Estimation of the power-law dependence between the $log_{10} - log_{10}$ SD of interaction frequencies and distance between interacting regions

John Stansfield, Mikhail Dozmorov

Contents

Introduction

To estimate the parameters for simulating individual Hi-C matrices, Hi-C data from Gm12878 cell line [@Rao:2014aa] were used (Supplementary Table 1, GSE63525). The first dataset was obtained with DpnII restriction enzyme, while the second dataset was obtained with the MboI enzyme. Data from chromosome 1 was used at resolutions of 1Mb, 500kb, 100kb, and 50kb. The data were converted in a sparse matrix format (see `HiCdiff-vignette.Rmd` for details). Additional Gm12878 Hi-C data from chrs 1, 18, and 19 at 1Mb resolution, cut using the DpnII and MboI enzymes were also included.

SD of IFs vs. distance dependence

First, we estimate the power-law approximation of the dependence between the standard deviation (SD) of interaction frequencies (IFs) and distance by fitting the power-law estimate and assessing the fit using Kolmogorov-Smirnov test. Because SDs at larger distances do not fit the power-law well the outlier values are iteratively removed, starting from largest distances, until Kolmogorov-Smirnov test results indicate the power-law fit is adequate. The α power-law parameter can then be used to approximate the decay of SD with distance.

As with the decay of IFs with distance, the power-law approximation can be affected by multiple factors, e.g., the resolution of the data, the enzymes used to obtain the data, the chromosomal differences (e.g., chromosome length (chromosome 1 being the longest), gene density (gene-poor chromosome 18 and gene-dense chromosome 19)).

The effect of resolution

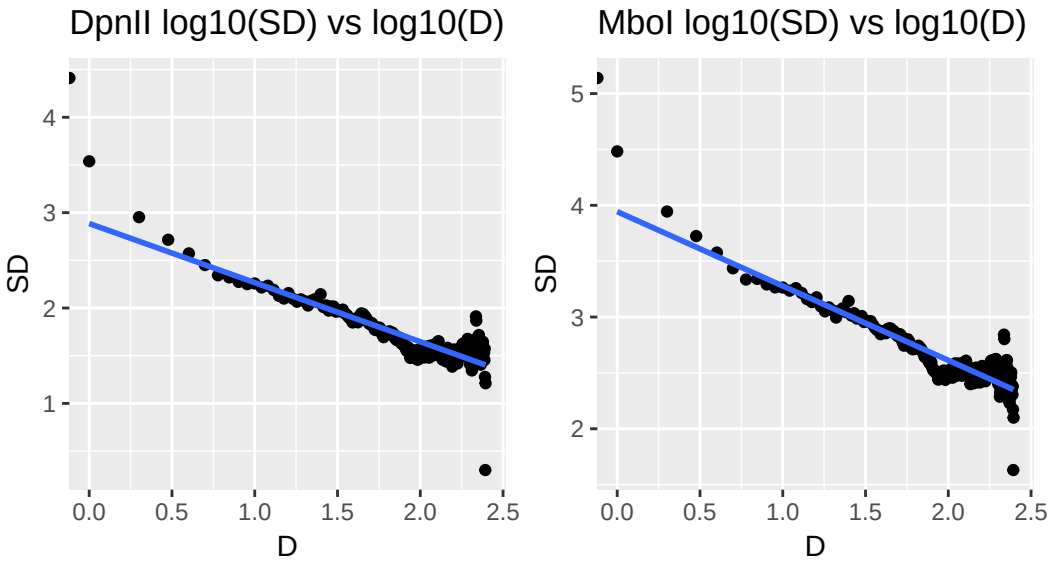
Here the fit to the power-law of the SD of IFs at each distance is tested for Chr 1 from GM12878 using cutting enzyme DpnII at 1MB, 500KB, 100KB, 50KB resolution, respectively.

