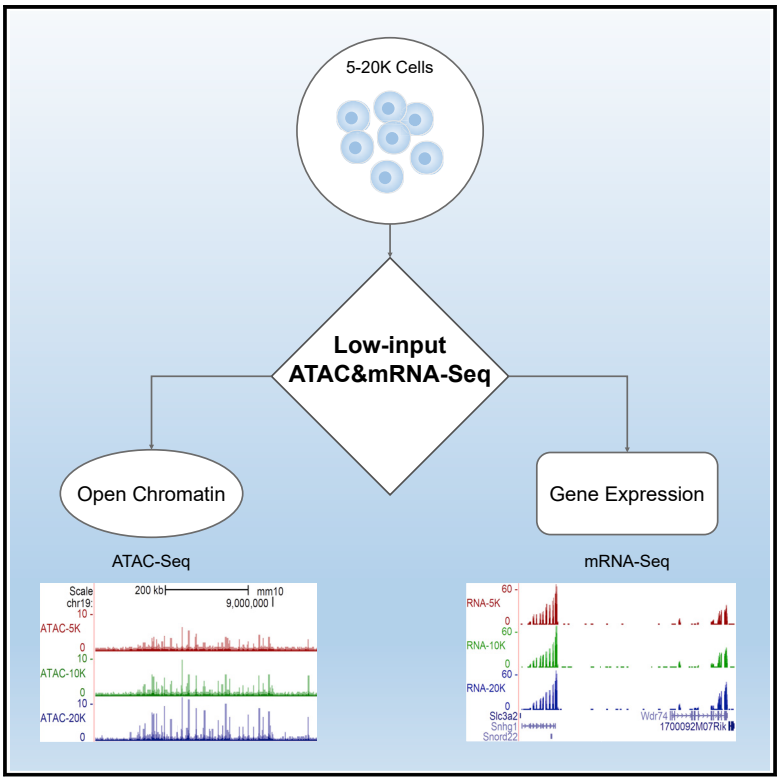
Cell Reports
Methods

Article

A simple and robust method for simultaneous dual-omics profiling with limited numbers of cells

Graphical abstract



Highlights

- Low-input ATAC&mRNA-seq is a simple, fast, robust, and reproducible dual-omics method
- It simultaneously profiles the epigenome and transcriptome of limited-input material
- The resulting data show high consistency with Omni-ATAC and conventional RNA-seq

Authors

Ruifang Li, Sara A. Grimm, Paul A. Wade

Correspondence

lir1@mskcc.org

In brief

Li et al. develop a simple and fast dual-omics method (low-input ATAC&mRNA-seq) for concurrent profiling of chromatin accessibility and gene expression in small numbers of cells with high robustness and reproducibility. It provides a unique solution to situations when joint epigenome and transcriptome analyses are required on limited sample material.

Li et al., 2021, Cell Reports Methods *1*, 100041
July 26, 2021
https://doi.org/10.1016/j.crmeth.2021.100041

CellPress

Cell Reports Methods

CellPress
OPEN ACCESS

Article

A simple and robust method for simultaneous dual-omics profiling with limited numbers of cells

Ruifang Li,[1,4,*] Sara A. Grimm,[2] and Paul A. Wade[3]

[1]Epigenetics Innovation Lab, Center for Epigenetics Research, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA
[2]Integrative Bioinformatics Support Group, Epigenetics and Stem Cell Biology Laboratory, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709, USA
[3]Eukaryotic Transcriptional Regulation Group, Epigenetics and Stem Cell Biology Laboratory, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709, USA
[4]Lead contact
*Correspondence: lir1@mskcc.org
https://doi.org/10.1016/j.crmeth.2021.100041

MOTIVATION Many cutting-edge studies require both transcriptome and epigenome analyses of very small amounts of input material. However, multi-omics profiling of low-input bulk samples remains challenging with existing mono-omics techniques. Here, we present a simple and fast dual-omics method (low-input ATAC&mRNA-seq) for simultaneous profiling of chromatin accessibility and gene expression from the same cells in limited cell numbers (5,000–20,000 cells) by using commercial off-the-shelf reagents and basic molecular biology equipment, with the resultant ATAC-seq and mRNA-seq data of acceptable quality comparable to that of the counterpart data generated with standard conventional mono-omics assays.

SUMMARY

Deciphering epigenetic regulation of gene expression requires measuring the epigenome and transcriptome jointly. Single-cell multi-omics technologies have been developed for concurrent profiling of chromatin accessibility and gene expression. However, multi-omics profiling of low-input bulk samples remains challenging. Therefore, we developed low-input ATAC&mRNA-seq, a simple and robust method for studying the role of chromatin structure in gene regulation in a single experiment with thousands of cells, to maximize insights from limited input material by obtaining ATAC-seq and mRNA-seq data simultaneously from the same cells with data quality comparable to that of conventional mono-omics assays. Integrative data analysis revealed similar strong association between promoter accessibility and gene expression when using the data of low-input ATAC&mRNA-seq as when using single-assay data, underscoring the accuracy and reliability of our dual-omics assay to generate both datum types simultaneously with just thousands of cells. We envision our method to be widely applied in many biological disciplines with limited materials.

INTRODUCTION

Joint profiling of the epigenome and transcriptome is needed to unravel epigenetic regulation of gene expression. Conventional high-throughput transcriptome profiling and epigenome mapping technologies, such as micrococcal nuclease digestion with deep sequencing (MNase-seq) (Schones et al., 2008), DNase I hypersensitive sites sequencing (DNase-seq) (Boyle et al., 2008), and formaldehyde-assisted isolation of regulatory elements (FAIRE-seq) (Giresi et al., 2007), typically require large amounts of input material (i.e., millions of cells or more); therefore, they are not suitable for many cutting-edge studies requiring transcriptome and epigenome analyses of very small amounts of input material. Assay for transposase-accessible chromatin using sequencing (ATAC-seq) (Buenrostro et al., 2013) is a state-of-the-art low-input technology widely used for studying chromatin structure. Combining ATAC-seq with RNA sequencing (RNA-seq) offers a powerful approach to understanding the role of chromatin structure in regulating gene expression. Recently, several single-cell multi-omics profiling assays have been developed for concurrent measurement of chromatin accessibility and gene expression in the same cell (Cao et al., 2018; Liu et al., 2019; Ma et al., 2020; Xing et al., 2020). These methods hold considerable promise for low-input material, but they cannot be easily adopted in regular biological laboratories because of high cost and complex workflows. In addition, the sparse and noisy nature of single-cell data remains a limitation.

