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

John Stansfield, Mikhail Dozmorov

Introduction

To estimate the parameters for simulating individual Hi-C matrices, Hi-C data from Gm12878 cell line (Rao et al. 2014) were used (Supplementary Table 1, GSE63525 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=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 upper triangular 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.

Median IF vs. distance dependence

First, we estimate the power-law approximation of the dependence between interaction frequencies (IFs) and distance (Lieberman-Aiden et al. 2009; Sanborn et al. 2015; Ay and Noble 2015; Fudenberg et al. 2016; Nagano et al. 2015) by fitting the power-law estimate and assessing the fit using Kolmogorov-Smirnov test. Because IFs at larger distances do not fit the power-law well (Supplementary Figure 1) 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 median IF 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)).

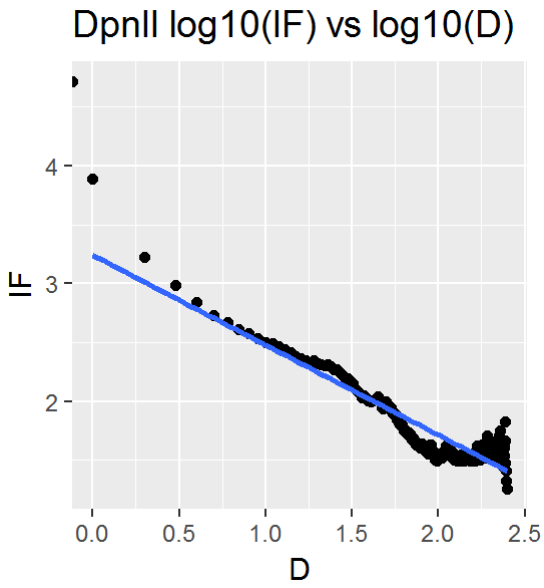
Effect of Resolution

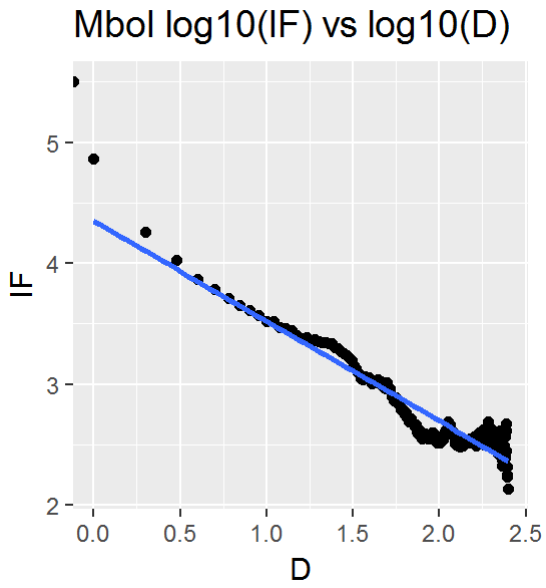
Here the fit to the power-law of median IF at each distance is tested for chromosome 1 from GM12878 cell line using cutting enzymes DpnII and MboI at 1MB, 500KB, 100KB, 50KB resolution, respectively.

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(median IF) and log10(Distance), one plot for each cutting enzyme.

1 MB

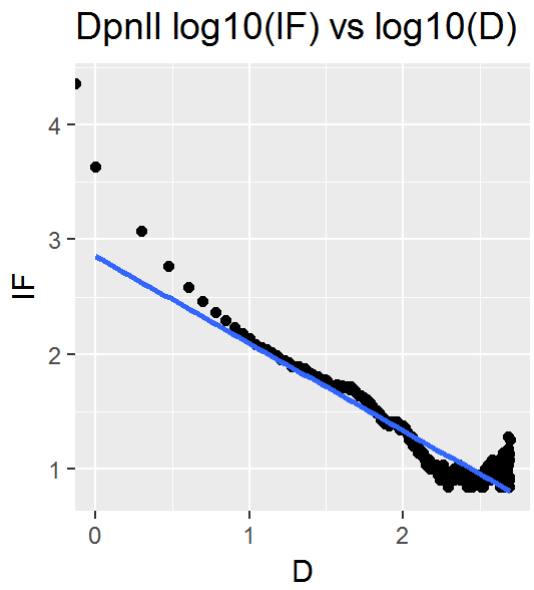
Enzyme	continuous	alpha	xmin	logLik	KS.stat	KS.p
DpnII	TRUE	2.03132	77	-364.3907	0.0616977	0.978235
MboI	TRUE	2.049246	924	-445.419	0.08447202	0.8519587

