Video Article

RNA-seq Analysis of Transcriptomes in Thrombin-treated and Control Human Pulmonary Microvascular Endothelial Cells

Dilyara Cheranova[1], Margaret Gibson[1], Suman Chaudhary[1], Li Qin Zhang[1], Daniel P. Heruth[1], Dmitry N. Grigoryev[1], Shui Qing Ye[1]

[1]Children's Mercy Hospital and Clinics, School of Medicine, University of Missouri-Kansas City

Correspondence to: Shui Qing Ye at sqye@cmh.edu

URL: https://www.jove.com/video/4393
DOI: doi:10.3791/4393

Keywords: Genetics, Issue 72, Molecular Biology, Immunology, Medicine, Genomics, Proteins, RNA-seq, Next Generation DNA Sequencing, Transcriptome, Transcription, Thrombin, Endothelial cells, high-throughput, DNA, genomic DNA, RT-PCR, PCR

Date Published: 2/13/2013

Citation: Cheranova, D., Gibson, M., Chaudhary, S., Zhang, L.Q., Heruth, D.P., Grigoryev, D.N., Qing Ye, S. RNA-seq Analysis of Transcriptomes in Thrombin-treated and Control Human Pulmonary Microvascular Endothelial Cells. *J. Vis. Exp.* (72), e4393, doi:10.3791/4393 (2013).

Abstract

The characterization of gene expression in cells via measurement of mRNA levels is a useful tool in determining how the transcriptional machinery of the cell is affected by external signals (*e.g.* drug treatment), or how cells differ between a healthy state and a diseased state. With the advent and continuous refinement of next-generation DNA sequencing technology, RNA-sequencing (RNA-seq) has become an increasingly popular method of transcriptome analysis to catalog all species of transcripts, to determine the transcriptional structure of all expressed genes and to quantify the changing expression levels of the total set of transcripts in a given cell, tissue or organism[1,2] . RNA-seq is gradually replacing DNA microarrays as a preferred method for transcriptome analysis because it has the advantages of profiling a complete transcriptome, providing a digital type datum (copy number of any transcript) and not relying on any known genomic sequence[3].

Here, we present a complete and detailed protocol to apply RNA-seq to profile transcriptomes in human pulmonary microvascular endothelial cells with or without thrombin treatment. This protocol is based on our recent published study entitled "RNA-seq Reveals Novel Transcriptome of Genes and Their Isoforms in Human Pulmonary Microvascular Endothelial Cells Treated with Thrombin,"[4] in which we successfully performed the first complete transcriptome analysis of human pulmonary microvascular endothelial cells treated with thrombin using RNA-seq. It yielded unprecedented resources for further experimentation to gain insights into molecular mechanisms underlying thrombin-mediated endothelial dysfunction in the pathogenesis of inflammatory conditions, cancer, diabetes, and coronary heart disease, and provides potential new leads for therapeutic targets to those diseases.

The descriptive text of this protocol is divided into four parts. The first part describes the treatment of human pulmonary microvascular endothelial cells with thrombin and RNA isolation, quality analysis and quantification. The second part describes library construction and sequencing. The third part describes the data analysis. The fourth part describes an RT-PCR validation assay. Representative results of several key steps are displayed. Useful tips or precautions to boost success in key steps are provided in the Discussion section. Although this protocol uses human pulmonary microvascular endothelial cells treated with thrombin, it can be generalized to profile transcriptomes in both mammalian and non-mammalian cells and in tissues treated with different stimuli or inhibitors, or to compare transcriptomes in cells or tissues between a healthy state and a disease state.

Video Link

The video component of this article can be found at https://www.jove.com/video/4393/

Protocol

A flowchart outlining this protocol is displayed in Figure 1.

1. Treatment of Cells with Thrombin, RNA Isolation, Quality Assessment and Quantification of RNA

1. Culture Human Lung Microvascular Endothelial Cells (HMVEC-LBl) to 90-100% confluence in 6-well plates in EGM-2 medium with 5% FBS, growth factors and antibiotics (Lonza, cat# CC-3202).
2. Change media to the starvation media (0% FBS) 30 min prior to treatment with thrombin.
3. Treat the cells with 0.05 U/ml thrombin or leave untreated as a control for 6 hr at 37 °C and 5% CO_2.
4. Isolate total RNA from the treated and control cells using the Ambion *mir*Vana kit according to manufacturer's instructions.
5. Assess the quality of the RNA with an Experion StdSense Eukaryotic RNA chip according to the standard protocol on the Experion Automated Electrophoresis Station (www.bio-rad.com).
6. Quantify the RNA using a standard spectrophotometric method.

2. Library Construction and Sequencing

1. Use 1 µg of high quality total RNA per sample as starting material.
2. To construct the library, follow the standard procedure from Illumina (protocol # 15008136 Rev. A). In this protocol, two rounds of poly(A) containing mRNA selections are performed to remove rRNA to minimize the rRNA sequencing.
3. Assess the quality of the libraries using an Experion DNA 1K chip according to the standard protocol on the Experion Automated Electrophoresis Station (www.bio-rad.com).
4. Quantify the library using qPCR: Use a library that has previously been sequenced as a standard curve and primers specific for the ligated adapters. Use a range of dilutions of the unknown libraries (*i.e.* 1:100, 1:500 and 1:1,000). Run the qPCR according to the SyberGreen MM protocol and calculate the original stock concentration of each library.
5. Dilute the library stocks to 10 nM and store at -20 °C until ready to cluster a flow cell.
6. When ready to cluster a flow cell, thaw the cBot reagent plate in a water bath. cBot is an Illumina instrument used to streamline the cluster generation process.
7. Wash the cBot instrument.
8. Denature the libraries: Combine 13 µl 1x TE and 6 µl 10 µM library and, to the side of the tube, add 1 µl 1 N NaOH (provided by Illumina). Vortex, spin down, incubate at room temperature for 5 min and place the denatured libraries on ice.
9. Dilute the libraries: Dilute the denatured libraries with pre-chilled hybridization buffer (HT1, provided by Illumina) by combining 996 µl HT1 and 4 µl denatured library for a final concentration of 12 pM. Place the denatured and diluted libraries on ice.
10. Invert each row of tubes of the cBot plate, ensuring that all the reagents are thawed. Spin down the plate, remove/puncture the foil seals and load onto the cBot.
11. Aliquot 120 µl of the diluted, denatured libraries to a strip tube, labeled 1-8. Add 1.2 µl diluted, denatured PhiX control library (from Illumina) into each tube as a spike-in control. Vortex and spin down the tubes and load them on the cBot in the correct orientation (tube #1 to the right).
12. Load a flow cell and manifold onto the cBot.
13. Complete the flow check and begin the clustering run.
14. After the run is complete, check reagent delivery across all lanes. Make note of any abnormalities. Either start the sequencing run immediately or store the flow cell in the provided tube at 4 °C.
15. Thaw the sequencing-by-synthesis (SBS, Illumina) reagents.
16. Load the reagents to the appropriate spots on the reagent trays, making sure not to touch the other reagents after touching the cleavage mix.
17. Using a non-sequencing flow cell (*i.e.* one that was sequenced previously), prime the reagent lines twice.
18. Thoroughly clean the sequencing flow cell with 70% EtOH and Kimwipes, followed by 70% EtOH and lens paper. Inspect the flow cell for any streaks. Re-clean it if necessary.
19. Load the flow cell onto the sequencer and perform a flow check to ensure that the seal between the manifolds and the flow cell is tight.
20. Start the sequencing run.
21. Assess the quality metrics (*e.g.* the cluster density, clusters passing filter, Q30, intensity) as they become available during the run.
22. Monitor intensity throughout the run.
23. After 101 cycles are completed, perform turnaround chemistry to complete the second read: Thaw the paired end reagents and the second read Incorporation Buffer (ICB, a component of the SBS reagents, Illumina) and load the reagents.
24. Continue the sequencing run, assessing 2nd read intensity, Q30 and other quality metrics as the run progresses.

