HSMMSingleCell: A single-cell RNA-Seq study of differentiating human skeletal muscle myoblasts

Cole Trapnell

Harvard University,
Cambridge, Massachussetts, USA
cole@broadinstitute.org

Davide Cacchiarelli

Harvard University,
Cambridge, Massachussetts, USA
davide@broadinstitute.org

April 14, 2014

Abstract

Contents

1 Introduction

Skeletal myoblasts undergo a well-characterized sequence of morphological and transcriptional changes during differentiation.

2 Cell Culture and Sequencing

In this experiment, primary human skeletal muscle myoblasts (HSMM) were expanded under high mitogen conditions (GM) and then differentiated by switching to low-mitogen media (DM). RNA-Seq libraries were sequenced from each of several hundred cells taken over a time-course of serum-induced differentiation. Between 49 and 77 cells were captured at each of four time points (0, 24, 48, 72 hours) following serum switch using the Fluidigm C1 microfluidic system. RNA from each cell was isolated and used to construct mRNA-Seq libraries, which were then sequenced to a depth of approximately 4 million reads per library, resulting in a complete gene expression profile for each cell.

For single-cell mRNA sequencing, dissociated cells were captured and processed with the C1 Single-Cell Auto Prep System (Fluidigm) following manufacturer's protocol 100-5950. Starting with a suspension of cells at a concentration of approximately 250 cells per microliter, up to 96 single cells are captured in each C1 microfluidic device. In this study, we used one C1 capture chip at 0, 24, 48, and 72 hours after switching to differentiation medium, for a total of four independent captures. After imaging with a microscope to identify which sites have captured a single cell, processing of the cells occurs within the C1 instrument to perform the steps of cell lysis, cDNA synthesis with reverse transcriptase, and PCR amplification of each cDNA library. The cDNA synthesis and PCR use reagents from the

SMARTer Ultra Low RNA Kit for Illumina Sequencing (Clontech 634936). The SMARTer chemistry utilizes a strand-switching mechanism so that both the 1st and 2nd strands of cDNA are synthesized in a single reaction. Following harvest from the C1 microfluidic device, each cDNA library is subjected to tagmentation (simultaneous fragmentation and tagging with sequencing adapters) using the Nextera XT DNA Sample Preparation Kit (Illumina FC-131-1096) as described in Fluidigm protocol 100-5950. PCR amplification of the tagmented cDNA uses Index Primers from the Nextera XT DNA Sample Preparation Index Kit (Illumina FC-131-1002). Following PCR, the cDNA libraries from individual cells are pooled and purified using AMPure XP beads (Agencourt Bioscience Corp A63880) as described in Fluidigm protocol 100-5950. Libraries that contained fewer than 1 million reads or for which less than 80% of fragments mapped to non-mitochondrial protein coding genes were excluded.

3 Estimating expression levels

Per-cell expression profiles were calculated in this study using the Tuxedo suite of tools [1]. The reads for each cell were mapped with TopHat [2] 2.0.9 and Bowtie [3] 2.0.6 against build 19 of the human genome, downloaded through the UCSC genome browser. TopHat was provided with GENCODE [4] gene annotations (build version 17). Mapped reads were analyzed with Cuffdiff [5] 2.2 to generate per-cell expression profiles.

4 The HSMM dataset

Gene expression values for each cell are stored in the `HSMM` object, which is a `CellDataSet`, as defined by *monocle*. To load the data, simply run the command below:

```
data(HSMM)
```

Because CellDataSet is based on Bioconductor's `ExpressionSet` class, CellDataSet inherits much of ExpressionSet's functionality. For example, you can quickly see how many features (i.e. genes) and samples (i.e. cells) are included in the dataset:

```
HSMM
```

```
## CellDataSet (storageMode: lockedEnvironment)
## assayData: 47192 features, 271 samples
##   element names: exprs
## protocolData: none
## phenoData
##   sampleNames: T0_CT_A01 T0_CT_A03 ... T72_CT_H12 (271 total)
##   varLabels: Library Well ... State (8 total)
##   varMetadata: labelDescription
## featureData
##   featureNames: ENSG00000000003.10 ENSG00000000005.5 ...
##     ENSGR0000270726.1 (47192 total)
##   fvarLabels: gene_short_name biotype
##   fvarMetadata: labelDescription
## experimentData: use 'experimentData(object)'
## Annotation:
```

