Article

The Cumulative Formation of R-loop Interacts with Histone Modifications to Shape Cell Reprogramming

Hanshuang Li [1,†], Chunshen Long [1,†], Yan Hong [1], Lemuge Chao [1], Yong Peng [2] and Yongchun Zuo [1,*]

[1] State Key Laboratory of Reproductive Regulation and Breeding of Grassland Livestock, College of Life Sciences, Inner Mongolia University, Hohhot 010070, China; lhshuang@mail.imu.edu.cn (H.L.); cslong@mail.imu.edu.cn (C.L.); 31709150@mail.imu.edu.cn (Y.H.); 32008182@mail.imu.edu.cn (L.C.)

[2] Department of Biochemistry and Molecular Biology, The University of Chicago, 929 East 57th Street, Chicago, IL 60637, USA; yongpeng@uchicago.edu

* Correspondence: yczuo@imu.edu.cn
† These authors contributed equally to this work.

Abstract: R-loop, a three-stranded RNA/DNA structure, plays important roles in modulating genome stability and gene expression, but the molecular mechanism of R-loops in cell reprogramming remains elusive. Here, we comprehensively profiled the genome-wide landscape of R-loops during cell reprogramming. The results showed that the R-loop formation on most different types of repetitive elements is stage-specific in cell reprogramming. We unveiled that the cumulative deposition of an R-loop subset is positively correlated with gene expression during reprogramming. More importantly, the dynamic turnover of this R-loop subset is accompanied by the activation of the pluripotent transcriptional regulatory network (TRN). Moreover, the large accumulation of the active histone marker H3K4me3 and the reduction in H3K27me3 were also observed in these R-loop regions. Finally, we characterized the dynamic network of R-loops that facilitates cell fate transitions in reprogramming. Together, our study provides a new clue for deciphering the interplay mechanism between R-loops and HMs to control cell reprogramming.

Keywords: R-loops; histone modifications; transcription regulation network; somatic cell reprogramming; cell fate transitions

1. Introduction

R-loop is a special three-stranded nucleic acid structure composed of a DNA:RNA hybrid double helix and a single-stranded DNA (ssDNA) molecule, which is ubiquitous in the whole organism. R-loops normally form during transcription [1], and were first found in the F1 phage in 1967 [2], and were subsequently characterized in *Escherichia coli* [3]. In mammals, the predominant R-loop formation shows a strong sequence preference, which usually occurs at multiple loci with high guanine–cytosine (GC) content, including the unmethylated CpG island [4] and strong G/C skew [calculated as $(G - C)/(G + C)$] promoters [5]. The formation events of R-loops also occur in conserved regions that are associated with specific epigenetic modification marks [6,7]. Several studies have revealed that R-loops play a part in many biological processes, such as DNA replication [8], chromatin modification [9], DNA damage response [10], and genomic stability [11–14]. The R-loops also play essential roles in transcriptional regulation [1,15], in which R-loops are inclined to co-localize with open chromatin and recruit transcription activation factor [16]. However, the role of R-loops in biological processes such as cell fate transition has not been fully studied.

Differentiated somatic cells can be reprogramed into induced pluripotent stem cells (iP-SCs) by forced expression of Yamanaka factors Oct4, Sox2, Klf4, and c-Myc (OSKM) [17,18]. Due to the molecular features of iPSCs being similar to those of embryonic stem cells

(ESCs) [19], the generation of iPSCs has a huge impact both on basic research and clinical applications [20,21]. Somatic cell reprogramming is a multilayered regulation process, in which key histone modifications (HMs) play a very important role for transcriptional regulation [22,23]. Extensive studies have shown that the activation of early transcriptional events in reprogramming are associated with extensive loss of histone H3 lysine 27 trimethylation (H3K27me3), which represents a general opening of the chromatin state [24]. Moreover, the deposition of the active histone marker H3 lysine 4 trimethylation (H3K4me3) enables a permissive chromatin environment for facilitating gene transcription [25,26]. Furthermore, a recent study revealed that R-loops act as epigenetic markers and coordinate with SOX2 in regulating reprogramming to pluripotency [27]. Moreover, R-loops have also been proven to be able to coordinate with other epigenetic marks, which not only contribute to the maintenance of pluripotency in human embryonic stem cells (hESCs), but also influence cell fate transition during multilineage differentiation [28].

Despite the fact that the important role of R-loops in gene regulation has been elucidated in detail, there is still a lack of general understanding about the formation dynamics of R-loops during reprogramming. In particular, the underlying mechanisms of R-loops driving cell fate transitions in reprogramming remain unexplored. In this study, we characterized the genome-wide formation of R-loops and found that the dynamic transition of R-loops was strongly linked to gene expression during the reprograming process. Moreover, the activation of the pluripotent transcriptional regulatory network requires the interplay between R-loops and histone modifications, in which the formation of R-loops is accompanied by the large accumulation of active histone marker H3K4me3, with a concomitant reduction in H3K27me3. In brief, our results reveal that there is collaborative action between R-loops and HMs for controlling cell fate transition during the reprogramming process.

2. Results

2.1. Stage-Specific R-loop Formation during the Reprograming Process

Based on the single-stranded DNA:RNA immunoprecipitation sequencing (ssDRIP-seq) data, we aimed to elucidate the characteristics of global R-loop formation in the process of mouse embryonic fibroblast (MEF, D0) reprogramming to iPSC, and found that R-loops showed a significant dynamic pattern during the reprogramming process. Stage-specific R-loop signals on genes were observed at each reprogramming stage (Figure 1A and Supplementary Table S1). By calculating the number of R-loop peaks and their covered genes in each indicated reprogramming point, we found that R-loop peaks were rapidly formed at the initial (D1) stage, and underwent extensive erasure in the D5 stage. Then, the R-loop peaks were gradually remodeled at the genome-wide regions until the iPSC stage reached more than 2.7 million peaks (Figure 1B). The same dynamic pattern was also observed in the number of target genes of the R-loop (Figure 1C). Next, we comparatively observed the formation preference of R-loops on different chromosomes; the results display that R-loop formation is a common event, and that their formation appears to be stabilized on different chromosomes (Figure 1D). By systematically analyzing the genome-wide R-loops in the process of reprogramming, we found that about 95% of the R-loops were mainly formed in promoter, gene body and intergenic regions. In the gene bodies in particular (including exon and intron regions), about 45% of R-loops were observed (Figure 1E), showing that R-loop formation in gene bodies is prevalent [6]. Furthermore, we observed that R-loop peaks have higher signals (FPKM, fragments per kb per million uniquely mapped reads) in the upstream regions of gene transcription start sites (TSS) and the downstream regions of transcription termination sites (TTS) (Figure 1F and Supplementary Table S2); this is consistent with the fact that R-loops play a role in all stages of gene expression, from transcription initiation to its termination [28].

Moreover, R-loops tend to occur on repetitive elements of the whole genome, and approximately 75% of R-loop peaks occupy repetitive elements during the whole reprogramming process (Figure 1G). Among them, the long-terminal repeat (LTR), long interspersed

