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

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

***Abstract*— Chromatin interaction data from protocols such as ChIA-PET and HiChIP provide valuable insights into genome organization and gene regulation, but can include spurious interactions that do not reflect underlying genome biology. We introduce a generalization of the Irreproducible Discovery Rate (IDR) method called IDR2D that identifies replicable interactions shared by experiments. IDR2D provides a principled set of interactions and eliminates artifacts from single experiments.**

I. Main

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. The irreproducible noise component is estimated and can 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. IDR has also been used in numerous projects outside of ENCODE [3–6].

Chromatin interaction experiments such as ChIA-PET [7] and HiChIP [8] 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, typically 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 how IDR2D can be applied to data from ChIA-PET [7] and HiChIP [8] experiments.

Like IDR, IDR2D independently ranks the results 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 can be estimated for each interaction pair.

If interaction x_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:

$$x_2^* = \underset{x_2 \in \Omega_{x_1}}{\operatorname{argmin}} f(x_1, x_2), \tag{1}$$

where Ω_{x_1} is the set of interactions in replicate 2 that overlaps with the interaction x_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 additive inverse of the sum of the interaction ranks.

IDR2D is used as the final step in chromatin interaction data workflows (see figure 1A). The input to IDR2D 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 [9–11], or based on a heuristic. Figure 1B breaks the IDR2D procedure into five steps.

To assess the performance and utility of IDR2D we analyzed the read data of 17 ChIA-PET experiments that had 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). However, the fraction of reproducible interactions is highest among interactions called by ChIA-PET2. The overall reproducibility of interactions is strongly dependent on the experiment and less dependent on the interaction caller used. For example, the ChIA-PET experiments `Snyder.GM19239.RAD21`