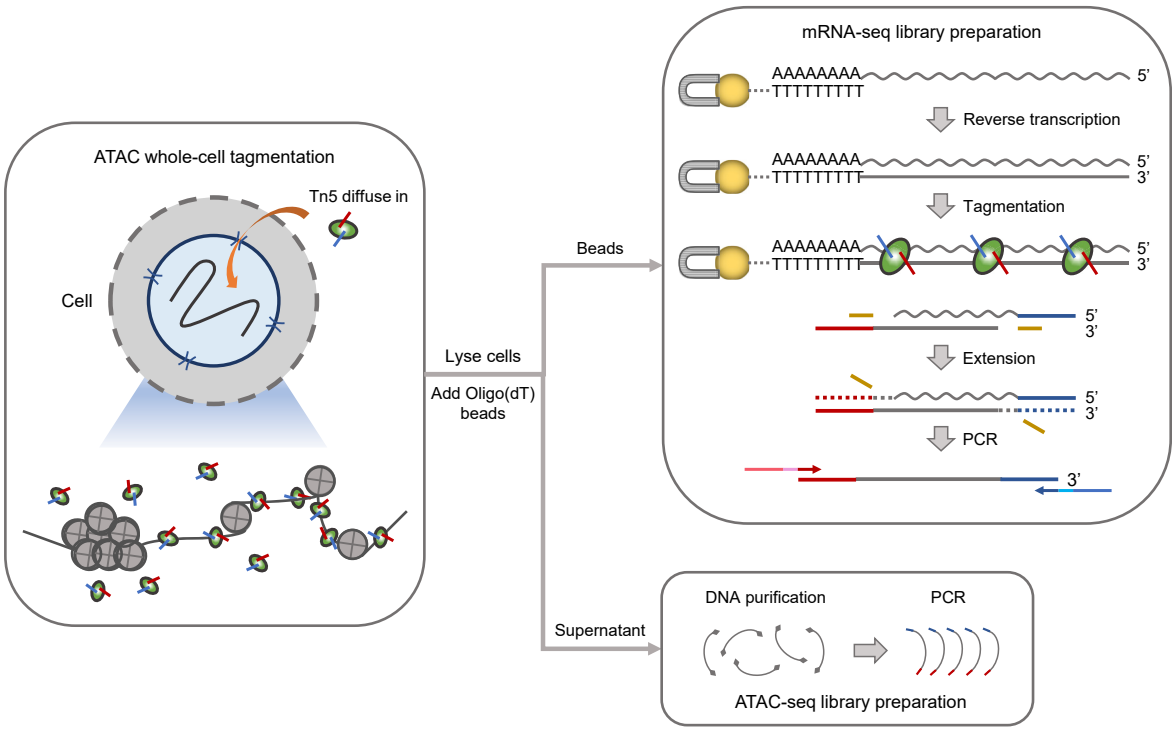


Figure 1. Schematic overview of low-input ATAC&mRNA-seq workflow

Harvested cells were washed and then permeabilized with mild detergent (indicated by holes in cell membrane) to facilitate the entry of Tn5 into the nuclei to tagment open chromatin regions. Tagmented cells were then lysed, and Dynabeads Oligo(dT)$_{25}$ were added into the cell lysate to capture mRNA. After magnetic separation, tagmented genomic DNA in the supernatant was purified and further amplified with indexed PCR to construct the ATAC-seq library, whereas mRNA captured on beads was reverse transcribed using the bead-bound oligo(dT) as primer. The mRNA/cDNA hybrids were then directly tagmented by Tn5, and after initial end extension the tagmented cDNA was amplified with indexed PCR to prepare the mRNA-seq library. Wavy and straight gray lines represent RNA and DNA, respectively. Dotted lines represent the extended fragment.

It remains a challenge to profile the epigenome and transcriptome simultaneously with a limited amount of material in many biological disciplines. Therefore, we developed low-input ATAC&mRNA-seq, a simple and low-cost method for fast simultaneous profiling of chromatin accessibility and gene expression with a small number of cells. As a proof of concept, we applied low-input ATAC&mRNA-seq to mouse embryonic stem cells (mESCs) with 5,000 (5K), 10,000 (10K), and 20,000 (20K) cells as input, and compared the resulting ATAC-seq and mRNA-seq data with conventional single-assay data. We found that our method generated comparable high-quality data even with just 5K cells, thus providing a unique solution to multi-omics profiling with limited material.

RESULTS

Low-input ATAC&mRNA-seq for simultaneous profiling of chromatin accessibility and gene expression

We developed low-input ATAC&mRNA-seq (Figure 1) to simultaneously profile chromatin accessibility and gene expression with a small number of cells. Instead of isolating the nuclei as in Omni-ATAC (Corces et al., 2017), our method uses a one-step membrane permeabilization and transposition of whole cells for ATAC-seq. As for mRNA sequencing, our method uses direct mRNA isolation from the cell lysate with Dynabeads Oligo(dT)$_{25}$ followed by solid-phase cDNA synthesis, then employs Tn5 transposase to directly fragment and tag mRNA/cDNA hybrids to form an amplifiable library for sequencing. Our mRNA-seq strategy enables a seamless on-bead process in one tube, which not only simplifies sample handling but also minimizes sample loss. With the greatly simplified workflow, our method takes only ~4 h from cells to sequencing-ready ATAC-seq and mRNA-seq libraries, dramatically reducing hands-on time. To benchmark our method, we performed low-input ATAC&mRNA-seq on mESCs with low-input cell numbers (5K, 10K, and 20K) and compared the resulting data with single-assay data from Omni-ATAC-seq and conventional bulk mRNA-seq to assess data quality.

The ATAC-seq data generated with low-input ATAC&mRNA-seq are comparable with Omni-ATAC-seq data

Omni-ATAC (Corces et al., 2017) is an improved protocol for ATAC-seq with dramatically reduced mitochondrial reads and enhanced signal-to-noise ratio. Even though our method used whole cells instead of isolated nuclei for transposition reaction, it achieved similarly low percentage of mitochondrial reads contamination (<10%) comparable with Omni-ATAC (Figure 2A). Notably, the duplication rate was low across our ATAC-seq libraries and increased only marginally with decrease in the

CellPress
OPEN ACCESS

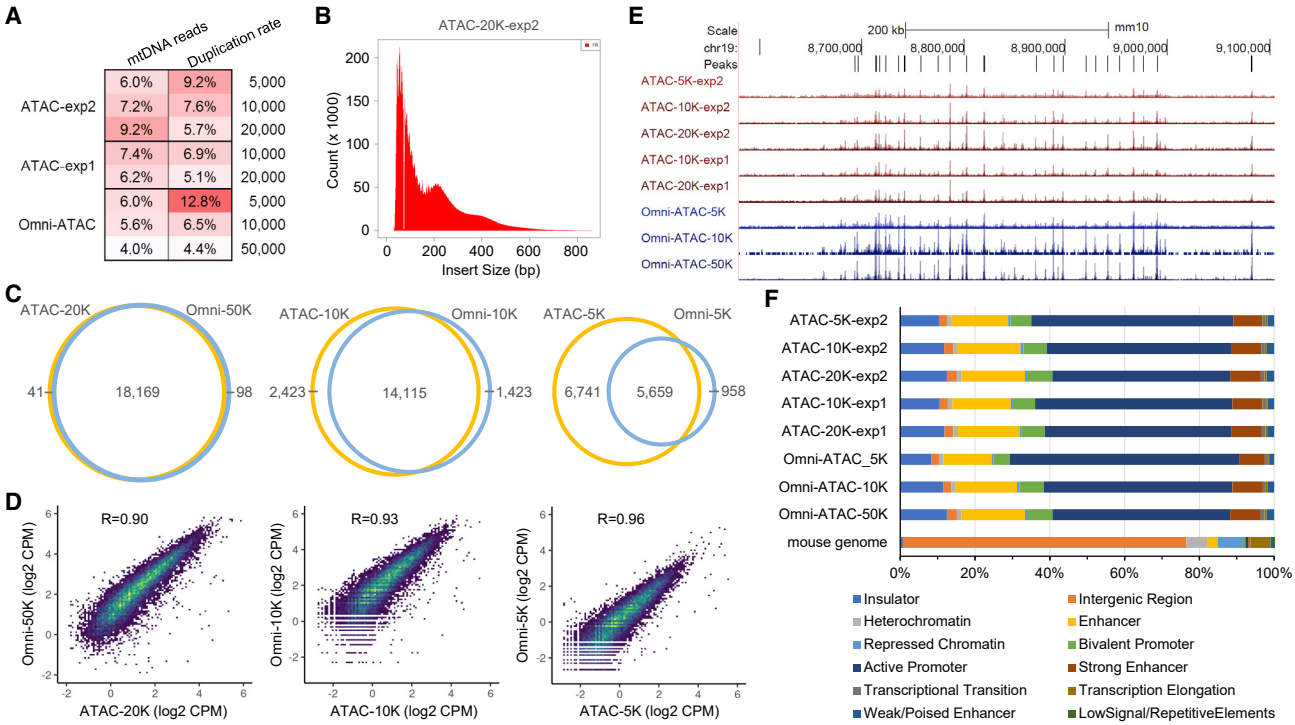


Figure 2. The ATAC-seq data generated with low-input ATAC&mRNA-seq are comparable with Omni-ATAC-seq data

Two independent low-input ATAC&mRNA-seq experiments (exp1 and exp2) as well as Omni-ATAC-seq were performed on E14 mESCs with different input cell numbers.

