

# Uncovering Mechanisms of Transcriptional Regulations by Systematic Mining of Cis Regulatory Elements with Gene Expression Profiles

Qicheng Ma<sup>1§</sup>, Gung-Wei Chirn<sup>1</sup>, Joseph D. Szustakowski<sup>1</sup>, Adel Bakhtiarova<sup>3</sup>,  
Penelope A Kosinski<sup>2</sup>, Daniel Kemp<sup>2</sup>, N.R. Nirmala<sup>1</sup>

<sup>1</sup>Bioinformatics Cambridge, Quantitative Biology, Developmental and Molecular Pathways, Novartis Institutes For Biomedical Research Inc, 250 Massachusetts Avenue, Cambridge, MA 02139 USA.

<sup>2</sup>Diabetes and Metabolism Disease Area, Novartis Institutes For Biomedical Research Inc, 250 Massachusetts Avenue, Cambridge, MA 02139 USA

<sup>3</sup>Genentech Inc, 1 DNA Way South San Francisco, CA 94080 USA

<sup>§</sup>Corresponding author

Email addresses:

QM: [Qicheng.Ma@novartis.com](mailto:Qicheng.Ma@novartis.com)

GWC: [Gung-Wei.Chirn@novartis.com](mailto:Gung-Wei.Chirn@novartis.com)

JDS: [Joseph.Szustakowski@novartis.com](mailto:Joseph.Szustakowski@novartis.com)

AB: [Bakhtiarova.Adel@gene.com](mailto:Bakhtiarova.Adel@gene.com)

PAK: [Penny.Kosinski@novartis.com](mailto:Penny.Kosinski@novartis.com)

DK: [Daniel.Kemp@novartis.com](mailto:Daniel.Kemp@novartis.com)

NRN: [Nanguneri.Nirmala@novartis.com](mailto:Nanguneri.Nirmala@novartis.com)



# **Abstract**

## **Background**

Contrary to the traditional biology approach, where the expression patterns of a handful of genes are studied at a time, microarray experiments enable biologists to study the expression patterns of many genes simultaneously from gene expression profile data and decipher the underlying hidden biological mechanism from the observed gene expression changes. While the statistical significance of the gene expression data can be deduced by various methods, the biological interpretation of the data presents a challenge.

## **Results**

A method, called CisTransMine, is proposed to help infer the unobserved underlying biological mechanisms for the observed gene expression changes in microarray experiments. Specifically, this method will predict potential cis-regulatory elements in promoter regions which could regulate gene expression changes. This approach builds on the MotifADE method published in 2004 and extends it with two modifications: up-regulated genes and down-regulated genes are tested separately and in addition, tests have been implemented to see whether combinations of two transcriptional factors could work synergistically.

## **Conclusions**

The method has been applied to a genome wide expression dataset intended to study the mouse C2C12 cell differentiation. The results shown here both confirm the prior biological knowledge and facilitate in the discovery of new biological insights. The

results validate that the CisTransMine approach is a robust method to uncover the hidden regulatory mechanisms at the transcriptional level and it can potentially facilitate in discovering new mechanisms of transcriptional regulation.

## **Background**

High-throughput microarray experiments have modernized biological experiments by enabling measurements of expression levels for genes on the genome scale under different conditions. Usually, hundreds or thousands of genes may change their expression values between conditions for many reasons, some of which could be either the up-regulation or down-regulation by transcriptional factors or their co-factors. It is challenging to be able to interpret these changes in the biological context. Understanding the transcription regulation mechanisms between transcriptional factors and their target genes is one of the key ways in which to formulate hypotheses governing the root causes of the observed changes.

Unveiling mechanisms of transcription regulation is an active bioinformatics research area. Different approaches have been proposed to discover mechanisms of transcription regulation. Bayesian network approaches have been applied [1] to integrate the motif discovery in the promoters and the analysis of gene expression data. Some approaches [2] split motifs and gene expression values of regulators and build a decision tree based on the combination of expression ratios of transcription factors and presence/absence of the motifs. Yet other approaches [3-4] fit gene expression data to a linear model using weights depending on whether a transcriptional factor is an inducer or repressor. Mootha *et. al.* [5] use a two-tailed non-parametric Mann-Whitney rank sum test to determine significances of motifs in promoter regions. The MotifADE method [5] assumes that if up-regulated or down-

regulated genes which contain certain transcriptional factors are tested to be statistically significant, their expression level changes could be explained by those transcriptional factors which regulate them; on the other hand, if genes which contain certain transcriptional factors are tested not to be statistically significant, there may not be any association between genes and transcriptional factors. In our hands, we have observed that two-tailed non-parametric Mann-Whitney rank sum tests used by MotifADE method cannot detect significances of transcriptional factors if they induce the transcription of some genes and repress the transcription of other genes at the same time. We have therefore extended the MotifADE method to investigate up-regulated and down-regulated genes separately since a transcriptional factor may simultaneously enhance the transcription of certain genes and inhibit the transcription of other genes. Synergistic effects of two transcriptional factors are also detected. The CisTransMine method is applied to the mouse C2C12 differentiation dataset [6]. It not only reveals novel transcription mechanisms for known transcriptional factors, but also assists in identifying a novel transcriptional factor binding site which regulates known target genes. These results demonstrate that the CisTransMine method is an important tool to discover unknown transcription regulation mechanisms, thus facilitating in extending biological knowledge.

## **Results**

### **Results for Known Transcriptional Factors**

We use the mouse C2C12 cell differentiation dataset as a test case [6]. In this experiment, the mouse C2C12 myoblast cells were induced to differentiate from myoblasts to myotubes in order to model the late stage myogenesis. Cells were cultured in 6-well plates. Induction of differentiation of the C2C12 myoblasts was initiated at Day 0 when cells were confluent by reducing the serum concentration in

the wells to 3% v/v. Upon induction of differentiation these mononucleate cells exited the cell cycle and fused to form myotubes. Cells were lysed for RNA preparation. The expression level was measured at eight time-points, with three replicates per time point at days -1, 0, 0.25, 1, 2, 3, 4, 5 post induction. The goal is to identify genes involved in myogenesis. Figure 1 shows gene expression profiles across all time points. It can be observed that the major switch in the expression profiles occurs between Day 1 and Day 2.

The CisTransMine algorithm was run on this dataset comparing expression profiles between different time points. Table 1 shows that the top 15 transcriptional factors(TF) for up-regulated genes that could potentially be statistically significant regulators in muscle differentiation between the day 1 and day 2 timepoints. The top TF among the up-regulated genes is MYOD, which is consistent with the observation that MYOD prepares myoblasts for efficient differentiation. [7]. Figure 2 shows the distribution of moderated t-values in up-regulated genes with the MYOD binding elements in their promoter regions. SRF (serum response factor), the second hit in table 1, is required for skeletal muscle growth and maturation [8]. The fourth hit is the C/EBP and C-Jun heterodimer denoted by CREBP1/CJUN, where the transcriptional factor C/EBP can activate differentiation-specific genes [9]. MEF2, which is implicated in the muscle contraction process [10], is also enriched since the muscle contraction pathway is up-regulated [6]. Similarly, other hits in this list can be explored through the literature to assess their known roles in muscle differentiation. In the case of enriched transcriptional factors which are not supported by the literature, one could hypothesize that these could be either new biological knowledge yet to be discovered or false positives, since any such algorithm has a certain false positive

rate. It must be noted that there are no synergistic transcription factors which are statistically significant in up-regulated genes from Day 1 to Day 2.