Tables show the output of `fitdistplus::power.law.fit` function. Key variables to note are `alpha` - the power of the $C * x^{-alpha}$ power-law formula, and `KS.p` - p-value of the Kolmogorov-Smirnov test, larger p-value means that the power-law fit is adequate. The first row is for DpnII and the second row is for MboI.

The plots represent the log10(SD) vs log10(Distance), one plot per cutting enzyme.

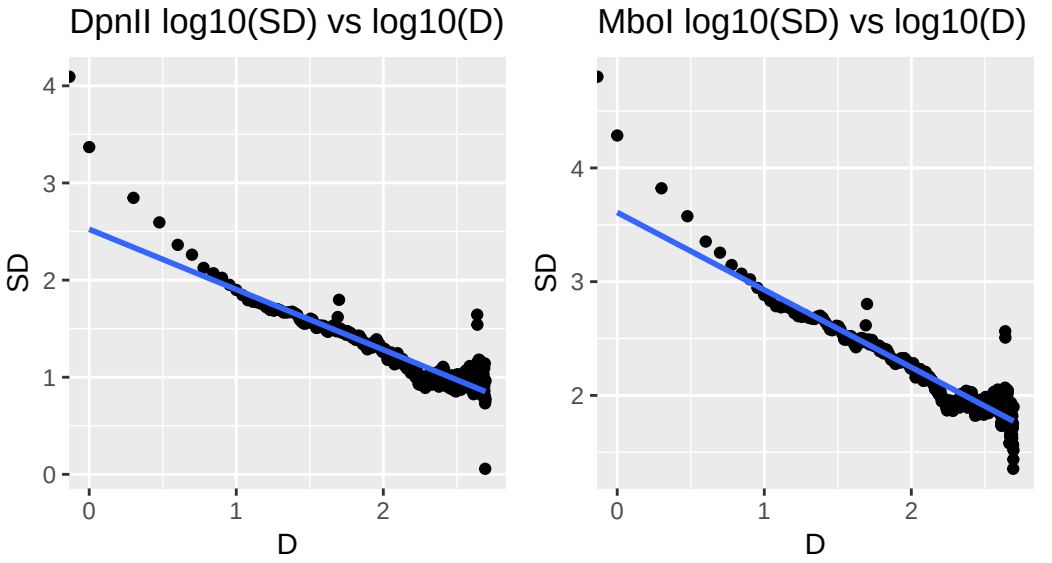
1 MB

Enzyme	continuous	alpha	xmin	logLik	KS.stat	KS.p
DpnII	TRUE	2.331709	42.20044	-494.6538	0.04391755	0.9925585
MboI	TRUE	2.365061	750.7263	-353.8661	0.07215841	0.9732388



