Visualizing Global Correlation in Large-Scale Molecular Biological Data

A.N.M. Imroz Choudhury, Kristin Potter, Theresa-Marie Rhyne, Yarden Livnat, Chris R. Johnson, Orly Alter

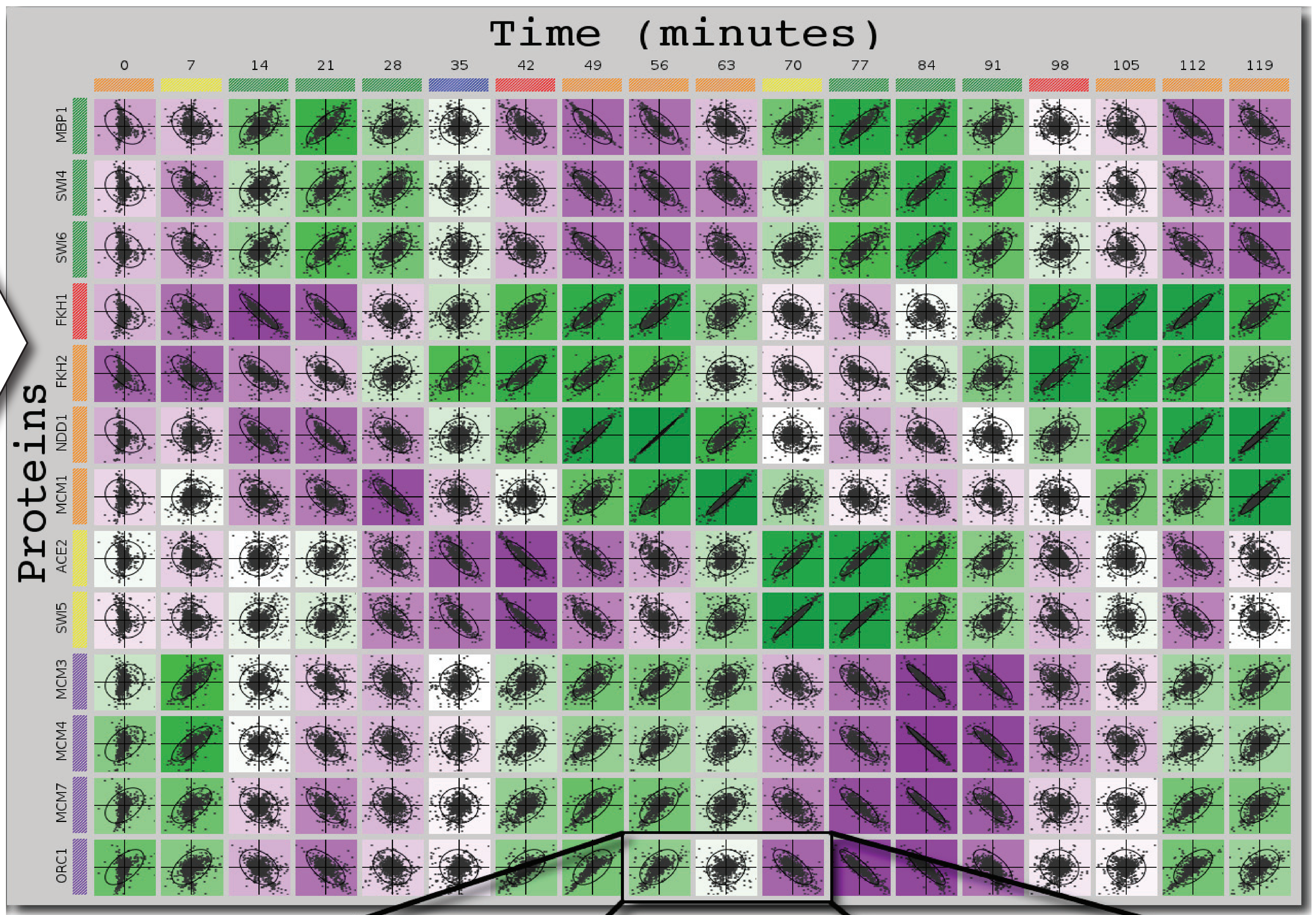
THE UNIVERSITY OF UTAH

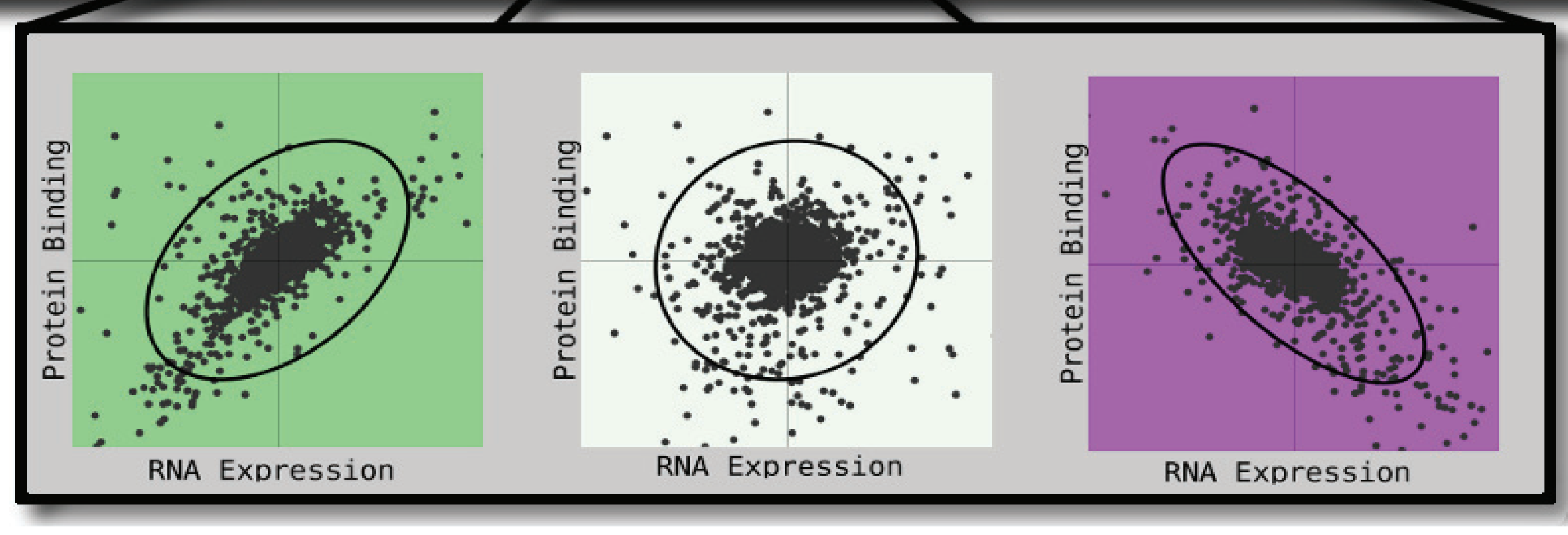
SCI www.sci.utah.edu

Detection of novel patterns of correlation in large-scale molecular biological data can hint at the existence of as yet unknown cellular regulatory mechanisms. For example, correlations were observed among the DNA binding of cell cycle transcription factors [4] and the mRNA expression levels of cell cycle-regulated genes [5]. These correlations correspond to a known causal coordination among these processes. Recent experimental results [3] verify a computationally predicted mechanism of regulation [2] correlating genome-wide binding of replication initiation proteins [6] with mRNA expression during the cell cycle. This has demonstrated for the first time that mathematical modeling of DNA microarray data can be used, beyond classification of genes and cellular samples, to correctly predict previously unknown global modes of regulation [1].

