doi: 10.1093/nar/gkaa030

IDR2D identifies reproducible genomic interactions

Konstantin Krismer[1,2], Yuchun Guo[1] and David K. Gifford[1,2,3,*]

[1]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 32 Vassar Street, Cambridge, MA 02139, USA, [2]Department of Biological Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA and [3]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

Received September 23, 2019; Revised December 19, 2019; Editorial Decision January 09, 2020; Accepted January 22, 2020

ABSTRACT

Chromatin interaction data from protocols such as ChIA-PET, HiChIP and Hi-C provide valuable insights into genome organization and gene regulation, but can include spurious interactions that do not reflect underlying genome biology. We introduce an extension of the Irreproducible Discovery Rate (IDR) method called IDR2D that identifies replicable interactions shared by chromatin interaction experiments. IDR2D provides a principled set of interactions and eliminates artifacts from single experiments. The method is available as a Bioconductor package for the R community, as well as an online service at https://idr2d.mit.edu.

INTRODUCTION

The Irreproducible Discovery Rate (1) (IDR) method identifies a robust set of findings that comprise the signal component shared by two replicate experiments. The IDR method is akin to the false discovery rate (FDR) in multiple hypothesis testing, but instead of alleviating alpha error accumulation within one replicate, IDR quantifies the reproducibility of findings using a copula mixture model with one reproducible and one irreproducible component. A finding's IDR is the probability it belongs to the irreproducible component. This permits findings that are likely in the irreproducible noise component to be eliminated for subsequent analyses. Assessing the IDR of genomic findings has been adopted by ENCODE (2), and is recommended for all ChIP-seq experiments with replicates (3). IDR has also been used in numerous projects outside of ENCODE (4–8).

Chromatin interaction experiments such as ChIA-PET (9), HiChIP (10) and Hi-C (11) provide important chromatin structure and gene regulation information, but the complexity of their results and the sampling noise present in their protocols makes the principled analysis of resulting data important. Single replicate false discovery rate (FDR) methods are often used to identify chromatin interactions, but questions can remain about the veracity of the interactions identified as significant as they may not be observed in replicate experiments.

Here, we generalize IDR from one dimensional analysis, performed on a single genome coordinate, to the analysis of interactions that are identified in two dimensions by a pair of genome coordinates. We call this extended method Irreproducible Discovery Rate for Two Dimensions (IDR2D) and it can be readily applied to any experimental data type that produces two-dimensional genomic results that admit appropriate distance metrics. We demonstrate the application of IDR2D to data from ChIA-PET, HiChIP and Hi-C experiments.

Like IDR, IDR2D independently ranks the findings from each replicate. This ranking reflects the confidence of the finding, with high-confidence interactions at the top and low-confidence interactions at the bottom of the list. In a subsequent step, corresponding interactions between replicates are identified. A genomic interaction from replicate 1 is said to correspond to an interaction in replicate 2, if both their interaction anchors overlap (see Figure 1C). After corresponding interactions are identified and ambiguous mappings of interactions between replicates are resolved (see equation 1 and Figure 1D), IDRs are computed for each replicated interaction.

If interaction $x_{i,\,1}$ in replicate 1 overlaps with more than one interaction in replicate 2, the ambiguous mapping is resolved by choosing $x_{*,\,2}$ in the following way:

$$x_{*,2} = \underset{x_{j,2} \in \Omega_{x_{i,1}}}{\operatorname{argmin}} f(x_{i,1}, x_{j,2}), \qquad (1)$$

where $\Omega_{x_{i,1}}$ is the set of interactions in replicate 2 that overlaps with the interaction $x_{i,1}$ in replicate 1, and $f(\,\cdot, \cdot)$ is the *ambiguity resolution value* (ARV) between an interaction in replicate 1 and an overlapping interaction in replicate 2. Depending on the ambiguity resolution method, this value corresponds to the genomic distance between anchor midpoints (see 1 in Figure 1D), the additive inverse of the relative anchor overlap (see 2 in Figure 1D), or the sum of the interaction ranks, where more significant interactions have lower ranks.

IDR2D is used as the final step in chromatin interaction data workflows (see Figure 1A). The input to IDR2D

[*]To whom correspondence should be addressed. Tel: +1 617 253 6039; Email: gifford@mit.edu

