Transcriptionally Profiling Cells Sorted by Transcript Abundance

Sandy Klemm[#,1,9], **Stefan Semrau**[#2], **Kay Wiebrands**[#3,4], **Dylan Mooijman**[3,4], **Dina A. Faddah**[5,6], **Rudolf Jaenisch**[5,6], and **Alexander van Oudenaarden**[2,3,4,5,7,9]

[1]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States. [2]Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States. [3]Hubrecht Institute – Royal Netherlands Academy of Arts and Sciences, Utrecht, The Netherlands. [4]University Medical Center Utrecht, Utrecht, The Netherlands. [5]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States. [6]Whitehead Institute for Biomedical Research, Cambridge, Massachusetts, United States. [7]Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States.

[#] These authors contributed equally to this work.

Abstract

We have developed a quantitative technique for sorting cells based on endogenous RNA abundance with a molecular resolution of 10-20 transcripts. We demonstrate efficient and unbiased RNA extraction from transcriptionally sorted cells and report a high fidelity transcriptome measurement of (mouse) induced pluripotent stem cells (iPSCs) isolated from a heterogeneous reprogramming culture. This method is broadly applicable to profiling transcriptionally distinct cellular states without requiring antibodies or transgenic fluorescent proteins.

A common challenge in biology is to identify and isolate transcriptionally distinct subpopulations within a single tissue or cell type. While a variety of techniques have been developed to discriminate among these alternative expression modes, the most widely used methods require either transgenic integration of fluorescent protein reporters or the availability of specific antibodies[1–3]. These approaches, however, are precluded for biological systems that are refractory to genetic manipulation (e.g., primary human tissue)

and for processes in which RNA – rather than protein – is the key discriminative marker (e.g., non-coding RNA). Recently, flow cytometry has been used to sort cells using a spectrum of fluorescent labeling techniques in whicholigonucleotide probes are hybridized to either DNA or RNA target sequences[4–7]. The principle limitation of these methods has been that RNA extracted from hybridized material is often highly degraded[8,9]. Although fragmented RNA can be reverse transcribed and analyzed by quantitative PCR (RT-qPCR)[10,11], full-length RNA is required for an unbiased representation of the transcriptome. While RNA degradation is partially mitigated by labeling RNA in live cells[7,12], extended hybridization in ex-vivo culture may obscure the molecular state of primary tissue. Given these limitations, we have developed a method for RNA labeling in cross-linked cells which permits full-length RNA isolation and unbiased transcriptional profiling.

The proposed RNA cell sorting technique uses flow cytometry to measure the fluorescence of individual cells labeled with a single molecule RNA FISH (smFISH) probe library[13,14]. As a proof-of-principle experiment, green fluorescent protein (GFP) transcripts were fluorescently labeled in cells which expressed the transgene under doxycycline control[15] (Supplementary Fig. 1a-c). To assess the sorting potential of the labeled RNA signal, single cell RNA fluorescence distributions were measured by flow cytometry, revealing a clear separation of high and low induction profiles (Fig. 1a and Supplementary Fig. 1b,c). Furthermore, the measured mRNA fluorescence was found to scale linearly with mRNA and protein abundance across a broad range of induction levels (Fig. 1b). We further confirmed the linearity of the labeled RNA fluorescence signal for a panel of endogenous genes by comparing the mean flow cytometry signal intensity with the average number of RNA molecules in iPSCs quantified by single cell transcript counting[14] (Fig. 1c).

