Update

Genome Analysis

Comparison of transcription regulatory interactions inferred from high-throughput methods: what do they reveal?

S. Balaji[1], Lakshminarayan M. Iyer[1], M. Madan Babu[2] and L. Aravind[1]

[1] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA
[2] MRC Laboratory of Molecular Biology, Cambridge CB2 2QH, UK

We compared the transcription regulatory interactions inferred from three high-throughput methods. Because these methods use different principles, they have few interactions in common, suggesting they capture distinct facets of the transcription regulatory program. We show that these methods uncover disparate biological phenomena: long-range interactions between telomeres and transcription factors, downstream effects of interference with ribosome biogenesis and a protein-aggregation response. Through a detailed analysis of the latter, we predict components of the system responding to protein-aggregation stress.

Reconstruction of transcriptional regulatory networks

Deciphering the complete transcriptional regulatory program of organisms is an important goal in molecular biology. Identification of the spatial and temporal regulatory interactions between transcription factors (TFs) and their target genes is an important step toward this goal (Box 1; Figure 1a). For this purpose, different high-throughput methods (see Figure S1), are currently used to infer transcription regulatory interactions in various organisms. Although these methods aim to identify regulatory interactions, they are based on different principles. Hence, it is not clear whether they capture the same or distinct facets, such as combinatorial regulation and backups, of the underlying regulatory program. Although numerous studies [1–7] have generated genome-scale transcriptional information, the results from the different studies have not been systematically compared. Therefore, we assembled and compared the genome-scale transcription regulatory networks (TRNs) for yeast, based on datasets from three high-throughput techniques: chromatin immunoprecipitation-chip (ChIP-chip), targeted gene disruption and overexpression of TFs (see Table S1 in the Online Supplementary Material). Although there was a significant overlap in TFs between the three reconstructed TRNs (Figure 1b), the number of common regulatory interactions shared by them was <1%. Furthermore, the extent of overlap of inferred regulatory interactions even between pairs of reconstructed TRNs was <5% (Figure 1b), suggesting that the high-throughput methods reveal different aspects of the actual regulatory process (Figure 1b). The level of agreement in regulatory interactions between the reconstructed TRNs did not change even when we restricted the analysis to the TFs shared between the TRNs. Likewise, we did not observe a significant increase in the overlap of interactions when we reconstructed TRNs using different *P* value thresholds (Figure S2; Table S2). This prompted us to further investigate the nature and significance of regulatory interactions in the three distinct TRNs: $TRN_{CC,}$ TRN_{GRD} and TRN_{GROE} (i.e. those generated by the three high-throughput methods; see Glossary). In particular, we address the following questions. Are there global and local structural differences among the different TRNs? Are the results of the high-throughput methods influenced by disparate biological phenomena? Do they provide novel biological insights apart from the description of the relevant regulatory programs?

Comparison of the local and global structure of the inferred networks

The three distinct TRNs have several interesting similarities and differences in terms of their global and local structure. At the global level, TFs in the TRN_{CC} and TRN_{GRD} have similar distributions in terms of the number of target genes regulated by a given TF (i.e. out-degree distribution), a trend best approximated by a power-law decay [2] (Figure 1a). This implies the presence of global regulators or hubs (traditionally defined as the top 20% of TFs with the greatest number of target genes) in the two TRNs. Interestingly, the out-degree distribution in the TRN_{GROE} has a more centralized distribution rather than a power-law decay, with a peak of 60–120 target genes (TGs) per TF (Figure 1a). This is suggestive of a downstream homeostatic process, such as increased RNA or

Glossary

TRG_{GROE}: The transcriptional network reconstructed from analysis of gene expression on overexpression of the relevant transcription factors (TFs). Nodes represent TFs or target genes (TGs). A TF is linked to a target gene if it is differentially expressed on overexpression of the TF.

TRN_{CC}: The transcriptional network reconstructed from large-scale chromatin immunoprecipitation-chip (ChIP-chip) experiments. Nodes represent TFs or TGs and edges represent direct binding of the TF in the promoter region of the TG.

TRN_{GRD}: The transcriptional network reconstructed from analysis of gene expression on deletion of the relevant TFs. Nodes represent TFs or target genes. A TF is linked to a target gene if it is differentially expressed on deletion of the TF.

Corresponding authors: Balaji, S. (sbalaji@ncbi.nlm.nih.gov); Babu, M.M. (madanm@mrc-lmb.cam.ac.uk).

