TRANSCRIPTOME PROFILING OF *RATTUS NORVEGICUS* EMBRYONIC STEM CELLS BY RNA-SEQUENCING

A Thesis

presented to

the Faculty of the Graduate School

of University of Missouri-Columbia

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

By

NATHAN TYLER JOHNSON

Dr. Elizabeth C. Bryda, Thesis Advisor

December 2014

The undersigned, appointed by the Dean of the Graduate School, have examined the thesis entitled

TRANSCRIPTOME PROFILING OF RATTUS NORVEGICUS EMBRYONIC STEM CELLS BY RNA-SEQ

presented by Nathan Tyler Johnson,

a candidate for the degree of Master of Science
and hereby certify that in their opinion it is worthy of acceptance.

Dr. Elizabeth C. Bryda

Dr. Kevin D. Wells

Dr. James M. Amos-Landgraf

DEDICATION

I would like to dedicate this thesis to my wife, Amanda Johnson. Without her love, patience, understanding, and simply putting up with my moods when things were not going right, made this work possible. Thank you for your continued support and journey.

ACKNOWLEDGEMENTS

There are many individuals that I would like to thank for their time and support while at the University of Missouri. First, I would like to thank the members of my committee for their continued suggestions, comments, and support. I would like to thank my extended lab mates Mary Shaw, and Anagha Sawant for being a sounding board for issues. I would like specifically thank my lab manager Miriam Hankins for her continued support in all that I did. I would like to thank my fellow graduate students simply for the pleasure of knowing them. I would like to thank my former committee members, Drs. Change Tan and Mark Kirk for their continued support.

Finally, last but definitely not the least, I would like to thank Dr. Bryda for her continued patience, guidance, and time. Dr. Bryda challenged me to exceed expectations in all that I do, encouraged me when failing, and taught me how to critically analyze all that goes on around me. I will forever be in Dr. Bryda's debt for everything that she has taught me as I know that her influence will always be obvious in how I conduct myself as a scientist.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

TRANSCRIPTOME PROFILING OF RATTUS NOREVEGICUS EMBRYONIC STEM CELLS BY RNA-SEQ

Nathan T. Johnson

Dr. Elizabeth C. Bryda, Thesis Advisor

ABSTRACT

Embryonic **S**tem **C**ells (ESCs) are a critical tool for producing targeted knockout animals and understanding development. ESCs were successfully isolated from rats in 2008 and have been used in producing several targeted knockout animal models. To date, little characterization of rat ESCs (rESCs) has been done. In order to establish a rESC transcriptome, RNA-Seq was done on mRNA from the rESC cell line DAc8, the first male germline competent rat ESC line to be described and the first to be used to generate a knockout rat model. RNA-Seq was chosen as it is currently the most sensitive transcriptome analysis method. In the studies described here, the genes expressed in rat ESCs were identified, and a subset of the undescribed isoforms and unannotated rat genes revealed by this analysis were confirmed by RT-PCR analysis. Importantly, the rESC data allowed comparison with previously reported data for mouse and human ESCs to begin to understand the similarities and differences of the transcriptomes of ESCs from different mammalian species.

CHAPTER I

INTRODUCTION

Embryonic **S**tem **C**ells (ESCs) are a critical tool for producing targeted knockout animals and understanding development. ESCs were successfully isolated from rats in 2008 and have been used in producing several targeted knockout animal models (Buehr et al. 2008; Li et al. 2008; Meek et al. 2010; Tong et al. 2010; Kawamata and Ochiya 2011; Tong et al. 2011; Yamamoto et al. 2011). However, despite their usefulness, detailed characterization of rat ESCs (rESCs) has been minimal and the transcriptome of rat ESCs has not been defined. Establishing the genetic expression pattern of normal rESCs will provide a base line for further exploring a variety of aspects of gene expression in rESCs for future experiments. In order to establish a rESC transcriptome, RNA-Sequencing (RNA-Seq) was performed with mRNA from the rESC cell line DAc8, the first male germline competent rat ESC line to be described and the first to be used to generate a knockout rat model. Undescribed isoforms and unannotated rat genes were identified and this data was confirmed by RT-PCR. Additionally, the expression data for rat ESCs was compared and contrasted to previously reported data for human and mouse ESC expressed genes.

Purpose of the Research

The objectives of this study are 1) to characterize the rESC transcriptome and determine what genes are expressed, and 2) to compare the rat ESC transcriptome with that of human and mouse ESC transcriptomes to gain insight into ESC expression patterns across species.

Significance of the Research

The rationale for this study is once we establish a normal expression pattern for rESCs, this expression pattern can be used as a baseline for gene expression in rESCs for comparisons with other rESC lines as well as human and mouse ESCs. It allows commonalities among species to be explored. Importantly, it will allow comparisons with other rESCs to be possible in order to address fundamental questions such as what genes are important for maintaining pluripotency in rESCs or what genes are important for germline competency.

Organization of the Thesis

This thesis is divided into five chapters. Chapter I describes the purpose of this thesis. Chapter II is a review of the current relevant literature as related to the definition and use of embryonic stem cells, the differences among human, mouse, and rat embryonic stem cells, RNA-Seq as a tool for gene expression analysis, and the definition of the embryonic stem cell state. Chapter III is a detailed description of the materials, methods, and experimental design. Chapter

IV contains the results. Chapter V is a discussion of the conclusions that can be drawn from the results of the research described in this thesis.

CHAPTER II

REVIEW OF THE LITERATURE

Embryonic Stem Cells

The defining characteristics of **E**mbryonic **S**tem **C**ells (ESCs) are 1) the ability to be maintained indefinitely *in vitro* in an undifferentiated state (Martin et al. 1977; Evans and Kaufman 1981; Martin 1981), 2) the ability to express pluripotency markers (Scholer et al. 1991; Ambrosetti et al. 1997), 3) the ability to contribute to all 3 germ layers (mesoderm, endoderm, and ectoderm) *in vitro* and *in vivo* (Kleinsmith and Pierce 1964; Brinster 1974; Martin and Evans 1975), and 4) the ability to contribute to the germline *in vivo* (Bradley et al. 1984; Schwartzberg et al. 1989; Smith 2001).

ESCs are isolated from the inner cell mass of a blastocyst and can be maintained in the artificial laboratory environment for an extended period of time (Jakob 1984). A key aspect of ESCs is the capability to contribute to the germline *in vivo* (Suzuki et al. 1997). Testing for this ability is analyzed by injecting ESCs into a blastocyst and following the genetic contribution of the ESCs in the resulting chimeric animal (Smith 2001). Successful contribution is generally determined by coat color. For example, ESCs carrying genes specifying for one coat color (i.e. black) will be injected into recipient blastocysts that carry genes specifying for another coat color (i.e. white). Therefore, if there is any contribution of the injected ESCs to the resulting animal, it will be evident in the coat color (i.e. black hairs on a primarily white coat, in this example, Figure 2.1). These chimeras are

then bred in order to verify if their progeny will inherit genetic material contributed by the injected ESCs. In the previous example, if the offspring of the chimera mated to a white coated animal have a black coat color it is evidence of germline transmission (Figure 2.1).

Figure 2.1. Rat Chimeras. The rat chimera (right) was generated by injection of Dark Agouti (DA) rat ES cells into a Fischer 344 (white) blastocyst and subsequent transfer of the embryo to a recipient Sprague-Dawley rat. The agouti coat color denotes the presence of DA ES cell–derived cells in the albino Fischer 344 host. The germline transmission of the DA ES cell genome in the offspring can be easily identified by the appearance of agouti coat color when the chimera is mated with albino Sprague-Dawley rats (Tong et al. 2011). Image reproduced with permission from the Nature Publishing group.

The study of ESCs has had wide implications as both a basic research and therapeutic tool (Ben-David et al. 2012). The increasingly growing interest in ESCs stems from their utility as tools for 1) genetic manipulation in order to produce animal models for studying human disease and/or gene function, and 2) asking general questions as related to genetics, epigenetics and/or cell biology, and stem cell therapy (Smith 2001; Bernstein et al. 2006; Fouse et al. 2008; Meissner et al. 2008; Guttman et al. 2009; Collin and Lako 2011; Huang et al. 2011; Chan and Gantenbein-Ritter 2012; Ong and da Cruz 2012; Serra et al.

2012). Generation of animal models for human disease and the creation of genetically modified animals to study the effects of knocking out genes has resulted in thousands of models that have revolutionized how scientists conduct research (Schofield et al. 2011). This allows questions concerning a disease etiology to be answered without harming a human being in addition to facilitating therapeutic testing.

ESCs have been used as a tool for identifying DNA demethylation enzymes, improving genomic techniques involving zinc finger nucleases and transcription activator like effector nucleases (TALENs), and long noncoding RNAs (lincRNAs), to name a few examples (Bhutani et al. 2010; Guttman et al. 2010; Hockemeyer et al. 2011; Tong et al. 2012). Additionally, ESCs are used as therapies for curing disease and as controls for induced pluripotent stem cells (iPSCs) (Bilic and Izpisua Belmonte 2012; Chan and Gantenbein-Ritter 2012; Ong and da Cruz 2012; Serra et al. 2012). iPSCs are thought to be the solution to morale issues concerning using ESCs since iPSCs share the same pluripotent characteristics as ESCs, but are produced from differentiated cells rather than embryos (Sohn et al. 2012). Additionally, since iPSCs can be isolated from the patient, the chance of immune rejection is less likely (Serra et al. 2012). In order to produce iPSCs, differentiated cells have to be “reprogrammed” by a cocktail of transcription factors that reactivate pluripotent genes allowing for a more “naïve ” or embryonic stem cell-like state (Kang et al. 2010). In order for this naïve state to be understood, it is necessary to understand the embryonic stem cell state (Bilic and Izpisua Belmonte 2012).

Whether iPSCs or ESCs are a better stem cell solution to curing disease is still not clear (Bilic and Izpisua Belmonte 2012). An understanding of ESCs is important to human health and to our understanding of basic biology.

Embryonic Stem Cell Isolation History

It is a matter of debate whether an ESC represents a transient cell state *in vivo* or is an artificial cell state made so due to the *in vitro* conditions imposed upon it (Smith 2001). Additionally, over the last several decades defining what an "embryonic stem cell state" is has been met with some difficulty due to differences in isolation or the inability to isolate authentic ESCs in some species (Blomberg and Telugu 2012). Difficulty in understanding the ESC state can be illustrated by the history of mouse, rat, and human ESCs (mESCs, rESCs, and hESCs respectively).

The first ESCs were isolated from the 129 strain of mouse in 1981 (Evans and Kaufman 1981; Martin 1981). The 129 strain proved consistently amenable to ESC derivation and genetic manipulation. However, this strain has the disadvantage of poor breeding efficiency and is seldom the genetic background of choice (Brook and Gardner 1997). Furthermore, multiple costly and time-consuming generations of backcrossing are required to transfer a genetic alteration created using 129 ESCs to a different desired genetic background (Blair et al. 2011). The original media used to isolate mESCs relied on undefined conditions, which included the use of fetal bovine serum (FBS) and growth of mESCs on a mitotically inactive mouse embryonic fibroblast feeder layer, which

were isolated at the same time-point of mESC isolation. There have been several reports of wide variation in the quality of FBS, which can have diverse effects on mESC culture (Boone et al. 1971). Due to this, defining the mechanism for mESC maintenance became a major goal. In 1988, the key contribution of feeders was determined to be the IL-6 family cytokine LIF and in 2003, the anti-neural cytokine BMP4 was found to substitute for serum (Smith et al. 1988; Williams et al. 1988; Ying et al. 2003). Based on these findings, a feeder-free, serum-free culture condition for mESC derivation and maintenance was developed (Ying et al. 2003; Nagy and Vintersten 2006). However, using these same conditions, researchers were unable to isolate rat or human ESCs (Daheron et al. 2004; Vallier et al. 2005).

A decade later in 1998, human embryonic stem cells were successfully derived from human blastocysts, and maintained on a mouse feeder fibroblast layer with the addition of fibroblast growth factor (FGF) (Thomson et al. 1998; Levenstein et al. 2006). Since it is considered unethical to test the ability of hESC to contribute to the germline, it is only possible to define human ESCs as such based on the expression of pluripotency genes and the ability to differentiate into all three germ layers. Determination of this differentiation capability is based on transplanting hESCs into immune-deficient mice and demonstrating the formation of differentiated tumors compromised of all three germ layers (Hentze et al. 2009). Furthermore, there are mESCs and rESCs cell lines that meet the criteria of differentiation of all three germ layers *in vivo*, but not capable of contribution to the germline (Suzuki et al. 1997; Keefer et al. 2007). However, it is not possible

to test whether hESCs contribute to the germline due to ethical concerns, because of this; the question is raised whether or not all hESCs are truly ESCs.

For several decades it was possible to isolate “embryonic stem cell like” rat cells, which met all the criteria of an “embryonic stem cell state” except for the ability to go germline (Stranzinger 1996; Ruhnke et al. 2003; Demers et al. 2007; Ueda et al. 2008). It was not until 2008, when a more basal understanding of the embryonic stem cell pluripotency network in mESCs allowed advances that enable the isolation of authentic germline competent rESCs (Sato et al. 2004; Chen et al. 2006; Buehr et al. 2008; Li et al. 2008; Li et al. 2009; Kanda et al. 2012). The key to this breakthrough appears to have been the development of an appropriate media. The new rat ESC media was originally termed 3 inhibitor (3i) media, but was later reformulated to include only 2 inhibitors (2i media) (Tong et al. 2011). The philosophy behind the 2i media was ESCs do not need to be induced to be pluripotent, but rather they needed to be “shielded” from differentiation signals (Sato et al. 2004; Chen et al. 2006; Buehr et al. 2008; Li et al. 2008; Li et al. 2009; Kanda et al. 2012).

While current knowledge of the ESC state has successfully allowed establishment of ESCs from rodents, primates, and fowl, attempts at isolating authentic ESCs from other species have been unsuccessful (Martins-Taylor and Xu 2010; Maruotti et al. 2012). Identified impediments to success include differences in pluripotency characteristics, timing of pre-implantation embryo development, pluripotency pathways, and culture conditions (Mitchell et al. 2008; Cao et al. 2009; Alahdal 2011; Blomberg and Telugu 2012). This leads to the

suggestion that there is more to the ESC state in different species than is currently understood.

RNA-Sequencing (RNA-Seq)

It is well established that one of the main components of phenotype is a direct result of genome wide differential gene expression and post transcriptional modification. In order to understand gene expression, a genome wide approach must be taken. Traditionally, this was and still is done with microarray technology in which a known set of fluorescently labeled oligonucleotide probes are immobilized to a solid substrate (Schena et al. 1995). This array is then hybridized to RNA, allowing fluorescence to be measured based on the absence or presence of hybridization. However, the major limitations to this approach are 1) expressed gene sequences must be known in order to design probes, 2) only a limited number of genes can be analyzed at any one given time, and 3) lack of sensitivity because of the limitations of hybridization-based methodology (Malone and Oliver 2011). Because of these limitations, it is estimated that microarray analysis can only detect medium to highly expressed genes, which account for only ~30% of the expressed transcripts or transcriptome (Evans et al. 2002).

The recent advent of ultra-high-throughput next generation sequencing (NGS) technology such as RNA-Sequencing (RNA-Seq) provides a great deal more sensitive method for characterizing transcriptomes (Marioni et al. 2008; Mortazavi et al. 2008; Wang et al. 2009; Nowrousian 2010) (Figure 2.2). RNA-Seq generates raw nucleotide sequence reads which are then efficiently mapped

to a corresponding reference genome, so that the overall cellular gene expression can be statistically determined (Mortazavi et al. 2008). With the introduction of various NGS data analysis platforms, it has become much easier to identify splice variants, single nucleotide polymorphisms, undescribed genes and transcripts, and predict gene fusions at a much higher sensitivity as compared to hybridization-based techniques such as conventional microarrays (Mortazavi et al. 2008; Wang and Bucan 2008; Wall et al. 2009; Trapnell et al. 2010; Edgren et al. 2011).

RNA-Seq overcomes the limitations of microarrays since 1) expressed gene sequences do not need to be known, 2) unlimited numbers of genes can be analyzed at one time, 3) sensitivity is such that it is possible to detect a single copy of an expressed transcript (Chen et al. 2011; Jiang et al. 2011). This technique enables efficient comparisons of expression levels among genes within a sample and among samples (McIntyre et al. 2011). Additionally, this technique allows a greater number of analyses to be performed including detection of alternative splice forms, gene fusions, long noncoding RNAs, single nucleotide polymorphisms (SNPs), and new genes and transcripts (Mortazavi et al. 2008; Chepelev et al. 2009; Maher et al. 2009; Wall et al. 2009; Guttman et al. 2010; Trapnell et al. 2010). This has not only opened up the opportunity to analyze in-depth allele-specific gene expression and changes in RNA editing, but also provides the broad information to thoroughly evaluate and accurately predict molecular, cellular, and functional processes (Sanchez-Pla et al. 2012).

RNA-Seq Analysis Pipeline

The general strategy for RNA-Seq is shown in Figure 2.2.

