diffloop: analysis of variable loops under different models

Caleb Lareau and Martin Aryee

August 1, 2017

Overview

The purpose of this document is to provide a high-level comparison of different models fit to the K562 and MCF7 ChIA-PET data from ENCODE. Specifically, we examine the association statistics from using voom, edgeR, and binned associations.

Mean-Variance Plot

The mean-variance plot of the ChIA-PET data analyzed in this manuscript is shown in **Figure S1**. RNA-Seq count data is often modeled as negative binomial, since the variance exceeds the mean resulting in an overdispersion relative to the Poisson model. We hypothesized that a similar model could be applied to counts data in the diffloop framework.

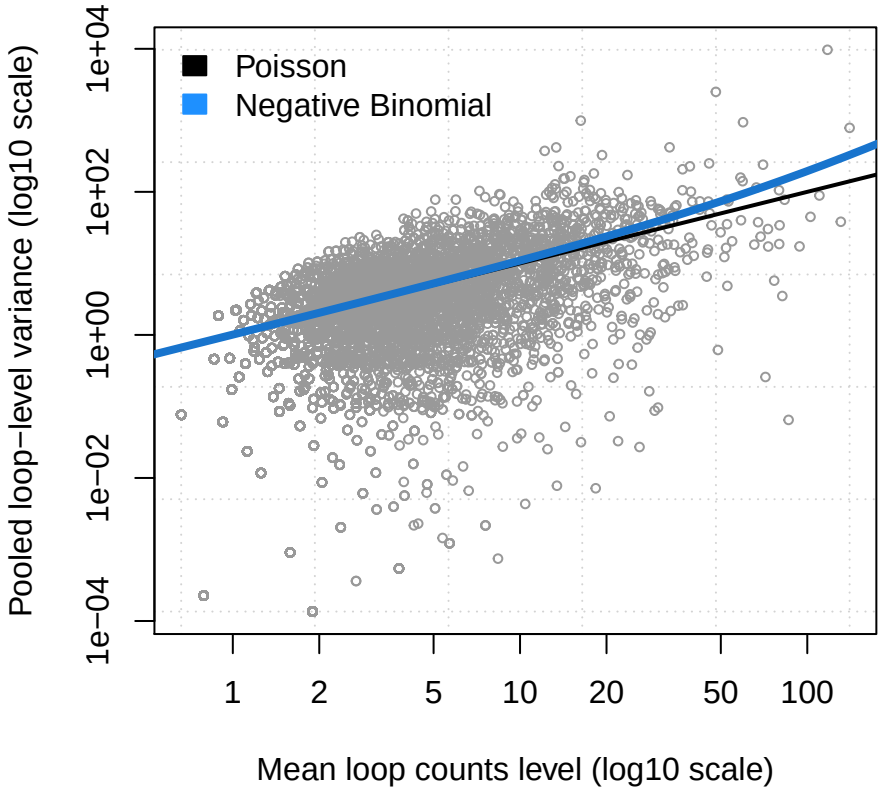

Figure S1. Mean-variance plot of 2x2 differential association between MCF7 and K562.

While the Negative Binomial line does not obviously deviate from the Poisson line for the loop counts data considered here, we note a similar pattern in RNA-Seq data for low counts in the displayed range while the variance deviates more clearly from the mean at larger count values. We expect that similar behaviour will be observed at higher ChIA-PET counts and therefore implemented the negative binomial as the default model in diffloop.

Loop Width Stratification

Though the size factor correction as implemented in diffloop adjusts for variable read depth between samples (Figure 2), we considered whether or not the size factor should vary with the loop width spanning the two anchors. For example, the plot below of differential associations calculated within loop width strata shows a potential weak dependence on loop width (**Figure S2**).

To assess the effect of variable loop widths, we binned loops based on their distance spanned using the groups above (increments of 0.3 on the $\log_{10}$ scale) and assessed differential looping within each stratum individually. **Figure S3** shows a per-loop comparison of the $-\log_{10}$ ratio of the q-values using the standard model compared to the stratified model. Loops annotated in red and blue are shown with summary statistics in **Table S1** and **Table S2**.

