genes

MDPI

*Article*

# enChIP-Seq Analyzer: A Software Program to Analyze and Interpret enChIP-Seq Data for the Detection of Physical Interactions between Genomic Regions

**Ashita Sarudate [1], Toshitsugu Fujita [2], Takahiro Nakayama [1] and Hodaka Fujii [2,*]**

[1] Research Institute of Bio-System Informatics, Tohoku Chemical Co., Ltd., 6-15-5 Mitake, Morioka 020-0122, Iwate, Japan; sarudate@t-kagaku.co.jp (A.S.); nakayama@t-kagaku.co.jp (T.N.)
[2] Department of Biochemistry and Genome Biology, Hirosaki University Graduate School of Medicine, 5 Zaifu-cho, Hirosaki 036-8562, Aomori, Japan; toshitsugu.fujita@hirosaki-u.ac.jp
* Correspondence: hodaka@hirosaki-u.ac.jp; Tel.: +81-(0)172-39-5018

**Abstract:** Accumulating evidence suggests that the physical interactions between genomic regions play critical roles in the regulation of genome functions, such as transcription and epigenetic regulation. Various methods to detect the physical interactions between genomic regions have been developed. We recently developed a method to search for genomic regions interacting with a locus of interest in a non-biased manner that combines pull-down of the locus using engineered DNA-binding molecule-mediated chromatin immunoprecipitation (enChIP) and next-generation sequencing (NGS) analysis (enChIP-Seq). The clustered regularly interspaced short palindromic repeats (CRISPR) system, consisting of a nuclease-dead form of Cas9 (dCas9) and a guide RNA (gRNA), or transcription activator-like (TAL) proteins, can be used for enChIP. In enChIP-Seq, it is necessary to compare multiple datasets of enChIP-Seq data to unambiguously detect specific interactions. However, it is not always easy to analyze enChIP-Seq datasets to subtract non-specific interactions or identify common interactions. To facilitate such analysis, we developed the enChIP-Seq analyzer software. It enables easy extraction of common signals as well as subtraction of non-specific signals observed in negative control samples, thereby streamlining extraction of specific enChIP-Seq signals. enChIP-Seq analyzer will help users analyze enChIP-Seq data and identify physical interactions between genomic regions.

**Keywords:** enChIP-Seq; CRISPR; intergenomic interactions; 3-D genomics; enChIP-Seq analyzer



**Citation:** Sarudate, A.; Fujita, T.; Nakayama, T.; Fujii, H. enChIP-Seq Analyzer: A Software Program to Analyze and Interpret enChIP-Seq Data for the Detection of Physical Interactions between Genomic Regions. *Genes* **2022**, *13*, 472. https://doi.org/10.3390/genes13030472

Academic Editors: Stefano Lonardi and Cenk Sahinalp

Received: 13 December 2021
Accepted: 2 March 2022
Published: 7 March 2022



## 1. Introduction

The dynamics of the three-dimensional structure of the genome play a critical role in the regulation of genome functions such as transcription, epigenetic regulation, genomic imprinting, and X-chromosome inactivation [1,2]. Methods to detect physical interactions between genomic regions have been developed in the last two decades. Chromosome conformation capture (3C) was the first biochemical method widely used to detect physical interactions between genomic regions [3]. Various derivatives of 3C have been developed [4]. 3C and its derivatives depend on cross-linking, fragmentation of chromatin by enzymatic digestion or other methods, and ligation of proximal DNA ends. The ligation step can be a source of artifactual "interactions", leading to occasional discrepancies relative to results derived from other methods, such as imaging analysis [5,6]. Therefore, several "ligation-free" biochemical methods were developed to detect physical interactions between genomic regions. Genome architecture mapping (GAM) is a method to measure the statistical proximities of genomic regions by sequencing the DNA extracted from ultrathin cryosectioned nuclear slices [7]. Split-pool recognition of interactions by tag extension (SPRITE) is a method that involves performing repeated rounds of splitting and barcoding of individual chromatin complexes followed by identification of the interacting genomic regions by matching the barcodes [8]. Chromatin interaction analysis via droplet-based

and barcode-linked sequencing (ChIA-Drop) tracks amplicons arising from gel-bead-in-emulsion (GEM) droplets of each chromatin complex by barcode sequencing [9].