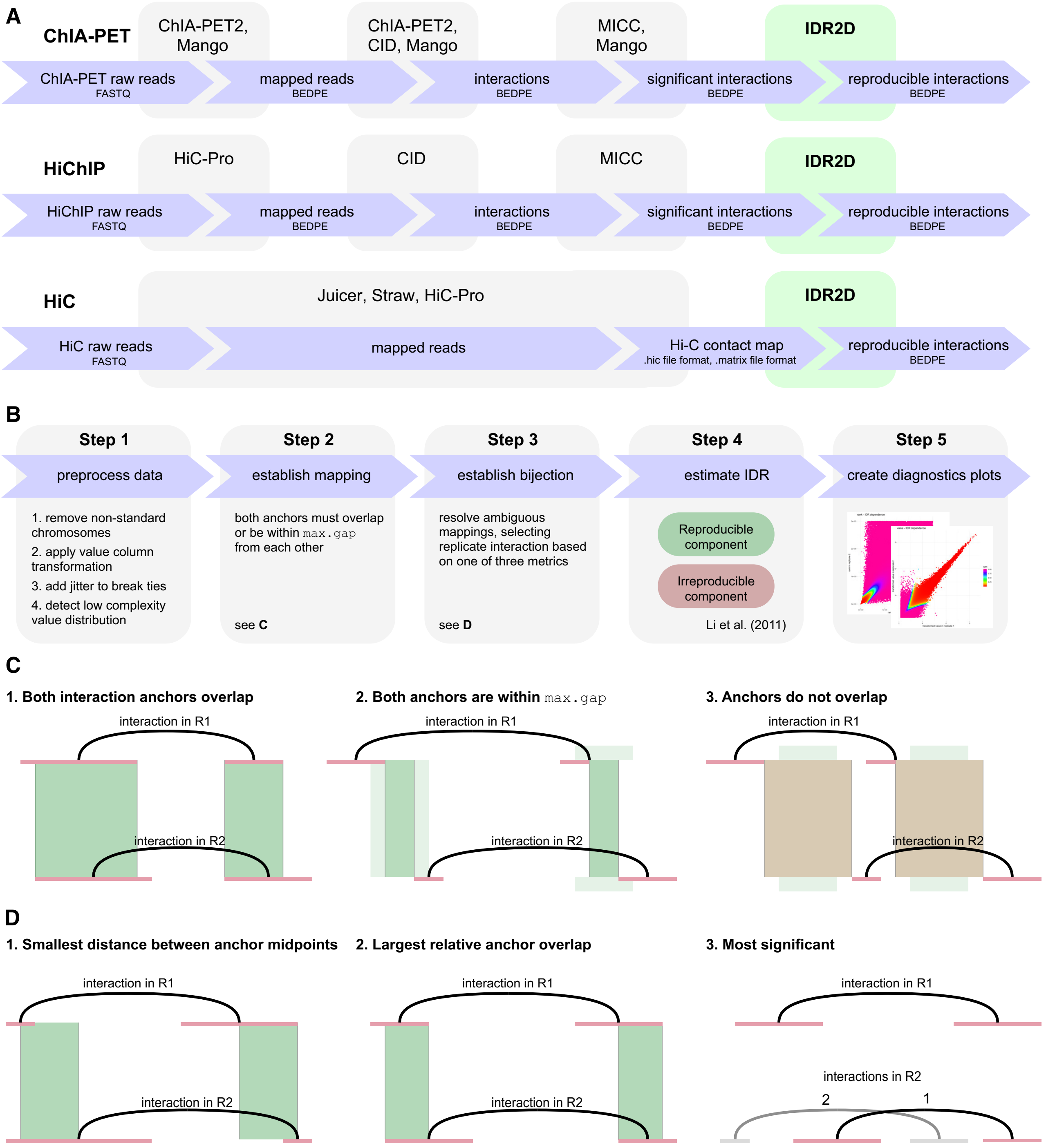

Figure 1. IDR2D identifies reproducible genomic interactions. (**A**) IDR2D is a potential post-processing step in the data analysis pipelines for ChIA-PET, HiChIP and Hi-C experiments that were done in replicate. It is compatible with a range of different interaction callers, such as ChIA-PET2, Mango, and CID. (**B**) This schematic depicts the five steps of the IDR2D procedure. In step 1, the data is prepared for IDR analysis, which includes the removal of interactions on non-standard chromosomes and a suitable transformation of the value column, which will be the basis of the ranking. In step 2, interactions in replicate 1 that overlap interactions in replicate 2 are identified, and in step 3 a one-to-one correspondence between overlapping interactions is established by resolving ambiguous cases. After this unambiguous mapping is established, in step 4 the irreproducible discovery rates are estimated for each interaction pair. Lastly, diagnostics plots are created in step 5. (**C**) An interaction in replicate 1 (R1) is assigned to all interactions in R2 for which both interaction anchors overlap or are within `maximum gap` of each other. (**D**) If more than one interaction in R2 overlaps with an interaction in R1, there are three methods to resolve this ambiguous mapping: select the interaction in R2 with (1) the smallest distance between the anchor midpoints (width of the green bars), (2) the largest relative overlap (width of the green bars divided by the sum of the anchor widths) or (3) the lowest rank sum of the interactions, which prioritizes more significant interactions.

are BEDPE formatted files of genomic interactions, where each genomic interaction has a score associated with it. This score is used to rank the interactions and can be probability-based, such as the scores from MICC-based methods (12–14), or based on a heuristic. For Hi-C experiments, IDR2D supports the .hic file format from the Juicer / Straw pipeline and the .matrix/.bed file formats from the Hi-C Pro pipeline. Figure 1B breaks the IDR2D procedure into five steps.

MATERIALS AND METHODS

IDR

IDR2D extends the reference implementation of IDR in R by Qunhua Li (1). All datasets were analyzed with the IDR2D package using default parameters. We used *overlap* as ambiguity resolution method and allowed no gaps between overlapping interactions (`max.gap = 0L`). The applied value transformations were dependent on the interaction calling method. The results were not sensitive to the initial values of the mean, standard deviation, correlation coefficient, or proportion of the reproducible component.

ChIA-PET datasets

We used 17 ChIA-PET datasets associated with protein factors that include POL2RA, CTCF and RAD21 from selected cell types (Supplementary Table S1). The datasets were downloaded from the ENCODE Project portal (https://www.encodeproject.org/). All FASTQ files of both biological replicates were pre-processed and aligned to the hg19 genome assembly using steps 1, 2 and 3 in the ChIA-PET data analysis software *Mango*.

