Article

HiMoRNA: A Comprehensive Database of Human lncRNAs Involved in Genome-Wide Epigenetic Regulation

Evgeny Mazurov [1], Alexey Sizykh [2] and Yulia A. Medvedeva [1,2,*]

[1] Institute of Bioengineering, Research Center of Biotechnology, Russian Academy of Sciences, 117312 Moscow, Russia; mazurovEV@gmail.com

[2] School of Biological and Medical Physics, Moscow Institute of Physics and Technology, Dolgoprudny, 141701 Moscow, Russia; sizykh.ad@phystech.edu

* Correspondence: ju.medvedeva@gmail.com

Abstract: Long non-coding RNAs (lncRNAs) play an important role in genome regulation. Specifically, many lncRNAs interact with chromatin, recruit epigenetic complexes and in this way affect large-scale gene expression programs. However, the experimental data about lncRNA-chromatin interactions is still limited. The majority of experimental protocols do not provide any insight into the mechanics of lncRNA-based genome-wide epigenetic regulation. Here we present the HiMoRNA (Histone-Modifying RNA) database, a resource containing correlated lncRNA–epigenetic changes in specific genomic locations genome-wide. HiMoRNA integrates a large amount of multi-omics data to characterize the effects of lncRNA on epigenetic modifications and gene expression. The current release of HiMoRNA includes more than five million associations in humans for ten histone modifications in multiple genomic loci and 4145 lncRNAs. HiMoRNA provides a user-friendly interface to facilitate browsing, searching and retrieving of lncRNAs associated with epigenetic profiles of various chromatin loci. Analysis of the HiMoRNA data suggests that several lncRNA including JPX might be involved not only in regulation of XIST locus but also in direct establishment or maintenance of X-chromosome inactivation. We believe that HiMoRNA is a convenient and valuable resource that can provide valuable biological insights and greatly facilitate functional annotation of lncRNAs.

Keywords: lncRNA; histone modification; database; X inactivation

1. Introduction

Transcription in higher organisms is pervasive and produces a wide variety of non-coding RNAs [1], with long non-coding RNAs (lncRNAs) being one of the most populated classes. Classically, lncRNAs are defined as transcripts longer than 200 nt in length lacking protein-coding capacity. LncRNAs are usually lowly expressed, highly tissue-specific and rarely evolutionary conserved leading to difficulties in functional annotation [2,3]. However, lncRNAs are heavily regulated [4,5] which supports their functionality. Indeed, it has been shown that lncRNAs function via surprisingly diverse molecular mechanisms [6,7] and have a vital role in various biological processes [8,9].

Many lncRNAs interact with chromatin and are essential for epigenetic control of specific genome loci as well as organization of the chromosomes [10–13]. Given that, identification of chromatin-interacting lncRNAs and their genome targets could be a critical first step for dissecting lncRNA function in epigenetic regulation. Many low-throughput experiments aimed at investigating lncRNA–chromatin interactions have been performed over the years [14–16]. With advances in next-generation sequencing, several high-throughput techniques (ChIRP-seq [17], CHART-seq [18], MARGI [19], GRID-seq [20], ChAR-seq [21], RADICL-seq [22], Red-C [23], have been employed to detect the binding sites of chromatin-interacting lncRNAs genome-wide. Experimentally validated lncRNA–chromatin interactions that could facilitate the understanding of the functional role of lncRNAs have been integrated in LnChrom [24], suggesting that more than 2300 lncRNAs could interact

with the chromatin in at least one cell type. However, high-throughput methods suffer from a high level of false-positives, since sporadic contacts could be also detected in the experiments. These methods also have issues with detecting contacts for lowly expressed lncRNAs. Additionally, high-throughput data on lncRNA-chromatin interactions is limited to a specific cell type and does not uncover a mechanism of lncRNA in regulation of histone modifications.

To address this problem computationally we developed an approach to search for lncRNAs whose presence in the cell is significantly correlated with the presence of epigenetic modification in a specific genome locus. Here we report the HiMoRNA (Histone-Modifying RNA) database, a resource containing coordinated lncRNA–epigenetic changes genome-wide. HiMoRNA integrates a large amount of multi-omics data to characterize the effects of lncRNA on epigenetic modifications and gene expression. The current release of HiMoRNA includes 5,640,250 associations in humans for the ten most studied histone modifications and 4,145 lncRNA. HiMoRNA provides a user-friendly interface to facilitate browsing, searching and retrieving of lncRNAs associated with specific chromatin loci. Analysis of the HiMoRNA data suggests that several lncRNA, including JPX, might be involved not only in regulation of XIST locus but also in direct establishment or maintenance of X-chromosome inactivation. We believe that HiMoRNA is a convenient and valuable resource that can provide valuable biological insights and greatly facilitate functional annotation of lncRNAs.

