Chapter Twelve

Transcription Regulatory Networks Analysis Using CAGE

Jesper Tegnér[1,2,*], Johan Björkegren[1,2,†], Timothy Ravasi[3−5,‡] and Vladimir B. Bajic[6,§]

[1] *King Gustaf V Research Institute, Karolinska Institutet, Sweden.*
[2] *Department of Physics, Linköping University, Sweden.*
[3] *Scripps NeuroAIDS Preclinical Studies (SNAPS), USA*
[4] *Jacobs School of Engineering, University of California, USA*
[5] *Computational Bioscience Research Centre (CBRC), King Abdullah University of Science and Technology (KAUST), Saudi Arabia*
[6] *South African National Bioinformatics Institute (SANBI), University of Western Cape, South Africa*
Email: [*] *jesper.tegner@ki.se,* [†] *johan.bjorkegren@ki.se,* [‡] *timothy.ravasi@kaust.edu.sa,* [§] *vladimir.bajic@kaust.edu.sa*

Mapping out cellular networks in general and transcriptional networks in particular has proved to be a bottle-neck hampering our understanding of biological processes. Integrative approaches fusing computational and experimental technologies for decoding transcriptional networks at a high level of resolution is therefore of uttermost importance. Yet, this is challenging since the control of gene expression in eukaryotes is a complex multi-level process influenced by several epigenetic factors and the fine interplay between regulatory proteins and the promoter structure governing the combinatorial regulation of gene expression. In this chapter we review how the CAGE data can be integrated with other measurements such as expression, physical interactions and computational prediction of regulatory motifs, which together can provide a genome-wide picture of eukaryotic transcriptional regulatory networks at a new level of resolution.

Cap Analysis Gene Expression (CAGE): The Science of Decoding Gene Transcription edited by P Carninci

www.panstanford.com
978-981-4241-34-2

12.1 CAGE DATA FOR NETWORK RECONSTRUCTION

Every molecular process occurring in a living cell is a concerted activity of numerous players. Networks, which are defined by the interactions between genes and proteins, govern critical cellular functions such as differentiation and cell death. A great challenge of modern biology is to elucidate these mechanisms and to decipher the corresponding actors in these molecular networks.[20,46] There are, however, numerous layers of interacting molecules and modern experimental and computational technologies have opened new possibilities for obtaining deeper insights into the intrinsic machinery governing the molecular interactions. From a global point of view, one can consider the plethora of interacting molecules as a network of molecular entities that dynamically form under specific intra- and extra-cellular demands and execute their programmed actions accordingly. These networks are characterized not only by the participating molecular entities and their interactions, but also by the direction and dynamics of the interactions.[31,35,47] From this perspective cellular networks are complex systems represented by numerous cause-consequence relationships.

Whole-genome technologies for profiling the molecules within these networks have been instrumental in generating cellular fingerprints obtained during different conditions Monitoring SNPs, mRNA, proteins and metabolites in this manner produces large amounts of data which requires an appropriate computational toolbox for the analysis.[20,38] In this context, the invention of CAGE technology[8,24] has opened yet another chapter into the generation of molecular data that can support biological network reconstruction at a new level of resolution. CAGE tags can also be produced massively under specific cellular and environmental conditions.[7,8,35] In contrast to micro-array gene expression technology monitoring only mRNA levels,[9] CAGE provides unprecedented data on the individual transcription start sites (TSSs) that are effectively utilized during the conditions of the experiment, thus linking the biological process/reaction to the actual control regions of the affected genes. Secondly, CAGE enables a quantitative measure of the gene expression level by utilizing the actual counts of the CAGE tags. In other words, CAGE allows for direct linking the expression events with the effectively utilized regulatory regions of the

gene, as well as the actual consequence of the use of these control regions through the measure of transcription/expression via CAGE tags.[7,8]

