Minireview

Genome-wide analysis of protein-DNA interactions in living cells

B Franklin Pugh and David S Gilmour

Address: Center for Gene Regulation, Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA 16802, USA.

Correspondence: David S Gilmour. E-mail: dsg11@psu.edu

Published: 4 April 2001

Genome ***Biology*** 2001, **2(4)**:reviews1013.1–1013.3

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2001/2/4/reviews/1013

Abstract

Understanding the regulation of gene expression requires an analysis of gene-specific transcription factors. This review highlights recent work that uses protein-DNA crosslinking, immunoprecipitation and DNA microarrays to determine the binding sites for specific transcription factors throughout the yeast genome.

Cells routinely alter their transcriptional program in response to a changing internal and external environment. These responses are mediated by the binding of transcription factors to specific sequences within the promoter of each responding gene. A major aim of studies of gene regulation is to ascertain which transcription factors control which genes. Simply identifying consensus binding sites by computerized searching of the genome is insufficient, because many transcription-factor binding sequences will occur at random in genomic sequences with some frequency. These fortuitous sites occur within both intergenic and intragenic regions; typically, the intragenic sites are not bound by their cognate factor, and are not functional.

Changes in gene expression profiles can be simultaneously monitored for every gene of an organism by hybridizing cDNAs to DNA microarrays (see Figure 1a) [1,2]. Unfortunately, such gene-expression profiling does not distinguish between direct effects of a transcription factor binding to target genes and indirect effects resulting from one transcription factor inducing the expression of a second. In an effort to measure the binding of transcription factors to their cognate sites, directly and on a genome-wide scale in the yeast *Saccharomyces cerevisiae*, two recent papers [3,4] describe the coupled use of the chromatin immunoprecipitation (ChIP) assay and DNA microarrays.

The ChIP assay involves the use of formaldehyde to covalently crosslink proteins to DNA *in vivo* (see Figure 1b) [5-8]; formaldehyde reacts with the lysine and arginine side chains of proteins and the purine and pyrimidine moieties of DNA. Antibodies against target proteins are used to purify the crosslinked DNA once it has been sheared into small fragments. After amplification of the enriched DNA fragments by PCR and labeling them with the green fluorescent dye Cy5, their identity is revealed by hybridization to DNA probes arrayed at specific locations on a glass slide. Each probe on such a microarray corresponds to a PCR-amplified intergenic region of the yeast chromosome. The study by Iyer *et al.* [4] also included intragenic (or open reading frame, ORF) probes.

Because of the nonuniform deposition of probe DNA during microarray fabrication (among other factors), more reliable results are achieved when a (red) Cy3-labeled reference sample is also included in each microarray hybridization. Using the two samples together provides a two-color readout, where the ratio of changes of the 'test' sample relative to the 'reference' sample is determined (expressed as '-fold', after local background subtraction). The most appropriate reference material to use is a matter of debate. Unenriched total genomic DNA was used in both of the recent studies [3,4], and this is expected to provide a constant reference level. The study by Iyer *et al.* [4] included additional references, such as DNA immunoprecipitated using an antibody to the Swi4 DNA-binding protein from a *swi4*-deletion strain. In practice, these additional references serve to control for the unavoidable nonuniform enrichment of DNA