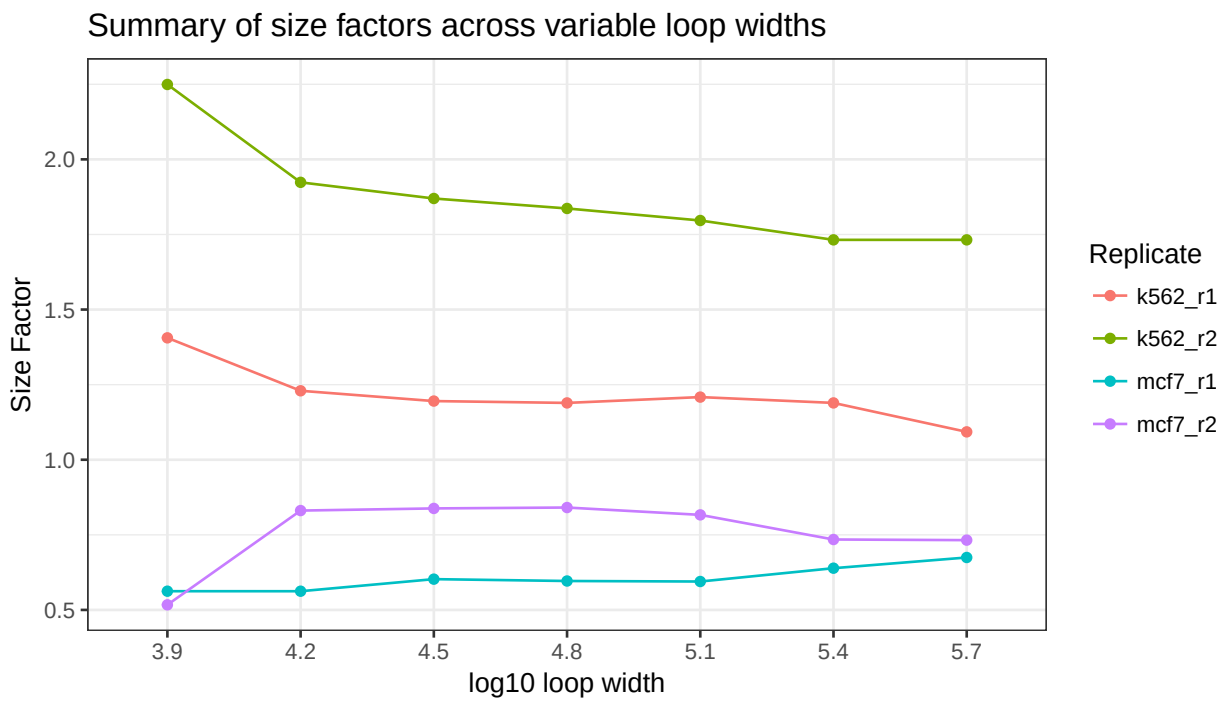

Figure S2. Variable size factor across different loop widths (distance spanned between the midpoint of the anchors).