Box 1. Reconstruction of transcription regulatory networks and high-throughput methods

Although the monumental task of reconstructing regulatory programs for whole organisms is far from complete, recent advances in high-throughput experimental techniques, together with conceptual and representational advances, have brought us closer to this objective. Independent experimental approaches enable the genome-scale reconstruction of the transcription regulatory program of an organism either by directly inferring *in vivo* binding to regulatory sequences or indirectly by identifying the set of genes which are differentially expressed on overexpression or deletion of the transcription factor (Figure 1a). This regulatory program is best represented as the transcriptional regulatory network (TRN) [15–17], where nodes represent transcription factors (TFs) or target genes, and edges represent inferred regulation of a target gene by a TF. As a result, the first assemblage of the TRN for both eukaryotic (*Saccharomyces cerevisiae*) and prokaryotic (*Escherichia coli*) model organisms have become available [1,3,4,6,7,18]. For instance, the high-throughput chromatin immunoprecipitation-chip (ChIP-chip) has helped in genome-scale reconstruction of the yeast TRN by identifying direct binding events for several TFs (TRN_{CC}) [3,4]. Similarly, large-scale gene expression analyses involving yeast strains with either deletions or overexpression of individual TFs have generated independent reconstructions of the yeast TRN [6,7]: TRN_{GRD} (for genetic reconstruction via deletion) and TRN_{GROE} (for genetic reconstruction through overexpression).

The three methods represent major technological landmarks; nevertheless, they have unique pros and cons in terms of experimental design. For example, it is not possible to directly establish the functional relevance of particular DNA-binding events detected in ChIP-chip experiments. The discrimination of direct regulatory interactions from indirect interactions or feedback mechanisms in genetic methods is also nontrivial (see Figure S1). Some of the technical issues concerning the design of these different experimental approaches have been given in Figure S1, but here we describe only the comparison of the reconstructed TRNs from these experiments. As a cautionary note, we state that it is not possible to completely discriminate noise (interactions with no biological relevance) from true regulatory interactions in the TRN reconstructions with the available information. Hence, there could still be some 'noise' in the TRN reconstructions used here.

protein decay, which channels the effects of overexpression of several functionally distinct TFs via a relatively constant number of responding TGs. Furthermore, the larger average number of inferred target genes per TF in the TRN_{GROE} compared with those in the TRN_{GRD} indicates the propagation of indirect downstream effects caused by TF overexpression.

We next identified TFs that are hubs in the TRN_{CC} and TRN_{GRD} and analyzed the extent of overlap in their inferred regulatory interactions. We found that only seven hubs (Abf1p, Ume6p, Aft1p, Swi4p, Cin5p, Cbf1p and Hsf1p), constituting less than one quarter of the total number of hubs, are shared between the networks (Figure S3a). Repeating this procedure using different thresholds to define hubs consistently revealed only a few shared hubs between TRN_{CC} and TRN_{GRD} (see Table S3). TRN_{CC} and TRN_{GRD} overlap to a larger extent in terms of number of regulatory interactions (normalized by respective network size) when the yeast TRNs were reconstructed from data accumulated from case-by-case biochemical studies [1] or from another comprehensive genetic study [8] (Figures S3b and S4). Hence, TRN_{GRD} might be underrepresenting condition-specific transcriptional responses, because all assays were conducted under standard conditions. Comparing the subnetwork for the ubiquitin conjugation system of TRN_{CC} with that of TRN_{GRD} showed that the differences mentioned above were also present at the level of this specific functional subsystem (see Supplementary Material).

We also discovered differences in the distribution of network motifs between the TRNs via a comprehensive search for different motif types (see Supplementary Material and Tables S4 and S5). Multiple-input motifs (MIMs) are most prominent in the TRN_{CC}, suggesting that these in part represent independent back-ups for regulatory interactions, which possibly contribute to the combinatorial robustness of the network [9]. Furthermore, the relative abundance of MIMs and FFMs (feed-forward motifs) in TRN_{GRD} and TRN_{GROE}, respectively (Table S1 and Supplementary Material) implies that (i) expression changes in response to TF deletions are less likely than TF overexpression to alter the expression levels of other TFs and (ii) overexpression of TFs probably tends to affect expression levels of other TFs both directly and indirectly, inducing further gene expression changes. The existence of transcription regulatory events that manifest only under certain conditions, such as stress response and cell cycle, could also account for some of the major qualitative differences in the motifs found in the three reconstructed TRNs (see Table S6).

