Visualization of the `loess` loint normalization over varying resolutions

John Stansfield, Mikhail Dozmorov

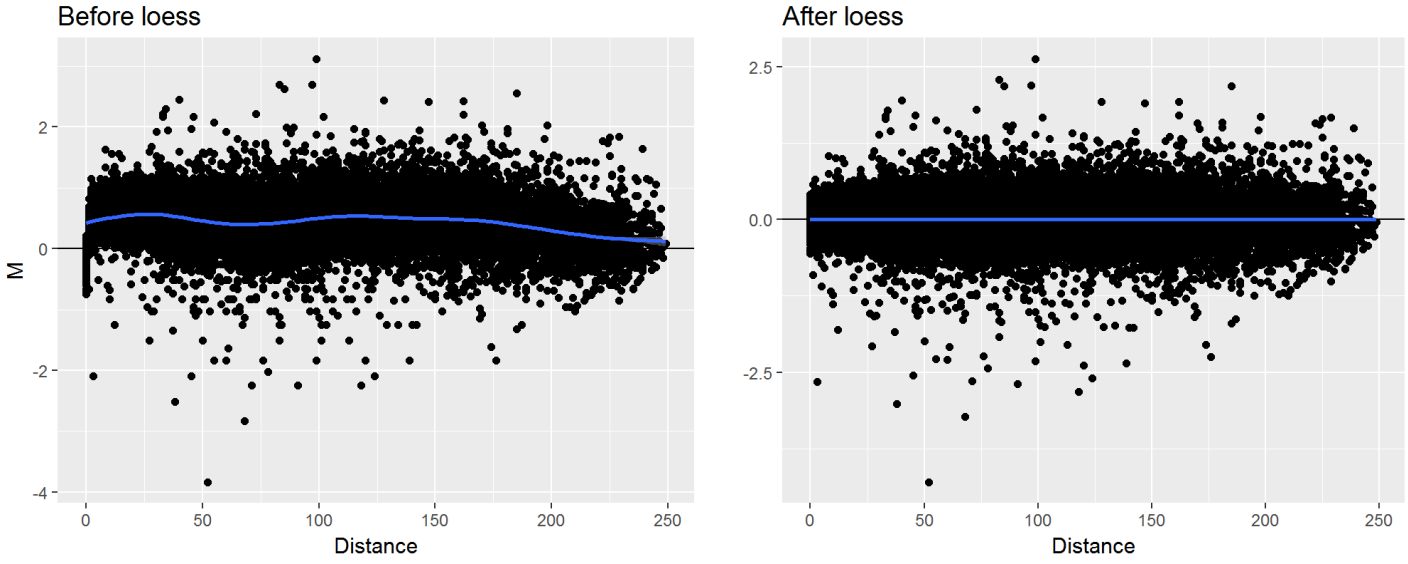
Introduction

Real Hi-C data from Gm12878 cell line were used. The data used were from chromosome 1 cut either using the DpnII enzyme or MboI enzyme at varying resolutions of 1MB, 500KB, 100KB, 50KB, and 5KB. The increased resolution (smaller length of genomic region) is accompanied by the increased proportion of zero interaction frequencies and the overall smaller dynamic range of IFs. The goal of this vignette is to observe the effect of resolution on the performance of joint `loess` normalization.

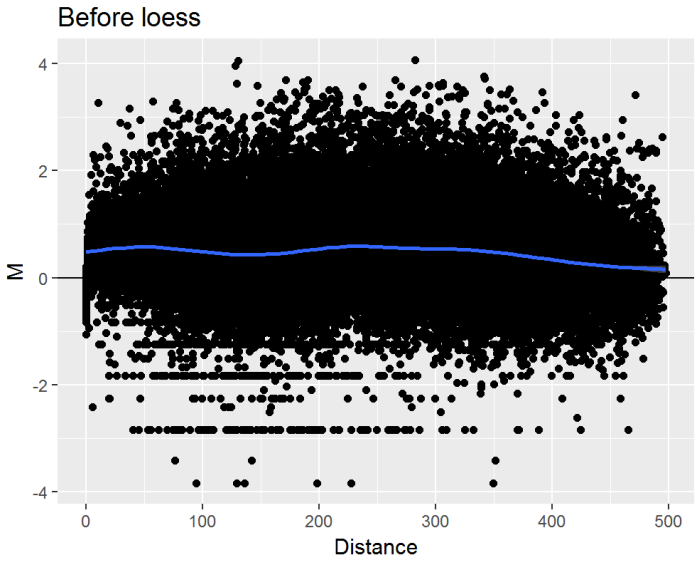
Perform joint loess normalization at varying resolutions