Table 2 shows a list of statistically significant transcriptional factors in down-regulated genes from Day 1 and Day 2. The transcriptional factors E2F1 and MYC, which regulate the cell cycle process, are the top enriched transcriptional factors among the down-regulated genes - which implicates the E2F1 and MYC involvement in down-regulation of cell cycle genes from Day 1 to Day 2 since the cell cycle pathway is down-regulated [6]. And p53-mediated transcriptional repression of cell cycle genes depends on the activities of E2F and NFY [11]. The Foxm1 gene is critical for G1/S transition and essential for mitotic progression [12]. Table 3 illustrates significant synergistic transcriptional factors in down-regulated genes from Day 1 to Day 2. The top interaction pair of transcriptional factors are CREB and MYC since CREB negatively regulates MYC in the cell cycle G1-S transition phase [13]. As discussed before, the transcription factors which cannot be explained by prior biological knowledge could be false positives or potentially new regulatory mechanisms that need to be validated by experiment. Some such factors are tested and presented in this paper as seen below.

The results above show that biologically significant transcription factors involved in muscle differentiation also show statistical significance in the gene expression profiling experiment. Thus one can use CisTransMine to tease out important regulatory processes that are in play under a given perturbation to a system.

### **Results for Unknown Transcriptional Factors**

This method was also used to discover novel regulatory elements from this experiment [6]. The elucidation of novel regulatory motifs in the context of a

specific cellular function may reveal new pathways and targetable mechanisms related to disease settings. In this paper, the terms “motifs” and “transcriptional factor binding sites” are used interchangeably. Motifs that emerged as potential regulatory elements with statistical significance were screened for functional relevance via application of luciferase reporter gene assay technology. Specifically, motifs were selected in the context of the genes that have a known role in myogenic differentiation and functional pathways that are regulated such as contractility, cell cycle, and mRNA splicing in addition to their statistical significances. The 400 bp DNA sequence surrounding the chosen motifs were analyzed using Transfac for additional transcription factor binding sites, that could potentially influence and complex with the transcription factor identified to bind the unknown novel motif. Table 4 lists the details for tested motifs and other known transcriptional factors within 400bp DNA sequences surrounding the chosen motifs.

To test for regulatory activity of selected motifs using a reporter gene assay approach, 400 bp sequences were generated by PCR using appropriate primers, and using XhoI restriction sites, these fragments were cloned into the pGL3 promoter reporter vector to assay their transcriptional activity. This relatively large promoter sequence was used due to the potential requirement for contextual surrounding elements for motif function/activity. A 400 bp fragment of the pck2 gene surrounding the motif, GCGGAGGC, was cloned from the pck2 promoter into pGL3 promoter firefly luciferase vector and was used to transfect C2C12 myoblasts along with the pGL4.75 renilla luciferase vector for transfection efficiency. The cells were then split into two plates, cells on one plate were induced to differentiate and the other plate was maintained as undifferentiated myoblasts. Cells transfected with the pGL3 promoter vector without the construct (control), expressed some reporter gene activity, and that

reporter activity increased eight fold over the control in the cells transfected with the same vector containing the 400 bp pck2 gene promoter fragment containing the motif, GCCGAGGC (Figure 3A). In order to assess the activity specifically mediated by the motif, the sequence was mutated by random nucleotide substitution, and two different mutant sequences were generated, mutant1 (acgctatc) and mutant2 (ctgcacgc). These mutations led to an increase in the reporter activity beyond that of the wildtype motif/promoter, up to twelve fold compared to control. The potential function of this motif, as a negative regulator of gene expression, is consistent with the expression pattern of the pck2 gene within the myogenic program. In contrast, reporter gene activity in C2C12 cells transfected with the pGL4.15 basic vector containing 400 bp of the myogenin promoter with the motif CGACCCGT did not change after mutations were introduced (Figure 3B). Thus, it was deemed that this particular motif has little if no functional role in the myogenin promoter.

This experiment demonstrated the potential of this method to successfully identify novel functional motifs. Such an approach may be extended to differential gene expression within a variety of disease-related settings and cell types, with potential relevance to disease pathway discovery.

## **Discussion**

In the post-genomics area, there is a sea of biological data including microarray experimental data. This provides an unprecedented opportunity and challenge to fully decipher the underlying biological system. One aspect of this analysis is to analyze significantly enriched pathways where coordinated subtle expression changes among genes within the pathway can be easily identified while these coordinated subtle expression changes can be overlooked by the gene-centric approach [14]. Though the

pathway analysis provides a way to see “forests, not individual trees”, it can not address the transcription regulation mechanisms which explain gene expression level changes that we have measured by microarray chips. Thus deciphering transcription regulation mechanisms help characterize the underlying biological process. Different approaches have been proposed to help decipher transcription regulation mechanisms including Bayesian networks, decision trees, and regression models. In this paper, the CisTransMine method has been implemented to identify transcriptional factors involved in biological processes through the analysis of microarray data. The CisTransMine method was applied to the in-house gene expression profiling of mouse C2C12 skeletal muscle myoblast differentiation to myotubes. It not only confirms some known biological knowledge but also reveals potentially novel biological insights. However, due to the filtering of low expressed genes and the limited information on curated transcription regulation relationships, only 6961 genes were included in the calculation process. Thus, some of the enriched transcriptional factors which are not supported by the literature could be either new biological knowledge yet to be discovered or false positives. As more and more transcriptional factors and their target genes are discovered, we will have more coverage on the transcriptional regulation relationships which will result in more comprehensive prediction results.

## **Conclusions**

In summary, preliminary results identified the relevant transcriptional factors involved and demonstrated the potential application of this method to the analysis of microarray data in order to identify the regulatory mechanisms involved in the biological experiment under investigation. Thus it illustrates that the CisTransMine

method can facilitate the discovery of novel regulation mechanisms and extend our knowledge of biological pathways.

## **Methods**

### **Preparation for promoter sequences.**

The human, mouse and rat promoter sequences were extracted from the genome assembly as of January 2006. The location of the transcriptional start site was approximated by the first nucleotide in the RefSeq mRNA transcript sequence. For each gene, promoter sequences with respect to their transcripts were extracted according to coordinates of first exons for corresponding transcripts. For each transcript, the region from -2000 bp to +300 bp with respect to the transcriptional start site was extracted. A gene may have several different transcripts, therefore several promoters.

The promoter sequences were masked against repetitive sequences, e.g., LINEs and SINEs with the RepeatMasker program to avoid any Transfac [15] matrix search hits in those repetitive regions. Then orthologous promoter sequences were aligned together with Wconsensus [16]. The orthologous relationships were defined in the NCBI Homologene database as of January 2006. For those promoters with orthologous promoters in human, mouse and rat, a sliding window of 10 nucleotides was used and non-conserved regions were masked out where promoter sequence identities among orthologous promoter sequences had a length of less than 5 nucleotides within a 10 nucleotide window.

### **Annotation of promoter sequences.**

Human-curated transcriptional factor binding sites from the Transfac database were used to record each transcription factor and its regulated genes for human, mouse and

rat promoter sequences. In addition, the GeneGo Metacore database [17], which also records transcription regulation relationships besides other relationships between a pair of genes as curated from the literature, were used to identify each transcriptional factor and its regulated genes. Totally, there are a total of 601 human transcriptional factors, 518 mouse transcriptional factors, and 302 rat transcriptional factors in our collection.

#### **Extraction of unknown transcriptional factor binding sites.**

Promoter sequence regions which have been annotated as known transcriptional factor binding sites were masked out. The remaining regions contain potentially novel transcriptional factor binding sites. All possible non-degenerative conserved 8-mer and 9-mer motifs which have at least 5 identical nucleotides within a 10 nucleotide window among human, mouse and rat promoter sequences were enumerated. Their true significances would be evaluated in biological experiments.