nuclear elements (LINE), short interspersed nuclear elements (SINE) and simple repeat are given priority by R-loops at each reprogramming stage. Notably, the R-loop formation on most different types of repetitive elements is stage-specific.
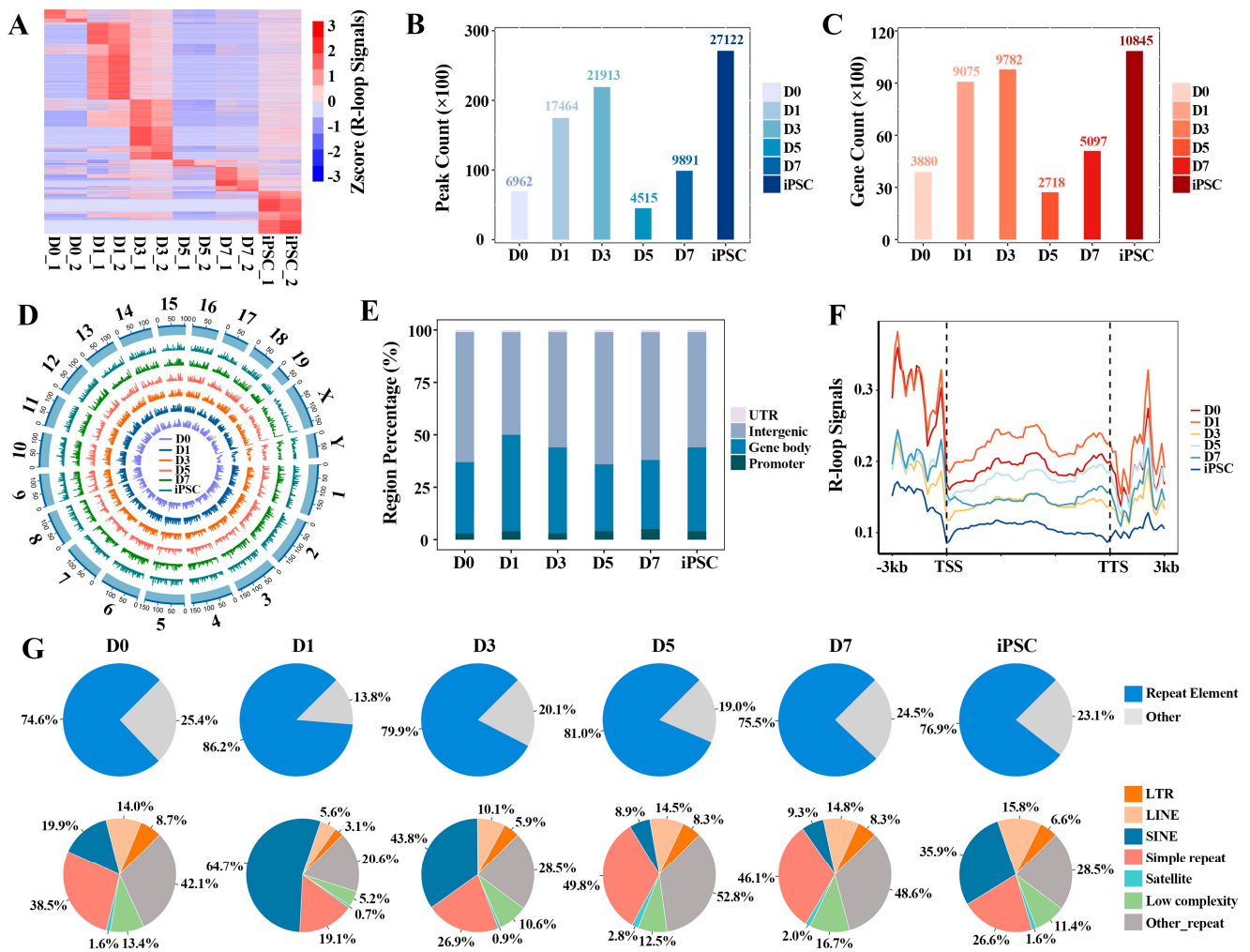
Figure 1. Genome-wide identification of R-loops during reprogramming: (**A**) heatmap displaying the global R-loop signals (FPKM) on genes at different reprogramming stages; (**B**) the number of R-loop peaks at different reprogramming stages; (**C**) the target genes' number of R-loop peaks at different reprogramming stages; (**D**) Circos plot showing the R-loop signals across all the chromosomes in reprogramming; (**E**) the distribution of R-loop peaks on different genomic features; (**F**) the average distribution of R-loop signals at 3 kb region flanking of TSS and transcription termination sites (TTS) during reprogramming progress; (**G**) pie plot showing the distribution of R-loop peaks in various types of repetitive elements.

To comprehensively explore whether the formation of R-loop has transcription factor (TF) family specificity (Supplementary Figure S1A), we counted the R-loop peaks observed in different TF families during the whole reprogramming process. It was found that R-loops were more inclined to appear in the nuclear receptor (NR) family than in the C2H2 family in reprogramming. Interestingly, about 60% of E-twenty six (ETS) family members may be regulated by R-loops in the iPSC stage, but the influence of R-loops on these TF families needs further exploration.

2.2. The Cumulative Effect of R-loops on Gene Expression Regulation in Reprogramming

More and more evidence reveal that R-loops are widespread in prokaryotic and eukaryotic organisms [1], and play an important role in many key biological processes [1,9,15,29].

However, whether the R-loop formation could affect the gene expression dynamics during reprogramming remains unclear. To address this, we collected the transcriptome data during the reprogramming process, and hierarchical clustering analysis showed that there was a high correlation between the duplicate samples. Principal-component analysis (PCA) of gene expression profiles allowed us to outline a "reprogramming pseudo time", which can clearly distinguish the reprogramming process into the MEF stage (D0), early reprogramming process (D1-D7) and iPSC phase (iPSC) (Figures 2A and S1B). By comparing the dynamic patterns of R-loop and gene expression during reprogramming, we found that the R-loop formation during reprogramming was more stage-specific than the gene expression pattern (Figure 2A), suggesting that R-loop levels and gene expression are not completely coupled in the process of reprogramming. Unlike R-loop dynamics, the genome-wide H3K4me3 and H3K27me3 signals shifted continuously during reprogramming until the specific pattern of iPSCs was formed (Supplementary Figure S1C,D). Moreover, we also noted that R-loops preferentially occurred within the upstream region of the transcription start site (TSS) compared with the other two histone modifications (H3K4me3 and H3K27me3) in this process, supporting previous observations [30] (Figure 2B).

In accordance with the above results, we next explored the relationship between R-loops and differential expression genes (DEGs) in the adjacent stages of reprogramming. The results revealed that the dynamic change pattern of R-loops within the genes was intense throughout the whole reprogramming process, whereas the prominent change wave of gene expression pattern only emerged between the iPSC and D7 (Figure 2C,D). These results suggest that the formation of R-loops may have a cumulative effect on gene expression. Among them, 58% (3689/6404) of genes including Nanog, Dppa4 and Jarid2 were up-regulated in the adjacent reprogramming stages, which was consistent with the up-regulated R-loop levels (Figures 2E and S2A), and 49% (3290/6668) down-regulated genes in the adjacent stages also had down-regulated R-loop signals (Figure 2F). Notably, these up-regulated genes (3689) with high R-loop levels were mainly characterized by signal pathways such as cell cycle, RNA transport and Stem cell pluripotency regulation, etc. (Supplementary Figure S2B), which coincided with the early event of mesenchymal epithelial transition (MET) in reprogramming. In contrast, the other down-regulated genes (3290) were mainly involved in Rap1, c-AMP and MAPK signal pathways, indicating that these genes play a part in the suppression of the lineage specific program (Supplementary Figure S2C).