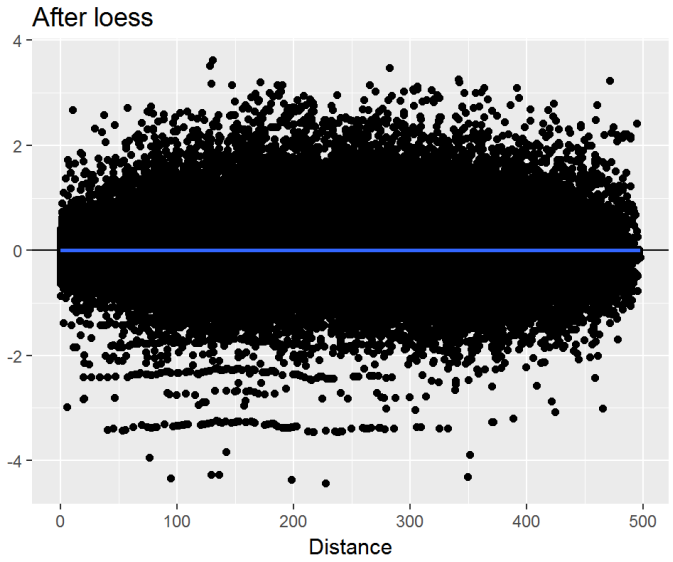
Here the `hic_loess` procedure is perforemd for the comparison of MboI and DpnII in GM12878 for chromosome 1 at varying resolutions.