Availability: http://himorna.fbras.ru (accessed on 19 November 2021).

2. Materials and Methods

All analyses were performed using human genome build hg38 (https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.26/, accessed on 19 November 2021) and GEN-CODE gene annotation version 31 (https://www.gencodegenes.org/human/release_31.html, accessed on 19 November 2021).

2.1. Histone Modification Data

We downloaded bam files (ChIP-seq, signals and controls) for ten histone modification marks (H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K9me3, H3K79me2, H4K20me1) having the highest number of biosamples (Supplementary Table S1) from the ENCODE portal [25,26]. We used only those biosamples for which RNA-seq data was also available. Using the NCIS package [27], we estimated a normalization factor for each ChIP-seq experiment and called peaks with MACS2 [28] (FDR-corrected *p*-value < 0.05) for matching pairs of signal-control experiments for the same biosample. ChIP-seq peaks detected in all biosamples for one histone modification were merged using bedtools [29] to standardize the peak boundaries across different cell types. Coordinates of the merged peaks were used as a universal genome annotation for this histone mark and read counts were obtained within each region of the universal annotation for each biosample using bedtools. Read counts were normalized using DESeq2 [30]. We associated peaks to genes from the gencode.v31.annotation if a peak was located within 1000 bp to a gene with ChIPpeakAnno [31] (Supplementary Figure S1).

2.2. Transcriptome Data

We downloaded RNA-seq data (gene counts) from the ENCODE portal for the same biosamples for which ChIP-seq data was used. Genes counts were normalized using DE-Seq2 separately for each group of RNA-seq samples matching ChIP-seq samples for every histone modification. As a list of lncRNAs we used gencode.v31.long_noncoding_RNAs and normalized the gene expression of each dataset with DESeq2. For each histone mark we further analyzed only a set of lncRNAs expressed in at least ten biosamples.

2.3. Correlation Analysis

To select lncRNAs potentially involved in epigenetics regulation of a specific locus across multiple cell types we calculated the Spearman correlation coefficient (SCC) between

all lncRNAs expression vectors and all ChIP-Seq peaks over all samples available for every histone modification. Only significant correlations (BH correction, $FDR < 0.05$) for each lncRNA are provided in the database.

Per chromosome distribution of correlated ChIP-seq peaks, for each lncRNA and histone modification in the database we tested if correlated peaks are non-uniformly distributed across different chromosomes by performing Fisher exact test for peaks on one chromosome against all others (BH correction, $FDR < 0.05$).

3. Results

3.1. Database Statistics

The current release of HiMoRNA contains 5,640,250 significantly correlated lncRNA-genomic loci in humans for ten histone modifications and 4145 lncRNA. In order to characterize the records of lncRNA-genome locus we also gathered a series of metadata for each interaction.

3.2. Web Interface

Data browsing, searching and retrieving HiMoRNA provides a user-friendly web interface that allows users to browse, query and download the lncRNA–locus correlation data Figure 1. In order to provide easy access we added a choice of filters for fast browsing and searching. The user can choose histone modification(s), lncRNA name(s), gene names, coordinates of the regions (in bed format), the SCC sign and value allowing users to perform a customized search for the contents of interest. The user is required to select at least a histone modification but to make the search faster we recommend selecting one or only a few RNAs or uploading the genomic regions. The server will return the related lncRNA-loci pairs. The user can also download the whole database content from the home page as well as the results of the custom search from the corresponding page.
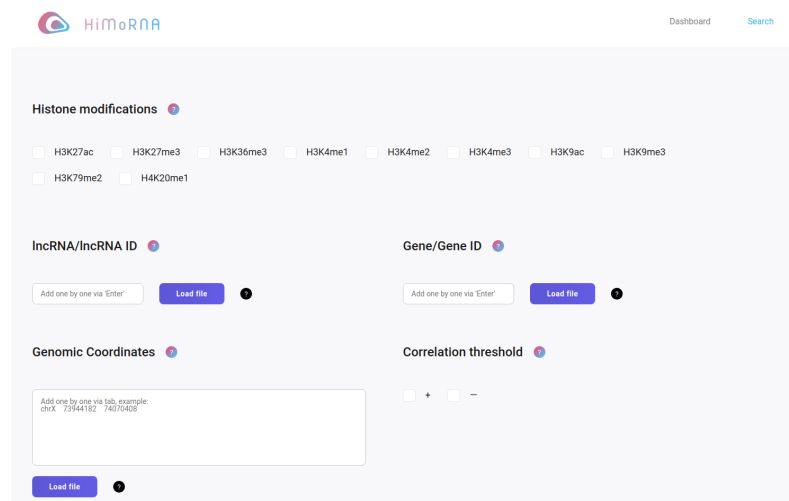

