Supplementary Information

Single-nucleus transcriptome sequencing of differentiating human myoblasts reveals the extent of fate heterogeneity in non-myotube cells

Weihua Zeng[1,2], Shan Jiang[1,2], Xiangduo Kong[3], Nicole El-Ali[1,2], Alexander R. Ball, Jr.[3], Christopher I-Hsing Ma[3], Naohiro Hashimoto[4], Kyoko Yokomori[3*], and Ali Mortazavi[1,2*].

[1] Department of Developmental and Cell Biology, University of California Irvine, Irvine, CA 92697-2300, USA.

[2] Center for Complex Biological Systems, University of California Irvine, Irvine, CA 92697-2280, USA

[3] Department of Biological Chemistry, School of Medicine, University of California Irvine, Irvine, CA 92697-1700, USA

[4] Department of Regenerative Medicine, National Center for Geriatrics and Gerontology, 7-430 Morioka, Oobu, Aichi 474-8522, Japan

*To whom correspondence should be addressed: Tel: +1 949 824 6762; E-mail: ali.mortazavi@uci.edu. Correspondence can also be addressed to: Tel: +1 949 824 8215; E-mail: kyokomor@uci.edu

Supplementary Figures S1-S11

Supplementary Fig 1. Immunostaining of myogenic genes on undifferentiated and 72hr differentiated KD3 together with bright field images. The immunostaining results of corresponding factors are in green fluorescence and the DAPI counter staining results are in blue. The scale bar is 100μm.

Supplementary Fig 2. Schema for isolation and capture of myoblast, myotube and MNC single nuclei. (A) In the left column, undifferentiated KD3 myoblasts were detached with Trypsin/EDTA and the cell pellet was treated with IGEPAL CA-630 to release nuclei. For the middle and right columns, differentiation media was added to myoblast culture when the cells reached 80% confluence. Myotubes and MNCs were harvested after 3 days, when 80-90% of the cells were fused into myotubes with 10-20% cells remained unfused as MNCs. The middle and right columns show isolation of myotubes and MNCs, respectively. The top middle image was contrast-enhanced to visualize multinuclei in myotubes while the top right image was focused on the interstitial MNCs. For myotube isolation, MNCs were first removed by brief Trypsin/EDTA treatment, and nuclei were subsequently released from the remaining myotubes using IGEPAL CA-630. Note the dispersion of myotube nuclei after IGEPAL CA-630 cell lysis. In the right column, MNCs separated from myotubes by brief Trypsin/EDTA treatment (see above) in the supernatant were pelleted by centrifugation and treated with IGEPAL CA-630 to release nuclei. Isolated

nuclei from myoblasts (left), myotubes (middle) and MNCs (right) were then subjected to the Fluidigm small size chip as indicated at the bottom. We found that the myotube nuclei recovery efficiency after lysis is 50-60%, compared to 70-80% from myoblast or MNC. The scale bar is 50μm. Representative cell images for each step are shown. (B) Representative image of nucleus capture on small-size Fluidigm C1 chip.

Supplementary Fig 3. Basic Metrics for scRNA-seq and snRNA-seq samples. **(A)** Mapped read numbers of all the libraries from myoblast scRNA-seq and myoblast, myotube, MNC snRNA-seq before filtration. **(B)** The number of genes with TPM >= 1.0 in each sample.

Supplementary Fig 4. Two-way hierarchical clustering of undifferentiated KD3 myoblast with 11,004 genes expressed at >= 1 TPM in 10 or more samples from **(A)** all single-cell samples versus all single-nucleus samples before sample filtration; **(B)** 37 single-cell samples versus 37 single-nucleus samples after removing outlying samples with less than 4,000 genes with TPM>=1.0 and selecting the single-cell samples with highest *MYF5* expression .

Supplementary Fig 5. qPCR verification of scRNA-seq and snRNA-seq data. 10 harvested cDNA samples after filtration were selected randomly for each sample type. The geometric average of *GAPDH* and *UBC* level was used as internal control to normalize the target gene expression level. The normalized

quantification of target genes was depicted in logarithmic scale. Samples without detectable qPCR signal are plotted at the bottom of the graphs labeled as N/D (non-detection). The p-value of gene expression level difference between different sample types was calculated by student's t-test and the parameters were set to two-tailed and two-sample with unequal variance. (A) qPCR validation of *XIST* and *H19* level in myoblast single-cell and single-nucleus cDNA samples. *NEUROD1* serves as negative control. (B) qPCR validation of *MYOD1* and *ID3* levels in myoblast, myotube and MNC single-nucleus cDNA samples. *NEUROD1* serves as negative control. (C) qPCR validation of *MIR222HG* (host gene for *MIR221/222*) and *MIR503HG* levels in myoblast, myotube and MNC single-nucleus cDNA samples.

