CLONING AND ANNOTATION OF NOVEL TRANSCRIPTS FROM HUMAN EMBRYONIC STEM CELLS

by

Jaswinder Khattra

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate Studies (Genetics)

THE UNIVERSITY OF BRITISH COLUMBIA

December 2007

© Jaswinder Khattra, 2007
ABSTRACT

Both cDNA tag-based and DNA chip hybridization assays have revealed widespread transcriptional activity across mammalian genomes, providing a rich source of novel protein-coding and non-coding transcripts. Annotation and functional evaluation of this undefined transcriptome space represents a major step towards the comprehensive definition of biomolecules regulating the properties of living cells, including embryonic stem cells (ESCs) and their derivatives.

In this study I analysed 87 rare mRNA transcripts from human ESCs that mapped uniquely to the human genome, in regions lacking evidence for known genes or transcripts. In addition, the transcripts appeared enriched in the hESC transcriptome as enumerated by serial analysis of gene expression (SAGE). Full-length transcripts corresponding to twelve novel LongSAGE tags were recovered and evaluated with respect to gene structure, protein-coding potential, and gene regulatory features. In addition, transcript abundance was compared between RNA isolated from undifferentiated hESCs and differentiated cells.

Analysis of full-length transcripts revealed that the novel ORFs did not exceed a size of 129 amino acids and no matches were observed to well characterized protein domains. Interesting protein level predictions included small disulfide-bonded proteins, known members of which are important in a variety of biological processes. Transcripts evaluated for differential expression by real-time RT-qPCR (Reverse Transcription followed by real-time quantitative Polymerase Chain Reaction) were found to be variably expressed (0.2- to 4.5-fold) in Day-2 or Day-4 retinoic acid-induced differentiation cultures compared to undifferentiated hESCs.
Relative quantitation using a universal reference RNA (derived from pooled adult tissues) showed large differences in novel transcript levels (0.002- to 35-fold) compared to hESCs.

Collectively, these results provide a detailed analysis of a set of novel hESC transcripts and their abundance in early and adult differentiated cell types, both of which may advance our understanding of the transcriptional events governing stem cell behavior.
TABLE OF CONTENTS

Abstract.......................................................................................................................... ii
Table of Contents.......................................................................................................... iv
List of Tables.................................................................................................................. v
List of Figures................................................................................................................ vi
List of Abbreviations..................................................................................................... vii
Acknowledgements..................................................................................................... viii
1. Introduction............................................................................................................... 1
   1.1 Human embryonic stem cell research................................................................. 1
   1.2 Small protein gene discovery ............................................................................ 1
   1.3 Transcriptome profiling using LongSAGE......................................................... 3
   1.4 Hypothesis........................................................................................................... 4
   1.5 Objectives........................................................................................................... 5
2. Amplification and cloning of novel cDNAs............................................................. 6
   2.1 Introduction........................................................................................................ 7
   2.2 Methods............................................................................................................ 7
      2.2.1 Reverse transcription and RACE chemistry................................................ 8
      2.2.2 Cloning RACE products and full-length transcripts.................................... 9
      2.2.3 Cloning full-length ORFs........................................................................... 11
   2.3 Results and Discussion....................................................................................... 11
3. Analysis of novel transcript sequences................................................................... 14
   3.1 Introduction........................................................................................................ 14
   3.2 Methods............................................................................................................ 14
   3.3 Results and Discussion....................................................................................... 18
      3.3.1 Candidate transcript 3012........................................................................... 21
4. Quantitation of transcript abundance using RT-qPCR........................................... 30
   4.1 Introduction........................................................................................................ 30
   4.2 Methods............................................................................................................ 30
4.2.1 Primer selection ................................................................. 32
4.2.2 qPCR data analysis ............................................................ 33
4.3 Results and Discussion ......................................................... 34

5. Conclusion and Further Work .................................................. 37

Bibliography ................................................................................. 41

Appendices .................................................................................. 52
  Appendix 1. Experimental work flowchart ..................................... 52
  Appendix 2. hESC culture and differentiation ............................... 53
  Appendix 3. RNA quality electropherogram ................................. 56
  Appendix 4. Primers for 3’ RACE, Full-Length cDNA, and ORF recovery ........................ 57
  Appendix 5. Novel transcript sequence analysis flowchart ........... 58
  Appendix 6. Candidate 3012 sequence ........................................ 59
  Appendix 7. Candidate 3210 structure and features ..................... 60
  Appendix 8. Candidate 3372 structure and features ..................... 63
  Appendix 9. Candidate 3373 structure and features ..................... 66
  Appendix 10. Example of relative quantitation calculation ............ 69
  Appendix 11. Biohazard Approval Certificate ............................... 72
**LIST OF TABLES**

Table 1. Human ESC lines used for LongSAGE................................................................. 4

Table 2. InterPro domain hits to novel predicted ORFs.................................................. 20

Table 3. Transcripts that may have resulted from oligo d(T) mispriming.......................... 26

Table 4. Transcripts in a region of segmental duplication................................................ 26

Table 5. Novel ORFs with predicted disulfide bridges..................................................... 27

Table 6. FASTA similarity search of 111 aa ORF from candidate 3210............................. 29

Table 7. Primers for qPCR assays..................................................................................... 33

Table 8. Relative quantitation of transcript abundance...................................................... 36

Supplementary Table 1. Transcript sequence data. ftp://ftp03.bcgsc.ca/public/hESC_novel
LIST OF FIGURES

Figure 1. RACE chemistry ......................................................... 7
Figure 2. Agarose gel electrophoresis of RACE amplicons ...................... 12
Figure 3. Agarose gel electrophoresis of colony PCR verification of cloned RACE products ................................. 12
Figure 4. Candidate 3012 structure and features ........................................... 23-25
Figure 5. Candidate 3210 alignment of FASTA similarity search results ................................. 29
# LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Stands for</th>
</tr>
</thead>
<tbody>
<tr>
<td>aa</td>
<td>Amino acids</td>
</tr>
<tr>
<td>BCCA</td>
<td>British Columbia Cancer Agency</td>
</tr>
<tr>
<td>BLAST</td>
<td>Basic local alignment search tool</td>
</tr>
<tr>
<td>bp</td>
<td>Base pairs</td>
</tr>
<tr>
<td>cDNA</td>
<td>Complementary DNA</td>
</tr>
<tr>
<td>ChIP-chip</td>
<td>Chromatin immunoprecipitation on DNA chip</td>
</tr>
<tr>
<td>CMOST</td>
<td>Comprehensive mapping of SAGE tags</td>
</tr>
<tr>
<td>CSPG4</td>
<td>Chondroitin sulfate proteoglycan 4</td>
</tr>
<tr>
<td>db</td>
<td>Database</td>
</tr>
<tr>
<td>DNMT3B</td>
<td>DNA (cytosine-5-)-methyltransferase 3 beta</td>
</tr>
<tr>
<td>dNTPs</td>
<td>Deoxynucleoside triphosphates</td>
</tr>
<tr>
<td>EGF-like</td>
<td>Epidermal growth factor-like</td>
</tr>
<tr>
<td>ESC</td>
<td>Embryonic stem cell</td>
</tr>
<tr>
<td>EST</td>
<td>Expressed sequence tag</td>
</tr>
<tr>
<td>GSC</td>
<td>Genome Sciences Centre</td>
</tr>
<tr>
<td>hESC</td>
<td>Human embryonic stem cell</td>
</tr>
<tr>
<td>irrMEFs</td>
<td>Irradiated mouse embryonic fibroblasts</td>
</tr>
<tr>
<td>kb</td>
<td>Kilo base-pairs</td>
</tr>
<tr>
<td>kDa</td>
<td>Kilo Dalton</td>
</tr>
<tr>
<td>LB</td>
<td>Luria-Bertani</td>
</tr>
<tr>
<td>LTR</td>
<td>Long terminal repeat</td>
</tr>
<tr>
<td>miRNA</td>
<td>Micro RNA</td>
</tr>
<tr>
<td>MPSS</td>
<td>Massively parallel signature sequencing</td>
</tr>
<tr>
<td>MW</td>
<td>Molecular weight</td>
</tr>
<tr>
<td>n/a</td>
<td>Not applicable</td>
</tr>
<tr>
<td>ORF</td>
<td>Open reading frame</td>
</tr>
<tr>
<td>pI</td>
<td>Isoelectric point</td>
</tr>
<tr>
<td>qPCR</td>
<td>Quantitative polymerase chain reaction</td>
</tr>
<tr>
<td>RA</td>
<td>Retinoic acid</td>
</tr>
<tr>
<td>RACE</td>
<td>Rapid amplification of cDNA ends</td>
</tr>
<tr>
<td>rpm</td>
<td>Revolutions per minute</td>
</tr>
<tr>
<td>RT</td>
<td>Reverse transcription</td>
</tr>
<tr>
<td>SAGE</td>
<td>Serial analysis of gene expression</td>
</tr>
<tr>
<td>SMART™</td>
<td>Switching mechanism at the 5’ end of RNA transcript</td>
</tr>
<tr>
<td>TFBS</td>
<td>Transcription factor binding site</td>
</tr>
<tr>
<td>UCSC</td>
<td>University of California at Santa Cruz</td>
</tr>
<tr>
<td>UTR</td>
<td>Untranslated region</td>
</tr>
</tbody>
</table>
ACKNOWLEDGEMENTS

Dr. Marco Marra for supporting and directing my research during graduate studies at the BCCA Genome Sciences Centre.

Dr. Connie Eaves and Dr. Pamela Hoodless for input and feedback as members of my thesis advisory committee.

Dr. Michael O'Connor and Dr. Connie Eaves for providing cultured human embryonic stem cells and differentiation cultures, and suggestions for qPCR assays.

Ryan Morin and Dr. Martin Hirst for providing assistance and orientation with starting this project.

The BCCA Genome Sciences Centre. I am indebted to numerous groups including Sequencing, Administration, Operations, and IT Systems, who have provided expert assistance and maintained state-of-the-art research laboratory facilities.

Genome British Columbia, Stem Cell Network Canada, and the National Cancer Institute (USA) for funding this research.

The Michael Smith Foundation for Health Research for a senior graduate trainee award.
1. INTRODUCTION

1.1 Human embryonic stem cell research

Since the first successful derivation and sustained culture of pluripotent human ESCs (Thomson et al. 1998), there has been growing interest in the research community (Guhr et al. 2006) to better understand and manipulate hESCs. Although the formal study of pluripotent cells can be traced back at least fifty years (Solter 2006), their defining properties of self-renewal and pluripotency have only recently raised expectations of myriad clinical and research applications. Potential uses exist for both undifferentiated hESCs and differentiated cell types derived from hESCs. There are an increasing number of reports into the latter and include the generation of cell types from all three primary germ layers. For example, chemical induction of hESCs with retinoic acid is an approach to make cells of the neural lineage for transplantation studies in animal models of nervous system damage (Ikeda et al. 2005).

1.2 Small protein gene discovery

Proteins and functional RNAs are important components defining cellular state and function, and so a thorough knowledge of the ESC transcriptome and proteome is a major goal of current research (Wang et al. 2006, Van Hoof et al. 2006, Baharvand et al. 2006, Unwin et al. 2003). The application of ESC products as therapeutic agents or models in laboratory research will be more effective with a comprehensive catalogue of the biomolecules present within such cells, along with an understanding of their expression patterns and interactions within or between individual cells.

Small proteins are important effectors in a variety of biological processes, including cell signalling, immunity, and metabolism. In regard to small-protein gene discovery, expert reviews
of large-scale cDNA sequencing projects showed that the longest ORFs did not necessarily correspond to the actual protein coding regions (Suzuki and Sugano 2006). Proteome-wide mass spectrometry has revealed large populations of small novel proteins, such as upstream open reading frames in 5' UTRs (Oyama et al. 2004). A genome-wide evaluation of short ORFs in eukaryotes for the purpose of discovering new proteins was first conducted in the yeast model system (Andrade et al. 1997) with the discovery of bonafide novel proteins. Thus, with important roles in basic cellular processes, and by extension in developing cell-based therapeutics, it is imperative that short proteins be well characterized.