Figure 1. HiMoRNA web-interface. Users can query the resource through selecting lncRNA, gene, genomic regions and setting a threshold for the sign and an absolute value of an SCC.

The initial output of the browsing or searching contains basic information about each correlated lncRNA-locus and is comprised of the histone modification name, lncRNA name, chromosome, start and end of the correlated loci (ChIP-seq peak), target gene name, and SCC value. HiMoRNA also provides a download button allowing users to retrieve all of the interaction data.

To further explore the interaction of interest, users can click on a histone modification, an lncRNA, a gene or a correlation in the table to launch a new page that contains the corresponding metadata and visualizations. LncRNA page provides information about its expression in all cell types used in HiMoRNA, per chromosome distribution of correlated

peaks for each histone modification separately and a summary table of all peaks correlated with this lncRNA. The histone modification page provides information about chromosomal distribution of ChIP-seq peaks and a table of all lncRNA correlated with at least one peak of this modification. The gene page provides information about its expression in all cell types used in HiMoRNA, a table of all histone modification peaks associated with the gene and a summary table of all peaks associated with this gene and at the same time correlated with any lncRNA. The correlation page provides detailed information about how this correlation has been obtained, including lncRNA expression and ChIP-seq peak signal values in all the cell types used to calculate SCC.

3.3. Implementation

LncRNA-locus correlations and the results of multi-omics data set analysis were stored and queried by neo4j. The web interface was implemented in JavaScript and Vue and has been tested on Firefox, Chrome and Safari.

Source code and database dumps in different formats are available at https://github.com/lab-medvedeva/himorna-frontend (accessed on 19 November 2021) (platform frontend) and https://github.com/lab-medvedeva/himorna-backend (accessed on 19 November 2021) (platform backend).

4. Discussion

HiMoRNA provides a comprehensive collection of lncRNA–genomic loci for which lncRNA expression is significantly correlated with the histone modification signal across multiple cell and tissue types. HiMoRNA currently contains over five million correlations for ten histone modifications and 4145 lncRNA in humans. For each lncRNA, histone modification and gene we summarized the metadata to facilitate the study of lncRNAs function in epigenetic regulation. HiMoRNA also provides a user-friendly web interface for customized searching, browsing and accessing the interaction data.

Deeper investigation of per chromosome distribution of ChIP-seq peaks associated with specific lncRNAs demonstrated significant nonuniformity. For seven out of ten studied histone modifications, at least one lncRNA demonstrated an enrichment of correlated ChIP-seq peaks at one of the chromosomes. The most striking cases include lncRNA JPX and TSIX for which the X-chromosome contains the majority of the correlated peaks for H3K9me3 and H3K27me3, respectively (Figure 2), while the length-normalized number of peaks for each modification is not dramatically different (SF9).
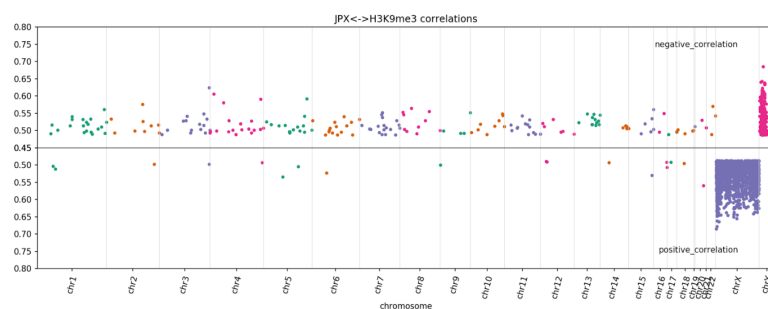

Figure 2. Per chromosomal distribution of lncRNA-correlated peaks of H3K9me3 for JPX.