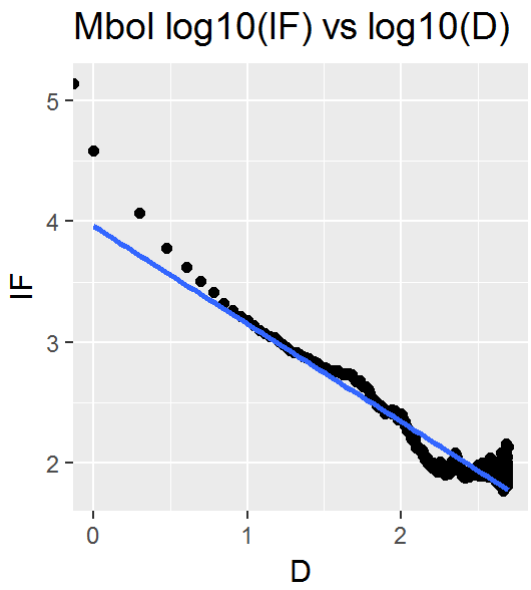




500KB

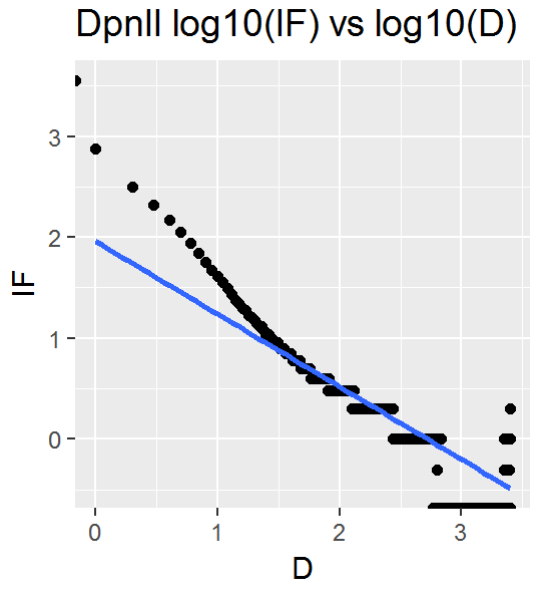
Enzyme	continuous	alpha	xmin	logLik	KS.stat	KS.p
DpnII	TRUE	2.064016	20	-536.0869	0.06545282	0.728491
MboI	TRUE	2.0716	218.5	-775.8375	0.0761566	0.5581262