(A) Heatmap representation of ATAC-seq quality control metrics including duplication rate and the percentage of reads mapped to mitochondrial DNA (mtDNA). Lighter color is used to depict the more desirable value of each metric, along with the numeric values shown for each sample. The numbers of input cells used in each sample are shown on the right of the heatmap.

(B) Fragment size distribution of a representative ATAC-seq library prepared with low-input ATAC&mRNA-seq.

(C) Venn diagrams showing overlap of ATAC-seq peaks identified by low-input ATAC&mRNA-seq and Omni-ATAC-seq.

(D) Density scatterplots displaying correlation of ATAC-seq data generated with low-input ATAC&mRNA-seq and Omni-ATAC-seq. Each dot represents an individual peak in the unified peak set with viridis color scale indicating density. Pearson's *r* value is shown at the top of each plot.

(E) University of California Santa Cruz (UCSC) genome browser view of ATAC-seq coverage tracks at chr19: 8,579,587 to 9,105,586.

(F) Bar graph showing the proportion of ATAC-seq peaks and the mouse genome falling into each chromatin state of mESCs.

number of input cells (Figure 2A), indicating high complexity of the libraries even when starting with just 5K cells. Our ATAC-seq libraries exhibited the characteristic nucleosome periodicity in fragment size distribution (Figures 2B and S1A) along with transcription start site (TSS) enrichment (Figure S1B), which are typical of a successful ATAC-seq experiment. The patterns were consistent across our ATAC-seq libraries with different input cell numbers, albeit with slight differences in the TSS enrichment score and visibility of the nucleosome periodicity (Figures S1A and S1B). In addition, the characteristic patterns of nucleosome-free region and nucleosome positioning at promoter regions were also clearly detected in our ATAC-seq data (Figure S1C). Overall, our ATAC-seq libraries were of quality comparable to that of Omni-ATAC libraries in terms of ATAC-seq quality control metrics.

Next, we compared the identified accessible chromatin regions (ATAC-seq peaks) by the two methods. Compared with standard Omni-ATAC using 50K cells as input, our method obtained a similar number of peaks with only 20K cells (Omni-50K: 18,267; ATAC-20K: 18,210); with 10K and 5K cells as input, our method detected more peak regions than Omni-ATAC (Figure 2C), suggesting that our method requires even fewer cells than Omni-ATAC, which is largely attributable to minimized sample loss with the one-step approach in our simplified ATAC workflow. Importantly, the majority of ATAC-seq peaks identified with our method overlapped with Omni-ATAC peaks (Figure 2C), and ATAC-seq signals at unified peak regions were highly correlated between our method and Omni-ATAC (Pearson's $r \geq 0.90$) over a range of input cell numbers (Figure 2D), indicating remarkably high consistency in enrichment between our ATAC-seq data and Omni-ATAC-seq data as exemplified in the coverage tracks (Figure 2E). Open chromatin regions encompass several key features of the epigenome, including active and poised regulatory regions. To verify that our ATAC-seq data correctly identified those regulatory features, we investigated the epigenomic contexts of ATAC-seq peaks with respect to chromatin states of mESCs as defined by the ChromHMM model (Pintacuda et al., 2017). We observed similar distribution of ATAC-seq peaks identified by our method and Omni-ATAC (Figure 2F); ~51% of peaks were located in "Active Promoter," ~6% in "Bivalent Promoter,"

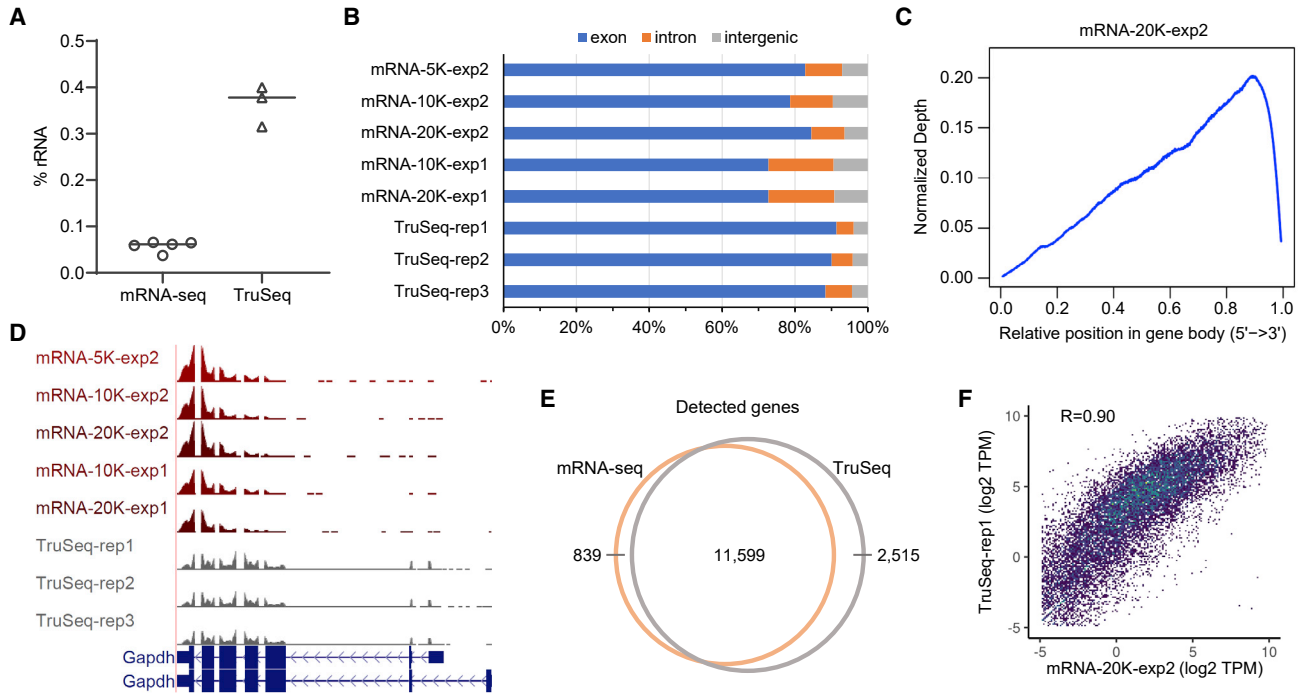


Figure 3. The mRNA-seq data generated with low-input ATAC&mRNA-seq are comparable with conventional bulk mRNA-seq data

(A) Column scattergraph showing the percentage of reads mapped to rRNA in each sample with the line indicating the median of each group.

(B) Distribution of mapped reads across known gene features (exons, introns, and intergenic regions).

(C) Gene body coverage profile of a representative mRNA-seq library prepared with low-input ATAC&mRNA-seq. The curve was smoothed over 15 consecutive points in the plot.

(D) UCSC genome browser view of mRNA-seq coverage tracks at Gapdh gene locus.

(E) Venn diagram showing overlap of detected genes (TPM $\geq$ 0.5) in low-input ATAC&mRNA-seq and conventional bulk mRNA-seq.

(F) Density scatterplot displaying correlation of gene expression measured by low-input ATAC&mRNA-seq and conventional bulk mRNA-seq. Each dot represents a gene with viridis color scale indicating density. Spearman's ρ value is shown at the top of the plot.

~8% in "Strong Enhancer," ~16% in "Enhancer," ~2% in "Weak/Poised Enhancer," and ~11% in "Insulator," consistent with previous findings that ATAC-seq peaks predominantly overlap with active and poised chromatin states while barely overlapping with repressed and inactive chromatin states (Tarbell and Liu, 2019). Taken together, the comprehensive analyses demonstrated that our method is comparable with Omni-ATAC in generating high-quality ATAC-seq data to identify the key regulatory regions controlling cell identity.

