Title
Quantitative Transcriptomics from Limiting Amounts of mRNA /

Permalink
https://escholarship.org/uc/item/609802sv

Author
Bhargava, Vipul

Publication Date
2013

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Quantitative Transcriptomics from Limiting Amounts of mRNA

A dissertation submitted in partial satisfaction of the requirements for the degree
Doctor in Philosophy

in

Bioinformatics and System Biology

by

Vipul Bhargava

Committee in charge:

Professor Shankar Subramaniam, Chair
Professor Mark Mercola, Co-Chair
Professor Vineet Bafna
Professor Steven Briggs
Professor Lawrence Goldstein

2013

The Dissertation of Vipul Bhargava is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

____

____

____

____
Co-Chair

____
Chair

University of California, San Diego

2013

DEDICATION

To my parents and my lovely wife, for all the sacrifices they have made over the years.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor Prof. Shankar Subramaniam and co-supervisor Mark Mercola for always believing in my ability and giving me an outstanding support to pursue my research ambitions. I joined Shankar's lab in late 2007 as a rotation student. His passion towards systems biology and the latest technological advances instantly caught my attention and led to conception of my PhD project. This project entailed considerable computational and molecular biology training. Given my engineering background, I took some time to familiarize myself with different aspects of molecular biology. Moreover, this project went though multiple rounds of optimization that lasted for more than three years. My supervisors showed tremendous patience during this period and provided me with much needed encouragement and guidance. There were times when I doubted my ability to lead this project to its completion, Shankar stood by me and approached his colleagues to ensure that I get all the assistance I needed. Both, Shankar and Mark, were highly approachable and always tried to accommodate me in their otherwise busy schedule. This project was initiated at the time when next generation sequencing was at early stages of its development and the sequencing costs were high. Shankar did not hesitate once to invest in this project and gave me complete independence to lead the project. On a similar note, Mark always reminded me of the "big picture" and how the focus of my project should be to address challenging biological problems. This ensured that I

do not digress from our main objective. I cherish my relationship with my supervisors and I hope their blessings stay with me forever.

I would also like to acknowledge Dr. Suresh Subramani for letting me use his lab resources. The majority of the molecular biology component of my PhD project was performed in his lab. He was very supportive of my project and treated me as one of his lab members. I was always invited to their lab celebrations and the Christmas parties. His humility and enthusiasm towards research are some of his traits that impressed me the most and I wish to inculcate these qualities as I move forward. I have worked in his lab for over four years and in this period I have made some exceptional friendships in his lab. I would like to particularly mention Dr. Ravi Manjithaya, Aparna, Saurabh Joshi, Gaurav Agrawal, Jean Claude and Taras for sharing their intellect with me over many coffee sessions and making my stay in Suresh’s lab comfortable. I would also like to thank the lab managers of his lab, Danielle and Sarah, for all of their support.

During my stint in UC San Diego as a graduate student, I met some exceptional scientists whose work and ethics are inspirational to me. I owe much appreciation to rest of my thesis committee (Dr. Vineet Bafna, Dr. Steven Briggs and Dr. Lawrence Goldstein) for their thoughtful comments and guidance on my work as well as my future ambitions. I thank Dr. Shyni Varghese and Kun Zhang for their much appreciated advice on my PhD project.

My camaraderie with the members of the Shankar and Mark labs was wonderful. I cherish all the stimulating discussions I had with them. My special thanks to Erik Willems, Wenqing Cai, Ke Wei, Ramon Diaz Trelles, Alexandre Colas and many more from Mark’s lab for providing me with their valuable feedback on my project during the lab meetings. Similarly, I would like to thank Pang Ko, Shakti Gupta, Mano Maurya, Merril Gersten, Kavitha Mukund, Thomas Hagan, Brian Saunders and Ashok from Shankar’s lab for helping me address the computational challenges of my project and advising me on the data analysis. Pang Ko was instrumental in designing the primers sets. He is an exceptional computer scientist and his contribution on my PhD project was immense. Shakti and Mano taught me various systems biology tools and their application to make meaningful biological interpretation from the large datasets.

The support that I received from the sequencing core facilities, Biogem and Scripps, was vital for successful completion of my projects. Dr. Steven Head and Phillip Ordoukhanian always prioritized my sequencing requests. There were times when my samples had to be run again because of some technical issues. They were accommodative of my request and tried to help me as much as they could in terms of resources and supplies. John Shimashita helped me prepare sequencing libraries in timely manner while Lana Schaffer provided bioinformatics support.

I feel extremely lucky to have made great friends over the years that ensured that I have work-life balance. My friends from the Bioinformatics and Systems Biology program (Harish Nagarajan, Nitin Udpa, Jian Wang, Winnie Yi Wen, Stan Luban and Josue Perez) provided me with much needed moral support all throughout my graduate studies. My peers in the Biology program

(Francisco, Gunnar, Valentino and Robert) and others (Shivani Singh, Marisol Chang, Julianne Huang, Vichy Zhong, Yuvraj Agarwal, Gunjan Agarwal, Anup Tapadia, Anand Mukhopadhyay and Nidhi Vashistha) never gave up on me and ensured that I maintain sanity in my difficult times. Finally, I owe great deal of gratitude to my wonderful wife, Surbhi. She has been one of my pillars of strength. Her delicious cooking kept me physically and mentally satiated. Her calm and composed disposition has always inspired me and I look forward to the rest of our future together.

Chapter 2, in full, is adapted from **Bhargava, V.,** Ko, P., Willems, E., Mercola, M. & Subramaniam, S. (2013) *Quantitative Transcriptomics using Designed Primer-based Amplification.* **Sci. Rep.** 3, 1740; DOI:10.1038/srep01740. The dissertation author was the primary author of this paper responsible for the research.

Chapter 3 is in full material submitted for publication from **Bhargava, V.**, Head, S., Ordoukhanian, P., Mercola, M., Subramaniam, S. (2013) *Technical Variations in Low Input RNA-seq Methodologies.* The dissertation author was the primary author of this paper responsible for the research.

Chapter 4, in full is currently being prepared for submission for publication of the material. Dinorah, F.M.*, **Bhargava, V.***, Gupta, S., Verma, I., Subramaniam, S.. The dissertation author was a joint first author of this paper, responsible for sequencing library generation and much of data analysis. *Equal contribution

VITA

2002 B.Tech. in Chemical Engineering, Indian Institute of Technology Bombay, India.

2004 M.Sc. in Bioinformatics, National University of Singapore, Singapore.

2004 – 2006 Research Associate, Bioinformatics Institute, Singapore.

2013 Ph.D. in Bioinformatics and Systems Biology, University of California San Diego, United States.

PUBLICATIONS

1. **Bhargava, V.,** Ko, P., Willems, E., Mercola, M. & Subramaniam, S. (2013) *Quantitative Transcriptomics using Designed Primer-based Amplification.* **Sci. Rep.** 3, 1740; DOI:10.1038/srep01740.

2. Gupta,N., Benhamida,J., **Bhargava, V.**, Goodman,D., Kain, E., Kerman.L., Nguyen, N., Ollikainen, N., Rodriguez, J., Wang, J., Lipton, M.S., Romine, M., Bafna, V., Smith, R.D. & Pevzner, P.A. (2008) *Comparative Proteogenomics: Combining Mass Spectrometry and Comparative Genomics to Analyze Multiple Genomes*. **Genome Res.**18: 1133-42.

3. Chiam, K.H., Tan, C.M., **Bhargava, V**., Rajagopal, G. (2006) *Hybrid Simulations of Stochastic Reaction – Diffusion Processes for Modeling Intracellular Signaling Pathways.* **Phys Rev E. Stat Nonlin Soft Matter Phys.** 74(5-1):051910.

4. **Bhargava, V.**, Head, S., Ordoukhanian, P., Mercola, M., Subramaniam, S. *Technical Variations in Low Input RNA-seq Methodologies.* (Submitted)

5. Dinorah, F.M.*, **Bhargava, V.***, Gupta, S., Verma, I., Subramaniam, S. *Functional Characterization of Dedifferentiated Neurons and Astrocytes Inducing Gliomas in Mice.* (In Preparation) * Equal Contribution.

FIELDS OF STUDY

Major Field: Bioinformatics and Systems Biology

ABSTRACT OF DISSERTATION

Quantitative Transcriptomics From Limiting Amounts Of mRNA

by

Vipul Bhargava

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California, San Diego, 2013

Professor Shankar Subramaniam, Chair

Quantification of global transcripts expression is a key step towards developing system-level understanding in biology. Probe independent RNA-seq provides digital estimation of transcript abundance with dynamic range large enough to accurately quantify the majority of complex mammalian transcriptomes. However, a reliable quantification of low abundant transcripts from limited amounts of mRNA has remained a challenge for RNA-seq. The widely used RNA-seq protocol requires 1-10 ng of mRNA to generate robust sequencing libraries restricting its application in disciplines where obtaining such amounts of mRNA is challenging, such as in developmental biology, stem cell biology and forensics. To address this issue, we developed a novel RNA-seq methodology (DP-seq) that uses a defined set of 44 heptamer primers to amplify majority of the mammalian transcripts from limiting amounts of mRNA, while preserving their relative abundance. DP-seq reproducibly yields high levels of amplification from as low as 50 pg of mRNA (50-100 mammalian cells) with a dynamic range of over five orders of magnitude in RNA concentrations. A novel two-step amplification step utilizing a combination of mesophilic and thermophilic polymerases was devised to achieve efficient amplification from the heptamer primers. Furthermore, we exploited PCR biases observed in our methodology to reduce the representation of highly expressed ribosomal transcripts by more than 70% in our sequencing libraries.

We validated DP-seq on lineage segregation model in early stem cell cultures achieved by modulating TGFβ pathway. DP-seq accurately quantified the majority of the low expressed transcripts and revealed novel lineage markers and putative TGFβ target genes. Similarly, by using DP-seq we functionally characterized dedifferentiated neurons and astrocytes and found the cell cycle, Wnt signaling and the focal adhesion pathways to be involved in the maintenance of their undifferentiated state.

Finally, we compared DP-seq with other amplification-based strategies and found similar transcriptome coverage and overlapping technical noise. Interestingly, the technical noise increased significantly when ultra-low amount of mRNA (single cell level) was used, irrespectively of the methodology. In conclusion, this study provides an economical and efficient solution for sequencing library generation using low amounts of mRNA thereby increasing the applicability of RNA-seq to a wider spectrum of biological systems.

Chapter 1

Introduction

Eukaryotic cells exhibit tremendous diversity in RNA expression and structure. It is the repertoire of RNA species expressed by a biological cell that differentiate it from other genetically identical cells sharing the same chromosomal DNA. Furthermore, most of the biological processes such as proliferation and differentiation involve systematic changes in expression levels of numerous RNA species. Hence, accurate quantification of whole RNA populations is necessary to define the cellular context and gain systems level understanding of the molecular mechanisms involved in the biological processes. A number of methodologies have emerged that can simultaneously analyze expression patterns of thousands of RNA species. Two of the most popular methods, microarrays and high-throughput RNA sequencing (RNA-seq) have greatly enhanced our understanding of transcription and post – transcriptional regulation of the mammalian genomes.

1.1 Microarrays vs. RNA-seq

Until recently, hybridization based microarray platforms used to be method of choice for simultaneous monitoring of expression levels of all annotated transcripts[1, 2]. However, the platform suffers with numerous drawbacks limiting its capability in deciphering the code of transcriptional machinery, especially for complex mammalian genomes[3]. Microarray platforms require at least microgram amounts of mRNA, which is equivalent to RNA content obtained from more than 100,000 mammalian cells) as opposed to 1-10 nanogram (ng) of mRNA required by the most popular RNA-seq method. Since, the microarray platform uses short oligos to capture mRNA expression, it requires a priori knowledge of transcripts expressed in a given cellular context and their sequences. Identification of differentially expressed transcripts under various biological conditions, are often marred by hybridization and cross hybridization bias introduced by variable GC content, length of the probes, dye bias etc. The limitations of microarray scanners in detecting low signal intensities restrict the accurate quantification of low abundant transcripts. The dynamic range of this platform is further undermined by saturation of the fluorescent intensities for the high expressed transcripts. In contrast, the sequencing based approach of transcriptome profiling does not require prior knowledge of transcripts and allows identification of novel transcripts and alternate splice site variants[4-6]. A typical deep sequencing based approach generates millions of sequencing reads, thus offering large dynamic range and subsequently better estimation of low abundant transcripts[7-9]. Higher dynamic range offered by sequencing based technologies implies identification of more number of transcripts (80 – 90%) as compared to those identified by microarrays (40 – 50%). Moreover, sequencing based protocols have digital output, as opposed to analog for hybridization based protocols obviating the need for complex algorithms for data normalization and summarization.

Comparison of expression profiles obtained from microarrays and deep sequencing based technologies like Illumina genome analyzer II revealed high correlation for transcripts expressed at moderate levels[6]. However, the correlations were poor for transcripts expressed at low or high levels. Finally, high reproducibility and sensitivity achieved by RNA-seq[10] and ever decreasing cost of sequencing has further reinforced their position as preferred platforms for mRNA expression analysis.

1.2 Protocols for RNA-seq

During the early stages of RNA-seq method development, two protocols were suggested for sequencing library generation: 3’ tag digital gene expression[6] and full-length RNA sequencing[10]. 3’ tag digital gene expression uses oligodT primers to synthesize first strand cDNA from polyadenylated mRNA. The first strand cDNA is later converted to double stranded cDNA using random hexamer primers. Next, the double stranded DNA is digested with DpnII enzyme followed by MmeI to generate 20 – 21 base pair cDNA tags. Later, the library is ligated to platform dependent adapter molecules and processed for massive parallel sequencing. As expected, the sequencing reads generated from this protocol are enriched for 3’ ends of the cDNA and give little information about structure of transcripts including exon usage, splice site variants etc.

Full-length RNA sequencing involves fragmentation of RNA (RNA hydrolysis or nebulization) into 100 – 300 bp fragments. This step is necessary to reduce the formation of stable RNA secondary structures, which hinders with the full-length cDNA synthesis via reverse transcriptase. Moreover, the fragmentation of RNA makes the cDNA library compatible for sequencing in Illumina platforms. The fragmented RNA library is primed with random hexamer primers to generate double stranded cDNA library. The library is later ligated to the standard Illumina adapters and processed for sequencing. Since the sequencing reads are generated from whole length of the mRNA, it allows investigation of the structure of the mammalian transcriptomes at unprecedented levels. However, this protocol creates a different bias in the outcome. The number of sequencing reads generated from a transcript depends upon the length of the transcript implying that the transcripts with longer lengths and high expression are preferred over all the other transcripts[11]. This protocol also maintains the relative order of the transcripts expression. The transcripts expression in mammalian genomes follows a power law distribution. This implies that most of the sequencing effort is spent on sequencing high expressed transcripts. In the majority of the cases, these high expressed transcripts are involved in maintenance of structural integrity of cells or cell viability (metabolic pathways). Importantly, these transcripts do not change their expression pattern in most of the cellular context. On the other hand, vast majority of the cell signaling molecules including transcription factors are expressed at low to moderate levels. These transcripts do not get enough representation in the sequencing libraries resulting in poor quantification even at high sequencing depths[12, 13]. The protocol also suffers with biases arising out of random hexamer priming and the random fragmentation of the mRNA[14]. This error is further propagated, as not all the fragments generated from this method will map uniquely to the transcripts. Theoretical assessment of mammalian transcriptome revealed existence of 74% of uniquely mapped 32 base pair reads[15]. This implies that only 74% of the time the random fragments generated will map uniquely to the transcripts. This systematic error gets more pronounced in low abundant transcripts which possess less unique regions within their sequences. Finally, this protocol requires 1 – 10 ng of mRNA for successful generation of the sequencing library. This restricts the application of this method in fields such as developmental biology, stem cell biology, forensics and even for FACs sorted cell populations where obtaining such large amounts of mRNA is impractical.

To address the issue of sequencing from limiting amounts of mRNA, a number of amplification-based methodologies were proposed. These methodologies generate large amount of amplified DNA, required for successful production of sequencing libraries, by performing either exponential or linear amplification of mRNA. Some of the initial work on the development of amplification-based approaches were demonstrated by ligation mediated PCR[16], multiple displacement amplification[17], single – primer isothermal amplification[18], in – vitro transcription based linear amplification[19]. The performance of these methods in deep sequencing based platforms has not been assessed. Other amplification-based methods have utilized the hybridization and extension potential of random hexamer, heptamer and/or octamer primers to amplify the majority of expressed transcripts[20-22]. However, they often result in low yield of good quality reads arising out of mis-hybridization of primers and primer dimerization. Also, the random priming methods do not discriminate regions of the transcriptome to amplify, specifically low abundance transcripts. This limitation is also seen in other uniform amplification strategies[23-27]. Another approach, involving targeted enrichment[28-30] requires longer sample preparation steps, larger amounts of RNA and high costs.

For our first RNA-seq experimentation, we used double random priming methodology that uses random octamer primers to amplify most of the mouse transcripts[22]. Our transcriptome data revealed a number of issues with library preparation protocol. Firstly, our sequencing libraries had high proportions of PCR spurious products and primer-dimerization. Consequently, only 64% of the sequencing reads mapped to the mouse genome. Moreover, the method generated sequencing reads that mapped to multiple mRNA species and only 18% of the reads were uniquely mapped to the mRNA database. This resulted in poor dynamic range and reduction of statistical power of the experiment with the quantitation of low abundant transcripts severely affected. In this method, the first eight sequencing cycles were used to sequence random octamer primers and since majority of the octamer primers displayed mis-priming, we had to truncate the first eight base pairs from the sequencing reads. Finally we had no control on the regions of the transcriptome that were amplifying and the extent of amplification.

1.3 Specific goals of the dissertation

Chapter 2 describes a novel RNA-seq methodology (DP-seq; Designed Primers based RNA-seq) where we designed a set of 44 heptamer primers to amplify the majority of the mouse transcriptome from as low as 50 pg of mRNA while maintaining the relative abundance of the transcripts[31]. Intensive computational analysis was performed to identify 44 heptamer primers that amplified >80% of the expressed transcripts. The primers were also designed to hybridize preferentially to the unique regions of the mouse transcriptome. Owing to low melting temperatures of the heptamer primers, a novel two-step amplification protocol was devised where a combination of mesophilic and thermophilic polymerases were used. The protocol was further optimized to reduce mis-hybridization of primers and primer dimerization. We further explored the potential of our primer design strategy to selectively suppress the amplification of the highly expressing transcripts such as ribosome encoding transcripts. Our sequencing data demonstrated a significant reduction in the representation of the ribosomal transcripts with multiple choices of primer sets thus demonstrating the potential of our methodology to perform “targeted amplification”. We later compared our methodology with a full-length cDNA amplification strategy (Smart-seq)[25] and observed comparable transcriptome coverage and similar technical noise. We validated our methodology on an in-vitro cell culture based model of early mouse embryogenesis to study lineage segregation achieved by modulating TGFβ signaling pathway. Our transcriptome data showed early expression of numerous lineage markers then previously anticipated, thus highlighting the sensitivity of our protocol.

Chapter 3 describes a comparative analysis of RNA sequencing libraries prepared from low amounts of mRNA using three different methodologies, namely Smart-seq[25], CEL-seq[23] and DP-seq. Two of these methodologies, Smart-seq and CEL-seq, have been utilized for single cell transcriptomics analysis. Our analysis of the sequencing libraries prepared from serial dilutions of mRNA revealed inefficient amplification of the majority of the low to moderately expressed transcripts. Enhanced stochasticity in primer hybridization and/or enzyme incorporation resulted in high technical noise particularly in the low expression regime. In the sequencing libraries prepared from 25 pg of mRNA, vast majority of the low expressed transcripts exhibited stochastic loss. Additionally, significant distortions in fold changes of the differentially expressed transcripts, irrespective of their average expression or level of differential regulation, were observed as the amount of mRNA was reduced. Our study demonstrated that the technical variations observed in these methodologies are profound which can mask subtle biological differences.

Chapter 4 discusses another implementation of our methodology where we performed transcriptome profiling of the dedifferentiated neurons and astrocytes to define their undifferentiated state. A recent study showed glioma formation in mouse brain when mature neurons and astrocytes were transduced with lentiviral vector containing shRNA targeting NF1 and p53 genes[32]. Our transcriptomics data revealed that these transduced neurons and astrocytes left their original state and dedifferentiated to undifferentiated neural stem cell like state. Our pathway enrichment analysis demonstrated the role of Wnt signaling, cell cycle and focal adhesion pathways in maintaining the undifferentiated states of these cells. Using cytoscape toolbox, we further identified a gene interaction network that was conserved between the two dedifferentiated cell-types. Finally, our analysis revealed the role of Spp1 gene in cell proliferation and migration, which needs further investigation.

Chapter 2

Quantitative Transcriptomics using Designed Primer-based Amplification

2.1 Abstract

We developed a novel Designed Primer-based RNA-sequencing strategy (DP-seq) that uses a defined set of heptamer primers to amplify the majority of expressed transcripts from limiting amounts of mRNA, while preserving their relative abundance. Our strategy reproducibly yields high levels of amplification from at least 50 picograms of mRNA while offering a dynamic range of over five orders of magnitude in RNA concentrations. We also demonstrated the potential of DP-seq to selectively suppress the amplification of the highly expressing ribosomal transcripts by more than 70% in our sequencing library. Using lineage segregation in embryonic stem cell cultures as a model of early mammalian embryogenesis, DP-seq revealed novel sets of low abundant transcripts, some corresponding to the identity of cellular progeny before they arise, reflecting the specification of cell fate prior to actual germ layer segregation.

2.2 Introduction

Next Generation Sequencing-based approaches for whole transcriptome analysis produce millions of sequencing reads, which represent the vast majority of the expressed transcripts. The high number of reads allows a digital estimation of transcript abundance, resulting in a large dynamic range and high sensitivity[5-8]. This is in contrast to microarray platforms, which rely on the hybridization efficiencies of transcript specific probes to their corresponding targets, and thus result in analog expression profiles and a low dynamic range. With the recent dramatic increase in sequencing depth and the decrease in cost per base sequenced, high throughput sequencing technologies have emerged as preferred platforms for mRNA expression analysis of the complex mammalian transcriptome[9, 33].

A major limitation of current gold standard RNA sequencing approach[10] is the large amount of starting material (10 – 100 ng of mRNA) required to generate a sequencing library. This limits the potential of this protocol when it is difficult to obtain such large amounts of RNA such as in the fields of developmental biology or forensics or even for FACS sorted cell populations. Also, the standard RNA-seq protocol[10] maintains the relative order of transcript expression with a few highly expressed transcripts occupying majority of the sequencing space. This results in a poor coverage of low abundant transcripts at current sequencing depths[12, 13]. Reliable quantitation of the low abundant transcripts within large mammalian transcriptomes is further hampered by multireads and biases introduced by the transcript length[11] and the random hexamer primer hybridization[14]. A number of amplification-based protocols have been developed to address these issues such as "random priming" strategies[20-22], which utilize the hybridization and extension potential of hexamer/heptamer primers to amplify the starting material (mRNA or cDNA). However, the random priming methods often result in a low yield of good quality reads, due to mis-hybridization of primers or primer dimerization. Furthermore, these methods do not discriminate the regions of the transcriptome to amplify, a feature also shared by other uniform amplification based strategies[23-26, 34].

Here we describe a novel quantitative cDNA expression profiling strategy, involving the amplification of a majority of the mouse transcriptome using a defined set of 44 heptamer primers. The amplification protocol allows an efficient amplification of the majority of the expressed transcripts from as low as 50 pg of mRNA and was optimized to reduce mis-hybridization of primers and primer dimerization. We further explored the potential of our primer design strategy to selectively suppress the amplification of the highly expressing transcripts such as ribosome encoding transcripts. Our sequencing data demonstrated a significant reduction in the representation of the ribosomal transcripts with multiple choices of primer sets. We compared our methodology with a full-length cDNA amplification strategy (Smart-seq)[25] and observed comparable transcriptome coverage and similar technical noise. We implemented DP-seq on a model of embryological lineage segregation, achieved by graded activation of Activin A/TGFβ signaling in mouse embryonic stem cells (mESCs). The fold changes in transcript expression were in excellent agreement with quantitative RT-PCR and we observed a dynamic range spanning more than five orders of magnitude in RNA concentration with a reliable estimation of low abundance transcripts. Our transcriptome data identified key lineage markers, while the high sensitivity indicated that novel lineage specific transcripts anticipate the differentiation of specific cell types.

2.3 Results

2.3.1 Sequencing-library generation using heptamer primers based amplification

A novel cDNA sequencing-library generation methodology was developed to reliably represent the relative abundance of transcripts using limited amounts of mRNA. DP-seq consisted of three distinct phases (**Figure 2.1a**). In the first phase, we developed a primer design strategy that identified a defined set of 44 heptamer primers amplifying >80% of the mouse transcriptome (**Figure 2.1a**, green panel). This strategy incorporated known biases in PCR, namely the secondary structure of primer-binding sites in single stranded cDNA, GC content and the proximity to the 3' end of the transcript to identify potential primer-binding sites. Of the 16384 input sequences of heptamer primers, we selected primers with annealing temperatures between 16 - 25°C. To minimize mis-priming, heptamer primers starting with adenines at the 5' end and/or purine rich primers were filtered out. Next, an iterative randomized algorithm was implemented to identify 44 heptamer primers, which preferentially amplified unique regions of mouse transcripts (**Supplementary Fig. S2.1**). The primers were split into multiple sets ensuring no two primers had a mutual interaction energy (Gibbs free energy) greater than -5 kcal/mol in order to reduce primer dimerization. Of the 26566 known transcripts in the mouse NCBI RefSeq mRNA database, our heptamer primers covered 15072 (56.7%) transcripts uniquely.

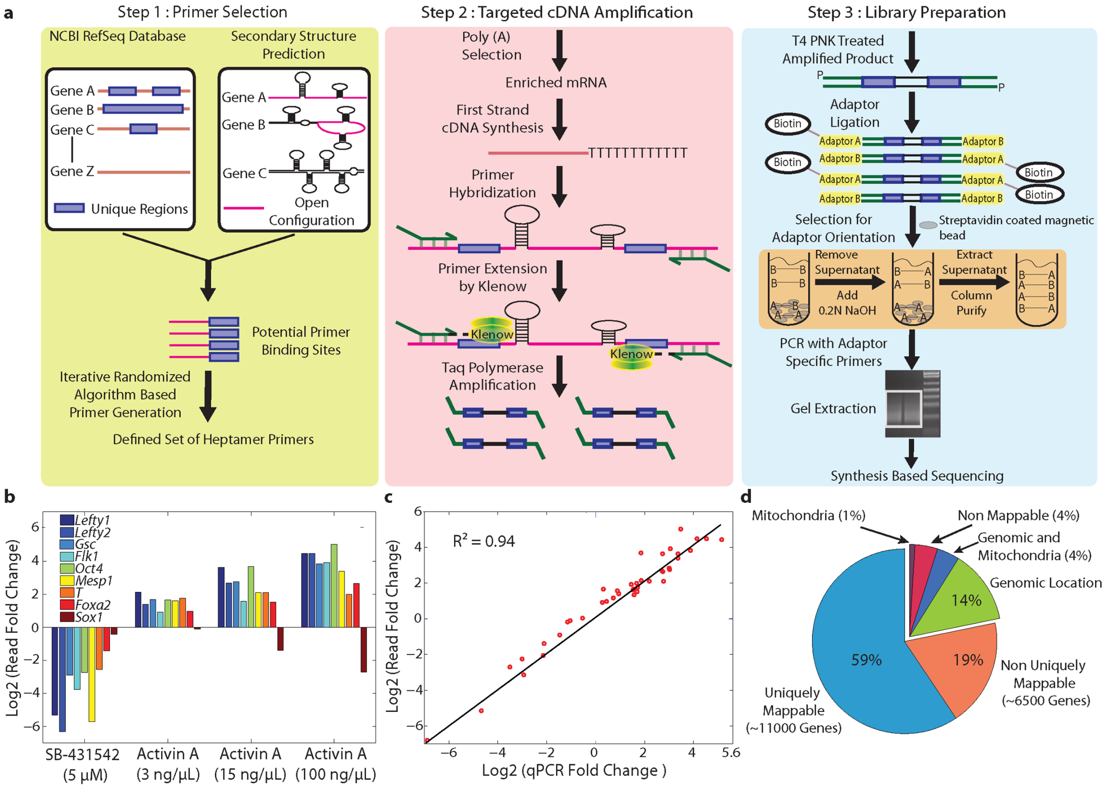


Figure 2.1: Schematic representation of sequencing library preparation using heptamer primers based amplification, DP-seq. (a) *Step 1: Primer selection* was based on identifying potential primer-binding sites that were less likely to form secondary structures and resided upstream to the unique regions on the mouse transcriptome. *Step 2: targeted cDNA amplification*. A Standard cDNA library was prepared and the primers selected from Step 1 were annealed to the single stranded cDNA library and were extended and amplified as indicated. *Step 3: Library preparation*. Illumina paired end adaptors were ligated to the ends of the amplicon library and the correct orientation of adaptors were selected. The library was further amplified using Illumina's paired end adaptor primers and were size selected for synthesis-based sequencing (b) Expression profiles of genes responding to graded activation of the Activin A/TGFβ signaling pathway in mouse embryoid bodies at day 4. Quantitative RT-PCR data was normalized with respect to untreated serum-free media controls. (c) The fidelity of amplification of the cDNA library using heptamer primers. Fold changes observed in 11 genes (from part (b), *Afp* and *Cer1*) across different dosages of Activin A showed perfect agreement with quantitative RT-PCR performed on cDNA (R^2=0.94; n=45). (d) Distribution of reads on the mouse genome.

In the second phase of the methodology, we performed a targeted amplification of the mouse transcriptome using the defined set of heptamer primers (**Figure 2.1a**, pink panel). This phase consisted of two components; (i) determination of the minimum length of the primer required to achieve efficient amplification and (ii) optimization of the amplification protocol to extend and amplify partially hybridized primers. We determined 14 bp (T_m~45-50°C) as the optimal length of the primers required to efficiently amplify regions of interest in the mouse transcriptome. As such the heptamer primers were extended by addition of a universal 7 bp sequence (5'-CCGAATA'-3') at the 5' end of heptamer primers. Standard PCR protocols failed to amplify partially hybridized primers because of low annealing temperatures of the last 7 bp, resulting in significant distortions in the expression level of low abundance transcripts. We therefore developed a novel protocol that uses a combination of mesophilic (Klenow polymerase) and thermophilic polymerases (Taq polymerase) to efficiently amplify regions of interest on cDNA. Klenow polymerase, which retains its optimal extension activity at 37°C, extends our partially hybridized primers (last 7 bp) at this temperature. The extended primers withstand the high temperature required for a Taq polymerase extension at 72°C, resulting in the formation of a double stranded amplicon library. These amplicons possess complementary sequences of the entire 14 bp of our primers at its ends. Since our 14 bp primers have a high Tm (45-50°C), they efficiently hybridize to the template and allow amplification of these amplicons during the subsequent cycles of Taq polymerase PCR.

In the last phase of the sequencing library generation, the amplicon library was 5' end phosphorylated and ligated to Illumina's adaptors (**Figure 2.1a**, blue panel). Since only distinct adaptor orientation fragments can be sequenced in Illumina's platform, we used a biotin-streptavidin chemistry to select the correct orientations of the adaptors. The fragments were later PCR amplified using Illumina's adaptor specific PCR primers and size selected for synthesis-based sequencing. The selection of fragments with a correct orientation of the adaptors can be skipped by ligating standard Illumina Y-adaptors to the amplicon library and using a custom sequencing primer that contains the universal tail sequence (5'-CCGAATA'-3') at its 3' end.

2.3.2 Evaluation of heptamer amplification-based transcriptomics

We implemented DP-seq on an *in vitro* cell culture based model of primitive streak (PS) induction in mESCs[35, 36]. Signaling by the TGFβ-family member Nodal through Activin receptor like kinase-4 is essential for mesoderm[37-39] and endoderm[40, 41] formation, and the dose-dependent induction of these tissues can be mimicked by treatment with Activin A. Various dosages of Activin A (3 ng/mL, AA3; 15 ng/mL, AA15; and 100 ng/mL, AA100) were therefore used to induce mesoderm and definitive endoderm while its inhibition by a small molecule inhibitor, SB-431542 (SB)[42], was used to induce neuro-ectoderm[43].

As expected, small doses of Activin A substantially induced mesodermal markers (e.g., *Kdr*, *Mesp1*) while higher doses of Activin A were required for the induction of anterior lineages including definitive endoderm (e.g., *Gsc*, *Foxa2*) (**Figure 2.1b**). On the other hand, complete inhibition of Activin A/TGFβ signaling caused an up-regulation of neuro-ectoderm markers (e.g., *Sox1*)[44]. Moreover, direct target genes (e.g., *Lefty1*, *Lefty2* and *T* also known as *Brachyury*)[45, 46] of the Activin A/TGFβ signaling pathway were regulated dose dependently. The differential expression of these low abundant genes DP-seq showed excellent concordance with quantitative RT-PCR (R^2=0.94, **Figure 2.1c**) validating the DP-seq approach.

For a typical transcriptome measurement, we obtained ~30 million reads per lane of Illumina's flowcell (**Table 2.1**). About 59% (18 million) reads uniquely mapped to more than 11000 transcripts with ≥10 reads. About 19% of the reads were non-uniquely mapped with a vast majority of them mapping to isoform groups. Another 18% of the reads (71% uniquely) mapped to genomic locations (excluding the open reading frames of known transcripts) and mitochondria transcripts (**Figure 2.1d**). Of these genomic reads, 72% mapped to intronic regions of transcripts while another 20% mapped within 5 Kb of the known transcripts. These reads most likely represent non-coding RNA, since we did not see a strong correlation between the fold changes in intronic reads with those from proximal exons.

The experimental data indicated expression of more than 100,000 different primer-binding sites representing ~ 18,000 known transcripts. This demonstrates the scale of massive multiplexing achieved by DP-seq. On average, we obtained expression of 10 different primer-binding sites for each expressed transcript. Notably, each site provided an independent measurement of relative abundance serving as technical replicates for the experiment (**Supplementary Fig. S2.2**).

Table 2.1: Mapping Summary of the sequencing experiment. Reads were first aligned to the NCBI mRNA RefSeq database allowing up to 2 mismatches. Unmapped reads were later aligned to the mouse genome including mitochondria. Multireads refer to reads that mapped to more than one transcript/genomic locations. TR refers to technical replicates.

	Lane 1 Serum Free Media	**Lane 2** SB-431542	**Lane 3** Activin A (3 ng/mL)	**Lane 4** Activin A (15 ng/mL) TR1	**Lane 6** Activin A (15 ng/mL) TR2	**Lane 7** Activin A (100 ng/mL)
Total reads	33.4M	35.2M	32.8M	29.4M	25.1M	30.0M
Unique reads (mRNA Refseq)	58.20%	56.80%	59.20%	59.10%	59.50%	58.20%
Multireads (Isoform group only, mRNA Refseq)	13.52%	13.37%	13.45%	13.20%	13.19%	13.05%
Multireads (mRNA refseq)	5.47%	5.63%	5.45%	6.20%	5.71%	5.45%
Genomic (Unique)	12.16%	13.33%	10.63%	10.76%	11.01%	12.16%
Genomic (Multireads)	2.10%	2.12%	1.98%	1.98%	2.03%	2.09%
Genomic and Mitochondria	2.49%	3.51%	4.52%	4.41%	3.92%	4.52%
Mitochondria (Unique)	0.59%	0.64%	1.06%	0.74%	0.75%	0.84%
Unmappable	5.38%	4.44%	3.61%	3.47%	3.73%	3.59%
Transcripts (Unique reads>=10)	11792	11565	11508	11409	11097	11401
Transcripts (Multireads >=10)	6401	6293	6329	6265	6167	6215
Binding Sites (Unique reads>=10)	126844	125775	117587	110069	96060	109109

More than 50% of the uniquely mapped reads came from perfectly matched primer-binding sites while the rest were the product of mis–priming or single nucleotide polymorphisms (SNPs) in the primer-binding sites. Fold changes observed in predicted and mis-primed binding sites were highly correlated (R^2=0.88) suggesting that mis-primed PCR products were able to conserve the relative abundance of transcripts (**Supplementary Fig. S2.2**). Mis-primed products were mainly stabilized by a favorable interaction between the last three bases of the universal tail of the heptamer primers (5'-ATA-3') and the upstream regions of the primer-binding sites (**Supplementary Fig. S2.3**). Finally, we observed no indication of primer – dimerization.

Analysis of the technical replicates revealed a strong correlation in quantitative transcript expression (R^2=0.96, **Figure 2.2a**). To assess the dynamic range, we spiked the untreated control (serum free media, SFM) with six artificial transcripts of the yeast POT1 promoter (~ 180 bp). The transcripts were flanked with different heptamer primer-binding sites and mixed in different dilutions, spanning six orders of magnitude in RNA concentration. The second most abundant transcript was similar in expression with the β-actin abundance in our biological samples. Our primers were able to effectively amplify all the six transcripts and maintained their relative abundance (R^2=0.99, **Figure 2.2b**). The distributions of fold changes (**Supplementary Fig. S2.2**) observed in all possible pairwise comparisons of the samples was broad (2^{-8} – 2^{10}) suggesting a much higher dynamic range in comparison to microarray platforms (few hundred folds)[8].

We next prepared serial dilutions of mRNA from 10 ng to 1 pg (10000 fold depth) of mRNA and constructed sequencing libraries to determine the lowest amount of mRNA required to prepare reliable sequencing library. The number of amplification cycles was increased for lower dilutions to achieve appropriate amounts of DNA for the library construction. The transcript measurements from the technical replicates consistently showed high correlations for the libraries prepared from 10 ng – 50 pg of mRNA (**Figure 2.2c**). The transcriptome coverage remained high even for libraries prepared from 1 pg of mRNA (~ 6000 transcripts; **Supplementary Table S2.1**), although the noise in the quantification of the transcripts increased substantially (**Supplementary Fig. S2.4**). We further investigated whether the transcript measurements were conserved within the dilution series of mRNA. Global transcript measurements of libraries constructed from at least 50 pg of mRNA showed high correlation with the libraries constructed from 10 ng of mRNA (**Supplementary Fig. S2.5**). Sequencing libraries constructed from 1 pg of mRNA showed significant deviations in measurements of low copy number transcripts from 10 ng of libraries and a considerable amount of spurious PCR artifacts were observed.

Our methodology exhibited few biases arising out of each stage of cDNA amplification (**Supplementary Fig. S2.3**). The most dominant bias came from local secondary structures of the single stranded cDNA. Regions with stable secondary structures prevented the heptamer primer-binding sites from hybridizing with their corresponding heptamer primers, resulting in their poor representation in the sequencing library. There was also an inherent bias towards preferential amplification of fragments with shorter lengths and lower GC content, which are known to be associated with Taq Polymerase amplification and have been reported in other multiplexed PCR strategies[47]. Finally, we observed that the majority of the experimental heptamer primer-binding sites resided in proximity to the 3’ end of the transcripts mainly because of the inability of the reverse transcriptase to produce full-length cDNA.

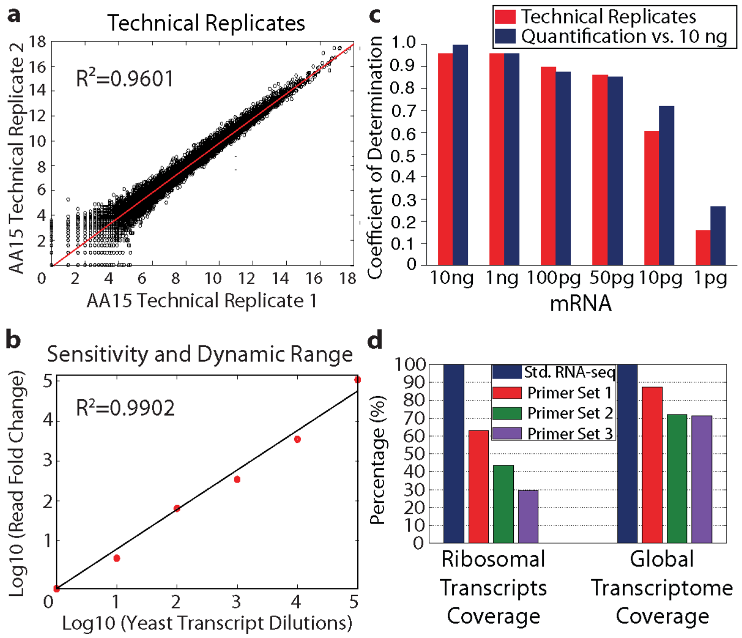


Figure 2.2: Performance of DP-seq. (a) Comparison of two Activin A 15 ng/ml dosage replicates (R^2=0.96). (b) Six *in vitro* synthesized transcripts derived from the yeast POT1 promoter with a length of 180 bp were added to untreated control cDNA at varying concentrations with six orders of magnitude. The reads obtained from the transcripts revealed a fold change of up to 10^5 (R^2=0.99) in comparison to the lowest abundant transcript. (c) Sequencing libraries constructed from serial dilutions of mRNA exhibited high correlations within the technical replicates. Libraries constructed from at least 50 pg of mRNA showed high correlations (R^2) in global expression measurements with the libraries made from 10 ng of mRNA. (d) Suppression of the ribosomal transcripts representation in the sequencing library generated from three different primer sets. The global transcriptome coverage remained high for all primer sets.

