scHiCNorm: A Software Package to Eliminate Systematic Biases in Single-Cell Hi-C Data
Supplementary data

1 Introduction

In this work, we present scHiCNorm, a method for eliminating systematic biases in single-cell Hi-C data, which stem from density of cutting sites with the restriction enzyme DpnII, GC content, and mappability (1). We assume that the observed count in the single-cell Hi-C contact matrix follows zero-inflated and hurdle models, and the normalized count is the observed count divided by the estimated mean of the regression model.

2 Normalization methods

We divide the individual chromosome of interest into hundreds of bins at the resolution of 1 Mb. we create a contact matrix $C = \{c_{ij}\}_{1 \leq i,j \leq n}$, where $c_{ij}$ is the number of Hi-C counts observed between the $i$th and $j$th bins, and $n$ is the number of bins in the current chromosome.

2.1 Generating local features

The three local features are generated for each bin in the chromosome. The density of cutting sites with the restriction enzyme DpnII for each bin is the number of cutting sites within the bin. The restriction enzyme DpnII splits the whole chromosome into fragments with different lengths. Each fragment has two ends; the fragment ends are defined differently when calculating GC content and mappability. GC content of a bin is the mean of all GC contents of fragment ends within the bin, and the fragment ends defined here are the 200-bp window from the cutting site toward the fragment (2). Mappability of a bin is the mean of all mappability scores of fragment ends within the bin, but the window length is 500 bp (1)(2). The mappability score of each fragment end is the fraction of artificial reads uniquely mapped onto the fragment end; the artificial reads are obtained by splitting the whole genome sequence into 50-bp subsequences starting by every 10 bp, and are mapped back to the genome using BWA (3).

2.2 Poisson and Negative Binomial models

Here we briefly describe the methods used in HiCNorm (4). It assumes that the Hi-C contact counts in $C$ follow Poisson and Negative Binomial (NB) distributions separately. When we assume $c_{ij}$ follows Poisson distribution, the log-linear relationship between mean ($\mu_{ij}$) and linear variables ($x, y, z$) is:

$$\log(\mu_{ij}) = \beta_0 + \beta_{density}\log(x_i x_j) + \beta_{GC}\log(y_i y_j) + \log(z_i z_j)$$

(1)

where $\beta_0$ is the intercept term; $\beta_{density}$ and $\beta_{GC}$ refer to the density and GC content biases, respectively; $x_i$, $y_i$, and $z_i$ represent the features of density, GC content, and mappability respectively;
\( \log(z_i z_j) \) is the offset term. The normalized \( c_{ij} (\text{HiCNorm}(c_{ij})) \) is equal to observed count \( (c_{ij}) \) divided by expected count \( (\text{estimated mean } \mu_{ij}) \):

\[
\text{HiCNorm}(c_{ij}) = \frac{\text{observed}}{\text{expected}} = \frac{c_{ij}}{\mu_{ij}}
\]

where \( \mu_{ij} \) is the estimated mean \( \mu_{ij} \) from the fitted Poisson regression model.

When we assume \( c_{ij} \) follows NB distribution with mean \( \mu \) and variance \( \mu + \alpha \mu^2 \). The dispersion parameter \( \alpha \) is to measure the magnitude of over-dispersion. Therefore, if the data we are trying to model exhibit over-dispersion, Poisson regression model cannot fit the data well, but NB model can explain the data better than Poisson when \( \alpha > 0 \) is significant. The log-linear relationship of NB regression model is the same as that of Poisson model (Eq. 1).

### 2.3 Zero-inflated models

When the count data of interest have excess zeros, this may result in over-dispersion, which means that NB will be better than Poisson in terms of fitting count data. However, NB is not designed for dealing with excess zero observations, and cannot fully explain the excess zeros. Therefore, here we use zero-inflated models (i.e., ZIP for Zero-inflated Poisson and ZINB for Zero-inflated Negative Binomial), which are designed for dealing with excess zero counts (4). And the zero counts come from two populations: always zero set and Poisson distribution. ZIP regression model defines the independent responses \( C = (C_{12}, C_{13}, \ldots, C_{n-1,n}) \):

\[
C_{ij} \sim \begin{cases} 
0 & \text{with probability } p_{ij}, \\
\text{Poisson}(\lambda_{ij}) & \text{with probability } 1 - p_{ij}.
\end{cases} \tag{3}
\]