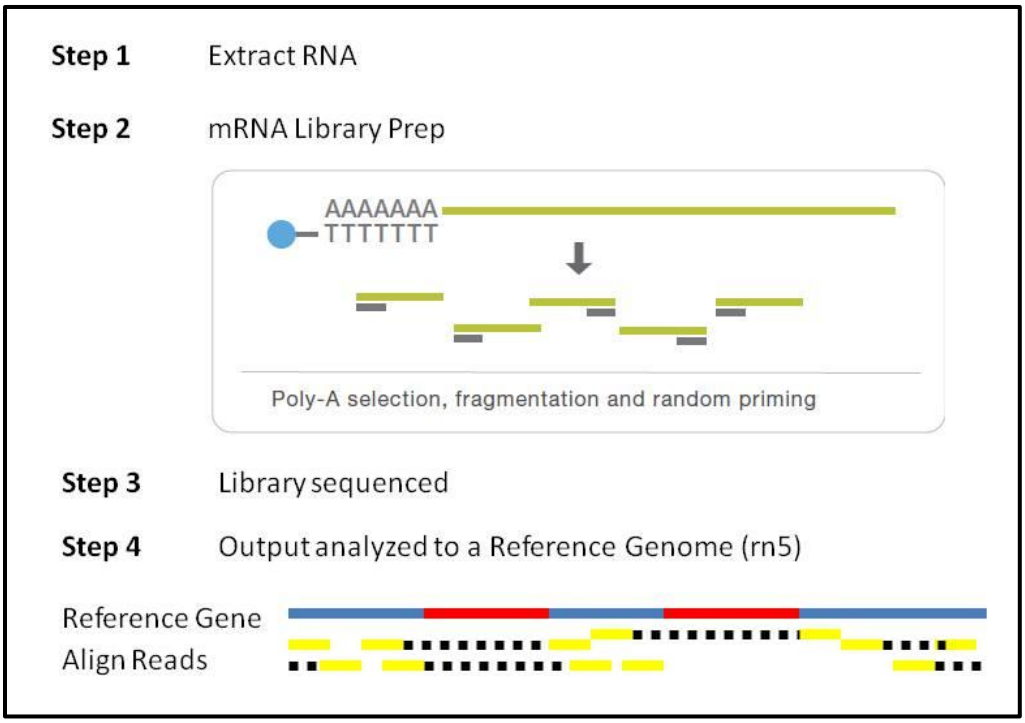


Figure 2.2. Steps in RNA-Seq pipeline used for establishing mRNA transcriptomes. Library preparation from total RNA involves poly $(A)^+$ mRNA selection, fragmentation, and random primer synthesis to make cDNA for each sample. The resulting library is sequenced and the output is analyzed by aligning the nucleotide sequence reads to a reference genome. Figure modified from protocol for TruSeq RNA™ sample preparation kit v2 (Illumina, San Diego, CA).

The first step in RNA-Seq is to isolate messenger RNA (mRNA) from a tissue or cell type of interest. This is typically done through poly $(A)^+$ selection as all mRNA have poly $(A)^+$ tails (Muller-McNicoll and Neugebauer 2013). The RNA is converted to cDNA and the cDNA is sequenced to generate a series of short nucleotide sequences or “reads”. The next step of analysis involves aligning sequences to a genome reference assembly and expression database to determine the identity of the transcripts present in the sample and calculate their

abundance (Torri et al. 2012). Post-sequence analysis of RNA-Seq data is referred to as an RNA-Seq pipeline summarized in Figure 2.3.

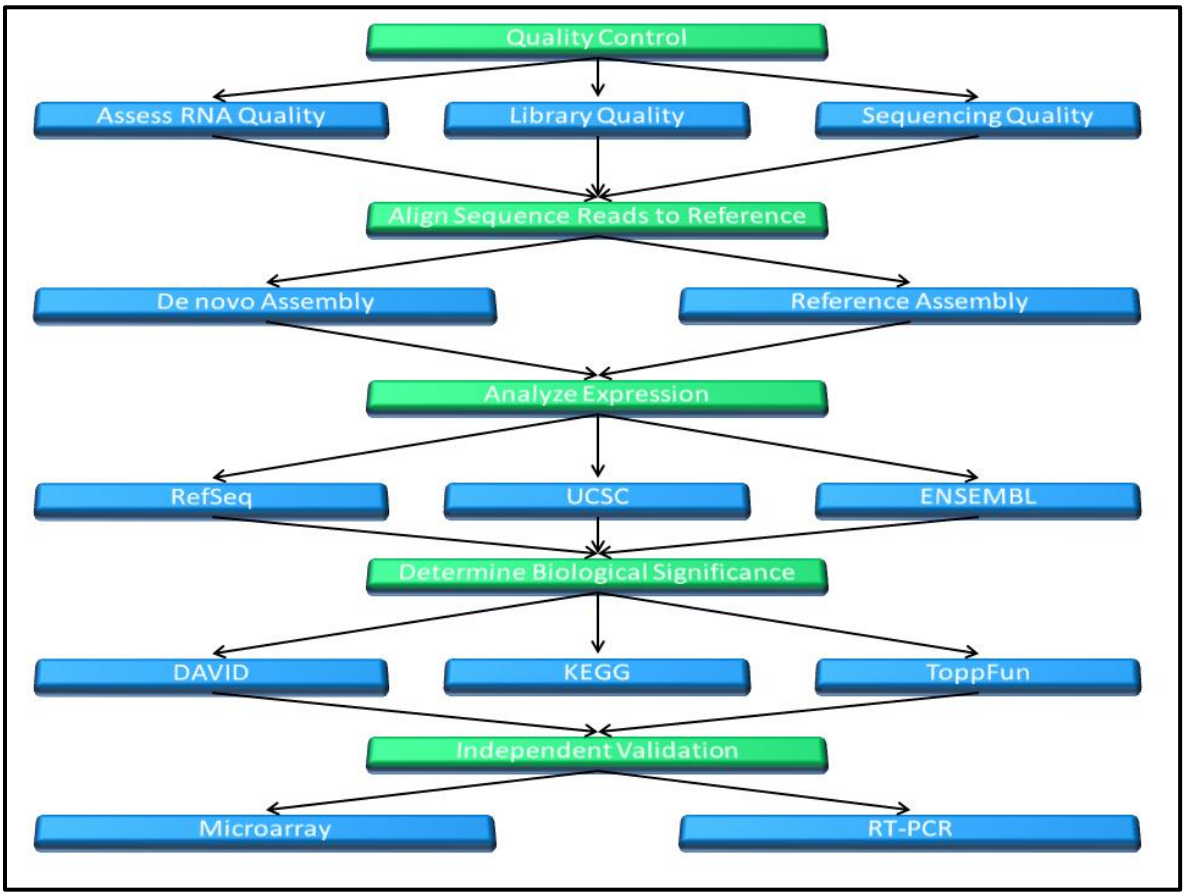


Figure 2.3. RNA-Seq Analysis Pipeline. Initial steps of RNA-Seq involve quality control steps prior to data alignment to reference genome sequences available in a variety of public databases. Once the gene expression profile is determined, bioinformatics' analysis is used to determine biological significance. Results of RNA-Seq analysis can be independently confirmed using alternative strategies such as RT-PCR and microarray analysis.

There are several layers of quality control necessary to ensure quality RNA-Seq data (Figure 2.3). The 1st quality control step is ensuring that high quality RNA is used. The definition of a "high" quality is based on the quantity of the 18S and 28S ribosomal units with minimal degradation. The quality of the RNA extracted directly relates to the quality of sequencing (Mortazavi et al. 2008). The next

step is referred to a library preparation. Prior to preparation of the library, the RNA must be fragmented by using a series of heat denaturation and cooling steps (Nagalakshmi et al. 2010). Once fragmentation has occurred, cDNA is produced and oligos of defined sequence (adapters) are ligated to the cDNA. Adapter sequences are complementary to the sequence of oligos anchored to the chip used for sequence analysis. During sequencing, each base is given a relative quality score highly dependent on the type of sequencer in order to depict sequence quality. These quality scores are based on peak intensity, shape, and resolution (Dillies et al. 2012). For example, Illumina uses Phred quality scores using the formula.

$$Q = -10 \log_{10} P$$

Where P stands for error probability and Q stands for Q score. A Q30 quality score is equivalent to the probability of an incorrect base call 1 in 1000 times or 0.01%. Using the Illumina Hi-seq 2000 next generation sequencer, in the case of the analysis reported here, the average size of the fragments was approximately 200 bp.

After sequence reads have been filtered for quality, determining where these reads align to a genomic sequence of an organism is necessary. There are two main ways to assess where these reads align within a genome; *de novo* assembly and alignment to a reference genome (Dillies et al. 2012) (Figure 2.3). *De novo* assembly uses the reads in order to assemble a new reference sequence. Issues with using RNA-Seq for *de novo* assembly are all current

algorithms such as commonly used Oases, Trinity, and trans-ABySSare are highly memory intensive and read preprocessing is absolutely necessary (Zhang et al. 2011; McGettigan 2013). There is a potential for every technique to create artifacts. In the case of RNA-Seq, these artifacts can be adapter contamination of sequence reads as well as low quality read scores. Since *de novo* assembly algorithms assemble transcripts based on overlapping sequences, any homologous sequences allow for alignment. Due to this, adapter contamination of sequence reads will produce false alignments. Therefore, trimming reads due to quality and adapter contamination from library preparation is essential in order to provide assembly of authentic reads from the organism. Alignment to a reference genome is a better solution than *de novo* assembly if the reference genome is an accurate build. In this case, a previously assembled genome is used as a reference to align sequences. This allows fewer misalignments due to sequencing artifacts (McGettigan 2013).

Once reads have been mapped to a location in a genome, it is important to assign these reads to genes and determine the relative abundance of each gene. In order to make this possible, it is important to have a high quality annotated database such as Refseq, ENSEMBL, or UCSC Genome Browser (Flicek et al. 2012; Pruitt et al. 2012; Meyer et al. 2013) (Figure 2.3). Additionally, undescribed genes and isoforms can be determined due to the unique nature of RNA-Seq. Since RNA-Seq is not a hybridization based analysis, anything being expressed can be detected (Garber et al. 2011). This type of analysis is done by comparing aligned sequences with known gene annotation. Any aligned

sequence not found in an annotation database is given the status of an undescribed gene if not “near” another gene. If “near” another gene it is given the status of undescribed isoform. The definition of “nearness” to another gene is dependent on the user’s conditions (Trapnell et al. 2010).

Determining abundance of transcript expression is an essential task. Transcript expression is typically reported as reads per kilobase per million (RPKM). This metric allows for an “apples to apples” comparison among genes independent of gene length by normalizing the number of mapped reads with the transcript length and the total number of reads from an experiment (Mortazavi et al. 2008). RPKM takes into account the total number of mapped reads for a particular transcript, the length of the transcript, and the total number of reads in an experiment. This is vital when trying to compare gene expression. This allows for transcript expression to be compared among transcripts. For example, if gene A is 100 bp long, and sequence reads are 50 bp, it is expected that if one transcript is expressed, two reads will be found. Likewise, if gene B has a length of 10,000 bp, then 200 reads would be expected for every copy of the transcript expressed. (Mortazavi et al. 2008).

After quantifying gene expression, it is then possible to assess biological functions for expressed genes. There are a wide variety of methods to assess biological function to a given set of genes. Generally speaking, databases utilizing Gene Ontology (GO) terms and association of genes with pathways and phenotypes are the hallmark across methods (Chen et al. 2011). A GO term is a term or group of terms, such as glycolysis, embryo, or enzyme, which describe

the function of a gene. By grouping genes based on thousands of terms, a network for function can be determined. In an ongoing effort (www.geneontology.org), GO terms are assigned whether manually or automatically to a given gene based on information gleaned from orthologs, documented involvement in biochemical or signaling pathways, and/or known function (Ashburner et al. 2000). Common databases using these terms are DAVID, KEGG, and Toppfun though the list is continually growing (Ogata et al. 1999; Kanehisa and Goto 2000; Chen et al. 2009; Huang da et al. 2009; Kanehisa et al. 2012) (Figure 2.3).

As with any method, analysis artifacts are possible when performing RNA-Seq therefore alternative strategies need to be used to validate the data. Typically this is done either by a microarray or reverse transcription polymerase chain reaction (RT-PCR) (Kogenaru et al. 2012) (Figure 2.3).

However, despite the many advantages of RNA-Seq over other genome wide analysis, there are some limitations. These limitations are short reads and a non-random distribution of reads. Reads for RNA-Seq can range from 50 to 400 bp depending on the instrument used (Marguerat and Bahler 2010). However, due to the demand for improvements for NGS technology, read lengths are expected to increase in the near future (Kircher and Kelso 2010). But currently due to the shortness of RNA-Seq reads, inappropriate alignment of reads can occur. For example, in cases where a pseudogene exists with significant homology to a functional gene, the programs used for analysis will not be able to determine whether the read should be assigned to the gene or the pseudogene.

(Balasubramanian et al. 2009). This results in inadequate representation of a gene's expression. However, one way to get around the shortness of these small reads is to use paired end reads. There are two types of reads for RNA-Seq; single and paired end reads. Single reads generate one single sequence read (Figure 2.4), whereas with paired end reads, a single read from each direction of a sequence is generated thus extending the length of the read (Garber et al. 2011).

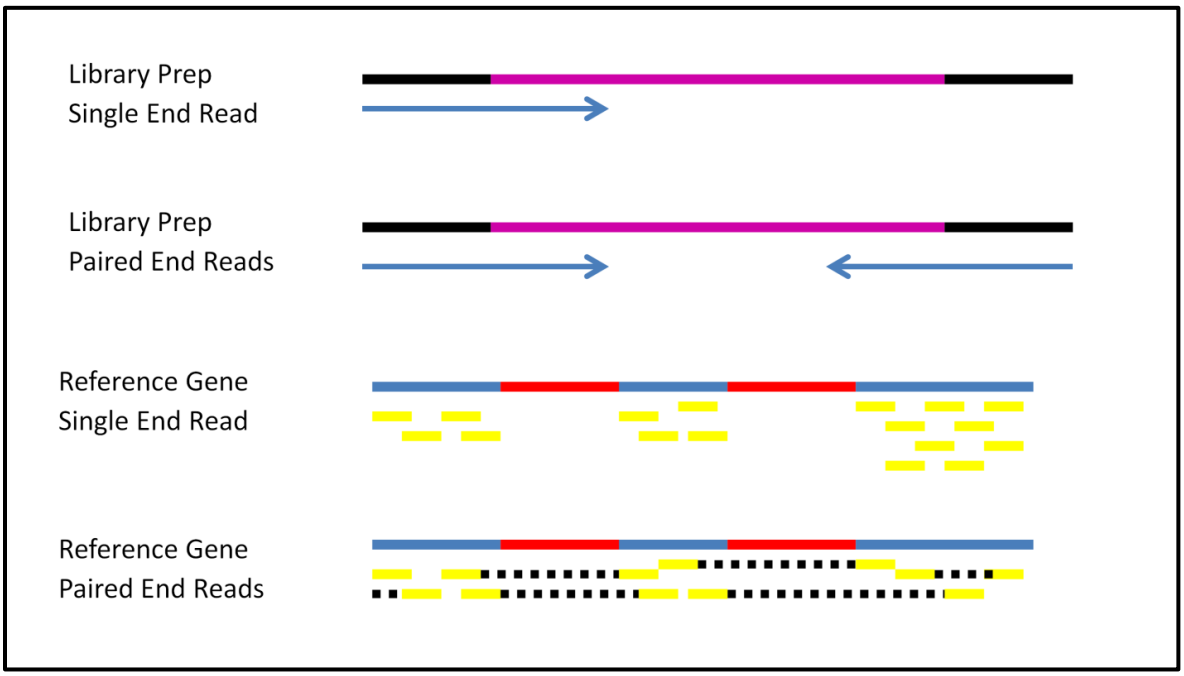


Figure 2.4. Single End Reads vs Paired End Reads. Black lines represent adapters. Purple represents cDNA. Blue arrows represent portion of a library prep sequenced. Blue lines refer to exons and red lines refer to introns of a gene. Yellow lines refer to sequence reads. Dashes among sequence reads indicate the sequences are paired. Due to this, alignment is more accurate with paired end reads as more sequence is known.

Previous studies have demonstrated that RNA-Seq reads do not follow a Guassian bell curve demonstrating a non-random distribution of reads or bias (Oshlack and Wakefield 2009; Li et al. 2010). If there was a random distribution of reads, it would be expected there is an equal number of reads across an entire

transcript. However, due to library preparation methods, transcript length, GC content, and nucleotide frequencies this is not the case (Linsen et al. 2009; Hansen et al. 2010; Li et al. 2010; Trapnell et al. 2010; Gao et al. 2011; Roberts et al. 2011).

Attempts to normalize for RNA-Seq bias are possible during library preparation and analysis. Library preparation involves mRNA pre-selection using beads that bind to the poly $(A)^+$ tail. RNA is then fragmented to generate small targets for the synthesis of cDNA. This allows for 1) cDNA synthesis of difficult RNA targets that tend to form secondary structures, and 2) generation of multiple small cDNA molecules representing all the regions within long transcripts. By improving fragmentation of RNA, a more random distribution of reads is possible (Roberts et al. 2011).

Another strategy to decrease RNA-Seq bias is through the post-sequencing analysis. One example is measurement of transcript expression based on assigning RPKM values (Mortazavi et al. 2008). RPKM normalizes for transcript length and total number of reads in order to avoid inaccurate representation of data.

However, it is reported that despite these attempts at normalizing, non-random distribution of reads still occurs. This is evident by a 5' bias of reads. Higher quality and quantity of reads tend to be present at the 5' end of reads. This is hypothesized to be due to PCR in the RNA-Seq protocol. In order to circumvent PCR bias, a technology needs to be developed that eliminates the PCR step.

