TRANSCRIPTOMICS OF EARLY HUMAN DEVELOPMENT

NAHENIWELA HERATH MUDIYANSELAGE

WISHVA BANDARA HERATH

NATIONAL UNIVERSITY OF SINGAPORE

2011

TRANSCRIPTOMICS OF EARLY HUMAN DEVELOPMENT

NAHENIWELA HERATH MUDIYANSELAGE

WISHVA BANDARA HERATH

BSc. (Hons.), Grad. Chem.

A THESIS SUBMITTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

NUS GRADUATE SCHOOL FOR INTEGRATIVE SCIENCES AND ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2011

Acknowledgement

My phD journey was an extremely eventful one.

I was trained as a microbiologist and a chemist, then became a biochemist working on proteomics and ended up doing my thesis project as a developmental / molecular biologist on transcriptomics.

Without the help of several special people this thesis would not have been possible.

I would like to thank...

Dr. Paul Robson, my supervisor, for giving me the opportunity to work under him, for his advice, insights and guidance throughout the project and for creating a stress-free environment in the lab which is a pleasure to work in.

Prof. Justine Burley, for her kindness and support when I needed it the most.

My wife Dulani, for being by my side, through thick and thin.

Thank you!

The trophoblast differentiation protocol which I use extensively in my thesis was first described by Dr. Luo Wenlong. Dr. Mikael Huss helped me with bioinformatic analysis and helped me learn python programming, Mr. Audi Harsono helped me in validating RNA-Seq data, Ms. Woon Chow Thai ran the microarrays, Ms. Jameelah Sheik Mohamed taught me stem cell culture and Dr. Guo Guoji helped me extract mouse embryo samples.

I would also like to acknowledge NUS Graduate School for Integrative Sciences and Engineering (NGS) for my scholarship and the Genome Institute of Singapore (GIS) where I carried out my research.

Table of Contents

Summary

In this thesis I present my attempt to further the knowledge on early human development with emphasis on trophoblast lineage, using RNA-Sequencing (RNA-Seq) technology. RNA-Seq leverages on high throughput next generation sequencing to profile entire transcriptomes with extreme sensitivity and resolution, providing data superior to that of conventional methods available for measuring gene expression.

Three major RNA-Seq datasets are presented in this thesis.

The first dataset contains information on transcriptomic dynamics of poly A mRNA from a time-course experiment with five time-points (day 0, 2 4 6 and 8), where human embryonic stem cells were differentiated along the trophoblast lineage using an improved differentiation protocol.

The second dataset contains transcriptomic data of smallRNA (all RNA transcripts less than 200 nucleotides) during the first three time-points of the above mentioned differentiation protocol.

The third dataset is on mouse early development and contains information on the transcriptomes of the 8-cell stage embryo, E3.5 blastocyst, E4.5 blastocyst and E4.5 inner cell mass. This mouse preimplantation dataset is used in a comparative capacity to find molecular mechanisms which are specific to the human system.

As an early adapter of the RNA-Seq technology during a time where there were no proper analysis software available, I created a series of programatic workflows in the form of scripts, written using the python programming language, meant to simplify the analysis of RNA-Seq data and to easily identify transcriptomic events such as alternative

splicing, novel exon - exon junctions, exon extensions and expression of novel transcripts. These workflows together with the results they provide are also presented in this thesis.

Using RNA-Seq datasets and results of programatic workflows mentioned above, this thesis presents a comprehensive view on the transcriptomics of early human trophoblast differentiation.

When comparing human and mouse preimplantation data, it was evident that the two systems have considerable differences at the transcriptome level concerning both the expression pattern and expression level of genes. This observation supports the hourglass model of development, where the species of the same animal phylum, for a brief period in their developmental timeline known as the phylotypic stage, show a remarkable similarity with each other, but show considerable differences during the rest of the developmental timeline. Trophoblast development occurs much earlier than the phylotypic stage and therefore shows great divergence in transcriptomics between mouse and human. This is important because it advocates the cautious extrapolation of biological observations made in the mouse system into human - as in the case of most data available for trophoblast differentiation.

Looking at novel (i.e. unannotated) transcribed regions of the human genome identified by RNA-Seq, it was evident that trophoblast differentiation induces the expression of a large number of endogenous retroviral sequences. There are instances where these retroviral elements modify transcripts by acting as extra exons or as new promoters resulting in the expression of new transcripts. Therefore this thesis argues that retroviral

elements are a major component responsible for the human / primate specific transcriptomic events in early development. Thus they are responsible for the interspecies diversity seen during the pre-phylotypic stages of development in human and mouse.

List of Tables

List of Figures

List of gene names

ABCG2	ATP-binding cassette sub-family G member 2;ABCG2
ADAMTS8	A disintegrin and metalloproteinase with thrombospondin motifs 8;ADAMTS8
AIF1	Allograft inflammatory factor 1;AIF1
AMOTL2	Angiomotin-like protein 2;AMOTL2
ANXA1	Annexin A1;ANXA1
ANXA3	Annexin A3;ANXA3
APOA4	Apolipoprotein A-IV;APOA4
ATCAY	Caytaxin;ATCAY
BCAN	Brevican core protein;BCAN
BLNK	B-cell linker protein;BLNK
C10orf10	Protein DEPP;DEPP
C14orf115	Uncharacterized protein C14orf115;C14orf115
C1orf105	Uncharacterized protein C1orf105;C1orf105
C1QL2	Complement C1q-like protein 2;C1QL2
C9orf135	Uncharacterized protein C9orf135;C9orf135
CASP4	Caspase-4 subunit 2;CASP4
CBLB	E3 ubiquitin-protein ligase CBL-B;CBLB
CCKBR	Gastrin/cholecystokinin type B receptor;CCKBR
CCR1	C-C chemokine receptor type 1;CCR1
CCR7	C-C chemokine receptor type 7;CCR7
CDKN1C	Cyclin-dependent kinase inhibitor 1C;CDKN1C
CEBPA	CCAAT/enhancer-binding protein alpha;CEBPA
CGA	Glycoprotein hormones alpha chain;CGA
CGB8	Choriogonadotropin subunit beta;CGB8
COL21A1	Collagen alpha-1(XXI) chain;COL21A1
CPEB1	Cytoplasmic polyadenylation element-binding protein 1;CPEB1
CRIP3	Cysteine-rich protein 3;CRIP3
CRLF1	Cytokine receptor-like factor 1;CRLF1
CST4	Cystatin-S;CST4
CTCFL	Transcriptional repressor CTCFL;CTCFL
CXCL5	ENA-78(9-78);CXCL5
CYP19A1	Cytochrome P450 19A1;CYP19A1
CYP3A7	Cytochrome P450 3A7;CYP3A7
DAPP1	Dual adapter for phosphotyrosine and 3-phosphotyrosine and 3-phosphoinositide;DAPP1
DCN	Decorin;DCN
DEFB1	Beta-defensin 1;DEFB1
DOCK2	Dedicator of cytokinesis protein 2;DOCK2
ELF5	ETS-related transcription factor Elf-5;ELF5
ENPEP	Glutamyl aminopeptidase;ENPEP
FAM124A	Protein FAM124A;FAM124A
FBLN7	Fibulin-7;FBLN7
FGA	Fibrinopeptide A;FGA
FGB	Fibrinopeptide B;FGB
FGF19	Fibroblast growth factor 19;FGF19

FHDC1	FH2 domain-containing protein 1;FHDC1
FOXD3	Forkhead box protein D3;FOXD3
FYB	FYN-binding protein;FYB
GADD45G	Growth arrest and DNA-damage-inducible protein GADD45 gamma;GADD45G
GATA2	Endothelial transcription factor GATA-2;GATA2
GCM1	Chorion-specific transcription factor GCMa;GCM1
GPRC5A	Retinoic acid-induced protein 3;GPRC5A
HERV-FRD	Transmembrane protein;ERVFRDE1
HMGCS2	Hydroxymethylglutaryl-CoA synthase, mitochondrial;HMGCS2
HOPX	Homeodomain-only protein;HOPX
HOXA1	Homeobox protein Hox-A1;HOXA1
HOXB3	Homeobox protein Hox-B3;HOXB3
HRH3	Histamine H3 receptor;HRH3
HSPB8	Heat shock protein beta-8;HSPB8
IGSF21	Immunoglobulin superfamily member 21;IGSF21
IL1R1	Interleukin-1 receptor type I;IL1R1
KANK4	KN motif and ankyrin repeat domain-containing protein 4;KANK4
KCNH6	Potassium voltage-gated channel subfamily H member 6;KCNH6
KCNQ2	Potassium voltage-gated channel subfamily KQT member 2;KCNQ2
KRT18	Keratin, type I cytoskeletal 18;KRT18
KRT19	Keratin, type I cytoskeletal 19;KRT19
KRT23	Keratin, type I cytoskeletal 23;KRT23
LCK	Proto-oncogene tyrosine-protein kinase LCK;LCK
LEFTY1	Left-right determination factor 1;LEFTY1
LGALS13	Galactoside-binding soluble lectin 13;LGALS13
LGR5	Leucine-rich repeat-containing G-protein coupled receptor 5;LGR5
LRP2	Low-density lipoprotein receptor-related protein 2;LRP2
LY6G6C	Lymphocyte antigen 6 complex locus protein G6c;LY6G6C
LY6H	Lymphocyte antigen 6H;LY6H
LYN	Tyrosine-protein kinase Lyn;LYN
MAGEA4	Melanoma-associated antigen 4;MAGEA4
MRGPRX1	Mas-related G-protein coupled receptor member X1;MRGPRX1
MUC15	Mucin-15;MUC15
NMU	Neuromedin-U-25;NMU
NPR3	Atrial natriuretic peptide clearance receptor;NPR3
NPTX2	Neuronal pentraxin-2;NPTX2
NRIP3	Nuclear receptor-interacting protein 3;NRIP3
NTRK3	NT-3 growth factor receptor;NTRK3
ODAM	Odontogenic ameloblast-associated protein;ODAM
OLFM1	Noelin;OLFM1
OPCML	Opioid-binding protein/cell adhesion molecule;OPCML
P11	Exosome complex exonuclease MTR3;EXOSC6
PADI3	Protein-arginine deiminase type-3;PADI3
PDZD4	PDZ domain-containing protein 4;PDZD4
PGC	Gastricsin;PGC
PLA2G2F	Group IIF secretory phospholipase A2;PLA2G2F
POU3F1	POU domain, class 3, transcription factor 1;POU3F1

POU5F1	POU domain, class 5, transcription factor 1;POU5F1
PPP1R16B	Protein phosphatase 1 regulatory inhibitor subunit 16B;PPP1R16B
PPY	Pancreatic icosapeptide;PPY
PRDM14	PR domain zinc finger protein 14;PRDM14
PRKCB	Protein kinase C beta type;PRKCB
PTPRN	Receptor-type tyrosine-protein phosphatase-like N;PTPRN
PTPRZ1	Receptor-type tyrosine-protein phosphatase zeta;PTPRZ1
RAB31	Ras-related protein Rab-31;RAB31
RASGRP4	RAS guanyl-releasing protein 4;RASGRP4
RCN1	Reticulocalbin-1;RCN1
REEP1	Receptor expression-enhancing protein 1;REEP1
RHBDL3	Rhomboid-related protein 3;RHBDL3
RHOU	Rho-related GTP-binding protein RhoU;RHOU
RTN4RL2	Reticulon-4 receptor-like 2;RTN4RL2
S100A14	Protein S100-A14;S100A14
S100P	Protein S100-P;S100P
SDC1	Syndecan-1;SDC1
SERPINB12	Serpin B12;SERPINB12
SERPINB9	Serpin B9;SERPINB9
SGMS1	Phosphatidylcholine:ceramide cholinephosphotransferase 1;SGMS1
SLC1A3	Excitatory amino acid transporter 1;SLC1A3
SLC22A11	Solute carrier family 22 member 11;SLC22A11
SLC38A1	Sodium-coupled neutral amino acid transporter 1;SLC38A1
SLC40A1	Solute carrier family 40 member 1;SLC40A1
SLC7A11	Cystine/glutamate transporter;SLC7A11
SLC7A3	Cationic amino acid transporter 3;SLC7A3
SLITRK3	SLIT and NTRK-like protein 3;SLITRK3
SLN	Sarcolipin;SLN
SMAD7	Mothers against decapentaplegic homolog 7;SMAD7
SMPX	Small muscular protein;SMPX
SNCB	Beta-synuclein;SNCB
TACSTD2	Tumor-associated calcium signal transducer 2;TACSTD2
TCL6	T-cell leukemia/lymphoma protein 6;TCL6
TFAP2B	Transcription factor AP-2 beta;TFAP2B
TGFBR2	TGF-beta receptor type-2;TGFBR2
TLR7	Toll-like receptor 7;TLR7
TMEM132D	Transmembrane protein 132D;TMEM132D
TMEM145	Transmembrane protein 145;TMEM145
TMEM151B	Transmembrane protein 151B;TMEM151B
TNS3	Tensin-3;TNS3
TPD52L1	Tumor protein D53;TPD52L1
TREM1	Triggering receptor expressed on myeloid cells 1;TREM1
ZIC2	Zinc finger protein ZIC 2;ZIC2
ZIC5	Zinc finger protein ZIC 5;ZIC5
ZNF750	Protein ZNF750;ZNF750

1.Introduction

1.1 Preimplantation development: from zygote to blastocyst

Fertilization occurs in the fallopian tube, 24 - 48 hrs after ovulation, leading to the production of the zygote, the new organism's first developmental stage. Then for 3 - 4 days in mouse and for 5 - 7 days in humans, it travels through the fallopian tube moving towards the uterus while producing new cells - the blastomeres, through mitotic division. During these cleavage stages the actual size of the embryo remains the same even though the number of cells within the structure increases. During the 8-cell stage of the embryo, the blastomeres are totipotent, clearly identifiable and are topologically symmetrical. The polarization events that take place during compaction create the morula, giving rise to two 'classes' of cells - inner blastomeres and outer blastomeres. The inner blastomeres are fully surrounded by the outer blastomeres while the outer blastomeres have a part of their cell surface exposed to the external environment. Maturation of the outer blastomeres into a functional epithelium combined with further cell divisions leads to the formation of the blastocoel, the defining feature of the blastocyst.

The blastocyst is composed of an outer layer of cells making the trophoectoderm (TE), which marks the perimeter of the blastocyst, and the inner cell mass (ICM), which initially exists as a small group of cells attached to the TE layer facing the blastocoel. At around E4.5 in mouse, the ICM differentiates into the primitive endoderm, which will produce the extraembryonic tissues and the epiblast which will create the embryo proper. Formation of trophoectoderm (TE) and inner cell mass

(ICM) during the genesis of blastocyst marks the first lineage segregation event of the embryo. The TE goes on to form the fetal component of the placenta.

1.2 Development of the placenta

The placenta is a complex organ which acts as the interface between two (partially) genetically diverse individuals. It is composed of both fetal and maternal tissue. The placenta is essential for the transport of nutrients, gases and waste products between the fetus and mother and acts as an endocrine organ facilitating the growth of the fetus. Though poorly understood, the placenta also plays an important role in maternal immune modulation preventing the mother from rejecting the semi-allograft embryo.

Placenta development gets underway just after implantation which takes place at around 8 - 9 days post fertilization in humans when the embryo is made up of around 107 - 256 cells (Benirschke, Kaufmann et al. 2006). Implantation is composed of three stages, *apposition*, *adhesion* and *invasion*.

