Single-cell quantification of a broad RNA spectrum reveals unique noncoding patterns associated with cell types and states

Alina Isakovaa, Norma Neffb, and Stephen R. Quakea,b,c

*Department of Bioengineering, Stanford University, Stanford, CA 94305; 1Chan Zuckerberg Biohub, San Francisco, CA 94158; and 5Department of Applied Physics, Stanford University, Stanford, CA 94305

Contributed by Stephen R. Quake; received November 8, 2021; accepted November 10, 2021; reviewed by Jay Shendure and Jonathan Weissman

The ability to interrogate total RNA content of single cells would enable better mapping of the transcriptional logic behind emerging cell types and states. However, current single-cell RNA-sequencing (RNA-seq) methods are unable to simultaneously monitor all forms of RNA transcripts at the single-cell level, and thus deliver only a partial snapshot of the cellular RNAome. Here we describe Smart-seq-total, a method capable of assaying a broad spectrum of coding and noncoding RNA from a single cell. Smart-seq-total does not require splitting the RNA content of a cell and allows the incorporation of unique molecular identifiers into short and long RNA molecules for absolute quantification. It outperforms current poly(A)-independent total RNA-seq protocols by capturing transcripts of a broad size range, thus enabling simultaneous analysis of protein-coding, long-noncoding, microRNA, and other noncoding RNA transcripts from single cells. We used Smart-seq-total to analyze the total RNAome of human primary fibroblasts, HEK293T, and MCF7 cells, as well as that of induced murine embryonic stem cells differentiated into embryoid bodies. By analyzing the coexpression patterns of both noncoding RNA and mRNA from the same cell, we were able to discover new roles of noncoding RNA throughout essential processes, such as cell cycle and lineage commitment during embryonic development. Moreover, we show that independent classes of short-noncoding RNA can be used to determine cell-type identity.

Results

Smart-seq-total relies on the ability of Escherichia coli poly(A) polymerase to add adenine tails to the 3’ prime of RNA molecules. Total polyadenylated RNA is then reverse-transcribed using anchored oligo dT in the presence of the template switch oligo (TSO) (20) (Fig. 1A). Compared to previous studies that explored similar approaches to construct libraries from total RNAs (21, 22), Smart-seq-total utilizes an optimized version of the TSO (19), specifically engineered to be rapidly eliminated from the reaction through enzymatic digestion directly following the reverse transcription. This allows us to remove the “contaminant” constructs, originating from polyA-tailing and mispriming of TSO, which otherwise dominate the resulting cDNA.

Significance

Each individual cell transcribes nearly 85% of the genome. However, when it comes to the analysis of cellular RNA, most studies are still only looking at the 3% that correspond to protein-coding transcripts. The role and abundance of the remaining RNAs across individual cells remain largely unknown. In the present study we describe Smart-seq-total, an RNA-sequencing method that delivers a broad picture of the total cellular RNA content. Using Smart-seq-total, we analyzed the content of hundreds of human and mouse cells and showed that the noncoding RNA content of cells significantly differs across cell types and dynamically changes throughout the vital processes of a cell, such as cell cycle and cell differentiation.
sequencing library and render the short RNA transcripts undetectable (SI Appendix, Fig. S1 A and B). Additionally, we employ a CRISPR-mediated removal of overrepresented sequences, which allows us to eliminate the majority of the sequences corresponding to ribosomal RNA from the final library in a single-pool reaction (23) (targeting rRNA regions and other abundant RNA molecules) (SI Appendix, Fig. S1 A and B and Table S1).

When applied to single HEK293T cells, Smart-seq-total identified the full complement of mRNA as well as a broad spectrum of ncRNA genes, such as small nuclear RNA (snRNA), small Cajal body-specific RNA ( scaRNA), histone RNA, and IncRNA. The majority of these molecules endogenously lack poly(A) tails and thus cannot be captured through a direct polyA-priming employed by Smart-seq2 (1) or other popular scRNA-seq methods (SI Appendix, Fig. S1 C and D). Among other ncRNA detected by Smart-seq-total, we found rRNA (rRNAs) and mature miRNAs (SI Appendix, Fig. S1 C and D). To facilitate the quantitative comparison across RNA biotypes within a cell, we also implemented a UMI-compatible version of Smart-seq-total (v2) (SI Appendix, Fig. S2 A–C). Like Smart-seq3, Smart-seq-total (v2) does not require a clean-up step after reverse transcription and thus is also potentially subject to random UMI incorporation during cDNA amplification (24). We show, however, that under the current protocol conditions these events are negligible (SI Appendix, Fig. S2 D–F). Both versions of Smart-seq-total are compatible with the Tn5-based fragmentation of cDNA in order to sequence full-length transcript coverage (SI Appendix, Fig. S3).