Thus, the ZIP distribution can be defined as:

\[
P(C_{ij} = c_{ij}) = \begin{cases} 
p_{ij} + (1 - p_{ij}) e^{-\lambda_{ij}} & c_{ij} = 0, \\
(1 - p_{ij}) \frac{e^{-\lambda_{ij}} \lambda_{ij}^{c_{ij}}}{c_{ij}!} & c_{ij} > 0.
\end{cases} \tag{4}
\]

Moreover, the relationship between Poisson parameter \( \lambda_{ij} \) and the three local features is given:

\[
\log(\lambda_{ij}) = \beta_0 + \beta_{\text{density}} \log(x_i x_j) + \beta_{\text{GC}} \log(y_i y_j) + \log(z_i z_j) \tag{5}
\]

where \( \beta_0 \) is the intercept term; \( \beta_{\text{density}} \) and \( \beta_{\text{GC}} \) refer to the density and GC content biases, respectively; \( x_i, y_i, \) and \( z_i \) represent the features of density, GC content, and mappability respectively; \( \log(z_i z_j) \) is the offset term. And \( p_{ij} \) satisfies:

\[
\logit(p_{ij}) = \log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \gamma_0 + \gamma_{\text{density}} \log(x_i x_j) + \gamma_{\text{GC}} \log(y_i y_j) + \gamma_{\text{map}} \log(z_i z_j) \tag{6}
\]

where \( \logit(p_{ij}) \) is the logit link function in binomial Generalized Linear Models (GLMs); \( \gamma_0 \) is the intercept term; \( \gamma_{\text{GC}}, \gamma_{\text{density}}, \) and \( \gamma_{\text{map}} \) are the density, GC content, and mappability biases respectively.

The mean \( \mu_{ij} \) of ZIP regression model is:

\[
\mu_{ij} = p_{ij} \cdot 0 + (1 - p_{ij}) \cdot \lambda_{ij} \tag{7}
\]

using the canonical log link [3]. ZINB model has the similar interpretation as ZIP model, but has the dispersion parameter \( \alpha \) and allows for over-dispersion.
2.4 Hurdle models

Hurdle models (i.e., PH for Poisson Hurdle and NBH for Negative Binomial Hurdle) are two-part models, which assumes that zeros and positive count data come from different processes. The first part models binary responses: zero and positive count data via density function $f_{\text{zero}}(c; x, y, z, \gamma)$, which can be a binomial model with logit link function. The second part models truncated count data ($>0$) via density function $f_{\text{count}}(c; x, y, z, \beta)$, which can be Poisson or NB models. The hurdle distribution can be defined as:

$$f_{\text{hurdle}}(c; x, y, z, \beta, \gamma) = \begin{cases} f_{\text{zero}}(0; x, y, z, \gamma) & c = 0, \\ 1 - f_{\text{count}}(0; x, y, z, \beta) & c > 0. \end{cases}$$ (8)

Specially, PH distribution can be described as:

$$P(C_{ij} = c_{ij}) = \begin{cases} p_{ij} & c_{ij} = 0, \\ 1 - p_{ij} e^{-\lambda_{ij}} \frac{\lambda_{ij}^{c_{ij}}}{c_{ij}!} & c_{ij} > 0. \end{cases}$$ (9)

The mean $\mu_{ij}$ of PH regression model is given by:

$$\mu_{ij} = \frac{1 - p_{ij}}{1 - e^{-\lambda_{ij}}} \lambda_{ij}$$ (10)

The truncated Poisson parameter $\lambda_{ij}$ and the probability $p_{ij}$ of $c_{ij} = 0$ have the relationships with the three local features as described in Equations 5 and 6, respectively. Also, NBH model can be interpreted as PH model, but has the dispersion parameter.

After the ZIP, ZINB, PH, and NBH models being fitted successfully, we can get the estimated $\mu'_{ij}$, and the normalized Hi-C contact count of $c_{ij}$ is given by:

$$\text{scHiCNorm}(c_{ij}) = \frac{\text{observed}}{\text{expected}} = \frac{c_{ij}}{\mu_{ij}}.$$ (11)

2.5 Model selection

We use likelihood ratio test, Vuong test, and Akaike information criterion (AIC) to compare the goodness of fit between the six regression models, and conclude that NBH achieves the best performance as shown in Table S1 for 74 human cells.