Supplementary Fig 6. Two-way hierarchical clustering of all KD3 single-nuclei from myoblast, myotube and MNC with **(A)** 11,004 genes expressed at >= 1 TPM in 10 or more samples; **(B)** 33 myogenic and cell cycle related genes. The sample clustering tree is the same as the analysis with 11,004 genes. Major clusters of nuclei are labeled 1-6 and discussed in the main text. The three myoblast nuclei outliers at the far left side, which have high detected numbers of genes (6,094, 8,066 and 9,675 genes from these three outliers versus 4,654 as the highest gene number from the other myoblast nuclei), are excluded from cluster 1. The color bar on top indicates the sample origin (green for myoblast, red for myotube and blue for MNC). The second row of color bar indicates which samples were chosen for differential gene analysis.

Supplementary Fig 7. Violin plots of the three genes (*hTERT*, *CDK4* and *CCND1*) to immortalize KD3 cell between KD3 and HSMMs from the monocle paper (6). We used the 63 selected single-nuclei from Fig2a, all the 0hr HSMM samples and the 72hr HSMM samples with positive *MYOG* (representative for myotubes) or *SPHK1* (putative candidates of mesenchymal stem cells) level for the comparison. We compare FPKMs instead of TPM because FPKM was used in the data matrix from the monocle paper. We added a 0.1 pseudocount for logarithm calculation.

Supplementary Fig 8. Violin plots of the MNC-related genes' expression level from the 63 single nuclei from Fig 2a. **(A)** The expression level of genes involved in different signaling pathways in myoblast, myotube and MNC snRNA-seq data. **(B)** The expression level of proinflammatory genes. The expression of other proinflammatory genes such as *CXCL*s, *IL6/8* is negative and is not shown in this figure.

Supplementary Fig 9. Violin plots of myogenic gene expression levels between KD3 and HSMMs used in the monocle paper. We used the 63 selected single-nuclei data from Fig 2a, all the 0hr HSMM samples and the 72hr HSMM samples with positive *MYOG* (representative for myotubes) or *SPHK1* (putative candidates of mesenchymal stem cells) level for comparison. We compare

FPKMs instead of TPM because FPKM was used in the data matrix from the monocle paper. We added a 0.1 pseudocount for logarithm calculation.

Supplementary Fig 10. (A) lncRNAs highly expressed in only one nucleus type. Top 20 highly expressed lncRNAs are listed on the right for each category. (red, high expression; blue, low expression). (**B)** lncRNAs highly expressed in any two of three nucleus types. Top 20 highly expressed lncRNAs are listed on the right for each category. (red, high expression; blue, low expression).

Supplementary Fig 11. (A) SMART-seq coverage and NanoString expression of 490 miRNAs with Pearson's correlation coefficient between SMART-seq coverage and NanoString miRNA expression. Both coverage and NanoString expression are normalized between -1 and 1 (red, high level; blue, low level). (**B)** Lists of miRNAs with correlation coefficient (between SMART-seq coverage and NanoString miRNA expression, *Pearson's*) higher than 0.5 and their neighboring lncRNAs. lncRNA is considered as pri-miRNA when miRNA and lncRNA are on the same strand (blue dot in the left column) and their locus overlap (blue dot in the right column). *MIR133B* and *MIR206* are two exceptions as *LINCMD1* are known to be their pri-miRNA but are on the opposite strand and *MIR206* does not overlap with *LINCMD1* in human unlike in mouse, which might be caused by mis-annotation in human.

Supplemental Tables S1-S22

Table S1. Sequences of primers used in qPCR validation assays

Table S2. Myotube versus myoblast differential genes. Genes with FDR <= 0.001 and minimal FC 4-fold using edgeR are identified as differential genes and labeled as “up” if up-regulated in myotube or “down” if down-regulated in myotube. Genes that pass neither FDR nor FC threshold are considered as non-differential genes (“nodiff”) and other genes are labeled as non-significant (“nonsig”). This data is used to generate Figure 2C.

Table S3. Myotube versus MNC differential genes. Genes with FDR <= 0.001 and minimal FC 4-fold using edgeR are identified as differential genes and labeled as “up” if up-regulated in myotube or “down” if down-regulated in myotube. Genes that pass neither FDR nor FC threshold are considered as non-differential genes (“nodiff”) and other genes are labeled as non-significant (“nonsig”). This data is used to generate Figure 2C.

Table S4. MNC versus myoblast differential genes. Genes with FDR <= 0.001 and minimal FC 4-fold using edgeR are identified as differential genes and labeled as “up” if up-regulated in MNC or “down” if down-regulated in MNC. Genes that pass neither FDR nor FC threshold are considered as non-differential genes (“nodiff”) and other genes are labeled as non-significant (“nonsig”). This data is used to generate Figure 2C.

