Supplementary Information for "PECA: a novel statistical tool for deconvoluting time-dependent gene expression regulation"

Guoshou Teo, Christine Vogel, Debashis Ghosh, Sinae Kim, Hyungwon Choi

November 19, 2013

Simulation scheme

We simulated gene expression data reflecting the burst of up- and down- regulation of mRNAs between the first two time points. These were generated from log-normal distribution with their respective mean parameters in each group as specified in Table 1 and the variance fixed at $\sigma^2 = 0.1$. To simulate protein expression data according to the mechanism of Figure 1, we set the translation and degradation rates (κ^s, κ^d) as tabulated in Table 2, in which the protein synthesis rate changes by a factor of r^*. We fixed κ^d at 1, and thus r^* essentially represents the rate ratio. This leads to the time-dependent mean concentration values following the relationship in Equation (2) in the main text. Using these mean parameters, we simulated protein expression data from log normal distribution, where varying variance parameters τ^2 were set to control the signal-to-noise ratio. Based on Equation (2), the ratio e^τ/r^* can be interpreted as a form of the coefficient of variation (CV), provided that the gene and protein expression data are properly scaled. We have evaluated the performance at different CVs, where we varied r^* from 1.5 to 2.0 and τ^2 from 0.01 to 0.2. In each scenario, we looked at three different probability thresholds $p^* = 0.5, 0.6, 0.7$.

Markov chain Monte Carlo

To estimate the model parameters, we constructed a MCMC sampler that combines standard Metropolis-Hastings updates and dimension switching updates in the form of the reversible-

Group	Size	μ_0	μ_1	μ_2	μ_3	μ_4	μ_5
1	500	1.00	1.25	1.20	1.10	1.00	1.00
2	500	1.00	1.25	1.20	1.10	1.10	1.10
3	500	1.00	0.75	0.80	0.90	0.90	0.90

Table 1: Mean parameters of gene expression data in the three groups.

Group	κ_0^s	κ_1^s	κ_2^s	κ_3^s	κ_4^s
1	1.00	1.00	1.00	1.00	1.00
2	r^*	1.00	1.00	1.00	1.00
3	1.00	1.00	r^{*-1}	r^{*-2}	r^{*-2}

Table 2: Protein-level rate parameters in the protein expression data in the three groups.

jump MCMC (1). The model parameters are updated in the following order:

$$\{\eta_{ji0}\}_{j=0}^{N} \to \tau_i^2 \to \{\kappa'_{it}\}_{t=0}^{T-1} \to \boldsymbol{C}_i$$

for all i. This whole cycle is repeated for 5,000 iterations for burn-in period and $M = 20,000$ iterations for the main iteration with thinning of 20 samples, in both simulation and data analysis sections that follow. We use hat and tilde symbols to denote current and proposal values respectively.

1. We first start with η_{ji0} by a Metropolis-Hastings step, with proposal value $\tilde{\eta}_{ji0}$ drawn from $\mathcal{N}(\hat{\eta}_{ji0}, 0.1^2)$, and compute the Metropolis-Hastings ratio to complete the update. Since this parameter is involved in the mean values at all time points, the likelihood has to be evaluated at all time points for updating each of these parameters.

2. Next, we draw the variance parameter τ_i^2 by Gibbs sampling from inverse gamma distribution $\mathcal{IG}(a_\tau + N(T+1)/2, b_\tau + \sum_{j,t}(y_{jit} - \eta_{jit})^2/2)$.

3. Next, we draw $\{\kappa'_{i\ell}\}$ for $\ell = 0, \ldots, |\boldsymbol{C}_i|$ under the fixed $\boldsymbol{C}_i$ for each protein i. We use random walk Metropolis-Hastings steps to update them, i.e. draw a proposal value $\tilde{\kappa}'_{i\ell}$ from $\mathcal{N}(\hat{\kappa}'_{i\ell}, 0.1^2)$ and accept or reject afterwards.

