Identifying SRF function with candidate cofactors in myeloid and lymphoid cells using a bioinformatics approach for gene expression analysis

Jason Bedford[1], Parker Sulkowski[2], Cong Liang[1], Gyong Ju (Liz) Min[2], Nur-Taz Rahman[2]

[1] Computational Biology and Bioinformatics (CBB), Yale University, New Haven, CT

[2] Molecular Cell Biology, Genetics and Development (MCGD) Track, Yale University, New Haven, CT

Abstract

Serum Response Factor (SRF) is a well-known transcription factor that is expressed in all cell types, turns on immediate early genes, and serves as a master regulator of cytoskeletal genes. In addition, SRF has been shown to be important for the maturation of multipotential cells in the blood, particularly in the myeloid lineage. In this study, we aim to examine the role of SRF in driving gene expression in myeloid and lymphoid blood cell lineages, using the cell lines: K562 and GM12878. Using publicly available RNA-seq and anti-SRF ChIP-seq datasets for these cells, we apply effective bioinformatic tools in order to analyze changes in gene expression corresponding to SRF binding. Our results suggest that SRF regulates pathways that are specific to each cell type and we identified candidate transcriptional cofactors that could potentially interact with SRF to mediate gene expression. The analysis provides a strong initial framework for future investigation continuing to resolve discrete roles of SRF in the hematopoietic lineage through experimental and computational work.

Introduction:

The Serum Response Factor (SRF) is a ubiquitously expressed transcription factor that serves as a master regulator of cytoskeletal genes. It binds to a consensus DNA sequence, CC(AT)6GG (CArG box), in the promoter of its target genes. Since SRF is widely expressed, it often interacts with cofactors that are tissue-specific or activated during a specific time in development. For example, SRF directly interacts with the Myocardin family of transcriptional coactivators to regulate muscle differentiation genes as well as the E26 transformation-specific (Ets) to regulate growth-induced genes. Recently, SRF was shown to interact with Megakaryoblastic Leukemia 1 (MKL1), a Myocardin family member, to drive megakaryocyte differentiation, but the role of SRF in cells of the blood lineage are largely unknown. The aim of this project is to use a bioinformatic approach to determine whether SRF regulates gene expression in K562 and GM12878 cells, which are erythroleukemia and lymphoblast cell lines, respectively, and to compare the genes that are regulated by SRF in the two different cell lines. To this end, we will analyze the ChIP-seq and RNA-seq data that are publicly available for each of these cell lines by mapping the SRF binding site and comparing the list of genes that are up- or down-regulated. We will use the DHS and ChIP-seq data for H3K27ac and H3K4me3 from the UCSC genome browser which mark transcriptionally active sites to correlate with the up-regulated genes. Though SRF is known to bind CArG sites, we can perform FIMO analysis to identify additional binding motifs. Also, we will use the GO analysis tool to examine the biological pathways in which the genes are involved and determine whether certain pathways are correlated with certain SRF binding motifs.

Methods:

ChIP-seq Analysis

K562 and GM12878 ChIP-seq data sets for SRF, and histone modifications H3K4me3 and H3K27ac were downloaded directly to LOUISE from ENCODE. Each data set had two biological replicates and all these data sets were in the FASTQ format. The BAM files for each sample was created using Bowtie and Samtools for mapping. Peaks were called for each sample using MACS14, which generated wiggle files and bed files.

BEDtools were used to integrate the SRF binding site with open regions of chromatin and active histone markers through DHS and histone modification ChIP peaks. A group from Duke University has uploaded analyzed DNase-seq data for K562 and GM12878. Peaks were called with F-seq. Here, we directly took advantage of the called peaks for each cell line as the open regions of chromatin. DHS data can be found under GSE32970. Histone modification and SRF ChIp-seq peaks were called as described above. For each cell line, only shared regions among SRF binding peaks, DHS, H3K27ac and H3K4me3 ChIP peaks were used for further analysis. For the sake of simplicity, we call these peaks "SRF active binding peaks" from here on. To make a comparison between K562 and GM12878, unique peaks and shared peaks were also identified.