During apposition the blastocyst orients itself so that its embryonic pole gets attached first. This is preceded by complex crosstalk between the blastocyst and the uterine wall. After attachment, the invasion phase begins when the trophoblasts in the attachment surface proliferate and produce cytotrophoblast cells and a syncytiotrophoblast layer. Syncytiotrophoblasts are a multi-nucleated layer of cells produced by the fusion of mono-nucleated cytotrophoblasts. Cytotrophoblasts have an active proliferative rate which enables them to increase their number while producing the syncytiotrophoblast. After some time, vacuoles start to appear in the syncytiotrophoblast layer. As development progresses these expand and forms a

system of lacunae, separated by ‘walls’ of syncytiotrophoblast - the trebaculae. At around day 12 the embryo is completely engulfed by the uterine epithelium, and due to the proliferation of trophoblasts, the embryo is completely covered by a syncytial layer of syncytiotrophoblasts and cytotrophoblasts. Physical and hormonal pressures put on the endometrium by trophoblasts causes the endometrium to form decidua.

Trophoblast proliferation together with lacunar formation divides the trophoblast layer into three layers - in the direction of fetus to maternal tissues - primary chorionic plate which is composed of cytotrophoblasts, the lacunary system and the trophoblasic shell made by syncytiotrophoblasts. Here the cytotrophoblasts which form the chorionic plate, invade the syncytiotrophoblast and continue their migration to the maternal endometrium. The invading trophoblasts penetrates maternal blood vessels. In the mature placenta this connection results in filling the lacuna with blood. These lacuna act as mini reservoirs enabling the diffusion of nutrients and gas to the fetus. The endometrium-invading trophoblast cells form villi, where each villus is composed of a column of cytotrophoblasts and a surrounding layer of syncytiotrophoblast. From the earliest stages of the placenta to the most mature stage, a layer of trophoblastic and fetal tissues, termed the placental barrier, separates maternal and fetal bloodstreams.

As the fetus grows and its demands for oxygen and nutrients increases, the maternal circulation system and the placental barrier get adapted, so that the blood flow to the placenta is enhanced and the efficiency of transfer through the placental barrier is maximized. Due to the plugging of maternal spiral arteries by trophoblasts, there is no detectable uteroplacental blood flow during the first trimester (Hustin and Schaaps

1987) which leads to a low oxygen environment. At around the twelfth week the plugs are removed and the maternal blood flow is increased.

1.3 Genes involved in the formation of the trophoectoderm (TE)

The vast majority of functional studies to determine essential gene function in preimplantation development and formation of trophoblast lineage relies mainly on data from the mouse. Currently Tead4 is considered to be the earliest transcription factor involved in TE lineage determination (Yagi, Kohn et al. 2007). Tead4 homozygous mutants die even before the formation of the blastocoel (Nishioka, Yamamoto et al. 2008). Even though Tead4 is expressed ubiquitously in the embryo (Nishioka, Yamamoto et al. 2008), its activity is modulated through the components of the Hippo signaling pathway (Nishioka, Inoue et al. 2010). Hippo signaling is made active by cell to cell contact, which results in phosphorylation of the Tead coactivator protein Yap. Yap protein inhibits the nuclear localization of Tead, essentially preventing it from acting as a transcription factor. Since cell to cell contact is high on the inside cells, Hippo signaling is more active there, leading to inactivation of Tead4. This is supported by the fact that Yap protein shows different sub-cellular localization in ICM versus TE (Nishioka, Inoue et al. 2010). In Lats 1/2 homozygous mutants, which are negative regulators of Yap, Yap accumulates in the nucleus and Cdx2 expression increases (Nishioka, Inoue et al. 2010). Even though Tead4 is expressed in the ICM it is not essential for the ICM (Yagi, Kohn et al. 2007; Nishioka, Yamamoto et al. 2008).

Activation of Tead4 in TE leads to the up-regulation of *Cdx2* (Yagi, Kohn et al. 2007). Cdx2 is recognized as an essential factor for the ICM/ TE lineage segregation and acts

by the repression of Oct4 and Nanog in the TE. It has been shown that Cdx2 homozygous mutant embryos, even if they produce the blastocyst, die before implantation due to the collapsing of the blastocoel, and that in these embryos there is no differential expression of *Oct4* and *Nanog* between the TE and ICM (Strumpf, Mao et al. 2005). The reciprocal relationship between Cdx2 and Oct4 has been further shown by the over-expression of Cdx2 in mouse ES cells, which down-regulates Oct4 leading to the differentiation of cells into TE lineage (Niwa, Toyooka et al. 2005). Cdx2 is first expressed in a nonspecific manner and then gets up-regulated in the outside cells which are the precursors of TE, suggesting that Cdx2 is not the trigger for lineage segregation (Ralston and Rossant 2008; Guo, Huss et al. 2010).

Apart from Tead4 and Cdx2 which based on current understanding, act as the main regulators of lineage shift to TE, there are other transcription factors which help in the maintenance and progression of TE state. Among these, Eomes is considered to be an important factor in trophoblast development and mesoderm formation (Russ, Wattler et al. 2000). Even though Eomes is expressed throughout early development (McConnell, Petrie et al. 2005), it is believed to be at least partially regulated by Cdx2 (Nishioka, Yamamoto et al. 2008).

Tcfap2c has been reported to differentiate ES cells into the trophoblast lineage independent of Cdx2, even though both Cdx2 and Tcfap2c are required for the up-regulation of Elf5 which helps in trophoblast cell maintenance (Kuckenberg, Buhl et al. 2010). Additionally Ets2 has been shown to be important in trophoblast development (Georgiades and Rossant 2006) and trophoblast stem cell self renewal (Wen, Tynan et al. 2007). Elf5 acts downstream of TE formation and aids in the

robust expression of Cdx2 and Eomes (Ng, Dean et al. 2008). Gata3, which is highly expressed in trophoblast cell lines, is expressed from the 8-cell stage but gets restricted to TE and is thought to regulate TE expression (Home, Ray et al. 2009; Ralston, Cox et al. 2010). Gata2 is also expressed in the blastocyst and is restricted to the TE. Thus redundancy between Gata 2/3 may explain why neither is early embryonic-lethal in the mouse.

1.4 Genes involved in the formation of the placenta

Even though placental mammals (and the placenta) came into being relatively recently compared to the timeframe of vertebrate evolution, the placenta as an organ is highly diverse among different species both at the tissue / cellular level and the sub cellular level (Rawn and Cross 2008). Surprisingly, this diversity is not caused mainly by new placenta-specific genes, as the number of such known genes are low. Instead placenta development involves genes with multiple functions in both placenta and in other tissues / organs (Cross, Baczyk et al. 2003). For example HAND1 is involved in heart and blood vessel formation along with placenta development (Riley, Anaon-Cartwight et al. 1998). Similar examples include both DLX3 (Beanan and Sargent 2000) and FGFR2 (Xu, Weinstein et al. 1998). However this lack of placenta specific genes is compensated mainly by transcriptional regulation. There are several examples of genes which undergo alternative splicing or have alternative start sites - regulated by placenta specific enhancers / promoters - thereby producing a placenta-specific or placenta-enriched isoform. Some of the placenta-specific promoter activity has been derived from historical retroviral infections.

1.4.1 GCM1 (Glial cells missing 1)

GCM1 is considered to be an essential transcription factor for placental development due to its ability to make active fusogenic and proangiogenic gene expression in the placenta, thereby leading to vasculogenesis and formation of the syncytiotrophoblast (Anson-Cartwright, Dawson et al. 2000; Lin, Chang et al. 2010). GCM1 is reported to positively regulate Syncytin (Yu 2002), placental growth factor (PGF) (Chang, Mukherjea et al. 2008) and Aromatase (CYP19A1) (Yamada, Ogawa et al. 1999), all genes essential for placental function. GCM1 acts as a regulator between the proliferative state and the cell cycle arrest / fusion of trophoblast cells (Baczyk, Drewlo et al. 2009). GCM1 has a highly placenta-specific expression and is known to be regulated at the post transcriptional level. GSK3B causes GCM1 to be phosphorylated which leads to it being detected by FBW2 and ultimately resulting in GCM1 degradation (Chiang, Liang et al. 2009). DUSP23 on the other hand has a protective effect, since it is involved in the dephosphorylation of GCM1 thereby preventing it from being degraded (Lin, Chang et al. 2010).

1.4.2 Chorionic gonadotropin (CG)

CG is a member of the glycogen hormone family where the rest of the members consist of the luteinizing hormone (LH), Follicle stimulating hormone (FSH) and the thyroid stimulating hormone (TSH) (Pierce and Parsons 1981). The CG protein is dimeric, consisting of an alpha subunit, encoded by *CGA*, which is shared among all the members of the glycogen hormone family and a beta subunit which is specific for CG. CG is only found in primates and horses and is expressed exclusively in the placenta (Nilson, Bokar et al. 1991; Rawn and Cross 2008). The beta subunit of CG,

derived from duplications of the gene encoding the beta subunit of LH, has 6 copies in the human genome. CG is involved in inducing progesterone secretion from the corpus lutem and preparing the uterus endometrium for pregnancy (Cameo, Srisuparp et al. 2004) and is essential for the maintenance of human pregnancy.

1.4.3 Growth Hormone cluster

The human growth hormone cluster is composed of five genes - *GH1, GH2, CSH1, CSH2* and *CSHL1*. Among these, all except GH1 is placenta specific, while GH1 is expressed both in the placenta and the pituitary (Su, Liebhaber et al. 2000). GH2 protein induces maternal lactogenic and growth promoting activities (Macleod , Worsley et al. 1991; Alsat, Guibourdenche et al. 1998).

1.4.4 ELF5 (E74-like factor 5)

ELF5 is believed to act as a "gatekeeper gene", to maintain the trophoblast lineage after the initial lineage commitment (Senner and Hemberger 2010) and is under epigenetic regulation (Hemberger, Udayashankar et al. 2010). In mouse, it has been reported that *Elf5* is methylated and repressed in the embryonic lineage and hypomethylated and expressed in the trophoblast lineage, and enforces maintenance of the trophoblast lineage through a positive feedback loop with Cdx2 and Eomes (Ng, Dean et al. 2008). In humans, *ELF5* expression is found in the villous cytotrophoblast cells of the placenta (Hemberger, Udayashankar et al. 2010). On the other hand, human embryonic stem cells, and the trophoblast cells derived either through spontaneous differentiation or BMP4 treatment have a hypermethylated and non - expressed *ELF5* (Hemberger, Udayashankar et al. 2010). This raises an issue with the conventional trophoblast differentiation protocols since ELF5 is an important

regulator of the trophoblast lineage and the conventional differentiation protocols do not induce *ELF5* expression similar to the actual system.

1.5 Transposable elements in the human genome

Transposable elements are mobile DNA sequences in the genome. They are able to change their location within the genome by using either a "cut - paste" or a "copy - paste" strategy. Transposable elements are divided into two classes - Retrotransposons (Class I) and DNA transposons (Class II). It has been estimated that 45% of the human genome is composed of sequence derived from transposable elements (Griffiths 2001) though many of these sequences are no longer "transposable".

Under the copy-paste strategy used by retrotranposons in the "migration" across genome, the element is first transcribed into an RNA intermediate, which then changes the original locale and finally gets reverse transcribed back into genomic DNA. Because of this mechanism they leave behind their original DNA footprint in the genome thus effectively amplifying their number over time. Endogenous retroviral elements (ERVs) belong to this class.

DNA transposons on the other hand do not have an RNA intermediate stage as they follow a cut - paste strategy. Their movement involves the excision of the transposon from the genomic DNA, movement to a new location and then the integration back into the genome.

1.5.1 Endogenous retroviral elements (ERVs)

Endogenous Retroviral Elements (ERVs) are the modern day genomic remnants of ancient germline infections of exogenous viruses. ERVs alone make up for around 8 - 10% of the human genome (Griffiths 2001; Goodier and Kazazian 2008; Black, Arnaud et al. 2010).

ERVs have the same genomic structure as their active exogenous counterparts. This includes the four viral genes, *gag*, *pro*, *pol* and *env*, sandwiched between two long tandem repeat (LTR) regions. The *gag* gene codes for structural components of the viral particle while *pro* and *pol* genes code for the enzymatic machinery. The *env* gene codes for viral capsid and envelope protein. The two LTR regions contain regulatory elements which could regulate the expression and the function of the ERV element (Black, Arnaud et al. 2010).

Previously, ERVs together with other members of the repeat elements group were thought to be nonfunctional and were labelled as “junk DNA”. In fact, most if not all human endogenous retroviral elements (HERVs) have acquired point mutations in the coding sequences, which disrupt the original function. That said the mere existence of these genes, with their original gene structure intact, indicates that they may still serve a biological purpose. If ERVs were truly “junk DNA” then at least the majority of them would simply cease to exist, removed through natural selection.

Creation of new ERVs by the integration of new retroviruses to the germline has happened throughout evolution. The recent ERVs known as modern ERVs still have functional viral pathogens such as the Jaagsiekte sheep retrovirus (JSRV), mouse

mammary tumor virus, feline leukemia virus and the avian leukemia virus which closely resemble their (i.e. ERV's) genomic structure and sequence (Black, Arnaud et al. 2010). However the common consensus is that most of the ERVs, due to point mutations, have completely lost one or more functional genes of the original viral particle.

1.6 Human genes originating from human ERVs which are highly expressed in placenta

The genes *HERV-W* (*ERVWE1*), *HERV-FRD* and *ERV-3* have intact *env* genes and are expressed in the human placenta (Venables, Brookes et al. 1995; Blond, Besème et al. 1999; De Parseval, Lazar et al. 2003).

1.6.1 ERV-3

This was the first retroviral protein to be associated with a physiological function. (Rote, Chakrabarti et al. 2004). *ERV-3* is coded by the env gene, which has a long open reading frame, and is expressed in syncytiotrophoblasts but not in villous cytotrophoblasts (Lin, Xu et al. 1999). In isolated cytotrophoblasts, the ERV-3 expression was up-regulated upon differentiation (Boyd, Bax et al. 1993). It has also been reported that its expression is associated with increased expression of hCG and cell cycle arrest prior to syncytiotrophoblast formation (Rote, Chakrabarti et al. 2004).

A mutation which introduces a stop codon in the ERV-3 coding *env* gene has been observed (Rasmussen and Clausen 1998; De Parseval, Lazar et al. 2003). The functional importance of ERV-3 has been questioned since 1% of the population which has this as a heterozygous mutation are healthy. However there is a chance that

ERV-3 still might be active due to a variety of reasons including the truncated protein retaining function, ERV-3 function being restored by other retroviral elements expressed during early development, and the stop codon being bypassed (Rote, Chakrabarti et al. 2004). However based on the structure of the full length ERV-3 protein, it has been reported that it lacks fusogenic ability and the ability for immunosuppression (Lin, Xu et al. 1999; Lin, Xu et al. 2000; Mi, Lee et al. 2000).

1.6.2 HERVE1 / Syncytin 1

Protein coded by the *env* open reading frame of *HERV-W* is known as Syncytin 1. *HERV-W* expression is restricted to syncytiotrophoblast (Rote, Chakrabarti et al. 2004). Syncytin 1 entered the primate genome 25 million years ago after the split of the new and old world monkeys which happened 40 million years ago (De Parseval, Lazar et al. 2003).

Interaction between Syncytin 1 and the D type mammalian retrovirus receptor leads to the formation of the syncytium (Blond, Besème et al. 1999; Handwerger 2009). It is believed that Syncytin 1 is involved in the fusion of mononuclear cytotrophoblasts to produce syncytiotropblasts. Syncytin 1 causes cell fusion in cell lines and this activity is reversed by an anti Syncytin 1 antibody (Blond, Lavillette et al. 2000; Mi, Lee et al. 2000) . Conversely when BeWo Cell fusion is induced by forskolin, Syncytin 1 is up-regulated (Mi, Lee et al. 2000). In addition Syncytin 1 contains a putative immunosuppressive region suggesting an immunological function as well (Black, Arnaud et al. 2010).