4. Finally, we update the change point configuration $\boldsymbol{C}_i$. There are two different moves: birth of a new change point and removal (death) of an existing change point. Since these two moves are reversible in notation, we just describe the birth move here. Suppose that $\hat{\kappa}'_{i\ell}$ covers a time period (h_t, h_{t+m}) that contains at least one observation time(s). Then we propose a birth of a new change point $h^* \in \{h_{t+1}, \ldots, h_{t+m-1}\}$ within the interval (chosen from one of the intermediate time points) and break the current rate parameter into two daughter parameters, namely $(\tilde{\kappa}'_{i\ell}, \tilde{\kappa}'_{i,\ell+1})$ where it is required to meet

$$(h^* - h_t) \cdot \text{logit}(\tilde{\kappa}'_{i\ell}) + (h_{t+m} - h^*) \cdot \text{logit}(\tilde{\kappa}'_{i,\ell+1}) = (h_{t+m} - h_t) \cdot \text{logit}(\hat{\kappa}'_{i\ell})$$

with a random perturbation such that

$$\frac{\tilde{\kappa}'_{i,\ell+1}}{1 - \tilde{\kappa}'_{i,\ell+1}} = \frac{1-u}{u} \frac{\tilde{\kappa}'_{i\ell}}{1 - \tilde{\kappa}'_{i\ell}},$$

with $u \sim \text{Uniform}(0,1)$. Under this transformation, the Jacobian is $\frac{(\tilde{\kappa}'_{i\ell}(1-\tilde{\kappa}'_{i\ell}) + \tilde{\kappa}'_{i,\ell+1}(1-\tilde{\kappa}'_{i,\ell+1}))^2}{\hat{\kappa}'_{i\ell}(1-\hat{\kappa}'_{i\ell})}$ for $(\hat{\kappa}'_{i\ell}, u) \to (\tilde{\kappa}'_{i\ell}, \tilde{\kappa}'_{i,\ell+1})$. Hence the Metropolis-Hastings ratio for the birth move just

equals the posterior ratio times the Jacobian since the acceptance probability of this proposal is

$$\min\{1, \text{likelihood ratio} \times \text{prior ratio} \times \text{proposal ratio} \times \text{Jacobian}\},$$

where the prior and proposal ratios are the ratios of Uniform distribution over unit intervals. Then the Metropolis-Hastings ratio becomes

$$\prod_{j,t}\left[\exp\left\{-\frac{1}{2\tau_i^2}(\ln(y_{jit})-\ln(\eta_{jit}))^2\right\}\right]\frac{\varphi}{1-\varphi}\frac{(\tilde{\kappa}'_{i\ell}(1-\tilde{\kappa}'_{i\ell})+\tilde{\kappa}'_{i,\ell+1}(1-\tilde{\kappa}'_{i,\ell+1}))^2}{\hat{\kappa}'_{i\ell}(1-\hat{\kappa}'_{i\ell})}.$$

In both simulation and yeast data analysis, we ran the MCMC for 20,000 iterations with thinning (every 10th sample) after 5,000 iterations for burn-in period. We elicited the same prior distributions used in the simulation studies, and the acceptance rates for the Metropolis-Hastings updates (13%) and reversible jump MCMC (21%) remained reasonably good (before thinning of the chain). We performed visual inspection of model fit by plotting the estimated level of protein expression $\{\eta_{jit}\}$ against the observed values and found that the fit was reasonably good. We also confirmed the convergence of the MCMC sampler to the posterior distribution by the trace plot of the log likelihood.

References

[1] P.J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.