Interference with translation, the telomere effect and a response to protein aggregation influence the different TRNs

We examined the TF hubs in TRN_{GRD} and found that ~40% of the regulatory interactions in this network were caused by the top four of the five major hubs (i.e. Gcr1p, Cst6p, Sfp1p and Mcm1p). None of these four was identified as a hub in TRN_{CC}. These hubs, with the exception of Mcm1p, have regulatory interactions with numerous target genes (~100) encoding ribosomal components (Figure S5a; Table S7). Furthermore, Sfp1p is a well-characterized major regulator of genes involved in ribosomal biogenesis [10,11]. Most ribosomal target genes (~86%; $P < 0.01$) are inferred to be upregulated on deletion of these TFs, implying that the TFs function as direct or indirect transcriptional repressors of ribosomal TGs. Consequently, these TF deletions might alter the stoichiometry of ribosomal components and thereby affect translation. Thus, the major hubs in the TRN_{GRD} seem to have acquired this status predominantly as a result of indirect translational defects. Genetic manipulation of translation has previously been shown to interfere with a large number of unconnected processes, including subsequent transcription [12].

We previously noted that a telomere-related effect acts as an influential factor in TRN_{CC}. TGs in the subtelomeric regions were inferred to have an unusually large number of binding events (i.e. incoming connections >13) with functionally diverse TFs. We proposed that this might result from TF–telomere interactions being captured in the ChIP-chip experiments owing to either the telomeres looping back and interacting with chromatin complexes on internal chromosomal sites or because of the interaction of chromosome ends with diverse TFs assembled at the inner nuclear envelope [13]. We tested this interpretation by comparing the number of incoming regulatory interactions of target