1.6.3 HERV-FRD / syncytin 2

Syncytin 2 was identified by a genome-wide screen for fusogenic retroviral envelopes (Blaise, de Parseval et al. 2003). Both Syncytin 1 and Syncytin 2 are structurally similar. Like Syncytin 1, Syncytin 2 is also reported to induce cell fusion and is believed to have immunosuppressive properties (Mangeney, Renard et al. 2007).

1.7 Functional implications of the existence and expression of ERVs

Despite their origins from infective exogenous retroviruses, ERVs have been associated with positive effects on their host's biology.

Most retroviruses contain an immunosuppressive region in their *env* protein which enables the viral particle to bypass the host immune defenses. While this is certainly detrimental to the host in the case of a retroviral infection, it could be considered as a benefit in rare instances where temporary immunosuppression is required. During implantation, the mother's reproductive system must accept the embryo, which is of a different genomic composition (i.e. a semi-allograft) to that of the maternal genome. An immune rejection at any stage of early development would be fatal to the fetus highlighting the requirement for immunosuppression.

While modulation of immunosuppressive effects has been argued as a main function of ERVs they have also been implicated in imparting antiviral resistance, maintaining genomic plasticity and introducing novel regulatory elements via LTRs (Nelson, Carnegie et al. 2003).

1.8 Genes which produce placenta-specific / placenta-enriched transcripts due to insertion of retroviral elements in their regulatory regions

Long terminal repeat (LTR) sequences flank retrotransposons and have their own promoters and enhancers. Therefore if an insertion of a retrotransposon occurs close to an existing gene there is a chance that the expression of the gene is influenced by the LTR promoters and enhancers. Several such examples have been reported and more instances were identified from the data presented in this thesis.

1.8.1 CYP19A1 (Cytochrome P450, family 19, subfamily A, polypeptide 1)

CYP19A1 codes for the protein Aromatase, which is a key enzyme in estrogen biosynthesis (Simpson, Mahendroo et al. 1994). It is involved in placental development as well as preparation for parturition (Fürbass, Selimyan et al. 2007). RefSeq annotation shows that aromatase has two isoforms, where the splice selects either the first or the second in a mutually exclusive manner. The isoform which encompasses the first exon is reported to be placental specific and this isoform is regulated by a placenta-specific promoter (Kamat and Mendelson 2001). The particular promoter exists in an LTR region, suggesting that its origin is retroviral (van de Lagemaat, Landry et al. 2003).

1.8.2 EDNRB (Endothelin receptor type B)

EDNRB has an active LTR region, derived from an HERV-E family retrovirus, as an alternative promoter, which produces a placental specific isoform (Medstrand, Landry et al. 2001). The LTR promoter is located ~52kb upstream of the "standard" promoter

of the gene. Unlike in the case of CYP19A1, the placental specific isoform of *EDNRB* accounts for only around 15% of the total transcripts in placenta (Sakurai, Yanagisawa et al. 1990).

1.9 MicroRNAs in early development

MicroRNAs are short non coding regulatory RNA, which regulate the translation of target mRNAs either by mRNA degradation or by translational repression (Lewis and Steel 2010). The microRNAs carry out their function by binding to the 3' UTR of the target mRNA and they add an additional layer of complexity to the transcriptome (Bartel 2004).

MicroRNA biogenesis consists of several steps. First, microRNA genes are transcribed by RNA polymerase II which produces a primary microRNA (pri-miRNA). Primary microRNA has a stable secondary stem loop structure consisting of a ~33 nucleotide stem. Then it is cleaved by the microprocessor complex consisting of Drosha and DGCR8 to produce a pre-microRNA. Pre-microRNA maintains the stem loop structure of pri-miRNA but has a shorter stem consisting of ~22 nucleotides. This stable stem loop structure is a key feature of microRNAs and can be used in microRNA prediction workflows. The pre-miRNA is then transported out of the nucleus by Exportin - 5 transporter protein, where the RISC loading complex subjects it to a further cleaving step by removing its loop thus resulting in a ~22 nucleotide double stranded RNA (Lee, Ahn et al. 2003). The double strand then gets separated where one goes on to act as a mature microRNA while the other (known as the "star" strand) gets degraded.

The expressed cohort of microRNA of a particular tissue / cell type has been shown to be highly specific (Bartel 2004). This implies that microRNA regulation is at least partially involved in defining the "state" of the tissue / cell type. Therefore the identification of microRNA expression dynamics during early development is very important.

Involvement of microRNA in early development has been reported. Mouse Oocytes with nonfunctional microRNA biogenesis machinery do not survive beyond the first cell division (Murchison, Stein et al. 2007; Tang, Kaneda et al. 2007) indicating that maternally derived microRNAs are essential for the very first steps of mammalian development (Lewis and Steel 2010). Mouse miR-125a expression has been shown to begin at the 2 cell stage, and is believed to regulate developmental timing (Byrne and Warner 2008). Mouse miR-92 has been shown to be specific for trophoectoderm and primitive endoderm (Takeda, Noguchi et al. 1997; Byrne and Warner 2008; Foshay and Gallicano 2009). In the case of human ES cells, mir-145 is reported to regulate *POU5F1*, *SOX2* and *KLF4* showing the importance of microRNA in regulating pluripotency (Xu, Papagiannakopoulos et al. 2009). It has been shown that placenta too expresses a unique set of microRNA and some even enter the mother's blood stream (Gilad, Meiri et al. 2008). Thus defining microRNAs in the human trophoblast lineage will not only provide resources for understanding the basic biology of placental formation but may also potentially provide biomarkers for placental function in maternal blood.

1.10 Epithelial - mesenchymal transition (EMT)

Epithelial-mesenchymal transition (EMT) is the process in which polarized epithelial cells convert themselves into a mesenchymal phenotype through structural and biochemical changes (Zeisberg and Neilson 2009). EMT transition, or its reverse (Mesenchymal to epithelial transition - MET) is dependent on the activity of specialized transcription factors, cell surface proteins, enzymes and even microRNAs. While epithelial cells are polar and stationary, mesenchymal cells show an increased capability for migration / invasion (Kalluri and Weinberg 2009).

EMTs are divided into three types. Type 1 includes EMT events that take place during implantation embryogenesis and organ development while type II includes EMT events during tissue regeneration and organ fibrosis (Kim, Kugler et al. 2006; Zeisberg, Tarnavski et al. 2007; Potenta, Zeisberg et al. 2008). EMTs which take place during cancer progression and metastasis are classified under type III (Hanahan and Weinberg 2000; Thiery 2002).

During trophoectoderm formation, cells of the morula undergo a transition to an epithelial phenotype. Furthermore, during implantation cytotrophoblast cells undergo an epithelial to mesenchymal transition which enable them to invade the maternal endometrium and act as an anchor and form an interface for gas and nutrient exchange (Aplin and Kimber 2004; Bischof, Aplin et al. 2006). These EMT events come under type I and are the least studied events among all EMT events. Most of the biochemistry relating to type I EMT comes from the studies done on embryogenesis or organ development, an event which occurs *after* the formation of the TE lineage.

When EMT events during embryogenesis is considered, canonical Wnt signaling is believed to be a critical factor, as embryos deficient in Wnt3 are unable to undergo EMT during gastrulation (Liu, Wakamiya et al. 1999; Skromne and Stern 2001). Formation of the primitive streak, which is the subsequent EMT event, requires Wnt8c (Popperl, Schmidt et al. 1997). Wnt proteins together with FGF receptors (Ciruna and Rossant 2001; Perea-Gomez, Vella et al. 2002) and the transcription factors Snail, Eomes and Mesps regulate gastrulation (Nieto 2002; Arnold, Hofmann et al. 2008; Lindsley, Gill et al. 2008; Kalluri 2009) .

1.11 Involvement of microRNA in the regulation of EMT

Using madin darby canine kidney (MDCK) clones, it has been shown that members of the microRNA 200 family (mir - 200a/b/c, miR- 141 and mir - 429) and mir - 205 are up regulated during EMT and regulate EMT through *ZEB1* and *ZEB2* (Gregory, Bert et al. 2008). mir-200b and 200c down-regulate the expression of *ZEB1* and *ZEB2* (Hurteau, Carlson et al. 2007; Nanna 2007). When expressed, ZEB1 and ZEB2 inhibit the expression of E-cadherin transcription, preserving the epithelial phenotype (Comijn, Berx et al. 2001; Eger, Aigner et al. 2005).

1.12 Importance of studying the development and function of trophoblast lineage

Trophoectoderm formation marks the first lineage commitment in early development. It also marks the creation of the first epithelial cell type of the new organism. The trophectoderm also plays a major role in implantation which is one of the most crucial steps for a successful pregnancy. The trophoblast create a majority of cells / tissues in the placenta, a unique organ which acts as the interface of two genetically different organisms (the mother and the child). Problems in trophoblast / placental biology are believed to contribute to poor pregnancy outcomes such as preeclampsia (incidence rate of 7 - 8%) and preterm labor (10%) (Goldenberg and Andrews 1996). Besides providing a better understanding of placental disorders, studying placental biology may provide an additional model to study cancer as there are similarities in the biology of the two systems. For instance, during the transformation of villous cytotrophoblasts to their extravillous state, trophoblasts change their morphology from that of epithelial to invasive mesenchymal type. This epithelial to mesenchymal transformation is seen in several types of cancer and studying trophoblast differentiation would shed more light on the underlying mechanisms. Therefore the study of trophectoderm is of vital importance not only from an early development stand point, but also from a clinical point of view.

1.13 Model systems currently available to study the trophoblast lineage

The study of trophectoderm formation and trophoblast differentiation, particularly in the human system, presents the researcher with a number of obstacles. Apart from the ethical issues raised by early human embryo research, TE formation takes place at an extreme early time point of embryo formation making it difficult to obtain clinical samples. Samples earlier than 6 weeks of gestation are not available (Golos, Giakoumopoulos et al. 2010). The available clinical samples from aborted fetuses are difficult to come by (Enders 2000). When available, the information obtained is limited to the later differentiation stages of the trophoblast. Therefore to circumvent the above mentioned issues, different types of model systems have been developed for the study of the early differentiation stages of the trophoblast.

1.13.1 Animal models to study trophoblast biology

One of the ways to study human trophoblast differentiation is to use a model system such as the mouse. Even now, the mouse and in some instances the primate system, are been used to study the TE formation and the differentiation of the trophoblasts. This approach while informative, has an inherent weakness, due to the genetic and biochemical differences between the model system (mouse) and human (Carter 2007) .

1.13.2 Cell lines to study trophoblast biology

Primary trophoblast cultures obtained from aborted placenta and a number of choriocarcinoma derived cell lines are available to study trophoblast biology (King, Thomas et al. 2000; Shiverick, King et al. 2001). Carcinoma cell lines are easy to

maintain and study as opposed to the primary cell cultures. However, the latter provides a more realistic view of the trophoblast biology (Genbacev and Miller 2000) as the cancer like properties of cell lines can be a handicap when using them to study trophoblast biology (Khoo, Bechberger et al. 1998). Since the primary trophoblast cells have been obtained from early pregnancy placenta or term placenta, they belong to a time point much later than the initial lineage commitment, making it difficult to use them to study TE formation and early TE differentiation.

1.13.3 Embryonic stem cells

The derivation of embryonic stem cells, specially human stem cells, from the inner cell mass of the blastocysts has enabled the study of pluripotency and differentiation, without using valuable and rare clinical samples. While trophoblast stem cells - stem cells derived from the trophoblast - are available for mouse, they are still not available for humans.

Under right conditions, stem cells can be differentiated into the trophoblast lineage, creating a model system, which starts from the earliest time points of trophoblast differentiation. Human embryonic stem cells, have been shown to spontaneously differentiate into the trophoblast lineage (Thomson, Kalishman et al. 1995; Thomson, Itskovitz-Eldor et al. 1998). However since this differentiation is not uniform, a variety of differentiation protocols, including controlling of gene expression, using chemical mediators and imparting physical stresses has been proposed to increase the efficiency of the differentiation.

It has been reported that ES cells can be differentiated along the trophoblast lineage by preventing the expression of pluripotency factors. This has been done by 'active' methods such as knocking down (Niwa, Miyazaki et al. 2000; Velkey and O'Shea 2003; Hay, Sutherland et al. 2004) or silencing *POU5F1*, *NANOG* or *SOX2* by siRNA (Hough, Clements et al. 2006; Ivanova, Dobrin et al. 2006; Loh, Wu et al. 2006) or through inducing ES cells to form embryoid bodies (EBs) and then selecting for trophoblast like cells (Gerami-Naini, Dovzhenko et al. 2004; Golos, Pollastrini et al. 2006) .

When it comes to the study of human trophoblast lineage, using these approaches on ES cells are preferred, due to its ability to show extreme early events in TE formation and differentiation. Since the differentiation starts from human stem cells, the observations obtained can be considered more realistic than what is gained when using materiel from different species.

1.13.4 Differentiating human embryonic stem cells to the trophoblast lineage through modification of hES media

ES cells grown in the presence of BMP4 differentiates into the trophoblast lineage (Xu, Chen et al. 2002; Liu, Dovzhenko et al. 2004). A similar observation has been done when BMP4 treatment was done without FGF2 (Schulz, Ezashi et al. 2008) . While these differentiation protocols do induce the expression of trophoblast related genes and suppress the expression of pluripotency factors they have certain flaws. For example it has been reported that the efficiency of BMP4 differentiation is cell line dependent and that certain IVF derived stem cell lines had poor trophoblast differentiation and that formation of endoderm / yolk sac like structures was also involved (Reubinoff, Pera et al. 2000; Pera, Andrade et al. 2004).

The BMP4 differentiation of human embryonic stem cells to the trophoblast lineage has been improved upon by Dr. Luo Wenlong under the supervision of Dr. Paul Robson (Dr. Luo Wenlong's thesis - https:// scholarbank.nus.edu.sg/handle/ 10635/18805, (Wenlong 2008)). This protocol used BMP4 treatment together with SU5402, an FGF receptor inhibitor, to produce a rapid, uniform differentiation of hES cells to the trophoblast lineage. This improved protocol works on multiple hES cell lines, and results in a more robust and rapid down-regulation of pluripotency factors and an up-regulation of trophoblast markers, compared to the standard BMP4 treatment (Wenlong 2008). My thesis relies on this particular improved differentiation protocol to study the transcriptome of the trophoblast lineage.

1.14 RNA-Sequencing as a tool for high-throughput transcriptomics

From the early Sanger sequencing methods, to the current high-throughput sequencing platforms, DNA sequencing technology has come a long way. The modern "Next generation" sequencing machines with their efficient chemistries and miniaturized technologies have the capacity to sequence millions of DNA reads per run. RNA-Sequencing (RNA-Seq) technology exploits this high-throughput sequencing capability, to sequence cDNA fragments from RNA extracts to study transcriptomes in great detail.

An RNA-Seq experiment has three main steps. The wet lab portion is where the RNA of the particular sample is extracted and the sequencing libraries generated. Then comes the sequencing part which results in a large amount of data describing all the

sequences of the libraries. The final step involves mapping the sequenced “fragments” to their original genes to identify the transcriptome.

The wet lab portion of an RNA-Seq experiment has several major steps.