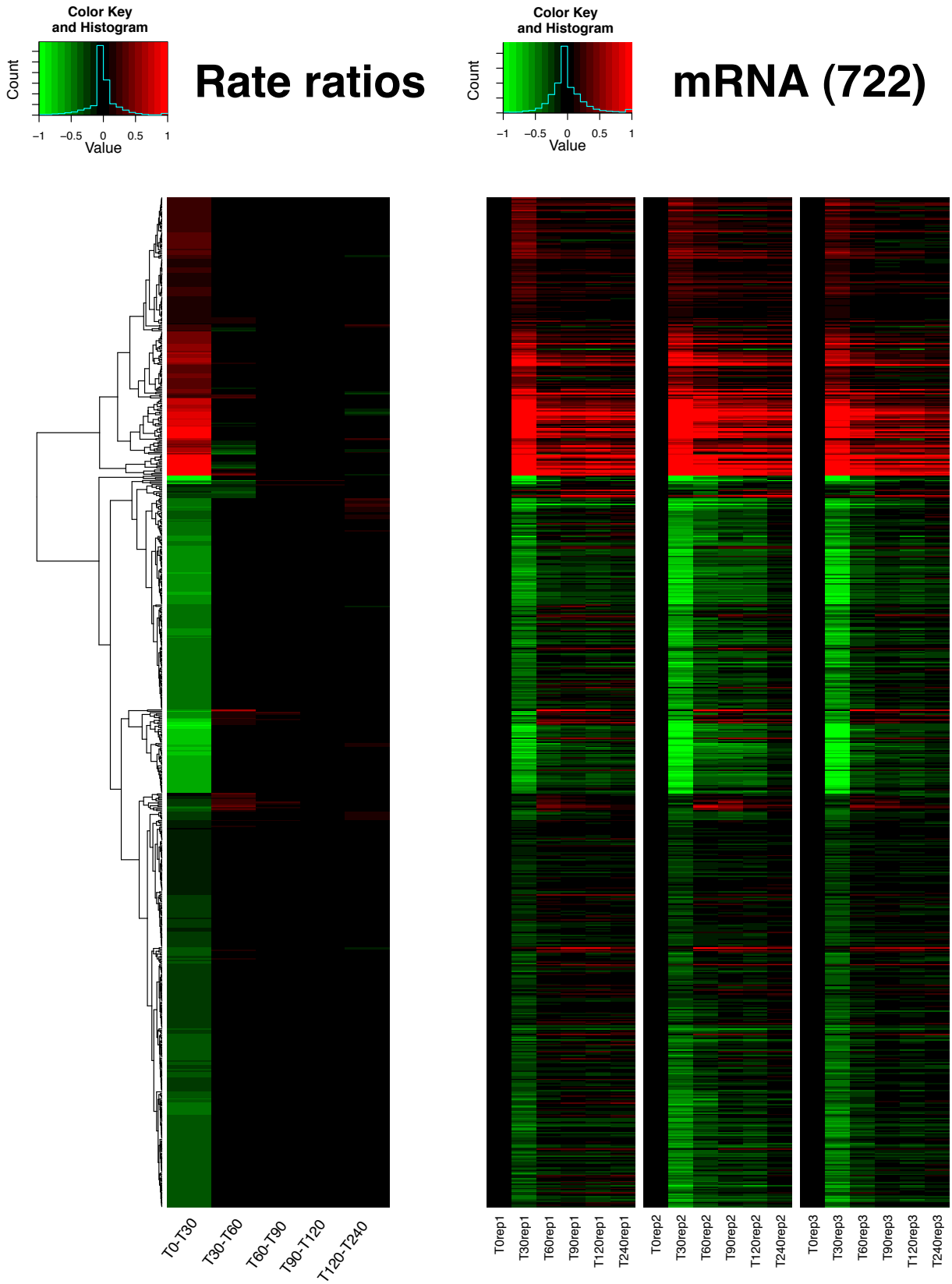

Figure 1: Heatmaps of the 722 stress induced and repressed proteins subject to RNA-level regulation. The left panel is the rate ratios at the RNA-level, estimated for each time interval (between two adjacent time points). The right panel is the mRNA data.