(Hansen et al. 2010; Zheng et al. 2011; Wagner et al. 2012; Trapnell et al. 2013).

Current State of ESC Transcriptomes

Understanding of the exact regulatory mechanisms of the ESC state remains to be fully understood even though great advances have occurred (Chickarmane et al. 2012). To this point, understanding the ESC state has predominantly been determined based on studies involving hESC and mESC. Previous studies have only explored the rat ESC state by examining pluripotency markers across several strains (Buehr et al. 2008; Li et al. 2008; Hirabayashi et al. 2010a; Hirabayashi et al. 2010b; Tong et al. 2011). hESC and mESC each have their own unique cell membrane markers (Calloni et al. 2013). It has been assumed that rESCs are like mESCs because they share the same cell markers. (Buehr et al. 2008; Li et al. 2008; Hirabayashi et al. 2010a; Hirabayashi et al. 2010b; Tong et al. 2011). Whether or not rESC are truly similar to mESC and the degree of similarity or dissimilarity to hESC still remains to be determined.

Current knowledge of the ESC state in both mESC and hESC relies on the core interaction of 3 transcription factors; *Oct4*, *Sox2*, and *Nanog* (Mitsui et al. 2003; Boyer et al. 2005; Niwa 2007). Previous studies have shown that knocking out or knocking down of *Oct4*, *Sox2*, or *Nanog* results in an inability to maintain pluripotent cells (Nichols et al. 1998; Avilion et al. 2003; Buckler et al. 2009). However, Nanog can be removed once pluripotent cells have been established (Mitsui et al. 2003). Additionally, these three transcription factors are found to

co-localize to numerous genomic sites to activate or silence gene expression in order to maintain pluripotency (Richards et al. 2004; Assou et al. 2007; Zhou et al. 2007; Cloonan et al. 2008; Rosenkranz et al. 2008; Tang et al. 2010). Based on this data, it is believed currently that maintenance of the ESC state is achieved by expression of these three transcription factors which 1) activate other pluripotency factors and repress lineage-specific genes, and 2) activate their own gene expression (Young 2011).

In conclusion, there have been no comprehensive studies to characterize the rat ESC state or transcriptome, therefore the studies described here were undertaken. The goals of this study were to 1) characterize gene expression in rESCs, and 2) to compare the rESC transcriptome to the human and mouse ESC transcriptomes to gain insight into ESC expression patterns across species.

CHAPTER III

MATERIALS AND METHODS

Cell Line

The DAc8 (RRRC#464 DA-EC8/Rrrc cell line) was obtained through the Rat Resource and Research Center (http://www.rrrc.us) and was previously demonstrated to be an authentic rat embryonic stem cell line (Tong et al. 2010). Three vials of 1 x 10^6 cells at passage 27 were thawed from liquid nitrogen then pooled and plated onto two, 60 mm plates with 2i media on mitotically inactive mouse embryonic fibroblasts as described previously (Tong et al. 2011). The following protocol is available at http://www.rrrc.us.

Three vials of 1 x 10^6 cells at passage 27 were thawed from liquid nitrogen by agitating the vial in a 37 °C water bath. Using a 5 mL serological pipette, 3 mLs of pre-warmed 2i media (Appendix A) was aseptically added to the rESCs and the resuspended cells were transferred into a 15 mL conical tube. Another 2 mLs of pre-warmed 2i media was used to rinse the vial and then added to the conical tube. Rat ESCs were resuspended by slowly pipetting up and down until cloudy. The cells were pelleted by centrifugation at 150 g for 5 min and the supernatant removed and discarded. The cells were resuspended in the appropriate volume (5 mL for 60 mm, 10 mL for 100 mm) of pre-warmed 2i media and pooled all three vials. Cells were plated onto 2, 60 mm cell culture plates (Fisher, Asheville, NC, BD Falcon 353002) with previously plated mouse

embryonic fibroblast (MEF) (Millipore, Billerica, MA PMEF-N) (see following section for MEF preparation). A hemocytometer was used to verify that 1.5 x 10^6 million cells were plated for each plate to ensure starting with identical number of cells. Plates were transferred to an incubator at 37 °C with 5% $C0_2$ and 90-100% humidity.

The media was changed every 48 hrs or sooner if the media was yellow. The media has a pH indicator, which turns yellow when acidic. Rat ESCs can start differentiating within a couple of hours if the media turns yellow, so media must be changed immediately once it begins to yellow. If rESCs colonies are at a high density (~70% confluency), then plating with 20 mLs (on 100 mm) of 2i media will circumvent this issue. Rat ESCs colonies can be at a high density, but not high enough to passage at the ~48 hour time point. Rat ESCs were passaged at 4 days after thawing from cryopreservation and every 3 days afterward. Rat ESCs were passaged by removal of all the 2i media and placing in a 50 mL conical tube. Two mL of the removed 2i media was picked up by a 1000 uL pipette tip and gently washed over the plate. When done correctly, rESCs will come off the MEF layer, but the MEFs will remain attached to the plate. Rat ESC colonies we centrifuged at 150 for 5 min and the supernatant was removed and discarded. Colonies were resuspended in 5 mLs of pre-warmed TrypLE (Invitrogen, Grand Island, NY A1217701) and incubated at 37 °C for 5 minutes or longer. Since rESC grow as colonies, they need to be disassociated from each other in order to create a one cell suspension. TrypLE is used as it is a gentle means of disassociating these rESC colonies as it digests what binds the cells together.

The length of time for incubation is increased (in 3 minutes intervals) until the colonies are disassociated into a one to four cell suspension. Rat ESCs grow in colonies, so one to four cells suspension is vital to ensure expansion due to one to four cells will give rise to 1 colony. If the rESCs were not appropriately disassociated then expansion of rESCs would not be as robust.

Rat ESCs were plated at a 1:3 ratio and transferred to an incubator at 37 °C with 5% $C0_2$ and 90-100% humidity. From thawing from liquid nitrogen to RNA extraction rESC expansion took a total of 16 days and resulted in 5, 100 mm plates per sample ($15x10^6$ cells) in order to have enough RNA for analysis.

Mouse Embryonic Fibroblast (MEF) Cell Culture

One day prior to bringing rESCs out of liquid nitrogen, MEFs (Millipore, Billerica, MA PMEF-N) are thawed and prepared by the same method as rESCs except MEF media (Appendix A) is used instead of 2i media. The following changes are implemented for MEF culture. 1 vial (5-6x10^6 cells) of MEFs is sufficient for 3, 100 mm or 8, 60 mm plates. MEF cells were maintained by changing the media every 48 hrs.

RNA Extraction

At passage 30, rat ESC colonies were gently detached from the mouse embryonic fibroblast layer. To reduce possible fibroblast and dead cell contamination, cells were allowed to sit on ice for 10 minutes which resulted in >90% of rat embryonic stem cell colonies settling to the bottom, while fibroblasts

and dead cells tended to float Cells were washed with 1xPBS (Invitrogen, Grand Island, NY 14190-136) by slowly pipetting PBS over cells and centrifuging for 5 minutes at 800 g. This was done twice. RNA was extracted using Trizol (Invitrogen, Grand Island, NY 15596-026) followed by further purification using a Qiagen RNeasy kit (Valencia, CA 74106) according to manufacturers' instructions. Briefly, 2.25 mL of Trizol reagent was added per 0.75 mL of samples (~20-30x10^6 cells) then lysed by pipetting up and down several times. The homogenized sample was incubated at room temperature (RT) for 5 minutes after which 0.45 mL of chloroform was added. The sample was shaken vigorously for 15 seconds, incubated at RT for 2-3 minutes, and centrifuged at 12,000 g for 15 minutes at 4 °C. The aqueous phase is pipetted into a new tube. RNA was precipitated by adding 1.2 mL of 100% isopropanol and incubated at RT for 10 minutes. Samples were centrifuged for 10 minutes at 4°C. The supernatant was removed and the pellet was washed with 2.25 mL of 75% ethanol. Samples were gently shaken to break up the pellet and centrifuged at 7500 g for 5 minutes at 4°C. The pellet was air dried for 8 minutes then resuspended in 100 µL DEPC treated water by incubating at 55°C until resuspended (~10 min). Additional purification was performed using Qiagen RNeasy (Valencia, CA 74104) kit using the RNA cleanup protocol. Briefly, 350 µL of Buffer RLT was added and the sample was mixed. 250 µL of ethanol was added then mixed by pipetting. Samples were added to an RNeasy Mini spin column and centrifuged for 15 seconds at 8000 g. The flow-through was discarded, and then 500 µL of Buffer RPE was added to the column and

centrifuged for 15 seconds at 8000 g. The flow-through was discarded, 500 µL of Buffer RPE was added to the column and the samples were centrifuged for 15 seconds at 8000 g. The flow-through was discarded, and then the samples were centrifuge for 2 minutes at 8000 g. 50 µL of RNase free water was added to the column; the samples were incubated at RT for 1 minute then centrifuged at 8000 g for 1 min. This was done twice. Samples were stored in 10 µL aliquots in 0.5 mL tubes at -80°C until further use.

RNA-Seq

RNA integrity from extracted total RNA was determined using a RNA nano chip on an Agilent 2100 Bioanalyzer to determine a RNA integrity number (RIN). Samples 1 and 2 had a RIN of 9.8 and 9.6 respectively. Approximately, 3.5 ng of total RNA was used for Illumina sequencing compatible library preparation using TruSeq RNA^{TM} sample preparation kit v2 (Illumina, San Diego, CA) according to the manufacturer's protocol. Briefly, input RNA was purified by two rounds of poly $(A)^+$ selection followed by chemical fragmentation. Random hexameric primers were used to generate cDNA from fragmented and primed RNA from the previous step using Superscript II reverse transcriptase. The cDNA was purified using Ampure XP beads and end repair was performed using a 3'-5' exonuclease enzyme, which removed the 3' overhang and filled the 5' overhang thereby producing blunt ends on either sides of the ds cDNA. Indexed adapter ligation was performed by adenylating the 3' end of this blunt cDNA to provide a complementary overhang to the corresponding 'T' nucleotide on the 3' end of the

adapters. PCR amplification (15 cycles) was used to enrich the adapter ligated ds cDNA molecules to generate single read libraries.

The final concentrations of the libraries were evaluated using a DNA 1000 chip on an Agilent 2100 Bioanalyzer. 5µl of each amplified library was diluted to 15nM stock in 1% Tween, and 2µl of this stock was used for quantification using a KAPA SYBR® Fast Universal qPCR kit (Kapa Biosystems, Inc., Woburn, MA). Stock libraries were diluted to a final 10nm concentration for cluster generation on a cBot v1.4.36.0 using Illumina's Truseq PE Cluster Kit v3.0. Massive parallel paired-end (PE) sequencing was performed using a 200 cycle TruSeq SBS HS v3 kit on a HiSeq2000, running HiSeq Control Software (HCS) v1.4.8. The clustered flowcell was sequenced for 106 cycles, broken down into 3 separate reads. The first read was 50 cycles in length, followed by a 6 cycle index read. Following the index read, paired end resynthesis was performed using Truseq PE Cluster Kit v3.0, which was then followed by another 50 cycles. Image analysis and base calling were performed using the standard Illumina Pipeline consisting of Real time Analysis (RTA) version v1.12.4.2 and Casava v1.8 using the default settings.

Post-Sequence Analysis

Raw reads were aligned against the rat genome RGSC 5.0/rn5 using the default settings for TopHat (v1.4.0) resulting in > 95% of reads mapping to the rat genome assembly (Gibbs et al. 2004; Havlak et al. 2004).

The bam files generated were uploaded onto a commercially available platform Avadis NGS (v 1.3) (Strand Scientific Intelligence, Inc.) for further downstream data analysis and alignment. Gene alignment was conducted against RefSeq (10/21/2012) (http://www.ncbi.nlm.nih.gov/refseq/). Duplicate reads as well as reads with low mapping quality were removed. The overall abundance of expressed genes was calculated as RPKM, Reads Per Kilobase of exon model per Million mapped reads as described earlier by Mortazavi et. al., 2008. Additionally, adapter contamination was evaluated and considered not to be a significant variable. The significance of adapter contamination was evaluated by using a script that takes the adapter sequence and determines the percentage of the reads that contain a portion of the adapter sequence. None of the samples had a greater then random distribution of the adapter sequence. Furthermore, when taking the adapter sequence and using BLAST against the rat genome there were >100 hits that matched 100% to genes.

Ortholog Processing

Output files from Avadis analysis were analyzed with g-profiler to determine the orthologs for each species (Reimand et al. 2011). Gene symbols and ENSEMBL id were taken into account when determining orthologs as using both gives a more concise ortholog list. A gene needed at least one ortholog in another species in order to be considered for downstream analysis. Typically, genes that did not have orthologs were uncharacterized proteins.

Gene Ontology and Pathway Analysis

Toppfun was used to analyze all Gene Ontology (GO) terms and perform pathway analysis in order to identify the most relevant biological term associated with a given gene list. Toppfun is a program for gene list enrichment analysis based on functional annotation and protein interactions network (Chen et al. 2009). For pathway cluster analysis, Toppcluster and Cytoscape were used to generate cluster maps (Kaimal et al. 2010; Praneenararat et al. 2012). All statistical analysis was conducted using Bonferonni correction value at cut off of <0.05 p-value. Due to the quantity of genes used in this analysis, it is typically considered good statistical practice to use a correction to reduce false positives. By using the Bonferonni correction value, the cutoff of the p-value is essentially changed to the threshold p-value of 0.000005.

RT-PCR

Total RNA was extracted from DAc8 samples as described previously. cDNA synthesis was performed using Invitrogen Superscript III First-Strand cDNA synthesis following the manufacturer's protocol. Briefly, 2 µg of RNA per 20 µL reaction was set up on ice with 2 µM random primers, 200 U of SuperScript III reverse transcriptase, and 40 U of RNase inhibitor using the following thermal conditions: 25 C° for 10 min, 50 C° for 50 min, and 85 C° for 5 min. 1 µL of cDNA reaction was used for each PCR reaction with a specified primer set (Appendix C) and Faststart *Taq* (Roche, Indianapolis, IN) following the manufacturer's protocol. Thermocycler conditions were 1 cycle of 95 C° for 3

min; 35 cycles of 94 C° for 30 seconds, 57 C° for 30 seconds, and 72 C° for either 30 seconds (*Nras* and *Dnmt3b*) or 1 min (*Lef1*). Amplicons were analyzed on 3% (*Nras* and *Dnmt3b*) or 1% (*Lef1*) 1X TBE agarose gels by electrophoresis. Amplicons were gel purified using Qiagen's gel extraction kit following the manufacturer's protocol. Amplicons were submitted to the University of Missouri, DNA Core for fragment analysis.

Nucleotide Sequencing

RT-PCR products were gel purified using Qiagen QIAuick Gel Extraction kit (Valencia, CA, 28704) using the manufacturer's protocol. Briefly, DNA fragments were excised with from the agarose gel using a scalpel. Three volumes of Buffer QG to 1 volume of gel were incubated at 50 C° until dissolved. Dissolved samples were applied to a QIAquick column and centrifuged at 15,000 g for 1 min. Flow-through was discarded and the samples were washed again with 0.5 mL of Buffer QG and 0.75 mL of Buffer PE followed by an additional spin for 1 min to remove residual ethanol. DNA was eluted by incubating 30 µL of Buffer EB for 1 min on the column at RT then centrifuging at 15,000 g for 1 min. Quality was assessed using a Nanodrop 8000 (Thermo Scientific, Wilmington, DE, for a 260/280 value of 1.80-2.00. 1500 ng of DNA was used for sequencing on a 3730xl 96-capillary DNA Analyzer with Applied Biosystems Big Dye Terminator cycle. All nucleotide sequence analysis was performed at the DNA Core (University of Missouri).

Data Access

The data discussed in this publication have been deposited in NCBI's Gene Expression Omnibus (Edgar et al. 2002) and are accessible through GEO Series accession number GSE44150 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE44150)

CHAPTER IV

RESULTS

To perform the analysis, a rat embryonic stem cell line (DAc8) derived from the Dark Agouti (DA) inbred rat strain was chosen. This male cell line is known to be germline competent and it has been used successfully for genetic manipulations, therefore, while characterization is minimal, it is one of only a few characterized rESC lines available and was a logical choice for generating the first rat ESC transcriptome. cDNA libraries from poly $(A)^+$ mRNA from two biological replicates of the DAc8 cell line were deep sequenced at 50 base paired-end reads in order to generate a data set of expressed genes. Sequencing generated a sum of more than 247 million reads for both samples. Reads were filtered to remove low quality and duplicate reads (Table 4.1). Adapter contamination was not filtered as it was determined not to be significant (see Materials and Methods). The reads were aligned against rat genome RGSC 5.0/rn5 using TopHat (v1.4.0) with > 95% of reads mapping to the rat genome assembly (Gibbs et al. 2004; Havlak et al. 2004). Expressed genes were calculated as reads per kilobase of exon model per million mapped reads (RPKM) as described previously by Mortazavi et. al, 2008. RPKM values estimated from the 2 biological replicates used for the analysis were plotted on a scatter plot, which indicated that gene expression was highly consistent across both samples (Appendix B). A total of 10,931 genes were detected based on a value of at least one RPKM per gene in both samples with a mean of 39.6 RPKM and a range from 1 to 2287 RPKM. When comparing

both biological replicates, only 34 genes showed expression differences of greater than 1 fold log change among the samples. 4 of these 34 genes were chosen for verification (Figure 4.1). The original analysis used the reference genome of rat genome RGSC 4.0/rn4. This resulted in only 4 genes that showed a difference in expression. Upon reanalysis using rn5 as a reference genome, the gene number was increased to 34. Since, these 4 genes confirmed the predicted biological variance among the samples, it was determined not necessary to confirm the other 30 genes.