Clearly, these two advantages position the CAGE technology for providing explicit information necessary for building molecular networks at a new level of resolution as compared to using regular micro-arrays. It is useful in this context to observe a broader schema of the events that lead to the formation and activities of molecular networks.[20] We will present this in a very rough and simplified form. When an external stimulus occurs, sensors on the cell surface react and transmit the chemical information to the interior of the cell. Signal receptors, normally a group of specialized proteins, do activate but could also interact with other down-stream molecules and initiate a set of chain reactions conducting signals to the nucleus, where transcription factors (TFs) and their protein-complexes interact with the DNA initiating transcription of genes. If we pause at this point, we may refer to the entire set of chemical reactions occurring in this process as a signaling pathway and the set of molecular reactions and interactions as a signaling network with the genes being the terminal entities.[46] From the viewpoint of CAGE technology, it provides us data that associates events at the level of interactions between TFs to transcription factor binding sites (TFBSs) on the regulatory regions of affected genes.[35,47] These interactions can be schematically viewed as directed links such as

$$\text{TF}\rightarrow\text{TFBS}\rightarrow\text{promoter}\rightarrow\text{gene}$$

By considering the gene as a terminal entity in networks we can apply a bottom-up approach for reconstructing the gene part of biologically meaningful molecular networks. Thus, we will not consider the metabolic components in this chapter. In Table 12.1, we put on view the type of entities we will consider for network reconstruction and the corresponding major informatics and genomics resources.

The use of the CAGE tag technology for improving the prediction of TFBSs is discussed in the previous chapter of this monograph. There, it is demonstrated how information provided by CAGE can be combined with the sequence analysis to produce improved predictions of TFBSs, thereby detecting more accurate links of the type.

$$\text{TF}\rightarrow\text{TFBS}\rightarrow\text{promoter}\rightarrow\text{gene}.$$

Type of molecules	Source of information
DNA-binding TFs	Genome annotations Expression measurements
Cofactors (co-activators/ repressors; chromatin remodeling)	Genome annotations Expression measurements
Regulatory regions TFBSs	CAGE TSS mapping Computational motifs search and discovery
Genes	CAGE TSS mapping Genome annotations Expression measurements
Type of interactions	
TF-DNA (TF-TFBS; TF-regulatory region)	Computational inferences Genome-wide location data (GWLA)
PPI (TF-TF; TF-Cofactor)	Protein-protein interaction databases

What is not transparent from CAGE data is how TFs interact with other proteins in the nucleus. It is well known that frequently TFs require other proteins known as co-factors to interact with them and form complexes that are capable of direct binding to DNA. One data-source that can provide such information is represented by protein-protein interactions (PPIs) repositories. By investigating all possible interactions between TFs and other proteins, we obtain a list of putative candidates of proteins that can operate as co-factors. Since the co-factors interact not only with TFs, but also with other proteins, we can gradually expand the network with more distant layers of putative regulatory proteins. Here we aim at the transcription regulatory network (TRN), extended by additional layers of co-factors and complemented by other PPIs.

12.2 METHODOLOGY

To make the presentation of the methodology simpler, we schematically depict it in Fig. 12.1 and describe in detail in the following sections the datasets required.

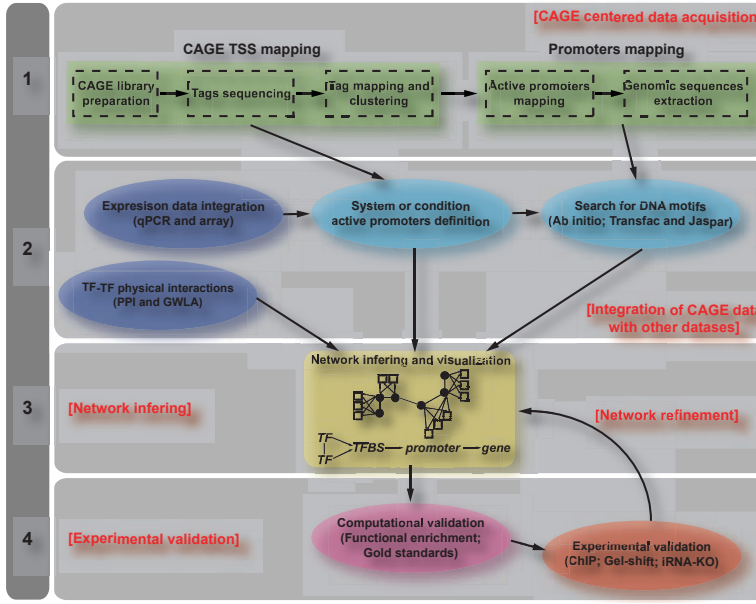

Figure 12.1. Flowchart of the method used to infer CAGE-based transcription regulatory networks (TRN). [Note by the editor: research is in progress to use CAGE data also for expression analysis to determine the expression at each promoter level].