The sensitivity of Smart-seq-total estimated based on external RNA control consortium (ERCC) capture is comparable to Smart-seq2 (SI Appendix, Fig. S3 E and F) (25). At the same time, Smart-seq-total detects a broader spectrum of RNA types than previous single-cell approaches and furthermore allows the incorporation of UMIs for absolute quantitation into both short and long RNA molecules (SI Appendix, Figs. S1 and S4).

To demonstrate the scalability of the method, we sequenced total RNA from individual human primary dermal fibroblasts (n = 277), HEK293T (n = 245), and MCF7 (n = 90) cells sorted in 384-well plates and processed in one-tenth of the standard Smart-seq2 volume (i.e., cells are sorted in 0.3 μL of lysis buffer and RNA is reverse-transcribed in 1 μL) (Materials and Methods) (19). We sequenced the libraries to obtain one to two million reads per cell, mapped, and counted the reads using the reference that contains both coding and noncoding transcripts (including mRNA, snoRNA, small nuclear RNA [snRNA], tRNA, and so forth). Within all three cell types, we identified a broad spectrum of transcripts, such as mRNA, miRNA, IncRNA, and snoRNA in each profiled cell (Fig. 1B and SI Appendix, Fig. S5). We found metazoan RNA7SK and RN7SL1, which are involved in regulation of transcription and translation, respectively (annotated as “miscellaneous RNA”-type [miscRNA] in the GENCODE database) to be the most abundant in our data comprising together ~40% of all mapped reads (Fig. 1B and SI Appendix, Fig. S6). We demonstrate that, if desired, these molecules could also be depleted from the
Histone RNA is another type of mainly nonpolyadenylated RNA, which we observed to be strongly correlated with the cell cycle. Consistent with prior studies (33, 34), histone RNA levels sharply rise during the S phase in all three profiled cell types (Fig. 2C). The ability to capture nonpolyadenylated histones also has a strong impact on cell clustering, by introducing a cell-cycle bias. Particularly, histones drive the separation of each cell type into two distinct populations (SI Appendix, Fig. S13A), marked by increased levels of the majority of histone genes during the DNA replication phase (SI Appendix, Fig. S13B).

In addition to being expressed in a cell-cycle–dependent manner, we also identified several histones to be cell-type specific. For example, HIST1H1B is expressed in fibroblasts but absent in HEK293T and MCF7 cells, while HIST1H4B is absent in HEK293T cells while present in the other two cell types (Fig. 2D). Given the importance of histones in establishing and maintaining a distinct chromatin landscape of a cell, we anticipate that the ability to measure corresponding transcripts could be valuable for predicting the epigenetic state of a cell.

In principle, Smart-seq-total is designed to broadly quantify total cellular RNA content. However, we also show that short, less abundant molecules, such as miRNAs (35), can be size-selected from a UMI-tagged and indexed Smart-seq-total library for further in-depth analysis (Fig. 2E). In the example of 24 HEK293T cells we demonstrate that the size-based enrichment strategy yields results comparable to the state-of-the-art single-cell small RNA-seq method (11) in terms of the number and type of miRNAs detected per cell (Fig. 2F and SI Appendix, Fig. S14 A–C). Among the most abundant miRNAs in profiled HEK293T cells, we identified multiple members of the three conserved paralog clusters—miR-17/92, miR106a/363, and miR-10b/25 (36)—as well as various members of the let-7 miRNA family (SI Appendix). We noted that levels of mature miRNAs generally correlated better within a cluster rather than across different clusters (SI Appendix, Fig. S14 D–F). However, the abundance of mature forms largely varied across cluster members. Specifically, we found that the levels of two mature miRNAs, miR-92a-3p and miR-25-3p, were several folds higher than those of any other cluster member (Fig. 2G). Selective retention of miR-92a was previously observed at the tissue level in vivo (36) and has been attributed to drive fetal posttranscriptional cell targeting of cluster members (37, 38). Our data indicate that the phenomenon of selective miRNA retention can be observed at the single-cell level and that it extends beyond the 17/92 cluster.