We recently developed a method to search for genomic regions interacting with a locus of interest in a non-biased manner that combines pull-down of the locus using engineered DNA-binding molecule-mediated chromatin immunoprecipitation (enChIP) and next-generation sequencing (NGS) analysis (enChIP-Seq). The CRISPR system, consisting of a nuclease-dead form of Cas9 (dCas9) and a guide RNA (gRNA), or transcription activator-like (TAL) proteins, can be used for enChIP. In enChIP-Seq, it is necessary to compare multiple datasets of enChIP-Seq data to unambiguously detect specific interactions. However, it is not always easy to analyze enChIP-Seq datasets to subtract non-specific interactions or identify common interactions. To facilitate such analysis, we developed the enChIP-Seq analyzer software. It enables easy extraction of common signals as well as subtraction of non-specific signals observed in negative control samples, thereby streamlining extraction of specific enChIP-Seq signals. enChIP-Seq analyzer will help users analyze enChIP-Seq data and identify physical interactions between genomic regions.

## 2. Materials and Methods

### 2.1. Implementation

enChIP-Seq analyzer is provided as free software at GitHub (https://github.com/TKY-SE/enChIP-Seq-Analyzer, accessed on 1 December 2021). All the program was executed on a computer with Intel® Core™ i5-7500 CPU @ 3.40 GHz 3.41 GHz, and 8 GB of RAM. Preparation of data sets, mode of analysis, and system handling for enChIP-Seq analyzer are shown in a step-by-step manner below.

### 2.2. Procedures for enChIP-Seq

The procedures to use enChIP-Seq to detect genomic regions interacting with a locus of interest depend on how the locus is tagged with an engineered DNA-binding molecule [10,11]. In "in cell" enChIP-Seq, an engineered DNA-binding molecule, such as a CRISPR complex, is expressed in the cells to be analyzed. After cross-linking with formaldehyde or other cross-linkers, if necessary, chromatin is fragmented using sonication or enzymatic digestion. Subsequently, the tagged locus is isolated using affinity purification, and the interacting genomic regions are identified by NGS [10]. In in vitro enChIP-Seq, chromatin is cross-linked, if necessary, fragmented, and then incubated with the CRISPR complex, consisting of a recombinant dCas9 protein and a synthetic or in vitro transcribed gRNA, for in vitro locus tagging. Subsequently, the tagged locus is isolated using affinity purification, and the interacting genomic regions are identified by NGS [11].

The CRISPR complex also binds to off-target sites in addition to the target genomic regions [12–15]. Therefore, to identify genomic regions truly interacting with the target site, it is necessary to analyze negative control samples, such as cells expressing only dCas9 or dCas9 plus irrelevant gRNAs. In addition, multiple gRNAs for the target locus should be analyzed to extract common peaks from the NGS analysis, which can then be identified as true positives with higher confidence. Consequently, it is necessary to compare multiple negative controls and multiple samples with on-target gRNAs. It is tedious to compare these samples manually, and this observation prompted us to develop a software program to automate comparison of multiple NGS peaks.

### 2.3. Filtering of the NGS Peaks to Identify Genomic Regions Interacting with a Target Locus

enChIP-Seq analyzer is a software program that enables easy consolidation of enChIP-Seq data, including extraction of common peaks observed for different gRNAs bound to the target genomic region and subtraction of peaks in negative control samples (Figure 1).

