Video Article

Identification of Key Factors Regulating Self-renewal and Differentiation in EML Hematopoietic Precursor Cells by RNA-sequencing Analysis

Shan Zong*[1], Shuyun Deng*[1], Kenian Chen[1], Jia Qian Wu[1]

[1] The Vivian L. Smith Department of Neurosurgery, Center for Stem Cell and Regenerative Medicine, University of Texas Health Science Center, The University of Texas Graduate School of Biomedical Sciences at Houston

* These authors contributed equally

Correspondence to: Jia Qian Wu at Jiaqian.Wu@uth.tmc.edu

URL: https://www.jove.com/video/52104
DOI: doi:10.3791/52104

Keywords: Genetics, Issue 93, EML Cells, Self-renewal, Differentiation, Hematopoietic precursor cell, RNA-Sequencing, Data analysis

Date Published: 11/11/2014

Citation: Zong, S., Deng, S., Chen, K., Wu, J.Q. Identification of Key Factors Regulating Self-renewal and Differentiation in EML Hematopoietic Precursor Cells by RNA-sequencing Analysis. *J. Vis. Exp.* (93), e52104, doi:10.3791/52104 (2014).

Abstract

Hematopoietic stem cells (HSCs) are used clinically for transplantation treatment to rebuild a patient's hematopoietic system in many diseases such as leukemia and lymphoma. Elucidating the mechanisms controlling HSCs self-renewal and differentiation is important for application of HSCs for research and clinical uses. However, it is not possible to obtain large quantity of HSCs due to their inability to proliferate *in vitro*. To overcome this hurdle, we used a mouse bone marrow derived cell line, the EML (Erythroid, Myeloid, and Lymphocytic) cell line, as a model system for this study.

RNA-sequencing (RNA-Seq) has been increasingly used to replace microarray for gene expression studies. We report here a detailed method of using RNA-Seq technology to investigate the potential key factors in regulation of EML cell self-renewal and differentiation. The protocol provided in this paper is divided into three parts. The first part explains how to culture EML cells and separate Lin-CD34+ and Lin-CD34- cells. The second part of the protocol offers detailed procedures for total RNA preparation and the subsequent library construction for high-throughput sequencing. The last part describes the method for RNA-Seq data analysis and explains how to use the data to identify differentially expressed transcription factors between Lin-CD34+ and Lin-CD34- cells. The most significantly differentially expressed transcription factors were identified to be the potential key regulators controlling EML cell self-renewal and differentiation. In the discussion section of this paper, we highlight the key steps for successful performance of this experiment.

In summary, this paper offers a method of using RNA-Seq technology to identify potential regulators of self-renewal and differentiation in EML cells. The key factors identified are subjected to downstream functional analysis *in vitro* and *in vivo*.

Video Link

The video component of this article can be found at https://www.jove.com/video/52104/

Introduction

Hematopoietic stem cells are rare blood cells that reside mainly in the adult bone marrow niche. They are responsible for the production of cells required to replenish the blood and the immune systems[1]. As a kind of stem cells, HSCs are capable of both self-renewal and differentiation. Elucidating mechanisms that control the fate decision of HSCs, toward either self-renewal or differentiation, will offer valuable guidance on the manipulation of HSCs for blood disease researches and clinical usage[2]. One problem faced by the researchers is that HSCs can be maintained and expanded *in vitro* to a very limited extent; the vast majority of their progeny are partially differentiated in culture[2].

In order to identify key regulators that control the processes of self-renewal and differentiation at a genome-wide scale, we used a mouse primitive hematopoietic progenitor cell line EML as a model system. This cell line was derived from murine bone marrow[3,4]. When fed with different growth factors, EML cells can differentiate into erythroid, myeloid, and lymphoid cells *in vitro*[5]. Importantly, this cell line can be propagated in large quantity in culture medium containing stem cell factor (SCF) and still retaining their multipotentiality. EML cells can be separated into subpopulations of self-renewing Lin-SCA+CD34+ and partially differentiated Lin-SCA-CD34- cells based on surface markers CD34 and SCA[6]. Similar to short-term HSCs, SCA+CD34+ cells are able of self-renewal. When treated with SCF, Lin-SCA+CD34+ cells can rapidly regenerate a mixed population of Lin-SCA+CD34+ and Lin-SCA-CD34- cells and continue to proliferate[6]. The two populations are similar in morphology and have similar levels of c-kit mRNA and protein[6]. Lin-SCA-CD34- cells are capable of propagating in media containing IL-3 instead of SCF[3]. Unveiling the key regulators in the EML cell fate decision will offer better understanding of cellular and molecular mechanisms in early developmental transition during hematopoiesis.

In order to investigate the underlying molecular differences between the self-renewing Lin-SCA+CD34+ and partially differentiated Lin-SCA-CD34- cells, we used RNA-Seq to identify differentially expressed genes. In particular, we focus on transcription factors, as transcription factors are crucial in determining cell fate. RNA-Seq is a recently developed approach that utilizes the capabilities of next-generation sequencing

(NGS) technologies to profile and quantify RNAs transcribed from genome[7,8]. In brief, total RNA is poly-A selected and fragmented as the initial template.The RNA template is then converted into cDNA using reverse transcriptase. In order to map full-length RNA transcripts, using intact, non-degraded RNA for constructing cDNA library is important. For the purpose of sequencing, specific adapter sequences are added to both ends of cDNA. Then, in most cases, cDNA molecules are amplified by PCR and sequenced in a high-throughput manner.

After sequencing, the resulting reads can be aligned to a reference genome and a transcriptome database. The number of reads that map to the reference gene is counted and this information can be used to estimate the gene expression level. The reads can also be assembled *de novo* without a reference genome, enabling the study of transcriptomes in non-model organisms[9]. RNA-seq technology has also been used to detect splice isoforms[10-12], novel transcripts[13] and gene fusions[14]. In addition to the detection of protein-coding genes, RNA-Seq can also be used to detect novel and analyze transcription level of non-coding RNAs, such as long non-coding RNA[15,16], microRNA[17], siRNA *etc.*[18]. Because of the accuracy of this method, it has been utilized for detection of single nucleotide variations[19,20].

Before the advent of RNA-Seq technology, microarray was the main method used for analyzing gene expression profile. Pre-designed probes are synthesized and subsequently attached to a solid surface to form a microarray slide[21]. mRNA is extracted and converted to cDNA. During the reverse transcription process, fluorescently labeled nucleotides are incorporated into the cDNA and the cDNA can be hybridized onto the microarray slides. The intensity of the signal collected from a specific spot depends on the amount of cDNA binding to the specific probe on that spot[21]. Compared with RNA-Seq technology, microarray has several limitations. First, microarray relies on the pre-existing knowledge of gene annotation, while RNA-Seq technology is able to detect novel transcripts at relative high background level, which limits its use when gene expression level is low. Besides, the RNA-Seq technology has much higher dynamic range of detection (8,000 fold)[7], whereas, due to background and saturation of signals, the accuracy of microarray is limited for both highly and lowly expressed genes[7,22]. Finally, microarray probes differ in their hybridization efficiencies, which make the results less reliable when comparing relative expression levels of different transcripts within one sample[23]. Although RNA-Seq has many advantages over microarray, its data analysis is complex. This is one of the reasons that many researchers still use microarray instead of RNA-Seq. Various bioinformatics tools are required for RNA-Seq data processing and analysis[24].

Among several next-generation sequencing (NGS) platforms, 454, Illumina, SOLID and Ion Torrent are the most widely used ones. 454 was the first commercial NGS platform. In contrast to the other sequencing platforms such as illumina and SOLID, the 454 platform generates longer read length (average 700 base reads)[25]. Longer reads are better for initial characterization of transcriptiome due to their higher assemble efficiency[25]. The main disadvantage of the 454 platform is its high cost per megabase of sequence. The Illumina and SOLID platforms generate reads with increased numbers and short lengths. The cost per megabase of sequence is much lower than the 454 platform. Due to the large numbers of short reads for the Illumina and SOLID platforms, data analysis is much more computationally intensive. The price of the instrument and reagents for sequencing for the Ion Torrent platform is cheaper and the sequencing time is shorter[25]. However, the error rate and the cost per megabase of sequence are higher compared to the Illumina and SOLID platforms. Different platforms have their own advantages and disadvantages and require different methods for data analysis. The platform should be chosen based on the sequencing purpose and the availability of funding.

In this paper, we take Illumina RNA-Seq platform as an example. We used EML cell as a model system to investigate the key regulators in EML cell self-renewal and differentiation, and provided a detailed methods of RNA-Seq library construction and data analysis for expression level calculation and novel transcript detection. We have shown in our previous publication that RNA-seq study in EML model system[2], when coupled with functional test (*e.g.* shRNA knockdown) provide a powerful approach in understanding the molecular mechanism of the early stages of hematopoietic differentiation, and can serve as a model for the analysis of cell self-renewal and differentiation in general.

Protocol

1. EML Cell Culture and Separation of Lin-CD34+ and Lin-CD34- Cells Using Magnetic Cell Sorting System and Fluorescence-activated Cell Sorting Method

1. Preparation of baby hamster kidney (BHK) cell culture medium for stem cell factor collection:
    1. Culture BHK cells in DMEM medium containing 10% FBS in 25 cm^2 flask (**Table 1**) at 37 °C, 5% CO_2 in a cell culture incubator.
    2. When cells grow to 80 - 90% confluence, wash cells once with 10 ml of PBS. Add 5 ml of 0.25% trypsin-EDTA solution to the monolayer and incubate the cells for 1-5 min at room temperature (RT) until the cells are detached.
    3. Pipet the solution up and down gently to break up clumps of cells. Add 5 ml of complete DMEM to the flask to stop trypsin activity. Collect cells by centrifugation at 200 x g for 5 min at RT.
    4. Remove the medium and resuspend the cell pellet in 10 ml of fresh BHK cell culture medium.
    5. Transfer 2 ml of the cell suspension from the **step 1.1.4** to a new 75 cm^2 flask and add 48 ml of fresh BHK cell culture medium to the flask.
    6. Culture the BHK cells for two days and collect the culture medium. Passage the medium through a 0.45 µm filter. Store the medium in -20 °C until further use.