3. Data Analysis

1. Use the latest version of CASAVA (Illumina, currently 1.8.2) to convert the base call files (.bcl) files to .fastq files, setting fastq-cluster-count to 0 to ensure the creation of a single fastq file for each sample. Unzip the fastq files for downstream analysis.
2. Perform paired end alignment using the latest versions of TopHat (1.4.1)[5], which aligns RNA-seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner(Bowtie,0.12.7)[6] and SAMtools (0.1.17)[7]. SAMtools implements various utilities for post-processing alignments in the SAM format. The reference human transcriptome can be downloaded from iGenomes (www.illumina.com). In running TopHat, we used all default parameter settings including the library type option as fr-unstranded (default).
3. Using the program CuffDiff, part of the CuffLinks (1.3.0)[8] software package, compare the thrombin-treated cells to the controls cells to screen out the differentially expressed gene transcripts in the former based on the human reference transcriptome. This comparison detects the differential expression of known transcripts. Use Microsoft Excel to visualize the result in table form. In running Cufflinks program, we used all default parameter settings. Those gene transcripts with FPKM<0.05 and p>0.05 are filtered out.
4. To detect novel isoforms, run Cufflinks without a reference transcriptome. Compare the sample transcript files to the reference genome using Cuffcompare and test the differential expression with Cuffdiff using the combined thrombin transcript files as the reference genome for one analysis and the combined control transcript files as the reference genome for a second analysis. Use Microsoft Excel to visualize the result in tabular format. Again, those gene transcripts with FPKM<0.05 and p>0.05 are filtered out. After this step, investigators may opt to upload a list of newly reported transcripts to the UCSC Genome Browser website (http://genome.ucsc.edu/) to verify their validity by a manual inspection.
5. Submit lists of differentially expressed genes to Ingenuity Pathway Analysis (IPA, www.ingenuity.com) for characterization of the genes and pathways affected by the thrombin treatment. In this step, investigators may opt to use CummeRbund (http://compbio.mit.edu/cummeRbund/), an R package that is designed to aid and simplify the task of analyzing Cufflinks RNA-seq output, to help manage, visualize and integrate all of data produced by a Cuffdiff analysis.

4. Validation of the RNA-seq Results by Quantitative Real-time-Polymerase Chain Reaction (qRT-PCR)

1. Perform total RNA isolation from control and thrombin-treated HMVEC-LBI cells, RNA quality assessment and RNA quantification described in Steps 1.4 to 1.6.

2. Generate complementary DNA from 1 µg total RNA of each sample with SuperScript III First-Strand Synthesis System RT Kits, following the manufacturer's instructions (Invitrogen, 18080-051).
3. Perform qRT-PCR analysis on a Applied Biosystems ViiA 7 Real-Time PCR System using Taqman Assay-on-Demand designed oligonucleotides for the detection of CUGBP, Elav-likefamilymember1(CELF1, Hs00198069_m1), Fanconianemia,complementationgroupD2 (FANDCD2, Hs00276992_m1), TNFreceptor-associated factor 1(TRAF1, Hs01090170_m1),and β-actin (ACTB,Hs99999903_m1). Each sample had a template equivalent to 5 ng of total RNA. Measure quantitation using the DDCt method and normalize to β-actin. Each assay was performed across at least three biological replicates.

Representative Results

For Step 1: The 28s:18s ratio is traditionally used as an indicator of RNA degradation. Ideally, the 28s peak should have approximately twice the area of the 18s band (a ratio of 2), however this ideal ratio is often not seen in practice. Furthermore, 28s:18s ratios obtained from spectrophotometric methods can underestimate the amount of degradation of the RNA. To more accurately quantify the degradation, and therefore the quality of the RNA sample, the Experion system calculates an RNA Quality Indicator (RQI) number. The RQI algorithm compares the electropherogram of RNA samples to data from a series of standardized, degraded RNA samples and automatically returns a number between 10 (intact RNA) and 1 (degraded RNA). The RNA quality should have an RQI of at least 7, ideally greater than 8. **Figure 2** shows the Experion results using a high-quality RNA sample with an RQI of 8.4.

For Step 2: The libraries should have a broad band at approximately 250-300 bp. **Figure 3** shows Experion results of a high-quality library. **Figure 4** shows the qPCR results of standard curve samples and one unknown sample (shown in dark blue). The progress and quality of the sequencing run should be constantly observed throughout the run. **Figure 5** shows appropriate cluster density during the first cycle imaging step; this is the first indication of the run quality. Clusters should be bright and focused. **Figure 6** shows the First Base Report generated after the first cycle is complete. It is important to assess the estimated cluster density, intensity levels, and focus quality at this point. The next quality checkpoint, after cycle 4, is shown in **Figure 7**. This shows the absolute cluster density for each lane. The cluster density should not be above 850 k/mm^2. After cycle 13, phasing (when a base is not added during a cycle) and prephasing (when two bases are added during a cycle) stats are calculated, as shown in **Figure 8**. Typical numbers are between 0.1 and 0.25. The major quality assessment is possible after cycle 24, when several quality metrics are calculated. The percent of reads above Q30, shown in **Figure 9**, is a measure of the confidence in the base calling. A read with a Q score of 30 means that there is a 1 in 1,000 chance that base call is wrong. The Q scores will decrease as the run progresses, but should start out with greater than 95% of the reads meeting or exceeding Q30. The clusters passing filter (PF), shown in **Figure 10**, are the clusters from which the actual sequence data will be taken. Ideally, this should be above 85%. The cluster PF is based on many factors, including phasing, prephasing, intensity and Q30. It will not change as the run progresses. The percent aligned (**Figure 11**) is a measure of the reads that align real-time to the PhiX genome. Since we spiked in approximately 1% PhiX library to the sample libraries, the percent aligned should be between 0.5 and 1. This statistic shows that the library content is represented well by the clusters and there was no cluster generation bias.

For Step 3: Table 1 presents expressed genes and isoforms in both control and thrombin-treated human pulmonary microvascular endothelial cells. Notably, there are about 26,000 novel isoforms detected, which illustrates the strength of RNA-seq—it can identify unknown RNAs, alternatively spliced transcripts and alternative promoter usage which are not detectable by microarray techniques[3]. RNA-seq can also measure the less abundant transcripts that are inaccurately quantified or not detected by microarrays. **Figure 12** and **Table 2** display differentially expressed genes in the thrombin signaling pathway. This is an example of the third generation knowledge base-driven pathway analysis[9]: pathway topology based approaches, using Ingenuity Pathway Analysis software. It shows that a six h thrombin treatment significantly up-regulates the thrombin receptor Par 4 and down-regulates the thrombin receptor Par 3 while there is no change in the expression of the thrombin receptor Par 1. Expression of NFkB1, NF kB2 and Src are also significantly up-regulated. Some of Rho family genes (Rho B, C, F and G) are up-regulated while others (Rho J, Q, T1, U and V) are down-regulated (**Table 4**). Myosin light chain gene 9 (MYL9) is up-regulated while MYL12B is down-regulated. Expression of other genes is either down-regulated or not affected.

For Step 4: To validate the RNA-seq results using an alternative approach, we performed a qRT-PCR experiment to assay three different genes (**Figure 13**). In RNA-seq data, TRAF1 was up-regulated by 7.96 fold; CELF1 was down-regulated by 1.16 fold; and FANCD2 was down-regulated by 1.70 fold. In qRT-PCR data, these corresponding numbers are +7.25 fold, -1.15 fold and -2.07 fold, respectively. The results of these three genes assayed by RNA-seq and qRT-PCR are in good agreement, which corroborates the RNA-seq results.

Genes		
	Control	**Thrombin**
Total Genes Expressed	16,636	16,357
Control Only	783	
Thrombin Only		504
Up-regulated (2-fold or greater difference)		152
Down-regulated (2-fold or greater difference)		2,190
Known Isoforms		
	Control	**Thrombin**
Total Known Isoforms Expressed	26,807	26,300
Control Only	1,492	
Thrombin Only		985
Up-regulated (2-fold or greater difference)		480

Down-regulated (2-fold or greater difference)		3,574
Novel Isoforms		
	Control	**Thrombin**
Total Novel Isoforms Expressed	25,880	25,886
Control Only	418	
Thrombin Only		424
Up-regulated (2-fold or greater difference)		1,775
Down-regulated (2-fold or greater difference)		12,202

Table 1. Gene/Isoform Expression Summary*. This table lists genes, known isoforms and novel isoforms expressed in control and thrombin-treated HMVEC-LBI cells. The fold change is the ratio of thrombin FragmentsPerKilobase of transcript perMillion fragments mapped (FPKM) to control FPKM. Those genes, known isoforms and novel isoforms with 2-fold or greater difference between two groups have statistically different expression levels (as determined after Benjamini-Hochberg correction). * This table is reproduced from **Table 2** in Reference 4.

