Consensus clustering for Bayesian mixture models: Supplementary materials

Stephen Coleman, Paul DW Kirk and Chris Wallace

October 7, 2020

Abstract

1 Yeast data

The "Yeast data" consists of three *S. cerevisiae* datasets with gene products associated with a common set of 551 genes. The datasets are:

- microarray profiles of RNA expression from Granovskaia et al. (2010) a cell cycle dataset that comprises measurements taken at 41 time points (the **Timecourse** dataset),
- Chromatin immunoprecipitation followed by microarray hybridization (**ChIP-chip**) data from Harbison et al. (2004), and
- Protein-protein interaction (**PPI**) data from BioGrid (Stark et al., 2006).

The datasets were reduced to 551 items by considering only the genes identified by Granovskaia et al. (2010) as having periodic expression profiles with no missing data in the PPI and ChIP-chip data, following the same steps as the original MDI paper (Kirk et al., 2012). The datasets were modelled using a base measure of a Gaussian process in the Timecourse dataset and Multinomial distributions in the ChIP-chip and PPI datasets.

1.1 Bayesian analysis

10 chains were run for 36 hours, resulting in 676,000 iterations per chain, thinned to every thousandth sample, resulting in 676 samples per chain.

1.1.1 Convergence

These chains were investigated for

- within-chain stationarity using the Geweke convergence diagnostic (Geweke et al., 1991), and

- across-chain convergence using the potential scale reduction factor ($\hat{R}$) and the Vats-Knudson extension (*stable* $\hat{R}$, Vats and Knudson, 2018).

The Geweke convergence diagnostic is a standard Z-score; it compares the sample mean of two sets of samples. It is calculated under the assumption that the two parts of the chain are asymptotically independent and if this assumption holds than the scores are expected to be standard normally distributed presenting evidence for within chain stationarity.

$\hat{R}$ is expected to approach 1.0 if the set of chains are converged. Low $\hat{R}$ is not sufficient in itself to claim chain convergence, but values above 1.1 are clear evidence for a lack of convergence (Gelman et al., 2013). Vats and Knudson (2018) show that this threshold is significantly too high (1.01 being a better choice) and propose extensions to $\hat{R}$ that enable a more formal rule for a threshold, and it is their method as implemented in the R package `stableGR` (Knudson and Vats, 2020) that is the final check of convergence.

In the case of clustering we are interested in stationarity of the continuous variables, in MDI this is the concentration parameter of the Dirichlet distribution for the component weights and the ϕ_{ij} parameter associated with the correlation between the i^{th} and j^{th} datasets.

We plot the Geweke-statistic for each chain in figure 1 and the series of the ϕ parameters alone in figure 2, excluding the most poorly behaved chain (chain 9). Very few of the chains appear to be truly stationary, but some behave far worse than others. Based upon this we exclude chains 1, 2, 4, 6 and 9, restricting the analysis to the 5 better, if not ideally, behaved chains. Further evidence that even these chains are not converged can be see in figure 3, where the values of $\hat{R}$ do not drop below 1.25 for the ϕ parameters. Stable $\hat{R}$ is also too high, with several million more samples recommended before behaviour suggesting convergence is expected.

Investigating the Posterior similarity matrices (PSMs) we can see that the Timecourse data appears to have only the mildest of disagreement between the PSMs from different chains. The lack of convergence between chains emerges in the ChIP-chip data and, to a far greater degree, in the PPI data.

1.2 Consensus clustering analysis

References

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

John Geweke et al. *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*, volume 196. Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN, 1991.

Marina V Granovskaia, Lars J Jensen, Matthew E Ritchie, Joern Toedling, Ye Ning, Peer Bork, Wolfgang Huber, and Lars M Steinmetz. High-resolution