To determine whether DP-seq is capturing the majority of the expressed transcripts, we performed standard RNA-seq (Std. RNA-seq) on the AA3 sample using the protocol adopted from Mortazavi et al., 2008[10]. We observed comparable transcriptome coverage with DP-seq libraries, representing > 80% of the expressed transcripts. Analysis of the technical replicates obtained from DP-seq and Std. RNA-seq revealed a similar noise structure (**Supplementary Fig. S2.6**). The PCR biases observed in our methodology distorted the order of transcript expression within a biological sample (**Supplementary Fig. S2.6**) resulting in a similar or enriched representation of the majority of low expressed transcripts (Reads Per Kilobase per Million mapped reads (RPKM) <=10 in the Std. RNA-seq library). However, the relative abundance of the transcripts across different biological samples was not affected (shown in **Figure 2.1c**) as these biases are expected to be similar for a given transcript across different biological samples. Furthermore, we observed an overlapping distribution of unique reads for the transcripts encoding transcription factors (http://genome.gsc.riken.jp/TFdb/) between the two protocols (**Supplementary Fig. S2.7**).

We then investigated a novel aspect of our primer design strategy where we incorporated the PCR biases observed in our protocol to suppress the representation of highly expressed ribosomal transcripts, while maintaining the overall transcriptome coverage. Transcripts encoding 81 ribosomal proteins occupied about 9% of the sequencing space in the Std. RNA-seq library prepared from the AA3 sample. Detailed analysis of the PCR biases led us to propose heuristics on favorable amplification by our heptamer primers. Amplicons with heptamer primer-binding sites in open configuration (<-4 Kcal/mol); significant tail interaction (>=2 bp interaction between the last four bases of the universal tail and the cDNA template); low GC content (<0.55) and short fragment lengths (<300 bp) were heavily penalized for the ribosomal transcripts. We designed three different primer sets and generated sequencing libraries from the AA3 sample. Our sequencing data revealed up to 70% reduction in the representation of the ribosomal transcripts while the global transcriptome coverage remained high for all primer sets (**Figure 2.2d**). Furthermore, the overall distribution of the reads coming from the transcription factor family also exhibited similar distribution for a representative primer set (**Supplementary Fig. S2.7**). This data demonstrates the potential of our designed primer based strategy to preferential suppress the representation of the transcripts of interest (e.g. highly expressed transcripts) and distinguishes it from other uniform amplification based strategies.

2.3.3 Comparison with a different PCR-based RNA-Seq method

We performed a thorough comparison of our methodology with Smart-seq[25], which performs full-length cDNA amplification from limiting amounts of mRNA. Sequencing libraries were generated from 50 picograms of mRNA derived from Activin A (3 ng/mL and 100 ng/mL) treated samples using DP-seq and Smart-seq. The same samples were also used to generate Std. RNA-seq libraries from 10 ng of mRNA. The libraries prepared from both methods were highly reproducible and displayed strong correlations in the expression measurements of the transcripts in the technical replicates (**Supplementary Fig. S2.8**). Both DP-seq and Smart-seq exhibited similar transcriptome coverage (**Supplementary Table S2.1**) and overlapping noise in the quantification of the transcripts (**Figure 2.3a**). However, the transcriptome coverage obtained in either method was significantly lower than that of Std. RNA-seq libraries with the majority of low expressed transcripts (average RPKM<3 in the Std. RNA-seq library) showing stochastic loss. Consequently, the distribution of unique reads for the low expressed transcripts was shifted towards a low read number (**Figure 2.3b**). A similar observation was made for moderately expressed transcripts (average RPKM between 3 and 300 in the Std. RNA-seq library) with DP-seq and Smart-seq libraries displaying an overlapping distribution of unique reads (**Supplementary Fig. S2.9**). Our mapping analysis revealed a significant length bias in Smart-seq sequencing libraries resulting from poor amplification of long cDNA species (>4 Kb). This was not observed with DP-seq as it performs amplification of selected regions of cDNA irrespective of its length, thus resulting in higher representation of a vast majority of the long transcripts (>77%; **Figure 2.3c**). Expression measurements of differentiating mESCs treated with a higher dose of Activin A (100 ng/mL) showed comparable up-regulation of mesendoderm markers (Cer1, Lefty1, Lefty2, Foxa2, Gsc etc.) and down-regulation of mesoderm and ectoderm genes, implying the conservation of the biological context in the sequencing libraries prepared from 50 pg of mRNA with either method (**Figure 2.3d**).

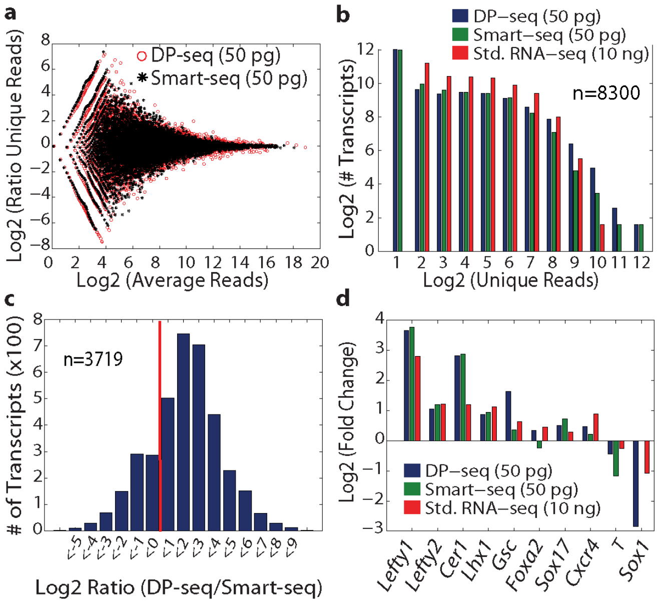


Figure 2.3: Comparison of DP-seq with Smart-seq on Activin A treated samples (AA3 and AA100). (a) MA plot of technical replicates obtained from AA100 sample showed similar technical noise in the two methods. (b) Distribution of unique reads for the low expressed transcripts (RPKM<3 in Std. RNA-seq library prepared from AA100 sample) obtained in the three methods. The majority of the low expressed transcripts did not show expression in the libraries constructed from 50 pg of mRNA in DP-seq and Smart-seq. (c) A length bias in Smart-seq resulted in higher reads for the long cDNA species (>4 Kb) in the DP-seq libraries. (d) Comparable fold changes were observed for the known lineage markers in the three methods between AA100 and AA3 samples. The amount of mRNA used for sequencing library generation is shown in parentheses.

We next sought to compare the differential gene expression observed in DP-seq, Smart-seq and Std. RNA-seq for the two Activin A dosages. Differentially expressed transcripts were identified by generating the null distribution from the technical replicates. The null distribution for Std. RNA-seq libraries showed little technical variation; as such a large proportion of differentially expressed transcripts were identified. The majority of these transcripts were expressed at low copy number. Smart-seq and DP-seq identified a comparable number of differentially regulated transcripts (1414 and 1297 respectively), however only a small proportion of them were common between the two methods (**Supplementary Fig. S2.9**). Pairwise comparison of these methods with Std. RNA-seq revealed 56% overlap of the differentially expressed transcripts. Only 191 differentially regulated transcripts (common set) were common in all three methods. We found however that the differentially regulated transcripts that were method-specific are low expressed and were prone to large noise as these transcripts showed lower RPKM distributions as compared to the common set (**Supplementary Fig. S2.9**). Further analysis of the fold changes observed for the common set in DP-seq and Smart-seq libraries showed strong correlations; however, they were poorly correlated with the fold changes observed in Std. RNA-seq libraries (R^2=0.6456 for DP-seq and R^2=0.5740 for Smart-seq). This highlights the issues caused by the increased noise in the quantification of low copy number transcript measurements, which is further amplified when using low amounts of input material.

2.3.4 Graded activation of the Activin A/TGFβ signaling pathway in mESCs

Mouse ESCs were differentiated in serum-free conditions in the presence of varying doses of Activin A and SB and the mRNA was profiled at day 4 (equivalent to 6.5 – 7.5 dpc) using DP-seq (**Figure 2.4a**). The differential gene expression analysis revealed a stepwise increase in the number of transcripts differentially regulated as mESCs responded to the gradient of Activin A. The most transcriptional diversity was observed between SB and AA100 samples corresponding to the two extreme states of pathway activation. By mapping those transcripts to known Activin A/TGFβ pathway components using Ingenuity pathway analysis (Ingenuity® Systems, www.ingenuity.com), we observed a substantial down-regulation of many of these genes in response to pathway inhibition via SB (**Figure 2.4b**) whereas Activin A up-regulated these genes.

Graded activation of Activin A/TGFβ signaling pathway allowed us to identify putative TGFβ regulated genes during early differentiation of mESCs (**Figure 2.4c**). Potential TGFβ target genes were predicted based on (i) the opposing modulations in SB and AA3 conditions (in comparison to untreated control) and (ii) the subsequent up/down regulation with higher dosages of Activin A. We identified many of the expected TGFβ target genes, including *Cer1*[48], *Lefty1*[46], *Lefty2*[46], *Foxa2*[49] and *T*[45] (**Figure 2.4c**, bold). Not all expected genes were found because they either did not meet our stringent classification criteria (e.g. *Nodal*[45], *Nanog*[50]) or they were not expressed in this cellular context. More importantly, we have identified transcripts that respond similarly to the graded Activin A/TGFβ pathway modulation, which have not been linked previously to the pathway. Promoter analysis of these transcripts revealed the presence of multiple FoxH1 binding sites[51-53] (Asymmetric Elements, ASE) within 10 Kb upstream and downstream of the transcription start site supporting our hypothesis that the Activin A/TGFβ signaling pathway regulates the expression of these transcripts.

2.3.5 Lineage segregation is achieved by regulation of Activin A/TGFβ signaling

Our preliminary experiments with T-GFP mESCs (GFP driven by Brachyury/T promoter) showed negligible induction of GFP^+ cells at day 4 of differentiation upon treatment with SB. The untreated control condition (SFM) naturally drives mESCs to neuro-ectoderm lineage with only 5-10% GFP^+ cells. However, in presence of mesoderm inducing factors such as Activin A (3 ng/mL), > 60% of the cells were GFP^+ demonstrating efficient induction of mesoderm (**Supplementary Fig. S2.10**). Neuro-ectoderm associated transcripts were classified as transcripts significantly up-regulated in SB and SFM in comparison to AA15 and comprised of known neuro-ectoderm markers (*Sox1*, *Sox2* and *Pax6,* **Figure 2.5a**). We then performed GO term (biological process annotation) enrichment and KEGG pathway enrichment to validate our classification (http://david.abcc.ncifcrf.gov/). Biological processes associated with neuron differentiation and morphogenesis (**Supplementary Table S2.2**) were enriched in the transcript list with the Wnt and Activin A/TGFβ pathway significantly represented (**Supplementary Table S2.3**).

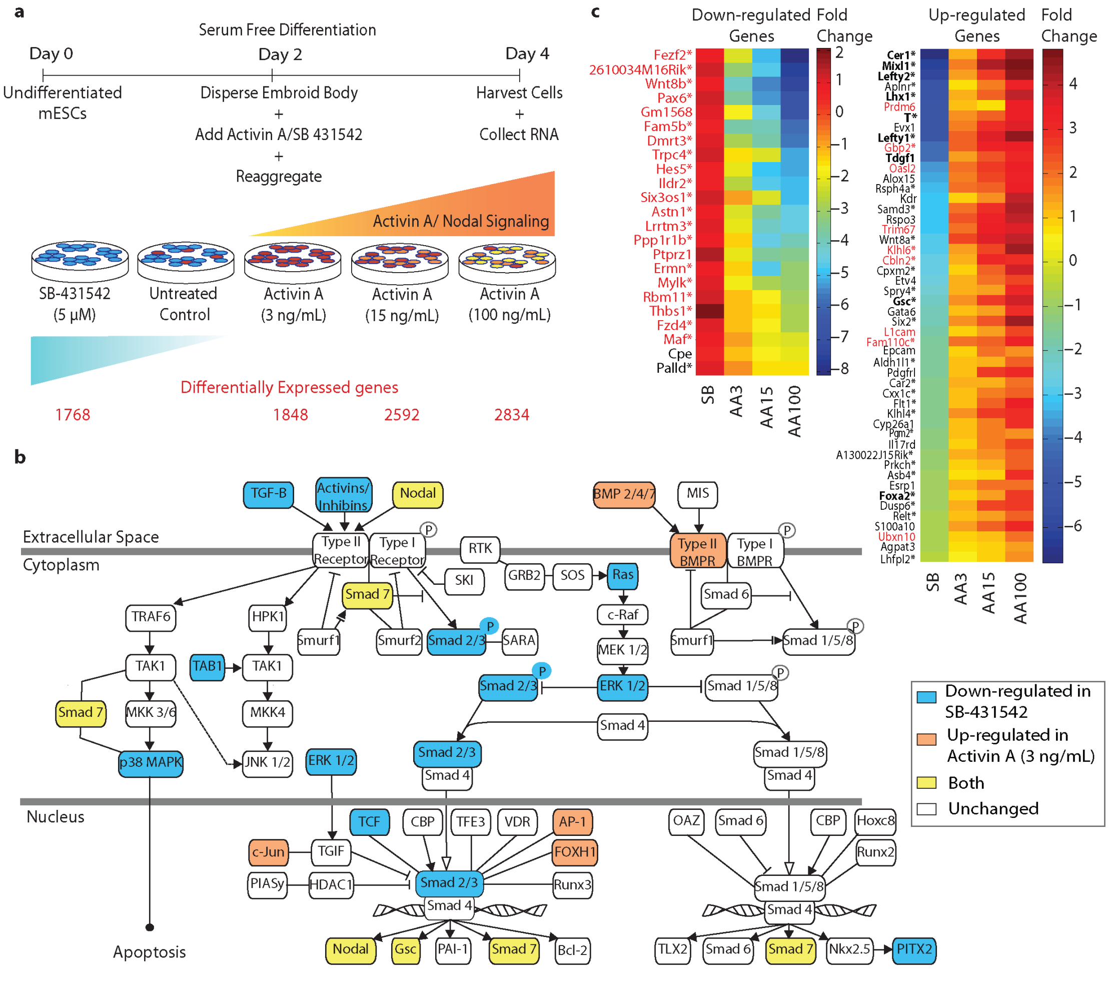


Figure 2.4: Graded expression of putative target genes of the Activin A/TGFβ signaling pathway in day 4 mESCs. (**a**) Schematic representation of the experimental setup. Mouse ESCs were differentiated in serum free conditions and different dosages of Activin A and SB-431542 were introduced to create a graded activation of the Activin A/TGFβ signaling pathway. Cells were harvested at day 4 for sequencing library generation. Differential gene expression analysis identified ~15 – 20% of expressed transcripts as differentially regulated in each sample in comparison with untreated controls (see **Supplementary Methods** online). (**b**) Regulation of Activin A pathway components in response to SB-431542 and Activin A. (**c**) Putative TGFβ target genes in differentiating mESCs at day 4. The heat map corresponds to fold changes observed for transcripts in comparison to untreated control. Putative target genes were classified as transcripts that followed opposite trends of regulation upon treatment with Activin A and SB. Fifty transcripts were successively up-regulated while 23 transcripts followed graded down-regulation with increasing dosages of Activin A. The majority of the TGFβ target genes (marked with *) had FoxH1 transcription factor binding sites separated by 30 – 200 bp (also called ASE) in 10 Kb upstream and downstream of the transcription start site. Known TGFβ target genes are highlighted in bold. Low copy number transcripts (RPKM<3 in AA3 sample) are displayed in red font.

To correlate some of the novel neuro-ectodermal transcripts with embryology, we searched the MGI gene expression database for the expression patterns of the identified transcripts throughout all stages of mouse embryonic development. Expression of the vast majority of the neuro-ectoderm associated transcripts were not reported in embryonic day 6.5 – 7.5 embryos, the stages that correspond to the studied mESC derived samples. A number of these transcripts, however, were expressed in neuro-ectoderm derivatives at later stages of development. To validate the early expression of these transcripts in the neuro-ectoderm lineage, we used Wnt pathway inhibition (IWR-1)[54] as an alternative to induce neuro-ectoderm and confirmed the up-regulation of a number of these neuro-ectoderm associated transcripts (**Figure 2.5b** and **Supplementary Fig. S2.11**). On the other hand, transcripts significantly up-regulated in AA15 in comparison to SB and SFM were designated as PS associated transcripts. The list included a number of known mesoderm and endoderm markers (*T*, *Mesp1*, *Foxa2* and *Sox17*). GO enrichment analysis (http://david.abcc.ncifcrf.gov/) revealed biological processes associated with gastrulation, tissue morphogenesis and tube development (**Supplementary Table S2.2 and S2.3**).

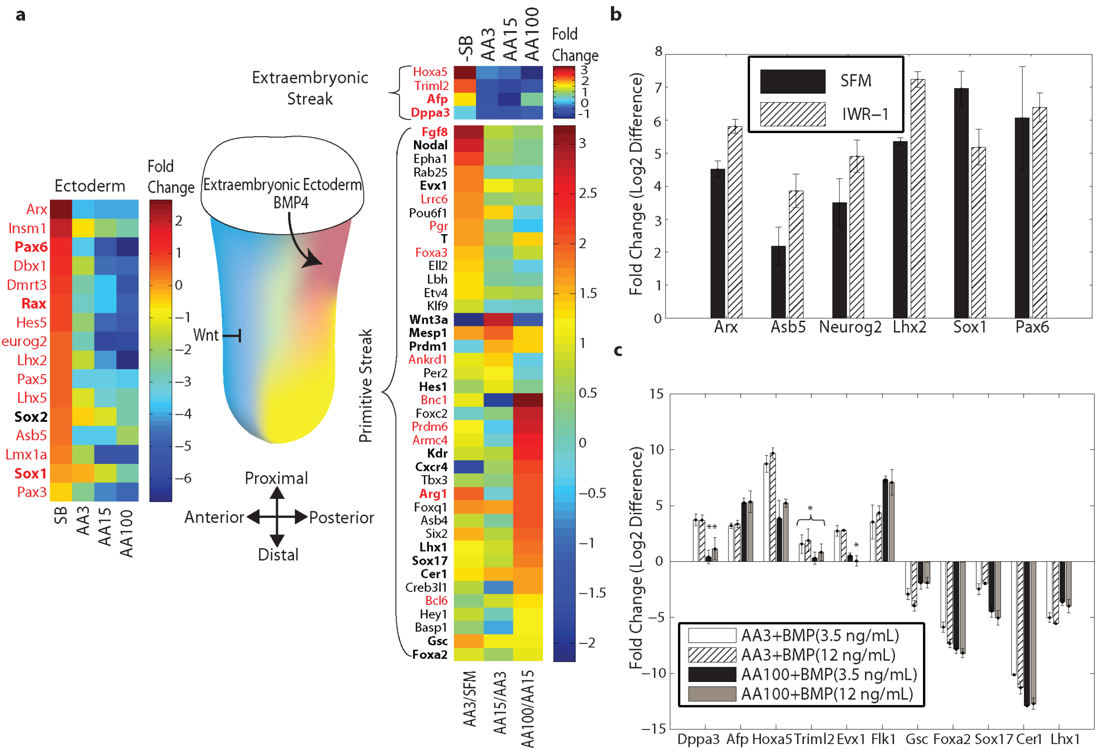


Figure 2.5: Lineage segregation between neuro-ectoderm and PS (mesoderm and definitive endoderm) achieved by modulation of Activin A/TGFβ signaling pathway. (a) Schematic of the mouse embryo at embryonic day 6.5-7.5 with the gradient of Nodal expression (yellow) with the maximum expression observed in the anterior tissue. Through inhibition of TGFβ signaling pathway cells commit to the neuro-ectoderm lineage (blue). A heat map of the neuro-ectoderm associated genes is depicted (left of the embryo) with their fold changes in different samples in comparison to untreated control. The heat map on the right of the embryo depicts successive fold changes of the PS markers with varying dosages of the Activin A. The transcripts with the highest fold change in AA100 in comparison with AA15 are enriched for definitive endoderm and other anterior tissue markers. Other PS transcripts are expected to have diffused expression pattern all throughout the streak. Genes with known expression in Theiler Stage 9-11 of mouse embryo are highlighted in bold (MGI database). Low copy number transcripts (RPKM<3 in the AA3 sample) are displayed in red font. (b) Small molecule inhibition of Wnt signaling pathway (IWR-1) induced the neuro-ectoderm lineage. The fold changes are normalized to the AA3 sample. (c) BMP4 enhanced expression of posterior and extraembryonic mesoderm markers at the expense of anterior markers. Quantitative RT-PCR fold changes for two BMP4 dosages are normalized with respect to Activin A alone induction. Error bars represent the standard deviation in biological replicates (n=3). Asterisks indicate p>0.05 (Student's t test) compared to controls.

Graded Activin A/TGFβ signaling has been shown to induce different mesoderm and endoderm tissues, correlating with the anteroposterior position of progenitors within the PS, with the highest levels of signaling corresponding to anterior most located progenitors[55-57]. Transcripts with a maximum fold change between AA100 and AA15 in comparison to other two fold changes (AA3/SFM and AA15/AA3) should mark anterior PS derivatives, and in our experiments indeed comprised definitive endoderm markers. Conversely, the majority of the transcripts with maximum fold changes in AA3/SFM and AA15/AA3 were expected to have a diffused expression pattern throughout the PS (**Figure 2.5a**), which was confirmed by reported *in-situ* hybridizations for some of these transcripts[58]. To further validate our classification, we studied some of these new transcripts by posteriorizing Activin A induced-mesoderm with BMP4[59-61]. Transcripts known to be expressed in the extra-embryonic mesoderm and the extreme posterior PS were indeed enriched and anterior PS transcripts were significantly down-regulated (**Figure 2.5c**). Pan-PS transcripts also exhibited down-regulation by BMP4 suggesting a dominant posteriorization effect of BMP4 signaling (**Supplementary Fig. S2.11**).

2.4 Discussion

Sequencing library generation from low amounts of starting material has remained a challenge for most of the existing RNA – seq protocols. Random priming strategies amplify from low amount of RNA, however, reliable quantitation of low abundant transcripts is not regularly obtained. In our initial experiments with a random priming strategy[22], primer-dimerization and mismatches in the primer-binding sites resulted in majority of the reads mapping to multiple mRNA species. Only 18% of the reads mapped uniquely to the transcriptome and low abundant transcripts were significantly under-represented because of a poor dynamic range. The methodology presented in this work addresses these issues by facilitating generation of reliable sequencing libraries from as low as 50 pg of mRNA. The dynamic range of our protocol exceeded five orders of magnitude in RNA concentrations allowing a more reliable detection of the majority of the low expressed transcripts.

Primer design was a critical component of DP-seq. The ubiquitous presence of heptamer primer-binding sites on the mouse transcriptome was utilized to amplify more than 80% of known transcripts (**Supplementary Fig. S2.2**) from a small set of 44 heptamer primers. We optimized PCR conditions for heptamer hybridization to achieve successful amplification of more than 50,000 different fragments representing ~18,000 transcripts in the mouse Refseq mRNA database. A number of considerations were made while determining the base composition of primers to reduce mis-priming and primer dimerization. As a result, majority of the reads (55%) came from perfect binding of the primers while another 38% had one mismatch in primer-binding site. This enabled us to use the entire read length for alignment to the mouse transcriptome.

Our transcriptome data demonstrated excellent reproducibility and sensitivity. We were able to reliably estimate up to a 2^{16}-fold change in transcript expression from limiting amounts of mRNA. Furthermore, fold changes observed in low abundant transcripts were in perfect agreement with quantitative RT-PCR.

Technical replicate data revealed comparable noise in the quantification of transcript expression with respect to standard RNA-seq protocols. Furthermore, the global measurements of transcript expression of libraries constructed from at least 50 pg of mRNA showed high correlations with the library made from 10 ng of mRNA.

A standard RNA-seq approach[10] requires at least 10 – 100 ng of mRNA for reliable library generation. To address this issue, a number of protocols[23-25] were recently developed. DP-seq offers a cost effective way of generating reliable sequencing library from limiting amounts of mRNA. The cost of amplification only includes a one-time purchase of 44 primers (14 bp) that are sufficient to generate hundreds of sequencing libraries. Our protocol is compatible with regular first strand cDNA synthesis kits and the polymerases used in our protocol (Taq and Klenow polymerase) are cheap and readily available. The processing time required for the generation of a sequencing library is also short, as DP-seq library preparation does not require fragmentation of the cDNA library or poly-adenylation of the 3’ end of the amplicon library. A direct comparison of DP-seq with Smart-seq revealed comparable transcriptome coverage and similar technical noise in the quantification of the low expressed transcripts. Furthermore, DP-seq does not suffer from length bias and provides higher representation; hence better quantification of the long cDNA species in the sequencing library. DP-seq primers amplify select regions of the known transcriptome as such the sequencing libraries are devoid of the information regarding RNA structure (exon usage, TSS, etc.) or uncharacterized transcripts.

Typical RNA-seq protocols do not discriminate against high abundant transcripts. Consequently, most of the sequencing effort is spent on a small number of highly abundant transcripts[62]. We exploited the PCR biases observed in our protocol to reduce the representation of ribosomal transcripts by designing primers that have less likelihood of hybridizing efficiently to these transcripts. Complete elimination of the ribosomal transcripts was not achieved because of the mis-priming of the heptamer primers. It would be desirable to reduce mis-priming seen in our approach, and further refinements in the design strategy to address above issues are in progress.

The increased sensitivity of our methodology allowed us to detect known transcripts that had only been associated with later stages of germ layer segregation. These findings are of interest since it supports the view that low-level expression of lineage specific transcripts precedes overt manifestation of lineage phenotype, at least as traditionally assayed. This might not be surprising, since lineage commitment probably involves making chromatin of lineage specific transcripts accessible to transcriptional machinery, and might result in low-level transcription. Indeed, recent work on the analysis of activation marks in the promoters of differentiation specific transcripts has demonstrated that promoter activity is detected well before established landmarks of differentiation are achieved[63, 64]. It will be very interesting to explore this idea further, at the single cell level, to determine when and how this early transcriptional activation determines germ line specification.

2.5 Materials and Methods

2.5.1 Primer Design

Heptamer primer-binding sites are ubiquitously present in the mouse transcriptome enabling the selection of a small set of heptamer primers to cover more than 80% of the mouse transcriptome. Moreover, while hexamer primers have a low range of annealing temperatures, heptamer primers hybridize with greater efficiency to allow Klenow polymerase to extend these primers and perform efficient amplification.

We first implemented a suffix array data structure to identify 32-mer unique regions in the mouse transcriptome. All suffixes in the suffix array were divided into disjoint segments using 32-mer sequences. For each segment, we then identified all related segments that possess up to 2 mismatches with the 32-mer sequence. If the segment and all of its related segments contained suffixes mapping to only one transcript, then the segment was designated as unique. Next, we predicted the local secondary structures of the known transcripts as stable secondary structures were expected to shield heptamer primers from hybridizing to their primer-binding site. For each transcript in the Mouse NCBI RefSeq mRNA database, we ran a window of 47 bp along the transcript length and determined its propensity to form stable secondary structure using UNAfold software[65]. Gibbs free energy (ΔG) was estimated at 37°C for standard PCR buffer conditions (2 mM $MgCl_2$ and 50 mM NaCl). Regions with a $\Delta G \geq -4$ kcal/mol were considered to be available for heptamer primer hybridization (open configuration).

We combined the two datasets and identified all heptamer primer-binding sites, (i) flanking unique regions on mouse transcriptome and (ii) residing in open configuration. We then implemented an iterative randomized algorithm (**Supplementary Fig. S2.1**) to identify a defined set of heptamer primers forming valid amplicons for >80% of the mouse transcriptome. We defined a valid amplicon as follows:

1. It has a length between 50 and 300 bp.
2. Both, forward and reverse primer-binding sites are in open configuration.
3. At least one of the primer-binding sites must have a ΔG ≥-2 Kcal/mol.
4. A 32 bp unique region should follow one of the primer-binding sites.
5. The GC content of the amplicon should not exceed 58%.
6. The amplicon must be within 5 Kb of the 3’ end.

Using this approach, we identified 44 unique primers, which were split into 3 sets to reduce primer-dimerization (**Supplementary Table S2.4**). This configuration covered ~80% of transcripts with 57% of transcripts covered uniquely. More than 170000 valid amplicons were predicted from 201242 primer-binding sites. The three primer sets used for suppressing the representation of the ribosomal transcripts are detailed in **Supplementary Table S2.5**.

2.5.2 cDNA preparation

Total RNA was extracted from harvested cells using Trizol (Invitrogen). About 1-5 ug of total RNA was later subjected to Oligo(dT) selection using Oligotex mRNA Mini Kit (Qiagen) according to the manufacturer's protocol. If the total RNA is less than 1 ug, we recommend using Dynabeads mRNA Purification

Kit (Invitrogen) for extraction of poly-adenylated RNA. Next, first strand cDNA was synthesized with oligo dT (20-mer) primers using QuantiTect Reverse Transcription Kit (Qiagen) according to manufacturer's instructions. This kit allows synthesis of full-length cDNA (as long as 10 Kb). The reaction was later purified using Agencourt AMPure XP system (Beckman Coulter) according to manufacture's protocol and eluted in 20 ul of elution buffer (EB).

2.5.3 Primer hybridization and extension

Heptamer primer hybridization and extension was achieved by using Klenow (exo-) polymerase, a mesophilic polymerase with strand displacement capability. Exo-nuclease deficient version of Klenow polymerase was used to avoid degradation of heptamer primers. Since the 44 heptamer primers were split into three different primer sets, a master mix was prepared comprising of 1 – 5 ng of cDNA, Taq polymerase buffer (10X) supplemented with 2.5 mM $MgCl_2$, 4% DMSO and 0.2 mM dNTP (10 mM stock). DNase free water was added to make the total reaction volume of 24 µl. The master mix was split equally into three PCR reaction tubes. Later 1 µl of heptamer primer mixes were added to their respective tubes. The reaction mix was incubated at 95°C for 5 mins to denature the cDNA template. Mis-hybridization of the heptamer primers was minimized by ramping down the temperature of reaction mix to 37°C at the rate of -0.2°C/sec. At this point, 1 unit of Klenow polymerase (exo-) was added to each reaction tube and incubated for 30 mins at 37°C and then 5 mins at 42°C. Klenow polymerase retained most of its activity in Taq polymerase buffer and its extension rate was not affected at 2.5 mM $MgCl_2$ concentration, as reported earlier[66].

2.5.4 Taq polymerase amplification

Taq polymerase possesses optimal affinity for DNA (K_m ~ 2 nM) allowing efficient amplification of the PCR products while avoiding primer dimerization. Moreover, Taq polymerase allowed the addition of tail dATP at the 3' end of most of the amplicons thus eliminating this step from sequencing-library generation. A PCR master mix was prepared containing: 2 μl of Taq reaction buffer (10X), 1.25 mM of $MgCl_2$, Buffer Q (5X, Qiagen), 2 μl of primer mix (2 μM stock), 0.2 mM of dNTPs (10 mM stock) and 2 units of Taq polymerase. DNase free water was later added to top up the reaction mix to 20 μl. Similar reaction mixes were prepared for the other reaction tubes. Later, the reaction mix was added to the Klenow reaction (30 μl of total volume) and a 14-cycle amplification was performed consisting of denaturation (95°C for 30 s), annealing (46°C for 30 s) and elongation (72°C for 40 s). The amplified libraries obtained from the three tubes were pooled together and purified using Agencount AMPure XP system. The amplicon library was eluted in 44 μl of EB.

2.5.5 End Repair

The 5' ends of the PCR products were phosphorylated using T4 Polynucleotide Kinase (PNK) enzyme (NEB) in the presence of T4 DNA Ligase buffer containing ATP. The T4 PNK treatment was set up as follows:

Amplicon library: 44 μl

T4 DNA ligase buffer: 5 μl

T4 PNK: 1 μl (10 units)

The reaction was incubated at 37°C for 30 mins. Later, the reaction was purified using Agencourt AMPure XP system and eluted in 15 µl of EB.

2.5.6 Ligation

Custom adaptor oligos were ordered in 100 µM concentration (Valuegene Inc.) with following modifications:

a) Adaptor_A_F

5'- Biotin-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCT-S-T -3’

(-S- represents Phosphorothioate Modification; 5’ end of the oligo is biotinylated)

b) Adaptor_A_R

5’- Phospho-AGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT -3’

(5’ end of the oligo is phosphorylated)

c) Adaptor_B_F

5’- CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCT-S-T – 3’

d) Adaptor_B_R

5’ – PhosphoAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTGCTTG -3’

Adaptor oligos referring to adaptor A (a, b) and adaptor B (c, d) were mixed in equi-molar concentrations and diluted to 2 µM final concentration. Both adaptors were later denatured at 95°C for 5 mins and then brought back to room temperature gradually at -0.2°C/s. This allowed hybridization of the two oligos of the adaptor with ‘T’ overhang. The adaptor mix was further diluted 1:10 to get a stock concentration of 200 nM. The Ligation reaction was set up as follows:

T4 PNK treated PCR product: 6 µl

Adapter A: 1 µl

Adapter B: 1 µl

T4 DNA Ligase Buffer: 1 µl

T4 DNA Ligase (NEB): 1 µl (400 units)

The reaction was performed at room temperature for 1 hr or at 16 °C overnight.

2.5.7 Selection of adaptor orientation

Ligation reaction resulted in fragments with either two identical (A-A and B-B) or two distinct (A-B and B-A) adapter orientations. However, only distinct adapter orientation fragments can be sequenced in Illumina's platform. We enriched desired ligation products by utilizing the biotin (adaptor A) – streptavidin chemistry. Streptavidin coated magnetic beads (Dynabeads MyOne Streptavidin C1, Invitrogen) were used to pull down A-A, A-B and B-A ligation products using manufacturer's protocol. The supernatant, containing B–B, was discarded. Later, 0.2 N NaOH was added to the beads. Incubation for 10 mins at room temperature denatured two strands of the ligation product. Only A'–B strand appeared in the supernatant while both strands of the A–A remained associated with the beads. The supernatant with distinct orientation was extracted and column purified using MinElute PCR Cleanup Kit (Qiagen). The pH of the supernatant was adjusted to allow maximal binding to the column. The single stranded DNA was eluted in 36 µl of EB.

2.5.8 Final PCR and size selection

The single stranded DNA obtained from previous step was amplified using adaptor specific primers. Following primers were ordered in 100 uM concentration:

a) Final_FP:

5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGAATA -3'

b) Final_RP:

5'-

CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATA-3'

A 50 µl PCR reaction was set up with 18 µl of single stranded template, 5 µl of primers (2 µM stock), 4% DMSO, 5 µl Pfu Turbo reaction buffer (10X), 0.2 mM dNTP (10 mM stock), 2.5 units of Pfu Turbo Polymerase. The amplification consisted of 14 cycles of denaturation (95°C – 30s), annealing (62°C – 30s) and extension (72°C – 40 s). The amplified product was run in 2% agarose gel at 80 – 100 volts for 1 hr. Using 50 bp ladder (NEB) a band corresponding to size range of 150 – 500 bp was cut out. The DNA was retrieved from the gel using MinElute Gel Extraction Kit (Qiagen) with 15 µl of elution.

2.5.9 Quantification of the sequencing library

Quantitative real time PCR was used to determine the concentration of the sequencing libraries prepared by our protocol. The standard curve for various dilutions of phiX control library was generated using the adapter specific primers recommended by Illumina. We later used the standard curve to determine the molarity of our sequencing libraries. The concentration of sequencing library loaded into the flowcell was calibrated by the sequencing facility. We typically obtained good cluster density with 5 pM of library concentration on HiSeq v3 kit.

2.5.10 Oligonucleotides

All of our heptamer primers were flanked by universal adapter sequence (5’-CCGAATA-heptamer-3’) and synthesized by Valuegene Inc. These primers were desalted and suspended in RNase/DNase free water to 100 μM concentration. Later, the primers were pooled together into three different tubes as described in **Supplementary Table S2.4** at equi-molar concentration to prepare a stock solution containing 2 μM of each heptamer primer.

2.5.11 Mouse embryonic stem cell culture and differentiation

Mouse R1 or T-GFP embryonic stem cells were cultured on mouse embryonic fibroblast (MEF) on gelatin-coated dishes in high glucose DMEM (Hyclone, Logan, UT) supplemented with 10% fetal calf serum (FCS) (Hyclone, Logan, UT), 0.1 mM b-mercaptoethanol (GIBCO), 1% non-essential amino acids (GIBCO), 2 mM L-glutamine (Sigma, St. Louis, MO), sodium pyruvate (Sigma), antibiotics (Sigma), and 1,000 U/ml of LIF (Sigma) and passaged with 0.25% Trypsin (GIBCO).

For embryoid body (EB) differentiation, MEF were stripped from the cultures by 15 minutes incubations on gelatin-coated dishes. mESCs were collected and washed in PBS to remove traces of serum. mESCs were differentiated in serum free media containing N2 and B27 supplements as described elsewhere[35, 36]. mESCs were aggregated at 50,000 cells/ml in non-coated polystyrene plates. After 2 days, EBs were dissociated by trypsin treatment and re-aggregated in fresh media in presence of different growth factors and small molecules. Activin A and BMP4 were obtained from R&D

Systems while SB-431542 was obtained from Sigma. IWR-1 was synthesized in house as described previously[54]. EBs were harvested at day 4 for RNA extraction and processing.

2.5.12 Library Generation for mRNA dilution series using DP-seq

Serial dilutions (10 ng, 1 ng, 100 pg, 50 pg, 10 pg, and 1 pg) were prepared for the mRNA derived from Activin A (3 ng/mL) sample. First strand cDNA synthesis was performed for all mRNA dilutions in duplicates to get the technical replicates. Later, the purified cDNA prepared from each dilution, was split into three tubes to perform amplification using our heptamer primers. The numbers of PCR cycles were increased for lower dilutions to get appropriate amounts of DNA for the library construction. The numbers of PCR cycles used are as follows:

10 ng and 1 ng – 13 cycles

100 pg and 50 pg – 16 cycles

10 pg – 19 cycles

1 pg – 23 cycles

The amplicon libraries thus constructed, were phosphorylated at the 5' end as mentioned above. Later, the libraries were ligated with Illumina's Y-adaptors and amplified using adaptor specific primers consisting of a different Illumina's Truseq barcode sequence for each library. The amplified libraries were run through the 2% agarose gel and size selected (150 – 500 bp) for sequencing. Similar methodology was used for the generation of sequencing libraries with ribosomal inhibition primers.

2.5.13 Library Generation using Std. RNA-seq protocol

Std. RNA-seq[10] libraries were constructed from about 10 ng of mRNA derived from Activin A (3 ng/mL) and Activin A (100 ng/mL) samples using Illumina's TruSeq RNA Sample Prep Kit v2.

2.5.14 Library Generation using Smart-seq

Smart-Seq cDNA generation and amplification was performed on 50 picograms of mRNA derived from Activin A (3 ng/mL and 100 ng/mL) treated samples using SMARTer Ultra Low RNA Kit for Illumina sequencing (Clontech). We performed 13 cycles of amplification to achieve about 1-10 ng of the amplified cDNA libraries. These libraries were later sheared using Covaris system to obtain 200-500 bp fragments. Later, standard Illumina library preparation protocol was followed to prepare the sequencing libraries using Illumina Paired-End DNA Sample Prep kit.