2. EML cell culture:
    1. Culture EML cells (in suspension) in EML basic medium containing BHK cell culture medium (**Table 1**) at 37 °C, 5% CO_2 in a cell culture incubator.
    2. Maintain the EML cells at low cell density (0.5-5 x 10^5 cells/ml) with the peak density less than 6 x 10^5 cells/ml. Split the cells every 2-3 days at the ratio of 1:5. Passage EML cells gently and discard the culture after passaging for 10 generations.

3. Depletion of lineage positive cells:
    1. Harvest the EML cells by centrifugation at 200 x g for 5 min and wash the cells once with PBS. Collect the cells by centrifugation at 200 x g for 5 min.

2. Resuspend the cells with PBS and count the cells with a hemocytometer. Determine the antibody concentration in the subsequent cell separation step according to the number of the cells (please refer to the instructions offered by the provider of the cell isolation system).
3. Isolate the lineage negative (Lin-) cells using lineage antibody cocktail (cocktail of biotin-conjugated monoclonal antibodies CD5, CD45R (B220), CD11b, Anti-Gr-1(Ly-6G/C), 7-4 and Ter-119) and a magnetic activated cell sorting system according to manufacturer's instructions.

4. Separation of Lin-CD34+ and Lin-CD34- cells:
   1. Spin down the Lin- cells from the **step 1.3.3** at 200 x g for 5 min. Resuspend the cell pellet with PBS and count the cells with a hemocytometer.
   2. Wash the cells twice with FACS buffer and pellet the cells at 200 x g for 5 min.
   3. Label five 1.5 ml microcentrifuge tubes with the number 1, 2, 3, 4, 5 respectively. Resuspend the cells with 100 µl FACS buffer per 10^6 cells (10^6 cells per tube).
   4. Add 1 µg of Anti-Mouse CD34 FITC antibody to tube 1 and tube 2 and mix the tubes gently.
   5. Incubate all tubes at 4 °C for 1 hr in the dark.
   6. Add 0.25 µg of PE-conjugated Anti-Sca1 antibody and 20 µl of APC-conjugated Lineage Cocktail antibodies to tube 1, 0.25 µg of PE-conjugated Anti-Sca1 antibody to tube 3, and 20 µl of APC-conjugated Lineage Cocktail antibodies to tube 4.
   7. Mix all the tubes gently and incubate the cells at 4 °C for an additional 30 min in the dark.
   8. Add 300 µl of FACS buffer to the cells and spin down the cells at 200 x g for 5 min.
   9. Wash the cells with 500 µl of FACS buffer for three times.
   10. Resuspend the cell pellet in 500 µl of FACS buffer.
   11. Use the cells in tubes 2, 3, 4, and 5 for setting up compensation. Isolate Lin-SCA+CD34+ and Lin-SCA-CD34- cells in tube 1 using FACS Aria.

2. RNA Preparation and Library Construction for High-throughput Sequencing

1. Isolation, quality analysis and quantification of RNA:
   1. Extract total RNA from Lin-CD34+ and Lin-CD34- cells respectively using TRIzol following the manufactures' protocol.
   2. Remove the contaminated DNA using deoxyribonuclease I (DNase I) following the manufacture's protocol. Optionally, store the RNA at -80 °C at this step for further use.
   3. Assess the quality of total RNA using Bioanalyzer according to the instructions offered by the supplier. Use RNA sample with RNA Integrity Number (RIN) lager than 9.

2. Library Construction and high-throughput sequencing:
   NOTE: This protocol describes RNA-Seq using Illumina platform. For other sequencing platforms, different library preparation methods are required.
   1. Use 0.1-4 µg of high quality total RNA per sample for library preparation. Normally 2 µg of total RNA can be extracted from 10^5 EML cells.
   2. Use a RNA-sequencing sample preparation system for RNA purification and fragmentation, first and second strand cDNA synthesis, end repair, 3' ends adenylation, adapter ligation and PCR amplification, following the detailed standard procedures from the provider's instructions.
      1. Positively select PolyA mRNA using oligo-dT magnetic beads and fragment the mRNA.
      2. Perform reverse transcription using random primers to obtain the cDNA and subsequently synthesize the second strand of cDNA to generate double stranded cDNA.
      3. Remove the 3' overhangs and fill the 5' overhangs by DNA polymerase. Adenylate 3' ends to prevent cDNA fragments from ligating to one another.
      4. Add multiplex indexing adapters to both ends of the dscDNA. Perform PCR for the enrichment of DNA fragments.

   3. Measure the A260/A280 to obtain information about the concentration of library using a spectrophotometer.
   4. Assess the library quality and measure the size range of DNA fragments using a Bioanalyzer.

3. Data Analysis

For reference of software used in this part, please see (**Table 2**).

1. Data file processing for downstream analysis:
   1. Convert .bcl (base call file) file to .fastq file using CASAVA software (Illumina, version 1.8.2).
      1. Fire up the 'Terminal' in Linux system. Go to the data folder that contains the data file from an Illumina HiSeq2000 sequencing machine. Suppose the result folder is 'NASboy1/JiaqianLabData/HiSeq_RUN/2013_07_11/130627_SN860_0309_A_2013-166_H0PW9ADXX/', type in the command in **Figure S1A**, and enter the data folder.
      2. Install CASAVA 1.8.2 in the Linux system. Suppose the outputfolder is 'Unaligned', use the command in **Figure S1B** to prepare the configuration file for converting. Use the option --fastq-cluster-count 0 to ensure only one .fastq file is created for each sample. The generated .fastq file is in .gz format. Unzip it for downstream analysis (**Figure S1B**).
      3. After the 'Unaligned' folder has been generated, go to the 'Unaligned' folder (**Figure S1C**).
      4. Use the command in **Figure S1D** to begin the converting process. The '-j' parameter supplies the cpu number that will be used.
      5. After the system finished the converting process, go to the result folder under 'Unaligned' folder (**Figure S1E**).

        6. Use the command in **Figure S1F** to decompress the .fastq.gz file into .fastq file under each sample folder.

2. Detect novel transcripts and evaluate the expression level using Tuxedo Suite[26]:
    1. Map the paired-end RNA-Seq reads to the mouse reference genome (UCSC version mm9, obtained from http://cufflinks.cbcb.umd.edu/igenomes.html) using Tophat software (version 1.3.3)[27], which uses the Bowtie read mapper (version 0.12.7)[28]. Tophat is supplied with "-no-novel-juncs" option to improve estimation accuracy of expression level.
        1. Put the .fastq files in a folder where the mapping process will be implemented. Suppose there are 2 .fastq files (rename to Example1.read1, Example1.read2) for a paired-end sequencing sample, use the command in **Figure S2** to do the mapping (adjust the parameters according to the system setting). The "-p" parameter supplies the cpu number that will be used. The "–r" and "–mate-std-dev" parameters can be obtained from library QC or inferred from a subset of aligned reads (**Figure S2**).

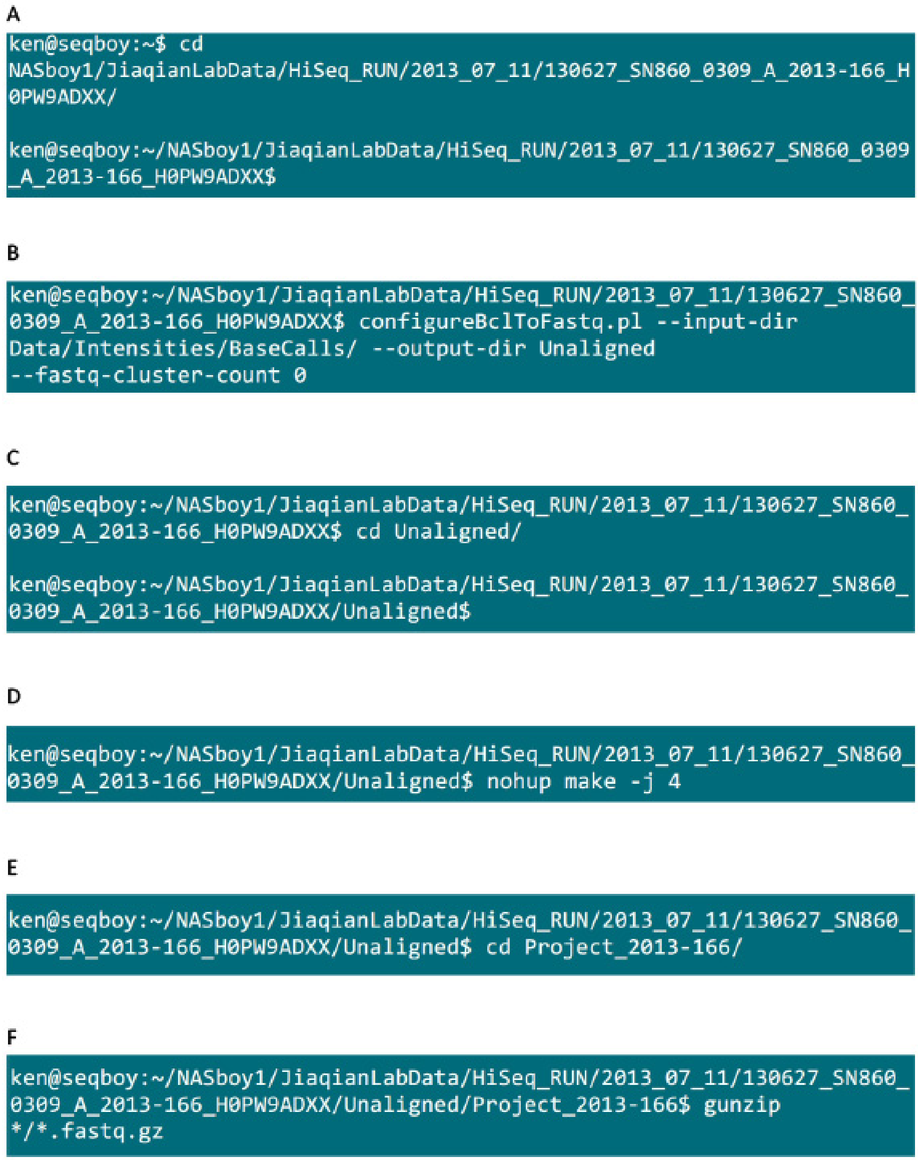
    2. Assemble the mapped reads into RNA transcripts using the Cufflinks software (version 1.3.0)[29]. Run Cufflinks using the annotation file of known genes (same .gtf file used by Tophat) and .bam file produced by Tophat.
        1. After Tophat finished running, in the same folder, use the command in **Figure S3A** to run cufflinks to construct transcriptome and estimate transcript expression level. The 'mm9_repeatMasker.gtf' and genome sequence files in the 'GenomeSeqMM9' folder can be obtained from UCSC Genome Browser.
        2. The resulting genes.expr and transcripts.expr files contain the expression value of genes and transcripts (isoforms). Copy and paste the file contents to an Excel file and manipulate with spreadsheet application (**Figure S3B**).
        3. Use the command in **Figure S3C** to compare the resulting 'transcripts.gtf' file to the reference 'mm9_genes.gtf' file in order to identify novel transcripts.
        4. The resulting .tmap file contains the comparison result. Copy and paste the file contents to an Excel file and manipulate with spreadsheet application. Transcripts with class code 'u' can be considered as 'novel' compared to the reference .gtf file provided (**Figure S3D**).
           NOTE: For downstream analysis convenience, set the FPKM values to 0.1 if the values are under 0.1.
           NOTE: **Step 3.2.3 - 3.2.6** is optional for those who wish to improve accuracy of novel transcripts' expression estimation. This will take a much longer time, because mapping and transcriptome construction need to be run more than once.