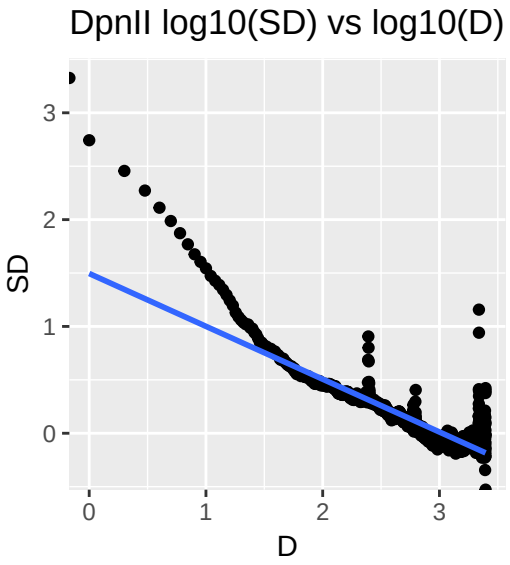
500KB

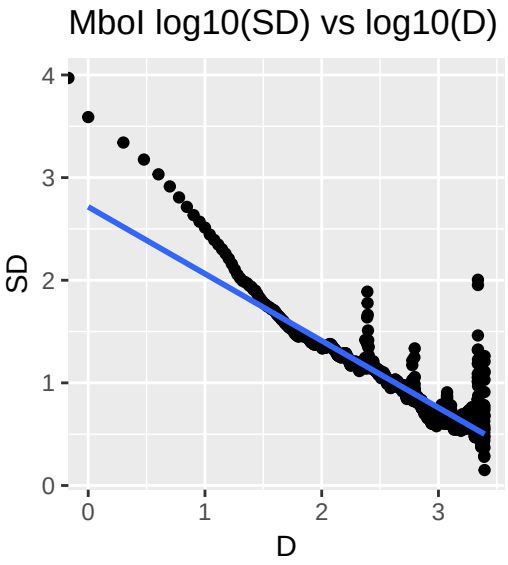
Enzyme	continuous	alpha	xmin	logLik	KS.stat	KS.p
DpnII	TRUE	2.334898	11.89302	-720.3278	0.0427384	0.8899287
MboI	TRUE	2.231429	105.525	-1133.574	0.06889731	0.3533398



100KB

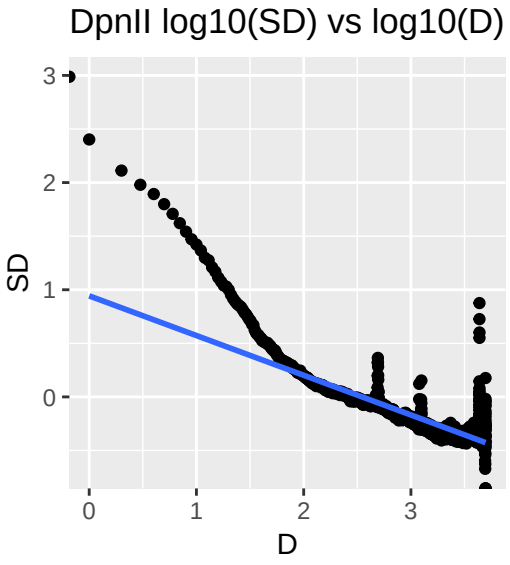
Enzyme	continuous	alpha	xmin	logLik	KS.stat	KS.p
DpnII	TRUE	1.729765	7.062814	-153.126	0.03494209	1
MboI	TRUE	1.857958	28.9925	-409.3773	0.04058937	0.9997465

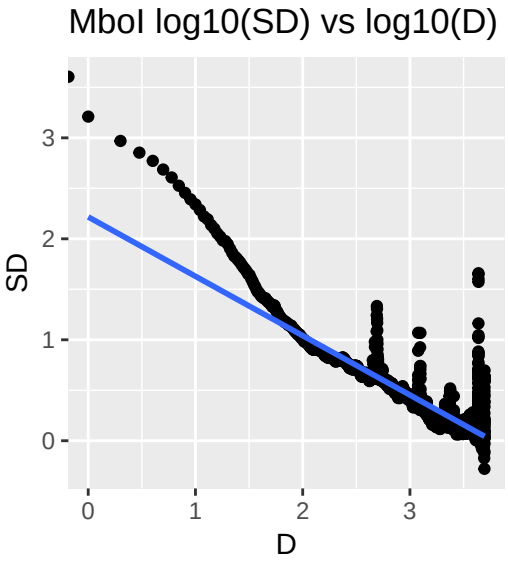




50KB

Enzyme	continuous	alpha	xmin	logLik	KS.stat	KS.p
DpnII	TRUE	1.757461	3.343024	-175.0236	0.03571082	1
MboI	TRUE	1.791732	14.98267	-400.6679	0.0302201	0.9999997





Summary

The SD of the IFs seems to fit the power-law adequately over the range of resolutions after the outliers are removed, however some of the plots still show some deviations from the ideal fit. α ranges from 1.73 to 2.68. There is more variability in the α parameter for modeling SD compared to the median IF.