However, the small size of short proteins has resulted in their being under-represented in large mammalian cDNA collections, which is evident by an artificial discontinuity in annotated protein-coding cDNAs below the 100 amino acid mark (Frith et al. 2006). This is in part a result of weak statistical confidence for short hits in sequence similarity searches, that can be compounded further for novel short proteins that lack similarity to known proteins or span a narrow phyletic range. Therefore, as a minimum, describing the short proteome needs to preclude artificial lower size-cutoffs in the transcriptome assay itself (e.g. cDNA size fractionation), or in cDNA annotation and de novo gene prediction methods. However, the analysis of protein-coding mRNA transcripts and, more recently, entire transcriptomes is a more feasible approach for the discovery of low abundance protein-coding genes. This is a consequence of our ability to manipulate nucleic acids in ways not yet possible with cellular proteins, including harvesting and assaying minute samples, molecular amplification, subtraction, and clonal propagation.
1.3 Transcriptome profiling using LongSAGE

Cross-platform comparisons for measuring transcript abundance have shown low correlation (Griffith et al. 2005, Oudes et al. 2005, van Ruissen et al. 2005) and thereby underscored the value of using multiple technologies to circumvent biases inherent in individual methods. In comparison to LongSAGE (Saha et al. 2002), or other cDNA tag-based methods to interrogate the growing complexity of mammalian transcriptomes (Carninci 2006), DNA microarray studies have been biased towards well characterized and relatively abundant transcripts, with little overlap observed in hESC-specific gene lists (Vogel 2003). It was observed that MPSS data (Brandenberger et al. 2004a), despite greater sampling depth, showed poorer coverage of transcriptomes when compared to well curated databases (Siddiqui et al. 2006). hESC EST data (Miura et al. 2004) was limited and not affordable on a scale comparable to that undertaken by shorter cDNA tag-based approaches such as SAGE. Because key molecular regulators may be of low abundance or transient in their expression, it is advantageous to utilize a comprehensive sampling assay like deep LongSAGE. A further improvement in sampling cDNA tag-based libraries is now achievable with next-generation sequencing technologies that have recently become affordable (Nielsen et al. 2006; Ng et al. 2006).

Bioinformatic resources for processing and mapping SAGE data have vastly improved since the first description of SAGE technology (Velculescu et al. 1995). Advances have included direct tag-to-genome mapping of LongSAGE tags (Saha et al. 2002), cross-species conservation analysis of novel tag regions (Schnerch 2005), and CMOST and p-value based quality filters (Siddiqui et al. 2006). These advances enable the processing of low-frequency tags to generate high confidence predictions for downstream functional studies. Furthermore, targeted rescue
experiments of transcripts corresponding to low abundance tags have demonstrated many such
tags to represent bonafide transcripts (Richards et al. 2006; Kim et al. 2006; Siddiqui et al. 2005;
Chen et al. 2002).

In prior work (Schnerch 2005), an analysis was performed on a meta-library of 2.6
million LongSAGE tags acquired from nine undifferentiated hESC lines which had been
independently derived in four different laboratories (Table 1). One aim of that analysis was to
identify tags enriched in the hESC metatlibrary, by comparison to publicly available human
SAGE libraries constructed from a variety of tissues. Enriched tags were filtered to select those
that mapped to human genomic sequence distant from known transcripts. Other criteria included
selecting tags from human genomic sequence that was orthologous to mouse or rat sequence.
The result was a filtered list of 301 novel hESC-enriched tags.

5' RACE experiments resulted in the rescue of 126 partial cDNA transcripts
corresponding to these novel tags (Hirst et al. 2007). My project was a continuation of these
hESC gene discovery efforts begun earlier in our department and complements similar efforts
elsewhere (Richards et al. 2006; Boheler and Tarasov 2006).

**Table 1.** Human embryonic stem cell lines used in LongSAGE

<table>
<thead>
<tr>
<th></th>
<th>Cell line</th>
<th>Number of tags</th>
<th>NIH registry</th>
<th>Passage</th>
<th>Feeder layer</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>H1 (Thomson, WI)</td>
<td>218,000</td>
<td>WA01</td>
<td>54</td>
<td>matrigel</td>
<td>XY</td>
</tr>
<tr>
<td>2</td>
<td>H1 (Thomson, WI)</td>
<td>276,000</td>
<td>WA01</td>
<td>31</td>
<td>irrMEFs, CF-1 mice</td>
<td>XY</td>
</tr>
<tr>
<td>3</td>
<td>H7 (Thomson, WI)</td>
<td>272,000</td>
<td>WA07</td>
<td>22</td>
<td>irrMEFs, CF-1 mice</td>
<td>XX</td>
</tr>
<tr>
<td>4</td>
<td>H9 (Thomson, WI)</td>
<td>468,000</td>
<td>WA09</td>
<td>38</td>
<td>irrMEFs, CF-1 mice</td>
<td>XX</td>
</tr>
<tr>
<td>5</td>
<td>H13 (Thomson, WI)</td>
<td>221,000</td>
<td>WA13</td>
<td>22</td>
<td>irrMEFs, CF-1 mice</td>
<td>XY</td>
</tr>
<tr>
<td>6</td>
<td>H14 (Thomson, WI)</td>
<td>212,000</td>
<td>WA14</td>
<td>22</td>
<td>irrMEFs, CF-1 mice</td>
<td>XY</td>
</tr>
<tr>
<td>7</td>
<td>ES 3 (Pera, Australia)</td>
<td>206,000</td>
<td>ES03</td>
<td>16</td>
<td>irrMEFS, B-81 mice</td>
<td>XX</td>
</tr>
<tr>
<td>8</td>
<td>ES 4 (Pera, Australia)</td>
<td>209,000</td>
<td>ES04</td>
<td>36</td>
<td>irrMEFS, B-81 mice</td>
<td>XY</td>
</tr>
<tr>
<td>9</td>
<td>HSF-6 (Firpo, CA)</td>
<td>190,000</td>
<td>UC06</td>
<td>50</td>
<td>irrMEFs, CF-1 mice</td>
<td>XX</td>
</tr>
<tr>
<td>10</td>
<td>hESBGN-01 (BresaGen, GA)</td>
<td>202,000</td>
<td>BG01</td>
<td>20</td>
<td>irrMEFS</td>
<td>XX</td>
</tr>
</tbody>
</table>
1.4 Hypothesis

Human ESC transcriptomes contain uncharacterized protein-coding transcripts that may be variably expressed in undifferentiated versus differentiated cell types and be candidate regulators of ESC function.

1.5 Objectives

My objectives were to discover novel ORFs encoded by rare mRNA transcripts detected in hESC transcriptomes, and assess regulation of transcript levels via measurement of transcript abundance in undifferentiated ES cells versus differentiated cell types.

The experimental approach was to perform the following:

One, conduct a bioinformatic analysis of 126 candidate partial transcripts (which, in earlier work, had yielded 5' RACE data from 301 novel hESC-enriched tags) and pick a subset for further study based on protein coding potential.

Two, recover 3' RACE products from the novel SAGE tags chosen for study in step one.

Three, perform a comprehensive analysis of full-length transcript sequences from the candidates chosen to evaluate their gene structure, gene regulatory features, and predicted protein structure.

Four, recover full-length cDNAs and ORFs for interesting candidates identified in step three.

The fifth and final aim was to determine whether any of the interesting candidates identified in steps three and four were present at different levels in differentiated cells as compared to undifferentiated hESCs.
2. AMPLIFICATION AND CLONING OF NOVEL cDNAs

2.1 Introduction

In earlier work (Hirst et al. 2007), 126 cDNA fragments were recovered by 5’ RACE from 301 novel hESC-enriched LongSAGE tags. Tags corresponding to these rescued fragments had a mean count of 2.5 tags per million (maximum 36 tpm), the rescued fragments had a mean length of 844 bp (maximum 2 kb), 47% were spliced, and 74% had independent EST support for transcription.

Twelve candidate transcripts were prioritized for 3’ RACE amplification after analysing the 5’ RACE sequence data using a subset of the features described in Section 3 (Analysis of novel transcript sequences). Criteria for selection included hits to known or poorly characterized protein domains, cross-species sequence conservation, sense-antisense tag pairs, putative single exon genes, and tags observed only in the deeply sampled H9 library (468,000 tags). Tags for these candidate transcripts were detected in one to six libraries, meta library tag counts were low (one to twelve tags in 2.6 million), and six transcripts were spliced.

My first task was to use standard RACE chemistry (Frohman et al. 1988; Figure 1) to rescue 3’ cDNA fragments overlapping 5’ RACE products generated earlier using Clontech 5’ SMART chemistry (complete experimental workflow is in Appendix 1).
**Figure 1.** Schematic of RACE chemistry. Outer (GSP2) and nested (NGSP2) 3' Gene-Specific Primers located upstream of a transcript's LongSAGE tag (and overlapping the 5' RACE product) were coupled with 3' oligo d(T)-based outer and inner primers to amplify the 3' ends of novel transcripts from reverse transcribed hESC total RNA.

2.2 Methods

Amplicons were generated using transcript-specific oligonucleotide primers (IDT Technologies, Coralville, IA; Appendix 4) and total RNA from the H9 cell line (WA09; Thomson 1998), which was one of the nine hESC lines that formed part of the original set for LongSAGE library construction and meta-analysis (Table 1). RNA was kindly provided by Dr. Michael O'Connor from the laboratory of Dr. Connie Eaves (Terry Fox Laboratory, BCCA). Quality was verified periodically over the course of this project using a 2100 Bioanalyzer and RNA 6000 NanoChips (Agilent Technologies, Palo Alto, CA). RNA integrity numbers were 9.0 or higher.
2.2.1 Reverse transcription and RACE chemistry

Reactions were conducted using a FirstChoice™ RLM-RACE kit (Ambion, Austin, TX) and comprised the following three steps:

(a) Reverse transcription

<table>
<thead>
<tr>
<th>Component</th>
<th>Concentration</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>H9 hESC total RNA</td>
<td>3.6 ug/uL</td>
<td>0.3 uL, 1.0 ug</td>
</tr>
<tr>
<td>dNTP mix</td>
<td>2.5 mM each</td>
<td>4.0 uL</td>
</tr>
<tr>
<td>3' RACE adapter</td>
<td>0.3 ug/uL</td>
<td>2.0 uL</td>
</tr>
<tr>
<td>10x RT buffer</td>
<td>10x</td>
<td>2.0 uL</td>
</tr>
<tr>
<td>RNase inhibitor</td>
<td>10 U/uL</td>
<td>1.0 uL</td>
</tr>
<tr>
<td>M-MLV Reverse Transcriptase</td>
<td>100 U/uL</td>
<td>1.0 uL</td>
</tr>
<tr>
<td>Nuclease-free water</td>
<td>n/a</td>
<td>9.7 uL</td>
</tr>
</tbody>
</table>

The 20 uL RT reactions were incubated at 42 °C for one hour using a Tetrad thermal cycler (MJ Research, Waltham, MA).

(b) Outer 3' RACE PCR

<table>
<thead>
<tr>
<th>Component</th>
<th>Concentration</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reverse Transcription reaction product</td>
<td>n/a</td>
<td>1.0 uL</td>
</tr>
<tr>
<td>10x PCR buffer</td>
<td>10x</td>
<td>5.0 uL</td>
</tr>
<tr>
<td>dNTP mix</td>
<td>2.5 mM each</td>
<td>4.0 uL</td>
</tr>
<tr>
<td>3' RACE gene-specific outer primer</td>
<td>10 uM</td>
<td>2.0 uL</td>
</tr>
<tr>
<td>3' RACE outer primer</td>
<td>10 uM</td>
<td>2.0 uL</td>
</tr>
<tr>
<td>SuperTaq™-Plus DNA polymerase</td>
<td>5 U/uL</td>
<td>0.25 uL</td>
</tr>
<tr>
<td>Nuclease-free water</td>
<td>n/a</td>
<td>35.75 uL</td>
</tr>
</tbody>
</table>
The 50 uL outer 3' RACE PCR reactions were thermally cycled as follows: 94°C for 3 minutes, followed by 35 cycles of (94°C for 30 seconds, 60°C for 30 seconds, 72°C for 90 seconds), and a final extension at 72°C for 7 minutes.

(c) Inner 3' RACE PCR

<table>
<thead>
<tr>
<th>Component</th>
<th>Concentration</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outer 3' RACE reaction product</td>
<td>n/a</td>
<td>1.0 uL</td>
</tr>
<tr>
<td>10x PCR buffer</td>
<td>10x</td>
<td>5.0 uL</td>
</tr>
<tr>
<td>dNTP mix</td>
<td>2.5 mM each</td>
<td>4.0 uL</td>
</tr>
<tr>
<td>3' RACE gene-specific inner primer</td>
<td>10 uM</td>
<td>2.0 uL</td>
</tr>
<tr>
<td>3' RACE inner primer</td>
<td>10 uM</td>
<td>2.0 uL</td>
</tr>
<tr>
<td>SuperTaq™-Plus DNA polymerase</td>
<td>5 U/uL</td>
<td>0.25 uL</td>
</tr>
<tr>
<td>Nuclease-free water</td>
<td>n/a</td>
<td>35.75 uL</td>
</tr>
</tbody>
</table>

The 50 uL inner 3' RACE PCR reactions were thermally cycled using the same parameters as the outer PCRs described above.