During the whole process of MEF reprogramming to iPSC, the R-loop formation within gene regions was markedly different, and most of the formation events of R-loops had obvious stage-specific selection (Figure 2G). For example, 2% of the R-loop formation events only occurred in the MEF stage (D0) and 41% of the events had induced pluripotent stem cell (iPSC) specificity, which also explained why the kinetic pattern of R-loops was more intense in the adjacent reprogramming stages. Intriguingly, the related genes occupied by these stage-specific R-loops did not show obvious stage-specific expression in the whole reprogramming process, indicating that, in addition to R-loop formation, the expression of these stage-specific genes was also synergistically regulated by other epigenetic modifications (Figure 2G). Moreover, obvious stage-specific R-loop events were observed on these stage-specific expression genes, indicating that the expression of these genes is closely related to the stage-specific formation of R-loops (Figure 2H). Hence, the above results reveal that gene expression is related to the formation of R-loops to a certain extent, while the gene transcription process is the result of multilayered collaborative regulation.
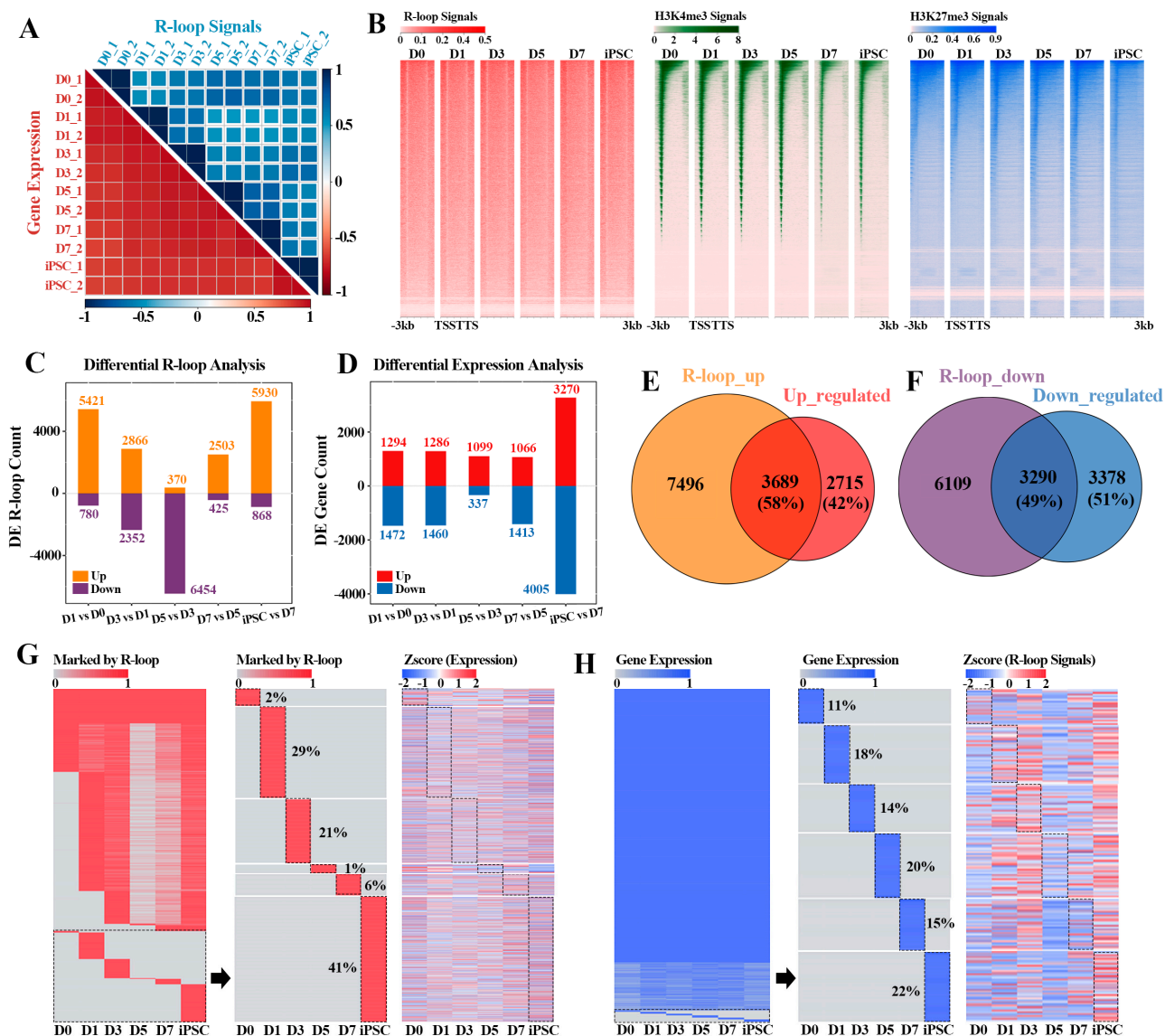
Figure 2. Reprogramming stage-specific R-loop modification and gene expression: (**A**) correlation heatmaps of different reprogramming stages according to R-loop signals (blue—top heatmap) and gene expression levels (red—bottom heatmap). Color scales represent Pearson correlation coefficients (PCC) between different stages; (**B**) heatmaps displaying the signals of R-loop, H3K4me3 and H3K27me3 in flanking TSSs at different reprogramming stages. In each panel, each row represents a 3 kb region flanking TSS; (**C**) analysis of differential R-loop signals in adjacent stages of reprogramming; (**D**) analysis of differential gene expression in adjacent stages of reprogramming; (**E**) Venn diagram revealing the overlap between genes with up-regulated R-loop signals (as shown in upper bar of (**C**)) and up-regulated genes (as shown in upper bar of (**D**)) between adjacent reprogramming stages; (**F**) Venn diagram showing the overlap between genes with down-regulated R-loop signals (as shown in lower bar of (**C**)) and down-regulated genes (as shown in lower bar of (**D**)) between adjacent reprogramming stages; (**G**) heatmaps showing the genes marked by R-loop peaks (left) during reprogramming progress, part of genes marked by R-loop peaks with reprogramming stage-specificity (middle) and the expression levels corresponding to these genes (right), respectively. Red—R-loop peaks; gray—no R-loop peaks; (**H**) heatmaps showing the dynamic expression profiles (left) of genes during reprogramming progress, part of gene expression with reprogramming stage-specificity (middle) and the R-loop signals corresponding to these genes (right), respectively. Blue—expression levels > 0; gray—no expression.

2.3. The Relationship between R-loop Dynamics and Gene Expression

To investigate the relationship between R-loop dynamics and associated gene activation during the acquisition of pluripotency, we integrally analyzed the relevant transcriptome and R-loop data of cell fate transition during reprogramming. Based on the differential R-loops analysis in iPSC compared with D0 (Figure 3A), 7084 genes with differential R-loop signals in iPSC were identified, in which, 92% genes had up-regulated R-loop signals (6497/7084). Moreover, 7845 differential expression genes (DEGs) were found in iPSC, including 3767 up-regulated and 4078 down-regulated genes (Figure 3B). Correspondingly, 2886 of these DEGs were directly related to R-loops (Figure 3C), in which the strength of the gene expression ($n = 1430$) correlated positively with the R-loop levels (Supplementary Table S3). In particular, 1271 up-regulated genes with high R-loop signals were identified. Functional analysis of these genes showed that they were mainly involved in stem cell pluripotency regulation and transcription-related signaling pathways, which were essential for iPSC pluripotency regulation and establishment (Figure 3C,D). Additionally, the down-regulated genes ($n = 159$) with low R-loop signals were highly enriched in MAPK, Hippo, RAS and other signaling pathways (Figure 3C,E). These results indicate that the dynamics of R-loops are closely correlated with gene expression, and play a regulatory role in the reprogramming of MEF to iPSC.

Fibroblast reprogramming to the pluripotent state is governed by multiple epigenetic marks, whereby active epigenetic modifications direct cell identity away from the fibroblast, and repressive epigenetic marks are inherited from the fibroblasts [31]. In order to further explore the influence of different histone modifications on these genes regulated by R-loops, we also calculated the signal intensities (Figure 3F) of active histone marker (H3K4me3) and inhibitory histone marker (H3K27me3) around the TSSs (TSS $\pm$ 2 kb) of DEGs associated with the R-loops. As expected, the average H3K27me3 signals decreased within the 2 kb of TSSs of these up-regulated genes with high R-loop signals. Likewise, the TSSs of down-regulated genes associated with low R-loop levels were accompanied by lower H3K4me3 signals and higher H3K27me3 signals. In conclusion, these results suggest that promoter R-loop formation is accompanied by the enhanced recruitment or depletion of unique epigenomic signatures, which is crucial for regulating the differential expression of genes in reprogramming.

2.4. Sequential Activation of Pluripotency Genes Linked to R-loops during Reprogramming

To gain insight into the mechanisms underlying the dynamic activation of the DEGs observed in iPSC compared to fibroblasts (D0), we profiled the sequential expression pattern of these genes. Based on the fuzzy C-means (FCM) clustering analysis, these genes can be categorized into six distinct clusters, labeled as Clusters 1 to 6 (C1–C6) (Figure 4A). These clusters can be further assigned to five main categories of different dynamic patterns. The genes in C1 ($n = 1817$) showed a transient increase at the iPSC stage, suggesting their key roles in pluripotency regulation. Alternatively, C3 ($n = 1619$) had an instantaneous decrease at the D0 stage, indicating that these genes are related to differentiation. Furthermore, the dynamic expression pattern of C2 ($n = 1220$) had an overall decreasing trend over the reprogramming process, while C4 ($n = 838$) had a dramatic increasing trend during reprogramming. Interestingly, the genes in C5 ($n = 1070$) and C6 ($n = 1069$) had specific expression patterns with different intermediate reprogramming stages. Then, the corresponding differential R-loops of these six clusters were identified; parts of the genes in each cluster had differential R-loop signals, so these clusters were labeled as R1—R6 (Figure 4B,C). This result further implies that the sequential activation of these genes is also regulated by other epigenetic modifications besides R-loops.