HiChIP datasets

The FASTQ SMC1A HiChIP data from GM12878 cells (10) were downloaded from the NCBI GEO portal (GSE80820). Raw read files were analyzed with HiC-Pro (15), and interactions were subsequently called by CID and *hichipper* (16).

Hi-C datasets and subsampling procedure

Preprocessed contact matrix files in *.hic* format were downloaded from the NCBI GEO portal (see Supplementary Table S1 for details) and parsed with *Straw*, a data extraction API for *.hic* files (17).

FASTQ files for Hi-C datasets from ENCODE were processed with the HiC-Pro pipeline (15) using default parameters for HindIII digested DNA. Contact matrices were normalized with the ICE procedure (18). Subsampling of Hi-C contact matrices was performed using uniform sampling of individual reads without replacement.

Mango pipeline

Mango 1.2.1 (19) was downloaded from https://github.com/dphansti/mango. Additionally, we installed the dependencies *R* 3.4.4, *bedtools* 2.26.0, *macs2* (version 2.1.1.20160309) (20), and *bowtie-align* 1.2 (21). Mango was executed with the default parameters and the flags `verboseoutput` and `reportallpairs` were set. For datasets that were generated with the ChIA-PET Tn5 tagmentation protocol, additional parameters recommended by the author were used: `-keepempty TRUE -maxlength 1000 -shortreads FALSE`. For subsequent IDR2D analyses, we used the `P` column in the Mango output files to establish the ranking of interactions. This column contains unadjusted *P*-values, which were transformed using the `log.additive.inverse` transformation to match the IDR semantics of the value column, where interactions with larger values are more likely to be true interactions.

The BEDPE files generated by Mango after step 3 were also used by the ChIA-PET2 and CID pipelines.

ChIA-PET2 pipeline

ChIA-PET2 0.9.2 (13) was obtained from https://github.com/GuipengLi/ChIA-PET2. The default setting was used for all parameters, except that the starting step was set to 4 to start the analysis from Mango-derived BEDPE files. The ranking for the IDR2D analysis was established by the untransformed `-log10(1-PostProb)` column, which is an output of *MICC* (12), a Bayesian mixture model used internally by ChIA-PET2 and *CID*.

CID pipeline

CID 1.0 (14) is part of the Java genomics software package *GEM* 3.4, which was downloaded from https://groups.csail.mit.edu/cgs/gem/versions.html. We used the default CID parameters. Before running MICC, we filtered all interactions that were supported by only one PET read. Same as with ChIA-PET2, we used the untransformed `-log10(1-PostProb)` column to rank interactions in IDR2D.

Package and web development

The R package development process was supported using *devtools*. We used *roxygen2* for inline function documentation, and *knitr* and *R Markdown* for package vignettes. With the R package we provide a platform-independent implementation of the methods introduced in this paper. The Hi-C analysis part of the package requires the Python package *hic-straw* (17), which is a data extraction API for Hi-C contact maps.

The website was developed in R with the reactive web application framework *shiny* from RStudio. The components of the graphical user interface were provided by shiny and *shinyBS*, which serve as an R wrapper for the components of the Bootstrap front-end web development framework. The analysis job queue of the website uses an *SQLite* database.

RESULTS

IDR2D identifies reproducible components of ENCODE ChIA-PET experiments

To assess the performance and utility of IDR2D we analyzed the read data of 17 ChIA-PET experiments that had