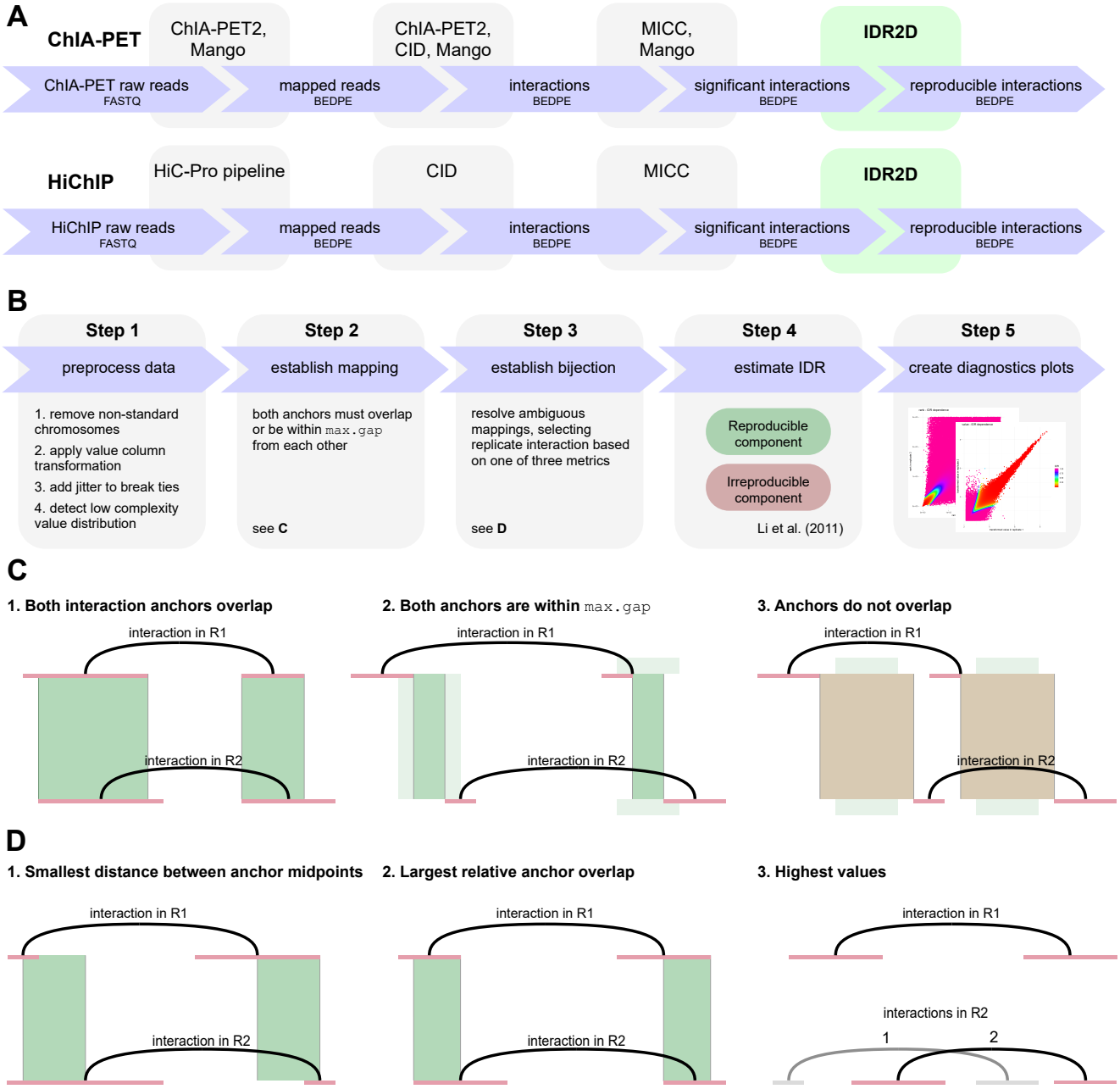

Fig. 1. **IDR2D identifies reproducible genomic interactions.** (**A**) IDR2D is a potential post-processing step in the data analysis pipelines for ChIA-PET and HiChIP experiments that were done in replicates. It is compatible with a range of different interaction callers, such as ChIA-PET2, Mango, 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 `max.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 highest value in the column that is used to establish the ranking, which prioritizes more significant interactions.

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 informs the conclusions drawn from the data. In addition, it can be used to help qualify new antibodies or experimental protocols for consistent results.

Furthermore, we used IDR2D to analyze experimental results from replicated HiChIP (see supplemental tables S5 and S6).

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 R package enables workflow integration with other data analysis pipelines. Additionally, we make the functionality of IDR2D is web-accessible at https://idr2d.mit.edu. The website offers a number of ways to transform the scores to match IDR requirements, and to map interactions between replicates.

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.

II. Methods

A. 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.

B. ChIA-PET datasets

We used 17 ChIA-PET datasets associated with protein factors such as POL2RA, CTCF and RAD21 from various 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*.

C. HiChIP datasets

The FASTQ SMC1A HiChIP data from GM12878 cells [8] were downloaded from the NCBI GEO portal (GSE80820). Raw read files were analyzed with HiC-Pro [12], and interactions were subsequently called by CID and *hichipper* [13].

D. Mango pipeline

Mango 1.2.1 [14] 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) [15], and *bowtie-align* 1.2 [16]. 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.

E. ChIA-PET2 pipeline

ChIA-PET2 0.9.2 [10] 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* [9], a Bayesian mixture model used internally by ChIA-PET2 and *CID*.

F. CID pipeline

CID 1.0 [11] is part of the Java genomics software package *GEM* 3.4, which was downloaded from http://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.

