RESEARCH ARTICLE

Elucidating functional context within microarray data by integrated transcription factor-focused gene-interaction and regulatory network analysis

Thomas Werner[1,2], Susan M. Dombrowski[3,4], Carlos Zgheib[5], Fouad A. Zouein[5], Henry L. Keen[6], Mazen Kurdi[5,7], George W. Booz[5]

[1] Genomatix Software GmbH, Munich, Germany
[2] University of Michigan, Internal Medicine-Nephrology Division & Center of Computational Medicine and Bioinformatics (CCMB), Ann Arbor, MI, USA
[3] Genomatix Software Inc., Ann Arbor, MI, USA
[4] Wayne State University School of Medicine, Detroit, MI, USA
[5] Department of Pharmacology and Toxicology, School of Medicine, and the Jackson Center for Heart Research, The University of Mississippi Medical Center, Jackson, Mississippi, USA
[6] Department of Pharmacology, University of Iowa College of Medicine, Iowa City, Iowa, USA
[7] Department of Chemistry and Biochemistry, Faculty of Sciences, Lebanese University, Rafic Hariri Educational Campus, Hadath, Lebanon

Correspondence: T. Werner, PhD, Genomatix Software GmbH, Bayerstrasse 85a, 80335 Munich, Germany
<wernersbc@me.com>

To cite this article: Werner T, Dombrowski SM, Zgheib C, Zouein FA, Keen HL, Kurdi M, Booz GW. Elucidating functional context within microarray data by integrated transcription factor-focused gene-interaction and regulatory network analysis. *Eur. Cytokine Netw.* 2013; 24(2): 75-90 doi:10.1684/ecn.2013.0336

ABSTRACT. Microarrays do not yield direct evidence for functional connections between genes. However, transcription factors (TFs) and their binding sites (TFBSs) in promoters are important for inducing and coordinating changes in RNA levels, and thus represent the first layer of functional interaction. Similar to genes, TFs act only in context, which is why a TF/TFBS-based promoter analysis of genes needs to be done in the form of gene(TF)-gene networks, not individual TFs or TFBSs. In addition, integration of the literature and various databases (e.g. GO, MeSH, etc) allows the adding of genes relevant for the functional context of the data even if they were initially missed by the microarray as their RNA levels did not change significantly. Here, we outline a TF-TFBSs network-based strategy to assess the involvement of transcription factors in agonist signaling and demonstrate its utility in deciphering the response of human microvascular endothelial cells (HMEC-1) to leukemia inhibitory factor (LIF). Our strategy identified a central core of eight TFs, of which only STAT3 had previously been definitively linked to LIF in endothelial cells. We also found potential molecular mechanisms of gene regulation in HMEC-1 upon stimulation with LIF that allow for the prediction of changes of genes not used in the analysis. Our approach, which is readily applicable to a wide variety of expression microarray and next generation sequencing RNA-seq results, illustrates the power of a TF-gene networking approach for elucidation of the underlying biology.

Key words: microarray data analysis, high-throughput (HT) approaches, transcription factor-gene networking, transcription factor binding sites, transcription factors

Microarrays record a snapshot of transcriptional changes caused by the administration of drugs or agonists to cells and define all changes, as far as the genome is covered by the microarray design, regardless of whether they have relevance to the functional actions of the drug or agonist [1]. They provide long lists of genes that show changes in steady-state RNA levels, but they do not yield direct evidence for functional connections between genes and miss even important genes if their steady-state RNA levels are not significantly changed. However, as recently demonstrated by results of the ENCODE project [2], functional interactions of genes depend on a variety of functional genomic elements with transcription factors (TFs) and their binding sites (TFBSs) in promoters and enhancers, and are important for inducing and coordinating changes in RNA levels. Moreover, multiple databases and the scientific literature provide huge amounts of functional information on genes and their interactions, including TFs. Therefore, an approach based on elucidation of TF/TFBS interactions (i.e. networks) by promoter analysis of genes with significantly changed transcripts is very well suited to elucidate functional connections between significantly changed genes in microarray data sets that might be missed in any individual gene or factor oriented analysis.

Attempts to include additional data frequently make use of pathways, GeneOntology (GO)-terms, or molecular features such as TFBSs in the vicinity of genes, e.g., an approach focusing on transcriptional regulation by transcription factor binding was recently described [3]. However, with the exception of pathways, all these approaches just produce more lists, while missing a structured biological context. Another clear-cut lesson from ENCODE, as well as many previous smaller scale studies, is that neither genes nor TFs or their corresponding TFBSs act in isolation, but are highly interconnected usually in the form of gene-gene networks. Biological functionality only

doi: 10.1684/ecn.2013.0336

becomes apparent at the network level (pathways representing small networks themselves). Moreover, integration of additional functional connections as taken from the literature and various databases (e.g. GO, MeSH, etc.) allows for inclusion of genes relevant to the functional context of the data even if they were initially missed because their RNA levels do not change significantly. An integrative approach has the additional advantage of compensating for the intrinsic weaknesses of individual methods; enrichment analyses are necessarily biased by uneven distribution of knowledge; co-citation literature networks face the same challenge and, in addition, inevitably contain variable numbers of false positive connections. However, by bringing several lines of evidence together outliers due to erroneous results of one method are readily identified and discarded. This rationale is based on "biological consistency", i.e. every finding in one area of analysis must also be reflected in the results of other lines of evidence in order to be accepted as real.

We developed a widely applicable strategy, entirely focusing on TF and TFBSs-centered networks, complemented by expression profiling information gathered from the literature. Other approaches report as the final results GO-terms, pathways, and associated TFBSs. These are only "stepping stones" in our strictly context/network-oriented approach. One of the most important principles of this strategy is to complement findings from expression data with conclusions drawn from our network approaches (biological consistency between data and knowledge-based analyses). We applied this strategy to elucidate the potential involvement of transcription factors in the regulation of genes in response to leukemia inhibitory factor (LIF) in human microvascular endothelial cells (HMEC-1). We were able to identify a central core of eight TFs based on multiple lines of evidence, most likely involved in the regulatory network of LIF-induced gene expression changes in HMEC-1 cells, although initially almost 100 TFs showed significant expression changes (one line of evidence). We also found potential molecular mechanisms of gene regulation in HMEC-1 cells upon stimulation with LIF that allowed prediction of changes of genes observed on the microarray, but not used in the analysis. This clearly demonstrates the power of a TF-gene networking approach for the elucidation of the underlying biology. Our approach is widely applicable to high-throughput analyses of transcriptional changes such as all expression microarrays, as well as all pertinent, next generation sequencing (NGS) applications (ChIP-seq, RNA-seq, bisulfite-resequencing), where the possibility of reducing the amount of data to a biologically-linked, small network is especially important.

MATERIALS AND METHODS[1]

Materials

Cell culture reagents were obtained from Invitrogen (Carlsbad, CA, USA). Epidermal growth factor was from BD Biosciences (Franklin Lakes, NJ, USA), hydrocortisone from Sigma-Aldrich (St. Louis, MO, USA), recombinant human LIF from Millipore (Billerica, MA, USA), and fetal bovine serum (SH30070.03) from Thermo Fisher Scientific (Waltham, MA, USA).

[1] Cf. Annex 1.

Experimental design

HMEC-1 cells were obtained from the Centers for Disease Control and Prevention (CDC), and grown in MCDB 131 with 15% fetal bovine serum (FBS), 10 ng/mL epidermal growth factor, 1 μg/mL hydrocortisone, 10 mM glutamine, and antibiotic-antimycotic. Cells were grown in 100 mm dishes to near confluency and incubated in medium with 0.5% FBS for 12-15 hours before being used in experiments. Cells were dosed with vehicle or 2 ng/mL LIF for 90 min at 37°C, placed on ice, and washed 2× with 10 mL ice-cold Hanks' buffered saline solution.

Microarray analysis

RNA was isolated using the RNAquous-4PCR Kit from Applied Biosystems (Foster City, CA, USA). RNA quality was established using the NanoDrop 3300 Fluorospectrometer (Thermo Scientific) and Agilent 2100 Bioanalyser. Only samples with a 260/280 ratio close to 2 and an RNA Integrity Number (RIN) value >9 were processed for microarray analysis. Microarray processing was performed by the core facility of the University of Mississippi School of Medicine using Agilent technology and whole human genome slides. Cy3 and Cy5 dye swap and background correction were applied. Genes were considered downregulated with treatment-to-control ratios <0.5, and up-regulated with treatment-to-control ratios >2. Image processing was performed using ImaGene (version 8.0.1), and statistical analysis done using the R statistical program (version 2.10.1). Array signals for six replicates (channel median values) were calculated by first subtracting the local background mean followed by normalization using LOESS (within array) and quantile (between arrays) algorithms. P values for differential expression were determined using the R/Bioconductor package limma, which incorporates both Bayesian and linear modeling methods and is routinely used in microarray data analyses [4]. In the calculation of signal values for each probe, there was a subtraction of the local background, which is the recommended procedure to remove bias (e.g., one array or part of an array was not washed as well after hybridization). This is thought to represent somewhat of a trade-off with reduced bias and lower variability for highly expressed genes, but with higher variability for genes with low expression. For that reason, we used an unadjusted p-value <0.05 as significance threshold. Annotation for the probe sets on the array was obtained from the Gene Expression Omnibus (GEO) at the NCBI, using accession number GPL4133 and from the Agilent internet site (http://www.chem.agilent.com/cag/bsp/gene_lists.asp).