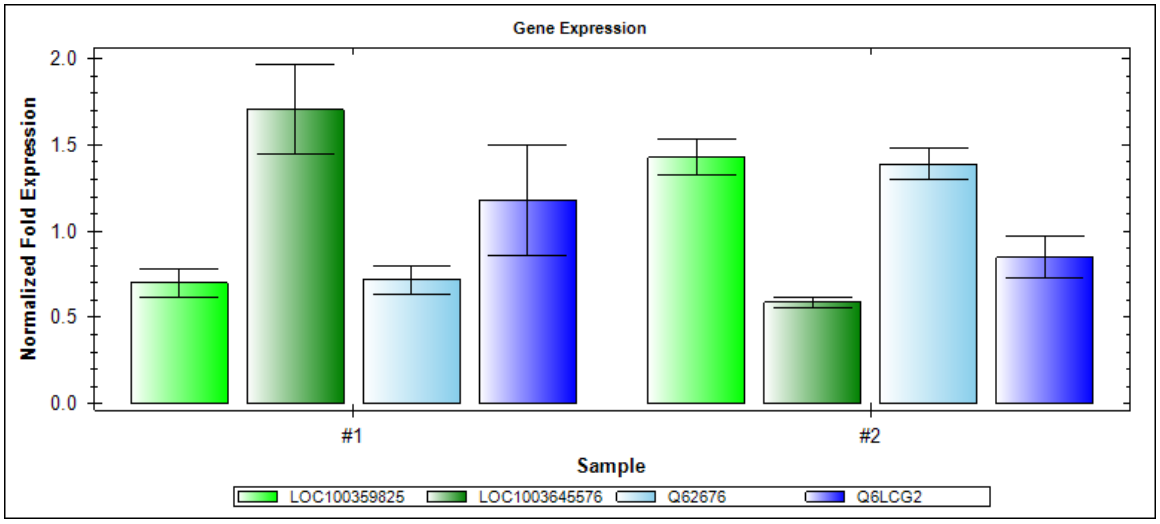


Figure 4.1 Verification of biological variance among samples. RNA-seq predicted 34 genes with a 1 log fold expression difference among sample 1 and 2. To verify the difference, expression for 4 genes (*LOC100359825*, *LOC1003645576*, *Q62676*, and *Q6LCG2*) was determined among sample 1 and 2 using quantitative RT-PCR. Fold change was determined based on ΔΔCt expression normalized to *B2m*. Lines above bar graph represent the standard deviation for each gene. Predicted biological variance among samples was confirmed for the genes tested.

Table 4.1 Number of filtered reads per species

Species	Genes Detected	Average Number of Reads*	Sequencer	Reference
rESC	10,931	94 million	Illumina Hi Seq 2000	This study
hESC	17,634	12 million	Illumina Genome Analyzer (GAI or GAIIx)	(Birney et al. 2007)
mESC	14,417	67 million	Illumina Genome Analyzer (GAI or GAIIx)	(Guttman et al. 2010)

***There are 2 biological replicates each for mouse and rat and 4 for human.

Gene Expression in rESCs

A total of 10,931 genes were detected based on at least one RPKM per gene. The 25 most highly expressed genes in the DAc2 rat ESCs are genes that are known to be involved with the glycolysis pathway, glucose regulation of insulin secretion, and genes involved in energy metabolism (Table 4.2).

Table 4.2 Twenty-five most highly expressed rat embryonic stem cell genes.

Gene Symbol	Protein Name	RPKM
Aldoa	Fructose-bisphosphate aldolase A	2286
Eef2	Elongation factor 2	2035
Gstp1	Glutathione S-transferase P	1478
ATP5B	ATP synthase subunit beta, mitochondrial	1424
Gpi	Glucose-6-phosphate isomerase	1343
Pgk1	Phosphoglycerate kinase 1	1291
Hsp90ab1	Heat shock protein HSP 90-beta	1268
Pgam1	Phosphoglycerate mutase 1	1223
Gnb2l1	Guanine nucleotide-binding protein subunit beta-2-like 1	1062
Slc2a1	Solute carrier family 2, facilitated glucose transporter member	1048
Pabpc1	Polyadenylate-binding protein 1	1046
Tubb5	Tubulin beta-5 chain	959
Bsg	Basigin	943
Trim28	Transcription intermediary factor 1-beta	937
Serbp1	Serpine1 mRNA binding protein 1	885
Pdpn	Podoplanin	861
Scd	Acyl-CoA desaturase 2	833
Cct5	T-complex protein 1 subunit epsilon	785
Akr1a1	Alcohol dehydrogenase [NADP+]	777
Actb	Actin, beta	719
Fn1	Fibronectin Anastellin	717
Ywhae	Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, epsilon polypeptide	658
Igfbp2	Insulin-like growth factor-binding protein 2	655
Utf1	Undifferentiated embryonic cell transcription factor 1	654
Mlf2	myeloid leukemia factor 2	652
Oaz1	ornithine decarboxylase antizyme 1	632

These genes were analyzed further using ToppFun to determine statistically significant pathways, phenotypes, and Gene Ontology (GO) terms for biological, molecular, and cellular components (Chen et al. 2009). A statistically significant

pathway, phenotype, and GO term are defined by the percentage of expressed genes out of the total number of genes associated with a certain function.

The top 5 statistically significant GO terms for cellular components are mitotic cell cycle, cellular response to stress, intracellular transport, cell cycle process, and cell cycle (Figure 4.2). The top 5 statistically significant GO terms for biological function are phosphotransferase activity, kinase activity, enzyme binding, transferase activity, and RNA binding (Figure 4.2). The top 5 statistically significant for molecular function GO terms are microtubule cytoskeleton, envelope, organelle envelope, ribonucleoprotein complex, and mitochondrial part (Figure 4.2).

The top 5 statistically significant phenotypes include genes involved in embryonic lethality, embryogenesis/development, the embryogenesis phenotype, prenatal growth, and embryonic growth. Due to phenotype similarity it should be noted that many genes are shared in common.

The top 5 statistically significant pathways not including the cell cycle are genes involved in the diabetes pathway, influenza, HIV infection, signaling by neuronal growth factors, and insulin synthesis and secretion. Pathways consist of groups of associated genes and pathway names often reflect the context in which the pathway was first identified and do not necessarily indicate global biological relevance.

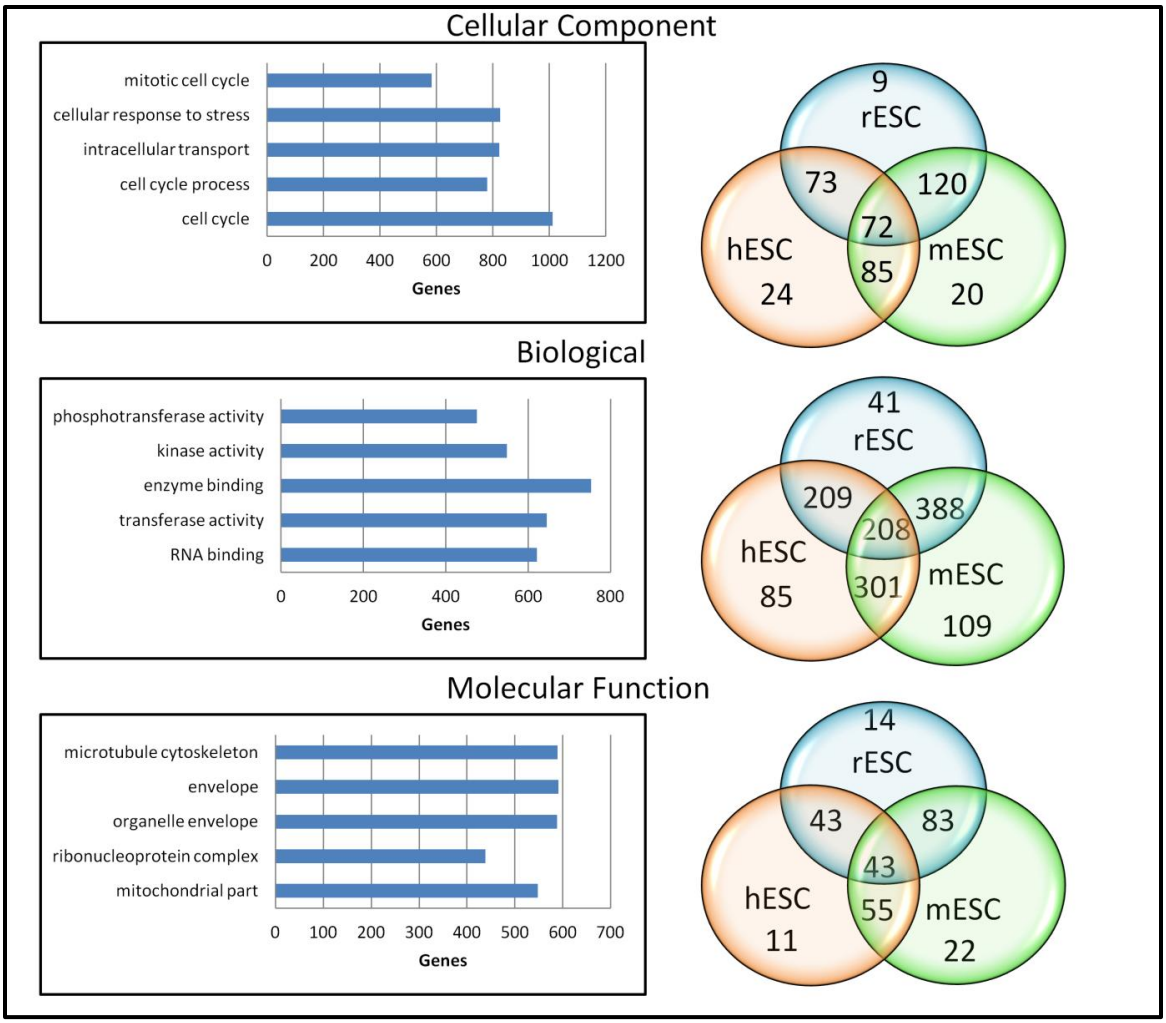


Figure 4.2 GO terms for rat ESCs. The bar graphs on the left depict the top 5 GO terms that correspond to the highest statistical significant expression in rat ESCs for each category (cellular, biological, and molecular). The Venn diagrams on the right include comparisons for each category of GO terms for all three species.

Undescribed Isoforms

Reads obtained by RNA-seq were aligned to known and predicted rat RNA sequences from RefSeq (/refseq/release/10/20/2012). Reads that mapped to undescribed exon-exon junctions indicate the potential to be undescribed isoforms. To verify the predicted undescribed isoforms have not been previously described ENSEMBL, UCSC, and NCBI were consulted. After this analysis was

completed, 27 genes were predicted to have the potential for undescribed isoforms (Table 4.3).

Table 4.3. Genes with predicted undescribed isoforms in rat ESCs.

Gene Symbol	Protein Name
Ap3b1	AP-3 complex subunit beta-1
Hmgcs1	Hydroxymethylglutaryl-CoA synthase, cytoplasmic
Nras	GTPase NRas
Lef1	Lymphoid enhancer-binding factor 1
Sptan1	Spectrin alpha chain, brain
Prrc2b	proline-rich coiled-coil 2B
Rif1	Telomere-associated protein RIF1
Dync1i2	Cytoplasmic dynein 1 intermediate chain 2
Dnmt3b	DNA (cytosine-5)-methyltransferase 3B
Tpd52l2	Tumor protein D54
Luc7l2	LUC7-like 2
Dctn1	Dynactin subunit 1
Clta	Clathrin light chain A
Mta1	Metastasis-associated protein MTA1
Azin1	Antizyme inhibitor 1
Ncaph2	Condensin-2 complex subunit H2
Ubp1	upstream-binding protein 1
Fn1	Fibronectin Anastellin
Nmral1	NmrA-like family domain-containing protein 1
Lig3	DNA ligase 3
Cisd3	CDGSH iron sulfur domain-containing protein 3, mitochondrial
Eif4h	Eukaryotic translation initiation factor 4H
Rsrc2	Arginine/serine-rich coiled-coil protein 2
RGD1304704	Uncharacterized Protein
RGD1304704	Uncharacterized Protein
Ubqln1	Ubiquilin-1
Dbn1	Drebrin

To test the accuracy of the RNA-seq-based predictions of the undescribed isoforms, three genes (*Nras, Dnmt3b*, and *Lef1*) were chosen for further validation by RT-PCR and nucleotide sequence analysis (Figure 4.3). These genes were chosen due to their biological significance for ESC biology.

Additionally, the homologues of these genes were compared to human and mouse. For every gene, human and mouse had a higher number of isoforms known.

The *Nras* gene is involved in the MEK/ERK signaling pathway and only one known rat isoform has been described (Gyorffy and Schafer 2010). The RNA-seq data predicted an undescribed isoform in which exon 6 was skipped. Primers were designed within exons 5 and 7 and based on RT-PCR and nucleotide sequence analysis, three amplicons were detected, including the full length isoform and the predicted isoform in which exon 6 was skipped. The third amplicon represented an artifact PCR product involving mispriming of the reverse primer to a duplicated sequence located in exon 11 of the *Nras* pseudogene (Figure 4.3).

Dnmt3b is implicated in *de-novo* DNA methylation (Jin et al. 2013). The *Dnmt3b* gene has 24 exons and 3 isoforms have been identified in the rat. Our RNA-seq-derived data set confirmed that all 3 previously described *Dnmt3b* isoforms are expressed in rESCs and predicted an additional isoform lacking exon 22. To verify the existence of the undescribed isoform, RT-PCR analysis with primers designed within exons 21 and 24 was performed followed by nucleotide sequence analysis of all resulting amplicons. This analysis, in combination with RT-PCR performed with additional primer sets within other exons (data not shown) confirmed that all 4 isoforms identified by RNA-seq analysis were present in the rESCs (Figure 4.3). Interestingly, the undescribed ESC isoform was not present in the control kidney sample and may represent an ESC-specific isoform.

Lef1 is involved in the Wnt signaling pathway which plays a role in determining cell lineage (Mao and Byers 2011). *Lef1* has a total of 11 exons and six isoforms have been described. The RNA-seq data predicted an additional isoform which skips exons 2, 3, and 6. Primers were designed to enable amplification of the entire gene from exon 1 to exon 11 by RT-PCR analysis. Based on RT-PCR and nucleotide sequence analysis of the resulting amplicons, three isoforms were detected: the full length isoform, one in which exon 6 was skipped, and one in which exons 2, 3, and 6 were skipped (Figure 4.3). The other 4 known isoforms differ by only a few base pairs and could not be identified effectively using this assay. However, the assay did confirm the presence of the undescribed isoform as intended.

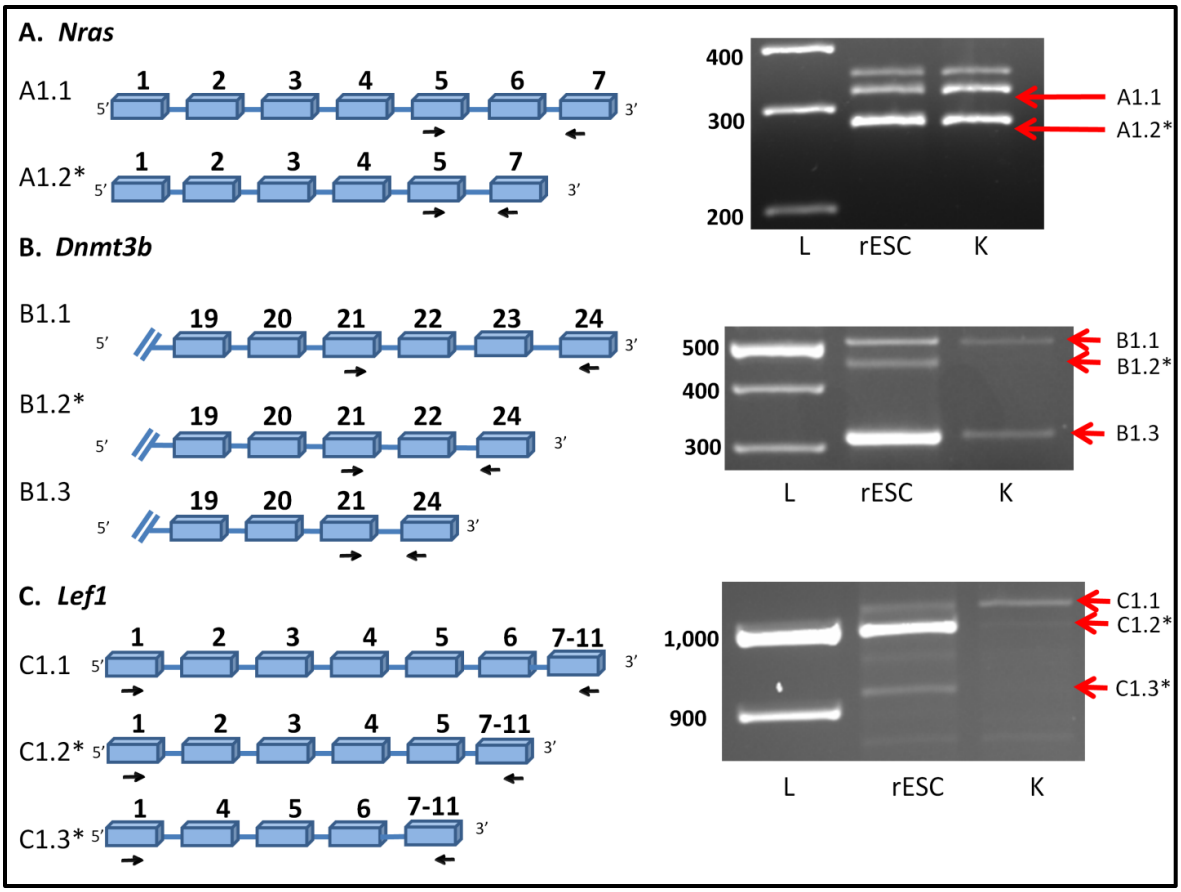


Figure 4.3. *Nras*, *Dnmt3b*, *Lef1* isoform confirmation. The schematics on the left represent the coding region of each isoform. Blue boxes represent exons. Exons are not drawn to scale. Arrows indicate location of primers (Appendix C). Gel images to the right represent RT-PCR results. For each RT-PCR gel image: L = DNA Ladder (with size in base pairs indicated to the left of the image), rESC = rat embryonic stem cell sample, K = adult rat kidney control, and * stands for predicted undescribed isoform. **(A)** *Nras,* **(B)** *Dnmt3b*, **(C)** *Lef1*.