RNA-seq Analysis

Data sets for RNA sequencing were obtained from NCBI GEO. K562 data sets were available from NCBI's GEO under the Sample ID GSM836272 and were published from the HudsonAlpha Institute for Biotechnology. These files were downloaded in a .sra format. The GM12878 data sets were also available from GEO under GSM754335 and was published by the Gerstein and Snyder labs at Yale . These files were available in .sra format and were converted to .fastq files using the SRA toolkit from NCBI. Reads were aligned to hg19 using the TopHat aligner. The alignment used a GTF file of UCSC genes as well as the bowtie index for hg19. After alignment differential expression data were collected using the CuffDiff program with two biological replicates for each data set, (K562, K562 with shRNA , and GM12878). This provided output in a gene_exp.diff file, which was used in the Perl script to compare gene expression to histone marks on the promoters of those genes. The Heatmaps for differential expressions were made using the cummRbund R software package, and the specific code and list of genes are attached in the supporting materials for this project.

Motif Analysis

MEME suite was used for motif analysis. First, SRF binding motif was downloaded from JASPAR motif database. A frequency matrix of binding pattern was used for known motif searching. Second, FIMO was utilized to take the motif matrix and search for the known motif among specified genomic regions. About 10% of the SRF active binding regions could be explained by SRF known motif. Finally, MEME-ChIP was used to find de novo motifs among those peak regions, and the motifs found were related to known motifs through the JASPAR database.

Correlating ChIP-seq and RNA-seq

Finally, to correlate SRF active binding sites to known promoters, a table of 1kbp upstream UCSC annotated TSS was generated through the "table browser" utility of the UCSC genome browser. BEDtools was used to extract the promoters that have an SRF active binding peak, and this was used as one of the inputs for Perl script to integrate the information with differential expression. Differential expression analysis with CuffDiff identified genes that were significantly differentially regulated in each cell line. We used Perl to extract those differentially expressed genes that also showed an SRF active binding peak in their promoter regions. A linking table of UCSC gene names to common alias gene names were used in the Perl script for gene name mapping. As the output, a list of genes associated with the SRF active binding sites for each cell line were identified with the Perl script in one run (more details can be found in readme.txt for the perl script). We ran this script on four SRF active binding peak files. This produced two lists of genes, each of which was significantly over-expressed in one cell type versus the other, that had the observed SRF active binding sites in their promoter region (K562_up_out.txt and GM12878_up_out.txt). Two other lists were also produced, each of which were significantly over-expressed genes in one cell type, that had uniquely SRF active binding sites in their promoter region (K562_unique_up_out.txt and GM12878_unique_up_out.txt). Finally, GO term analysis using the lists of highly expressed genes in each cell type showed the significantly enriched pathways or functions that are related to SRF regulation.

Results:

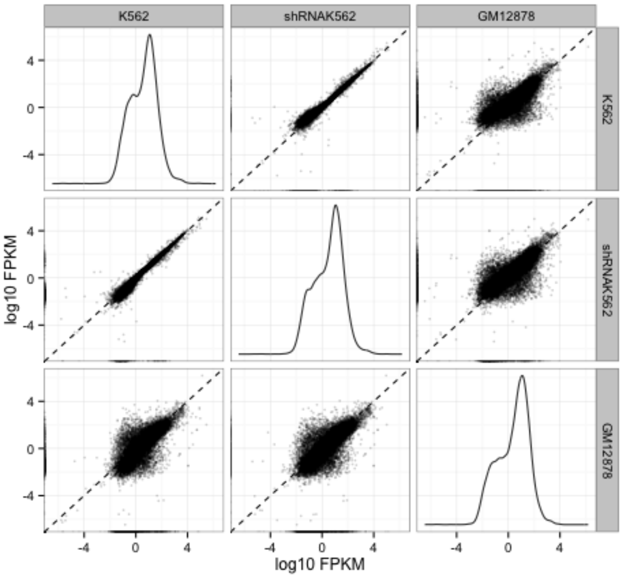
ChIP Peak Motif Results

SRF binding peaks in open chromatin and active histone marked regions are called for each cell type. There are a total of 10,382 peaks in K562 and 25,281 peaks in GM12878. Among those peaks, around 10% (1,147 peak regions for K562 and 2,055 peak regions for GM12878) were found to have CArG sites, the SRF binding motif. To get a better understanding of how SRF binding patterns are different between K562 and GM12878, a motif analysis on unique SRF binding peaks in open chromatin and active histone marked regions in each cell line was carried out. Among those unique peaks, 354 out of 7,720 regions could be explained by known motif in K562, 1,185 out of 9,725 regions could be explained by known motif in GM12878.