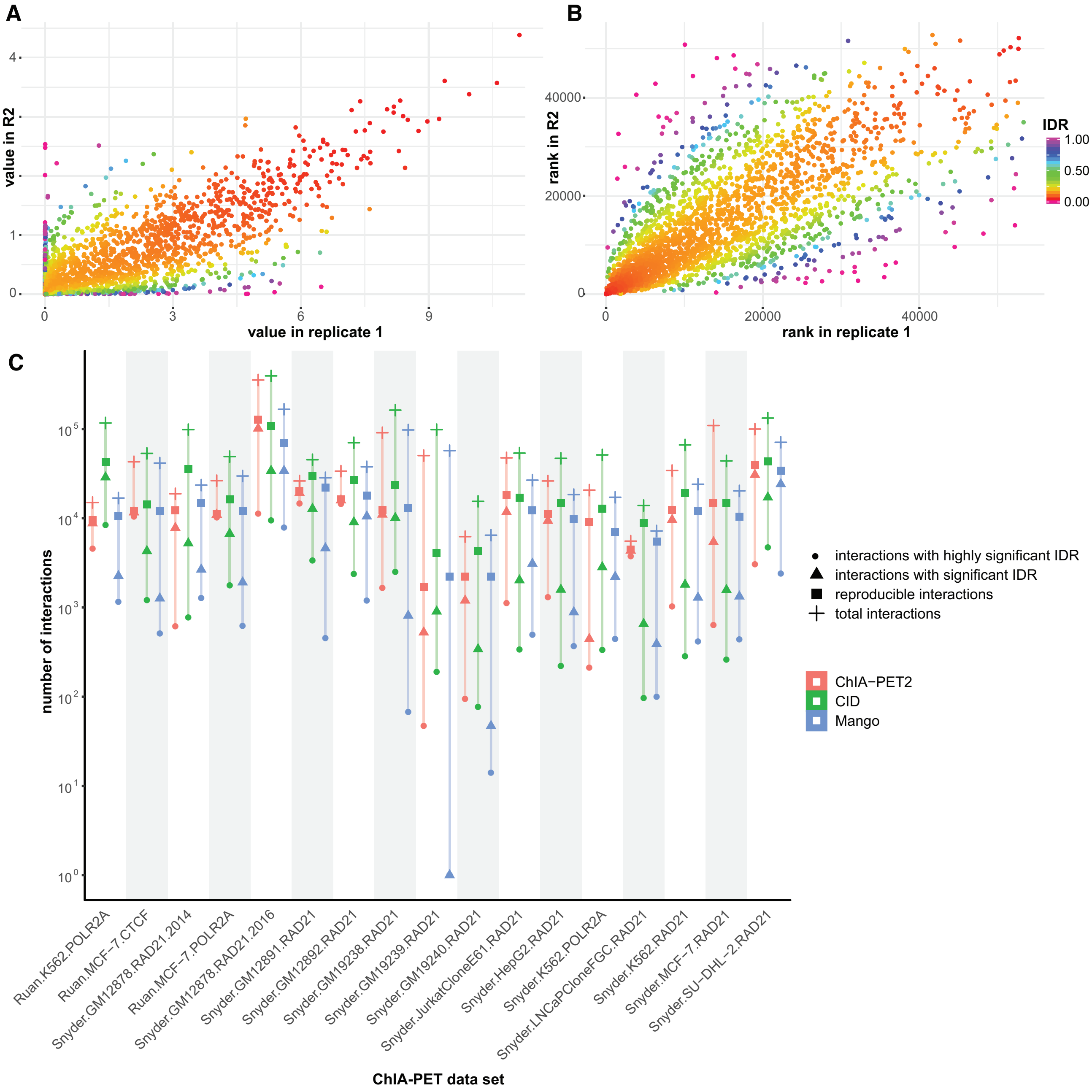

Figure 2. IDR2D analysis of 17 replicated ChIA-PET experiments identifies reproducible components. (**A**) Diagnostic scatterplot of IDRs of genomic interactions called by CID from replicated ChIA-PET experiments targeting RAD21 in HepG2 cells. Plotted are replicated interactions with their estimated IDR (color) and their scores in the two replicates (position). As expected, interactions with low IDRs that have a low probability of belonging to the irreproducible component, are along the diagonal (similar scores in both replicates). (**B**) Similar to panel A, but plots interaction ranks instead of scores (higher score results in lower rank). (**C**) A comparison of ChIA-PET interaction callers ChIA-PET2, CID, and Mango across 17 ChIA-PET experiments. Significant IDR <0.05, highly significant IDR <0.01, total interactions is the number of interactions in replicate 1.

replicates (see Supplementary Table S1 for details). Mango was used for data preprocessing such as linker removal, read mapping (via bowtie), and peak calling (via macs2). We called interactions with three different methods (ChIA-PET2, CID, and Mango) and then used IDR2D to identify reproducible interactions across replicates. The number of identified interactions varies greatly between the three interaction callers, with on average CID identifying the most, and Mango the fewest interactions (see Figure 2C and Supplementary Tables S2–S4). As a result, the overall reproducibility of interactions is also dependent on the interaction caller. For example, the ChIA-PET experiments `Snyder.GM19239.RAD21` and `Snyder.GM19240.RAD21` show poor reproducibility across all three interaction calling methods. By identifying the reproducible component within each of the replicated experiments, IDR2D helps to assess the overall reproducibility of each experiment, as well as the reproducibility of individual findings, which in turn