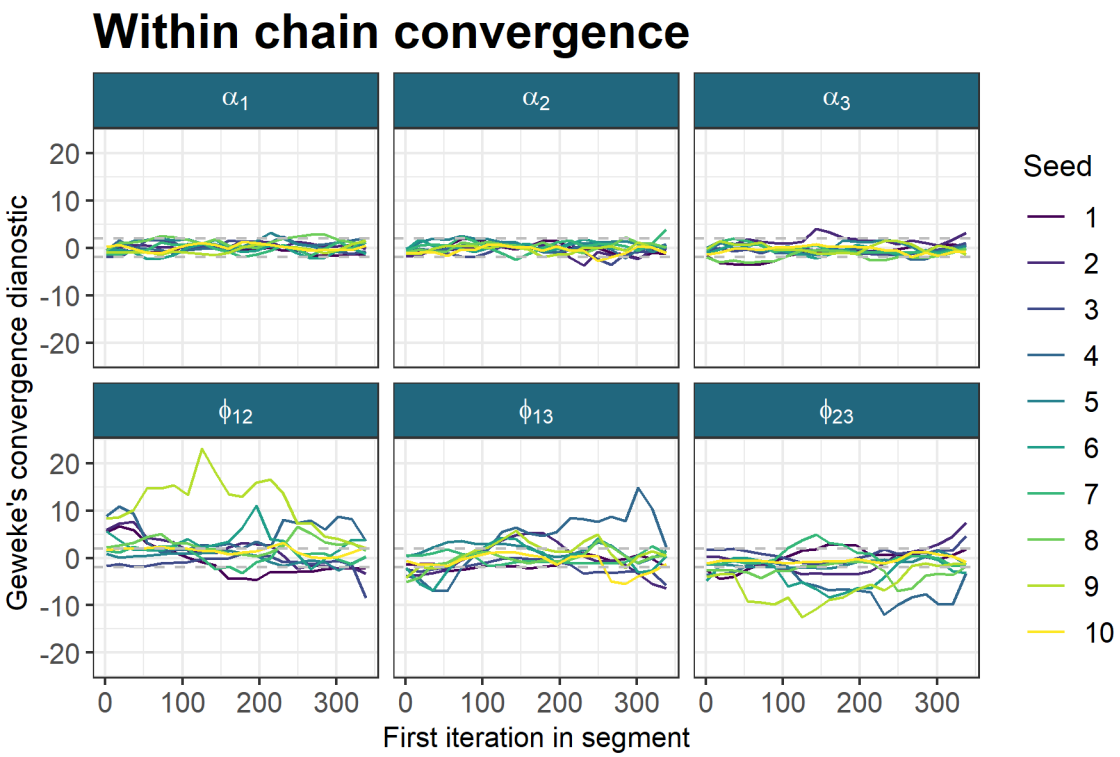

Figure 1: Chain 9 can be seen to have the most extreme behaviour in the distribution of the Geweke diagnostic for ϕ_{12}, ϕ_{13} and ϕ_{23}. We remove this chain from the analysis. We also see that is in these same variables that the chains reveal poor behaviour and focus on these.