Undescribed Poly (A)$^+$ Transcripts

Reads aligned to the rn5 reference genome, but lacking any known or predicted transcript according to RefSeq (/refseq/release/10/20/2012) were assigned undescribed poly (A)$^+$ transcript status. To verify the undescribed poly (A)$^+$ transcript have not been previously described, ENSEMBL, UCSC, and NCBI databases were consulted. In total, 133 undescribed poly (A)$^+$ transcripts were predicted.

To verify this prediction, 2 undescribed poly $(A)^+$ transcripts (71 & 80) were predicted to have 2 exons each was chosen for RT-PCR verification. Neither of these predicted transcripts had homology to any known transcript. The nucleotide sequences of these predicted transcripts were used for a BLASTN analysis using the RefSeq database. The closest match for 71 was a 79% identity match for the gene *C1p2a-like* from *Apis flora*. The closest match for 80 was a 85% identity to an uncharacterized rat gene *LOC100911535*. Of note, *LOC100911535* was not detected in the rESC samples.

Based on this analysis, primers were designed to amplify each "exon" of each transcript. Genomic DNA was chosen as a control to demonstrate the assay was possible. However, the conditions are not optimized for genomic DNA for the putative full length transcript due to the difference in size. Of the four exons tested, only exon 2 for predicted poly $(A)^+$ 80 was confirmed (Figure 4.4).

It is unclear why there was lack of confirmation for the 4 "exons" chosen for confirmation. A possible explanation could be the lack of sensitivity for the assay. Taking into account the numbers of reads and gene length for 71 and 80, only 2 and 14 full length transcripts were present in the data set, respectively.

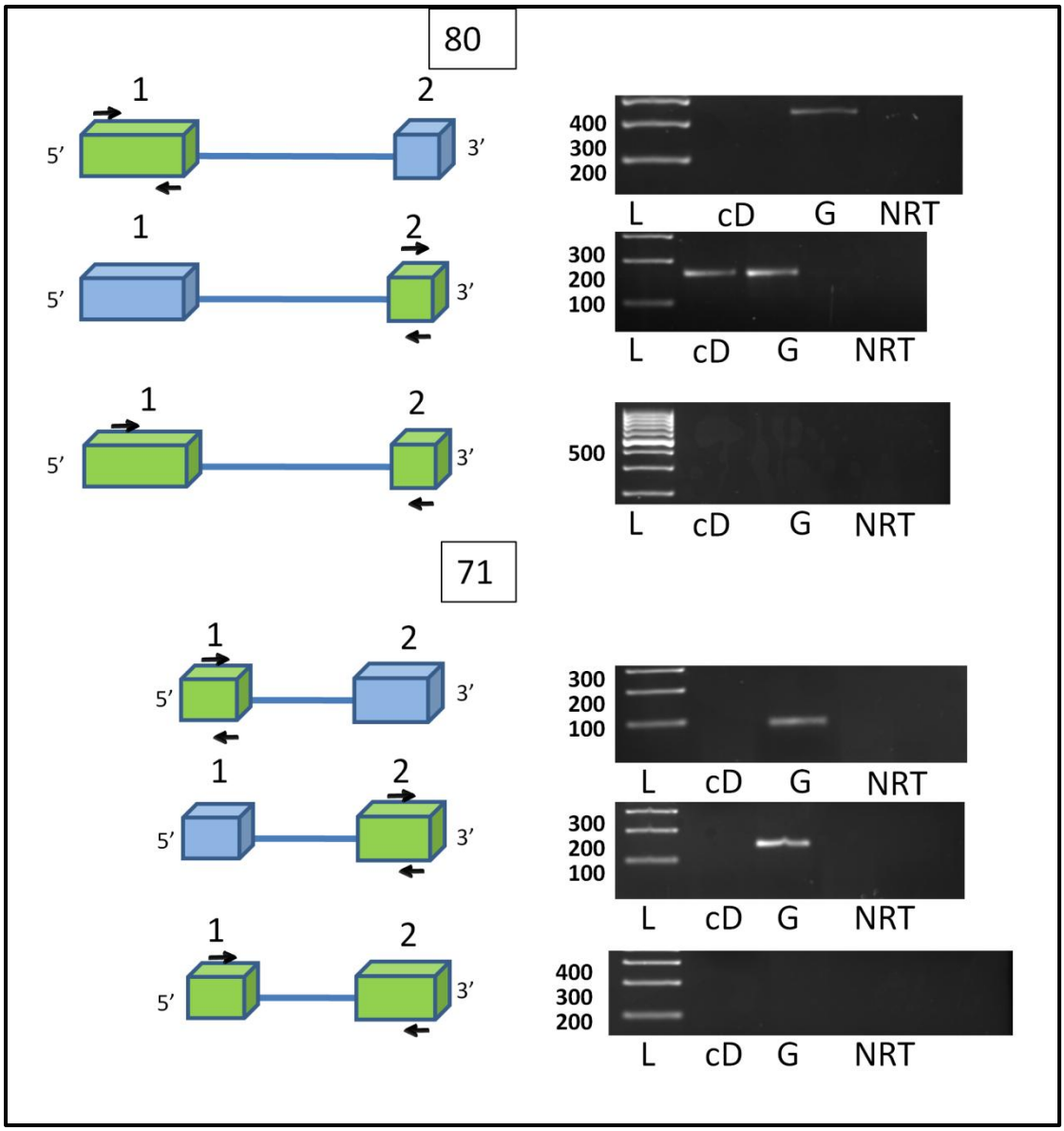


Figure 4.4. **Undescribed predicted poly (A)$^+$ transcript confirmation.** The diagram on the left represent the predicted undescribed poly A transcripts for 80 and 71. Boxes represent exons. Arrows indicate location of primers (Appendix C). For each RT-PCR gel image: L = DNA ladder (with size in base pairs indicated to the left of the image), cD = cDNA for rESC, G = genomic DNA for rESC (positive control), NRT = no reverse transcriptase (negative control).

Rat, Human, and Mouse ESC Paired End RNA-seq Comparison

The rESC data set was compared to publicly available paired end RNA-seq data sets for a human ESC line (GSM958733) and a mouse ESC line (GSM521650) (Birney et al. 2007; Guttman et al. 2010). All post sequence analyses were

conducted in the same manner for all three data sets. The number of genes detected for all three species in addition to the number of reads and the type of sequencer used are summarized in Table 4.1.

In order to compare expression patterns among species, it was first necessary to insure that orthologs could be accurately identified. To compare the data sets, orthologs for all expressed genes for each species were determined by using g-profiler (Reimand et al. 2011). Any gene that did not have an ortholog in all three species was removed: this resulted in 18% rat, 26% mouse, and 34% human genes removed from further analysis (Table 4.4). It is important to note that most, if not all, of these genes were eliminated due to deficiencies in annotation and not because they are unique to a given species (Table 4.5). The higher percentage of genes eliminated in the human and mouse data set is a reflection of the fact that the rat genome was sequenced later than both the mouse and human genomes and the annotation in the rat is not as robust. Using the set of orthologous genes, the number of genes expressed in ESCs among all three species and among species was determined (Figure 4.5).

Table 4.4. Summary of orthologs by species

		Shares an ortholog			
Species	**Genes Detected**	**Mouse**	**Rat**	**Human**	**All Species**
hESC	17,874	11,739	11,739	-	11,739
mESC	14,417	-	10,611	10,726	10,611
rESC	10,931	9,370	-	9,465	8,968

Table 4.5 Comparison of genomes and annotated transcripts in available databases

	Rat (Rnor 5.0)	Mouse (GRCm38.p1)	Human (GrCH37.p10)
Genome Size	2.7 Gbp	2.7 Gbp	3.2 Gbp
# of Genes	29,100	36, 506	41,607
Reference	(Gibbs et al. 2004)	(Gnerre et al. 2011)	(Gnerre et al. 2011)

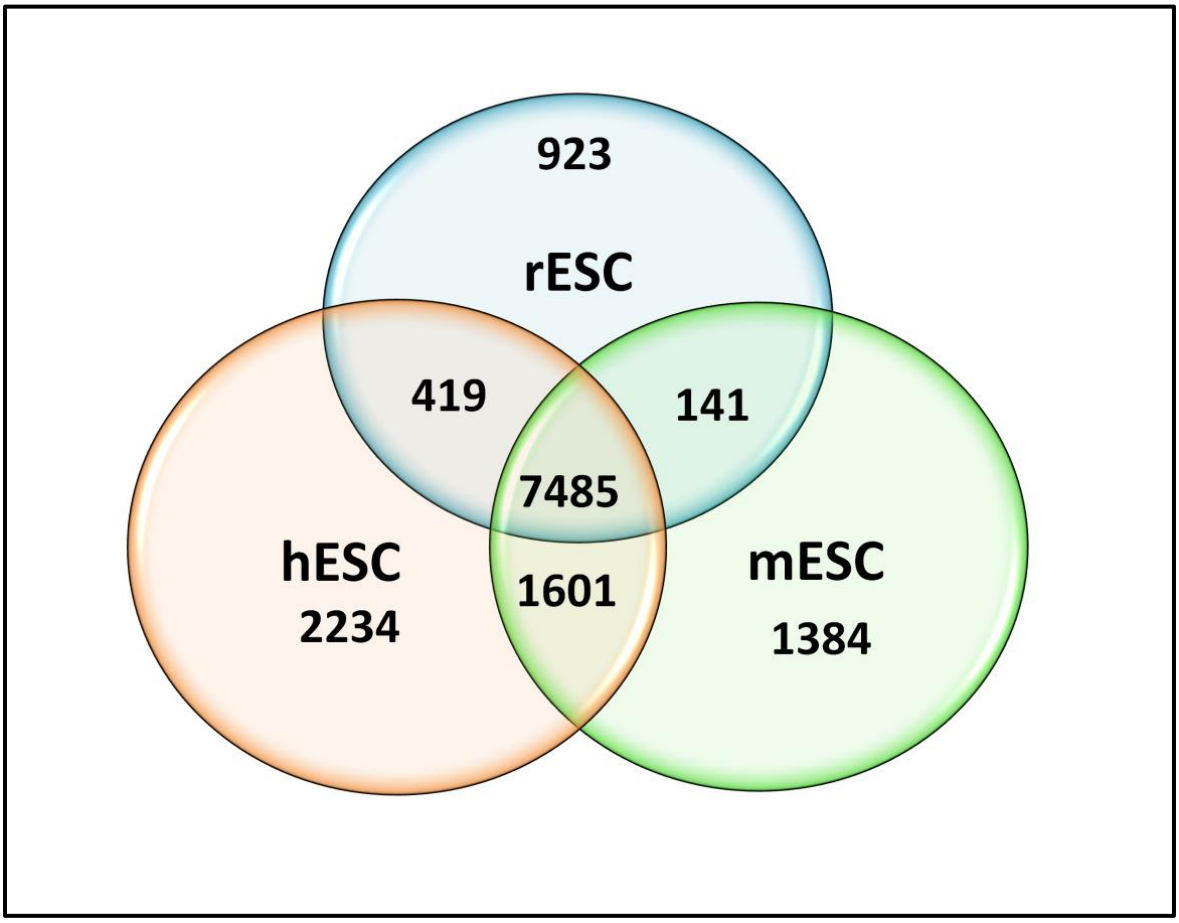


Figure 4.5. Comparison of expressed genes in rat, mouse, and human ESCs. ESC's gene comparison among species was done only with orthologous genes. Total number of genes is 11,739, 10,611, and 8,968 for human, mouse, and rat, respectively.

Analysis was performed using Toppcluster to determined statistically significant pathways for all three species in order to draw comparisons (Figure 4.6) (Praneenararat et al. 2012). In summary, mouse and rat share more common

pathways than any other species combination. Sixty-nine pathways are rodent specific, 23 shared among all three species, 1 among human and mouse only, and 0 among rat and human only. Additionally, 42 pathways are specific for rat, 23 are specific for mouse, and 1 is specific for human. The top 10 statistically significant pathways as defined by a <0.05 p-value suggest that all three species have statistically significant expression of HIV pathways, NGF signaling, and Erbβ downstream signaling. Included in the 21 pathways shared by all three species are cancer pathways, Tgf- β, and Wnt signaling pathways.

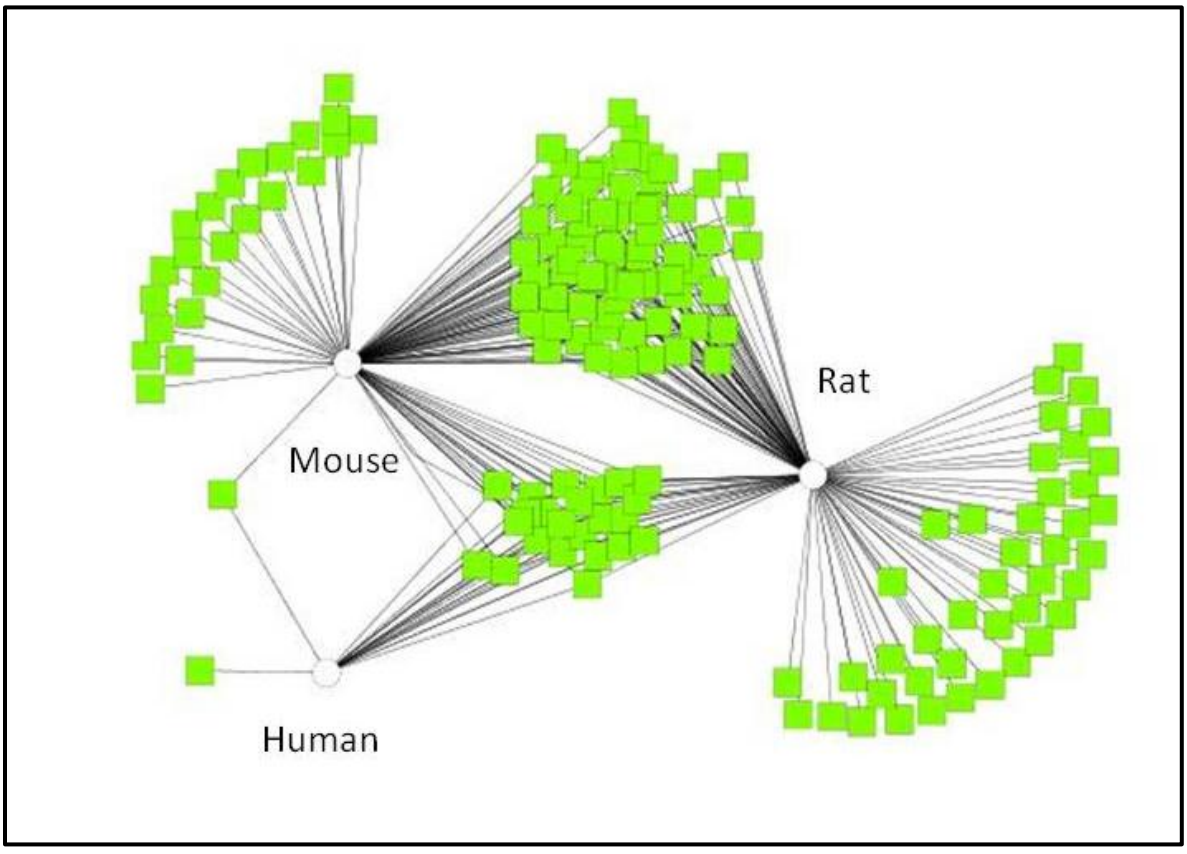


Figure 4.6 Toppcluster pathway analysis of rat, human, and mouse ESCs. Pathways are represented by boxes. Sixty-nine pathways are shared exclusively among mouse and rat, 23 are shared among all three species, 1 is shared exclusively among human and mouse, and 0 are shared exclusively among rat and human. Forty-two pathways were specific for rat, 23 are specific for mouse, and 1 is specific for human.