Regulatory network analysis

Figure 1 summarizes the strategies used for the analysis of the significantly regulated genes. We separated up- and down-regulated genes by GO and pathway-analysis in order to find TFs specifically associated with up- or down-regulation. The whole strategy is a combination of five results originating from three independent lines of evidence: a) mRNA values and their relative changes, b) literature and pathway analysis, c) sequence-based promoter analysis (*figure 1* top "lines of evidence"). The only experiment-specific data used were the list of significantly regulated genes and their expression values. Our main

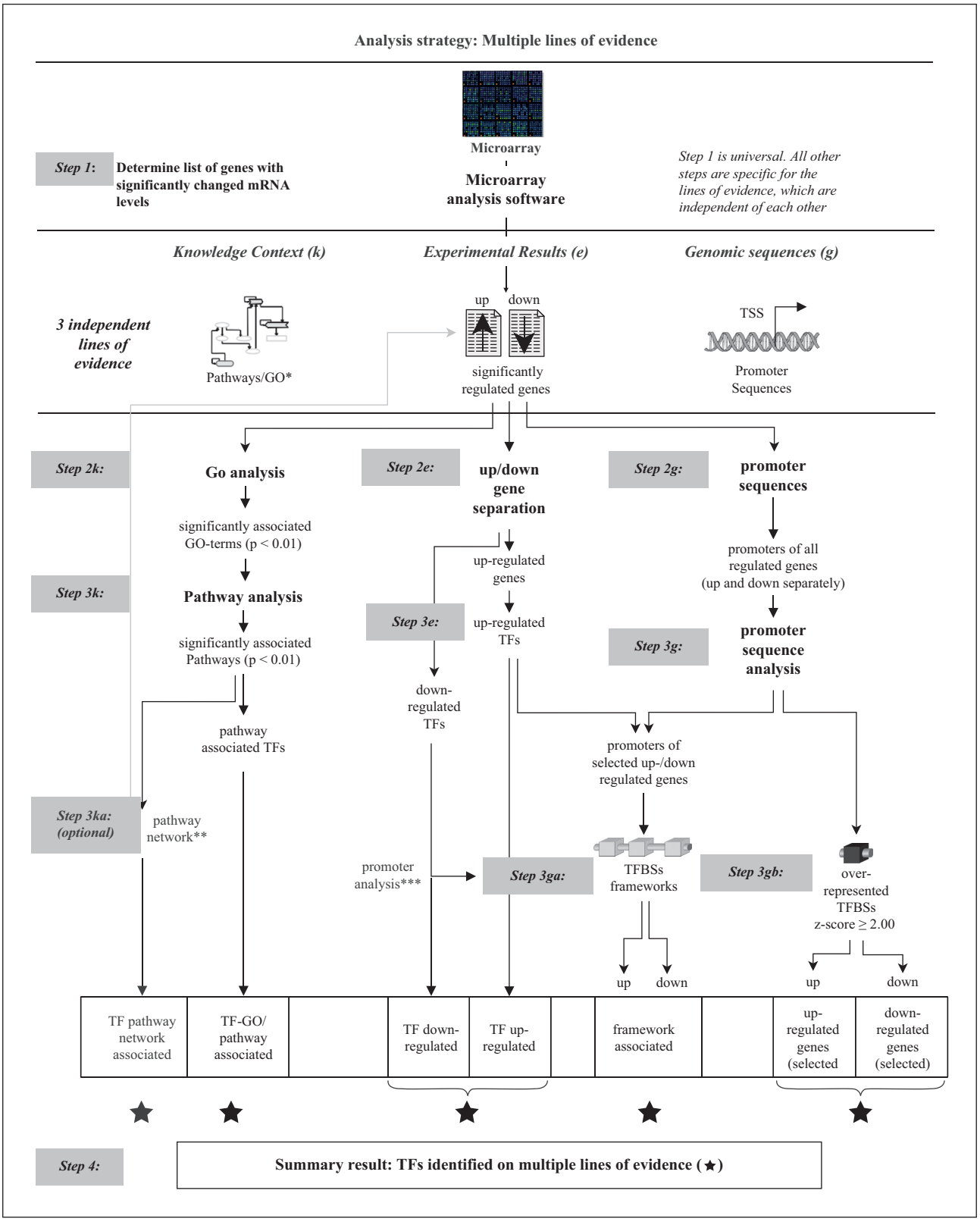

Figure 1

Analysis strategy and summary of results. The upper part of the figure indicates the three major lines of evidence used in the subsequent analysis. Three parallel threads of analysis were carried out from the associated lines of evidence, each using results from all lines of evidence to focus and restrict the next analysis step. This is indicated by the cross-connections. The whole strategy focused on transcription factors (TFs) throughout and collected all positive evidence for involvement of a TF. *In case no pathways are available, GO categories can be used in the same way. **The pathway network essentially produces a reduced initial list, which can be treated exactly the same way as the initial list.***Promoter analysis for the down-regulated genes is carried out exactly as for the up-regulated, if a down-regulated TF is thought to be responsible for the down-regulation.

focus was the analysis of TF genes and their potential targets in order to understand the transcriptional effects of LIF treatment.

Analysis, downstream of the significant microarray signals, was carried out using the standard integrated analysis package Genomatix Software Suite (Genomatix Software GmbH, Munich, Germany) and the various databases and software tools within this package, including : Gene-ontology *(*GO)-analysis which was carried out with the program GeneRanker using default parameters recommended by the supplier. All literature-based analyses were carried out using the Genomatix Pathways System (GePS), which combines co-citation analysis from the whole PubMed database with canonical pathway analysis. GePS was used with the default parameters recommended by the supplier. Promoters used for TFBSs analysis were all extracted from the ElDorado genome database (Release 12/2010) using the program Gene2Promoter. The various promoter collections were then analyzed using the program RegionMiner, which contains pre-compiled databases of TFBSs match-numbers for whole genomes and whole-genome promoter collections, and for which over-representations and z-scores are automatically calculated. We refer to whole-genome promoter collections as the relevant background throughout this study.

Promoter context is defined as sets of TFBSs that show a specific organization within sequences: the individual TFBSs (e.g. TFBSs A, B, and C) and their relative order is conserved (A-B-C only, A-C-B rejected), and a flexible, but limited, distance range is allowed between the individual TFBSs, which also must have a conserved strand-orientation. In this way a complete framework of three TFBSs would have the annotation A(+) - distance range 1 - B(-) - distance range 2 - C(+) where + and - symbolize the strand orientation of the individual TFBSs. Such a framework needs to be found conserved in a minimum number of sequences (sequence quorum), which can be set as a user parameter. Throughout this study we used the following parameters: minimum number of TFBSs in a framework; 3, variation of distance range; 20 (in case no results were found, this was increased to 30), minimum distance; 10, maximum distance; 200 (between TFBSs). The sequence quorum was set high initially (no results), and then reduced step-wise until frameworks of three elements were found or the minimum quorum was reached without finding frameworks. For each search, single TFBSs identified as important in previous analyses were set as mandatory elements, and all frameworks found with the described settings were collected as framework sets and the sets were then evaluated.

Evaluation for association of the frameworks with the respective promoter sets was carried out using the program ModelInspector as follows: each set was looked at for matches in the promoters of the specific set that the frameworks were derived from, various larger subsets from the significantly regulated genes (such as network genes, 3-and higher up-regulated genes, etc.). This was compared to matching results obtained either from all microarray-derived promoters or all promoters from the human genome (automatically carried out by ModelInspector). The over-representation of the framework sets in the specific promoter sets, as compared to random sampling of the genome, was calculated. These are the results shown in the tables. For more detailed description of the methodologies see [5].

RESULTS

Differentially expressed genes: steady-state mRNA levels of HMEC-1 cells were analyzed by microarray assays for genes with significantly changing mRNA levels in response to LIF treatment. LIF-treated cells were compared to untreated control cells. Microarray files were analyzed as described in Methods using the Bioconductor package limma in order to find the significantly regulated genes. We found a total of 1,171 genes significantly regulated between the LIF-treated cells and the control: 589 genes were up-regulated and 582 genes were found to be down-regulated. Out of the 1,171 genes 1,107 were annotated, allowing GO and pathway analysis, which were the first steps in our data analysis.

GO-term and pathway analysis: we had a total of 368 GO-terms from the significantly (p-value $\leq e^{-03}$) associated biological processes. *Table 1* shows the top ten GO-terms according to their p-value. There is a clear preference for kinase-cascade signaling in GO-terms, which is a hallmark of multiple signal transduction pathways. Therefore, we went on to pathway analysis as the third step using the GenomatixPathwaySystem (GePS, Genomatix Software, Munich) database/tool. *Table 2* shows the six pathways that were significantly associated with the 1,107 regulated (and annotated) genes. Again, JAK-STAT regulation is evident (IL7 signaling pathway). However, several other signaling pathways are also found. There were seven transcription factor families, i.e., TFs that are very similar and bind to the same motifs, directly implicated by the six pathways (AP1, ETS, STAT, HNF, CREB, CEBP, DDIT3). This step concluded the analysis of the knowledge–based and GO- and pathway-based line of evidence.

TF-regulation analysis: this is another line of evidence independent from the literature-based analyses shown above, except for the literature-derived TF-gene annotation. The only common starting point is the list of significantly changed genes. GePS is also able to identify genes for TFs and we used this feature to evaluate the number of TF genes that showed altered expression. We found 50 TF genes to be up-regulated among the 1,107 genes, and 45 TF genes that were down-regulated. Merging results from pathway and TF-regulation analysis showed that from the pathway-associated TF genes, ETS and CEBP factors were up-regulated, while AP1 and Jun (a CREB family factor) were down-regulated, yielding a total of four differentially expressed TFs so far supported by two lines of evidence (expression data and pathway analysis). However, as many more TFs were regulated we also looked for additional evidence of association of these factors with regulated genes. This step concluded the analysis of the knowledge–based lines of evidence.