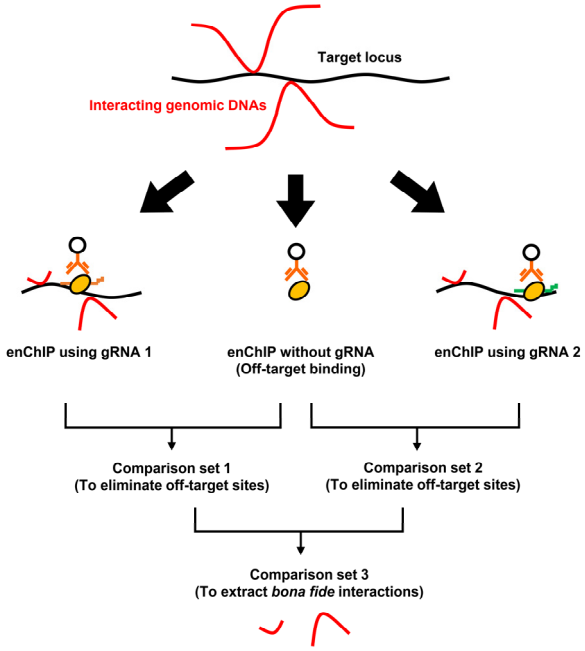


**Figure 1.** An example of using enChIP-Seq to identify interacting genomic regions. enChIP-Seq was performed in the presence of different gRNAs or in the absence of a gRNA (a negative control). Peaks detected by enChIP without gRNA were subtracted from those for enChIP-Seq with gRNAs to eliminate off-target sites (Comparison sets 1 and 2). Subsequently, the resultant data could be compared to extract *bona fide* interactions.

### *2.4. Preparation of the Data Set*

To extract enChIP-specific NGS peaks, Model-based Analysis of ChIP-Seq (MACS) [16] is used to compare NGS data from enChIP with that for input DNA. MACS data for enChIP-specific NGS peaks is exported as a tab file (Figure 2), which is used by enChIP-Seq analyzer.

```
# This file is generated by MACS version 1.4.2 20120305
# ARGUMENTS LIST:
# name = MACS14_HS5_6_NaB_IP
# format = BED
# ChIP-seq file = /illumina_disk2/runs3/fujita_ChIP/140318_Hiseq3A/HS5_6_NaB_IP_I7_20_bed
# control file = /illumina_disk2/runs3/fujita_ChIP/140318_Hiseq3A/HS5_6_NaB_input_I8_08_bed
# effective genome size = 2.70e+09
# band width = 300
# model fold = 10,30
# pvalue cutoff = 1.00e-05
# Large dataset will be scaled towards smaller dataset.
# Range for calculating regional lambda is: 1000 bps and 10000 bps
# tag size is determined as 35 bps
# total tags in treatment: 25696905
# tags after filtering in treatment: 23889895
# maximum duplicate tags at the same position in treatment = 1
# Redundant rate in treatment: 0.07
# total tags in control: 33148962
# tags after filtering in control: 31888685
# maximum duplicate tags at the same position in control = 1
# Redundant rate in control: 0.04
# d = 35
chr start end length summit tags -10*log10(pvalue) fold_enrichment FDR(%)
chr1 2053504 2053594 91 24 8 67.20 22.60 1.79
chr1 2298994 2299079 86 52 7 50.95 10.17 3.67
chr1 2399589 2399712 124 84 9 65.88 12.92 1.81 8165:NM_014638:PLCH2|
chr1 5646669 5646769 101 48 9 72.06 18.45 1.70
chr1 5731697 5731849 153 79 16 54.61 6.02 3.14
chr1 7347532 7347648 117 64 10 78.46 19.37 1.66
chr1 7347671 7347770 100 67 8 63.83 19.37 1.88
chr1 7626357 7626443 87 22 7 58.10 19.37 2.91
chr1 7812202 7812293 92 30 10 66.14 13.35 1.84
chr1 9292338 9292423 86 31 7 56.42 15.13 2.91 2525:NM_004285:H6PD|
chr1 10286082 10286164 83 60 10 75.07 14.83 1.72
chr1 11072883 11072966 84 32 7 53.84 16.34 3.48 -204:NM_007375:TARDBP|
chr1 11610725 11611206 482 216 78 580.85 29.87 0.00
chr1 11696034 11696148 115 57 8 57.17 18.45 3.03
chr1 11919451 11919604 154 64 14 90.93 14.86 1.05 613:NM_002521:NPPB|
```

A

B

**Figure 2.** An example of a tab file used in enChIP-Seq analyzer. (**A**) Lines starting with # are comments and are not processed by the software. (**B**) Lines separated by tags are used for calculations in enChIP-Seq analyzer.