1) Extraction of RNA and selecting for the RNA component of interest. The RNA extraction method should ensure the extraction of the RNA of interest, for example to study small RNA, the method used should be able to efficiently extract the smallRNA available in the sample.
2) Removal of rRNA. In an extracted RNA sample (unless the extraction was done so that only small RNA was extracted), ribosomal RNA would be the major component. Since rRNA show limited change in biology, they need to be removed to better use the available sequencing depth. For this reason commercial kits which deplete rRNA or extract mRNA using their polyA tail are available.
3) Fragmentation - Current sequencing technologies have a limited sequencing length. Therefore to accommodate this requirement the RNA (or in some cases the reverse transcribed cDNA) needs to be fragmented. Sonication methods as well as enzymatic methods are used in fragmentation.
4) Reverse transcribing of RNA. Depending on the protocol used, this step comes before or after fragmentation.
5) Adapter ligation. For the sequencing machine to process a particular read, it should contain two adapters either side of it (This is so that the reads can be incorporated into the specific sequencing chemistry used by the sequencer). In some protocols one of the adapters is used as a “barcode” for multiplexing of samples.

1.14.1 Available sample preparation strategies

Since RNA-Seq technology has been around for some time now, there are different protocols available for different samples. Choice of protocol is mainly determined by the amount of RNA available and the segment of the transcriptome which is of interest.

If the sample amount is not an issue, then the most common sequencing method is the fragment sequencing, which (if ABI sequencers are used) gives strand specific reads. There is also another protocol by Nugen (http://www.nugeninc.com/nugen/index.cfm/products/amplification-systems/ovation-rna-seq-system/), which uses a lesser amount of sample, and which does rRNA depletion and sample amplification within the same protocol. If the sample amount is really low there is also a single cell RNA-Seq protocol (Tang, Barbacioru et al. 2010) which takes in a single cell's worth of RNA and amplifies it so that sufficient material is obtained to construct RNA-Seq libraries.

1.14.2 Applied Biosystems (ABI) SOLiD Sequencing

When we were selecting technologies for RNA-Seq, ABI offered the best sequencing depth and fragment length combination. SOLiD technology provides 50bp reads which are most importantly, strand specific, i.e. one can take a sequenced read and not only say which portion of the genome it came from but also say which *strand* it originated from. In the case of transcriptomics this feature is extremely useful.

SOLiD technology has a different method of "reading" bases when it sequences a read compared to other available technologies. Instead of reading one base at a time (base space), the SOLiD method reads two bases at the same time and this is done in a staggered manner so that each base is read twice. This results in increased accuracy of the read and higher mappable reads. However the disadvantage of this is that, compared to standard FASTA-like sequence outputs given by other sequencing technologies, the SOLiD platform results in sequence data encoded in "color space", where each base pair is represented not by their actual names (e.g. ATGC) but a number representing two neighboring bases, determined by each base pair. This adds another layer of complexity to the data. As a result there are fewer tools available to analyze color space data compared to standard base space data.

1.15 Results from RNA-Seq compared to traditional methods

Compared with traditional transcriptomics techniques such as real time PCR and microarray, RNA-Seq experiments tend to cost more, require a greater effort both during sample preparation and data analysis and requires specialized and expensive equipment. However despite these drawbacks the quality and depth of information provided by an RNA-Seq experiment is far superior to that obtained from any other conventional transcriptomics method.

In contrast to hybridization-based methods such as microarrays, RNA-Seq is a sequencing method. Therefore the technique is highly accurate and sensitive and not confounded by cross-hybridization effects. The technology is now mature enough to make available different protocols for different sample amounts (cell lines to embryos) which looks at different RNA types (mRNA to smallRNA).

Furthermore, unlike a microarray or a qPCR experiment where the sequence of the gene is critical to measure its expression level, RNA-Seq data is independent of known annotations. Due to this, RNA-Seq technology provides information on the entire transcriptome including both known and unknown entities. This enables the easy identification of new transcripts / genes. In addition, since RNA- Seq data contains information on all the exons of the transcripts, accurate expression levels can be calculated and also alternative splicing and alternative start events can also be studied.

Moreover, depending on the protocol, the RNA-Seq data can be strand specific. This means that in addition to the expression level of a particular transcript, its coding

strand can also be identified. All the data except the mouse embryo RNA-Seq data is strand specific.

Due to these advantages RNA-Seq is currently the best tool available for the study of transcriptomes.

2 Materials and Methods

2.1 Cell culture

Only the cells grown in feeder free conditions were used in the study to prevent the samples from being contaminated by mouse embryonic fibroblasts (MEFs). The WiCell H1 human embryonic stem cell line (WiCell research institute) was grown on conditioned hES media, at 37 ºC, in a humidified atmosphere with 5% CO_2 . The cells were routinely passaged every 7 days.

2.1.1 Preparation of conditioned human embryonic stem cell (hESC) media

Human embryonic stem cell (hESC) media was prepared by combining 800 ml of DMEM F12 (Gibco, #11330-032), 200 ml (or 20%) of Knockout serum (Invitrogen #10828028), 10 ml of 100 mM L- Glutamine (Gibco, #25030) with 7 µl of 2 - mercaptoethanol (Gibco, #21985-023), 10ml of non essential amino acid (Gibco, #11140) and 45 µl of 10% bFGF (Invitrogen, # 13256-029).

A 15 cm cell culture plate was coated with 0.1% gelatin (Stem cell technologies, #07903) overnight, and the plate was seeded with 4 million inactivated MEFs in MEF media.

On the second day, the MEF media was replaced with hES media, and from the third day to the tenth day the conditioned hES media was collected, and a new volume of hES media added daily. Finally the collected conditioned media was filter sterilized and 90µl of bFGF added for a final bFGF concentration of 4 ng / ml.

2.1.2 Passaging cells

First, the plates to which the split cells were to be added, were coated with Matrigel (BD, #354234) diluted with knockout DMEM (Gibco, #10829) to a dilution ratio of 1: 30.

During passaging, the cells were first incubated with a 1 mg / ml solution of type IV collagenase (Gibco, #17104019), for 5 - 7 minutes at 37 ºC . After the incubation and after ensuring that the edges of the cell colonies appear to be curled, the collagenase solution was replaced by hES medium. Then using a sterile 5 ml pipette, the cells were gently scraped and the cell suspension was centrifuged (Eppendorf, #5810R) at 800 rpm, for 1 minute at room temperature. After the centrifugation step the supernatant was removed and the cells were resuspended in conditioned media. The suspension was gently pipetted up and down to break the large cell clumps and the cells were added to the Matrigel coated plates.

2.1.3 Treatment of cells

SU5402 (Calbiochem, #572630) was dissolved in DMSO (Sigma, #D2650) before diluting in conditioned hES media for a final concentration of 20 µM and BMP4 (R & D Systems, # 314-BP /CF) was diluted in DPBS (Gibco, #14190) to a concentration of 100 µg / ml before being diluted to a concentration of 100 ng / ml. The media containing 20 µM of SU5402 and 100 ng / ml of BMP4 was added to cells during treatment and the media was changed daily.

2.1.4 RNA extraction

Extraction of RNA was done by using a combination of standard TRIzol (Invitrogen, #15596-018) method and the RNA extraction using RNAeasy mini kit (Qiagen, #74106). Please note that unless otherwise stated the centrifugation steps and incubation steps were done at room temperature.

Each 15cm dish containing H1 hESC colonies were first washed twice with 10ml of PBS (Gibco, 14190 - 144) followed by the addition of 6 ml TRIzol. After incubating for 5min with TRIzol, the lysed cells were mixed well by pipetting up and down and divided, 1 ml each, into 1.5 ml micro-centrifuge tubes (Eppendorf, MCT-175-C). For each tube (containing 1 ml of TRIzol), 200 µl of chloroform was added and incubated for 3 minutes. This was followed by a centrifugation step at 4 ºC for 15 minutes at 12,000 rpm (Eppendorf, #5415R). After the centrifugation the aqueous layer was carefully placed into another 1.5 ml micro-centrifuge tube and 500 µl of isopropanol was added to it. This was incubated for 10 minutes and centrifuged at 12,000 rpm for 4 ºC for 10 minutes. The supernatant was discarded and the remaining pellet was washed by adding 1 ml of 75% ethanol followed by a centrifugation step at 10,000 rpm for 5 minutes at 4ºC. The remaining washed pellet was dissolved in 100 µl of RNase free water (Ambion, #AM9937) and the Qiagen mini RNAeasy kit (Qiagen, # 74106) was used to process the resulting RNA solution.

350 µl of Buffer RLT (with 1%, 2- mercaptoethanol (Gibco, # 21985 - 023)) was added to the 100 µl RNA solution followed by 250 µl of 100% ethanol. The resulting solution was mixed by gently pipetting up and down and then applied to a RNeasy mini column. This was then centrifuged for 30 seconds at 12,000 rpm and the flow-

through was discarded. After changing the collection tube, 500 µl of buffer RPE was added to the spin column and centrifuged for 30 seconds at 12,000 rpm and the flow-through discarded. Then the spin column was again centrifuged for 1 minute at 12,000 rpm. The RNeasy mini column was placed in a new sterile 1.5 ml micro-centrifuge tube. The RNA was eluted out by adding 20 µl of RNase free water directly onto the filter membrane of the spin column, incubating for 1min and centrifuging for 12,000 rpm for 1 minute. This step was repeated once to elute the remaining RNA from the membrane.

2.1.5 Checking RNA concentration, purity and integrity

RNA concentration and purity was measured by the NanoDrop spectrophotometer (Thermoscientific, #ND-1000). The NanoDrop uses absorbance at 260nm wavelength to predict the RNA concentration based on the Beer - Lambert law. The 260 / 280 absorbance ratio was used as a measure of RNA purity and a value above 2.0 was considered to be pure.

The RNA integrity was evaluated by performing a Agilent RNA 6000 pico assay (#5067-1513) on the Agilent bioanalyzer. The RNA integrity number (RIN) gives the integrity of the RNA sample in a scale of 0 - 10 where 10 is the highest. All samples used had a RIN value of more than 9.

2.1.6 Poly (A) RNA purification

The Poly (A) Purist MAG kit (Ambion, #AM1922) was used for extracting RNA transcripts with a poly A tail from the total RNA extract. The poly (A) purist mag kit

uses oligo (dT) magnetic beads and a magnet to capture RNA transcripts with a poly A tail. The capture process involves placing the micro-centrifuge tube with the sample / wash solution in the holder of the magnet for 2 minutes letting the magnetic beads attach to the surface closest to the magnet and carefully removing the liquid portion. 100 µg of total RNA was used for the mRNA extraction for each sample.

The total RNA concentration was adjusted to 600 µg / ml by adding RNase free water and to this diluted RNA solution, an equal volume of 2X binding buffer was added and mixed. 10 µl of oligo (dT) beads were used per 100 µg of RNA. The beads were first washed twice with wash solution 1 prior to use. The RNA in binding buffer was then mixed with the beads and the mixture was incubated for 5 minutes at 70 ºC. After this it was incubated for 60 minutes on a shaker (Labnet, #S2030 - RC - 220) at room temperature, with gentle rocking. The beads were captured and washed twice with wash solution 1 and wash solution 2, respectively. The volume of wash solutions used was equal to the volume of the initial diluted total RNA. The poly (A) RNA was eluted from the beads by two 200 µl washes of the RNA storage solution heated to 75ºC. The RNA was then precipitated using an incubation step of 1 hour at -80ºC with 0.1 volumes of 5M ammonium acetate, 1 µl Glycogen and 1.1 ml of 100% ethanol. After incubation the poly (A) RNA pellet was isolated by a centrifugation step of 30 minutes at 12,000 g at 4ºC. The pellet was then washed with 1ml of 70% ethanol followed by a centrifugation step for 10 minutes at 4ºC . Finally the poly (A) RNA pellet was resuspended in 30 µl of RNase free water.

The concentration and the efficiency of poly (A) extraction was measured using a bioanalyzer trace.

2.1.7 Whole transcriptome library preparation for SOLiD sequencing

The whole transcriptome library preparation consisted of fragmenting the poly (A) RNA, adapter ligation, reverse transcription, size selection and amplification. For these steps the contents of the ABI whole transcriptome library preparation kit (ABI, #4425680) was used.

2.1.7.1 Fragmentation

750 ng of poly (A) RNA diluted in 8 µl of RNase free water was used for fragmentation. 1 µl each of 10X RNase III reaction buffer and RNase III was added to the diluted poly (A) RNA solution. It was mixed by gently pipetting up and down and incubated for 10 minutes at 37 ºC in a thermocycler (BioRad, tetrad 2). Immediately after the incubation 90 µl of nuclease free water was added to the reaction mix.

Following the fragmentation step an RNA cleanup step was performed using the Ribominus concentration module (Invitrogen, #K155005). 100 µl of binding buffer L3 and 250 µl of 100% ethanol was added to the fragmentation reaction mix with water. The mixture was placed in a spin column and centrifuged for 1 minute at 12,000 g. After discarding the flow-through, 500 µl of buffer W5 was added to the spin column and it was centrifuged for 1 minute followed by another 2 minute centrifugation, after removing the flow-through. To elute the RNA, the spin column was then placed on a recovery tube. The elution was done by two 20 µl wash steps using RNase free water, an incubation step of 1 minute, and a centrifugation of 1 min at 12,000 g.

The Nanodrop spectrophotometer was used to quantify the resulting fragmented RNA and the Bioanalyzer was used to measure the size distribution.

2.1.7.2 Hybridization

75 ng of fragmented RNA in 3 µl of RNase free water, 2 µl of adapter mix A and 3 µl of the hybridization solution was mixed on ice. The resulting mixture was incubated 65 ºC for 10 minutes and 16 ºC for 5 minutes, using a thermocycler with a heated lit. After the two incubation steps 10 µl of 2X ligation buffer and 2 µl of the ligation enzyme mix was added and the resulting ligation mix was incubated for 16 hours at 16 ºC on a thermocycler with the heated lid turned off.

2.1.7.3 Reverse transcription

The reverse transcription master mixture (per sample), was prepared by mixing together 13 µl of Nuclease water, 4 µl of 10X RT buffer, 2 µl of 2.5mM dNTP mix and 1 µl of array script reverse transcriptase enzyme, on ice. This RT mix was added to the 20 µl ligation mix after the 16 hour incubation. After gently mixing, the reverse transcription was carried out in a thermocycler at 42 ºC with a heated lid for 30 minutes.

2.1.7.4 cDNA purification

For cDNA purification the Qiagen PCR purification kit was used (cat #28106). The centrifugation steps were carried out at room temperature.

The resulting cDNA from the reverse transcription was transferred to a 1.5ml micro-centrifuge tube and was mixed with 60 µl of RNase free water and 500 µl of Buffer PB. The resulting 600 µl solution was added to a mini elute column and centrifuged for 13,000 g for 1 minute. The followthrough was discarded and the spin column was placed on a new centrifuge tube followed by another centrifugation step of 1 minute at 13,000 g. The spin column was placed on another clean micro-centrifuge tube and 10 µl of buffer EB was added to the spin membrane. After an incubation of 1 minute the purified cDNA was extracted by a centrifugation step of 13,000 g for 1 min.

2.1.7.5 Size selection

The size selection of the cDNA is done using a gel purification step which uses Novel reagents and NuPage gels (Invitrogen, # EC6865BOX).

5 µl of purified cDNA was run on a Novex 6% TBE-Uread gel (using 1X TBE running buffer on the Xcell surelock mini-cell electrophoresis system (Invitrogen, EI0001)). Once the gel has run for 15 minutes it was stained with SYBR gold nucleic acid stain (Invitrogen, #S11494) and the gel band corresponding to the range 100 - 200bp was excised and cut into four equal sized vertical bands.

