ExPlain™ Analysis of TAL1 ChIP-seq Intervals

Introduction

Next generation sequencing (NGS) technologies have opened up novel research possibilities in many areas including cancer research, gene regulation, epigenetics, personalized medicine and pharmacogenomics. ChIP-seq is a remarkable NGS-based approach to uncover functional features of the DNA on a genome-wide scale and in great detail. A major application area of ChIP-seq is the discovery of transcription factor binding sites.

In this application note, we present several possibilities to explore and analyze a publicly available set of TAL1-bound genomic regions using the ExPlain™ software [1,2]. Here we focus on three important directions and investigate the biological function of nearby genes, enrichment of transcription factor binding sites as well as composite modules.

The ExPlain™ system

Systems biology and genomic data demand a broad spectrum of computational methods that address different biological aspects such as the functional role of genes or proteins, transcription regulation, or connections between proteins and other molecules via molecular networks and pathways.

The ExPlain™ tool integrates all of these aspects into a single platform and seamlessly interlinks different types of computational biology analyses as well as high-quality databases with in-depth information on a wide scope of scientific subjects. As a web application, ExPlain™ requires no client-side setup or installation, is available at any place with internet access, and makes complex computational analyses achievable even for researchers with limited system resources. Moreover, ExPlain's utility is reflected by a growing list of published scientific studies that have made use of it.

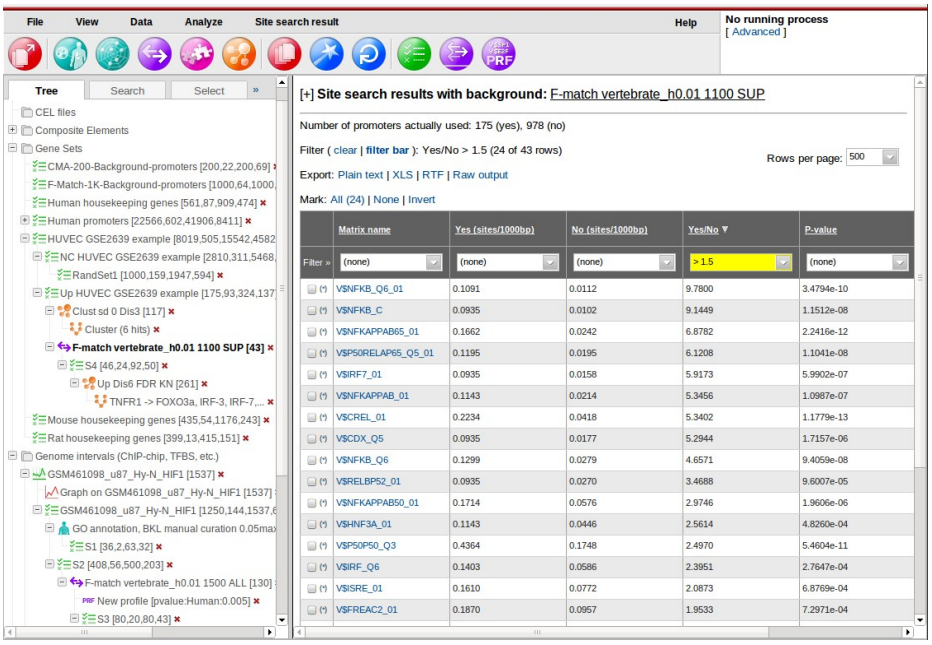

Figure 1. *The ExPlain™ interface.*

Genome-wide TAL1 occupancy profiles in two haematopoietic lineages

The data employed in this analysis were originally published by Palii et al. [3], who kindly made them available via the GEO database [4] (series GSE25000). Our case study is based on the ChIP-seq peaks derived by the authors and can be downloaded from GEO in BED format from the respective website [5].