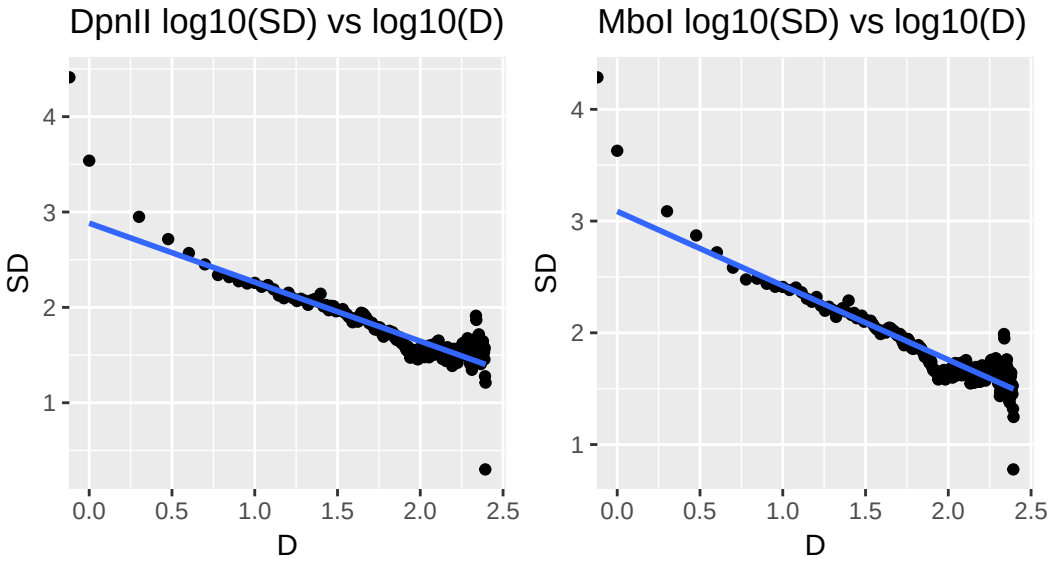
The effect of chromosomes

Here the fit to the power-law of SD of IF at each distance is tested for chromosome 1, 18, 19 from GM12878 cell line using cutting enzymes DpnII and MboI at 1MB resolution.

As above, the table shows the output of `fitdistplus::power.law.fit` function. Key variables to note are `alpha` - the power of the $C * x^{-alpha}$ power-law formula, and `KS.p` - p-value of the Kolmogorov-Smirnov test, larger p-value means that the power-law fit is adequate. The plots represents the log10(SD) and log10(Distance), one plot for each cutting enzyme.