We then asked whether the observed RNA fluorescence signal provides an accurate measurement of single cell transcript levels. For this purpose, we measured both GFP protein and labeled mRNA fluorescence in single cells and found a strong correlation ($\rho = 0.77$, Supplementary Fig. 1d), which is consistent with the expectation that mRNA and protein are broadly correlated in single cells. Additionally this correlation between mRNA and protein was confirmed across a broad range of GFP induction levels (Supplementary Fig. 1e). Next, we tested the single cell precision of the proposed RNA measurement by differentially labeling the 3′ and 5′ ends of a single transcriptional target (*Oct4-IRES-GFP* fusion mRNA) and found that the labels were strongly correlated at the single cell level ($\rho = 0.90$, Supplementary Fig. 2a). We then measured a mixture of cells labeled with differentially coupled probes for the gene Oct4 and confirmed that the two labels were strongly anti-correlated ($\rho = -0.81$, Supplementary Fig. 2b). Having established measurements of both positive and negative correlation, we hypothesized that doxycycline induced GFP expression would in principle be uncorrelated with every endogenous mRNA species in the genome. We confirmed this prediction across all measured GFP induction levels for the gene Oct4 ($\rho = 0.1$, Supplementary Fig. 2c). Finally, we sorted cellsabove and below each quartile of the Sox2 RNA fluorescence distribution. Single cell transcript counting in each fraction quantitatively recapitulated the flow cytometry quartile measurements (Fig. 1d). Taken together, these data suggest that flow cytometry can be used

to quantitatively measure the abundance of mRNA in single cells hybridized with a complementary smFISH library.

To assess the resolution of the labeled RNA signal, we measured the expected number of transcripts required for a cell to be statistically resolved from the background (see **Online Methods**). An estimated resolution of 57 transcripts was measured for a library of 20 nucleotide probes hybridized to the Sox2 gene in derived iPSCs (Supplementary Fig. 4a). Under more stringent hybridization conditions, a 30 nucleotide library improved the resolution to 24 transcripts (Supplementary Fig. 4b). This estimate was confirmed by sorting cells and directly measuring the difference in the number of transcripts required for a pair of cells to be resolved with 95% confidence (Supplementary Fig. 4c,d). The sorting resolution was further improved by measuring RNA abundance and sorting in specific cell cycle phases (Supplementary Fig. 5a-b). For G1 cells, the molecular resolution was estimated as 30 molecules for the 20 nucleotide probe library and 13 molecules for the 30 nucleotide library (Fig. 1d and Supplementary Fig. 5c,d).

In order to measure the transcriptome of sorted subpopulations, RNA is extracted from hybridized cells by cross-link reversal (see **Online Methods**). Under standard hybridization conditions[14], the molecular integrity of the extracted RNA is attenuated in a time and temperature dependent manner, due to enzymatic RNA degradation as well as hydrolysis mediated RNA fragmentation (Supplementary Fig. 6a). We have addressed this by developing an RNA preserving hybridization buffer (RPHB) which facilitates efficient isolation of full-length RNA following hybridization (Supplementary Fig. 6b). RPHB employs a nearly saturating salt concentration to eliminate enzymatic RNA degradation by precipitating proteins, and includes a high concentration of the chelating agent EDTA, which inhibits RNA degradation by sequestration of metal ions involved in RNA hydrolysis. We tested RPHB by extracting RNA from live and RPHB hybridized mouse embryonic stem cells (mESCs) and comparing relative expression levels by RT-qPCR and microarrays (Fig. 2a,b), establishing that RNA extracted from RPHB hybridized material is quantitatively unbiased. Furthermore, we found that the error distribution between technical microarray replicates is identical for live and hybridized samples (Supplementary Fig. 6c). We then examined the genome-wide expression fold change between NIH-3T3 fibroblasts and J1 mES cells. A comparison of fold change measurements for both live and hybridized samples revealed a strong correlation ($\rho = 0.94$) over the full dynamic range of the microarray (Fig. 2c). Finally, we isolated mES cells and fibroblasts by RNA FACS from an artificial mixture of these cells types; subpopulation transcriptome measurements on the sorted fractions recapitulated therespective cell type signatures (Supplementary Fig. 6d and Supplementary Fig. 7a,b).

One of the motivating applications for RNA based sorting has been to transcriptionally profile induced pluripotent stem cells during the process of cellular reprogramming[16]. Following disruption of the somatic state during reprogramming, individual cells stochastically reactivate the pluripotency machinery at widely different rates[17] and contribute to a diverse collection of co-existing subpopulations[18]. While iPS cells define a subpopulation which is independent of ectopic reprogramming factors, incompletely reprogrammed cells are dependent on sustained reprogramming factor expression in order to

be competitively maintained in culture. To interrogate these subpopulations, we isolated cells which had reactivated the endogenous *Sox2* locus – a definitive marker for reprogramming[18,19] – from a background of partially reprogrammed cells. The *Sox2*[+] and *Sox2*[−] transcriptomes were then compared with iPS cells derived by reprogramming factor withdrawal.