The mRNA-seq data generated with low-input ATAC&mRNA-seq are comparable with conventional bulk mRNA-seq data

To evaluate the quality of our mRNA-seq data, we compared them with previously published E14-mESC mRNA-seq data (Ramisch et al., 2019) generated by using an Illumina TruSeq Stranded mRNA kit, which is the prevailing library preparation kit for conventional bulk mRNA-seq. Compared with TruSeq libraries, our mRNA-seq libraries showed even lower percentage of rRNA contamination (average 0.06% versus 0.37%) (Figure 3A) and exhibited similar read distribution across known gene features with ~82% of reads mapped to exons, ~10% to introns, and ~8% to intergenic regions (Figure 3B), validating the mRNA origin of the libraries and negligible genomic DNA contamination with direct mRNA isolation from the cell lysate. Inspection of read coverage over gene bodies revealed a bias toward the 3′ end of genes (3′ bias) in our mRNA-seq data (Figures 3C, 3D, andS2A). Nevertheless, in total 12,438 genes were detected in our mRNA-seq data with a minimum expression threshold of 0.5 transcripts per million (TPM). The number of detected genes was slightly lower in our mRNA-seq data than in TruSeq data (12,438 versus 14,114), which was expected given that the input cell numbers were two orders of magnitude less in our experiment than in the TruSeq experiment and the genes detected only in TruSeq data were expressed at very low levels (Figure S2B). Importantly, 93.25% of the expressed genes (11,599 out of 12,438) were concordantly detected in TruSeq data (Figure 3E), and the measured gene expression levels exhibited strong correlation between our method and TruSeq across the transcriptome (Spearman's ρ = 0.90, Figure 3F), suggesting that our method performs comparably with conventional bulk mRNA-seq in terms of gene expression measurement but requires substantially fewer cells.

Integrative analysis of promoter accessibility and gene expression

mRNA-seq reveals the expression levels of genes whereas chromatin accessibility mapping with ATAC-seq uncovers the

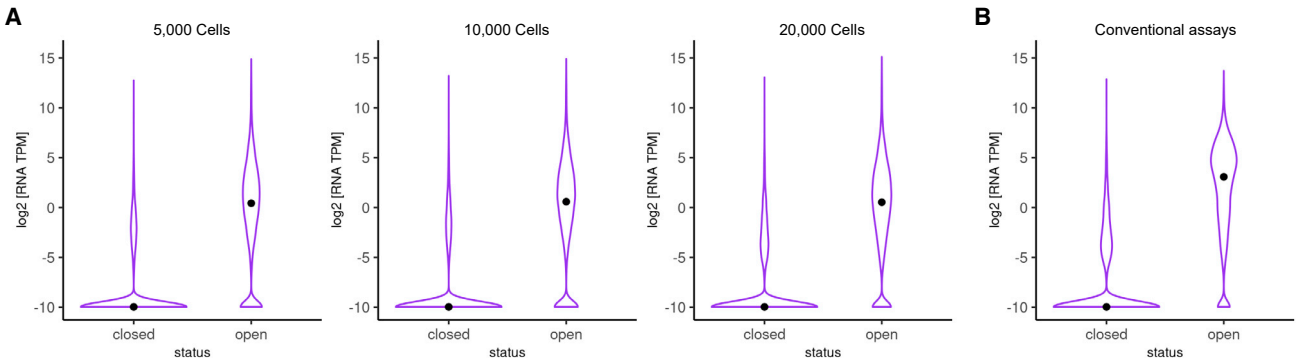


Figure 4. Integrative analysis of promoter accessibility and gene expression

(A and B) Violin plots showing mRNA expression levels of genes with "open" versus "closed" promoter by integrating ATAC-seq and mRNA-seq data generated with (A) low-input ATAC&mRNA-seq and (B) conventional mono-omics assays (Omni-ATAC and TruSeq mRNA sequencing) using substantially larger numbers of input cells in the same context (publicly available data). Black points indicate the median of mRNA expression levels in each category. For plotting purposes, a floor value of 0.001 TPM was applied.

associated regulatory landscape. With joint profiling of accessible chromatin and mRNA expression in the same cells, our method would enable a direct link of transcriptional regulation to its output. Indeed, integrative analysis of promoter accessibility and gene expression showed that "open" promoters were correlated with relatively high gene expression whereas "closed" promoters were associated with low gene expression (Figure 4A), which was similar to that observed by integrating conventional single-assay mRNA-seq and ATAC-seq data generated with substantially more input cells (Figure 4B), suggesting that our dual-omics profiling method performed comparably well in integrative data analysis. Notably, consistent results were obtained across different input cell numbers (Figure 4A), demonstrating the reliability of our method to detect the association between chromatin accessibility and gene expression with just thousands of cells.

Low-input ATAC&mRNA-seq is robust and reproducible

Lastly, to demonstrate the robustness of our method, we compared samples from the same batch but using different numbers of input cells. Remarkably, ATAC-seq peaks identified in those samples largely overlapped (Figures 5A and S4A), and ATAC-seq signals at peak regions showed very high correlation in pairwise comparisons of those samples (Pearson's r = 0.96–0.99) (Figures 5B and S4B), indicating high consistency in ATAC-seq enrichment despite different input amount. Likewise, a similar number (~11,000) of genes were detected across the mRNA-seq samples regardless of input cell numbers, and the vast majority (~90%) of the detected genes were shared among the mRNA-seq samples (Figures 5C and S4C). Moreover, the measured gene expression levels were highly consistent across the samples with different input cell numbers (Pearson's $r > 0.99$) (Figures 5D and S4D). Collectively, these analyses showed the robustness of our method against variation in the number of input cells. Next, to assess reproducibility of the data generated with our method, we compared samples from two independent experiments performed on different days. Irrespective of the number of input cells (10K or 20K), the independent samples consistently exhibited large overlap of ATAC-seq peaks (Figure 6A) and high correlation of ATAC-seq signals at unified peaks (Pearson's $r \geq 0.98$) (Figure 6B); similarly, they also showed high concordance in mRNA-seq data with ~92.5% of the detected genes in common (Figure 6C) and highly consistent gene expression levels (Pearson's $r \geq 0.99$) (Figure 6D), altogether indicating high reproducibility of the ATAC-seq and mRNA-seq data generated with our method. In conclusion, our method is robust and reproducible in obtaining both datum types simultaneously with low cell input.

DISCUSSION

We developed a dual-omics profiling method (low-input ATAC&mRNA-seq) to simultaneously measure chromatin accessibility and mRNA expression in the same cells with low cell number. Although multi-omics profiling of low-input material is challenging, we succeeded in that by improving both ATAC-seq and mRNA-seq procedures to minimize sample loss during the process. Our method uses whole cells instead of nuclei for chromatin tagmentation in ATAC (Figure 1). By removing the nuclei isolation step, it not only simplifies the experimental procedure but also avoids cell loss without significantly affecting the quality of ATAC-seq data. As a result, our method requires even fewer cells than Omni-ATAC to identify more accessible chromatin regions encompassing key regulatory features in the epigenome, including active and poised promoters, enhancers, and insulators (Figure 2). The advantage became more noticeable especially when starting with much fewer cells, as demonstrated by the superior performance of our method with 5K cells in terms of ATAC-seq library complexity, peak number, and consistency with larger-input samples (Figure 2). In addition, the improvement in the ATAC procedure enables downstream mRNA-seq to measure both cytoplasmic and nuclear transcripts. With whole-cell transcriptome profiling, it greatly increases the amount of mRNA available for library preparation and ensures fair comparison with publicly available mRNA-seq datasets, as most use RNA isolated from whole cells.