X-chromosome inactivation is an essential epigenetic process which is regulated by several lncRNAs, including XIST, the master regulator of X-inactivation initiation. Other lncRNAs have also been implicated in regulating XIST, including TSIX, JPX, and FTX. Tian et al. [32] proposed that XIST is controlled by two RNA switches: TSIX, which represses XIST on the active X chromosome, and JPX, which activates XIST on the inactive X chromosome. Our pipeline shows a strong link between TSIX and H3K27me3, as well as between JPX and H3K9me3. Both H3K27me3 and H3K9me3 are repressive marks and are known to contribute to different stages of X inactivation. Our observations suggest

that TSIX and JPX may also contribute not only to regulation of XIST per se but also to the repression of X-chromosome directly by affecting different histone modifications at different stages of repression.

It has been found recently that Jpx acts as a CTCF release factor and in this way affects a 3D structure of the genome by regulating anchor site usage [33]. CTCF is a DNA-binding protein with multiple roles in chromatin architecture and gene regulation [34]. Apart from being an architectural protein CTCF serves as a transcription factor. We hypothesize that the JPX-induced release of CTCF from promoters or enhancers, where this transcription factor serves as an activator, may contribute to the establishment of the repressed chromatin mark H3K9me3, observed in our study.

Finally, we observed several other lncRNAs significantly correlated with the ChIP-seq peaks of a particular histone modification on the X-chromosome (ST3) for which experimental evidence of participation in X-chromosome inactivation is still required.

All the above suggest a possible mechanistic explanation of the role of several lncRNAs in establishing and maintaining the epigenetic profile of inactive X-chromosome. We realize that correlation of the ChIP-seq peaks and lncRNA expression does not guarantee causal relationship between these two events. Still, we hypothesize that at least some of these lncRNA could participate in epigenetic complex formation or attract them to specific X-chromosome loci. This example shows nicely how HiMoRNA enables researchers to translate associations between lncRNA and histone modifications into novel biologic insights.

5. Future Development

The area of functional lncRNA annotation is moving fast. We are fully committed to the maintenance and update of HiMoRNA, and we will monitor new data on the role of lncRNA in epigenetic regulation and expand the database as data becomes available. Specifically, we will add experimental data on lncRNA-chromatin interactions available databases to support our predicted lncRNA-genomic loci interactions. We also will incorporate data on lncRNA knockdown effects of gene expression from FANTOM6 [35]. As more experimentally validated data become available, we hope to build several predictive models for lncRNA–histone modification. Furthermore, we will continue to improve the performance of our computer servers for storing and analysing the newly generated data. We expect that these continuous efforts and our commitment to develop and improve HiMoRNA will contribute to the understanding of lncRNA-mediated chromatin regulation.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/ncrna8010018/s1, Figure S1: Genes to histone modification peaks association. Table S1: An overview of the histone modification and expression data used in HiMoRNA.

Author Contributions: Conceptualization, Y.A.M.; methodology, Y.A.M. and E.M.; software, E.M. and A.S.; validation and formal analysis, E.M.; resources, E.M. and A.S.; data curation, E.M.; writing, Y.A.M. and E.M.; visualization, Y.A.M. and E.M.; supervision and project adminstration, Y.A.M.; funding acquisition, Y.A.M. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Carninci, P.; Kasukawa, T.; Katayama, S.; Gough, J.; Frith, M.C.; Maeda, N.; Oyama, R.; Ravasi, T.; Lenhard, B.; Wells, C.; et al. The transcriptional landscape of the mammalian genome. *Science* **2005**, *309*, 1559–1563. [CrossRef] [PubMed]
2. Cabili, M.N.; Trapnell, C.; Goff, L.; Koziol, M.; Tazon-Vega, B.; Regev, A.; Rinn, J.L. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **2011**, *25*, 1915–1927. [CrossRef] [PubMed]