For transcription factor binding site analysis of TAL1-occupied regions, we sampled, for each cell type, 500 peaks from the 1000 most strongly enriched locations that were not more than 1kb long and length-standardized them to 1000 bases. Corresponding sequences were extracted from the genome and loaded into ExPlain™.

Furthermore, we analyzed Gene Ontology (GO) biological processes [6] enriched in the genes located near TAL1-bound regions, where we accounted for all genes with a transcription start site (TSS) not farther than 100kb away from a ChIP-seq peak.

Comparison of enriched GO categories

ExPlain™ presents results of a Gene Ontology analysis in a table like the one shown in Figure 2 which was obtained for genes associated with TAL1-binding in Jurkat cells.

File View Data Analyze Functional classification

Go to

[+] **Functional analysis:** GO annotation, BKL manual curation 1max 1min on 'GSM614003_jurkat'

Filter (filter bar): none (total 2752 rows)

Export: Plain text | XLS | RTF

Mark: Page (100) | All (2752) | None | Invert

1 2 3 4

GO Identifier	GO Term	Ontology ▲	#Hits in group	Group size	#Hits expected	p-value
GO:0002376	immune system process	Biological process	103	1892	57	8.95049e-10
GO:0001775	cell activation	Biological process	54	803	25	2.26528e-08
GO:0045321	leukocyte activation	Biological process	47	688	21	1.24529e-07
GO:0048534	hemopoietic or lymphoid organ development	Biological process	44	652	20	4.50573e-07
GO:0002520	immune system development	Biological process	45	676	21	4.84838e-07
GO:0030097	hemopoiesis	Biological process	43	636	20	5.83168e-07
GO:0050793	regulation of developmental process	Biological process	119	2594	78	6.66728e-07
GO:0006461	protein complex assembly	Biological process	62	1140	35	3.54707e-06
GO:0070271	protein complex biogenesis	Biological process	62	1140	35	3.54707e-06
GO:0002521	leukocyte differentiation	Biological process	28	364	11	5.41608e-06
GO:0010941	regulation of cell death	Biological process	91	1946	59	9.35522e-06
GO:0006812	cation transport	Biological process	41	700	22	3.45986e-05
GO:0008219	cell death	Biological process	100	2266	68	3.47716e-05
GO:0016265	death	Biological process	100	2266	68	3.47716e-05
GO:0022607	cellular component assembly	Biological process	78	1663	50	4.15991e-05
GO:0042981	regulation of apoptosis	Biological process	87	1920	58	4.95295e-05
GO:0044093	positive regulation of molecular function	Biological process	66	1349	41	5.06832e-05
GO:0046649	lymphocyte activation	Biological process	33	525	16	5.34874e-05
GO:0012501	programmed cell death	Biological process	96	2187	66	6.38093e-05

Figure 2. *Gene Ontology functional analysis output in ExPlain™.*

In order to compare GO category P-values in erythroid and Jurkat cells, we made use of the possibility to extract ExPlain™ results for further analysis. A comparison of GO category enrichment P-values derived for gene sets that are associated with different biological conditions can lead to interesting results. Of most

interest are those categories that feature the strongest difference with regard to enrichment. Here we ordered GO categories by the difference of their log10-P-values obtained for erythroid and Jurkat gene sets.

Figures 3 and 4 show the 20 GO categories with strongest differential enrichment in Jurkat and erythroid cells, respectively. The results clearly reveal distinctive features of the two study cell types. Genes near TAL1 binding sites in Jurkat cells are involved in processes like leukocyte activation or lymphocyte differentiation (Figure 3).

