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 [5]. 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 [6]. 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:

$$p(y_i) = \begin{cases} \alpha(1-\gamma) & y_i = 0 \\ \alpha\gamma & y_i = 1 \\ \frac{y_i^{\mu\phi-1}(1-y_i)^{(1-\mu)\phi-1}}{B(\mu\phi,(1-\mu)\phi)} & y_i \notin \{0,1\} \end{cases} \tag{1}$$

α is the zero-one-inflation probability (probability that a zero or one occurs), γ is the conditional one-inflation probability (probability that one occurs rather than zerio), $B(.)$ is the beta function [4], and ϕ is a positive precision parameter. For the link function, we used the logit: $g(\eta) = \log(\frac{\eta}{1-\eta})$. Zero-one-inflation was used because standard beta regression expects $y \in (0, 1)$, but not at the boundaries.

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. Additionally, zeros may be observed. We used this scale of the data to better standardize ChIP-seq dynamics across cells and better detect consistent patterns compared to the normalized ChIP-seq data. However, it represents a challenging scale in the context of a GLMM, and an abuse of assumptions of beta regression specifically. The zeros and ones are informative, but would strongly violate the expectations of standard beta regression, resulting in anti-conservative, extreme parameter estimates. Instead, we used ZOIB regression, which conservatively drops out zeros and ones in terms of parameter estimation, resulting in less extreme estimates for the time effect. Alternatively, we modeled the non-proportional, normalized ChIP-seq data.

Zero-inflated negative binomial regression model of H3K36me3 proportions

Time models

Continuous

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

$$\eta_{ij} = \mu + u_i + (\beta_{\text{time}} + v_i)x_{ij}, \tag{2}$$

where μ is a shared intercept term, x_{ij} is the j^{th} timepoint (in minutes) for the i^{th} sample, and β_{time} is the log odds ratio (OR) for change in proportion H3K36me3 (relative to the maximum) per minute. u_i and v_i are random, or group-level [5], effects that account for the longitudinal nature of the data, and modeled with the following priors: $\text{N}(0, \tau_u^2)$ and $\text{N}(0, \tau_v^2)$, respectively. For all genes and conditions, we recorded the posterior mean ($\widehat{\beta}_{\text{time},kl}$) as a point estimate for the change in H3K36me3 proportion with time for gene k and condition l, 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. $\widehat{\beta}_{\text{time},kl}$ were plotted and correlated with additional factors of interests in order to identify interesting trends.

Categorical

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:

$$\begin{aligned}\eta_{ijl} = & \underbrace{\mu + u_i}_{\text{WL time0}} + \underbrace{(\delta_{\text{WEL}} + w_i) I\{l = \text{WEL}\}}_{\text{WEL time0}} \\ & + \underbrace{\sum_{p=1}^{3} (\beta^p_{\text{timepoint}} + v_i^p) I\{j = p\}}_{\text{WL timepoint } p} \\ & + \underbrace{\sum_{p=1}^{3} (\delta^p_{\text{timepoint}} + z_i^p) I\{j = p, l = \text{WEL}\}}_{\text{WEL timepoint } p},\end{aligned} \tag{3}$$

where μ is the intercept representing log odds for the proportion of maximum H3K36me3 at time0 for WL, δ_{WL} is the log OR at time0 comparing WEL to WL, $\beta^p_{\text{timepoint}}$ is the log OR comparing timepoint p to 0 for WL, and $\delta^q_{\text{timepoint}}$ represents log OR for timepoint p for WEL compared to WL. u_i, v_i^p, z_i^p are group-level effects that model the correlations within replicates, and have the following prior distributions: $\text{N}(0, \tau_u^2)$, $\text{N}(0, \tau_{v,p}^2)$, and $\text{N}(0, \tau_{z,p}^2)$, respectively. $I\{A\}$ represents the indicator function, returning 1 if condition A is satisfied, and 0 if not.

Sampling and model convergence

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] G. Casella and R.L. Berger. *Statistical Inference*. Duxbury advanced series in statistics and decision sciences. Thomson Learning, 2002.

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

[6] R Core Team. RSoftware2019, 2019.