While short RNAs are selected from a fully constructed UMI-indexed Smart-seq-total library, the remaining total library can be used to quantify miRNA counterpart of the same cells. We applied this strategy to investigate the relationship of the retained miR-92a-3p and miR-25-3p with the rest of the transcriptome at the physiological cell state. Specifically, we generated a matching miRNA–mRNA profile for ∼300 HEK293T cells by sequencing the complete Smart-seq-total library, as well as its small RNA fraction (SI Appendix). We used the obtained data to perform a pair-wise correlation of miR-92a-3p and miR-25-3p with the rest of the genes across profiled cells. Among the most correlated and anticorrelated genes (Pearson’s r < −0.5 or > 0.5, adjusted P < 0.01), we identified several validated targets of the respective miRNAs (Fig. 2H and SI Appendix, Fig. S15A). Interestingly, we also found that several other genes, which are neither known nor predicted to directly interact with the respective miRNAs, correlate with either miR-92a-3p, miR-25-3p or both ([|bs(r)|] > 0.5, adjusted P value <0.01) (SI Appendix, Fig. S15B and Dataset S2). Gene ontology (GO) analysis on top correlated and anticorrelated genes (Materials and Methods) revealed energy metabolism as well as RNA translation among their main molecular functions (SI Appendix, Fig. S15B). The MiR-17/92 cluster has been previously shown to regulate tumor metabolic reprogramming (39). The results of our miRNA–mRNA sequencing libraries at the rRNA depletion step using dedicated CRISPR guides (SI Appendix, Fig. S1 A and B and Table S1). Among cell-type–specific transcripts, we found well-characterized marker genes for either fibroblasts (COL1A2, FN1, MEG3), HEK293T (CKB, AMOT, HEY1), or MCF7 cells (KRT8, TFF1) (Fig. 1C and SI Appendix, Fig. S8A), as well as transcripts that belong to various types of ncRNA, such as microRNA, snoRNA, and lncRNA (Fig. 1C and SI Appendix, Fig. S8). For example, we found high levels of MIR222 in fibroblasts while we could not detect it in MCF7 cells. We also observed that oncogenic miRNA cluster MIR17HG is specific to HEK293T cells, while not found in either fibroblasts or MCF7 cells. In contrast, MCF7-specific transcripts include lncRNA, such as LINC00052, as well as snoRNA, such as SNORD71 and SNORD104.

Given the observed differences in the levels of ncRNA across profiled cells, we next asked whether ncRNA alone could be used to distinguish cell types. To answer this question, we performed principal component analysis (PCA) followed by the dimensionality reduction through t-distributed stochastic neighbor embedding (t-SNE) on the genes corresponding to one or multiple ncRNA types. Evaluation of the similarity between cells in two-dimensional space revealed that, in addition to lncRNA (26), miRNA taken alone separate the investigated cell types into three distinct clusters. Combining snoRNA, scaRNA, snRNA, and tRNA together allowed us to achieve similar results (Fig. 1D). While the exact function of individual snoRNA and scaRNA remains largely uncharacterized, these RNAs are believed to play a vital role in posttranscriptional and posttranslation control (27–29). Here we show that their abundance is also cell-type specific (SI Appendix, Fig. S8A).

After binning all the cells according to cell-cycle phases (30), we observed that in addition to cell-type–dependent differences in ncRNA, the abundance of certain noncoding transcripts also changed throughout the cell cycle (Fig. 2A). In agreement with previous bulk studies, which suggested the involvement or miRNA in cell-cycle regulation (31, 32), we found that levels of a subset of miRNAs in a cell dynamically change through the cell cycle, peaking at either S, G2M, or G1 phase (Fig. 2A). For example, our data showed that the levels of MIR16-2 in fibroblasts are high during the S phases and then sharply drop during G2M and G1 phases (SI Appendix, Fig. S9). The opposite holds true for MIR222, in both fibroblasts and HEK293T cells, which is more abundant during cell proliferation (G1) and decays during DNA replication (S) and cell division (G2M) phases (Fig. 2A and SI Appendix, Fig. S10). Among miRNAs more abundant during G2M phase, we identified MIR27A, MIR103A2, and MIR877 (SI Appendix, Figs. S9–S11). In addition to miRNA, a large number of lncRNA, snRNA, scaRNA, snRNA, and miscRNA were also up-regulated (log2 fold-change [FC] > 1, adjusted P < 0.01) during the G2M phase (Fig. 2A and SI Appendix, Figs. S9–S11). Given the active role of these RNA types in splicing and ribosome biogenesis, we suggest that they are produced by the cell in response to a rapid demand for protein synthesis and cell growth during the G2M phase.