    3. Run Tophat using default parameters and then run cufflinks to generated .gtf file using the command in **Figure S3E**.
    4. Compare the resulting .gtf file to the reference genome .gtf file using the command in **Figure S3F**.
    5. Parse the resulted .tmap file as described in the **step 3.2.2.4**. Copy and paste the file contents to an Excel file and manipulate with spreadsheet application. Transcripts with class code 'u' can be considered as 'novel' compared to the reference .gtf file provided.
    6. After the **step 3.2.5**, there is a .combined.gtf file in the folder which can be used as the reference .gtf file. A second run of Tophat and cufflinks can be performed as described in the **step 3.2.1** and **3.2.2** to obtain a more accurate FPKM estimation of novel transcripts.

3. Detect differentially expressed genes using DESeq package[30].
    1. The input of DESeq is a raw read counts table. To obtain such a table, use the htseq-count script distributed with the HTSeq Python package which can be downloaded from HTSeq website (http://www-huber.embl.de/users/anders/HTSeq/doc/count.html).
        1. Ensure that samtools, python, and htseq-count programsare installed in the system. Obtain raw read count numbers from tophat output by using the command in **Figure S4A**.
        2. Prepare 'Raw_Count_Table.txt', 'ExperimentDesign.txt' files using Excel. Copy and save the content in .txt format for the DESeq R package (**Figure S4B**).
        3. Install R program in the system. In the terminal, type 'R' and press ENTER.A screen message will appearas showed in **Figure S4C**.
        4. Read 'Raw_Count_Table.txt', 'ExperimentDesign.txt' into R using the command in **Figure S4D**.
        5. Load DESeq package using the command in **Figure S4E**.
        6. Factorize conditions in R (**Figure S4F**).
        7. Use the command in **Figure S4G** to run negative binominal test on the normalized count table.
        8. Use the command in **Figure S4H** to output significant differential expressed genes in a .csv file.

4. Lookup transcription factors' (TFs) FPKM values across samples using Excel. Intersect DE gene table and TFs table. Genes belong to both table are differentially expressed transcription factors.
    1. Go to the website http://www.bioguo.org/AnimalTFDB/download.php and download the transcription factors. Then lookup the DE transcription factors in the Excel (**Figure S5**).

5. Generating .bigwig file for UCSC genome browser visualization.
    1. Download 'bedtools' software package from the website https://github.com/arq5x/bedtools2 and install the software in the system[31]. Download the UCSC tools 'bedGraphToBigWig' from the website http://hgdownload.cse.ucsc.edu/admin/exe/ and install the software in the system.
    2. In the folder containing the .bam file, use the command in **Figure S6A** to convert .bam file generated by tophat into .bed file.
    3. After the .bed file is produced, use the command in **Figure S6B** to generate .bigwig file. The file 'ChromInfo.txt' can be obtained from following url: http://hgdownload.cse.ucsc.edu/goldenPath/mm9/database/chromInfo.txt.gz.
    4. Observe a custom track on UCSC Genome Browser. Refer to the website http://genome.ucsc.edu/goldenPath/help/customTrack.html on how to display a custom track using UCSC genome browser.



A

```
ken@seqboy:~$ cd
NASboy1/JiaqianLabData/HiSeq_RUN/2013_07_11/130627_SN860_0309_A_2013-166_H
0PW9ADXX/

ken@seqboy:~/NASboy1/JiaqianLabData/HiSeq_RUN/2013_07_11/130627_SN860_0309
_A_2013-166_H0PW9ADXX$
```

B

```
ken@seqboy:~/NASboy1/JiaqianLabData/HiSeq_RUN/2013_07_11/130627_SN860_
0309_A_2013-166_H0PW9ADXX$ configureBclToFastq.pl --input-dir
Data/Intensities/BaseCalls/ --output-dir Unaligned
--fastq-cluster-count 0
```

C

```
ken@seqboy:~/NASboy1/JiaqianLabData/HiSeq_RUN/2013_07_11/130627_SN860_
0309_A_2013-166_H0PW9ADXX$ cd Unaligned/

ken@seqboy:~/NASboy1/JiaqianLabData/HiSeq_RUN/2013_07_11/130627_SN860_
0309_A_2013-166_H0PW9ADXX/Unaligned$
```

D

```
ken@seqboy:~/NASboy1/JiaqianLabData/HiSeq_RUN/2013_07_11/130627_SN860_
0309_A_2013-166_H0PW9ADXX/Unaligned$ nohup make -j 4
```

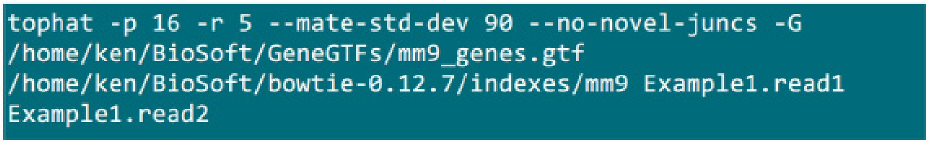
E

```
ken@seqboy:~/NASboy1/JiaqianLabData/HiSeq_RUN/2013_07_11/130627_SN860_
0309_A_2013-166_H0PW9ADXX/Unaligned$ cd Project_2013-166/
```

F

```
ken@seqboy:~/NASboy1/JiaqianLabData/HiSeq_RUN/2013_07_11/130627_SN860_
0309_A_2013-166_H0PW9ADXX/Unaligned/Project_2013-166$ gunzip
*/*.fastq.gz
```

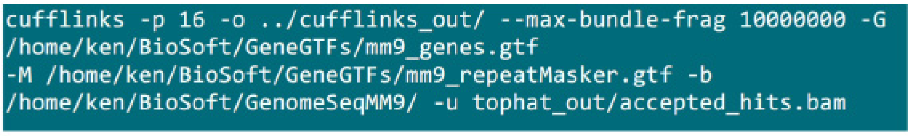
Figure S1: Converting .bcl file to .fastq file using CASAVA software.



```
tophat -p 16 -r 5 --mate-std-dev 90 --no-novel-juncs -G
/home/ken/BioSoft/GeneGTFs/mm9_genes.gtf
/home/ken/BioSoft/bowtie-0.12.7/indexes/mm9 Example1.read1
Example1.read2
```

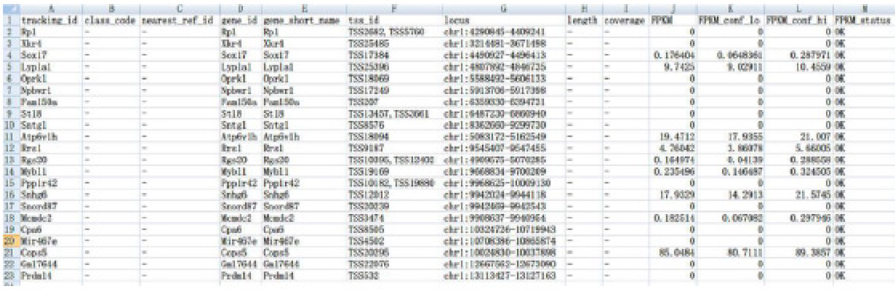
Figure S2: Mapping reads to reference genome using Tophat.