Statistical promoter analysis for TFBSs: sequence-based analyses have the advantage of being largely independent of the above mentioned, heavily knowledge-dependent methods. The genomic sequence (and thus the promoters) is universal, entirely independent of literature, and the detection of TFBSs is based on sequence patterns

Table 1
Top 10 GO "biological process" categories associated with genes differentially regulated by LIF treatment.

GO process	GO-ID	p-value	Go	Ge	Gt
Enzyme-linked receptor protein signaling pathway	GO:0007167	$8.58 \times e^{-08}$	57	27.22	472
Transmembrane receptor protein tyrosine kinase signaling pathway	GO:0007169	$1.56 \times e^{-07}$	41	17.07	296
Prostate gland growth	GO:0060736	$2.13 \times e^{-07}$	7	0.58	10
Phosphate metabolic process	GO:0006796	$3.98 \times e^{-07}$	117	74.45	1291
Phosphorus metabolic process	GO:0006793	$3.98 \times e^{-07}$	117	74.45	1291
MAPKKK cascade	GO:0000165	$4.31 \times e^{-07}$	39	16.44	285
Phosphorylation	GO:0016310	$7.32e^{-07}$	104	64.82	1124
Regulation of cellular component movement	GO:0051270	$8.72 \times e^{-07}$	34	13.73	238
Regulation of MAPKKK cascade	GO:0043408	$9.07 \times e^{-07}$	26	8.99	156
Regulation of phosphorus metabolic process	GO:0051174	$9.46 \times e^{-07}$	59	30.68	532

All associated GO-processes were ranked by their p-value. Go = number of genes observed in the significantly regulated genes belonging to the respective biological process, Ge = number of genes expected in the significantly regulated genes belonging to the respective biological process by a random selection of the same size, Gt = total number of genes belonging to the respective biological process.

Table 2
Six pathways associated with the differentially regulated genes.

Pathway	p-value	Input genes in pathway	Gene IDs
PDGFR-alpha signaling pathway	1.39E-03	ITGAV, IFNG, SHF, JUN, CSNK2A1, PDGFRA, CAV1	3685, 3458, 90525, 3725, 1457, 5156, 857
pertussis toxin-insensitive ccr5 signaling in macrophage	2.42E-03	CCL2, CCR5, JUN, CXCL12	6347, 1234, 3725, 6387
E-cadherin signaling events	5.25E-03	EPHA2, EXOC3, AKT1, HGF, IGF1, IGF1R, EFNA1	1969, 11336, 207, 3082, 3479, 3480, 1942
IL-7 signaling pathway(JAK1 JAK3 STAT5)	6.74E-03	IL7, RIPK3, AKT1, SYK, ZAP70, MAPK13, KIT, BRAF, LCK, FGFR2, IRAK4, PRKCD, PIK3CD, FLT4, IGF1R, PAK2, CSNK1A1, CAMK2G, AKT2, PDGFRA, MAP3K2, ITK	3574, 11035, 207, 6850, 7535, 5603, 3815, 673, 3932, 2263, 51135, 5580, 5293, 2324, 3480, 5062, 1452, 818, 208, 5156, 10746, 3702
ATF-2 transcription factor network	6.94E-03	IFNG, POU2F1, SOCS3, JUN, CCND1, DUSP8, PDGFRA, BCL2, NOS2	3458, 5451, 9021, 3725, 595, 1850, 5156, 596, 4843
TCR signaling in naive CD4+ T cells	8.93E-03	VAV1, AKT1, ZAP70, LAT, FYB, LCK, LCP2, PTPRC, DBNL, PTEN, ITK	7409, 207, 7535, 27040, 2533, 3932, 3937, 5788, 28988, 5728, 3702

All associated pathways were ranked by their p-value as determined by the program GePS/GeneRanker (Genomatix Software, Munich). Input genes in pathways: these genes were part of the list of regulated genes, as well as the pathway.

derived by sequence analysis. The only part where knowledge comes into play is the completeness of the library, i.e. TF identification. TFs may act directly or indirectly on genes and some may change transcriptional activity without any apparent change in their own mRNA levels. In order to estimate direct regulation by TFs we decided to look at the other end of TF-mediated transcriptional regulation namely the TFBSs in the promoters of differentially regulated genes. If any particular TF is directly involved in the regulation of a set of genes, then those genes should contain at least one TFBS for such TFs. Thus, TFBSs for factors prominently involved in mediating transcriptional signaling might be statistically enriched in the regulated promoters. Lack of overrepresentation does not preclude a functional connection, but a positive result is additional evidence for inclusion. We extracted all 5,371 promoters associated with the 1,107 regulated genes using the Gene2Promoter tool (Genomatix Software GmbH, Munich) and analyzed them for statistical overrepresentation of TFBSs with the MatBase Matrix Family Library (Version 8.3, Genomatix Software GmbH, Munich). A total of 53 TFBSs families were found to be overrepresented (as compared to a random sampling from all promoters in the human genome, using a cutoff threshold of a z-score of 2.00), 47 TFBSs families were in those promoters that were up-regulated and six TFBSs families were associated with up-regulated TF genes (HOMF (HMX1), FKHD (FOXD1), BCDF (OTX1), CEBP (CEBPD), IRFF (IRF1, IRF8), DMRT (DMRTB1).

In promoters from down-regulated genes, 35 TFBSs were found to be significantly associated, six of which were also associated with down-regulated TF genes FKHD (FoxP4, FOXJ2), PARF (HLF), VTBP (TBP), NKXH (NKX2-2, NKX2-3), HOXF (HOXD8), OCT1 (POU2F1). It became evident that different factors belonging to the same TF family (e.g. forkhead, FKHD) and their respective TFBSs were associated with up- and down-regulated genes. It also became evident that eight transcription factor families showed up in at least two out of three analyses (*table 3*). Of the three that were not associated with a differentially expressed TF gene (STAT, HOMF, HOXF), only STAT was directly associated with one of the six associated pathways,

Table 3
TFs prominently associated with significantly regulated genes.

TFBS family	TF/up (+) or down (-) regulated	Pathway association	z-score all regulated promoters	z-score up-regulated promoters	z-score down-regulated promoters
OCT1	POUF2 +	+	5.5	5.07	2.55
FKHD	FOXD1 + FOXP4 - FOXJ2 -	-	7.4	5.31	4.92
IRF	IRF1 + IRF8 +	-	5.59	3.32	4.9
CEBP	CEBPD +	+	3.65	4.67	-
BCDF	OTX1 +	-	3.18	4.85	-
STAT	-	+	3.59	3.57	-
HOMF	-	-	8.03	7.3	4.33
HOXF	-	-	5.75	5.64	2.63

Column 1 shows the TFBS family of which the individual TFs shown in column 2 are members. Column 3 indicates whether the TF was directly implicated by an associated pathway and columns 4 to 6 indicate the statistical over-representation of the respective TFBS family as compared to all promoters in the human genome. Only factors that show at positive values in at least three columns are shown.

as well as being co-cited with LIF in the context of vascular endothelium [6], resulting in a short list of six TFs: FKHD, IRF, OCT1, CEBP, BCDF, and STAT (*table 3*).

So far the selection was based on a combination of classical analyses essentially focusing on individual TFs. Next we focused on functional connections between TFs not necessarily restricted to these eight TFs in *table 3*, but using them as a starting set.

Promoter context analysis of TFBSs (frameworks): the presence of TFBSs is a physical phenomenon while the organization of TFBSs into clearly defined groups (frameworks) is connected to transcriptional function. Thus, frameworks establish another line of evidence in addition to the presence of TFBSs. Thus, we extended our analysis to find such TFBSs networks in regulated promoters. *Table 3* shows three forkhead factors, one of which was up-regulated transcriptionally (FOXD1), while two (FOXP4 and FOXJ2) were down-regulated. As all three factors are able to bind to the same FKHD binding sites (MatBase, Matrix Family Library Version 8.3, Genomatix Software GmbH), this suggests that the transcription factors most likely act in different contexts with other factors. Such contexts can be specifically addressed and elucidated by promoter analysis for conserved TFBSs frameworks (strand-, order- and distance-correlated sets of TFBSs) [5]. However, as there are 2,744 promoters associated with the up-regulated genes (Gene2Promoter, Genomatix Software GmbH, Munich), systematic analysis of all up-regulated promoters could not be carried out owing to the technical limitations of the software (limit is 1000 promoters due to the combinatorial explosion of possible TFBSs combinations). Therefore, we decided to select the subset of 764 promoters of three-fold or more up-regulated genes.

We analyzed these 764 promoters for frameworks of at least three TFBSs (essentially representing regulatory networks with one molecular mechanism), where one of TFBS was mandatory (exhaustively for all six TFBSs families corresponding to the six most important TFs identified in this study). *Table 4* summarizes the results of these context searches. Most framework sets show a modest association with the selected promoter set (Z-score cutoff 2.00, promoters of three-fold or more up-regulated genes) except for one FKHD-group (3.13) and the STAT-group, which has the highest association (>8-fold overrepresented). However, none show an association with all regulated microarray promoters (the STAT group being borderline with 2.03). However, restriction to one model that also contained a second associated TFBS (CEBP) resulted in more selective results (*table 4*, last row). Interestingly, the two TFBSs families HOMF and HOXF originally found but discarded based on few lines of evidence, showed up numerous times in the context of significant factors. Thus, all six previously selected TFs, OCT1, FKHD, IRF, CEBP, BCDF, and STAT were also supported by associated TFBSs framework context (3-fold or more up-regulated promoters).