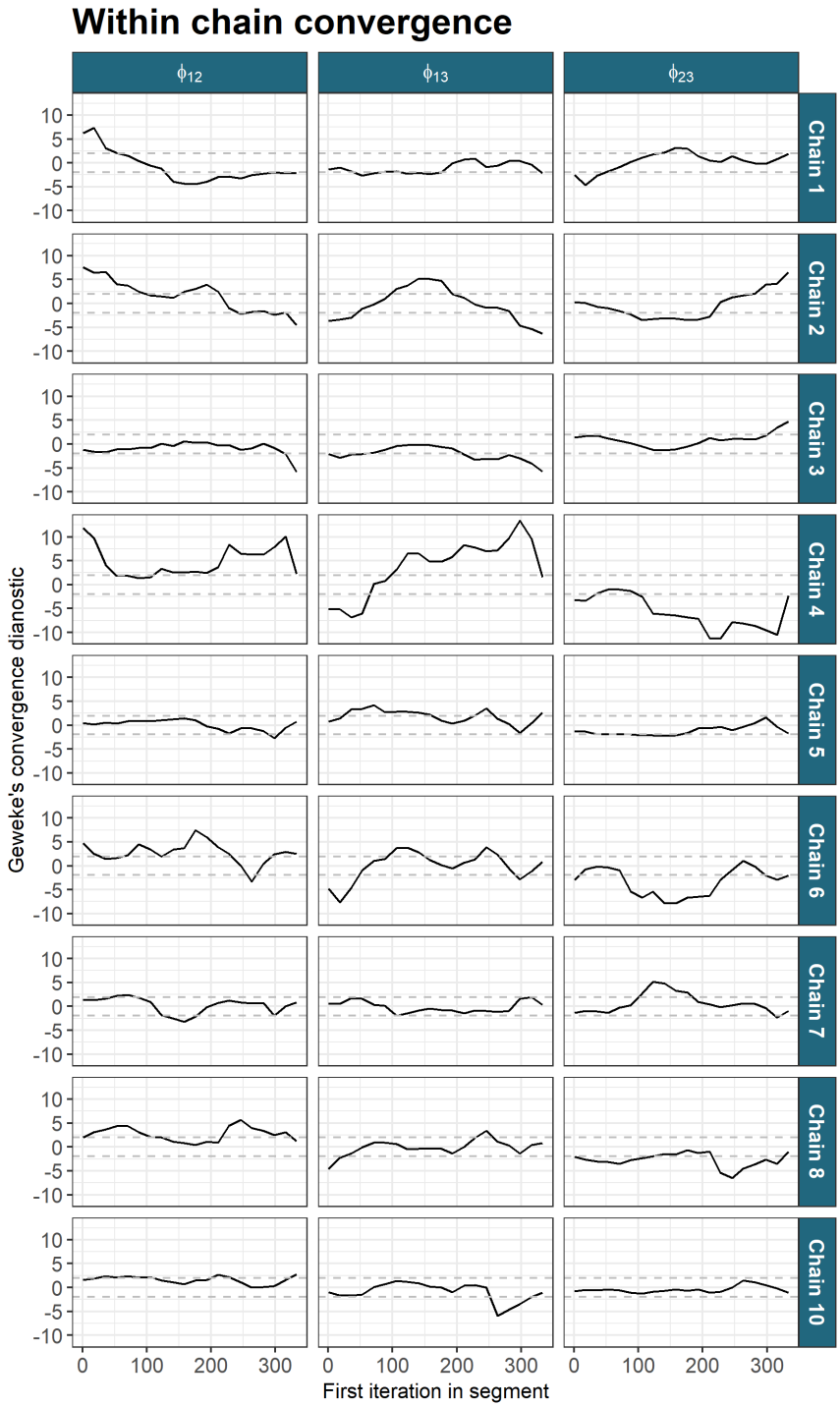

Figure 2: None of the chains appear to be standard normal in their distribution. Chain 4 behaves very strangely and is also dropped from the analysis. Of the remaining chains there is less clear distinctions, but chains 1, 2, and 6 appear most extreme and thus are dropped.