A



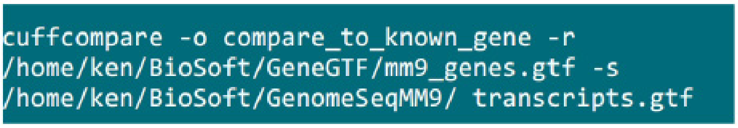
```
cufflinks -p 16 -o ../cufflinks_out/ --max-bundle-frag 10000000 -G
/home/ken/BioSoft/GeneGTFs/mm9_genes.gtf
-M /home/ken/BioSoft/GeneGTFs/mm9_repeatMasker.gtf -b
/home/ken/BioSoft/GenomeSeqMM9/ -u tophat_out/accepted_hits.bam
```

B



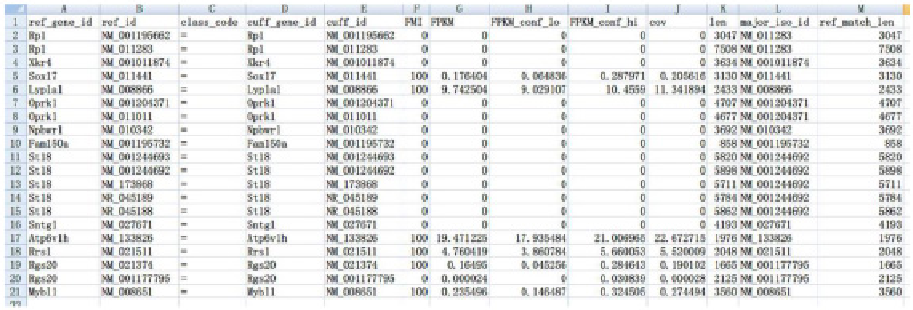
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	tracking_id	class_code	nearest_ref_id	gene_id	gene_short_name	tss_id	locus	length	coverage	FPKM	FPKM_conf_lo	FPKM_conf_hi	FPKM_status
2	Rp1	-	-	Rp1	Rp1	TSS2682, TSS5760	chr1:4280645-4409241	-	-	0	0	0	OK
3	Xkr4	-	-	Xkr4	Xkr4	TSS25485	chr1:3214481-3671498	-	-	0	0	0	OK
4	Sox17	-	-	Sox17	Sox17	TSS17384	chr1:4490927-4496413	-	-	0.176404	0.0648361	0.287971	OK
5	Lypla1	-	-	Lypla1	Lypla1	TSS25396	chr1:4807892-4846735	-	-	9.7425	9.02911	10.4559	OK
6	Oprk1	-	-	Oprk1	Oprk1	TSS18069	chr1:5588492-5606133	-	-	0	0	0	OK
7	Npbwr1	-	-	Npbwr1	Npbwr1	TSS17249	chr1:5913706-5917398	-	-	0	0	0	OK
8	Fam150a	-	-	Fam150a	Fam150a	TSS207	chr1:6359330-6394731	-	-	0	0	0	OK
9	St18	-	-	St18	St18	TSS13457, TSS3061	chr1:6487230-6860940	-	-	0	0	0	OK
10	Sntg1	-	-	Sntg1	Sntg1	TSS8576	chr1:8362960-9299730	-	-	0	0	0	OK
11	Atp6v1h	-	-	Atp6v1h	Atp6v1h	TSS18094	chr1:5083172-5162549	-	-	19.4712	17.9355	21.007	OK
12	Rrs1	-	-	Rrs1	Rrs1	TSS9187	chr1:9545407-9547455	-	-	4.76042	3.86078	5.66005	OK
13	Rgs20	-	-	Rgs20	Rgs20	TSS10095, TSS12402	chr1:4909575-5070285	-	-	0.164974	0.04139	0.288558	OK
14	Mybl1	-	-	Mybl1	Mybl1	TSS19109	chr1:9668834-9700209	-	-	0.235496	0.146487	0.324505	OK
15	Ppp1r42	-	-	Ppp1r42	Ppp1r42	TSS10182, TSS19890	chr1:9968625-10009130	-	-	0	0	0	OK
16	Snhg6	-	-	Snhg6	Snhg6	TSS12012	chr1:9942024-9944118	-	-	17.9329	14.2913	21.5745	OK
17	Snord87	-	-	Snord87	Snord87	TSS20239	chr1:9942469-9942543	-	-	0	0	0	OK
18	Mcmdc2	-	-	Mcmdc2	Mcmdc2	TSS3474	chr1:9908637-9940964	-	-	0.182514	0.067082	0.297946	OK
19	Cpa6	-	-	Cpa6	Cpa6	TSS8505	chr1:10324726-10719943	-	-	0	0	0	OK
20	Mir467e	-	-	Mir467e	Mir467e	TSS4502	chr1:10708396-10865874	-	-	0	0	0	OK
21	Cops5	-	-	Cops5	Cops5	TSS20295	chr1:10024830-10037898	-	-	85.0484	80.7111	89.3857	OK
22	Gm17644	-	-	Gm17644	Gm17644	TSS22076	chr1:12667562-12673090	-	-	0	0	0	OK
23	Prdm14	-	-	Prdm14	Prdm14	TSS532	chr1:13113427-13127163	-	-	0	0	0	OK

C



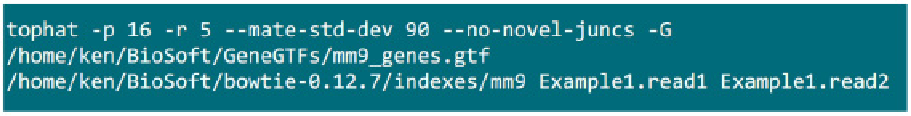
```
cuffcompare -o compare_to_known_gene -r
/home/ken/BioSoft/GeneGTF/mm9_genes.gtf -s
/home/ken/BioSoft/GenomeSeqMM9/ transcripts.gtf
```

D



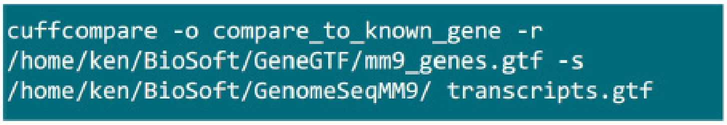
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	ref_gene_id	ref_id	class_code	cuff_gene_id	cuff_id	FMI	FPKM	FPKM_conf_lo	FPKM_conf_hi	cov	len	major_iso_id	ref_match_len
2	Rp1	NM_001195662	=	Rp1	NM_001195662	0	0	0	0	0	3047	NM_011283	3047
3	Rp1	NM_011283	=	Rp1	NM_011283	0	0	0	0	0	7508	NM_011283	7508
4	Xkr4	NM_001011874	=	Xkr4	NM_001011874	0	0	0	0	0	3634	NM_001011874	3634
5	Sox17	NM_011441	=	Sox17	NM_011441	100	0.176404	0.064836	0.287971	0.205616	3130	NM_011441	3130
6	Lypla1	NM_008866	=	Lypla1	NM_008866	100	9.742504	9.029107	10.4559	11.341894	2433	NM_008866	2433
7	Oprk1	NM_001204371	=	Oprk1	NM_001204371	0	0	0	0	0	4707	NM_001204371	4707
8	Oprk1	NM_011011	=	Oprk1	NM_011011	0	0	0	0	0	4677	NM_001204371	4677
9	Npbwr1	NM_010342	=	Npbwr1	NM_010342	0	0	0	0	0	3692	NM_010342	3692
10	Fam150a	NM_001195732	=	Fam150a	NM_001195732	0	0	0	0	0	858	NM_001195732	858
11	St18	NM_001244693	=	St18	NM_001244693	0	0	0	0	0	5820	NM_001244692	5820
12	St18	NM_001244692	=	St18	NM_001244692	0	0	0	0	0	5898	NM_001244692	5898
13	St18	NM_173868	=	St18	NM_173868	0	0	0	0	0	5711	NM_001244692	5711
14	St18	NR_045189	=	St18	NR_045189	0	0	0	0	0	5784	NM_001244692	5784
15	St18	NR_045188	=	St18	NR_045188	0	0	0	0	0	5862	NM_001244692	5862
16	Sntg1	NM_027671	=	Sntg1	NM_027671	0	0	0	0	0	4193	NM_027671	4193
17	Atp6v1h	NM_133826	=	Atp6v1h	NM_133826	100	19.471225	17.935484	21.006966	22.672715	1976	NM_133826	1976
18	Rrs1	NM_021511	=	Rrs1	NM_021511	100	4.760419	3.860784	5.660053	5.520009	2048	NM_021511	2048
19	Rgs20	NM_021374	=	Rgs20	NM_021374	100	0.16496	0.045256	0.284643	0.190102	1665	NM_001177795	1665
20	Rgs20	NM_001177795	=	Rgs20	NM_001177795	0	0.000024	0	0.030839	0.000028	2125	NM_001177795	2125
21	Mybl1	NM_008651	=	Mybl1	NM_008651	100	0.235496	0.146487	0.324505	0.274494	3560	NM_008651	3560