### *2.5. Mode of Analysis*

enChIP-Seq analyzer extracts common peak information or eliminates negative peak information, as shown in Figure 3.

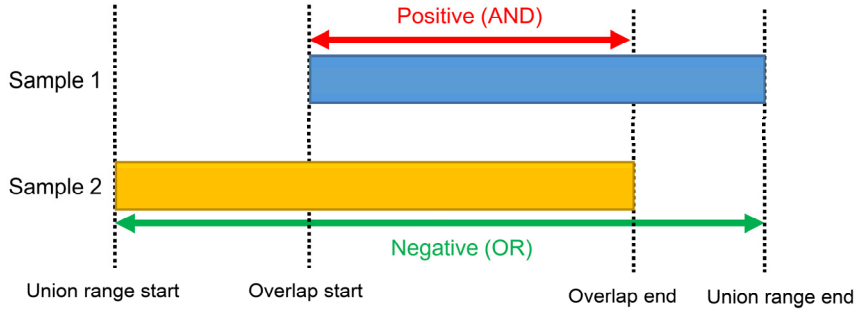


**Figure 3.** Mode of analysis. enChIP-Seq analyzer extracts overlapped regions (Positive (AND)) as common peak information or eliminates union range regions (Negative (OR)) as negative peak information.

Extraction of common peak information. When two or more tab files are analyzed, overlapped peak regions (red double-headed arrow) are extracted as common peak information (region).

Elimination of negative peak information. When two or more tab files are analyzed, union range regions (green double-headed arrow) are eliminated as negative peak information (region).

### *2.6. Handling of the System*

enChIP-Seq analyzer, a Java-based software program, is designed for use with Microsoft Windows 10. The software can be downloaded from GitHub (https://github.com/TKY-SE/enChIP-Seq-Analyzer, accessed on 1 December 2021). The browser for the software is shown in Figure 4A. The detailed procedures are outlined below.