12.2.1 Step 1 of the Process

The first step is the production of CAGE libraries for the system under investigation (see previous chapters). After deep sequencing of these CAGE libraries, the tags are mapped to the genome to determine CAGE defined transcription starting sites (CTSS; Chapter 10). This enables the identification of active starting sites, and hence the promoters for which the genome sequences can be extracted. Furthermore, the number of tags corresponding to the CTSS reflects the expression of the gene associated with vicinity of the CTSS.

12.2.2 Step 2 of the Process

CAGE expression and mapping data are integrated with other expression data in order to infer all TFs and the regulated genes which are expressed by the system, representing the nodes of the network. The promoters of the expressed nodes are then scanned

for enriched TFBSs using model based or *de novo* methods. This defines the promoter architecture of the nodes (see the previous Chapter 11). Physical interaction data, such as PPI and protein-DNA interactions, can be also used to increase confidence of the bioinformatics predictions.

12.2.3 Step 3 of the Process

The global network is then inferred by combining the nodes using the inferred (*TF→TFBS*) and physical (PPI and TF-DNA) interactions (edges of the network). The network can be represented graphically where the nodes are usually illustrated by solid shapes and edges are denoted by dashed lines for TF-DNA interaction and as blue lines for PPI. See Fig. 12.2 for an example of the human cerebellum CAGE-derived transcription regulatory network.

12.2.4 Step 4 of the Process

The inferred network is then validated using bioinformatics and experimental approaches. The results of validation can be used to refine the original network model.

12.3 GENE EXPRESSION DATA COMPLEMENTARY TO CAGE FOR NETWORK RECONSTRUCTION

Although CAGE provides for digital counting of gene expression, expression profiling using micro-array chips is by far the most popular genome-wide technology for capturing genomic reaction of a cell.[9,41] Expression micro-array is an RNA based method that allows the simultaneous measurement of virtually all the transcripts in a cell. This has been and is still a powerful and wide-spread technique thanks to the relative technical simplicity, low cost, and short turn-over time, which make expression micro-arrays a standard molecular biology technique available to any laboratory. Computational methods used for the analysis of these large collections of data have also been improved and standardized, making the interpretation of micro-array data more accessible to those without a strong computational background.[6,29,38] Although with less throughput than CAGE and chip-based technologies, quantitative real-time PCR

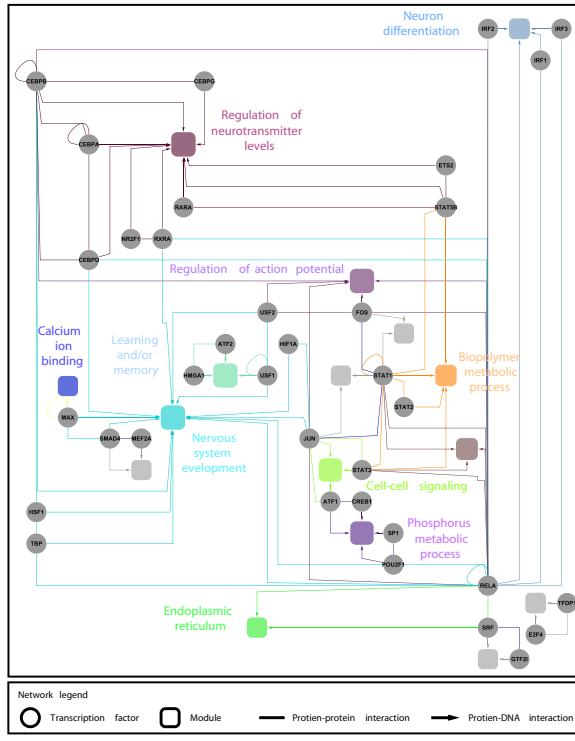

Figure 12.2. A Cerebellar TRN was inferred using CAGE active promoters as described in the text and illustrated in Fig. 12.1 In this particular view, genes expressed in cerebellum are grouped into functional modules and represented as single node (squared nodes) whose size is proportional to the number of genes in the module.