E



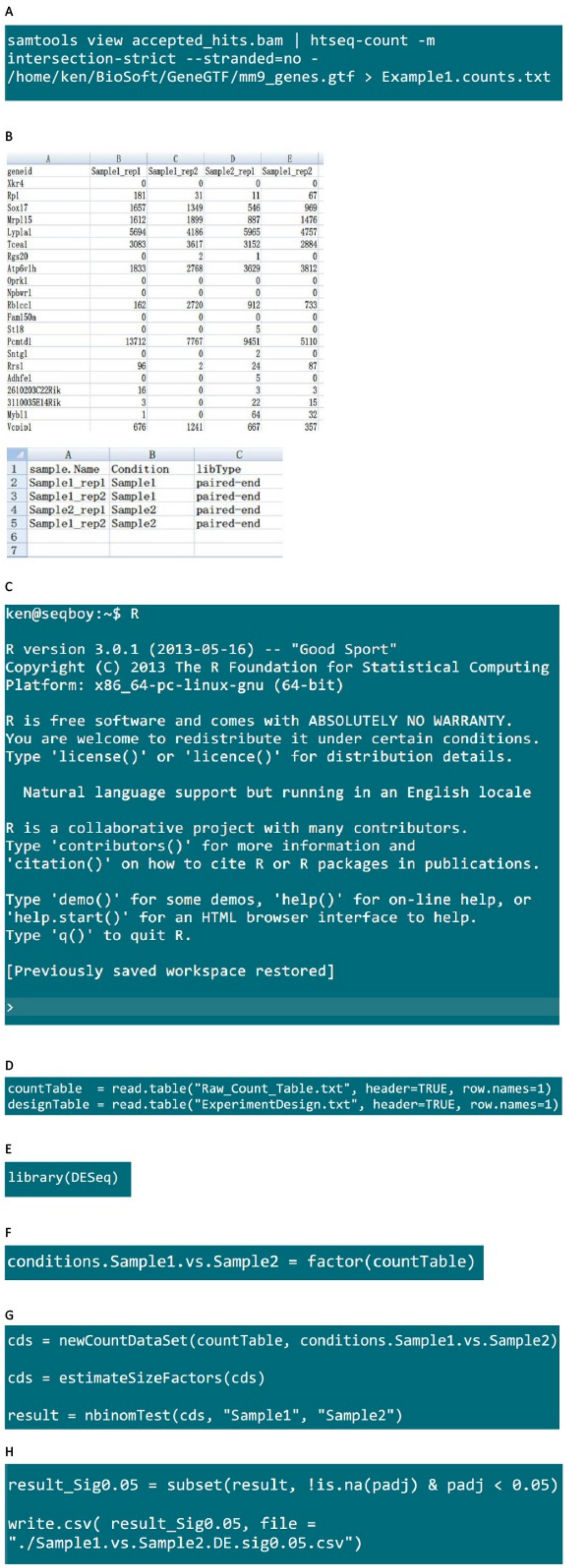
```
tophat -p 16 -r 5 --mate-std-dev 90 --no-novel-juncs -G
/home/ken/BioSoft/GeneGTFs/mm9_genes.gtf
/home/ken/BioSoft/bowtie-0.12.7/indexes/mm9 Example1.read1 Example1.read2
```

F



```
cuffcompare -o compare_to_known_gene -r
/home/ken/BioSoft/GeneGTF/mm9_genes.gtf -s
/home/ken/BioSoft/GenomeSeqMM9/ transcripts.gtf
```

Figure S3: Detection of novel transcripts and expression level estimation.



A

```
samtools view accepted_hits.bam | htseq-count -m
intersection-strict --stranded=no -
/home/ken/BioSoft/GeneGTF/mm9_genes.gtf > Example1.counts.txt
```

B

A	B	C	D	E
geneid	Sample1_rep1	Sample1_rep2	Sample2_rep1	Sample1_rep2
Xkr4	0	0	0	0
Rp1	181	31	11	67
Sox17	1657	1349	546	969
Mrpl15	1612	1899	887	1476
Lypla1	5694	4186	5965	4757
Tcea1	3083	3617	3152	2884
Rgs20	0	2	1	0
Atp6v1h	1833	2768	3629	3812
Oprk1	0	0	0	0
Npbwr1	0	0	0	0
Rb1cc1	162	2720	912	733
Fam150a	0	0	0	0
St18	0	0	5	0
Pcmtd1	13712	7767	9451	5110
Sntg1	0	0	2	0
Rrs1	96	2	24	87
Adhfe1	0	0	5	0
2610203C22Rik	16	0	3	3
3110035E14Rik	3	0	22	15
Mybl1	1	0	64	32
Vcpip1	676	1241	667	357

	A	B	C
1	sample.Name	Condition	libType
2	Sample1_rep1	Sample1	paired-end
3	Sample1_rep2	Sample1	paired-end
4	Sample2_rep1	Sample2	paired-end
5	Sample1_rep2	Sample2	paired-end
6			
7			

C

```
ken@seqboy:~$ R

R version 3.0.1 (2013-05-16) -- "Good Sport"
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]

>
```

D

```
countTable  = read.table("Raw_Count_Table.txt", header=TRUE, row.names=1)
designTable = read.table("ExperimentDesign.txt", header=TRUE, row.names=1)
```

E

```
library(DESeq)
```

F

```
conditions.Sample1.vs.Sample2 = factor(countTable)
```

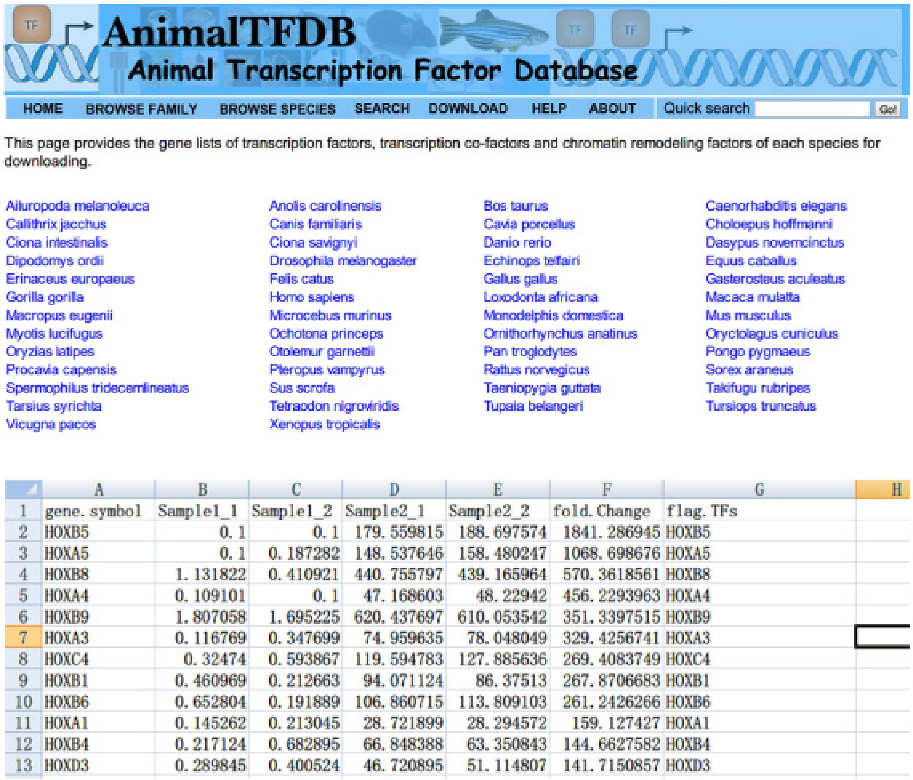
G

```
cds = newCountDataSet(countTable, conditions.Sample1.vs.Sample2)

cds = estimateSizeFactors(cds)

result = nbinomTest(cds, "Sample1", "Sample2")
```

H

```
result_Sig0.05 = subset(result, !is.na(padj) & padj < 0.05)

write.csv( result_Sig0.05, file =
"./Sample1.vs.Sample2.DE.sig0.05.csv")
```

Figure S4: Calling differential expressed gene using DESeq package.

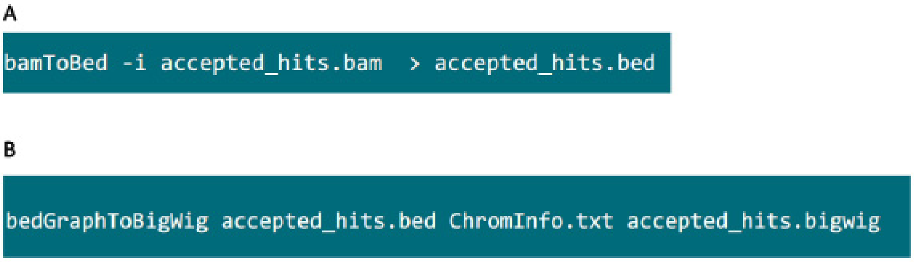


This page provides the gene lists of transcription factors, transcription co-factors and chromatin remodeling factors of each species for downloading.

Ailuropoda melanoleuca
Anolis carolinensis
Bos taurus
Caenorhabditis elegans
Callithrix jacchus
Canis familiaris
Cavia porcellus
Choloepus hoffmanni
Ciona intestinalis
Ciona savignyi
Danio rerio
Dasypus novemcinctus
Dipodomys ordii
Drosophila melanogaster
Echinops telfairi
Equus caballus
Erinaceus europaeus
Felis catus
Gallus gallus
Gasterosteus aculeatus
Gorilla gorilla
Homo sapiens
Loxodonta africana
Macaca mulatta
Macropus eugenii
Microcebus murinus
Monodelphis domestica
Mus musculus
Myotis lucifugus
Ochotona princeps
Ornithorhynchus anatinus
Oryctolagus cuniculus
Oryzias latipes
Otolemur garnettii
Pan troglodytes
Pongo pygmaeus
Procavia capensis
Pteropus vampyrus
Rattus norvegicus
Sorex araneus
Spermophilus tridecemlineatus
Sus scrofa
Taeniopygia guttata
Takifugu rubripes
Tarsius syrichta
Tetraodon nigroviridis
Tupaia belangeri
Tursiops truncatus
Vicugna pacos
Xenopus tropicalis

	A	B	C	D	E	F	G	H
1	gene. symbol	Sample1_1	Sample1_2	Sample2_1	Sample2_2	fold. Change	flag. TFs	
2	HOXB5	0. 1	0. 1	179. 559815	188. 697574	1841. 286945	HOXB5	
3	HOXA5	0. 1	0. 187282	148. 537646	158. 480247	1068. 698676	HOXA5	
4	HOXB8	1. 131822	0. 410921	440. 755797	439. 165964	570. 3618561	HOXB8	
5	HOXA4	0. 109101	0. 1	47. 168603	48. 22942	456. 2293963	HOXA4	
6	HOXB9	1. 807058	1. 695225	620. 437697	610. 053542	351. 3397515	HOXB9	
7	HOXA3	0. 116769	0. 347699	74. 959635	78. 048049	329. 4256741	HOXA3	
8	HOXC4	0. 32474	0. 593867	119. 594783	127. 885636	269. 4083749	HOXC4	
9	HOXB1	0. 460969	0. 212663	94. 071124	86. 37513	267. 8706683	HOXB1	
10	HOXB6	0. 652804	0. 191889	106. 860715	113. 809103	261. 2426266	HOXB6	
11	HOXA1	0. 145262	0. 213045	28. 721899	28. 294572	159. 127427	HOXA1	
12	HOXB4	0. 217124	0. 682895	66. 848388	63. 350843	144. 6627582	HOXB4	
13	HOXD3	0. 289845	0. 400524	46. 720895	51. 114807	141. 7150857	HOXD3	

Figure S5: Identification of differentially expressed transcription factors.



```
bamToBed -i accepted_hits.bam  > accepted_hits.bed
```

```
bedGraphToBigWig accepted_hits.bed ChromInfo.txt accepted_hits.bigwig
```

Figure S6: Converting mapping result for data visualization.