Down-Regulated Genes						
Gene Symbol	**Overall Fold Change**	**Members**	**Individual Fold Change**	**q_value****	**FPKM Control**	**FPKM Thrombin**
PLC	-2.49					
		PLCB1	-2.97	0	6.75585	2.27611
		PLCB2	-1.32	0.00183634	1.80892	1.36957
		PLCB4	-10.04	6.28386E-14	0.682668	0.0679837
		PLCL1	-2.53	5.26728E-09	0.602879	0.238017
Gaq	-1.87					
		GNAQ	-1.87	0	24.0907	12.8996
Gai	-1.56					
		GNAI1	-1.86	0	9.88417	5.31795
		GNAI3	-1.49	0	35.7592	24.0198
PKC			-2.15			
		PRKCE	-1.65	0	9.36364	5.67305
		PRKCI	-2.14	0	6.63566	3.09818
		PRKD3	-2.61	0	16.0215	6.12878
IP3R	-2.60					
		ITPR1	-1.39	0.000018734	1.51592	1.09009
		ITPR2	-3.19	0	7.23283	2.26944
CREB	-2.36					
		CREB1	-2.36	0	5.19117	2.2004
GATA	-2.33					
		GATA3	-2.33	0.0228108	0.716773	0.307131
TBP	-1.35					
		TBP	-1.35	2.66E-05	8.48689	6.30476
G-protein Alpha	-1.64					
		GNA14	-1.49	0.000122496	2.78076	1.86573
		GNAI1	-1.86	0	9.88417	5.31795
		GNAI3	-1.49	0	35.7592	24.0198
		GNAQ	-1.87	0	24.0907	12.8996
PAR3	-1.87					
		F2RL2	-1.87	1.02474E-06	1.51286	0.810286
SOS1						

		SOS1	-2.27	0	8.94353	3.94679
Ras	-1.69					
		KRAS	-2.03	0	11.2766	5.5659
		NRAS	-1.61	0	38.0874	23.6491
AKT	-1.77					
		AKT3	-1.77	0	15.8228	8.94223
MLCP	-2.38					
		PPP1CB	-2.10	0	39.585	18.8199
		PPP1R12A	-3.56	0	15.2274	4.27516
		PPP1R12B	-2.66	5.28466E-13	1.92123	0.722188
FAK	-1.31					
		PTK2	-1.31	4.3856E-10	39.7935	30.3044
p70 S6K						
		RPS6KB1	-1.99	0	8.46312	4.25129
ROCK	-3.68					
		ROCK1	-3.54	0	19.5921	5.53112
		ROCK2	-3.86	0	16.0054	4.14759
CAMK	-1.17					
		CAMK1	1.30	0.000255587	7.90794	10.296
		CAMK2D	-1.74	2.22911E-12	11.4497	6.56804
		CAMK4	-2.58	0.000619045	0.62714	0.243277
Up-Regulated Genes						
Symbol	**Overall Fold Change**	**Members**	**Individual Fold Change**	**q_value****	**FPKM Control**	**FPKM Thrombin**
PAR4	1.59					
		F2RL3	1.59	9.19043E-13	4.99867	7.9253
Src	1.47					
		Src	1.47	0	14.1085	20.675
NF-kB	1.65					
		NFKB1	1.66	0	19.5113	32.3457
		NFKB2	2.02	0	28.7563	58.1293
		RELA	1.45	0	55.3482	80.28
Partial Up-/Partial Down-Regulated Genes						
Symbol	**Overall Fold Change**	**Members**	**Individual Fold Change**	**q_value****	**FPKM Control**	**FPKM Thrombin**
G-protein gamma	1.22					
		GNG10	-1.34	2.17216E-08	27.192	20.2206
		GNG11	1.31	5.66209E-05	770.922	1011.05
		GNG12	-1.44	5.11591E-13	112.397	78.3023
		GNG2	2.43	6.38453E-09	0.465705	1.13132
G-protein beta	1.27					
		GNB2	1.36	1.91268E-10	137.786	187.43
		GNB3	-2.69	4.71259E-11	2.58703	0.962504
		GNB4	-1.61	0	15.8275	9.85484
Rho GEF	-1.16					
		ARHGEF12	-1.67	0	30.6424	18.3015

		ARHGEF2	1.33	5.80478E-08	27.6288	36.758
		ARHGEF3	-1.48	0	20.3279	13.7813
		ARHGEF6	-2.08	0	2.67944	1.28635
		ARHGEF9	-1.54	0.00469498	1.56367	1.0159
PI3K	-1.80					
		ATM	-6.84	0	4.98972	0.729483
		PIK3C2A	-5.09	0	17.7894	3.49234
		PIK3C3	-1.49	4.66294E-15	12.6504	8.50852
		PIK3CA	-3.14	0	16.8398	5.36228
		PIK3CB	-1.36	8.4357E-09	9.68516	7.12129
		PIK3CD	1.86	0	5.6579	10.5507
		PIK3CG	-1.76	2.32945E-10	1.86962	1.06419
		PIK3R1	-1.79	0	5.48118	3.06256
		PIK3R3	-1.59	1.50915E-05	5.24549	3.30763
		PIK3R4	-1.40	7.18425E-12	8.34176	5.97942
Rho	1.19					
		RHOB	1.31	9.48429E-06	232.232	304.587
		RHOC	1.38	0	458.267	630.235
		RHOF	1.65	0	8.29676	13.7268
		RHOG	1.40	1.02141E-14	58.6003	82.0172
		RHOJ	-1.57	0	127.446	80.9317
		RHOQ	-1.52	0	16.4802	10.8122
		RHOT1	-1.96	3.00071E-12	8.48834	4.33755
		RHOU	-1.54	0.00184367	0.900141	0.586092
		RHOV	-3.32	0.00564671	0.413912	0.124591
		RND3	-1.95	0	58.0564	29.7075
MLC	-1.06					
		MYL12B	-1.43	4.87832E-11	872.075	608.472
		MYL9	1.48	0	202.636	299.559

Table 2. Differentially Expressed Genes and Isoforms in Thrombin Signaling Pathway*. *, This table is reproduced from **Table S4** in Reference 4 with a minor modification. **, q-value, a false discovery rate adjusted p-value.

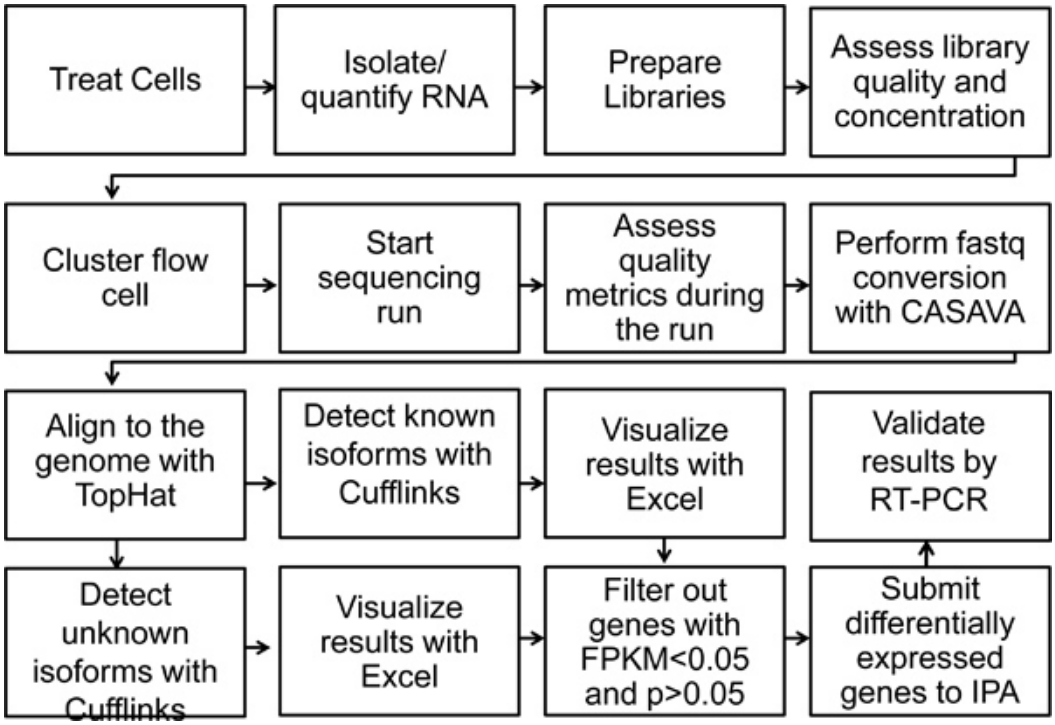


Figure 1. Flowchart of the protocol for RNA-seq profiling of the thrombin-mediated transcriptome in human pulmonary microvascular endothelial cells. First, treat cells and isolate and quantify RNA. Prepare libraries from the high-quality RNA and assess their concentration (via qPCR) and quality (via a bioanalyzer), then cluster on a flow cell. Start the sequencing run and analyze the quality of the run throughout. The major quality checkpoints are cycles 1, 4, 13 and 24. Using CASAVA, convert the bcl files to fastq files and then align those to the genome with TopHat. Detect known and unknown isoforms using Cufflinks and determine differential expression with CuffDiff. View the output files in Excel and filter out genes that have an expression level less than 0.05 FPKM or a p-value larger than 0.05. Submit the remaining genes to Ingenuity Pathway Analysis for further information about the gene functions and pathways affected by the thrombin treatment. Usually, selected set of differentially expressed genes are validated by an alternative approach such as qRT-PCR.