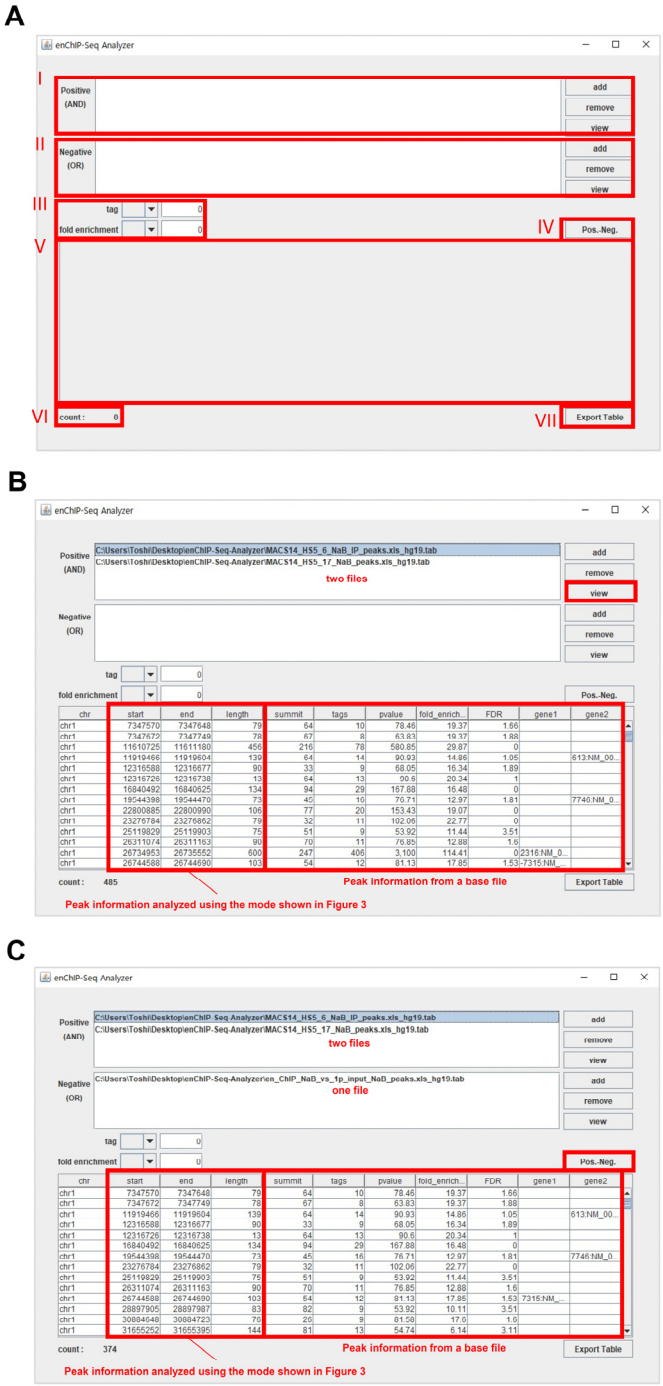


**Figure 4.** The main screen of enChIP-Seq analyzer. (**A**–**C**) Examples of the screens of enChIP-Seq analyzer. Handling processes are shown in the main text.

2.6.1. Extraction of Peak Information Commonly Included in Two or More Tab Files

(1-1) Add tab files that include enChIP-specific NGS peaks in the field "I". To this end, click on "add" to open a selection window and select a tab file from your computer. Repeat this step to add additional tab files.

(1-2) (If necessary) To remove an added file, select a file(s) and click on "remove".

(1-3) (Mandatory) Select an added file as a base file, which is then highlighted.

(1-4) To execute the program, click on "view".

(1-5) The result is shown in "V", which is based on information from the base file. The number of common peaks is shown in "VI" as "count". An example is shown in Figure 4B.

#### 2.6.2. Elimination of Negative Peak Information

(2-1) Perform steps (1-1) through (1-3). For this analysis, use of only one tab file is also acceptable.

(2-2) Add a tab file(s) that includes NGS peaks with negative peak information (e.g., NGS peaks from enChIP experiments performed without a gRNA) in the field "II". To this end, click on "add" to open a selection window and select a tab file(s) from your computer. Repeat this step to add two or more tab files.

(2-3) (If necessary) To remove an added file, select a file(s) and click on "remove" in "II".

(2-4) (Mandatory) Select an added file in "I" as a base file, which is then highlighted.

(2-5) To execute the program, click on "Pros.-Neg. (IV)".

(2-6) The result is shown in "V", which is based on the information from the base file. The number of peaks remaining after elimination of negative peak information is shown in "VI" as "count". An example is shown in Figure 4C.

(2-7) (If necessary) If tag number and/or fold enrichment are necessary criteria for filtering, set tag number and/or fold enrichment criteria in "III". The following symbols can be used for filtering: $\geq$, =, and $\leq$.

#### 2.6.3. Data Export

Data shown in "V" can be exported as a csv file and easily utilized by other software/web tools. To this end, click on "Export Table (VII)". Select a folder to save the file from the opened window.

#### 2.6.4. View Information within a Tab File

(1) The information within a tab file can be viewed directly in the software program. To this end, add one tab file that includes enChIP-specific NGS peaks in the field "I" (or "II").
(2) Click on "view" in the field "I" (or "II") to view the information within the file.
(3) The information is shown in "V".