To further link the observed ncRNA dynamics with the expression of well-characterized cell-cycle miRNA markers, we searched for correlated coding and noncoding genes throughout the cell cycle. We identified 24 clusters comprised of coexpressed coding and noncoding genes specific to either one or multiple cell types (Fig. 2B, SI Appendix, Fig. S12, and Dataset S1). Two of these mixed-gene clusters (33 genes up-regulated in the S phase and 53 genes up-regulated in the G2M phase) showed identical patterns in all three profiled cell types. Interestingly, both clusters are marked by landmark cell-cycle genes—such as CCNA2, MCM6, and TOP2A—but also include miRNAs, lncRNAs, and snRNAs previously unknown to follow a distinct expression pattern upon transition between phases.
coanalysis align with those findings and indicate that miR-92a-3p and miR-25-3p miRNA levels are linked to the expression of a metabolic gene set at the physiological cell state.

Finally, we sought to understand whether the unique noncoding signature acquired by different cell types is established during early stages of cell development, and if so, how dynamic it is with respect to cellular transcriptome. To address this question, we referred to an in vitro model of early lineage commitment: the differentiation of pluripotent stem cells into embryoid bodies (EBs) (40). The role of ncRNA in maintaining stem cell pluripotency and lineage commitment has been demonstrated previously through bulk experiments (41, 42). Thus, we hypothesized that applying Smart-seq-total to single cells at different stages of EB formation would allow us to identify coexpressed coding and noncoding transcripts within emerging lineages. As such, we analyzed the RNAome of primed pluripotent stem cells and that of individual cells obtained from dissociated EB at days 4, 8, and 12 of culture (∼200 cells per each time point, 913 cells in total) (Fig. 3A). Consistent with previous studies (43), the number of coding genes expressed by pluripotent stem cells was also higher compared to differentiated progenitors (SI Appendix, Fig. S16A). This was also the case for several
Fig. 3. Coding and noncoding signature of differentiated single mESCs. (A) Microscope images and corresponding schematic representations of EB formation at four sampled time points. PIE charts represent distribution of mapped reads across RNA types. Genes were assigned to a specific biotype based on GENCODE M23 annotation for the reference chromosomes. tRNA was quantified by mapping the reads, nonmapping to any other RNA type, to the high-confidence gene set obtained from GtRNAdb. (B) Exemplary coding and noncoding genes that are up- or down-regulated during EB formation. Subpanels are grouped according to RNA type. (C) UMAP plot of collected cells colored by timepoint. Cells were clustered using a k-nearest-neighbor algorithm and cell lineages were annotated based on the expression of marker genes within the identified clusters. (D) Lineage tree of EB differentiation. Each dot represents a cell colored according to the assigned lineage. Cells are arranged according to the computed pseudotime. (E) UMAP plot of collected cells colored by pseudotime. (F) Heatmap showing the variability in coding and noncoding gene expression across identified clusters. (G) Temporal and lineage-specific expression of selected protein-coding, IncRNA, and miRNA genes. Each column from left to right shows genes specific to: pluripotency state, ectoderm, mesoderm, or endoderm lineages.
ncRNA types, such as lncRNA, miRNA, and scaRNA (SI Appendix, Fig. S16B). Specifically, we observed that the levels of certain snoRNAs (such as Snord17, Snora23, and Snord87), scaRNAs (such as Scarna13 and Scarna6), IncRNAs (Plat3, Lincel1, Snhg9, Gm31659, and so forth), and miRNAs (Mir92-2, Mir302b, and Mir19b-2) go down (log2-FC > 1, adjusted P < 0.01) after cells exit pluripotency (Fig. 3B). In contrast, we also identified that the levels of several IncRNAs (Tieg1,Meg3, Lockd) and miRNAs (Mir298, Mir351, Mir370) increase with differentiation (Fig. 3B).

Louvain clustering of all collected cells revealed the presence of six molecularly distinct populations, which we assigned to: primed mouse embryonic stem cells (mESC), pre-ectoderm, ectoderm, endoderm, ectomesoderm, and mesoderm (Fig. 3C), based on the expression of known lineage-specific marker genes (e.g., Nanog and Pou5f1 for pluripotent cells, Pax6 and Olf787 for ectoderm, Afp and Shh for endoderm, Acta2 and Col3a1 for mesoderm (44)) (SI Appendix, Fig. S17 and Dataset S3). The analysis of genes differentially expressed between primed mESCs and each of the identified clusters showed that in addition to well-characterized lineage-specific mRNAs (SI Appendix, Fig. S18A) (45, 46) and IncRNAs (Tieg1 in ectodermal and Meg3 mesodermal lineages, respectively) (6), other ncRNA genes—such as miRNAs, scaRNAs, snoRNAs, tRNAs, and histone RNAs—are either specifically expressed or down-regulated within a certain lineage (log2-FC > 1, adjusted P < 0.05) (SI Appendix, Fig. S18B and Dataset S3).