Representative Results

In order to analyze differentially expressed genes in Lin-CD34+ and Lin-CD34- EML cells, we used RNA-Seq technology. **Figure 1** shows the workflow of the procedures. After isolation of lineage negative cells by magnetic cell sorting, we separated Lin-SCA+CD34+ and Lin-SCA-CD34- cells using FACS Aria. Lin-enriched EML cells were stained with anti-CD34, anti-Sca1 and lineage cocktail antibodies. Only Lin- cells were gated for analysis of Sca1 and CD34 expression. Two populations (SCA+CD34+ and SCA-CD34- EML cells) could be observed by FACS analysis (**Figure 2**)[6].

After cell separation, we extracted total RNA from CD34+ and CD34- cells respectively and analyzed the quality of RNA. The accuracy of RNA-Seq data largely relies on the quality of RNA-Seq library and the quality of total RNA is vital for preparing a high quality library. High quality RNA sample should have an OD 260/280 value between 1.8 and 2.0. In addition to using the spectrophotometer, RNA quality was further assessed with greater accuracy by Bioanalyzer. **Figure 3** shows a result of a high quality RNA sample with the RIN equal to 9.4. Only high quality total RNA sample with RIN value greater than 9 was used for mRNA extraction and subsequent library construction procedures.

Ribosomal RNA is the most abundant type of RNA in cell. Currently two main strategies, depletion of rRNA or positively selection of polyadenylated mRNA (poly-A mRNA), are used for enrichment of target RNA before library construction. Non polyadenylated RNA species are lost during the selection of poly-A mRNA. In contrast, rRNA depletion methods such as RiboMinus could preserve non polyadenylated RNA species. The purpose of our study is to look for differentially expressed coding genes in two cell types, thus we used the poly-A mRNA selection method for enrichment of target RNAs before library construction. When library construction was finished, the size of DNA fragments in the library was checked before sequencing using Bioanalyzer. **Figure 4** shows a good quality library with the fragment size peaks at about 300 bp.

In the subsequent step, the library was subjected to high-throughput sequencing. In principle, longer read length will be helpful for read mapping. It can reduce the probability that the read is mapped to multiple locations due to similarity among duplicate genes or gene family members.

As the pair-end sequencing sequences are from both ends of the fragments, the read length chosen should be less than half of the average fragments length. If the main goal of the experiment is to measure the expression level instead of constructing transcript structure, single-end read (75 or 100 bp) can reduce the cost without losing too much information. Paired-end sequencing is more useful for transcript structure construction and shorter read length can be used to reduce cost. Certainly, when sufficient funding is available, longer read length is preferred.

For differential expression analysis, there are many alternative algorithms other than DESeq. There is also one included in cufflinks package named cuffdiff[32]. DESeq is one of the most widely used count based DE gene analysis algorithms. DESeq method is based on a well characterized statistics model — negative binomial distribution. In our experience, DESeq is more stable compare to cuffdiff. Early versions of cuffdiff often give significantly different numbers of DE genes. Therefore we used DESeq for DE analysis here.

Because transcription factors are crucial for cell fate determination, we focused on the significantly differentially expressed transcription factors[33]. The TFs changed >1.5 fold between Lin-CD34+ and Lin-CD34- were found and are shown on the heatmap (**Figure 5**)[2]. Notably, the relative expression level of *Tcf7* in Lin-CD34+ cells is more than 100 fold higher than that in Lin-CD34- cells. Thus *Tcf7* was chosen for further ChIP-Sequencing (Chromatin Immunoprecipitation and sequencing) analysis and functional test to confirm *Tcf7*'s function in regulation of EML cell self-renewal and differentiation[2].

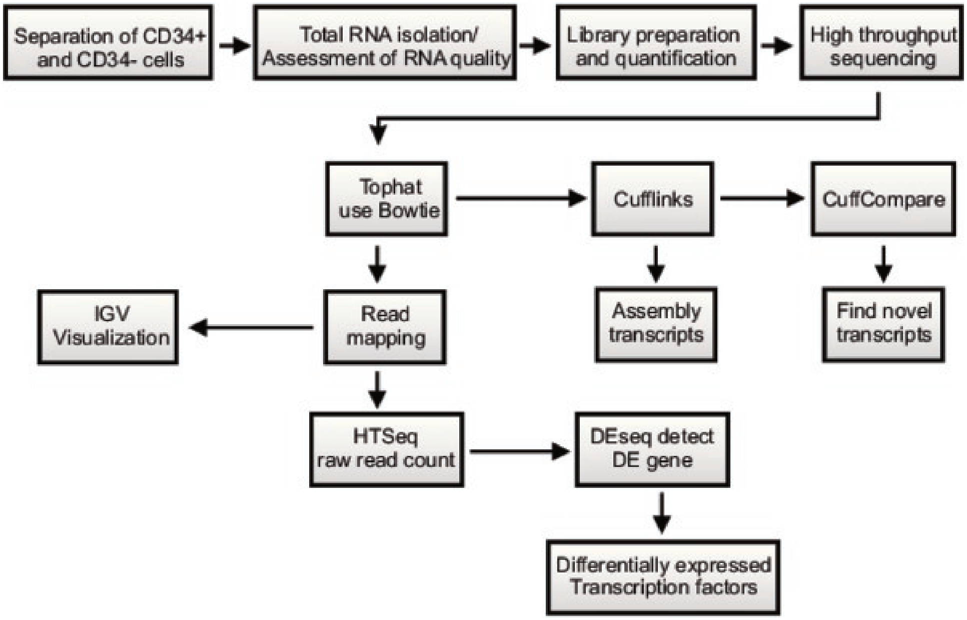

Separation of CD34+ and CD34- cells
Total RNA isolation/ Assessment of RNA quality
Library preparation and quantification
High throughput sequencing
Tophat use Bowtie
Cufflinks
CuffCompare
IGV Visualization
Read mapping
Assembly transcripts
Find novel transcripts
HTSeq raw read count
DEseq detect DE gene
Differentially expressed Transcription factors

Figure 1: Workflow of the procedures. Lin-CD34+ and Lin-CD34- cells were separated by magnetic cell separation system and fluorescence-activated cell sorting method. Total RNA was extracted followed by mRNA purification and library construction. After analysis of library quality, samples were subjected to high throughput sequencing. Data were analyzed and differentially expressed transcription factors were identified.

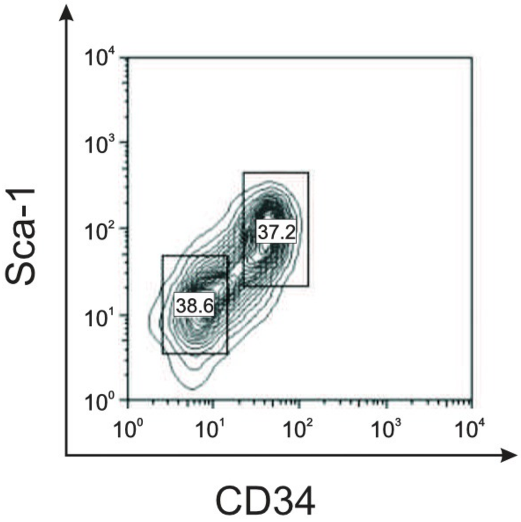


Figure 2: Separation of Lin-CD34+ and Lin-CD34- EML cells[6]. Lin- EML cells were enriched by magnetic cell sorting. Lin- cells were stained with anti-CD34, anti-Sca1 and lineage mixture antibodies. Lin- cells were gated for expression of CD34 and Sca1. Lin-CD34+SCA+ and Lin-CD34-SCA- EML cell populations were sorted.