Conventional bulk mRNA-seq library construction involves many laborious and time-consuming steps, including RNA

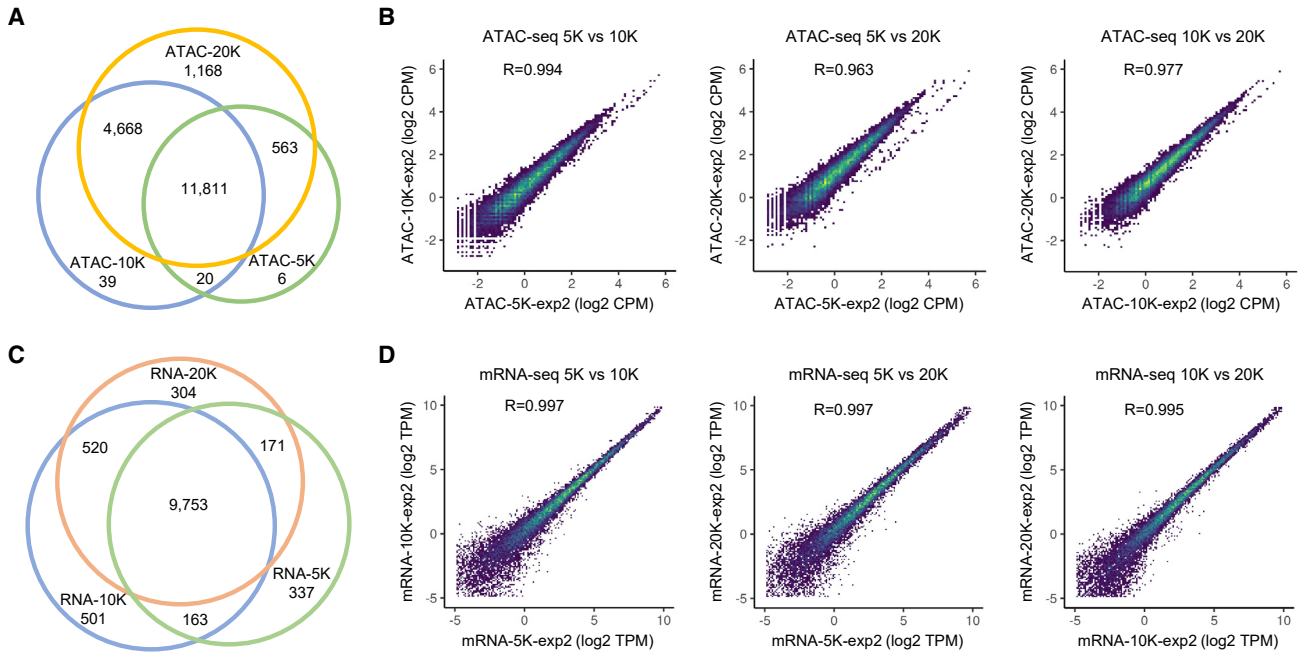


Figure 5. Low-input ATAC&mRNA-seq is robust against variation in the number of input cells

(A) Venn diagram showing overlap of ATAC-seq peaks identified with 20K, 10K, and 5K input cells.

(B) Density scatterplots displaying pairwise correlations of ATAC-seq data generated with 20K, 10K, and 5K input cells. Each dot represents an individual peak in the unified peak set with viridis color scale indicating density. Pearson's *r* value is shown at the top of each plot.

(C) Venn diagram showing overlap of detected genes (TPM $\geq$ 0.5) with 20K, 10K, and 5K input cells.

(D) Density scatterplots displaying pairwise correlations of gene expression measured with 20K, 10K, and 5K input cells. Each dot represents a gene with viridis color scale indicating density. Pearson's *r* value is shown at the top of each plot.

extraction, mRNA purification with poly(A) selection or ribosomal RNA depletion, mRNA fragmentation, reverse transcription, second-strand cDNA synthesis, end repair, adaptor ligation, and PCR amplification (Cui et al., 2010; Mortazavi et al., 2008); purification procedures needed between the enzymatic steps also cause inevitable sample loss. To minimize sample loss during mRNA-seq library preparation for low-cell-input samples, we replaced the complex traditional method with a simple "one-tube" method that consists of only three seamless steps, namely direct mRNA isolation from cell lysate with Dynabeads Oligo$(dT)_{25}$, on-bead cDNA synthesis, and tagmentation of mRNA/cDNA hybrids and PCR amplification (Figure 1). The whole procedure was performed on beads, enabling simple and rapid wash and buffer change between the steps without requiring any laborious purification. Our mRNA-seq approach greatly simplifies the workflow and minimizes sample loss. With just thousands of input cells, it generated acceptable mRNA-seq data that was comparable with those of conventional bulk mRNA-seq in terms of read distribution across gene features, number of detected genes, and gene expression levels, albeit with inferior coverage uniformity over the gene body (Figure 3). The 3′ bias in gene body coverage is a common phenomenon also observed in other similar approaches such as sequencing hetero RNA-DNA-hybrid (SHERRY) (Di et al., 2020) and transposase-assisted RNA/DNA hybrids co-tagmentation (TRACE-seq) (Lu et al., 2020), two recently published RNA-seq methods similar to our mRNA-seq approach in using oligo(dT) primed cDNA synthesis and Tn5-mediated direct tagmentation of RNA/cDNA hybrids for library preparation. However, this limitation might be overcome by using template-switching reverse transcription to generate full-length cDNA (Di et al., 2020), which could also enable identification of TSSs.

In summary, by coupling the simplified ATAC procedure with the novel mRNA-seq approach, low-input ATAC&mRNA-seq can simultaneously profile both chromatin accessibility and gene expression with low cell numbers ranging from 5K to 20K cells. The ATAC-seq and mRNA-seq data generated with our method were highly reproducible and strongly correlated with counterpart data of conventional mono-omics methods (Figures 2, 3, 4, 5, and 6), demonstrating the accuracy and reliability of our method to generate both datum types simultaneously with just thousands of cells. Our method uses commercially available off-the-shelf reagents and requires only basic molecular biology equipment. The speed, low-input requirement, technical simplicity, and robustness of our method make it appealing, especially in situations when it is challenging to collect large numbers of cells. Hence, we envision our method to be widely useful in many biological disciplines with limited materials. Furthermore, as a scalable method, low-input ATAC&mRNA-seq holds promise for even fewer input cells with further optimization of the protocol, which would also reduce reaction costs by using proportionally less amounts of reagents. Moreover, we speculate that by adapting the cleavage under targets and tagmentation (CUT&Tag) strategy (Kaya-Okur et al., 2019), our approach could be further

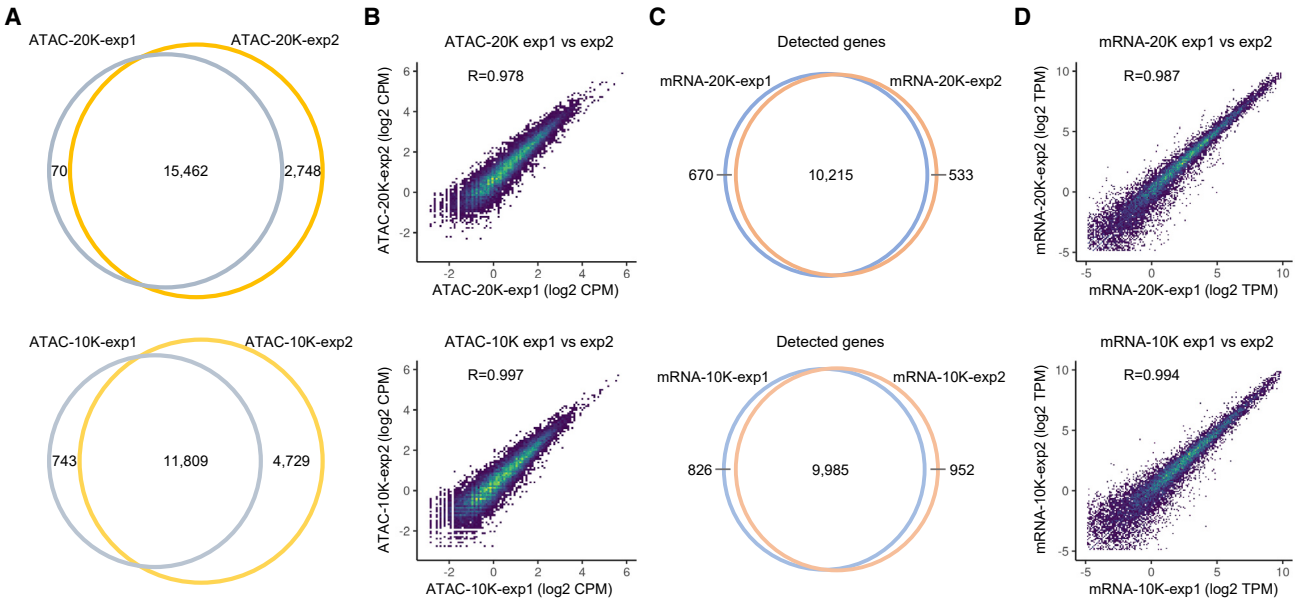


Figure 6. The ATAC-seq and mRNA-seq data generated with low-input ATAC&mRNA-seq are reproducible

(A) Venn diagrams showing overlap of ATAC-seq peaks identified in two independent experiments using the same numbers of input cells.