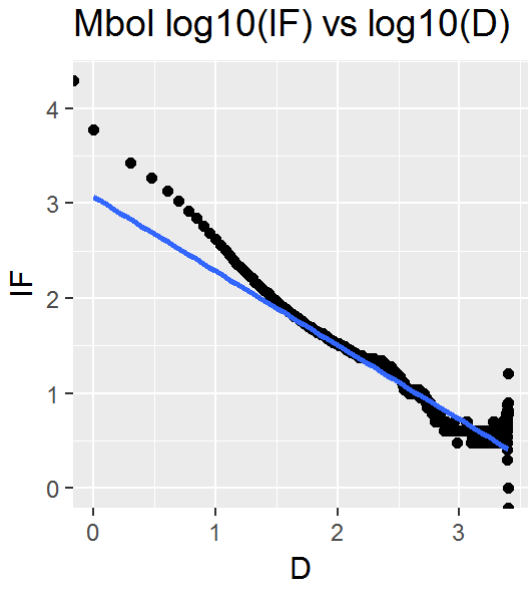




100KB

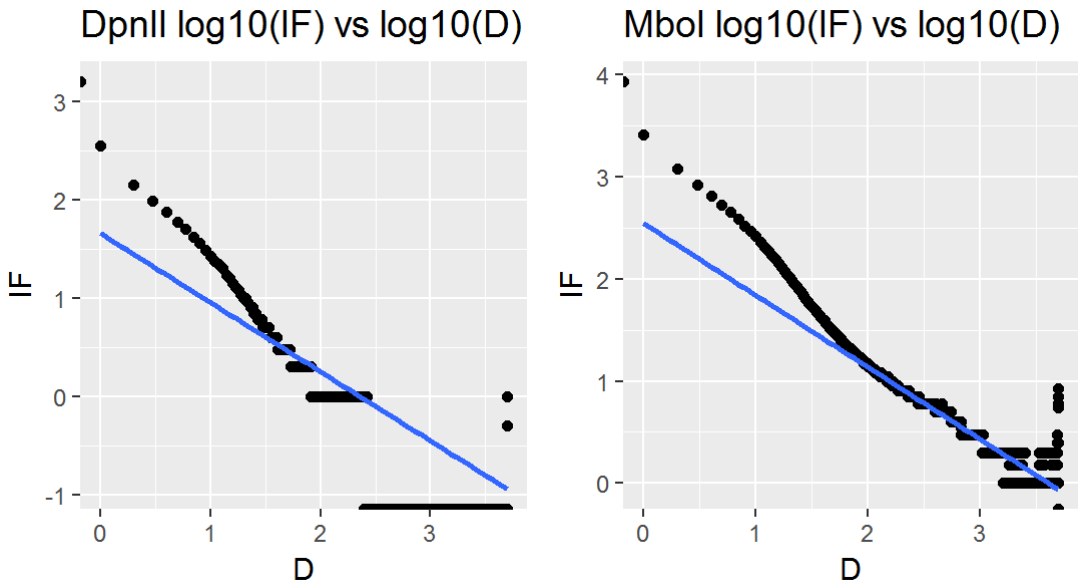
Enzyme	continuous	alpha	xmin	logLik	KS.stat	KS.p
DpnII	TRUE	1.731601	10	-144.4771	0.03851044	1
MboI	TRUE	1.807315	68	-286.9056	0.03292535	1





50KB

Enzyme	continuous	alpha	xmin	logLik	KS.stat	KS.p
DpnII	TRUE	1.813011	16	-83.35353	0.04528414	1
MboI	TRUE	1.790413	24	-352.0617	0.03109884	1



Summary

The power-law seems to fit the data fairly well over the varying resolutions once the outliers at the furthest distances are removed. The estimates of α range from 1.73 to 2.07.

