Cell

Resource

Large-Scale Genetic Perturbations Reveal Regulatory Networks and an Abundance of Gene-Specific Repressors

Patrick Kemmeren,[1,3] Katrin Sameith,[1,3] Loes A.L. van de Pasch,[1,3] Joris J. Benschop,[1,3] Tineke L. Lenstra,[1,3] Thanasis Margaritis,[1,3] Eoghan O'Duibhir,[1] Eva Apweiler,[1] Sake van Wageningen,[1] Cheuk W. Ko,[1] Sebastiaan van Heesch,[1] Mehdi M. Kashani,[1] Giannis Ampatziadis-Michailidis,[1] Mariel O. Brok,[1] Nathalie A.C.H. Brabers,[1] Anthony J. Miles,[1] Diane Bouwmeester,[1] Sander R. van Hooff,[1] Harm van Bakel,[1] Erik Sluiters,[1] Linda V. Bakker,[1] Berend Snel,[2] Philip Lijnzaad,[1] Dik van Leenen,[1] Marian J.A. Groot Koerkamp,[1] and Frank C.P. Holstege[1,*]

[1]Molecular Cancer Research, University Medical Centre Utrecht, Universiteitsweg 100, Utrecht 3584 CG, the Netherlands

[2]Theoretical Biology and Bioinformatics, Department of Biology, Utrecht University, Utrecht 3584 CG, the Netherlands

[3]Co-first author

*Correspondence: f.c.p.holstege@umcutrecht.nl

http://dx.doi.org/10.1016/j.cell.2014.02.054

SUMMARY

To understand regulatory systems, it would be useful to uniformly determine how different components contribute to the expression of all other genes. We therefore monitored mRNA expression genome-wide, for individual deletions of one-quarter of yeast genes, focusing on (putative) regulators. The resulting genetic perturbation signatures reflect many different properties. These include the architecture of protein complexes and pathways, identification of expression changes compatible with viability, and the varying responsiveness to genetic perturbation. The data are assembled into a genetic perturbation network that shows different connectivities for different classes of regulators. Four feed-forward loop (FFL) types are overrepresented, including incoherent type 2 FFLs that likely represent feedback. Systematic transcription factor classification shows a surprisingly high abundance of gene-specific repressors, suggesting that yeast chromatin is not as generally restrictive to transcription as is often assumed. The data set is useful for studying individual genes and for discovering properties of an entire regulatory system.

INTRODUCTION

Cells depend on many intricate molecular interactions to successfully perform a myriad of functions in an integrative manner. One of the current challenges of molecular biology is to determine and study all interactions important for cellular function (Ideker et al., 2001). This is inspired by increased awareness that complex properties can emerge from combinations of relatively few simple interactions. Systematic interaction analyses are being realized through high-throughput approaches and are required to understand many aspects of living organisms, including disease (Vidal et al., 2011). Whereas some interactions are physically direct, e.g., protein-protein interactions (Walhout and Vidal, 2001), others can be more abstract, e.g., genetic interactions (Costanzo et al., 2010). Both are informative, either for the function of individual components or for properties of the entire system. Various data sets, generated to different degrees of accuracy and completion, have successfully been applied to study cellular systems. One such system is mRNA expression. To study the regulatory network underlying mRNA expression, it would be useful to determine how different cellular components influence mRNA expression genome-wide.

It is well established that perturbation of individual factors, followed by genome-wide expression analysis, can yield insight into function (DeRisi et al., 1997; Holstege et al., 1998). Regulatory pathways (Roberts et al., 2000) and protein complexes (van de Peppel et al., 2005) can be similarly studied, additionally revealing functional relationships between components. Focusing on functionally uncharacterized genes, a pioneering study of 276 mutants in the yeast *Saccharomyces cerevisiae* first demonstrated the utility of much larger collections of genetic perturbation expression signatures (Hughes et al., 2000). This has been followed by studies of many factors individually, as well as of entire classes of regulators (Hu et al., 2007; van Wageningen et al., 2010; Lenstra et al., 2011) also incorporating other types of perturbation (Chua et al., 2006; Weiner et al., 2012).

Despite many other advances, the number of genetic perturbations analyzed within such studies has not increased significantly since the first compendium (Hughes et al., 2000), likely for logistical reasons. Although many genetic perturbations have been analyzed, analysis of entire systems has been hampered, in particular because of difficulties inherent to properly comparing gene expression data generated across the different conditions, genetic backgrounds, technology platforms, types of controls, and degrees of replication in different studies. Here, we report mRNA expression profiles uniformly generated for deletion of one-quarter of all protein-coding genes

in *S. cerevisiae*. By making particular use of data uniformity and the causal relationships inherent to genetic perturbation, the data are analyzed at different levels of complexity to study fundamental properties of the underlying regulatory system.