## 3. Results and Discussion

We previously identified genomic regions associated with the 5′HS5 locus on a genome-wide scale in K562 cells under undifferentiated or sodium butyrate (NaB)-mediated differentiated conditions [10]. In that study, we performed enChIP using gRNAs #6 and #17, which targeted the 5′HS5 locus and used the purified DNA for NGS analysis. As a negative control, we also performed enChIP in the absence of a gRNA and used the purified DNA for NGS analysis. Next, we compared the NGS peaks from each enChIP experiment to those for input DNA and used MACS analysis to extract the list of enriched peaks ("enChIP #6 peaks", "enChIP #17 peaks", and "Off-target sites") (Data sets in Figure 5A). Using enChIP-Seq analyzer, we compared the lists of extracted peaks in a step-by-step manner as follows (Figure 5A).

Step 1:

(1) We compared the data from "enChIP #6 peaks" and "Off-target sites" to eliminate off-target binding sites. The resultant information was named "enChIP #6-specific sites".
(2) We compared the data from "enChIP #17 peaks" and "Off-target sites" to eliminate off-target binding sites. The resultant information was named "enChIP #17-specific sites".
(3) To identify peaks with confidence, we adopted two criteria for choosing peaks based on NGS information from the target 5′HS5 locus: (i) Tag number $\geq$5% of that of

the target 5′HS5 locus (which can be considered as an interacting ratio of ≥5%) and (ii) fold enrichment relative to input genomic DNA ≥10. In this regard, 19 and 228 peaks for "enChIP #6-specific sites" and "enChIP #17-specific sites", respectively, passed the two criteria.

Step 2:

We compared the data from "enChIP #6-specific sites" and "enChIP #17-specific sites" to extract "enChIP #6/#17-common sites". We extracted six peaks, which can be considered *bona fide* physically interacting genomic regions (Figure 5B). These results were consistent with those from our previous study [10] (please note that in Ref. [10], the symbols > should have been ≥).