G. 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 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.

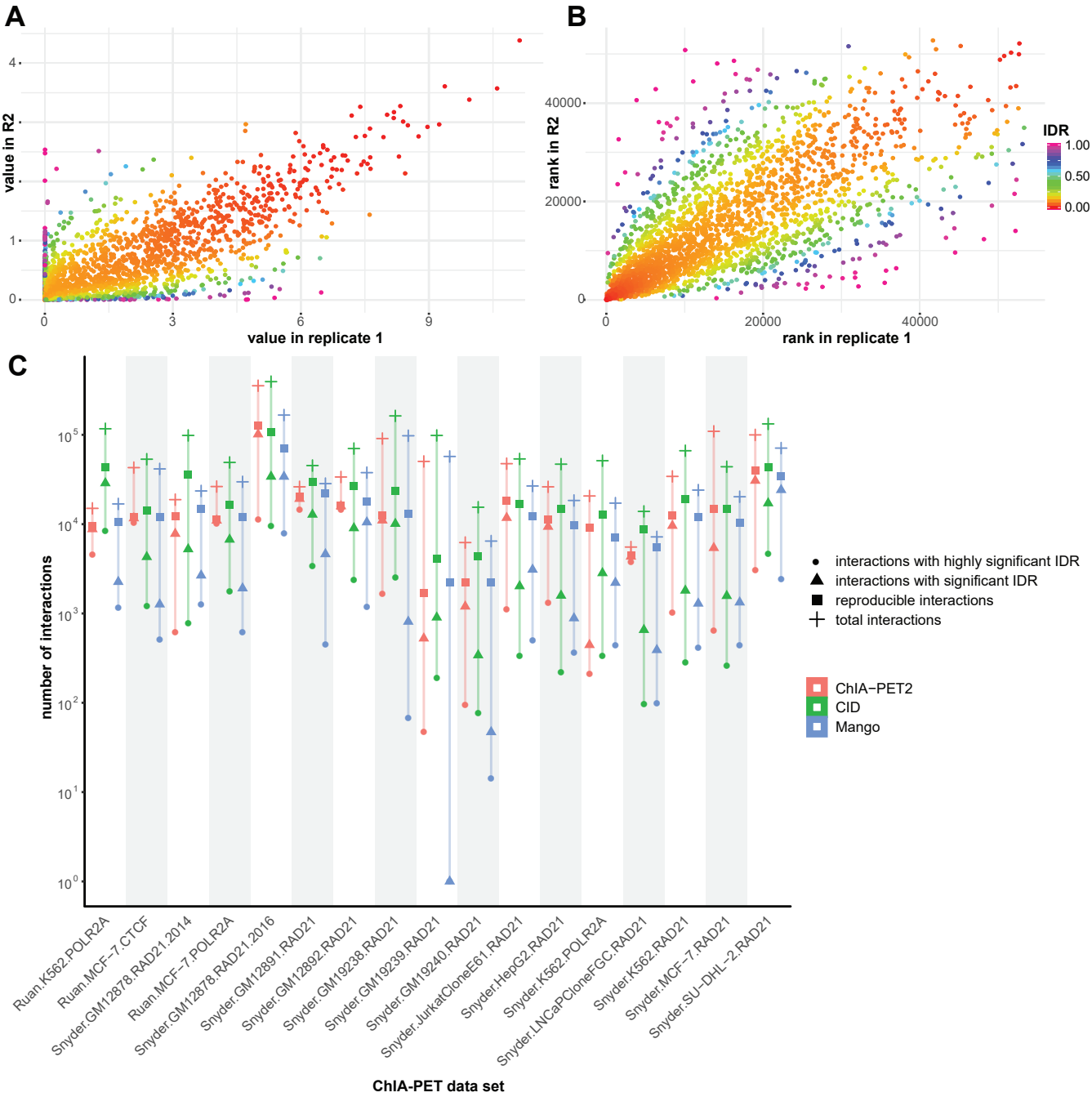

Fig. 2. **IDR2D analysis of 17 replicated ChIA-PET experiments identifies reproducible components.** (**A**) Diagnostic scatterplot of IDRs of genomic interactions called by CID from replicated ChIA-PET experiments targeting RAD21 in HepG2 cells. Plotted are replicated interactions with their estimated IDR (color) and their scores in the two replicates (position). As expected, interactions with low IDRs, i.e., having a low probability of belonging to the irreproducible component, are along the diagonal (similar scores in both replicates) and are generally scored higher. (**B**) Similar to panel **A**, but plotting the ranks instead of the scores. Lower rank means higher score. (**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.

AVAILABILITY

The functionality of IDR2D is available as a website (https://idr2d.mit.edu). For workflow integration and advanced analysis, the method is also available as an R/Bioconductor package. The source code of the R package is hosted on GitHub (https://github.com/gifford-lab/idr2d).

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.

ACKNOWLEDGEMENTS

We thank members of the Gifford lab for suggestions.

FUNDING

National Institutes of Health (NIH) grant 1R01HG008363 to D.K.G.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

[1] Li, Q., Brown, J. B., Huang, H., and Bickel, P. J. (09, 2011) Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.,* **5**(3), 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., Chen, Y., DeSalvo, G., Epstein, C., Fisher-Aylor, K. I., Euskirchen, G., Gerstein, M., Gertz, J., Hartemink, A. J., Hoffman, M. M., Iyer, V. R., Jung, Y. L., Karmakar, S., Kellis, M., Kharchenko, P. V., Li, Q., Liu, T., Liu, X. S., Ma, L., Milosavljevic, A., Myers, R. M., Park, P. J., Pazin, M. J., Perry, M. D., Raha, D., Reddy, T. E., Rozowsky, J., Shoresh, N., Sidow, A., Slattery, M., Stamatoyannopoulos, J. A., Tolstorukov, M. Y., White, K. P., Xi, S., Farnham, P. J., Lieb, J. D., Wold, B. J., and Snyder, M. (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research,* **22**(9), 1813–1831.