2.1.7.6 cDNA amplification

The PCR mastermix was made by adding 171.6 µl of Nuclease free water, 22 µl of 10X PCR buffer, 4.4 µl of SOLiD PCR primer 1, 17.6 µl of 2.5mM dNTP and 4.4 µl of Amplitaq DNA polymerase, per sample. Two of the gel pieces cut in the above step were individually put on 2 PCR tubes and 100 µl of the PCR master mix was added to each. The program for the PCR was set as follows. A holding step of 95 ºC for 5

minutes. 15 cycles of of 95ºC for 30 seconds, 62 ºC for 30 seconds and 72 ºC for 30 seconds. And a final holding step of 72 ºC for 7 minutes.

2.1.7.7 Purification of the amplified cDNA

The Purelink PCR micro kit (Invitrogen, #A11199) was used for the purification of the amplified cDNA. All centrifugation steps were done at room temperature.

The PCR reaction solution in both of the tubes were pooled into a 1.5ml microcentrifuge tube and mixed with 800 µl of binding buffer B2. This was added to a Purelink column, centrifuged for 1 min at 10,000 g and flow through discarded. The centrifugation step was repeated and then the purified cDNA was eluted with two washes of 10 µl of elution buffer by incubating for 1 minute and spinning for 14,000 g for 1 minute.

A bioanalyzer trace was obtained for the finished DNA and the sample was submitted for sequencing.

2.2 smallRNA RNA-Seq

2.2.1 Extraction of smallRNA enriched RNA

The mirVana miRNA Isolation kit (Ambion, # AM1560) was used for the extraction of RNA, enriched with small RNA. This kit allows the extraction of RNA less than 200 nucleotides. The treated cells and the control cells were lysed with 600 µl of Lysis / Binding buffer, inside the culture dish and the cell lysate was collected. 1/10 volume of miRNA homogenate additive was added to the lysate and it was incubated for 10min on ice. An equal volume of Acid-Phenol:Chloroform was then added to the mixture, and it was centrifuged at 10,000 g for 5 minutes. After the centrifugation the upper (aqueous) phase was carefully transferred to a new tube and mixed with 1/3 volumes of 100% ethanol. The mix was then transferred to a filter cartridge and after a centrifugation step (10,000g,~15sec) the filtrate was collected. A 2/3 volume of 100% ethanol was added to the filtrate, and it was again filtered using a filter cartridge. The flow through was discarded and the filter was washed with 700 µl of miRNA wash Solution 1 followed by two washes of 500 µl of wash solution. The RNA enriched with small RNA was then eluted with 100µl of heated (95ºC) nuclease free water.

The RNA integrity and the size distribution of the RNA was evaluated by using the Agilient bioanalyzer.

2.2.2 Library preparation

For the library preparation for smallRNA RNA - Seq, the SOLiD Total RNA-Seq kit for small RNA libraries protocol (ABI, # 4452439) was used. This protocol is in principal similar to the standard RNA-Seq library preparation, but excludes the

enzymatic fragmentation step. 200ng total RNA enriched for smallRNA was directly hybridized with SOLiD adaptors and the library generation was done as per the above stated protocol. The library generation steps are omitted here to avoid repetition with the standard RNA-Seq library preparation section.

2.2.3 RNA-Seq library generation system with the Ovation system

The standard ABI RNA-Seq library preparation protocol requires a minimum of 100ng - 200ng of rRNA depleted or poly A RNA. This RNA requirement becomes an issue when limited samples such as mouse embryos are concerned. Therefore to analyze the transcriptome of the mouse embryos, which yield very little RNA, the Ovation RNA-Seq system by Nugen (Nugen, # 7100-08) was used. The Ovation kit can amplify RNA (in a linear manner) starting from as little as 500 pg and produce around 3µg of RNA. Apart from the impressive amplification the other advantage of this method is that its amplification does not solely depend on the poly A tail of the transcripts. It uses random priming for amplification, where the primers are specifically designed to bind to all RNA except ribosomal RNA. The amplified product from the Ovation kit was further processed using the ABI fragment library preparation protocol to produce the libraries.

2.2.4 RNA extraction

RNA extraction of mouse embryos was done using the pico pure RNA extraction kit.

2.2.5 RNA amplification by the Nugen Ovation kit

2.2.5.1 First strand cDNA synthesis

500 pg of RNA in a 5µl solution was mixed,on ice, with 2µl of A1 solution, 2.5µl A2 solution and 0.5µl A3 solution. The mix was then put in a thermocycler and the following program was run - (4ºC 1 min, 25ºC 10min, 42ºC 10min, 70ºC 15min, 4ºC hold).

2.2.5.2 Second strand cDNA synthesis

9.7µl of B1 solution and 0.3µl of B2 solution was added to the products of the first strand cDNA synthesis step. It was then placed in a thermocycler and the following program was run (4ºC 1 min, 25ºC 10min, 50ºC 10min, 80ºC 20min, 4ºC hold).

2.2.5.3 Purification of double stranded cDNA

RNAClean beads were used for this step. The beads were first resuspended and allowed to return to room temperature. 32µl of the bead mix was added to the products of the previous step and incubated for 10 min at room temperature. The beads were then aggregated using a magnet and 42µl of the cleared buffer was removed. Then the beads were washed three times with 200µl 70% ethanol and air-dried for 20 min.

2.2.5.4 SPIA Amplification

The SPIA mastermix was prepared by mixing on ice 20µl of C2 solution, 10µl of C1 solution and 10µl C3 solution. This was added to the air dried RNAClean beads

containing the cDNA. The mix was then put in a thermocycler and the following program was run (4ºC 1 min, 47ºC 60min, 95ºC 5min, 4ºC hold).

2.2.5.5 Post SPIA Modification

The RNAClean beads were aggregated using a magnet and the supernatant (35µl) was put into a new tube. To the supernatant 5µl of E1 primer was added and the mix was incubated for 3 min at 98ºC in a thermocycler. After the incubation, 5µl of E2 solution and 5µl of E1 was added and the resulting mix was put in the thermocycler and the following program was run (4ºC 1 min, 30ºC 10min, 42ºC 15min, 75ºC 10min, 4ºC hold). This produces the final amplified cDNA. The cDNA was purified using QIAquick PCR purification kit (Qiagen, # 28104).

2.2.5.6 Library preparation

The SOLiD fragment library kit (ABI, #S3100101) was used to prepare the small RNA-Seq libraries using the amplified cDNA.

2.3 Bioinformatic analysis of RNA-Seq

During an RNA-Sequencing run, the sequencer records the nucleotide sequence of all the sequenced reads. In order to make biologically relevant interpretations these sequenced reads must be first mapped to the genome to identify the region it originated from, and then the reads should be counted so that the expression level of the particular region they map to, can be measured. For these above mentioned steps Bioscope software (version 1, ABI) was used.

2.3.1 Alignment / mapping

Bioscope uses the software - mapreads (also known as, SOLiD system colour space mapping tool)(http://solidsoftwaretools.com/gf/project/mapreads/) for the alignment of reads to the genome. Mapreads uses a seed and match strategy for mapping. In this approach, the software first tries to find an initial alignment of 25 bases between the read and the genome (the seed) with a maximum of 2 mismatches. Once it finds such a place the alignment is extended to the entire length of the read, and an alignment score is calculated by giving a score of +1 for each correctly aligned base and -2 for each misaligned base. During alignment the mapreads software looks at up to ten positions each read aligns to and only considers a read as uniquely aligned if the read maps to only one position or if the difference in score between the best alignment and the next best is more than 4. During the mapping step the reads are mapped to the filter sequences (described below), splice junctions and the genome in parallel.

2.3.2 Mapping to filter sequences

If the reads which originate from the repeat sequences of the genome is used for the genomic alignment they can cause unnecessary computational overhead and incorrect results, as they would match to multiple locations of the genome with virtually the same alignment score. To counter this problem, the reads are mapped to a database of known repeat sequences which in effect filter them out and prevent them from being mapped to the genome. In our case the repeat database contained ribosomal sequences, tRNA sequences and other common repeat sequences. By looking at the total number of reads mapping to rRNA sequences, the efficiency of the rRNA removal method can be validated.

2.3.3 Mapping to the reference genome

This is the single most computationally intensive and most time consuming process of the RNA-Seq data analysis. The data described here were aligned to the hg18 build of the human genome. This step enables the identification of the genomic regions which are being transcribed, and since the alignment is done to the whole genome even the unannotated but transcribed regions can be identified.

2.3.4 Mapping to the splice junctions

The seed and extension method used by mapreads works well for a majority of reads, which originate from exon bodies. However the reads which originate from splice junctions do not get mapped to the genome due to the presence of introns in the genome which in this case is the reference sequence. To counter this problem Bioscope tries to map the reads to a database of all possible exon exon junctions

within a given RefSeq gene. This step manages to recover the splice junction reads. Normally splice junction reads are much less in number as compared to the number of reads which mapped to their corresponding exons. This is understandable as the splice junction reads represent only a small portion of the entire footprint of the transcript. However they are important in identifying and quantifying the alternative splicing events of genes, as they act as markers of linkages between two exons.

2.3.5 Counting known transcripts

The counting step quantifies the reads which map to a particular transcript or a gene. It should be noted that the final counts file shows read counts per exon. Post processing is needed to obtain the total number of reads which map to a particular exon or a gene. During counting the reads which get aligned to the genome get counted if they have less than 3 bases outside the given exon. As for reads which get mapped to splice junctions, they contribute to the count of the exon if it starts or ends at the boundary of the exon.

2.3.6 File formats

Bioscope software produces several files in several formats which contain the RNA-Seq data.

Counts file : This is a tab delimited text file which contains the total number of reads mapping to individual exons of RefSeq annotation.

wig file format: These files contains data on the expression level of each base in the genome. These can be uploaded to the UCSC genome browser for visualizing the data.

GTF file format: This contains the sequence and the mapping location of each and every read which align to the genome.

Filter files: Contains statistics on the number of reads which align to the repeat sequences.

2.4 Calculating expression levels

The simplest measure of the expression level of a particular gene is the total number of reads which align to it. However the raw read count is not a good indicator of expression level as it is dependent on the length of the sequence as well as the sequencing depth. (*i.e.* longer the transcript, the higher the total read count, the higher the total sequenced reads, the higher the read count). The RPKM value (Mortazavi, Williams et al. 2008) was introduced to nullify the effects of these two factors on the expression level of a particular transcript or a gene. The RPKM value was calculated by normalizing the total number of reads which fall on all the exons of the gene, with the length of all its exons and the sequencing depth. Total mapped reads for the entire genome was used as a representative value for the sequencing depth.

RPKM value for a particular gene was calculated using the following equation:

$$= [\ \sum(\text{read counts of exons}) / \{\ \sum\ (\text{exon length})/1000 \text{ X total mapped reads}\}\ ] \text{ X } 1000000$$

2.4.1 Microarray Data

The Illumine Single Color Human Ref-8 Version 2 microarray data was first imported into GenomeStudio (Illumine) for background correction. Then the data was imported into GeneSpring GX (Version 11.0, Agilent Technologies Inc.) where it was

normalized (Shift to 75 percentile, Baseline transformation - median of all samples) and analyzed.

2.4.2 Comparing expression levels of RNA-Seq data and microarray data

To compare the expression levels obtained by RNA-Seq and microarray, the fold change of all the RefSeq genes (day 0 vs day 8 of treatment) were calculated using RPKM data values for RNA-Seq and normalized probe intensity values for microarray data. To keep the fold change values accurate only genes showing RPKM values of more than one and probe intensity values of more than 20 at both time points were used in the comparison. The fold changes were converted into signal log ratios (SLRs) by converting the fold change into its $\log_2$ value and the resulting SLR values of RNA-Seq and microarray data were plotted against each other and the coefficient of determination (R^2) value was calculated for the two datasets.

2.4.3 Gene ontology analysis

Gene ontology analysis was performed using the Panther classification system (http://www.pantherdb.org/, genome biology article). The geneIDs were uploaded and the enriched human gene ontology terms were extracted.

2.4.4 Hierarchical clustering

To gauge the differentiation to the trophoblast lineage brought about by the treatment, the microarray expression levels of the treated samples were compared against a compilation of published microarray data of normal tissues and cell lines (Ge, Yamamoto et al. 2005; Burleigh, Kendziorski et al. 2007; Bilban, Tauber et al. 2010).

All the data were normalized using default parameters using Genespring GX, and exported into Genesis (Sturn, Quackenbush et al. 2002). There the data was used to perform a Hierarchical Clustering using Pearson correlation.

2.5 Interpreting UCSC browser views

The main usable dataset of any RNA-Seq experiment includes the locations of the genome where each and every sequenced read map to. In this thesis the UCSC genome browser (www.genome.ucsc.edu) (Kent, Sugnet et al. 2002) was used to visualize the data. The UCSC genome browser has the capacity to take in the large RNA-Seq dataset and display it as peaks, which denote the expression of a particular region. The USCS browser is also capable of showing existing annotation data (exons, genes, ESTs, microRNAs etc.) together with the above mentioned RNA-Seq peaks.

2.5.1 The organization of data in the UCSC genome browser

UCSC genome browser displays data and annotation information based on the genomic coordinates. Each browser view, which is the image the user sees, is composed of 1) a data section, where the user provided data is displayed - in this case RNA-Seq peaks and 2) an annotation section which comes built into the browser, where the user can select the information to be displayed - such as RefSeq genes. Both data and annotations are organized as tracks which are "strips" of either data or annotation whose appearance and the order can be customized by the user (see Figure 2.1).

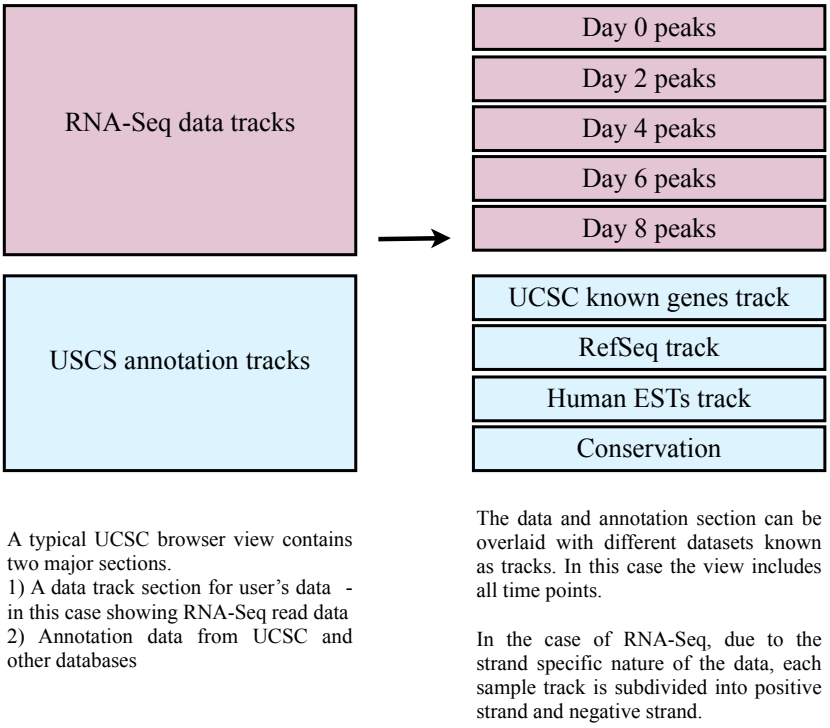


Figure 2.1: Organization of data and annotation in the UCSC genome browser. UCSC genome browser displays data based on the genomic co-ordinates. Annotations (shown in blue) for a particular genomic region and the RNA-Seq peaks (shown in red) corresponding to the region are overlaid on-top of each other as "tracks". The user has the ability to upload custom data (in this case the RNA-Seq data) and also to select which annotation types are selected.