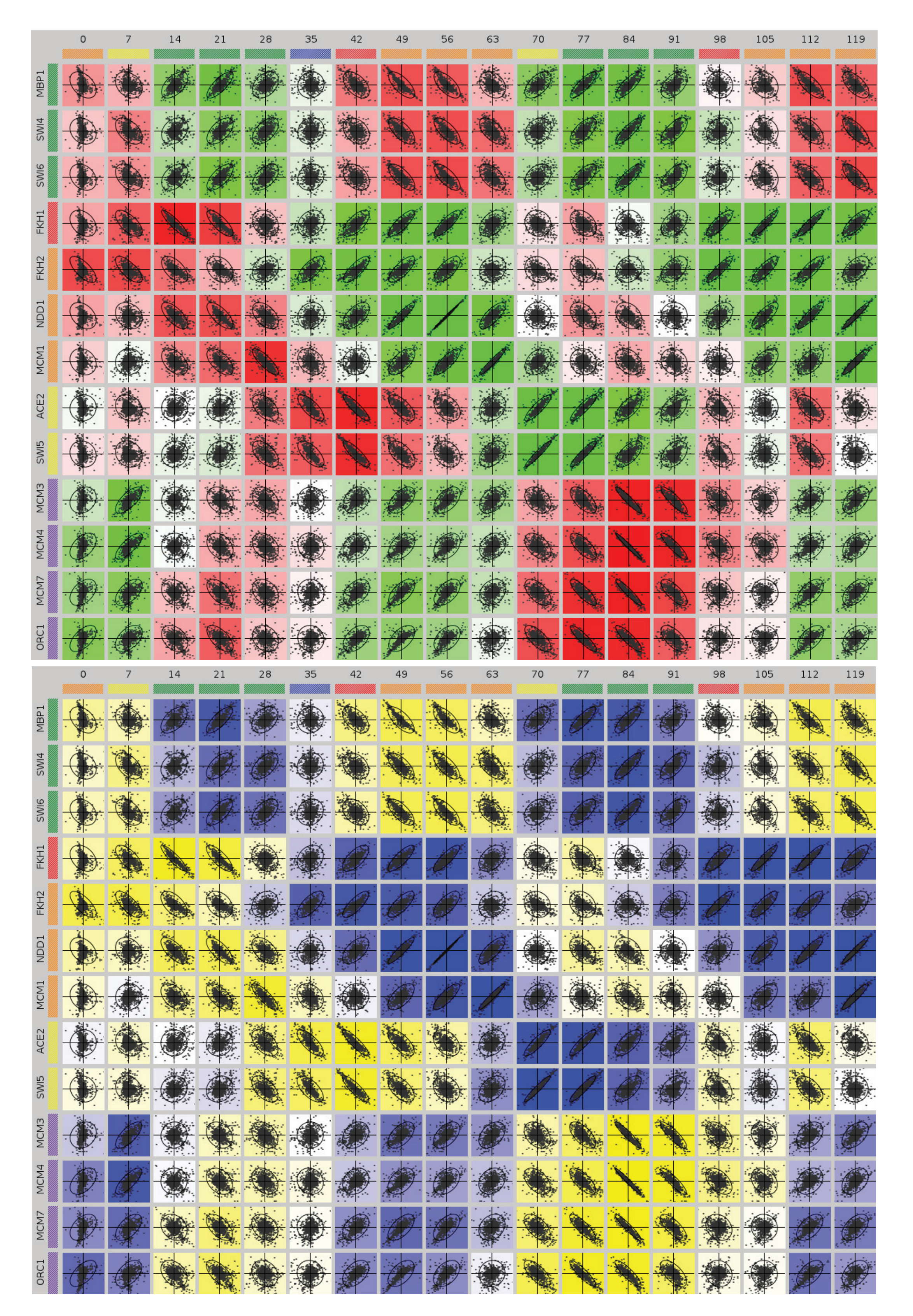
In this work, we propose a visualization approach that facilitates exploration and identification of patterns of correlation in biological data. Our method provides a global view of temporal relationships between biological variables and local views of underlying data. This approach empowers researchers to discover global patterns and possible regulatory mechanisms while supporting visual verification of data quality and maintaining confidence in the visualization.

The global view depicts the temporal aspect of protein activity, tiling several local views, one per protein per time point, into a table. Each row represents a particular protein and each column a time point. The panel of scatter plots shows patterns on a larger scale, such as coordinated activation of proteins in time, highlighted in the background colors, which reflects the correlation of each local view. In this way we can depict correlations for both local and global views within the same visual space.