#### **Normalization of Affymetrix GeneChip Arrays.**

High density oligonucleotide microarrays are often used to measure gene expression values on a genome scale. Affymetrix GeneChip arrays are the most popular microarray platform. On an Affymetrix GeneChip array, each transcript may be represented by one or more than one probe set. Each probe set may contain 11-20 pairs of oligonucleotide 25-mer probes. The first group of these probes are designed to match certain region of target genes and are called Perfect Match. The second group of these probes are designed to capture the noise and are called Mismatch. Because of the non-specific binding and noise introduced by the scanner, it is typical to remove the non-biological noise by normalization of chips within the experiment. Normalization in our analysis was carried out using the GC-RMA normalization method which has relatively good performance among different normalization

methods [18]. GC-RMA normalization results are on the log scale. These results are then trim-scaled so that the median raw value from the 2nd percentile and the 98th percentile for all the probe sets is 150. We removed probe sets which have their average raw values among replicates less than 100 for both conditions.

### **Calculation of the moderated t statistic for each probe set**

The traditional student t-test statistic is often used to assess the significance of individual probe sets between two conditions, e.g., treatment group versus control group. However, there are usually only a few replicates (usually three) within each group. Given such a small sample size, it is difficult to estimate the variance reliably. This makes the estimation of the t-statistic problematic. To address this problem, the moderated t-test [19] implemented in the Limma package within the Bioconductor package [20] is adopted to evaluate the significance of individual probe sets between the two groups. The moderated t-test assumes the same distribution for the error variance of all genes in order to estimate the variance of an individual gene with an empirical Bayes method, using posterior residual standard deviations instead of traditional standard deviations, to accommodate for the low number of replicates for each group[19]. Up-regulated genes and down-regulated genes have positive and negative moderated t-values respectively. If a gene is represented by several probe sets, the moderated t-statistic with the highest absolute value is used to represent the moderated t- statistic for that gene.

### **Evaluation of the significance of a single motif**

The CisTransMine method is built on MotifADE. MotifADE provides a framework to identify significant transcriptional factor binding sites enriched between two microarray conditions. Particularly, it uses a two-tailed non-parametric Mann-Whitney (Wilcoxon) rank sum U statistic to evaluate the significance of a motif.

Specifically, for each motif, moderated t-statistics for all the genes are divided into two samples: one sample containing moderated t-statistics for genes having the motif of interest in their promoter region and the other sample for genes not having the motif in their promoter regions. The null hypothesis is that there is no difference between the means of the ranks of these two sets of moderated t-statistics; the alternative hypothesis is that the means of the ranks of these two sets are not equal, i.e., genes containing the motif are either up-regulated or down-regulated (Figure 4) .

In the case where a transcriptional factor may enhance the transcription of certain genes and repress the transcription of other genes at the same time, the two-tailed Mann-Whitney test might obscure such contexts. Under this situation, a two-tailed Mann-Whitney test cannot detect the significance of that motif since the two-tailed Mann-Whitney test calculates for a given motif, the rank sum for all genes having that motif whether they are up-regulated genes, down-regulated genes or non-regulated genes. If there is an approximately equal number of up- and down-regulated genes with a particular motif, the statistical significance of the up-regulated genes will be more or less cancelled out by the statistical significance of the down-regulated genes. As a result the motif contained in those genes will compute to be statistically insignificant. For example, in Figure 5, Motif 1 and Motif 3 would have the same p-values with the two-tailed Mann-Whitney test since only the t-value 0.9 is important and all other t-values from Motif 1 or Motif 3 are symmetric with respect to 0 contributing the same to the rank sum as does t-value 0 even though Motif 1 is more significant than Motif 3, as there are several genes containing motif 1 that are more highly down- or up-regulated relative to the extremes of the genes containing motif 3. An approach using absolute values was implemented to solve this problem [21] where the absolute enrichment can identify important gene sets that may not be identified by

two-tailed methods. The CisTransMine method is proposed to test up-regulated genes and down-regulated genes separately for statistical significance by using the one-tailed non-parametric Mann-Whitney test. For up-regulated genes, the null hypothesis is that the mean of the ranks in the up-regulated genes containing the motif is equal to the mean of ranks in the up-regulated genes not containing the motif; the alternative hypothesis is that the mean of ranks in the up-regulated genes containing the motif is greater than the mean of ranks in the up-regulated genes not containing the motif. Similarly, for down-regulated genes, the null hypothesis is that the mean of the ranks in the down-regulated genes containing the motif is equal to the mean of ranks in the down-regulated genes not containing the motif; the alternative hypothesis is that the mean of ranks in the down-regulated genes containing the motif is less than the mean of ranks in the down-regulated genes not containing the motif. Thus, significances for motifs in up-regulated genes and down-regulated genes are tested separately.

### **Synergistic motifs.**

In eukaryotic genomes, the synergistic relationship is present when multiple transcriptional factors work in concert to regulate target genes, e.g., combinatorial activities of multiple transcriptional factors regulate the B cell lineage commitment and differentiation [22]. In the CisTransMine method, synergistic relationships between two transcriptional factors can be detected in a two-step process. The first step, all the genes containing transcriptional factor A binding sites ( $TF_A$ ) and transcriptional factor B binding sites ( $TF_B$ ) in the promoter regions can be denoted by  $TF_A \cap TF_B$ , which is a subset of genes containing both types of binding sites. All the genes containing transcriptional factor A binding sites but not transcriptional factor B binding sites can be denoted by  $TF_A - TF_B$ . All the genes containing transcriptional factor B binding sites but not transcriptional factor A binding sites can be denoted by

$TF_B - TF_A$ . For up-regulated genes, the necessary conditions for the true synergy between two transcriptional factors to exist are that (1) one-tailed Wilcoxon rank sum test P-value between genes in the set of  $TF_A \cap TF_B$  and the genes in the set of  $TF_A - TF_B$  is less than 0.05, (2) one-tailed Wilcoxon rank sum test P-value between genes in the set of  $TF_A \cap TF_B$  and the genes in the set of  $TF_B - TF_A$ , is less than 0.05. For down-regulated genes, the necessary conditions for the true synergy between two transcriptional factors to exist are that (1) one-tailed Wilcoxon rank sum test P-value between genes in the set of  $TF_A \cap TF_B$  and the genes in the set of  $TF_A - TF_B$  is less than 0.05, (2) one-tailed Wilcoxon rank sum test P-value between genes in the set of  $TF_A \cap TF_B$  and the genes in the set of  $TF_B - TF_A$  is less than 0.05. If the necessary condition is satisfied, the algorithm proceeds to the second step where the significance of the synergistic relationship between the two transcriptional factors is tested with the same method as that for the single motif with the one-tailed Wilcoxon rank sum test.

### **Multiple testing correction.**

This method usually tests the significances of several hundred single motifs, and more than ten thousand synergistic motifs simultaneously. Thus several thousand null hypotheses are tested at the same time. In order to reduce the false positive rate, multiple testing correction method must be applied to take into account that thousands of null hypotheses are tested at the same time. The multiple testing correction method we adopt is the False Discovery Rate (FDR) q-value [23]. The FDR q-value is a measure of the rate of false discovery from the distribution of p-values. The FDR q-value method is chosen since it can balance between the specificity and the sensitivity without a priori p-value cutoff (see reference for details).

## Competing interests

The authors declare that they do not have competing interests.

## Authors' contributions

QM carried out the design and implementation of the algorithm and wrote the manuscript. GWC provided the mapping of the Affymetrix probeset to the NCBI refseq sequence. JDS did the quality control of the Affymetrix chips. AB, PAK, and DK did the web lab work. NRN directed and participated in the project. All authors involved in reviewing and revising the manuscript and approved the final manuscript.

## Acknowledgements

We thank Liam O'Connor and Richard Cai for reviewing the manuscript and Leah Martell for the statistics help.