Binding of SRF to other motifs may indicate that in these cell lines, SRF interacts with other transcription factors and/or transcription co-factors, which alters the motif to which SRF preferentially binds. We are aware that the large number of peaks might be an artifact from the peak-calling threshold. Further analysis, such as imposing more stringent thresholds are suggested. With the help of MEME-ChIP online tool, we were able to perform a de novo motif finding for each cell line. Though co-activating motifs are highly conservative in both cell lines, some of them are pretty interesting.

RNA-seq Analysis Results

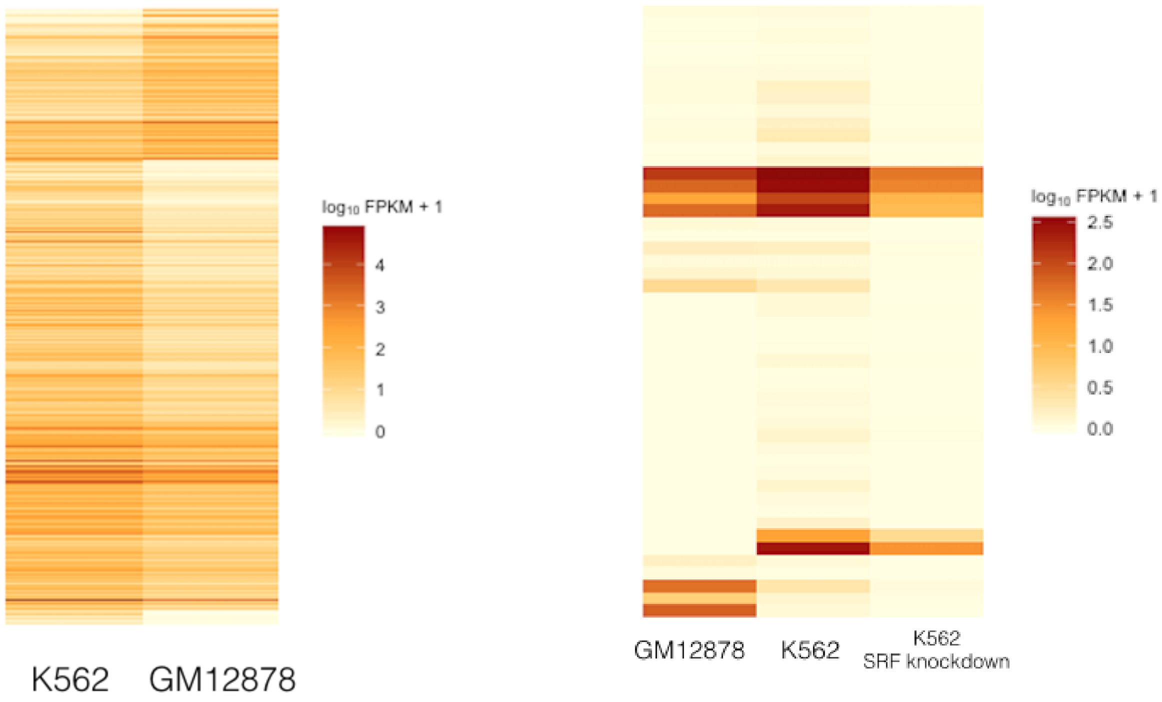
Figure1:Scatter plot of RNA-seq data



The RNA sequence analysis showed the K562 cell line to be highly correlated with knockdown of SRF in K562 cells as compared to the GM12878 cell line as expected, and shown in the scatterplot shown in Figure 1. Genes on or the y=x line are highly correlated in the two samples plotted against each other, whereas those deviating from the y=x line show differential expression. The differential expression is further visualized in the heatmaps in Figure2.

The differential expression between genes with a SRF site as well as an activating histone mark shown in the left Heatmap below, with genes clustered using both the row and column data, where indices are re-ordered based on the pairwise Jensen-Shannon distance of FPKM values using the csheatmap function in the cummeRbund R package. The heatmap to the right below, constructed using the same methodology, shows the effect in gene expression from the shRNA knockdown in K562 cells as compared to normal K562 and GM12878.