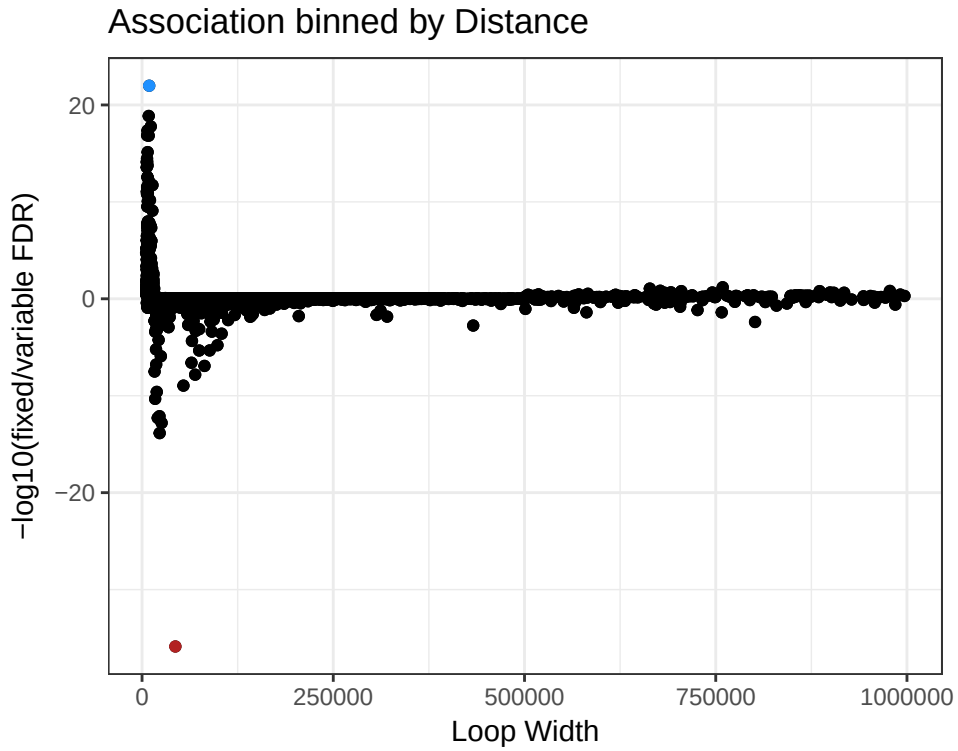

Figure S3. Comparison of per-loop statistical significance with using the standard unstratified model vs. a stratified model based on loop width. Loops marked in red are more significant (> 20 orders of magnitude) when using the unstratified model between the tests whereas the blue loops are more significant (> 20 orders of magnitude) when using the stratified model. The Y axis shows the $-\log_{10}$ ratio of the q-values.

k562_r1	k562_r2	mcf7_r1	mcf7_r2	LoopWidth	Regular.FDR	Variable.SF.FDR
9	19	158	194	43600	1.07e-64	1.37e-104

Table S1. Summary statistics of the loops marked in red in **Figure S3.** These loops were highlighted due to large differences in the summary statistic measures such that the unstratified model yielded a less significant test statistic. Individual replicate PET counts, the distance the loop spans, as well as the q-values for each test are displayed.

k562_r1	k562_r2	mcf7_r1	mcf7_r2	LoopWidth	Regular.FDR	Variable.SF.FDR
0	0	75	114	9360	1.3e-61	7.12e-43

Table S2. Summary statistics of the loops marked in blue in **Figure S3.** These loops were highlighted due to large differences in the summary statistic measures such that the unstratified model yielded a more significant test statistic. Individual replicate PET counts, the distance the loop spans, as well as the q-values for each test are displayed.

We note that all the highlighted loops are deeemed highly statistically significant in both stratified and unstratified analyses. To explore whether the two models identify different sets of differential loops we plotted the $-\log_{10}$ q-values against each other after applying a ceiling of 4 (q = 0.0001) (**Figure S4**). Of note, few loops tended to deviate significantly in this range indicating that either approach produces a very similar set of differential loops. Thus, in the present implementation of diffloop, a single unstratified model is applied and the loop width is not considered. As more topology libraries become available, novel methods of association, including those that account for loop width, may be incorporated in the package as the evidence demands.