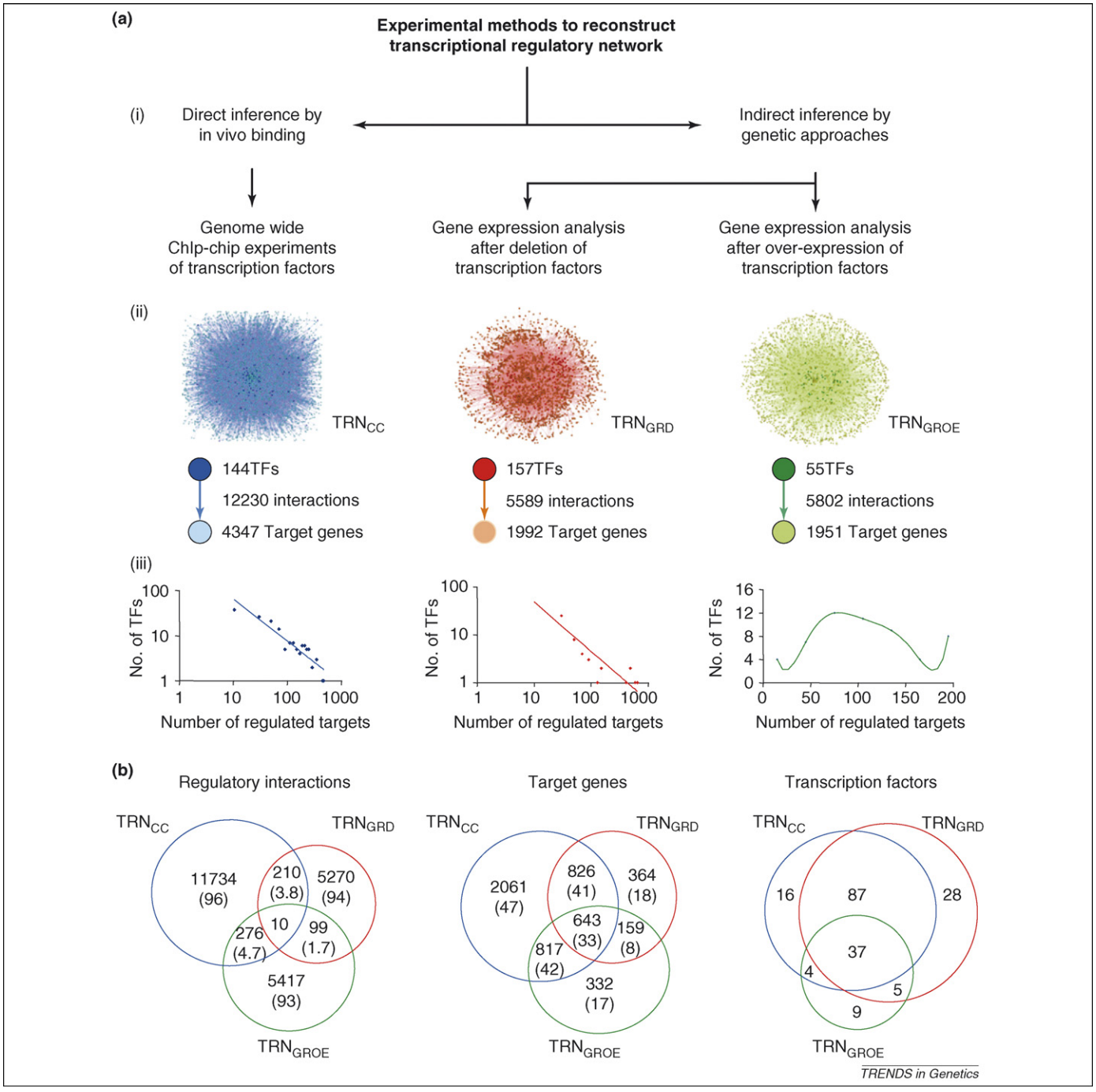

Figure 1. Comparison of transcription regulatory networks (TRNs) reconstructed based on data from *in vivo* binding and genetic studies: (**a**) (i) Experimental methods and a description of the corresponding high-throughput datasets used in this study. (ii) In TRN_{GRD} and TRG_{GROE}, nodes represent transcription factors (TFs) or target genes and edges represent differential expression of a target gene on deletion or overexpression of a particular TF. (iii) The graphs below represent the distribution of the number of TFs and the number of target genes (out-degree distribution) in the TRN_{CC}, TRN_{GRD} and TRN_{GROE}. The out-degree distributions in the TRN_{CC} and TRN_{GRD} are approximated by a power-law equation, suggesting the existence of scale-free structure in these networks. However, TRN_{GROE} shows a centralized out-degree distribution. (**b**) Venn diagrams showing the extent of overlaps in the regulatory interactions, target genes (TGs) and transcription factors between the TRN_{CC}, TRN_{GRD} and TRN_{GROE}. Respective percentage values are shown in parentheses. Although the number of shared transcription factors and target genes are high, the number of common regulatory interactions between the three TRNs is strikingly low. There are 200 predicted TFs and >6000 TGs in yeast. As the number of TFs and TGs in TRN_{CC} and TRN_{GRD} are comparable to the total number of the predicted TFs, the results are likely to be representative on the whole genome scale.

genes associated with the subtelomeric regions in the three TRNs. In most cases, the TGs in the subtelomeric regions show a much greater normalized in-degree (number of distinct TFs regulating a target gene) in the TRN_{CC} than in TRN_{GRD} or TRN_{GROE} (Figure S5b). This further supports the proposal that the ChIP-chip studies captured genuine, potentially long-range, interactions between telomeres and TFs.

A systematic search for high in-degree TGs in TRN_{GRD} or TRN_{GROE} (see Supplementary Material and Figure S6a) identified 42 and 56 such genes, respectively. We classified them into functional categories based on sequence analysis and evidence from the literature (Figure S6b and Supplementary Material). We found that 16 of 56 TGs with high in-degree in the TRN_{GROE} were related to a stress response pertaining to protein unfolding and oxidative damage

($P < 0.02$; Figure S7 and Supplementary Material). In particular, we found that three paralogous genes of the DJ-1/ThiJ/PfpI superfamily, [Hsp31, Hsp33 and Hsp34 (Sno4)] have high in-degrees, suggesting that their expression is affected by overexpression of several unrelated TFs. Disruption of DJ1, the human ortholog of these proteins, was implicated with a protein aggregation defect in Parkinson's disease [14]. Hence, we suggest that overexpression of several TFs and subsequent overproduction of certain proteins causes an increase of aggregated misfolded polypeptides, in turn triggering a specific stress-response pathway. We conjecture that many of the other high in-degree TGs in the TRN_{GROE} are likely to be functionally associated with such a stress response. We also predict that products of these TGs, which include other chaperones, such as Hsp26, Hsp42 and Hsp12, along with the nitrosative stress response protein Yhb1p (all of which have statistically significant high in-degree in the TRN_{GROE}), are likely to cooperate with the DJ-1/ThiJ/PfpI superfamily proteins in a protein-aggregation stress response system. Hence, the analysis of overexpression of TFs could be used as a model to uncover the program that underlies protein misfolding and/or aggregation responses in different cellular systems (see Supplementary Material for further details).