(qRT-PCR) is becoming an increasingly important complementary tool in particular in the construction of TRNs,[35,47] due to its quantitative nature and higher sensitivity which allows for more accurate measurements of low abundant transcripts such as those encoding for transcription factors.[19]

12.4 USING PHYSICAL INTERACTIONS

The edges of a transcriptional regulatory network contain two types of physical interactions: hidden i.e. those between the regulatory proteins and their DNA binding sequences (PDIs) and

those between regulatory proteins (PPIs). In eukaryotes, the regulation of gene expression often requires more than one TF to ensure a proper expression of a single gene. TFs interact to form protein complexes and in many cases this is a requirement for the binding of DNA regulatory elements.[4,13,18,23,27,3] For example, this is the case for homo-dimers binding palindromic transcription factor binding sites (TFBS).[49] In the genome, TFBSs tend to cluster together in specific and conserved regions whereas TFs interact at the protein level forming protein complexes that include chromatin remodeling factors.[2,14,30,48] Last, but no less important, are the transcriptional initiation complexes, which despite the fact that they are composed of more than 30 proteins will bind specific regulatory elements via just a few core components such as the TATA box Binding Protein (TBP).[10,25,26,42] The interplay between TFs is often referred to as the combinatorial regulation of gene expression. Therefore capturing all possible combinatorial interactions between TFs is an essential step toward the construction of mammalian transcription regulatory networks. For this purpose the complete maps of PPI are of uttermost value as a first step to map putative pair-wise interactions. PPIs are usually generated by two-hybrid technologies (Y2H).[21] PPI maps can also be constructed using co-immunoprecipitation followed by mass spectrometry.[1,5,12,15,17,28] This technology is more specific than Y2H (less false positive rate) and less scalable. Since the technology relies on co-immunoprecipitation it is more suitable to identify protein complexes with indirect interactions, in contrast to Y2H which instead measures pair-wise, binary and direct interactions.

In recent years the number of binary non-redundant human PPIs has increased dramatically thanks to extensive literature mining (36,617 in the HPRD database)[33,37] and also to large scale experimentally determined human PPIs such as the work from Rual and colleagues and Ewing and colleagues.[12,40] However, one of the limitations with the current human PPI maps is the low coverage of TF interactions because the experimental techniques generally bias toward large macromolecular complexes (i.e. ribosome, splicesosome, membranes channels etc.) and due to a low abundance of TFs compared to cytosolic proteins. Suzuki and colleagues of the RIKEN Genome Science Center in Japan have generated for the first time nuclear specific PPI maps for mouse[44] and now they are focusing on the human nuclear PPI maps (personal

communication). Such maps will be very useful resources for the construction of mammalian TRNs. To regulate gene expression, either individual TFs or complexes of TFs need to first bind specifically to cis-regulatory DNA sequences, which are usually at the 5′ end of genes. The most common methods to infer TF-DNA binding events are computational ones (see previous sections).

Technologies have emerged that enable *in vivo* genome-wide experimental mapping of TF-DNA binding events. The most wide-spread of these techniques is the Genome Wide Location Analysis (GWLA) also known as ChIP-chip or ChIP-PET.[11,16,39,51] In GWLA analysis, TF-DNA binding events are captured and frozen in a specific cellular state by *in vivo* crosslinking. Then the genomic DNA is fragmented and the TF of interest is isolated with a specific antibody, along with those genomic fragments bounded by the TF. After crosslinking reversal and protein digestion, the pulled down DNA is labeled in a manner analogous to a cDNA microarray experiment, but hybridized to an oligo micro-array chip whose content is directed towards regulatory regions rather than exons. GWLA are powerful techniques, since they capture *in vivo* and in a high-throughput fashion empirical binding events, thus the TF binding events can be compared across several cellular conditions (drug stimulation, developmental stage, etc.), but contrary to CAGE, they are restricted to one TF at a time.

12.5 TRNS RECONSTRUCTION

The reconstruction of TRN is based on combining the edges as the basic network building blocks. An edge is a link of the form entity1 — entity2, or entity1→entity2. The first type of link is non-directional and is characteristic of PPIs. The second type of link is directional and several types can be derived based on CAGE. As discussed earlier, CAGE can help us elucidate links TF→TFBS→promoter→gene. These can be split into several simpler types, such as TF→TFBS, TF→promoter, TFBS→promoter, TFBS→gene, etc., depending on what details of the interaction we are interested exploring. But we can also expand these by TF→cofactor, TF→protein, cofactor→protein, protein→protein and TF-DNA (ChIP-chip) physical interactions, as mentioned earlier. Using these constituents, blocks of complex TRNs can be reconstructed. Figure 12.2 displays an example of a

reconstructed mammalian TRN using data from the CAGE technology (Fig. 12.1).