Secondary mouse embryonic fibroblasts (2° MEFs, see **Online Methods**) were reprogrammed by exposure to doxycycline induced expression of Oct4, Sox2, Klf4, and c-Myc (OSKM) for 32 consecutive days (Fig. 3a), at which point iPS cells were passaged in co-culture with MEFs (iPSMEF) and in feeder-free 2i media (iPS-2i) as OSKM independent colonies (Fig. 3a). The distribution of endogenous *Sox2* expression – measured by flow cytometry using a smFISH probe library designed against the non-coding 3′UTR of *Sox2* – was bimodal (Fig. 3a) for day 32 OSKM[+] cells, reflecting an underlying diversity of reprogramming depth among individual cells. The upper and lower *Sox2* expression tails were sorted and transcriptionally profiled in comparison with OSKM independent iPSCs, revealing an unambiguous pluripotency signature for the positive fraction (Fig. 3b). While many somatic marks were repressed in both *Sox2*[+] and *Sox2*[−] cells, a broad spectrum of iPSC specific genes were differentially upregulated in *Sox2*[+] cells, including transcription factors, RNA binding proteins, chromatin regulators, and cell surface markers. Reciprocally, *Sox2*[−] cells expressed a class of differentiation associated genes which were repressed in iPS cells. The striking similarity between *Sox2*[+] and iPS cells suggests that cells within the *Sox2*[+] subpopulation are reprogrammed and give rise to stable iPS cells under OSKM withdrawal. By leveraging a non-coding transcriptional element (*Sox2* 3′U TR), these experiments illustrate the flexibility of RNA FACS and suggest the potential for a broader subpopulation analysis of cellular reprogramming.

Single cell heterogeneity in gene expression is a common phenomenon for a variety of developmental and homeostatic processes. The principle focus of this work has been to develop a fluorescent measure of RNA in single cells, which facilitates high resolution sorting as well asefficient and unbiased RNA isolation. This technique extends flow cytometry to a new class of applications based on direct quantification of RNA.

ONLINE METHODS

Molecular resolution

The molecular resolution is defined as the expected number of mRNA molecules required for a cell to be statistically resolved from the background. The resolution depends on the measured signal, the single cell distribution of mRNA, and the background fluorescence. The transcript distribution was measured by counting smFISH labeled mRNA in single cells as previously described[14]. The background fluorescence – which accounts for both cellular autofluorescence and non-specific probe binding – was estimated by hybridizing a mock sample with a 1:20 mixture of fluorescently labeled and unlabeled probe libraries (yielding a 95% attenuation of the specific signal). The resolution R_p was calculated for the significance level p using the mock signal as a null model (see Supplementary Note).

Fluorescence in-situ hybridization buffers and probes

Oligonucleotide libraries with 20 nucleotide probes were designed and fluorescently labeled as previously described[14]. Probes for the 30 nucleotide *Sox2* library were similarly designed such that the predicted melting temperature of individual probes deviates from the median by no more than $5°_C$. See Supplementary Data for all probe libraries used in this study. The following buffers were used in this study: (RPH buffer) 300 mM Sodium chloride, 30 mM Sodium citrate, 2.1 M Ammonium sulfate, 10 mM EDTA, 1 mg/ml E. Coli tRNA, 500 µg/ml BSA, 25% (40%) formamide for 20 (30) nucleotide probe library; (Wash buffer) 25% (40%) formamide for 20 (30) nucleotide probe library, 2x SSC; (Flow buffer) 2x SSC; (Sorting buffer) 50% RNALater Solution (Ambion), 2x SSC; (Reverse cross-linking buffer) 100 mM NaCl, 10 mM pH 8.0 Tris, 1 mM EDTA, 0.5% SDS, 500 µg/ml Proteinase K. The dependence of RNA degradation on hybridization conditions (Supplementary Fig. 4a) was measured using the previously reported (classical) smFISH hybridization buffer[14].