2.2.2 Topo™-TA cloning of RACE products and full-length transcripts

PCR products were electrophoresed in preparative gels using Ultrapure™ agarose (Invitrogen, Mississauga, ON) and DNA bands were visualized for excision using a non-UV Dark Reader transilluminator (Clare Chemical Research, Dolores, CO). Excised bands were purified with a MinElute Gel Extraction Kit (Qiagen, Mississauga, ON) using a standard microcentrifuge-based protocol from the manufacturer.
Purified amplicons were cloned into the pCR4-TOPO vector using a TOPO™-TA Cloning Kit for Sequencing (Invitrogen) and the following reaction setup (optimized for transformation using electrocompetent *E. coli*):

<table>
<thead>
<tr>
<th>Component</th>
<th>Concentration</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fresh PCR product</td>
<td>n/a</td>
<td>2.5 uL</td>
</tr>
<tr>
<td>Dilute salt solution</td>
<td>0.25x</td>
<td>1 uL</td>
</tr>
<tr>
<td>pCR4-TOPO vector</td>
<td>10 ng/uL</td>
<td>1 uL</td>
</tr>
<tr>
<td>Nuclease-free water</td>
<td>n/a</td>
<td>1.5 uL</td>
</tr>
</tbody>
</table>

The 6 uL cloning reactions were incubated at room temperature for 30 minutes. A 2 uL aliquot was then used for transforming 50 uL of One Shot™ TOP10 electrocompetent cells with 0.1 cm cuvettes and a GenePulser™ machine (Bio-Rad, Hercules, CA). Electroporated cells were immediately resuspended in 250 uL of SOC medium and incubated at 37°C for 1 hour with shaking at 225 rpm. Serial dilutions of each transformation were then plated individually on LB-Kanamycin (50 ug/mL) agar plates and incubated overnight at 37°C.

Resulting colonies were evaluated directly by colony PCR using Platinum™ Blue PCR SuperMix (Invitrogen) to verify insert size and quality of each cloning-transformation reaction. Colony PCR reaction conditions were as follows:

<table>
<thead>
<tr>
<th>Component</th>
<th>Concentration</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Platinum™ Blue PCR SuperMix</td>
<td>n/a</td>
<td>24 uL</td>
</tr>
<tr>
<td>M13 F primer</td>
<td>100 uM</td>
<td>0.1 uL</td>
</tr>
<tr>
<td>M13 R primer</td>
<td>100 uM</td>
<td>0.1 uL</td>
</tr>
<tr>
<td>Template - colony spiked</td>
<td>n/a</td>
<td>n/a</td>
</tr>
</tbody>
</table>
The reactions were thermally cycled as follows: 94°C for 10 minutes, followed by 30 cycles of (94°C for 30 seconds, 55°C for 30 seconds, 72°C for 60 seconds), and a final extension at 72°C for 10 minutes. Inserts were sized using analytical agarose gel electrophoresis, after accounting for the contribution of vector sequence (~130 bp) to the PCR amplicon size.

Twelve colonies per amplicon, from each of the 12 candidates targeted for full-length recovery, were picked into 384-well LB-Kanamycin plates and submitted to the Genome Sciences Centre (GSC) Sequencing Production Group for DNA sequencing.

2.2.3 Cloning full-length ORFs

Full-length ORFs were recovered using ORF-specific primers designed with flanking Gateway™ technology (Invitrogen) attB sites. Amplicons were cloned into the pDONR221 Gateway™ donor vector utilizing standard BP recombination reaction chemistry. Cloned ORFs were subsequently prepped and readied for transfer into the pEXP3-DEST vector using Gateway LR recombination chemistry. The expression-ready clones will be used in subsequent work for in vitro transcription and translation assays using the Expressway™ Lumio™ Cell-Free Expression and Detection System (Invitrogen).

2.3 Results and Discussion

The 3' RACE procedure required both outer and nested PCRs (Figure 2), given the observation that outer PCRs alone rarely yielded distinct amplicons. That was not surprising, given the low abundance of tags for the candidate transcripts in the hESC LongSAGE data (counts ranged from one to twelve in a meta-library of 2.6 million tags). Approximating a few hundred thousand transcripts per cell, these tag counts approach a high of one copy per cell.
**Figure 2.** Agarose gel showing nested PCR amplicons generated from four candidates using H9 hESC RNA and transcript-specific 3’ RACE primers. Sizing was achieved using a 1 kb Plus DNA Ladder (Invitrogen).

I recovered a total of 36 amplicons from the 12 candidates, with one to five distinct amplicons (75 bp to 2 kb) per candidate. Clones were checked for inserts by colony PCR (Figure 3) and then sequenced.

**Figure 3.** Agarose gel electrophoresis from colony PCR verification of cloned RACE products. Six colonies were tested per RACE amplicon cloned. The three sizes of colony PCR amplicons shown for candidate 3012 represent 3’ RACE products of 75 bp, 175 bp, and 200 bp. Sizing was achieved using a 1 kb Plus DNA Ladder (Invitrogen).

* 3’ RACE amplicon size. For colony PCR, add ~130 bp of vector
Full-length transcript sequences were reconstructed *in silico* for all candidate transcripts. Following a detailed bioinformatic analysis of these sequences (as described below in Section 3), I prioritized four candidates (3012, 3210, 3372, 3373) for further study and used transcript-specific primers to amplify and clone full-length cDNAs using the methods described above.
3. ANALYSIS OF NOVEL TRANSCRIPT SEQUENCES

3.1 Introduction

This project began with an analysis of 5' RACE data from the set of 126 novel hESC gene candidates that were successfully amplified by 5' RACE (Hirst et al. 2007). My objective was to shortlist a set of putative protein-coding transcripts for further study. The initial analysis included annotation of relevant empirical ChIP-chip data (Boyer et al. 2005) and scoring synonymous versus non-synonymous substitution rates ($K_a/K_s$) from pairwise alignments to the mouse genome (Nekrutenko et al. 2001). I was assisted in this regard by Mr. Ryan Morin (a CIHR Bioinformatics M.Sc. student at the GSC), who had tabulated hESC 5' RACE clone statistics. I subsequently conducted 3' RACE data processing and additional analyses that consisted of the nine activities described below (flowchart in Appendix 5). The strategy was to evaluate a variety of sequence features without restrictions on the size of transcripts or ORFs.

3.2 Methods

3.2.1. Raw 3' RACE DNA sequence was processed by trimming contaminating vector and RACE linker sequence, and bases of quality less than phred-20 (Ewing and Green 1998a; Ewing et al. 1998b) were also excluded. The presence and correct location of the LongSAGE tag was confirmed and clones that passed these criteria were assembled using overlapping 5' and 3' RACE sequence reads. The resulting full-length transcript sequences for each target region were analyzed for sequence polymorphism and alternative polyadenylation.

3.2.2. Full-length transcript sequences were subjected to sequence composition analysis, including features such as word repeats, in-frame triplets, and nucleotide frequencies.
3.2.3. Transcripts were analysed extensively in the context of the human genome using the UCSC Genome Browser (Kuhn et al. 2007) and supporting data tracks (http://genome.ucsc.edu/, March 2006 assembly). That activity included:

(a) BLAT (Kent 2002) analysis of transcripts and corresponding ORFs.

(b) assessing potential oligo d(T) primer mispriming of genomic DNA from trace DNA contamination of RNA preparations.

(c) checking regions of pseudogenes (Ashurst et al. 2005; Karro et al. 2007) for overlap with the novel transcribed regions under study.

(d) documenting genomic repeat elements according to the RepBase library (Jurka 2000).

(e) checking for segmental duplications, defined as duplications over 1 kb of non-RepBase masked sequences (Bailey et al. 2001).

(f) assessing cross-species sequence conservation using 17 vertebrate species, based on the phastCon model (Siepel et al. 2005).

(g) evaluating proximal regulatory sequences using the Transfac database (Matys et al. 2006), CpG island (Gardiner-Garden and Frommer 1987) and DNaseI-hypersensitivity (Crawford et al. 2004) tracks, canonical TATA and CAAT box signals in upstream regions, and polyadenylation signals at transcript 3' ends. Additional regulatory feature data included aligned transcription factor binding sites predicted from ChIP-chip experiments in ES cells (Boyer et al. 2005) for two key stem cell transcription factors, NANOG and SOX2.

(h) documenting splicing with canonical GT/AG splice signals (Clark and Thanaraj 2002).
3.2.4. Open reading frames

(a) ORF discovery was performed using the NCBI ORF Finder program (http://www.ncbi.nlm.nih.gov/gorf/gorf.html).

(b) Predicted ORFs were assessed for codon usage bias using the Graphical Codon Usage Analyzer program (Fuhrmann et al. 2004), which computed the frequency of usage of each codon versus a reference codon usage table derived from well characterized human protein-coding genes (http://gcua.schoedl.de).

3.2.5. Sequence similarity searches were conducted for both full-length transcripts and predicted ORFs using primarily the BLAST family (Altschul et al. 1990; McGinnis and Madden 2004) of algorithms (http://www.ncbi.nlm.nih.gov/BLAST/).

(a) TBLASTX, searching the translated human EST db using translated transcripts.

(b) TBLASTN, searching the translated human EST db using novel ORF aa sequences.

(c) BLASTX, searching the non-redundant protein db using translated transcripts.

(d) BLASTP, search the non-redundant protein db using predicted ORF aa sequences.

To find regions of low similarity for diverged sequences and detect distant evolutionary relationships, the following exhaustive search algorithms were applied:

(e) PSI-BLAST (Altschul et al. 1997), searching the non-redundant protein db using novel ORF aa sequences.

(f) MPsrch (http://www.ebi.ac.uk/MPsrch/), searching the UniProt Knowledgebase (Bairoch et al. 2005) using novel ORF aa sequences.

3.2.6. Domain and motif searches were performed for predicted ORFs using the Conserved Domain Database (Marchler-Bauer et al. 2005) and the InterPro collection (Mulder et al. 2007)
of protein domains comprising 16 major data sources (http://www.ebi.ac.uk/interpro/). Short functional sequences were evaluated using the Eukaryotic Linear Motif resource (Puntervoll et al. 2003), which included motifs for peptide cleavage, ligand binding, PTMs, and subcellular targeting (http://elm.eu.org/).

3.2.7. The predicted 5' and 3' UTR regions of novel transcripts were checked for regulatory RNA sequence signals using the RegRNA tool (http://regrna.mbc.nctu.edu.tw/). That involved searching a collection of motifs from UTRdb (Mignone et al. 2005), including translational control signals and miRNA target sites.

3.2.8. Protein-level analyses were performed on predicted ORFs, primarily using software tools hosted by the ExPASy proteomics resource (Gasteiger et al. 2005) at the Swiss Institute of Bioinformatics (http://expasy.org/). Each ORF was queried using the comprehensive PredictProtein server (Rost et al. 2004; http://www.predictprotein.org/) and I evaluated the following features:

(a) protein physicochemical parameters, using the ExPASy ProtParam tool.
(b) secondary structure and transmembrane regions, using the PredictProtein programs PHD and PROF, TopPred (Claros and von Heijne 1994; http://bioweb.pasteur.fr/seqanal/interfaces/toppred.html), and PSIPRED (Bryson et al. 2005; http://bioinf.cs.ucl.ac.uk/psipred/).
(c) coiled segments, using Coils (Lupas 1996) in PredictProtein.
(d) globularity, using the PredictProtein tool Globe.
(e) cysteine bridges, using Disulfind (Ceroni et al. 2006) in PredictProtein.
(f) leucine zippers, using the program 2Zip (Bornberg-Bauer et al. 1998; http://2zip.molgen.mpg.de/).

(g) signaling and subcellular localization signatures, using the SignalP 3.0 and TargetP 1.1 programs (Emanuelsson et al. 2007; http://www.cbs.dtu.dk/services/).

(h) post-translational modifications, using ScanProsite (de Castro et al. 2006).

3.2.9. The final step in data analysis was to screen novel transcript sequences against non-coding RNA gene datasets to filter out hits to potential non-coding sequences. This involved BLAST searches against RNAdb (a mammalian non-coding RNA database; Pang et al. 2005), and querying Rfam, a database of non-coding RNA family alignments (Griffiths-Jones et al. 2005).

3.3 Results and Discussion

To strengthen the evidence that these novel tags and corresponding transcripts were bonafide products of hESC transcriptional activity and not technical artifacts of the experimental process, transcript sequences were studied in detail with respect to gene structure, protein-coding potential, and gene regulatory features. This comprehensive manual evaluation included features generally used in gene finding approaches for eukaryotes (Stormo 2000), such as splice donor and acceptor sites, start and stop codons, TATA boxes, poly(A) addition signals, ORF annotation, codon usage bias, predicted protein features, upstream genomic regulatory potential, conservation, and sequence similarity.

Following are seven key observations. One, novel predicted ORFs ranged in size from 22 to 129 amino acids (mean 66 aa). Two, there were no hits to well described InterPro protein domains and the three observed hits were to two uncharacterized domains and one domain of
unclear function (Table 2). Three, the most frequent independent support for expression at many of these novel transcribed regions was from an EST dataset (Brandenberger et al. 2004b) profiling both undifferentiated and differentiated human ESC samples. Four, GC content ranged from 26 to 54 % and genomic repeat elements (LINEs, SINEs, LTRs) comprised 0 to 88 % of transcript sequence. Five, there was no overlap observed with known pseudogenes. Six, splicing was evident in six of twelve candidates with exon counts ranging from two to eight and canonical splicing signals (GT/AG) in three candidates. Seven, the mean difference in codon usage frequency ranged from 18 to 26 % for selected ORFs.