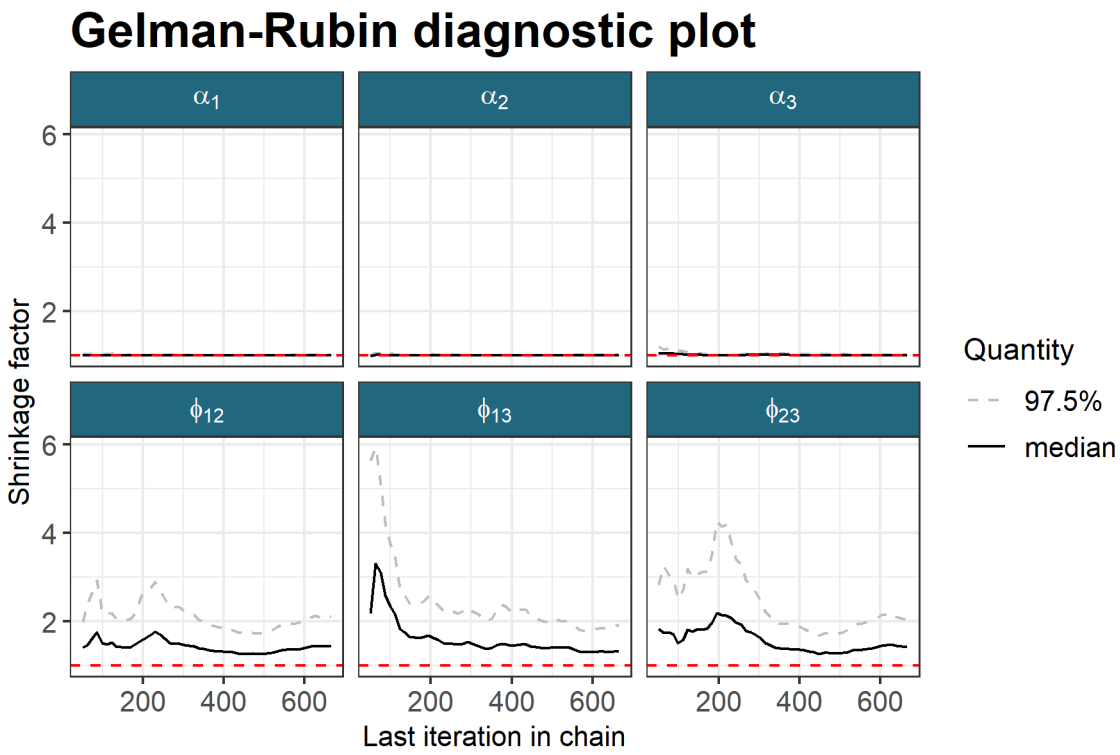

Figure 3: The chains still appear to be unconverged with $\hat{R}$ remaining above 1.25 for the ϕ_{12}, ϕ_{13} and ϕ_{23} parameters. Stable $\hat{R}$ is also too high with values of 1.049, 1.052 and 1.057.

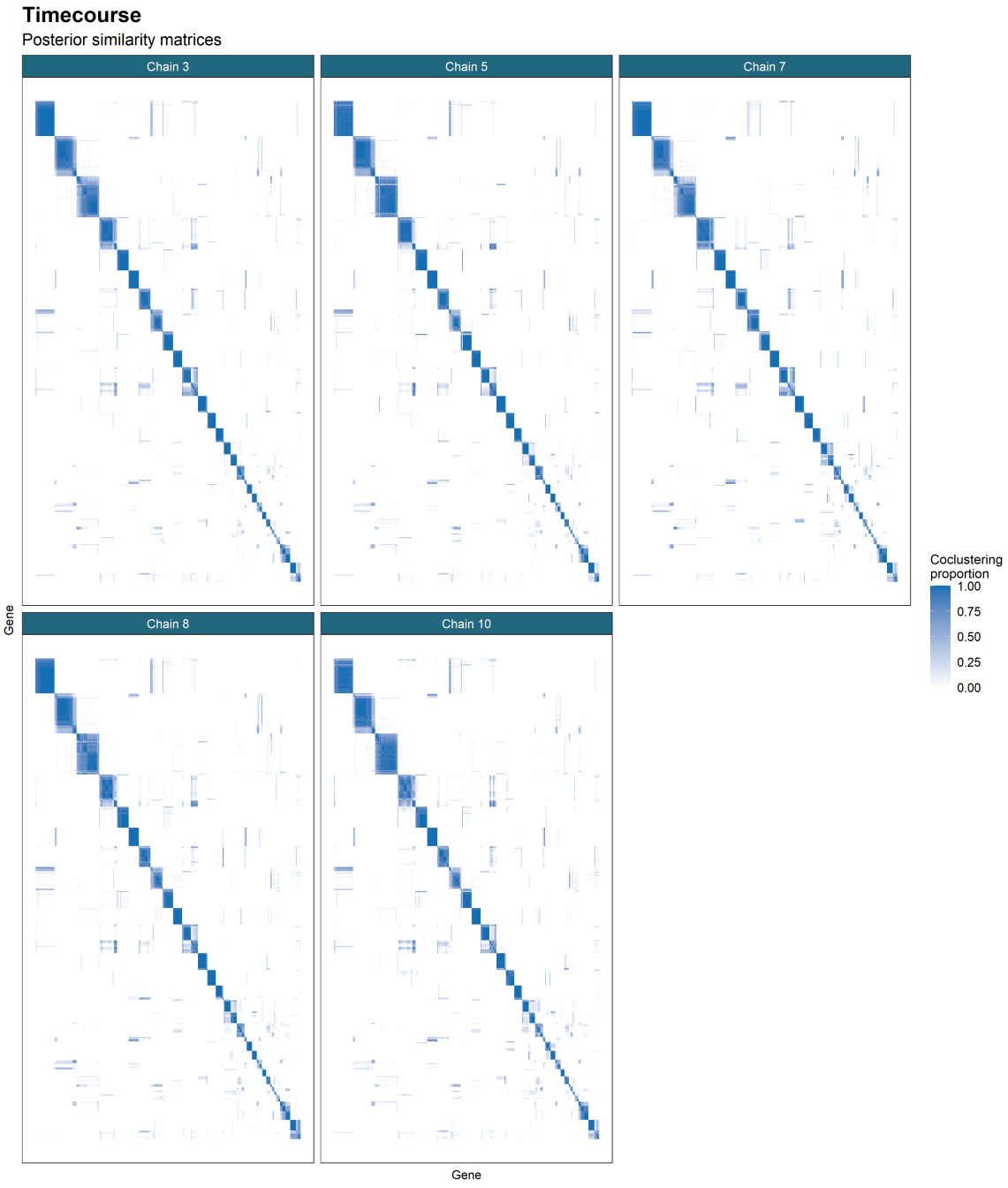

Figure 4: No marked difference.