Oct4 (*Pou5F1*) is a transcription factor that plays an essential role in the self-renewal capacity of ESCs. An *Oct4* centric protein network had been previously described for mouse ESC (Pardo et al. 2010; van den Berg et al. 2010). Using the RNA-seq-derived data sets for rat, mouse and human, we examined the expression of the 169 genes that are part of the *Oct4* network. Of these 169 genes, only 22 were not expressed in all three species (Figure 4.7). Of these 22 genes, 16 were expressed in mouse and human but not rat, 1 was expressed in rat and mouse but not human, 1 was expressed only in human, and 4 were expressed in rat and human but not mouse. This later observation was unexpected as the network was based on expression data from mouse ESCs therefore all of the genes were expected to be represented in the mouse data set we used. However, it is possible that differences among culture conditions, the fact that different cell lines were used in the different studies, or that the detection methods used to assess expression may account for the discrepancy.

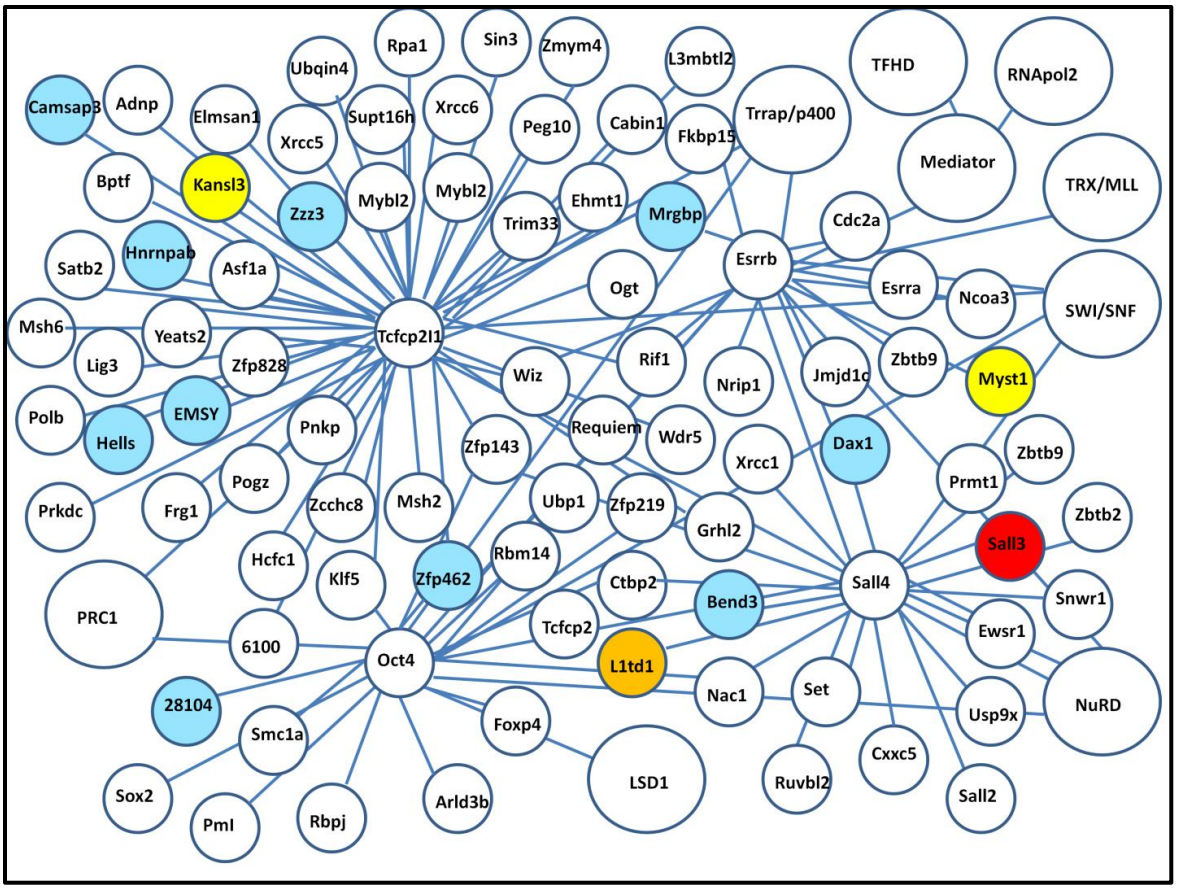


Figure 4.7. *Oct4* centric protein network. The network represents proteins that interact with *Oct4* based on previous work using mESCs (Pardo et al. 2010; van den Berg et al. 2010). Large circles represent protein complexes. White = genes expressed in all three species, Orange = genes expressed in mouse and rat, Teal = genes expressed only in rat, Yellow = genes expressed only in mouse, Red = genes expressed only in human.

CHAPTER V

DISCUSSION

In this study, we characterized the mRNA transcriptome of a Dark Agouti rat embryonic stem cell line in order to provide a publicly available normal reference for future experiments. Towards this goal, we generated 50-mer paired end reads for 2 samples at a high sequence depth (247 million reads), mapped the sequences against the UCSC RGSC v5.0 database (March, 2012), and generated a list of expressed genes and their relative expression values as measured in RPKMs. Because our results were highly repeatable based on the observation that only 32 genes had a >1 fold expression difference among the replicates and the correlation of expression was high (R^2 = 0.99), we limited our data set to 2 replicates. After strict filtering by removal of duplicate and low quality reads, and setting the criteria of inclusion to those genes with >1 RPKM expression, a total of 10,931 genes were used for further analysis. We chose to use more stringent criteria in an effort to eliminate analysis artifacts.

In order to validate the observation that 32 genes showed a >1 fold expression difference among biological replicates, primers for 4 genes were chosen. Based on our results the predicted RNA-seq fold change among biological replicates appears to be accurate. This correlates with previously reported findings that RNA-seq expression has a high correlation with quantitative RT-PCR results (Fang and Cui 2011).

The most highly expressed rESC genes correlated with pathways and GO terms associated with insulin metabolism, embryogenesis, and neural growth factor signaling. Since insulin is a component of rESC media, finding upregulated genes for insulin metabolism was not surprising. ESCs are isolated at the blastocyst stage, so upregulation of genes involved with embryonic development is also not unexpected. However, it is interesting that neuronal growth factor signaling is upregulated in rESC as well as in human and mouse ESCs. Upregulation of NGF signaling is commonly used to induce embryonic stem cells to differentiate into neurons, therefore its role in the embryonic stem cell state is not clear (Wobus et al. 1988; Schuldiner et al. 2000; Bibel et al. 2004).

Additional analysis was performed to detect undescribed isoforms. Alternative splicing is an important mechanism for enabling a single gene to code for multiple proteins and it results in functional diversity of these proteins in different tissues and cell types. Given the unique nature of ESCs, the presence of undescribed isoforms that have not been previously reported in other tissues or cell types was not unexpected. To validate our data sets using alternative methodology, three predicted undescribed isoforms were chosen for confirmation by RT-PCR and nucleotide sequence analysis. Primers were designed to allow amplification of key regions of predicted isoforms. All predicted undescribed isoforms chosen for analysis were confirmed demonstrating that is possible to identify undescribed isoforms using RNA-Seq. This does not however, explain what the biological explanation for the presence of these undescribed isoforms.

RNA-seq has the capability to detect undescribed poly $(A)^+$ transcripts. It is not surprising that there are undescribed predicted transcripts in rESCs as 1) annotation in rat is not as robust as human and mouse and 2) noncoding RNAs with poly $(A)^+$ tails such as lincRNA were identified in mESCs (Guttman et al. 2010). Primers were designed to allow for RT-PCR amplification of the entire transcript and each "exon". Of the 4 exons tested, only 1 was confirmed. The reason for this lack of confirmation is hypothesized as due to the sensitivity of the assay rather than an error in the RNA-seq analysis. RNA-seq is capable of detecting a transcript down to one copy (Jiang et al. 2011). It is predicted that only 2 and 14 full length transcripts are present for the predicted poly A^+ transcript 71 and 80. In order to verify the validity of these findings, further testing capable of detecting 1 copy of a transcript will be necessary.

In order to potentially learn more about the commonalities and differences among ESCs from different species, the rat embryonic stem cell data set was compared to publicly available paired-end RNA-seq mouse and human ESC data (Birney et al. 2007; Guttman et al. 2010). Despite being performed at different sequencing depths, all three data sets were paired end in order to provide a more accurate comparison. The algorithms associated with aligning reads consider a single read length in order to gain specificity in alignment. By using both ends of a read it allows for greater accuracy of an alignment by essentially extending the length of read. If a dataset with single end reads were used in an analysis with paired end reads, the dataset with single end reads would have a less accurate alignment when compared with a paired end dataset. Since each species was

performed at a different sequence depth, we did not make any comparisons related to relative expression levels among the three species.

In our analysis, expressed genes that were shared in common among all three species can help establish common ESC expression profiles that may uniquely define ESCs. Genes that were expressed only in one species or only among two species may reflect fundamental differences in the ESC transcriptomes of different species or alternatively, they may be artifacts of differences in culture conditions. Of note, all three species cell lines were cultured in different media, and the type of media influences expression of different signaling pathways (Kunath et al. 2007; Buehr et al. 2008; Li et al. 2008; Hirai et al. 2011). Because of this, when genes are expressed only in one or two species, it is difficult to speculate about the significance of that finding.

The analysis for pathway enrichment and GO terms was done using only the set of orthologous genes. Many of these pathways are labeled according to the context in which they were discovered such as HIV infection and influenza. Genes that are involved in these particular pathways are involved in normal cellular functions. It is not clear why these particular genes are upregulated in ESCs. One possible explanation is that these genes are vital in order for these viruses to “rewire” the host cell in order to propagate. This information provides a starting point for exploring the gene network and the biological role of this molecular pathway in ESCs.

Further analysis of a protein interaction network based on *Oct4* expression in mESC revealed differences across all three species (Pardo et al. 2010; van den Berg et al. 2010). *Oct4* is fundamental in maintaining the pluripotency network for mESC, so exploring differences in all three species may assist in defining the ESC state (Niwa et al. 2000). In total 22 genes were differentially expressed among the three species. Possible explanations for these differences could be related to dissimilarity in ESC media or they could truly represent species-specific gene expression. A few genes such as the transcript *2810474O19Rik* have no known function, but have been repeatedly identified as being expressed in mESCs (Ko et al. 2000; Diez-Roux et al. 2011). A few genes such as *Act16a* which is involved in TNF-alpha signaling could be linked to media differences (Gotschel et al. 2008). rESC media uses inhibition of GSK-3, which modulates TNF-alpha signaling. Several genes are involved in epigenetic regulation such as *Nr0b1* (*Dex-1*). *Nr0b1* is of particular interest because it has been previously shown that removal or down regulation of this gene results in loss of pluripotency in mESC (Khalfallah et al. 2009). All of these genes were not detected in rat, but were in mouse and human, so their role in maintaining the rESC state is unclear.

ESCs for all three species have a high number of expressed genes assigned to cancer pathways. It has been proposed that stem cells have the potential to give rise to cancer (Reya et al. 2001; Polyak and Hahn 2006; Dalerba et al. 2007; Friedmann-Morvinski et al. 2012) and the data would support this. Consistent with previous reports implicating the Erbβ, TGF-β, and Wnt signaling pathways in

ESC pluripotency (Alvarez et al. 2012; Yeo and Ng 2013), these same pathways were statistically significant for the ESCs from all three species in our analysis.

This data is only a snapshot at one time point of one strain of rESCs. In order to provide a deeper examination of the normal transcriptome of rESCs, RNA-Seq would need to be performed on 1) multiple strains of rESCs and 2) multiple time points.

Examining multiple strains of rESCs would help elucidate what the key components for rESC maintenance are. From this dataset alone, it will be impossible to determine what gene expression patterns are unique to the Dark Agouti strain that these rESCs were isolated from. Different rat strains have different phenotypes and presumably their rESCs would express different gene patterns. By overlaying these gene patterns, it would be possible to reveal what the consistent rESC pattern. From this pattern, a test could be provided for identification of future rESC lines.

Exploring multiple time points would help answer a couple of questions concerning rESCs. The longer ESCs are passaged it becomes more difficult to produce germline competent chimeras. Examining multiple time points should help elucidate what critical change happens to causes this difficulty. This would lead to 1) improvements in media conditions and 2) help provide the key to understanding how germline transmission from ESC occurs. Another question concerning ESC biology is whether ESCs are a natural phenomenon or a product of *in vitro* conditions. By documenting what the changes are to ESCs as they

“age” would provide insight to the degree ESCs change from the environment they are isolated.

Additional experiments to understand rESC biology are essential. Examination on a global scale micro RNAs and the epigenome of rESC would provide insight to the regulation of rESC transcriptome. Examining only mRNA does not provide a complete picture of rESC biology on a global scale. It could be argued that the factors controlling mRNA expression are just as important as the expression levels of mRNA.

In conclusion, we have presented the first transcriptome for rat ESCs using RNA-seq analysis. It is envisioned that this data set will be of use by serving as a control for future experiments. To this end, the dataset for this paper have been made publicly available (See Methods).

Appendix A

Mouse Embryonic Fibroblast (MEF) Media

GMEM ((Sigma, St. Louis, MO G5154)
10% FBS (HyClone SH30070.03)
1% GlutaMAX™-I (2 mM) (Invitrogen, Grand Island, NY (Gibco) 35050-061)
1% penicillin/streptomycin (Invitrogen, Grand Island, NY 15140-122)

MEF Preparation: For 250 mL MEF medium, add 25 mL FBS, 2.5 mL GlutaMAX™-I solution and 2.5 mL penicillin/streptomycin solution to 220 mL GMEM and filter. Store at 4 °C and use within one month.

Rat Embryonic Stem Cell Media (2i) Stock Solution Preparation

All stock reagents are made at least one day prior to making media and filtered sterilized individually using pore size of 0.2 µM (Fisher, Ashevile, NC (Thermo Scientific) 09-740-39A, SCGP00525, or (BD Falcon) 301603) and stored in amber Eppendorf tubes.

Apo-Transferin (Sigma, St.Louis, MO, T1147) (100 mg/mL) stock solution: Dissolve 500 mg in 5 mL sterile water overnight at 4 °C. Prepare in 1 mL aliquots and store at -20 °C for up to 1 year.

BSA (Invitrogen, Grand Island, NY 15260-037) 7.5% solution stock solution: Prepare in 1 mL aliquots in 1.5 mL tubes and store at -20 °C for up to 1 year.

Insulin (Sigma, St.Louis, MO, I1882) 25 mg/mL stock solution: Dissolve 100 mg insulin in 4 mL sterile 0.01 M HCl (Sigma, St.Louis, MO, H9892) overnight at 4 °C. Prepare 100 µl aliquots in 0.5 mL tubes. Store at -20 °C for up to 1 year.

Progesterone (Sigma, St.Louis, MO, P8783) 0.6 mg/mL stock solution: Dissolve 6 mg in 10 mL high-grade ethanol. Prepare aliquots (volume is of your own preference) and store at -20 °C for up to 1 year.

Putrescine (P5780) 160 mg/mL stock solution: Dissolve 1.6 g in 10 mL sterile water. Prepare 0.5 mL or 1 mL aliquots and store at -20 °C up to 1 year.

Sodium selenite (Sigma, St.Louis, MO, S5261) 3 mM stock solution: Must be prepared in fume hood. Due to the air circulation in a fume hood, it is necessary to prepare 2 dilutions in order to obtain a 3 mM stock solution. First dissolve 25.9 mg in 5 mL sterile water to make a 30 mM stock and then add 0.5 mL of this stock into 4.5 mL sterile water to obtain a 3 mM stock solution. Prepare 0.5 mL aliquots and store at -20 °C for up to 1 year.

CHIR99021 (Stemgent 04-0004) 3 µM stock solution: Dissolve 2 mg into 1.43 mL of Dimethyl sulfoxide (DMSO) (Sigma, St.Louis, MO, D248). Prepare aliquots in 200 µL and store at -80 °C for up to 1 year.

PD0325901 (Stemgent 04-0006) 0.5 µM stock solution: Dissolve 2 mg into 8.2 mL of Dimethyl sulfoxide (DMSO) (Sigma, St.Louis, MO, D248). Prepare aliquots in 200 µL and store at -80 °C for up to 1 year.

N2 stock solution

1 ml apo-Transferrin
0.67 mL BSA
33 µl progesterone
100 µl putrescine
10 µl sodium selenite
8.187 mL DMEM/F12

Prepare 1 mL aliquots and store at -20 °C for up to 1 year.