## References

1. Segal E, Yelensky R, Koller D: **Genome-wide Discovery of Transcriptional Modules from DNA Sequence and Gene Expression.** *Bioinformatics* 2003, 19(Suppl 1): 273-282
2. Middendorf M, Kundaje A, Wiggins C, Freund Y, Leslie C: **Predicting Genetic Regulatory Response Using Classification.** *Bioinformatics* 2004, 20(Suppl 1): 1232-1240
3. Chen KC, Wang TY, Tseng HH, Huang CYF, Kao CY: **A stochastic differential equation model for quantifying transcriptional regulatory**

- network in *Saccharomyces cerevisiae*.** *Bioinformatics* 2005, 21(12) 2883-2890
4. Stephen Yeung MK, Tegnér J, and Collins JJ: **Reverse engineering gene networks using singular value decomposition and robust regression.** *Proc Natl Acad Sci U S A* 2002, 99 (9) 6163-6168
  5. Mootha VK, Handschin C, Arlow D, Xie X, St Pierre J, Sihag S, Yang W, Altshuler D, Puigserver P, Patterson N, Willy PJ, Schulman IG, Heyman RA, Lander ES, Spiegelman BM: **Erralpha and Gabpa/b specify PGC-1alpha-dependent oxidative phosphorylation gene expression that is altered in diabetic muscle.** *Proc Natl Acad Sci U S A* 2004, 101(17):6570-6575
  6. Szustakowski, J.D., Lee, J, Marrese, C.A., Kosinski, P.A., Nirmala, N.R. and Kemp, D.M: **Identification of novel pathway regulation during myogenic differentiation.** *Genomics* 2006, 87(1):129-138
  7. Ishibashi J, Perry RL, Asakura A, Rudnicki MA: **MyoD induces myogenic differentiation through cooperation of its NH<sub>2</sub>- and COOH-terminal regions.** *J Cell Biol* 2005, 171(3):471-82
  8. Li S, Czubryt MP, McAnally J, Bassel-Duby R, Richardson JA, Wiebel FF, Nordheim A, Olson EN: **Requirement for serum response factor for skeletal muscle growth and maturation revealed by tissue-specific gene deletion in mice.** *Proc Natl Acad Sci U S A* 2005, 102(4):1082-7
  9. Johnson PF: **Molecular stop signs: regulation of cell-cycle arrest by C/EBP transcription factors.** *J Cell Sci* 2005, 118(12):2545-55
  10. Silva JL, Giannocco G, Furuya DT, Lima GA, Moraes PA, Nacheff S, Bordin S, Britto LR, Nunes MT, Machado UF: **NF-kappaB, MEF2A, MEF2D and HIF1-a involvement on insulin- and contraction-induced regulation of**

- GLUT4 gene expression in soleus muscle.** *Mol. Cell Endocrinol* 2005, 240(1-2):82-93
11. Tabach Y, Milyavsky M, Shats I, Brosh R, Zuk O, Yitzhaky A, Mantovani R, Domany E, Rotter V, Pilpel Y: **The promoters of human cell cycle genes integrate signals from two tumor suppressive pathways during cellular transformation.** *Mol Syst Biol* 2005, 1:0022
12. Wang IC, Chen YJ, Hughes D, Petrovic V, Major ML, Park HJ, Tan Y, Ackerson T, Costa RH: **Forkhead box M1 regulates the transcriptional network of genes essential for mitotic progression and genes encoding the SCF (Skp2-Cks1) ubiquitin ligase.** *Mol Cell Biol* 2005, 25(24):10875-94
13. Rajabi HN, Baluchamy S, Kolli S, Nag A, Srinivas R, Raychaudhuri P, Thimmapaya B: **Effects of depletion of CREB-binding protein on c-Myc regulation and cell cycle G1-S transition.** *J Biol Chem* 2005, 280(1):361-74
14. Curtis RK, Oresic M and Vidal-Puig A: **Pathways to the analysis of microarray data.** *Trends Biotechnol* 2005, 23(8):429-35
15. Matys, V., Fricke, E., Geffers, R., Gößling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V., Kloos, D.U., Land, S., Lewicki-Potapov, B., Michael, H., Münch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S. and Wingender, E: **TRANSFAC: transcriptional regulation, from patterns to profiles.** *Nucleic Acids Res* 2003, 31(1): 374-378
16. Hertz GZ, Stormo GD: **Identifying DNA and protein patterns with statistically significant alignments of multiple sequences.** *Bioinformatics* 1999, 15(7-8):563-77

17. Ekins S, Nikolsky Y, Bugrim A, Kirillov E, Nikolskaya T: **Pathway mapping tools for analysis of high content data.** *Methods Mol Biol* 2007, 356:319-50
18. Irizarry RA, Wu JZ, Jaffee HA: **Comparison of Affymetrix GeneChip expression measures.** *Bioinformatics* 2007, 22 (7): 789-794
19. Smyth GK: **Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments.** *Statistical Applications in Genetics and Molecular Biology* 2004, 3(1): Article 3
20. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol* 2004, 5(10): R80
21. Saxena V, Orgill D, Kohane I: **Absolute enrichment: gene set enrichment analysis for homeostatic systems.** *Nucleic Acids Res* 2006, 34(22): e151
22. Nutt SL, Kee BL: **The transcriptional regulation of B cell lineage commitment.** *Immunity* 2007, 26(6):715-25
23. Storey JD, Tibshirani R: **Statistical significance for genomewide studies.** *Proc Natl Acad Sci U S A* 2003, 100(16): 9440-9445

# Figures

Figure 1 - Gene expression profiles across all time points

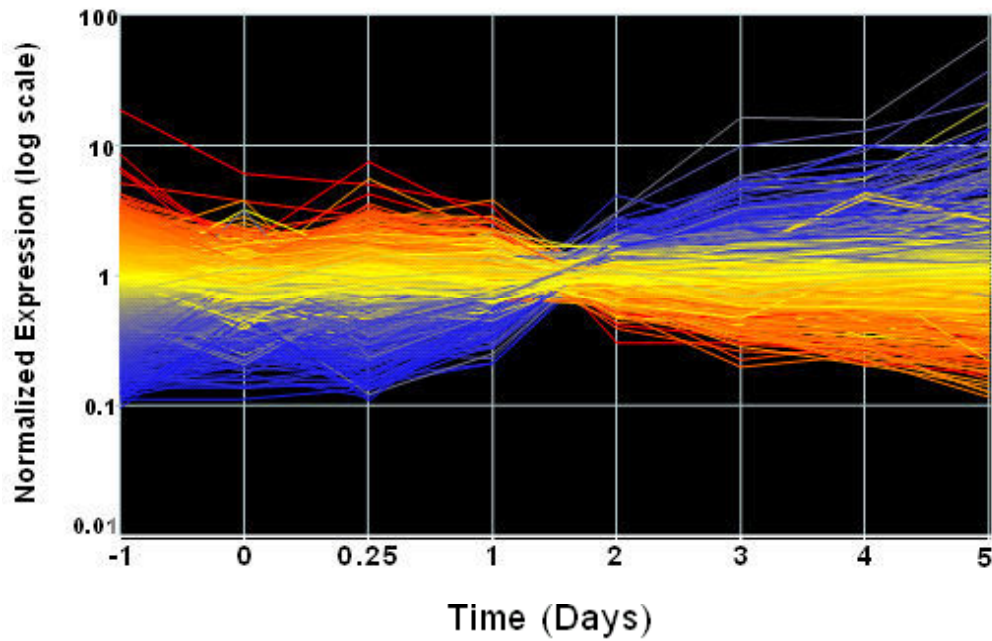


Figure 1. Gene expression profiles across eight time points. It can be observed that major changes occur from Day 1 to Day 2.