Figure 2 : Heatmaps of Gene Expression



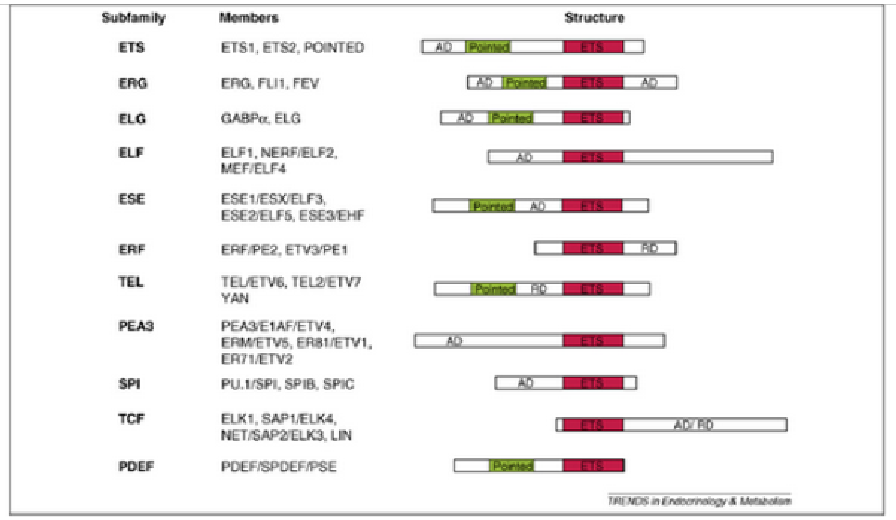
By combining the ChIP-seq data and expression data sorting for genes that are in the top 20% up-regulated range for K562 and GM12878 we revealed a large number of genes that were both positive for an activating mark in the promoter and highly differentially expressed. A difference of gene expression in 2,398 genes was seen when filtering for genes with SRF binding active histone marks and high differential expression. The differential expression of these genes is shown in the Heatmap in Figure 2.

Comparing an SRF knockdown to the SRF expressing K562 cell line there were 49 genes that showed a 50% reduction in expression ($\log_2$FPKM) less than or equal to -1. A Gene Ontology (GO) analysis of these genes revealed genes associated with metal ion bindings, specifically cation ion binding as well as ion channels and integral membrane proteins, suggesting regulation of ion transport as a potential role for SRF in K562 cells.

Discussion:

SRF is a ubiquitously expressed transcription factor that is known to interact with two families of cofactors which provide spatial and temporal specificity to SRF function. One family of SRF cofactors that regulates many growth-related genes is the E26 transformation-specific (Ets) family, which includes 11 diverse subfamilies (Figure 3). The Ets proteins form a complex with CArG-bound SRF through direct protein-protein interaction with SRF as well as binding to a DNA sequence directly upstream of the CArG box (Figure 4). This upstream sequence contains a conserved GGA core and each Ets protein has a different affinity to the sequences flanking the GGA core.

Another family of SRF cofactors is the Myocardin family. All members have the basic-rich domain which is required for protein-protein interaction with SRF. It is not yet known whether there is a DNA consensus motif analogous to the GGA core that promotes DNA binding of Myocardin family members. Studies have shown that Ets and Myocardin



TRENDS in Endocrinology & Metabolism

Figure 3. The Ets family of transcription factors and its 11 subfamily members (4) cofactors compete for binding to the same hydrophobic pocket of SRF, and thus their interaction with SRF is mutually exclusive. Unlike the Ets factors which can bind DNA without SRF, it is thought that the Myocardin family of cofactors cannot bind to DNA on their own; SRF must be bound to DNA first before the cofactors can interact with the SRF-DNA complex (1).

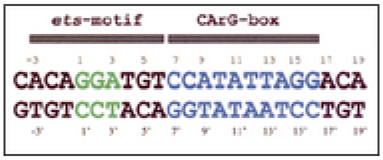


Figure 4: Binding Motifs of ETS and SRF Transcription Factors (9)

Though SRF has been shown to play an important role in many cell types, its function in cells of the hematopoietic lineage is poorly understood. The aim of our study was to determine whether and how SRF is involved in driving gene expression in myeloid and lymphoid blood cell lineages, K562 and GM12878, respectively. Our results suggest that SRF regulates pathways that are specific to each cell type and we identified candidate transcriptional cofactors that could potentially interact with SRF to mediate gene expression. From the list of genes that were up-regulated in each cell line based on our combined ChIP-seq and RNA-seq analysis, we used DAVID to generate a list of biological processes in which the gene are involved, which are shown in Table 1. As we expected, the pathways were specific to each cell line. For example, SRF was implicated in biological processes such as erythrocyte differentiation and homeostasis in the K562 cells in contrast to processes such as immune response, lymphocyte, and T cell activation in the GM12878 cells. An additional interesting observation comes from the RNA-seq analysis where the ion channel and ion binding genes were identified by gene ontology analysis as down regulated in SRF knockdown in the K562 cell line, this was not enriched in the other ontology analyses on genes upregulated by SRF. This provides an interesting potential future direction, as SRF may indirectly regulate these genes or direct regulation of these pathways may be masked by the large amount of binding by this ubiquitous transcription factor.