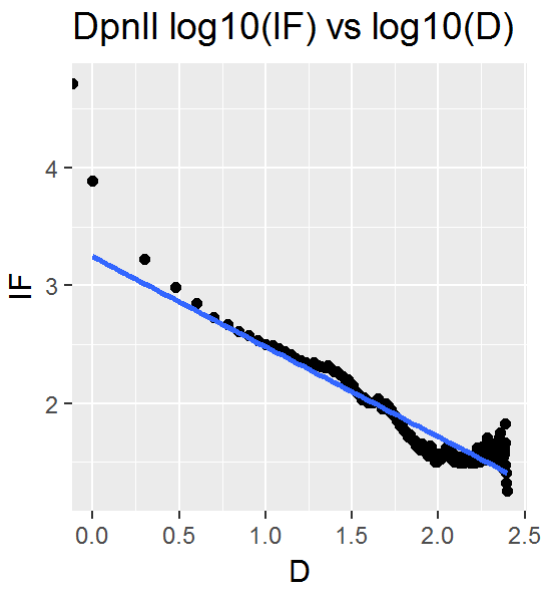
Effect of Chromosome

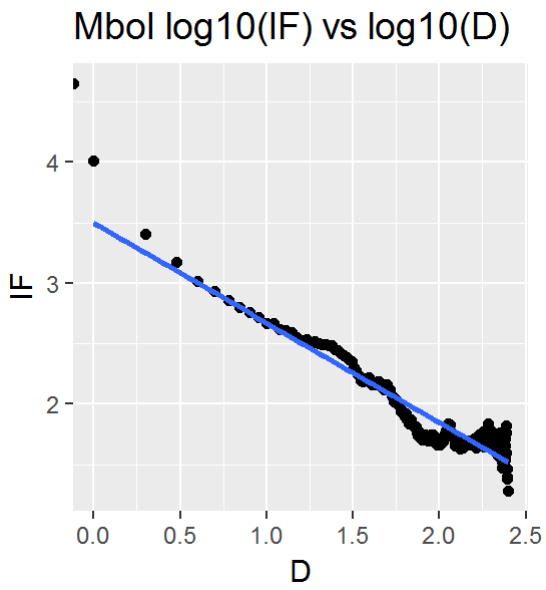
Here the fit to the power-law of median IF at each distance is tested for chromosome 1, 18, 19 from GM12878 cell line using cutting enzymes DpnII and MboI at 1MB resolution.

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(median IF) and log10(Distance), one plot for each cutting enzyme.

Chr 1

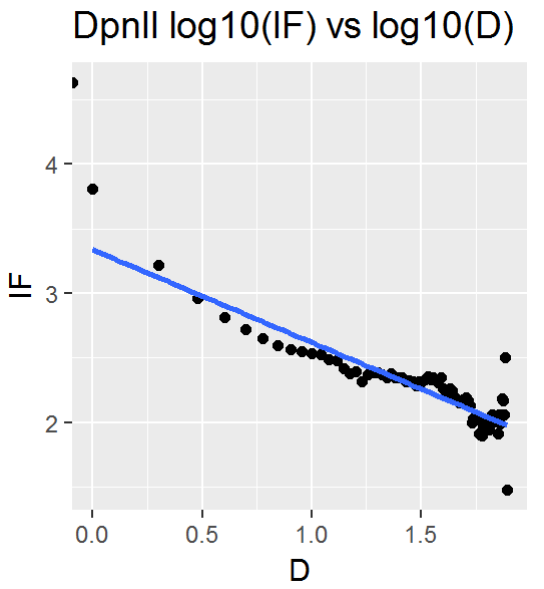
Enzyme	continuous	alpha	xmin	logLik	KS.stat	KS.p
DpnII	TRUE	2.051869	83	-347.5429	0.06175009	0.9832048
MboI	TRUE	2.053774	130.5844	-345.1992	0.08307155	0.8655548