2.5.15 Reverse Transcription and Quantitative RT-PCR (qPCR)

Total RNA was extracted from cells using Trizol (Invitrogen) according to the manufacturer's instructions. About 1 µg of total RNA was treated for DNA removal and converted into first strand cDNA using Quantitect Reverse Transcription kit (Qiagen). SYBR Green qPCR was run on a LightCycler 480 (Roche) using the LightCycler 480 SYBR Green Master Kit (Roche). All primers were designed with a T_m of 60°C. Data was analyzed using the $\Delta\Delta C_t$ method, using β-actin as normalization control, which was determined as a valid reference in mouse ESC differentiation. The primer sequences are listed in **Supplementary Table S2.6**.

2.5.16 Flow Cytometry

Day 4 embryoid bodies from T-GFP mESC were dissociated with trypsin to single cell suspensions and analyzed on a FACSCanto (BD Biosciences). Prior to analysis, cells were stained with propidium iodide to label dead cells. Data analysis was performed using FlowJo (Treestar Inc.) where measured events were gated for PI negative populations (exclusion of dead cells) and forward/side scatter (exclusion of debris and aggregates) before generating dot plots.

2.5.17 Mapping reads

Our libraries were sequenced on Illumina's GIIx Analyzer and HISEQ2000 platforms. We performed single end 36 sequencing cycles on version 5.0 of flowcell (FC-104-5001 | TruSeq SBS Kit v5 – GA (36-cycle)). The raw reads were truncated as 32-mer with the first and last 2 base pairs of the reads removed. The 32-mer reads were aligned to the RefSeq mRNA database allowing up to 2 mismatches using our in-house software which uses suffix array implementation. Reads that did not align to the mouse RefSeq mRNA database were later aligned to mouse genome using Bowtie[67].

Libraries constructed from serial dilutions of mRNA were sequenced in Illumina's HiSeq2000 systems (TruSeq SR Cluster Kit v3-cBot-HS and TruSeq SBS Kit v3-HS). The libraries were sequenced as 100 bp single-end reads. The first 14 sequences came from our heptamer primers including the universal tail sequence (5'-CCGAATA-3') as such the first 14 bps were truncated and next 32 bp sequence was aligned to the mouse transcriptome allowing ≤ 2 mismatches.

2.5.18 Mapping of Smart-seq reads

The number of reads obtained from Smart-seq was double the number of reads for DP-seq. Previous studies[68, 69] have demonstrated that the transcriptome coverage and the technical noise in expression measurements vary with the sequencing depth and global normalization of the reads across different samples is heavily affected by few highly expressing transcripts. In order to perform an objective comparison between Smart-seq and DP-seq, we downsized the Smart-seq libraries by generating multiple random sets, consisting of a similar number of reads obtained from DP-seq. The reads in these datasets were mapped to the mouse transcriptome allowing ≤ 2 mismatches. The analysis of these random sets showed similar transcriptome coverage and technical noise. In this study, we present the comparison of DP-seq with one of the random sets generated from the Smart-seq library.

2.5.19 Differential gene expression analysis

We employed a local pooled variance test similar to LPE[70] to identify differentially regulated transcripts. For each transcript, unique reads coming from predicted and non-predicted primer-binding sites were combined in all samples. Prior to identifying the differentially expressed transcripts, the fold changes between control and treatment conditions (including technical replicates) were lowess normalized to eliminate average read dependent variations in the fold changes. The noise in the technical replicates reflected variability arising out of sample preparation and the sequencing platform. As such we used the expression measurements obtained from the technical replicates to determine the baseline (null) distribution where no differential expression of the transcripts was expected. The null distribution was determined by plotting M and A quantities for technical replicates, which are defined as:

$M_{i,j}=\mathrm{Log}_2(\frac{Reads,i}{Reads,j})$

$A_{i,j}=0.5\times \mathrm{Log2}(\mathrm{Reads},_{i}\ \times \mathrm{Reads},_{j})$

where ‘i’ and ‘j’ represents any two samples. M corresponds to log ratio in unique reads for a transcript between samples ‘i’ and ‘j’ while A corresponds to average reads for the transcript in the two samples.

To quantify the technical noise, we pooled the expression of ~200 transcripts in the null distribution with similar reads and estimated the standard deviation in their fold change. We assumed that all transcripts with similar expression measurements possess similar noise. Also, the distribution of the fold changes was assumed to follow a Gaussian distribution. Next, a threshold for differentially expressed transcripts was determined as 1.96 times the standard deviation, corresponding to a less than 5% chance of the transcript being called differentially expressed by random chance. The experimental MA plot, which was defined as treatment/control, was overlaid on the technical replicate MA plot and any transcript representing a fold change above/below the threshold was designated as differentially expressed. Higher thresholds (blue curve) were used for the low expressing transcripts as demonstrated in **Supplementary Fig. S2.12.**

2.5.20 Identification of Activin A/TGFβ target genes

Putative Activin A/TGFβ target genes were determined as genes exhibiting opposite mode of regulation in AA3 and SB samples as compared to serum free media control. The target genes were further classified into three categories of expression as shown in **Figure 6**. A p-value cutoff of 0.05 was used to determine differentially expressed transcripts.

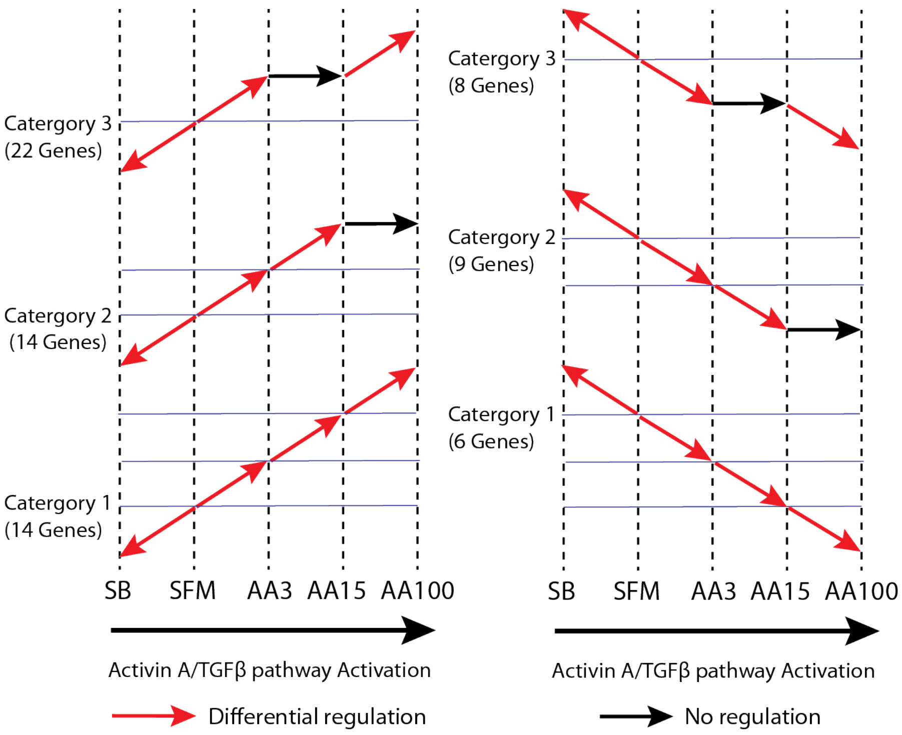


Figure 2.6: Identification of TGFβ target genes.

2.6 Supplementary Figures

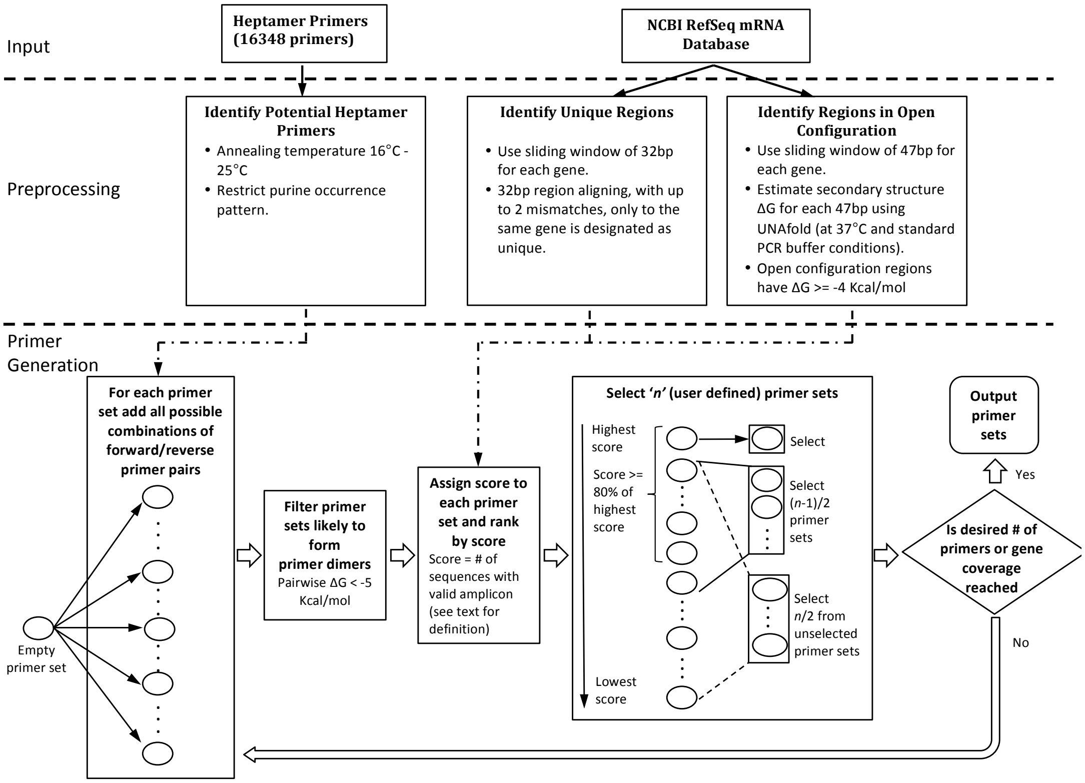

Input
Heptamer Primers (16348 primers)
NCBI RefSeq mRNA Database
Preprocessing
Identify Potential Heptamer Primers
• Annealing temperature 16°C - 25°C
• Restrict purine occurrence pattern.
Identify Unique Regions
• Use sliding window of 32bp for each gene.
• 32bp region aligning, with up to 2 mismatches, only to the same gene is designated as unique.
Identify Regions in Open Configuration
• Use sliding window of 47bp for each gene.
• Estimate secondary structure ΔG for each 47bp using UNAfold (at 37°C and standard PCR buffer conditions).
• Open configuration regions have ΔG >= -4 Kcal/mol
Primer Generation
For each primer set add all possible combinations of forward/reverse primer pairs
Empty primer set
Filter primer sets likely to form primer dimers
Pairwise ΔG < -5 Kcal/mol
Assign score to each primer set and rank by score
Score = # of sequences with valid amplicon (see text for definition)
Select 'n' (user defined) primer sets
Highest score
Score >= 80% of highest score
Lowest score
Select
Select (n-1)/2 primer sets
Select n/2 from unselected primer sets
Is desired # of primers or gene coverage reached
Yes
No
Output primer sets

Figure S2.1: Flowchart of heptamer primer generation using an iterative randomized algorithm.

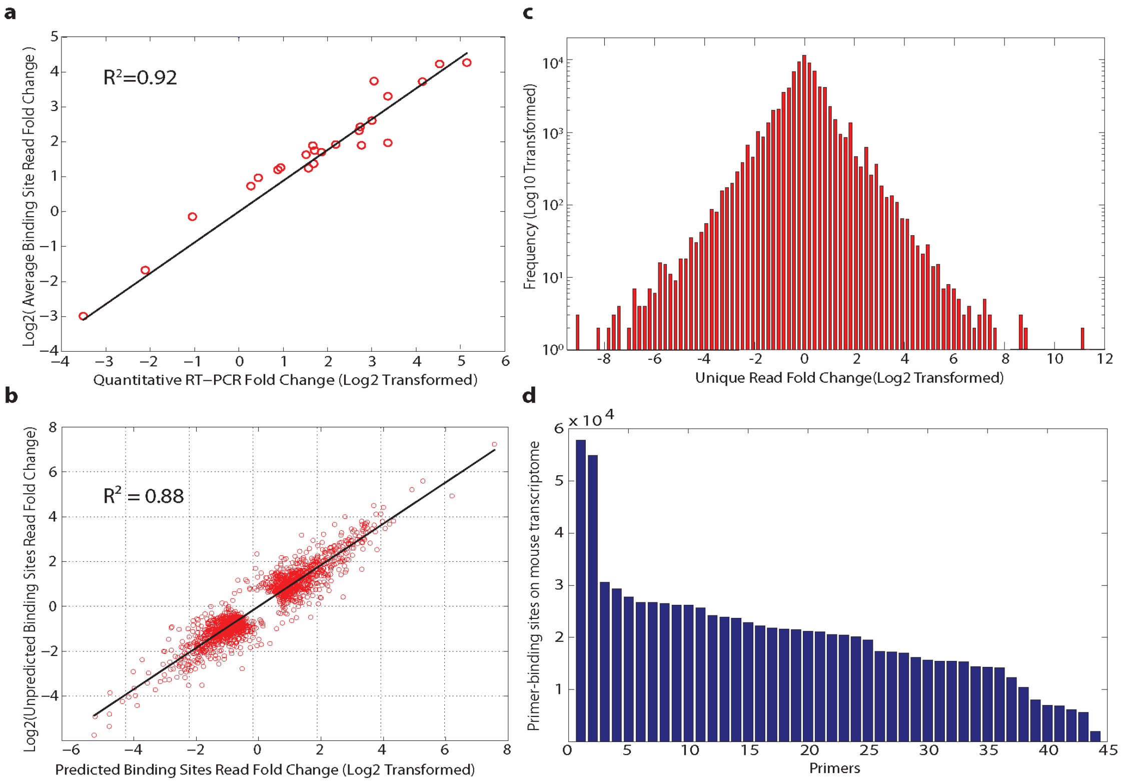


Figure S2.2: Performance of heptamer primers based amplification. (a) Multiple heptamer primer-binding sites on a transcript provided independent measurements of relative abundance of the transcript. The average fold change obtained from multiple primer-binding sites for a transcript was in concordance with quantitative RT-PCR (n=24). (b) Mis-primed PCR products maintained relative abundance of gene expression. Fold changes observed in predicted vs. mis-primed binding sites for differentially expressed transcripts (in SB-431542 vs. AA100) showed strong correlation. (c) Distribution of fold changes observed in unique reads of the transcripts across all of the samples. The majority of the trancripts were not differentially regulated. Our methodology captured fold changes in range of 2-8 – 210 demonstrating broad dynamic range. (d) Distribution of heptamer primer-binding sites on the mouse transcriptome.

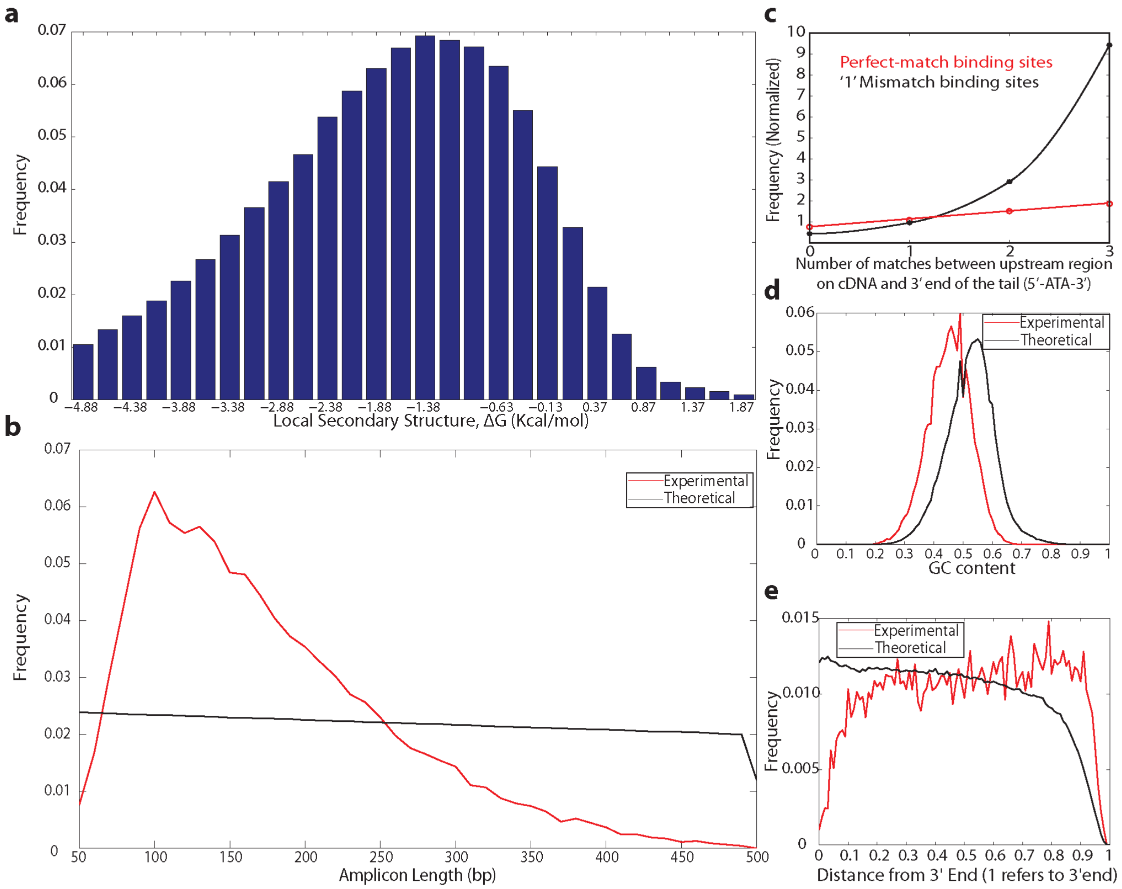


Figure S2.3: PCR biases observed in our methodology. (a) PCR bias caused by the secondary structure of the cDNA. The distribution is shifted towards high Gibbs free energy (ΔG) implying that the primer-binding sites forming stable secondary structure shielded heptamer primers from annealing to their target sequences. (b) Bias towards shorter PCR fragments. The black curve represents the distribution estimated for all theoretically possible amplicons from the 44 heptamer primers in the mouse transcriptome. The experimental curve dropped sharply around 100bp because of the size selection step performed at the last stage of the sequencing library generation. (c) Tail Interaction. Heptamer primer binding sites with '1' mismatch had significantly higher tail interaction as compared to perfectly matched primer-binding sites. (d) GC bias. The amplicons with lower GC content are preferentially amplified. (e) PCR bias caused by reverse transcriptase. Majority of the primer-binding sites came from 3' end of the genes mainly because of the inability of the reverse transcriptase to produce full-length first strand cDNA.

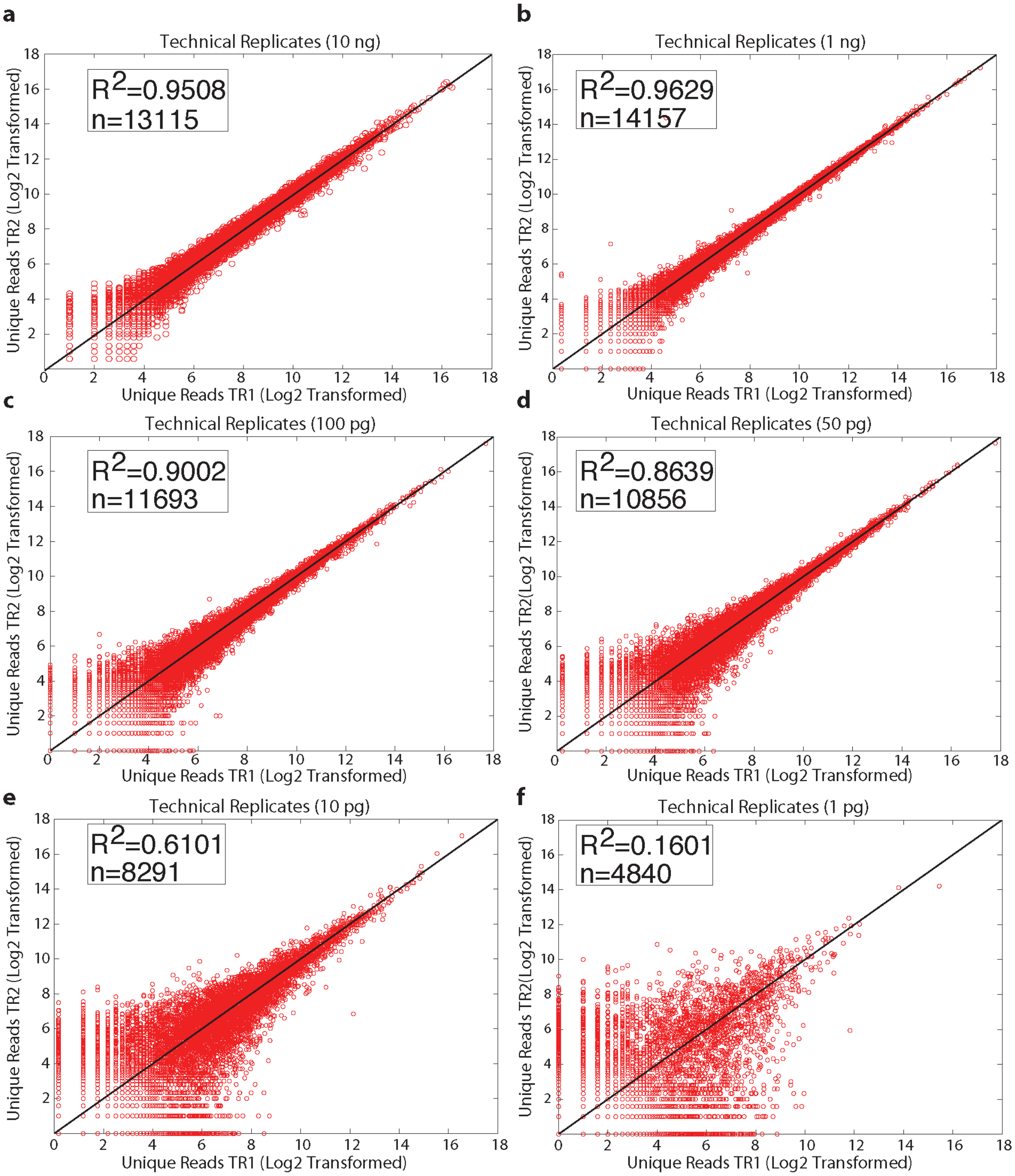


Figure S2.4: Techincal Replicates for sequencing libraries prepared from various amounts of starting material (mRNA). The transcriptome coverage dropped with lower amounts of mRNA. Significant technical noise was observed for the sequencing libraries prepared from 1 pg of mRNA.

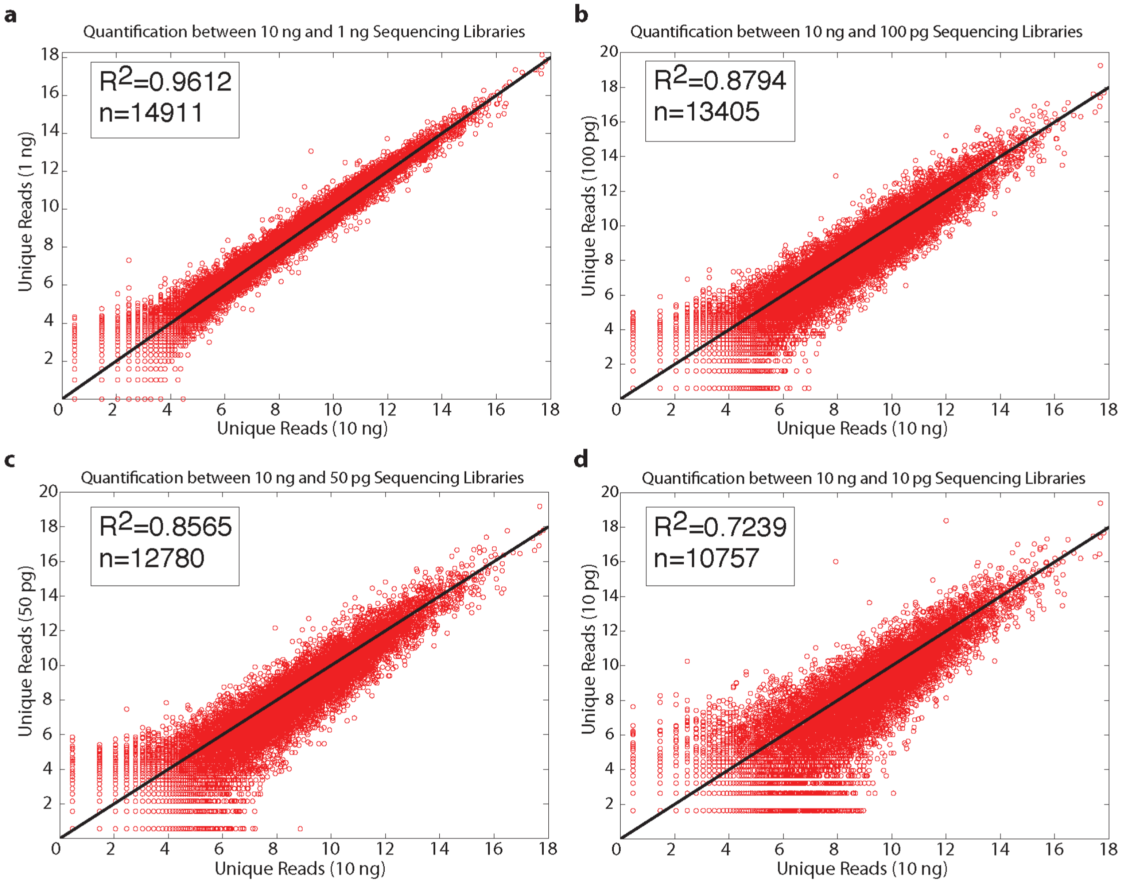


Figure S2.5: Transcript representation is conserved with serial dilutions of the starting material (mRNA). Transcripts abundance obtained from dilutions (1 ng, 100 pg, 50 pg, 10 pg) were compared with respect to highest concentration of 10 ng.

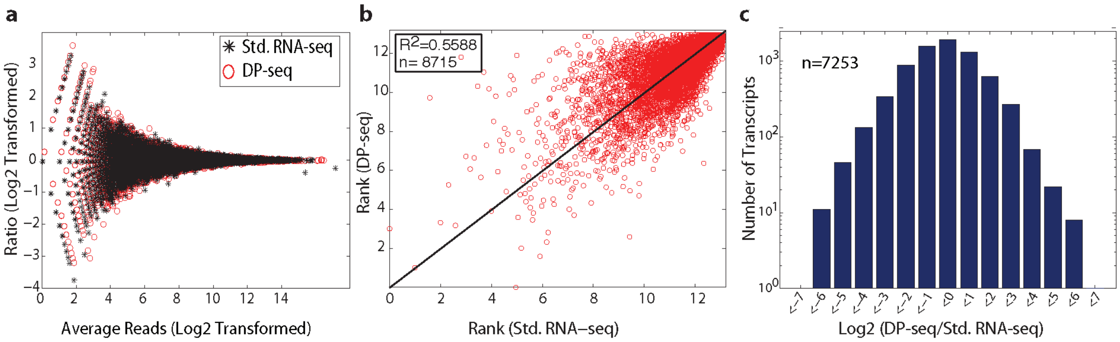


Figure S2.6: DP-seq vs. Std. RNA-seq. (a) Std. RNA-seq exhibited similar technical noise in the technical replicates as DP-seq. (b) PCR biases observed in our protocol distorted the order of transcript expression resulting in poor Rank Correlation with respect to the Std. RNA-seq. (c) Distribution of the ratio of unique reads obtained for the low expressed transcripts (RPKM<=10) in DP-seq and Std. RNA-seq.

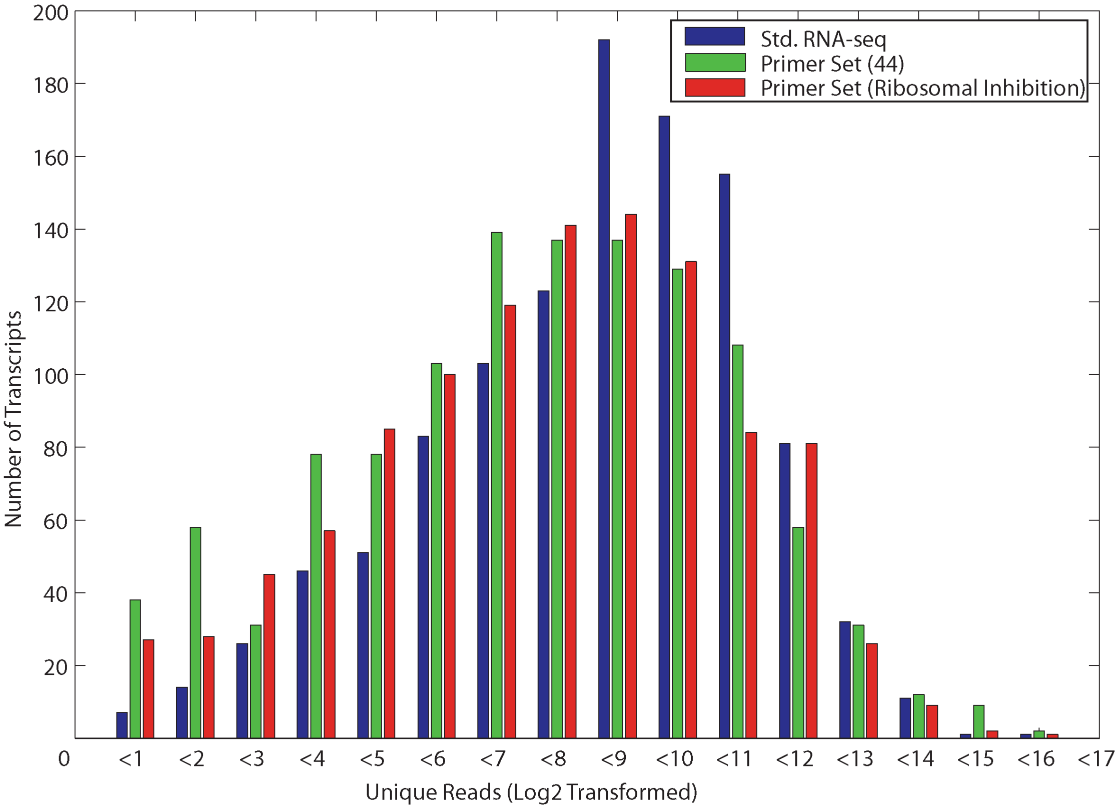


Figure S2.7: Distribution of reads. Sequencing libraries prepared from Std. RNA-seq and DP-seq (44 primer set and a primer set used for suppression of the ribosomal transcripts) displayed overlapping distributions of reads mapping to the mouse transcription factors (n=1148; AA3 sample).

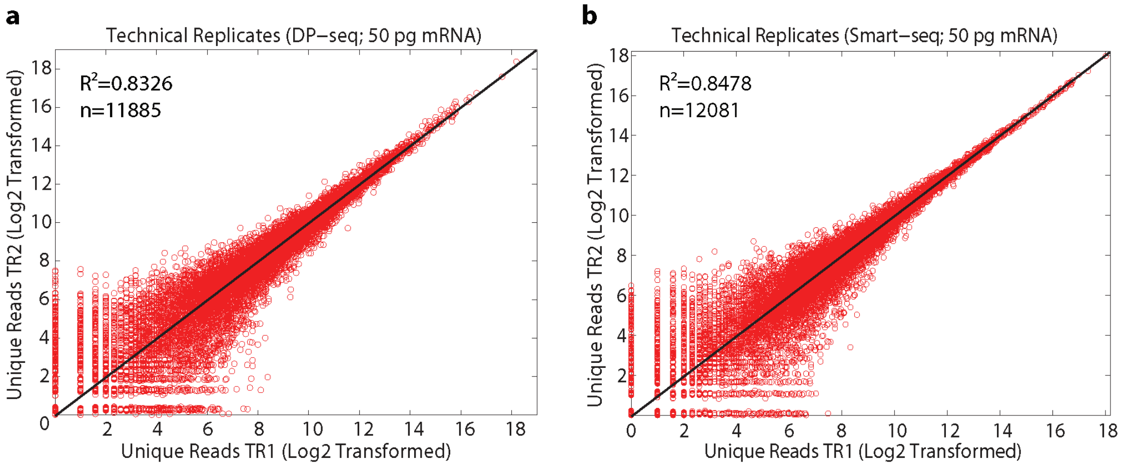


Figure S2.8: Technical Replicates for DP-seq and Smart-seq. Technical replicates prepared from 50 picograms of mRNA derived from Activin A 100ng/mL dosage exhibited high correlation in expression measurements for DP-seq and Smart-seq.

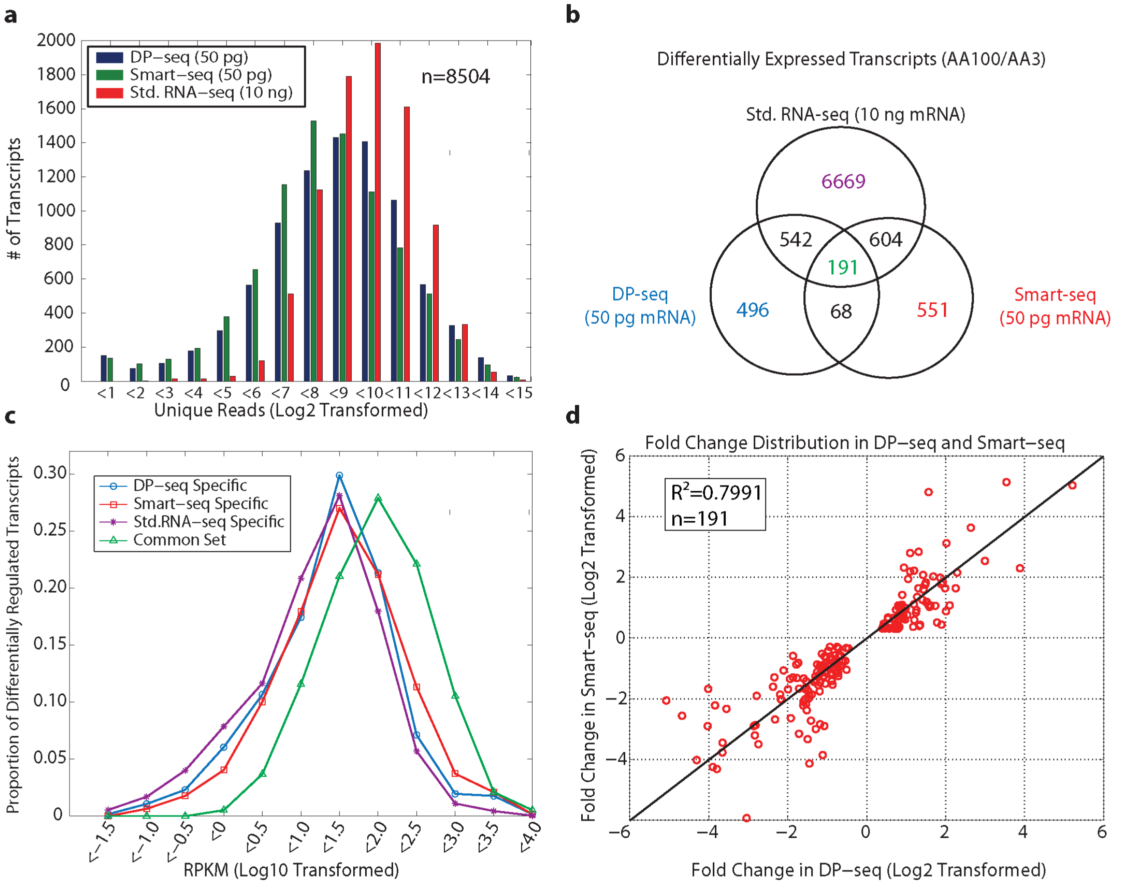


Figure S2.9: Comparison of the sequencing libraries prepared from DP-seq, Smart-seq and Std. RNA-seq methods. (a) Histogram of unique reads obtained for the moderately expressed transcripts (3<RPKM<300) in the three methods. The amounts of mRNA used for the sequencing library generation are mentioned in the parentheses. (b) Venn diagram depicting the overlap of the differentially expressed transcripts between Activin A 100ng/mL and 3ng/mL dosages identified in the three methods. (c) The expression profile of the common set (green) is shifted towards higher RPKM as compared to the method specific differentially expressed transcripts. (d) Correlation in fold changes for the common set between DP-seq and Smart-seq. The RPKM measurements were made from Std. RNA-seq experiment performed on AA100 sample.

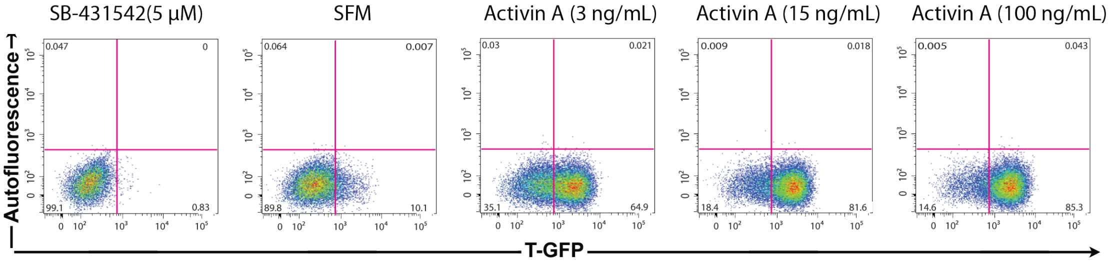


Figure S2.10: Flow Cytometry. Flow cytometry on T-GFP mESCs at day 4 of differentiation upon treatment with SB and Activin A. Graded activation of Activin A/TGFβ signaling pathway led to increased expression of mesoderm marker, T.

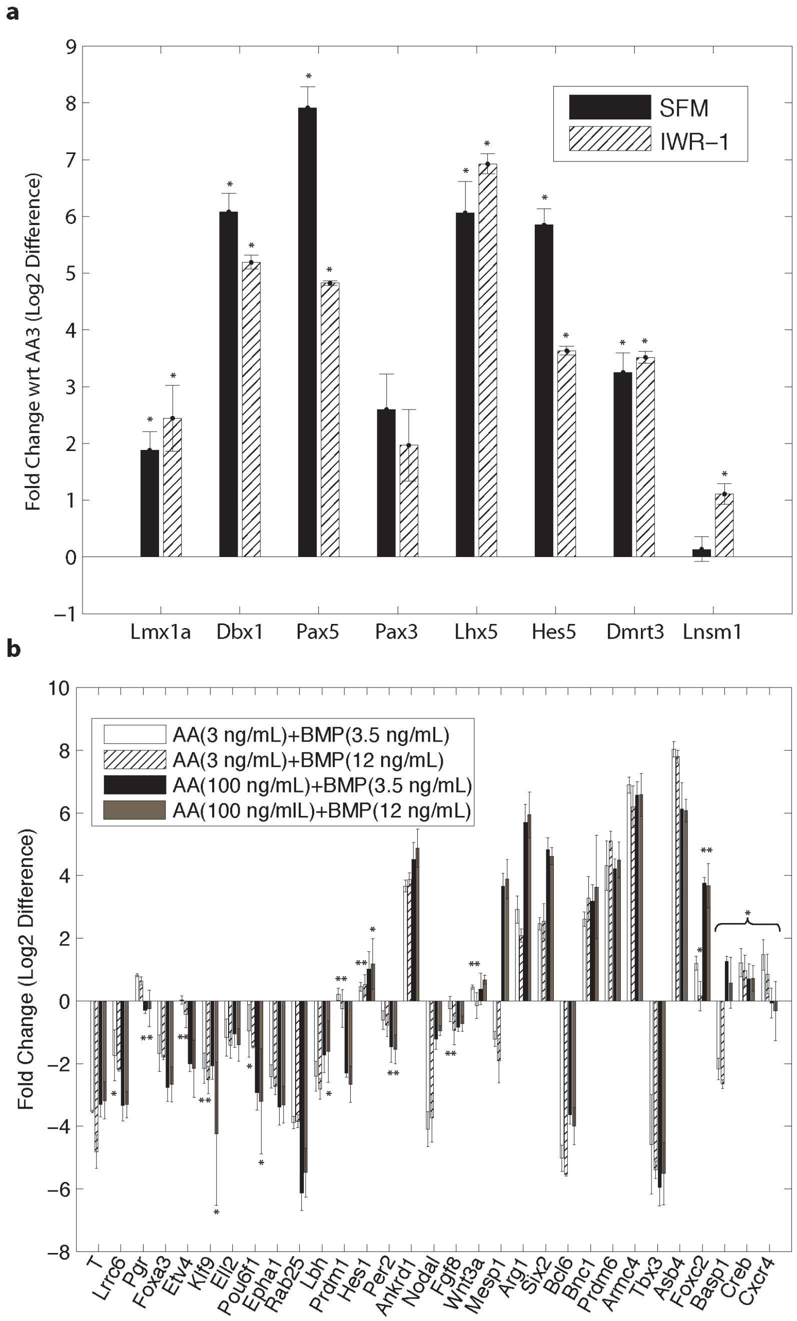


Figure S2.11: qPCR Validation. (a) Validation of neuro-ectoderm specific genes by using small molecule inhibitor of Wnt Signaling pathway, IWR-1 to efficiently induce neuro-ectoderm in an in-vitro differentiation model. The quantitative RT-PCR fold changes were normalized to Activin A (3 ng/mL) dosage. Error bars represent standard deviation in biological replicates (n=3). Asterisks indicates p>0.05 (Student's t-test) compared with controls. (b) Expression profiles of Primitive Streak markers in response to BMP4 signaling. Quantitative RT-PCR fold changes for two BMP4 dosages (3.5 and 12 ng/mL) were normalized with respect to Activin A alone induction. Error bars represent standard deviation in biological replicates (n=3). Asterisks indicate p>0.05 (Student's t-test) compared with controls.