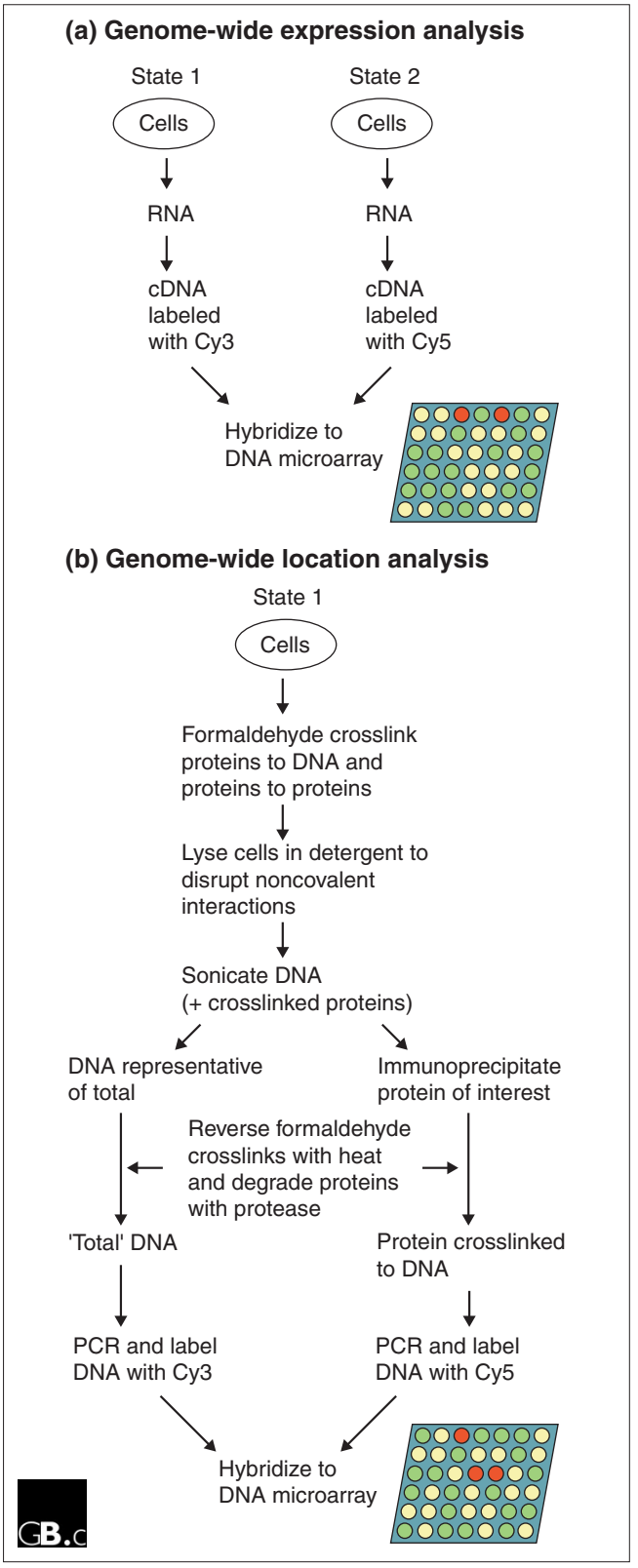

Figure 1
Summaries of the methods for genome-wide analysis of **(a)** gene expression and **(b)** transcription factor binding sites. See the text for further details.

that nonspecifically affects the immunoprecipitations. In an ideal immunoprecipitation with no genomic DNA contamination, such controls would not be appropriate, because the denominator in the two-color ratio scheme would be zero.

The genome-wide analysis by Ren *et al.* [3] examined binding of the galactose-utilization transcription factor Gal4 in the presence and absence of galactose, and binding of the mating-pathway transcription factor Ste12 in response to the mating pheromone α factor. Ten out of approximately 6,400 yeast intergenic regions appeared to be bound by Gal4; 29 were bound by Ste12. In the study by Iyer *et al.* [4], 163 regions were bound by Swi4, a subunit of the SBF transcription factor, and 87 by the MBF transcription factor. Both SBF and MBF have been implicated in control of the cell cycle, so it is not surprising that Iyer *et al.* found that half of the regions bound by MBF were also bound by SBF.

Interestingly, genome-wide analysis reveals that SBF and MBF appear to control a number of non-cell-cycle-regulated genes involved in cell-wall biogenesis and DNA metabolism, respectively, so they might also function in distinct pathways separate from the cell cycle. Given that cell-wall biogenesis and DNA replication occur simultaneously under one state (mitotic growth) and separately under others (pseudohyphal or invasive growth and meiotic S phase), it is reasonable to expect that these pathways are regulated by separate transcription factors that function coordinately during the cell cycle.

