Bioinformatics, 36(9), 2020, 2902–2904
doi: 10.1093/bioinformatics/btaa048
Advance Access Publication Date: 31 January 2020
Applications Note

Genome analysis

GiniQC: a measure for quantifying noise in single-cell Hi-C data

Connor A. Horton†, Burak H. Alver and Peter J. Park*

Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA

*To whom correspondence should be addressed.
†Present address: Department of Genetics, Stanford University, Stanford, CA 94305, USA
Associate Editor: Alfonso Valencia

Received on July 22, 2019; revised on January 6, 2020; editorial decision on January 19, 2019; accepted on January 24, 2019

Abstract

Summary: Single-cell Hi-C (scHi-C) allows the study of cell-to-cell variability in chromatin structure and dynamics. However, the high level of noise inherent in current scHi-C protocols necessitates careful assessment of data quality before biological conclusions can be drawn. Here, we present GiniQC, which quantifies unevenness in the distribution of inter-chromosomal reads in the scHi-C contact matrix to measure the level of noise. Our examples show the utility of GiniQC in assessing the quality of scHi-C data as a complement to existing quality control measures. We also demonstrate how GiniQC can help inform the impact of various data processing steps on data quality.

Availability and implementation: Source code and documentation are freely available at https://github.com/4dn-dcic/GiniQC.

Contact: peter_park@hms.harvard.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Recent advances in sequencing- and microscopy-based assays have dramatically increased our ability to probe chromatin structure and dynamics. Chromatin conformation capture methods, for example, have revealed variability in nuclear compartmentalization and topologically associated domains (Rowley and Corces, 2018). Among these, single-cell Hi-C (scHi-C), a method based on high-throughput chromatin conformation capture at the single cell level, has been applied to assess cell-to-cell variability in chromatin structure across the cell cycle and the oocyte-to-zygote transition (Flyamer *et al.*, 2017; Nagano *et al.*, 2017). However, despite the potential of scHi-C to probe chromatin structure of individual cells in high-throughput, the high level of noise inherent in sequencing-based single-cell assays remains a barrier to wider adoption.

Here, we present GiniQC, a novel quality control measure designed to quantify noise in scHi-C data. In designing our measure, we employed the observation, corroborated by decades of microscopy data, that each chromosome contacts a limited number of other chromosomes at a given moment (Cremer and Cremer, 2001). Consequently, GiniQC measures a different aspect of data quality than measured by the percentage of contacts in *cis*, a commonly used measure of scHi-C data quality. Our tool calculates GiniQC as well as other frequently used scHi-C quality control metrics, including percentage of contacts in *cis*.

2 Materials and methods

The tendency of chromosomes to occupy distinct territories in the nucleus has been observed in previous scHi-C studies as distinct, well-supported clusters of *trans* contacts (Lando *et al.*, 2018; Nagano *et al.*, 2017). In contrast, random ligation noise is distributed more uniformly throughout the genome. In Figure 1A, contact maps from two datasets are plotted on the same matrix: the upper triangle is characterized by 'clumps' of reads, thus representing higher quality data, than the lower triangle, which does not have discrete clusters of *trans* contacts.

To quantify these observations, we begin by counting the number of reads that support a contact between two given inter-chromosomal, or *trans*, regions. When these read counts vary greatly between bins, such as in the upper triangle of Figure 1A, they reflect the tendency of chromosomes to contact a limited number of other chromosomes. When these read counts are similar across *trans* bins, such as in the lower triangle of Figure 1A, we consider the data to be noisy and less representative of actual chromatin structure. In our distribution of read counts, we exclude *cis* (intra-chromosomal) bins, which contain far greater reads than *trans* bins, in order to resolve smaller differences between *trans* bins.