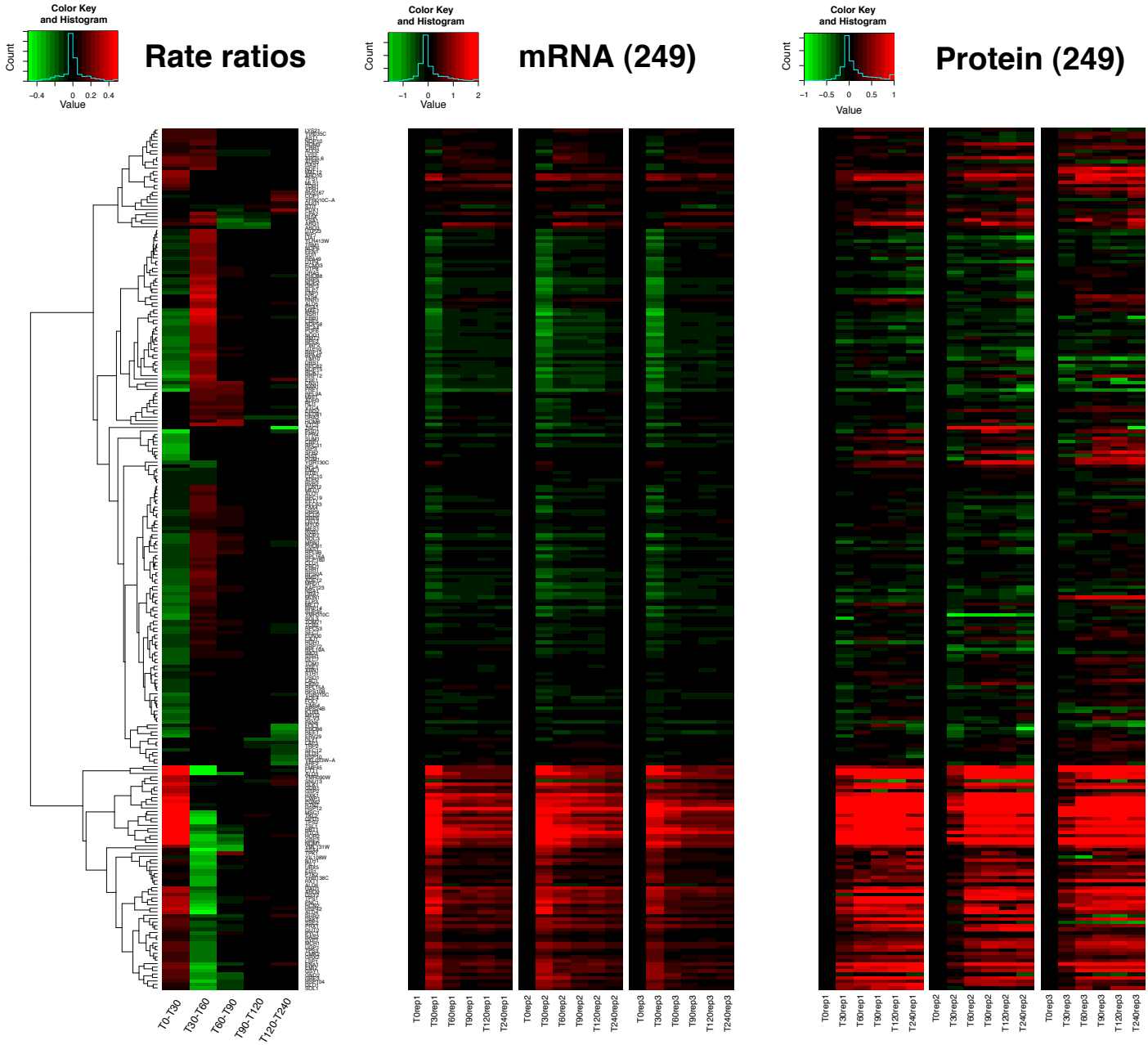

Figure 2: Heatmaps of the 249 stress induced and repressed proteins subject to protein-level regulation. The left panel is the rate ratios at the protein-level, estimated for each time interval (between two adjacent time points). The middle and right panels are the mRNA and protein expression data.

