**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 (Gelman & Hill 2006).

The Stan statistical platform (Carpenter et al. 2017) is one such computational tool for fitting complex Bayesian hierarchical models. We used the brms software package (Bürkner 2017), which acts as a wrapper of Stan in the R statistical programming language (R Core Team 2019). Let be the H3K36me3 measurement for the ith replicate sample () at the jth timepoint () for the kth gene () for lth condition (). Briefly, we model y with a GLM by defining , where is the expected value of a variable, is the inverse link function, and 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.

**Beta regression model of H3K36me3 proportions**

For the H3K36me3 data transformed to the proportion scale, we modeled the data with a beta distribution and the following parameterization: , where is the beta function (Casella & Berger 2002) and is a positive precision parameter. For the link function, we used the logit: .

The H3K36me3 data scaled to the proportion of the maximum H3K36me3 measured within a replicate (three replicates per gene per condition) results in three values of 1 for each gene and condition pairing. Beta regression expects proportional outcomes between 0 and 1, and fails with values at 0 and 1. We explored two approaches to accommodating the large number of 1s and some number of 0s in these data. [+/- 0.0001 and zero-one-inflated beta].

Continuous time model: We modeled the gain or loss of H3K36me3 with a continuous time variable in the linear predictor of the model,

,

where μ is a shared intercept term, is the *j*th timepoint (in minutes) for the *i*th replicate, and is the log odds ratio (OR) for change in proportion H3K36me3 (compared to the maximum) per minute. For all genes and conditions, we recorded the posterior mean ( as a point estimate for the change in H3K36me3 proportion with time, the standard error on the estimate, and the 95% Credible Interval (CrI), which we use to define “confident” genes for a given condition, which possess CrI that do not cover 0. were plotted and correlated with additional factors of interests in order to identify interesting trends.

Categorical time model: Notably, writer loss and writer eraser loss had similar H3K36me3 dynamics with time. To more directly compare the dynamics of writer loss and writer eraser loss, we simultaneously analyzed their data per gene within a categorical time model, allowing for greater flexibility and non-linearity with respect to time:

where μ is the intercept representing log odds for the proportion H3K36me3 at time0 for writer loss, is the log OR comparing timepoint to 0 for writer loss, and represents deviations at each timepoint for writer eraser loss.

**Bibliography**

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models (Analytical Methods for Social Research)*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511790942

﻿Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, 76(1). http://doi.org/10.18637/jss.v076.i01

﻿Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, 80(1). <http://doi.org/10.18637/jss.v080.i01>

﻿Casella, G., & Berger, R. L. (2002). *Statistical Inference*. Thomson Learning.