnell C, Pop M, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome [J]. Genome Biol, 2009,10(3):R25.

[54] Zhang Y, Liu T, Meyer CA, et al. Model-based analysis of ChIP-Seq (MACS) [J]. Genome Biol, 2008,9(9):R13.

[55] Cooper SJ, Trinklein ND, Nguyen L, et al. Serum response factor binding sites differ in three human cell types [J]. Genome Res, 2007,17(2):136 - 144.

[56] Pérot G, Derré J, Coindre JM, et al. Strong smooth muscle differentiation is dependent on myocardin gene amplification in most human retroperitoneal leiomyosarcomas [J]. Cancer Res, 2009,69(6):2269–2278.

[57] Labbé E, Letamendia A, Attisano L. Association of Smads with lymphoid enhancer binding factor 1/T cell-specific factor mediates cooperative signaling by the transforming growth factor-β and Wnt pathways [J]. Proc Natl Acad Sci U S A, 2000,97(15):8358–8363.

[58] Gandhirajan RK, Staib PA, Minke K, et al. Small molecule inhibitors of Wnt/β-Catenin/Lef-1 signaling induces apoptosis in chronic lymphocytic leukemia cells in vitro and in vivo [J]. Neoplasia, 2010,12(4):326–335.

[59] Celetti A, Cerrato A, Merolla F, et al. H4 (D10S170), a gene frequently rearranged with RET in papillary thyroid carcinomas: functional characterization [J]. Oncogene, 2004,23(1):109–121.

[60] Oster SK, Ho CS, Soucie EL, et al. The myc oncogene: MarvelouslY Complex [J]. Adv Cancer Res, 2002,84 81–154.

[61] Wang Z, Wang DZ, Pipes GC, et al. Myocardin is a master regulator of smooth muscle gene expression [J]. Proc Natl Acad Sci U S A, 2003,100(12):7129–7134.

[62] Philippar U, Schratt G, Dieterich C, et al. The SRF target gene Fhl2 antagonizes RhoA/MAL-dependent activation of SRF [J]. Mol Cell, 2004,16(6):867–880.

[63] Cen B, Selvaraj A, Prywes R. Myocardin/MKL family of SRF coactivators: key regulators of immediate early and muscle specific gene expression [J]. J Cell Biochem, 2004,93(1):74 - 82.

[64] Miettinen M, Lasota J. Gastrointestinal stromal tumors: review on morphology, molecular pathology, prognosis, and differential diagnosis [J]. Arch Pathol Lab Med 2006,130(10):1466–1478.