We next used PAGA (47) to infer a developmental trajectory and compute pseudotime coordinates for each cell in our dataset (Materials and Methods and Fig. 3 D and E). Aligning cells in pseudotime within each lineage further confirmed the existence of expression gradient within different RNA types (Fig. 3F). Furthermore, we found that the majority of identified variable noncoding transcripts were germ-layer specific. Examples of such transcripts include Mir2137, Mir320, Gm49024, and Gm38708 in ectoderm; Mir351, Mir370, and Meg3 in mesoderm; as well as Neat1 in endoderm. Mir296 and Mir298 were expressed in both mesoderm and endoderm but were absent in ectoderm (Fig. 3G, SI Appendix, Fig. S18B, and Dataset S3).

Finally, to understand the relationship between mRNA and ncRNA genes, we performed a pairwise correlation analysis of gene expression across all sampled cells. This analysis showed that germ layer-specific miRNAs are correlated with gene sets associated with the proliferation of certain cell lineages (SI Appendix, Fig. S19A and Dataset S4). For example, miR-370 positively correlated with genes involved in the regulation of nervous system development and miR-351 correlated with genes associated with smooth muscle cell migration and osteoblasts differentiation (SI Appendix, Fig. S19). In addition, the expression of ~50% of identified histone-coding genes correlated with the expression of other protein-coding genes (Spearman’s r p > 0.5) (SI Appendix, Fig. S19). Overall, we found that multiple ncRNAs from all assayed RNA types (e.g., miRNA, snoRNA, snRNA, and so forth) are either positively or negatively correlated with the expression of protein-coding genes (SI Appendix, Fig. S19). Most of these ncRNAs represent putative uncharacterized regulators of lineage commitment.

Discussion

Altogether, Smart-seq-total enables an unbiased exploration of a broad spectrum of coding RNA and ncRNA transcripts in individual cells. Current limitations of Smart-seq-total are: 1) the inability to assay circRNA and 2) the loss of the endogenous polyadenylation status of transcripts. Further modifications to Smart-seq-total can include the selection for a specific transcript length (short vs. long) and depletion of a wider range of overrepresented RNA. We anticipate that Smart-seq-total will facilitate the identification of noncoding regulatory patterns and their functional roles in regulating cellular functions and shaping cellular identity. This could also shift the current protein-centered view of gene regulation toward comprehensive maps featuring both protein and RNA regulators.

Materials and Methods

Cell Culture. HEK293T cells were cultured in complete DMEM high-glucose medium (ThermoFisher, 11600-064) supplemented with 5% FBS (ThermoFisher, 16000-044), 1 mM sodium pyruvate (ThermoFisher, 11360070), and 100 μg/mL penicillin/streptomycin (ThermoFisher, 15070063). Human primary dermal fibroblasts were obtained from ATCC (PCS-201-012). Cells were cultured and passaged four times in fibroblast basal medium (ATCC, PCS-201-030) supplemented with 5 ng/mL human recombinant FGF (rhFGF) β, 7.5 mM—glutamine, 50 μg/mL ascorbic acid, 5 μg/mL human r recombinant insulin, and 1% FBS (Fibroblast Growth Kit low serum, ATCC PCS-201-041). MCF7 cells (ATCC, HTB22) were cultured in complete DMEM high-glucose medium (ThermoFisher, 11605092) supplemented with 10% FBS (ThermoFisher, 16000044), 1 mM sodium pyruvate (ThermoFisher, 11360070), and 100 μg/mL penicillin/streptomycin (ThermoFisher, 15070063). Cells were collected 2 to 4 h after passaging, dissociated using 0.25% trypsin-EDTA (ThermoFisher, 25000065) for 2 to 4 min at 37 °C, and sorted in either 96-well plates containing 3 μL lysis buffer or 384-well plates containing 0.3 μL of lysis buffer in each well. mESCs were maintained and differentiated as described previously (40, 48).

Briefly, mESCs were grown in serum-free 2i-LIF medium (complete medium: DMEM/F12 (Gibco, 10836010), 1% N2 supplement (Gibco, 16800050), and 110 μM 2-mercaptoethanol (Fermentas); supplemented with MEK inhibitor PD0325901 (0.8 μM, GS101) and IGF1 receptor inhibitor CHIR99021 (33 μM), and 10 ng/mL mouse LIF (Gibco, P0454548) in tissue culture (TC) dishes pretreated with 7.5 μg/mL polya-ornithine (Sigma) and 5 μg/mL laminin (BD). To induce spontaneous EB formation, cells were washed with PBS, dissociated with StemPro Accutase (ThermoFisher, A1110501) following the manufacturer’s protocol, transferred to serum-rich medium (complete medium: DMEM/F12 (Gibco, 10560018), 1% N2 supplement (Gibco, 11004050), and 110 μM 2-mercaptoethanol (Fermentas); supplemented with 10% FBS (ThermoFisher, 10439001), and diluted to 106 cells/mL. Each 10 μL of cell suspension was plated as a hanging drop in 10 mm² TC dishes (15 to 20 drops per dish). Ten microliters of fresh serum-rich media was added to each drop on the day 4 postseeding. Primed mESCs were collected 6 h after seeding. EBs were collected and dissociated at days 4, 8, and 12 of culture.