2.5.2 A typical view of the UCSC browser

The RNA-Seq data on the trophoblast differentiation (mRNA and small RNA) is strand specific which means that by looking at the alignment of a particular read the location as well as the strand which it originates from can be identified. During visualization of RNA-Seq data using the UCSC genome browser the strand specificity is represented by two tracks (one for each strand) per sample.

Figure 2.2 describes a typical view of the UCSC browser loaded with RNA-Seq data.

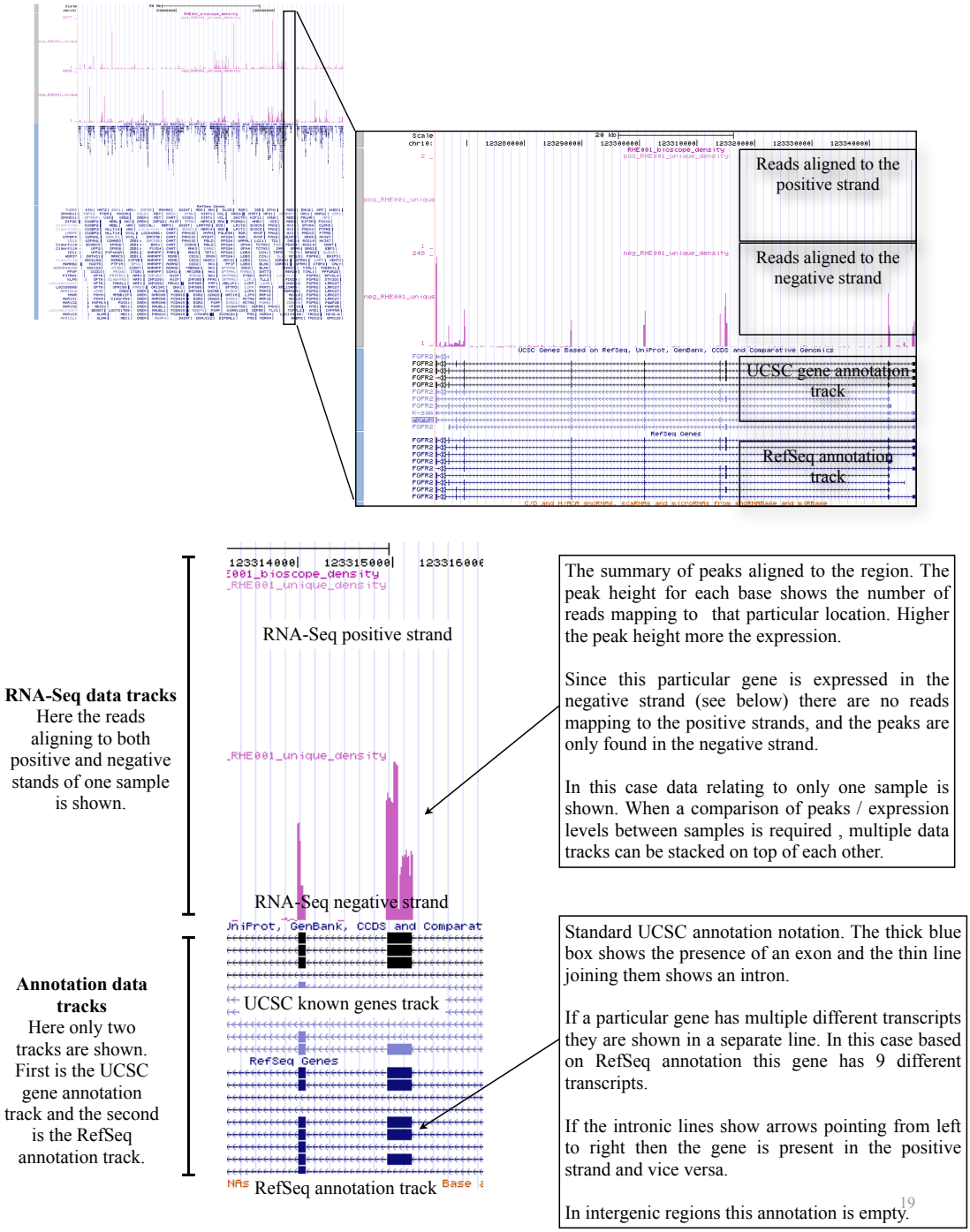


Figure 2.2: Visualizing RNA-Seq data using UCSC genome browser
UCSC browser provides co-ordinate specific information of the genome onto which user data (in this case data from RNA-Seq) can be overlaid. The panel on the left shows the read data of the entire chromosome 10, while the panel on the right shows a small enlarged view marked by the black rectangle. For clarity positive and negative stand tracks from only one sample are shown.

3.0 Results 1
Programatic workflows designed for the analysis of RNA-Seq data

The function of Bioscope software from ABI, which is provided together with the sequencing platform, is primarily to align sequence reads to the reference genome. Thus to perform analysis of the RNA-Seq dataset beyond expression levels, I developed a set of programmatic workflows. These workflows, coded using the python programming language (www.python.org), were designed to use the data / files produced by Bioscope as input and produce results files on various aspects of transcriptome dynamics.

Over the course of this thesis project, programmatic workflows were designed for the following tasks.

- Identification of genes which show changes in their splicing profile during treatment
- Identification of genes with exons which show mutually exclusive expression
- Identification of novel transcribed regions
- Identification of novel transcripts
- Identification of extensions of annotation

All the workflows described here, when combined, make up a suite of software utilities which enables the rapid identification of interesting transcriptomics phenomena from an RNA-Seq dataset. Since these were developed during the course of the thesis and used for the analysis of data presented here, the workflows are shown in the results section.

3.1 Workflow for identifying genes for which the splicing pattern is altered during treatment

This workflow identifies genes which undergo alternative splicing events of a RNA-Seq dataset. Apart from altering the expression level of a gene, a treatment can also cause a change in its splicing profile (i.e. induce or suppress the expression of different isoforms of the same gene). In cases where alternative splicing takes place, the overall expression of the gene might not change significantly, even though a considerable change in function could occur.

A straightforward approach to identifying alternative splicing events from an RNA-Seq experiment would be to use junction reads (reads which originate from the exon - exon boundary). This approach works well provided that there is a significant amount of junction reads available in the dataset. Unfortunately in the case of most RNA-Seq datasets this is not so, as the commonly used fragment library protocol produces fragments of 50bp and this short sequencing length reduces the likelihood of a junction read being mapped to the genome. This reduces the number of junction reads discovered.

In order to compensate for the lack of junction reads, this workflow was designed to identify splicing events purely based on individual exon counts and not on junction reads. Despite not using split reads, with the correct settings, this workflow produces good quality predictions with a minimum number of false positives.

A change in the number of reads mapping to a particular exon during a time course, can be due to either a change in the expression level of the gene or due to a splicing

event or both. Thus when trying to identify genes which show an altered splicing profile it is vital to negate the effects of changes in expression levels. In the workflow this is done by comparing the expression level of an exon relative to its neighbors. The assumption here is that while expression level of the gene is proportional to the read counts of its individual exon, a change in gene expression should not change the proportion of contribution made by individual exons to the gene expression level if no splicing event takes place. For example, in the case of the three exons shown in Figure 3.1 - left panel, when the splicing profile remains unchanged, each has a read count of 100 which goes down to 50 during treatment. Even though their expression levels change, the ratio of expression levels remains constant (100:100:100 = 1:1:1 = 50:50:50) because the splicing profile remains the same. However when there is a change in splicing the ratio cannot be maintained, as splicing selectively increases or decreases the read counts of a exon (100:100:100 vs 50:25:50). By detecting this phenomenon the workflow can predict the alteration of the splicing profile of a gene.

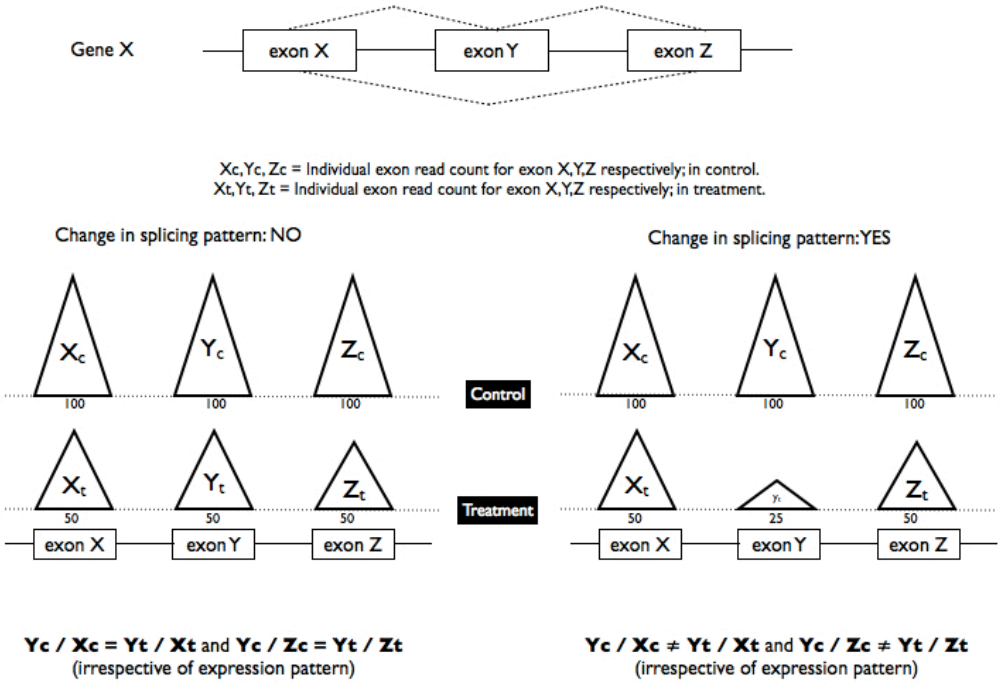


Figure 3.1: The workflow for identifying genes whose splicing profile is altered during treatment.

If there is a change in the splice profile it would be seen as a marked increase or decrease in read counts of a particular exon with respect to its neighboring exons. This method of normalization negates the effect of changes in read count due to up or down-regulation of the gene.

3.2 Workflow for identifying genes which show mutually exclusive exon patterns

This workflow is an extension of the (above mentioned) method for detecting genes which show changes in their splicing profile during a treatment. Here the expression pattern (up-regulation or down-regulation) of individual exons belonging to a particular gene is monitored during treatment to identify a pair or more of exons which show an opposite regulation pattern. For example in the case of Figure 3.2, genes with the expression pattern on the left will be discarded as all of them show a similar type of regulation. However genes showing an expression pattern on the right will be identified as having mutual exclusively expressed exons as it has one down-regulated and one up-regulated exon.

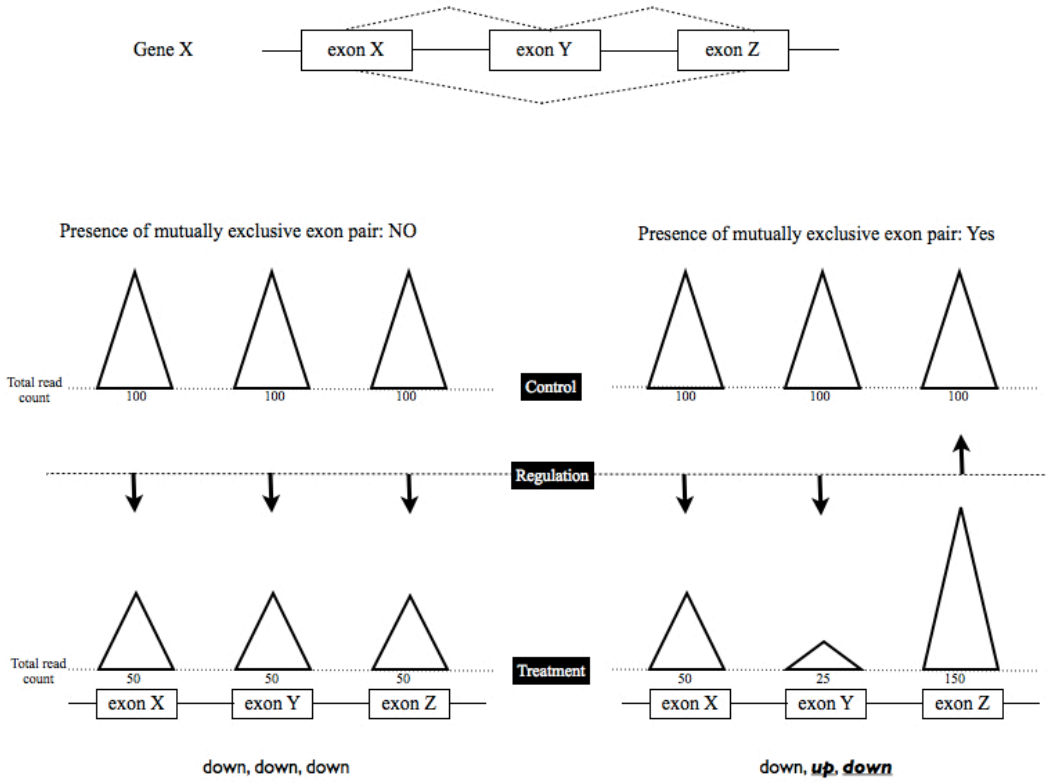


Figure 3.2: Workflow for the identification of genes with mutually exclusive exons.
The workflow looks for a pair or more exons which are regulated in an opposing manner in two time points.

3.3 Workflow for the Detection of novel transcribed regions (NTRs)

Novel transcribed regions (NTRs) are defined as unannotated regions which are transcribed, as shown by RNA-Seq data. Throughout this study, RefSeq annotations were used to identify NTRs, although it should be mentioned that the NTR detection workflow was designed to take in any annotation. The identified NTRs could be new exons of known transcripts, extensions of known genes or totally new transcripts.

The first step of the NTR detection process involves identifying unannotated regions (i.e gaps between annotated regions) of the genome. In the case of RefSeq annotation these gaps include introns as well as regions between gene footprints. In the next stage, regions identified as unannotated are probed, to find locations which are shown to be expressed by having reads mapping to it. These regions are identified as NTRs provided that they satisfy user provided expression criteria which includes the minimum height of the peak and the read count of the NTR peak.