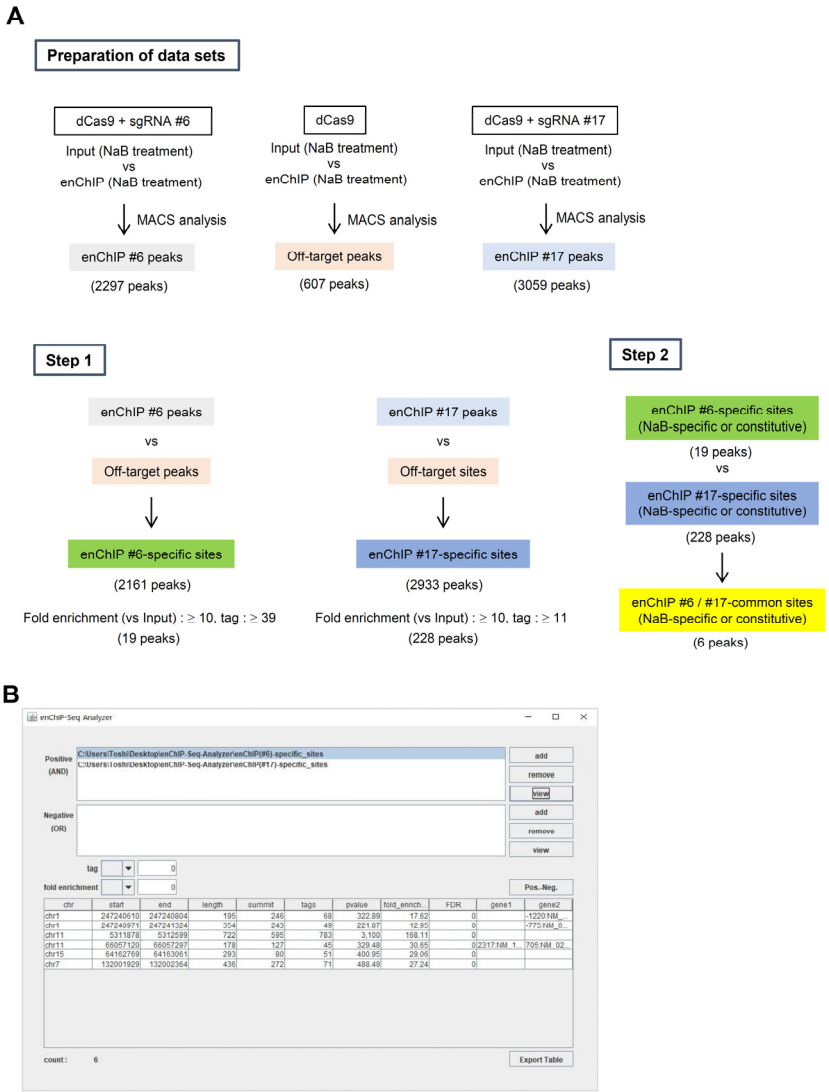


**Figure 5.** An example of results from enChIP-Seq analyzer. (**A**) Step-by step procedures to extract *bona fide* genomic regions interacting with a target genomic region. (**B**) An example of the results from enChIP-Seq analyzer.

## 4. Conclusions

In this study, we developed enChIP-Seq analyzer to compare multiple enChIP-Seq datasets to unambiguously detect specific interactions. enChIP-Seq analyzer is simple and easy to use. We believe that enChIP-Seq analyzer will help users to analyze enChIP-Seq data and detect physical interactions between genomic regions. In addition, the software can be used to compare ChIP-Seq datasets to extract common peaks.

## 5. Availability and Requirements

Project name: enChIP-Seq analyzer
Software homepage (GitHub): https://github.com/TKY-SE/enChIP-Seq-Analyzer, accessed on 1 December 2021
Programming language: Java
Other requirements: Windows machine
License: None
Any restrictions to use by non-academics: None.
Notes: We have a plan to adapt the software for Linux and Mac. As soon as they are available, we will upload the code in GitHub.

**Author Contributions:** T.N. and A.S. developed and used the software. T.F. curated the data and wrote the manuscript. H.F. conceived and supervised the project, and wrote the manuscript. All authors have read and approved the final manuscript, and ensure that this is the case.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The software is available at: https://github.com/TKY-SE/enChIP-Seq-Analyzer, accessed on 1 December 2021. The data is publicly available and described at our previous paper [10].

**Acknowledgments:** We thank Naohisa Goto at Research Institute for Microbial Diseases, Osaka University, for valuable comments and discussion.

**Conflicts of Interest:** T.F. and H.F. are inventors of granted patents and a patent pending that are owned by Osaka University for the technology of purification and subsequent analysis of specific types of DNA, including genomic DNA with chromatin structure, using an engineered DNA-binding molecule, including the CRISPR complex binding to the target DNA (patent name: "Method for isolating specific genomic regions using molecule binding specifically to endogenous DNA sequence". Patent numbers: Japan 5,954,808 and EP 2,963,113. Patent application number: WO2014/125668). T.F. and H.F. are co-founders of Epigeneron, Inc. and own stock in the company. H.F. is one of the directors of Epigeneron, Inc. A.S. and T.N. are employees of Tohoku Chemical Co. Ltd. Tohoku Chemical Co. Ltd. played no role in this study.