Figure 2 - The distribution of moderated t-values for up-regulated genes containing Myod binding elements in the promoter regions

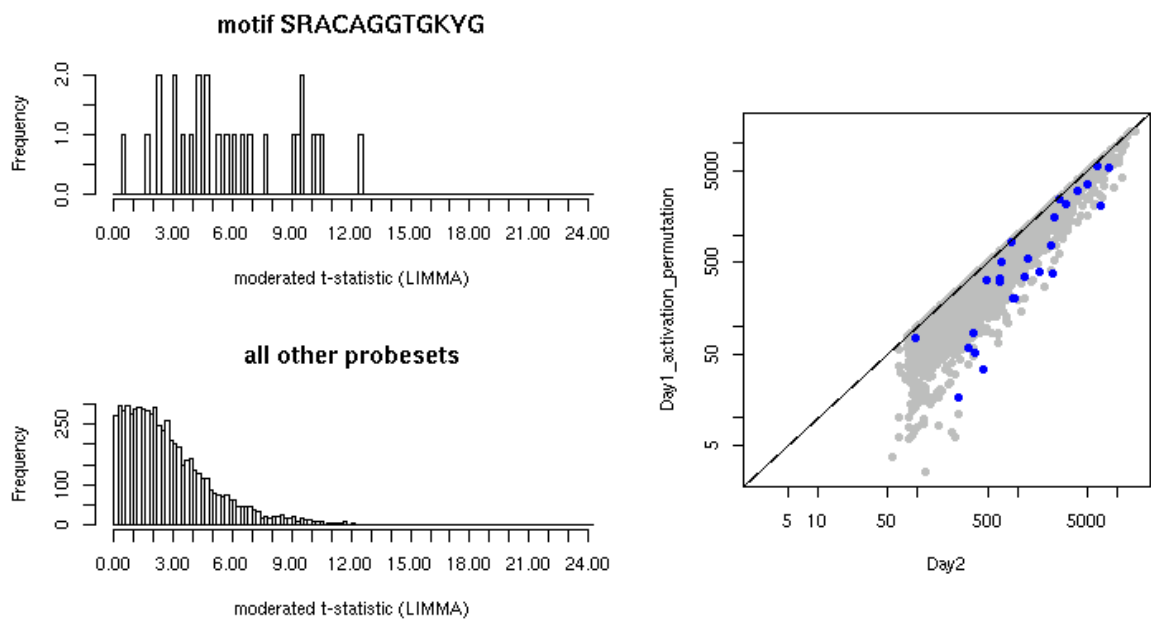


Figure 2. The top histogram shows the distribution of moderated t-values for up-regulated MYOD target genes (also depicted as blue dots in the scatter plot), and the bottom histogram shows the distribution of moderated t-values gene expression profiles across all time points for all other up-regulated genes (also depicted as grey dots in the scatter plot).

### Figure 3 - Luciferase reporter assay results

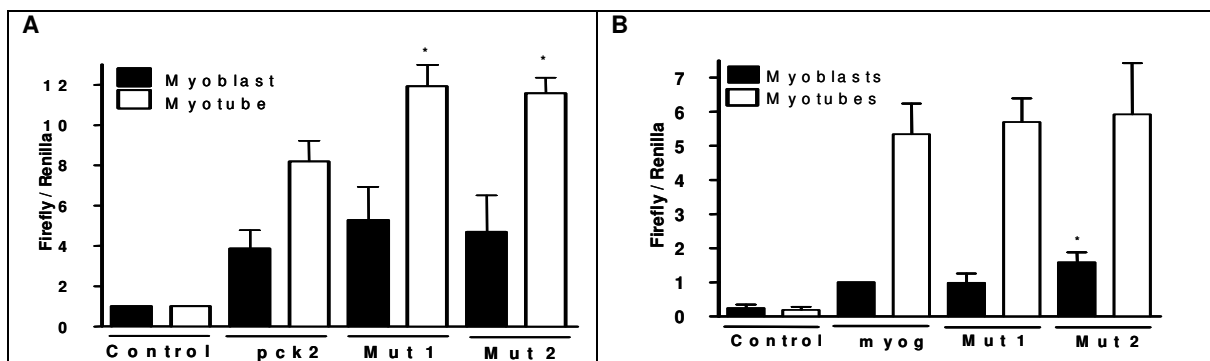
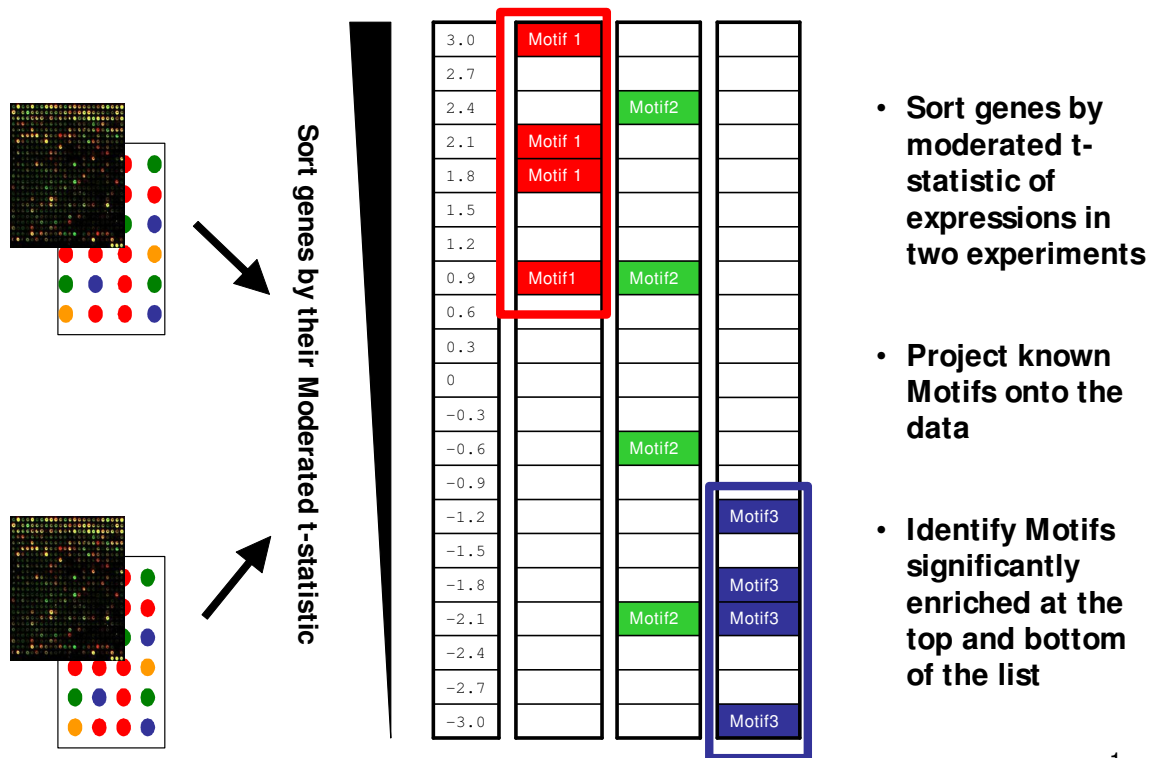


Figure 3. Reporter gene assay of pck2 400 bp fragment containing GCGGAGGC motif (A) and myogenin fragment containing CGACCCGT motif (B). There is a change in the reporter activity upon mutagenesis in pck2 construct and there is no change in myogenin construct. Data normalized to corresponding myoblasts or myotubes transfected with pGL3 promoter vector in the case of pck2 assay (A), Myogenin data was normalized to myoblasts transfected with pGL4.15 containing myogenin construct, because pGL4.15 alone does not have any basal activity. Data represents at least three replicates  $\pm$  s.e.m. (\*,  $p < 0.05$ , t-test).

Figure 4 - MotifADE overview



1

Figure 4. Overview of MotifADE method: Genes are sorted by their moderated t-test statistic values. Motifs in the promoter regions in these genes are identified. Two-tailed Wilcoxon rank sum statistics is applied. In this schematic view, Motif 1 is significant in the up-regulated genes. Motif 2 is not significant in either the up-regulated genes or down-regulated genes and Motif 3 is significant in the down-regulated genes.