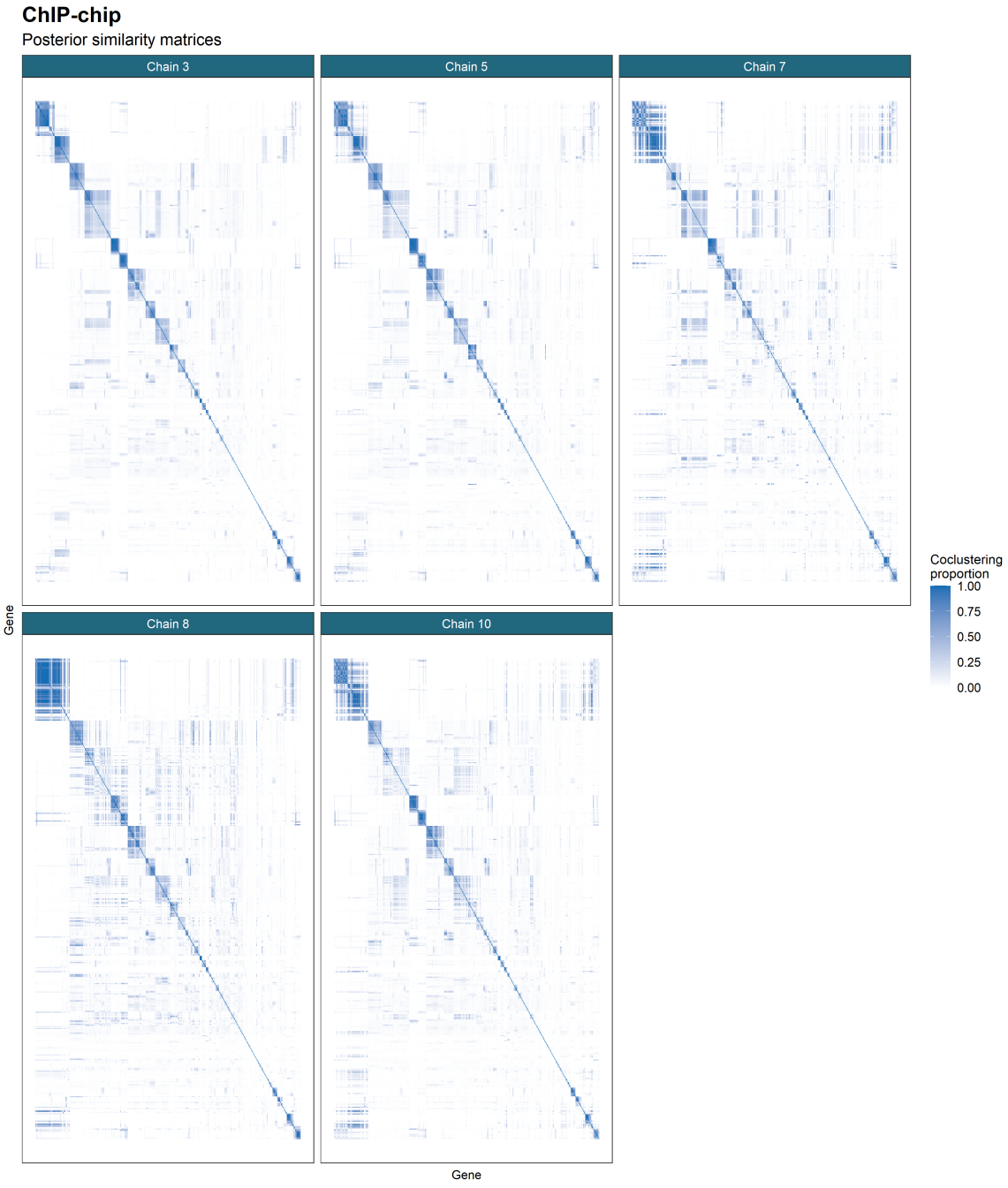

Figure 5: Some difference.