(B) Density scatterplots displaying correlation of ATAC-seq data of the two independent experiments. Each dot represents an individual peak in the unified peak set with viridis color scale indicating density. Pearson's *r* value is shown at the top of each plot.

(C) Venn diagrams showing overlap of detected genes (TPM $\geq$ 0.5) in the two independent experiments.

(D) Density scatterplots displaying correlation of gene expression measured in the two independent experiments. Each dot represents a gene with viridis color scale indicating density. Pearson's *r* value is shown at the top of each plot.

extended to simultaneous profiling of transcriptome and other epigenomic layers such as histone modification and transcription factor binding from the same cells.

Limitations of the study

Despite its unique advantages, our method has limitations. First, it requires fresh live cells and is thus not suitable for frozen samples. Second, the mRNA-seq data generated with this approach are not strand specific and are limited by the uneven read coverage along the gene body.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Low-input ATAC&mRNA-seq
  - Omni-ATAC-seq
  - Publicly available data used in this work
  - ATAC-seq data analysis
  - Peak calling and peak overlap across samples
  - Overlap of peaks with chromatin states
  - Promoter accessibility stratification
  - ATAC-seq nucleosome signal
  - RNA-seq data analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.crmeth.2021.100041.

ACKNOWLEDGMENTS

We thank Dr. Hua Wang at the Helin lab for providing the E14 mESCs and Dr. Kristian Helin for critical reading of the manuscript. This work was supported by donations to the Center for Epigenetics Research from The Metropoulos Family Foundation and The Ambrose Monell Foundation and by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (ES101965 to P.A.W.).

AUTHOR CONTRIBUTIONS

R.L. conceived the project and performed the experiments; R.L. and S.A.G. analyzed and interpreted the data; R.L. wrote the manuscript; S.A.G. and P.A.W. reviewed and edited the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 9, 2021
Revised: May 4, 2021
Accepted: May 24, 2021
Published: June 21, 2021

REFERENCES

Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S., and Crawford, G.E. (2008). High-resolution mapping and characterization of open chromatin across the genome. Cell *132*, 311–322.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods *10*, 1213–1218.

Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y., and Greenleaf, W.J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. Nature *523*, 486–490.

Cao, J., Cusanovich, D.A., Ramani, V., Aghamirzaie, D., Pliner, H.A., Hill, A.J., Daza, R.M., McFaline-Figueroa, J.L., Packer, J.S., Christiansen, L., et al. (2018). Joint profiling of chromatin accessibility and gene expression in thousands of single cells. Science *361*, 1380–1385.

Chen, K., Xi, Y., Pan, X., Li, Z., Kaestner, K., Tyler, J., Dent, S., He, X., and Li, W. (2013). DANPOS: dynamic analysis of nucleosome position and occupancy by sequencing. Genome Res *23*, 341–351.

Corces, M.R., Trevino, A.E., Hamilton, E.G., Greenside, P.G., Sinnott-Armstrong, N.A., Vesuna, S., Satpathy, A.T., Rubin, A.J., Montine, K.S., Wu, B., et al. (2017). An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. Nat Methods *14*, 959–962.

Cui, P., Lin, Q., Ding, F., Xin, C., Gong, W., Zhang, L., Geng, J., Zhang, B., Yu, X., Yang, J., et al. (2010). A comparison between ribo-minus RNA-sequencing and polyA-selected RNA-sequencing. Genomics *96*, 259–265.

Di, L., Fu, Y., Sun, Y., Li, J., Liu, L., Yao, J., Wang, G., Wu, Y., Lao, K., Lee, R.W., et al. (2020). RNA sequencing by direct tagmentation of RNA/DNA hybrids. Proc. Natl. Acad. Sci. U S A *117*, 2886–2893.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Giresi, P.G., Kim, J., McDaniell, R.M., Iyer, V.R., and Lieb, J.D. (2007). FAIRE ((F)under-barormaldehyde-(A)under-barssisted (I)under-barsolation of (R)under-baregulatory (E)under-barlements) isolates active regulatory elements from human chromatin. Genome Res. *17*, 877–885.

Kaya-Okur, H.S., Wu, S.J., Codomo, C.A., Pledger, E.S., Bryson, T.D., Henikoff, J.G., Ahmad, K., and Henikoff, S. (2019). CUT&Tag for efficient epigenomic profiling of small samples and single cells. Nat. Commun. *10*, 1930.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079.

Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930.

Liu, L., Liu, C., Quintero, A., Wu, L., Yuan, Y., Wang, M., Cheng, M., Leng, L., Xu, L., Dong, G., et al. (2019). Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity. Nat. Commun. *10*, 470.

Lu, B., Dong, L., Yi, D., Zhang, M., Zhu, C., Li, X., and Yi, C. (2020). Transposase-assisted tagmentation of RNA/DNA hybrid duplexes. eLife *9*, e54919.

Ma, S., Zhang, B., LaFave, L.M., Earl, A.S., Chiang, Z., Hu, Y., Ding, J., Brack, A., Kartha, V.K., Tay, T., et al. (2020). Chromatin potential identified by shared single-cell profiling of RNA and chromatin. Cell *183*, 1103–1116.e20.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal *17*, 10–12.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-seq. Nat. Methods *5*, 621–628.

Pintacuda, G., Wei, G., Roustan, C., Kirmizitas, B.A., Solcan, N., Cerase, A., Castello, A., Mohammed, S., Moindrot, B., Nesterova, T.B., et al. (2017). hnRNPK recruits PCGF3/5-PRC1 to the xist RNA B-repeat to establish polycomb-mediated chromosomal silencing. Mol. Cell *68*, 955–969.e10.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Ramisch, A., Heinrich, V., Glaser, L.V., Fuchs, A., Yang, X., Benner, P., Schopflin, R., Li, N., Kinkley, S., Romer-Hillmann, A., et al. (2019). CRUP: a comprehensive framework to predict condition-specific regulatory units. Genome Biol. *20*, 227.

Schones, D.E., Cui, K., Cuddapah, S., Roh, T.Y., Barski, A., Wang, Z., Wei, G., and Zhao, K. (2008). Dynamic regulation of nucleosome positioning in the human genome. Cell *132*, 887–898.

Tarbell, E.D., and Liu, T. (2019). HMMRATAC: a Hidden Markov ModeleR for ATAC-seq. Nucleic Acids Res. *47*, e91.

Xing, Q.R., Farran, C.A.E., Zeng, Y.Y., Yi, Y., Warrier, T., Gautam, P., Collins, J.J., Xu, J., Droge, P., Koh, C.G., et al. (2020). Parallel bimodal single-cell sequencing of transcriptome and chromatin accessibility. Genome Res. *30*, 1027–1039.