In accordance with our gene expression data, our motif analysis revealed candidate cofactors that may be mediating the differential gene expression; the transcription factors implicated by the most significant binding motifs in each cell line had known functions that were specific to myeloid or lymphoid lineages. For example, in K562 cells, the two most significant binding motifs are bound by Specificity Protein 1 (SP1) and GATA2. SP1 directly interacts with GABPα, which is an important transcription factor during the bifurcation of the megakaryocyte and erythroid lineages (3, 15). The GATA transcription factors also play an important role in the fate decision making of a Megakaryocyte/Erythroid Progenitor (MEP); when an MEP commits to become an erythrocyte, GATA1 is up-regulated while GATA2 is down-regulated and vice versa when it commits to the megakaryocyte lineage (11). In GM12878 cells, the most significant binding motif is bound by Ets Homologous Factor (EHF) and SPIB/PU.1, which are both Ets proteins. As shown in Figure X, EHF is a member of the ESE subfamily and SPIB/PU.1 is a member of the SPI subfamily, which acts as lymphoid-specific enhancer and is exclusively expressed in mature B-cells, T-cell progenitors, and plasmacytoid dendritic cells (10). Though we do not have evidence demonstrating direct interaction between SRF and these cofactors, we can speculate that SRF may be interacting with these cofactors implicated in the motif analysis since 1) in literature, most known functions of SRF are described in the context of a specific cofactor and 2) many cofactors on our list, though we only discussed a few, are Ets proteins which have the GGA core binding motif that is only several bases upstream of the CArG box (9).

Table 1. Top 10 biological processes in which the up-regulated genes are involved

	Unique to K562 cell line	**Unique to GM12878 cell line**
1	Chordate embryonic development	Leukocyte activation
2	embryonic development ending in birth or egg	immune response
3	hatching	cell activation
4	TGFB signaling pathway	lymphocyte activation
5	homeostasis of number of cells	regulation of small GTPase mediated signal transduction
6	pattern specification process	regulation of lymphocyte activation
7	enzyme linked receptor protein signaling pathway	protein kinase cascade
8	muscle cell differentiation	T cell activation
9	in utero embryonic development	regulation of leukocyte activation
10	erythrocyte differentiation	positive regulation of immune system process
	erythrocyte homeostasis	

Our computational analysis could have been improved if we had more time and disk space to run different ChIP-seq peak calling algorithm each with a range of different stringency cut-off values. This would help reduce the noise and reduce the chance that our findings are just artifacts of our peak calling strategy. In addition, our computational analysis would have been more significant if we used more technical and biological replicates, as this would help reduce noise and erroneous findings, such as false positives. In addition gathering to gathering replicates we could add additional cell types to our study, to generate more informative data. However GM12878 and K562 cell lines were appropriate for this type of project as they both have a large amount of publicly available data and we had to rely on data generated by other research groups for this study. Finally, this analysis could have been bolstered by having more control over the experimental conditions under which these cell lines were cultured. Having as uniform conditions as possible, between GM12878 and K562, would decrease noise and erroneous findings.

References:

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

2. Charles E. Grant, Timothy L. Bailey, and William Stafford Noble, "FIMO: Scanning for occurrences of a given motif", *Bioinformatics* 27(7):1017–1018, 2011.

3. Galvagni F, Capo S, Oliviero S. Sp1 and Sp3 physically interact and co-operate with GABP for the activation of the utrophin promoter. J Mol Biol. 2001 Mar 9;306(5):985-96.

4. Gutierrez-Hartmann A, Duval DL, Bradford AP. ETS transcription factors in endocrine systems. Trends Endocrinol Metab. 2007 May-Jun;18(4):150-8.

5. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002, 12(6): 996-1006.

6. Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012, 9: 357-359.

7. Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup. The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics. 2009, 25: 2078-2079.

8. Miano JM, Long X, Fujiwara K. Serum response factor: master regulator of the actin cytoskeleton and contractile apparatus. Am J Physiol Cell Physiol. 2007 Jan;292(1):C70-81.

9. Mo Y, Ho W, Johnston K, Marmorstein R. Crystal structure of a ternary SAP-1/SRF/c-fos SRE DNA complex. J Mol Biol. 2001 Nov 30;314(3):495-506.