Functional context analysis (TFBSs-frameworks) already linked several TFBSs, even when based only on a statistical selection ($\geq$ 3-fold up regulated). Therefore, we expected an approach based on a subset based on biologically linked genes to confirm the results and perhaps be even more successful.

The following analysis is currently only possible using the Genomatix solution, which is commercial. However, as also indicated in *figure 1*, this analysis is optional and essentially supports the findings achieved without it, albeit in a much faster time and with many fewer interactive steps.

Pathway network analysis: we used another selection method that is more biology-oriented. Based on the initially associated pathways and the regulated genes, the new pathway-network tool determines a subset of genes that link those pathways into a network with optimal co-citation connectivity, i.e. the network of genes has the highest number of co-citation-based edges (normalized for gene count). This is motivated by best-knowledge based biological connections, bypassing any fold-change-based criteria and should be more biologically correlated to LIF action than the 3-fold or higher sub-sections, as expression values represent only one of three selection criteria (pathways, co-citations, and expression changes). The network method is entirely data-driven, and requires no more input than the complete list of all regulated genes (Hahn *et al.* in preparation). A network of 335 genes was defined (as detailed in Methods) by this method, 190 of which were

Table 4
Framework analysis of the six associated TFBSs families.

Framework set mandatory TFBSs in bold	3 and more up-regulated promoters (764)	All microarray promoters (5371)	All genome promoters (82703)	3 up overrepresentation	All microarray overrepresentation
DMRT-HOMF-**OCT1**	36	153	1990	1.96	1.2
PDX1-**OCT1**-HOXF MYT-**OCT1**-HOXF	39	129	1655	2.55	1.2
CDXF-HOMF-**FKHD**	25	76	870	3.13	1.26
IRFF-HOMF-BRNF	23	84	897	2.80	1.44
X-**CEBP**-**FKHD**-X	144	531	6316	2.46	1.29
BCFD-OCT-FKHD	119	457	5855	2.33	1.26
STAT-set	36	60	454	8.78	2.03
CEBP-BRNF-**STAT**	9	11 (allup)	120	Na	2.76

Column 1 shows the main TFBSs determining the Framework sets as automatically determined by FrameWorker (Genomatix Software GmbH, Munich). Columns 2 to 4 show the number of promoters matched by the whole sets of frameworks in the three respective promoter collections, and columns 5 and 6 show the respective over-representation with respect to all human promoters.

up-regulated, connecting all six, significantly associated pathways into one network. We then applied the exact same strategy as for the unselected and the 3-fold-up-regulated genes to the analysis of the network-selected genes.

GO-term analysis comparison: all together the network was significantly associated with 988 GO-terms (as compared to 368 for all regulated genes). *Table 5* shows that several GO/Medical Subject Heading (MeSH) terms significantly associated with both gene groups (all regulated and network-selected genes) show a dramatically lower p-value in the network genes than in all regulated genes, suggesting a sharper focus on the corresponding biology by the network selection.

Pathway analysis

The 190 up-regulated genes of the network were significantly associated with 10 pathways (*table 6*). These 10 pathways are related/overlap as can be seen from the fact that there were six genes shared by five out of 10 pathways (SOCS3 ZAP70, ITK, PDGFRA, PRKCD, SYK). Promoter modeling of this set of six genes most common to the 10 pathways revealed also a strong association with STAT and FKHD TFBSs (data not shown).

Statistical promoter analysis for TFBSs

The 190 up-regulated genes in the network were associated with 18 TFBSs (data not shown), and although there were only 49 down-regulated genes, they were associated with 17 TFBSs (data not shown). As shown in *table 7*, the network analysis so far identified eight TFs supported by at least two out of four lines of evidence (TF mRNA regulation, network pathway association, TFBSs association with up and/or down-regulated network promoters). Notably, there is an overlap of five factors (in bold) already identified by the same approach in all regulated genes. Joining all lines of evidence, including the network analysis, all together a list of eight TFs emerged, confirming the initially detected OCT1 and adding SP1 to the list (*table 8*).

Promoter context analysis of TFBSs (frameworks): an analogous approach as described for the 3-fold or more up-regulated promoters based on network-derived up-regulated promoters yielded framework sets that also were associated with the up-regulated network promoters as well as with the three and more up-regulated promoters (data not shown).

TFBS-frameworks in promoters are associated with transcriptional regulation of the corresponding genes and can be located by computational search in promoters of genes not involved in the detection of those frameworks. Hence, they are also suitable for predicting transcriptional up-regulation for genes that contain such frameworks in their promoters.

Framework-predicted gene regulation is confirmed by mircoarray data: we selected the FKHD-CREB-SORY framework (defined from promoters of ITK, PDGFRA, SYK) as it associates two relevant TFBSs (FKHD and CREB) with the central genes of the gene-interaction network-derived pathways. All promoters of up-regulated genes on the whole microarray were analyzed for presence of this framework. Any matching promoter is supposed to be associated with an up-regulated transcript, which in turn can be verified using the microarray data for these genes. It is important to note, that none of these microarray results have been used at any time to generate the framework, which makes them independent data. The framework was overrepresented in the promoters of the up-regulated genes on the microarray (6.41-fold enriched) matching just 11 promoters (*table 9*). The only down-regulated gene was skipped as it was not annotated and was thus not suitable for further evaluation. We then used GePS to construct a co-citation linked network from the 206 genome-wide matches. A central area connected five genes including the three input genes and consisting of: ITK-SYK-KDR (vascular endothelial growth factor receptor 2 VEGFR2)-PDGFRA-BRAF (*figure 2*). BRAF was also associated with four of the 10 network-up-regulated genes associated pathways.

DISCUSSION

We applied a predominantly data-driven and strictly network-focused strategy to the analysis of microarray data - in our case HMEC-1 cells treated with LIF. Several attempts have already been published employing

Table 5
GO/MeSH term comparison all regulated genes / network genes.

GO-term	p-value 1107 regulated genes	p-value 335 network genes
Top ranked GO term	e^{-8}	e^{-29}
MapKKK cascade	e^{-7}	$1.32 \times e^{-15}$
Signal transmission via phosphorylation event	$1.11 \times e^{-6}$	$2.80 \times e^{-19}$
Inflammation (MeSH disease)	$1.93 \times e^{-11}$	$2.75 \times e^{-64}$

Selected GO-processes were compared by their p-value. All three selected individual GO-terms (rows 2 to 4) showed a much lower p-value for the network association than for all of the regulated genes.

Table 6
Ten pathways associated with the 190 up-regulated genes contained in the network.

Pathway	p-value	Input genes in pathway
Cytokine receptor degradation signaling	2.84E-04	ILA1, MAP3K2, IRAK4, IL4R, AKT2, IGF1R, FLT4, ITK, IL7, IL1B, PDGFRA, IFNG, SOCS3, BRAF, PRLR, PRKCD, SYK, FGFR2, ZAP70
IL-7 signaling pathway(JAK1 JAK3 STAT5)	5.82E-04	MAP3K2, IRAK4, PIK3CD, AKT2, IGFR1, FLT4, ITK, IL7, PDGFRA, BRAF, PRKCD, SYK, FGFR2, ZAP70
pertussis toxin-insensitive ccr5 signaling in macrophage	1.86E-03	CCL2, CCR5, JUN, CXCL12
AKT(PKB)-Bad signaling	1.95E-03	MAP3K2, IRAK4, PIK3CD, AKT2, IGFR1, FLT4, ITK, PDGFRA, BRAF, PRKCD, SYK, FGFR2, TZAP70
Migration	2.15E-03	MAP3K2, IRAK4, PIK3CD, AKT2, IGFR1, FLT4, ITK, PDGFRA, BRAF, PRKCD, SYK, FGFR2, ZAP70
ATF-2 transcription factor network	2.89E-03	DUSP8, I BCL2, NOS2, PDGFRA , IFNG, SOCS3
Signaling events mediated by PTP1B	3.92E-03	ITGB3, LAT, LYN, SOCS3, PRLR, CSF1R
IL23-mediated signaling events	4.23E-03	CCL2, NOS2, IL1B, IFNG, SOCS3
Class I PI3K signaling events	8.96E-03	ITK, LYN, VAV1, SYK, ZAP70
IL-6-mediated signaling events	9.85E-03	CEBPD, IRF1, VAV1, SOCS3, PRKCD

All associated pathways were ranked by their p-value as determined by the program GePS/GeneRanker (Genomatix Software, Munich). Input genes in pathways: these genes were part of the list of regulated genes as well as the pathway.

Table 7
Seven TFBSs associated with genes in the network of LIF-associated pathways.

TFBS family	TF / up + or down - regulated	Network pathway association	z-score up-regulated network promoters	z-score down-regulated network promoters
SP1	KLF11 +	+	3.89	2.97
CEBP	CEBPD +	+	2.79	-
FKHD	FOXD1 + FOXP4 - FOXJ2 -	+	2.38	-
IRF	IRF1 + IRF8 +	-	-	2.54
STAT	-	+	2.54	-
ETS	SPI1 + PBRM1 +	+	-	-
ZBP	ZNF219 +	-	5.22	-
BCDF	OTX1 +	-	2.00	-

Column 1 shows the TFBS family of which the individual TFs shown in column 2 are members of. Column 3 indicates whether the TF was directly implicated by an associated pathway, and columns 4 and 5 indicate the statistical over-representation of the respective TFBS family in network promoters as compared to all promoters in the human genome.

more data-driven strategies, such as identification of co-expression of transcription factors and their putative target genes [7], which worked best in yeast. A more recent approach was aimed at the identification of functionally coordinated TF-clusters also in human and Arabidopsis microarray data [8]. These and many other approaches are truly data-driven analyses, but focus on expression data only, while our approach was designed to include as many sources of information as possible in a data-driven and network-focused analysis. Even the simplest analysis

Table 8
Final results: core set of TFs involved in response to LIF.