Figure 5 - Problems with the two-tailed Mann-Whitney test

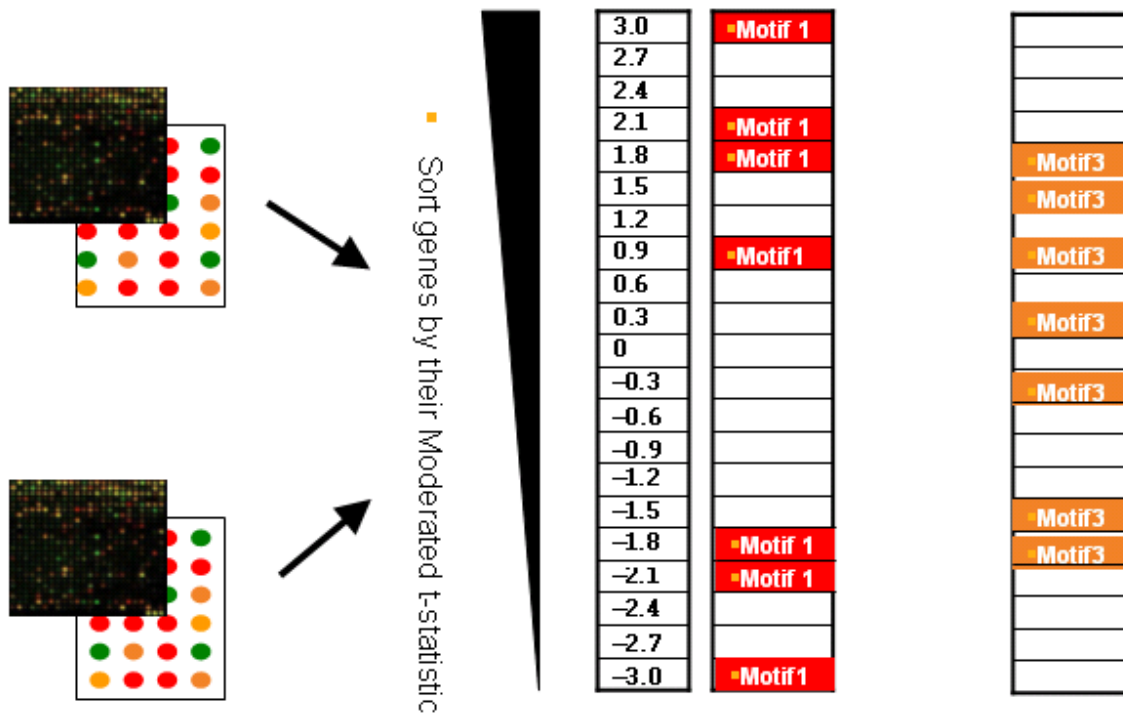


Figure 5. Motif 1 and Motif 3 would have the same p-values with the two-tailed Mann-Whitney test since only the t-value 0.9 is important and all other t-values from Motif 1 or Motif 3 are symmetric with respect to 0 contributing the same to the rank sum as does t-value 0 even though genes in Motif 1 show higher magnitude changes than genes in Motif 3.

## Tables

Table 1 - Significant transcriptional factors in up-regulated genes from Day 1 to Day 2

Motif	Occurrence	p-value	q-value	Transcription Factors
SRACAGGTGKYG	20	2.76E-05	0.00376	MYOD
ATGCCCATATATGGWNNT	37	0.000126	0.00858	SRF
GGTACAANNTGTYCTK	13	0.000227	0.00909	GRE
TGACGYA	35	0.000267	0.00909	CREBP1/CJUN
RRCAGGTGNCV	13	0.000421	0.0115	E12
GGGGCGGGGT	171	0.000509	0.0115	SP1
RSTGACTNANW	56	0.00209	0.0405	AP1
CKSNYAAAAAWRMICY	4	0.00271	0.0417	MMEF2
NNRYCACGTGRYNN	29	0.0029	0.0417	USF
AGATADMAGGGA	15	0.00307	0.0417	GATA4
NNNNNNGGNACRNNNTGTTCTNN	3	0.00364	0.045	PR
GGACATGCCCGGGCATGTCY	104	0.00502	0.052	P53
NCACSTGNCN	4	0.00559	0.052	EBOX
GGGGAGGG	3	0.00619	0.052	MAZ
RGCAGSTG	9	0.00621	0.052	MYOGENIN

Table 1. Significant transcriptional factors in up-regulated genes from Day 1 to Day 2.

The Motif column shows the consensus binding site sequence for the transcriptional factor. The second column lists the total number genes containing that transcriptional factor binding sites in the promoter regions. The p-value column illustrates the Wilcoxon rank sum p-value. The q-value column shows the multiple testing corrected FDR q-value. The transcriptional factor column lists the name of the transcriptional factor which is known to bind to that motif.

**Table 2 - Significant transcriptional factors in down-regulated genes from Day 1 to Day 2**

Motif	Occurrence	p-value	q-value	Transcription Factors
NKTSSCGC	72	9.22E-11	1.18E-08	E2F1
NNACCACGTGGTNN	187	2.53E-07	1.48E-05	MYC/MAX
GGGGCGGGGT	195	3.47E-07	1.48E-05	SP1
GGACATGCCCGGGCATGTCY	143	5.06E-06	0.000157	P53
TRRCCAATSRN	82	6.14E-06	0.000157	NFY
ARATKGAST	14	6.31E-05	0.00135	FOXO1
TGACGYA	64	0.000174	0.00318	CREBP1/CJUN
NDDNNCACGTGNNNNN	13	0.000966	0.0155	ARNT
NNTTGGCNNNNNCCNNN	4	0.0015	0.0214	NF-1
RSTGACTNANW	49	0.00171	0.0218	AP1
ACWTCK	14	0.00338	0.0393	PEA3
TWSGCGCGAAAAYKR	9	0.00371	0.0395	E2F
NBTGGGTGGTCN	10	0.00455	0.0448	GLI
ASMCTTGGGSRGGG	4	0.00878	0.077	SP3
TCATGTGN	5	0.00902	0.077	TFE

Table 2. Significant transcriptional factors in down-regulated genes from Day 1 to Day 2. The Motif column shows the consensus binding site sequence for the transcriptional factor. The second column lists the total number of genes containing that motif in the promoter regions. The p-value column illustrates the Wilcoxon rank sum p-value. The q-value column shows the multiple testing corrected FDR q-value. The last column lists the name of the transcription factor.

**Table 3 - Significant synergistic transcriptional factors in down-regulated genes from Day 1 to Day 2**

Motif	Occurrence	p-value	q-value	Transcription Factors
TGACGTMA_NNACCACGTGGTNN	187	2.53E-07	6.32E-07	CREB MYC/MAX
VTGAACTTTGMMB_TRRCCAATSRN	82	6.14E-06	7.68E-06	HNF4ALPHA NFY
RNRTKDNGMAAKNN_ANNCACTTCCTG	32	0.0747	0.0747	CEBPB ETS

Table 3. Significant synergistic transcriptional factors in down-regulated genes from Day 1 to Day 2. The Motif column shows the consensus binding site sequence for the transcriptional factor where two motifs are separated by an underscore. The second column lists the total occurrence number of genes containing that motif in the promoter regions. The p-value column illustrates the Wilcoxon rank sum p-value. The q-value column shows the multiple testing corrected FDR q-value. The last column lists the name of the transcription factors.

**Table 4 - Tested novel motifs with mutagenesis**

Motif	Occurrence number	p-value	Gene symbol	Fold change Ratio	Gene description	Known nearby Transcriptional factor binding sites
-------	-------------------	---------	-------------	-------------------	------------------	---------------------------------------------------

gcggaggc	1238	2.57E-06	pck2	0.2	Phosphoenol- pyruvate carboxykinase 2 (mitochondrial)	Oct-1, TFIIA
cgaccggt	95	3.60E-06	myog	5.2	myogenin	SREBP-1, MEF2, MEF3

Table 4. Novel motifs tested with mutagenesis and their surrounding known transcriptional factor binding sites.