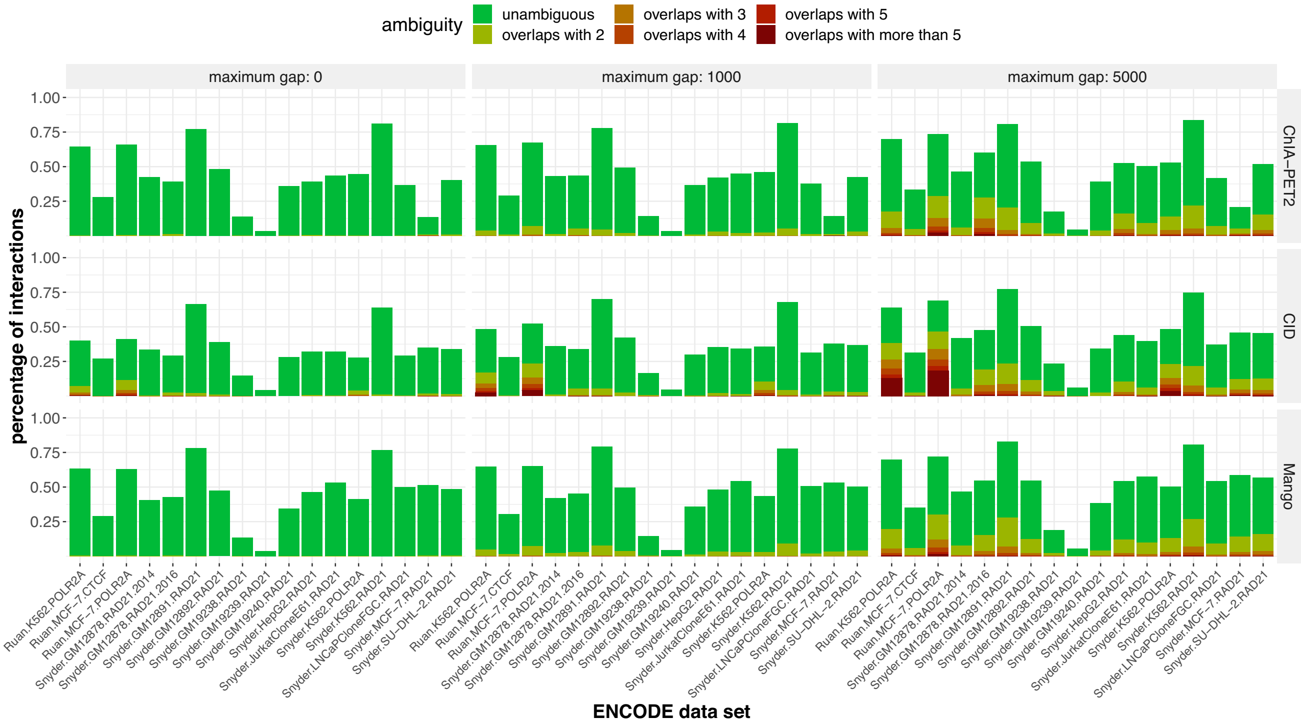

Figure 3. Mappings of genomic interactions between replicated ChIA-PET experiments are predominantly unambiguous. The great majority of interactions in replicate 1 that overlapped with interactions in replicate 2 only overlapped with one interaction, leading to an unambiguous assignment of corresponding replicated interactions (green bars). Unsurprisingly, the number of ambiguous mappings (interactions in replicate 1 that overlap with more than one interaction in replicate 2) increases when the maximum acceptable gap is increased, the tolerated distance between anchors to still be considered overlapping.

informs the conclusions drawn from the data. In addition, it can be used to help qualify new experimental protocols for consistent results. Venn diagrams depicting the overlap of identified interactions between ChIA-PET2, CID and Mango are shown in Supplementary Figure S1.

Furthermore, we used IDR2D to analyze experimental results from replicated HiChIP (see Supplementary Tables S5 and S6). Similar to ChIA-PET, IDR2D can identify reproducible HiChIP interactions and expose poorly replicated experiments, which is valuable information for subsequent analysis steps.

Mappings of genomic interactions between replicated ChIA-PET experiments are predominantly unambiguous

The great majority of interactions in replicate 1 that overlapped with interactions in replicate 2 overlapped with only one interaction, leading to an unambiguous assignment of corresponding replicated interactions (see Figure 3). Unsurprisingly, the number of ambiguous mappings (interactions in replicate 1 that overlap with more than one interaction in replicate 2) increases when the *maximum gap* is increased, the tolerated distance between anchors that are considered to overlap. On average, only 2.66% of interactions are ambiguous in the case of zero *max gap*, whereas this number increases to 8.00% and 24.11% with maximum gaps of 1000 and 5000 bp, respectively.

There are more ambiguous mappings between replicated interactions that were called with CID (14.73% for CID, 9.90% for ChIA-PET2 and 10.14% for Mango). We expect this is because (i) CID on average calls significantly more interactions than ChIA-PET2 and Mango and (ii) interactions called with CID exhibit a wider range of anchor lengths, and longer anchors naturally increase the probability of overlap.

Assessing reproducibility of Hi-C experiments

When analyzing pairs of Hi-C experiments with IDR2D, blocks from Hi-C contact matrices are used as observations. The resolution of contact matrix values typically ranges between 5 kb (kilo base pairs) to 2.5 Mb blocks. With the fixed grid of contact map observations, finding corresponding observations in the second replicate is straightforward. Each block in replicate 1 is simply matched with the block spanning the same genomic regions in replicate 2. Blocks are subsequently ranked by their read counts and analyzed using the same procedure that was used for ChIA-PET and HiChIP data.

In addition to computing IDR values, IDR2D produces diagnostic plots that help interpret the overall reproducibility of a pair of Hi-C experiments, as well as identify reproducible parts of Hi-C contact matrices for a more focused, downstream analysis.

In Figure 4, we show IDR2D results for three pairs of Hi-C experiments. The first pair of Hi-C experiments consists of true replicate experiments in GM12878 cells (22). The second pair of experiments were obtained in phased