Clearly the network structure suggests novel mechanistic hypotheses which must be experimentally tested as a final validation step. However, before this step is taken it is mandatory to consider that networks are condition and state-dependent, that is, different parts of the network will be active during different conditions.[31,35] For example, a cell which is exposed to a particular compound or a physiological condition such as stress, will produce two different activity patterns. A network can be generated without using information from expression, but only from 'static' sequence analysis. Such a network we can denote as a static network Therefore, a static network has to be evaluated and projected onto the specific condition of interest. Such a network projection can be performed in space (over different organs/tissues) and/or in time (in response to a stimulation for example) as shown in Fig. 12.2 for human cerebellum.

12.6 USING PATHWAY INFORMATION

Pathways are nothing else but a collection of molecular reactions occurring collectively under specific conditions.[23] In the context of TRNs they are very useful as TRNs can be matched to the pathways and specific segments of TRNs (nodes) could be found in an enriched manner in specific pathways.[34,43] TRN itself will provide a broader context of the pathway functioning and it can suggest possible additional pathway members based on network properties. Vice versa, TRN can be expanded by the other members of the significantly hit pathways: these other not been included in the information from which TRN was reconstructed. On the other hand can pathways provide for the interpretat of the biological role of the TRN and its constitutive elements (Fig. 12.2).

12.7 VALIDATION OF THE RECONSTRUCTED NETWORKS

Since reconstruction of TRNs is a complex process that requires integration and processing of data originating from a variety of resources, it is necessary to make an assessment of the quality of the reconstructed TRNs. At the end, the predictions based

on the network analysis have to be evaluated experimentally, but before proceeding to the laboratory there are several useful computational validation steps that can and should be employed. These include statistics, benchmarking against current knowledge, and biological relevance of the annotations associated with the extracted parts lists and pathways.

A proper use of statistics usually goes beyond regular parametric testing including a t-test. As an example, it has become increasingly clear that to evaluate the significance of a particular network motif such as a feed-forward structure, it is necessary to compare the occurrence of the feed-forward loop against a null distribution. Such a statistical randomization procedure can be adapted to the analysis of other features of the reconstructed network, and basically informs us about the significance of a particular finding. Next, it is useful to compare the reconstructed TRN with what is currently known. For example, we can construct a "gold standard list" which we can use for comparison with the predicted TRN. Such a gold standard list is composed of a set of interactions (edges) that have been extensively and experimentally validated and some are available in the literature. A successful validation can be measured using a combined measure of the number of true positive (TP) gold standard interactions recovered (present) in the inferred network as well as those not captured (false negatives, FN) by the predicted TRN. Finally, TRNs can be also computationally validated by biological context relevance, by applying Gene Ontology and Pathways enrichment analyses[3] in order to detect those sub-networks of the TRN that "make sense" in the biology or systems under study. For example, if we are studying a system resembling the brain development, we expect to find regions in the network that are enriched in genes which are involved in processes such as in "nervous systems development", "neuron differentiation", etc.

Although computational validations can be very useful to increase the confidence of the inferred TRNs, it is necessary to assess the novel predicted interactions or regulatory events through functional validation by conducting experiments in the laboratory. Unfortunately, up to date, such an experimental validation is only possible for a handful of targets, because the experiments required to comprehensively assess the biological role and context of a regulatory event are laborious and therefore not yet scalable to a larger number of targets.

References

[1] R. Aebersold and M. Mann. Mass spectrometry-based proteomics, *Nature* **422**, 198 (2003).

[2] J. A. Armstrong and B. M. Emerson. Transcription of chromatin: These are complex times, *Curr. Opin. Genet. Dev.* **8**, 165 (1998).

[3] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium, *Nat. Genet.* **25**, 25 (2000).