Cell Sort. Lysis plates were prepared by dispensing 0.3 μL lysis buffer (4 U recombinant RNase inhibitor (RRI; Takara, 23138), 0.12% TritonX-100 (Sigma, 93441-100ML) in dH2O, 1 μM Smart-seq-total oligo-dT primer (5’-Biotin-5’-BioGCACTATCGTGCGGTGAAGATGTAGTAAAGACAGGATG3’; IDT) in TE buffer (IDT [10 mM Tris, 0.1 mM EDTA, IDT] (see SI Appendix, Fig. S2B) for a full list of oligos used in the present study) into 384-well hard-shell PCR plates (Bio-Rad HS93901) using Mantis liquid handler (Formulatrix). For the comparison with Smart-seq2 (see Comparison of Smart-seq2 and Smart-seq-total, below), 96-well lysis plates were prepared with 3 μL lysis buffer. All plates were sealed with Alumaseal CS Films (Sigma-Aldrich, ZT22634), spun down, and snap-frozen on dry ice.

Cells were stained with calcein-AM and ethidium homodimer-1 (LIVE/DEAD Viability/Cytotoxicity Kit, ThermoFisher, L32240) following the manufacturer’s protocol and individual live cells were sorted in 384-well lysis plates using Sony sorters (SH800S) with a 100-μm nozzle chip. Plates were spun down and stored at ~80 °C immediately after sorting.

Generation of Smart-seq-total v1 Libraries. To facilitate cell lysate and denaturation of the RNA, 384-well plates were incubated at 72 °C for 3 min, and immediately placed on ice afterward. Next, 0.2 μL of polya tailing mix, containing 1.25 μL E. coli Polya (New England Biolabs, M0276S), 1.25 μL Polya buffer (New England Biolabs), 1.25 mM ATPs (New England Biolabs) and 4 U of RII (Takara) were added to each sample. Polya tailing was carried out for 15 min at 37 °C followed by 72 °C for 30 sec. After polya tailing were completed immediately on ice for 2 to 5 min, 1 μL of reverse-transcription mix, containing 15 U SuperScript (ThermoFisher), 4 U RII (Takara), 1.5x First-Strand Buffer, 1.5 μM TSO (Exiqon, 5’-Biotin-5’-UGUCGAGGCACAGGUGUAGUCGACGCAGACAUAGGCU-G3’), 7.5 mM DTT, 1.5 M Betaine (Sigma, 80300-SVL), 10 mM MgCl2 (Sigma, M1028-10X1ML), and 1.5 mM dNTPs (ThermoFisher, 18427013) was added to each well. Reverse transcription was carried out at 42 °C for

6 of 9 | PNAS
https://doi.org/10.1073/pnas.2113568118
Isakov et al.
Single-cell quantification of a broad RNA spectrum reveals unique noncoding patterns associated with cell types and states
generation of TSO (depletion of abundant sequences by hybridization) and tRNA, all reads mapping to arms or the stem loop were summed to quantify the expression at the gene level.

Comparison of Smart-seq2 and Smart-seq-total. HEK293T cells were sorted into 96-well plates containing 3 μL of lysis buffer (as described above). The reaction volumes for Smart-seq-total were scaled up 10 times compared to 384-well plate format (i.e., RNA from each cell was polyadenylated in 5 μL, reverse-transcribed in 15 μL, and cDNA was preamplified in 15 μL total volume). We retrieved Smart-seq2 data from Picelli et al. (19) (GSE49521). Smart-seq2 and Smart-seq-total reads were mapped using STAR and counted using featureCounts, as described above. Comparison between protocols in 2D (open ended) Fig. S1C was generated on depth-normalized libraries, i.e., the 5 million randomly selected reads per adaptor-trimmed library (or all reads for libraries that had less than 2.5 million reads) to compute expression levels (in million, cpm).