RESULTS

mRNA Expression Profiles of 1,484 Deletion Mutants

To systematically investigate the regulatory network of a model organism, expression changes were determined genome-wide for haploid *S. cerevisiae* strains bearing single gene deletions (Giaever et al., 2002). Selection was based on the deleted gene having a (putative) role in regulating gene expression. Selection also included characteristics such as nuclear location or the capacity to modify other proteins. The 1,484 mutants cover many different functional categories, including gene-specific and global transcription factors (TFs), RNA processing and export, ubiquitin(-like) modifications, protein kinases/phosphatases, protein trafficking, cell cycle, meiosis, and DNA replication and repair (Figure S1A and Table S1 available online).

Various strategies were incorporated to ensure a high degree of accuracy and precision (Experimental Procedures). This included four replicates per responsive mutant, robotic procedures optimized with external calibration controls (van Bakel and Holstege, 2004), a common reference design with wild-type (WT) reference RNA applied in dye-swap to each microarray (Figure S1B), as well as dye-bias correction (Margaritis et al., 2009) and spike-in controls to monitor global changes (van de Peppel et al., 2003). Additional WT cultures were processed alongside batches of mutants, with day-specific effects countered by regrowing the entire batch. Statistical modeling results in an average expression profile for each mutant. Each profile consists of p values and average transcript level changes in the mutant relative to 428 WTs. Further controls for consistency, aneuploidy, and correct gene deleted resulted in 101 deletion strains being remade and reprofiled (Table S1). Consistency controls included analysis alongside strains from the same protein complex or pathway, resulting in remaking strains with suspected secondary mutations (Teng et al., 2013). These technical aspects were uniformly applied to the entire data set, some of which has been used previously (Table S1). With coverage of one-quarter of all genes and one-third of all genes not required for viability, this constitutes the largest collection of uniformly generated expression signatures for genetic perturbations.

Response to Genetic Perturbation

The data set consists of approximately 40 million expression measurements including WTs and replicates. Hierarchical clustering is presented in Figure 1. Although low-magnitude fold-changes [FCs] may have biological relevance, a stringent threshold ($FC > 1.7$, $p < 0.05$) was applied throughout the study to ensure a focus on robust changes more likely to be biologically meaningful. This threshold was based on WT variation. When analyzed collectively, the number of transcripts robustly affected in at least one mutant ($FC > 1.7$, $p < 0.05$) starts leveling off at two-thirds (Figure 2A). Transcripts that do not change are highly enriched for dubious open reading frames (ORFs; $p = 2.6 \times 10^{-9}$) and for genes essential for viability ($p = 7.8 \times 10^{-31}$). Most dubious ORFs are lowly or not expressed in WT (Figure 2B). Combined with their low degree of change, this agrees with their classification as dubious, with most not likely to encode functional proteins (Fisk et al., 2006). Essential genes show much higher WT transcript levels (Figure 2B). The low degree of change observed for essential genes (Figure 2B) indicates that larger changes in their expression are too deleterious for survival. Plateauing of transcripts with altered expression (Figure 2A) suggests that most of the robust gene expression changes compatible with viable genetic perturbation have been covered for this growth condition.

As observed before, strains with reduced growth generally have more transcripts affected, and not all genetic perturbations result in transcriptome changes (Hughes et al., 2000). To focus on mutants with stronger changes, signatures were classified as different from WT (responsive) when at least four transcripts show robust changes. Excluded are a set of 58 transcripts with highly variable behavior in WTs (WT variable genes; Experimental Procedures). These criteria ensure that almost all WTs are classified as having no change and indicate that 53% of mutants are similar to WT (nonresponsive). This is concordant with the previous determination of 43% on a smaller set of deletions using different thresholds (Hughes et al., 2000). Redundancy likely contributes to nonresponding deletions. This is demonstrated by a strong enrichment for genes with a close paralog (Figure S1E). Growth condition-dependency likely also contributes. This is indicated by the larger number of genes with low transcript and undetectable protein levels within the group of nonresponder deletions (Figures S1C and S1D). The information that loss of a gene does not strongly affect expression of other genes is useful for several purposes, including modeling regulatory networks (Macneil and Walhout, 2011). Taking essential genes into account (Giaever et al., 2002), the fraction of genes that can be individually removed under a single growth condition with no strong effects on gene expression is 43%.

Protein Complex and Pathway Organization

Functional relationships are revealed by hierarchical clustering of deletion signatures (Figure 1, columns; dendrogram in Data S1). Previous analyses indicate protein complex and pathway membership as the main factors contributing to profile similarity (Hughes et al., 2000; Lenstra et al., 2011). In contrast to coexpression across different conditions, the degree of deletion-profile similarity for different types of interactions has so far not been systematically addressed. We therefore determined signature similarity for all complexes and pathways, including metabolic pathways as well as signaling factors such as protein kinases, ubiquitin(-like) enzymes and their targets. Signature correlation is highest for protein complexes (Figure 2C), in particular for smaller complexes with four or less subunits (examples in Figure 3A). All transcripts that change significantly in any single mutant are depicted in such figures, rather than a subset selected for similar behavior. Highly similar profiles (Figure 3A) indicate disruption of the entire complex upon deletion of any individual subunit. As shown previously for the transcription coregulator Mediator (van de Peppel et al., 2005) and more comprehensively for 30 chromatin complexes (Lenstra et al.,