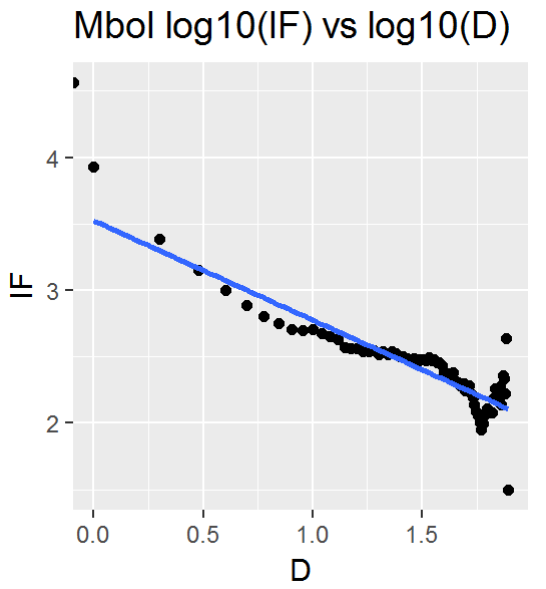




Chr 18

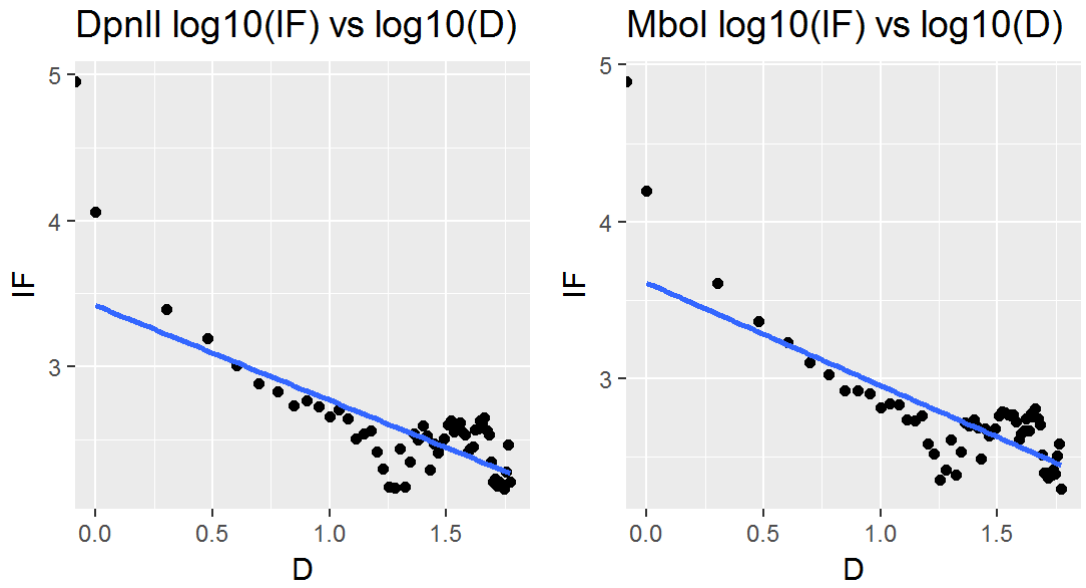
Enzyme	continuous	alpha	xmin	logLik	KS.stat	KS.p
DpnII	TRUE	1.861295	352.5	-81.75412	0.1185144	0.9978284
MboI	TRUE	1.814663	510.7638	-78.0155	0.1201584	0.9987165





Chr 19

Enzyme	continuous	alpha	xmin	logLik	KS.stat	KS.p
DpnII	TRUE	1.745079	538	-80.31837	0.1201525	0.9987176
MboI	TRUE	1.701832	841.1262	-76.10917	0.1222033	0.9992946



Summary

The power-law fits the data over varying chromosomes fairly well after the outliers are removed. α ranges from 1.7 to 2.05. For the simulation methods it seems that choosing an α between 1.7 and 2.05 should be adequate.

References

Ay, Ferhat, and William S Noble. 2015. “Analysis Methods for Studying the 3D Architecture of the Genome.” *Genome Biol* 16 (September): 183. doi:10.1186/s13059-015-0745-7 (https://doi.org/10.1186/s13059-015-0745-7).

Fudenberg, Geoffrey, Maxim Imakaev, Carolyn Lu, Anton Goloborodko, Nezar Abdennur, and Leonid A Mirny. 2016. “Formation of Chromosomal Domains by Loop Extrusion.” *Cell Rep* 15 (9): 2038–49. doi:10.1016/j.celrep.2016.04.085 (https://doi.org/10.1016/j.celrep.2016.04.085).

Lieberman-Aiden, Erez, Nynke L van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. 2009. “Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome.” *Science* 326 (5950): 289–93. doi:10.1126/science.1181369 (https://doi.org/10.1126/science.1181369).

Nagano, Takashi, Csilla Várnai, Stefan Schoenfelder, Biola-Maria Javierre, Steven W Wingett, and Peter Fraser. 2015. “Comparison of Hi-c Results Using in-Solution Versus in-Nucleus Ligation.” *Genome Biol* 16 (August): 175. doi:10.1186/s13059-015-0753-7 (https://doi.org/10.1186/s13059-015-0753-7).

Rao, Suhas S P, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, et al. 2014. “A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping.” *Cell* 159 (7): 1665–80. doi:10.1016/j.cell.2014.11.021 (https://doi.org/10.1016/j.cell.2014.11.021).

Sanborn, Adrian L, Suhas S P Rao, Su-Chen Huang, Neva C Durand, Miriam H Huntley, Andrew I Jewett, Ivan D Bochkov, et al. 2015. “Chromatin Extrusion Explains Key Features of Loop and Domain Formation in Wild-Type and Engineered Genomes.” *Proc Natl Acad Sci U S A* 112 (47): E6456–65. doi:10.1073/pnas.1518552112 (https://doi.org/10.1073/pnas.1518552112).