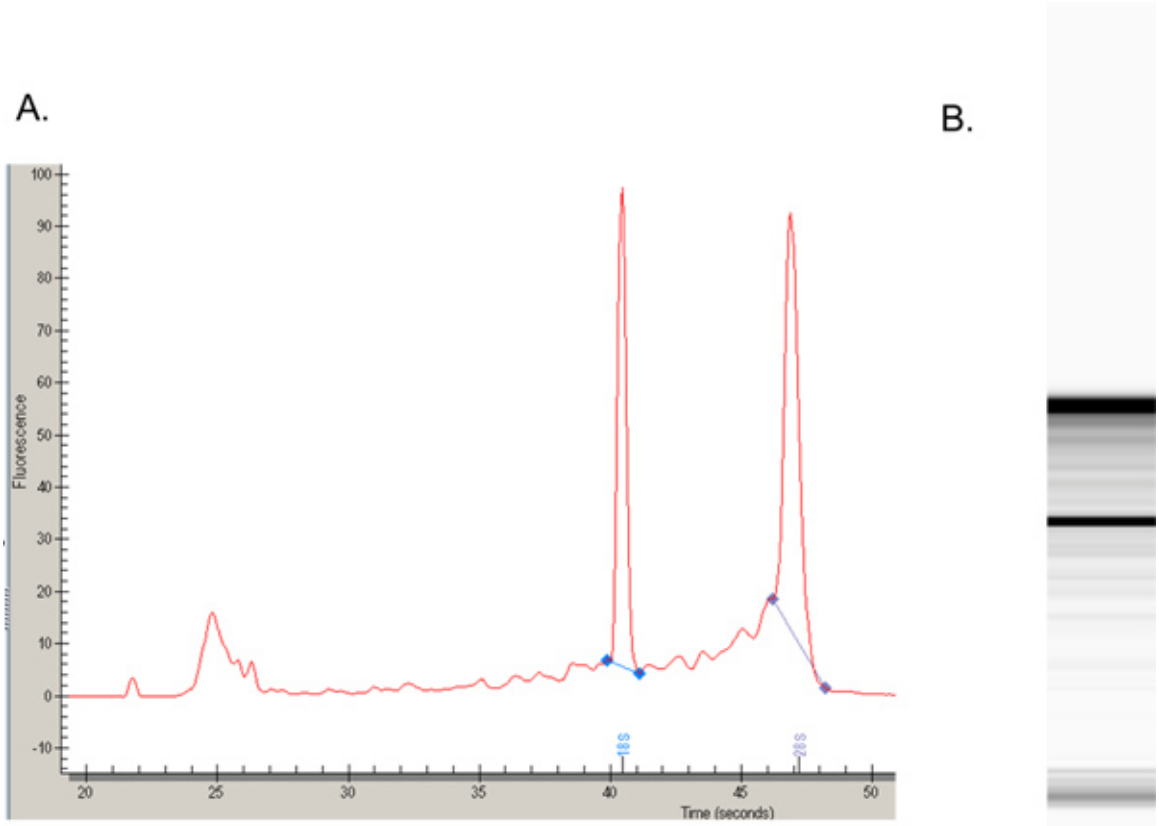


Figure 2. A representative RNA sample with an RQI of 8.4 as analyzed by the Experion Automated Electrophoresis Station. A) The individual trace of high-quality RNA - the x-axis depicts time and the y-axis depicts fluorescent signal. **B)** The virtual gel picture of a high-quality RNA sample. The intensity of the 28S peak is greater than 18S peak and no contamination is seen. Both bands are sharp and defined. Click here to view larger figure.

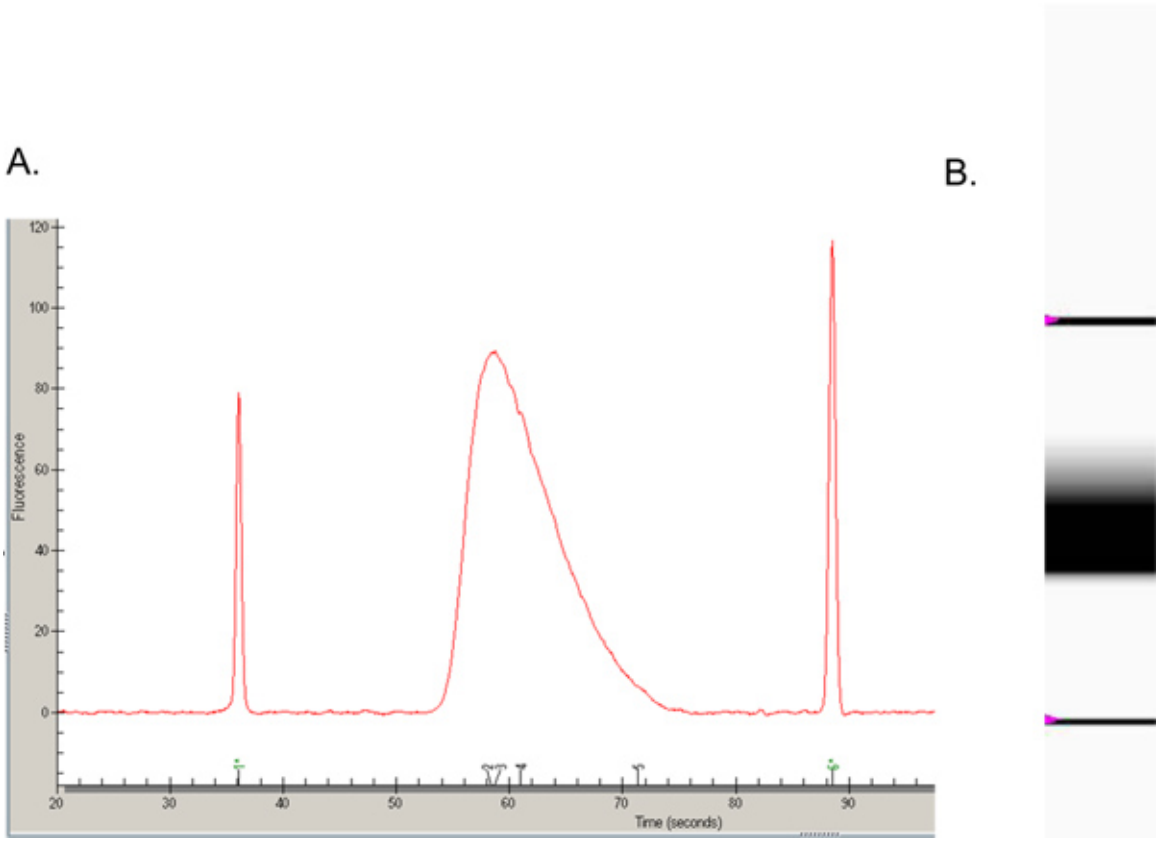


Figure 3. High quality library prepared from RNA as analyzed by the Experion Automated Electrophoresis Station. A broad peak between 250 and 300 bp is detected and no high molecular weight DNA (contaminating DNA) is observed. **A)** The individual trace of a high-quality library - the x-axis depicts time and the y-axis depicts fluorescent signal **B)** The virtual gel picture of a high-quality library. A broad peak between 250 and 300 bp is detected and no high molecular weight DNA (contaminating DNA) is observed. Click here to view larger figure.