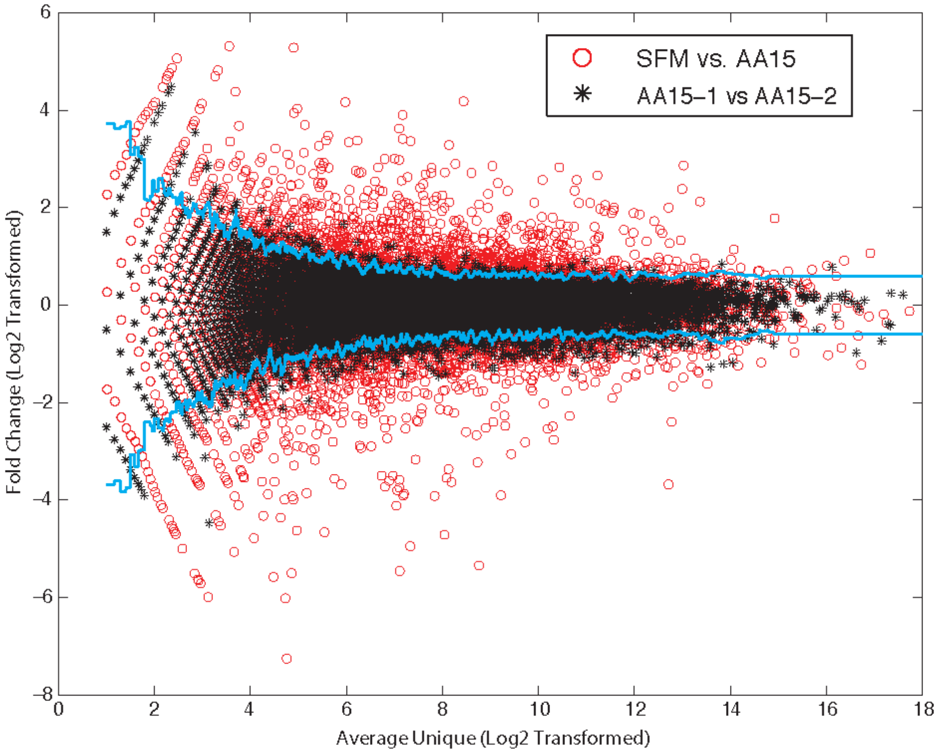


Figure S2.12: Identification of the differentially expressed transcripts. Baseline distribution was determined from MA plot of the technical replicates. Experimental MA plot of untreated control vs. Activin A (15 ng/mL) was overlaid on top of the baseline distribution. The blue curve represents p-value threshold of 0.05 and experimental ratios above/below the curve were designated as differentially regulated.

2.7 Supplementary Tables

Table S2.1A: Comparison of sequencing libraries made from various dilutions of mRNA derived from Activin A (3ng/mL; AA3) sample using DP-seq.

Amount of mRNA	Total Reads	% Of reads aligned to Refseq Transcripts	% Of unmapped reads aligned to genomic locations	Number of Transcripts >=1 unique reads	R^2 for Technical Replicates	R^2 with 10ng Library
10 ng, TR1	6251585	67.79	18.46	13547	0.9508	
10 ng, TR2	20404270	68.70	18.01	15236		
1 ng, TR1	19807355	55.78	18.43	15151	0.9643	0.9615
1 ng,TR2	25119387	55.75	18.20	15306		
100 pg,TR1	13913778	59.35	17.92	12955	0.9016	0.8794
100 pg,TR2	13522446	61.24	17.99	12648		
50 pg, TR1	13378971	59.03	18.22	11986	0.8640	0.8565
50 pg, TR2	15297046	60.06	17.87	12002		
10 pg, TR1	14189544	27.46	16.88	9603	0.6102	0.7239
10 pg, TR2	13971891	31.03	12.62	9589		
1 pg, TR1	16038243	4.45	11.51	6531	0.1901	0.1002
1 pg, TR2	14281289	5.22	9.81	6465		

Table S2.1B: Mapping Summary. Mapping summary of sequencing libraries made from different protocols using two different dosages of Activin A 3ng/mL (AA3) and 100ng/mL (AA100). Smart-seq mapping summary is given for one of the random sets obtained from all of the reads. Ribosome inhibition libraries were made from Activin A (3ng/mL; AA3) sample using DP-seq.

Amount of mRNA	Total Reads	% Of reads aligned to Refseq Transcripts	% Of unmapped reads aligned to genomic locations	Number of Transcripts >=1 unique reads	R^2 for Technical Replicates
Std. RNA-seq					
AA3; 10ng, TR1	18196250	81.21	15.82	17455	0.9755
AA3; 10ng, TR2	17638530	81.18	15.74	17380	
AA100; 10ng	17905346	79.50	16.77	17026	
DP-seq					
AA3; 50pg	24633672	58.59	17.65	13138	
AA100; 50pg, TR1	26108501	58.56	13.10	12910	0.8326
AA100; 50pg, TR2	27486701	65.27	14.53	13271	
Smart-seq					
AA3; 50pg, TR1	24272863	87.24	7.53	13798	0.8640
AA3; 50pg, TR2	26014738	86.89	7.30	13715	
AA100; 50pg, TR1	22298719	86.35	7.71	13400	0.8478
AA100; 50pg, TR2	24284435	87.25	7.43	13568	
Ribosome Inhibition (DP-seq)					
Primer Set 1; 500pg	21816975	67.93	21.90	14616	
Primer Set 2; 500pg	19668914	71.03	25.61	13246	
Primer Set 3; 500pg	10267103	68.49	27.05	11654	

Table S2.2: GO (Biological Process) Enrichment for genes differentially regulated in SB/AA15. Genes up-regulated in SB are enriched for ectoderm related terms while genes up-regulated in AA15 are enriched for mesoderm and endoderm related terms. P-values were determined from background set of genes that showed expression in SB/AA15 samples.

Up-regulated in SB in comparison to AA15

Term	PValue	Bonferroni	Benjamini
Neuron differentiation	1.79E-23	4.76E-20	4.76E-20
Neuron development	2.48E-17	6.59E-14	3.30E-14
Neuron projection development	5.72E-17	2.95E-13	9.81E-14
Forebrain development	4.04E-16	1.18E-12	2.95E-13
Axonogenesis	1.05E-13	2.79E-10	5.58E-11
Cell projection organization	4.54E-13	1.20E-09	2.01E-10
Neuron projection morphogenesis	1.53E-12	4.06E-09	5.80E-10
Axon guidance	1.71E-12	4.55E-09	5.68E-10
Cell motion	2.02E-12	5.35E-09	5.94E-10
Cell projection morphogenesis	2.36E-12	6.25E-09	6.25E-10
Neuron migration	4.17E-12	1.11E-08	1.01E-09
Cell morphogenesis involved in neuron differentiation	4.78E-12	1.27E-08	1.06E-09
Cell morphogenesis involved in differentiation	1.29E-11	3.42E-08	2.63E-09
Cell part morphogenesis	1.29E-11	3.42E-08	2.63E-09
Sensory organ development	6.41E-11	1.70E-07	1.21E-08
Cell morphogenesis	1.49E-10	3.96E-07	2.64E-08
Embryonic morphogenesis	5.74E-10	1.52E-06	9.52E-08
Pattern specification process	6.64E-10	1.76E-06	1.04E-07
Cell migration	2.70E-09	7.17E-06	3.98E-07
Up-regulated in AA15 in comparison to SB			
Tissue morphogenesis	5.66E-10	1.86E-06	1.86E-06
Tube morphogenesis	1.43E-08	4.68E-05	2.34E-05
Tube development	1.75E-08	5.74E-05	1.91E-05
Regulation of cell proliferation	4.47E-08	1.47E-04	3.67E-05
Muscle organ development	1.02E-07	3.34E-04	6.68E-05
Epithelium development	1.09E-07	3.59E-04	5.99E-05
Morphogenesis of a branching structure	7.34E-07	0.002407	3.44E-04
Embryonic development in birth or egg hatching	7.95E-07	0.002606	3.26E-04
Gastrulation	8.05E-07	0.002639	2.94E-04
Chordate embryonic development	1.27E-06	0.004150	4.16E-04
Muscle tissue morphogenesis	1.37E-06	0.004472	4.07E-04
Cardiac muscle tissue morphogenesis	1.37E-06	0.004472	4.07E-04
Cardiac muscle tissue development	1.64E-06	0.005377	4.49E-04
Blood vessel morphogenesis	1.79E-06	0.005852	4.51E-04
Epithelial cell differentiation	1.87E-06	0.006115	4.38E-04
Embryonic morphogenesis	2.26E-06	0.007406	4.95E-04
Formation of primary germ layer	2.54E-06	0.008290	5.20E-04
Endoderm development	2.68E-06	0.008762	5.18E-04
Striated muscle tissue development	3.44E-06	0.011238	6.28E-04
Heart morphogenesis	3.63E-06	0.011851	6.27E-04

Table S2.3: Kegg Pathways enriched in SB/AA15 samples. P-values were determined from background set of genes that showed expression in SB/AA15 samples.

Up-regulated in SB in comparison to AA15		
Term	PValue	Fold Enrichment
Axon guidance	1.57E-08	3.70
Pathways in cancer	1.51E-05	2.14
Focal adhesion	1.18E-04	2.35
Wnt signaling pathway	3.19E-04	2.50
Basal cell carcinoma	5.50E-04	3.73
Colorectal cancer	5.60E-04	3.04
Pancreatic cancer	0.001349	3.11
Notch signaling pathway	0.004505	3.36
TGF-beta signaling pathway	0.006142	2.57
ErbB signaling pathway	0.006142	2.57
Melanogenesis	0.006550	2.42
Adherens junction	0.006669	2.70
Chronic myeloid leukemia	0.006669	2.70
Hedgehog signaling pathway	0.007261	3.11
Non-small cell lung cancer	0.007261	3.11
Biosynthesis of unsaturated fatty acids	0.012782	4.15
Small cell lung cancer	0.014350	2.41
Endometrial cancer	0.019416	2.87
Prostate cancer	0.020781	2.28
Regulation of actin cytoskeleton	0.021865	1.72
Chondroitin sulfate biosynthesis	0.027081	4.24
ABC transporters	0.030854	2.90
Renal cell carcinoma	0.031684	2.40
MAPK signaling pathway	0.043071	1.55
VEGF signaling pathway	0.048220	2.21
Up-regulated in AA15 in comparison to SB		
Glioma	0.001299	2.74
Pathways in cancer	0.002808	1.59
Melanoma	0.003446	2.47
Alanine, aspartate and glutamate metabolism	0.007788	3.34
Arginine and proline metabolism	0.007922	2.60
Cysteine and methionine metabolism	0.013292	3.04
p53 signaling pathway	0.019104	2.18
Amino sugar and nucleotide sugar metabolism	0.020808	2.56
ABC transporters	0.023613	2.51
Fatty acid metabolism	0.023613	2.51
Non-small cell lung cancer	0.024969	2.32
MAPK signaling pathway	0.029124	1.46
Endocytosis	0.029288	1.55
Nitrogen metabolism	0.031517	3.27
Tight junction	0.037866	1.67
Focal adhesion	0.040549	1.52
Glycolysis / Gluconeogenesis	0.040867	2.03
Bladder cancer	0.045432	2.39

Table S2.4: List of heptamer primers used for our sequencing-library generation. 44 unique primers were split into three tubes with some primers repeated in different tubes to get coverage ≥80% mouse transcriptome.

1. cccagtg	1. caaagcc	1. cacacac
2. ccccaga	2. caacccc	2. cagcagc
3. cccccaa	3. cccagca	3. ccaccag
4. ctcccca	4. cccccaa	4. cccagca
5. cttcacg	5. ctcgtcc	5. cccccaa
6. gcaacag	6. cttcccc	6. ccttccc
7. tgacagc	7. gcctctc	7. cttcccc
8. tggctct	8. gcctctg	8. gcaacag
9. tggcttc	9. gcgaact	9. gcctcag
10. tccctcc	10. tcagccc	10. tccctcc
11. ccttccc	11. tctccga	11. tgaccca
12. cagaccc	12. tgccatc	12. tgagcct
13. gcaaacc	13. tgccttg	13. cagcact
14. ccaggac	14. tgagcct	14. gcgaact
15. cacacac	15. tcctcgt	15. ctcccag
16. tctccga	16. tctgcct	16. gccaaag
17. cctccca	17. ctgccct	17. ccccaga
18. tgaccca	18. tgccact	18. tcagcca
	19. cttcacg	19. gaagcca
	20. gcaacag	20. tgacagc
	21. cctctgc	
	22. gcaaacc	
	23. ccccaga	
	24. ctcagca	
	25. tgacagc	

Table S2.5: Primer Sets for Ribosomal Inhibition.

Primer Set 1	Primer Set 2	Primer Set 3
CCTCCTG	GGACAGC	GAAAGCC
GCAGCCT	CACACAC	CACACAC
TCCCACA	GCAACAA	CCACACA
CACACTG	GCATGTG	TGCTGTG
CTTCCCC	GTGACCT	GACAACC
CCACCAC	CATCAGC	GTCACAC
CCTCCCC	CTTGAGC	GACACAC
CTTGCAG	TACAGCC	GCGTTTT
CCCACAC	GTTCTCG	GAGCCTC
CCTTCCC	CAAGCAC	GTGATGC
CACCCCA	TCAGCAC	CCGTCTT
CTCTCCC	TCGTTCC	TCCCTCA
CAGAGCC	GCGTCTG	GTTTCCG
CCCCAAA	CAAACCG	TCCAACC
CTCCCCA	CCGTGAC	CGAATGG
CAAGAGC	TGTCTCG	CCGTGTA
CCCTGGA	GCGTCAG	CAAACCG
CCCCCTC	CCCCTAC	CGAGTGT
CCCCTCA	CCGTGTA	GACTCCG
CTGAGCT	CCGTTGA	GCGAATT
CCCCCAG	GATCCCG	GGTGCCC
	CCGACTT	CGAGAGC
	GCGACAC	GATGCGT
	CTGAGCG	CGACTCA
		GCGTTAG
		CAGTACG
		GAATGCG
		CCGTGCT
		CAACCGA
		TGCTACG
		GTAACCG
		TGCCGAT
		CCCGTTA
		TAGAGCG
		CAAGCGT
		CGATAGC
		GACCGAC
		CGATCCC
		CGAGTGC
		CGATTGC

Table S2.6: List of quantitative RT-PCR primers used in the study.

Gene	Forward Primer	Reverse Primer
Lefty1	CGCTGAATCTGGGCTGAGTCCC	GCCTAGGTTGGACATGTTTGCCCA
Lefty2	TGCAAGTAGCCGACTTCGGAGC	CCTATTCCCAGGCCTCTGGCCA
Gsc	GGGGGTCGAGAAAGCAACGAGG	ACGAGGCTCACGCAGGCAGC
Flk-1	AGAGGAAGTGTGCGACCCCAA	CACTGGCCGGCTCTTTCGCTT
Oct4	TGAAGTGCCCGAAGCCCTCCCTA	GCCCTTCTGGCGCCGGTTACA
Mesp1	TCTAGAAACCTGGACGCCGCC	TCCGTTGCATTGTCCCCTCCAC
T	CTCCGATGTATGAAGGGGCTGCT	GCTATGAGGAGGCTTTGGGCCG
Foxa2	CCCCATGCCAGGCAGCTTGG	AAGTGTCTGCAGCCAGGGGC
Sox1	TTCCCCAGGACTCCGAGGCG	GTTCAGTCTAAGAGGCCAGTCTGGT
Arx	AAGCATAGCCGCGCTGAGGC	TTCGGGGAACGCCCTAGGGG
Lnsm1	TACAGCTCCCCGGGCCTGAC	ACTCTAGCAGGCCGGACGCA
Pax6	ACCTCCTCATACTCGTG	ACTGATACCGTGCCTT
Dbx1	GACGTGCAGCGGAAAGCCCT	CGCTAGACAGGAGCTCGCGC
Dmrt3	AACCGGCCACCCCTGGAAGT	GTCGCCCCCGCAACCTTTCA
Hes5	TCCGACCCCGTGGGGTTGTT	TCTACGGGCTGGGGTGAGCC
Neurog2	ACACGAGACTCGGGCGAGCT	CCGGAACCGAGCACGGTGTC
Lhx2	TGGGCTCAGCCGGGGCTAAT	ACAGCTAAGCGCGGCGTTGT
Pax5	ACACTGTGCCCAGCGTCAGC	GCACTGGGGGACGTGATGCC
Lhx5	GAGCTCAACGAAGCGGCCGT	CCGAGAAATTGCGCAGGCGC
Sox2	GCACATGAAGGAGCACCCGGA	GGTTCACGCCCGCACCCAG
Asb5	GGGACACGCCACTGCATGCT	GCCAAGTCGACAGGCCGCAA
Lmx1a	TGACGTCATGCCCGGGACCA	GCCCCCTACACCCGCCTCAT
Pax3	CCCCCACCTATAGCACCGCAGG	ACATGCCTCCAGTTCCCCGTTCT
Hoxa5	AGGGAACCGAGTACATGTCCCAGT	TGCAACTGGTAGTCCGGGCCA
Triml2	TGCGCAGCCTCCAGACGATG	TCTGGAGCAGTGCAACGGCA
Afp	TTCCTCCCAGTGCGTGACGGA	TCCTCGGTGGCTTCCGGAACA
Dppa3	CCGGCGCAGTCTACGGAACC	ACCGACAACAAAGTGCGGACCC
Fgf8	GCGAAGCTCATTGTGGAGAC	CACGATCTCTGTGAATACGCA
Nodal	ACCAACCATGCCTACATCCAGAG	CCCTGCCATTGTCCACATAAAGC
Epha1	TACGCCTGCCCAGCCTGAGT	GGTGTCCAGCCCAGCCGAAC
Rab25	TCAGCCAGGCCCGAGAGGTC	GATGGCACTGGTCCGGGTGC
Evx1	GAGTGGCGTCACCAGCGGTACT	TCACCTTGTGATGCGAGCGC
Lrrc6	GGGAAATCCTGCCTGCCGGTC	CTGTGATTCGGCCCATGGTGCTT
Pou6f1	CGCCTTTCCTGCCTGGTGGG	GCTAGCAGTGGGCAGTGGCC
Pgr	CGCCATCTACCAGCCGCTCG	ACTGTGGGCTCTGGCTGGCT
Foxa3	TTTGGGGGCTACGGGGCTGA	TGCAGCCCACGCCCATCATG
Ell2	TGCAGGCCTCCTACCACCCC	TCCCCAGGCCTTCTGGAGTGC
Lbh	ACGTTGGGGCAAGAGCGTGG	GAGACGGGGGAGGGGGTGAC
Etv4	GAAGGTGGCTGGCGAACGCT	GCGGGGCCAGTGAGTTCTGG
Klf9	CCGCGTACTCGGCTGATGCC	CACACGTGGCGGTCGCAAGT
Wnt3a	ACCAAGACCTAACAAACCC	CATGGACATCACGGACC
Prdm1	GCCGAGGTGCGCGTCAGTAC	GGGGCAGCCAAGGTCGTACC
Ankrd1	ACGCAGACGGGAACGGAAGC	TGCGGCACTCCTGACGTTGC
Per2	GGTGGCCTCTGCAAGCCAGG	CCTCCGTGCTCAGTGGCTGC
Hes1	CCCTGCAAGTTGGGCAGCCA	CGAAGGCCCCGTTGGGGATG
Bnc1	GCTGGAGCACCTGGGTGAGC	CCTCCACTGTGCACGCGTGT
Foxc2	AGGGACTTTGCTTCTTTTTCCGGGC	CCCGCAGCGTCAGCGAGCTA
Prdm6	CCGGCCTTTCAAGTGCGGCT	GGCATGCGCTGGTGTCGACT
Armc4	GCATCCCCTTGCTGGCTCGG	GGCCATGGCACAGTGCTCCT

Table S2.6: List of quantitative RT-PCR primers used in the study. (Continued)

Gene	Forward Primer	Reverse Primer
Cxcr4	TACCCCGATAGCCTGT	GCACGATGCTCTCGAA
Tbx3	CCAAGCGATCACGCAACGTGG	CTCTGACGATGTGGAACCGCGG
Arg1	GCGAGACGTAGACCCTGGGG	GGTCGCCGGGGTGAATGCTG
Foxq1	GGAGCCGCCGCAGGGTTATATTG	TGGCGCACCCGCTACTTTTGAG
Asb4	TCACCTCCGTGCGTCCTGCT	TTCGGGCAAGAGTGGCAAGCC
Six2	ACTCGTCGTCCAGTCCCGCTC	CAAGGTTGGCCGACATGGGGT
Lhx1	ACTAGGGACCGAGGGACGCG	CAGTTTGGCGCGGATTGCCG
Sox17	GAGCCAAAGCGGAGTCTC	TGCCAAGGTCAACGCCTTC
Cer1	AGAGGTTCTGGCATCGGTTCA	TCTCCCAGTGTACTTCGTGGC
Creb3l1	ACAGGACGGACACCCTGGCA	GGTCAGCCCAGGGGAGCAGT
Bcl6	AAGCACGGCGCCATCACCAA	TTTGGGGAGCTCCGGAGGCA
Hey1	AATGGCCACGGGAACGCTGG	CACCACGGGAAGCACCGGTC
Basp1	AGGGGGCGGGGAGAATCCAAA	GGAGCCTAGGGGACAGCGGTT
β-Actin	GCTGTATTCCCCTCCATCGTG	CACGGTTGGCCTTAGGGTTCAG

2.8 Acknowledgements

This work was supported by National Institutes of Health (NIH), National Institute of Diabetes and Digestive and Kidney Diseases Grant P01-DK074868 (S.S.), National Heart, Lung and Blood Institute Grant 5 R33 HL087375-02 (S.S. and M.M.), and NIH/National Institute of General Medical Sciences Grant GM078005-05 (S.S.). V.B. was recipient of a California Institute for Regenerative Medicine (CIRM) Graduate Fellowship (T1-00003 and TG2-01154, Interdisciplinary Stem Cell Training Program at UCSD). E.W. was a fellow supported by a CIRM postdoctoral training grant T2-00004 at SBMRI and by an AHA postdoctoral fellowship. The authors especially thank Wenqing Cai for sample preparation, data analysis and useful discussions, Dr. Suresh Subramani, Dr. Brian Saunders, Harish Nagarajan, Shakti Gupta and Dr. Gaurav Agarwal for many helpful discussions and Dr. Juan Carlos Izpisúa Belmonte for providing T-GFP mESCs.

Chapter 2, in full, is adapted from **Bhargava, V.,** Ko, P., Willems, E., Mercola, M. & Subramaniam, S. (2013) *Quantitative Transcriptomics using Designed Primer-based Amplification.* **Sci. Rep.** 3, 1740; DOI:10.1038/srep01740. The dissertation author was the primary author of this paper responsible for the research.

Chapter 3

Technical Variations in Low Input RNA-seq Methodologies

3.1 Abstract

Amplification-based strategies are essential to generate RNA-sequencing libraries from ultra-low amounts of mRNA, such as sequencing from a single or a few cells. However, the transcriptomics data obtained from single or a few cell RNA is often noisy resulting in poor quantification of the majority of low expressed transcripts. Here, we generated sequencing libraries from serial dilutions of mRNA using three such methods, viz., Smart-seq, CEL-seq and DP-Seq, to perform whole transcriptome comparative analysis and characterize technical variations intrinsic to each method. Regardless of the method used, reduction in mRNA levels resulted in inefficient amplification of a majority of low to moderately expressed transcripts. Stochasticity in primer hybridization and/or enzyme incorporation was further enhanced by an amplification step resulting in greater uncertainty in transcript quantification. Additionally upon comparison with standard RNA-seq and real time quantitative PCR; we noted significant distortions in fold changes of the transcripts as the amount of mRNA was reduced. Consequently, the majority of the differentially expressed transcripts were high expressed and/or exhibited high fold changes. Our analysis demonstrates that technical noise is substantially increased particularly at limiting

amounts of mRNA that could mask subtle biological differences mandating development of improved amplification-based strategies for single cell transcriptomics.

3.2 Introduction

Complex Mammalian transcriptomes display a power-law distribution in transcripts abundance with transcript expression ranging over six orders of magnitude in RNA concentrations[71, 72]. RNA-seq with its large dynamic range and high sensitivity has facilitated accurate quantification of vast majority of these transcripts[8, 10, 73]. The most widely used RNA-seq protocol relies on fragmentation of mRNA into short 100 – 200 bp fragments which are later converted to double stranded cDNA and processed to prepare a sequencing library[10]. Since, there is no pre-amplification step involved, this method requires at least 1 – 10 ng amounts of mRNA making it difficult to apply the method to instances such as stem cell and cancer biology where it is difficult to obtain large quantities of mRNA.

To address this issue of sequencing from limiting amounts of mRNA, a number of amplification-based methodologies[23-25, 31, 34, 74-76] were recently proposed. These methodologies generate large amount of amplified cDNA, required for successful production of sequencing libraries, by performing either exponential or linear amplification of mRNA. In Smart-seq[25], exponential amplification of the mRNA is achieved by associating universal primer sequences to either ends of the cDNA library followed by global amplification of all the transcripts using complementary sequences of the universal primers. In another instance of exponential amplification, DP-seq[31], the hybridization and extension potential of heptamer and octamer primers are utilized to amplify majority of the transcripts. These strategies generate large amounts of amplified DNA within few hours although with high proportions of primer dimerization and/or PCR spurious products[77]. Linear amplification of the mRNA, in CEL-seq[23] method, requires incorporation of T7 promotor sequence to the cDNA template followed by in-vitro transcription (IVT) by T7 RNA polymerase that performs over 1000-fold amplification of the DNA in one round of amplification. Owing to stringent binding of the T7 RNA polymerase to its promotor region, IVT strategy results in reduced accumulation of PCR spurious products and fewer PCR biases. However, it requires at least 400 pg of total RNA for successful linear amplification, which is obtained by associating unique barcodes to individual RNA samples and pooling them together before the IVT step[23].

Sequencing library generation from these methodologies involve multiple steps that are susceptible to technical variations. During the amplification step, these variations get further amplified, often non-linearly, resulting in an increased uncertainty in quantifying low expressed transcripts[25, 78]. In this study, we investigated technical variations arising out of library preparation protocols and the sequencing platform as the amount of mRNA is reduced. We generated sequencing libraries in replicates from serial dilutions of mRNA ranging from 50 ng to 25 pg using three amplification-based methods; Smart-seq[25], DP-seq[31] and CEL-seq[23]. Two of these methods, Smart-seq and CEL-seq, have demonstrated generation of robust sequencing libraries from ultra low amounts of mRNA obtained from single cells. Our whole transcriptome analysis of these methods revealed increased technical noise particularly in the low expression regime along with stochastic loss of vast majority of low expressed transcripts as the amount of mRNA was reduced to 25 pg (25 – 50 mammalian cells). Significant distortions in fold changes of the differentially expressed transcripts, irrespective of their average expression or level of differential regulation, were observed as the amount of mRNA was reduced. Our study demonstrates that technical variations observed in these methodologies are profound which can mask subtle biological differences with only highly expressed and/or highly differentially regulated transcripts reliably estimated at reduced mRNA levels.

3.3 Results

3.3.1 Experimental Design

Smart-seq performs full-length cDNA amplification by utilizing universal primers attached to either ends of the cDNA library and a thermophilic polymerase capable of performing long distance amplifications. DP-seq uses a defined set of 44 heptamer primers to amplify >80% of the mouse transcripts by using a combination of mesophilic and thermophilic polymerases in two stages. CEL-seq, on the other hand, generates amplified RNA via in-vitro transcription by T7 RNA polymerase. These methodologies represent both exponential and linear mode of amplification of the cDNA libraries derived from low amounts of mRNA.

To directly compare these methods, we generated sequencing libraries for each method using the same mRNA source (**Figure 3.1**). The mRNA was derived from an *in vitro* cell culture based model of primitive streak (PS) induction in mouse embryonic stem cells (mESCs)[35, 36]. Activation of Activin A/TGFβ pathway is necessary for successful induction of mesoderm tissue[37-39] and endoderm tissue[40, 41]. This is achieved by introducing high dosage of Activin A (100 ng/mL; AA100) during the early stages of mESCs differentiation. Omission of Activin A, however, results in negligible activation of Activin A/TGFβ pathway leading to neuro-ectoderm induction[43]. Mouse ESCs were differentiated in serum-free conditions and the mRNA was collected at day 4 (equivalent to 6.5 – 7.5 days per coitum) from embroid bodies maintained in control serum free media (SFM) and those subjected to Activin A treatment. Next, serial dilutions of mRNA ranging from 50 ng – 25 pg were prepared. Standard RNA-seq libraries (Std. RNA-seq)[10] were prepared from 50 ng of mRNA while for rest of the dilutions (1 ng, 100 pg, 50 pg and 25 pg) sequencing libraries were prepared from the amplification-based methods. For all methods, technical replicates were prepared for each dilution to access technical variations in the library preparation protocol.

Libraries obtained from Std. RNA-seq, Smart-seq and DP-seq were subjected to single-end 100 bp sequencing in Illumina platform. Paired-end sequencing was performed for CEL-seq libraries where read 1 was used to determine the barcodes of the pooled samples while read 2 was mapped to the mouse transcriptome (**Supplementary Table S3.1**).

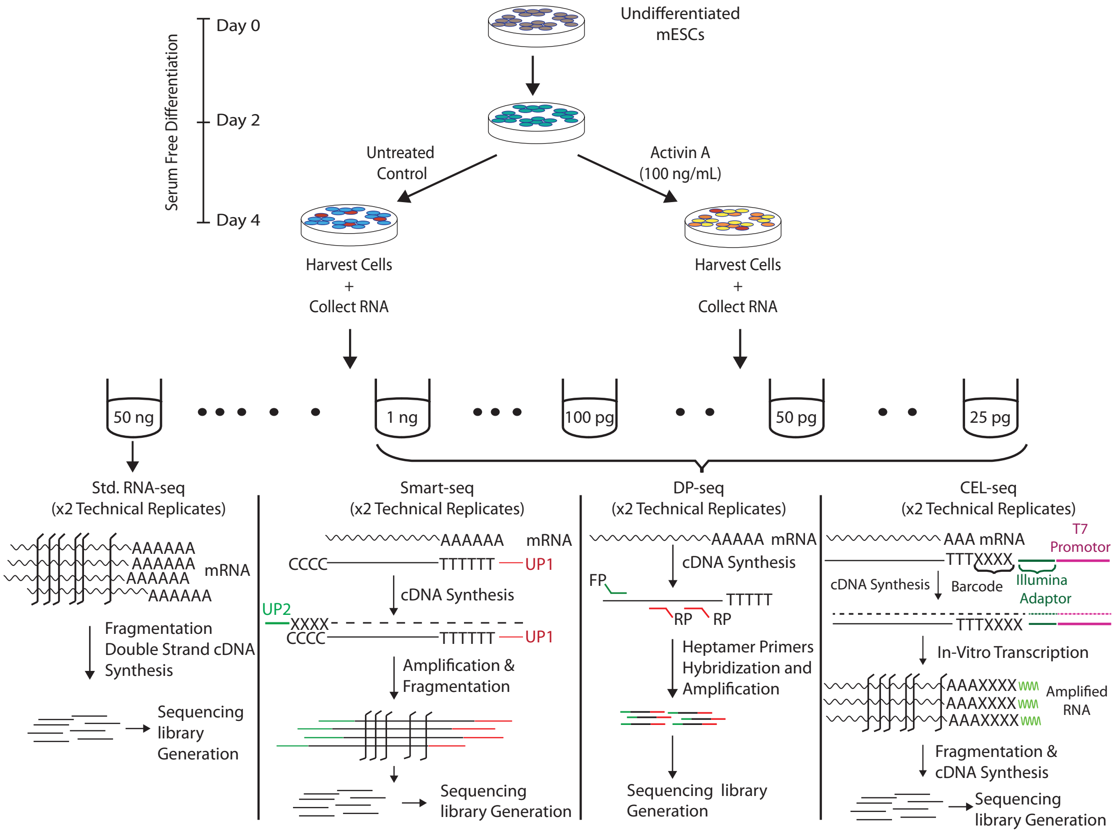

Serum Free Differentiation
Day 0
Day 2
Day 4
Undifferentiated mESCs
Untreated Control
Activin A (100 ng/mL)
Harvest Cells + Collect RNA
Harvest Cells + Collect RNA
50 ng
1 ng
100 pg
50 pg
25 pg
Std. RNA-seq (x2 Technical Replicates)
AAAAAA
AAAAAA
AAAAAA
AAAAAA
mRNA
Fragmentation Double Strand cDNA Synthesis
Sequencing library Generation
Smart-seq (x2 Technical Replicates)
AAAAAA
mRNA
CCCC
TTTTTT
UP1
cDNA Synthesis
UP2
XXXX
CCCC
TTTTTT
UP1
Amplification & Fragmentation
Sequencing library Generation
DP-seq (x2 Technical Replicates)
AAAAA mRNA
cDNA Synthesis
FP
TTTTT
RP
RP
Heptamer Primers Hybridization and Amplification
Sequencing library Generation
CEL-seq (x2 Technical Replicates)
AAA mRNA
T7 Promotor
TTTXXXX
cDNA Synthesis
Barcode
Illumina Adaptor
TTTXXXX
In-Vitro Transcription
AAAXXXX
AAAXXXX
AAAXXXX
Amplified RNA
Fragmentation & cDNA Synthesis
Sequencing library Generation

Figure 3.1: Schematic representation of the experimental design. Mouse ESCs were differentiated in serum free conditions for four days. At day 2 of differentiation, embroid bodies were dispersed and Activin A was added to the culture media to stimulate Activin A/TGFβ signaling pathway. Cells were harvested at day 4 from control (SFM) and Activin A containing well (AA100) and mRNA was isolated. The mRNA was later subjected to serial dilutions ranging from 50 ng – 25 pg. Std. RNA-seq libraries were prepared from 50 ng of mRNA derived from control and AA100 samples. Sequencing libraries were prepared from serial dilutions (1 ng, 100 pg, 50 pg and 25 pg) of mRNA using Smart-seq, DP-seq and CEL-seq. All sequencing libraries were prepared with two technical replicates where same mRNA source was used and the library preparation steps were replicated. Salient details of all the methods are shown.

3.3.2 Comparative transcriptomics analysis of the three amplification-based methods

Since, sequencing libraries were sequenced at different depths, we randomly selected 16 million reads from each library to perform comparative analysis. Transcriptome coverage obtained from all amplification-based methods was high for libraries prepared from 1 ng of mRNA. However, the transcriptome coverage dropped as the amount of mRNA was successively reduced (**Figure 3.2A**). Smart-seq libraries exhibited the highest transcriptome coverage at all amounts of mRNA explored. DP-seq was designed to amplify >80% of the transcripts and as such it exhibited marginally less transcriptome coverage as compared to Smart-seq. CEL-seq's transcriptome coverage showed the greatest reduction in coverage as the amount of mRNA was reduced. We further determined that the transcripts that lost their representation in the sequencing libraries prepared from 25 pg of mRNA by all methods were low expressed (**Supplementary Fig. S3.1A**).

Exponential amplification of mRNA has previously been shown to accumulate primer-dimers and PCR spurious products as the number of amplification cycles are increased[77]. Mapping statistics of the libraries revealed high proportions of PCR spurious products in DP-seq libraries specifically at low amounts of mRNA (**Figure 3.2B**). On the other hand, Smart-seq libraries possessed the smallest proportions of unmapped reads. CEL-seq libraries demonstrated about 20% unmapped read for all dilutions of mRNA although a slightly higher proportion of reads mapped to genomic (excluding NCBI Refseq database) locations in comparison to the other methods.

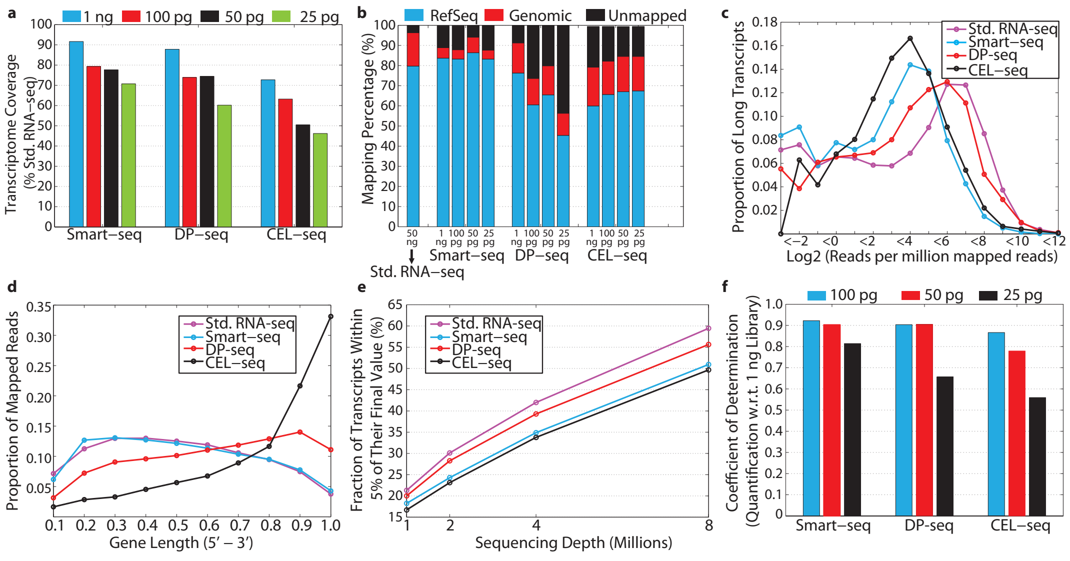


Figure 3.2: Comparative transcriptomics analysis between all methods. (A) Transcriptome Coverage. Transcriptome coverage obtained by amplification-based methods was normalized to the coverage obtained in std. RNA-seq libraries. (B) Mapping Statistics. DP-seq exhibited higher proportions of primer dimerization and PCR spurious products at low amounts of mRNA. (C) Length Bias. Smart-seq failed to efficiently amplify transcripts with length > 4Kb. (D) Distribution of mapped reads along the transcript length. Majority of the CEL-seq reads mapped to the last exon of the transcripts. (E) Robustness of unique reads measurements as a function of transcript expression levels and depth of sequencing. 16 million reads were taken from AA100 sequencing libraries to ascertain the expression of the transcripts. These reads were successively reduced by factor of two and number of transcripts falling within ± 5% of the final expression was determined. (F) Coefficient of determination (R^2) was estimated in global expression measurements in sequencing libraries constructed from lower dilutions of mRNA (100 pg, 50 pg, 25 pg) with the libraries made from 1 ng of mRNA.