The preceding observations suggest that finding novel predicted ORFs larger than 100-150 aa are unlikely in human ES cells with current cDNA databases and transcript sampling protocols. Perhaps deeper sampling for rare transcripts using next generation sequencing technologies and assaying highly purified stem cell populations may reveal longer novel proteins. What is also apparent in current cDNA and protein databases are many sparsely described or predicted proteins that require better characterization. Two of the three InterPro hits in this study belonged to the ‘uncharacterized’ category (Table 2). There was a wide range in genomic repeat content of novel transcripts, with some transcripts having high repeat content. For example, 56 % of the 609 bp transcript of candidate 3373 was comprised of an LTR element. Such transcripts may prove to be interesting in studies of genomic repeat sequences and gene evolution. Splicing sites were observed in half the transcripts studied and they represent examples of difficult de novo predictions from genomic DNA, underscoring the utility of cDNA sampling from rare cell populations for novel gene discovery. The unspliced transcripts may be good candidates for novel single exon genes, generally thought to be under-represented in
multicellular eukaryotic gene databases. The mean difference in codon usage frequency in predicted ORFs was within typical values for known genes. For example, human POU5F1 and NANOG values were 11% and 24%, respectively, using the same analysis method.

Table 2. InterPro domain hits to novel predicted ORFs

<table>
<thead>
<tr>
<th>Transcript</th>
<th>ORF size</th>
<th>InterPro entry</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3210</td>
<td>66 aa</td>
<td>PTHR12138</td>
<td>uncharacterized</td>
</tr>
<tr>
<td>3210</td>
<td>111 aa</td>
<td>PTHR16213</td>
<td>uncharacterized</td>
</tr>
<tr>
<td>3373</td>
<td>72 aa</td>
<td>PS01186</td>
<td>EGF-like</td>
</tr>
</tbody>
</table>

As noted above, a detailed analysis of twelve hESC novel transcripts did not yield transcripts encoding large ORFs or well described protein domains. This observation may not be surprising, given the context of searching for novel protein coding sequences in a species whose genome and transcriptome datasets have been subject to exhaustive gene prediction algorithms. Thus, my focus shifted to the evaluation of novel short proteins. I was, however, cognizant of the fact that hypothetical short ORFs can result from faulty annotation (Linial 2003) and thus require expert manual review. The scope of transcript and gene features that I reviewed exceeded those of automated prediction algorithms used in projects annotating large cDNA collections. For example, one recent international collaborative annotation project (Yamasaki et al. 2006) used three features (ORF prediction, similarity search, and protein motif prediction).

From the set of twelve candidates, four (3012, 3210, 3372, 3373) were shortlisted as most likely representing bonafide protein-coding transcripts. This selection was based on the presence of canonical ORF and splicing signals, EST support, $K_s/K_a$ and codon usage scores, upstream
genomic regulatory sequences, predicted protein structures, and hits to InterPro protein domain entries. Structures for these candidates are presented in Figure 4 (for 3012) and in the Appendices (7, 8, and 9, for 3210, 3372, and 3373, respectively), showing gene and transcript structure, predicted protein features, and gene regulatory features. As an example, I describe candidate transcript 3012 below.

### 3.3.1 Candidate transcript 3012

This transcript had a meta-library frequency of 3 LongSAGE tags from 2 libraries (WA01 and WA09), including one tag from a deeply sampled H9 hESC library (468,000 tags). Two exons (163 and 1606 bp) were separated by a 62 kb intron with transcript orientation in the sense direction (Figure 4b). Shown in the UCSC Genome Browser view (Figure 4d) are data tracks of known genes (ZNF281, 1.5 kbp upstream), pseudogenes (none), predicted exons (one upstream), RNA secondary structure prediction (mostly intronic), human and non-human ESTs (including a match to independent hESC EST data), predicted transcription factor binding sites (TFBS, mostly upstream in a block of conservation), cross-species conservation, and repeat elements (one LTR comprising 9% of the transcript sequence at the 5' end). There were CpG island and DNase-I hypersensitive sites 1 to 2 kb upstream, indicative of potential gene regulatory sequences.

The 3012 transcript (Figure 4a) was 1769 bp long with a predicted 65 amino acid ORF, a 41 bp 5’ UTR and 1530 bp 3’ UTR (sequences in Appendix 6). The canonical AUG initiation signal at the 42 bp position was the 5’-most AUG and the only one predicted with a strong sequence context. There were also canonical GT/AG splicing sites at the termini of the predicted 62 kb intron. The 65 aa ORF had a GC content (49%) which was higher relative to the complete
transcript, and a codon usage frequency mean difference score of 22%. There was empirical ChIP-chip evidence from hESCs (Boyer et al. 2005) for upstream NANOG and SOX2 transcription factor binding sites. The upstream promoter region had TATA and CAAT boxes, and there was a canonical polyadenylation signal 25 bp upstream of the 3' poly(A) tail. Furthermore, these gene sequence signals were found within canonical distances (Mathews and van Holde 1990) to the predicted gene or transcript.

ORF translation and protein structure predictions for 3012 suggested a 7.6 kDa single-transmembrane protein (Figure 4c), with an N-terminal signal anchor or peptide region, and four phosphorylation sites. There were a total of 13 short motifs in the protein coding region, including an N-arginine dibasic convertase cleavage site (an endopeptidase processing secreted proteins). In addition, regulatory motifs were predicted in the 5' and 3' UTRs, including four miRNA target sites. One of these (hsa-miR-128a) was found to be expressed in human HL-60 leukemia cells. Another (hsa-miR-197) was cloned from the Saos-2 human epithelial-like osteosarcoma cell line. And a third miRNA (hsa-miR-331) was a predicted homologue of a miRNA cloned from rat neuronal tissue and later found to be seven-fold down-regulated in human CNS tumor cell lines compared to normal tissue (Gaur et al. 2007). These predicted miRNA target sites may be of use in functional studies using RNA interference-mediated transcript knockdown.
Figure 4. Candidate 3012 (a) transcript (b) gene (c) predicted protein (d) Genome Browser view.

4 (a) 3012 transcript

4 (b) 3012 gene

Summary transcript-gene information

- 3 tags in 2 LongSAGE libraries (WA01, WA09)
- 64 kb predicted transcription unit
- GT/AG intron splice signals
- 40 % GC (ORF = 49 %)
- 22 % codon frequency mean difference
- Cytoplasmic polyadenylation element, K-Box, and Gamma Interferon Activated Inhibitor element in 3' UTR
- 100-fold less abundant in reference adult RNA versus hESCs
4 (c) 3012 predicted protein structure using the program TopPred. *P* denotes a site of phosphorylation. The N-terminal is cytoplasmic with a signal peptide or anchor region.

**Summary protein information**

- MW = 7665 Daltons, pI = 7.0
- ELM motifs = 13, including an NDR class endopeptidase cleavage site
- PTMs = 4 sites of phosphorylation
- PROFacc = 60 % aa solvent accessibility
- PROFsec protein class = mixed
- PHDhtm = one TM region (N cytoplasmic)
- GLOBE = globular protein
4 (d) 3012 UCSC Genome Browser view. The 1769 bp transcript and 65 aa ORF are marked. Data tracks shown include empirically derived NANOG and SOX2 binding sites, known and predicted genes, pseudogenes, predicted exons, RNA secondary structure predictions, mRNA and EST tracks, TFBS, cross-species sequence conservation tracks, and repeat elements.
Detailed analyses were similarly performed using cDNA sequences (Supplementary Table 1) from the other rescued novel transcripts. Five points of particular note were:

(i) Two candidates (3114 and 3630) were discredited due to evidence for artifactual transcripts arising from oligo d(T) mispriming of genomic DNA along a tract of adenines immediately downstream of the transcript sequence when aligned against human genome sequence (Table 3). This observation highlights the need for acquiring full-length novel cDNA sequences in order to exclude such artifacts which, based on this study, may account for one in six putative transcripts.

Table 3. Transcripts that may have resulted from oligo d(T) mispriming

<table>
<thead>
<tr>
<th>Transcript</th>
<th>Number of adenines</th>
<th>Context</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>3114</td>
<td>13</td>
<td>Genomic</td>
<td>Immediately 3' of aligned transcript</td>
</tr>
<tr>
<td>3630</td>
<td>21</td>
<td>Genomic</td>
<td>Immediately 3' of aligned transcript</td>
</tr>
</tbody>
</table>

(ii) Regions of segmental duplication were observed for candidate transcripts 3149 and 3630 (Table 4). Akin to the novel transcripts containing genomic repeat sequences, these segmentally duplicated regions may be interesting candidates for studying gene evolution.

Table 4. Transcripts in a region of segmental duplication

<table>
<thead>
<tr>
<th>Transcript</th>
<th>Chromosome (alignment score)</th>
<th>Size</th>
<th>Overlapping sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>3149</td>
<td>7q (100 %) 1 duplication in 14- (96.1 %)</td>
<td>4 kb</td>
<td>Intron of predicted TF and hypothetical thymus protein</td>
</tr>
<tr>
<td>3630</td>
<td>15q (99.8 %) 27 duplications in Y &amp; 15</td>
<td>1.3 to 83 kb</td>
<td>Segment of CSPG4</td>
</tr>
</tbody>
</table>

(iii) $K_d/K_s$ data yielded only two high-scoring segments, in transcripts 3149 and 3372. One of these segments overlapped a 101 aa ORF in 3372. Thus, this popular approach for recognizing
potential protein coding sequences by scoring regions of purifying selection, provided few clues in this dataset of novel transcripts.

(iv) The 72 aa ORF of candidate 3373 (Appendix 9c) contained five cysteine residues with two predicted disulfide bridges (Table 5). This is an interesting example, with disulfide bridge complexity reminiscent of SDPs (small disulfide-bonded proteins), a class including hormones, toxins and inhibitors. All are characterized by small size and complex disulfide connectivity (Kong et al. 2004). This protein also contained a PCSK cleavage site, belonging to the family of subtilisin-like proprotein convertases. These are expressed extensively in mammalian neural and endocrine cells, and play a key role in the proteolytic processing of neuropeptide and peptide hormone precursors.

Table 5. Novel ORFs with predicted disulfide bridges

<table>
<thead>
<tr>
<th>Transcript</th>
<th>ORF length</th>
<th>Number of cysteines</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>3372</td>
<td>101 aa</td>
<td>2</td>
<td>C23, C89</td>
</tr>
<tr>
<td>3373</td>
<td>72 aa</td>
<td>5</td>
<td>C17, C19, C30, C53, C72</td>
</tr>
</tbody>
</table>

(v) Sequence similarity searches, the mainstay of many gene finding approaches (Stormo 2000), did not contribute substantially to strengthening evidence for the candidate transcripts in this study. As noted earlier, this outcome is not unexpected when querying short sequences. It was a key reason why large cDNA annotation projects used size cutoffs of 100 aa when conducting gene finding exercises (Frith et al. 2006), to circumvent the weaker statistical confidence of short sequence similarity hits. Nevertheless, I performed comprehensive sequence similarity searches using both the novel transcript sequences and their predicted ORFs.
One example of an interesting result was the predicted 111 aa ORF of transcript 3210 (Table 6 and Figure 5 below). Searching the large UniProt Knowledgebase of proteins (Bairoch et al. 2005) returned a collection of high scoring hits, albeit only across a narrow phyletic range of primate sequences. The sequences mapped to distinct genomic loci. Thus, this set of predicted protein sequences dispersed over several chromosomes may represent an expansion in primate genomes. Interestingly, variation amongst these sequences occurred primarily at the N-terminus, suggestive of variant signal sequences. These have been found to be a common feature contributing to the diversity of protein output in the mouse transcriptome (Davis et al. 2006).

This 111 aa ORF also had a high (96 %) aa identity hit to a patented full-length human cDNA sequence (NCBI accession ABP27450), the subject of US patent number 7193069 (Isogai et al. 2007).
Table 6. FASTA similarity search results of the 111 aa ORF from candidate 3210 queried against UniProt protein database.