[3] Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Roder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Bar, N. S., Batut, P., Bell, K., Bell, I., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Falconnet, E., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O. J., Park, E., Persaud, K., Preall, J. B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L. H., Shahab, A., Skancke, J., Suzuki, A. M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Ruan, X., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T., Reymond, A., Antonarakis, S. E., Hannon, G., Giddings, M. C., Ruan, Y., Wold, B., Carninci, P., Guigo, R., and Gingeras, T. R. (Sep, 2012) Landscape of transcription in human cells. *Nature,* **489**(7414), 101–108.

[4] Chen, X., Iliopoulos, D., Zhang, Q., Tang, Q., Greenblatt, M. B., Hatziapostolou, M., Lim, E., Tam, W. L., Ni, M., Chen, Y., Mai, J., Shen, H., Hu, D. Z., Adoro, S., Hu, B., Song, M., Tan, C., Landis, M. D., Ferrari, M., Shin, S. J., Brown, M., Chang, J. C., Liu, X. S., and Glimcher, L. H. (Apr, 2014) XBP1 promotes triple-negative breast cancer by controlling the HIF1-alpha pathway. *Nature,* **508**(7494), 103–107.

[5] Wang, H., Maurano, M. T., Qu, H., Varley, K. E., Gertz, J., Pauli, F., Lee, K., Canfield, T., Weaver, M., Sandstrom, R., Thurman, R. E., Kaul, R., Myers, R. M., and Stamatoyannopoulos, J. A. (Sep, 2012) Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res.,* **22**(9), 1680–1688.

[6] Zanconato, F., Forcato, M., Battilana, G., Azzolin, L., Quaranta, E., Bodega, B., Rosato, A., Bicciato, S., Cordenonsi, M., and Piccolo, S. (Sep, 2015) Genome-wide association between YAP/TAZ/TEAD and AP-1 at enhancers drives oncogenic growth. *Nat. Cell Biol.,* **17**(9), 1218–1227.

[7] Fullwood, M. J. and Ruan, Y. (May, 2009) ChIP-based methods for the identification of long-range chromatin interactions. *J. Cell. Biochem.,* **107**(1), 30–39.

[8] Mumbach, M. R., Rubin, A. J., Flynn, R. A., Dai, C., Khavari, P. A., Greenleaf, W. J., and Chang, H. Y. (Nov, 2016) HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods,* **13**(11), 919–922.

[9] He, C., Zhang, M. Q., and Wang, X. (Dec, 2015) MICC: an R package for identifying chromatin interactions from ChIA-PET data. *Bioinformatics,* **31**(23), 3832–3834.

[10] Li, G., Chen, Y., Snyder, M. P., and Zhang, M. Q. (01, 2017) ChIA-PET2: a versatile and flexible pipeline for ChIA-PET data analysis. *Nucleic Acids Res.,* **45**(1), e4.

[11] Guo, Y., Krismer, K., Gifford, D. K., Wichterle, H., and Closser, M. (02, 2019) High resolution discovery of chromatin interactions. *Nucleic Acids Res.,*.

[12] Servant, N., Varoquaux, N., Lajoie, B. R., Viara, E., Chen, C.-J., Vert, J.-P., Heard, E., Dekker, J., and Barillot, E. (Dec, 2015) HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biology,* **16**(1), 259.

[13] Lareau, C. A. and Aryee, M. J. (02, 2018) hichipper: a preprocessing pipeline for calling DNA loops from HiChIP data. *Nat. Methods,* **15**(3), 155–156.

[14] Phanstiel, D. H., Boyle, A. P., Heidari, N., and Snyder, M. P. (Oct, 2015) Mango: a bias-correcting ChIA-PET analysis pipeline. *Bioinformatics,* **31**(19), 3092–3098.

[15] Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., and Liu, X. S. (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.,* **9**(9), R137.

[16] 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**(3), R25.