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

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, peptides, and recombinant proteins		
Digitonin	Promega	Cat# G9441
TWEEN® 20	Sigma-Aldrich	Cat# 11332465001
DPBS, no calcium, no magnesium	Gibco™	Cat# 14190250
SUPERase●In™ RNase Inhibitor (20 U/μL)	Invitrogen™	Cat# AM2696
Water, nuclease-free	Thermo Scientific™	Cat# R0581
Lithium chloride solution (8M)	Sigma-Aldrich	Cat# L7026-100ML
Ethylenediaminetetraacetic acid solution (0.5M EDTA)	Sigma-Aldrich	Cat# 03690-100ML
Sodium Acetate Solution (3 M), pH 5.2	Thermo Scientific™	Cat# R1181
AMPure XP beads	Beckman Coulter	Cat# A63881
Illumina Tagment DNA Enzyme and Buffer Large Kit	Illumina	Cat# 20034198
NEBNext® High-Fidelity 2X PCR Master Mix	NEB	Cat# M0541S
Critical commercial assays		
Dynabeads™ mRNA DIRECT™ Micro Purification Kit	Invitrogen™	Cat# 61021
SuperScript™ IV First-Strand Synthesis System	Invitrogen™	Cat# 18091050
Nextera XT DNA Library Preparation Kit	Illumina	Cat# FC-131-1024
Nextera XT Index Kit (24 indexes, 96 samples)	Illumina	Cat# FC-131-1001
MinElute PCR Purification Kit (250)	QIAGEN	Cat# 28006
Qubit™ dsDNA HS Assay Kit	Invitrogen™	Cat# Q32851
Deposited data		
Low-input ATAC&mRNA-seq data	This paper	GEO: GSE165478
Omni-ATAC-seq data	This paper	GEO: GSE165478
Publicly available TruSeq mRNA-seq and ATAC-seq data	NCBI GEO	GEO: GSE120376
Experimental models: cell lines		
E14 mouse embryonic stem cells	Kristian Helin lab	RRID: CVCL_C320
Oligonucleotides		
Universal i5 primer: AATGATACGGCGACCACCGAGATC TACACTCGTCGGCAGCGTCAGATGTG	Buenrostro et al., 2013	Ad1_noMX
index i7 primer_Ad2.31: CAAGCAGAAGACGGCATACG AGATTAACTTATGTCTCGTGGGCTCGGAGATGTG	Buenrostro et al., 2015	v2_Ad2.31_ATAAGTTA
index i7 primer_Ad2.32: CAAGCAGAAGACGGCATACG AGATCGAGTGATGTCTCGTGGGCTCGGAGATGTG	Buenrostro et al., 2015	v2_Ad2.32_ATCACTCG
index i7 primer_Ad2.34: CAAGCAGAAGACGGCATAC GAGATCTACCATTGTCTCGTGGGCTCGGAGATGTG	Buenrostro et al., 2015	v2_Ad2.34_AATGGTAG
index i7 primer_Ad2.35: CAAGCAGAAGACGGCATACG AGATACGTGCTCGTCTCGTGGGCTCGGAGATGTG	Buenrostro et al., 2015	v2_Ad2.35_GAGCACGT
index i7 primer_Ad2.36: CAAGCAGAAGACGGCATACG AGATTGACGAAAGTCTCGTGGGCTCGGAGATGTG	Buenrostro et al., 2015	v2_Ad2.36_TTTCGTCA
Software and algorithms		
Cutadapt v1.12	Martin, 2011	N/A
Bowtie2 v2.1.0	Langmead and Salzberg, 2012	N/A
samtools v1.3.1	Li et al., 2009	N/A
Picard tools MarkDuplicates.jar (v1.110)	http://broadinstitute.github.io/picard	N/A
BEDtools v2.24.0	Quinlan and Hall, 2010	N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
bedGraphToBigWig	http://hgdownload.soe.ucsc.edu/admin/exe/	N/A
MACS2 v2.1.1	Zhang et al., 2008	N/A
STAR v2.5	Dobin et al., 2013	N/A
featureCounts (Subread v1.5.0-p1)	Liao et al., 2014	N/A

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Ruifang Li (lir1@mskcc.org).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The datasets generated during this study are available at NCBI Gene Expression Omnibus (GEO) with accession number GSE165478.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Mouse E14 embryonic stem cells (129/Ola background) were maintained in Glasgow Minimum Essential Medium (GMEM, Sigma) containing 15% fetal bovine serum, hemented with 1× Pen-Strep (Gibco), 2 mM Glutamax (Gibco), 50 μM β-mercaptoethanol (Gibco), 0.1 mM nonessential amino acids (Gibco), 1 mM sodium pyruvate (Gibco), and Leukemia Inhibitory Factor (LIF, 1000U/ml, Millipore).

METHOD DETAILS

Low-input ATAC&mRNA-seq

Harvested E14 mouse ESCs were washed twice with cold PBS and then aliquoted into three Eppendorf tubes with 20K, 10K, and 5K cells, respectively. The cells were then spun down at 500 x g, 4°C for 5 min in a pre-chilled fixed-angle microcentrifuge. The supernatant was removed carefully without disturbing the cell pellet. The cell pellets were then resuspended in 20 μl, 10 μl, and 8 μl of transposition mix (25 μl 2×TD buffer, 2.5 μl Tn5 (Illumina), 16.5 μl PBS (Gibco), 0.5 μl 1% digitonin (Promega), 0.5 μl 10% Tween-20 (Sigma), 2.5 μl RNase inhibitor (Invitrogen), and 2.5 μl nuclease-free water (Thermo Scientific)), respectively, by pipetting up and down six times. Permeabilization/transposition reactions were incubated at 37 °C for 30 min in a thermomixer with shaking at 1000 rpm. Once the incubation was complete, EDTA (Sigma) and LiCl (Sigma) were added to a final concentration of 10 mM and 0.5 M, respectively, followed by adding 100 μl of Lysis/Binding Buffer from Dynabeads® mRNA DIRECT™ Micro Kit (Invitrogen) and pipetting up and down to lyse the cells. After complete cell lysis, 20 μl pre-washed Dynabeads® Oligo (dT)25 were added into the cell lysate and the mixture was incubated at room temperature for 5 min with continuous rotation to allow the mRNA to anneal to the oligo(dT) on Dynabeads. The sample tubes were then placed on a Dynal magnet for 1 min to separate the mRNA and genomic DNA (gDNA). The supernatant containing tagmented gDNA was transferred into a new Eppendorf tube followed by DNA purification with MinElute PCR purification kit (Qiagen). Meanwhile, the mRNA captured on beads was washed extensively according to the manual of Dynabeads® mRNA DIRECT™ Micro Kit (Invitrogen). The Dynabeads-mRNA complex was then resuspended in 20 μl of reverse transcription reaction mix without any primer prepared using SuperScript™ IV First-Strand Synthesis System (Invitrogen), and the reaction was incubated initially at 50°C for 5 min and then at 55°C for 10 min. The bead-bound oligo(dT) was used as primer for first strand cDNA synthesis, and as a result, the mRNA/cDNA hybrids were covalently bound to the Dynabeads. Once reverse transcription was complete, the PCR tubes were immediately placed on the magnet for 30 seconds and the supernatant was then removed. The Dynabeads-mRNA/cDNA complex was washed twice in 100 μl of ice-cold 10 mM Tris-HCl and then resuspended in 5 μl of ice-cold 10 mM Tris-HCl. Direct tagmentation of the mRNA/cDNA hybrids and PCR amplification were performed on beads using Nextera XT DNA Library Prep Kit (Illumina); the Reference Guide was followed exactly for 20K and 10K samples, while the amounts of reagents used for 5K sample were scaled down by half. In the meantime, previously purified ATAC-DNA was amplified using NEBNext® High-Fidelity 2X PCR Master Mix (NEB) and universal i5 and index i7 primers (Buenrostro et al., 2015) with 10 cycles of PCR. The ATAC-seq and mRNA-seq libraries were purified with AMPure XP beads (Beckman Coulter) and then quantified with Qubit dsDNA HS Assay Kit (Invitrogen). The libraries were pooled and sequenced on Illumina NextSeq550 with paired-end 75-bp reads.