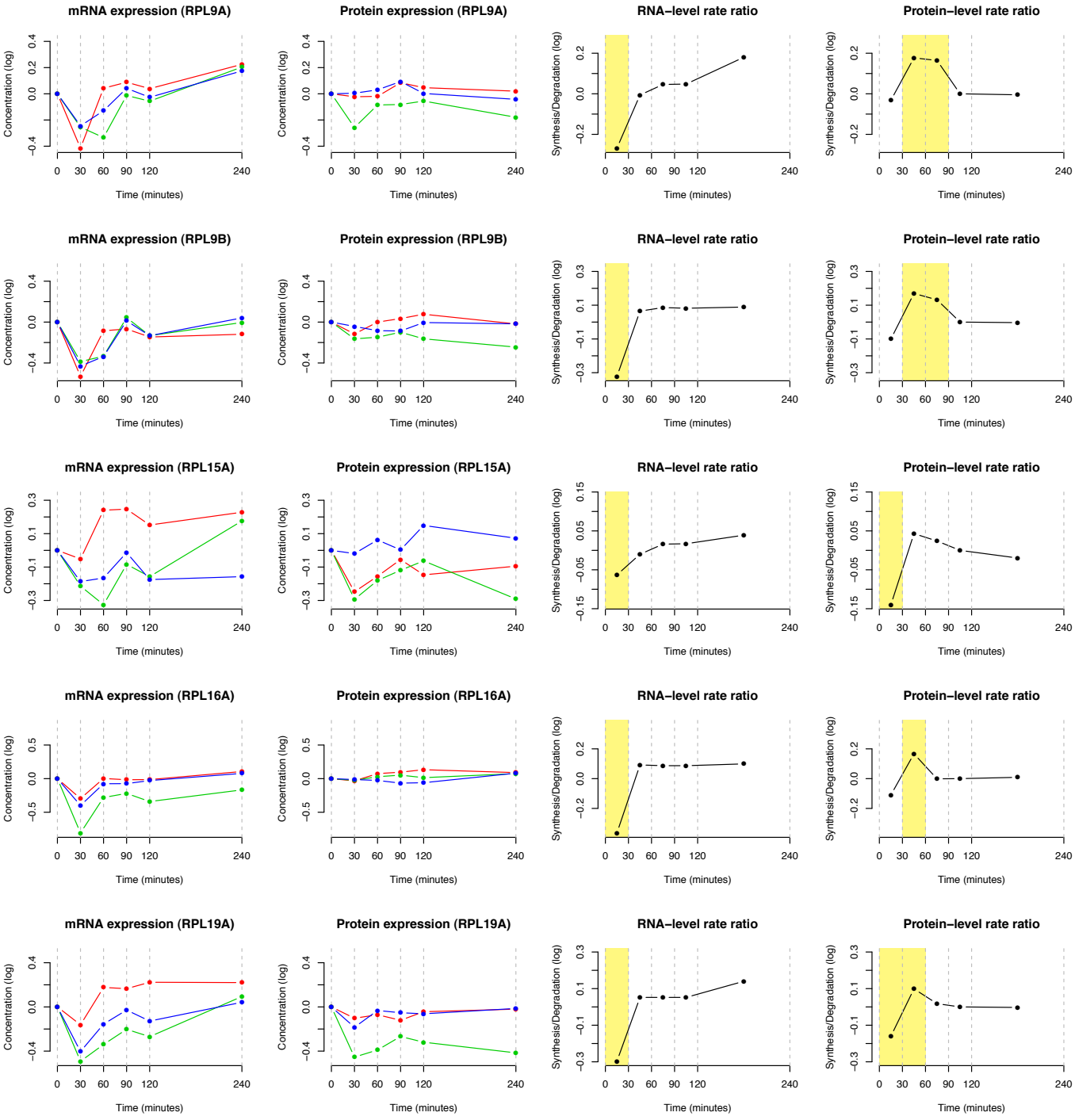

Figure 3: The mRNA and protein expression and estimated rate ratios at both levels of regulation for RPL9A, RPL9B, RPL16A, RPL19A, which are members of the large subunit of ribosome.