To measure the clumping of reads, we used the Gini coefficient. Commonly used as a measure of economic inequality (Gini, 1912), the Gini coefficient allows one to quantify the inequality in the

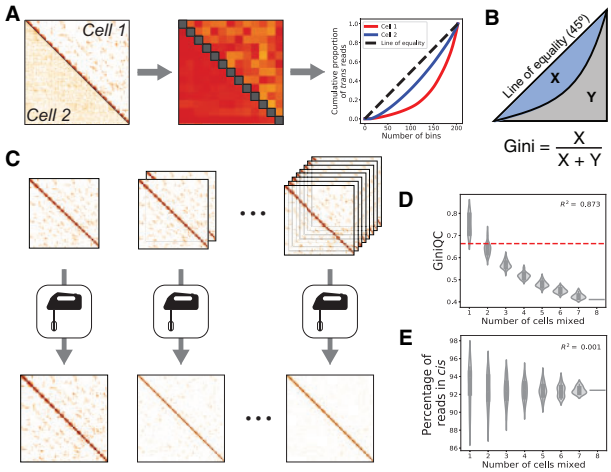

Fig. 1. (**A**) Upper triangle and lower triangle are contact maps from two diploid cells from Nagano *et al.* (2017). To compute GiniQC, *cis* reads are discarded and *trans* reads are tallied by chromosomal bin. The cumulative distribution function (CDF) can be used to calculate the Gini coefficient. 1CDX1-335 (upper) has read-count-adjusted GiniQC of 0.506; 1CDU-524 (lower) has read-count-adjusted GiniQC of 0.386. (**B**) A visual demonstration of GiniQC calculation from the line of equality and the CDF. (**C**) To test GiniQC and the percentage of reads in *cis*, we selected 50 000 reads from every combination of 1, 2, … 8 cells from Stevens *et al.* (2017) and then computed these QC measures. (**D**) GiniQC versus of the number of cells from which 50 000 reads are mixed together. Dashed line shows 90th percentile of GiniQC values from mixing two cells. (**E**) Percentage of contacts in *cis* versus the number of cells from which 50 000 reads are mixed together

distribution of *trans* contacts and, by extension, the amount of noise. Specifically, our tool takes a Hi-C contact matrix and normalizes read counts by iterative correction to account for sampling bias at the experimental or sequencing stages (Imakaev *et al.*, 2012). After *cis* bins are discarded, the list of normalized read counts per bin is then sorted. This sorted list can be used to calculate the Gini coefficient (G) using the following formula (adapted from Sen, 1973; further detail in Supplementary Material):

$$G = \frac{\sum_{i=1}^{n} (2i - n - 1)x_i}{n \sum_{i=1}^{n} x_i},$$

where x_i is the read count per bin and n is the total number of 2D bins.

We call this specific application of the Gini coefficient as GiniQC. A higher value of GiniQC, approaching 1, corresponds to an unequal distribution of reads per bin, while a lower value of GiniQC, approaching 0, corresponds to a more uniform distribution of reads per bin. Higher quality data are associated with higher values of GiniQC. Due to a systematic bias in calculated values of the Gini coefficient associated with the number of reads, we adjust the value to account for this correlation (Supplementary Material; Supplementary Fig. S2).

3 Results and comparison to existing metrics

As one example of GiniQC identifying noisy data, the GiniQC values are 0.506 and 0.386 for the upper and lower triangles in Figure 1A, respectively. To systematically test the ability of GiniQC to quantify signal and noise, we simulate low-quality data by mixing together reads from different cells (Fig. 1C; Supplementary Material). This particular approach is meant to represent suspected sources of noise in the scHi-C protocol—such as nuclear rupture during sample preparation or misapplication of barcodes during library preparation or sequencing—that could lead to mixing reads from different cells. When this is applied to the eight haploid cells from Stevens *et al.* (2017), we find that GiniQC decreases with the number of cells mixed (Fig. 1D). We arbitrarily use the 90% quantile of the distribution of GiniQC values from mixing two cells to identify a quality control threshold in our tool (Fig. 1D).