<table>
<thead>
<tr>
<th>Hit</th>
<th>Database ID</th>
<th>Source</th>
<th>Tissue</th>
<th>Human Chr. BLAT</th>
<th>Length aa</th>
<th>Identity %</th>
<th>Overlap aa</th>
<th>E value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Q9H387_Human</td>
<td>cDNA PRO2550</td>
<td>fetal liver</td>
<td>14+</td>
<td>118</td>
<td>75.3</td>
<td>93</td>
<td>1.1e-26</td>
</tr>
<tr>
<td>2</td>
<td>Q6ZSV2_Human</td>
<td>cDNA FLJ45189 fis</td>
<td>brain</td>
<td>22+</td>
<td>130</td>
<td>75.3</td>
<td>93</td>
<td>1.2e-26</td>
</tr>
<tr>
<td>3</td>
<td>Q6ZUU5_Human</td>
<td>cDNA FLJ43326 fis</td>
<td>NT2 cell line</td>
<td>5-</td>
<td>138</td>
<td>76.3</td>
<td>93</td>
<td>1.2e-25</td>
</tr>
<tr>
<td>4</td>
<td>Q9NX85_Human</td>
<td>cDNA FLJ20378 fis</td>
<td>ileal mucosa</td>
<td>6-</td>
<td>152</td>
<td>73.9</td>
<td>92</td>
<td>2.5e-25</td>
</tr>
<tr>
<td>5</td>
<td>Q8N266_Human</td>
<td>cDNA FLJ33873 fis</td>
<td>tongue tumor</td>
<td>3-</td>
<td>156</td>
<td>72.3</td>
<td>94</td>
<td>1.8e-23</td>
</tr>
<tr>
<td>6</td>
<td>Q6ZVX8_Human</td>
<td>cDNA FLJ41941 fis</td>
<td>pericardium</td>
<td>22+</td>
<td>193</td>
<td>69.2</td>
<td>91</td>
<td>2.1e-23</td>
</tr>
<tr>
<td>7</td>
<td>Q6ZUJ9_Human</td>
<td>cDNA FLJ43311 fis</td>
<td>NT2 neurons</td>
<td>X-</td>
<td>222</td>
<td>72.6</td>
<td>95</td>
<td>2.3e-23</td>
</tr>
<tr>
<td>8</td>
<td>Q4RSJ8_Macfa</td>
<td>cDNA QfA-12119</td>
<td>brain</td>
<td>X+</td>
<td>145</td>
<td>69.3</td>
<td>88</td>
<td>3.3e-23</td>
</tr>
<tr>
<td>9</td>
<td>Q6ZWD5_Human</td>
<td>cDNA FLJ42626 fis</td>
<td>amygdala</td>
<td>10+</td>
<td>142</td>
<td>71.0</td>
<td>93</td>
<td>8.9e-23</td>
</tr>
<tr>
<td>10</td>
<td>Q6ZTF6_Human</td>
<td>cDNA FLJ44704 fis</td>
<td>cerebellum</td>
<td>21+</td>
<td>165</td>
<td>68.5</td>
<td>92</td>
<td>1.2e-22</td>
</tr>
<tr>
<td>11</td>
<td>Q9P195_Human</td>
<td>cDNA PRO1722</td>
<td>fetal liver</td>
<td>11-</td>
<td>118</td>
<td>71.9</td>
<td>89</td>
<td>2.5e-22</td>
</tr>
<tr>
<td>23</td>
<td>Q9BGW3_Macfa</td>
<td>cDNA QfA-12187</td>
<td>frontal lobe</td>
<td>11-</td>
<td>122</td>
<td>66.3</td>
<td>89</td>
<td>1.3e-20</td>
</tr>
</tbody>
</table>

Figure 5. MView (Brown et al. 1998) sequence alignment of FASTA search results using the 111 aa ORF from candidate 3210 queried against the UniProt protein database. Shown are 80 of 111 residues starting at the 5' end. Residue highlighting is based on a consensus of amino acids with similar physicochemical properties.
4. RELATIVE QUANTITATION OF TRANSCRIPT ABUNDANCE

4.1 Introduction

As a first attempt to functionally characterize these novel transcripts, I evaluated transcript abundance in RNA samples from undifferentiated hESCs and differentiated cell types. This was conducted with real-time RT-qPCR, using relative quantitation assays with SYBR Green dye and transcript-specific primers. This nucleic acid quantitation approach is amenable to verifiable specificity and high sensitivity, with low input template requirements, scalable assays, and replicability of data for statistical analysis (Bustin 2002).

4.2 Methods

RNA samples utilized in this part of my study included qPCR human universal reference total RNA (Clontech, Mountain View, CA), undifferentiated H9 hESC total RNA, and total RNA samples from three timepoints in a retinoic acid-induced (RA-induced) differentiation protocol (Day 0, undifferentiated; Days 2 and 4, differentiated; Appendix 2). Cell samples were kindly provided by Dr. Michael O'Connor from the laboratory of Dr. Connie Eaves (Terry Fox Laboratory, BCCA). Total RNA was extracted using TRIzol reagent (Invitrogen, Burlington, ON) and quality was verified on a 2100 Bioanalyzer with RNA 6000 NanoChips (Agilent Technologies, Palo Alto, CA). RNA integrity numbers were 9.7 or higher (Appendix 3). Concentrations were verified using a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE) and equal masses of each RNA sample were used for DNase I treatment with DNA-free™ reagent (Ambion, Austin, TX).

A two-step RT-PCR approach was used, where equal amounts of DNase I treated total RNA were first subjected to reverse transcription using random hexamers and MultiScribe
Reverse Transcriptase as per the manufacturer’s protocol (Applied Biosystems, Foster City, CA). To prepare a dilution series of 1st strand cDNA products for generating relative standard curves, I assumed a 1:1 mass conversion in the reverse transcription reaction when labeling the diluted aliquots. Thus, a 2.0 microgram and 100 microlitre reverse transcription reaction was assumed to yield 2.0 micrograms of 1st strand cDNA per 100 microlitre of reaction product.

*Power SYBR™ Green* 2x dye based qPCR detection was performed on a 7900 HT SDS machine with SDS v2.2 software (Applied Biosystems). Assays were performed in 20 uL reaction volumes, in triplicate. Optimizations were conducted using a primer dilution series ranging from 50 to 300 nM, a template dilution series spanning four orders of magnitude from 10 pg to 100 ng, dissociation curve analysis, measurements of PCR efficiency and precision of replicates. I kept input template amounts for each qPCR reaction equivalent by ensuring equal amounts of total RNA added in the DNase I treatments and reverse transcription reactions, identical dilutions of the 1st strand cDNA products, and equal volumes of diluted 1st strand cDNA template per qPCR reaction. One large preparation of H9 hESC 1st strand cDNA was synthesized and diluted for use as a reference standard in all qPCR experiments conducted over the course of this study. qPCR reactions comprised 10 uL of 2x *Power SYBR™ Green* dye (Applied Biosystems), 5 uL of 0.6 uM each primer pair mix (IDT Technologies), and 5 uL (10 ng equivalent) of 1st strand cDNA template. Thermal cycling was perfomed as follows: 95°C for 10 minutes, then 40 cycles of (95°C for 15 seconds followed by 60°C for 1 minute), and a final dissociation curve stage comprising 95°C for 15 seconds followed by a 60-to-95°C slow ramp, (at a rate of 2%).
4.2.1 Primer selection

Up to four primer pairs were selected for each novel transcript, endogenous control gene, and marker of undifferentiated or differentiated cellular states using Primer3 software (Rozen and Skaletsky 2000) and published literature or primer databases (PrimerBank, http://pga.mgh.harvard.edu/primerbank/; qPrimerDepot, http://primerdepot.nci.nih.gov/; and RTPrimerDB, http://medgen.ugent.be/rtprimerdb/). For endogenous controls, I reviewed relevant hESC literature (Synnergren et al. 2007; Willems et al. 2006) to include genes considered better normalization controls in ESC-to-EB differentiation protocols. Likewise, markers of the undifferentiated and differentiated cellular states were selected based on published literature, including findings from a comprehensive study that characterised 59 hESC lines from 17 laboratories worldwide (The International Stem Cell Initiative; Adewumi et al. 2007). Novel transcript-specific primers were picked following an assessment of each transcript for regions of unique sequence. This involved a preliminary BLAST search of each novel sequence against human genomic, transcript, and EST DNA sequence databases to flag regions lacking hits of high similarity to any known transcribed sequences or multiple sites in the human genome. Following primer design, each primer pair was rechecked using the above specificity criteria. Other prioritized features in primer design included picking amplicons that spanned introns, avoiding pseudogenes or regions of repetitive genomic sequence, and avoiding sites of known sequence polymorphism. The final sets of primer pairs for each target used in this study are listed in Table 7 below.
### Table 7. Primers for qPCR assays

<table>
<thead>
<tr>
<th>Primer group</th>
<th>Forward</th>
<th>Reverse</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Endogenous controls</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAPDH</td>
<td>CTCTGCTCCTCCTGGTCGAC</td>
<td>TTAAGCGAGGGCTGGTGAC</td>
</tr>
<tr>
<td>ACTB</td>
<td>GCACAGGCTGGCCCTTTT</td>
<td>GGTGGGACGAGCGCACG</td>
</tr>
<tr>
<td>NUBP1</td>
<td>TGGCCAGCAGGAGGATGAAA</td>
<td>GTCTCCACGTCACCTGGAGA</td>
</tr>
<tr>
<td><strong>Novel transcripts</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3005</td>
<td>GACTCTTTGATGCCGGCTTTG</td>
<td>TCCCAAGCACGTGGTAGGATA</td>
</tr>
<tr>
<td>3010</td>
<td>ACTGCAAAAAGGAGCACAACAA</td>
<td>TAAGCCACTTGCGGGACTCCAG</td>
</tr>
<tr>
<td>3012</td>
<td>AGACCTTTCTCCTGGCTCTCTT</td>
<td>AAAATCGCCTGGCCCTTTTCTT</td>
</tr>
<tr>
<td>3042</td>
<td>AACATTTGACAGGCTTGGGA</td>
<td>TGGAGCTGGCAACTGACCT</td>
</tr>
<tr>
<td>3052</td>
<td>TACCTTCCGATGGTCGCTG</td>
<td>CATGTGACAAAGGAGGCAAG</td>
</tr>
<tr>
<td>3061</td>
<td>CCGTGCGCCTTTTTCTCCTC</td>
<td>TCAGCACAGACCGGAAAG</td>
</tr>
<tr>
<td>3116</td>
<td>GAAAGACGGCTGGCTGCCTCAG</td>
<td>CGTCAAGAAATAAAGGAGCAG</td>
</tr>
<tr>
<td>3194</td>
<td>TCTGTTGCGCAATAGACAG</td>
<td>TATCTGGGACCTACACACAG</td>
</tr>
<tr>
<td>3210</td>
<td>CAGGGGAGGCGCTAAGTACAGG</td>
<td>GGCAAGGAGGCAACAAACT</td>
</tr>
<tr>
<td>3210 F02</td>
<td>CACCCGAGATTTCTCGCTCTT</td>
<td>TGCAACATGAAAGGGTTCAG</td>
</tr>
<tr>
<td>3372</td>
<td>CGGAAGAGGAGCAGCTCATACA</td>
<td>TGCTCAAGGCTCCATCCAC</td>
</tr>
<tr>
<td>3373</td>
<td>AAAAGCATTGACCTCCTCTTGG</td>
<td>CCCCACTAGGATCAGCCCTC</td>
</tr>
<tr>
<td><strong>Markers of undifferentiated state</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>POU5F1</td>
<td>CTTGCTGCAGAAGTGCTGGAGGA</td>
<td>CTGCAAGTGTTGGGCTGGCA</td>
</tr>
<tr>
<td>NANOG</td>
<td>CAAAGGCAAAAACACACACCTT</td>
<td>TCTGCTGGAGGCTGAGTAT</td>
</tr>
<tr>
<td>LIN28</td>
<td>GATCAAAGGAGCACAGTGCTAC</td>
<td>CTTGCAAGGCAATGGCTGAA</td>
</tr>
<tr>
<td>DNMT3B</td>
<td>CCAATCTGAGGAGGCTCATCAG</td>
<td>ACTGGGGGCTGCAAGCCAT</td>
</tr>
<tr>
<td>GDF3</td>
<td>TGCCTACATTAGAGGACTGGG</td>
<td>CAGGAGGAAGCTGGAGAAAT</td>
</tr>
<tr>
<td><strong>Markers of differentiated state</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AFP</td>
<td>ACTGAATCCAGAACACTCGCATAG</td>
<td>GCTTCTTGAAAGGACTGAG</td>
</tr>
<tr>
<td>T (BRACHYURY)</td>
<td>CCCTATGCTCTACACCGAAACAA</td>
<td>CAATTGCTATGAGGACTCAT</td>
</tr>
<tr>
<td>NEOFH</td>
<td>CAGGACCTGCTCATAATGCA</td>
<td>CAAAGCAGCAGACACCTT</td>
</tr>
<tr>
<td>NEUROD1</td>
<td>GCCCAGGCTTGTAGGAGC</td>
<td>CTGCTGGACTAGTTGGTCA</td>
</tr>
<tr>
<td>CDX2</td>
<td>GGCAGCAAGGGTAAGCAACAG</td>
<td>GGTGATGTAGCGGAGACTGGA</td>
</tr>
</tbody>
</table>

#### 4.2.2 qPCR data analysis

Raw qPCR data was processed for relative quantitation using a relative standard curve method (Performing Relative Quantitation of Gene Expression Using Real-Time Quantitative PCR, http://www.appliedbiosystems.com/support/apptech/). This approach does not require PCR.
efficiencies of target and endogenous control primers to be equivalent and gives more accurate quantitative results, especially for low-expression-level fold changes. Triplicate data points were averaged and standard deviations calculated. The same preparation of undifferentiated H9 hESC total RNA was used as the reference sample for generating a relative standard curve for every primer set on each qPCR reaction plate. The Day-0 (undifferentiated) hESC sample used in the retinoic acid-induced differentiation experiment served as the calibrator sample and the basis for comparing changes in relative expression during the differentiation time course.