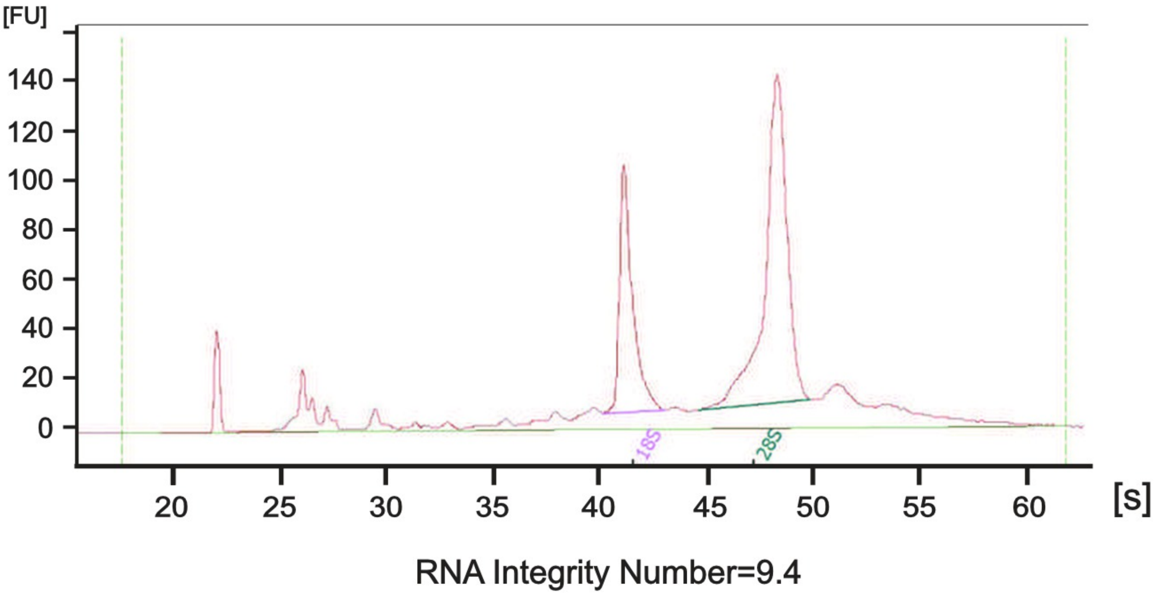


Figure 3: A representative of high-quality total RNA sample. The quality of total RNA was assessed by Bioanalyzer. The RNA Integrity Number is 9.4 (FU, Fluorescence Units).

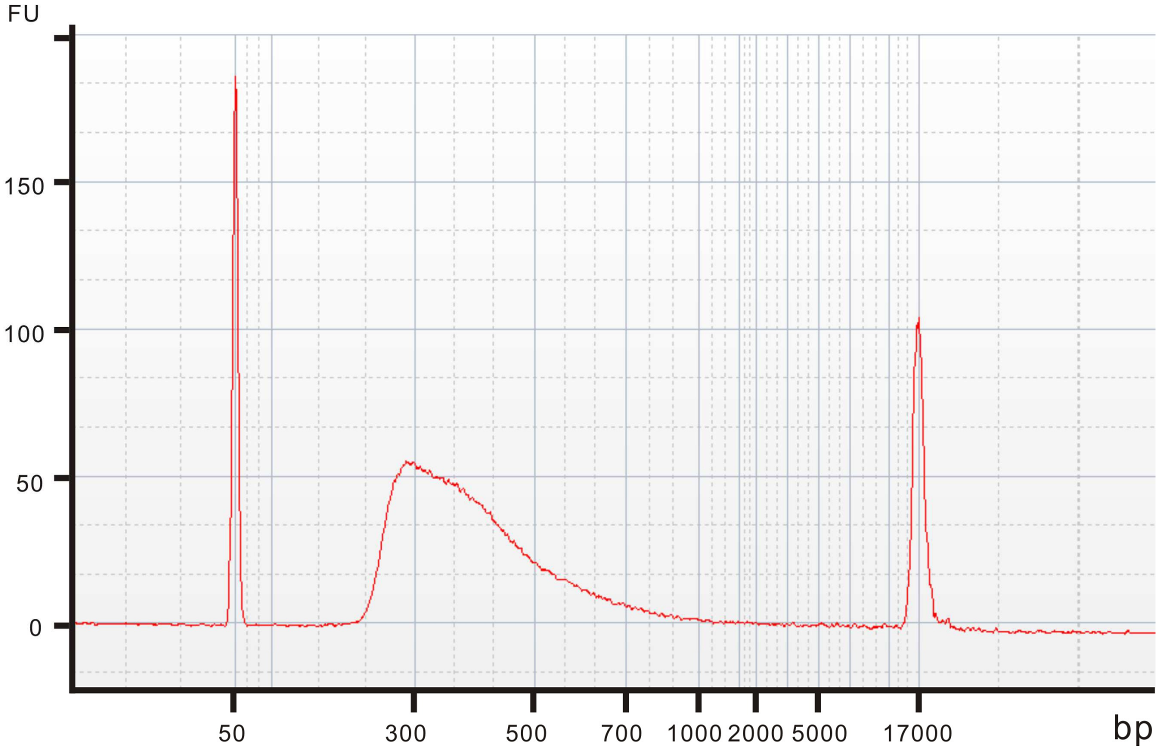


Figure 4: Fragments size range of Paired-End library. The DNA size distribution of the library was analyzed using Bioanalyzer. Most fragments are within the size range of 250-500 bp.

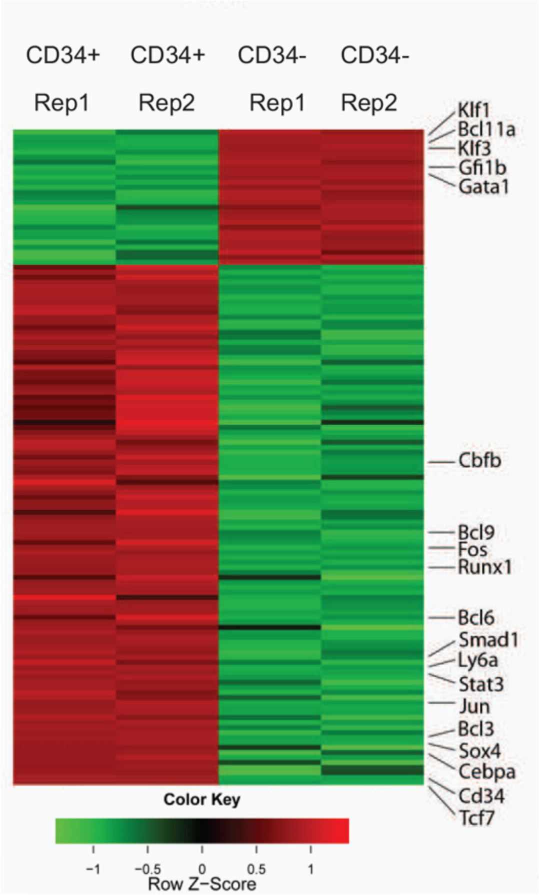


Figure 5: Differentially expressed transcription factors (>1.5 fold) between Lin-CD34+ cells and Lin-CD34- cells[2]. For each cell type, two independent experiments were performed. Up-regulated genes are indicated as red color and down-regulated genes are indicated as green color.

BHK medium	
100x Antibiotic-Antimycotic	10 ml
200 mM L-Glutamine	10 ml
FBS	100 ml
DMEM	880 ml
Total volume	1,000 ml
EML Basic medium	
IMDM	390 ml
HI horse serum	100 ml
100 x Penicillin-Streptomycin	5 ml
200 mM L-Glutamine	5 ml
BHK medium	75 ml
Total Volume	575 ml
Filtrate through 0.45 µM filter	
FACS buffer	
BSA	0.50%
EDTA	1 mM
Dissolved in PBS and filtrate through 0.45 µM filter	

Table 1: Buffers and Cell culture mediums.

Software	Usage	Reference
Bowtie 1.2.7	Used by Tophat for mapping	[28]
Tophat 1.3.3	Mapping reads back to reference genome	[27]
Cufflinks 1.3.0	Transcripts construction and expression level estimation	[29]
DESeq 1.16.0	Differential expression analysis	[30]
Bedtools 2.18	Convert .bam file into .bed file	[31]
bedGraphToBigWig	Convert .bed file to .bigwig file	http://genome.ucsc.edu/

Table 2: List of software for data analysis.

Discussion

Mammalian transcriptome is very complex[34-38]. RNA-Seq technology plays an increasingly important role in the studies of transcriptome analysis, novel transcripts detection and single nucleotide variation discovery *etc.* It has many advantages over other methods for gene expression analysis. As mentioned in the introduction, it overcomes the hybridization artifacts of microarray and can be used to identify novel transcripts *de novo*. One limitation of RNA-sequencing is relative short read length comparing to Sanger sequencing. However, with the rapid improvement of sequencing technology, read length is increasing constantly. In this paper, we provide detailed methods of using this technology to identify potential key regulators in mouse EML cell self-renewal and differentiation.

The first key step for this protocol is EML cell culture. Although EML is a hematopoietic precursor cell line and it can be propagated in large quantity with SCF. The culturing condition of EML cells requires more attention than the usual immortalized cell lines. The cells should be fed and passaged at a regular basis with gentle operation; otherwise the cells could change in their properties of self-renewal and differentiation and undergo cell death. As the first step after collecting enough cells, we isolated lineage negative cells using a magnetic activated cell sorting system. Then we separated CD34+ and CD34- cells using fluorescence-activated cell sorting. The EML cells are normally passaged less than 10 generations before using for RNA extraction and the numbers of CD34+ and CD34- cells should be similar after separation. If the two populations vary greatly in cell number, it is advisable to discard the culture and re-thaw another tube of cell stock for culture.

After separation of CD34+ and CD34- cell, total RNA extraction was performed, another important step for this study. High quality RNA is the base for construction of a high quality library, which promises the accuracy of the sequencing data. In this critical step, any contact with RNase should be avoided. All reagents should be RNase free. It is important to wear gloves at all times while handling RNA. High quality RNA sample has an OD 260/280 value between 1.8 and 2.0. When collecting the aqueous phase containing RNA, be careful not to carry any organic phase with the RNA sample. Any residual organic solvents such as phenol or chloroform in the RNA would result in an OD260/280 value lower than 1.65. If the OD260/280 value is lower than 1.65, precipitate RNA again with ethanol. After washing with 75% ethanol, do not overdry RNA pellet. Drying RNA pellet completely will affect the solubility of RNA and lead to low yield of RNA.