Unsupervised Clustering and Dimensionality Reduction Analysis of Human Cell Types. Standard procedures for filtering, variable gene selection, dimensional-ity reduction, and clustering were performed using the Seurat package v3.1.4 (54). Cells with fewer than 2,000 detected genes and those with more than twice the Mio reads were excluded from the analysis. Counts were log-normalized for each cell using the natural logarithm of 1 + cpm. Variable genes were selected based on overdispersion analysis and projected onto a low-dimensional subspace using PCA. The number of PCs was selected on the basis of inspection of the plot of variance explained. Cells were visualized using a two-dimensional t-SNE of the PC-projected data and identified using t-SNE (resolution and number of PCs) were adjusted on a per cell type and per biological sample basis. Clustering was performed using the Seurat package v3.1.4 (30).

Clustering of Coding and Noncoding Genes. Clusters of coding and noncoding genes shown in Fig. 2B were computed and visualized using the DEGreportR package (55). The top 250 marker genes for each cell cycle phase and all non-coding genes with average expression ln(pcm + 1) > 0.05 in at least one phase were used for this analysis. Gene-expression values were normalized using variance stabilizing transformation (56) before clustering. Further details of the analysis can be viewed in the Rmd files available on GitHub. Cells were assigned a cell cycle score using Seurat's CellCycleScoring() function using cell cycle markers described in Tirioh et al. (30).
Correlation between Mature miRNAs and mRNA. Total RNA fraction and a corresponding fraction of total RNA were sequenced separately on NextSeq 5000 sequencer. Sequencing data from both libraries was processed as described above (also https://github.com/aisakova/smarter-seq-total/blob/master/README.md). Small RNA and total RNA count tables were normalized and log-transformed separately. Pearson coefficients were computed for all pair-wise correlations between miRNAs and all other genes detected in ~300 HEK293T cells (present in at least 30 cells, InFlank > 1). Predicted and validated targets were annotated using multimiR pipeline (scoring predictions from various databases, such as TargetScan, mirTarBase, miRanda, and so forth) (57). Predicted targets were limited to those listed in two or more external databases. Validated targets with “weak” evidence were excluded from the analysis (labs(r)>0.5, adjusted P value <0.01) (SI Appendix, Fig. S15A). GO was performed on r ~100 to 200 top correlated (r > 0.3, adjusted P < 0.01) or ~100 anticorrelated (r < −0.3, adjusted P < 0.01) genes. The 10 GO terms with lowest P value in Kolmogorov–Smirnov test were used (SI Appendix, Fig. S15B).

Preprocessing and Clustering of mESCs. Standard procedures for filtering, variable gene selection, dimensionality reduction, and clustering were performed using the Seurat package v3.1.4 (54). Cells with fewer than 1,000 detected genes and those with more than two Mio reads were excluded from the analysis. Counts were log-normalized for each cell using log(p_counts) and 1e4 scale factor. Variable genes were projected onto a low-dimensional subspace using PC analysis. The number of PCs was selected counts) and 1e4 scale factor. Variable genes were projected onto a low-dimensional subspace using PC analysis. The number of PCs was selected.

Data Availability. The data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, https://www.ncbi.nlm.nih.gov/geo (accession no. GSE151334) (61). All code used for analysis is available on GitHub (https://github.com/aisakova/smarter-seq-total).

ACKNOWLEDGMENTS. We thank Giana Cirolla from Chan Zuckerberg Biohub for insightful advice on deletion of abundant sequences by hybridization protocol, as well as for providing rRNA guides; Jennifer Okamoto and Rene Sit for the assistance in sequencing of the libraries; Prof. Joanna Wysocka and her team members for guidance in embryoid body differentiation procedure; and Gita Mahmoudabadi for her constructive comments on the manuscript. This study was supported by Chan Zuckerberg Biohub. A.I. was supported by the Swiss National Foundation PostDoc Mobility Fellowship.