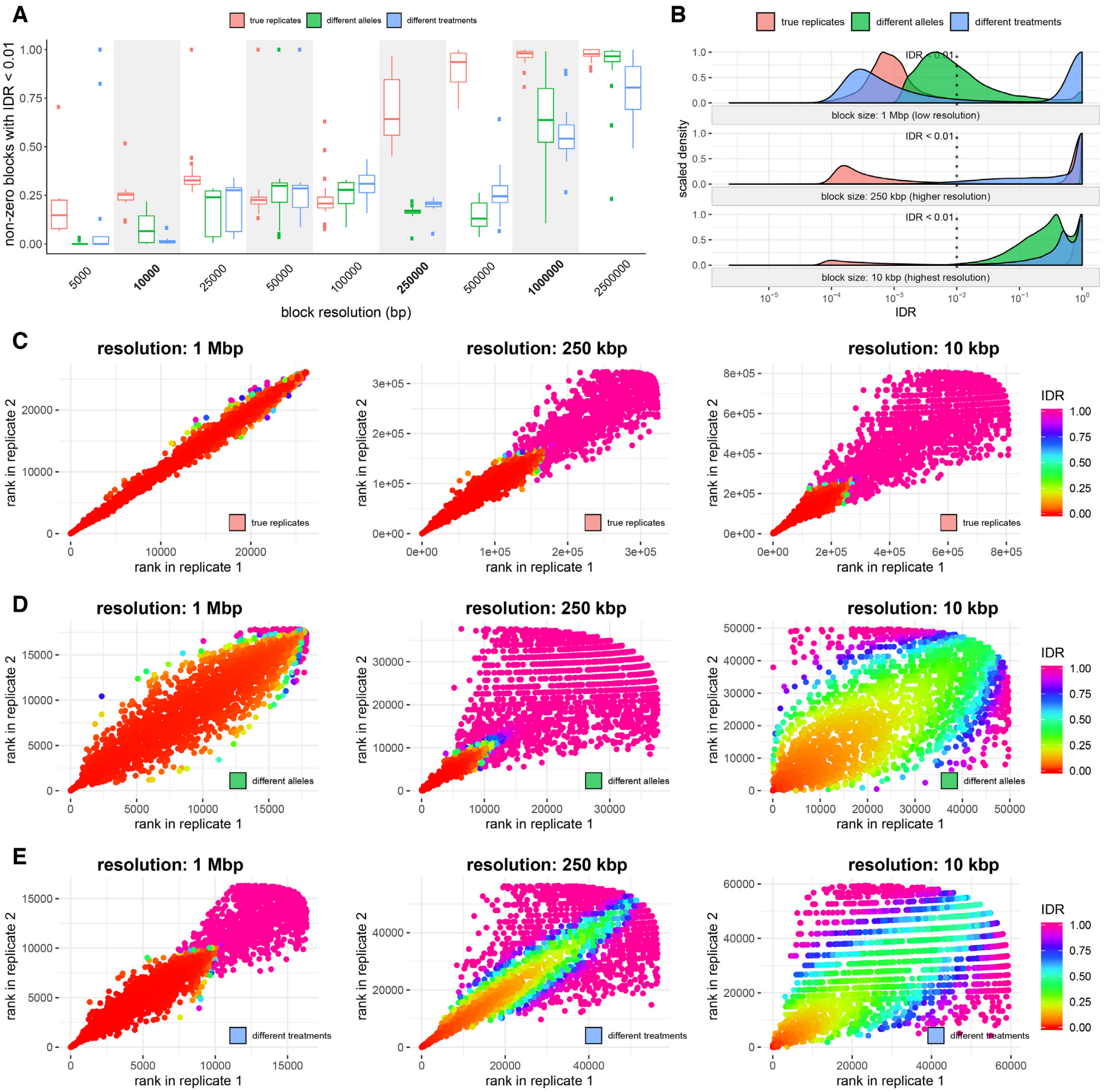

Figure 4. Reproducibility analysis of Hi-C experiments. (**A**) Summary of IDR2D results on individual chromosomes of three pairs of Hi-C experiments, True replicate Hi-C experiments (Lieberman.GM12878) are compared to IDR2D analysis of Hi-C experiments of different alleles (Lieberman.Patski) and different treatments (Skok.NSD2). (**B**) Histograms of the IDR distribution of IDR values for all blocks of chromosome 1 for the three pairs of Hi-C experiments. (**C**) Scatterplots of block ranks of chromosome 1 of the two Hi-C replicate experiments, colored by IDR. (**D**) Analagous to C, for Hi-C experiments of paternal and maternal alleles. (**E**) Analagous to C, for Hi-C experiments before and after overexpression of NSD2. Axis scales are not fixed between scatterplots.

murine embryonic kidney fibroblasts, where allele-specific Hi-C reads (22) were available (different alleles in Figure 4) (23), and the third pair of Hi-C experiments were obtained before and after the overexpression of NSD2 (different treatments) (24). GEO identifiers of all data sets are listed in the Supplementary Table S1 and detailed results in Supplementary Table S7. Figure 4A gives an overview of all data sets and all resolutions, showing that, as expected, the reproducibility is highest between true replicates, and in general higher at lower resolutions (larger blocks). Figure 4B depicts the distribution of IDR values for chromosome 1 of each of the Hi-C pairs at block resolutions of 1 Mb, 250 kb and 10 kb. The largest fraction of reproducible blocks is found between replicated experiments. Figures 4C–E are scatterplots of interaction pairs (corresponding blocks in the contact matrices) of the two experiments, where the color denotes the IDR value of the interaction pair. Given a Hi-C experiment with a fixed sequencing depth, the higher the resolution of the Hi-C analysis the less reproducible the individual interactions will be as a consequence of sampling noise.

Not all Hi-C interaction pairs that lie on the diagonal and have similar ranks in both replicates are deemed reproducible by IDR2D. For example, see the upper right panel of Figure 4C. This lack of reproducibility is intended and is justified by taking into account the poor reproducibility of other interaction pairs with similar ranks. Hi-C contacts close to the diagonal can be found irreproducible when they are in rank neighborhoods of irreproducibility. We note that while experiment level methods may find a Hi-C experiment reproducible, IDR2D may find a specific interaction irreproducible because it is in a rank neighborhood that is not reproducibly ordered. IDR2D may require increased sequencing depth to consistently rank interactions to ascertain reproducibility of such interactions.

IDR2D is largely insensitive to sequencing depth when it is sufficient to recover contacts, thus reproducible pairs of experiments are identified as such even if their sequencing depths are different. Reproducibility as measured by IDR2D only starts to degrade significantly at extremely low coverage, with only very few reads (single to low double digits) per block. Subsampling experiments were performed to illustrate this behavior (see Supplementary Figure S2).