As mentioned above, we further characterized the detailed dynamics of different HMs patterns (H3K4me3, H3K27me3, Bivalent and other) on these gene clusters associated with R-loops (Figures 4D and S3A). Different HMs were classified into "H3K4me3-only" (H3K4me3), "H3K27me3-only" (H3K27me3), "both H3K4me3/H3K27me3" (Bivalent) and "no H3K4me3 or H3K4me3" (Other) categories based on the H3K4me3 and H3K27me3

modification signals detected on the promoter region of each gene. Interestingly, these genes are more and more modified by H3K4me3 and bivalent during reprogramming (Figure 4D). In particular, the genes in R1 and R4 were preferentially modified by H3K4me3 rather than H3K27me3, while bivalent histone modifications were more likely to be co-occupied in R2 and R3 (Supplementary Figure S3A). The results display that there may be a varying degree of interdependency between HMs and R-loops in regulating genes transcription.
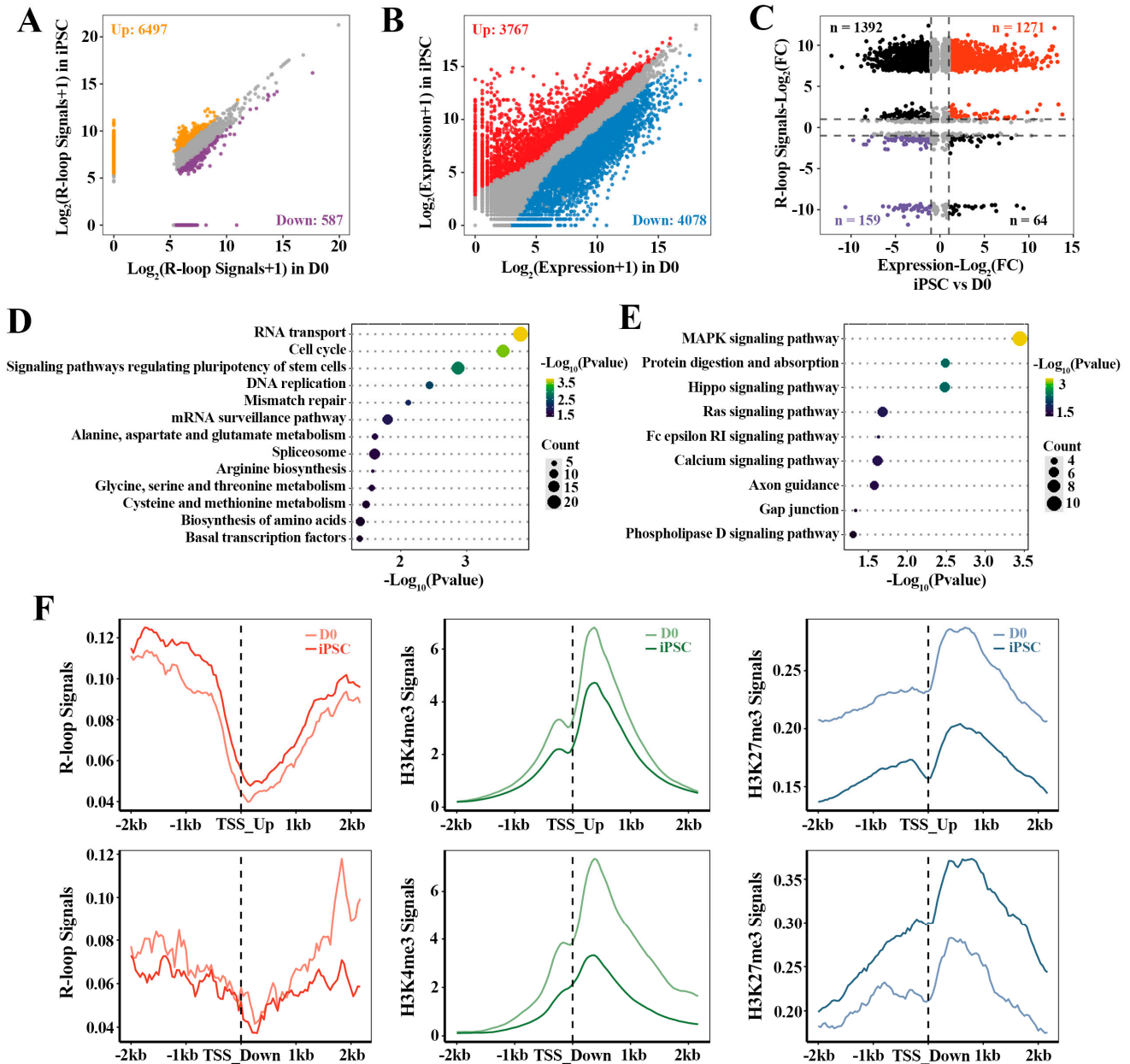
Figure 3. The dynamic effect of R-loops on gene expression in reprogramming: (**A**) analysis of differential R-loop signals in iPSC vs. D0; (**B**) analysis of differential gene expression in iPSC vs. D0; (**C**) scatterplots showing the relationship between R-loops and gene expression during reprogramming process. Red dots are the genes whose expression levels and R-loop signals are up-regulated in iPSC compared to D0. Purple dots are the genes whose expression levels and R-loop signals are down-regulated in iPSC; (**D,E**) KEGG pathway analysis of up-regulated genes with higher R-loop signals (*n* = 1271) and down-regulated genes with lower R-loop signals (*n* = 159); (**F**) the read density profiles of R-loops, H3K4me3 and H3K27me3 at the 2 kb flanking TSSs of the up-regulated genes (*n* = 1271, upper three panels) and down-regulated genes (*n* = 159, lower three panels).
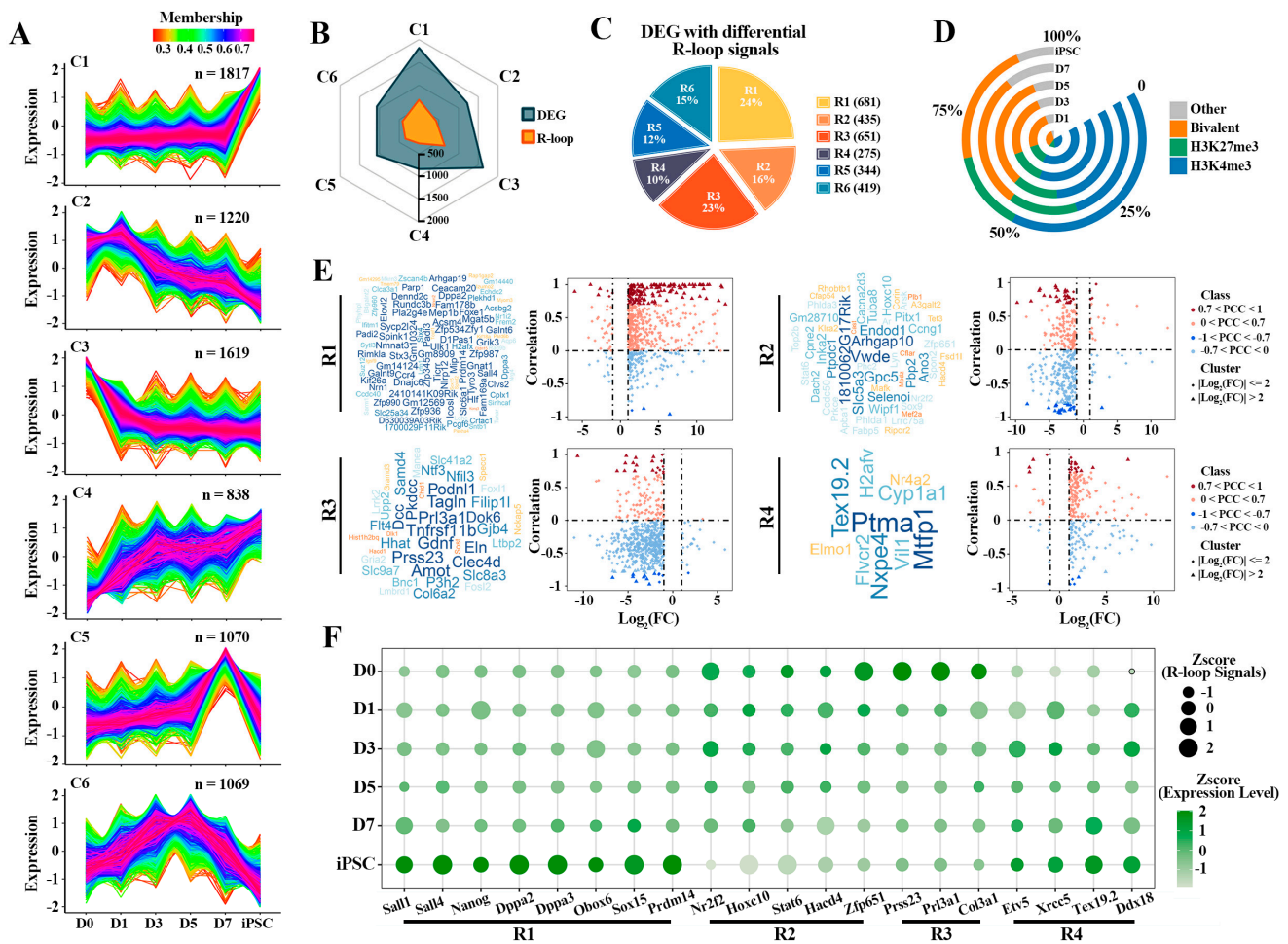
Figure 4. Correlation of R-loop levels with gene expression of different clusters: (**A**) fuzzy C-means (FCM) clustering analysis according to the expression profiles of DEGs. Average scaled FPKM values are represented by each line. Colors indicate the membership value of each gene in the current cluster; (**B**) the radar chart displaying the overlap of genes with differential R-loop signals and six different clusters; (**C**) the percentage of DEGs with differential R-loop signals in each cluster is indicated. These clusters have differential R-loop signals, labeled as R1-R6; (**D**) proportion of co-differential genes (the genes involved in R1-R6) bound by "H3K4me3-only" (H3K4me3—blue), "H3K27me3-only" (H3K27me3—green), "both H3K4me3/H3K27me3" (Bivalent—orange) and "no H3K4me3 or H3K4me3" (Other—grey) modification; (**E**) integrated analysis of DEGs and differential R-loop signals in R1—R4. Word clouds represent the key genes (0.7 < |PCC| < 1; |Log2(FC)| > 2) in each group, both size and color of the word are proportional to the membership value of each gene. Y-axis of the scatter plot indicates the correlation between expression levels and R-loop signals, X-axis indicates logarithmic transformation of the fold change (FC) of the expression level (iPSC vs. D0); (**F**) dynamic changes of gene expression levels and R-loop signals of representative genes in each cluster. The dot sizes indicate expression levels and colors display R-loop signals.