2i Media Preparation

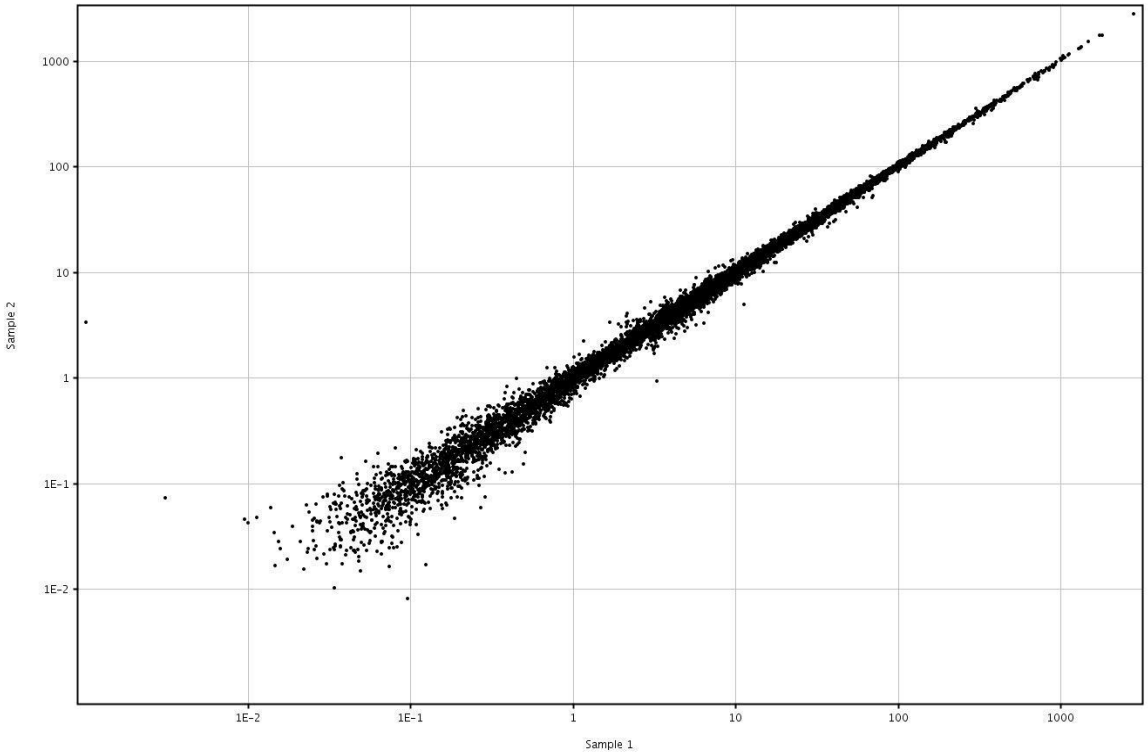
1. For 200 mL of media, thaw out 1 mL of N2 stock, 2 mL of B27 (Invitrogen, Grand Island, NY, 17504-044), 1 mL of Glutamax, 200 µL of CHIR99201, 200 µL of PD0325901, and 100 µL of insulin.
2. In one 500 mL beaker add 100 mL of DMEM/F12 and 1 mL N2 stock.
3. In another 500 mL beaker add 100 mL Neurobasal media, 2 mL B27, and 1 mL Glutamax.
4. Mix both beakers by pouring into one another.
5. Immediately rinse out the empty beaker 10x with Milli-Q water to avoid residue from forming on glass.
6. While stirring media with 5 mL serological pipet tip add drop by drop, 100 µL of insulin to avoid precipitation of the insulin.
7. Add 200 µL of CHIR99201, 200 µL of PD0325901 and 2 mL of β-mercaptoethanol (Millipore, Billerica, MA ES-007-E).
8. Mix well and pour into 250 mL filter unit and filter.
9. Immediately rinse out the empty beaker 10x with Milli-Q water to avoid residue from forming on glass.
10. Aliquot media at 40 mLs.
11. Store at 4 °C and use within a month.

Cryopreservation of rat ESCs

1. rESCs were passaged for >3 passages as previously described in the Materials and Methods section.
2. On the day to cryopreserve the cells, prepare cryovials and media prior to preparing cells.

3. Cryopreservation media is 90% 2i media and 10% DMSO. Prepare media on ice by slowly dropping DMSO in the 2i media as the reaction is exothermic and keep on ice.
4. To prepare the cells, obtain single cell suspension as described in the Materials and Methods section and count using a hemocytometer.
5. Cells are pelleted at 150 g for 5 min.
6. While in the hood on ice, pipette off the 2i media and resuspend the cells in cryopreservation media to a density of 1x10^6 cells/mL. Aliquot 1 mL of cells per cryotube and place the vials into a NALGEN Cryo 1 °C freezing container (Thermo Scientific 5100-0001).
7. Place the container into a -80 °C overnight and transfer the vials to liquid nitrogen the following day for long term storage.

Appendix B



F 1- Scatter plot of gene expression. Demonstrates the high correlation of gene expression among rESC samples 1 & 2. $R^2 = 0.9919$.

Appendix C

Name	5’ -Sequence -3’
Rat Dnmt3b Ex 21 F	CGCCATCAAGGTTTCTGCTG
Rat Dnmt3b Ex 24 R	CCCCACACAGGTGAGCTAAG
Rat Lef1 Ex 1 F-2	CGAGATCAGTCACCCCGAAG
Rat Lef1 Ex 11 R	TGTAGGCAGCTGTCATTCTGGG
Rat Nras Ex 5 F	GGGTGTGGAGGATGCCTTTT
Rat Nras Ex 7 R	AGCCGAGTGAGGAGGTAGTT
NewGene 80 Ex1F	ACCAGGGAATGCCTGCTACTA
NewGene 80 Ex1R	TTTGCCCAACTCATCCCACT
NewGene80 Ex2F	CCACCCAGATCTGAAGGGAC
NewGene80 Ex2R	CCAGATGGTGCTAGGCGTTT
NewGene71 Ex 1F	TGAGCATTCTTTGGTTGCTGT
NewGene71 Ex 1 R	CGAGCCAGAATCTGCAGTCA
NewGene71 Ex2 F	CCTTGGCTATGGGCAACTGA
NewGene 71 Ex2 R	CTGCCATGGAGACCCAGTTT

BIBLIOGRAPHY

Alahdal HM. 2011. Deriving Bovine Embryonic Stem-Like Cells in Defined Conditions. p. 179. University of Waikato.

Alvarez CV, Garcia-Lavandeira M, Garcia-Rendueles ME, Diaz-Rodriguez E, Garcia-Rendueles AR, Perez-Romero S, Vila TV, Rodrigues JS, Lear PV, Bravo SB. 2012. Defining stem cell types: understanding the therapeutic potential of ESCs, ASCs, and iPS cells. *J Mol Endocrinol* **49**(2): R89-111.

Ambrosetti DC, Basilico C, Dailey L. 1997. Synergistic activation of the fibroblast growth factor 4 enhancer by Sox2 and Oct-3 depends on protein-protein interactions facilitated by a specific spatial arrangement of factor binding sites. *Mol Cell Biol* **17**(11): 6321-6329.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**(1): 25-29.

Assou S, Le Carrour T, Tondeur S, Strom S, Gabelle A, Marty S, Nadal L, Pantesco V, Reme T, Hugnot JP et al. 2007. A meta-analysis of human embryonic stem cells transcriptome integrated into a web-based expression atlas. *Stem Cells* **25**(4): 961-973.

Avilion AA, Nicolis SK, Pevny LH, Perez L, Vivian N, Lovell-Badge R. 2003. Multipotent cell lineages in early mouse development depend on SOX2 function. *Genes Dev* **17**(1): 126-140.

Balasubramanian S, Zheng D, Liu YJ, Fang G, Frankish A, Carriero N, Robilotto R, Cayting P, Gerstein M. 2009. Comparative analysis of processed ribosomal protein pseudogenes in four mammalian genomes. *Genome Biol* **10**(1): R2.

Ben-David U, Kopper O, Benvenisty N. 2012. Expanding the boundaries of embryonic stem cells. *Cell Stem Cell* **10**(6): 666-677.

Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K et al. 2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**(2): 315-326.

Bhutani N, Brady JJ, Damian M, Sacco A, Corbel SY, Blau HM. 2010. Reprogramming towards pluripotency requires AID-dependent DNA demethylation. *Nature* **463**(7284): 1042-1047.

Bibel M, Richter J, Schrenk K, Tucker KL, Staiger V, Korte M, Goetz M, Barde YA. 2004. Differentiation of mouse embryonic stem cells into a defined neuronal lineage. *Nat Neurosci* **7**(9): 1003-1009.

Bilic J, Izpisua Belmonte JC. 2012. Concise review: Induced pluripotent stem cells versus embryonic stem cells: close enough or yet too far apart? *Stem Cells* **30**(1): 33-41.

Birney E Stamatoyannopoulos JA Dutta A Guigo R Gingeras TR Margulies EH Weng Z Snyder M Dermitzakis ET Thurman RE et al. 2007. Identification

and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**(7146): 799-816.

Blair K, Wray J, Smith A. 2011. The liberation of embryonic stem cells. *PLoS Genet* **7**(4): e1002019.

Blomberg LA, Telugu BP. 2012. Twenty years of embryonic stem cell research in farm animals. *Reproduction in domestic animals = Zuchthygiene* **47 Suppl 4**: 80-85.

Boone CW, Mantel N, Caruso TD, Jr., Kazam E, Stevenson RE. 1971. Quality control studies on fetal bovine serum used in tissue culture. *In Vitro* **7**(3): 174-189.

Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, Zucker JP, Guenther MG, Kumar RM, Murray HL, Jenner RG et al. 2005. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**(6): 947-956.

Bradley A, Evans M, Kaufman MH, Robertson E. 1984. Formation of germ-line chimaeras from embryo-derived teratocarcinoma cell lines. *Nature* **309**(5965): 255-256.

Brinster RL. 1974. The effect of cells transferred into the mouse blastocyst on subsequent development. *J Exp Med* **140**(4): 1049-1056.

Brook FA, Gardner RL. 1997. The origin and efficient derivation of embryonic stem cells in the mouse. *Proc Natl Acad Sci U S A* **94**(11): 5709-5712.

Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, Ersoz E, Flint-Garcia S, Garcia A, Glaubitz JC et al. 2009. The genetic architecture of maize flowering time. *Science* **325**(5941): 714-718.

Buehr M, Meek S, Blair K, Yang J, Ure J, Silva J, McLay R, Hall J, Ying QL, Smith A. 2008. Capture of authentic embryonic stem cells from rat blastocysts. *Cell* **135**(7): 1287-1298.

Calloni R, Cordero EA, Henriques JA, Bonatto D. 2013. Reviewing and Updating the Major Molecular Markers for Stem Cells. *Stem Cells Dev*.

Cao S, Wang F, Chen Z, Liu Z, Mei C, Wu H, Huang J, Li C, Zhou L, Liu L. 2009. Isolation and culture of primary bovine embryonic stem cell colonies by a novel method. *Journal of experimental zoology Part A, Ecological genetics and physiology* **311**(5): 368-376.

Chan SC, Gantenbein-Ritter B. 2012. Intervertebral disc regeneration or repair with biomaterials and stem cell therapy--feasible or fiction? *Swiss Med Wkly* **142**: w13598.

Chen G, Wang C, Shi T. 2011. Overview of available methods for diverse RNA-Seq data analyses. *Sci China Life Sci* **54**(12): 1121-1128.

Chen J, Bardes EE, Aronow BJ, Jegga AG. 2009. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res* **37**(Web Server issue): W305-311.

Chen S, Do JT, Zhang Q, Yao S, Yan F, Peters EC, Scholer HR, Schultz PG, Ding S. 2006. Self-renewal of embryonic stem cells by a small molecule. *Proc Natl Acad Sci U S A* **103**(46): 17266-17271.

Chepelev I, Wei G, Tang Q, Zhao K. 2009. Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. *Nucleic Acids Res* **37**(16): e106.

Chickarmane V, Olariu V, Peterson C. 2012. Probing the role of stochasticity in a model of the embryonic stem cell - heterogeneous gene expression and reprogramming efficiency. *BMC Syst Biol* **6**(1): 98.

Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G et al. 2008. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods* **5**(7): 613-619.

Collin J, Lako M. 2011. Concise review: putting a finger on stem cell biology: zinc finger nuclease-driven targeted genetic editing in human pluripotent stem cells. *Stem Cells* **29**(7): 1021-1033.

Daheron L, Opitz SL, Zaehres H, Lensch MW, Andrews PW, Itskovitz-Eldor J, Daley GQ. 2004. LIF/STAT3 signaling fails to maintain self-renewal of human embryonic stem cells. *Stem Cells* **22**(5): 770-778.

Dalerba P, Cho RW, Clarke MF. 2007. Cancer stem cells: models and concepts. *Annual review of medicine* **58**: 267-284.

Demers SP, Yoo JG, Lian L, Therrien J, Smith LC. 2007. Rat embryonic stem-like (ES-like) cells can contribute to extraembryonic tissues in vivo. *Cloning Stem Cells* **9**(4): 512-522.

Diez-Roux G, Banfi S, Sultan M, Geffers L, Anand S, Rozado D, Magen A, Canidio E, Pagani M, Peluso I et al. 2011. A high-resolution anatomical atlas of the transcriptome in the mouse embryo. *PLoS Biol* **9**(1): e1000582.

Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J et al. 2012. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in bioinformatics*.

Edgar R, Domrachev M, Lash AE. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**(1): 207-210.

Edgren H, Murumagi A, Kangaspeska S, Nicorici D, Hongisto V, Kleivi K, Rye IH, Nyberg S, Wolf M, Borresen-Dale AL et al. 2011. Identification of fusion genes in breast cancer by paired-end RNA-sequencing. *Genome Biol* **12**(1): R6.

Evans MJ, Kaufman MH. 1981. Establishment in culture of pluripotential cells from mouse embryos. *Nature* **292**(5819): 154-156.

Evans SJ, Datson NA, Kabbaj M, Thompson RC, Vreugdenhil E, De Kloet ER, Watson SJ, Akil H. 2002. Evaluation of Affymetrix Gene Chip sensitivity in rat hippocampal tissue using SAGE analysis. Serial Analysis of Gene Expression. *Eur J Neurosci* **16**(3): 409-413.

Fang Z, Cui X. 2011. Design and validation issues in RNA-seq experiments. *Briefings in bioinformatics* **12**(3): 280-287.

Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S et al. 2012. Ensembl 2012. *Nucleic Acids Res* **40**(Database issue): D84-90.

Fouse SD, Shen Y, Pellegrini M, Cole S, Meissner A, Van Neste L, Jaenisch R, Fan G. 2008. Promoter CpG methylation contributes to ES cell gene

regulation in parallel with Oct4/Nanog, PcG complex, and histone H3 K4/K27 trimethylation. *Cell Stem Cell* **2**(2): 160-169.

Friedmann-Morvinski D, Bushong EA, Ke E, Soda Y, Marumoto T, Singer O, Ellisman MH, Verma IM. 2012. Dedifferentiation of neurons and astrocytes by oncogenes can induce gliomas in mice. *Science* **338**(6110): 1080-1084.

Gao L, Fang Z, Zhang K, Zhi D, Cui X. 2011. Length bias correction for RNA-seq data in gene set analyses. *Bioinformatics* **27**(5): 662-669.

Garber M, Grabherr MG, Guttman M, Trapnell C. 2011. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Methods* **8**(6): 469-477.

Gibbs RA Weinstock GM Metzker ML Muzny DM Sodergren EJ Scherer S Scott G Steffen D Worley KC Burch PE et al. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**(6982): 493-521.

Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S et al. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci U S A* **108**(4): 1513-1518.

Gotschel F, Kern C, Lang S, Sparna T, Markmann C, Schwager J, McNelly S, von Weizsacker F, Laufer S, Hecht A et al. 2008. Inhibition of GSK3 differentially modulates NF-kappaB, CREB, AP-1 and beta-catenin signaling in hepatocytes, but fails to promote TNF-alpha-induced apoptosis. *Exp Cell Res* **314**(6): 1351-1366.

Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP et al. 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**(7235): 223-227.

Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C et al. 2010. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* **28**(5): 503-510.

Gyorffy B, Schafer R. 2010. Biomarkers downstream of RAS: a search for robust transcriptional targets. *Current cancer drug targets* **10**(8): 858-868.

Hansen KD, Brenner SE, Dudoit S. 2010. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res* **38**(12): e131.

Havlak P, Chen R, Durbin KJ, Egan A, Ren Y, Song XZ, Weinstock GM, Gibbs RA. 2004. The Atlas genome assembly system. *Genome Res* **14**(4): 721-732.

Hentze H, Soong PL, Wang ST, Phillips BW, Putti TC, Dunn NR. 2009. Teratoma formation by human embryonic stem cells: evaluation of essential parameters for future safety studies. *Stem cell research* **2**(3): 198-210.

Hirabayashi M, Kato M, Kobayashi T, Sanbo M, Yagi T, Hochi S, Nakauchi H. 2010a. Establishment of rat embryonic stem cell lines that can participate in germline chimerae at high efficiency. *Mol Reprod Dev* **77**(2): 94.

Hirabayashi M, Kato M, Sanbo M, Kobayashi T, Hochi S, Nakauchi H. 2010b. Rat transgenesis via embryonic stem cells electroporated with the Kusabira-orange gene. *Mol Reprod Dev* **77**(6): 474.

Hirai H, Karian P, Kikyo N. 2011. Regulation of embryonic stem cell self-renewal and pluripotency by leukaemia inhibitory factor. *Biochem J* **438**(1): 11-23.

Hockemeyer D, Wang H, Kiani S, Lai CS, Gao Q, Cassady JP, Cost GJ, Zhang L, Santiago Y, Miller JC et al. 2011. Genetic engineering of human pluripotent cells using TALE nucleases. *Nat Biotechnol* **29**(8): 731-734.

Huang da W, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**(1): 44-57.

Huang G, Ashton C, Kumbhani DS, Ying QL. 2011. Genetic manipulations in the rat: progress and prospects. *Curr Opin Nephrol Hypertens* **20**(4): 391-399.

Jakob H. 1984. Stem cells and embryo-derived cell lines: tools for study of gene expression. *Cell Differ* **15**(2-4): 77-80.