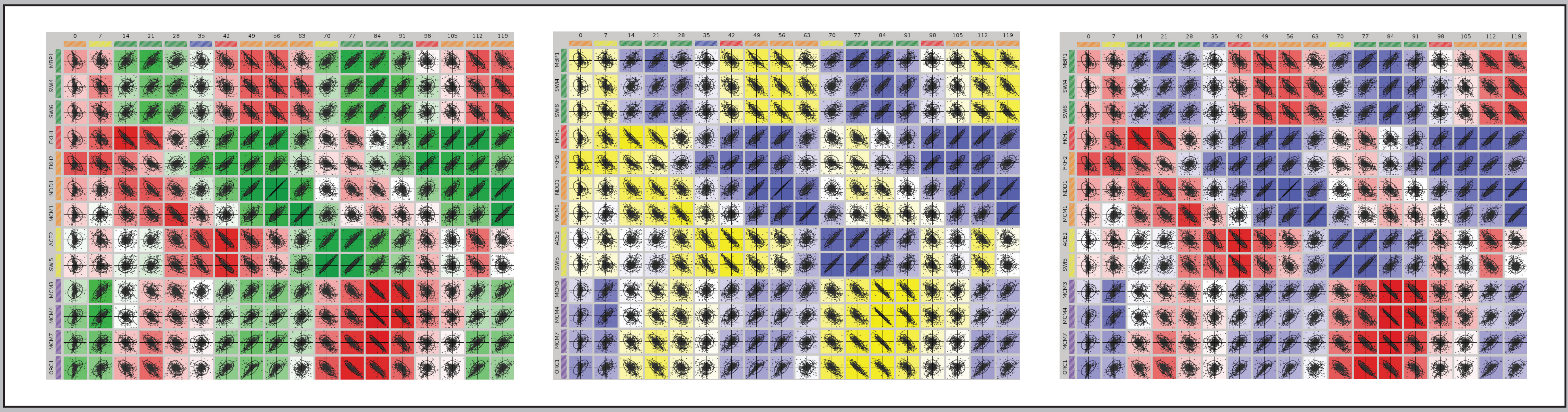


The local view provides a direct view of the data with a concise representation of statistical measures. Each focuses on a single protein's activity at a specific time and consists of a scatter plot where each point represents the level of protein binding adjacent to and gene expression level of a given gene. To show possible (anti-)correlation between the values, we depict the data's principal components using an ellipse. An elongated ellipse indicates a high correlation while its orientation indicates positive or negative correlation. The ellipse provides a concise view of the correlation and the associated uncertainty for the data, allowing researchers to determine their own level of confidence in the data. Left: Color analysis using the Adobe Kuler tool.



Two **traditional color maps** used by biologists, with fully saturated colors showing non-uniform luminance. The red-green color map in particular is challenging for colorblind viewers to examine.

[1] O. Alter. Discovery of principles of nature from mathematical modeling of DNA microarray data. PNAS, 103:16063, 2006.
[2] O. Alter and G. H. Golub. Integrative analysis of genome-scale data by using pseudoinverse projection predicts novel correlation between DNA replication and RNA transcription. PNAS, 101:16577, 2004.
[3] L. Omberg et al. Global effects of DNA replication and DNA replication origin activity on eukaryotic gene expression. Mol. Syst. Biol., 5(312), 2009.
[4] I. Simon et al. Serial regulation of transcriptional regulators in the yeast cell cycle. Cell, 106:697, 2001.
[5] P. T. Spellman et al. Comprehensive identification of cell cycleregulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol. Biol. Cell, 9:3273, 1998.
[6] J. J. Wyrick et al. Genome-wide distribution of ORC and MCM proteins in Saccharomyces cerevisiae: high-resolution mapping of replication origins. Science, 294:2357, 2001.



We have **designed several color maps** to address these concerns. We use muted color tones with complementary hues that retain strong contrast between extremes and are optimized to contrast with our scatter plots.

ACKNOWLEDGEMENTS

This research was supported by grants from the National Human Genome Research Institute R01 Grant HG-004302, NSF CAREER Award DMS-0847173 (to OA) as well as NSF, DOE, Award No. KUS-C1-016-04, made by King Abdullah University of Science and Technology (KAUST), and the NIH/NCRR Center for Integrative Biomedical Computing, 2P41 RR0112553-12 (to CRJ).