[4] M. Bellorini, D. K. Lee, J. C. Dantonel, K. Zemzoumi, R. G. Roeder, L. Tora and R. Mantovani. CCAAT binding NF-Y-TBP interactions: NF-YB and NF-YC require short domains adjacent to their histone fold motifs for association with TBP basic residues. *Nucleic Acids Res.* **25**, 2174 (1997).

[5] B. Blagoev, I. Kratchmarova, S. E. Ong, M. Nielsen, L. J. Foster and M. Mann. A proteomics strategy to elucidate functional protein-protein interactions applied to EGF signaling. *Nat. Biotechnol.* **21**, 315 (2003).

[6] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo and M. Vingron. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* **29**, 365 (2001).

[7] P. Carninci. Tagging mammalian transcription complexity. *Trends Genet.* **22**, 501 (2006).

[8] P. Carninci, A. Sandelin, B. Lenhard, S. Katayama, K. Shimokawa, J. Ponjavic, C. A. Semple, M. S. Taylor, P. G. Engstrom, M. C. Frith, A. R. Forrest, W. B. Alkema, S. L. Tan, C. Plessy, R. Kodzius, T. Ravasi, T. Kasukawa, S. Fukuda, M. Kanamori-Katayama, Y. Kitazume, H. Kawaji, C. Kai, M. Nakamura, H. Konno, K. Nakano, S. Mottagui-Tabar, P. Arner, A. Chesi, S. Gustincich, F. Persichetti, H. Suzuki, S. M. Grimmond, C. A. Wells, V. Orlando, C. Wahlestedt, E. T. Liu, M. Harbers, J. Kawai, V. B. Bajic, D. A. Hume and Y. Hayashizaki. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* **38,** 626 (2006).

[9] J. DeRisi, L. Penland, P. O. Brown, M. L. Bittner, P. S. Meltzer, M. Ray, Y. Chen, Y. A. Su and J. M. Trent. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat. Genet.* **14,** 457 (1996).

[10] A. Dvir, J. W. Conaway and R. C. Conaway. Mechanism of transcription initiation and promoter escape by RNA polymerase II. *Curr. Opin. Genet. Dev.* **11**, 209 (2001).

[11] G. M. Euskirchen, J. S. Rozowsky, C. L. Wei, W. H. Lee, Z. D. Zhang, S. Hartman, O. Emanuelsson, V. Stolc, S. Weissman, M. B. Gerstein, Y. Ruan and M. Snyder. Mapping of transcription factor binding regions in mammalian cells by ChIP: Comparison of array- and sequencing-based technologies. *Genome Res.* **17**, 898 (2007).

[12] R. M. Ewing, P. Chu, F. Elisma, H. Li, P. Taylor, S. Climie, L. McBroom-Cerajewski, M. D. Robinson, L. O'Connor, M. Li, R. Taylor, M. Dharsee, Y. Ho, A. Heilbut, L. Moore, S. Zhang, O. Ornatsky, Y. V. Bukhman, M. Ethier, Y. Sheng, J. Vasilescu, M. Abu-Farha, J. P. Lambert, H. S. Duewel, II. Stewart, B. Kuehl, K. Hogue, K. Colwill, K. Gladwish, B. Muskat, R. Kinach, S. L. Adams, M. F. Moran, G. B. Morin, T. Topaloglou and D. Figeys. Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol. Syst. Biol.* **3**, 89 (2007).

[13] J. V. Falvo, A. M. Uglialoro, B. M. Brinkman, M. Merika, B. S. Parekh, E. Y. Tsai, H. C. King, A. D. Morielli, E. G. Peralta, T. Maniatis, D. Thanos and A. E. Goldfeld. Stimulus-specific assembly of enhancer complexes on the tumor necrosis factor alpha gene promoter. *Mol. Cell. Biol.* **20**, 2239 (2000).

[14] M. Gaestel. Molecular chaperones in signal transduction. *Handbook Exp. Pharmacol*. 93 (2006).

[15] A. C. Gingras, R. Aebersold and B. Raught. Advances in protein complex analysis using mass spectrometry. *J. Physiol.* **563**, 11 (2005).

[16] N. D. Heintzman, R. K. Stuart, G. Hon, Y. Fu, C. W. Ching, R. D. Hawkins, L. D. Barrera, S. Van Calcar, C. Qu, K. A. Ching, W. Wang, Z. Weng, R. D. Green, G. E. Crawford and B. Ren. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39**, 311 (2007).