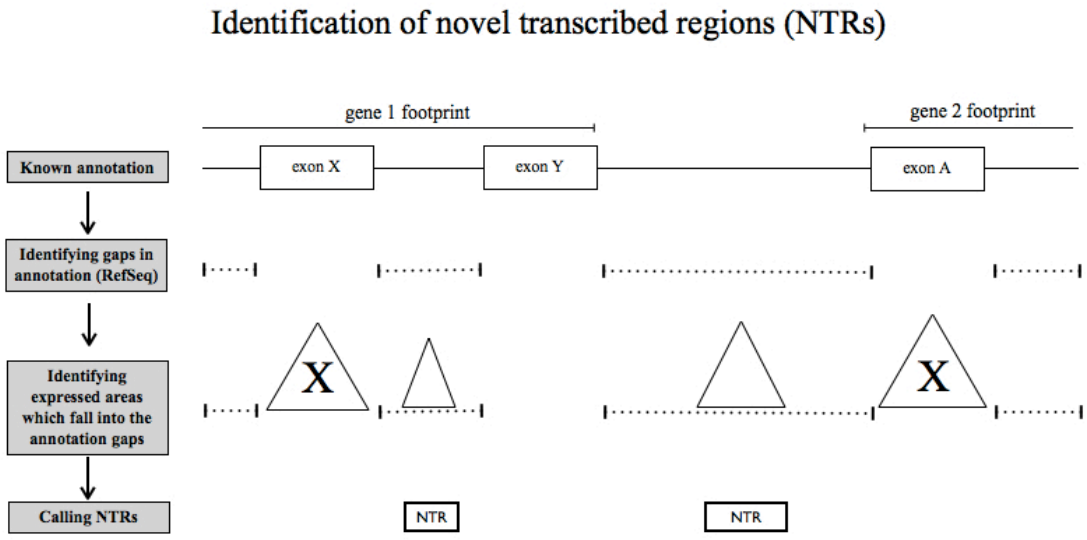


Figure 3.3: Novel transcribed regions (NTRs) identification workflow.
NTRs are identified by finding expressed regions in un annotated regions.

3.4 Workflow for the detection of novel transcripts using NTR data

Identification of novel transcripts is done by clustering the novel transcribed regions based on the distance between them. The premise is that if there is a series of novel transcribed regions with close proximity to each other then there is a very good likelihood that they belong to the same transcript (NTRs acting as exons of the new transcript). While NTRs are dispersed throughout the genome and can be found in intronic regions, close to known genes and far away from genes, the NTRs which are important in identifying novel genes should ideally exist as a cluster of peaks located a considerable distance away from known genes.

Thus the transcript identification workflow first filters out NTR regions which lie close to known exon and gene footprints. In the first pass it filters out NTRs which fall within 25 nucleotides before and after an exon boundary. NTRs filtered at this step are used to redefine the known exon boundary based on expression data. Then in the second pass it removes NTRs which fall within 10,000 nucleotides before and after a known gene boundary. The NTRs removed here could potentially be novel exons of known genes. Resulting NTRs are used for novel transcript identification. A set of NTRs are recognized as part of a novel transcript if they are within 10,000 nucleotides of each other. Once the potentially novel genes are identified they are grouped based on whether they are expressed at all time points, a few or at only one.

3.5 Workflow for the identification of Extended exon footprints

RNA-Seq data, when aligned to the entire genome, is not dependent on existing annotation. While in most cases the existing RefSeq annotation matches with the footprints of the expressed regions as obtained by RNA-Seq, there are some striking examples where expression occurred beyond the existing annotation. These 'extensions' of expression beyond RefSeq annotation, were seen in exons, 3' UTR and 5 'UTR regions. A workflow was designed to identify regions where there is a significant difference between RNA-Seq data and existing annotation.

To identify extensions in ordinary exons (non UTR exons), the workflow first identifies exon boundaries from RefSeq and extends the exon footprint until it covers all the expressed bases on either side of it. This process essentially corrects the exon annotation, based on the RNA-Seq data.

To identify extensions of the 3' and 5' UTR, the workflow starts at the end of the UTR and tries to extend the footprint using the expressed regions. Since most UTRs contain regions where mapping efficiency is low, the workflow allows the extended region to have gaps of less than 100 base pairs.

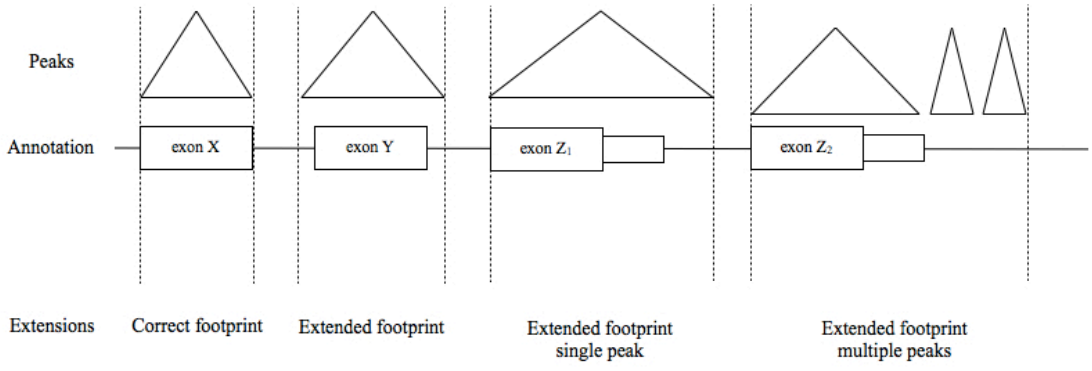


Figure 3.4: The workflow for the identification of exon extensions.
The extensions are identified by comparing RNA-Seq data and RefSeq data.

3.6 Workflow for the discovery of expressed repeat regions

Repeat regions usually consist of long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), long terminal repeat elements (LTRs), DNA repeat elements, satellite repeats and RNA repeats.

Bioscope software automatically does basic filtering of some repeat regions during alignment. However the filter process is not extensive and, based on the unique alignment, it seems that a significant proportion of repeat regions do show expression.

Here the main interest was to identify expressed repeat regions which did not have any overlap with RefSeq annotations. Thus during this workflow the repeat regions (which were downloaded from UCSC repeat masker track) that did not have an overlap with RefSeq genes were first identified. Subsequently the reads which map to these individual regions were counted.

3.7 Workflow for the identification of novel splice sites

As mentioned earlier, the efficiency of junction reads mapping is low in the case of 50bp fragment RNA-Seq. While this reduces its effectivity in identifying alternative splicing events, these can be used to identify un-annotated exon - exon junctions (i.e.. novel splicing between known RefSeq exons). The inherent flaw in this method is that it only picks up novel exon junctions of highly expressed transcripts as they have the highest chance of generating a significant number of junction reads.

To identify novel junctions, the novel junction reads identified during alignment are filtered to obtain only the reads which are of best quality and align perfectly to the junction. These reads are then grouped based on the gene they belong to.

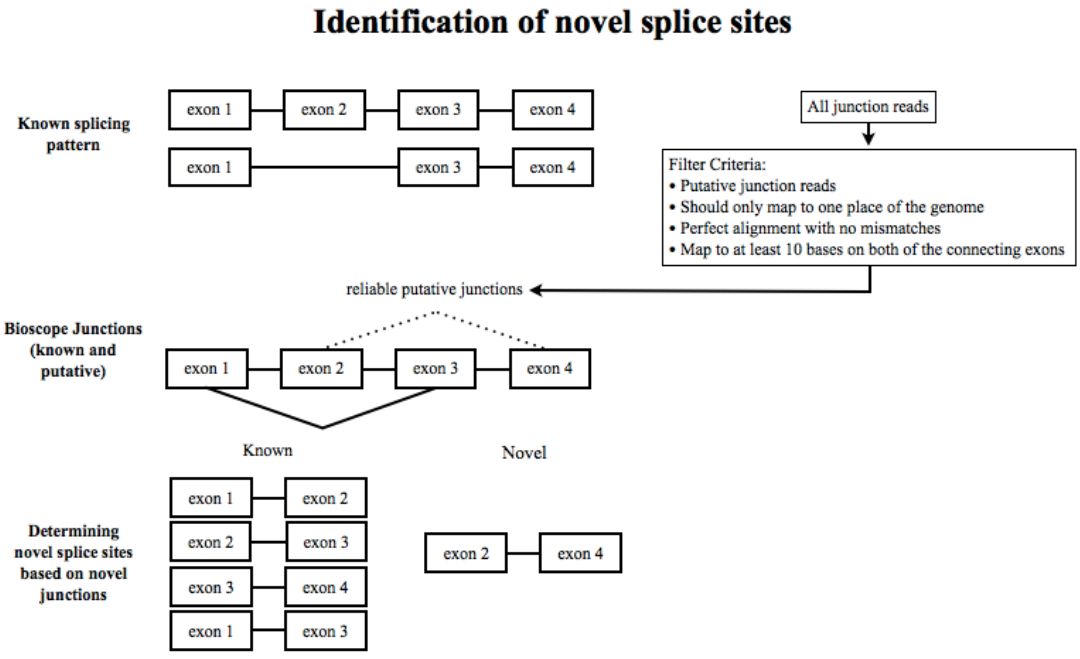


Figure 3.5 : Novel junction identification workflow
The workflow for identifying novel splice sites uses the Bioscope junction reads which are conveniently labelled to be either known (from a known splice site) or putative (unknown splice site). To identify reliable splice sites the junction reads are first filtered to remove any potentially mis - aligned reads. Then the filtered putative junction reads are used to identify the corresponding exons connected by the junction, thereby defining novel splice sites.

3.8 Workflow for the identification of novel microRNA from smallRNA RNA-Seq

From a theoretical point of view, the smallRNA-Seq experiment should be able to capture all the smallRNA expression events that takes place during the experiment. By using the novel transcribed region (NTR) detection workflow described above, all the potentially novel small RNA in the transcriptome can be identified. Looking at the data, as shown in the second results section, most of the novel smallRNA turn out to be within the size range of mature microRNAs. Therefore a workflow was developed to identify novel microRNA from small RNA-Seq data.

In this workflow (Figure 3.6), the unique properties of the microRNA stem loop structure was used to confirm whether an NTR was a microRNA. In this workflow, an NTR discovery was done using the smallRNA sample, and then the footprint of the NTR was expanded to encompass the sequence of the stem loop structure. This DNA sequence was then programmatically folded using RNAfold (Hofacker 2003) using minimum free energy to see if it was capable of producing a stable stem loop.

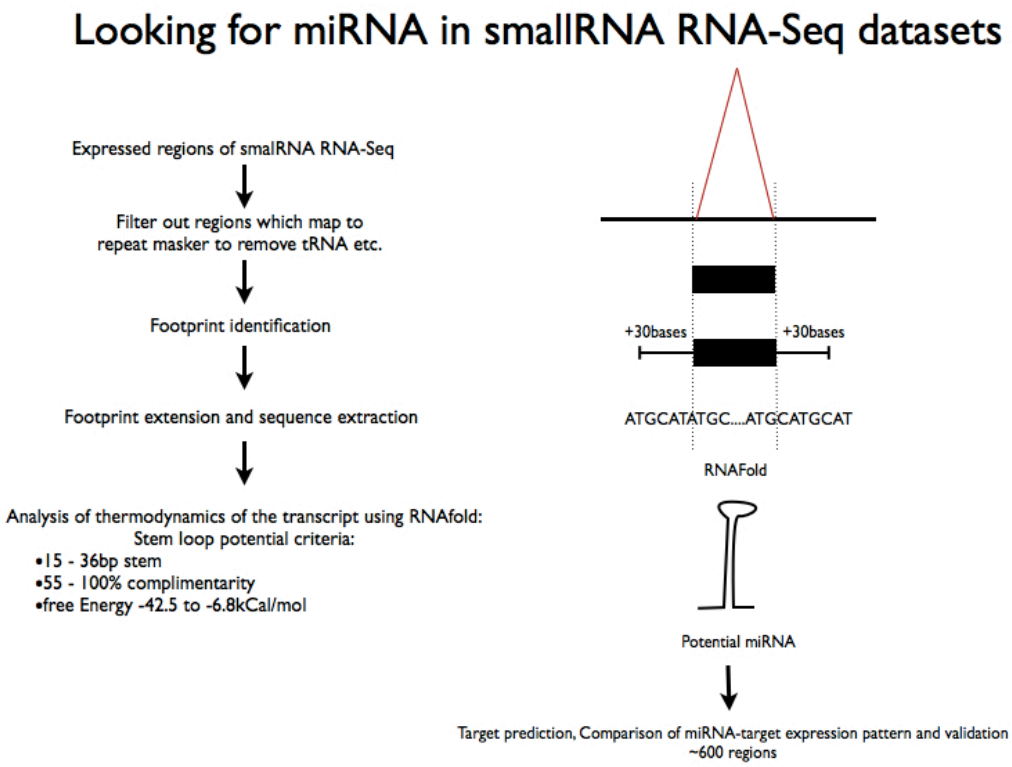


Figure 3.6: Workflow for the identification of novel microRNA.
Novel microRNA were identified by looking at the thermodynamic stability of their stem-loop structure.

4.0 Results 2

4.1 Trophoblast differentiation

The differentiation protocol using SU5402 and BMP4, which pushes a human embryonic stem cell to the trophoblast lineage, gives access to a unique and scarcely studied cell type. The unbiased nature of RNA-Sequencing is ideally suited to study the transcriptome of such a product. This is because RNA-Seq provides much more information than the expression levels provided by traditional technologies such as microarrays and quantitative PCR.

4.2 hESC derived trophoblast gene expression strongly correlates with that of placental derived tissue

Though we were confident that SU5402+BMP4 treatment directs human embryonic stem cells to the trophoblast lineage (as described in Dr. Wenlong Luo's thesis (Wenlong 2008)), a recent report suggested that hESC-based protocols did not form true trophoblast (Hemberger, Udayashankar et al. 2010), particularly because there was a lack of expression of *ELF5*, a key gene in the trophoblast that is repressed in embryonic stem cells through promoter DNA methylation. Thus I aimed to comprehensively compare the outcome of our novel differentiation protocol (which was not taken into account in Hemberger et. al.) to that of its natural counterpart.

Trophoblasts are the major zygotically-derived cell type which contributes to the placenta from the fetus. In order to find out the cell lines / tissues(s) that show a similar expression pattern to that brought about by the differentiation, unsupervised hierarchal clustering was used to compare a dataset containing microarray expression

levels of published tissues / cell types and the SU5402 + BMP4 differentiation microarray data. Importantly, besides a broad array of human tissues including the placenta (Ge, Yamamoto et al. 2005) this comparison also include sorted extravillous trophoblast and cytotrophoblast cells from first trimester human placentae (Bilban, Haslinger et al. 2009), the earliest possible placental cells from post-implantation human development. Hierarchal clustering is designed to cluster together similar datasets. Therefore, tissues / samples / cell lines which are clustered together can be considered as having closely matching global expression profiles.

Reassuringly, the closest tissue type to day 6 and day 8 of treatment was the placenta, and the closest cell types were the extravillous trophoblast and cytotrophoblast cells. In addition, our microarray data indicated that ELF5 was indeed expressed during the differentiation. These evidence support that the products of our trophoblast differentiation protocol is being representative of the true trophoblast.

