Statistical Methods

Description of the data

We sought to characterize the dynamics of trimethylation at histone H3 at lysine 36 (H3K36me3) in a large number of genes in three biological conditions: after light activation of Set2 (writer add), and after dark inactivation of Set2 with and without the presence of Rpd3 (writer loss and writer eraser loss, respectively). H3K36me3 levels were measured through ChIP-seq across four timepoints for each condition (0/20/40/60 min for writer add and 0/30/60/90 min for the loss conditions) for 5,368 genes. Seq data was measured from three sample replicates for each condition across the timepoints. We used two different approaches to process the seq data for analysis. First, within each replicate, H3K36me3 levels for a gene and condition were scaled to the proportion of maximum H3K36me3 level observed (for said replicate, gene, and condition), resulting in a common scale across all replicates, genes, and conditions. Second, we took the measurements, previously standardized based on spike-in samples and normalized to adjust for gene length, multiplied them by 1000, and rounded to an integer, resulting in count data with a distribution proportional to the starting data.

General statistical model

Though the H3K36me3 data are large and well-balanced, with the vast majority of genes having complete measurements of three replicates from each condition across four timepoints, they possess challenging features from a statistical modeling perspective. Namely, the statistical model should accommodate the non-normal distribution of the H3K36me3 measurements, as well as the time course, *i.e.* longitudinal, nature of measurements within a replicate. This broad range of features can be flexibly handled using Bayesian inference [4]. The Stan statistical platform [3] is one such computational tool for fitting complex Bayesian hierarchical models. We used the brms software package [1, 2], which acts as a wrapper of Stan in the R statistical programming language [5]. Let y_{ijkl} be the H3K36me3 measurement for the i^{th} replicate sample ($i = 1, 2, 3$) at the j^{th} timepoint ($j = 0, 1, 2, 3$) for the k^{th} gene ($k = 1, 2, \ldots, 5368$) for l^{th} condition ($l = 1, 2, 3$). Briefly, we model y with a GLM by defining $E(y) = g^{-1}(\eta)$, where $E(.)$ is the expected value of a random variable, $g^{-1}(.)$ is the inverse link function, and η is the linear predictor that relates the outcome to factors of interest. A Bayesian analysis could in principle simultaneously model all genes and conditions, though in practice, such an approach would likely be computationally infeasible, particularly when considering the need for sufficient sampling in order for the parameter estimates to converge. To avoid these challenges, we instead model more manageable subsets of H3K36me3 data. First, we fit a model of H3K36me3 data for a specific gene and condition. Second, to make more direct comparisons between the writer loss and writer eraser loss conditions, we modeled their data jointly within a gene.

Zero-one-inflated Beta regression model of H3K36me3 proportions

For the H3K36me3 data transformed to the proportion scale, we modeled the data with a zero-one-inflated beta distribution (ZOIB) and the following parameterization:

References

[1] Paul-Christian Bürkner. Advanced Bayesian Multilevel Modeling with the R Package brms. 10(July):395–411, 2017.

[2] Paul-Christian Bürkner. brms : An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, 80(1), 2017.

[3] Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan : A Probabilistic Programming Language. *Journal of Statistical Software*, 76(1), 2017.

[4] Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2006.

[5] R Core Team. RSoftware2019, 2019.