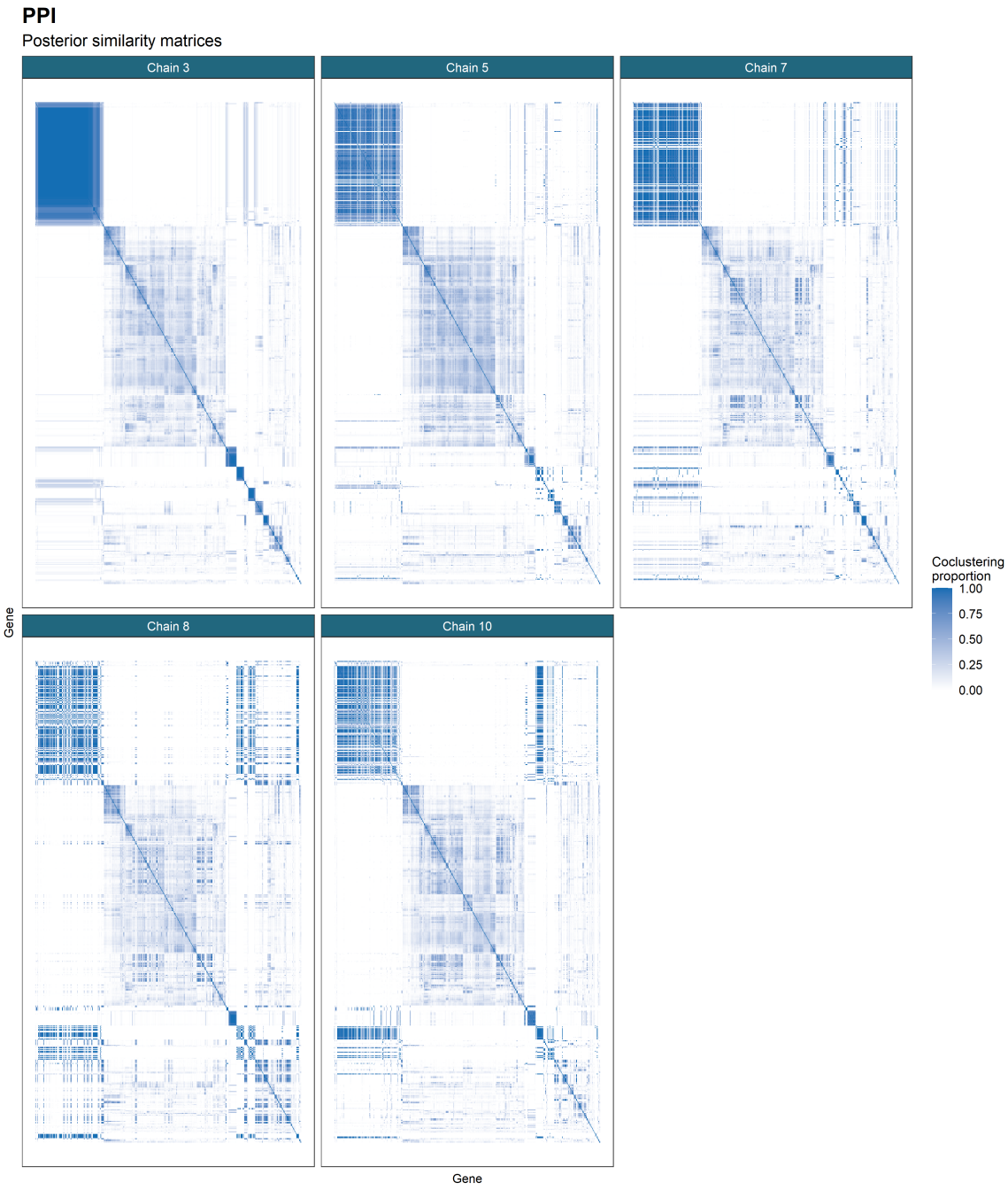

Figure 6: These PSMs have very large disagreements between eachother. There is some common agreement in the square in the centre of each plot. However, the other sections (which consist of the most confident allocations) appear to completely fail to overlap. These sections appear to be approximately random in the partition defined.