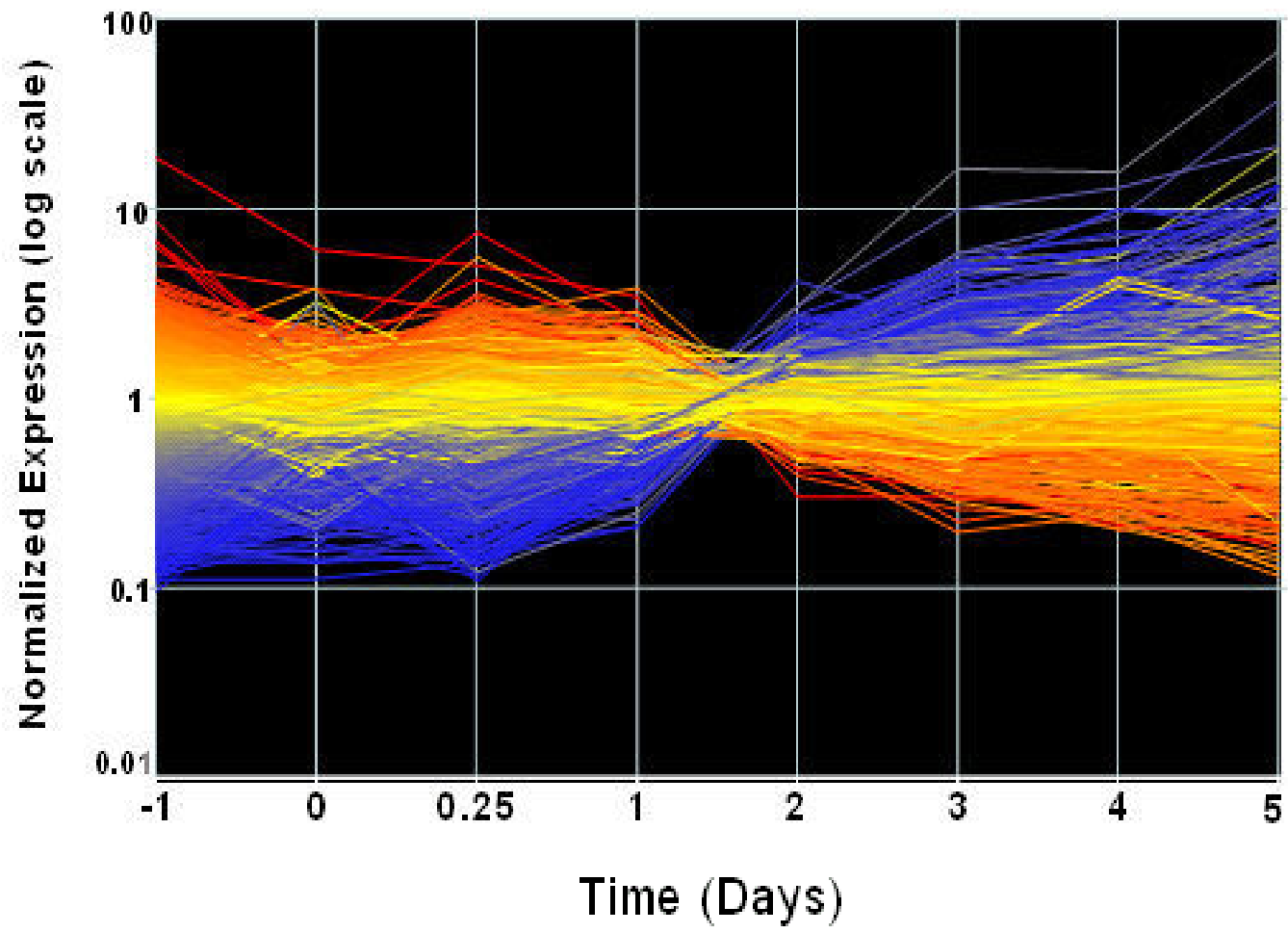
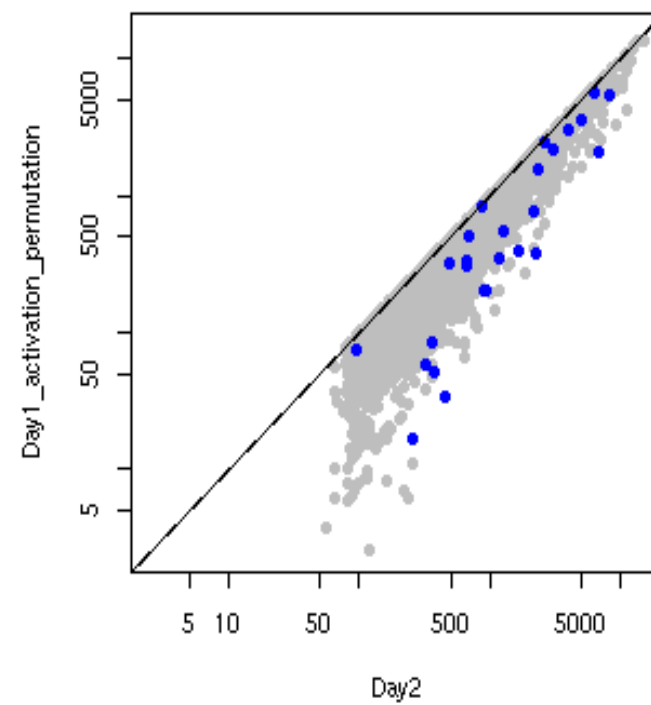
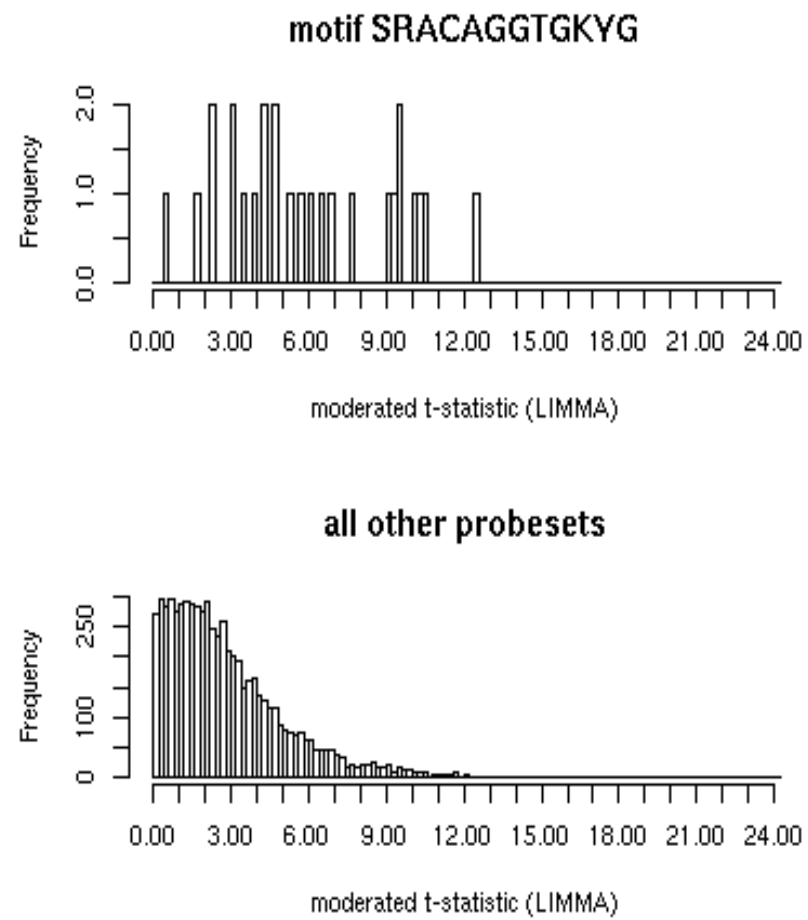