2.6 Non-parametric method

We use non-parametric (NP) regression to predict expected Hi-C contacts. The variates are still the three bias features. The kernel regression estimator we used here is local-linear; and Kullback Leibler cross-validation is used to select bandwidths.

2.7 Matrix-balancing methods

We evaluate three matrix-balancing methods, ICE, VC, and KR, that have been previously used to normalize massive-cell Hi-C data in the literature. To test whether matrix-balancing methods can eliminate single-cell systematic biases as efficiently as bias-based methods (i.e., HiC-Norm and scHiCNorm), we implement these three methods in R and compare their performance with bias-based methods (see Supplementary results). Our software package also includes the R implementations of these three matrix-balancing methods.
3 Results and discussion

3.1 Bias existence

To test whether the three local features (i.e., density, GC content, and mappability) directly affect the enrichment of Hi-C contacts, we perform the same benchmarking as described in (1)(2): (1) calculating the three local features for all of 1 Mb bins in the whole genome; (2) for each feature, stratifying the bins into 20 sets based on their feature values; (3) calculating the average number of Hi-C contact counts between each pair of bins belonging to different sets. We test this on two Hi-C data sets: all 74 cells (Figure S1) and one single cell (Figures S2 and S3). The results indicate that both Hi-C data sets have clear biases for all of the three local features.

3.2 Evaluation of normalization

To evaluate the efficiency of eliminating biases, we calculate the Pearson’s correlation coefficient between Hi-C data and the three local features. A good normalization methods should make the normalized Hi-C data have a smaller Pearson’s correlation coefficient (i.e., close to zero) with the three local features (i.e., less bias). We test ten methods, including HiCNorm (i.e., Poisson and NB), schiCNorm (i.e., ZIP, ZINB, PH, and NBH), three matrix-balancing methods (i.e., ICE, VC, and KR), and non-parametric regression. The results in Figure S4 show that NBH achieves the best performance on 74 cells. We compare the computational cost (i.e., time cost and memory usage) of the ten methods except ICE and NP, because they are too time-consuming compared with other methods. Figure S5 shows time cost and memory usage when we try to normalize chromosome 1 of the 74 cells, from which we can conclude that VC and KR are faster and take smaller amount of memory. The time and memory usage data are obtained on our own shared memory Linux server equipped with 4 CPUs (each with 12 cores) with speed 2500 MHz and 378 gigabytes of memory.

To confirm that our methods can be used to normalize Hi-C data for a broad range of single cells, we generate another cell set by randomly selecting 200 cells from ML1 in (12). The model selection results for the 200 cells can be found in Table S2; and the evaluation results for normalization are shown in Figure S6. We can make the same conclusion as for with 74 cells, indicating that our methods can be applied for general single cells.

3.3 Cell-to-cell variances

We next explore whether the normalized Hi-C data contribute to cell-to-cell variances in contact maps. We calculate the variances for different bin separations (i.e., 2, 3, 4, 6, 9, 14, 22, 35, 56, 91, 149, and 247). The contact probability for each bin separation is calculated as follows: (1) calculating the cumulative Hi-C contacts between two consecutive bin separations (e.g., for the bin separation of 9, 9 ≤ |i − j| < 14); (2) the cumulative Hi-C contacts are first normalized by the length between two consecutive bin separations (e.g. the length between 9 and 14 is 5), and then further normalized by the sum of all cumulative Hi-C contacts between any of two consecutive bin separations.

The results shown in Figure S7 indicate that the variances of normalized Hi-C data are larger than those of raw Hi-C data. Notice that for population Hi-C method (e.g., based on 10^5 cells), removing bias usually decreases the variance in Hi-C contacts. This is because the normalization method removes outliers and makes the data more consensus and unified. However, we think it is opposite for the single-cell Hi-C method because one of the purposes of single-cell Hi-C method is to reveal the cell-to-cell variance. All the three local features (i.e., density, GC content, and mappability) are calculated based on the same reference genome. Therefore, all of the cells in the single-cell Hi-C experiment share the same bias, which uncover their cell-to-cell variances. The single-cell normalization method, that removes the bias, therefore helps reveal the cell-to-cell variance, a
contribution of the single-cell normalization algorithm. We think this is why after normalization the cell-to-cell variance is higher.