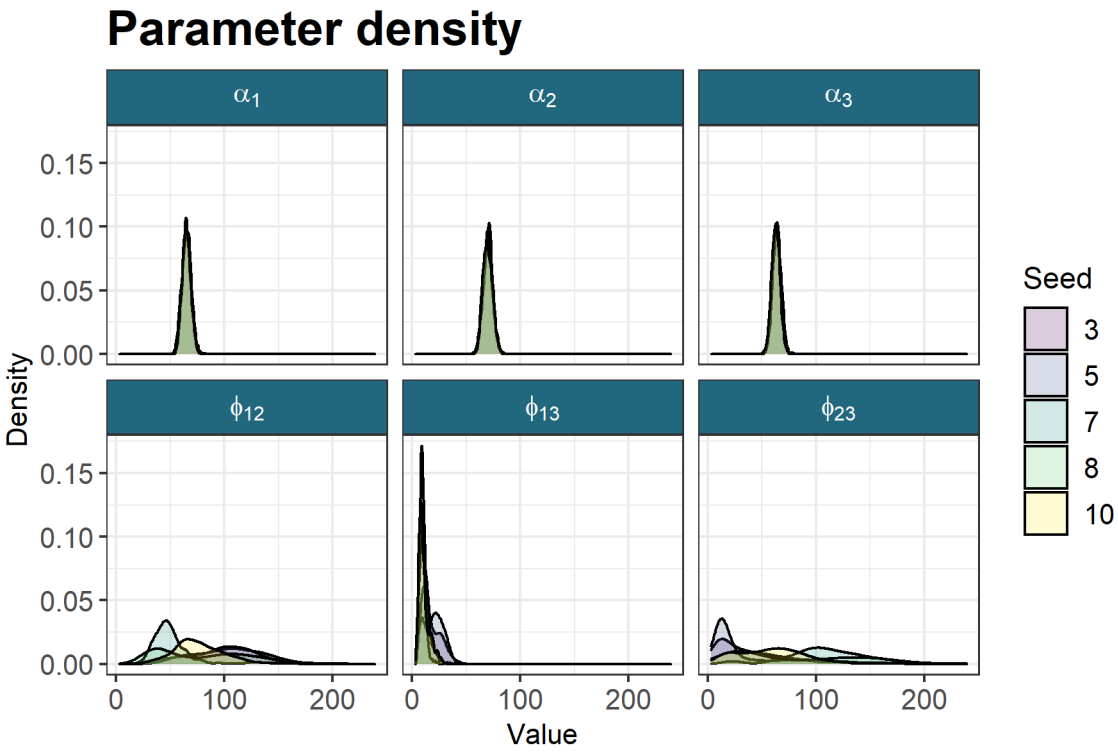

Figure 7: The densities of the continuous variables across the 5 chains kept for analysis. The mean sampled values are $\alpha_1 = 64.84$, $\alpha_2 = 69.85$, $\alpha_2 = 63.22$, $\phi_{12} = 81.76$, $\phi_{13} = 13.87$, and $\phi_{23} = 65.03$. It can be seen that different modes are being sampled for the ϕ parameters in each chain.

transcription atlas of the mitotic cell cycle in budding yeast. *Genome biology*, 11(3):1–11, 2010.

Christopher T Harbison, D Benjamin Gordon, Tong Ihn Lee, Nicola J Rinaldi, Kenzie D Macisaac, Timothy W Danford, Nancy M Hannett, Jean-Bosco Tagne, David B Reynolds, Jane Yoo, et al. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104, 2004.

Paul Kirk, Jim E Griffin, Richard S Savage, Zoubin Ghahramani, and David L Wild. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*, 28(24):3290–3297, 2012.

Christina Knudson and Dootika Vats. *stableGR: A Stable Gelman-Rubin Diagnostic for Markov Chain Monte Carlo*, 2020. URL `https://CRAN.R-project.org/package=stableGR`. R package version 1.0.

Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. Biogrid: a general repository for interaction datasets. *Nucleic acids research*, 34(suppl_1):D535–D539, 2006.

Dootika Vats and Christina Knudson. Revisiting the gelman-rubin diagnostic. *arXiv preprint arXiv:1812.09384*, 2018.