Each CellDataSet includes expression levels, along with metadata about cells and genes. You can view metadata for cells as follows:

```
head(pData(HSMM))
```

```
##                Library Well Hours Media Mapped.Fragments
## T0_CT_A01 SCC10013_A01  A01     0    GM          1958074
## T0_CT_A03 SCC10013_A03  A03     0    GM          1930722
## T0_CT_A05 SCC10013_A05  A05     0    GM          1452623
## T0_CT_A06 SCC10013_A06  A06     0    GM          2566325
## T0_CT_A07 SCC10013_A07  A07     0    GM          2383438
## T0_CT_A08 SCC10013_A08  A08     0    GM          1472238
##           num_genes_expressed Pseudotime State
```

```
## T0_CT_A01                7383      7.200     1
## T0_CT_A03                7252      2.716     1
## T0_CT_A05                7410      2.272     1
## T0_CT_A06                5868      6.461     1
## T0_CT_A07                6472      3.402     1
## T0_CT_A08                6446     20.300     2
```

```
head(fData(HSMM))

##                     gene_short_name        biotype
## ENSG00000000003.10           TSPAN6 protein_coding
## ENSG00000000005.5              TNMD protein_coding
## ENSG00000000419.8              DPM1 protein_coding
## ENSG00000000457.8             SCYL3 protein_coding
## ENSG00000000460.12         C1orf112 protein_coding
## ENSG00000000938.8               FGR protein_coding
```

5 Further analysis

6 Citation

If you use Monocle to analyze your experiments, please cite:

```
citation("HSMMSingleCell")

##
## To cite package 'HSMMSingleCell' in publications use:
##
##   Cole Trapnell (2014). HSMMSingleCell: Single-cell RNA-Seq for
##   differentiating human skeletal muscle myoblasts (HSMM). R
##   package version 0.99.2.
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {HSMMSingleCell: Single-cell RNA-Seq for differentiating human skeletal muscle
## myoblasts (HSMM)},
##     author = {Cole Trapnell},
##     year = {2014},
##     note = {R package version 0.99.2},
##   }
##
## ATTENTION: This citation information has been auto-generated from
## the package DESCRIPTION file and may need manual editing, see
## 'help("citation")'.
```

7 Acknowledgements

Monocle was built by Cole Trapnell and Davide Cacchiarelli, with substantial design input John Rinn and Tarjei Mikkelsen. We are grateful to Sharif Bordbar, Chris Zhu, Amy Wagers and the Broad RNAi platform for technical assistance, and Magali Soumillon for helpful discussions. Cole Trapnell is a Damon Runyon Postdoctoral Fellow. Davide Cacchiarelli is a Human Frontier Science Program Fellow. Cacchiarelli and Mikkelsen were supported by the Harvard Stem Cell Institute. John Rinn is the Alvin and Esta Star Associate Professor. This work was supported by NIH grants 1DP2OD00667, P01GM099117, and P50HG006193-01. This work was also supported in part by the Single Cell Genomics initiative, a collaboration between the Broad Institute and Fluidigm Inc. This vignette was created

from Wolfgang Huber's Bioconductor vignette style document, and patterned after the vignette for *DESeq*, by Simon Anders and Wolfgang Huber.

8 Session Info

```
sessionInfo()

## R version 3.0.3 (2014-03-06)
## Platform: x86_64-apple-darwin13.1.0 (64-bit)
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
##  [1] grid      stats4    splines   parallel  stats     graphics  grDevices
##  [8] utils     datasets  methods   base
##
## other attached packages:
##  [1] monocle_0.99.0        cluster_1.14.4        HSMMSingleCell_0.99.2
##  [4] plyr_1.8.1            Hmisc_3.14-3          Formula_1.1-1
##  [7] survival_2.37-7       lattice_0.20-27       fastICA_1.2-0
## [10] combinat_0.0-8        igraph_0.7.0          matrixStats_0.8.14
## [13] irlba_1.0.3           VGAM_0.9-3            ggplot2_0.9.3.1
## [16] reshape2_1.2.2        Biobase_2.22.0        BiocGenerics_0.8.0
## [19] knitr_1.5
##
## loaded via a namespace (and not attached):
##  [1] colorspace_1.2-4    dichromat_2.0-0     digest_0.6.4
##  [4] evaluate_0.5.1      formatR_0.10        gtable_0.1.2
##  [7] highr_0.3           labeling_0.2        latticeExtra_0.6-26
## [10] MASS_7.3-29         Matrix_1.1-2        munsell_0.4.2
## [13] proto_0.3-10        R.methodsS3_1.6.1   RColorBrewer_1.0-5
## [16] Rcpp_0.11.1         scales_0.2.3        stringr_0.6.2
## [19] tools_3.0.3
```

References

[1] Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, 7(3):562–578, March 2012.

[2] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. TopHat2: accurate alignment of transcriptomes inthe presence of insertions, deletions and gene fusions. *Genome biology*, 14(4):R36, April 2013.

[3] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4):357–359, March 2012.

[4] Jennifer Harrow, France Denoeud, Adam Frankish, Alexandre Reymond, Chao-Kung Chen, Jacqueline Chrast, Julien Lagarde, James G R Gilbert, Roy Storey, David Swarbreck, Colette Rossier, Catherine Ucla, Tim Hubbard, Stylianos E Antonarakis, and Roderic Guigo. GENCODE: producing a reference annotation for ENCODE. *Genome biology*, 7 Suppl 1:S4.1–9, 2006.

[5] Cole Trapnell, David G Hendrickson, Martin Sauvageau, Loyal Goff, John L Rinn, and Lior Pachter. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology*, pages 1–9, December 2012.