In our previous study[31], we demonstrated the limitation of Smart-seq to efficiently amplify long transcripts (>4 Kb). DP-seq performs targeted amplification of selected regions of the transcripts; as such it does not suffer with the transcript length bias. Expectedly, the majority of the long transcripts in Smart-seq libraries possessed lower read counts in comparison to DP-seq and Std. RNA-seq (**Figure 3.2C**). Interestingly, CEL-seq also demonstrated low read counts for long transcripts. Next, we investigated the distribution of mapped reads across the length of the mRNA. Smart-seq and Std. RNA-seq libraries displayed overlapping distribution of reads across the length of the mRNA (**Figure 3.2D**). DP-seq libraries showed bias towards 3' end of the transcripts presumably because of inability of reverase transcriptase to generate full-length cDNA libraries. CEL-seq libraries, on the other hand, preferentially amplified last exons of the transcripts with vast majority of the reads mapping close to the 3' end of the transcripts.

Amplification-based methods possess variety of PCR biases. Consequently, a subset of transcripts is preferentially amplified resulting in reduced representation of the rest of the transcripts. We examined the % of unique reads occupied by top 100 highly expressed/amplified transcripts in the sequencing libraries prepared from all methods. Std. RNA-seq protocol does not involve any pre-amplification step before the library PCR amplification and as such the top 100 highly expressed transcripts occupied only 20% of the mapped reads. Amplification-based methods, on the other hand, occupied high proportions of mapped reads with CEL-seq and Smart-seq libraries showing significant enrichment of few transcripts (**Supplementary Fig. S3.1B**). The top 100 transcripts occupied 39% and 51% of the mapped reads in Smart-seq and CEL-seq libraries respectively while DP-seq occupied only 29% of the mapped reads. We further investigated the robustness in measurements of transcript expression for all methods as a function of sequencing depth. We estimated the number of transcripts within ±5% of the final expression as the sequencing reads were reduced by a factor of 2 (**Figure 3.2E**). Std. RNA-seq libraries demonstrated robust quantification for the highest number of transcripts followed by DP-seq. This observation remained unchanged for sequencing libraries prepared from varying amounts of mRNA. Finally, global transcript measurements of libraries constructed from at least 50 pg of mRNA showed high correlation with the libraries constructed from 1 ng of mRNA for all methods. However, the coefficient of determination (R^2) dropped significantly as the amount of mRNA was further reduced to 25 pg, with CEL-seq libraries showing the highest distortions in transcript expression measurements.

3.3.3 Technical variations

For all methods, we generated technical replicates to access the variations arising out of the library preparation protocols and the sequencing platform. Std. RNA-seq libraries prepared from 50 ng of mRNA showed little technical variations. For the amplification-based methods, libraries prepared from 1 ng of mRNA were highly reproducible (**Figure 3.3A**). However, the technical variations increase substantially as the amount of mRNA was reduced (**Supplementary Fig. S3.2**). DP-seq libraries prepared from 25 pg mRNA exhibited high technical variations presumably because of accumulation of PCR spurious products (**Supplementary Fig. S3.3**). CEL-seq libraries displayed significant technical variations/noise in the libraries prepared from 50 pg or less amounts of mRNA (**Supplementary Fig. S3.4**).

We hypothesized that with reduction of the starting material (mRNA), the amplification step involved in all methods becomes highly inefficient resulting in stochastic loss of a vast majority of the low expressed transcripts (RPKM<10, in Std. RNA-seq library). Expectedly, the distributions of low expressed transcripts were successively shifted towards low read counts with a majority of the transcripts losing their representation in the sequencing libraries, as the amount of mRNA was reduced (**Figure 3.3B**). Similar observations were made for DP-seq and CEL-seq. We also observed similar trends even for moderately expressed transcripts (200>RPKM>10, in std. RNA-seq library) with a majority of these transcripts failing to amplify efficiently (**Supplementary Fig. S3.5**).

Next, we estimated the technical variations in the replicate libraries by measuring the standard deviations in fold changes of the transcripts as a function of average read counts. Transcripts are not expected to be differentially regulated between the technical replicates implying that the fold changes should be close to zero. All amplification-based methods showed characteristic profiles of variations/noise as a function of average read counts with high variations reported for transcripts with low expression. Regardless of the method used, we noticed significant increase in technical variations in the libraries prepared from low amounts of mRNA (**Figure 3.3D,E and F**). This resulted in a poor quantification of the vast majority of moderate to low expressed transcripts including the transcription factor family of genes (**Figure 3.3C**).

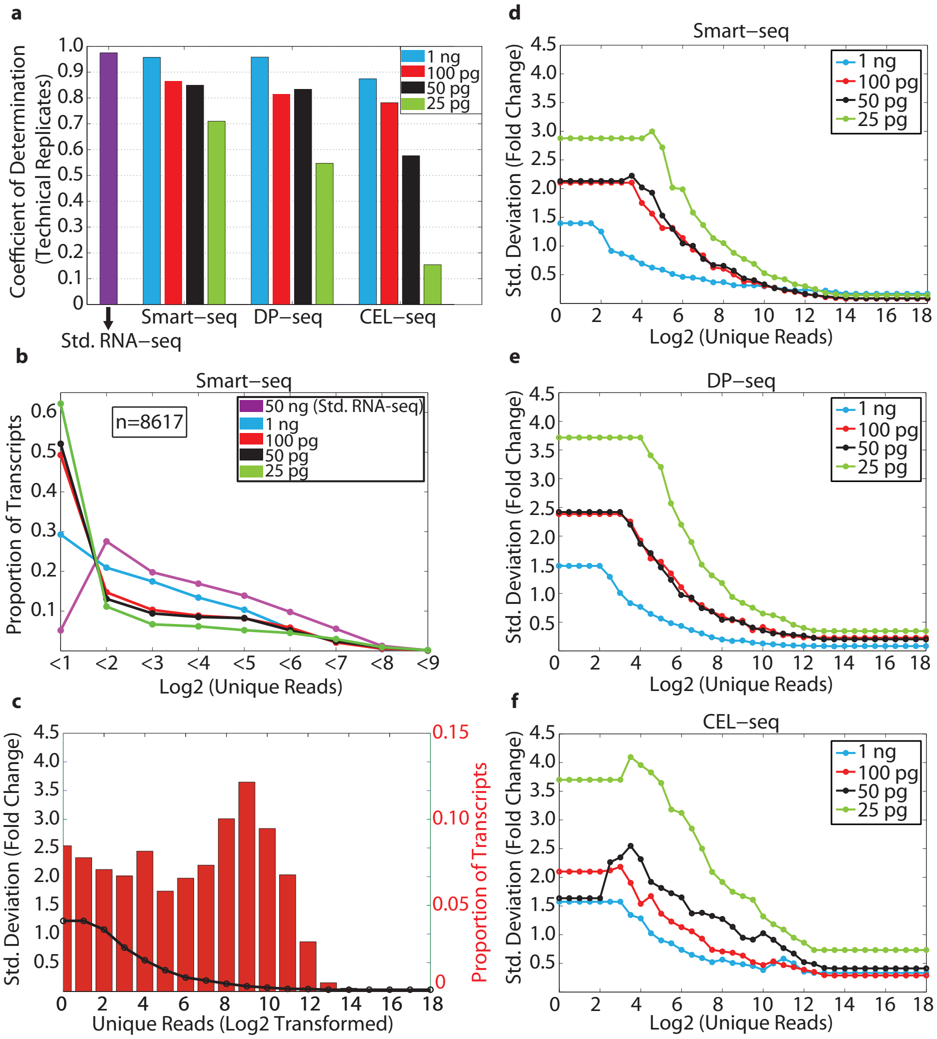


Figure 3.3: Technical Variations as a function of amount of starting material (mRNA). (A) Coefficient of determination (R^2) observed between the technical replicates in global transcriptome measurements. (B) Distribution of unique reads obtained for low expressed transcripts in Smart-seq libraries generated from different amounts of mRNA (average RPKM <10 in std. RNA-seq libraries prepared from control and AA100 samples). Similar distributions were observed for libraries prepared from DP-seq and CEL-seq. (C) Distribution of unique reads mapping to known mouse transcription factors (n=1596) for AA100 sample. The black curve represents standard deviation in fold changes observed in technical replicates of std. RNA-seq libraries as a function of average reads. Standard deviations in fold changes observed in technical replicates as a function of average reads in libraries prepared from different amounts of mRNA using (D) Smart-seq (E) DP-seq (F) CEL-seq.

3.3.4 Differential gene expression analysis

The biological system considered in our study was highly divergent (meso-endoderm vs. ectoderm) allowing us to perform detailed analysis of differential gene expression and fold changes in the sequencing libraries prepared from all methods. We used DEseq[79] to normalize the sequencing libraries and perform differential gene expression analysis. In libraries prepared from 50 ng of mRNA using Std. RNA-seq method, we identified more than 8400 differentially expressed genes (DEG). The pathway and GO term (Biological Processes) enrichments for genes up-regulated in AA100 samples contained terms specific to mesoderm/endoderm formation (**Supplementary Table S3.2, 3.3**). On the contrary, down-regulated genes were enriched for pathways and GO terms specific for ectoderm lineage. The amplification-based methods, with the exception of CEL-seq, identified a large set of DEGs for libraries prepared from 1 ng of mRNA with vast majority of them shared with those identified by Std. RNA-seq. DEGs that were not common to Std. RNA-seq were low-expressed and consequently were prone to large noise (**Supplmentary Fig. S3.6**). Additionally, the number of DEGs drastically reduced for all methods as the amount of mRNA was reduced (**Figure 3.4A**). On the other hand, CEL-seq libraries consistently identified low numbers of DEGs with only 26 differentially regulated transcripts identified for libraries prepared from 25 pg of mRNA.

PCR biases associated with Smart-seq resulted in preferential amplification of highly expressed and short transcripts. Owing to high technical variations in the low read counts, DEGs identified in the Smart-seq libraries generated from low amounts of mRNA, were highly expressed and shorter in length (**Supplementary Fig. S3.7**). DP-seq, on the other hand, performed targeted amplification of selected regions of the mouse transcriptome and hence did not suffer from the transcript length bias.

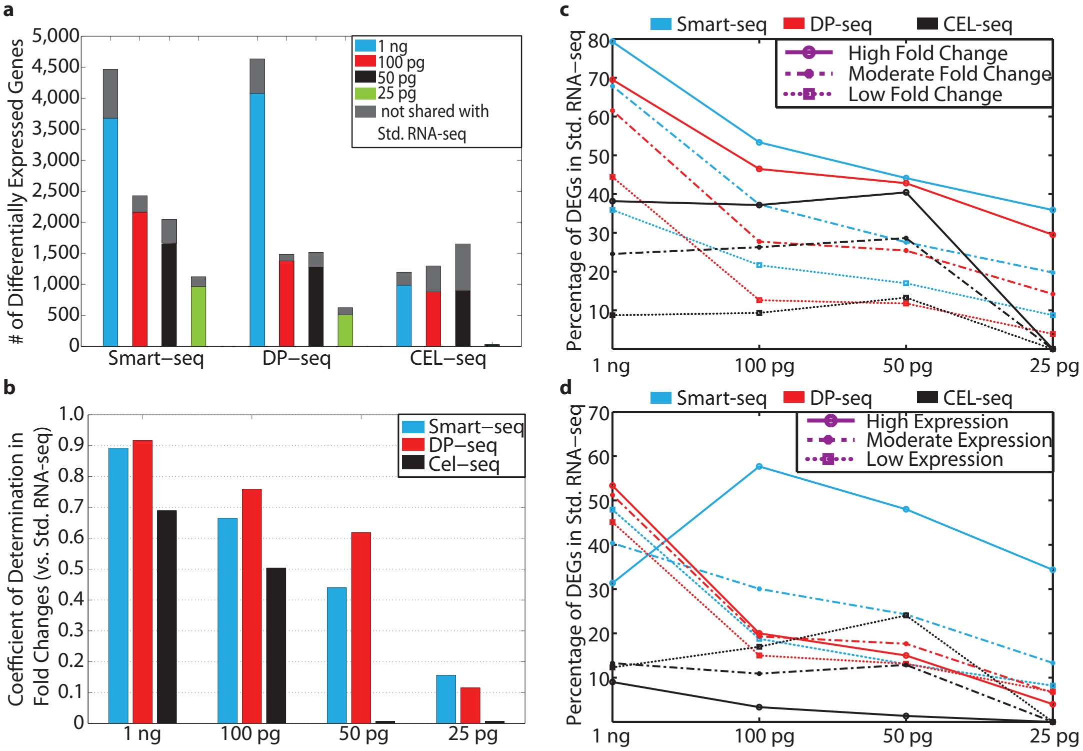


Figure 3.4: Differential gene expression analysis. (A) Differentially expressed genes identified from the sequencing libraries prepared from different amounts of mRNA. (B) R^2 between the fold changes of differentially expressed genes observed between amplification-based method and std. RNA-seq. (C) Differentially expressed genes identified from std. RNA-seq libraries were classified into three categories of differential expression: High (fold change>4, log2 scale), Moderate (4>fold change>2, log2 scale) and Low (fold change<2, log2 scale). Proportions of these genes identified by amplification-based methods as a function of the amount of mRNA used for library preparation, are plotted. (D) Differentially expressed genes identified from std. RNA-seq libraries were classified into three categories of transcript expression: High (RPKM>200), Moderate (200>RPKM>10) and Low (RPKM<10). Proportions of these genes identified by amplification-based methods as a function of the amount of mRNA used for library preparation, are plotted.

The fold change distributions observed for DEGs in libraries prepared from high amounts of mRNA revealed high proportion of transcripts with small fold changes. However, as the amount of mRNA was reduced, more transcripts with high fold changes were identified as differentially regulated and transcripts with low fold changes were lost because of high technical variations (**Supplementary Fig. S3.8**). We next sought to compare the fold changes of the DEGs identified for each amplification-based method to the fold changes obtained in Std. RNA-seq libraries. DP-seq demonstrated higher correlations in the fold changes in comparison to Smart-seq (**Figure 3.4B**). CEL-seq libraries showed significant distortions in the fold changes. Interestingly, these correlations dropped significantly for all methods as the amount of mRNA was reduced (**Supplementary Fig. S3.9, 3.10 and 3.11**).

Identification of DEGs was severely affected for amplification-based methods owing to high technical variations and significant fold change distortions. We next investigated which characteristics are necessary for a DEG to be identified by the amplication-based methods. DEGs identified by the standard RNA-seq method were classified into different categories based on their fold changes. We noted that the category consisting of highly differentially regulated genes (>16 fold change) were consistently identified by all the methods. However, the identification for moderate (16>fold change>4) and low (fold change<4) DEGs was poor. Importantly, all three categories of DEGs suffered heavy loss as the amount of mRNA was reduced, irrespective of the method used (**Figure 3.4C**). A similar analysis was performed where DEGs identified in std. RNA-seq were classified into different categories based on their average expression. Smart-seq identified larger proportions of highly expressed (RPKM>200) DEGs as compared to moderate (200>RPKM>10) and low (RPKM<10) expressed genes (**Figure 3.4D**). Since DP-seq distorts the relative order of gene expression, it did not discriminate based on the gene expression and identified similar proportions of DEGs for the three categories of expression. CEL-seq, because of high technical noise even at high expression, failed to identify a vast majority of highly expressed DEGs. We again noticed that the proportions of DEGs identified by all methods dropped significantly as the amount of mRNA was reduced. This analysis demonstrates that with low amounts of starting material (mRNA), only highly expressed and/or highly differentially regulated transcripts are expected to be identified by these methods.

3.3.5 Distortion in Fold Changes

Activation of Activin A/TGFβ pathway by introduction of Activin A in differentiating mESCs is well documented[35, 36]. Our sequencing libraries showed differential regulation of a majority of TGFβ target genes[31]. Overall, Smart-seq and DP-seq displayed similar profiles for both up and down-regulated TGFβ target genes. CEL-seq, on the other hand, displayed similar trends although with suppressed fold changes. Moreover, heterogeneity in the fold changes of the TGFβ target genes was apparent for the libraries prepared from low amounts of mRNA (**Figure 3.5A**).

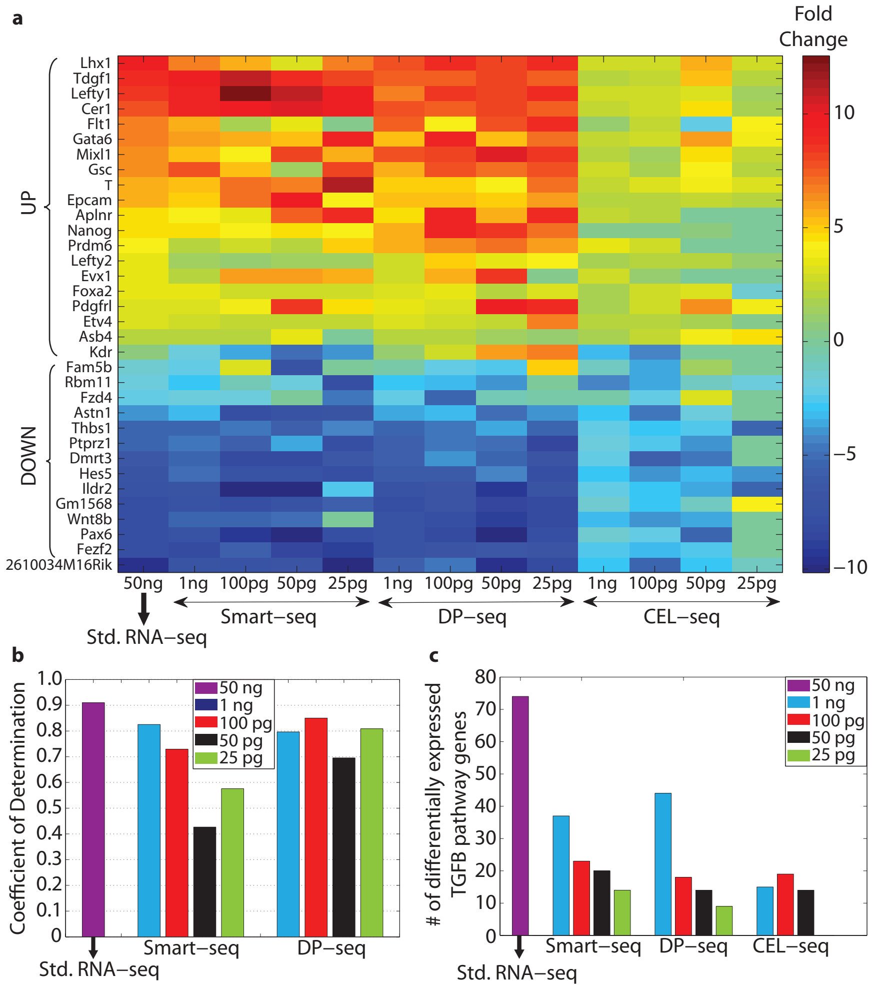


Figure 3.5: Expression of Activin A/TGFβ pathway target genes in day 4 mouse embryoid bodies. (A) Heatmap displaying up/down regulation of Activin A/TGFβ pathway target genes upon introduction of Activin A in the culture media in comparison to the control. (B) Number of Activin A/TGFβ pathway genes identified as differentially regulated. (C) R^2 between the fold changes observed in the sequencing libraries and quantitative real time PCR for 40 transcripts including the TGFβ target genes and lineage markers.

We next compared fold changes in transcripts expression observed in our sequencing libraries to the gold standard measurements obtained from real time quantitative PCR (qPCR). For this analysis, we selected 40 transcripts representing TGFβ target genes and known lineage markers (ectoderm and mesendoderm). Majority of these genes have moderate to low expression in the Std. RNA-seq libraries. Libraries prepared from Std. RNA-seq method conserved the relative abundance of these transcripts. However, Smart-seq libraries displayed significantly lower R^2 as the amount of mRNA was reduced. Interestingly, DP-seq showed strong correlations for all amounts of mRNA used (**Figure 3.5B**). Fold changes obtained from CEL-seq libraries showed significant distortions resulting in a poor correlation with the qPCR fold changes.

Finally, we investigated whether or not increased technical noise and distortions in fold changes result in loss of subtle biological differences as the amount of mRNA is reduced. Out of the 181 Activin A/TGFβ pathway associated genes, 74 genes were differentially regulated in mESCs treated with a high dosage of Activin A in Std. RNA-seq libraries. Regardless of the method used, the number of DEGs associated with the Activin A/TGFβ pathway reduced significantly as the amount of mRNA was reduced (**Figure 3.5C**). This highlights the issue of increased technical variations in the sequencing libraries prepared from limiting amounts of mRNA that could potentially result in loss of biological context.

3.4 Discussion

Current sequencing technologies require nanogram quantities of RNA before it could be processed and made compatible for high-throughput sequencing. This motivated the development of amplification-based strategies to generate whole transcriptome profiles from limited number of cells. The transcriptomics data obtained from these strategies have shown expression of thousands of transcripts, although accurate quantification of these transcripts has been marred by high technical variations. In this study we compared the performance of three amplification-based strategies in generating robust sequencing libraries from limiting amounts of mRNA. Two of these methodologies, Smart-seq and CEL-seq, amplify mRNA using two different modes of amplification - exponential and linear respectively. Moreover, both of them have been demonstrated to generate libraries from mRNA derived from a single cell. DP-seq, on the other hand, performs exponential amplification of the transcripts from as low as 50 pg of mRNA by utilizing a defined set of 44 heptamer primers. Serial dilutions of mRNA were prepared ranging from 1 ng – 25 pg to characterize technical variations arising out of library preparation protocols and access the consequences of these variations on fold change estimations of the transcripts and biological interpretation of the datasets.

Comparative analysis revealed that Smart-seq and DP-seq exhibited similar transcriptome coverage and comparable technical variations in the libraries prepared from up to 50 pg of mRNA. Smart-seq preferentially amplified highly expressed and/or short transcripts resulting in a larger proportion of such transcripts being identified as differentially regulated. DP-seq showed high technical variations and accumulation of PCR spurious products at 25 pg libraries. One of the challenges with DP-seq is that the 44 heptamer primers are split into three tubes which implies that only 8.33 pg of mRNA was amplified by each tube. A better primer design where more primers can be accommodated into a single tube to get similar transcriptome coverage while ensuring that no two primers have ΔG <-4 Kcal/mol, is expected to reduce the technical noise. DP-seq libraries were more consistent in maintaining relative abundance of the transcripts in comparison to Smart-seq and provided better quantification of the transcripts as a function of sequencing depth. In terms of cost, both Smart-seq and CEL-seq had higher cost for library generation than DP-seq with CEL-seq requiring paired-end sequencing. The first read obtained from CEL-seq libraries was used only for barcode identification while read 2 was used for mapping purposes. Additionally, CEL-seq required longer time to construct sequencing libraries and more time was spent handling less stable RNA.

CEL-seq has been shown to produce highly reproducible libraries from limiting amounts of mRNA[23], however, in our hands the libraries exhibited high technical variations and significant fold change distortions in comparison to the other methods. Even though the CEL-seq libraries showed expression of thousands of transcripts, the transcriptome coverage was considerably lower than that of Smart-seq and DP-seq. CEL-seq requires at least 400 pg of total RNA for successful IVT reaction. In our library preparation, we associated different barcodes to cDNA libraries prepared from same amounts of mRNA and pooled them for IVT reaction. For instance in the case of lowest dilution, the pooled cDNA was prepared from 100 pg of mRNA (25 pg x 2 biological samples x 2 technical replicates) which is substantially higher than 400 pg of total RNA requirement. We also noted a high proportions of reads (>80%) mapping to the mouse genome thus ruling out contamination. The authors of CEL-seq protocol noted that transcripts expressed at low levels were not efficiently amplified during the IVT step and we suspect that the incorporation of T7 RNA polymerase to its promotor region is subjected to high noise resulting in substantial technical variations.

Previous analysis of variations in RNA-seq libraries revealed little technical variations in libraries prepared from large amounts of mRNA[7, 10, 68, 69]. However, in amplification-based methods technical variations intrinsic to the library preparation steps prior to the amplification step get significantly amplified resulting in high variations especially in low read counts. These variations are expected to remain high even at high sequencing depths. Regardless of the method used, technical variations substantially increased as the amount of mRNA was reduced. It was accompanied by poor amplification of majority of low to moderately expressed transcripts with the distributions of the transcripts shifted towards low read counts. High proportions of low expressed transcripts were consequently lost as the amount of mRNA was reduced.

We compared fold changes in transcript expression obtained in all amplification-based methods to those in Std. RNA-seq libraries. This facilitated the transcriptome-wide analysis of fold changes of the transcripts, spanning the entire dynamic range of transcript expression. We further selected 40 transcripts, which included TGFβ target genes or lineage makers, and performed qPCR to determine precise measurements of their relative abundance. Majority of these transcripts exhibited moderate to low expression. This analysis was sufficient to draw conclusions on dynamic range and sensitivity of the amplification-based methods and as such we decided not to use spike-in controls including ERCC libraries[80].

Expectedly, Std. RNA-seq libraries prepared from 50 ng of mRNA conserved relative abundance of the transcripts with high correspondence to the qPCR readouts. As the amount of mRNA was reduced, distortions in fold change estimations of the transcripts became evident for all the methods. Smart-seq libraries prepared from lower amounts of mRNA (<= 50pg) showed considerable drop in correlations to std. RNA-seq and qPCR. DP-seq libraries performed significantly better than Smart-seq in conserving the fold changes of the transcripts expression. Fold change distortions were significant for libraries generated from CEL-seq with a vast majority of differentially regulated transcripts displaying suppressed fold change estimations.

As a consequence of the increased technical noise, loss of low abundant transcripts and significant distortions in the fold change estimations, the number of transcripts identified as differentially regulated dropped significantly in the libraries constructed from low amounts of mRNA. Our analysis of DEGs further demonstrated that transcripts, which were highly expressed and/or differentially regulated with high fold changes, were identified in low input libraries. Subtle fold changes (<4 fold) in transcript abundance increasingly fell into noisy regime and accurate quantification of transcripts expression was severely undermined. Our transcriptome data showed that at low amounts of mRNA, TGFβ target genes expression[31] followed expected trend, although with high fold change variations. More importantly, the number of differentially regulated TGFβ pathway-associated transcripts dropped considerably in these libraries. This implies that subtle biological differences between the experimental conditions are likely to be lost or diluted as the amount of mRNA is reduced. We expect biological interpretation of the transcriptome data to suffer further as the amounts of mRNA are reduced to single cell levels and biological variations[81-83] are incorporated.

Based on our analysis, we recommend implementing these methods on biological samples where transcriptomics data is expected to be highly divergent. Finally, development of new amplification-based methodologies that perform amplification with high fidelity is warranted. Quatz-seq[74] has shown potential to generate robust sequencing libraries from low amounts of mRNA by performing suppression PCR to eliminate PCR spurious products and reducing the loss of material by performing multiple enzymatic reactions in the same tube. Similar improvements in designing new enzymes that work at low temperatures with greater fidelity, reducing volume of the reactions and minimizing the loss of mRNA at different steps leading up to the amplification step will substantially reduce the technical variations in the low-input libraries.

3.5 Materials and Methods

3.5.1 Mouse embryonic stem cell culture and differentiation

Mouse R1 embryonic stem cells were cultured on mouse embryonic fibroblast (MEF) on gelatin-coated dishes in high glucose DMEM (Hyclone, Logan, UT) supplemented with 10% fetal calf serum (FCS) (Hyclone, Logan, UT), 0.1 mM b-mercaptoethanol (GIBCO), 1% non-essential amino acids (GIBCO), 2 mM L-glutamine (Sigma, St. Louis, MO), sodium pyruvate (Sigma), antibiotics (Sigma), and 1,000 U/ml of LIF (Sigma) and passaged with 0.25% Trypsin (GIBCO).

For embryoid body (EB) differentiation, MEF were stripped from the cultures by 15 minutes incubations on gelatin-coated dishes. mESCs were collected and washed in PBS to remove traces of serum. mESCs were differentiated in serum free media containing N2 and B27 supplements as described elsewhere[35, 36]. mESCs were aggregated at 50,000 cells/ml in non-coated polystyrene plates. After 2 days, EBs were dissociated by trypsin treatment and re-aggregated in fresh media in presence of Activin A at a dosage of 100 ng/mL (AA100). Activin A was obtained from R&D. EBs were harvested at day 4 for RNA extraction and processing.

3.5.2 mRNA purification and dilution series

Total RNA was extracted from the harvested cells using Trizol (Invitrogen). Total RNA was later subjected to Oligo(dT) selection using Dynabeads mRNA Purification Kit (Invitrogen) for extraction of poly-adenylated RNA. The enriched mRNA was later quantified using Nanodrop 2000 and serial dilutions were made ranging from 50 ng – 25 pg of mRNA.

3.5.3 Library Generation using Std. RNA-seq protocol

Std. RNA-seq[10] libraries were constructed from about 50 ng of mRNA derived from the serum free media control and Activin A (100 ng/mL) samples using Illumina's TruSeq RNA Sample Prep Kit v2.

3.5.4 Library Generation using Smart-seq

Smart-Seq cDNA library generation and amplification was performed on mRNA dilutions (1 ng, 100 pg, 50 pg and 25 pg) derived from serum free media control and Activin A (100 ng/mL) in duplicates using SMARTer Ultra Low RNA Kit for Illumina sequencing (Clontech). Following PCR cycles were used for amplification:

1 ng – 12 cycles

100 pg – 14 cycles

50 pg – 14 cycles

25 pg – 15 cycles

These libraries were later sheared using Covaris system to obtain 200-500 bp fragments. Later, standard Illumina library preparation protocol was followed to prepare the sequencing libraries using Illumina Paired-End DNA Sample Prep kit.

3.5.5 Library Generation using CEL-seq

CEL-seq libraries were constructed using the protocol described earlier. We used CEL-seq primers # 37, 38, 39 and 40 to generate double stranded cDNA libraries from same amount of mRNA (including the technical replicates). The libraries were later pooled together for in-vitro transcription reaction. Similar strategy was implemented for all mRNA dilutions. The PCR cycles used for varying amounts of mRNA are:

1 ng – 13 cycles

100 pg – 15 cycles

50 pg – 15 cycles

25 pg – 15 cycles

To avoid loss of the material, we replaced the column purification steps involved in the protocols with Agencourt RNAClean XP purification system.

3.5.6 Library Generation using DP-seq

mRNA dilutions (1 ng, 100 pg, 50 pg and 25 pg) prepared from the serum free media control and Activin A (100 ng/mL) were subjected to DP-seq library preparation as described. First strand cDNA synthesis was performed for all mRNA dilutions in duplicates to get the technical replicates. Later, the purified cDNA prepared from each dilution, was split into three tubes to perform amplification using our heptamer primers. The numbers of PCR cycles were increased for lower dilutions to get appropriate amounts of DNA for the library construction. The numbers of PCR cycles used are as follows:

1 ng – 14 cycles

100 pg – 17 cycles

50 pg – 17 cycles

25 pg – 18 cycles

The amplicon libraries thus constructed, were phosphorylated and ligated with Illumina's Y-adaptors and amplified using adaptor specific primers consisting of a different Illumina's Truseq barcode sequence for each library. The amplified libraries were run through the 2% agarose gel and size selected (150 – 500 bp) for sequencing.

3.5.7 Quantification of the sequencing library

Quantitative real time PCR was used to determine the concentration of the sequencing libraries prepared from DP-seq method. The standard curve for various dilutions of phiX control library was generated using the adapter specific primers recommended by Illumina. We later used the standard curve to determine the molarity of our sequencing libraries. Libraries prepared from std. RNA-seq, Smart-seq and CEL-seq were quantified using Qubit Flurometer (Invitrogen) according to the manufacturer's protocol.

The concentration of sequencing library loaded into the flowcell was calibrated by the sequencing facility. We typically obtained good cluster density with 5 pM of library concentration on HiSeq v3 kit.

3.5.8 Reverse Transcription and Quantitative RT-PCR (qPCR)

Total RNA was extracted from cells using Trizol (Invitrogen) according to the manufacturer's instructions. About 1 μg of total RNA was treated for DNA removal and converted into first strand cDNA using Quantitect Reverse Transcription kit (Qiagen). SYBR Green qPCR was run on a LightCycler 480 (Roche) using the LightCycler 480 SYBR Green Master Kit (Roche). All primers were designed with a T_m of 60°C. Data was analyzed using the $\Delta\Delta C_t$ method, using Gapdh as normalization control, which was determined as a valid reference in mouse ESC differentiation. The primer sequences are listed in **Supplementary Table S3.6**.

3.5.9 Mapping reads

Our libraries were sequenced on HISEQ2000 platforms ((TruSeq SR Cluster Kit v3-cBot-HS and TruSeq SBS Kit v3-HS). The libraries obtained from std. RNA-seq, Smart-seq and DP-seq were sequenced as 100 bp single-end reads. For CEL-seq libraries paired-end 100 bp sequencing libraries were generated. For all the reads obtained from the methods (Read 2 for CEL-seq sequencing library), the first 7 bp were truncated and next 32 bp sequences were first aligned to the mouse NCBI Refseq database allowing upto 2 mismatches. The reads that did not map to the database were further aligned to mouse genomic locations (Build 37) using Bowtie while allowing ≤ 2 mismatches.

3.5.10 Differential gene expression analysis

DEseq method was used for sequencing library normalization and identification of differentially expressed genes. To estimate dispersions, we used “pooled-CR” method with “fit-only” sharing mode.

3.6 Supplementary Figures

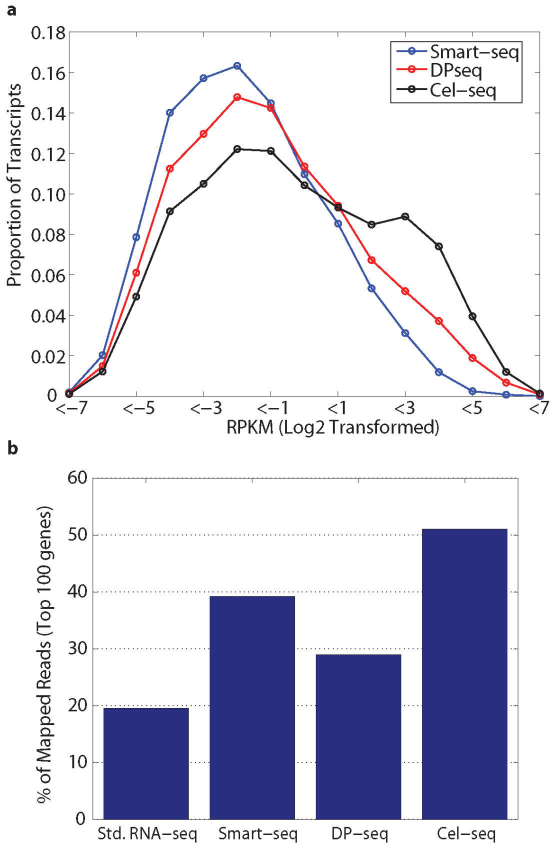


Figure S3.1: Comparative Analysis. (A) The vast majority of the transcripts that underwent stochastic loss were low expressed. (B) Percentage of unique reads represented by the top 100 highly expressed/amplified transcripts in each method. Higher % reflects more PCR bias resulting in high representation of few transcripts.

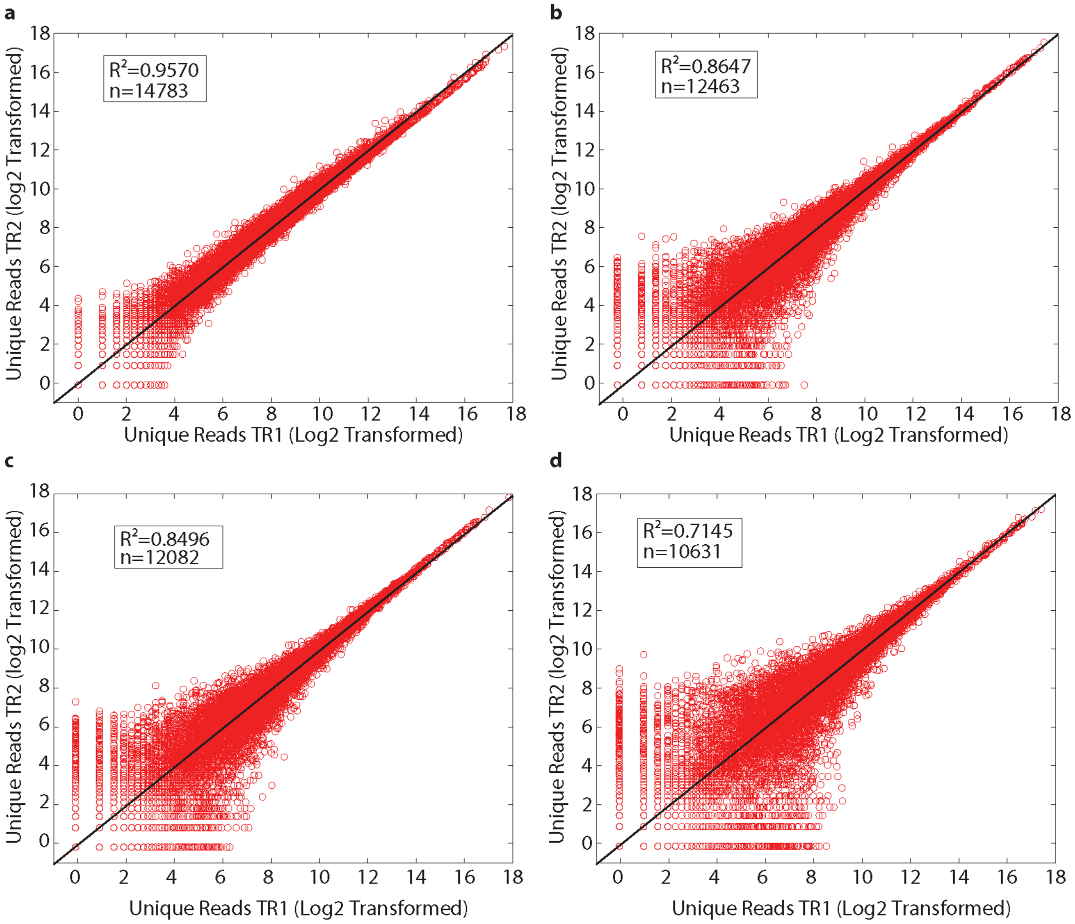


Figure S3.2: Smart-seq Technical Replicate. Coefficient of Determination (R2) between the technical replicate libraries prepared by Smart-seq from (A) 1 ng (B) 100 pg (C) 50 pg and (D) 25 pg of mRNA.

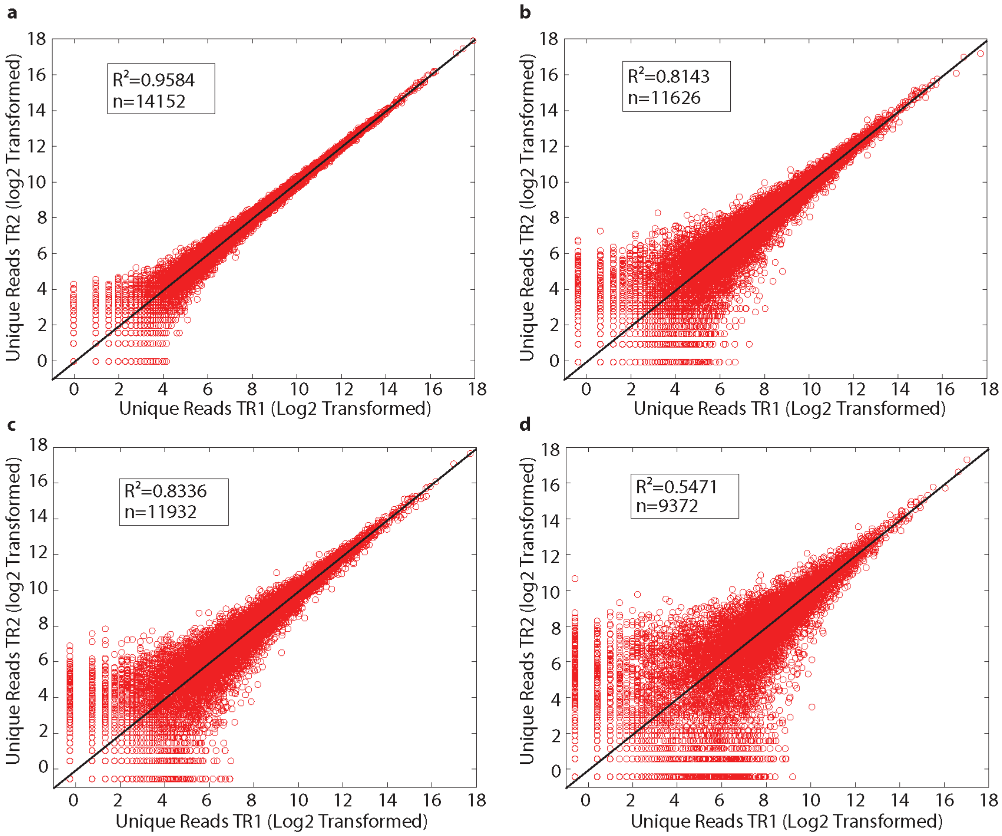


Figure S3.3: DP-seq Technical Replicates. Coefficient of Determination (R2) between the technical replicates libraries prepared by DP-seq from (A) 1 ng (B) 100 pg (C) 50 pg and (D) 25 pg of mRNA.