RNA fluorescence in-situ hybridization

Cells were fixed in 4% paraformaldehyde for 5 minutes, centrifuged at 1000 RCF for 5 minutes, and washed in 70% ethanol (EtOH). Following overnight ethanol permeabilization (70% EtOH) at $4°_C$, cells were resuspended in RPH buffer with labeled probes (0.5-1 ng/µl) and incubated at $30°_C$ for 12 hours. An equal volume of wash buffer was then added and mixed thoroughly with the sample. Cells were then pelleted by centrifugation and resuspended in wash buffer for 30 minutes at $30°_C$. After repeating the previous wash step, cells were resuspended in flow buffer and maintained at $4°_C$ in preparation for sorting.

Flow cytometry, FACS and RNA extraction

Cells were sorted by FACS into $4°_C$ sorting buffer using a BD Biosciences Aria II flow cytometer. Unless otherwise noted, cell size was controlled for by co-staining cells with Hoechst 33342 (Invitrogen) and selecting for diploid DNA content. FACS sorted cells were centrifuged as before and resuspended in reverse cross-linking buffer at $50°_C$ for 1 hour. Total RNA was isolated by phenol-chloroform (Trizol, Invitrogen) using the manufacturerfs protocol.

Microarrays

Microarray assays were performed by the Massachusetts Institute of Technology (MIT) BioMicro Center. NIH-3T3 fibroblasts and mESCs were assayed using Eukaryotic Exon 1.0 ST arrays from Agilent, while Mouse 430A 2.0 Affymetrix chips were used for MEF and iPSCs.

Cell lines and media

Embryonic and induced pluripotent stem cells were grown as indicated: (1) co-cultured with irradiated MEF cells (Global Stem) or (2) cultured in 2i conditions with both glycogen synthase kinase 3β (Stemgent) and mitogen-activated protein kinase kinase inhibitors (Stemgent). ES and iPS cells were grown with leukemia inhibitory factor (10^3 units/ml, Millipore) and 15% (10% for E14 cells) heat-inactivated fetal bovine serum (Hyclone) together with Knockout DMEM (Gibco), L-glutamine (Gibco), MEM non-essential amino

acids (Gibco), penicillin-streptomycin (Gibco) and β-mercaptoethanol (Sigma). NIH-3T3 fibroblast and primary MEF cells were cultured without inhibitors in 10% serum. E14 mES cells were used for the post-hybridization RNA extraction controls and J1 mES cells (2i media) were used for the NIH-3T3 & mESC sorting experiments. KH2:GFP cells[15] – with constitutive R26-M2rtTA expression and a tetracycline-inducible eGFP construct targeted to the ColA1 locus – were used for the titrated GFP induction experiments.

Secondary somatic cell generation and reprogramming

Secondary mouse embryonic fibroblasts (2° MEFs) were isolated from chimeric embryos as previously described[20], providing doxycycline inducible Oct4, Sox2, Klf4, and c-Myc expression; Neomycin selection at the Nanog locus; and a fluorescent transcriptional reporter for Oct4. The parental pluripotent stem cells used to generate the secondary line were derived by replacing an Oct4 allele with an Oct4-IRES-GFP sequence in Nanog-Neo iPS cells[20]. Secondary MEFs were plated at optimal density[20] and passaged after 48 hours. Doxycycline (2μg/ml, Stemgent) was added 24 hours after replating, marking the start of reprogramming. Neomycin selection (Stemgent G418, 1μg/ml) was applied beginning at day 6 and maintained throughout the reprogramming timecourse. Reprogrammed iPS cells were stabilized in co-culture with MEFs (iPSMEF) and in feeder-free conditions (iPS-2i).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