By investigating the influence of R-loops on the dynamic expression patterns of these genes (Figures 4E and S3B and Supplementary Table S4), we found that the expression of most of these genes was positively correlated with R-loop signals, whether in transiently up-regulated gene cluster (R1) at the iPSC stage or in the gene cluster which have a continuous up-regulated trend during reprogramming (R4). For instance, the expression of most pluripotent genes (such as *Dppa2*, *Dppa3*, *Sall4*, etc.) in R1 is directly related to the regulation of R-loops, and the PCC can reach 0.7 (Figures 4E and S3B). Nevertheless, in the other clusters, the effect of R-loops on the expression of these genes did not show a positive or negative regulatory preference, further suggesting that the R-loops had a selective

preference for gene expression regulation (Supplementary Figure S3B and Supplementary Table S4). Then we selected some representative genes from R1-R4 and found that the presence of R-loops was indeed related to the expression of genes (Figure 4F). In R1, with the instantaneous up-regulation of pluripotency genes in the iPSC stage, the R-loop signals also present a transient up-regulation pattern. All in all, these results demonstrate that there is a subset of R-loops which plays an important role in the sequential activation of pluripotency genes in reprogramming.

2.5. The Collaborative Regulation of R-loops and HMs Facilitates the Pluripotency Network Activation during Reprogramming

To investigate how the expression dynamics of these gene clusters are reflected in HM changes, we tracked the binding abilities of H3K4me3 (activating) and H3K27me3 (repressing) on these clusters at D0 and iPSC, respectively (Figure 5A). Interestingly, the genes in R1 and R4 enriched higher H3K4me3 and lower H3K27me3 signals in iPSC than in D0, which is in accordance with the higher expression level of these genes in iPSC, suggesting that H3K4me3 and H3K27me3 keep their antagonistic relationship in regulating gene expression during the reprogramming process. At the same time, compared with D0, we observed lower H3K4me3 levels and higher levels H3K27me3 in R2 and R3 at the iPSC stage, which coincides with the fact that these genes were inhibited in iPSC. Additionally, in the remaining two gene clusters (R5 and R6), the H3K27me3 level at D0 was higher than that in iPSC, while H3K4me3 levels showed no significant difference in D0 and iPSC. These results indicate that these gene clusters exhibit different dynamic expression patterns under the collaborative regulation of H3K4me3 (activating) and H3K27me3 (repressing).

Next, we attempted to explore the relationship between different HM patterns and R-loops on gene expression in R1 and R3 (Figure 5B). The result showed that these genes, highly expressed in iPSC (R1) or in D0 (R3), were attributed to the positive regulation of H3K4me3 and bivalent histone modifications. On the contrary, the decreased expression of these genes in D0 or iPSC was associated with the H3K27me3 epigenetic mark (Figure 5B, right). Interestingly, different HM patterns do not directly affect the R-loop modification signals on gene clusters (Figure 5B, left). To verify this hypothesis that the deposition of these histone markers may be related to the formation of R-loops, we further explored the modification differences between H3K4me3 and H3K27me3 in the formation region of R-loops in different gene clusters (R1 and R3) (Figure 5C). It was found that the gene activation in iPSC (R1) was consistent with the deposition of H3K4me3 in the R-loop formation regions, accompanied by the decrease in H3K27me3 signals. In agreement, the inhibition of genes (R3) at the iPSC stage is related to the decrease in H3K4me3 and the high H3K27me3 levels in the R-loop formation region. Finally, we mapped the dynamic transformation of the regulatory networks of pluripotent genes associated with R-loops during reprogramming, to determine whether the R-loop transition and network activation simply followed the same dynamic pattern (Figure 5D,E). The results showed that the activation of pluripotent transcriptional regulatory networks (TRNs), including *Dppa2*, *Dppa3* and *Zscan4b*, was accompanied by the dynamic transformation of R-loops. Together, the above results indicate that collaborative interactions between R-loops and HMs can facilitate the activation of the pluripotency regulation network, thereby driving cell fate transformation during reprogramming.
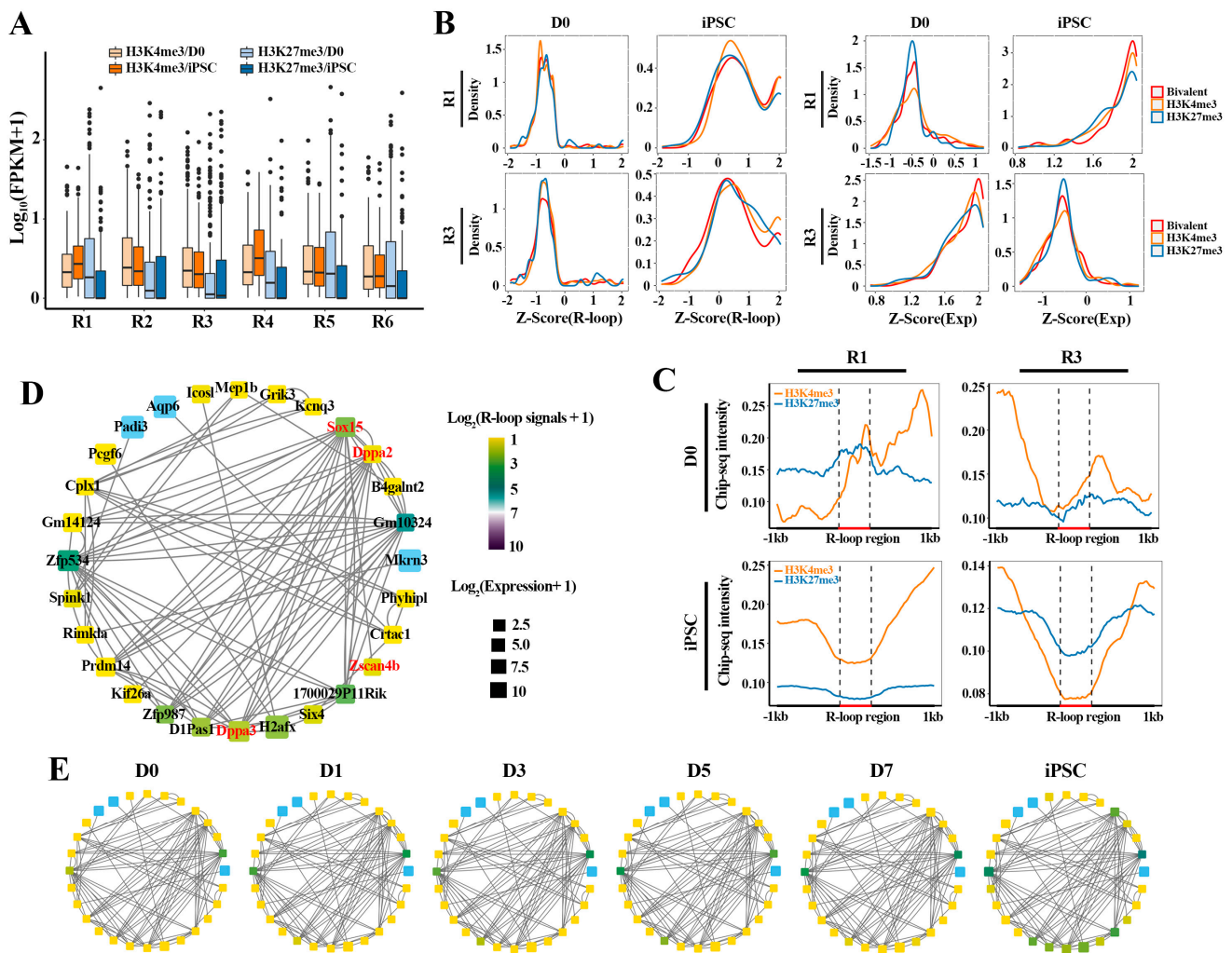
Figure 5. Dynamic transition of TRNs related to R-loops and HMs in reprogramming: (**A**) boxplot showing H3K4me3 and H3K27me3 binding signals (FPKM) on genes associated with bivalent histone modifications in D0 and iPSC; (**B**) the density distribution of R-loop signals (FPKM) and expression levels for genes (involved in R1 and R3, respectively) bound by H3K4me3, H3K27me3 and bivalent histone modifications; (**C**) average H3K4me3, H3K27me3 signals in 1 kb flanking R-loop regions for R1 ($n = 681$) and R3 genes ($n = 651$) at D0 and iPSC stages. These R-loops were located in promoter regions of genes; (**D**,**E**) dynamic transition of transcriptional networks from MEF (D0) to iPSC in R1. Color indicates R-loop signals and box size indicates gene expression levels. Genes symbols as shown in (**D**).