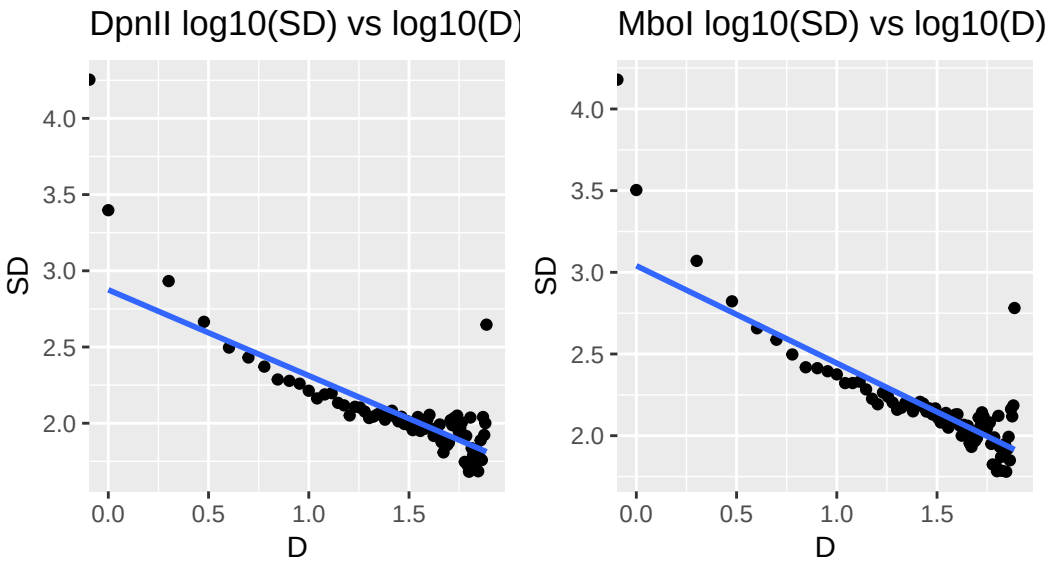
Chr 1

Enzyme	continuous	alpha	xmin	logLik	KS.stat	KS.p
DpnII	TRUE	2.336356	42.17819	-494.0248	0.04407543	0.9922233
MboI	TRUE	2.247001	55.16459	-514.4183	0.06870571	0.7722857



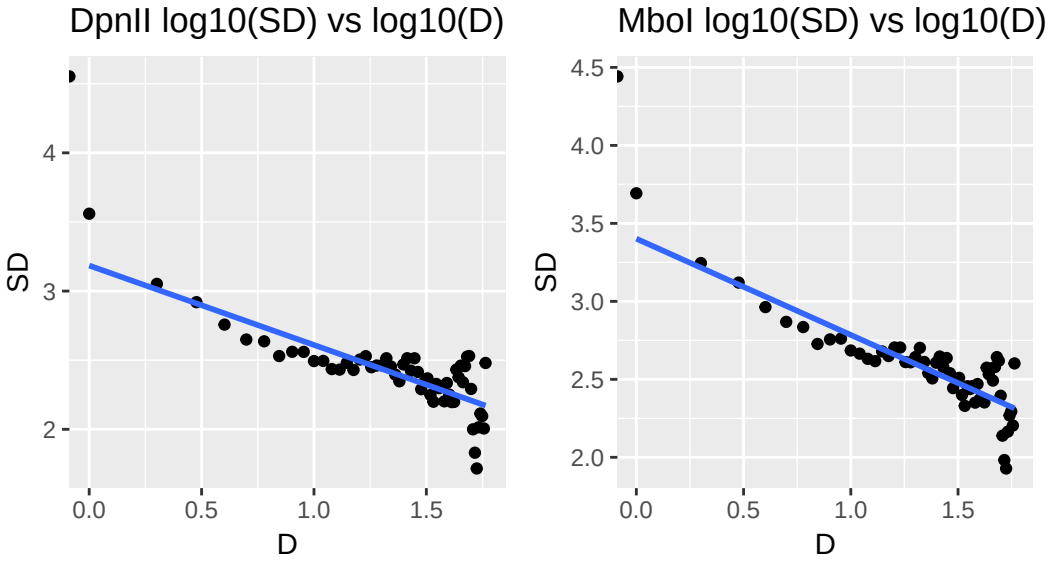
Chr 18

Enzyme	continuous	alpha	xmin	logLik	KS.stat	KS.p
DpnII	TRUE	1.797929	193.6994	-69.70759	0.09225675	0.9999956
MboI	TRUE	1.822987	262.2809	-71.81386	0.09429987	0.9999921



Chr 19

Enzyme	continuous	alpha	xmin	logLik	KS.stat	KS.p
DpnII	TRUE	1.826717	363.6522	-66.36875	0.102455	0.9999826
MboI	TRUE	1.904656	576.7541	-68.50406	0.1064836	0.9999559



Summary

The power-law fit is better over the varying chromosomes at 1MB resolution after the outliers were removed. α ranges from 1.79 to 2.25. The plots of the fits show less deviations compared to the plots in the effect of resolution section. For simulations an α between 1.7 and 2.7 should provide give a reasonable approximation to the data.