TF	Matrix family	Pathway network	Associated pathway	TF regulated (+)/(-)	TF framework associated	TFBS over-represented (+)/(-)
FOXD1	**FKHD**	+		-	+	-
FOXP4 - FOXJ2	**FKHD**	+	+	+		+
STAT 1 / 3 / 4 / 5a	**STAT**	+	+		+	+
CEBPD	**CEBP**	+	+	+		+
	SP1	+		+	+	+
IRF1 IRF8	**IRF**		+	+	+	-
	CREB		+	+	+	
POU2F1	**OCT1**			-	+	-
OTX1	**BCDF**			+	+	+

The table summarizes the results from five analyses (pathway network, pathway association, TF gene up/down regulation, framework association, and TFBS overrepresentation in promoters of up/down-regulated genes) derived from three independent lines of evidence: generic knowledge databases, experimental measurements, and promoter sequence analysis. Final selection was made with a cutoff of 3/5, i.e. only factors supported by at least three of the five analyses are shown. (+) and (-) indicate association with up (+) or down (-) regulated genes and are for the purpose of sum scores treated as equivalent.

Table 9
FKHD-CREB-SORY-containing promoters are all up-regulated with one exception.

All microarray promoters (5371)	All up-regulated promoters (764)	All genome promoters (101233)	All microarray promoters	All up-regulated promoters
Matches	Matches	Matches	Overrepresentation	Overrepresentation
11	10	206	1.01	6.41

Overrepresentation analysis was carried out in the same way as for the data in *table 4*.

of the ENCODE data as published recently in Nature [2], provided overwhelming evidence of how strong network-oriented gene regulation is.

The actions of LIF include several mRNA-independent steps such as kinase cascades, which can never be observed directly in microarray data [6]. However, we were not only able to identify STAT as a central factor in LIF action solely by data analysis, but we also determined a short-list of eight TFs, most of which were not known to be important for LIF action (*table 8*). IRF8, STAT3, SP1, IRF from that list are significantly associated with myeloid leukemia ($p = 1.21e^{-10}$), yielding further support for the validity of the TF selection.

The most compelling part of regulatory network-oriented analyses is the ability to predict changes in RNA of other genes not used in the definition of the TFBSs frameworks defining regulatory networks. We ran the prediction using a network-associated framework containing two of the best associated TF/TFBSs (FKHD-CREB-SORY), found 11 promoters of genes interrogated on the microarray, and 10 of these matched the prediction derived from the framework analysis. At this point, verification by other experimental methods such as RT-PCR, NGS or the like would be required to turn most likely candidates into verified transcriptional regulators or transcriptional targets (by ChIP-seq, ChIP-on-chip, siRNA or vector-driven over-expression approaches), but this is clearly beyond the scope of this study that focused on strategies for the computational data analysis. However, supporting evidence can also be collected from existing knowledge: four of the core TFs are part of the androgen receptor pathway (STAT3, SP1, POU2F1, and ATF2) and three are part of the IL-6 and the c-Myc signaling pathways respectively (STAT3, CEPBD, IRF1, and CEBPD, IRF8, SP1). This may allow selective inhibition of such pathway-oriented downstream reactions, which might even enable the differentiation of inflammatory responses from others, such as angiogenesis.

Our strategy focused on TFs, their TFBSs and the potential functional network-context by combining knowledge-based measures (GO-terms, pathways, co-citations) with experimental data (expression changes) and genomics-based sequence analysis (TFBSs and promoters) as outlined before [9]. The almost perfect agreement of framework-derived predictions with the actual microarray readings on genes is another line of supporting evidence. We used specific, prior knowledge solely to judge our results not to generate them, e.g. we used the knowledge about STAT and SOCS3 involvement to qualify our results as valid, but both factors were identified without the explicit use of this knowledge.

A TF involved in the regulation should bind to its target genes and would naturally act together with other factors in this context, which is modeled by the framework approach [5, 10]. Each line of evidence basically provides quantitative results of some kind (scores, expression values etc.). But it is almost impossible to normalize knowledge-based [11] and genomics based data in any way that would allow a quantitative comparison. Therefore, we count a line of evidence as supportive (i.e. associated significantly with the data) or not, without any internal ranking or order. This safeguarded against the bias of "more" evidence (e.g. from literature) available for particularly popular factors and premature filtering. For example, STAT factors turned

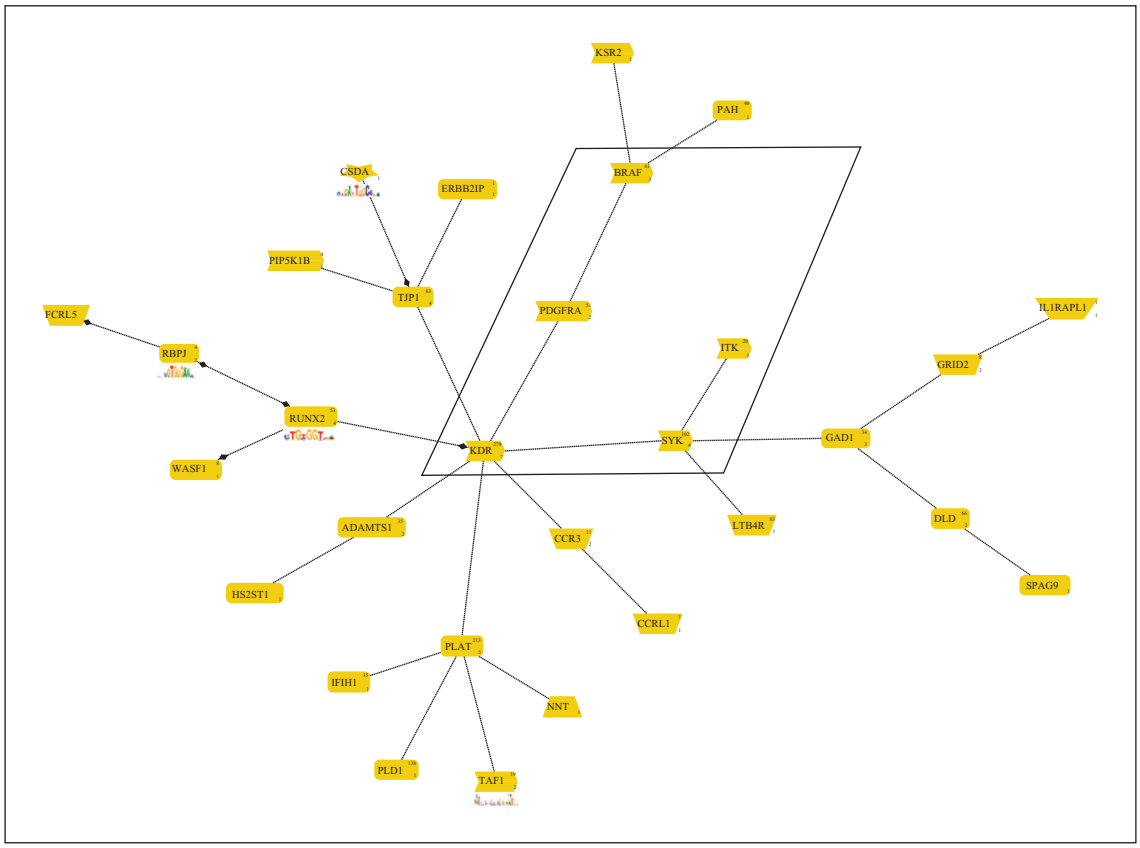

Figure 2
Literature-derived co-citation network based on the 206 genes selected by genome-wide search for the FKHD-CREB-SORY promoter TFBS-framework. This network represents the largest contiguous network detectable in the set of 206 genes. The central area containing the framework-founding genes ITK, SYK and PDGFRA is boxed.

out to be among the most important TFs in the end, despite the fact that we did not observe a significant mRNA regulation in the microarray data as STAT is finally activated by phosphorylation even if transcriptionally up-regulated [12]. The collection of multiple lines of evidence made the results robust with respect to missing lines of evidence, as long as enough lines remained supportive.

We have successfully used a highly systematic, network-focused approach, which can be applied to almost all high-throughput data sets such as microarrays, NGS-based experiments (e.g. RNA-Seq, and ChIP-seq), as well as protein-interaction maps with very few adaptations. The general process contains steps with quantitative limitations requiring some pre-selections by the scientist, that cannot always be strictly motivated from the data, as in our case the selection of 3-fold or higher, induced genes. Here, a best guess approach is required, but it is possible to test a few alternatives. This is one of the reasons why we also used a novel pathway-network oriented approach that does not suffer from such limitations and essentially confirmed results obtained on the arbitrarily selected gene subset. The network tool can take an unlimited number of pathways and genes, and always results in a single network, optimal in terms of co-citation-based connectivity. The biggest advantage is that the network is constructed in a fully automatic process within less than a minute, requiring no user-defined parameters. The results appeared to be more focused on the LIF-relevant biology as indicated by the much lower p-values of pertinent GO-terms. SOCS3 featured prominently as a central gene in the network-associated pathways, and is already known to be involved in the actions of LIF [13]. All in all, we hope that this strategy can contribute another building block for standardized data analysis of experimental, high-throughput methods aimed at rapid selection of subsets of data relevant to the experimental question at hand.