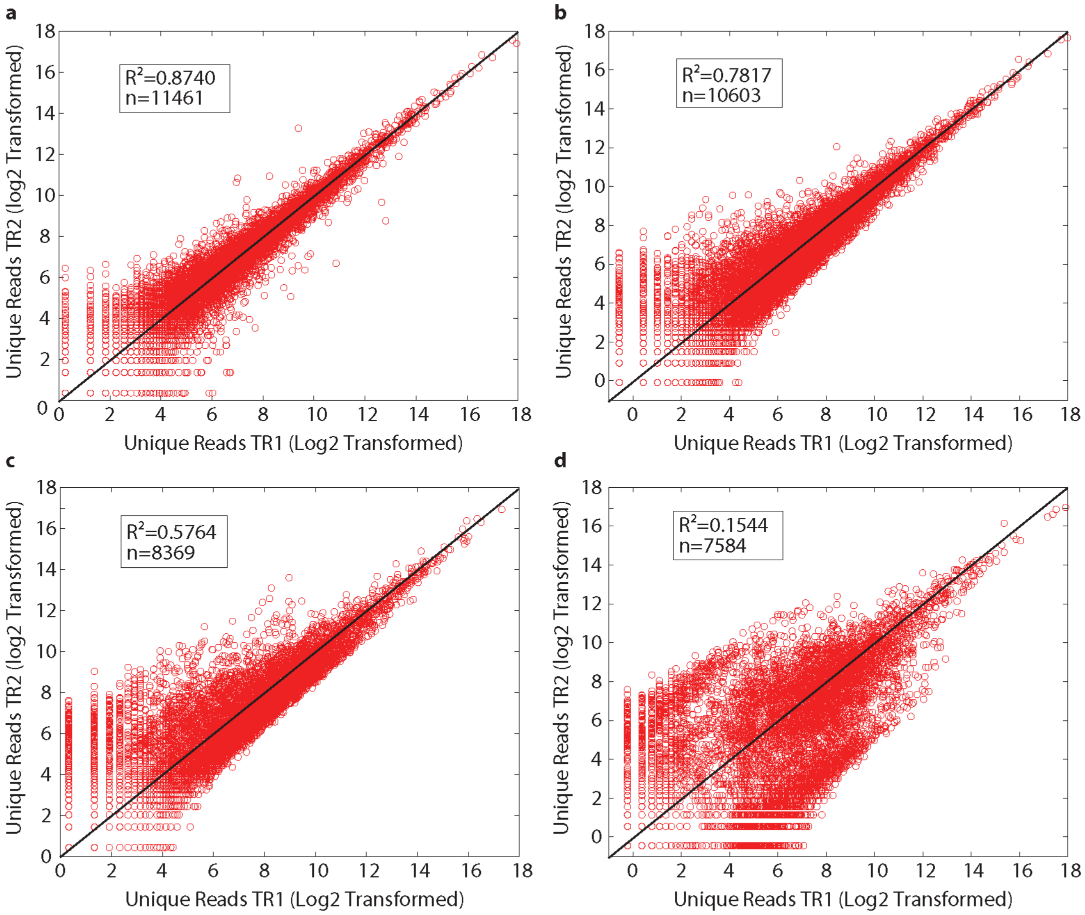


Figure S3.4: CEL-seq technical replicates. Coefficient of Determination (R2) between the technical replicates libraries prepared by CEL-seq from (A) 1 ng (B) 100 pg (C) 50 pg and (D) 25 pg of mRNA.

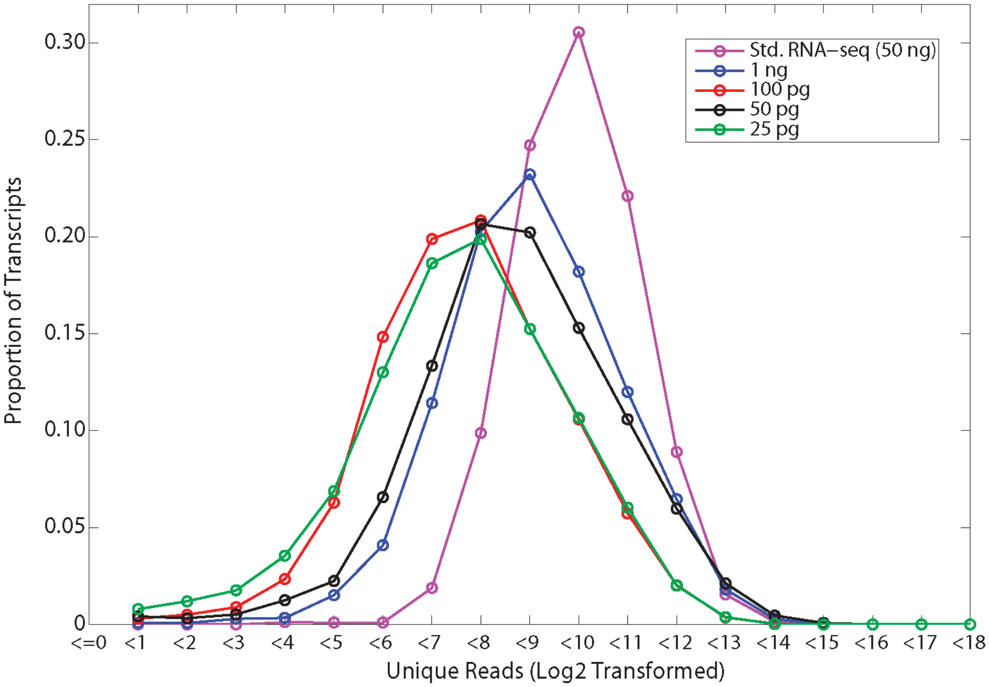


Figure S3.5: Distribution of the moderately expressed transcripts (200>RPKM>10 in Std. RNA-seq library prepared from AA100 sample) as a function of amount of mRNA used for the library preparation using Smart-seq.

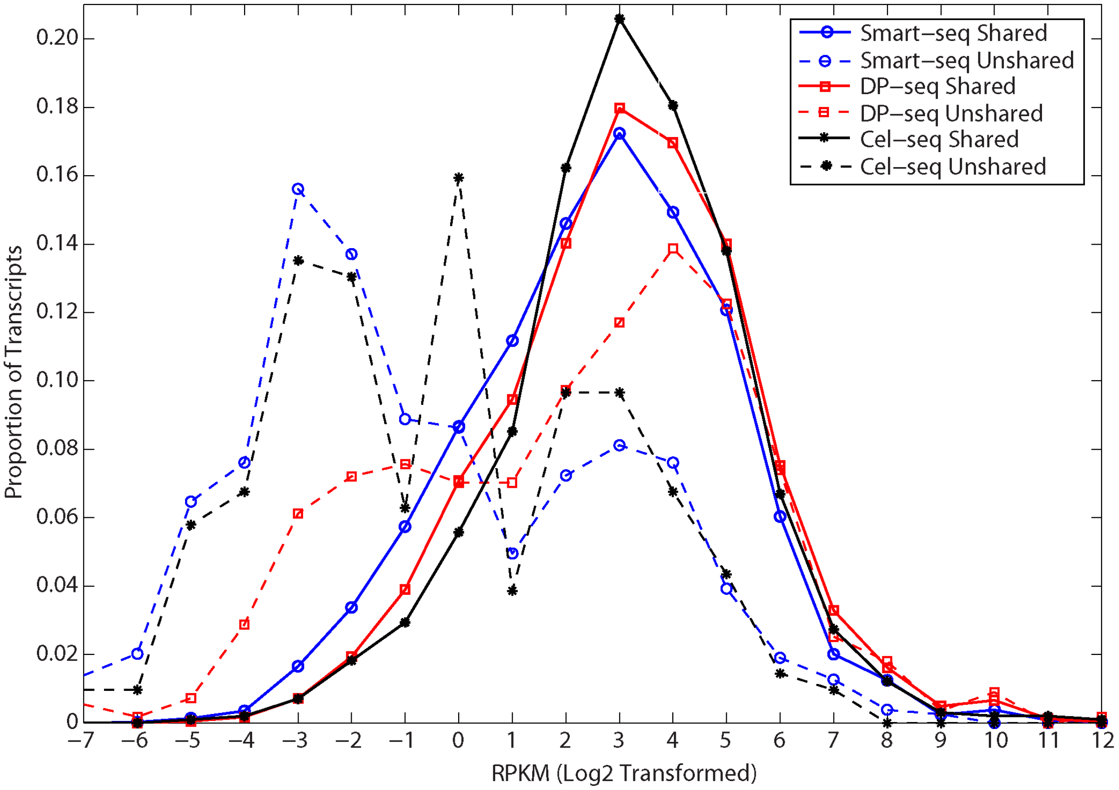


Figure S3.6: Differential gene expression analysis. Differentially expressed genes identified in the amplification-based methods (1 ng sequencing libraries) that were not shared with Std. RNA-seq method, exhibited lower RPKMs. This demonstrates noise in the quantification of the low expressed genes.

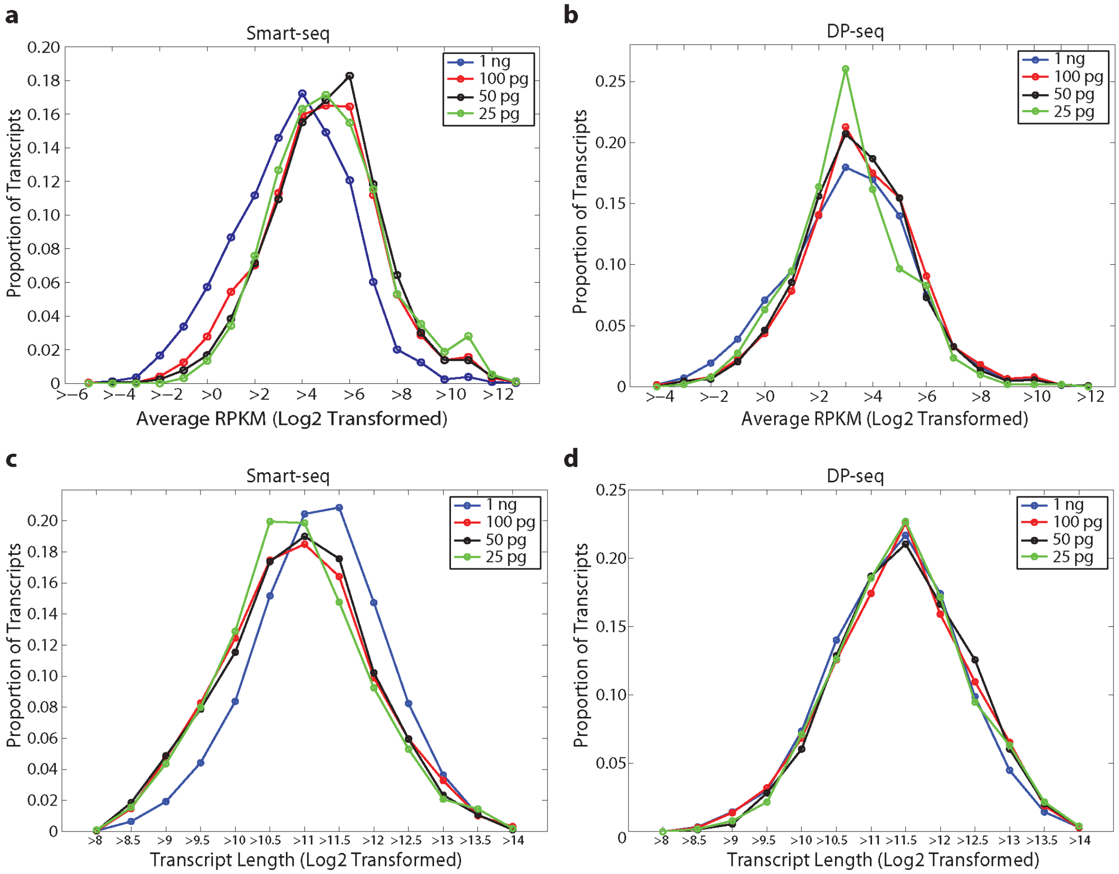


Figure S3.7: Characteristics of differentially expressed genes (DEG) identified in Smart-seq and DP-seq. (A) RPKM (obtained from std. RNA-seq) distribution of DEGs in the Smart-seq libraries. Genes with higher RPKMs were identified as differentially regulated when the amount of mRNA is reduced. (B) RPKM (obtained from std. RNA-seq) distribution of DEGs in the DP-seq libraries. (C) Transcript length distribution of DEGs in the Smart-seq libraries. Genes with shorter length are preferentially identified as differentially regulated. (D) Transcript length distribution of DEGs in the DP-seq libraries.

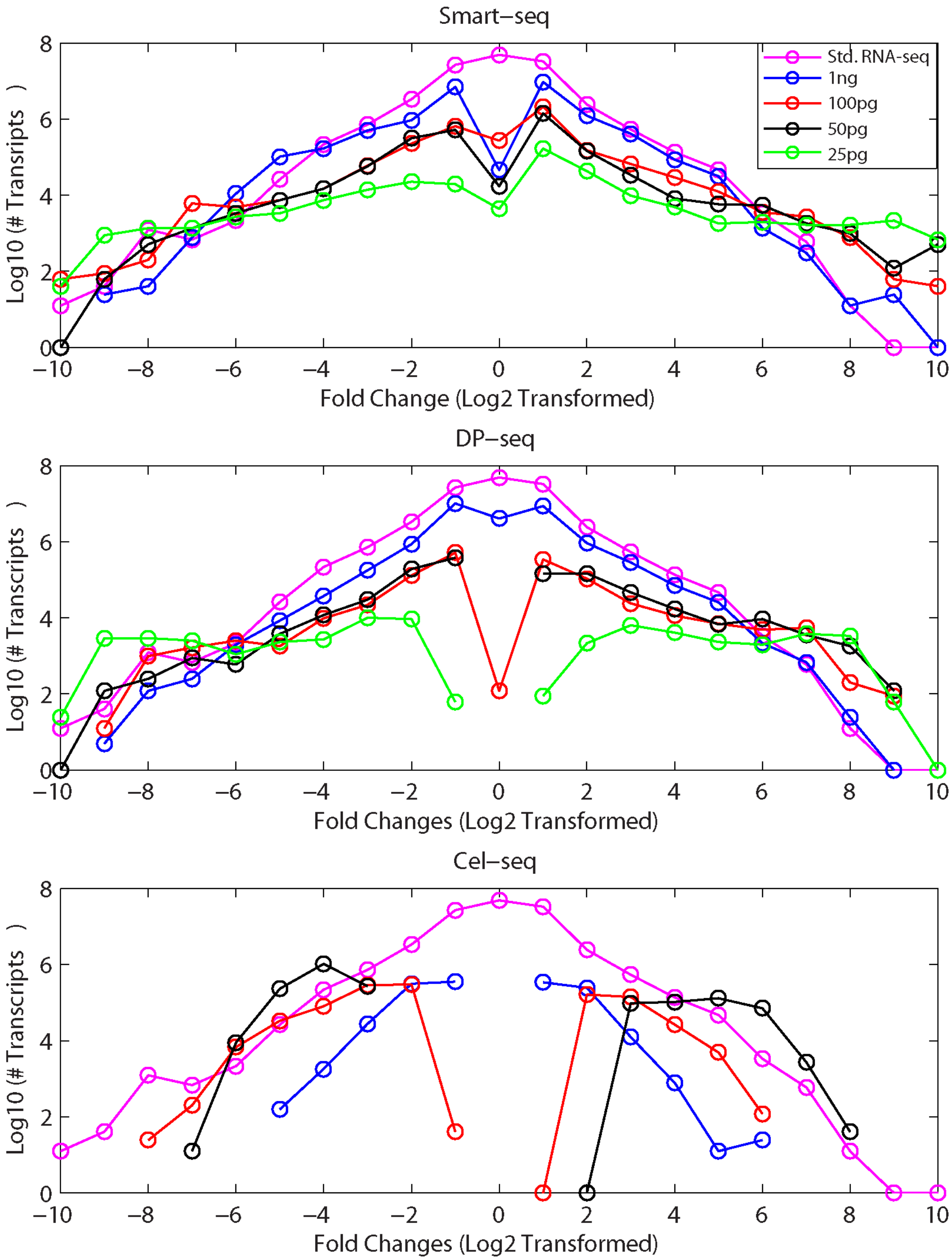


Figure S3.8: Distribution of the fold changes observed for differentially expressed transcripts in the sequencing libraries prepared from varying amounts of mRNA.

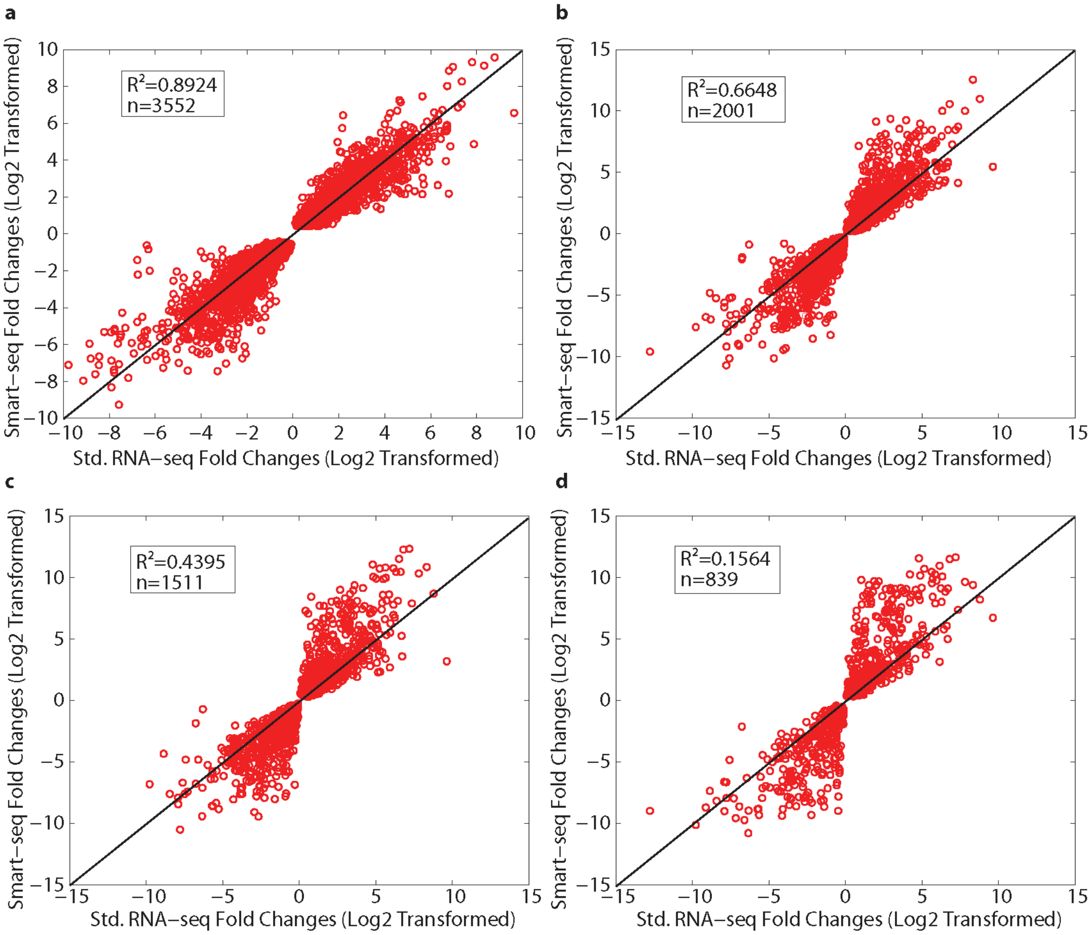


Figure S3.9: Smart-seq fold change comparison with Std. RNA-seq. Comparison of the fold changes of differentially expressed genes identified in the Smart-seq libraries prepared from (A) 1 ng (B) 100 pg (C) 50 pg and (D) 25 pg of mRNA.

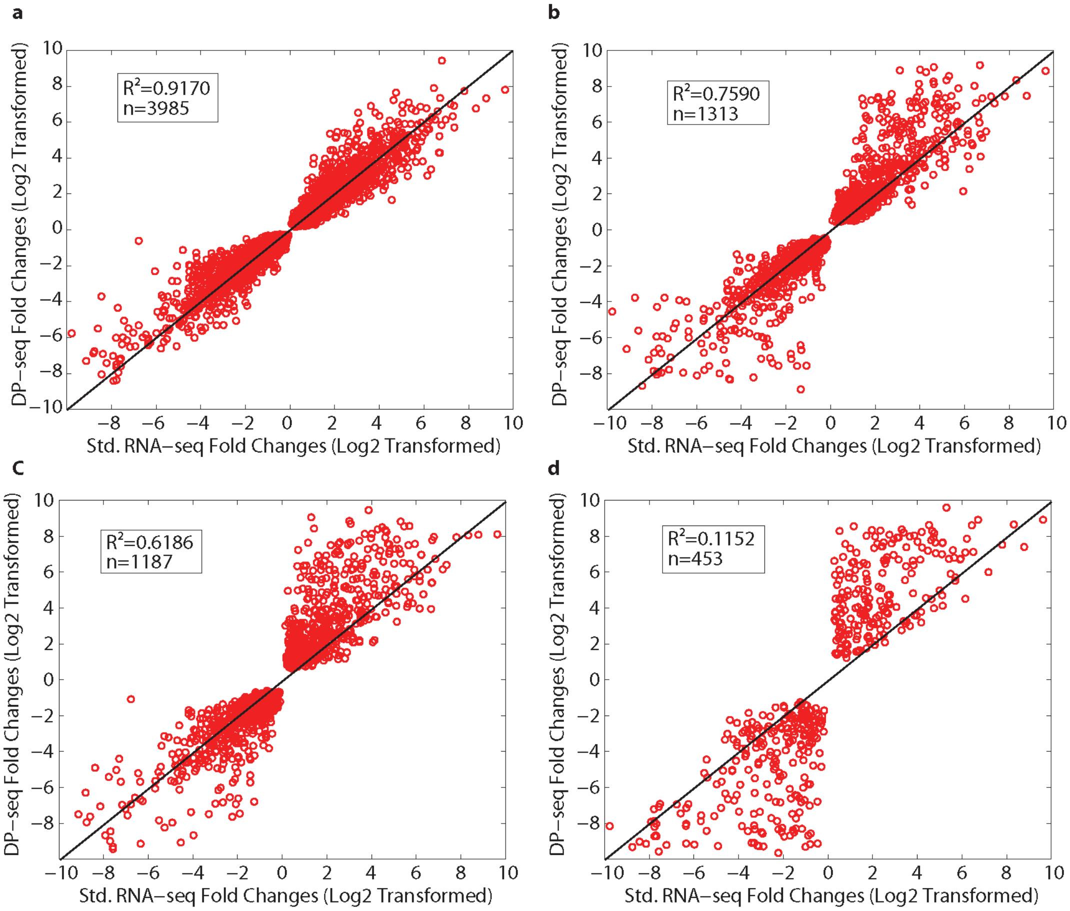


Figure S3.10: DP-seq fold change comparison with Std. RNA-seq. Comparison of the fold changes of differentially expressed genes identified in the DP-seq libraries prepared from (A) 1 ng (B) 100 pg (C) 50 pg and (D) 25 pg of mRNA.

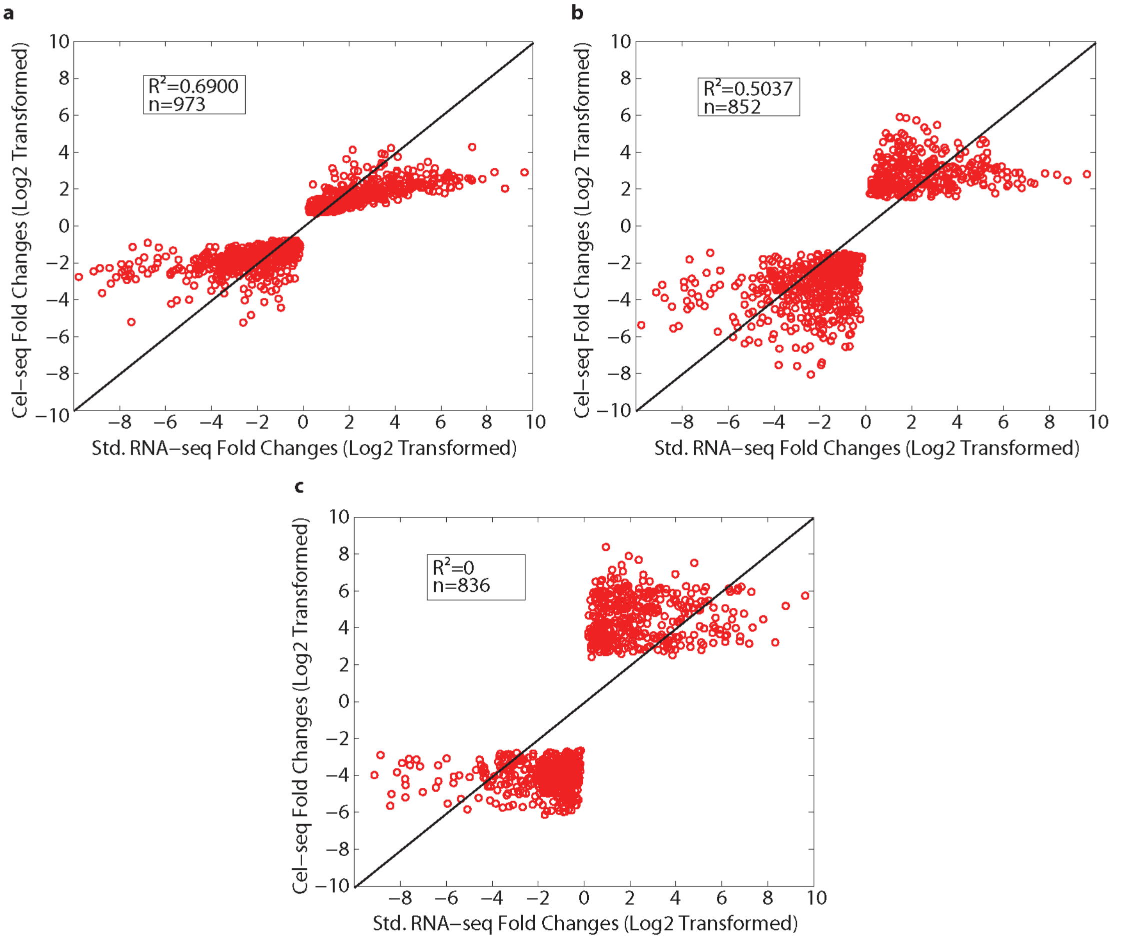


Figure S3.11: CEL-seq fold change comparison with Std. RNA-seq. Comparison of the fold changes of differentially expressed genes identified in the CEL-seq libraries prepared from (A) 1 ng (B) 100 pg (C) 50 pg and (D) 25 pg of mRNA.

3.7 Supplementary Tables

Table S3.1: Mapping Summary. For both samples, SFM and AA100, technical replicate libraries were prepared. Refseq represents the percentage of reads that mapped to NCBI Refseq database. Genomic refers to percentage of reads that were unmapped to the NCBI refseq database but mapped to the genomic locations of mouse. Unmapped refers to percentage of reads that did not map to either of the databases. Transcriptome coverage represents the number of transcripts with at least 1 uniquely mapped read.

Std. RNA-seq					
Sample	**Total Reads**	**Refseq**	**Genomic**	**Unmapped**	**Coverage>=1**
SFM_1	24453925	68.01	28.13	3.60	18870
SFM_2	23697503	67.90	27.89	3.93	18791
AA100_1	21129136	79.74	16.53	3.64	17332
AA100_2	14681557	79.74	16.54	3.64	16750
Smart-seq					
Sample	**Total Reads**	**Refseq**	**Genomic**	**Unmapped**	**Coverage>=1**
SFM_11	24349951	70.92	16.06	12.76	17124
SFM_12	23428763	70.70	16.55	12.49	17107
AA100_11	25118762	84.02	4.85	10.87	15856
AA100_12	23978377	83.67	5.20	10.87	16064
SFM_1001	21367303	69.79	17.69	12.25	14865
SFM_1002	21453069	70.34	17.61	11.79	14781
AA100_1001	26945482	84.40	4.76	10.57	13887
AA100_1002	25212172	83.25	4.58	11.89	13873
SFM_501	22094248	67.06	18.61	14.05	13785
SFM_502	20943831	67.31	18.63	13.79	13451
AA100_501	22298719	86.35	7.71	5.86	13400
AA100_502	24284435	87.25	7.43	5.24	13568
SFM_251	24257173	68.84	17.32	13.57	12924
SFM_252	21923234	68.21	17.28	14.23	12770
AA100_251	22655528	83.21	4.43	12.08	12291
AA100_252	23410644	83.67	4.40	11.66	12059
DP-seq					
Sample	**Total Reads**	**Refseq**	**Genomic**	**Unmapped**	**Coverage>=1**
SFM_11	28474965	67.92	20.65	11.17	15207
SFM_12	18244022	70.52	19.83	9.38	14481
AA100_11	21291454	76.29	14.97	8.48	15169
AA100_12	20993366	76.03	14.97	8.73	15178
SFM_1001	25718262	48.53	15.71	35.49	12660
SFM_1002	14793596	47.11	15.72	36.91	11798

Table S3.1: Mapping Summary. For both samples, SFM and AA100, technical replicate libraries were prepared. Refseq represents the percentage of reads that mapped to NCBI Refseq database. Genomic refers to percentage of reads that were unmapped to the NCBI refseq database but mapped to the genomic locations of mouse. Unmapped refers to percentage of reads that did not map to either of the databases. Transcriptome coverage represents the number of transcripts with at least 1 uniquely mapped read. (Continued)

Sample	Total Reads	Refseq	Genomic	Unmapped	Coverage>=1
AA100_1001	23378520	60.51	13.11	26.11	12744
AA100_1002	18366027	60.17	12.88	26.72	12933
SFM_501	29346385	40.36	14.57	44.83	11871
SFM_502	19512341	41.78	14.62	43.37	11377
AA100_501	26108501	58.72	13.01	28.18	12983
AA100_502	27486701	65.44	14.42	20.04	13302
SFM_251	22923417	24.86	9.51	65.40	9949
SFM_252	24601138	25.63	10.09	64.05	10021
AA100_251	27740980	40.20	9.72	49.85	10935
AA100_252	22478624	45.33	10.97	43.46	10516
CEL-seq					
Sample	**Total Reads**	**Refseq**	**Genomic**	**Unmapped**	**Coverage>=1**
SFM_11	30275877	51.23	27.35	20.83	13653
SFM_12	16195257	44.08	20.45	34.93	13035
AA100_11	13946599	59.96	19.30	20.16	12284
AA100_12	13570701	53.32	18.40	27.76	12497
SFM_1001	16171294	52.21	28.46	18.74	10735
SFM_1002	15621913	49.50	22.07	27.88	11453
AA100_1001	23773752	65.63	16.61	17.17	10893
AA100_1002	18074154	60.18	16.98	22.29	11443
SFM_501	33161547	54.08	28.91	16.44	9904
SFM_502	15037880	50.72	26.81	21.93	10117
AA100_501	13033759	67.04	17.56	14.83	8524
AA100_502	12209593	61.00	18.98	19.49	9787
SFM_251	9975315	44.38	36.77	18.28	7196
SFM_252	8843522	50.50	24.53	24.47	8096
AA100_251	25480845	67.41	17.15	14.87	8056
AA100_252	13152956	61.90	18.75	18.88	8192

Table S3.2: Kegg pathways enriched in AA100/SFM samples. This list represents a subset of pathways that were enriched in the two samples. P-value was determined from background set of mouse genes.

Pathways	**P-value**
Up-regulated in AA100 in comparison to SFM	
mmu03010:Ribosome	5.92E-21
mmu00190:Oxidative phosphorylation	7.34E-18
mmu05016:Huntington's disease	4.54E-14
mmu04530:Tight junction	0.001179122
mmu04114:Oocyte meiosis	0.008939401
mmu04150:mTOR signaling pathway	0.014273056
mmu04350:TGF-beta signaling pathway	0.032027106
Down-regulated in AA100 in comparison to SFM	
mmu04360:Axon guidance	3.47E-14
mmu04310:Wnt signaling pathway	7.27E-13
mmu04510:Focal adhesion	1.37E-08
mmu04722:Neurotrophin signaling pathway	2.32E-07
mmu04110:Cell cycle	3.82E-07
mmu04012:ErbB signaling pathway	4.45E-07
mmu04340:Hedgehog signaling pathway	9.03E-07
mmu04330:Notch signaling pathway	4.25E-05
mmu04910:Insulin signaling pathway	5.63E-05
mmu04660:T cell receptor signaling pathway	7.57E-04

Table S3.3: GO (Biological Process) enrichments for genes differentially regulated in AA100 sample in comparison to SFM. P-value was determined from background set of mouse genes.

Biological Processes	P-value
Up-regulated in AA100 in comparison to SFM	
GO:0006412~translation	1.57E-27
GO:0006091~generation of precursor metabolites and energy	3.37E-22
GO:0042254~ribosome biogenesis	3.14E-18
GO:0007369~gastrulation	3.34E-04
GO:0042074~cell migration involved in gastrulation	4.67E-04
GO:0007005~mitochondrion organization	5.50E-04
GO:0051301~cell division	6.05E-04
GO:0048332~mesoderm morphogenesis	9.94E-04
GO:0001701~in utero embryonic development	0.001228
GO:0001824~blastocyst development	0.002370
GO:0007492~endoderm development	0.003770
GO:0001892~embryonic placenta development	0.011984
Down-regulated in AA100 in comparison to SFM	
GO:0045449~regulation of transcription	4.30E-38
GO:0030182~neuron differentiation	5.41E-21
GO:0051252~regulation of RNA metabolic process	1.53E-19
GO:0007389~pattern specification process	1.71E-18
GO:0048666~neuron development	8.52E-16
GO:0003002~regionalization	2.07E-14
GO:0031175~neuron projection development	2.59E-14
GO:0048598~embryonic morphogenesis	3.27E-14
GO:0032990~cell part morphogenesis	2.96E-13
GO:0048858~cell projection morphogenesis	3.04E-13
GO:0007409~axonogenesis	3.50E-13
GO:0048667~cell morphogenesis involved in neuron differentiation	9.25E-13
GO:0030030~cell projection organization	1.17E-12
GO:0006357~regulation of transcription from RNA polymerase II promoter	1.52E-12
GO:0043009~chordate embryonic development	1.59E-12
GO:0009792~embryonic development ending in birth or egg hatching	1.65E-12
GO:0000902~cell morphogenesis	3.56E-12
GO:0048812~neuron projection morphogenesis	3.57E-12
GO:0000904~cell morphogenesis involved in differentiation	7.05E-12

Table S3.4: List of quantitative RT-PCR primers used in the study.

Gene	Forward Primer	Reverse Primer
Lefty1	CGCTGAATCTGGGCTGAGTCCC	GCCTAGGTTGGACATGTTTGCCCA
Lefty2	TGCAAGTAGCCGACTTCGGAGC	CCTATTCCCAGGCCTCTGGCCA
Gsc	GGGGGTCGAGAAAGCAACGAGG	ACGAGGCTCACGCAGGCAGC
Oct4	TGAAGTGCCCGAAGCCCTCCCTA	GCCCTTCTGGCGCCGGTTACA
T	CTCCGATGTATGAAGGGGCTGCT	GCTATGAGGAGGCTTTGGGCCG
Sox1	TTCCCCAGGACTCCGAGGCG	GTTCAGTCTAAGAGGCCAGTCTGGT
Pax6	ACCTCCTCATACTCGTG	ACTGATACCGTGCCTT
Dmrt3	AACCGGCCACCCCTGGAAGT	GTCGCCCCCGCAACCTTTCA
Sox2	GCACATGAAGGAGCACCCGGA	GGTTCACGCCCGCACCCAG
Fgf8	GCGAAGCTCATTGTGGAGAC	CACGATCTCTGTGAATACGCA
Nodal	ACCAACCATGCCTACATCCAGAG	CCCTGCCATTGTCCACATAAAGC
Epha1	TACGCCTGCCCAGCCTGAGT	GGTGTCCAGCCCAGCCGAAC
Rab25	TCAGCCAGGCCCGAGAGGTC	GATGGCACTGGTCCGGGTGC
Pgr	CGCCATCTACCAGCCGCTCG	ACTGTGGGCTCTGGCTGGCT
Foxa3	TTTGGGGGCTACGGGGCTGA	TGCAGCCCACGCCCATCATG
Ell2	TGCAGGCCTCCTACCACCCC	TCCCCAGGCCTTCTGGAGTGC
Lbh	ACGTTGGGGCAAGAGCGTGG	GAGACGGGGGAGGGGGTGAC
Etv4	GAAGGTGGCTGGCGAACGCT	GCGGGGCCAGTGAGTTCTGG
Klf9	CCGCGTACTCGGCTGATGCC	CACACGTGGCGGTCGCAAGT
Wnt3a	ACCAAGACCTAACAAACCC	CATGGACATCACGGACC
Per2	GGTGGCCTCTGCAAGCCAGG	CCTCCGTGCTCAGTGGCTGC
Hes1	CCCTGCAAGTTGGGCAGCCA	CGAAGGCCCCGTTGGGGATG
Foxc2	AGGGACTTTGCTTCTTTTTCCGGGC	CCCGCAGCGTCAGCGAGCTA
Prdm6	CCGGCCTTTCAAGTGCGGCT	GGCATGCGCTGGTGTCGACT
Cxcr4	TACCCCGATAGCCTGT	GCACGATGCTCTCGAA
Asb4	TCACCTCCGTGCGTCCTGCT	TTCGGGCAAGAGTGGCAAGCC
Lhx1	ACTAGGGACCGAGGGACGCG	CAGTTTGGCGCGGATTGCCG
Sox17	GAGCCAAAGCGGAGTCTC	TGCCAAGGTCAACGCCTTC
Creb3l1	ACAGGACGGACACCCTGGCA	GGTCAGCCCAGGGGAGCAGT
Basp1	AGGGGGCGGGGAGAATCCAAA	GGAGCCTAGGGGACAGCGGTT
β-Actin	GCTGTATTCCCCTCCATCGTG	CACGGTTGGCCTTAGGGTTCAG
Lhx2	TGGGCTCAGCCGGGGCTAAT	ACAGCTAAGCGCGGCGTTGT
Gapdh	AATGGATACGGCTACAGC	GTGCAGCGAACTTTATTG
Nanog	AGGACAGGTTTCAGAAGCAGA	CCATTGCTAGTCTTCAACCACTG
Flt1	CTCAGACAAGTCAAACCTGGAG	GGGAACTTCATCTGGGTCCATAA
Foxh1	ACTTGCCCATCTATACGCCC	GATTCAGTGCCTACGCTCCA
Fst	CTGAGAAAGGCCACCTGCTT	GCCGCCACACTGGATATCTT
Id1	TGGGAAAGACACTACCGCAG	CTCTGGAGGCTGAAAGGTGG
En1	CTACCACCACGGTTCAGGAC	ATAGCGATCGTCTCTGCGTG
Fzd3	GTACCCGTTGCACTCTTGGA	CACTGAGGGGCATCACTGAG
Sox3	CCCTGAGCACCACTCCGAC	CACGGGGTTCTTGAGTTCAGT

3.8 Acknowledgements

This work was supported by National Institutes of Health (NIH), National Institute of Diabetes and Digestive and Kidney Diseases Grant P01-DK074868 (S.S.), National Heart, Lung and Blood Institute Grant 5 R33 HL087375-02 (S.S. and M.M.), and NIH/National Institute of General Medical Sciences Grant GM078005-05 (S.S.). V.B. was recipient of a California Institute for Regenerative Medicine (CIRM) Graduate Fellowship (T1-00003 and TG2-01154, Interdisciplinary Stem Cell Training Program at UCSD). The authors would like to thank John Shimashita who helped them with sequencing libraries preparation, Lana Schaffer for providing bioinformatics support, Andrew Richards and Gaurav Agrawal for many useful discussions.