Jiang L, Schlesinger F, Davis CA, Zhang Y, Li R, Salit M, Gingeras TR, Oliver B. 2011. Synthetic spike-in standards for RNA-seq experiments. *Genome Res* **21**(9): 1543-1551.

Jin L, Wang W, Hu D, Min S. 2013. Effects of protonation and c5 methylation on the electrophilic addition reaction of Cytosine: a computational study. *The journal of physical chemistry B* **117**(1): 3-12.

Kaimal V, Bardes EE, Tabar SC, Jegga AG, Aronow BJ. 2010. ToppCluster: a multiple gene list feature analyzer for comparative enrichment clustering and network-based dissection of biological systems. *Nucleic Acids Res* **38**(Web Server issue): W96-102.

Kanda A, Sotomaru Y, Shiozawa S, Hiyama E. 2012. Establishment of ES cells from inbred strain mice by dual inhibition (2i). *The Journal of reproduction and development* **58**(1): 77-83.

Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**(1): 27-30.

Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. 2012. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* **40**(Database issue): D109-114.

Kang L, Kou Z, Zhang Y, Gao S. 2010. Induced pluripotent stem cells (iPSCs)-a new era of reprogramming. *J Genet Genomics* **37**(7): 415-421.

Kawamata M, Ochiya T. 2011. Gene-manipulated embryonic stem cells for rat transgenesis. *Cell Mol Life Sci* **68**(11): 1911-1915.

Keefer CL, Pant D, Blomberg L, Talbot NC. 2007. Challenges and prospects for the establishment of embryonic stem cell lines of domesticated ungulates. *Anim Reprod Sci* **98**(1-2): 147-168.

Khalfallah O, Rouleau M, Barbry P, Bardoni B, Lalli E. 2009. Dax-1 knockdown in mouse embryonic stem cells induces loss of pluripotency and multilineage differentiation. *Stem Cells* **27**(7): 1529-1537.

Kircher M, Kelso J. 2010. High-throughput DNA sequencing--concepts and limitations. *BioEssays : news and reviews in molecular, cellular and developmental biology* **32**(6): 524-536.

Kleinsmith LJ, Pierce GB, Jr. 1964. Multipotentiality of Single Embryonal Carcinoma Cells. *Cancer Res* **24**: 1544-1551.

Ko MS, Kitchen JR, Wang X, Threat TA, Wang X, Hasegawa A, Sun T, Grahovac MJ, Kargul GJ, Lim MK et al. 2000. Large-scale cDNA analysis reveals phased gene expression patterns during preimplantation mouse development. *Development* **127**(8): 1737-1749.

Kogenaru S, Qing Y, Guo Y, Wang N. 2012. RNA-seq and microarray complement each other in transcriptome profiling. *BMC Genomics* **13**: 629.

Kunath T, Saba-El-Leil MK, Almousailleakh M, Wray J, Meloche S, Smith A. 2007. FGF stimulation of the Erk1/2 signalling cascade triggers transition of pluripotent embryonic stem cells from self-renewal to lineage commitment. *Development* **134**(16): 2895-2902.

Levenstein ME, Ludwig TE, Xu RH, Llanas RA, VanDenHeuvel-Kramer K, Manning D, Thomson JA. 2006. Basic fibroblast growth factor support of human embryonic stem cell self-renewal. *Stem Cells* **24**(3): 568-574.

Li C, Yang Y, Gu J, Ma Y, Jin Y. 2009. Derivation and transcriptional profiling analysis of pluripotent stem cell lines from rat blastocysts. *Cell Res* **19**(2): 173-186.

Li J, Jiang H, Wong WH. 2010. Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biol* **11**(5): R50.

Li P, Tong C, Mehrian-Shai R, Jia L, Wu N, Yan Y, Maxson RE, Schulze EN, Song H, Hsieh CL et al. 2008. Germline competent embryonic stem cells derived from rat blastocysts. *Cell* **135**(7): 1299-1310.

Linsen SE, de Wit E, Janssens G, Heater S, Chapman L, Parkin RK, Fritz B, Wyman SK, de Bruijn E, Voest EE et al. 2009. Limitations and possibilities of small RNA digital gene expression profiling. *Nat Methods* **6**(7): 474-476.

Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, Sam L, Barrette T, Palanisamy N, Chinnaiyan AM. 2009. Transcriptome sequencing to detect gene fusions in cancer. *Nature* **458**(7234): 97-101.

Malone JH, Oliver B. 2011. Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biol* **9**: 34.

Mao CD, Byers SW. 2011. Cell-context dependent TCF/LEF expression and function: alternative tales of repression, de-repression and activation potentials. *Critical reviews in eukaryotic gene expression* **21**(3): 207-236.

Marguerat S, Bahler J. 2010. RNA-seq: from technology to biology. *Cell Mol Life Sci* **67**(4): 569-579.

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* **18**(9): 1509-1517.

Martin GR. 1981. Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *Proc Natl Acad Sci U S A* **78**(12): 7634-7638.

Martin GR, Evans MJ. 1975. Differentiation of clonal lines of teratocarcinoma cells: formation of embryoid bodies in vitro. *Proc Natl Acad Sci U S A* **72**(4): 1441-1445.

Martin GR, Wiley LM, Damjanov I. 1977. The development of cystic embryoid bodies in vitro from clonal teratocarcinoma stem cells. *Dev Biol* **61**(2): 230-244.

Martins-Taylor K, Xu RH. 2010. Determinants of pluripotency: from avian, rodents, to primates. *J Cell Biochem* **109**(1): 16-25.

Maruotti J, Munoz M, Degrelle SA, Gomez E, Louet C, Diez C, de Longchamp PH, Brochard V, Hue I, Caamano JN et al. 2012. Efficient derivation of bovine embryonic stem cells needs more than active core pluripotency factors. *Mol Reprod Dev* **79**(7): 461-477.

McGettigan PA. 2013. Transcriptomics in the RNA-seq era. *Current opinion in chemical biology*.

McIntyre LM, Lopiano KK, Morse AM, Amin V, Oberg AL, Young LJ, Nuzhdin SV. 2011. RNA-seq: technical variability and sampling. *BMC Genomics* **12**: 293.

Meek S, Buehr M, Sutherland L, Thomson A, Mullins JJ, Smith AJ, Burdon T. 2010. Efficient gene targeting by homologous recombination in rat embryonic stem cells. *PLoS One* **5**(12): e14225.

Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB et al. 2008. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**(7205): 766-770.

Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, Wong M, Sloan CA, Rosenbloom KR, Roe G, Rhead B et al. 2013. The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res* **41**(Database issue): D64-69.

Mitchell RT, Cowan G, Morris KD, Anderson RA, Fraser HM, McKenzie KJ, Wallace WH, Kelnar CJ, Saunders PT, Sharpe RM. 2008. Germ cell differentiation in the marmoset (Callithrix jacchus) during fetal and neonatal life closely parallels that in the human. *Hum Reprod* **23**(12): 2755-2765.

Mitsui K, Tokuzawa Y, Itoh H, Segawa K, Murakami M, Takahashi K, Maruyama M, Maeda M, Yamanaka S. 2003. The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. *Cell* **113**(5): 631-642.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**(7): 621-628.

Muller-McNicoll M, Neugebauer KM. 2013. How cells get the message: dynamic assembly and function of mRNA-protein complexes. *Nat Rev Genet* **14**(4): 275-287.

Nagalakshmi U, Waern K, Snyder M. 2010. RNA-Seq: a method for comprehensive transcriptome analysis. *Current protocols in molecular biology / edited by Frederick M Ausubel [et al]* **Chapter 4**: Unit 4 11 11-13.

Nagy A, Vintersten K. 2006. Murine embryonic stem cells. *Methods Enzymol* **418**: 3-21.

Nichols J, Zevnik B, Anastassiadis K, Niwa H, Klewe-Nebenius D, Chambers I, Scholer H, Smith A. 1998. Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4. *Cell* **95**(3): 379-391.

Niwa H. 2007. How is pluripotency determined and maintained? *Development* **134**(4): 635-646.

Niwa H, Miyazaki J, Smith AG. 2000. Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. *Nat Genet* **24**(4): 372-376.

Nowrousian M. 2010. Next-generation sequencing techniques for eukaryotic microorganisms: sequencing-based solutions to biological problems. *Eukaryot Cell* **9**(9): 1300-1310.

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. 1999. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **27**(1): 29-34.

Ong JM, da Cruz L. 2012. A review and update on the current status of stem cell therapy and the retina. *Br Med Bull* **102**: 133-146.

Oshlack A, Wakefield MJ. 2009. Transcript length bias in RNA-seq data confounds systems biology. *Biol Direct* **4**: 14.

Pardo M, Lang B, Yu L, Prosser H, Bradley A, Babu MM, Choudhary J. 2010. An expanded Oct4 interaction network: implications for stem cell biology, development, and disease. *Cell Stem Cell* **6**(4): 382-395.

Polyak K, Hahn WC. 2006. Roots and stems: stem cells in cancer. *Nat Med* **12**(3): 296-300.

Praneenararat T, Takagi T, Iwasaki W. 2012. Integration of interactive, multi-scale network navigation approach with Cytoscape for functional genomics in the big data era. *BMC Genomics* **13 Suppl 7**: S24.

Pruitt KD, Tatusova T, Brown GR, Maglott DR. 2012. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res* **40**(Database issue): D130-135.

Reimand J, Arak T, Vilo J. 2011. g:Profiler--a web server for functional interpretation of gene lists (2011 update). *Nucleic Acids Res* **39**(Web Server issue): W307-315.

Reya T, Morrison SJ, Clarke MF, Weissman IL. 2001. Stem cells, cancer, and cancer stem cells. *Nature* **414**(6859): 105-111.

Richards M, Tan SP, Tan JH, Chan WK, Bongso A. 2004. The transcriptome profile of human embryonic stem cells as defined by SAGE. *Stem Cells* **22**(1): 51-64.

Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. 2011. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol* **12**(3): R22.

Rosenkranz R, Borodina T, Lehrach H, Himmelbauer H. 2008. Characterizing the mouse ES cell transcriptome with Illumina sequencing. *Genomics* **92**(4): 187-194.

Ruhnke M, Ungefroren H, Zehle G, Bader M, Kremer B, Fandrich F. 2003. Long-term culture and differentiation of rat embryonic stem cell-like cells into

neuronal, glial, endothelial, and hepatic lineages. *Stem Cells* **21**(4): 428-436.

Sanchez-Pla A, Reverter F, Ruiz de Villa MC, Comabella M. 2012. Transcriptomics: mRNA and alternative splicing. *J Neuroimmunol* **248**(1-2): 23-31.

Sato N, Meijer L, Skaltsounis L, Greengard P, Brivanlou AH. 2004. Maintenance of pluripotency in human and mouse embryonic stem cells through activation of Wnt signaling by a pharmacological GSK-3-specific inhibitor. *Nat Med* **10**(1): 55-63.

Schena M, Shalon D, Davis RW, Brown PO. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**(5235): 467-470.

Schofield PN, Sundberg JP, Hoehndorf R, Gkoutos GV. 2011. New approaches to the representation and analysis of phenotype knowledge in human diseases and their animal models. *Brief Funct Genomics* **10**(5): 258-265.

Scholer HR, Ciesiolka T, Gruss P. 1991. A nexus between Oct-4 and E1A: implications for gene regulation in embryonic stem cells. *Cell* **66**(2): 291-304.

Schuldiner M, Yanuka O, Itskovitz-Eldor J, Melton DA, Benvenisty N. 2000. Effects of eight growth factors on the differentiation of cells derived from human embryonic stem cells. *Proc Natl Acad Sci U S A* **97**(21): 11307-11312.

Schwartzberg PL, Goff SP, Robertson EJ. 1989. Germ-line transmission of a c-abl mutation produced by targeted gene disruption in ES cells. *Science* **246**(4931): 799-803.

Serra M, Brito C, Correia C, Alves PM. 2012. Process engineering of human pluripotent stem cells for clinical application. *Trends Biotechnol* **30**(6): 350-359.

Smith AG. 2001. Embryo-derived stem cells: of mice and men. *Annu Rev Cell Dev Biol* **17**: 435-462.

Smith AG, Heath JK, Donaldson DD, Wong GG, Moreau J, Stahl M, Rogers D. 1988. Inhibition of pluripotential embryonic stem cell differentiation by purified polypeptides. *Nature* **336**(6200): 688-690.

Sohn YD, Han JW, Yoon YS. 2012. Generation of induced pluripotent stem cells from somatic cells. *Progress in molecular biology and translational science* **111**: 1-26.

Stranzinger GF. 1996. Embryonic stem-cell-like cell lines of the species rat and Bovinae. *Int J Exp Pathol* **77**(6): 263-267.

Suzuki H, Kamada N, Ueda O, Jishage K, Kurihara Y, Kurihara H, Terauchi Y, Azuma S, Kadowaki T, Kodama T et al. 1997. Germ-line contribution of embryonic stem cells in chimeric mice: influence of karyotype and in vitro differentiation ability. *Experimental animals / Japanese Association for Laboratory Animal Science* **46**(1): 17-23.

Tang F, Barbacioru C, Bao S, Lee C, Nordman E, Wang X, Lao K, Surani MA. 2010. Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. *Cell Stem Cell* **6**(5): 468-478.

Thomson JA, Itskovitz-Eldor J, Shapiro SS, Waknitz MA, Swiergiel JJ, Marshall VS, Jones JM. 1998. Embryonic stem cell lines derived from human blastocysts. *Science* **282**(5391): 1145-1147.

Tong C, Huang G, Ashton C, Li P, Ying QL. 2011. Generating gene knockout rats by homologous recombination in embryonic stem cells. *Nat Protoc* **6**(6): 827-844.

Tong C, Huang G, Ashton C, Wu H, Yan H, Ying QL. 2012. Rapid and Cost-Effective Gene Targeting in Rat Embryonic Stem Cells by TALENs. *J Genet Genomics* **39**(6): 275-280.

Tong C, Li P, Wu NL, Yan Y, Ying QL. 2010. Production of p53 gene knockout rats by homologous recombination in embryonic stem cells. *Nature* **467**(7312): 211-213.

Torri F, Dinov ID, Zamanyan A, Hobel S, Genco A, Petrosyan P, Clark AP, Liu Z, Eggert P, Pierce J et al. 2012. Next Generation Sequence Analysis and Computational Genomics Using Graphical Pipeline Workflows. *Genes* **3**(3): 545-575.

Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. 2013. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* **31**(1): 46-53.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**(5): 511-515.

Ueda S, Kawamata M, Teratani T, Shimizu T, Tamai Y, Ogawa H, Hayashi K, Tsuda H, Ochiya T. 2008. Establishment of rat embryonic stem cells and making of chimera rats. *PLoS One* **3**(7): e2800.

Vallier L, Alexander M, Pedersen RA. 2005. Activin/Nodal and FGF pathways cooperate to maintain pluripotency of human embryonic stem cells. *J Cell Sci* **118**(Pt 19): 4495-4509.

van den Berg DL, Snoek T, Mullin NP, Yates A, Bezstarosti K, Demmers J, Chambers I, Poot RA. 2010. An Oct4-centered protein interaction network in embryonic stem cells. *Cell Stem Cell* **6**(4): 369-381.

Wagner GP, Kin K, Lynch VJ. 2012. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in biosciences = Theorie in den Biowissenschaften* **131**(4): 281-285.

Wall PK, Leebens-Mack J, Chanderbali AS, Barakat A, Wolcott E, Liang H, Landherr L, Tomsho LP, Hu Y, Carlson JE et al. 2009. Comparison of next generation sequencing technologies for transcriptome characterization. *BMC Genomics* **10**: 347.

Wang K, Bucan M. 2008. Copy Number Variation Detection via High-Density SNP Genotyping. *CSH Protoc* **2008**: pdb top46.

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

Williams RL, Hilton DJ, Pease S, Willson TA, Stewart CL, Gearing DP, Wagner EF, Metcalf D, Nicola NA, Gough NM. 1988. Myeloid leukaemia inhibitory

factor maintains the developmental potential of embryonic stem cells. *Nature* **336**(6200): 684-687.

Wobus AM, Grosse R, Schoneich J. 1988. Specific effects of nerve growth factor on the differentiation pattern of mouse embryonic stem cells in vitro. *Biomedica biochimica acta* **47**(12): 965-973.

Yamamoto S, Nakata M, Sasada R, Ooshima Y, Yano T, Shinozawa T, Tsukimi Y, Takeyama M, Matsumoto Y, Hashimoto T. 2011. Derivation of rat embryonic stem cells and generation of protease-activated receptor-2 knockout rats. *Transgenic Res*.

Yeo JC, Ng HH. 2013. The transcriptional regulation of pluripotency. *Cell Res* **23**(1): 20-32.

Ying QL, Nichols J, Chambers I, Smith A. 2003. BMP induction of Id proteins suppresses differentiation and sustains embryonic stem cell self-renewal in collaboration with STAT3. *Cell* **115**(3): 281-292.

Young RA. 2011. Control of the embryonic stem cell state. *Cell* **144**(6): 940-954.

Zhang W, Chen J, Yang Y, Tang Y, Shang J, Shen B. 2011. A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. *PLoS One* **6**(3): e17915.

Zheng W, Chung LM, Zhao H. 2011. Bias detection and correction in RNA-Sequencing data. *BMC Bioinformatics* **12**: 290.

Zhou Q, Chipperfield H, Melton DA, Wong WH. 2007. A gene regulatory network in mouse embryonic stem cells. *Proc Natl Acad Sci U S A* **104**(42): 16438-16443.