Acknowledgements. *We would like to thank Ruth Brack-Werner for helpful comments and discussion on this manuscript.*
Disclosure. *Financial support: this work was supported by grants from the National Heart, Lung, and Blood Institute to G. W. Booz (R01HL088101-06 and R01HL088101-02S1) from the Lebanese University and the National Council for Scientific Research, Lebanon (# 01-10-12) to M. Kurdi, and by grant 01EX1021L (M4 Personalized Medicine Ring funding project) to Genomatix (TW and SD). Conflict of interest: none.*

Annex 1. Preface

This whole paper is about HT-data analysis not about setting up and carrying out the experimental part. This is why we start at the expression data as provided by microarray readers or NGS mapping and annotation of RNA-seq data. The description of the experimental system in this manuscript is for scientific completeness but of no consequence to the strategy.

The strategy outlined in this paper and summarized in *figure 1* in a general manner and can be followed without a Genomatix license in all but two steps: the pathway-network analysis is not possible by other means right now; However, this part is optional and similar results can be reached by going manually through all the steps. The

second step is the promoter framework analysis, which is possible without Genomatix but carries a prohibitive workload (to our knowledge).
For every other step alternative methods (commercial as well as public domain) are available. However, to our knowledge there is no other package that would offer everything is one integrated system, which is why we used Genomatix.
We will not recommend any particular other tools for two reasons: First we have not tested other tools sufficiently well to justify recommendations and second the field is developing so fast that we expect more tools to become available for individual tasks quickly after publication of this manuscript. Therefore, as a more durable alternative, we clearly describe the required results for each step In order to facilitate following our strategy, so scientists can look for alternative tools if they choose to do so.

Rationale of the overall strategy

Data-driven analysis. Although one line of evidence (knowledge-based context analysis) clearly involves prior knowledge it does so in a generic manner: No experiment- or experience-motivated pre-selection or prioritizing of any of the knowledge database content is done. The only selections are made by applying direct experimental results (list of significantly changed genes) or results of prior analyses of this strategy. At no point specific prior knowledge of the experimental system is used to direct analysis. All prior knowledge is solely used to judge data-driven results.

Multiple lines of evidence. The overall aim of our strategy is to make use of as many sources of information as possible in order to develop a network-based representation of the biology observed in any high-throughput expression analysis (microarrays, RNA-seq, ChIP-seq, etc). We focus on the identification of central transcription factors (TFs) via networks involving their binding sites (TFBSs) and any other functionally motivated gene-gene-interaction (GGI), including protein-protein-interactions (PPI) as such transcription factors are among the most important actors in gene regulation. Once the networks have been developed other features can also be looked at (such as signaling transduction pathways or other gene groups of specific interest).

Optimization is at the level of overall results not individual analyses. Our strategy follows a paradigm slightly different from the usual approaches: We do not aim to optimize the results of individual tools and steps (such as minimizing the rate of false positives at every step) but use one of the most important principles in biology, which is *biological consistency*, i.e all knowledge and all results from analyses (lines of evidence) essentially represent the same biology (there is only one) and as a consequence any true finding must be reflected in one or several other results or prior knowledge. Due to gaps in our knowledge and limitations of the tools this will not be the case for all lines of evidence but at least two should coincide. If this is the case then a third finding distinct from the two in agreement is regarded as a false positive and not considered in order to make errors on the safe side. Of course, the more lines of evidence can be introduced to complement findings from expression data the better this selection by biological consistency works.

Lines of evidence need to be truly independent. There is one important point that needs to be observed carefully: Only independent lines of evidence contribute to this decision process. Two different expression analyses are good, replicates on one experiments are not (with respect to independent lines of evidence) as they only contribute to the improvement of the same line of evidence. Two different literature mining tools using the same basis (such as PubMed) are not independent lines of evidence, one PubMed based and one pathway-database based are considered sufficiently independent. Of course, knowledge-based methods, experimental results, and purely genomic sequence based analyses are naturally as independent as possible, which is why we chose exactly this combination.

Individual results need to be good enough not perfect. Each and every method has shortcomings including automatic annotation or any other method relying on prior knowledge (enrichment analysis, TFBSs matrix libraries, promoter databases etc.). Therefore, we do not rely on ANY single analysis result but collect several lines of evidence in independent analyses. This helps to identify erroneous individual results even in the absence of specific knowledge about the error due to inconsistencies with the results of two or more methods in agreement.
As a consequence of this approach, perfect optimization of individual tools might even be counterproductive: For example, the association of two genes by "expert curation" is virtually free of false positives (except for errors of the experts). However, this comes at the price of a considerable number of false negatives, excluded not because the experts were sure about the results being negative but simply because they did not find conclusive positive evidence. "Two genes being associated by co-citation" on the other hand does have much less false negatives, which in this case comes at the expense of a considerable number of false positives. However, in the light of the overall strategy false positives will be eliminated by the fact that the will not find supporting other lines of evidence while false negatives are simply eliminated from this line of evidence without the chance to collect additional support. So in this strategy it pays off to be more lenient on each individual line of evidence as the pileup will take care of most false positives brought in by individual lines of evidence. Of course, this only works within limits and requires some minimal quality standards. False positives should always be a minority of the results and tools and parameters need to be adjusted to guarantee this. However, the default parameters offered with the tools usually take care of this point already.

The strategy step by step

Step 1: *List of significantly changed genes in the HT-expression analysis*

Required results: A list of gene IDs (preferably) and/or gene symbols of genes with changed steady-state levels of mRNAs, in case of RNA-seq transcripts, selected by:
– cutoff p-value for statistics
– cutoff (log) folds-change (up and down regulated)

We have used the Limma package here, but any tool delivering similar results can be used as long as the desired list of significantly regulated genes is provided.

Line of evidence 1 (k): Knowledge-based context analysis

It is important that at this point we took advantage of the associative nature of enrichment analyses. We collected all TF genes that were linked to the associated GO-terms and pathways, regardless if they were also significantly regulated or not. They belong to the context of the regulated genes.

***Step 2(k):** GO- / MeSH / tissue enrichment analysis*

Required results: A list of GO-terms (and/or MeSH-terms, and/or tissue terms) significantly associated with gene IDs (preferably) and/or gene symbols from the list resulting in step 1:

– cutoff p-value for statistics

– identification and download of the genes belonging to each significantly associated term.

We have used Genomatix tool GeneRanker to analyze GO-terms for biological process as we found this category to be most informative. However, other categories may also proof to be helpful in specific cases. This needs to be decided by the researcher (details on parameters in the "methods details" section).

As our focus was on TFs we scanned the resulting gene lists for each associated GO-term for TF genes to get an initial TF-list supported by GO-terms.

***Step 3(k):** Pathway enrichment analysis*

Required results: A list of canonical pathways significantly associated with gene IDs (preferably) and/or gene symbols from the list resulting in step 1:

– list of pathway databases used (to spot omissions)

– cutoff p-value for statistics

– list of significantly associated pathways

– identification and download of the genes belonging to each significantly associated pathway.

We used the Genomatix tool GePS (Genomatix pathway system) for this task. However, any other tool yielding the desired results can be used interchangeably (details on parameters in the "methods details" section).

GePS already automatically identifies all TF-genes, so collection was easy in this case. With other tools this might require an additional annotation step. We collected all TFs that were associated with the gene list from step 1 regardless if they were on the list or in the pathways.

With this we completed the first line of evidence analysis resulting in a list of TFs that were associated with the gene list from step one either via GO-terms or canonical pathways.

The optional step 3(k)a: pathway network analysis will be described at the end of the strategy section.

Line of evidence 2 (e): Experimental data-based analysis

***Step 2(e)**: Separation of the genes from the list obtained in step 1 into up- and downregulated genes.*

Required results: Two sublists of gene IDs (preferably) and/or gene symbols from the list resulting in step 1:

– list sorting by attached parameter

– list of down-regulated genes

– list of up-regulated genes

Since this is only a very simple sorting we used Excel for convenience. Of course, every tool allowing to sort lists by an attached parameter is suitable.

***Step 3(e)**: Extraction of regulated TFs from the sublists obtained in step 2(e)*

Required results: Two sublists of TF gene IDs (preferably) and/or TF gene symbols from the list resulting in step 1:

– identification of TF genes

– list of down-regulated TF genes

– list of up-regulated TF genes

We used GePS to identify the genes that code for TFs. Essentially any tool capable of extracting TF genes based on the gene annotation for any genome-wide gene ID or gen symbol database can be used.

The TF-gene sublists were then extracted using Excel again. Important step is to carry the fold-change values along as they will be needed later on. See above (step 2(e)).

With this we completed the second line of evidence analysis resulting in a list of TFs that were up- or down-regulated in the gene list from step.

Please note that at this point GO-/pathway based information and expression-change based information is collected absolutely independently.

Line of evidence 3 (g): Genomic sequence-based analysis

***Step 2(g)**: Extraction of promoter sequences associated with the genes from the list obtained in step 1.*

Required results: Two sublists of promoter sequences of the gene IDs (preferably) and/or gene symbols from the list resulting in step 1:

– identification of promoter sequences of genes

– list of promoter sequences from down-regulated genes

– list of promoter sequences from up-regulated genes

We used the Genomatix tool Gene2Promoter as it allows one-step extraction of all promoters belonging to one list of genes in batch mode. However, any tool capable of finding transcriptional start sites (TSS) in a genome browser and extracting the appropriate promoter sequence (we used Genomatix defaults, see "method details" section) can be used for this task as well.