DISCUSSION

The appropriate choice of significance values for the computation of interaction ranks depends on the method used to identify contacts. As a general rule, larger values should reflect higher confidence and there should be as few ties as possible. IDR2D operates on the ranks of the interactions in both replicates and therefore is invariant to order-preserving transformations of the original significance values. If *P*-values are used as significance values for interactions, the additive inverse or the log additive inverse of uncorrected *P*-values is recommended. Unadjusted *P*-values are preferred over *P*-values adjusted for multiple hypothesis testing, because uncorrected *P*-values reduce rank ties.

Other methods assess the overall reproducibility of genome interaction experiments but do not characterize the reproducability of each reported contact. Such methods include HiCRep (25), HiC-spector (26), and GenomeDISCO (27). GenomeDISCO also supports experimental data from ChIA-PET and HiChIP. HiCRep calculates a score of experiment reproducibility between contact matrices based on aggregated stratified Pearson correlation coefficients, while HiC-spector determines experiment reproducibility by comparing the eigenvectors of a spectral decomposition of the contact maps, and GenomeDISCO's concordance score is based on random walks on a graph representation of contact maps. These methods have in common that they assess the overall reproducibility of replicated experiments with a global measure of similarity between contact matrices. IDR2D provides a measure of reproducibility for each reported contact and then summarizes these findings to characterize experiment reproducability (see Supplementary Figure S3). IDR2D's fine-grained analysis of reproducibility identifies contacts that are invariant across experimental replicates and those that are not, which is a unique capability. Thus, IDR2D is intended to complement, rather than replace previous Hi-C reproducibility assessment methods.

IDR2D, and the methods mentioned above, are limited to comparisons of two replicates at a time. If more replicates are available, multiple pairwise analysis can be performed and the results combined.

While IDR2D is a compatible post-processing step for the tested interaction callers, it cannot recover true interactions that were discarded by the interaction caller and therefore the identified set of reproducible interactions is always limited by the sensitivity of the caller.

IDR2D can potentially support single-cell or single-molecule chromatin interaction data obtained by methods such as Sci-Hi-C (28) and ChIA-Drop (29). However, the sparsity of interaction data from single cells will necessitate data imputation or cell clustering as preprocessing steps to IDR2D, similar to strategies applied to single-cell ATAC-seq data (30).

IDR2D offers a complementary way to evaluate the results of chromatin interaction experiments for significance, and provides a foundation for subsequent analysis such as enhancer-gene mapping that incorporates the important concept of experimental replicability.

DATA AVAILABILITY

The implementation of IDR2D facilitates workflow integration with other data analysis pipelines, and is also web-accessible at https://idr2d.mit.edu. IDR2D is implemented in R and bundled as an R/Bioconductor package (idr2d), supporting observations with both one-dimensional and two-dimensional genomic coordinates. The IDR2D website implementation offers a number of ways to transform the scores to match IDR requirements, and to map interactions between replicates. The source code of the R package is hosted on GitHub (https://github.com/gifford-lab/idr2d).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank members of the Gifford lab for insightful suggestions and discussions. We also thank William Stafford Noble for his thoughtful comments on an earlier version of the manuscript.

Author contributions: Conceptualization, D.K.G.; Methodology, K.K.; Software, K.K.; Formal Analysis, K.K. and Y.G.; Investigation, K.K. and Y.G.; Resources, D.K.G.; Data Curation, K.K. and Y.G.; Writing - Original Draft, K.K.; Writing - Review & Editing, K.K., Y.G. and D.K.G.; Visualization, K.K.; Supervision, D.K.G.; Funding Acquisition, D.K.G.

FUNDING

National Institutes of Health (NIH) [1R01HG008363 to D.K.G., 1R01NS078097 to H.W. and D.K.G.]. Funding for open access charge: NIH.

Conflict of interest statement. None declared.