Furthermore, we randomly select 200 pairs of cells and calculate the Pearson’s correlations between each pair’s raw Hi-C matrices and normalized Hi-C matrices by ten methods. Figure S8 shows the distribution of the correlation values. The average correlation between the raw Hi-C matrices is high as expected. However, ICE-normalized and KR-normalized Hi-C data are similarity high (> 0.75) indicating bias is not removed (cell-to-cell variance is not observable). In contrast, the normalization by HiCNorm and particularly our scHiCNorm successfully reduce the correlation to about 0.5, indicating they help reveal cell-to-cell variance. Because the Hi-C contact map is eventually indicating the three-dimensional chromosome structure, this Hi-C matrix variance also indicate the variance of 3D structural conformation between different single cells.

Figures and Tables

Table S1: Model selection and goodness of fit among six different models using likelihood ratio test, Vuong statistics, and Akaikes information criterion (AIC) for 74 human cells

| Model1 | Model2 | Test of significant (Model2 > Model1) | Methods | Results | P values (<0.05) | AIC |
|--------|--------|---------------------------------------|---------|---------|-----------------|-----|
| Poisson | NB     | Likelihood ratio 99.94%                | 100%    | 100%    |                  | 100%|
| ZIP    | ZINB   | Likelihood ratio 99.88%                | 98.52%  | 98.88%  |                  |     |
| PH     | NBH    | Likelihood ratio 99.88%                | 98.57%  | 98.88%  |                  |     |
| Poisson | ZIP    | Vuong 100%                             | 100%    | 100%    |                  | 100%|
| NB     | ZINB   | Vuong 97.81%                           | 32.12%  | 30.53%  |                  |     |
| Poisson | PH     | Vuong 100%                             | 100%    | 100%    |                  | 100%|
| NB     | NBH    | Vuong 100%                             | 69.19%  | 67.81%  |                  |     |
| ZIP    | PH     | Vuong 89.05%                           | 20.07%  | 89.05%  |                  |     |
| ZINB   | NBH    | Vuong 89.59%                           | 60.56%  | 89.59%  |                  |     |

Table S2: Model selection and goodness of fit among six different models using likelihood ratio test, Vuong statistics, and Akaikes information criterion (AIC) for 200 human and mouse cells

| Model1 | Model2 | Test of significant (Model2 > Model1) | Methods | Results | P values (<0.05) | AIC |
|--------|--------|---------------------------------------|---------|---------|-----------------|-----|
| Poisson | NB     | Likelihood ratio 91.56%                | 94.08%  | 89.73%  |                  |     |
| ZIP    | ZINB   | Likelihood ratio 68.01%                | 50.78%  | 41.63%  |                  |     |
| PH     | NBH    | Likelihood ratio 63.68%                | 48.56%  | 37.95%  |                  |     |
| Poisson | ZIP    | Vuong 99.90%                           | 78.28%  | 87.41%  |                  |     |
| NB     | ZINB   | Vuong 99.77%                           | 26.07%  | 22.42%  |                  |     |
| Poisson | PH     | Vuong 100%                             | 72.83%  | 85.63%  |                  |     |
| NB     | NBH    | Vuong 100%                             | 16.09%  | 19.87%  |                  |     |
| ZIP    | PH     | Vuong 70.86%                           | 4.87%   | 70.84%  |                  |     |
| ZINB   | NBH    | Vuong 66.78%                           | 7.08%   | 89.59%  |                  |     |
Figure S1: Hi-C data biases for Hi-C data from an ensemble of 74 cells.

Figure S2: Hi-C data biases for Hi-C data of one single cell with barcodes TGAGCCAC-AGCATATG: raw vs normalized. The normalization models are Poisson-based: d, e, and f for Poisson; g, h, and i for ZIP; and j, k, and l for PH.
Figure S3: Hi-C data biases for Hi-C data of one single cell with barcodes TGAGCCAC-AGCATATG: raw vs normalized. The normalization models are NB-based: d, e, and f for NB; g, h, and i for ZINB; and j, k, and l for NBH.
Figure S4: Pearson’s correlation coefficients between Hi-C data (raw vs normalized) and the three local features for 74 human cells. The normalization methods we test are HiCNorm, scHiCNorm, ICE, VC, KR, and NP.