## References

1. E. Mereu et al., Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. Nat. Biotechnol. 38, 747–755 (2020).
2. Tabula Muris Consortium; Overall coordination; Logistical coordination; Organ coordination; Library preparation and sequencing; Computational data analysis; Cell type annotation; Writing group; Supplemental text writing group; Principal investigators; Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature 562, 367–372 (2018).
3. S. Aheger et al., Human Cell Atlas Meeting Participants, The human cell atlas. elife 6, e27041 (2017).
4. L. Kester, A. van Oudenaarden, Single-cell transcriptomics meets lineage tracing. Cell Stem Cell 23, 166–178 (2019).
5. C. Melton, R. L. Judson, R. Blelloch, Opposing microRNA families regulate self-renewal in mouse embryonic stem cells. Nature 463, 621–626 (2010).
6. R. A. Flynn, H. Y. Chang, Long noncoding RNA in cell-fate programming and reprogramming. Cell Stem Cell 14, 752–761 (2014).
7. A. Iasakova, T. Fehmahn, A. Keller, S. R. Quake, A mouse atlas of small non-coding RNA. bioRxiv [Preprint] (2018). https://doi.org/10.1101/430561 (Accessed 20 May 2020).
8. N. Gill, I. Ulltiniy, Regulation of gene expression by cis-acting long-non-coding RNAs. Nat. Rev. Genet. 21, 102–117 (2020).
9. E. Pennisi, Genomics. ENCODE project writes eulogy for junk DNA. Science 337, 1159–1161 (2012).
10. N. Wang et al., Single-cell microRNA-mRNA co-sequencing reveals non-genetic heterogeneity and mechanisms of microRNA regulation. Nat. Commun. 10, 95 (2019).
11. O. R. Fariadi et al., Single-cell sequencing of the small RNA transcriptome. Nat. Biotechnol. 34, 1264–1266 (2016).
12. Z. Xiao et al., Holo-seq: Single-cell sequencing of holo-transcriptome. Genome Biol. 19, 163 (2018).
13. K. Sheng, W. Cao, Y. Niu, Q. Deng, C. Zong, Effective detection of variation in single-cell transcriptomes using MATQ-seq. Nat. Methods 14, 267–270 (2017).
14. X. Fan et al., Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. Genome Biol. 16, 148 (2015).
15. T. Hashay et al., Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. Nat. Commun. 9, 619 (2018).
16. Y. Wang et al., Endogenous miRNA sponge lincRNA-RoR regulates Oct4, Nanog, and Sox2 in human embryonic stem cell self-renewal. Dev. Cell 25, 69–80 (2013).
17. S. J. Liu et al., Single-cell analysis of long non-coding RNAs in the developing human neocortex. Genome Biol. 17, 67 (2016).
18. K. Verboom et al., SMARTer single cell total RNA sequencing. Nucleic Acids Res. 47, e93 (2019).
19. S. Picelli et al., Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat. Methods 10, 1906–1909 (2013).
20. Y. Y. Zhu, E. M. Machleder, A. Chenkich, R. L. P. Siebert, Reverse transcriptase plate switching: A SMART approach for full-length cDNA library construction. Biotechniques 30, 892–897 (2001).
41. M. Guttman et al., lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* **477**, 295–300 (2011).

42. S. Lee et al., Human long noncoding RNA regulation of stem cell potency and differentiation. *Stem Cells Int.* **2017**, 6374504 (2017).

43. G. S. Gulati et al., Single-cell transcriptional diversity is a hallmark of developmental potential. *Science* **367**, 405–411 (2020).

44. M. Pekkanen-Marttila et al., Spatial and temporal expression pattern of germ layer markers during human embryonic stem cell differentiation in embryoid bodies. *Histoch. Cell Biol.* **133**, 595–606 (2010).

45. S. C. van den Brink et al., Single-cell and spatial transcriptomics reveal somitogenesis in gastruloids. *Nature* **582**, 405–409 (2020).

46. X. Han et al., Mapping human pluripotent stem cell differentiation pathways using high throughput single-cell RNA-sequencing. *Genome Biol.* **19**, 47 (2018).

47. F. A. Wolf et al., PAGA: Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 59 (2019).

48. A. Rada-Iglesias et al., A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279–283 (2011).

49. Tabula Muris Consortium, A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature* **583**, 590–595 (2020).

50. A. Dobin et al., STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

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

52. A. Frankish et al., GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**(D1), D766–D773 (2019).

53. A. Kozomara, S. Griffiths-Jones, miRBase: Annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* **42**, D68–D73 (2014).

54. T. Stuart et al., Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902.e21 (2019).

55. L. Pantano, DEGreport: Report of DEG analysis, R package version 1.20.0 (2019).

56. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014). https://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8.

57. Y. Ru et al., The multMIIR R package and database: Integration of microRNA-target interactions along with their disease and drug associations. *Nucleic Acids Res.* **42**, e133 (2014).

58. C. McInnes, J. Healy, J. Melville, UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv* [Preprint] (2018). https://arxiv.org/abs/1802.03426 (Accessed 27 April 2020).

59. W. Saelens, R. Cannoodt, H. Todorov, Y. Saeys, A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).

60. W. F. Marzluff, E. J. Wagner, R. J. Duronio, Metabolism and regulation of canonical histone mRNAs: Life without a poly(A) tail. *Nat. Rev. Genet.* **9**, 843–854 (2008).

61. A. Isakova, Single cell profiling of total RNA using Smart-seq-total. Gene Expression Omnibus (GEO). https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE151334. Deposited 27 May 2020.