1MB

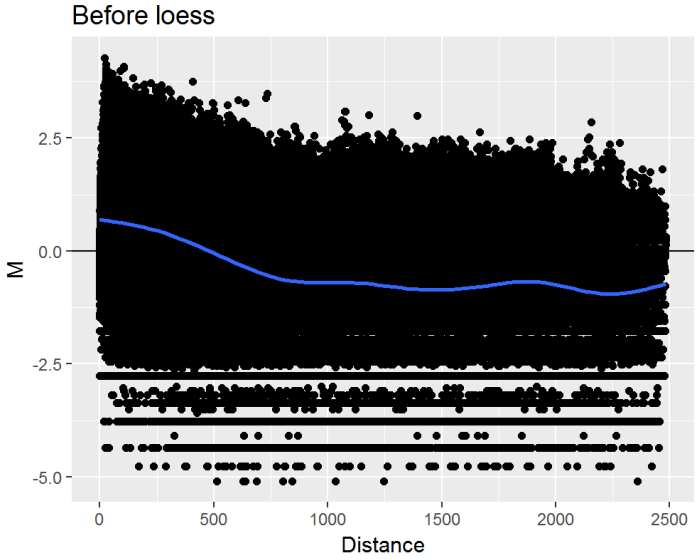


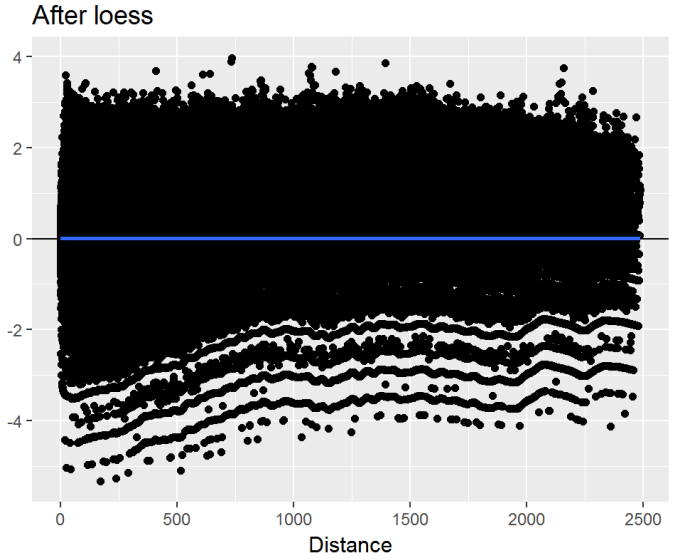
500KB


Before loess
M
Distance

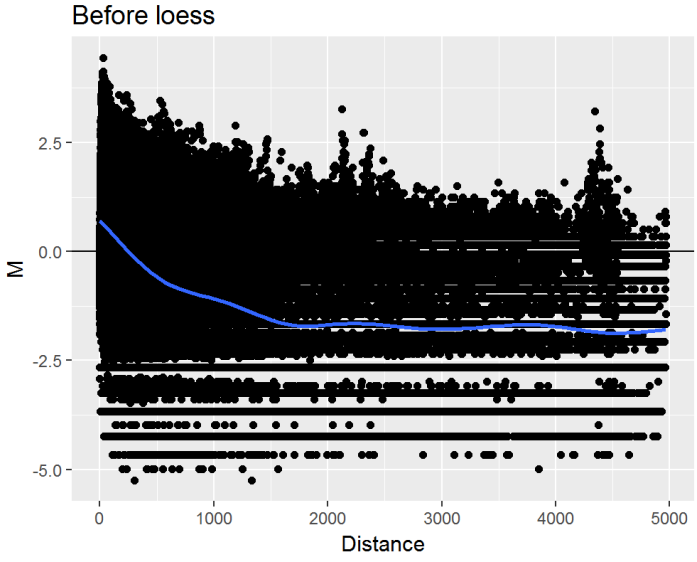

After loess
Distance

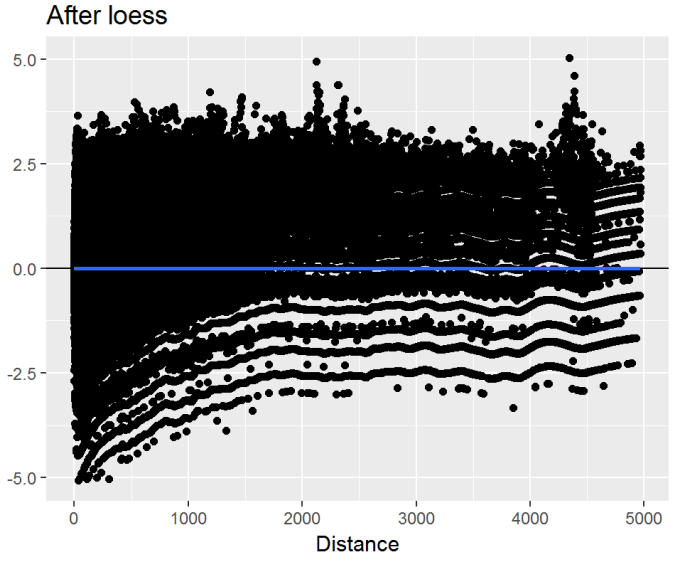
100KB


Before loess
M
Distance


After loess
Distance

50KB


Before loess
M
Distance


After loess
Distance

Summary

`loess` works well for removing biases at resolutions between 1MB and 100KB. Once the resolution is higher than 100KB, the procedure begins to fail due to the sparsity of the data. At high resolutions Hi-C data becomes very sparse with most values in the matrix being 0 or a small number. Thus when plotted on the MD plot the sparsity begins to show as the straight horizontal lines of points representing very small differences existing between the two datasets due to the sparsity of the sequencing coverage.