Figure S5: Time cost and memory usage for normalizing Chromosome 1 of 74 human cells at the resolution 1Mb.
Figure S6: Pearson’s correlation coefficients between Hi-C data (raw vs normalized) and the three local features for 200 human and mouse cells. The normalization methods we test are HiCNorm and scHiCNorm.
Figure S7: With the increase of bin distances, the average contact probability with standard deviation keeps decreasing. The normalized Hi-C data by the four methods (i.e., ZIP, ZINB, PH, and NBH) achieve larger standard deviations (variances) than raw Hi-C data.

Figure S8: The Pearson’s correlation between randomly selected 200 pairs of cells. The raw Hi-C data for each pair are very similar. However, the normalized Hi-C data indicate cell-to-cell variances, especially for HiCNorm and scHiCNorm methods.
References

[1] E. Yaffe and A. Tanay. Probabilistic modeling of hi-c contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat Genet, 43(11):1059–65, 2011. ISSN 1546-1718 (Electronic) 1061-4036 (Linking). doi: 10.1038/ng.947. URL http://www.ncbi.nlm.nih.gov/pubmed/22001755

[2] M. Hu, K. Deng, S. Selvaraj, Z. Qin, B. Ren, and J. S. Liu. Hicnorm: removing biases in hi-c data via poisson regression. Bioinformatics, 28(23):3131–3, 2012. ISSN 1367-4811 (Electronic) 1367-4803 (Linking). doi: 10.1093/bioinformatics/bts570. URL http://www.ncbi.nlm.nih.gov/pubmed/23023982

[3] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrowswheeler transform. Bioinformatics, 25(14):1754–1760, 2009. ISSN 1546-1718.

[4] Diane Lambert. Zero-inflated poisson regression, with an application to defects in manufacturing. Technometrics, 34(1):1–14, 1992. ISSN 0040-1706.

[5] Achim Zeileis, Christian Kleiber, and Simon Jackman. Regression models for count data in r. Journal of statistical software, 27(8):1–25, 2008. ISSN 1548-7660.

[6] John Mullahy. Specification and testing of some modified count data models. Journal of econometrics, 33(3):341–365, 1986. ISSN 0304-4076.

[7] Tristen Hayfield and Jeffrey S Racine. Nonparametric econometrics: The np package. Journal of statistical software, 27(5):1–32, 2008.

[8] Maxim Imakaev, Geoffrey Fudenberg, Rachel Patton McCord, Natalia Naumova, Anton Goloborodko, Bryan R Lajoie, Job Dekker, and Leonid A Mirny. Iterative correction of hi-c data reveals hallmarks of chromosome organization. Nature methods, 9(10):999–1003, 2012. ISSN 1548-7091.

[9] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Guirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science, 326(5950):289–93, 2009. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi: 10.1126/science.1181369. URL http://www.ncbi.nlm.nih.gov/pubmed/19815776

[10] Philip A Knight and Daniel Ruiz. A fast algorithm for matrix balancing. IMA Journal of Numerical Analysis, 33(3):1029–1047, 2013. ISSN 0272-4979.

[11] Suhas SP Rao, Miriam H Huntley, Neva C Durand, Elena K Stanenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, and Eric S Lander. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell, 159(7):1665–1680, 2014. ISSN 0092-8674. URL http://ac.els-cdn.com/S0092867414014974/1-s2.0-S0092867414014974-main.pdf?_tid=b9e4b50a-8b1f-11e6-a058-00000aab0f26&acdnat=1475688886_c50f4f613587bd486467fbeeaf8f18f457

[12] V. Ramani, X. Deng, R. Qiu, K. L. Gunderson, F. J. Steemers, C. M. Disteche, W. S. Noble, Z. Duan, and J. Shendure. Massively multiplex single-cell hi-c. Nat Methods, 14(3):263–266, 2017. ISSN 1548-7105 (Electronic) 1548-7091 (Linking). doi: 10.1038/nmeth.4155. URL http://www.ncbi.nlm.nih.gov/pubmed/28135255