Figure 1



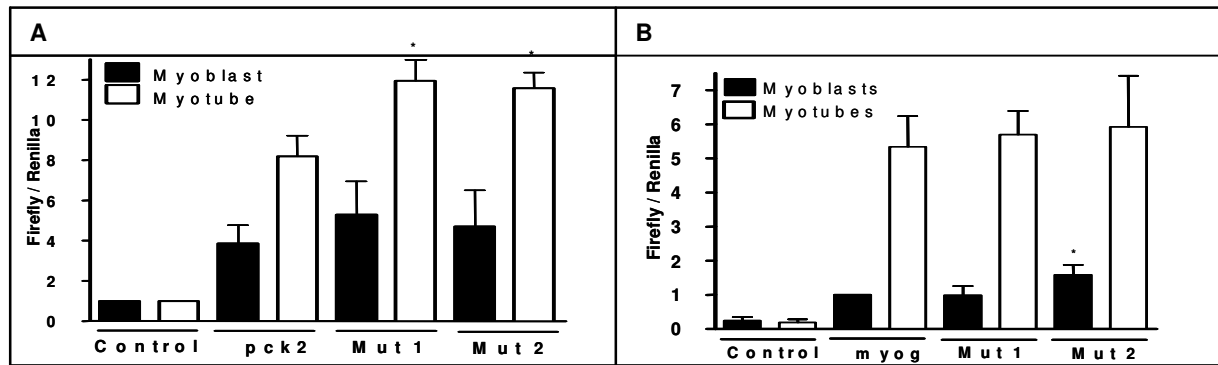
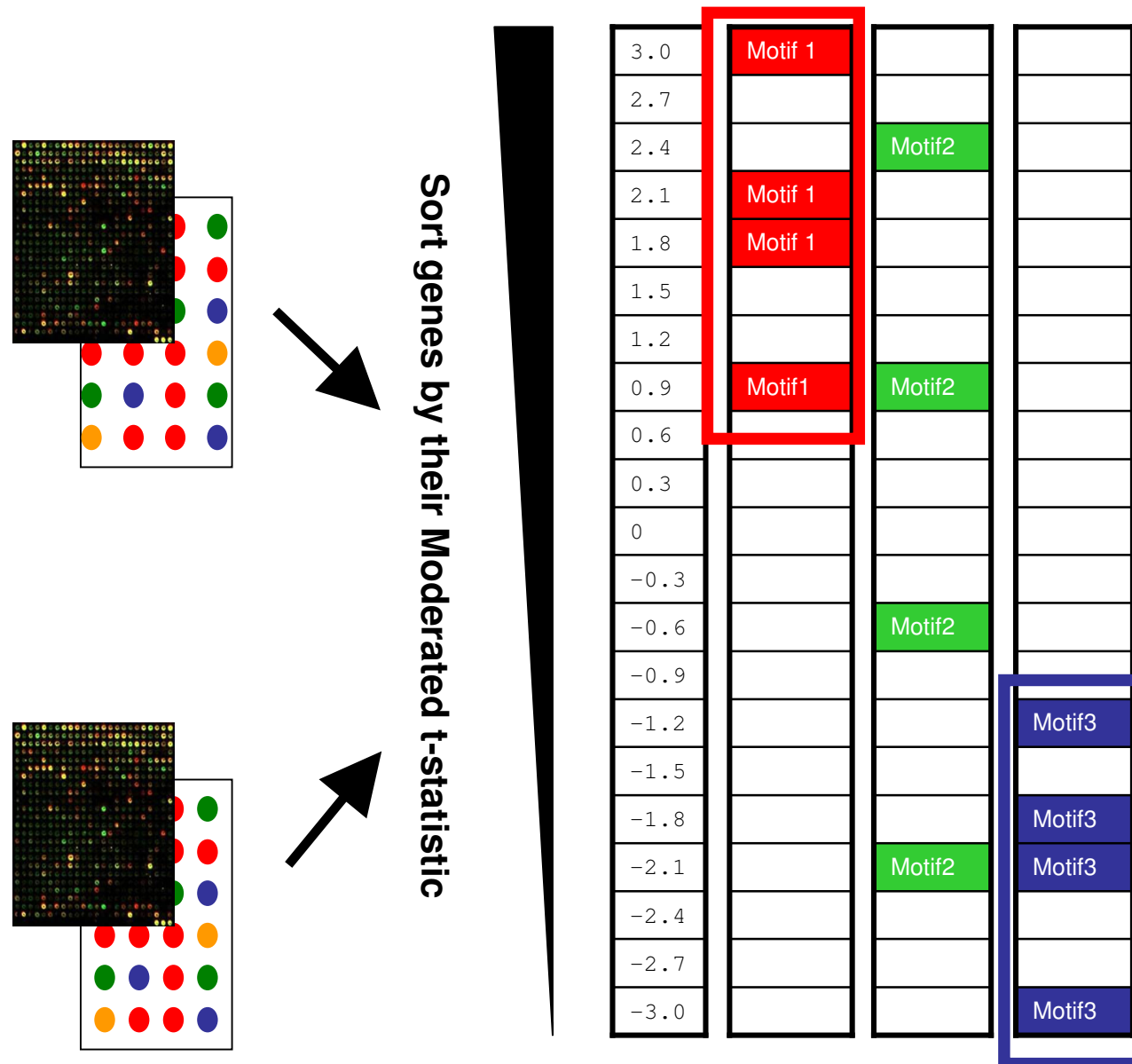
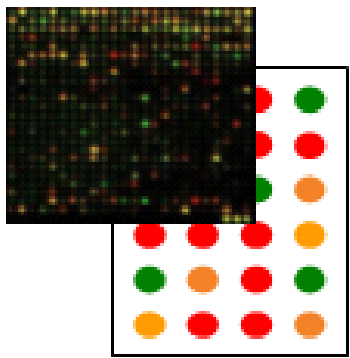
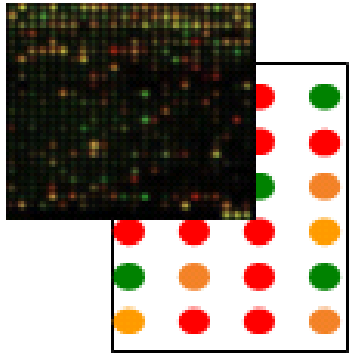


Figure 3

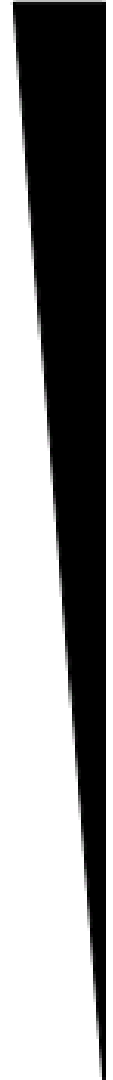
# MotifADE Overview



- Sort genes by moderated t-statistic of expressions in two experiments
- Project known Motifs onto the data
- Identify Motifs significantly enriched at the top and bottom of the list



Sort genes by their Moderated t-statistic



3.0	■ Motif 1
2.7	
2.4	
2.1	■ Motif 1
1.8	■ Motif 1
1.5	
1.2	
0.9	■ Motif 1
0.6	
0.3	
0	
-0.3	
-0.6	
-0.9	
-1.2	
-1.5	
-1.8	■ Motif 1
-2.1	■ Motif 1
-2.4	
-2.7	
-3.0	■ Motif 1

	■ Motif 3
	■ Motif 3
	■ Motif 3
	■ Motif 3
	■ Motif 3
	■ Motif 3

Figure 5