We focused on promoters despite the well-known importance of enhancers and other regulatory regions (such as Locus control regions or matrix attachment regions) as promoters are directly linked to transcriptional control of transcripts observed by HT-expression analysis and all the other regions act through promoters. This will miss some important relations but essentially capture enough for the purpose of this strategy.

Step 3(g)a: TFBSs framework analysis

Required results: Frameworks linking specific promoter sequences of the gene IDs (preferably) and/or gene symbols from the list resulting in step 1:

– identification of TFBSs frameworks

- finding matches to individual TFBSs in promoters based on a TFBSs library
- Analysis subsets of promoters for the occurrence of TFBSs frameworks characterized as follows:
  - Individual TFBSs matches that are members of the framework

- ○ Strand orientation of the respective TFBSs matches
- ○ Determination of the distance range of these selected TFBSs

– list of frameworks and their corresponding promoter sequences from list of regulated down-regulated genes
– list of frameworks and their corresponding promoter sequences from list of regulated up-regulated genes

This step of the analysis may be the least familiar for scientists not deeply involved in the mechanism of transcription control. Therefore here is a brief description of the underlying biological principles:

Individual TFBSs are physical units capable of binding TFs. However, binding of an isolated TF does not elicit any transcriptional control, which always requires complexes of more than one TF to be bound simultaneously to a promoter. In order to ensure specific regulation, the individual TFBSs cooperating (e.g. TFBSs A, B, and C) and their relative order is conserved (A-B-C only, A-C-B rejected), a flexible but limited distance range is allowed between the individual TFBSs, which also in most cases have a conserved strand-orientation. In this way a complete framework of three TFBSs would have the annotation A(+) - distance range 1 - B(-) - distance range 2 - C(+) where + and - symbolize the strand orientation of the individual TFBSs. Such a framework is associated with a specific regulation (if it is complete, i.e. no more TFs are required) or a group of gene regulations (if additional TFs are required that were not found in the analysis). In order to contribute to our strategy, such frameworks need to be found conserved in a minimum number of sequences (sequence quorum, see "method details"). Since up- and down-regulation definitely will employ different frameworks and mechanisms we separated the up- and down-regulated promoter sequences for this purpose.

In case of the down-regulated genes it should be noted, that this strategy will only detect frameworks that are basically up-regulatory but one or more of the important factor's activity is down-regulated (not necessarily the mRNA level!). Cases where a different, specific repressor element is responsible will not be accessible to this strategy.

We used the Genomatix tool FrameWorker for this purpose, which carries out the whole analysis fully automatically for one set of promoters at a time. Since this tool uses combinatorial TFBSs analysis there is a limit of 1,000 promoters due to the combinatorial explosion of possible TFBSs combinations. We are not aware of any other tool that would currently do the same. However, any tool that will help to determine the specific TFBSs organization described in the results required, could be used to carry out this step.

Due to the number of promoters exceeding limit FrameWorker requires some pre-selection of promoters before a framework analysis can be started. Basically, any biologically motivated selection process can be used provided it selects ≥ 3 and $< 1{,}000$ promoters: By GO-term group, by pathways, by expression profile, by network analysis. The important selection criterion is that the genes in the list should have a high likelihood of being functionally connected. As becomes evident from the results of the optional step 3(k)a (pathway networks) the more biological connection information is used the better the selection. We chose to go for highly up-regulated gene promoters in the basic strategy as we did not have expression profiles (time-series) available and wanted to demonstrate viability of the strategy even without the pathway network analysis.

Sometimes FrameWorker delivers long lists of frameworks necessitating a ranking. This ranking can be done either post-analysis requiring to go through all the long lists and crossing them with other lines of evidence or pre-analysis by the setting of mandatory elements. These are pre-selected TFBSs that MUST be part of any framework reported. This is only a filter, and the same frameworks would be found without the mandatory element but somewhere down the list in unfortunate cases.

Step (3g)a continued: Verification of data association of the frameworks

Required results: Lists of matches for individual TFBSs frameworks in selected promoter lists:
– promoter sequences used as training set for framework detection
– promoter sequences of the gene IDs (preferably) and/or gene symbols from the lists resulting in step 2(g):
– all promoter sequences of the human genome (or the genome of interest)
– table of over-representations of the above results vs the last (whole-genome match list)

As described in the text, we validated the association of the frameworks found by comparing the number of matches in various subsets of promoters with the matches in all promoters in the genome. A framework that is not over-represented in the experiment-specific subsets may still describe biological functionality but too broad to be relevant for the interpretation of the experimental data. This is no new line of evidence but only a safeguard that the framework and its TFBSs are relevant for the study.

We used Genomatix' tool ModelInspector for this tasks as it carries out the whole analysis fully automatically and is already linked to the genomic promoter databases of many organisms. However, any tool capable of locating matches to frameworks together with appropriate promoter sequence selection is suitable for this task.

***Step 3(G)b**: TFBSs overrepresentation analysis*

Required results: Lists of TFBSs overrepresented in promoter sequences of the gene IDs (preferably) and/or gene symbols from the list resulting in step 1:
– identification of TFBSs matches to individual TFBSs in promoters based on a TFBSs library
– p-value / z-score for the overrepresentation of TFBSs in the promoters of the promoter list:
– list of TFBSs over-represented in promoter sequences from list of regulated down-regulated genes
– list of TFBSs over-represented in promoter sequences from list of regulated up-regulated genes

The statistical over-representation of TFBSs only takes the number of matches compared to expectation from the genomic total match numbers into account and does not look at individual matches or any context. Therefore, this is rather different from the framework analysis. However, TFBSs involved in the regulation of gene groups are known to be over-represented in the corresponding promoters sometimes, which makes this a helpful result for selecting mandatory elements for step 3(g)a in case too many frameworks are found. TFBSs over-represented are often involved in gene regulation; however, statisti-

cal over-representation is not mandatory for functional involvement. Therefore, only positive results are considered and negative results are no exclusion criterion.
We used the Genomatix tool RegionMiner for this purpose, as it conveniently carries out the whole analysis automatically for each promoter set. Any tool producing the required results is suitable as this is basically mostly a statistical analysis (except for the location of the TFBSs matches, for which several solutions exist).
With this we completed the third line of evidence analysis resulting in a list of TFs that were found in a putatively functional context in the promoters of up- or down-regulated in the gene list from step.
Please note that at this point framework-based TF identification is carried out absolutely independently from the GO-/pathway based information and expression-change based analysis.

Step 4: *Compilation of the results into one final table*
Now that all results from three independent lines of evidence are in, the final step is rather easy: A table is compiled simply listing all potentially involved TFs in the first column and tabulating the supporting lines of evidence in further columns. Then support is counted and the list is ranked by level of support.
We have refrained from weighting individual results in the table. However, researchers may decide that some lines of evidence appear to be stronger than others in their mind, weighting counts accordingly. We do not recommend this as it deviates form the principle of strictly data-driven analysis maintained so far.

Optional Step 3(k)a: *Pathway network analysis*
Genomatix Synopsis: Extraction of pathway-centered optimal networks.
Required results: Sublist of genes from gene lists obtained in steps 1 or 2(g) with the following properties:
– all genes are either members of associated pathways or
– linked by statistically significant co-citation (against co-citation background) with two associated pathways
– the extent of the network is determined so that the connectivity (co-citation weight / number of genes) is optimal.
– The network is de novo constructed based on the experimental data and not any predefined interactome
– Parameters: none
To our knowledge there is currently no other method available fulfilling the described requirements for the results.
We used Synopsis on the complete list of annotated genes (1,105) to derive a maximum connectivity sub-network (335), which was then subjected to the complete strategy as outlines in steps 2 to 4. This is a better selection strategy than the 3-fold or more fold-change used before, as Synopsis is closer to biology due to the pathway connections. Also it is fully automatic and requires about a minute for the analysis. However, as the results obtained without it showed, it is optional and not absolutely necessary; however, it yielded more structured results due to the underlying network structure.

What is to be gained from this analysis in addition to TF identification?
First of all, the TF identification and their ranking is based on solid data-driven analyses with several lines of evidence and should be more reliable than any of the individual analyses. The other, probably even more important point is that TF-oriented microarray-analysis as far as it includes functionally related analyses such as framework analysis allows for predictions. These can be used for direct verification in other parts of the HT-data as we demonstrated, or as blueprints for experimental design. Especially Next Generation Sequencing expression analysis (RNA-seq) produces large amounts of data, where knowledge-based analysis fails due to lack of knowledge. However, the promoter-sequence based TFBSs analysis including the framework approach can be also used on entirely anonymous sequences outside any known genes, which we consider a major advantage. Once this approach links unknowns transcripts to known ones (such as demonstrated here by finding other genes on the microarray) the knowledge-based analyses can be applied based on a guilt-by-association principle.

Method details including rationale for setting and optimizing parameters

Preface
Wherever possible we used default parameters as suggested by the programs. In such cases no further explanations are given. Wherever parameters have been adjusted, the rationale for the adjustment is given.

Limma package: *significantly changed gene list*
Parameters:
– p-value threshold set to: <0.05 (unadjusted)
Preprocessing: Array signals for 6 replicates (channel median values) were calculated by first subtracting the local background mean followed by normalization using loess (within array) and quantile (between arrays) algorithms.
Limma analysis: P values for differential expression were determined using the R/Bioconductor package limma, which incorporates both Bayesian and linear modeling methods and is routinely used in microarray data analyses.

