HiCNorm: removing biases in Hi-C data via Poisson regression

Users' Guide

Ming Hu

Last update: 09.17.2012

1. Introduction

We propose a parametric model, HiCNorm, to remove systematic biases in the raw Hi-C contact maps. It relates chromatin interactions and systemic biases at the desired resolution level, resulting in a simple, yet accurate normalization procedure. Compared to the exiting Hi-C normalization method, our model has only a few parameters, is much easier to implement, can be interpreted intuitively, and achieves higher reproducibility in real Hi-C data.

2. Normalization of Hi-C *cis* contact map

2.1 Command

The command to normalize Hi-C *cis* contact map is:
R CMD BATCH rcode_glm_normalization_cis.txt

2.2 Input file format

Users need to provide two input files: the raw Hi-C *cis* contact map and the local genomic features (We also provide script to generate local genomic features).

input_hic_cis_contact_map.txt

Assume the chromosome of interest contains N loci. The input file of Hi-C *cis* contact map is a $N \times N$ symmetric matrix separated by the table delimiter. All off-diagonal numbers should be non-negative integers. All diagonal numbers should be zero. The number in the (i, j) th cell is the total number of Hi-C reads spanning the i th locus and the j th locus.

Example:

0	33	11	8	9	…
33	0	47	39	30	…
11	47	0	102	78	…
8	39	102	0	253	…
9	30	78	253	0	…
…	…	…	…	…	…

input_local_genomic_features.txt

The input file of local genomic features is a $N \times 6$ matrix separated by the table delimiter. For the i th row ($i = 1, \ldots, N$):
Column 1: chromosome name for the i th locus.
Column 2: start position for the i th locus.
Column 3: end position for the i th locus.
Column 4: effective length in the i th locus (positive real number).
Column 5: mean GC content in the i th locus (positive real number).
Column 6: mean mappability score in the i th locus (positive real number).

Example:

chr	start	end	len	gcc	map

1	1e+06	2e+06	115988	0.5368	0.8581
1	2e+06	3e+06	121123	0.5573	0.9641
1	3e+06	4e+06	129296	0.5241	0.9651
1	4e+06	5e+06	213221	0.4678	0.9561
1	5e+06	6e+06	208584	0.4673	0.9623
…	…	…	…	…	…

2.3 Output file format

The R code "rcode_glm_normalization_cis.txt" provides a normalized Hi-C *cis* contact map "**output_normalized_hic_cis_contact_map.txt**". It is a $N \times N$ symmetric matrix separated by the table delimiter. All off-diagonal numbers should be non-negative integers. All diagonal numbers should be zero. The number in the (i, j) th cell is the normalized Hi-C *cis* contact between the i th locus and the j th locus.

Example:

0	29.894	8.7905	3.2113	3.6853	…
29.894	0	32.5497	13.5672	10.6459	…
8.7905	32.5497	0	31.3023	24.4179	…
3.2113	13.5672	31.3023	0	39.7841	…
3.6853	10.6459	24.4179	39.7841	0	…
…	…	…	…	…	…

3. Normalization of Hi-C *trans* contact map

3.1 Command

The command to normalize Hi-C *trans* contact map is:
R CMD BATCH rcode_glm_normalization_trans.txt

3.2 Input file format

Users need to provide three input files: the raw Hi-C *trans* contact map between chromosome *a* and chromosome *b*, the local genomic features of chromosome *a* and the local genomic features of chromosome *b* (We also provide script to generate local genomic features).

input_hic_trans_contact_map_chr_a_chr_b.txt

Assume we are interested in the *trans* contact between chromosome *a* and chromosome *b*, and chromosome *a* and chromosome *b* contain N and M loci, respectively. $N \times M$ matrix separated by the table delimiter. All numbers should be non-negative integers. The number in the (i, j) th cell is the total number of Hi-C reads spanning the i th locus in chromosome *a* and the j th locus in chromosome *b*.

Example:

0	0	0	0	0	…
0	0	1	0	0	…
0	1	1	0	0	…
2	1	0	1	1	…
1	0	1	1	0	…
…	…	…	…	…	…

input_local_genomic_features_chr_a.txt

The input file of local genomic features is a $N \times 6$ matrix separated by the table delimiter. The format is the same as the format of "**input_local_genomic_features.txt**".

input_local_genomic_features_chr_b.txt
The input file of local genomic features is a $M \times 6$ matrix separated by the table delimiter. The format is the same as the format of "**input_local_genomic_features.txt**".

3.3 Output file format
The R code "rcode_glm_normalization_trans.txt" provides a normalized Hi-C *trans* contact map "**output_normalized_hic_trans_contact_map_chr_a_chr_b.txt**". It is a $N \times M$ matrix separated by the table delimiter. All off-diagonal numbers should be non-negative real numbers. The number in the (i, j) th cell is the normalized Hi-C *trans* contact between the i th locus in chromosome *a* and the j th locus in chromosome *b*.

Example:

0	0	0	0	0	…
0	0	1.751	0	0	…
0	1.8974	1.5778	0	0	…
1.7154	0.9442	0	1.008	0.6883	…
0.8778	0	0.8037	1.0317	0	…
…	…	…	…	…	…

4. Note
We observed strong over-dispersion in the Hi-C contact maps. In the source code, user can choose to fit the Poisson regression model or the negative binomial regression model. The negative binomial regression model fits the over-dispersed Hi-C count data better than the Poisson regression, and provides more accurate variance estimates. These two models provide similar point estimates, and therefore achieve similar Hi-C normalization results. Fitting the negative binomial regression model is slightly slower than fitting the Poisson regression model.

5. Contact
Comments, suggestions, questions are welcomed, and should be directed to Ming Hu.
Email: minghu@fas.harvard.edu.

6. Citation
If you use HiCNorm, please cite our paper:
Hu M, Deng K, Selvaraj S, Qin ZS, Ren B and Liu JS. (2012) HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics*. To appear.