4.3 Results and Discussion

To evaluate gene regulation at the level of transcript abundance, I conducted real-time RT-qPCR relative quantitation assays using SYBR Green dye and transcript-specific primers. First, the assays were optimized using undifferentiated H9 hESC total RNA and qPCR universal reference human adult RNA to pick the best primer pair per target, after analysis of dissociation curves, PCR efficiency, precision, and dynamic range.

Next, to assess the functional significance of these uncharacterized transcripts in a biological context, I measured relative transcript abundance in an experiment involving the differentiation of human ES cells using a RA-induced differentiation protocol. A comparison was also conducted using human adult universal reference RNA versus undifferentiated H9 hESC total RNA.

I observed five trends of differential expression when evaluating the abundance of novel transcripts in undifferentiated (Day-0 and reference H9 hESC) versus differentiated (Days 2, 4, and human adult universal reference RNA) cell types (example calculation in Appendix 10). These trends are summarized in Table 8 below, with the most dramatic changes being a 500-fold
lower abundance of two transcripts (3372, 3373) in human adult reference RNA, and a 35-fold higher abundance of one transcript (3005) in human adult reference RNA.

Interestingly, some novel transcripts differed in abundance even between the two differentiated cell samples (RA-induced days 2 or 4 versus adult reference RNA). One set (3012, 3052, 3372, 3373) had expression patterns suggestive of transcripts restricted to early cellular differentiation events, and the other (3061) may be relevant in differentiated adult cells.

In contrast to changing expression patterns seen for the majority of transcripts, two transcripts (3194, 3210_F02) had relatively stable levels of expression across all samples and so may be regulated by an alternative mechanism. 3210_F02 represents an isoform whose poly(A) signal was located at a canonical distance upstream from the poly(A) tail, in contrast to the two other isoforms of candidate transcript 3210.

To address the question of reliability of a single housekeeping gene as sufficient for an endogenous control in ESC differentiation experiments, a panel of three such genes (GAPDH, ACTB, NUBP1) were used in all relative quantitation assays. I observed similar performance across this panel of genes, with the GAPDH assay yielding more consistent measurements.
Table 8. Relative quantitation of transcript abundance via RT-qPCR, using GAPDH as an endogenous control. An ESC cellular differentiation experiment was sampled at Days 0, 2, and 4. RNA from independently cultured H9 hES cells served as the reference sample for generating relative standard curves, and the Day-0 timepoint was used as the calibrator sample against which fold changes in transcript abundance were measured. Novel transcript levels were also compared between human adult reference RNA (Clontech) and H9 hESCs.

<table>
<thead>
<tr>
<th>Transcript</th>
<th>Day-2 fold difference from Day-0 hESCs</th>
<th>Day-4 fold difference from Day-0 hESCs</th>
<th>Universal reference adult RNA fold difference from H9 hESC</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>3372</td>
<td>2.0 ±0.3</td>
<td>1.0 ±0.1</td>
<td>0.002 ±0.000</td>
<td>much lower in adult differentiated cells</td>
</tr>
<tr>
<td>3373</td>
<td>0.9 ±0.1</td>
<td>0.3 ±0.0</td>
<td>0.002 ±0.001</td>
<td>much lower in adult differentiated cells</td>
</tr>
<tr>
<td>3012</td>
<td>2.1 ±0.2</td>
<td>1.5 ±0.1</td>
<td>0.04 ±0.01</td>
<td>lower in Days 2 and 4, but higher in adult differentiated cells</td>
</tr>
<tr>
<td>3052</td>
<td>4.0 ±0.6</td>
<td>1.5 ±0.1</td>
<td>4.7 ±1.0</td>
<td>much higher in adult differentiated cells</td>
</tr>
<tr>
<td>3061</td>
<td>0.2 ±0.0</td>
<td>0.8 ±0.2</td>
<td>35.2 ±3.9</td>
<td>higher in differentiated cells</td>
</tr>
<tr>
<td>3005</td>
<td>1.6 ±0.2</td>
<td>3.0 ±0.6</td>
<td>2.1 ±0.3</td>
<td>higher in differentiated cells</td>
</tr>
<tr>
<td>3010</td>
<td>4.1 ±0.6</td>
<td>4.4 ±0.2</td>
<td>7.9 ±0.9</td>
<td>similar across cell types</td>
</tr>
<tr>
<td>3042</td>
<td>0.9 ±0.1</td>
<td>2.3 ±0.2</td>
<td>4.1 ±1.1</td>
<td></td>
</tr>
<tr>
<td>3210</td>
<td>1.0 ±0.2</td>
<td>2.6 ±0.8</td>
<td>3.4 ±0.9</td>
<td></td>
</tr>
<tr>
<td>3116</td>
<td>1.7 ±0.2</td>
<td>4.5 ±0.5</td>
<td>1.5 ±0.2</td>
<td></td>
</tr>
<tr>
<td>3194</td>
<td>1.5 ±0.4</td>
<td>0.4 ±0.1</td>
<td>0.8 ±0.3</td>
<td></td>
</tr>
<tr>
<td>3210_F02</td>
<td>0.7 ±0.1</td>
<td>1.2 ±0.1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
5. CONCLUSION AND FURTHER WORK

As has been generally observed in the field of transcriptional profiling over the last several years, there can be abundant diversity in the transcriptional output of mammalian genomic loci. Although tested on a much smaller scale, this phenomenon was evident from the extent of sequence variation observed in the 3’ RACE data from this study. Nine of twelve candidate transcripts showed alternative polyadenylation and eleven exhibited sequence polymorphisms, ranging in size from single nucleotide variants to indels of over 100 bp amongst transcript isoforms. Functionally evaluating such transcriptional diversity in a cell-specific and developmental context remains a formidable challenge.

When conducting detailed analyses of novel transcribed sequences, it can be difficult scoring evidence for bonafide transcripts. Consider, for example, the case of codon usage scores. On the one hand, assigning cut-off values for codon usage scores may help exclude apparent transcripts with poor predicted translation efficiency. But problematically, there may be bonafide novel ORFs having low to intermediate scores, with a less-than-average codon translation efficiency being a way of modulating translational output in a cell. Thus, novel protein-coding gene discovery today in the heavily mined human genome and transcriptome datasets is a particularly difficult task, given that more readily discernible clues like deep sequence conservation or regions with canonical gene signals have already been annotated.

On assessing novel transcript abundance, I found much higher fold differences when comparing human adult universal reference RNA to undifferentiated hESCs than in an RA-induced differentiation time course. A broader interpretation of the former comparison was rendered difficult from insufficient (proprietary) information about the types and proportions of
adult tissues contributing to the reference RNA preparation. The product was marketed as a human adult reference RNA sample optimal for qPCR studies and purported to represent transcripts from a wide variety of adult tissues. Nevertheless, the reference RNA can be considered to be representative of many adult differentiated cell types and devoid of large contaminating pools of stem cells. By comparison, resolution in the RA-induction assay may suffer from heterogeneity in the cell sample populations, since neither are hESC cultures comprised entirely of undifferentiated cells, nor are differentiating cultures devoid of undifferentiated cells. Conducting qPCR assays on more homogeneous and well defined cell samples, such as a neural cell line, would be a logical next step to resolve the above issues.

Candidate transcripts 3372 and 3373 were particularly interesting in that both exhibited similar expression profiles (500-fold less abundant in adult cells) and had protein structure predictions that included cysteine disulfide bridges. In addition, both had proprotein peptide cleavage sites, with an NDR endopeptidase cleavage site in 3372 and a PCSK cleavage site in 3373 (of functional importance in mammalian neural and endocrine cells). Perhaps these novel proteins form cysteine bridges with one another or other proteins to form larger order functional structures. A protein-protein interaction study would be particularly informative in this regard.

Another compelling example of a novel transcript with altered expression upon cell differentiation was candidate 3042 (2- to 8-fold increase). In earlier work (Schnerch 2005), the genomic coordinate orthologous to 3042 in mouse was found to transcribe an uncharacterized cDNA in bone marrow (Riken 9430067K14), and the corresponding tag was present in mouse ESC SAGE data. Cross-species conserved gene expression in orthologous genomic coordinates,
and in the same cell type, is strong validation for a bonafide transcript. This candidate transcript would be well suited for further functional studies using a mouse model.

Similarly, the mouse genomic regions orthologous to candidates 3052 and 3116 were found to be transcribed in mouse (although without evidence for a tag in mouse ESC SAGE data). In addition, a segment of the 97 aa ORF from 3052 scored a strong BLASTP hit to ECEL1, a member of the peptidase family M13. 3052 expression was also supported by independent hESC EST data (Brandenberger et al. 2004b).

Thus, the collective evidence from successful cDNA rescue of transcripts corresponding to novel hESC tags, transcript and gene structure signals, gene regulatory features, predicted protein features, and empirical data from quantitation of transcript abundance, all suggest that many of these novel candidates are bonafide transcripts that may have roles in stem cell biology. Subsequent experiments using cloned expression-ready novel ORFs in an \textit{in vitro} transcription and translation assay will provide further evidence for these cryptic transcripts present in human ES cells. Other possible functional assays include transcript knockdown using RNA interference protocols, targeting the predicted miRNA sites found in the 3’ UTRs of some of these novel transcripts.

However, despite the utility of the experimental approach for novel gene discovery adopted in this study, there are many technical and biological limitations to consider when interpreting such findings. These include the impact of cellular heterogeneity in ESC cultures defined as ‘undifferentiated’, mouse feeder cell contamination, false negatives for known and novel transcripts due to the absence of an NlaIII tagging site, oligo(dT) mispriming in the LongSAGE and 3’ RACE protocols, noise from sampling RNA molecules other than mature
mRNAs (e.g. pre-mRNAs) that are present in total RNA preparations, presence of truncated transcripts captured by the 5' RACE SMART chemistry, inability to distinguish between transcripts generating the same tag, tags spanning splice junctions, transcribed pseudogenes, ambiguities in tag-to-transcript mapping due to RNA editing and fusion transcripts, conserved but non-coding transcripts, scoring non-canonical signals of gene structure in putative transcripts, and challenges with sampling transiently expressed transcripts.

Nevertheless, the discovery and validation of novel transcribed sequences from cDNA tag-based assays is an important contributor to our understanding of the transcriptional output of mammalian genomes. The approach is gaining power with much deeper sampling capabilities offered by next-generation sequencing technologies. The resulting biomolecular knowledge resource should help refine protocols for the controlled genetic manipulation and evaluation of stem cells and their derivatives in research and therapeutic applications.
BIBLIOGRAPHY

embryonic stem cell lines by the International Stem Cell Initiative. Nat Biotechnol 25:
803-816.

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment

Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman.

Characterization of new proteins found by analysis of short open reading frames from the


duplications: organization and impact within the current human genome project
Nucleic Acids Res 33: D154-159.

Boheler, K.R. and K.V. Tarasov. 2006. SAGE analysis to identify embryonic stem cell-


Kumar, H.L. Murray, R.G. Jenner et al. 2005. Core transcriptional regulatory circuitry in

Brandenberger, R., I. Khrebtukova, R.S. Thies, T. Miura, C. Jingli, R. Puri, T. Vasicek, J.
Lebkowski, and M. Rao. 2004a. MPSS profiling of human embryonic stem cells. BMC

Brandenberger, R., H. Wei, S. Zhang, S. Lei, J. Murage, G.J. Fisk, Y. Li, C. Xu, R. Fang, K.
Guegler et al. 2004b. Transcriptome characterization elucidates signaling networks that

multiple alignment viewer. Bioinformatics 14: 380-381.

structure prediction servers at University College London. Nucleic Acids Res 33: W36-
38.

Bustin, S.A. 2002. Quantification of mRNA using real-time reverse transcription PCR (RT-


Hio, K. Otsuka et al. 2007. Full-length cDNA, sequence 2465, NCBI accession
ABP27450. US Patent 7193069.

Trends Genet 16: 418-420.


comparative modeling server for small disulfide-bonded proteins. Nucleic Acids Res 32:
W356-359.

Sugnet, M. Stanke, K.E. Smith, A. Siedel et al. 2007. The UCSC genome browser

21: 298-300.

525.