Genomatix GeneRanker: *GO-term enrichment analysis*
Parameters:
– organism: human
– p-value threshold set to: <0.01 (unadjusted, default)
We used the option to analyze GO-term enrichment, described on the help page as follows:
Biological Processes (GO): The ontology "biological process" from the Gene Ontology Consortium. Here is a short description of the p-value concept: Let q be the number of genes in the input set; Let m be the number of genes from the input set having annotation A assigned; Then the p-value is the probability (using Fisher's Exact Test) of finding at least m genes in a input list of length q having annotation A (under the assumption that belonging to the input list is independent of having this annotation).

Genomatix Pathway System (GePS): *Enrichment of canonical pathways, determination of TF genes*
The databases behind GePS are collected from public domain sources as well as by licensing other commercial databases (e.g. NetPro for expert-curated PPI).
The pathway database of GePS is compiled from four publicly available databases (Pathway interaction database,

NCI, Biocarta, Cancer cell map, and the INOH database). It contained a total of 512 pathways at the time of analysis.
The gene-gene interaction database was constructed in two ways:
All PubMed abstracts (with few exclusions) are automatically annotated to convert all gene and protein synonyms used into the NCBI preferred gene symbols using a expert curated synonym and homonym database (more than 600,000 synonyms just for mammals). From this annotation co-citations are determined on three different levels: i) in the same abstract, ii) in the same sentence, iii) in the same sentence with connecting functions words (e.g. regulates, inhibits etc). This ensured to cover the whole literature on each gene without missing entries that use other synonyms. This resulted in a basic interactome database containing > 6.7 million interactions at the time of analysis.
On top of this automatic effort, GGIs and PPIs are verified by a team of PhD-level scientists who have verified the connection in the abstracts of every paper that is part of an expert-level interaction, which totaled more that 64,000 at the time of analysis. This was complemented by expert curated PPIs from the NetPro database which added another 67,000 expert level interactions to bring the total of expert curated interactions to more than 130,000.
All information in the GePS interactome database is curated from the published literature. Thus, every gene interaction in this manuscript is supported by evidence extracted from the underlying publications (abstract level. On top of that GePS complements specially TF-gene interactions by literature independent verification of the presence of TF binding sites in the promoters of connected genes adding already a second line of evidence to such cases. GePS is available online through Genomatix.
We used the option: Characterization of gene sets: Input a gene list, optionally with expression values of GePS.

Parameters:

– organism: humans
– p-value threshold set to: <0.01 (unadjusted, default)
– co-citation level: sentence

TF gene identification by GePS is done by exporting the input gene list using the Export/import option: Export advanced gene list (filter genes). This results in a tab-delimited text with one column indicating whether the gene is a TF or not.
Pathway association:
This part does not make use of the co-citation analysis and is solely based on the overlap between the pathway genes and the input gene list. P-value determination as described in GeneRanker.
Literature-based network construction:
Here GGIs and PPIs are constructed between the genes of the input list using the interactome database described above. Each network is constructed based on the input genes and there is no projection to any precompiled interactome map.

Excel: *Handling and sorting of gene lists with expression values:*

Parameters:

– no parameters

Excel was used to compile and maintain all gene lists used throughout the analysis.

Genomatix Gene2Promoter: *Extraction of promoters for gene lists*

Gene2Promoter utilizes the Genomatix promoter databases; at the time of analysis the human database contained a total of about 120,000 human promoters. Promoters are extracted either by fixed format (user-defined) or by the Genomatix-defined default, which is flexible depending on the number of TSS known for each promoter. The default is extraction of a sequence that reaches 500 bp upstream of the most 5‘ TSS in the promoter to 100 bp‘ downstream of the most 3’ TSS of the same promoter. The -500/+100 range has been motivated by whole genome-analyses and was recently confirmed by whole genome-DNAse hypersensitive site analysis to encompass the bulk of accessible promoter sequences.
We used the batch version of the program: Extraction of larger sets of promoters and/or filtering promoters for TF sites

Parameters:

– organism: human
– promoter length: Genomatix variable (default)
– sequence format: FASTA

Gene2Promoter automatically produces a file of sequences of the promoters for the selected genes (upload list).

Genomatix FrameWorker: *Definition of TFBSs frameworks in subsets of promoters*

FrameWorker uses the concept of matrix families, which groups TFBSs that are known or very likely to be able to replace each other functionally into families and used as such. The family concept is supported by numerous experimental results.

Parameters:

– Matrix family group: vertebrates (all)
– Matrix filters: none (default)
– Framework analysis: exhaustive combination (default)
– Sequence quorum constraint: adjusted, see below.
– Sequence constraints: none (default)
– Minimum and maximum distance between TFBSs: 10 / 200 (see below)
– Maximum distance variation: 20 (see below)
– Minimum number of elements: 3 (see below)
– Mandatory elements: used according to TFBSs overrepresentation analysis
– Determine p-value of models: none (default)

We always started FrameWorker on a sequence set using all default parameters. However, very often this results in no frameworks reported. In such cases only we adjusted the parameters in the following order:
– Reduce the sequence quorum (default >80%) to lower numbers in increments of 10% until results were found.
– In case step one did not yield results, then increase the distance variation from 10 to 20 or even 30 bps in order to relax the requirements for similarity of the distance ranges.
– Restrict the distance range from 5-200 to 10-200 in case overlapping TFBSs were found causing combinatorial explosions. A distance of 10 usually excludes overlapping sites.

Mandatory elements were used when too many frameworks for easy inspection were found. Mandatory elements filter the results but do not produce different frameworks.

Mandatory elements were selected from the results of other analyses (TFBSs overrepresentation, other lines of evidence).

***Genomatix ModelInspector**: Locating matches to frameworks in sequences*

Parameters:

– Model group: user defined model

– Maximum number of matches: 1,000 (default)

ModelInspector analyzes both strands of all sequences selected and can do so also for whole genome sequences or whole genome promoter collections. It reports model matches only if all elements of the model match with the thresholds defined in the model (distances, order, individual TFBSs scores). The maximum number of matches was only extended when an initial search resulted in more than 1,000 matches. Usually more than 2,000 matches are an indication that no over-representation of the model with the experimental data set is to be expected.

The resulting match lists were used to calculate over-representation of individual models or model sets against the whole genome collection of promoters in our case. We did not calculate p-values for such over-representations since we usually had small numbers where statistics are not advisable.

***Genomatix RegionMiner:** Overrepresentation of TFBSs in promoter sets*

Parameters:

– User-defined sequence set: promoter sets

– Matrix description: matrix families

– Background selection for over-representation analysis: human promoters

RegionMiner is a tool for large-scale sequence analysis (up to whole genome), where one of the tasks possible is determination of the over-representation of TFBSs in particular sets of sequences vs a background distribution.

In our case we used the whole genome promoter collection as background as described in the help page of the program: Here the background is selected, which is used for the calculation of the overrepresentation values and the Z-Score (see below). You can select from

– genomic background: Genomic background comprises all chromosomes of the selected organism.

– promoter background: The promoter background comprises all Genomatix defined promoters of optimized length (about 500/100bp up/downstream of the TSS, details)

user-defined background: If this option is selected, please supply either a sequence file or a BED file with genomic positions. These sequences will be then searched for TFBS to get the background match numbers.

The result is a list of TFBSs with all the match numbers and a z-score to indicate validity of the over-representation. We used a z-score of 2.0 as cutoff. We also ignore any underrepresentation (negative z-score) as this can be the consequence of various artifacts and focused entirely on positive z-scores.

REFERENCES

1. Altman RB, Raychaudhuri S. Whole-genome expression analysis: challenges beyond clustering. *Curr Opin Struct Biol* 2001; 11: 340-7.
2. Dunham I, Kundaje A, Aldred SF, *et al*. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012; 489: 57-74.
3. Barrera LO, Ren B. The transcriptional regulatory code of eukaryotic cells–insights from genome-wide analysis of chromatin organization and transcription factor binding. *Curr Opin Cell Biol* 2006; 18: 291-8.
4. Wettenhall JM, Smyth GK. limmaGUI: a graphical user interface for linear modeling of microarray data. *Bioinformatics* 2004; 20: 3705-6.
5. Cartharius K, Frech K, Grote K, *et al*. MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics* 2005; 21: 2933-42.
6. Kubota Y, Hirashima M, Kishi K, Stewart CL, Suda T. Leukemia inhibitory factor regulates microvessel density by modulating oxygen-dependent VEGF expression in mice. *J Clin Invest* 2008; 118: 2393-403.
7. Zhu Z, Pilpel Y, Church GM. Computational identification of transcription factor binding sites via a transcription-factor-centric clustering (TFCC) algorithm. *J Mol Biol* 2002; 318: 71-81.
8. Nie J, Stewart R, Ruan F, *et al*. TF-Cluster: A Pipeline For Identifying Functionally Coordinated Transcription Factors Via Network Decomposition of the Shared Coexpression Connectivity Matrix (SCCM). *BMC Syst Biol* 2011; 5: 53.
9. Werner T. Bioinformatics applications for pathway analysis of microarray data. *Curr Opin Biotechnol* 2008; 19: 50-4.
10. Werner T, Fessele S, Maier H, Nelson PJ. Computer modeling of promoter organization as a tool to study transcriptional coregulation. *Faseb J* 2003; 17: 1228-37.
11. Scherf M, Epple A, Werner T. The next generation of literature analysis: integration of genomic analysis into text mining. *Brief Bioinform* 2005; 6: 287-97.
12. Kurdi M, Booz GW. JAK redux: a second look at the regulation and role of JAKs in the heart. *Am J Physiol Heart Circ Physiol* 2009; 297: H1545-H1556.
13. Forrai A, Boyle K, Hart AH, *et al*. Absence of suppressor of cytokine signalling 3 reduces self-renewal and promotes differentiation in murine embryonic stem cells. *Stem Cells* 2006; 24: 604-14.