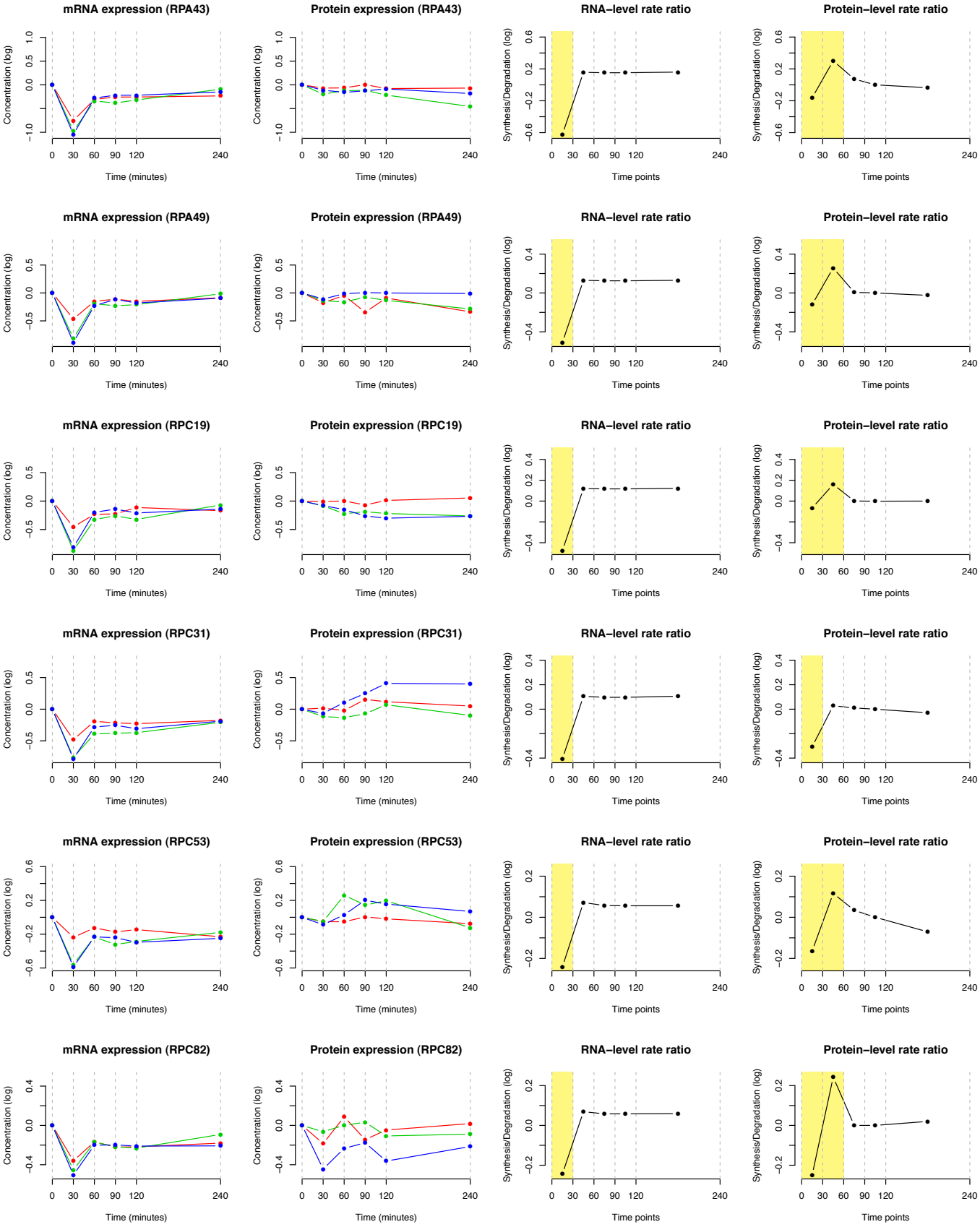

Figure 4: The mRNA and protein expression and estimated rate ratios at both levels of regulation for RPA43, RPA49, RPC19, RPC53, and RPC82, which are subunits of RNA polymerase I and III.