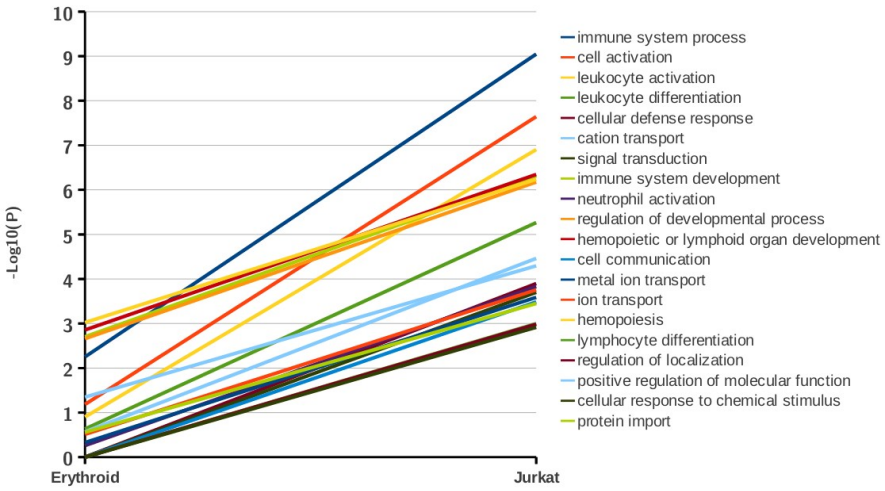

Figure 3. *GO biological processes with strongest differential enrichment in Jurkat cells.*

Likewise, GO biological processes related to erythrocyte differentiation or homeostasis as well as heme biosynthesis are more strongly enriched in the erythroid gene set (Figure 4). Notably, the results point to differences in metal ion transport.

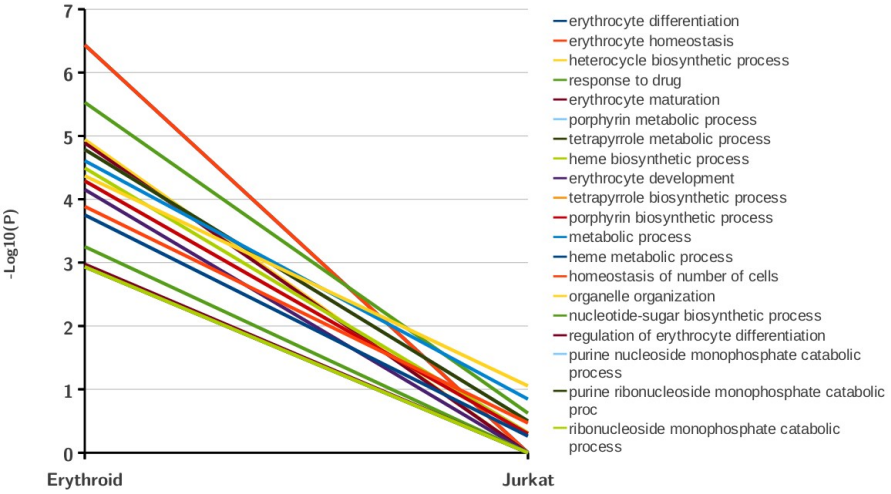

Figure 4. *GO biological processes with strongest differential enrichment in erythroid cells.*

Analysis of transcription factor binding sites

Transcription factor binding site analysis is a key strength of the ExPlain™ system and can be carried out with a defined set of algorithms from classical pattern analysis to complex composite module analysis. All of these methods use primarily the collection of TF binding sites and positional weight matrices (PWMs) from the TRANSFAC® database [7], but custom PWMs are supported as well.

We applied the Match™ program [8] of ExPlain™ to predict binding sites in the sequence sets consisting of 500 sequences described above as well as in a sequence set assembled from 1000 promoter regions of genes that were not in the vicinity of a TAL1 binding site in either cell type. The Match results were then subjected to F-Match analysis. The F-Match algorithm [9] was designed to discover transcription factors whose binding sites are enriched in a sequence set of interest, the so-called "Yes" set, compared to a sequence set that represents genomic background, the so-called "No" set.