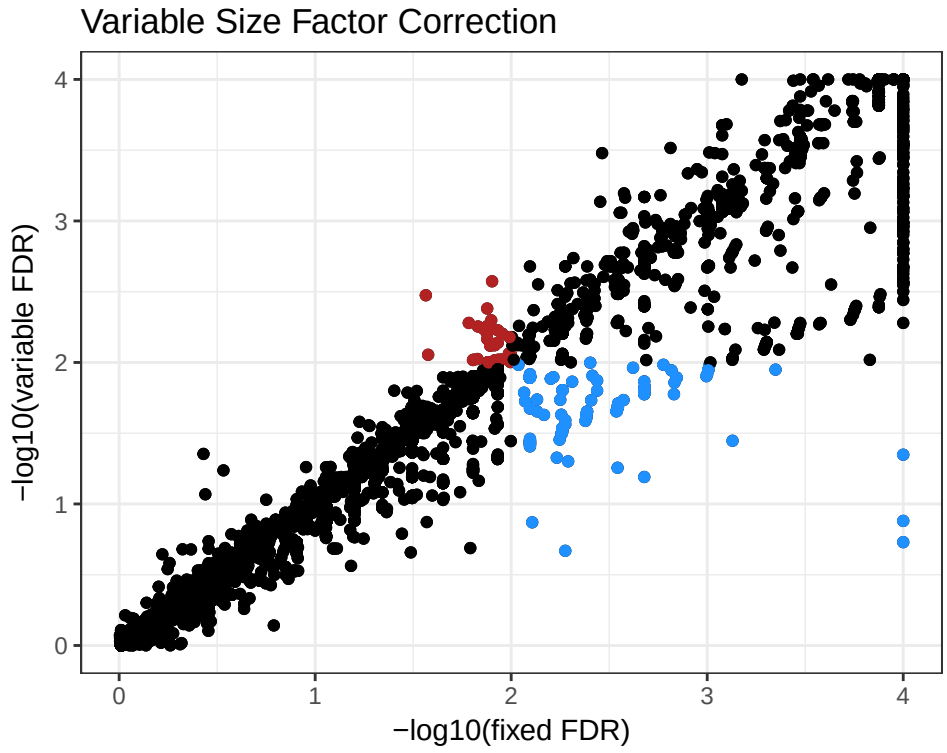

Figure S4. A scatter plot comparing the $-\log_{10}$ FDR q-values between 0 and 4 of the fixed and binned association tests. For either test, $-\log_{10}$ FDR q-values greater than 4 were shrunk to 4 for visualization. Loops marked in red were differential (q-value < 0.01) in the binned association test.

Comparison with Voom

As an alternative model to Negative Binomial regression for RNA-Seq data, the limma-voom method approximates the mean-variance relationship of the log-counts and applies a moderated t-test. We implemented a similar approach for finding differential loops analogous to differential transcripts in this model. **Figure S5** shows a per-loop comparison of the Negative Binomial regression (edgeR) and the voom q-values as a function of loop width. We also highlight the largest deviants in **Table S3**.

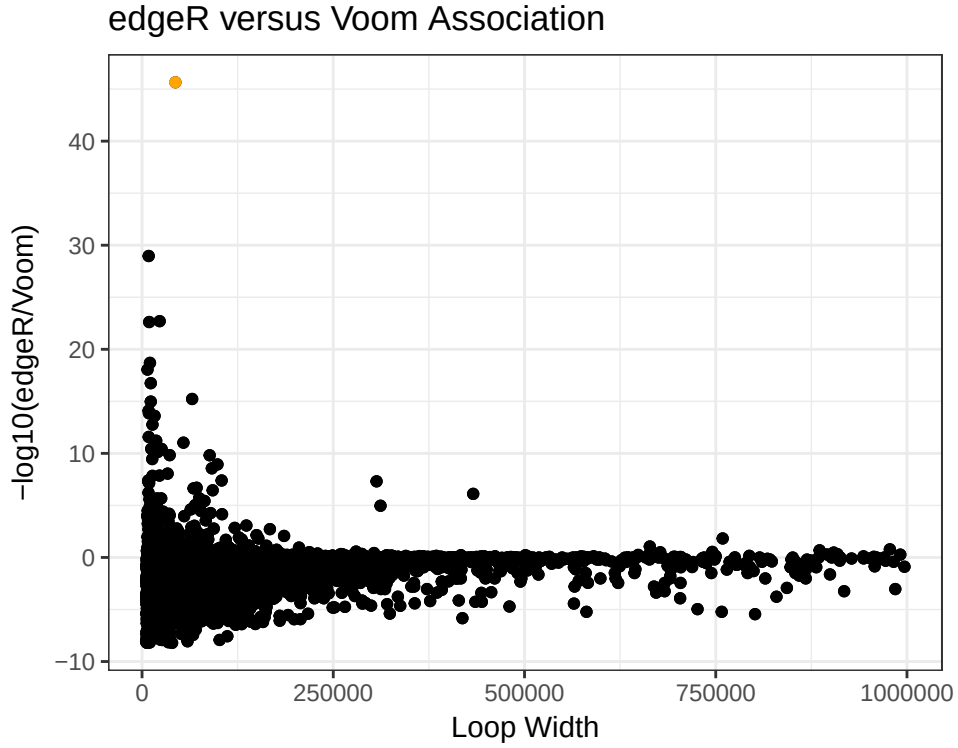

Figure S5. Comparison of per-loop statistical significance using models from edgeR and Voom. Loops highlighted in orange were assigned a much more statistically significant q-value in edgeR (> 30 orders of magnitude) than voom. These loops are shown in **Table S3**.

k562_r1	k562_r2	mcf7_r1	mcf7_r2	LoopWidth	edgeR.FDR	Voom.FDR
9	19	158	194	43600	1.07e-64	4.78e-19

Table S3. Summary statistics of the loops marked in orange in **Figure S5.** These loops were highlighted due to large differences (> 30 orders of magnitude) in the summary statistic measures between the edgeR and Voom association models.