We compare this with the percentage of reads in *cis*, the most commonly used metric for scHi-C data, using the same cell-mixing procedure (Fig. 1E). We find that the percentage of reads in *cis* does not decrease as reads from different cells are mixed, which suggests that percentage of reads in *cis* is less suited to detect certain kinds of noise. We also note that using scHi-C contact matrices to model genome structures has been proposed as a method for quality control in the literature (Lando *et al.*, 2018). While this approach largely aims to address a similar question of whether sequencing reads support plausible clusters of contacts, it is far more computationally expensive (Supplementary Material). Finally, we compare GiniQC with QuASAR (Sauria and Taylor, 2017; Supplementary Figs S5 and S6).

Another source of noise in scHi-C data is the arbitrary fluctuations in sequencing coverage by chromosome, which could bias GiniQC values. We address this challenge by applying iterative correction to our contact matrices and by calculating the maximum deviation in chromosomal coverage from the median chromosome as an additional measure of data quality.

GiniQC can also be used to assess modifications to the data processing pipeline. To validate this approach, we compare GiniQC values when multimapping reads are discarded to GiniQC values when multimapping reads are retained. As expected, we find that the inclusion of multimapping reads, whose ambiguity in alignment should increase noise, is associated with lower GiniQC values (Supplementary Fig. S3). We extend this approach to assess whether discarding contacts observed only once ('singletons') improves data quality, a data filter that has been debated in the scHi-C literature (Lando *et al.*, 2018; Nagano *et al.*, 2017; Stevens *et al.*, 2017). We find that discarding these reads marginally improves data quality, but that the size of the effect depends on the dataset (Supplementary Fig. S3).

Finally, we assess whether GiniQC has any association with cell cycle, ploidy, or percentage of contacts in *cis*. We find that higher values of GiniQC, indicating higher quality data, are only marginally associated with a higher percentage of reads in *cis* (Supplementary Fig. S4A), suggesting that the two quantities provide non-overlapping information. We find no correlation with ploidy and a negligible correlation with cell cycle (Supplementary Fig. S4B–D).

Our results demonstrate that GiniQC is a useful measure of data quality. It complements other measures such as the percentage of *cis* contacts, as each captures a separate aspect of data quality and as no single metric can sufficiently describe a complex dataset.

4 Implementation

GiniQC is implemented in Python. GiniQC generates the total number of reads, percentage of reads in *cis*, raw GiniQC value, GiniQC value adjusted for read counts, and maximum chromosomal coverage aberration for each cell. When a list of Cooler files is passed to GiniQC, the application produces a tab-separated values table of the metrics listed above and calculates a suggested data quality threshold for GiniQC.

Funding

This work was supported by the National Institutes of Health Common Fund 4D Nucleome Project [U01CA200059]; C.A.H. acknowledges support from the Harvard College Research Program and the Pechet Family Research Fund.

Conflict of Interest: none declared.

References

Cremer,T. and Cremer,C. (2001) Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat. Rev. Genet.*, **2**, 292–301.

Flyamer,I.M. *et al*. (2017) Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature*, **544**, 110–114.

Gini,C. (1912) Variabilità e Mutuabilità. In: Pizetti, E. and Salvemini, T. (eds) *Memorie di Metodologica Statistica* (Reprinted). Libreria Eredi Virgilio Veschi, Rome.

Imakaev,M. *et al*. (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods*, **9**, 999–1003.

Lando,D. *et al*. (2018) Calculation of 3D genome structures for comparison of chromosome conformation capture experiments with microscopy: an evaluation of single-cell Hi-C protocols. *Nucleus*, **9**, 190–201.

Nagano,T. *et al*. (2017) Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature*, **547**, 61–67.

Rowley,M.J. and Corces,V.G. (2018) Organizational principles of 3D genome architecture. *Nat. Rev. Genet*., **19**, 789–800.

Sauria,M.E.G. and Taylor,J. (2017) QuASAR: Quality Assessment of Spatial Arrangement Reproducibility in Hi-C Data. *bioRxiv*, 204438.

Sen,A. (1973) *On Economic Inequality*. Oxford: Clarendon Press.

Stevens,T.J. *et al*. (2017) 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature*, **544**, 59–64.