Chapter 3 is in full material submitted for publication from **Bhargava, V.**, Head, S., Ordoukhanian, P., Mercola, M., Subramaniam, S. (2013) *Technical Variations in Low Input RNA-seq Methodologies.* The dissertation author was the primary author of this paper responsible for the research.

Chapter 4

Functional Characterization of dedifferentiated Neurons and Astrocytes using DP-seq

4.1 Abstract

Dedifferentiation of mature neurons and astrocytes was recently demonstrated as a mechanism for glioma formation in mice. Expression analysis of the known markers (of the differentiated cell-types) showed diminished expression while the expression of the undifferentiated state markers were significantly up-regulated in the dedifferentiated neurons and astrocytes. Here, we performed whole transcriptome analysis of these cells along with mouse pluripotent embryonic stem cells (mESC), neural stem cells (NSC), neurons and astrocytes to characterize the undifferentiated state of these cells. Our analysis revealed that dedifferentiated cell-types shared traits with neurons and NSCs at global transcriptome level suggesting that they have not completely regressed to an undifferentiated state of NSCs. Functional analysis of the transcriptomics data revealed involvement of the Wnt signaling, cell cycle and the focal adhesion pathways in defining the state of the dedifferentiated cell-types. Our analysis further revealed conservation of a gene interaction network in both dedifferentiated cell-types. This network exhibited modular architecture; connecting components of the cell cycle pathway to Wnt signaling and the focal adhesion pathway. Genetic perturbations of the interacting partners and/or the

abolishment of the interactions will elucidate the regulatory mechanism of this network in maintaining the dedifferentiated state of the neurons and the astrocytes.

4.2 Introduction

Dedifferentiation of terminally differentiated cells to a less differentiated state within its own lineage has generated a lot of interest in the field of regenerative medicine. Expansion of the pool of rapidly dividing progenitor stem cells and their subsequent re-differentiation could replace the tissue lost as a result of injury. In fact, some vertebrate species have demonstrated remarkable regenerative capacity resulting in a complete limb replacement[84] to full regeneration of their heart[85, 86]. In contrast, mammals have exhibited a limited capacity to regenerate and maintain their tissues and organs. However, recently some evidence of dedifferentiation in mammals has emerged. Schwann cells have been reported to dedifferentiate and proliferate upon injury to a nerve[87, 88]. It has been demonstrated that upon loss of contact with the axon that they are myelinating, schwann cells begin expressing markers of precursor stem cells, proliferate and differentiate to give rise to mature myelinating or non-myelinating schwann cells. Similarly, upon brain injury, astrocytes have shown the ability to re-enter the cell cycle and undergo long – term self – renewal and multipotency by forming neurospheres[89-91]. Skeletal muscle cells in an injured mouse model have also demonstrated their dedifferentiation capacity resulting in cell proliferation and myogenesis[92]. Additionally, dedifferentiation has been achieved by experimental induction wherein extract isolated from regenerative newt limbs reduced the expression of myoblast genes MyoD and myogenin in the mouse myotubes and subsequently led to their proliferation[93]. Introduction of chemical compounds has also resulted in dedifferentiation of lineage-committed myoblasts to multipotent mesenchymal progenitor cells[94]. Finally, in the extreme form of dedifferentiation, researchers have shown that overexpression of specific transcription factors in any differentiated cell type could result in complete reprogramming of the cells to a pluripotent state[95-97].

Regression of cells to a less differentiated state has also been demonstrated in progression of cancer. Recent studies have suggested that dedifferentiation of non-stem cells in the cancer mass results in the generation of cancer stem cells that greatly enhances the pool of rapidly proliferating cells[32, 98, 99]. Epithelial to mesenchymal transition , which correlates with β-catenin expression, has been associated with dedifferentiation of invading cells[100]. Moreover, TGFβ has been demonstrated to promote dedifferentiation during the squamous-cell carcinoma resulting in most aggressive form of skin cancer[101]. In another study, human mammary epithelial cells have been shown to spontaneously dedifferentiate to cancer stem cells like state in-vitro[102]. Genetic models of tumor initiation by dedifferentiating non-stem cell population in intestinal epithelial cells have implicated the role of NF-κB in enhancing Wnt-signaling leading to the dedifferentiation and proliferation of epithelial cells into tumor-initiating cells[98]. A recent study showed transduction by oncogenic lentiviral vectors containing short hairpin RNA (shRNA) targeting NF1 and p53 genes of mature neurons, astrocytes and neural stem cells gave rise to malignant tumors[32]. All the tumors hence generated, showed high expression of progenitor markers and low expression of differentiation markers. Even though a number of signaling pathways have been identified to play a role in the dedifferentiation process, the exact molecular mechanism remains elusive[103].

Here, we investigated dedifferentiation achieved by mature neurons and astrocytes in mouse hippocampus upon transduction with lentiviral vector containing shRNA targeting NF1 and p53[32]. Both of these genes have shown high % of mutations, 18% and 35% respectively, in glioblastoma multiforme (GBM). Previous studies have indicated that the loss of NF1 results in high cell proliferation while loss of p53 results in genomic instability, some of the hallmarks of cancer progression. Stereotaxic injection of the lentivirus in different sites of the mouse brain resulted in gliomas formation. While, glial cells[104], oligodendrocyte precursor cells[105] and NSCs[106] have been suggested as possible candidates of cell of origin, this study showed that cortical neurons also exhibited dedifferentiation and generated malignant gliomas. The gliomas were heterogeneous and matched the histo-pathological traits of the gliomas obtained from other cell types. To further validate their hypothesis, the authors isolated neurons that were Map2-positive, GFAP (Glial Fibrillary Acidic Protein) and doublecortin negative, and Ki67 negative and transduced them with the lentiviral vector in-vitro. The transduced neurons where transplanted into NOD-SCID mice and the resulting tumors were analyzed. The tumors showed similar characteristics to in-vivo transduced tumors. These tumors further showed high levels of Sox2 and Nestin (NSC markers) expression. Similar observations were made for cortical astrocytes that showed no expression of NSC markers but exhibited elevated expression of progenitor markers upon transduction with the same lentiviral vector.

We performed whole genome transcriptome analysis of the dedifferentiated neurons (Tr. Neuron) and astrocytes (Tr. Astrocyte) along with the enriched populations of mESCs, NSCs, neurons and astrocytes to characterize the point of regression of these dedifferentiated cells on the differentiation axis. Our transcriptome data revealed that the dedifferentiated cells have significantly lower expression of known markers of their parental cell-types. They also exhibited increased expression of progenitor NSC markers. However, at whole transcriptome level, dedifferentiated cells retained high expression of some of the neuronal markers, suggesting that these cells have not completely dedifferentiated to the NSC like state. Enrichment analysis of the differentially regulated genes in the dedifferentiated cell-types revealed up-regulation of the cell cycle, Wnt signaling and the focal adhesion pathways in comparison to the differentiated cell-types. Furthermore, we identified a gene interaction network that was conserved in the dedifferentiated neurons and astrocytes thus revealing significant interactions between the genes responsible for the phenotype observed in the dedifferentiated cell-types.

4.3 Results

4.3.1 Experimental design and transcriptome analysis

To understand the molecular mechanism involved in the dedifferentiation of mature neurons and astrocytes upon transduction with the lentiviral vector, we made use of an in-vitro culture system. Cortical neurons and astrocytes were derived from postnatal day 11 mice. They were maintained in in-vitro conditions in the presence of serum to maintain their identity. These cells were later transduced with the lentiviral vector with the transduction efficiency of >90%. The transduced neurons and astrocytes were later transferred in a stem cell media that was devoid of serum and was supplemented with FGF-2. Within one week, these cells became proliferative and aggregated to form free-floating neurospheres. These cells were later harvested and mRNA was collected for sequencing library generation using DP-seq[31]. To access the regression of these cells to an undifferentiated state along the differentiation axis, enriched populations of mESCs and NSCs were also grown in in-vitro system and mRNA obtained from these cells were subjected to library preparation (**Figure 4.1**).

Sequencing libraries prepared from these samples exhibited high transcriptome coverage with a vast majority of the reads mapping to the NCBI Refseq database (**Supplementary Table 4.1**). Analysis of the biological replicates showed little biological variations, except for the neurons (**Supplementary Fig. S4.1**). To validate our sequencing libraries, we investigated the expression of known markers of different cell-types. Mouse ESC markers[107], which were significantly enriched in mESCs libraries, showed low expression in other cell types. Similarly, expression of NSC markers, Sox2, Dll3 and Meis1, were significantly up-regulated in the NSC populations. Known markers of neurons and astrocytes[108] also showed high expression in their respective sequencing libraries thus validating the sequencing libraries. In case of dedifferentiated neurons and astrocytes, majority of the mESCs markers had low expression. Additionally, these cells exhibited diminished expression of their parental cell-types while the expression of known NSC markers, Nestin and Sox2 was significantly high in these cells. This demonstrated that the dedifferentiated cells partially abandoned the expression pattern of their parental cell-types and acquired an undifferentiated progenitor stem cell state (**Figure 4.2**).

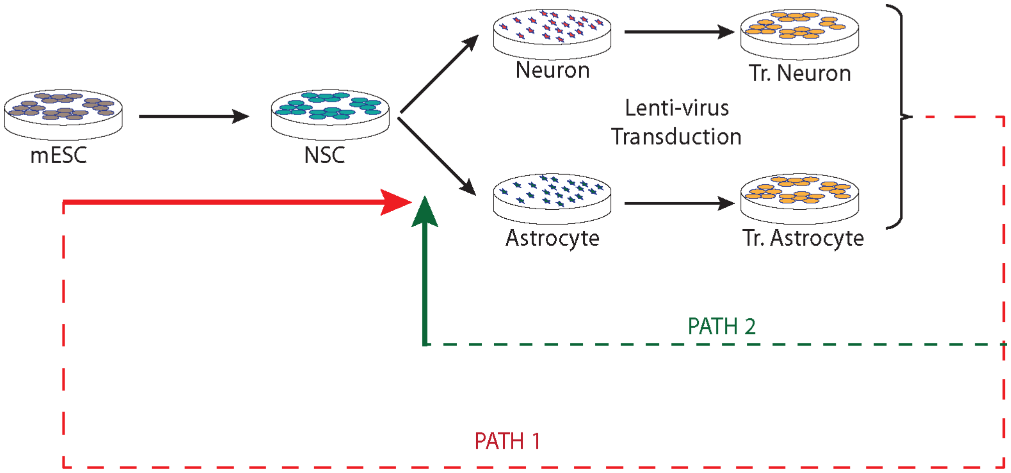


Figure 4.1: Schema of experimental design. mRNA was collected from enriched populations of mESCs, NSCs, primary culture of neurons, primary culture of astrocytes and dedifferentiated neurons and astrocytes. Dedifferentiation of neurons and astrocytes was achieved by transducing the primary cultures of neuron and astrocytes by lentiviral vector comprising of shRNA targeting NF1 and p53. The transduced neurons and astrocytes were maintained in stem cell media devoid of serum and supplemented with FGF-2 for three weeks. The mRNA was also derived from transduced neurons that were maintained in DMEM media.

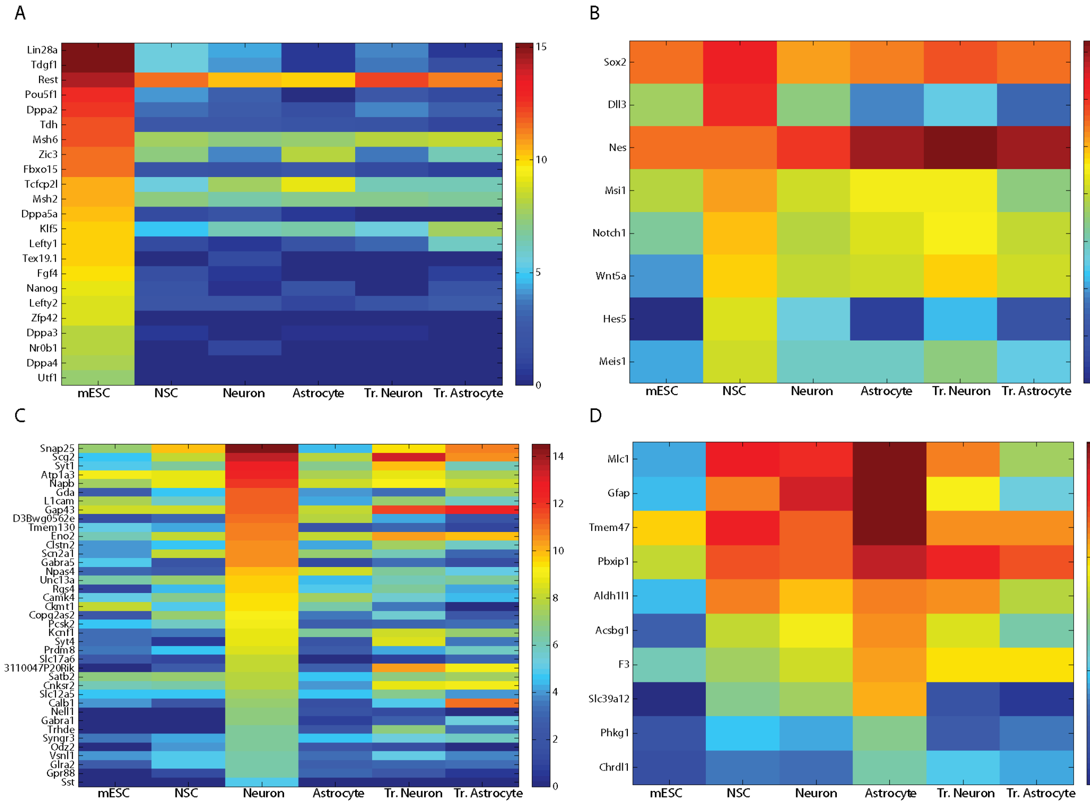


Figure 4.2: Expression analysis of the known markers. Heatmap displaying expression of the known markers of different cell-types. (A) mESC markers (B) NSC markers (C) Neuron markers (D) Astrocyte markers.

4.3.2 Differential gene expression analysis

The biological cell-types considered in this study were highly divergent with many housekeeping genes exhibiting differential expression. Therefore, we normalized the sequencing libraries using quantile normalization. Differential expression analysis identified 463 genes up-regulated in Tr. Neurons in comparison to mature neurons. Tr. Astrocyte samples showed higher differential expression (1966 genes up-regulated in comparison to astrocytes) owning to high biological variations in the neuron samples. Majority of the 463 genes up-regulated in Tr. Neurons were also up-regulated in Tr. Astrocytes (**Figure 4.3**) highlighting that the genetic alterations introduced by the lentiviral vector affected same set of genes in the differentiated cell-types. Similar observations were made for the down-regulated genes in the dedifferentiated neurons and astrocytes (**Figure 4.4**).

We next performed pathway enrichment analysis on the differentially regulated genes identified in the dedifferentiated cell-types. In both cell-types, canonical Wnt signaling, cell cycle and the focal adhesion pathways were significantly up-regulated. Aberrant regulation of Wnt signaling has been implicated in progression of many cancers[109] and many of its components have been associated with maintenance of cancer stem cells[110]. Expectedly, cell cycle related genes were up-regulated in dedifferentiated cell-types as these cells were highly proliferative as opposed to their parental cell-types. The dedifferentiated cell-types underwent drastic transformation where they left their flattened morphology and acquired an aggregated free-floating neurosphere like structure.

This transformation resulted in differential expression of many focal adhesion genes. Interestingly, focal adhesion genes exhibited bifurcated expression pattern where a unique set of genes were enriched in the dedifferentiated cell-types while some genes lost their expression (**Figure 4.5**). Both, neurons and astrocytes, displayed conserved regulation of many of the focal adhesion associated genes.

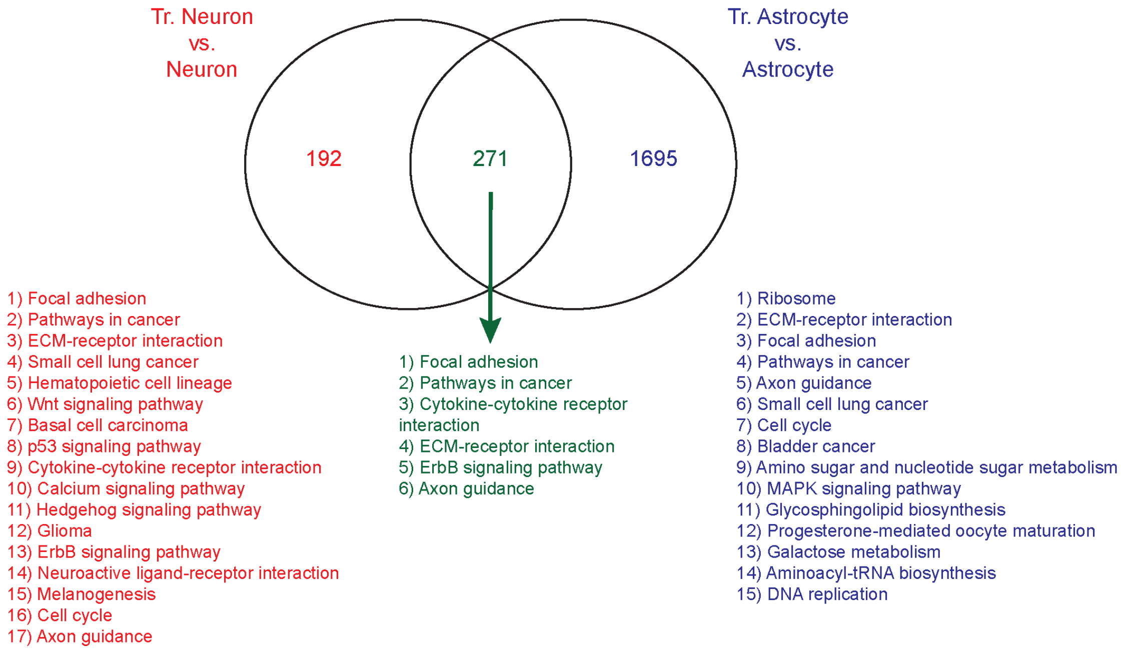
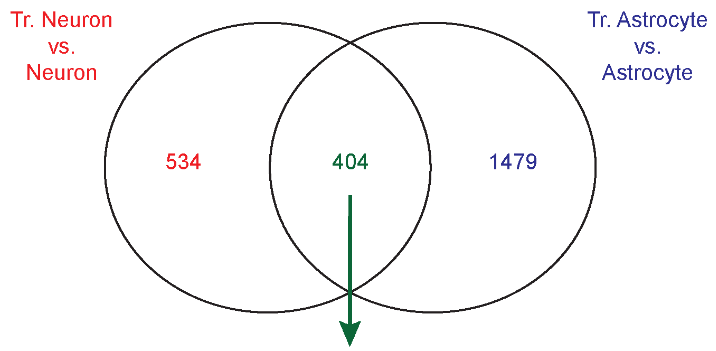

Tr. Neuron
vs.
Neuron
Tr. Astrocyte
vs.
Astrocyte
192
271
1695
1) Focal adhesion
2) Pathways in cancer
3) ECM-receptor interaction
4) Small cell lung cancer
5) Hematopoietic cell lineage
6) Wnt signaling pathway
7) Basal cell carcinoma
8) p53 signaling pathway
9) Cytokine-cytokine receptor interaction
10) Calcium signaling pathway
11) Hedgehog signaling pathway
12) Glioma
13) ErbB signaling pathway
14) Neuroactive ligand-receptor interaction
15) Melanogenesis
16) Cell cycle
17) Axon guidance
1) Focal adhesion
2) Pathways in cancer
3) Cytokine-cytokine receptor interaction
4) ECM-receptor interaction
5) ErbB signaling pathway
6) Axon guidance
1) Ribosome
2) ECM-receptor interaction
3) Focal adhesion
4) Pathways in cancer
5) Axon guidance
6) Small cell lung cancer
7) Cell cycle
8) Bladder cancer
9) Amino sugar and nucleotide sugar metabolism
10) MAPK signaling pathway
11) Glycosphingolipid biosynthesis
12) Progesterone-mediated oocyte maturation
13) Galactose metabolism
14) Aminoacyl-tRNA biosynthesis
15) DNA replication

Figure 4.3: Differentially expressed genes (Up-regulated) identified in the dedifferentiated neurons and astrocytes (Tr. Neurons and Tr. Astrocytes) in comparison to their parental cell-type. Majority of the up-regulated genes identified in the Tr. Neurons were also differentially regulated in Tr. Astrocytes. Astrocytes samples displayed more differential regulation of the transcripts owing to less biological variability in their biological replicates. Enriched KEGG signaling pathways represented by differentially expressed genes are also depicted.



1) ECM-receptor interaction
2) Axon guidance
3) Neuroactive ligand-receptor interaction
4) Focal adhesion
5) Cell adhesion molecules (CAMs)
6) Gap junction
7) Leukocyte transendothelial migration
8) Vascular smooth muscle contraction
9) Tight junction
10) Complement and coagulation cascades

1) Complement and coagulation cascades
2) Systemic lupus erythematosus

1) Lysosome
2) Valine, leucine and isoleucine degradation
3) Fatty acid elongation in mitochondria
4) Biosynthesis of unsaturated fatty acids
5) Propanoate metabolism
6) ABC transporters
7) Other glycan degradation
8) Primary bile acid biosynthesis
9) Glycosaminoglycan degradation
10) Glycerophospholipid metabolism
11) Fatty acid metabolism
12) Leukocyte transendothelial migration
13) PPAR signaling pathway
14) Complement and coagulation cascades
15) Regulation of actin cytoskeleton

Figure 4.4: Differentially expressed genes (Down-regulated) identified in dedifferentiated neurons and astrocytes (Tr. Neurons and Tr. Astrocytes) in comparison to their parental cell-type. Enriched KEGG signaling pathways represented by differentially expressed genes are also depicted.

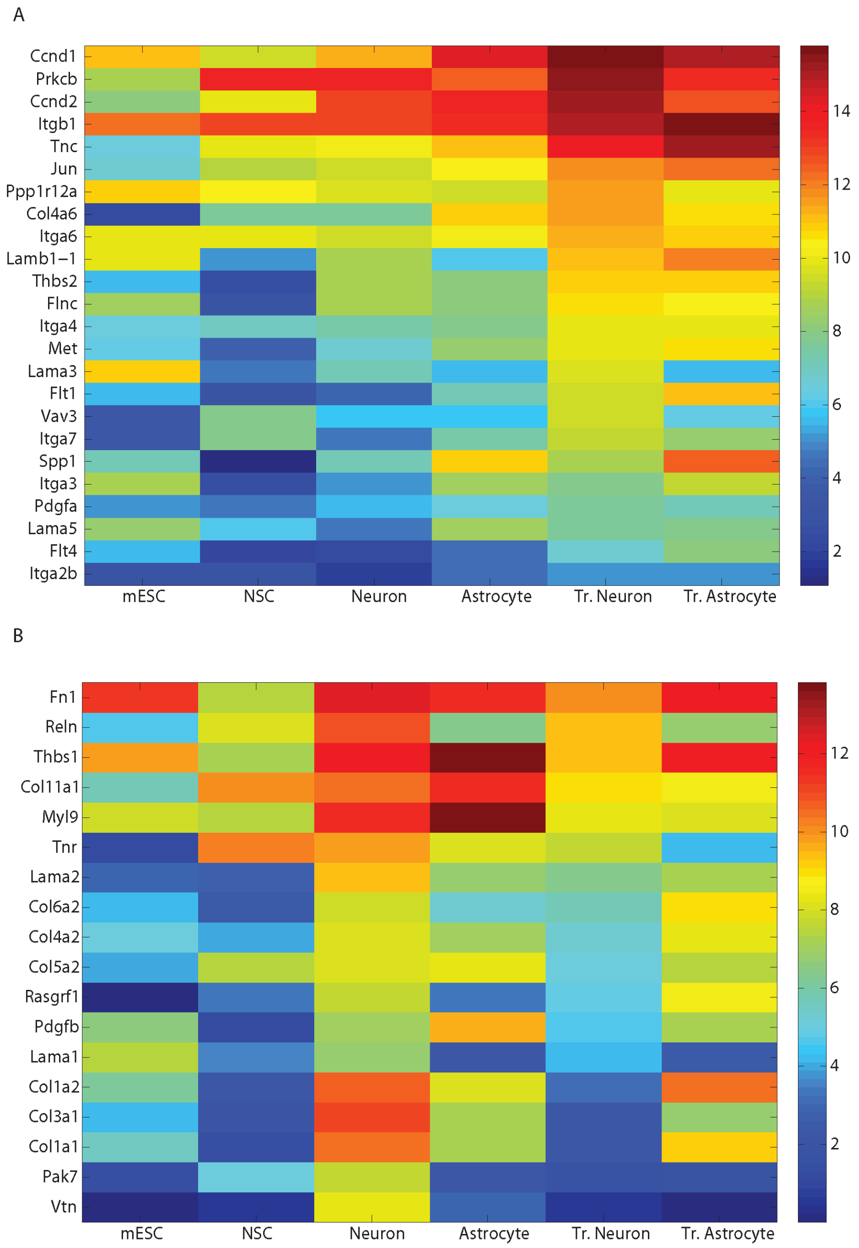


Figure 4.5: Heatmap displaying expression of focal adhesion related genes differentially regulated in Tr. Neurons in comparison to neurons. (A) Up-regulated focal adhesion related genes (B) Down-regulated focal adhesion related genes.

Pathways down-regulated in dedifferentiated cell-types were necessary for maintenance of terminally differentiated cell-types (neurons and astrocytes). This further highlights that the dedifferentiated cell-types have distanced themselves from the differentiated state and acquired a progenitor stem cell like state.

4.3.3 Gene set enrichment analysis

We performed single sample gene set enrichment analysis (ssGSEA)[32, 111] to access the path adopted by mature neurons and astrocytes to dedifferentiate to a less differentiated state. For this analysis, we first compiled a list of known markers of the enriched populations, viz., mESC, NSC, neurons and astrocytes. This list was short and the enrichment analysis was prone to high noise. As such, we constructed gene-list specific to each population by using our transcriptomics datasets. We identified genes that were significantly up-regulated in one population as opposed to all other enriched populations and designated them as "putative" markers of that population.

Expectedly, the ssGSEA analysis performed on the enriched populations showed strong enrichment scores for their "putative" markers. The enrichment was also significant for the known markers of the enriched populations. In case of dedifferentiated neurons, positive enrichment was observed for known neuronal markers. The "putative" markers of neurons also showed positive enrichment although the significance was poor. Surprisingly, similar observations were made for dedifferentiated astrocytes where positive enrichment was observed for known neuronal markers. The known astrocyte markers also displayed positive enrichment in the dedifferentiated astrocytes, however, the enrichment was not statistically significant (**Table 4.1**). A number of known neuronal markers showed high expression in both dedifferentiated neurons and astrocytes and the GO term enrichment analysis of those genes showed biological processes associated with neuronal function (**Table 4.2**). On the other hand, the known neuronal markers that showed low expression in dedifferentiated neurons and astrocytes were mostly associated with ion transport and ligand interaction (**Table 4.3**). The enrichment scores for the mESC markers in the dedifferentiated cell-types were negative with the majority of the mESC-associated genes exhibiting low expression in these cells. Additionally, dedifferentiated cell-types showed low enrichment score for “putative” NSC markers. This suggests that at whole transcriptome level, dedifferentiated cell-types have not completed regressed to NSC like state and they still possess similarities in gene expression to their parental cell-types.

Table 4.1: Single sample gene set enrichment analysis. Dedifferentiated neurons and astrocytes showed similar expression of neuronal markers as opposed to NSC markers.

Tr. Neurons		
Gene List	**# of genes**	**PValue**
Focal Adhesion in Tr. Neurons	22	0.005
Known Neuron Markers	64	0.034
Neuron Specific	373	0.277
Known Astrocyte Markers	78	0.402
NSC Specific	232	0.498
Tr. Astrocytes		
Focal Adhesion in Tr. Neurons	22	0.021
Known Neuron Markers	64	0.081
Known Astrocyte Markers	78	0.224
Neuron Specific	373	0.275
Astrocyte Specific	416	0.502

Table 4.2: GO Enrichment (Biological Process) of neuron markers that exhibited high expression in the Tr. Neurons and Tr. Astrocytes.

Tr. Neurons (16 genes)	
Term (Biological Process)	**PValue**
GO:0031175~neuron projection development	0.0006
GO:0048666~neuron development	0.0014
GO:0030030~cell projection organization	0.0018
GO:0031133~regulation of axon diameter	0.0024
GO:0032536~regulation of cell projection size	0.0024
GO:0045664~regulation of neuron differentiation	0.0029
GO:0045110~intermediate filament bundle assembly	0.0032
GO:0030182~neuron differentiation	0.0034
GO:0050767~regulation of neurogenesis	0.0048
Tr. Astrocytes (11 genes)	
GO:0031133~regulation of axon diameter	0.0013
GO:0032536~regulation of cell projection size	0.0013
GO:0045110~intermediate filament bundle assembly	0.0017
GO:0060052~neurofilament cytoskeleton organization	0.0039
GO:0045109~intermediate filament organization	0.0057
GO:0045104~intermediate filament cytoskeleton organization	0.0096
GO:0031099~regeneration	0.0101
GO:0045103~intermediate filament-based process	0.0109
GO:0050770~regulation of axonogenesis	0.0171
GO:0010975~regulation of neuron projection development	0.0205

We performed similar analysis on an *in-vivo* cancer tissue obtained from the stereotaxic injection of the lentiviral vector in the hippocampus of the mice. Positive enrichments for both neuronal markers and focal adhesion molecules up-regulated in the dedifferentiated neurons, were observed. This implies that even the cancer formed by dedifferentiation of neurons and astrocytes, share neuronal traits and exhibit similar expression of focal adhesion molecules as observed in the dedifferentiated cell-types.

Table 4.3: GO Enrichment (Biological Process) of neuron markers that exhibited low expression in the Tr. Neurons and Tr. Astrocytes.

Tr. Neurons	
Term	**PValue**
GO:0006836~neurotransmitter transport	0.00005
GO:0006811~ion transport	0.00894
GO:0007268~synaptic transmission	0.01063
GO:0006821~chloride transport	0.01072
GO:0001505~regulation of neurotransmitter levels	0.01106
GO:0032940~secretion by cell	0.01197
GO:0046903~secretion	0.01899
GO:0015698~inorganic anion transport	0.01925
GO:0019226~transmission of nerve impulse	0.02009
GO:0007267~cell-cell signaling	0.03810
GO:0006814~sodium ion transport	0.04100
GO:0006820~anion transport	0.04401
GO:0007214~gamma-aminobutyric acid signaling pathway	0.04539
Tr. Astrocytes	
GO:0006836~neurotransmitter transport	0.000003
GO:0007268~synaptic transmission	0.00014
GO:0019226~transmission of nerve impulse	0.00044
GO:0006821~chloride transport	0.00070
GO:0001505~regulation of neurotransmitter levels	0.00073
GO:0006811~ion transport	0.00079
GO:0007267~cell-cell signaling	0.00135
GO:0015698~inorganic anion transport	0.00171
GO:0032940~secretion by cell	0.00192
GO:0046903~secretion	0.00358
GO:0006928~cell motion	0.00378
GO:0007269~neurotransmitter secretion	0.00475
GO:0006814~sodium ion transport	0.00545
GO:0006820~anion transport	0.00608
GO:0030001~metal ion transport	0.00824
GO:0006813~potassium ion transport	0.01073
GO:0015672~monovalent inorganic cation transport	0.01079
GO:0006812~cation transport	0.01525

4.3.4 Identification of functional network

We next sought to identify functional connectivity between genes that were differentially regulated in the dedifferentiated cell-types. We compiled a database of known as well as predicted direct and functional gene and protein interactions from three different sources including TRANFAC[112], STRINGS (Search Tool for the Retrieval of Interacting Genes/Proteins 8.3; http://string-db.org/) and HPRD (Human Protein Reference Database, http://www.hprd.org). The resulting network consisted of more than 8000 nodes/genes that were connected by > 40000 edges/interactions. We projected the up-regulated genes in dedifferentiated neurons (in comparison to neurons) on this network and identified a functional connectivity between 38 nodes that were connected by more than 53 edges (**Figure 4.6**). The network further demonstrated sub-networks representing the genes associated with the three signaling pathways, viz., Wnt signaling, cell cycle, and the focal adhesion pathway. These pathways were also identified as significantly enriched in dedifferentiated neurons.

Since, neurons and astrocytes were infected with same lentiviral vector and the dedifferentiated cell-types were phenotypically similar, we postulated that the functional network of the up-regulated genes should also be conserved between the dedifferentiated neurons and astrocytes. Indeed, the core functional connectivity was maintained in these cells (**Figure 4.7**). The network comprised of 17 nodes and 20 edges and genes associated with the Wnt signaling, cell cycle and focal adhesion pathways were represented.

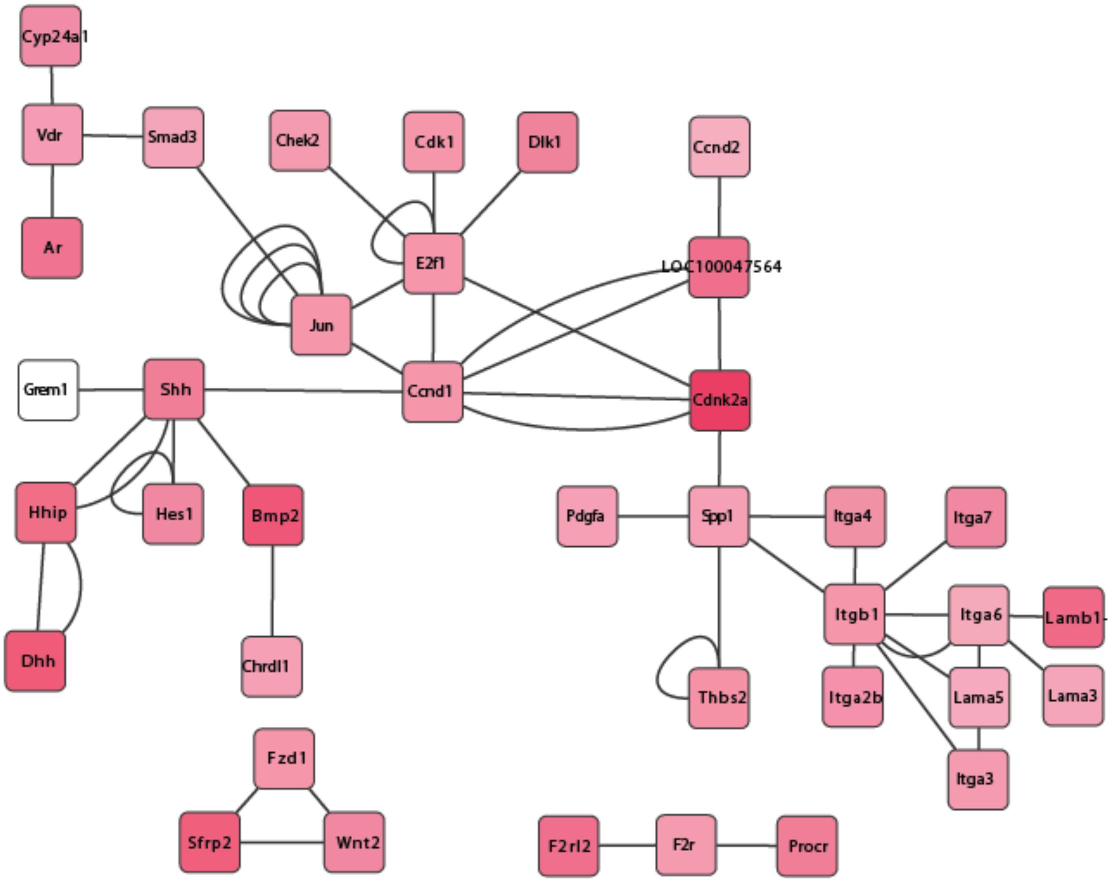


Figure 4.6: Gene interaction network of differentially expressed genes in Tr. Neurons in comparison to neurons. The module shows association of genes related to the focal adhesion, cell cycle and the Wnt signaling pathways.

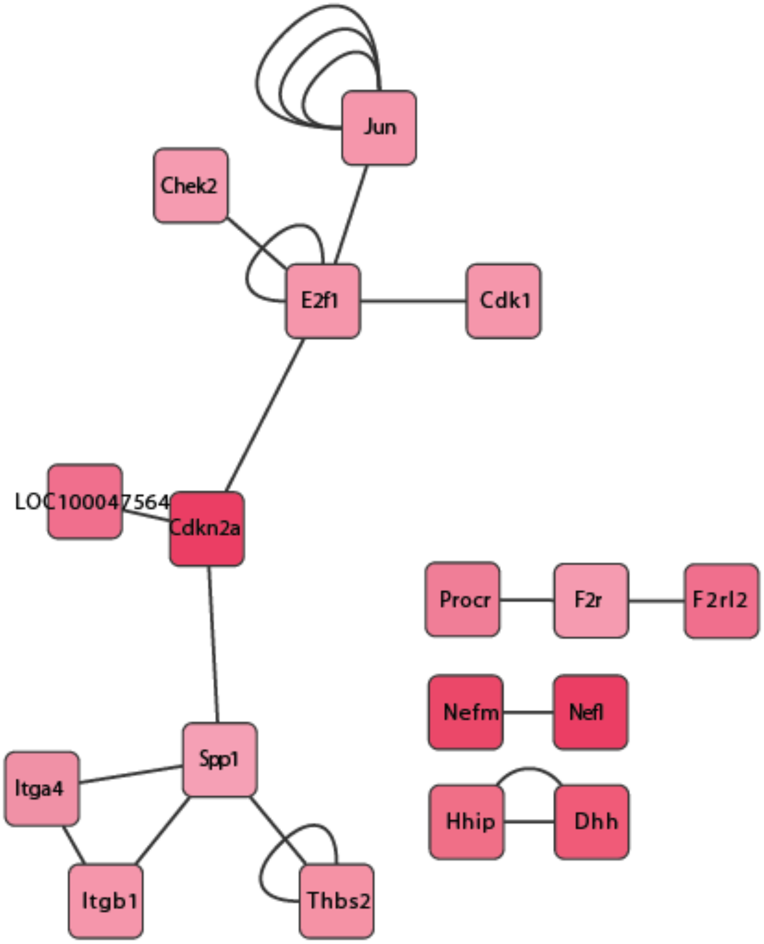


Figure 4.7: Interaction network conserved in both dedifferentiated neurons and astrocytes in comparison to their parental cell-types. The network shows connectivity between the focal adhesion and cell cycle related genes.

4.4 Discussion

Genetic alterations of mature neurons and astrocytes have been implicated in gliomagenesis[32]. High expressions of NSC markers in these tumors and the transformation of the primary cultures of neurons and astrocytes to a proliferative state leading to formation of neurospheres upon transduction with lentiviral vector, gave strong evidence of dedifferentiation of these cells to an undifferentiated stem cell like state. Dedifferentiated neurons and astrocytes also exhibited high expressions of known markers of pluripotent mESCs including SSEA1, c-myc and Nanog. They also possessed open and more relaxed chromatin structure. This raised the possibility that these cells may have regressed to an undifferentiated state that shares the characteristics of both ESCs and NSCs. To further characterize the undifferentiated state of these cells, we performed whole transcriptome analysis of the dedifferentiated cell-types along with the enriched populations of mESCs, NSCs and terminally differentiated neurons and astrocytes.

ssGSEA analysis of the gene-lists specific to the enriched populations of ESC, NSC, neurons and astrocytes revealed that dedifferentiated neurons retained the expression of some of the known neuronal markers (**Table 3**). Majority of the ESC markers had low expression in these cells. Our transcriptome data revealed high expression for some of the NSC markers (Nestin, Sox2, Wnt5a, Notch1, Msi1). However, a number of NSC specific genes showed reduced expression in these cells and consequently the enrichment scores for NSC specific gene-list was poor. While dedifferentiated neurons were far apart from ESCs at whole transcriptome level, they bore similarities to NSC and neurons suggesting that these cells did not regress completely to NSC like state. In case of dedifferentiated astrocytes, the known markers of astrocyte exhibited low expression with 311 out of 423 “putative” astrocyte markers showing down-regulation in the dedifferentiated astrocytes. On the contrary, dedifferentiated astrocytes showed positive enrichment scores for known neuron markers. GO term enrichment of these genes showed biological processes related to neuron function. This implies that these cells have left their astrocyte state and acquired a state similar to dedifferentiated neurons where they share traits similar to NSC and neurons. Interestingly, when the dedifferentiated astrocytes were placed back in the NOD SCID mice; the resulting tumors exhibited expression of some neuronal markers (Tuj1)[32]. Recent study further corroborates the cell fate plasticity observed between astrocytes and neurons[113]. We reckoned that the genetic alterations introduced by lentiviral vector containing shRNA against NF1 and p53, predisposes these cells to an undifferentiated state that lies between NSC and neurons hence the dedifferentiated cells seemed to have followed path 2 (**Figure 4.1**).