Table S5. GO enrichment of myotube versus myoblast differential genes. The list of differential genes is from Table S2. The analysis was performed with Metascape and a subset of representative GO terms with p-value lower than 0.05 are shown.

Table S6. GO enrichment of myotube versus MNC differential genes. The list of differential genes is from Table S3. The analysis was performed with Metascape and a subset of representative GO terms with p-value lower than 0.05 are shown.

Table S7. GO enrichment of MNC versus myoblast differential genes. The list of differential genes is from Table S4. The analysis was performed with Metascape and a subset of representative GO terms with p-value lower than 0.05 are shown.

Table S8. GO enrichment of protein-coding genes associated with differentially expressed lncRNAs in myoblast nucleus and cell comparison. The list of differential genes is from Table S12. The analysis was performed with Metascape and a subset of representative GO terms with p-value lower than 0.05 are shown.

Table S9. GO enrichment of protein-coding genes associated with differentially expressed lncRNAs in MNC and myoblast nucleus comparison. The list of differential genes is from Table S13. The analysis was performed with

Metascape and a subset of representative GO terms with p-value lower than 0.05 are shown.

Table S10. GO enrichment of protein-coding genes associated with differentially expressed lncRNAs in myotube and myoblast nucleus comparison. The list of differential genes is from Table S14. The analysis was performed with Metascape and a subset of representative GO terms with p-value lower than 0.05 are shown.

Table S11. GO enrichment of protein-coding genes associated with differentially expressed lncRNAs in myotube and MNC nucleus comparison. The list of differential genes is from Table S15. The analysis was performed with Metascape and a subset of representative GO terms with p-value lower than 0.05 are shown.

Table S12. Protein-coding genes associated with differentially expressed lncRNAs in the comparison between myoblast nucleus and cell. Protein-coding genes are associated with lncRNAs if the maximum distance between gene bodies <10kb and they share the same regulation direction (up or down) with differential lncRNAs.

Table S13. Protein-coding genes associated with differentially expressed lncRNAs in the comparison between MNC and myoblast nucleus. Protein-coding

genes are associated with lncRNAs if the maximum distance between gene bodies <10kb and they share the same regulation direction (up or down) with differential lncRNAs.

Table S14. Protein-coding genes associated with differentially expressed lncRNAs in the comparison between myotube and myoblast nucleus. Protein-coding genes are associated with lncRNAs if the maximum distance between gene bodies <10kb and they share the same regulation direction (up or down) with differential lncRNAs.

Table S15. Protein-coding genes associated with differentially expressed lncRNAs in the comparison between myotube and MNC nucleus. Protein-coding genes are associated with lncRNAs if the maximum distance between gene bodies <10kb and they share the same regulation direction (up or down) with differential lncRNAs.

Table S16. Differentially expressed lncRNAs and their neighboring miRNAs in the comparison between MNC and myoblast nucleus. LncRNAs serving as pri-miRNAs are labeled with “yes”.

Table S17. Differentially expressed lncRNAs and their neighboring miRNAs in the comparison between myotube and myoblast nucleus. LncRNAs serving as pri-miRNAs are labeled with “yes”.

Table S18. Differentially expressed lncRNAs and their neighboring miRNAs in the comparison between myotube and MNC nucleus. LncRNAs serving as pri-miRNAs are labeled with “yes”.

Table S19. LncRNAs differentially expressed in only one nucleus type, including myoblast (“myob”), MNC (“mono”), myotube (“myot”). This data is used to generate Figure S10A.

Table S20. LncRNAs differentially expressed in any two nuclei types, including MNC and myoblast (“mono_myob”) or myotube and myoblast (“myot_myob”) or myotube and MNC (“myot_mono”). This data is used to generate Figure S10B.

Table S21. Smart-seq coverage and NanoString expression of 490 miRNAs having NanoString count>=50 in at least one nucleus type (Myoblast, “myob”; MNC, “mnc”; Myotube, “myot”). NanoString expressions for each nucleus type are shown as duplicates and are averaged as “nano_avg”. Smart-seq coverage for each nucleus type is normalized into RPM (Reads Per Million) and labeled as “coverage_norm”. miRNAs are sorted by Pearson’s correlation coefficient between coverage and NanoString expression. This data is used to generate Figure 4C, 4D and S11A.

Table S22. Neighboring lncRNAs for 490 miRNAs having NanoString cout>=50 in at least one nucleus type. LncRNAs serving as pri-miRNAs are labeled with “yes”. This data is used to generate Figure 4C and S11B.