A trend emerges as shown in **Figure S5** where differential loops spanning smaller distances were more statistically significant in the edgeR association model. We note however that the loops with the largest deviation (orange, **Table S4**) are in fact highly significant in both models. To assess whether any loops might switch from 'significant' to 'not significant' we again plotted the $-\log_{10}$ q-values with a cap at 4 (**Figure S6**). Loops that are significant in one model but not the other are marked by color. Summary statistics for the orange loops, where edgeR indicates a highly significancant difference (FDR < 0.01) whereas voom suggests no/little difference (FDR > 0.1) are shown in **Table S4**.

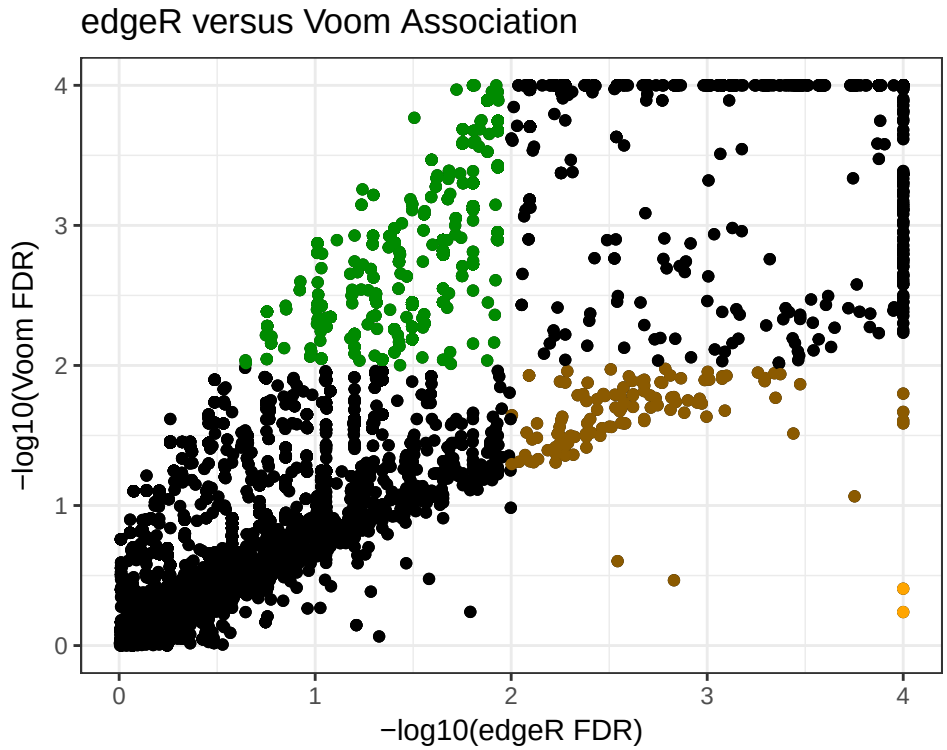

Figure S6. Comparison of per-loop statistical significance using models from edgeR and voom. Loops highlighted in orange were indicated as the edgeR model provided more statistically significant FDR value. These loops are shown in **Table S4**.

k562_r1	k562_r2	mcf7_r1	mcf7_r2	LoopWidth	edgeR.FDR	Voom.FDR
165	61	7	12	8600	3.1e-15	0.393
2	1	0	50	65500	3.44e-16	0.576

Table S4. Summary statistics of the loops marked in orange in **Figure S6.** These loops were highlighted due to differences in the summary statistic measures such that they were differential (FDR < 0.01) in the edgeR model but not differential (FDR > 0.1) in the voom model.