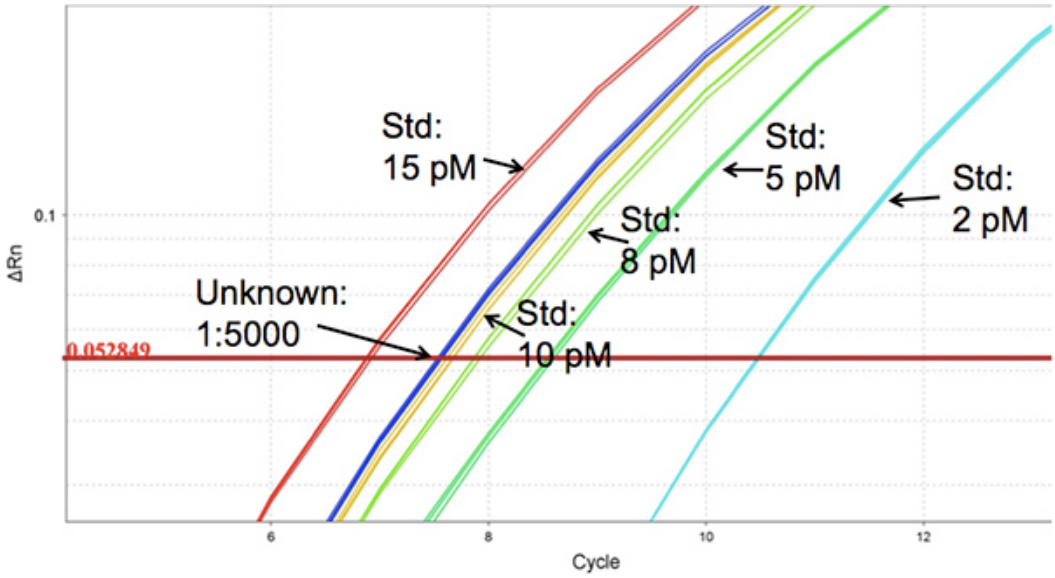


Figure 4. qPCR results using SyberGreen and primers specific to Illumina ligated adapters. A previously clustered library is used as standard curve to determine the optimal clustering concentration for the newly prepared library. The dilution of the unknown library falls within the standard curve concentrations. The standard curve samples are indicted by the arrows and the unknown sample is dark blue and also indicated by an arrow.

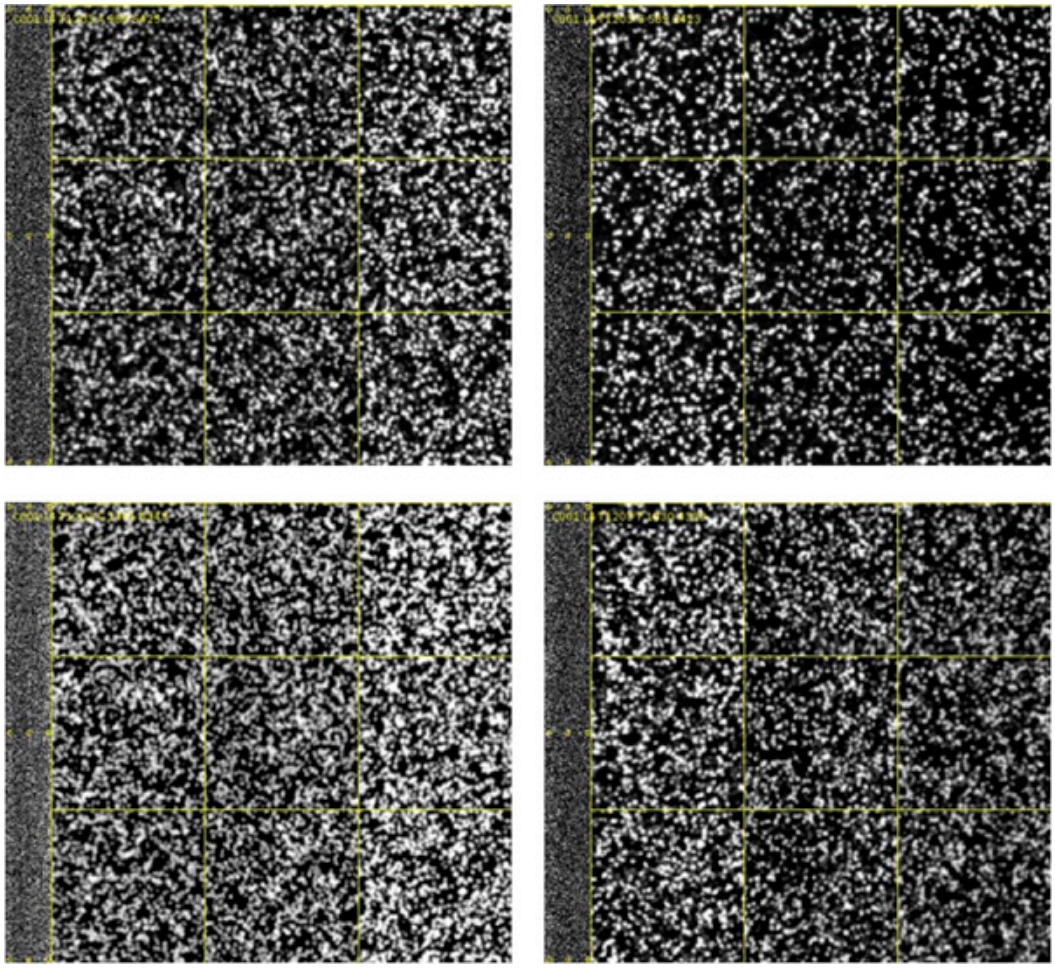


Figure 5. Appropriate cluster density during the first cycle imaging step. The clusters are clear, focused and bright. **A)** Base A; **B)** Base C; **C)** Base G; **D)** Base T.

Metric	Lane 1	Lane 2	Lane 3	Lane 4	Lane 5	Lane 6	Lane 7	Lane 8
Cluster Density (k/ mm^2)	420.94	412.95	410.81	408.54	416.71	416.56	410.02	411.84
A Intensity	25870.5	23886	23891.38	23735.83	23949.38	24933.08	24305.79	23950.29
C Intensity	21559.62	19822.33	19759.33	19472	19954.21	20611.75	19892	19438.29
G Intensity	13619.46	11756.67	11486.44	10498.38	12186.62	12195.5	11826.88	11010.5
T Intensity	19402.67	17663.79	17854.42	17353.75	17545.54	18060.67	18457.92	17397.12
A Focus Score	68.2	67.46	67.67	67.75	67.41	67.43	67.63	67.05
C Focus Score	67.93	67.33	67.56	67.66	67.33	67.39	67.46	66.9
G Focus Score	65.4	64.14	63.98	63.92	63.29	63.41	63.47	63.74
T Focus Score	66.45	65.57	65.5	65.49	65.03	65.23	65.57	65.59

Figure 6. The First Base Report generated after the first cycle is complete. The cluster density (although underestimated at this stage) is appropriate. The intensities look good (10,000-26,000, with G having the lowest intensity). The focus scores are also appropriate, falling around 65-70, although higher values are also appropriate.

Lane	Tiles	Density (K/mm2)
1	48	526 +/- 137
2	48	725 +/- 127
3	48	576 +/- 139
4	48	534 +/- 132
5	48	523 +/- 146
6	48	559 +/- 136
7	48	553 +/- 127
8	48	570 +/- 127

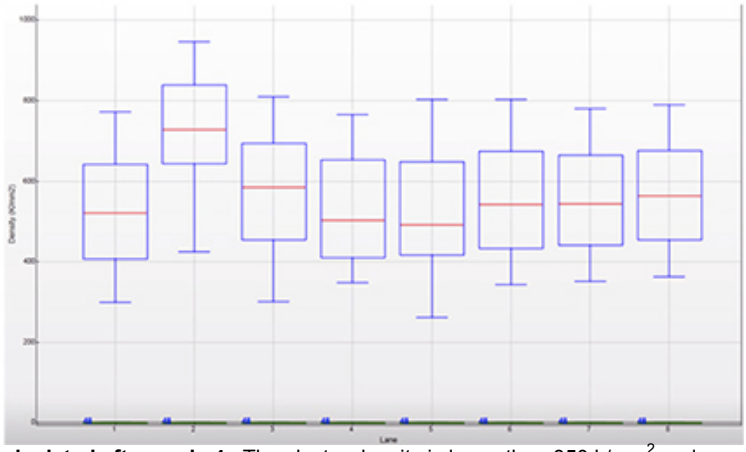


Figure 7. The cluster density calculated after cycle 4. The cluster density is lower than 850 k/mm^2 and even across all lanes. **A)** Tabular form; **B)** Graph form. Click here to view larger figure.