[17] Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S. L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. R. Willems, H. Sassi, P. A. Nielsen, K. J. Rasmussen, J. R. Andersen, L. E. Johansen, L. H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. D. Sorensen, J. Matthiesen, R. C. Hendrickson, F. Gleeson, T. Pawson, M. F. Moran, D. Durocher, M. Mann, C. W. Hogue, D. Figeys and M. Tyers: Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415**, 180 (2002).

[18] A. Hoffmann, T. Oelgeschlager and R. G. Roeder. Considerations of transcriptional control mechanisms: Do TFIID-core promoter complexes recapitulate nucleosome-like functions? *Proc. Natl. Acad. Sci. USA* **94**, 8928 (1997).

[19] M. J. Holland. Transcript abundance in yeast varies over six orders of magnitude. *J. Biol. Chem.* **277**, 14363 (2002).

[20] T. E. Ideker. Network genomics. *Ernst. Schering. Res. Found. Workshop.* 89 (2007).

[21] J. K. Joung, E. I. Ramm and C. O. Pabo. A bacterial two-hybrid selection system for studying protein-DNA and protein-protein interactions. *Proc. Natl. Acad. Sci. USA* **97**, 7382 (2000).

[22] M. Kanehisa and S. Goto. KEGG. Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27 (2000).

[23] T. K. Kim and T. Maniatis. The mechanism of transcriptional synergy of an in vitro assembled interferon-beta enhanceosome. *Mol. Cell.* **1**, 119 (1997).

[24] R. Kodzius, M. Kojima, H. Nishiyori, M. Nakamura, S. Fukuda, M. Tagami, D. Sasaki, K. Imamura, C. Kai, M. Harbers, Y. Hayashizaki and P. Carninci. CAGE: Cap analysis of gene expression. *Nat. Methods* **3**, 211 (2006).

[25] M. Kozak. Initiation of translation in prokaryotes and eukaryotes. *Gene.* **234**, 187 (1999).

[26] M. Kunzler, C. Springer and G. H. Braus. The transcriptional apparatus required for mRNA encoding genes in the yeast Saccharomyces cerevisiae emerges from a jigsaw puzzle of transcription factors. *FEMS Microbiol. Rev.* **19**, 117 (1996).

[27] A. B. Lassar, P. L. Martin and R. G. Roeder. Transcription of class III genes: Formation of preinitiation complexes. *Science* **222**, 740 (1983).

[28] D. Lin, D. L. Tabb and J. R. Yates III. Large-scale protein identification using mass spectrometry. *Biochim. Biophys. Acta.* **1646**, 1 (2003).

[29] D. W. Lin and P. S. Nelson. Microarray analysis and tumor classification. *N. Engl. J. Med.* **355**, 960; author reply 960 (2006).

[30] R. X. Luo and D. C. Dean. Chromatin remodeling and transcriptional regulation, *J. Natl. Cancer. Inst.* **91**, 1288 (1999).

[31] N. M. Luscombe, M. M. Babu, H. Yu, M. Snyder, S. A. Teichmann, and M. Gerstein. Genomic analysis of regulatory network dynamics reveals large topological changes, *Nature* **431**, 308 (2004).

[32] M. Mann, R. C. Hendrickson and A. Pandey. Analysis of proteins and proteomes by mass spectrometry. *Annu. Rev. Biochem.* **70**, 437 (2001).

[33] S. Mathivanan, B. Periaswamy, T. K. Gandhi, K. Kandasamy, S. Suresh, R. Mohmood, Y. L. Ramachandra and A. Pandey. An evaluation of human protein-protein interaction data in the public domain. *BMC Bioinformatics* **7** (Suppl 5), S19 (2006).

[34] V. K. Mootha, C. M. Lindgren, K. F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstrale, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander,

J. N. Hirschhorn, D. Altshuler and L. C. Groop. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267 (2003).

[35] R. Nilsson, V. B. Bajic, H. Suzuki, D. di Bernardo, J. Bjorkegren, S. Katayama, J. F. Reid, M. J. Sweet, M. Gariboldi, P. Carninci, Y. Hayashizaki, D. A. Hume, J. Tegner and T. Ravasi. Transcriptional network dynamics in macrophage activation. *Genomics* **88**, 133 (2006).