APPENDICES

Appendix 1. Experimental work flowchart

1. Selection of 12 novel candidates from set of 126 with 5' RACE data
2. 5' RACE sequence-based primer design for 3' RACE
   - Outer & nested 3' RACE PCRs
   - Gel & column purification, 36 amplicons
   - TA cloning & insert validation by colony PCR, 36 amplicons
   - DNA sequencing, 32 amplicons by GSC Sequencing Production Group (12 clones/amplicon/target)
3. Reverse transcribe H9 hESC total RNA
   - 3' RACE-ready cDNA
4. Raw data processing and bioinformatic analysis of 12 full-length cDNAs
5. Analysis of other candidates from set of 126 for (short) protein coding potential
6. FL cDNA recovery and cloning 4 protein-coding candidates
7. Differential expression analysis
   - Undifferentiated H9 vs. Universal reference adult RNA using RT-qPCR, 12 candidates
8. Expression-ready ORF clones for cell-free protein translation assay
Appendix 2. hESC culture and differentiation

Contributions

Cell culture, differentiation experiments, and cell-based assays were done by Dr. Michael O'Connor in the laboratory of Dr. Connie Eaves (Terry Fox Laboratory, BC Cancer Research Centre), who kindly provided samples and protocol notes. I conducted RT-qPCR assays for markers of undifferentiated and differentiated cellular states.

(a) Cells

H9 and H1 cells were purchased from WiCell (Madison, WI, USA). Mouse embryo fibroblasts (MEFs) were a generous gift from StemCell Technologies, Inc. (Vancouver, BC, Canada). Approval for use of these cells as described was obtained from the Canadian Stem Cell Oversight Committee. Unless otherwise stated, all reagents were obtained from StemCell Technologies.

(b) Maintenance hESC cultures

Undifferentiated hESCs were cultured in maintenance medium consisting of DMEM/F12 supplemented with 20% Knockout Serum Replacer (Invitrogen, Carlsbad, CA), 0.1 mM beta-mercaptoethanol (Sigma, St. Louis, MO), 0.1 mM non-essential amino acids, 1 mM glutamine and 4 ng/mL human FGF2, either in contact with mitotically inactivated MEFs or in MEF-conditioned maintenance medium on dishes coated with Matrigel (BD Biosciences, San Jose, CA), and passaged enzymatically using 1 mg/mL collagenase (Invitrogen) every 7 days as previously described (Thomson et al. 1998; Xu et al. 2001).
(c) hESC differentiation cultures

For all-trans retinoic acid (RA) induced differentiation, the medium on adherent hESCs in 5- to 7-day old maintenance cultures was replaced with fresh maintenance medium lacking FGF2 and supplemented with $10^{-5}$ M RA (Draper et al. 2002; Henderson et al. 2002). The cells were then cultured for up to 5 days with daily medium changes. RNA samples were collected in TRIzol (Invitrogen) from the initial undifferentiated hESCs and then from the differentiating cells, each day up to 5 days post-initiation of RA differentiation.

All the undifferentiated and RA-differentiated samples collected were analyzed for pluripotency-related antigens via flow cytometry as well as colony-forming cell assays.

(d) Flow cytometry

Antigen expression was analysed by flow cytometry as described (Ungrin et al., in press) using the following unconjugated antibodies: anti-stage-specific embryonic antigen-3 (SSEA3), anti-SSEA4, and anti-SSEA1, all from the Developmental Studies Hybridoma Bank (Iowa City, IA); TRA-1-60 and TRA-1-81 (Abcam, Cambridge, MA); and anti-Oct3/4 (BD Biosciences). The respective secondary antibodies used were: APC-conjugated anti-IgM (Jackson ImmunoResearch, West Grove, PA), FITC-conjugated anti-IgG3, and FITC-conjugated anti-Ig (BD Biosciences); FITC-conjugated anti-IgM (BD Biosciences); and FITC-conjugated anti-IgG1 (BD Biosciences). Cells were analysed using a FACSCalibur (BD) and FlowJo software (Tree Star, Ashland, OR) with gates for positive cells set such that they excluded > 99% of events detected when the same cells were stained only with the appropriate fluorochrome-labelled secondary antibody.
(e) Colony assays

For primary colony-forming cell (CFC) assays, cultures were incubated with 0.05% trypsin or TrypLE reagent (Invitrogen) for 10 minutes at 37°C. Cell suspensions were filtered through a 40 μm cell strainer and counts of viable cells performed in a hemacytometer after staining the cells in 0.1% nigrosine (Sigma). Cells were plated onto MEFs in maintenance medium containing 4 ng/mL FGF2 or Matrigel-coated plates with MEF-conditioned maintenance medium containing 4 ng/mL FGF2, and were then cultured for 7 days. For AP detection, cells were fixed and then stained using an Alkaline Phosphatase kit (Sigma) as recommended by the manufacturer.

(f) RT-qPCR

The abundance of mRNA markers (Table 7), routinely used in assessing pluripotent or differentiated cellular states, were measured to complement flow cytometry assays for pluripotency-related antigens and colony-forming cell assays.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Day-2 fold difference from Day-0 hESC</th>
<th>Day-4 fold difference from Day-0 hESC</th>
<th>Predicted trend</th>
<th>Observed trend</th>
</tr>
</thead>
<tbody>
<tr>
<td>POU5F1</td>
<td>0.53 ±0.02</td>
<td>0.40 ±0.02</td>
<td>Downregulated upon differentiation</td>
<td>As predicted, except for LIN28</td>
</tr>
<tr>
<td>NANOG</td>
<td>0.32 ±0.01</td>
<td>0.18 ±0.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIN28</td>
<td>1.46 ±0.06</td>
<td>1.64 ±0.08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNMT3B</td>
<td>0.70 ±0.03</td>
<td>0.45 ±0.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GDF</td>
<td>0.60 ±0.08</td>
<td>0.27 ±0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AFP</td>
<td>44.68 ±3.22</td>
<td>58.5 ±3.4</td>
<td>Upregulated upon differentiation</td>
<td>As predicted, except for BRACHYURY</td>
</tr>
<tr>
<td>BRACHYURY (T)</td>
<td>0.05 ±0.01</td>
<td>0.03 ±0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NEFH</td>
<td>2.36 ±0.64</td>
<td>1.82 ±0.24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NEUROD1</td>
<td>0.72 ±0.41</td>
<td>4.34 ±0.60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDX2</td>
<td>55.84 ±4.50</td>
<td>160.2 ±8.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CD30 N-terminus</td>
<td>0.40 ±0.03</td>
<td>0.14 ±0.04</td>
<td>Unknown</td>
<td>Day-0 level elevated relative to days 2 or 4</td>
</tr>
<tr>
<td>CD30 C-terminus</td>
<td>0.48 ±0.08</td>
<td>0.16 ±0.03</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix 3. RNA quality electropherogram of Day-2 and Day-4 RA-induced cells generated using an Agilent Bioanalyzer

### Overall Results for sample 2: Day 2
- RNA Area: 203.7
- RNA Concentration: 141 ng/µl
- rRNA Ratio [28S / 18S]: 1.6
- RNA Integrity Number (RIN): 10.0 (8.02.03)
- Result Flagging Color: RIN:10
- Result Flagging Label:

### Fragment table for sample 2: Day 2
<table>
<thead>
<tr>
<th>Name</th>
<th>Start Size [nt]</th>
<th>End Size [nt]</th>
<th>Area</th>
<th>% of total Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>18S</td>
<td>1,512</td>
<td>1,994</td>
<td>41.6</td>
<td>20.4</td>
</tr>
<tr>
<td>28S</td>
<td>3,250</td>
<td>4,098</td>
<td>68.4</td>
<td>33.6</td>
</tr>
</tbody>
</table>

### Overall Results for sample 3: Day 4
- RNA Area: 209.8
- RNA Concentration: 200 ng/µl
- rRNA Ratio [28S / 18S]: 1.8
- RNA Integrity Number (RIN): 9.7 (8.02.03)
- Result Flagging Color: RIN:9.70
- Result Flagging Label:

### Fragment table for sample 3: Day 4
<table>
<thead>
<tr>
<th>Name</th>
<th>Start Size [nt]</th>
<th>End Size [nt]</th>
<th>Area</th>
<th>% of total Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>18S</td>
<td>1,488</td>
<td>1,989</td>
<td>53.8</td>
<td>18.8</td>
</tr>
<tr>
<td>28S</td>
<td>2,860</td>
<td>4,171</td>
<td>90.0</td>
<td>33.1</td>
</tr>
</tbody>
</table>
### Appendix 4. Primers for 3’ RACE, Full-Length cDNA, and ORF recovery

<table>
<thead>
<tr>
<th>Novel transcript</th>
<th>3’ RACE Outer</th>
<th>3’ RACE inner</th>
</tr>
</thead>
<tbody>
<tr>
<td>3012</td>
<td>TTGCTTTAGAATGCGCGATT</td>
<td>TAAAAGGGGCTTGAGCCATCA</td>
</tr>
<tr>
<td>3114</td>
<td>CTTCATTTACATTTGCCATGC</td>
<td>CTTCATTTACATTTGCCATGC</td>
</tr>
<tr>
<td>3149</td>
<td>TATATCCCATCCCATCTG</td>
<td>ATCCCATCTCCTGAGTCA</td>
</tr>
<tr>
<td>3210</td>
<td>CGCAGGCTTTGCATGAAAT</td>
<td>GGCCCTTTAATCTTCTGTT</td>
</tr>
<tr>
<td>3372</td>
<td>AAGGCGCTTACATTGGTT</td>
<td>AGGAAGGCTTTGCCCCAA</td>
</tr>
<tr>
<td>3373</td>
<td>TCCTTTAATGCGGATCTTT</td>
<td>GCAATTGCTCTTCTGCTTT</td>
</tr>
<tr>
<td>3399</td>
<td>AACATCCTTAGCACATCTGGGAAAC</td>
<td>GCATCTGGGAGACACCTAGT</td>
</tr>
<tr>
<td>3443</td>
<td>TCACAGTGGACAGCTCAGC</td>
<td>CAGTGCGACTGACGCCC</td>
</tr>
<tr>
<td>3461</td>
<td>TGGGCATCTCCACACATAAA</td>
<td>CATCATACGCCACCCAA</td>
</tr>
<tr>
<td>3465</td>
<td>TGGGAGAGGAGGGTTCCTTA</td>
<td>CAGCTTGGGATATTTTT</td>
</tr>
<tr>
<td>3518</td>
<td>TTTGATTTCTATGGGAACATTTG</td>
<td>TGGGAACCTTTGCAAGTAGATATAGG</td>
</tr>
<tr>
<td>3630</td>
<td>GCCACAGTACCCACATACACT</td>
<td>CACATGTACTTGAACAGGAG</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Novel FL transcript</th>
<th>FL-Forward</th>
<th>FL-Reverse</th>
</tr>
</thead>
<tbody>
<tr>
<td>3012</td>
<td>GGAGAGCTTCTCTGATAGCAGAA</td>
<td>CCAAGGGCTAGAATCAAGTT</td>
</tr>
<tr>
<td>3210</td>
<td>GATGCGGGGAGGGGTTTGA</td>
<td>AAATATAAAATATAGGCTTTTATAG</td>
</tr>
<tr>
<td>3372</td>
<td>GATCTCGGCTCCGTGAC</td>
<td>GGAATTCATGAAATTTTGAAGAGA</td>
</tr>
<tr>
<td>3373</td>
<td>TACTAAGCAGAAAAAGGAAACTG</td>
<td>GGTTTCAAAATATAGGCACTATTTT</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Novel ORF</th>
<th>ORF-Forward</th>
<th>ORF-Reverse</th>
</tr>
</thead>
<tbody>
<tr>
<td>3012 65aa</td>
<td>ATGGCGCACCAGGAAAGG</td>
<td>TCAACATGAAATCTGCTCTG</td>
</tr>
<tr>
<td>3210 111aa</td>
<td>ATGGTGGAAGGAGCTCGAGCCCGAG</td>
<td>TCAATGCGGGGAGGAGCTGCTCTG</td>
</tr>
<tr>
<td>3372 101aa</td>
<td>ATGAAATTTGCGGTCCGTGAC</td>
<td>TCAATCTCTCTACTTGGAC</td>
</tr>
<tr>
<td>3373 72aa</td>
<td>ATGCGGGCCCTTACCTG</td>
<td>TCAACAGGTTGATCATATTTT</td>
</tr>
</tbody>
</table>

57
Appendix 5. Novel transcript sequence analysis flowchart

**1. Process RACE sequence data**
1. Raw data processing, LongSAGE tag validation
2. Join 5' and 3' RACE sequences
3. Annotate alternative polyA & sequence polymorphism
   - 1. Representative FL transcript
   - 2. Alternative poly(A) transcripts
   - 3. Polymorphic sequence variants

**2. Nucleotide sequence level analysis**
1. Sequence composition
2. Word repeats

**3. Analyses in genomic context**
1. Fidelity of tag/transcript mapping
2. Genomic/internal oligo d(T) mispriming
3. Pseudogenes
4. Genomic repeats
5. Segmental duplications
6. Cross-species conservation
7. Proximal features: TFBS, CpG islands, DNAse I hypersensitive, TATA, CAAT
8. Splicing signals
9. Overlapping gene, mRNA, EST, ncRNA

**5. Sequence similarity (FL transcript and ORF)**
- BLAST family
- FASTA; MPsrch

**6. Domain and motif search** of ORFs
- InterProScan, CDD, ELM

**7. Regulatory RNA features** in nc region
- 5' / 3' UTRs & introns

**8. Protein level predictions:**
- Physicochemical parameters
- 2° structure, TM regions: PHD, PROF, TopPred, PSIPRED
- Coils and globularity: Coils, Globe
- Cysteine bridges: Disulfbrid
- Leucine zippers: 2Zip
- Signal/subcellular localization peptides: SignalP, TargetP
- PTMs: PROSITE