3. Andersson, R.; Refsing Andersen, P.; Valen, E.; Core, L.J.; Bornholdt, J.; Boyd, M.; Heick Jensen, T.; Sandelin, A. Nuclear stability and transcriptional directionality separate functionally distinct RNA species. *Nat. Commun.* **2014**, *5*, 5336. [CrossRef] [PubMed]
4. Alam, T.; Medvedeva, Y.A.; Jia, H.; Brown, J.B.; Lipovich, L.; Bajic, V.B. Promoter analysis reveals globally differential regulation of human long non-coding RNA and protein-coding genes. *PLoS ONE* **2014**, *9*, e109443. [CrossRef]
5. Hon, C.C.; Ramilowski, J.A.; Harshbarger, J.; Bertin, N.; Rackham, O.J.L.; Gough, J.; Denisenko, E.; Schmeier, S.; Poulsen, T.M.; Severin, J.; et al. An atlas of human long non-coding RNAs with accurate 5′ ends. *Nature* **2017**, *543*, 199–204. [CrossRef] [PubMed]
6. Böhmdorfer, G.; Wierzbicki, A.T. Control of Chromatin Structure by Long Noncoding RNA. *Trends Cell Biol.* **2015**, *25*, 623–632. [CrossRef] [PubMed]
7. Jandura, A.; Krause, H.M. The New RNA World: Growing Evidence for Long Noncoding RNA Functionality. *Trends Genet.* **2017**, *33*, 665–676. [CrossRef] [PubMed]
8. Wapinski, O.; Chang, H.Y. Long noncoding RNAs and human disease. *Trends Cell Biol.* **2011**, *21*, 354–361. [CrossRef] [PubMed]
9. Morris, K.V.; Mattick, J.S. The rise of regulatory RNA. *Nat. Rev. Genet.* **2014**, *15*, 423–437. [CrossRef] [PubMed]
10. Goff, L.A.; Rinn, J.L. Linking RNA biology to lncRNAs. *Genome Res.* **2015**, *25*, 1456–1465. [CrossRef] [PubMed]
11. Khalil, A.M.; Guttman, M.; Huarte, M.; Garber, M.; Raj, A.; Rivea Morales, D.; Thomas, K.; Presser, A.; Bernstein, B.E.; van Oudenaarden, A.; et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 11667–11672. [CrossRef] [PubMed]
12. Grote, P.; Herrmann, B.G. The long non-coding RNA Fendrr links epigenetic control mechanisms to gene regulatory networks in mammalian embryogenesis. *RNA Biol.* **2013**, *10*, 1579–1585. [CrossRef]
13. Mondal, T.; Subhash, S.; Vaid, R.; Enroth, S.; Uday, S.; Reinius, B.; Mitra, S.; Mohammed, A.; James, A.R.; Hoberg, E.; et al. MEG3 long noncoding RNA regulates the TGF-β pathway genes through formation of RNA–DNA triplex structures. *Nat. Commun.* **2015**, *6*, 7743. [CrossRef] [PubMed]
14. Hu, X.; Feng, Y.; Zhang, D.; Zhao, S.D.; Hu, Z.; Greshock, J.; Zhang, Y.; Yang, L.; Zhong, X.; Wang, L.P.; et al. A functional genomic approach identifies FAL1 as an oncogenic long noncoding RNA that associates with BMI1 and represses p21 expression in cancer. *Cancer Cell* **2014**, *26*, 344–357. [CrossRef]
15. Marchese, F.P.; Huarte, M. Long non-coding RNAs and chromatin modifiers: their place in the epigenetic code. *Epigenetics* **2014**, *9*, 21–26. [CrossRef]
16. Wang, K.C.; Chang, H.Y. Molecular mechanisms of long noncoding RNAs. *Mol. Cell* **2011**, *43*, 904–914. [CrossRef] [PubMed]
17. Chu, C.; Qu, K.; Zhong, F.L.; Artandi, S.E.; Chang, H.Y. Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. *Mol. Cell* **2011**, *44*, 667–678. [CrossRef] [PubMed]
18. Simon, M.D.; Pinter, S.F.; Fang, R.; Sarma, K.; Rutenberg-Schoenberg, M.; Bowman, S.K.; Kesner, B.A.; Maier, V.K.; Kingston, R.E.; Lee, J.T. High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation. *Nature* **2013**, *504*, 465–469. [CrossRef]
19. Sridhar, B.; Rivas-Astroza, M.; Nguyen, T.C.; Chen, W.; Yan, Z.; Cao, X.; Hebert, L.; Zhong, S. Systematic Mapping of RNA-Chromatin Interactions In Vivo. *Curr. Biol.* **2017**, *27*, 610–612. [CrossRef] [PubMed]
20. Li, X.; Zhou, B.; Chen, L.; Gou, L.T.; Li, H.; Fu, X.D. GRID-seq reveals the global RNA–chromatin interactome. *Nat. Biotechnol.* **2017**, *35*, 940. [CrossRef] [PubMed]
21. Bell, J.C.; Jukam, D.; Teran, N.A.; Risca, V.I.; Smith, O.K.; Johnson, W.L.; Skotheim, J.M.; Greenleaf, W.J.; Straight, A.F. Chromatin-associated RNA sequencing (ChAR-seq) maps genome-wide RNA-to-DNA contacts. *Elife* **2018**, *7*, e27024. [CrossRef]
22. Bonetti, A.; Agostini, F.; Suzuki, A.M.; Hashimoto, K.; Pascarella, G.; Gimenez, J.; Roos, L.; Nash, A.J.; Ghilotti, M.; Cameron, C.J.F.; et al. RADICL-seq identifies general and cell type–specific principles of genome-wide RNA-chromatin interactions. *Nat. Commun.* **2020**, *11*, 1018. [CrossRef] [PubMed]
23. Gavrilov, A.A.; Zharikova, A.A.; Galitsyna, A.A.; Luzhin, A.V.; Rubanova, N.M.; Golov, A.K.; Petrova, N.V.; Logacheva, M.D.; Kantidze, O.L.; Ulianov, S.V.; et al. Studying RNA-DNA interactome by Red-C identifies noncoding RNAs associated with various chromatin types and reveals transcription dynamics. *Nucleic Acids Res.* **2020**, *48*, 6699–6714. [CrossRef] [PubMed]
24. Yu, F.; Zhang, G.; Shi, A.; Hu, J.; Li, F.; Zhang, X.; Zhang, Y.; Huang, J.; Xiao, Y.; Li, X.; et al. LnChrom: A resource of experimentally validated lncRNA-chromatin interactions in human and mouse. *Database* **2018**, *2018*, bay039. [CrossRef]
25. Maher, B. ENCODE: The human encyclopaedia. *Nature* **2012**, *489*, 46–48. [CrossRef]
26. Sloan, C.A.; Chan, E.T.; Davidson, J.M.; Malladi, V.S.; Strattan, J.S.; Hitz, B.C.; Gabdank, I.; Narayanan, A.K.; Ho, M.; Lee, B.T.; et al. ENCODE data at the ENCODE portal. *Nucleic Acids Res.* **2016**, *44*, D726–D732. [CrossRef] [PubMed]
27. Liang, K.; Keleş, S. Normalization of ChIP-seq data with control. *BMC Bioinform.* **2012**, *13*, 199. [CrossRef]
28. 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* **2008**, *9*, 1–9. [CrossRef] [PubMed]
29. Quinlan, A.R.; Hall, I.M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **2010**, *26*, 841–842. [CrossRef] [PubMed]
30. Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **2014**, *15*, 550. [CrossRef]
31. Zhu, L.J.; Gazin, C.; Lawson, N.D.; Pagès, H.; Lin, S.M.; Lapointe, D.S.; Green, M.R. ChIPpeakAnno: A Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinform.* **2010**, *11*, 237. [CrossRef]