The next key step for this protocol is library preparation. After total RNA extraction, a step of using DNase for removal of contaminated DNA is highly recommended, since DNA contamination might result in wrong estimation of the amount of total RNA used. It is recommended to perform the downstream procedure immediately after RNA isolation, since after long-term storage and freeze-thawing procedure, RNA will degrade to some degree. If the subsequent steps after RNA isolation can not be performed immediately, store the RNA in -80 °C. Before total RNA is used for mRNA purification and cDNA synthesis, the quality should always be checked. Only high quality RNA can be used for library preparation. Using low quality or degraded RNA might lead to over-representation of 3' ends. Before sequencing, library quality was assessed to ensure maximum sequencing efficiency.

In the data analysis part, after performing a run of Cufflinks without a reference transcriptome, we combined the novel transcripts with known transcripts to form a reference .gtf file and run Tophat and cufflinks for the second time. This two-run procedure is recommended, since this provide more accurate FPKM estimation than running only once. After data analysis, the differentially expressed genes were identified. Downstream experiments can be performed to validate the function of genes *in vitro* and *in vivo*. In our previous publication[2], we chose the significantly differentially expressed transcription factors and identified the genome binding site of these factors by performing chromatin immunoprecipitation and sequencing (ChIP-Seq). In addition, we applied shRNA knockdown assay to test the functional effect of *Tcf7*. We found that in *Tcf7* knockdown cells, up-regulated genes were the genes highly enriched in CD34- cells, while down-regulated genes were found to be significantly enriched in CD34+ cells. Therefore, the gene expression profile of *Tcf7* knockdown cells shifted toward a partially differentiated CD34- state.Overall, using EML cell as a model system coupled with RNA-Sequencing technology and functional assays, we identified and confirmed *Tcf7* as an important regulator of EML cell self-renewal and differentiation.

Disclosures

The authors have nothing to disclose.

Acknowledgements

JQW, SZ, SD and KC are supported by grant from the National Institutes of Health and the Staman Ogilvie Fund—Memorial Hermann Foundation.

References

1. Chambers, S. M., & Goodell, M. A. Hematopoietic stem cell aging: wrinkles in stem cell potential. *Stem Cell Rev.* **3**, 201-211, doi:SCR:3:3:201 [pii] (2007).
2. Wu, J. Q. *et al.* Tcf7 is an important regulator of the switch of self-renewal and differentiation in a multipotential hematopoietic cell line. *PLoS genetics.* **8**, e1002565, doi:10.1371/journal.pgen.1002565 (2012).
3. Ye, Z. J. *et al.* Complex interactions in EML cell stimulation by stem cell factor and IL-3. *Proceedings of the National Academy of Sciences of the United States of America.* **108**, 4882-4887, doi:10.1073/pnas.1018002108 (2011).
4. Tsai, S., Bartelmez, S., Sitnicka, E., & Collins, S. Lymphohematopoietic progenitors immortalized by a retroviral vector harboring a dominant-negative retinoic acid receptor can recapitulate lymphoid, myeloid, and erythroid development. *Genes Dev.* **8**, 2831-2841 (1994).
5. Weiler, S. R. *et al.* D3: a gene induced during myeloid cell differentiation of Linlo c-Kit+ Sca-1(+) progenitor cells. *Blood.* **93**, 527-536 (1999).
6. Ye, Z. J., Kluger, Y., Lian, Z., & Weissman, S. M. Two types of precursor cells in a multipotential hematopoietic cell line. *Proc Natl Acad Sci U S A.* **102**, 18461-18466, doi:0509314102 [pii] 10.1073/pnas.0509314102 (2005).
7. Wang, Z., Gerstein, M., & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics.* **10**, 57-63, doi:10.1038/nrg2484 (2009).
8. Chu, Y., & Corey, D. R. RNA sequencing: platform selection, experimental design, and data interpretation. *Nucleic acid therapeutics.* **22**, 271-274, doi:10.1089/nat.2012.0367 (2012).
9. Hornett, E. A., & Wheat, C. W. Quantitative RNA-Seq analysis in non-model species: assessing transcriptome assemblies as a scaffold and the utility of evolutionary divergent genomic reference species. *BMC genomics.* **13**, 361, doi:10.1186/1471-2164-13-361 (2012).
10. Eswaran, J. *et al.* RNA sequencing of cancer reveals novel splicing alterations. *Scientific reports.* **3**, 1689, doi:10.1038/srep01689 (2013).
11. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature.* **456**, 470-476, doi:10.1038/nature07509 (2008).
12. Wu, J. Q. *et al.* Dynamic transcriptomes during neural differentiation of human embryonic stem cells revealed by short, long, and paired-end sequencing. *Proceedings of the National Academy of Sciences of the United States of America.* **107**, 5254-5259, doi:10.1073/pnas.0914114107 (2010).
13. Loraine, A. E., McCormick, S., Estrada, A., Patel, K., & Qin, P. RNA-seq of Arabidopsis pollen uncovers novel transcription and alternative splicing. *Plant physiology.* **162**, 1092-1109, doi:10.1104/pp.112.211441 (2013).
14. Edgren, H. *et al.* Identification of fusion genes in breast cancer by paired-end RNA-sequencing. *Genome biology.* **12**, R6, doi:10.1186/gb-2011-12-1-r6 (2011).
15. Ilott, N. E., & Ponting, C. P. Predicting long non-coding RNAs using RNA sequencing. *Methods.* **63**, 50-59, doi:10.1016/j.ymeth.2013.03.019 (2013).
16. Sun, L. *et al.* Prediction of novel long non-coding RNAs based on RNA-Seq data of mouse Klf1 knockout study. *BMC bioinformatics.* **13**, 331, doi:10.1186/1471-2105-13-331 (2012).
17. Luo, S. MicroRNA expression analysis using the Illumina microRNA-Seq Platform. *Methods in molecular biology.* **822**, 183-188, doi:10.1007/978-1-61779-427-8_12 (2012).
18. Bolduc, F., Hoareau, C., St-Pierre, P., & Perreault, J. P. In-depth sequencing of the siRNAs associated with peach latent mosaic viroid infection. *BMC molecular biology.* **11**, 16, doi:10.1186/1471-2199-11-16 (2010).
19. Chepelev, I., Wei, G., Tang, Q., & Zhao, K. Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. *Nucleic acids research.* **37**, e106, doi:10.1093/nar/gkp507 (2009).
20. Djari, A. *et al.* Gene-based single nucleotide polymorphism discovery in bovine muscle using next-generation transcriptomic sequencing. *BMC genomics.* **14**, 307, doi:10.1186/1471-2164-14-307 (2013).
21. Murphy, D. Gene expression studies using microarrays: principles, problems, and prospects. *Advances in physiology education.* **26**, 256-270 (2002).
22. Chen, K. *et al.* RNA-seq characterization of spinal cord injury transcriptome in acute/subacute phases: a resource for understanding the pathology at the systems level. *PloS one* **8**, e72567, doi:10.1371/journal.pone.0072567 (2013).
23. Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., & Gilad, Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research.* **18**, 1509-1517, doi:10.1101/gr.079558.108 (2008).
24. Ramskold, D., Kavak, E., & Sandberg, R. How to analyze gene expression using RNA-sequencing data. *Methods in molecular biology.* **802**, 259-274, doi:10.1007/978-1-61779-400-1_17 (2012).
25. Glenn, T. C. Field guide to next-generation DNA sequencers. *Mol Ecol Resour.* **11**, 759-769, doi:10.1111/j.1755-0998.2011.03024.x (2011).
26. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols.* **7**, 562-578, doi:10.1038/nprot.2012.016 (2012).
27. Trapnell, C., Pachter, L., & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics.* **25**, 1105-1111, doi:10.1093/bioinformatics/btp120 (2009).
28. Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology.* **10**, R25, doi:10.1186/gb-2009-10-3-r25 (2009).
29. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology.* **28**, 511-515, doi:10.1038/nbt.1621 (2010).
30. Anders, S., & Huber, W. Differential expression analysis for sequence count data. *Genome biology.* **11**, R106, doi:10.1186/gb-2010-11-10-r106 (2010).
31. Quinlan, A. R., & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* **26**, 841-842, doi:10.1093/bioinformatics/btq033 btq033 [pii] (2010).
32. Cheranova, D. *et al.* RNA-seq analysis of transcriptomes in thrombin-treated and control human pulmonary microvascular endothelial cells. *J Vis Exp.* doi:10.3791/4393 4393 [pii] (2013).
33. Zhang, H. M. *et al.* AnimalTFDB: a comprehensive animal transcription factor database. *Nucleic acids research.* **40**, D144-149, doi:10.1093/nar/gkr965 (2012).
34. Wu, J. Q. *et al.* Systematic analysis of transcribed loci in ENCODE regions using RACE sequencing reveals extensive transcription in the human genome. *Genome Biol.* **9**, R3, doi:10.1186/gb-2008-9-1-r3gb-2008-9-1-r3 [pii] (2008).

35. Wu, J. Q. *et al.* Large-scale RT-PCR recovery of full-length cDNA clones. *Biotechniques.* **36**, 690-696, 698-700 (2004).
36. Wu, J. Q., Shteynberg, D., Arumugam, M., Gibbs, R. A., & Brent, M. R. Identification of rat genes by TWINSCAN gene prediction, RT-PCR, and direct sequencing. *Genome Res.* **14**, 665-671, doi:10.1101/gr.195960414/4/665 [pii] (2004).
37. Dewey, C. *et al.* Accurate identification of novel human genes through simultaneous gene prediction in human, mouse, and rat. *Genome Res.* **14**, 661-664, doi:10.1101/gr.193980414/4/661 [pii] (2004).
38. Wu, J. *Characterize Mammalian Transcriptome Complexity.* LAP Lambert Academic Publishing (2011).