REFERENCES

1. Chalfie M, Tu Y, Euskirchen G, Ward WW, Prasher DC. Science. 1994; 263:802–805. [PubMed: 8303295]

2. Chambers I, et al. Nature. 2007; 450:3–8.

3. Dietrich J-E, Hiiragi T. Development. 2007; 134:4219–31. [PubMed: 17978007]

4. Rufer N, Dragowska, Wieslawa Thorbury G, Roosnek E, Lansdorp P. Nat. Biotechnol. 1998; 16:743–747. [PubMed: 9702772]

5. Prigodich AE, et al. Anal. Chem. 2012; 84:2062–2066. [PubMed: 22288418]

6. Robertson KL, Verhoeven AB, Thach DC, Chang EL. RNA. 2010; 16:1679–1685. [PubMed: 20584898]

7. Rhee WJ, Bao G. BMC Biotechnol. 2009; 9:30. [PubMed: 19341452]

8. Masuda N, Ohnishi T, Kawamoto S, Monden M, Okubo K. Nucleic Acids Res. 1999; 27:4436–43. [PubMed: 10536153]

9. Specht K, et al. Am. J. Pathol. 2001; 158:419–29. [PubMed: 11159180]

10. Yamada H, et al. Cytometry A. 2010; 77:1032–7. [PubMed: 20872886]

11. Maruo R, et al. Mol. Biotechnol. 2011; 49:42–7. [PubMed: 21246309]

12. Larsson HM, et al. PLoS One. 2012; 7:e49874. [PubMed: 23209609]

13. Femino AM, Fay FS, Fogarty K, Singer RH. Science. 1998; 280:585–90. [PubMed: 9554849]

14. Raj A, Bogaard PVD, Rifkin SA, van Oudenaarden A, Tyagi S. Nat. Methods. 2008; 5:877–879. [PubMed: 18806792]

15. Beard C, Hochedlinger K, Plath K, Wutz A, Jaenisch R. Genesis. 2006; 44:23–8. [PubMed: 16400644]

16. Takahashi K, Yamanaka S. Cell. 2006; 126:663–76. [PubMed: 16904174]

17. Hanna J, et al. Nature. 2009; 462:595–601. [PubMed: 19898493]

18. Buganim Y, et al. Cell. 2012; 150:1209–22. [PubMed: 22980981]

19. Golipour A, et al. Cell Stem Cell. 2012; 11:769–82. [PubMed: 23217423]

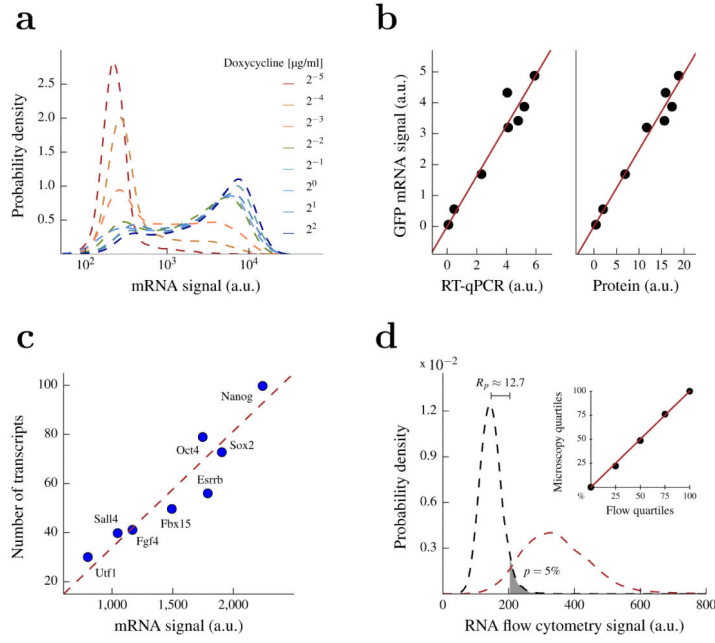20. Wernig M, et al. Nat. Biotechnol. 2008; 26:916–24. [PubMed: 18594521]