10. Montes-Moreno S, Ramos-Medina R, Martínez-López A, Barrionuevo Cornejo C, Parra Cubillos A, Quintana-Truyenque S, Rodriguez Pinilla SM, Pajares R, Sanchez-Verde L, Martinez-Torrecuadrada J, Roncador G, Piris MA. SPIB, a novel immunohistochemical marker for human blastic plasmacytoid dendritic cell neoplasms: characterization of its expression in major hematolymphoid neoplasms. Blood. 2013 Jan 24;121(4):643-7.
Philip Machanick and Timothy L. Bailey, "MEME-ChIP: motif analysis of large DNA datasets",*Bioinformatics* 27(12):1696-1697, 2011.

11. Philipsen S. A new twist to the GATA switch. Blood. 2013 Nov 14;122(20):3391-2.

12. Quinlan, A. R. and I. M. Hall (2010). "BEDTools: a flexible suite of utilities for comparing genomic features." Bioinformatics **26**(6): 841-842.

13. Rosenbloom KR, Sloan CA, Malladi VS, Dreszer TR, Learned K, Kirkup VM, Wong MC, Maddren M, Fang R, Heitner SG, Lee BT, Barber GP, Harte RA, Diekhans M, Long JC, Wilder SP, Zweig AS, Karolchik D, Kuhn RM, Haussler D, Kent WJ. ENCODE data in the UCSC Genome Browser: year 5 update. Nucleic Acids Res. 2013. 41: D56-63.

14. Sun Q, Chen G, Streb JW, Long X, Yang Y, Stoeckert CJ Jr, Miano JM. Defining the mammalian CArGome. Genome Res. 2006 Feb;16(2):197-207.

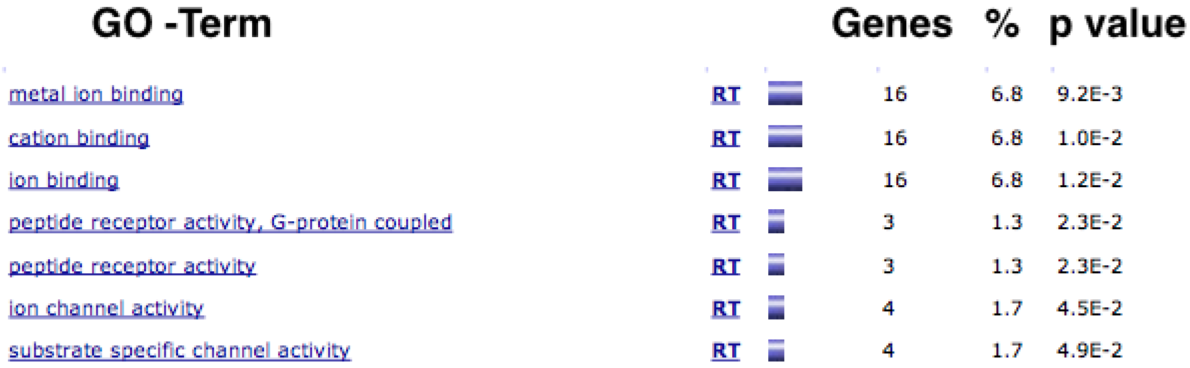
15. Tijssen MR, Cvejic A, Joshi A, Hannah RL, Ferreira R, Forrai A, Bellissimo DC, Oram SH, Smethurst PA, Wilson NK, Wang X, Ottersbach K, Stemple DL, Green AR, Ouwehand WH, Göttgens B. Genome-wide analysis of simultaneous GATA1/2, RUNX1, FLI1, and SCL binding in megakaryocytes identifies hematopoietic regulators. Dev Cell. 2011 May 17;20(5):597-609.

16. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics doi:10.1093/bioinformatics/btp120

17. Zhang et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biol. 2008, 9(9): R137.

Supplementary figure:

Table1. DAVID gene ontology analysis of genes downregulated in K562 cells after knockdown of SRF by shRNA



GO -Term			Genes	%	p value
metal ion binding	RT	▬	16	6.8	9.2E-3
cation binding	RT	▬	16	6.8	1.0E-2
ion binding	RT	▬	16	6.8	1.2E-2
peptide receptor activity, G-protein coupled	RT	▪	3	1.3	2.3E-2
peptide receptor activity	RT	▪	3	1.3	2.3E-2
ion channel activity	RT	▪	4	1.7	4.5E-2
substrate specific channel activity	RT	▪	4	1.7	4.9E-2