REFERENCES

1. Li,Q., Brown,J.B., Huang,H. and Bickel,P.J. (2011) Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.*, **5**, 1752–1779.
2. Landt,S.G., Marinov,G.K., Kundaje,A., Kheradpour,P., Pauli,F., Batzoglou,S., Bernstein,B.E., Bickel,P., Brown,J.B., Cayting,P. *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, **22**, 1813–1831.
3. Bailey,T., Krajewski,P., Ladunga,I., Lefebvre,C., Li,Q., Liu,T., Madrigal,P., Taslim,C. and Zhang,J. (2013) Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS Comput. Biol.*, **9**, e1003326.
4. Djebali,S., Davis,C.A., Merkel,A., Dobin,A., Lassmann,T., Mortazavi,A., Tanzer,A., Lagarde,J., Lin,W., Schlesinger,F. *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.
5. Chen,X., Iliopoulos,D., Zhang,Q., Tang,Q., Greenblatt,M.B., Hatziapostolou,M., Lim,E., Tam,W.L., Ni,M., Chen,Y. *et al.* (2014) XBP1 promotes triple-negative breast cancer by controlling the HIF1-alpha pathway. *Nature*, **508**, 103–107.
6. Wang,H., Maurano,M.T., Qu,H., Varley,K.E., Gertz,J., Pauli,F., Lee,K., Canfield,T., Weaver,M., Sandstrom,R. *et al.* (2012) Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res.*, **22**, 1680–1688.
7. Zanconato,F., Forcato,M., Battilana,G., Azzolin,L., Quaranta,E., Bodega,B., Rosato,A., Bicciato,S., Cordenonsi,M. and Piccolo,S. (2015) Genome-wide association between YAP/TAZ/TEAD and AP-1 at enhancers drives oncogenic growth. *Nat. Cell Biol.*, **17**, 1218–1227.
8. Madrigal,P. (2015) CexoR: an R/Bioconductor package to uncover high-resolution protein-DNA interactions in ChIP-exo replicates. *EMBnet.journal*, **21**, 837.
9. Fullwood,M.J. and Ruan,Y. (2009) ChIP-based methods for the identification of long-range chromatin interactions. *J. Cell. Biochem.*, **107**, 30–39.
10. Mumbach,M.R., Rubin,A.J., Flynn,R.A., Dai,C., Khavari,P.A., Greenleaf,W.J. and Chang,H.Y. (2016) HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods*, **13**, 919–922.
11. Lieberman-Aiden,E., van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
12. He,C., Zhang,M.Q. and Wang,X. (2015) MICC: an R package for identifying chromatin interactions from ChIA-PET data. *Bioinformatics*, **31**, 3832–3834.
13. Li,G., Chen,Y., Snyder,M.P. and Zhang,M.Q. (2017) ChIA-PET2: a versatile and flexible pipeline for ChIA-PET data analysis. *Nucleic Acids Res.*, **45**, e4.
14. Guo,Y., Krismer,K., Gifford,D.K., Wichterle,H. and Closser,M. (2019) High resolution discovery of chromatin interactions. *Nucleic Acids Res.***47**, e35.
15. Servant,N., Varoquaux,N., Lajoie,B.R., Viara,E., Chen,C.J., Vert,J.P., Heard,E., Dekker,J. and Barillot,E. (2015) HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.*, **16**, 259.
16. Lareau,C.A. and Aryee,M.J. (2018) hichipper: a preprocessing pipeline for calling DNA loops from HiChIP data. *Nat. Methods*, **15**, 155–156.
17. Durand,N.C., Robinson,J.T., Shamim,M.S., Machol,I., Mesirov,J.P., Lander,E.S. and Aiden,E.L. (2016) Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.*, **3**, 99–101.
18. Imakaev,M., Fudenberg,G., McCord,R.P., Naumova,N., Goloborodko,A., Lajoie,B.R., Dekker,J. and Mirny,L.A. (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods*, **9**, 999–1003.
19. Phanstiel,D.H., Boyle,A.P., Heidari,N. and Snyder,M.P. (2015) Mango: a bias-correcting ChIA-PET analysis pipeline. *Bioinformatics*, **31**, 3092–3098.
20. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
21. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
22. Rao,S.S., Huntley,M.H., Durand,N.C., Stamenova,E.K., Bochkov,I.D., Robinson,J.T., Sanborn,A.L., Machol,I., Omer,A.D., Lander,E.S. *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.
23. Darrow,E.M., Huntley,M.H., Dudchenko,O., Stamenova,E.K., Durand,N.C., Sun,Z., Huang,S.C., Sanborn,A.L., Machol,I., Shamim,M. *et al.* (2016) Deletion of DXZ4 on the human inactive X chromosome alters higher-order genome architecture. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, E4504–E4512.
24. Lhoumaud,P., Badri,S., Hernaez,J.R., Sakellaropoulos,T., Sethia,G., Kloetgen,A., Cornwell,M., Bhattacharyya,S., Ay,F., Bonneau,R. *et al.* (2019) NSD2 overexpression drives clustered chromatin and transcriptional changes in a subset of insulated domains. *Nat Commun*, **10**, 4843.
25. Yang,T., Zhang,F., Yardımcı,G.G., Song,F., Hardison,R.C., Noble,W.S., Yue,F. and Li,Q. (2017) HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome Res.*, **27**, 1939–1949.
26. Yan,K.K., Yardimci,G.G., Yan,C., Noble,W.S. and Gerstein,M. (2017) HiC-spector: a matrix library for spectral and reproducibility analysis of Hi-C contact maps. *Bioinformatics*, **33**, 2199–2201.
27. Ursu,O., Boley,N., Taranova,M., Wang,Y. X.R., Yardimci,G.G., Stafford Noble,W. and Kundaje,A. (2018) GenomeDISCO: a concordance score for chromosome conformation capture experiments using random walks on contact map graphs. *Bioinformatics*, **34**, 2701–2707.
28. Ramani,V., Deng,X., Qiu,R., Lee,C., Disteche,C.M., Noble,W.S., Shendure,J. and Duan,Z. (2020) Sci-Hi-C: a single-cell Hi-C method for mapping 3D genome organization in large number of single cells. *Methods*, **170**, 61–68.
29. Zheng,M., Tian,S.Z., Capurso,D., Kim,M., Maurya,R., Lee,B., Piecuch,E., Gong,L., Zhu,J.J., Li,Z. *et al.* (2019) Multiplex chromatin interactions with single-molecule precision. *Nature*, **566**, 558–562.
30. Xiong,L., Xu,K., Tian,K., Shao,Y., Tang,L., Gao,G., Zhang,M., Jiang,T. and Zhang,Q.C. (2019) SCALE method for single-cell ATAC-seq analysis via latent feature extraction. *Nat. Commun.*, **10**, 4576.