3. Discussion

The regulatory roles of R-loops in yeast, plants and animals have been extensively investigated. However, the potential effect of R-loop dynamics on cell fate transition in somatic cell reprogramming remains unclear. In this study, we profiled the genome-wide formation of R-loops during the reprograming process and found that the formation of R-loops mainly occurs in the promoter, gene body and intergenic regions. Importantly, the formation events of R-loops mainly occur in repetitive elements and the R-loop formation on these repetitive elements has stage-specificity in cell reprogramming. Additionally, R-loop occupation also has TF family specificity, in which they prefer to appear in the NR family than C_2H_2 family during the whole reprogramming process. However, the regulatory mechanism of R-loops on this TF family needs to be further explored.

Previous studies have indicated that R-loops as new candidate regulators play important roles in gene expression [15,27,28]. Consistently, we found that R-loop dynamics were

also closely coupled with gene expression in reprogramming, and that the cumulative effect of R-loops on gene expression is extremely significant in this process. This is most easily interpreted to explain why the kinetic pattern of R-loops was intense for the entire duration of reprogramming, whereas the genes were activated on a large scale in the last stage of reprogramming. However, the gene transcription in the reprogramming process is the result of multilayered collaborative regulation [32], in which the dynamics of R-loops were accompanied by the enhanced recruitment of H3K4me3 and the depletion of H3K27me3 for regulating gene expression in this process. These results support the fact that R-loops and HMs collaboratively shape cell fate transition in reprogramming. Furthermore, the activation of pluripotent TRNs followed the same dynamic pattern with R-loop formation, further displaying that R-loops indeed play an important regulatory role for accelerating cell fate transition in reprogramming.

4. Materials and Methods

4.1. Dataset Collection

Both the single-stranded DNA:RNA immunoprecipitation sequencing (ssDRIP-seq) dataset and RNA sequencing (RNA-seq) data of mouse embryonic fibroblast (MEF) reprogramming to the iPSC state were downloaded from the Gene Expression Omnibus (GEO) database under accession number GSE125644 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE125644, accessed on 12 November 2020) [27], including six different reprogramming stages: MEF/day 0, day 1, day 3, day 5, day 7, induced pluripotent stem cells (D0, D1, D3, D5, D7, iPSC). For chromatin immunoprecipitation sequencing (ChIP-seq) a dataset of six different reprogramming stages was also downloaded from GEO database and GEO accession no. GSE67520 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE67520, accessed on 13 November 2020) [20], which included two histone modifications of H3K4me3 and H3K27me3.

4.2. ssDRIP-seq Data Processing

The ssDRIP-seq raw reads were filtered by Cutadapt software (version 2.10, EMBnet.journal, Uppsala, Sweden) [27,33] and reads less than 50 bases in length were discarded. Trimmed reads were mapped to the mouse genome assembly mm10 by using Bowtie2 (version 2.2.9, Nature Methods, New York, NY, USA) short read alignment software with default parameters. Then, R-loop peak calling was performed using MACS2 (version 2.1.0) [34–36] using the following parameters: macs2 callpeak $-$t \$treatmentsam $-$c \$controlsam $-$f SAM $-$keep$-$dup 1 $-n$ \$name $-$g 1.87e9 $-$B $-$q 0.01. Finally, R package ChIPseeker [37] was used to annotate with the position of the peaks in the genome, in which $-$2 kb to 1 kb of gene transcription start site (TSS) were defined as gene promoter. We also calculated the occupancy of each R-loop as FPKM (fragments per kb per million uniquely mapped reads) for their signals. Genome-wide R-loop signals were calculated from absolute mapped read density over 300 bp sliding windows using 'ngs.plot.r' from ngsplot (version 2.63, BMC Genomics, Basingstoke, UK) [38].

4.3. ChIP-seq Data Processing

For ChIP-seq data processing, the original data were controlled by Cutadapt software to remove low-quality reads. Next, the clean reads were aligned to mm10 by using Bowtie2 (version 2.2.9) with default parameters [39]. Then we used MACS2 (version 2.1.0, Current Protocols in Bioinformatics, Hoboken, NJ, USA) (with the parameters setting: macs2 callpeak $-$t \$treatmentsam $-$c \$controlsam $-$f SAM $-$keep-dup 1 $-$n \$name $-$g 1.87e9 $-$B $-$q 0.01) to call binding peaks [34,35]. The ChIP-seq coverage densities of H3K4me3 and H3K27me3 were plotted by mapping reads to regions that were 2-kb up/downstream of the TSS for each gene, respectively. Each gene region was split into 300 equally sized bins. The average ChIP-Seq density of H3K4me3 and H3K27me3 in the R-loop regions were calculated using ngsplot (version 2.63) [38].

4.4. RNA-seq Analysis

The pair-end raw RNA-seq reads were trimmed by the Trimmomatic (version 0.38, Bioinformatics, Oxford, UK) [40] and mapped to mm10 reference genome using Hisat2 (version 2.1.0, Nature Protocols, New York, NY, USA) aligner with default parameters [41]. The retained reads were subsequently assembled by using Stringtie (version 1.3.5, Nature Biotechnology, New York, NY, USA) [41,42]. To eliminate the effects of sequencing depth and transcript length, normalized FPKM (fragments per kilobase of exon model per million mapped reads) for each gene were conducted. Read counts for each gene were calculated using HTseq (version 0.11.0, Bioinformatics, Oxford, UK) [43].

4.5. Identification of R-loop Peaks on Repetitive Elements and Transcription Factor (TF) Family

For alignments to repetitive regions in the genome, repeat annotations were downloaded from the University of California at Santa Cruz (UCSC) browser (RepeatMasker, mm10) [44]. R-loop peaks were assigned to TF families according to the JASPAR database [45]. TF families with fewer than two members were grouped under "Other."

4.6. Differential Gene Expression Analysis

Differential expression analysis was performed by R package DEseq2 [46,47]. For each comparison, genes with a Benjamini and Hochberg-adjusted p value (false discovery rate, FDR) < 0.05 and the absolute of Log2(fold change, FC) > 1 were regarded as differential expression genes (DEGs) [48–50].