Omni-ATAC-seq

Omni-ATAC-seq was performed on E14 mESCs following Omni-ATAC protocol (Corces et al., 2017) with slight modifications. Briefly, harvested mESCs were counted and then aliquoted into three Eppendorf tubes with 50K, 10K, and 5K cells, respectively. The cells were washed once with cold ATAC-seq resuspension buffer (RSB; 10 mM Tris-HCl pH 7.4, 10 mM NaCl, and 3 mM $MgCl_2$ in water) followed by lysing the cells on ice for 3 min in 50 μl of ATAC-seq RSB containing 0.1% NP40 (Roche), 0.1% Tween-20 (Sigma), and 0.01% Digitonin (Promega). After cell lysis, 1 mL of cold ATAC-seq RSB containing 0.1% Tween-20 (without NP40 or digitonin) was added and then mixed by inverting the tubes. Isolated nuclei were spun down at 500 x g, 4 °C for 10 min in a pre-chilled fixed-angle centrifuge. After removing the supernatant, the nuclei of 50K, 10K, and 5K samples were resuspended in 50 μl, 10 μl, and 5 μl of transposition mix (25 μl 2x TD buffer, 2.5 μl Tn5 (Illumina), 16.5 μl PBS (Gibco), 0.5 μl 1% digitonin (Promega), 0.5 μl 10% Tween-20 (Sigma), 5 μl nuclease-free water (Thermo Scientific)), respectively, by pipetting up and down six times. Transposition reactions were incubated at 37 °C for 30 min in a thermomixer with shaking at 1,000 rpm. The reactions were cleaned up with MinElute PCR Purification Kit (Qiagen), and the purified DNA was amplified using NEBNext® High-Fidelity 2X PCR Master Mix (NEB) with 10 cycles of PCR. The ATAC-seq libraries were purified with AMPure XP beads (Beckman Coulter) and quantified with Qubit dsDNA HS Assay Kit (Invitrogen). The libraries were then pooled and sequenced on Illumina NextSeq550 with paired-end 35-bp reads.

Publicly available data used in this work

Previously published datasets of conventional mRNA-seq and ATAC-seq on E14 mESCs were downloaded as raw fastq files from GEO series GSE120376 (GSM3399470, GSM3399471, GSM3399472, GSM3399494) (Ramisch et al., 2019).

ATAC-seq data analysis

ATAC-seq fastq files were filtered to remove all entries with a mean base quality score below 20 for either read in the read pair. Adapters were removed via Cutadapt v1.12 (Martin, 2011) with parameters "-a CTGTCTCTTATA -O 5 -q 0", and the trimmed reads were further filtered to exclude those with length less than 30 bp. The remaining filtered and trimmed read pairs were mapped against the mm10 reference assembly via Bowtie2 v2.1.0 (Langmead and Salzberg, 2012) with parameters "-X 2000 –fr –end-to-end –very-sensitive", followed by filtering with samtools v1.3.1 (Li et al., 2009) at MAPQ5. Reads mapped to chrM were ignored in all downstream analysis. Duplicate mapped read pairs were removed by Picard tools MarkDuplicates.jar (v1.110) (http://broadinstitute.github.io/picard). Only the 9 bp at the 5′ end of each read was retained for downstream analysis. Coverage tracks for genome browser views were generated using BEDtools v2.24.0 (Quinlan and Hall, 2010) genomeCoverageBed with depth normalized to 10 million read ends per sample and then converted to bigWig format with UCSC utility bedGraphToBigWig (http://hgdownload.soe.ucsc.edu/admin/exe/).

Peak calling and peak overlap across samples

MACS2 v2.1.1 (Zhang et al., 2008) was used for initial peak calls per sample with parameters "callpeak -g mm -q 0.0001 –keep-dup=all –nomodel –extsize 9", followed by merging peaks within 200bp with BEDtools v2.24.0 mergeBed. To facilitate comparisons across samples, a single set of unified peaks was generated by collapsing called peaks and identifying only peak regions that overlap from at least 3 of 8 samples via BEDtools v2.24.0 unionBedGraphs, again followed by merging peaks within 200bp with BEDtools v2.24.0 mergeBed. Unified peaks less than 50 bp in width were discarded, as were peaks outside canonical chromosomes (chr1-19, X, Y). Analysis of shared peaks among samples was performed using subsets of unified peaks that overlap the MACS2 peak calls of given samples. Briefly, the unified peak set was intersected individually with MACS2 peak calls from each sample in comparison. The subset of unified peaks in the intersection represented the total peaks in the sample. These subsets of unified peaks were then intersected with each other to determine the overlap of peaks among samples.

Overlap of peaks with chromatin states

Chromatin states of mESC (Pintacuda et al., 2017) as defined by ChromHMM were downloaded from https://github.com/guifengwei/ChromHMM_mESC_mm10. To investigate the epigenomic contexts of ATAC-seq peaks, each peak was assigned to the chromatin state with which it showed the most overlap.

Promoter accessibility stratification

Promoter was defined as the region +/− 1Kb of the annotated TSS. To determine promoter accessibility, ATAC-seq signal was quantified by counting number of 9-mer read ends per promoter region with BEDtools v2.24.0 multiBamCov and then converted to CPM (counts per million). Promoters were stratified as either "open" or "closed" with a hard cutoff of 1.45 CPM, which was derived from the bimodal distribution of ATAC-seq signals at promoters (Figure S3). For genes with multiple TSS, we selected the one with the maximum ATAC-seq CPM that was averaged over all evaluated samples.

ATAC-seq nucleosome signal

Due to limited sequencing depth, the processed ATAC-seq data of all low-input ATAC&mRNA-seq samples were merged for nucleosome signal analysis. Read pairs were binned according to insert size as nucleosome-free (up to 100 bp), mono-nucleosome (180-250 bp), and di-nucleosome (330-450 bp). Size ranges were established based on overlap of exponential (for sub-nucleosome) and

Gaussian (for mono-, di-, and tri-nucleosome) functions fit to the observed insert size distribution. Fragments from the di-nucleosome bin were split into two fragments then aggregated with the mono-nucleosome bin. Smoothed background-subtracted nucleosome signal was determined via the 'dpos' function of Danpos v2.2.2 (Chen et al., 2013) at parameters "-p 1 -a 1 -jd 20 -m 1", with the nucleosome-free fragments submitted as background. The nucleosome signal analysis was limited to canonical autosomes, chrX, and chrY.

RNA-seq data analysis

RNA-seq fastq files were filtered to remove all entries with a mean base quality score below 20 for either read in the read pair. Filtered read pairs were mapped against the mm10 reference assembly via STAR v2.5 (Dobin et al., 2013) with parameters "–outSAMattrIHstart 0 –outFilterType BySJout –alignSJoverhangMin 8 –limitBAMsortRAM 55000000000 –outSAMstrandField intronMotif –outFilterIntronMotifs RemoveNoncanonical". Coverage tracks were generated with STAR v2.5 with parameters "–runMode inputAlignmentsFromBAM –outWigType bedGraph –outWigStrand Unstranded –outWigNorm RPM", followed by conversion to bigWig format via UCSC utility bedGraphToBigWig.

Read counts per gene were determined via featureCounts (Subread v1.5.0-p1) (Liao et al., 2014) with parameters "-s0 -Sfr -p" and then converted to TPM (transcripts per million). The gene models used in this study were from NCBI RefSeq annotations limited to only curated transcripts, with a GTF format version downloaded from the UCSC Genome Browser (http://hgdownload.soe.ucsc.edu/goldenPath/mm10/bigZips/genes/, dated January 10, 2020).

Gene body coverage profile was generated by calculating the number of mapped reads that overlap with genomic bins tiled over exonic regions of a given gene model (where each bin covers 0.1%) via BEDtools v2.24.0 intersectBed, then aggregating over all gene models, and finally normalizing by total counts.

QUANTIFICATION AND STATISTICAL ANALYSIS

Pearson correlation coefficient (Pearson's r) and Spearman correlation coefficient (Spearman's ρ) were calculated using R (https://www.r-project.org/).