## References

1. Misteli, T. The Self-organizing genome: Principles of genome architecture and function. *Cell* **2020**, *183*, 28–45. [CrossRef] [PubMed]
2. Krumm, A.; Duan, Z. Understanding the 3D genome: Emerging impacts on human disease. *Semin. Cell Dev. Biol.* **2019**, *90*, 62–77. [CrossRef] [PubMed]
3. Dekker, J.; Rippe, K.; Dekker, M.; Kleckner, N. Capturing chromosome conformation. *Science* **2002**, *295*, 1306–1311. [CrossRef] [PubMed]
4. de Wit, E.; de Laat, W. A decade of 3C technologies: Insights into nuclear organization. *Genes Dev.* **2012**, *26*, 11–24. [CrossRef] [PubMed]
5. Williamson, I.; Berlivet, S.; Eskeland, R.; Boyle, S.; Illingworth, R.S.; Paquette, D.; Dostie, J.; Bickmore, W.A. Spatial genome organization: Contrasting views from chromosome conformation capture and fluorescence in situ hybridization. *Genes Dev.* **2014**, *28*, 2778–2791. [CrossRef] [PubMed]
6. Finn, E.H.; Pegoraro, G.; Brandão, H.B.; Valton, A.L.; Oomen, M.E.; Dekker, J.; Mirny, L.; Misteli, T. Extensive heterogeneity and intrinsic variation in spatial genome organization. *Cell* **2019**, *176*, 1502–1515.e10. [CrossRef] [PubMed]
7. Beagrie, R.A.; Scialdone, A.; Schueler, M.; Kraemer, D.C.A.; Chotalia, M.; Xie, S.Q.; Barbieri, M.; de Santiago, I.; Lavitas, L.M.; Branco, M.R.; et al. Complex multi-enhancer contacts captured by genome architecture mapping. *Nature* **2017**, *543*, 519–524. [CrossRef] [PubMed]
8. Quinodoz, S.A.; Ollikainen, N.; Tabak, B.; Palla, A.; Schmidt, J.M.; Detmar, E.; Lai, M.M.; Shishkin, A.A.; Bhat, P.; Takei, Y.; et al. Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus. *Cell* **2018**, *174*, 744–757.e24. [CrossRef]
9. Zheng, M.; Tian, S.Z.; Capurso, D.; Kim, M.; Maurya, R.; Lee, B.; Piecuch, E.; Gong, L.; Zhu, J.J.; Li, Z.; et al. Multiplex chromatin interactions with single-molecule precision. *Nature* **2019**, *566*, 558–562. [CrossRef] [PubMed]
10. Fujita, T.; Yuno, M.; Suzuki, Y.; Sugano, S.; Fujii, H. Identification of physical interactions between genomic regions by enChIP-Seq. *Genes Cells* **2017**, *22*, 506–520. [CrossRef] [PubMed]

11. Fujita, T.; Kitaura, F.; Yuno, M.; Suzuki, Y.; Sugano, S.; Fujii, H. Locus-specific ChIP combined with NGS analysis reveals genomic regulatory regions that physically interact with the Pax5 promoter in a chicken B cell line. *DNA Res.* **2017**, *24*, 537–548. [CrossRef] [PubMed]
12. Cencic, R.; Miura, H.; Malina, A.; Robert, F.; Ethier, S.; Schmeing, T.M.; Dostie, J.; Pelletier, J. Protospacer adjacent motif (PAM)-distal sequences engage CRISPR Cas9 DNA target cleavage. *PLoS ONE* **2014**, *9*, e109213. [CrossRef] [PubMed]
13. Kuscu, C.; Arslan, S.; Singh, R.; Thorpe, J.; Adli, M. Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. *Nat. Biotechnol.* **2014**, *32*, 677–683. [CrossRef] [PubMed]
14. Wu, X.; Scott, D.A.; Kriz, A.J.; Chiu, A.C.; Hsu, P.D.; Dadon, D.B.; Cheng, A.W.; Trevino, A.E.; Konermann, S.; Chen, S.; et al. Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nat. Biotechnol.* **2014**, *32*, 670–676. [CrossRef] [PubMed]
15. O'Geen, H.; Henry, I.M.; Bhakta, M.S.; Meckler, J.F.; Segal, D.J. A genome-wide analysis of Cas9 binding specificity using ChIP-seq and targeted sequence capture. *Nucleic Acids Res.* **2015**, *43*, 3389–3404. [CrossRef] [PubMed]
16. 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. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **2008**, *9*, R137. [CrossRef] [PubMed]