**9. Non-coding RNA gene** filtering:
- RNAdb BLAST
- Rfam hits

→ Novel protein-coding candidates
Appendix 6. Candidate transcript 3012 sequence

>3012_1769bp
GGAGAGCTTCTCTGTAAAGCGAAACCGGTGAAGGTCCAGAGAGGATGCGCACCCGGAAAGGGCGTAAGCTCCGCGCCCCTTCCCACAGACTTTCCCCCATCCATCCATCTCTTCATCTGGTGTTATCTGTATTTTGGTAGCAGCGGATGCATGGGCAAAGCATTA
CACAGCCGCTCCATGGACATGAGCTTAAGCTTTGCTTTCACAGCTAACCCAGAACCAGCGACCTCCAAAAGGACCCGCATGATA
AGTGGCAGCCAGGCGAAACATCACTGCTATCGTCAGAAGAAGCGAGGAGTTATTTGGGATAGCAGCGGACAAGCTGGGAG

>3012_1769bp_65aaORF (frame 3, 42 to 239)
MAHPERAWKLRAPSHRLSPIHLFIWCSSIAFVMSLIVNRKTVEEEENHHISYLEVLQKGQSEFH

>3012_1769bp_65aaORF (frame 3, 42 to 239)
MAHPERAWKLRAPSHRLSPIHLFIWCSSIAFVMSLIVNRKTVEEEENHHISYLEVLQKGQSEFH
Appendix 7.  **Candidate 3210** (a) transcript (b) gene (c) predicted protein (d) UCSC Genome Browser view

7 (a) 3210 Transcript structure

7 (b) 3210 Gene structure

**Summary transcript-gene information**

- 1 tag, in 468,000-tag H9 hESC library
- 47 kb predicted transcription unit
- GT/AG intron splice signals
- Polymorphism, 3 major transcript variants
- Poly(A) signal, best context in F02 isoform
- 54 % GC (ORF = 55 %)
- 18 % codon frequency mean difference
- Internal Ribosome Entry Site and upstream ORF in 5’ UTR
- K-Box, GY-Box, and Gamma Interferon Activated Inhibitor element in 3’ UTR
- InterPro hit = PTHR16213, uncharacterized
7 (c) 3210 predicted protein structure using PSIPRED. The top bars denote confidence of the prediction. Protein structural features are represented by a straight line for coils, a cylinder for helical regions, and an arrow for beta sheets. There are two sites of myristoylation and ten sites of phosphorylation (denoted by a *P*).

**Summary protein information**

- **MW** = 12.1 kDa, **pI** = 6.5
- **ELM motifs** = 13, including PCSK class proprotein cleavage sites
- **PTMs** = phosphorylation and myristoylation
- **PROFacc** = 71 % aa solvent accessibility
- **PROFsec protein class** = mixed
- **PHDhtm** = non-membrane (strong prediction)
- **GLOBE** = not globular
- **InterPro hit** = PTHR16213, uncharacterized protein
7 (d) 3210 UCSC Genome Browser view. The 2227 bp transcript and 111 aa ORF are marked. Data tracks shown include known and predicted genes, pseudogenes, predicted exons, RNA secondary structure predictions, mRNA and EST tracks, TFBS, cross-species sequence conservation tracks, and repeat elements.
Appendix 8. **Candidate 3372** (a) transcript (b) gene (c) predicted protein (d) UCSC Genome Browser view

8 (a) 3372 Transcript structure

8 (b) 3372 Gene structure

**Summary transcript-gene information**

- 4 tags in 3 libraries (WA07, WA09, WA14)
- 6930 bp predicted transcription unit
- GT/AG intron splice signals
- 43 % GC (ORF = 44 %)
- 20 % codon frequency mean difference
- BLAST = strong hits to hESC ESTs & hypothetical proteins
- Cytoplasmic polyadenylation element in 3' UTR
- 500-fold less abundant in reference adult RNA versus hESCs
8 (c) 3372 predicted protein structure using PSIPRED. The top bars denote confidence of the prediction. Protein structural features are represented by a straight line for coils, a cylinder for helical regions, and an arrow for beta sheets. There is one predicted disulfide bridge between two cysteines (positions 24 and 89), two sites of glycosylation, and eight sites of phosphorylation (denoted by a *P*).

**Summary protein information**

- MW = 7866 Daltons, pl = 8.76
- ELM motifs = 11, including NDR class endopeptidase cleavage sites
- PTMs = disulfide bridge, phosphorylation, glycosylation
- PROFace = 71 % aa solvent accessibility
- PROFsec protein class = mixed
- PHDhtm = non-membrane (strong prediction)
- GLOBE = may be globular, but not compact
8 (d) 3372 UCSC Genome Browser view. The 652 bp transcript and 101 aa ORF are marked. Data tracks shown include a high-scoring $K_a/K_s$ region overlapping the 101 aa ORF, known and predicted genes, pseudogenes, predicted exons, RNA secondary structure predictions, mRNA and EST tracks, TFBS, cross-species sequence conservation tracks, and repeat elements.
Appendix 9. **Candidate 3373** (a) transcript (b) gene (c) predicted protein (d) UCSC Genome Browser view

### 9 (a) 3373 Transcript structure

![Transcript structure diagram]

### 9 (b) 3373 Gene structure

![Gene structure diagram]

**Summary transcript-gene information**

- 6 tags in 3 libraries (WA01, WA07, WA09)
- 609 bp transcription unit
- 43% GC (ORF = 46%)
- 21% codon frequency mean difference
- InterPro hit = PS01186, EGF-2 (EGF-like region)
- Gamma Interferon Activated Inhibitor element in 3' UTR
- 500-fold less abundant in reference adult RNA versus hESCs
9 (c) 3373 predicted protein structure using PSIPRED. The top bars denote confidence of the prediction. Protein structural features are represented by a straight line for coils, a cylinder for helical regions, and an arrow for beta sheets. With five cysteine residues there is a strong likelihood of disulfide bridging, an EGF-like domain occurs near the N-terminal, there is one site of myristoylation, and three sites of phosphorylation (denoted by a *P*).

**Summary protein information**

- MW = 8163 Daltons, pI = 7.6
- ELM motifs = 8, including a PCSK class proprotein cleavage site
- PTMs = disulfide bridges, phosphorylation, myristoylation
- PROFace = 60% aa solvent accessibility
- PROFsec protein class = all-alpha
- PHDhtm = weak prediction of 2 TM, N cytoplasmic
- GOLOBE = compact and globular
- InterPro hit = PS01186, EGF-2 (EGF-like region)
9 (d) 3373 UCSC Genome Browser view. The 609 bp transcript and 72 aa ORF are marked. Data tracks shown include known and predicted genes, pseudogenes, predicted exons, RNA secondary structure predictions, mRNA and EST tracks, TFBS, cross-species sequence conservation tracks, and repeat elements (one LTR comprises 56% of the transcript sequence).
Appendix 10. Example calculation for relative quantitation of transcript abundance using the standard curve method.

Definitions
Reference standard = H9 hESC.
Calibrator sample (baseline) = Day-0 sample in retinoic acid-induced differentiation.
Test samples = Day-2 and Day-4 of retinoic acid differentiation.
Endogenous reference gene = GAPDH.
Target (query) = Novel transcript 3012.

(i) Calculate the mean and standard deviation values of replicate reactions.

(ii) Generate a relative standard curve for the endogenous control (GAPDH) in the reference sample (H9 hESC).

<table>
<thead>
<tr>
<th>Input total RNA H9 hESC (ng)</th>
<th>Log of input</th>
<th>GAPDH-1f/r average Ct</th>
<th>GAPDH-1f/r Ct Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>2.0</td>
<td>15.72</td>
<td>0.31</td>
</tr>
<tr>
<td>10</td>
<td>1.0</td>
<td>17.84</td>
<td>0.10</td>
</tr>
<tr>
<td>1</td>
<td>0.0</td>
<td>21.45</td>
<td>0.23</td>
</tr>
<tr>
<td>0.1</td>
<td>-1.0</td>
<td>24.80</td>
<td>0.14</td>
</tr>
<tr>
<td>0.01</td>
<td>-2.0</td>
<td>28.41</td>
<td>0.11</td>
</tr>
</tbody>
</table>

\[
y = -3.234x + 21.644
\]
(iii) Generate a relative standard curve for the target (3012) in the reference sample (H9 hESC).

<table>
<thead>
<tr>
<th>Input total RNA H9 hESC (ng)</th>
<th>Log of input</th>
<th>3012-1f/r average Ct</th>
<th>3012-1f/r Ct Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>2.0</td>
<td>21.38</td>
<td>0.01</td>
</tr>
<tr>
<td>10</td>
<td>1.0</td>
<td>24.29</td>
<td>0.07</td>
</tr>
<tr>
<td>1</td>
<td>0.0</td>
<td>27.66</td>
<td>0.05</td>
</tr>
<tr>
<td>0.1</td>
<td>-1.0</td>
<td>31.77</td>
<td>0.45</td>
</tr>
<tr>
<td>0.01</td>
<td>-2.0</td>
<td>35.39</td>
<td>0.91</td>
</tr>
</tbody>
</table>

(yii) Determine relative quantity of the target (3012) in the calibrator (Day-0) and test (Days 2 and 4) samples from the standard curve.

<table>
<thead>
<tr>
<th></th>
<th>3012-1f/r average Ct</th>
<th>3012-1f/r Ct Std. Dev.</th>
<th>3012 quantity (ng) interpolated from std. curve</th>
</tr>
</thead>
<tbody>
<tr>
<td>RA Day-0</td>
<td>25.00</td>
<td>0.05</td>
<td>7.49</td>
</tr>
<tr>
<td>RA Day-2</td>
<td>23.56</td>
<td>0.09</td>
<td>18.97</td>
</tr>
<tr>
<td>RA Day-4</td>
<td>25.02</td>
<td>0.02</td>
<td>7.39</td>
</tr>
</tbody>
</table>
(v) Calculate the normalized value and standard deviation of the target (3012 normalized to GAPDH) in the calibrator (Day-0) and test (Days 2 and 4) samples.

(vi) Calculate the fold difference and standard deviation of the target (3012) between each test sample (Days 2 and 4) and the calibrator sample (Day-0).

<table>
<thead>
<tr>
<th></th>
<th>Calibrator Day-0</th>
<th>Normalized Day-0</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAPDH</td>
<td>10.29</td>
<td>1.00</td>
</tr>
<tr>
<td>3012</td>
<td>7.49</td>
<td>0.73</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GAPDH</td>
<td>12.26</td>
<td>1.00</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>3012</td>
<td>18.97</td>
<td>1.55</td>
<td>0.17</td>
<td>2.1</td>
<td>0.23</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Test sample Day-4</th>
<th>Normalized Day-4</th>
<th>Std. Deviation of normalized Day-4</th>
<th>Day-4 fold diff. from Day-0</th>
<th>Std. Deviation of Day-4 fold diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAPDH</td>
<td>6.67</td>
<td>1.00</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>3012</td>
<td>7.39</td>
<td>1.11</td>
<td>0.04</td>
<td>1.5</td>
<td>0.05</td>
</tr>
</tbody>
</table>

**Interpretation**

The abundance of novel transcript 3012 is $2.1 \pm 0.2$ and $1.5 \pm 0.1$ times higher on Day-2 and Day-4, respectively, compared to Day-0 of retinoic acid-induced differentiation.

The University of British Columbia

Biohazard Approval Certificate

PROTOCOL NUMBER: H06-0109

INVESTIGATOR OR COURSE DIRECTOR: Marra, Marco

DEPARTMENT: Medical Genetics

PROJECT OR COURSE TITLE: Development of Technologies for the Derivation, Propagation and Differentiation of hESC

APPROVAL DATE: 07-06-20

APPROVED CONTAINMENT LEVEL: 2

FUNDING AGENCY: Networks of Centres of Excellence (NCE)

The Principal Investigator/Course Director is responsible for ensuring that all research or course work involving biological hazards is conducted in accordance with the Health Canada, Laboratory Biosafety Guidelines, (2nd Edition 1996). Copies of the Guidelines (1996) are available through the Biosafety Office, Department of Health, Safety and Environment, Room 50 - 2075 Wesbrook Mall, UBC, Vancouver, BC, V6T 1Z1, 822-7596, Fax: 822-6650.

Approval of the UBC Biohazards Committee by one of:
Chair, Biosafety Committee
Manager, Biosafety Ethics
Director, Office of Research Services

This certificate is valid for one year from the above start or approval date (whichever is later) provided there is no change in the experimental procedures. Annual review is required.

A copy of this certificate must be displayed in your facility.

Office of Research Services
102, 6190 Agronomy Road, Vancouver, V6T 1Z3
Phone: 604-827-5111 FAX: 604-822-5093