[36] T. Oelgeschlager, C. M. Chiang and R. G. Roeder. Topology and reorganization of a human TFIID-promoter complex. *Nature* **382**, 735 (1996).

[37] S. Peri, J. D. Navarro, R. Amanchy, T. Z. Kristiansen, C. K. Jonnalagadda, V. Surendranath, V. Niranjan, B. Muthusamy, T. K. Gandhi, M. Gronborg, N. Ibarrola, N. Deshpande, K. Shanker, H. N. Shivashankar, B. P. Rashmi, M. A. Ramya, Z. Zhao, K. N. Chandrika, N. Padma, H. C. Harsha, A. J. Yatish, M. P. Kavitha, M. Menezes, D. R. Choudhury, S. Suresh, N. Ghosh, R. Saravana, S. Chandran, S. Krishna, M. Joy, S. K. Anand, V. Madavan, A. Joseph, G. W. Wong, W. P. Schiemann, S. N. Constantinescu, L. Huang, R. Khosravi-Far, H. Steen, M. Tewari, S. Ghaffari, G. C. Blobe, C. V. Dang, J. G. Garcia, J. Pevsner, O. N. Jensen, P. Roepstorff, K. S. Deshpande, A. M. Chinnaiyan, A. Hamosh, A. Chakravarti and A. Pandey. Development of human protein reference database as an initial platform for approaching systems biology in humans, *Genome. Res.* **13,** 2363 (2003).

[38] J. Quackenbush. Extracting biology from high-dimensional biological data. *J. Exp. Biol.* **210**, 1507 (2007).

[39] B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell and R. A. Young. Genome-wide location and function of DNA binding proteins. *Science.* **290**, 2306 (2000).

[40] J. F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. Li, J. S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. E. Cusick, D. E. Hill, F. P. Roth and M. Vidal. Towards a proteome-scale map of the human protein-protein interaction network, *Nature* **437**, 1173 (2005).

[41] M. Schena, D. Shalon, R. W. Davis and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467 (1995).

[42] C. A. Spencer and M. Groudine. Transcription elongation and eukaryotic gene regulation. *Oncogene.* **5**, 777 (1990).

[43] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, *Proc. Natl. Acad. Sci. USA* **102,** 15545 (2005).

[44] H. Suzuki, R. Saito, M. Kanamori, C. Kai, C. Schonbach, T. Nagashima, J. Hosaka and Y. Hayashizaki. The mammalian protein-protein interaction database and its viewing system that is linked to the main FANTOM2 viewer. *Genome. Res.* **13**, 1534 (2003).

[45] T. Tamura, Y. Makino and T. Kishimoto. [Regulation of gene expression and recent advance on transcription studies]. *Nippon Rinsho.* **53**, 1033 (1995).

[46] K. Tan, J. Tegner, and T. Ravasi. Integrated approaches to uncovering transcription regulatory networks in mammalian cells. *Genomics*. **91**(3), 219–231, Epub 2008 Jan 8 Review., (2008).

[47] J. Tegner, R. Nilsson, V. B. Bajic, J. Bjorkegren and T. Ravasi: Systems biology of innate immunity, *Cell. Immunol*. **244,** 105 (2006).

[48] T. Tsukiyama and C. Wu. Chromatin remodeling and transcription. *Curr. Opin. Genet. Dev.* **7**, 182 (1997).

[49] S. C. Tucker and R. Wisdom. Site-specific heterodimerization by paired class homeodomain proteins mediates selective transcriptional responses. *J. Biol. Chem*. **274**, 32325 (1999).

[50] M. W. Van Dyke, M. Sawadogo and R. G. Roeder. Stability of transcription complexes on class II genes. *Mol. Cell. Biol.* **9**, 342 (1989).

[51] C. L. Wei, Q. Wu, V. B. Vega, K. P. Chiu, P. Ng, T. Zhang, A. Shahab, H. C. Yong, Y. Fu, Z. Weng, J. Liu, X. D. Zhao, J. L. Chew, Y. L.Lee, V. A. Kuznetsov, W. K. Sung, L. D. Miller, B. Lim, L. Lee, E. T. Liu, Q. Yu, H. H. Ng and Y. Ruan. A global map of p53 transcription-factor binding sites in the human genome. *Cell*. **124**, 207 (2006).