The intergenic regions identified as bound by a transcription factor using the ChIP assay are likely to be only the strongest binding regions, so they tell only part of the story. The avidity of binding of a transcription factor from the strongest site to the weakest is a continuum. Statistical analysis is essential for determining the confidence level (p value) associated with each binding. For example, a binding event that has a p value of 0.001 indicates only a 0.1% chance of this level of binding being due to random data fluctuation. Even at this high level of confidence, for 6,400 intergenic regions it is expected that approximately six of the binding events will be 'false positives' (results obtained by chance). In situations like that of Gal4, for which the number of detected binding events is similar to the number expected by chance at this p value, it is critical to incorporate gene-expression information into the analysis, to help filter out false positives. For example, Ren *et al.* [3] established a cut-off of three-fold or greater difference between the test and reference samples to indicate 'real' binding (with $p < 0.001$), and a cut-off of two-fold or greater for gene expression. By these criteria, ten genes were determined to be regulated by Gal4. Seven of these genes were known from other studies to be regulated by Gal4, and all ten appear to play a role in galactose utilization. Lowering the binding ratio cut-off to 2.5 (still with $p < 0.001$) identified another 23 genes, all of which showed less than a 1.8-fold increase in gene expression.

Because the biology of these additional genes did not suggest roles in galactose utilization, a gene's biological function, if known, could serve as an additional subjective criterion to be used in establishing appropriate cut-off values for binding and expression data.

The relationship between the results obtained in these arrays and the behavior of transcription factors *in vivo* is not a simple one. For example, the use of gene-expression profiles is important when ChIP analysis detects binding of a factor to a region that lacks any discernible consensus binding site for the factor. In the study by Iyer *et al.* [4], approximately half of the Swi4-bound regions lacked a consensus site for binding SBF. Although some of these are likely to be false positives, it is possible that SBF can in fact recognize additional sequences, or that promoter specificity at the consensus site is achieved through interactions with other promoter-bound transcription factors. Also, binding of transcription factors to promoter sites does not necessarily result in transcriptional activation. Genome-wide analysis confirmed previous findings that Gal4 is associated with the promoters of the genes *gal1* and *gal10* under glucose-repressed conditions; but for these promoters, displacement of the Gal80 repressor is necessary to achieve activation by Gal4. Finally, although SBF appears to be bound to many intergenic sites, it is perplexing that deletion of its Swi4 subunit had little effect on the expression of putative SBF target genes [4]. It is possible that additional or redundant transcriptional programs direct the expression of these genes, possibly throughout the cell cycle. If so, the current analysis using asynchronous cells would reveal only a composite expression profile, and studies of synchronized cell populations will be required to resolve this issue.

Genome-wide location analysis offers the promise of being able to identify the complete set of genomic regions to which transcription factors are bound *in vivo*. When coupled with gene-expression profiling and searches for consensus binding sites, it has the potential to identify the direct effectors of complex gene expression programs. Application of these techniques to additional transcription factors as cells respond to changing internal and external environments should lead to a broader understanding of the physical regulatory networks governing cellular behavior.

References

1. DeRisi JL, Iyer VR, Brown PO: **Exploring the metabolic and genetic control of gene expression on a genomic scale.** *Science* 1997, **278:**680-686.
2. Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA: **Dissecting the regulatory circuitry of a eukaryotic genome.** *Cell* 1998, **95:**717-728.
3. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, *et al.*: **Genome-wide location and function of DNA binding proteins.** *Science* 2000, **290:**2306-2309.
4. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO: **Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF.** *Nature* 2001, **409:**533-538.
5. Solomon MJ, Larsen PL, Varshavsky A: **Mapping protein-DNA interactions *in vivo* with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene.** *Cell* 1988, **53:**937-947.
6. Dedon PC, Soults JA, Allis CD, Gorovsky MA: **Formaldehyde cross-linking and immunoprecipitation demonstrate developmental changes in H1 association with transcriptionally active genes.** *Mol Cell Biol* 1991, **11:**1729-1733.
7. Orlando V, Paro R: **Mapping Polycomb-repressed domains in the bithorax complex using *in vivo* formaldehyde cross-linked chromatin.** *Cell* 1993, **75:**1187-1198.
8. Braunstein M, Rose AB, Holmes SG, Allis CD, Broach JR: **Transcriptional silencing in yeast is associated with reduced nucleosome acetylation.** *Genes Dev* 1993, **7:**592-604.