The top 20 motifs, according to statistically corrected fold enrichment, identified in erythroid and Jurkat TAL-1-bound sequences are shown in Figures 5 and 6, respectively. Like the original authors we found enrichment of GATA motifs in both cell types as well as ETS and AML/RUNX1 motifs (represented by VAML_Q6, VCOREBINDING-FACTOR_Q6 and V$PEBP_Q6) in Jurkat cells [3].

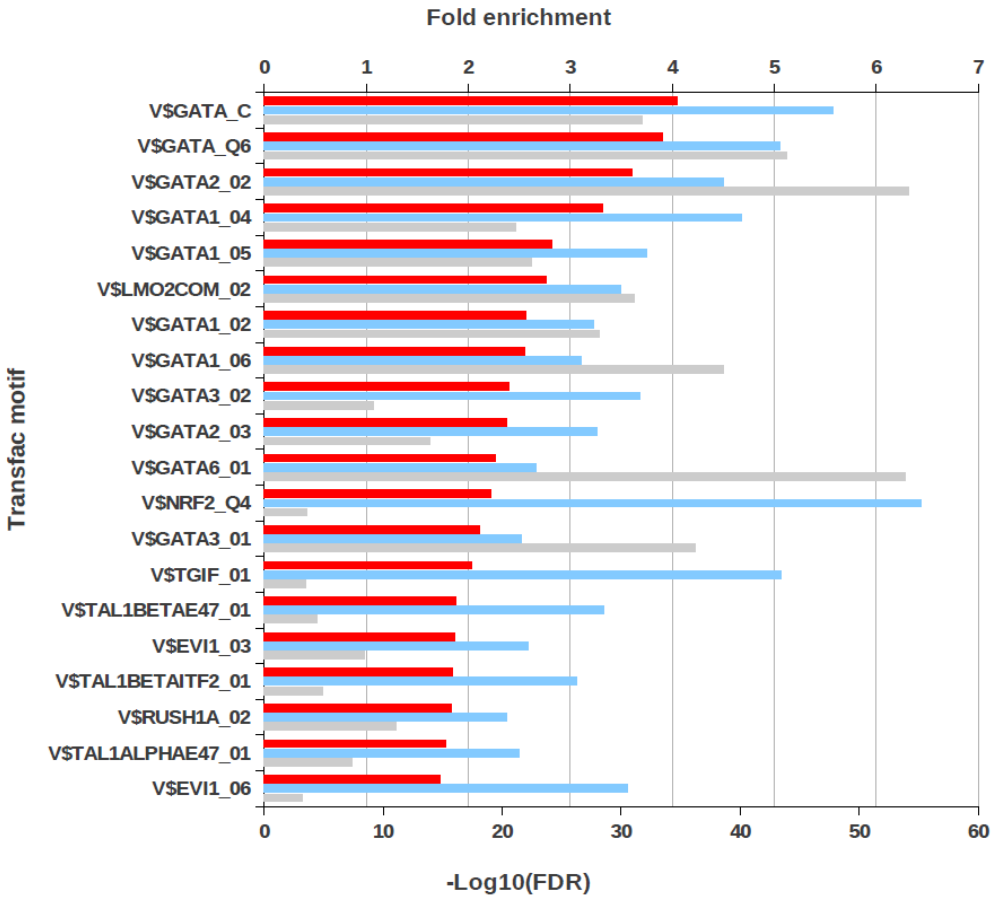

Figure 5. *Top 20 TRANSFAC motifs overrepresented in TAL1-bound regions from erythroid cells. Blue bars: raw fold enrichment of binding sites; Red bars: statistically corrected fold enrichment; Gray bars: negative log10-FDR of the one-tailed binomial test.*