Lane	Tiles	Density (K/mm2)	Phas/Prephas (%)
1	48	526 +/- 137	0.134 / 0.179
2	48	725 +/- 127	0.135 / 0.176
3	48	576 +/- 139	0.140 / 0.181
4	48	534 +/- 132	0.142 / 0.183
5	48	523 +/- 146	0.139 / 0.181
6	48	559 +/- 136	0.139 / 0.180
7	48	553 +/- 127	0.139 / 0.179
8	48	570 +/- 127	0.147 / 0.178

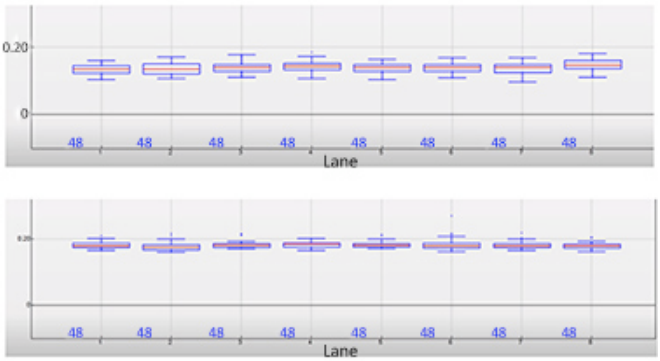
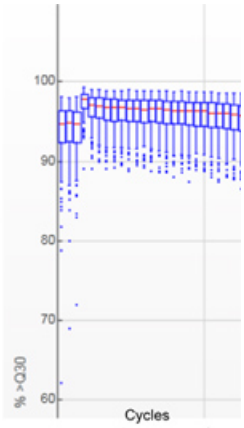


Figure 8. Phasing (not adding a base during a cycle) and Pre-phasing (adding two bases during a cycle). Both values are at 0.25 or below, indicating appropriate phasing/pre-phasing levels. **A)** The phasing and pre-phasing numerical values. **B)** The phasing values in graph form. **C)** The pre-phasing values in graph form. Click here to view larger figure.

A.

Lane	Tiles	Density (K/mm2)	% >= Q30
1	48	526 +/- 137	91.5
2	48	725 +/- 127	86.4
3	48	576 +/- 139	89.8
4	48	534 +/- 132	90.4
5	48	523 +/- 146	91.1
6	48	559 +/- 136	90.3
7	48	553 +/- 127	90.4
8	48	570 +/- 127	89.8

B.



C.

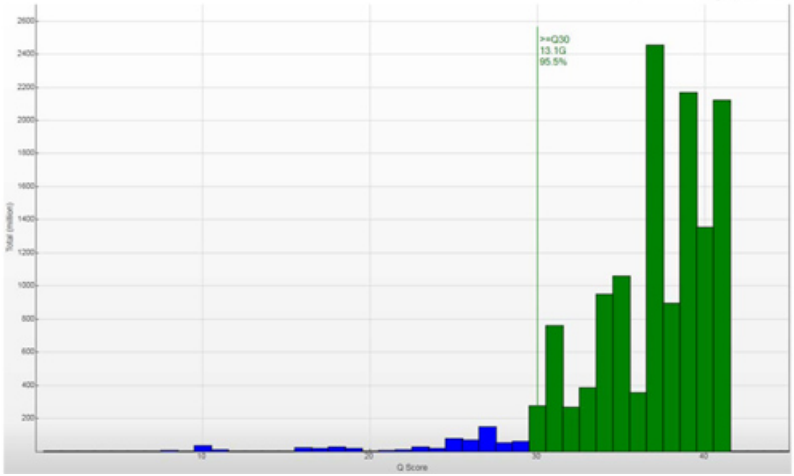


Figure 9. The different representations of Q30 scores after cycle 24. A score of 30 or higher indicates a 1 in 1,000 chance that base call is wrong. Around 90% of the Q scores are above 30. **A)** tabular form (Q30 scores here are from the entirety of the run, through cycle 24); **B)** Q30 scores by cycle. Each bar represents the distribution of the reads falling at Q30 or above for that particular cycle. **C)** Q30 score distributions through cycle 24. The distribution of Q scores on a cumulative basis through cycle 24. Q score is on the x-axis and millions of reads is on the y-axis. Reads with a Q30 above 30 are represented in green. Click here to view larger figure.

A.

Lane	Tiles	Density (K/mm2)	Cluster PF (%)
1	48	526 +/- 137	92.06 +/- 3.41
2	48	725 +/- 127	85.00 +/- 6.88
3	48	576 +/- 139	90.48 +/- 3.97
4	48	534 +/- 132	91.14 +/- 3.97
5	48	523 +/- 146	91.79 +/- 3.69
6	48	559 +/- 136	90.64 +/- 4.09
7	48	553 +/- 127	90.99 +/- 3.54
8	48	570 +/- 127	90.43 +/- 3.80

B.

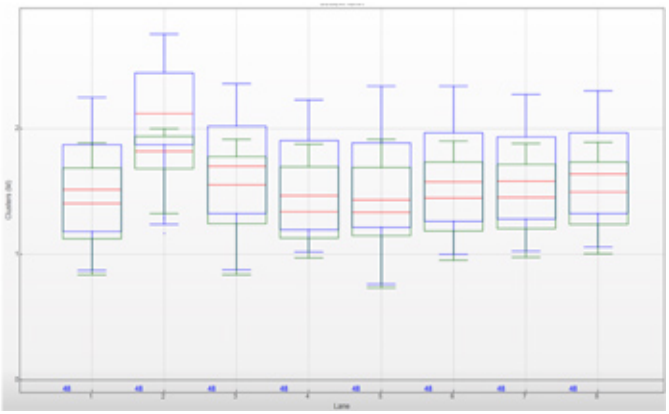


Figure 10. The different representations of clusters passing filter. Clusters passing filter is based on several parameters, including Q30, intensity and phasing/pre-phasing. At least 85% of the clusters are passing filter in each lane. **A)** Tabular form; **B)** Graph form, blue boxes represent the total number of clusters, green boxes represent the clusters passing filters. Click here to view larger figure.

A.

Lane	Tiles	Density (K/mm2)	Aligned (%)
1	48	526 +/- 137	0.5 +/- 0.0
2	48	725 +/- 127	0.2 +/- 0.0
3	48	576 +/- 139	0.5 +/- 0.0
4	48	534 +/- 132	0.6 +/- 0.1
5	48	523 +/- 146	0.5 +/- 0.0
6	48	559 +/- 136	0.5 +/- 0.1
7	48	553 +/- 127	0.5 +/- 0.1
8	48	570 +/- 127	0.5 +/- 0.0

B.

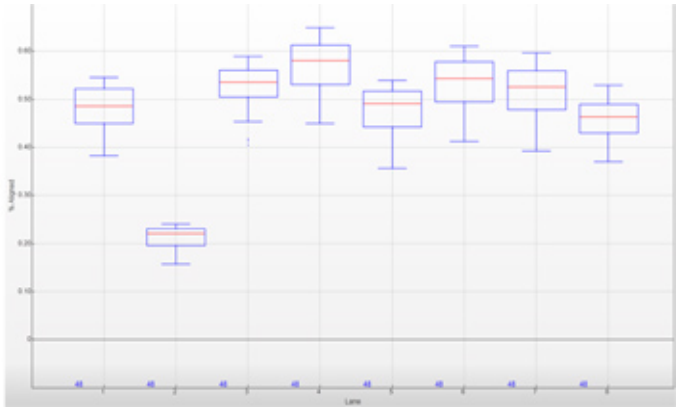


Figure 11. The percent of the clusters aligned to the PhiX genome. Approximately 0.5% of the clusters align to the PhiX genome, appropriate for the amount of PhiX spiked into the samples. **A)** Tabular form; **B)** Graph form. Click here to view larger figure.