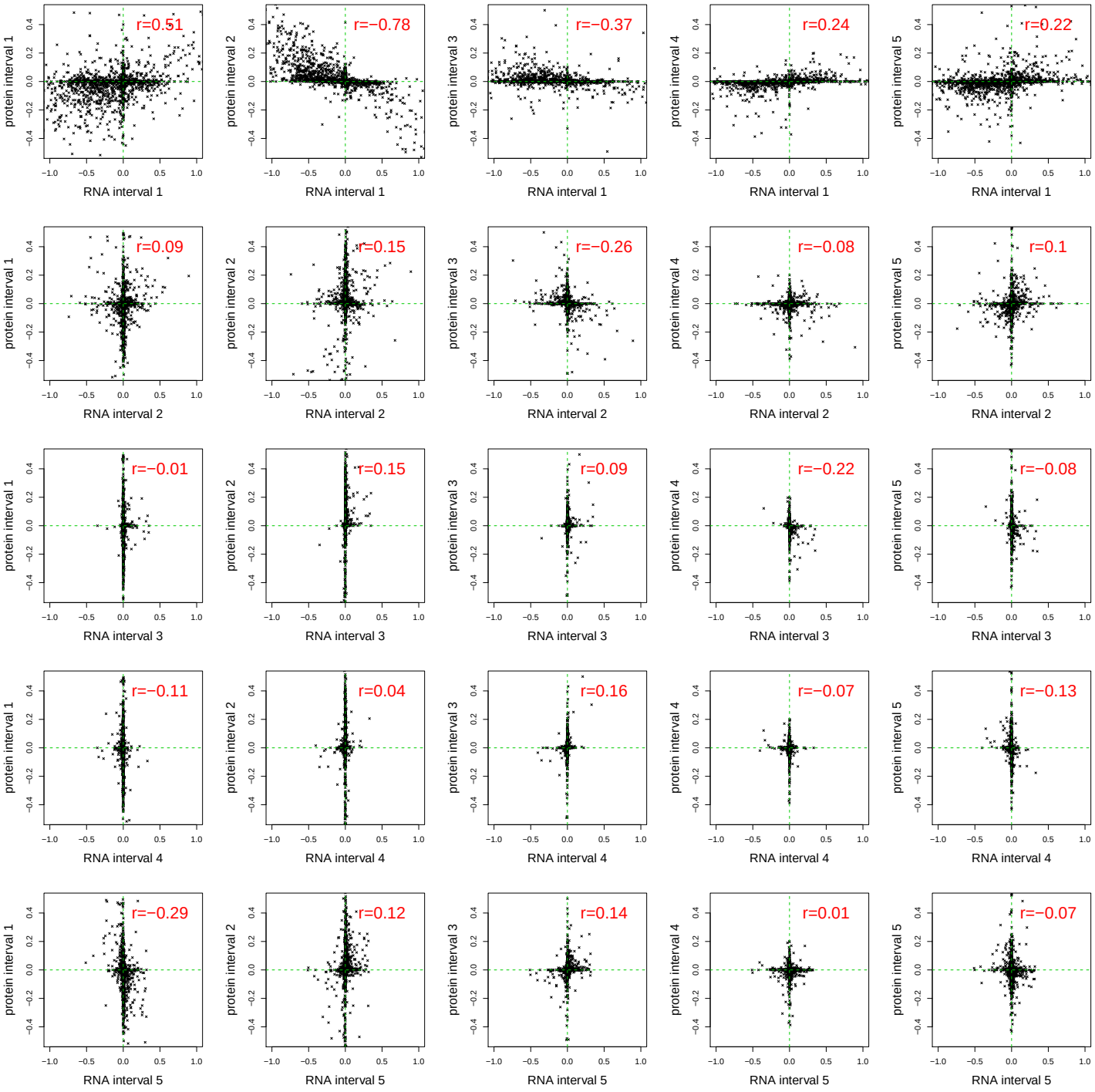

Figure 5: *S. Cerevisiae* data with osmotic stress. The panels were arranged so that each row and column corresponds to each time point respectively. In each panel, the protein rate ratios were plotted against the RNA rate ratios (transformed by log base 2, then centered by median in each protein). The panels on diagonal positions show coupling at the same time point, whereas the panels on off-diagonal positions show buffering at different time points or time-delayed correlation.