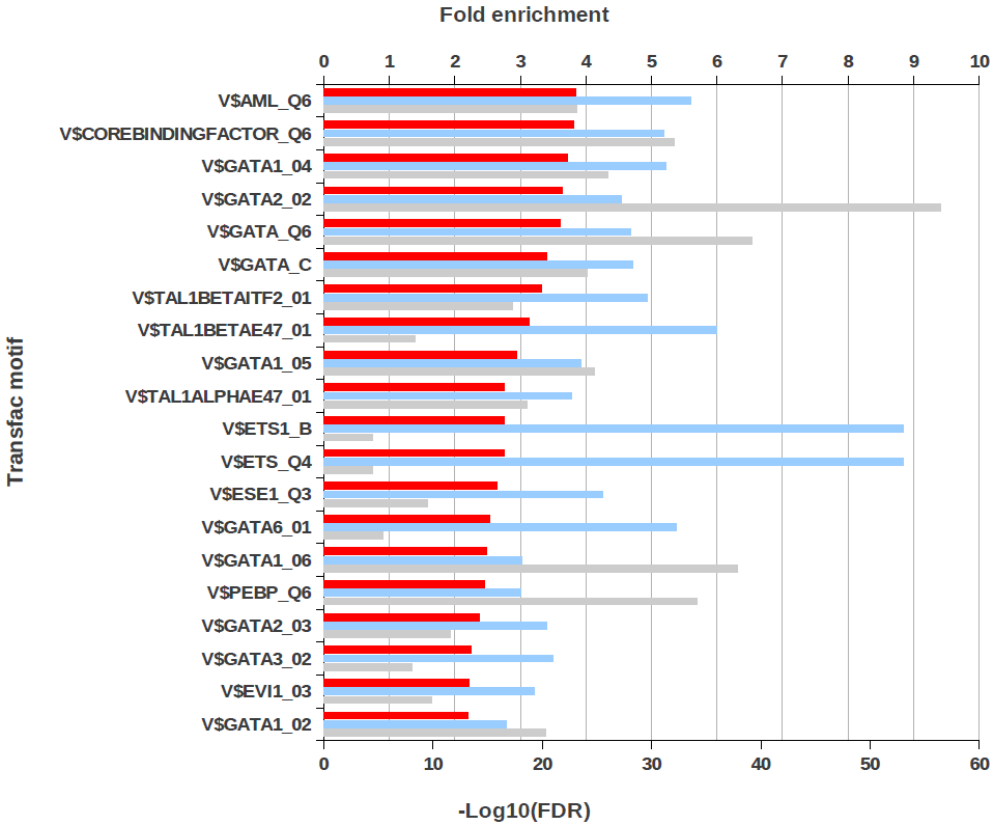

Figure 6. *Top 20 TRANSFAC motifs overrepresented in TAL1-bound regions from Jurkat cells. Blue bars: raw fold enrichment of binding sites; Red bars: statistically corrected fold enrichment; Gray bars: negative log10-FDR of the one-tailed binomial test.*

Composite module analysis

To take the exploration of co-regulators of TAL1 a step further, we subjected the subset of sequences from the Jurkat set that contained a site of the V$AML_Q6 or of the V$COREBINDINGFACTOR_Q6 motif to composite module analysis. The CMA [10] tool of ExPlain™ searches for combinations of transcription factor binding sites that best discriminate a *Yes* sequence set against a *No* sequence set.

With our described setup we sought to identify co-regulators that function together with TAL1 and RUNX1 in respective genomic regions. For the given data set, the CMA algorithm converged on a module composed of AML/RUNX1, GATA and MYB motifs. Figure 7 shows the interactive ExPlain™ visualization for several sequences with composite modules predicted by CMA. An expanded view with additional details about the individual binding site elements and the genomic sequence is shown at the top.