4.7. Fuzzy C-Means (FCM) Clustering Analysis

Fuzzy cluster analysis was performed by using mFuzz [51] for cluster analysis of the gene expression pattern, the normalized expression levels of all DEGs between iPSC and D0 were analyzed.

4.8. Transcriptional Regulatory Networks Analysis

Transcriptional regulatory network (TRNs) analysis of representative genes associated with R-loops (as shown in Figure 5D) were implemented by STRING database (version 11.0) [52]. Then, the TRNs were visualized by Cytoscape (version 3.7.0, Genome Research, New York, NY, USA) [53,54]. In TRNs, the color indicates R-loop signals and box size indicates gene expression levels.

4.9. Functional Enrichment and Statistical Analysis

Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis was performed based on the R package clusterProfiler (version 3.14.3, OMICS: A Journal of Integrative Biology, New York, NY, USA) [55]. Statistical analyses were implemented with R (version 3.6.0, http://www.r-project.org, accessed on 1 June 2020). Representative KEGG pathways with p values < 0.05 were summarized in each gene cluster. The Pearson correlation coefficients (PCC) between biological replicates were calculated using the 'cor' function with default parameters [56,57]. A Student's t test was performed using the 't.test' function with default parameters, and p values < 0.05 were considered statistically significant [58,59].

5. Conclusions

Taken together, our results not only profiled the unique dynamic patterns of R-loops in reprogramming, but also deciphered that a crosstalk exists between R-loops and HMs to regulate cell fate transition during the reprogramming process.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/ijms23031567/s1.

References

1. Skourti-Stathaki, K.; Proudfoot, N.J. A double-edged sword: R loops as threats to genome integrity and powerful regulators of gene expression. *Genes Dev.* **2014**, *28*, 1384–1396. [CrossRef] [PubMed]
2. Milman, G.; Langridge, R.; Chamberlin, M.J. The structure of a DNA-rna hybrid. *Proc. Natl. Acad. Sci. USA* **1967**, *57*, 1804–1810. [CrossRef] [PubMed]
3. Thomas, M.; Davis, W. Hybridization of rna to double-stranded DNA: Formation of r-loops. *Proc. Natl. Acad. Sci. USA* **1976**, *73*, 2294–2298. [CrossRef] [PubMed]
4. Ginno, P.; Lott, P.; Christensen, H.; Korf, I.; Chédin, F. R-loop formation is a distinctive characteristic of unmethylated human cpg island promoters. *Mol. Cell* **2012**, *45*, 814–825. [CrossRef] [PubMed]
5. Ginno, P.A.; Lim, Y.W.; Lott, P.L.; Korf, I.; Chedin, F. Gc skew at the 5′ and 3′ ends of human genes links r-loop formation to epigenetic regulation and transcription termination. *Genome Res.* **2013**, *23*, 1590–1600. [CrossRef]
6. Sanz, L.A.; Hartono, S.R.; Lim, Y.W.; Steyaert, S.; Rajpurkar, A.; Ginno, P.A.; Xu, X.; Chédin, F. Prevalent, dynamic, and conserved r-loop structures associate with specific epigenomic signatures in mammals. *Mol. Cell* **2016**, *63*, 167–178. [CrossRef]
7. Chen, L.; Chen, J.Y.; Zhang, X.; Gu, Y.; Xiao, R.; Shao, C.; Tang, P.; Qian, H.; Luo, D.; Li, H. R-chip using inactive rnase h reveals dynamic coupling of r-loops with transcriptional pausing at gene promoters. *Mol. Cell* **2017**, *68*, 745–757.e5. [CrossRef]
8. García-Muse, T.; Aguilera, A. R loops: From physiological to pathological roles. *Cell* **2019**, *179*, 604–618. [CrossRef]
9. Skourti-Stathaki, K.; Kamieniarz-Gdula, K.; Proudfoot, N.J. R-loops induce repressive chromatin marks over mammalian gene terminators. *Nature* **2014**, *516*, 436–439. [CrossRef]
10. Hamperl, S.; Bocek, M.J.; Saldivar, J.C.; Swigut, T.; Cimprich, K.A. Transcription-replication conflict orientation modulates r-loop levels and activates distinct DNA damage responses-sciencedirect. *Cell* **2017**, *170*, 774–786. [CrossRef]
11. Wahba, L.; Gore, S.K.; Koshland, D. The homologous recombination machinery modulates the formation of rna-DNA hybrids and associated chromosome instability. *eLife* **2013**, *2*, e00505. [CrossRef] [PubMed]
12. García-Pichardo, D.; Cañas, J.C.; García-Rubio, M.L.; Gómez-González, B.; Rondón, A.G.; Aguilera, A. Histone mutants separate r loop formation from genome instability induction. *Mol. Cell* **2017**, *66*, 597–609.e595. [CrossRef] [PubMed]
13. Santos-Pereira, J.M.; Aguilera, A. R loops: New modulators of genome dynamics and function. *Nat. Rev. Genet.* **2015**, *16*, 583–597. [CrossRef] [PubMed]
14. Lv, H.; Dao, F.Y.; Zulfiqar, H.; Su, W.; Ding, H.; Liu, L.; Lin, H. A sequence-based deep learning approach to predict ctcf-mediated chromatin loop. *Brief. Bioinform.* **2021**, *22*, 1–13.
15. Niehrs, C.; Luke, B. Regulatory r-loops as facilitators of gene expression and genome stability. *Nat. Rev. Mol. Cell Biol.* **2020**, *21*, 167–178. [CrossRef] [PubMed]
16. Nadel, J.; Athanasiadou, R.; Lemetre, C.; Wijetunga, N.A.; Broin, P.; Sato, H.; Zhang, Z.; Jeddeloh, J.; Montagna, C.; Golden, A. Rna:DNA hybrids in the human genome have distinctive nucleotide characteristics, chromatin composition, and transcriptional relationships. *Epigenet. Chromatin* **2015**, *8*, 1–19. [CrossRef] [PubMed]
17. Takahashi, K.; Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **2006**, *126*, 663–676. [CrossRef]
18. Takahashi, K.; Tanabe, K.; Ohnuki, M.; Narita, M.; Ichisaka, T.; Tomoda, K.; Yamanaka, S. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **2007**, *131*, 861–872. [CrossRef]

19. Chin, M.H.; Pellegrini, M.; Plath, K.; Lowry, W.E. Molecular analyses of human induced pluripotent stem cells and embryonic stem cells. *Cell Stem Cell* **2010**, *7*, 263–269. [CrossRef]

20. Chen, J.; Chen, X.; Li, M.; Liu, X.; Gao, Y.; Kou, X.; Zhao, Y.; Zheng, W.; Zhang, X.; Huo, Y.; et al. Hierarchical oct4 binding in concert with primed epigenetic rearrangements during somatic cell reprogramming. *Cell Rep.* **2016**, *14*, 1540–1554. [CrossRef]

21. Gomes, K.M.; Costa, I.C.; Santos, J.F.; Dourado, P.M.; Forni, M.F.; Ferreira, J.C. Induced pluripotent stem cells reprogramming: Epigenetics and applications in the regenerative medicine. *Rev. Assoc. Med. Bras.* **2017**, *63*, 180–189. [CrossRef] [PubMed]

22. Godini, R.; Lafta, H.Y.; Fallahi, H. Epigenetic modifications in the embryonic and induced pluripotent stem cells. *Gene Expr. Patterns GEP* **2018**, *29*, 1–9. [CrossRef] [PubMed]

23. Li, H.; Ta, N.; Long, C.; Zhang, Q.; Li, S.; Liu, S.; Yang, L.; Zuo, Y. The spatial binding model of the pioneer factor oct4 with its target genes during cell reprogramming. *Comput. Struct. Biotechnol. J.* **2019**, *17*, 1226–1233. [CrossRef]

24. Ying, Z.; Xiang, G.; Zheng, L.; Tang, H.; Duan, L.; Lin, X.; Zhao, Q.; Chen, K.; Wu, Y.; Xing, G.; et al. Short-term mitochondrial permeability transition pore opening modulates histone lysine methylation at the early phase of somatic cell reprogramming. *Cell Metab.* **2018**, *28*, 935–945.e935. [CrossRef] [PubMed]