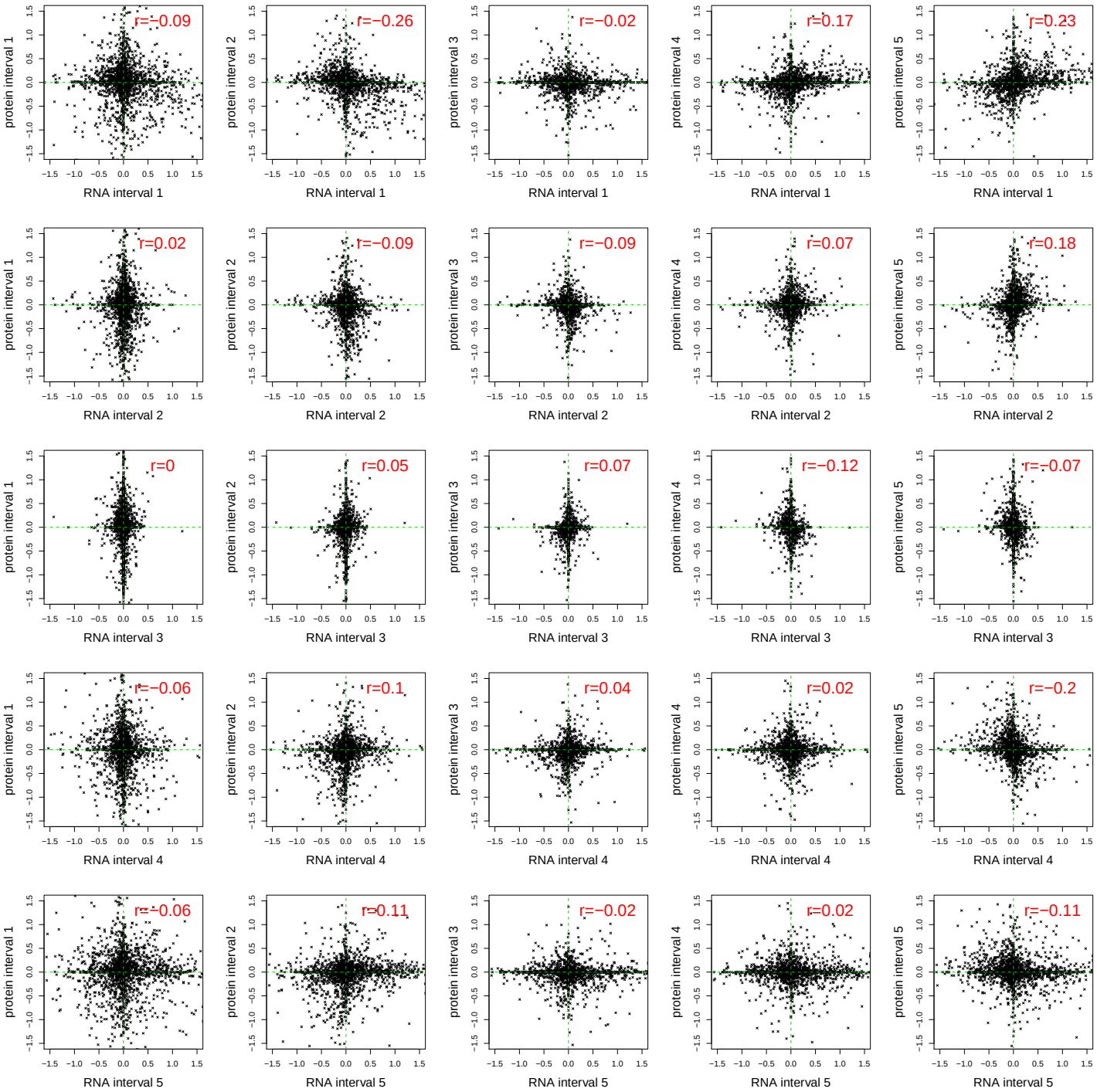

Figure 6: *S. Bombe* data with oxidative stress. The panels are arranged the same way as Supplementary Figure 5.