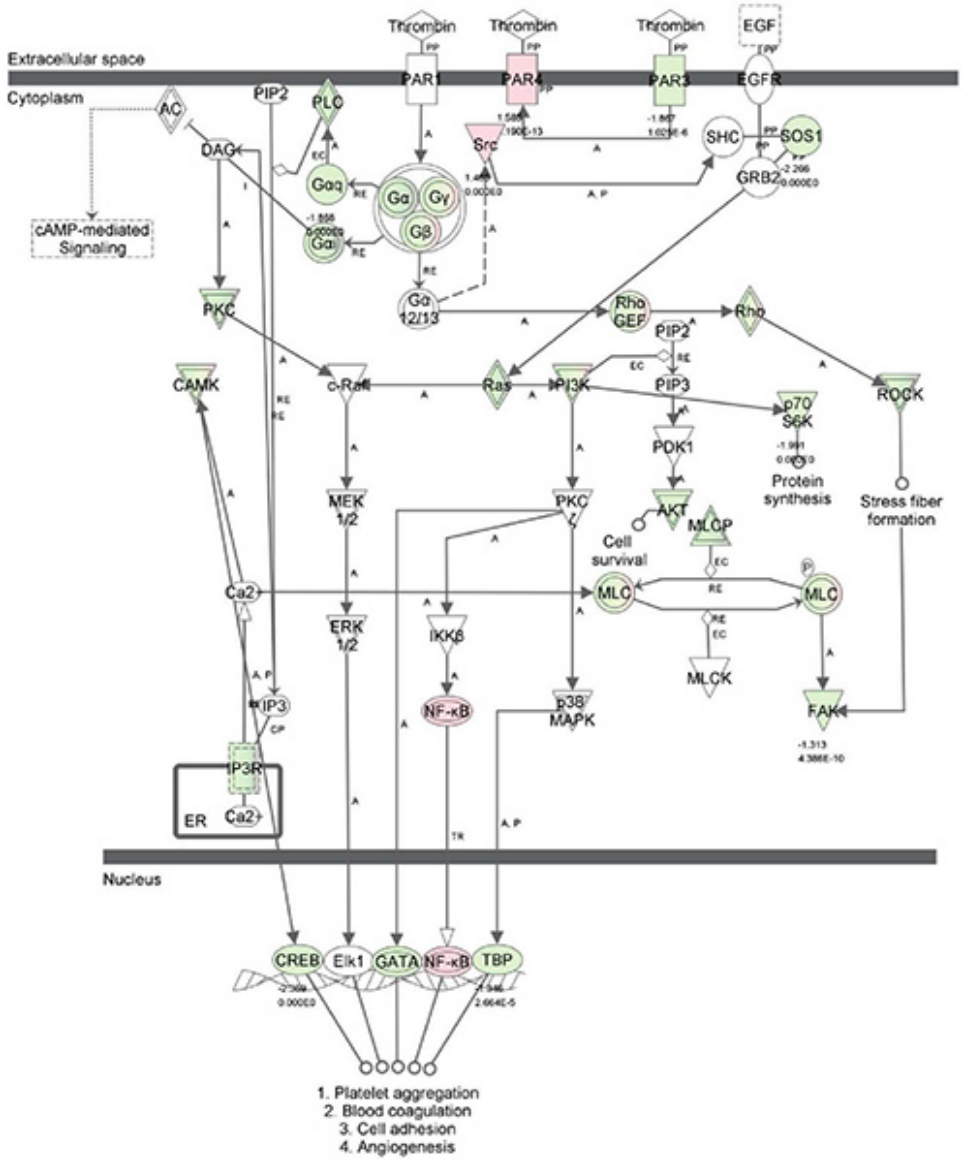


Figure 12. Genes differentially expressed by the thrombin treatment of HMVEC-LBI in the Thrombin Signaling pathway. The Thrombin Signaling pathway was constructed by Ingenuity. In the pathway, red color indicates up-regulated genes(^{3} 1.3 fold) and green down-regulated genes (^{3} 1.3 fold) by the thrombin treatment while white indicates no change of gene expression compared to the control cells. The detailed fold changes and full name of each gene or isoform are presented in **Table 2**. This figure is reproduced from **Figure 4** in Reference 4. Click here to view larger figure

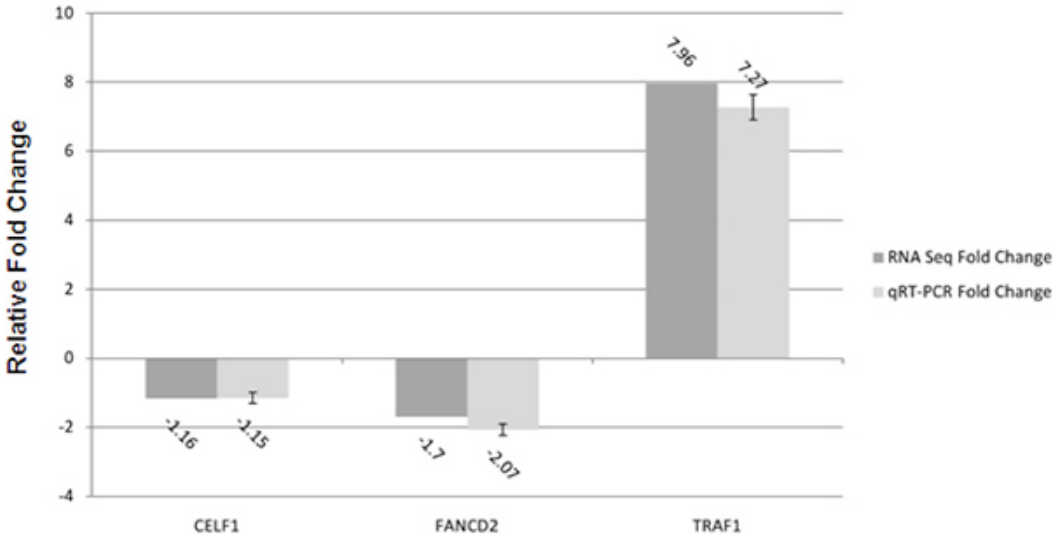


Figure 13. qRT-PCR validation of three differentially expressed genes from thrombin-treated HMVEC-LBI RNA-seq data. qRT-PCR was carried out as described in the protocol. Fold changes determined from the relative Ct values of the TaqMan Gene Expression assay for CUGBP, Elav-like family member 1 (CELF1), Fanconi anemia, complementation group D2 (FANCD2) and TNF receptor-associated factor 1 (TRAF1) were compared to those detected by RNA-seq. Replicates (n=4) of each sample were run and the Ct values averaged. All Ct values were normalized to β-actin. The error bars represent the range of the fold change as determined by the Data Assist software. $p<0.05$ was considered statistically significant in relative fold changes between thrombin-treat group and control group by both the RNA-seq and the qRT-PCR assays. The mRNA level in control groups by each assay was arbitrarily set as one, which are not shown. *, $p<0.05$; **, $p<0.01$. This figure is reproduced from **Figure 6** in Reference 4.

Discussion

Key steps

RNA Handling: RNases will degrade even the most high-quality RNA, therefore care must be taken during the isolation, storage and use of RNA[10]. Gloves are always worn to prevent contamination by RNases found on human hands. Gloves should be changed often, particularly after touching skin, doorknobs or other common surfaces. A set of pipettes should be dedicated solely to RNA work and all tips and tubes should be RNase-free. RNA isolation and downstream application should be performed in low-traffic areas that are routinely decontaminated with a product such as RNase-Zap. Degradation can also occur with repeated freeze-thaw cycles. These can be minimized by aliquoting RNA for downstream applications immediately upon isolation. Alternatively, cDNA can be made for downstream application such as qRT-PCR directly after isolation. RNA should be stored at -80 °C and kept on ice during use. It is also important to thoroughly thaw and vortex RNA samples before and during use.

The starting quality of the RNA is essential to the success of this protocol. Use of degraded or otherwise low-quality RNA can cause the library preparation to fail or result in low yield that will be insufficient for sequencing. Even if enough library can be made from degraded or low-quality RNA, it will not accurately represent the transcripts present in the sample, resulting in inaccurate quantification of expression. Also, any unremoved genomic DNA during the RNA isolation can cause an over-estimation of the amount of RNA used in the protocol, but if an amount (*e.g.* 1-2 µg) in the middle of the suggested range (0.1-4 µg) is used, the actual amount of RNA will be sufficient for the protocol. In addition, any genomic DNA is not likely to be picked up by oligo(dt) primer based library construction in the standard Illumina protocol and thus it will not cause downstream issues with expressed mRNA transcript levels. The quantification of the starting RNA using regular spectrophotometric readings should be sufficient and there is no need to use highly accurate quantitative methods such as RiboGreen since the libraries will be accurately quantified after preparation by qPCR and gene expression levels are normalized relative to the total number of reads during the data analysis steps.