Concluding remarks

We identified additional effects captured by the high-throughput methods, highlighting for the need of *post facto* analysis to discriminate functionally relevant regulatory interactions from such effects. The major secondary effects in the three networks, the telomere effect (in TRN_{CC}), the ribosomal gene effect (in TRN_{GRD}) and the role of the protein misfolding and/or aggregation response (in TRN_{GROE}), provide leads, some of which were previously unsuspected, to understand disparate biological processes. These observations along with low number of shared interactions between the three networks point to the existence of distinct features in the transcription regulatory program, such as combinatorial regulation and backups. Hence, we envisage that a careful combination of the TRNs reconstructed from more-complete versions of such datasets might enable us to decouple genuine combinatorial regulation from regulatory back-up and provide an estimate of the robustness in the regulatory program. It is therefore important that the results of future high-throughput experiments that aim to reconstruct regulatory networks are analyzed with awareness of these secondary effects. Our identification of these effects facilitates two distinct directions of study: (i) a deeper understanding of specific biological phenomena, such as protein aggregation response or the telomere effect, and (ii) improved experimental designs to subtract the additional effects and obtain more accurate network reconstructions.

Acknowledgements

S.B., L.M.I., and L.A. are funded by the Intramural research program of National Institutes of Health, USA. M.M.B. is funded by the Medical Research Council UK, Darwin College and Schlumberger. We thank Arthur Wuster, colleagues at the Laboratory of Molecular Biology, the editor and the anonymous referees for helpful feedback on previous versions of this manuscript.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.tig.2008.04.006.

References

1 Svetlov, V.V. and Cooper, T.G. (1995) Review: compilation and characteristics of dedicated transcription factors in *Saccharomyces cerevisiae*. *Yeast* 11, 1439–1484
2 Guelzim, N. *et al.* (2002) Topological and causal structure of the yeast transcriptional regulatory network. *Nat. Genet.* 31, 60–63
3 Harbison, C.T. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature* 431, 99–104
4 Lee, T.I. *et al.* (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* 298, 799–804
5 Horak, C.E. *et al.* (2002) Complex transcriptional circuitry at the G1/S transition in Saccharomyces cerevisiae. *Genes Dev.* 16, 3017–3033
6 Hu, Z. *et al.* (2007) Genetic reconstruction of a functional transcriptional regulatory network. *Nat. Genet.* 39, 683–687
7 Chua, G. *et al.* (2006) Identifying transcription factor functions and targets by phenotypic activation. *Proc. Natl. Acad. Sci. U. S. A.* 103, 12045–12050
8 Hughes, T.R. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell* 102, 109–126
9 Balaji, S. *et al.* (2006) Uncovering a hidden distributed architecture behind scale-free transcriptional regulatory networks. *J. Mol. Biol.* 360, 204–212
10 Marion, R.M. *et al.* (2004) Sfp1 is a stress- and nutrient-sensitive regulator of ribosomal protein gene expression. *Proc. Natl. Acad. Sci. U. S. A.* 101, 14315–14322
11 Jorgensen, P. *et al.* (2004) A dynamic transcriptional network communicates growth potential to ribosome synthesis and critical cell size. *Genes Dev.* 18, 2491–2505
12 Komili, S. *et al.* (2007) Functional specificity among ribosomal proteins regulates gene expression. *Cell* 131, 557–571
13 Babu, M.M. *et al.* (2006) Estimating the prevalence and regulatory potential of the telomere looping effect in yeast transcription regulation. *Cell Cycle* 5, 2354–2363
14 Wilson, M.A. *et al.* (2003) The 1.1-A resolution crystal structure of DJ-1, the protein mutated in autosomal recessive early onset Parkinson's disease. *Proc. Natl. Acad. Sci. U. S. A.* 100, 9256–9261
15 Babu, M.M. *et al.* (2004) Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol.* 14, 283–291
16 Barabasi, A.L. and Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5, 101–113
17 Schlitt, T. and Brazma, A. (2006) Modelling in molecular biology: describing transcription regulatory networks at different scales. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 361, 483–494
18 Huerta, A.M. *et al.* (1998) RegulonDB: a database on transcriptional regulation in *Escherichia coli*. *Nucleic Acids Res.* 26, 55–59

doi:10.1016/j.tig.2008.04.006 Available online xxxxxx