25. Hussein, S.M.; Puri, M.C.; Tonge, P.D.; Benevento, M.; Corso, A.J.; Clancy, J.L.; Mosbergen, R.; Li, M.; Lee, D.S.; Cloonan, N.; et al. Genome-wide characterization of the routes to pluripotency. *Nature* **2014**, *516*, 198–206. [CrossRef] [PubMed]

26. Wang, Z.R.; Liu, D.Y.; Xu, B.F.; Tian, R.X.; Zuo, Y.C. Modular arrangements of sequence motifs determine the functional diversity of kdm proteins. *Brief. Bioinform.* **2021**, *22*, bbaa215. [CrossRef]

27. Li, Y.; Song, Y. R-loops coordinate with sox2 in regulating reprogramming to pluripotency. *Sci. Adv.* **2020**, *6*, eaba0777. [CrossRef]

28. Yan, P.; Liu, Z.; Song, M.; Wu, Z.; Liu, G.H. Genome-wide r-loop landscapes during cell differentiation and reprogramming. *Cell Rep.* **2020**, *32*, 107870. [CrossRef]

29. Gaillard, H.; Aguilera, A. Transcription as a threat to genome integrity. *Annu. Rev. Biochem.* **2016**, *85*, 291–317. [CrossRef] [PubMed]

30. Liu, X.; Wang, C.; Liu, W.; Li, J.; Li, C.; Kou, X.; Chen, J.; Zhao, Y.; Gao, H.; Wang, H.; et al. Distinct features of h3k4me3 and h3k27me3 chromatin domains in pre-implantation embryos. *Nature* **2016**, *537*, 558–562. [CrossRef]

31. Wang, Y.; Zhao, C.; Hou, Z.; Yang, Y.; Bi, Y.; Wang, H.; Zhang, Y.; Gao, S. Unique molecular events during reprogramming of human somatic cells to induced pluripotent stem cells (ipscs) at naïve state. *eLife* **2018**, *7*, e29518. [CrossRef] [PubMed]

32. Li, H.S.; Song, M.M.; Yang, W.R.T.; Cao, P.B.; Zheng, L.; Zuo, Y.C. A comparative analysis of single-cell transcriptome identifies reprogramming driver factors for efficiency improvement. *Mol. Ther.-Nucl. Acids* **2020**, *19*, 1053–1064. [CrossRef]

33. Chen, C.; Khaleel, S.S.; Huang, H.; Wu, C.H. Software for pre-processing illumina next-generation sequencing short read sequences. *Source Code Biol. Med.* **2014**, *9*, 8. [CrossRef] [PubMed]

34. Feng, J.; Liu, T.; Zhang, Y. Using macs to identify peaks from chip-seq data. *Curr. Protoc. Bioinform.* **2011**, *2*, 2.14.1–2.14.14. [CrossRef] [PubMed]

35. De Iaco, A.; Coudray, A.; Duc, J.; Trono, D. Dppa2 and dppa4 are necessary to establish a 2c-like state in mouse embryonic stem cells. *EMBO Rep.* **2019**, *20*, e47382. [CrossRef]

36. Li, H.; Long, C.; Xiang, J.; Liang, P.; Li, X.; Zuo, Y. Dppa2/4 as a trigger of signaling pathways to promote zygote genome activation by binding to cg-rich region. *Brief. Bioinform.* **2020**, *22*, bbaa342. [CrossRef]

37. Yu, G.; Wang, L.G.; He, Q.Y. Chipseeker: An r/bioconductor package for chip peak annotation, comparison and visualization. *Bioinformatics* **2015**, *31*, 2382–2383. [CrossRef]

38. Shen, L.; Shao, N.; Liu, X.; Nestler, E. Ngs.Plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. *BMC Genom.* **2014**, *15*, 284. [CrossRef]

39. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with bowtie 2. *Nat. Methods* **2012**, *9*, 357–359. [CrossRef]

40. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef]

41. Pertea, M.; Kim, D.; Pertea, G.; Leek, J.; Salzberg, S. Transcript-level expression analysis of rna-seq experiments with hisat, stringtie and ballgown. *Nat. Protoc.* **2016**, *11*, 1650–1667. [CrossRef] [PubMed]

42. Pertea, M.; Pertea, G.M.; Antonescu, C.M.; Chang, T.-C.; Mendell, J.T.; Salzberg, S.L. Stringtie enables improved reconstruction of a transcriptome from rna-seq reads. *Nat. Biotechnol.* **2015**, *33*, 290. [CrossRef] [PubMed]

43. Anders, S.; Pyl, P.T.; Huber, W. Htseq—a python framework to work with high-throughput sequencing data. *Bioinformatics* **2015**, *31*, 166–169. [CrossRef] [PubMed]

44. Tarailo-Graovac, M.; Chen, N. Using repeatmasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinform.* **2009**, *25*, 4.10.1–4.10.14. [CrossRef] [PubMed]

45. Mathelier, A.; Fornes, O.; Arenillas, D.J.; Chen, C.Y.; Denay, G.; Lee, J.; Shi, W.; Shyr, C.; Tan, G.; Worsley-Hunt, R.; et al. Jaspar 2016: A major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **2016**, *44*, D110–D115. [CrossRef] [PubMed]

46. 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]

47. Long, C.; Li, W.; Liang, P.; Liu, S.; Zuo, Y. Transcriptome comparisons of multi-species identify differential genome activation of mammals embryogenesis. *IEEE Access* **2019**, *7*, 7794–7802. [CrossRef]

48. Liu, W.; Liu, X.; Wang, C.; Gao, Y.; Gao, R.; Kou, X.; Zhao, Y.; Li, J.; Wu, Y.; Xiu, W.; et al. Identification of key factors conquering developmental arrest of somatic cell cloned embryos by combining embryo biopsy and single-cell sequencing. *Cell Discov.* **2016**, *2*, 16010. [CrossRef] [PubMed]

49. Costa-Silva, J.; Domingues, D.; Lopes, F.M. Rna-seq differential expression analysis: An extended review and a software tool. *PLoS ONE* **2017**, *12*, e0190152. [CrossRef]

50. Liang, P.; Zheng, L.; Long, C.; Yang, W.; Yang, L.; Zuo, Y. Helpredictor models single-cell transcriptome to predict human embryo lineage allocation. *Brief. Bioinform.* **2021**, *22*, bbab196. [CrossRef]

51. Kumar, L.; Futschik, M.E. Mfuzz: A software package for soft clustering of microarray data. *Bioinformation* **2007**, *2*, 5–7. [CrossRef] [PubMed]

52. Damian, S.; Morris, J.H.; Helen, C.; Michael, K.; Stefan, W.; Milan, S.; Alberto, S.; Doncheva, N.T.; Alexander, R.; Peer, B. The string database in 2017: Quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.* **2017**, *45*, D362–D368.

53. Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N.S.; Wang, J.T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **2003**, *13*, 2498–2504. [CrossRef] [PubMed]

54. Long, C.; Li, H.; Li, X.; Yang, W.; Zuo, Y. Nuclear transfer arrest embryos show massive dysregulation of genes involved in transcription pathways. *Int. J. Mol. Sci.* **2021**, *22*, 8187. [CrossRef] [PubMed]

55. Yu, G.; Wang, L.G.; Han, Y.; He, Q.Y. Clusterprofiler: An r package for comparing biological themes among gene clusters. *Omics-A J. Integr. Biol.* **2012**, *16*, 284–287. [CrossRef]

56. Ahlgren, P.; Jarneving, B.; Rousseau, R. Requirements for a cocitation similarity measure, with special reference to pearson's correlation coefficient. *J. Am. Soc. Inf. Sci. Technol.* **2003**, *54*, 550–560. [CrossRef]

57. Tang, Q.; Nie, F.; Kang, J.; Chen, W. Mrnalocater: Enhance the prediction accuracy of eukaryotic mrna subcellular localization by using model fusion strategy. *Mol. Ther. J. Am. Soc. Gene Ther.* **2021**, *29*, 2617–2623. [CrossRef]

58. Bandyopadhyay, S.; Mallik, S.; Mukhopadhyay, A. A survey and comparative study of statistical tests for identifying differential expression from microarray data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2014**, *11*, 95–115. [CrossRef]

59. Greenland, S.; Senn, S.J.; Rothman, K.J.; Carlin, J.B.; Poole, C.; Goodman, S.N.; Altman, D.G. Statistical tests, p values, confidence intervals, and power: A guide to misinterpretations. *Eur. J. Epidemiol.* **2016**, *31*, 337–350. [CrossRef]