Our functional analysis of the transcriptome profiles revealed signaling pathways that were necessary for neurons and astrocytes to maintain their dedifferentiated state. Many components of cell cycle and Wnt signaling pathways were up-regulated in dedifferentiated neurons and astrocytes. This was expected, as these cells were highly proliferative in the in-vitro cultures. Focal adhesion pathway was also differentially regulated in these cell-types. The transformation in morphology from flattened profiles to free-floating three-dimensional neurospheres resulted in loss of expression of a set of focal adhesion molecules in the dedifferentiated cell-types. On the other hand, these cells acquired expression of a distinct set of focal adhesion molecules that had conserved expression profiles in the two dedifferentiated cell-types. Interestingly, these focal adhesion molecules are not highly expressed in NSCs but have high expressions in *in-vivo* cancer tissue, suggesting their unique role in cancer progression. SSGSEA analysis of the focal adhesion molecules up-regulated in Tr. Neurons showed significant positive enrichment score in Tr. Astrocytes (**Table 4.1**). Additionally, a number of focal adhesion molecules that lost their expression in these cells, were also down-regulated in NSCs in comparison to neurons.

We next projected genes up-regulated in dedifferentiated neurons on protein – protein interaction network to access their functional connectivity. The resulting network revealed distinct modules representing the three differentially regulated pathways, viz. cell cycle, Wnt signaling and focal adhesion pathway. Interestingly, many components of this network were also observed in the dedifferentiated astrocytes network, implying that the core network of genes responsible for the maintaining the dedifferentiated state are conserved in the two cell-types.

This network analysis revealed several known interactions (**Figure 4.7**). E2F1, a transcription factor, has recently been shown to support NSC proliferation and its down-regulation is required for differentiation of NSCs to neurons[114]. Spp1 is a secretory protein that has been associated with cell migration and proliferation via interaction with its receptor, α5β3 integrin[115]. Recent studies have revealed high expression of Spp1 in glioblastoma and its role in cell adhesion and invasiveness of the cancer cell[116, 117]. In our network, this protein connects cell cycle module to the focal adhesion pathway. To validate our network connectivity and elucidate possible role of focal adhesion pathway in maintenance of the dedifferentiated state of neurons and astrocytes, it will be desirable to inhibit the expression of this gene or neutralize the protein expression via antibody. Furthermore, our analysis revealed up-regulation of a number of transcription factors (**Figure 4.8**) in the dedifferentiated cell-types, many of whom are well known cell cycle related transcription factors. Some of these transcription factors (Jun and Prrx1) are known to regulate expression of focal adhesion molecules up-regulated in the dedifferentiated neurons. Genetic perturbation of these transcription factors will facilitate identification of gene regulatory networks responsible for maintenance of the undifferentiated state of the neurons and astrocytes. Finally, comparative analysis of these gene regulatory networks across other dedifferentiation models will elucidate molecular mechanisms responsible for cell-fate plasticity.

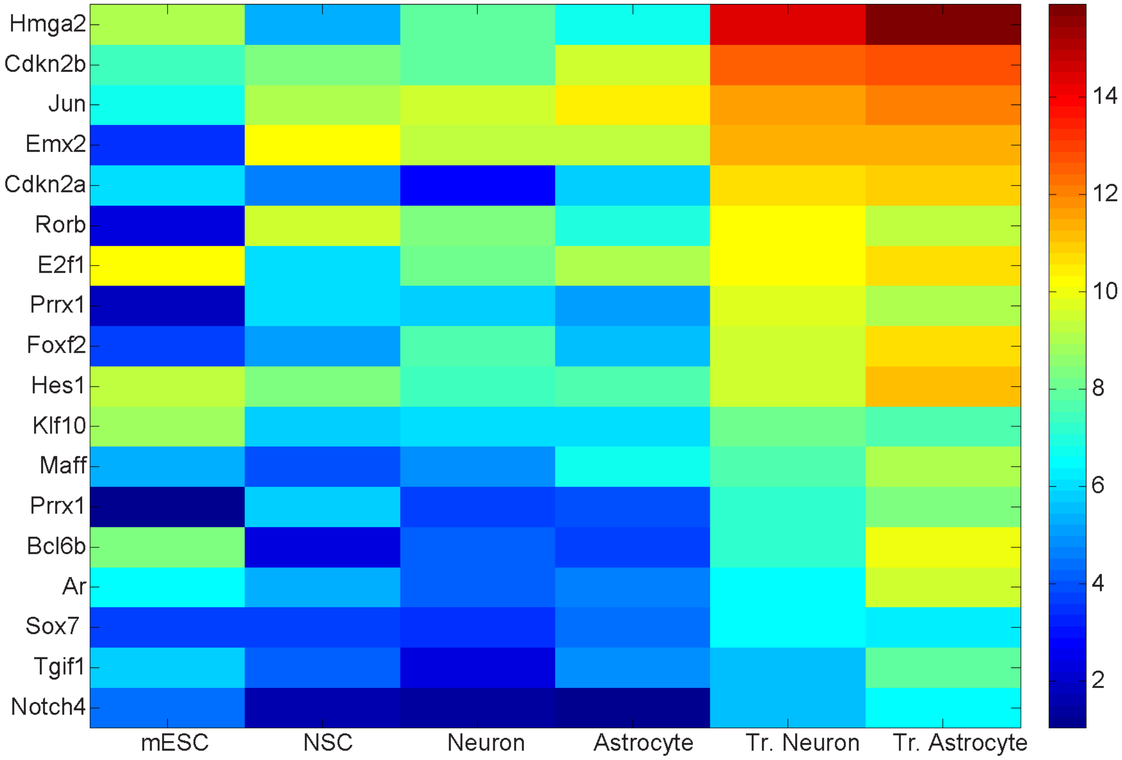


Figure 4.8: Heatmap of the transcription factors that were significantly up-regulated in dedifferentiated neurons and astrocytes in comparison to their respective parental cell-types.

4.5 Materials and Methods

4.5.1 Cell Culture

Primary astrocytes and neurons were obtained from 11 days postnatal pups from GFAP-Cre and SynapsinI-Cre transgenic mice respectively, and prepared according to published methods[118, 119].

Astrocytes were maintained in DMEM containing 10% FBS and neurons were cultured in Neurobasal™-A Medium (Gibco) containing Glutamax™ (Gibco) and B-27 supplement (Gibco). Following transduction of either primary astrocytes or neurons with the lentivirus, at the early passages, cells were either cultured in the medium described above or in parallel cultured in NSCs medium containing FGF-2. NSCs media was prepared using the following reagents: DMEM/F-12 (Gibco), $NaHCO_3$, Insulin (Sigma), apo-Transferrin (Sigma), Putrescin (Sigma, Sodium Selenite (Sigma), Progesterone (Sigma), and supplemented with 20 ng/ml fibroblast growth factor-2 (Prepotech).

NSCs were isolated from E14.5 mouse embryos, the brains were microdissected to harvest the ganglionic eminences, dissociated to harvest the tissue in NSC media to gain a single cell suspension for plating in coated poly-ornithin and laminin tissue culture plates[120]. Neurospheres were passaged by dissociation of the spheres into single cells using TripLE™ Express (Gibco).

4.5.2 RNA Extraction

Total RNA was isolated using Trizol (Ambion). For extraction of poly-adenylated mRNA Dynabeads mRNA Purification Kit (Invitrogen) was used.

4.5.3 Sequencing library preparation

All mRNA samples were subjected to sequencing library preparation using DP-seq[31].

4.5.4 Quantification of the sequencing library

Quantitative real time PCR was used to determine the concentration of the sequencing libraries prepared by our protocol. The standard curve for various dilutions of phiX control library was generated using the adapter specific primers recommended by Illumina. We later used the standard curve to determine the molarity of our sequencing libraries. The concentration of sequencing library loaded into the flowcell was calibrated by the sequencing facility. We typically obtained good cluster density with 5 pM of library concentration on HiSeq v3 kit.

4.5.5 Mapping reads

All libraries were sequenced by Illumina's HiSeq2000 systems (TruSeq SR Cluster Kit v3-cBot-HS and TruSeq SBS Kit v3-HS). The libraries were sequenced as 50 bp single-end reads, except for astrocytes samples, which were sequenced as 100 bp single-end reads. The first 7 sequencing cycles were truncated as they came from our heptamer primers and the following 32 bp reads were mapped to the mouse NCBI refseq database allowing up to 2 mismatches, using our in house mapping software which implements suffix array data structure. The reads that did not align to the NCBI Refseq database were later aligned to the mouse genomic locations using bowtie software[67] (≤ 2 mismatches).

4.5.6 Differential gene expression analysis

Unique reads obtained from different samples were quantile normalized. This method normalizes the sequencing libraries assuming that the distribution of reads for all the transcripts come from the same underlying distribution, regardless of the cell-type. Differentially expressed genes were identified using local pooled error test (LPE)[70]. A p-value cut-off of 0.05 was used to assign significance to the differentially expressed genes.

4.5.7 Single Sample Gene Set Enrichment Analysis

The sequencing data was arranged into matrix, where rows and column represented the number of genes and different cells lines respectively. The sequencing measurement of biological replicates were averaged and assigned to each column. There were four primary cell lines, three transformed cell lines and one cancer cell line. Gene counts for each gene (row) were normalized to get the gene rank order among the cell lines. The normalization parameters, mean, standard deviation and mean absolute deviation, were calculated based on four primary cell lines and whole row was normalized as:

$$x_i = \frac{(x_i - \overline{x})}{\sigma_x} * \frac{x_{mad}}{\overline{x}}$$

Where, $\overline{x}$, x_{mad} and σ_x are mean, standard deviation and mean absolute deviation, respectively, calculated based on four primary cell lines. An arbitrary MAD cutoff of >= 20 was applied to eliminate the genes with low expression and low variation from the analysis. The rescaled gene rankings for each cell line (column) were used for single sample gene set analysis (SSGSEA)

4.6 Supplementary Figures

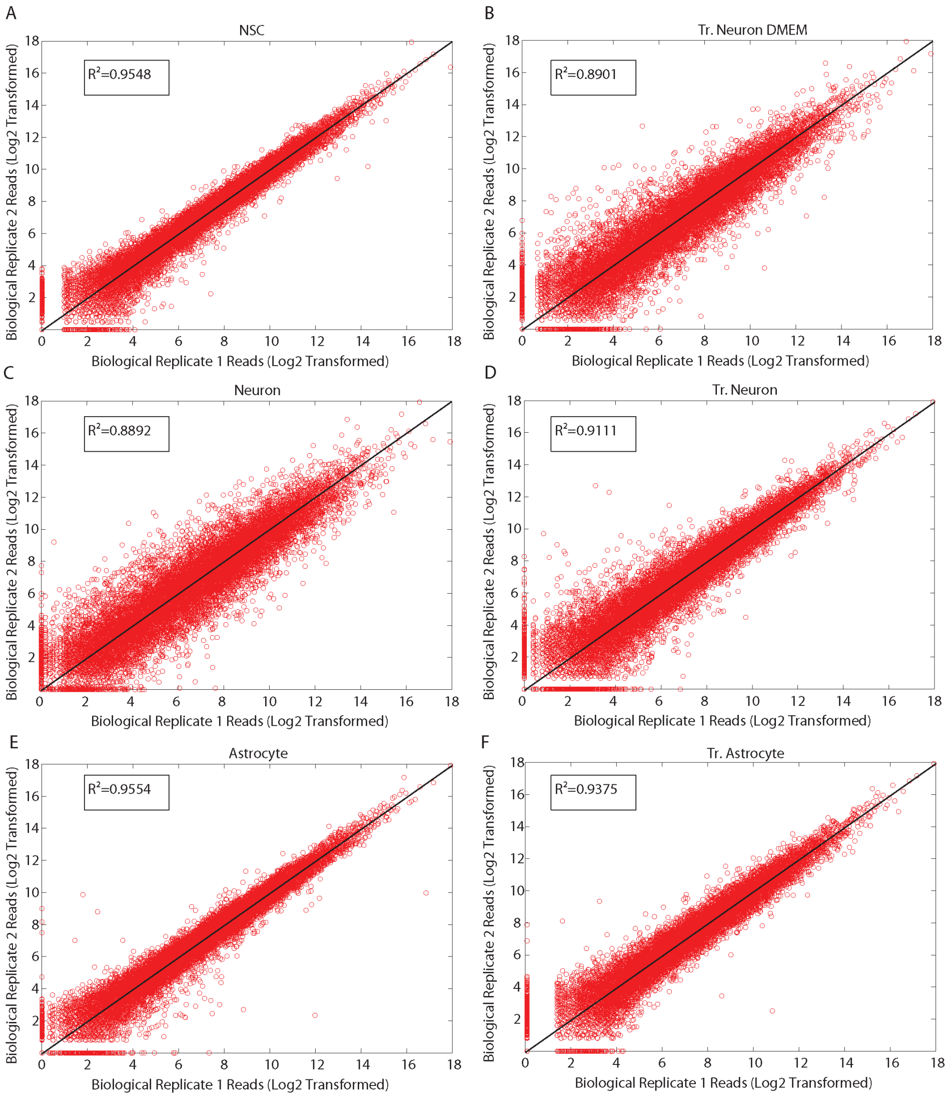


Figure S4.1: Biological Replicates. The variations between the biological replicates reflect the biological variability as well as technical variations arising from the sequencing library generation and the sequencing platform. The neuron biological replicates possessed high variations.

4.7 Supplementary Tables

Table S4.1: Mapping statistics. Tr. refers to dedifferentiated Neurons and Astrocytes. BR refers to biological replicates.

Samples	Total Reads	Uniquely Mapped	Non Uniquely mapped	Transcripts >=1 unique reads
mESC	19895309	60.08	22.54	14721
NSC BR1	18729943	45.44	19.31	14553
NSC BR2	31350382	43.69	18.69	15149
Neuron BR1	17932355	51.21	19.28	15370
Neuron BR2	19564005	47.50	19.53	15964
Astrocyte BR1	23744582	41.85	14.09	15214
Astrocyte BR2	18369803	42.46	14.43	14789
Tr. Neuron BR1	18838348	53.19	19.13	14748
Tr. Neuron BR2	25653027	56.95	19.46	14822
Tr. Neuron DMEM BR1	26149612	49.29	18.13	15129
Tr. Neuron DMEM BR2	19120915	50.16	18.07	14762
Tr. Astrocyte BR1	17524415	47.60	15.86	13802
Tr. Astrocyte BR2	29102865	49.08	15.58	14800
In-vivo Cancer	14247457	54.27	18.49	15226

4.8 Acknowledgement

Chapter 4, in full is currently being prepared for submission for publication of the material. Dinorah, F.M.*, **Bhargava, V.***, Gupta, S., Verma, I., Subramaniam, S.. The dissertation author was a joint first author of this paper, responsible for sequencing library generation and much of data analysis. *Equal contribution

Chapter 5

Future Directions

In this thesis, I described a novel amplification-based strategy that uses a defined set of heptamer primers to amplify the majority of the mouse transcripts from as low as 50 pg of mRNA. Some of the useful features of our methodology are:

1. The library construction requires a very small amount of purified poly “A” mRNA (~ 25 - 50 pg).
2. The amplification of genes ensures better representation of low abundant genes.
3. Since the heptamer-primer binding sites and their corresponding amplicon lengths are already known, we do not expect transcript length bias as observed in other RNA-seq strategies.
4. In our method, multiple amplicons are generated from each mRNA; thereby providing technical replicates to access statistically relevant gene expression.
5. Our methodology can be adapted to provide strand - specific gene expression profiles.
6. Our strategy has potential for “targeted amplification” to perform preferential amplification of genes involved in a given pathway (eg. Wnt signaling pathway) or a phenotype (eg. pluripotency related genes).

7. Designing primers specific for known “hotspots” in the mammalian chromosomes could detect chromosomal abnormalities and structural variations.
8. The strategy could be extended to sequence promoter sites of all genes to study epigenetic mark up by subjecting the mammalian genomes to bi-sulphite treatment.
9. Another application of this strategy would be to design minimal set of primers to specifically sequence regions of the genome mutated in genetic diseases. These primers set could act as a diagnostic kit for assessing the susceptibility of individuals towards certain diseases.

These applications would require further improvements in our primer generation pipeline and optimization of the targeted cDNA amplification protocol. Some of the natural directions for future research are listed below:

A) Quantitative Prediction Model: Heptamer primers are highly promiscuous with thousands of primer binding sites on the mouse transcriptome. However, only a small proportion of these primer-binding sites are experimentally observed. The PCR biases associated with our protocol causes preferential amplification of these sites. Quantification of these PCR biases will help formulate a *quantitative prediction model* that would access amplification efficiency of all possible amplicons given a set of heptamer primers and the transcriptome sequence. This model will further facilitate designing heptamer primers that selectively/preferentially amplify “genes of interest” while minimizing the representation of unwanted transcripts. One of such instances was demonstrated in Chapter 2 where we constructively exploited PCR biases and designed multiple primer sets that reduced the representation of highly expressing ribosomal transcripts by more than 70% while maintaining the overall transcriptome coverage. This methodology could also be used to design primers specific to a particular phenotype (cancer related genes) or a biological process (stem cell differentiation). Since, only subsets of the transcripts are amplified, multiple samples can be combined, thus bringing cost effectiveness.

B) Primers for Human Transcriptomes: The prediction model can also be utilized to design primers specific to splice junctions residing within the isoform groups of the human transcriptome. About 90% of human genes exhibit some form of alternative splicing. Of these genes, 50-80% show tissue-specific splicing. This supports the hypothesis that alternative splicing plays an important role in the development of phenotypic complexity in mammals[121]. Reliable quantitation of isoform transcripts has remained a challenge because of high sequence identity between the isoforms. Using 30bp reads with one mismatch, approximately 70% of the human transcriptome was found to be unique[15]. However, within an isoform groups, only 13% of the sequences are unique implying that the vast majority of the sequencing reads prepared from the most popular RNA-seq method will map non-uniquely to these transcripts. Our methodology will perform targeted amplification of the splice junctions that distinguishes the isoforms within an isoform group, thus providing quantitative information while occupying less sequencing space. Moreover, our method will provide sequence information of the splice junctions in the human transcriptome.

This will facilitate identification of disease associated mutations/single nucleotide polymorphisms in and around the splice junctions. A number of point mutations have been identified in the vicinity of mRNA splice junctions that alter the efficiency of mRNA splicing resulting in various disease phenotypes[122]. We believe our dataset could be used to accurately predict these mutations and provide a basis for designing a diagnostic kit.

C) SNPs/Disease causing mutation detection: The need for personalized drugs has been steadily gaining momentum over last decade. The observation that two individuals respond differently to a given treatment implies inherent genetic heterogeneity between them. The genetic heterogeneity is often manifested in form of single nucleotide polymorphisms (SNPs) and copy number variations introduced by genomic rearrangements. Most of the SNPs are located in the translated regions of the transcriptome and affect the gene expression by manipulating the mechanisms by which cis-regulatory elements interact with transcription factors/activators. Similarly, numerous cancer-related mutations were identified in diseased human tissues using whole genome/exome/transcriptome sequencing[123-127]. More than 100x sequencing depth is typically required to achieve high confidence in calling mutations/SNPs. Since, present protocols of sequencing library generation do not discriminate the regions to sequence, a vast majority of sequencing cost is wasted while achieving such a depth. Moreover, the requirement for large amount of starting material further restricts the applicability of the existing strategies in accurate detection of mutations/SNPs in human tissues. Here we propose, designing heptamer primers that preferentially amplify regions carrying disease related SNPs/mutations in the human genome. We will use known databases on cancer and other genetic disorder related SNPs/mutations to design primers that hybridize upstream to these regions thus providing us with high quality sequencing reads and high coverage to accurately call SNPs/mutations. Similar strategy could be utilized to study DNA methylation patterns to identify pattern of cytosine residue methylation upon bisulphite treatment of CpG islands located in the mammalian genomes.

D) Single cell transcriptomics - DP-seq showed high technical variations and accumulation of PCR spurious products at 25 pg libraries. One of the challenges with DP-seq is that the 44 heptamer primers are split into three tubes which implies that only 8.33 pg of mRNA was amplified by each tube. A better primer design where more primers can be accommodated in a single tube to get similar transcriptome coverage while ensuring that no two primers have ΔG <-4 Kcal/mol, is expected to reduce the technical noise and make our strategy compatible with single cell transcriptomics.

Bibliography

1. Shi, L. et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol* **24**, 1151-1161 (2006).

2. Guo, L. et al. Rat toxicogenomic study reveals analytical consistency across microarray platforms. *Nat Biotechnol* **24**, 1162-1169 (2006).

3. Shendure, J. The beginning of the end for microarrays? *Nat Methods* **5**, 585-587 (2008).

4. Licatalosi, D.D. & Darnell, R.B. RNA processing and its regulation: global insights into biological networks. *Nat Rev Genet* **11**, 75-87 (2010).

5. Marguerat, S. & Bahler, J. RNA-seq: from technology to biology. *Cell Mol Life Sci* **67**, 569-579 (2010).

6. Asmann, Y.W. et al. 3' tag digital gene expression profiling of human brain and universal reference RNA using Illumina Genome Analyzer. *BMC Genomics* **10**, 531 (2009).

7. 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).

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

9. Metzker, M.L. Sequencing technologies - the next generation. *Nat Rev Genet* **11**, 31-46 (2010).

10. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621-628 (2008).

11. Oshlack, A. & Wakefield, M.J. Transcript length bias in RNA-seq data confounds systems biology. *Biol Direct* **4**, 14 (2009).

12. Fang, Z. & Cui, X. Design and validation issues in RNA-seq experiments. *Brief Bioinform* **12**, 280-287 (2011).

13. Bloom, J.S., Khan, Z., Kruglyak, L., Singh, M. & Caudy, A.A. Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. *BMC Genomics* **10**, 221 (2009).

14. Hansen, K.D., Brenner, S.E. & Dudoit, S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res* **38**, e131 (2010).

15. Koehler, R., Issac, H., Cloonan, N. & Grimmond, S.M. The uniqueome: a mappability resource for short-tag sequencing. *Bioinformatics* **27**, 272-274 (2011).

16. Pfeifer, G.P., Steigerwald, S.D., Mueller, P.R., Wold, B. & Riggs, A.D. Genomic sequencing and methylation analysis by ligation mediated PCR. *Science* **246**, 810-813 (1989).

17. Dean, F.B. et al. Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci U S A* **99**, 5261-5266 (2002).

18. Dafforn, A. et al. Linear mRNA amplification from as little as 5 ng total RNA for global gene expression analysis. *Biotechniques* **37**, 854-857 (2004).

19. Eberwine, J. et al. Analysis of gene expression in single live neurons. *Proc Natl Acad Sci U S A* **89**, 3010-3014 (1992).

20. Adli, M., Zhu, J. & Bernstein, B.E. Genome-wide chromatin maps derived from limited numbers of hematopoietic progenitors. *Nat Methods* **7**, 615-618 (2010).

21. Armour, C.D. et al. Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. *Nat Methods* **6**, 647-649 (2009).

22. Li, H. et al. Determination of tag density required for digital transcriptome analysis: application to an androgen-sensitive prostate cancer model. *Proc Natl Acad Sci U S A* **105**, 20179-20184 (2008).

23. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep* **2**, 666-673 (2012).

24. Gertz, J. et al. Transposase mediated construction of RNA-seq libraries. *Genome Res* **22**, 134-141 (2012).

25. Ramskold, D. et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol* **30**, 777-782 (2012).

26. Hoeijmakers, W.A., Bartfai, R., Francoijs, K.J. & Stunnenberg, H.G. Linear amplification for deep sequencing. *Nat Protoc* **6**, 1026-1036 (2011).

27. Tang, F. et al. Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. *Cell Stem Cell* **6**, 468-478 (2010).

28. Levin, J.Z. et al. Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. *Genome Biol* **10**, R115 (2009).

29. Li, J.B. et al. Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science* **324**, 1210-1213 (2009).

30. Zhang, K. et al. Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. *Nat Methods* **6**, 613-618 (2009).

31. Bhargava, V., Ko, P., Willems, E., Mercola, M. & Subramaniam, S. Quantitative transcriptomics using designed primer-based amplification. *Sci Rep* **3**, 1740 (2013).

32. Friedmann-Morvinski, D. et al. Dedifferentiation of neurons and astrocytes by oncogenes can induce gliomas in mice. *Science* **338**, 1080-1084 (2012).

33. Ozsolak, F. & Milos, P.M. RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet* **12**, 87-98 (2011).

34. Tang, F. et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* **6**, 377-382 (2009).

35. Gadue, P., Huber, T.L., Paddison, P.J. & Keller, G.M. Wnt and TGF-beta signaling are required for the induction of an in vitro model of primitive streak formation using embryonic stem cells. *Proc Natl Acad Sci U S A* **103**, 16806-16811 (2006).

36. Willems, E. & Leyns, L. Patterning of mouse embryonic stem cell-derived pan-mesoderm by Activin A/Nodal and Bmp4 signaling requires Fibroblast Growth Factor activity. *Differentiation* **76**, 745-759 (2008).

37. Armes, N.A. & Smith, J.C. The ALK-2 and ALK-4 activin receptors transduce distinct mesoderm-inducing signals during early Xenopus development but do not co-operate to establish thresholds. *Development* **124**, 3797-3804 (1997).

38. Gurdon, J.B., Harger, P., Mitchell, A. & Lemaire, P. Activin signalling and response to a morphogen gradient. *Nature* **371**, 487-492 (1994).

39. Jones, C.M., Kuehn, M.R., Hogan, B.L., Smith, J.C. & Wright, C.V. Nodal-related signals induce axial mesoderm and dorsalize mesoderm during gastrulation. *Development* **121**, 3651-3662 (1995).

40. Sulzbacher, S., Schroeder, I.S., Truong, T.T. & Wobus, A.M. Activin A-induced differentiation of embryonic stem cells into endoderm and pancreatic progenitors-the influence of differentiation factors and culture conditions. *Stem Cell Rev* **5**, 159-173 (2009).

41. Tam, P.P., Kanai-Azuma, M. & Kanai, Y. Early endoderm development in vertebrates: lineage differentiation and morphogenetic function. *Curr Opin Genet Dev* **13**, 393-400 (2003).

42. Inman, G.J. et al. SB-431542 is a potent and specific inhibitor of transforming growth factor-beta superfamily type I activin receptor-like kinase (ALK) receptors ALK4, ALK5, and ALK7. *Mol Pharmacol* **62**, 65-74 (2002).

43. Vallier, L. et al. Early cell fate decisions of human embryonic stem cells and mouse epiblast stem cells are controlled by the same signalling pathways. *PLoS One* **4**, e6082 (2009).

44. Pevny, L.H., Sockanathan, S., Placzek, M. & Lovell-Badge, R. A role for SOX1 in neural determination. *Development* **125**, 1967-1978 (1998).

45. Dahle, O., Kumar, A. & Kuehn, M.R. Nodal signaling recruits the histone demethylase Jmjd3 to counteract polycomb-mediated repression at target genes. *Sci Signal* **3**, ra48 (2010).

46. Guzman-Ayala, M. et al. Graded Smad2/3 activation is converted directly into levels of target gene expression in embryonic stem cells. *PLoS One* **4**, e4268 (2009).

47. Zajac, P., Oberg, C. & Ahmadian, A. Analysis of short tandem repeats by parallel DNA threading. *PLoS One* **4**, e7823 (2009).

48. Katoh, M. CER1 is a common target of WNT and NODAL signaling pathways in human embryonic stem cells. *Int J Mol Med* **17**, 795-799 (2006).

49. Zhang, Y. et al. High throughput determination of TGFbeta1/SMAD3 targets in A549 lung epithelial cells. *PLoS One* **6**, e20319 (2011).

50. Vallier, L. et al. Activin/Nodal signalling maintains pluripotency by controlling Nanog expression. *Development* **136**, 1339-1349 (2009).

51. Labbe, E., Silvestri, C., Hoodless, P.A., Wrana, J.L. & Attisano, L. Smad2 and Smad3 positively and negatively regulate TGF beta-dependent transcription through the forkhead DNA-binding protein FAST2. *Mol Cell* **2**, 109-120 (1998).

52. Norris, D.P., Brennan, J., Bikoff, E.K. & Robertson, E.J. The Foxh1-dependent autoregulatory enhancer controls the level of Nodal signals in the mouse embryo. *Development* **129**, 3455-3468 (2002).

53. Shiratori, H. et al. Two-step regulation of left-right asymmetric expression of Pitx2: initiation by nodal signaling and maintenance by Nkx2. *Mol Cell* **7**, 137-149 (2001).

54. Chen, B. et al. Small molecule-mediated disruption of Wnt-dependent signaling in tissue regeneration and cancer. *Nat Chem Biol* **5**, 100-107 (2009).

55. Hoodless, P.A. et al. FoxH1 (Fast) functions to specify the anterior primitive streak in the mouse. *Genes Dev* **15**, 1257-1271 (2001).

56. Rossant, J. & Tam, P.P. Blastocyst lineage formation, early embryonic asymmetries and axis patterning in the mouse. *Development* **136**, 701-713 (2009).

57. Yamamoto, M. et al. The transcription factor FoxH1 (FAST) mediates Nodal signaling during anterior-posterior patterning and node formation in the mouse. *Genes Dev* **15**, 1242-1256 (2001).

58. Faust, C., Schumacher, A., Holdener, B. & Magnuson, T. The eed mutation disrupts anterior mesoderm production in mice. *Development* **121**, 273-285 (1995).

59. Kattman, S.J. et al. Stage-specific optimization of activin/nodal and BMP signaling promotes cardiac differentiation of mouse and human pluripotent stem cell lines. *Cell Stem Cell* **8**, 228-240 (2011).

60. Kishigami, S. & Mishina, Y. BMP signaling and early embryonic patterning. *Cytokine Growth Factor Rev* **16**, 265-278 (2005).

61. Nostro, M.C., Cheng, X., Keller, G.M. & Gadue, P. Wnt, activin, and BMP signaling regulate distinct stages in the developmental pathway from embryonic stem cells to blood. *Cell Stem Cell* **2**, 60-71 (2008).

62. Labaj, P.P. et al. Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling. *Bioinformatics* **27**, i383-391 (2011).

63. Wamstad, J.A. et al. Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage. *Cell* **151**, 206-220 (2012).

64. Paige, S.L. et al. A temporal chromatin signature in human embryonic stem cells identifies regulators of cardiac development. *Cell* **151**, 221-232 (2012).

65. Markham, N.R. & Zuker, M. UNAFold: software for nucleic acid folding and hybridization. *Methods Mol Biol* **453**, 3-31 (2008).

66. Zhao, G. & Guan, Y. Polymerization behavior of Klenow fragment and Taq DNA polymerase in short primer extension reactions. *Acta Biochim Biophys Sin (Shanghai)* **42**, 722-728 (2010).

67. Langmead, B. Aligning short sequencing reads with Bowtie. *Curr Protoc Bioinformatics* **Chapter 11**, Unit 11 17 (2010).

68. Bullard, J.H., Purdom, E., Hansen, K.D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11**, 94 (2010).

69. McIntyre, L.M. et al. RNA-seq: technical variability and sampling. *BMC Genomics* **12**, 293 (2011).

70. Jain, N. et al. Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays. *Bioinformatics* **19**, 1945-1951 (2003).

71. Furusawa, C. & Kaneko, K. Zipf's law in gene expression. *Phys Rev Lett* **90**, 088102 (2003).

72. Ueda, H.R. et al. Universality and flexibility in gene expression from bacteria to human. *Proc Natl Acad Sci U S A* **101**, 3765-3769 (2004).

73. Nookaew, I. et al. A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. *Nucleic Acids Res* **40**, 10084-10097 (2012).

74. Sasagawa, Y. et al. Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biol* **14**, R31 (2013).

75. Pan, X. et al. Two methods for full-length RNA sequencing for low quantities of cells and single cells. *Proc Natl Acad Sci U S A* **110**, 594-599 (2013).

76. Islam, S. et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res* **21**, 1160-1167 (2011).

77. Tang, F., Lao, K. & Surani, M.A. Development and applications of single-cell transcriptome analysis. *Nat Methods* **8**, S6-11 (2011).

78. Qiu, S. et al. Single-neuron RNA-Seq: technical feasibility and reproducibility. *Front Genet* **3**, 124 (2012).

79. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol* **11**, R106 (2010).

80. Baker, S.C. et al. The External RNA Controls Consortium: a progress report. *Nat Methods* **2**, 731-734 (2005).

81. Elowitz, M.B., Levine, A.J., Siggia, E.D. & Swain, P.S. Stochastic gene expression in a single cell. *Science* **297**, 1183-1186 (2002).

82. Bar-Even, A. et al. Noise in protein expression scales with natural protein abundance. *Nat Genet* **38**, 636-643 (2006).

83. Kim, J.K. & Marioni, J.C. Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biol* **14**, R7 (2013).

84. Brockes, J.P. & Kumar, A. Plasticity and reprogramming of differentiated cells in amphibian regeneration. *Nat Rev Mol Cell Biol* **3**, 566-574 (2002).

85. Poss, K.D., Wilson, L.G. & Keating, M.T. Heart regeneration in zebrafish. *Science* **298**, 2188-2190 (2002).

86. Raya, A. et al. Activation of Notch signaling pathway precedes heart regeneration in zebrafish. *Proc Natl Acad Sci U S A* **100 Suppl 1**, 11889-11895 (2003).

87. Chen, Z.L., Yu, W.M. & Strickland, S. Peripheral regeneration. *Annu Rev Neurosci* **30**, 209-233 (2007).

88. Mirsky, R. et al. Novel signals controlling embryonic Schwann cell development, myelination and dedifferentiation. *J Peripher Nerv Syst* **13**, 122-135 (2008).

89. Buffo, A. et al. Origin and progeny of reactive gliosis: A source of multipotent cells in the injured brain. *Proc Natl Acad Sci U S A* **105**, 3581-3586 (2008).

90. Seri, B., Garcia-Verdugo, J.M., McEwen, B.S. & Alvarez-Buylla, A. Astrocytes give rise to new neurons in the adult mammalian hippocampus. *J Neurosci* **21**, 7153-7160 (2001).

91. Lang, B. et al. Astrocytes in injured adult rat spinal cord may acquire the potential of neural stem cells. *Neuroscience* **128**, 775-783 (2004).

92. Mu, X., Peng, H., Pan, H., Huard, J. & Li, Y. Study of muscle cell dedifferentiation after skeletal muscle injury of mice with a Cre-Lox system. *PLoS One* **6**, e16699 (2011).

93. McGann, C.J., Odelberg, S.J. & Keating, M.T. Mammalian myotube dedifferentiation induced by newt regeneration extract. *Proc Natl Acad Sci U S A* **98**, 13699-13704 (2001).

94. Rosania, G.R. et al. Myoseverin, a microtubule-binding molecule with novel cellular effects. *Nat Biotechnol* **18**, 304-308 (2000).

95. Okita, K., Ichisaka, T. & Yamanaka, S. Generation of germline-competent induced pluripotent stem cells. *Nature* **448**, 313-317 (2007).

96. Park, I.H. et al. Reprogramming of human somatic cells to pluripotency with defined factors. *Nature* **451**, 141-146 (2008).

97. Yu, J. et al. Induced pluripotent stem cell lines derived from human somatic cells. *Science* **318**, 1917-1920 (2007).

98. Schwitalla, S. et al. Intestinal tumorigenesis initiated by dedifferentiation and acquisition of stem-cell-like properties. *Cell* **152**, 25-38 (2013).

99. Debeb, B.G. et al. Histone deacetylase inhibitors stimulate dedifferentiation of human breast cancer cells through WNT/beta-catenin signaling. *Stem Cells* **30**, 2366-2377 (2012).

100. Thiery, J.P. Epithelial-mesenchymal transitions in tumour progression. *Nat Rev Cancer* **2**, 442-454 (2002).

101. Cui, W. et al. TGFbeta1 inhibits the formation of benign skin tumors, but enhances progression to invasive spindle carcinomas in transgenic mice. *Cell* **86**, 531-542 (1996).

102. Chaffer, C.L. et al. Normal and neoplastic nonstem cells can spontaneously convert to a stem-like state. *Proc Natl Acad Sci U S A* **108**, 7950-7955 (2011).

103. Jopling, C., Boue, S. & Izpisua Belmonte, J.C. Dedifferentiation, transdifferentiation and reprogramming: three routes to regeneration. *Nat Rev Mol Cell Biol* **12**, 79-89 (2011).

104. Zhu, H. et al. Oncogenic EGFR signaling cooperates with loss of tumor suppressor gene functions in gliomagenesis. *Proc Natl Acad Sci U S A* **106**, 2712-2716 (2009).

105. Liu, C. et al. Mosaic analysis with double markers reveals tumor cell of origin in glioma. *Cell* **146**, 209-221 (2011).

106. Alcantara Llaguno, S. et al. Malignant astrocytomas originate from neural stem/progenitor cells in a somatic tumor suppressor mouse model. *Cancer Cell* **15**, 45-56 (2009).

107. Zhao, W., Ji, X., Zhang, F., Li, L. & Ma, L. Embryonic stem cell markers. *Molecules* **17**, 6196-6236 (2012).

108. Cahoy, J.D. et al. A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. *J Neurosci* **28**, 264-278 (2008).

109. Polakis, P. Wnt signaling in cancer. *Cold Spring Harb Perspect Biol* **4** (2012).

110. Holland, J.D., Klaus, A., Garratt, A.N. & Birchmeier, W. Wnt signaling in stem and cancer stem cells. *Curr Opin Cell Biol* **25**, 254-264 (2013).

111. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550 (2005).

112. Matys, V. et al. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* **31**, 374-378 (2003).

113. Corti, S. et al. Direct reprogramming of human astrocytes into neural stem cells and neurons. *Exp Cell Res* **318**, 1528-1541 (2012).

114. Palm, T. et al. A systemic transcriptome analysis reveals the regulation of neural stem cell maintenance by an E2F1-miRNA feedback loop. *Nucleic Acids Res* **41**, 3699-3712 (2013).

115. Chen, Y.J. et al. Osteopontin increases migration and MMP-9 up-regulation via alphavbeta3 integrin, FAK, ERK, and NF-kappaB-dependent pathway in human chondrosarcoma cells. *J Cell Physiol* **221**, 98-108 (2009).

116. Yamaguchi, Y. et al. Thrombin-cleaved fragments of osteopontin are overexpressed in malignant glial tumors and provide a molecular niche with survival advantage. *J Biol Chem* **288**, 3097-3111 (2013).

117. Jan, H.J. et al. Osteopontin regulates human glioma cell invasiveness and tumor growth in mice. *Neuro Oncol* **12**, 58-70 (2010).

118. McCarthy, K.D. & de Vellis, J. Preparation of separate astroglial and oligodendroglial cell cultures from rat cerebral tissue. *J Cell Biol* **85**, 890-902 (1980).

119. Meyer-Franke, A., Kaplan, M.R., Pfrieger, F.W. & Barres, B.A. Characterization of the signaling interactions that promote the survival and growth of developing retinal ganglion cells in culture. *Neuron* **15**, 805-819 (1995).

120. Azari, H., Sharififar, S., Rahman, M., Ansari, S. & Reynolds, B.A. Establishing embryonic mouse neural stem cell culture using the neurosphere assay. *J Vis Exp* (2011).

121. Wang, E.T. et al. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470-476 (2008).

122. Krawczak, M., Reiss, J. & Cooper, D.N. The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Hum Genet* **90**, 41-54 (1992).

123. Dahl, F. et al. Multigene amplification and massively parallel sequencing for cancer mutation discovery. *Proc Natl Acad Sci U S A* **104**, 9387-9392 (2007).

124. Ashktorab, H. et al. Distinct genetic alterations in colorectal cancer. *PLoS One* **5**, e8879 (2010).

125. Wang, K. et al. Exome sequencing identifies frequent mutation of ARID1A in molecular subtypes of gastric cancer. *Nat Genet* **43**, 1219-1223 (2011).

126. Forshew, T. et al. Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. *Sci Transl Med* **4**, 136ra168 (2012).

127. Banerji, S. et al. Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **486**, 405-409 (2012).