32. Tian, D.; Sun, S.; Lee, J.T. The long noncoding RNA, Jpx, is a molecular switch for X chromosome inactivation. *Cell* **2010**, *143*, 390–403. [CrossRef] [PubMed]

33. Oh, H.J.; Aguilar, R.; Kesner, B.; Lee, H.G.; Kriz, A.J.; Chu, H.P.; Lee, J.T. Jpx RNA regulates CTCF anchor site selection and formation of chromosome loops. *Cell* **2021**, *184*, 6157–6173. [CrossRef] [PubMed]

34. Luan, J.; Xiang, G.; Gómez-García, P.A.; Tome, J.M.; Zhang, Z.; Vermunt, M.W.; Zhang, H.; Huang, A.; Keller, C.A.; Giardine, B.M.; et al. Distinct properties and functions of CTCF revealed by a rapidly inducible degron system. *Cell Rep.* **2021**, *34*, 108783. [CrossRef] [PubMed]

35. Ramilowski, J.A.; Yip, C.W.; Agrawal, S.; Chang, J.C.; Ciani, Y.; Kulakovskiy, I.V.; Mendez, M.; Ooi, J.L.C.; Ouyang, J.F.; Parkinson, N.; et al. Functional annotation of human long noncoding RNAs via molecular phenotyping. *Genome Res.* **2020**, *30*, 1060–1072. [CrossRef] [PubMed]