Library Quantification. Accurate quantification of the libraries is vital to the appropriate generation of clusters. Only DNA molecules with successfully ligated adapters will form clusters. Spectrophotometric readings will not differentiate between DNA with adapters and DNA without, which will result in an over-estimation of the concentration of the library and ultimately result in under-clustering of the flow cell. The amount of DNA without ligated adapters will differ from library to library, even if the same RNA was used as starting material, so one cannot accurately assume that a standard percentage of the spectrophotometric reading is DNA with ligated adapters. Performing qPCR with primers specific to the adapters will detect only those DNA molecules with adapters successfully ligated to both ends of the molecule. This allows an absolute quantification of the libraries and in turn an accurate cluster density. Assessment of the libraries with an Experion instrument or a bioanalyzer is also important. This allows a visualization of the size range of the library, which is important during the data analysis steps, and ensures that there is no contamination that may interfere with cluster generation.

Data Analysis. In this protocol, we applied TopHat to align RNA-seq reads to human reference transcriptome and Cufflinks to assemble transcripts, estimate their abundances, and test for differential expression in thrombin-treated HMVEC-LBI cells. TopHat and Cufflinks are two popular tools, and they are free, open-source software[11]. A recent study showed that Cufflinks had a high sensitivity and specificity in detecting previously annotated introns[12]. However, TopHat and Cufflinks do not address all applications of RNA-seq nor are they the only tools for RNA-seq analysis[11]. TopHat and Cufflinks require a sequenced reference transcriptome or genome. They are particularly suitable for the analysis of RNA-seq data generated from either Illumina or SOLiD sequencing machines but not 454 or the classic capillary electrophoresis sequencing

platform. Users working without a sequenced reference genome may consider performing *de novo* transcriptome assembly using one of several tools such as Trinity (http://trinityrnaseq.sourceforge.net/), Trans-Abyss (http://www.bcgsc.ca/platform/bioinfo/software/trans-abyss/releases/0.2) or Oases (http://www.ebi.ac.uk/~zerbino/oases/). Many alternative transcriptome analysis programs now exist. For a survey of different programs, readers may wish to read the study by Garber *et al.*[13] and the review by Martin and Wang[14], which describe the comparative advantages and disadvantages and the theoretical considerations of most currently and commonly used programs. The choice of transcriptome analysis strategies (reference-based strategy, *de novo* strategy, and combined strategy) and programs depends on many factors, including the existence or completeness of a reference genome, the availability of sequencing and computing resources, the type of data set generated and, most importantly, the overarching goal of the sequencing project[14].

Another point should be made: computer programs for the transcriptome analysis may not yield a 100% accuracy of all transcript reporting. Sequencing error is another concern. It is always advisable that any differentially expressed transcript identified by RNA-seq be validated by real-time PCR. TaqMan probe-based chemistry is considered the gold standard for real time PCR. Also, newly identified transcripts should be validated by the Sanger method of DNA sequencing before any further experimentation.

Troubleshooting

Library Preparation: A smear of high molecular weight genomic DNA after the library preparation steps could be caused by a carry-over from the final bead purification step during the library preparation. Returning the libraries to the magnetic stand for a full five minutes may resolve the issue. If not, the issue may stem from incomplete fragmentation of the RNA during the first steps of the library preparation. This could occur if low-quality RNA is used or if the protocol is not followed exactly. For example, altering the incubation times or temperature, adding reagents in the incorrect order or pausing between fragmentation and the synthesis of the first strand of the cDNA. If returning the libraries to the magnetic stand does not remove the high MW DNA, it is advisable to repeat the library preparation with fresh RNA samples.

Low intensity: If the first cycle intensity is lower than expected, it is possible that the primers did not hybridize to the clusters properly during the last step of cluster generation or the flow cell was stored for too long between clustering and the start of the sequencing run. In this case, a primer rehybridization should be performed using a primer rehyb kit from Illumina. Most often, this rehyb solves the intensity issues and the run can be started over without any problem. If, after a rehyb, the intensities are still low, the issue may be with the libraries or the sequence by synthesis (SBS) reagents and Illumina technical service should be contacted for further troubleshooting. If the first read's intensities are good, but the intensities are low after turn around, a primer rehyb of the second read primer may be necessary. This is performed on the sequencing instrument and Illumina technical service should be contacted for the procedure.

High phasing/prephasing: If higher than normal phasing or prephasing is observed, a possible cause is the contamination of the other SBS reagents with cleavage buffer. During the SBS preparation steps, it is essential to handle the cleavage buffer last and change gloves between the handling of the cleavage buffer and any other reagents. If cleavage buffer contamination is suspected, a primer rehyb of the flow cell should be performed and new SBS reagents should be prepared. Alternatively, there could be contamination in the lines of the instrument. If this is suspected, the instrument should undergo a maintenance wash (a water wash followed by a NaOH wash, followed by a final water wash), a primer rehyb should be performed on the flow cell and new SBS reagents should be prepared. If the phasing/prephasing values are moderately high (~0.5), the run may be continued until the Q30 scores and clusters passing filter are calculated. These levels may still be in acceptable limits even with high phasing and/or prephasing.

Generalizability of this protocol

While it is specific to the Illumina HiScanSQ, this protocol is applicable to any of the HiSeq family or the Genome Analyzer II of Illumina instruments (www.illumina.com) with minor modifications of the cluster generation steps and sequencing reagents. Other next-generation DNA sequencing platforms such as the SOLiD series (www.lifetechnologies.com), the GS systems (www.454.com) as well as some emerging newer systems are also employed for the purpose of RNA-seq [11]. Although their library construction and sequencing procedures may be slightly different, the RNA handling tips, the data analysis portions (downstream of the CASAVA steps) and the validation by RT-PCR presented in this protocol can be of reference value to their RNA-seq applications.

It should also be pointed out that this protocol is specific to RNA-seq at one time point of thrombin-treated HMVEC-LBl cells, but it could easily be adapted to a multi-time point study or studies in other cells, tissues treated with different stimuli or inhibitors, or comparisons of transcriptomes in cells or tissues between a healthy state and a disease state.

Disclosures

No conflicts of interest declared.

Acknowledgements

The authors would like to thank Dr. Stephen Kingsmore and the Pediatric Genome Medicine Center at Children's Mercy Hospitals and Clinics for the use of their computing clusters for our data analysis, Illumina's field service team (Elizabeth Boyer, Scott Cook and Mark Cook) and technical consultant team for their quick responses and helpful suggestions on the running of the next generation DNA sequencing instrument, HiScanSQ, and data quality analysis. This work was supported in part by National Institutes of Health Grant HL080042 (to S.Q.Y.) and start-up fund and endowment of Children's Mercy Hospitals and Clinics, University of Missouri at Kansas City (to S.Q.Y.).

References

1. Wang, Z., Gerstein, M., & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57-63 (2009).

2. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26**, 1135-1145 (2008).
3. 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 Res.* **18**, 1509-1517 (2008).
4. Zhang, L.Q., *et al.* RNA-seq Reveals Novel Transcriptome of Genes and Their Isoforms in Human Pulmonary Microvascular Endothelial Cells Treated with Thrombin. *PLoS One.* **7**, e31229 (2012).
5. Trapnell, C., Pachter, L., & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics.* **25**, 1105-1111 (2009).
6. Langmead, B., Trapnell, C., Pop, M., & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
7. Li, H., *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* **25**, 2078-2079 (2009).
8. Trapnell, C., *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511-515 (2010).
9. Khatri, P., Sirota, M., & Butte, A.J. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.* **8**, e1002375 (2012).
10. Nielsen, H. Working with RNA. *Methods. Mol. Biol.* **703**, 15-28 (2011).
11. Trapnell, C., *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562-578 (2012).
12. Robertson, G., *et al. De novo* assembly and analysis of RNA-seq data. *Nat. Methods.* **7**, 909-912 (2010).
13. Garber, M., Grabherr, M.G., Guttman, M., & Trapnell, C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods.* **8**, 469-477 (2011).
14. Martin, J.A. & Wang, Z. Next-generation transcriptome assembly. *Nat. Rev. Genet.* **12**, 671-682 (2011).