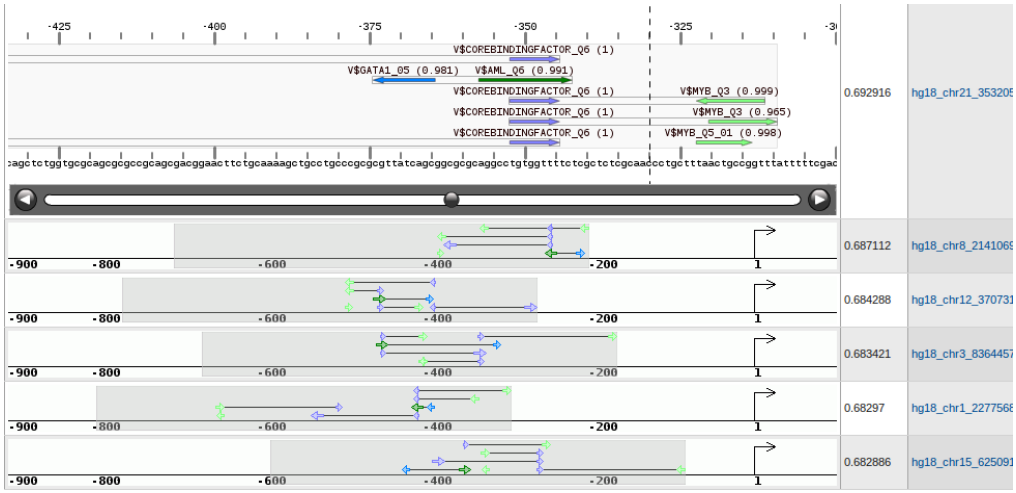

Figure 7. *Sequence view with composite modules predicted by CMA.*

We extracted two diagnostic plots from ExPlain™ (Figure 8). According to the upper plot the score distributions of *Yes* (red) and *No* (blue) set are well separated, which demonstrates that the reported TF combination provides for a distinctive feature of the study sequences.

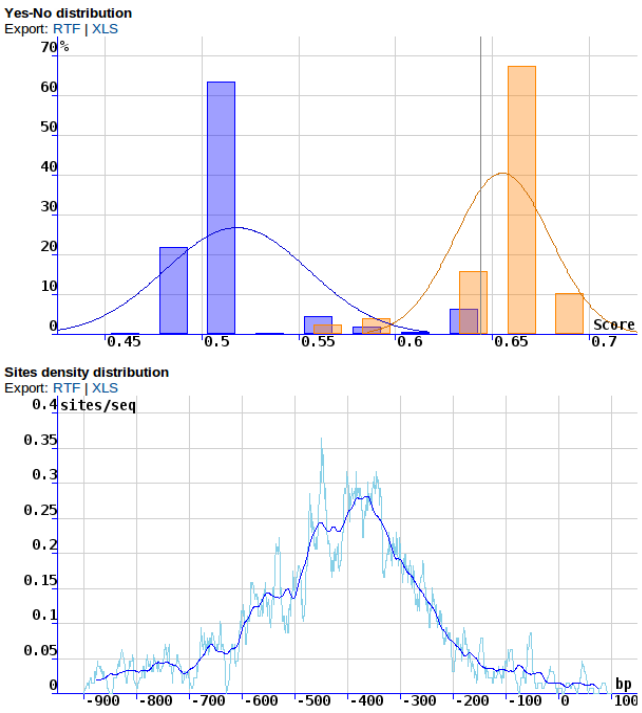

Figure 8. *Diagnostic plots of ExPlain's CMA tool showing the separation of Yes and No score distributions (top) as well as the location of predicted binding sites in sequences of the Yes set.*

Furthermore, the lower plot reveals that the predicted modules are primarily located in the middle of the (ChIP-seq derived) sequences, which supports the hypothesis that corresponding TFs cofunction with TAL1.

Notably, MYB is known to play a role in leukemia. Its expression is correlated with leukemia disease progression and it has been investigated as a drug target [11, 12], fitting well with the GO biological process profile of TAL1-associated genes in Jurkat cells. MYB may therefore constitute an additional transcription regulator that works coordinately with TAL1, GATA, and AML/RUNX1 in Jurkat cells.