Figure 1. A quantitative single cell measurement of transcription

(a) Single cell distribution of GFP mRNA fluorescence for indicated doxycycline induction levels. **(b)** Linear and unbiased scaling of GFP mRNA fluorescence with RT-qPCR and GFP protein fluorescence. **(c)** Linear scaling of mean mRNAfluorescence with single cell transcript quantification by classical (microscopy) smFISH for a panel endogenous genes in iPS-2i cells (see **Online Methods**). **(d)** *Sox2* mRNA signal (red) and background (black) with shaded 95% quantile (gray) and molecular resolution R_p reported for the $p = 5\%$ significance level (see **Online Methods**). Recapitulation of first (25%), second (50%), and third (75%) quartile sorting by classical smFISH transcript counting (inset).
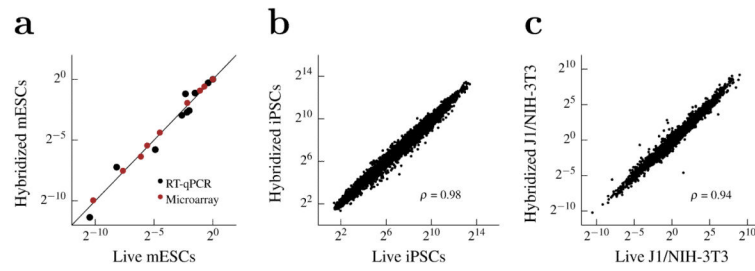
Figure 2. RNA extraction and transcriptional sorting
(a) Hybridized and live mES cell expression (normalized to *Gapdh*) for a panel of genetic markers measure by RT-qPCR (black) and microarray (red). **(b)** Genome-wide microarray expression measurements for live and hybridized cells. **(c)** NIH-3T3 and J1 mESC expression fold change measurements for hybridized and live samples.
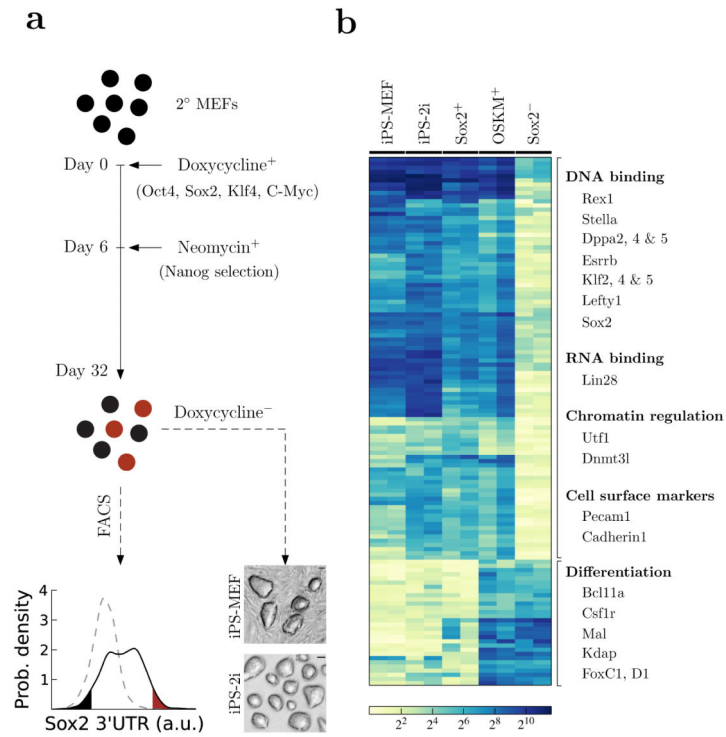
Figure 3. Isolation and transcriptional profiling of iPS cells
(a) Secondary MEF reprogramming; isolation of iPS cells by OSKM withdrawal in MEF co-culture (iPS-MEF) and 2i media (iPS-2i); RNA FACS for reprogrammed *Sox2*+ (red) and non-reprogrammed *Sox2*− (black) cells (25 µm scale bars). The bimodal *Sox2* 3′UTR signal (black) relative to the non-specific background signal (gray) reflects transcriptional and phenotypic heterogeneity in day 32 OSKM+ cells. **(b)** Hierarchical clustering of genes differentially expressed between *Sox2*+ and *Sox2*− cells (top 5% shown).