Conclusion

The ExPlain™ system enables easy and powerful computational analysis of systems biological and genomics data. We have presented several possibilities to gain useful insight from ChIP-seq data using the ExPlain™ system. Here we addressed the exploration of functional roles of genes near TAL1 binding sites, identification of transcription regulators with enriched binding sites in TAL1-bound regions, and the proposition of a composite module of potential interest.

The functional roles associated with erythroid and Jurkat gene sets reflected distinct properties of these cell types. Transcription factor binding site analyses identified an enrichment of GATA motifs in both cell types, along with the unique enrichment of AML/RUNX1 and ETS motifs in Jurkat cells. More detailed composite module analysis additionally identified MYB as a potential co-regulator with TAL1, GATA, and AML/RUNX1 in Jurkat cells. Importantly, the results delivered by ExPlain™ are in agreement with those previously published on the data [3] and also suggest additional avenues for follow-up studies.

References

1. **ExPlain: finding upstream drug targets in disease gene regulatory networks.** Kel, A., Voss, N., Valeev, T., Stegmaier, P., Kel-Margoulis, O., and Wingender, E.; SAR QSAR Environ Res. 2008; 19:481-94.

2. **Advanced computational biology methods identify molecular switches for malignancy in an EGF mouse model of liver cancer.** Stegmaier, P., Voss, N., Meier, T., Kel, A., Wingender, E., and Borlak, J.; PLoS One. 2011; 6:e17738.

3. **Differential genomic targeting of the transcription factor TAL1 in alternate haematopoietic lineages.** Palii, C.G., Perez-Iratxeta, C., Yao, Z., Cao, Y., Dai, F., Davison, J., Atkins, H., Allan, D., Dilworth, F.J., Gentleman, R., Tapscott, S.J., and Brand, M.; EMBO J. 2011; 30:494-509.

4. **NCBI GEO: archive for high-throughput functional genomic data.** Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Muertter, R.N., and Edgar, R.; Nucleic Acids Res. 2009; 37:D885-890.

5. http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSE25000

6. **Gene Ontology: tool for the unification of biology.** The Gene Ontology Consortium; Nature Genet. 2000; 25:25-29.

7. **TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes.** Matys, V., Fricke, E., Geffers, R., Gößling, E., Haubrock, M., Hehl, R., Hornischer, K., 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.; Nucleic Acids Res. 2006; 34:D108-D110.

8. **MATCH: a tool for searching transcription factor binding sites in DNA sequences.** Kel, A.E., Gößling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O.V., and Wingender E.; Nucleic Acids Res. 2003; 31:3576–3579.

9. **Beyond microarrays: Find key transcription factors controlling signal transduction pathways.** Kel, A., Voss, N., Jauregui, R., Kel-Margoulis, O. and Wingender, E.; BMC Bioinformatics. 2006; 7: S13.

10. **Composite Module Analyst: A fitness-based tool for identification of transcription factor binding site combinations.** Kel, A., Konovalova, T., Valeev, T., Cheremushkin, E., Kel-Margoulis, O., and Wingender, E.; Bioinformatics. 2006; 22:1190-1197.

11. **In vivo treatment of human leukemia in a scid mouse model with c-myb antisense oligodeoxynucleotides.** Ratajczak, M.Z., Kant, J.A., Luger, S.M., Hijiya, N., Zhang, J., Zon, G., and Gewirtz, A.M.; Proc Natl Acad Sci USA. 1992; 89:11823-11827.

12. **Target site search and effective inhibition of leukaemic cell growth by a covalently closed multiple anti-sense oligonucleotide to c-myb.** Moon, I.J., Lee, Y., Kwak, C.S., Lee, J.H., Choi, K., Schreiber, A.D., and Park, J.G.; Biochem J. 2000; 346.2:295-303.

Philip Stegmaier
philip.stegmaier@biobase-international.com

BIOBASE GmbH
Halchtersche Strasse 33
D-38304 Wolfenbuettel
Germany