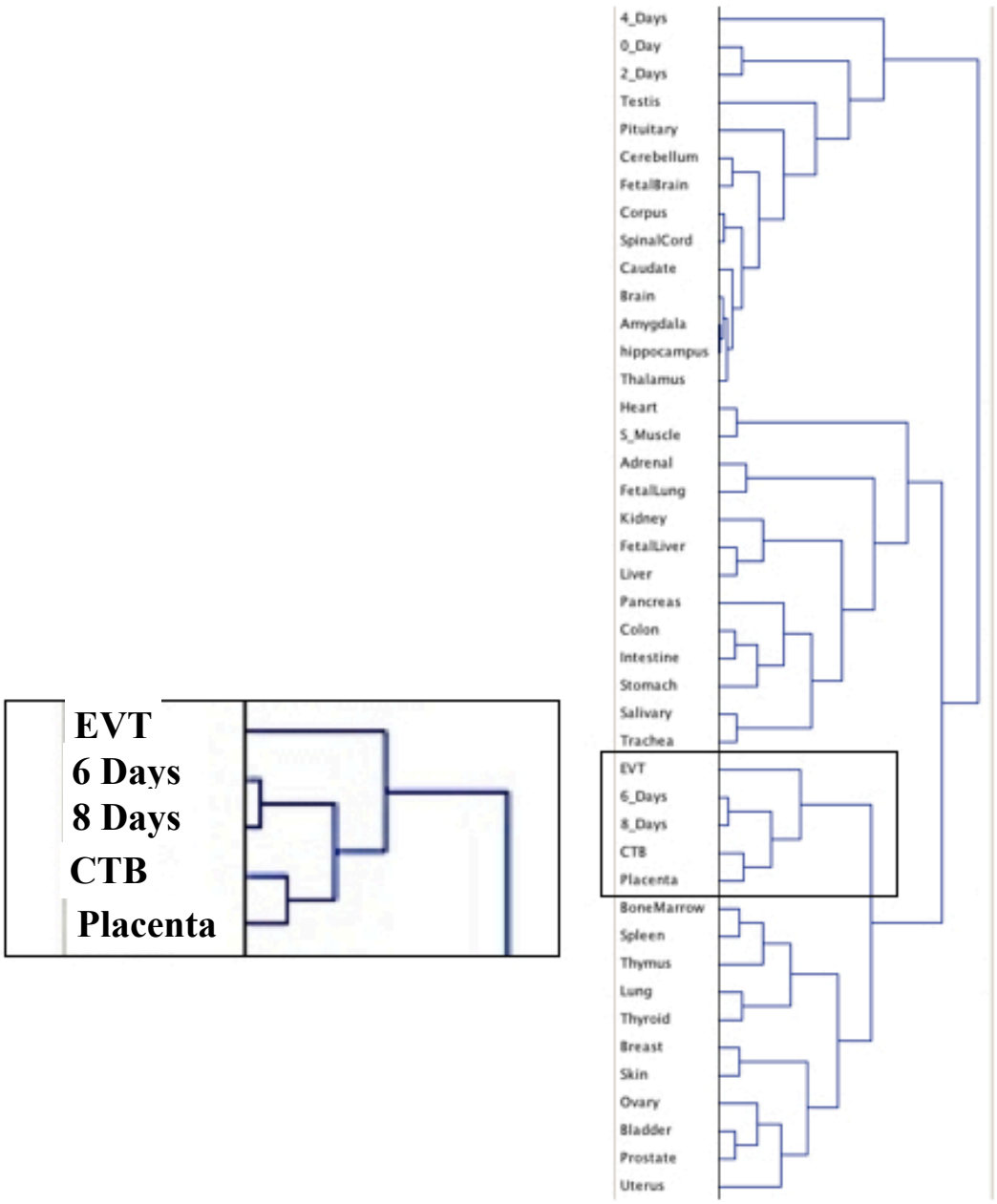


Figure 4.1: Trophoblast differentiation products cluster closely with placenta and related cell types.

This shows the hierarchical clustering result of the five time-points of the differentiation protocol with a list of tissue / organ expression profiles. Day 6 and Day 8 time-points cluster with Placenta, EVT - Extravillous trophoblast and CTB - Cytotrophoblast samples.

4.3 Poly A extraction of total RNA effectively removes ribosomal RNA to increase the dynamic range of the transcriptomic data

RNA-Seq involves extraction of RNA from a particular sample and converting it to a library which can be sequenced so that the transcripts can be later reconstructed, and their expression levels measured. Ribosomal RNA (rRNA) is the single most abundant species (>90%) in any total RNA sample extracted using conventional methods. If a sample contains a significant amount of rRNA, it reduces the sequencing depth of messenger RNA (mRNA). This hinders the study of the dynamics of the transcriptome as rRNA levels are mostly static. Therefore, for the trophoblast differentiation experiments, an mRNA extraction step (based on the polyA tail) was performed to remove rRNA from the sample. The Agilent bioanalyzer, which measures the length distribution with concentration of RNA, was used to asses the reduction of rRNA for each sample.

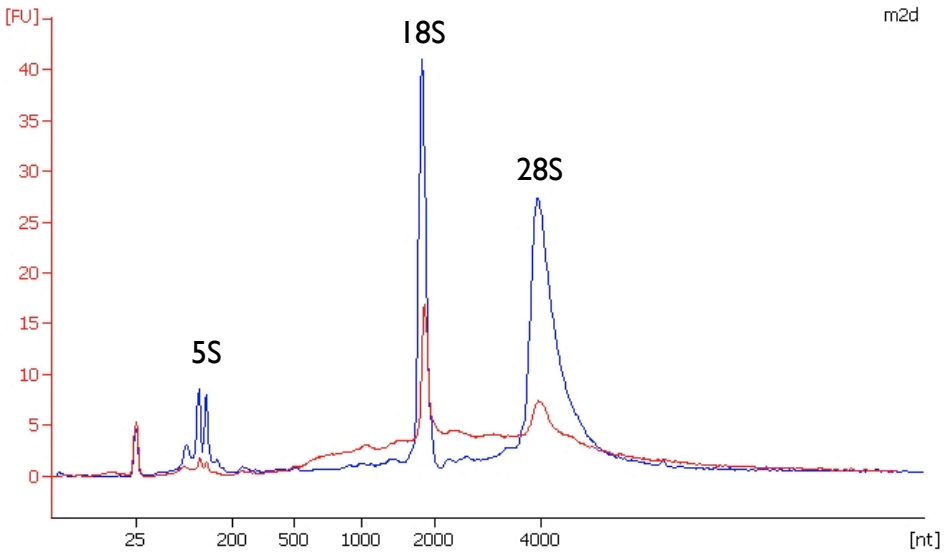


Figure 4.2: Removal of rRNAs from polyA RNA.
Overlapped Agilent Bioanalyzer trace of total RNA (blue) and one time - poly A extracted mRNA (red). The reduction of rRNA species (Reduction in height of rRNA peaks) is evident.

Above observation was confirmed after sequencing, where the reads which mapped to known rRNA sequences (and some other filter sequences) were less than 10% of the total sequenced reads (Table 1).

Sample Name	Percentage of reads mapping to rRNA
SB - Day 0	9%
SB - Day 2	6.5%
SB - Day 4	6.8%
SB - Day 6	7.9%
SB - Day 8	6.2%
Mm - E 3.5 BL	4.2%
Mm - E 4.5 BL	4.2%
Mm - E 4.5 ICM	2.7%
Mm - 8 cell	2.1%

Table 1: Percentage of reads which map to rRNA.
If a sequencing run has a high percentage of reads mapping to rRNA regions of the genome, it reduces the number of reads representing mRNAs. All the RNA-Seq samples presented in this thesis show a low percentage of reads mapping to rRNA. SB - SU5402 +BMP4 treatment, Mm - mouse embryo samples, showing that the rRNA removal was successful.

4.4 Expression levels obtained by RNA-Seq for known genes show a good correlation with microarray data

Parallel to the RNA-Seq experiment, a microarray run was performed using the same samples (These samples were also used for the hierarchical clustering mentioned above). Microarray run was carried out by Ms. Woon Chow Thai from our group. The intention of running the microarrays was to observe how expression levels of all genes compared between RNA-Seq and the more conventional microarray technologies. The comparison was done on differential expression values of RNA-Seq and microrray data. To do the comparison *en masse* the Signal Log Ratio values

(SLR = $\log_2$ of fold change) of the two datasets were plotted with each other. To obtain reliable fold change / signal log ratio values from the microarray data, only genes which had a raw signal value of more than 10 in both day 0 and day 8 time-points were used for the comparison. No filtering was done based on RNA-Seq RPKM values. The value of coefficient of determination (R^2) between the two datasets was 0.8055. Expected R^2 value for the two identical datasets is 1 (on a range of 0 - 1), thus the two expression datasets with a R^2 value of 0.8 can be considered to be significantly comparable with each other. This is further confirmed by the fact that 94% of the genes had a difference of less than 1 SLR (signal log ratio) value between the two datasets and only 5 of the genes showed contradictory expression pattern (i.e up-regulated according to one dataset and down-regulated in the other) (Figure 4.3).

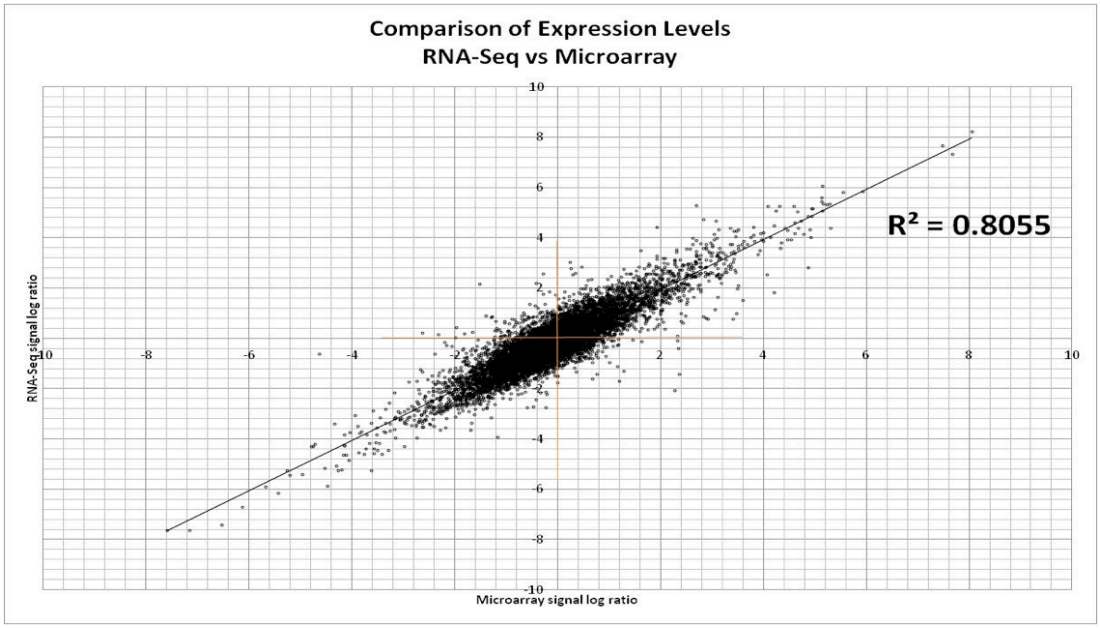


Figure 4.3 : Comparison of signal log ratio values of microarray and RNA-Seq dataset.
The fold change values of day 0 vs day 8 were converted to Signal Low Ratios (SLRs) by converting them to $\log_2$ form. The RNA-Seq and microarray expression values showed a good correlation with each other.

There were only 5 genes which showed opposite expression patterns (i.e. up-regulation in one dataset and a down-regulation on the other). Upon closer examination these were found to be due to issues with the placement of the Illumina microarray probe. For example in the case of *BAX*, microarray data indicated the gene is up-regulated during differentiation while RNA-Seq data showed otherwise. Looking at the RNA-Seq peak profile of *BAX*, clearly a short and a long isoform of the gene is being expressed, and the longer one is down-regulated while the shorter one is up-regulated (Figure 4.4). However the Illumina microarray probe in this case, only picked up the shorter isoform, marking the gene as being up-regulated. This clearly shows the advantage of looking at the expression of the entire transcript(s) of any gene (as in the case of RNA-seq), instead of merely considering a small portion to represent the expression of the entire gene as done by microarray technology.

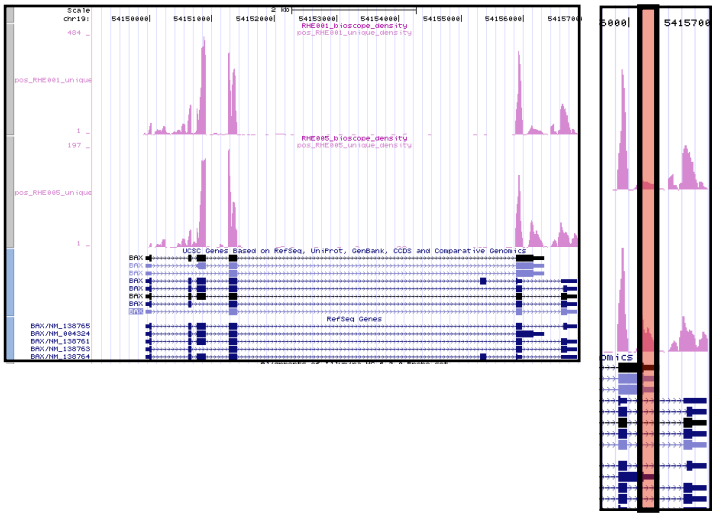


Figure 4.4: Differences in methods used in RNA-Seq and microarrays for measuring gene expression.
Figure shows the BAX RNA-Seq UCSC view. There is a conflict between RNA-Seq and microarray expression patterns for this gene. The 3' portion of the gene has been enlarged for clarity on the panel on the right, and the red bar marks the position where the Illumina microarray probe binds. RNA-Seq data shows that in the case of BAX, the longer isoform is being up-regulated while the shorter isoform is down-regulated. Since the microarray probe only detects the shorter isoform, it marks the gene as being down-regulated.

4.5 The trophoblast differentiation protocol brings about drastic changes in the hES cell transcriptome as identified by RNA-Seq

One of the advantages of RNA-seq is its ability to provide a digital count of expression levels through the RPKM value. As described in the methods section, RPKM value represents the number of reads mapping to a particular gene normalized to it is length and the total sequencing depth. A gene was defined as significantly expressed when it had an RPKM value higher than a set number. Based on the numbers presented in Table 2, it is evident that there is an increase in expressed genes upon treatment, which is maintained up to day 6. In other words the treatment seems to be inducing a number of genes which are not expressed in human embryonic stem cells under normal conditions. The decrease of expressed transcripts between day 6 and 8, could be due to the clearance of pluripotent genes. Indeed, this differentiation protocol had been characterized to co-express both pluripotent and trophoblast genes over the first couple of days of differentiation and become committed to the trophoblast lineage only after approximately 48 hours of treatment (Wenlong 2008)

Treatment Duration	Total number of genes with RPKM > 0	Total number of genes with RPKM > 2	Total number of genes with RPKM > 5	Total number of genes with RPKM > 10
Day 0	17005	11094	8688	6452
Day 2	16965	11138	8896	6583
Day 4	17127	11306	9084	6728
Day 6	17132	11359	9155	6729
Day 8	17238	11279	8967	6575

Table 2: RefSeq genes expressed during the trophoblast differentiation time-course from a total of 21296 RefSeq genes.
Day 0 represents undifferentiated H1 human ES cells. The number of total expressed genes increases during differentiation and peaks at day 6.

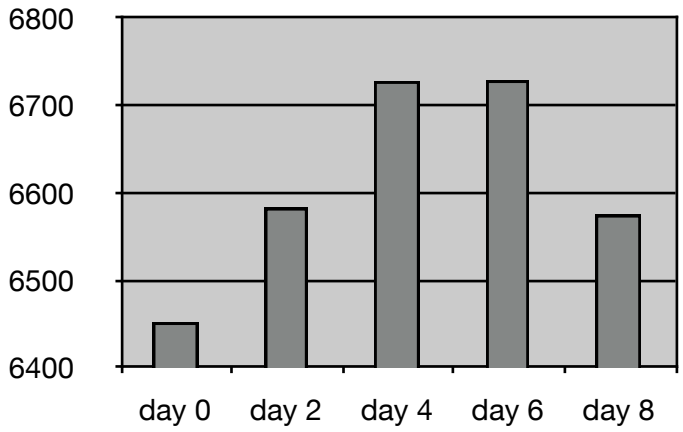


Figure 4.5: Genes expressed at very high level (RPKM > 10) at each time point.
There is a clear increase in expressed genes in all time points compared to day 0. Please note that the baseline of the graph has been raised to 6400 to highlight the difference in Y axis.

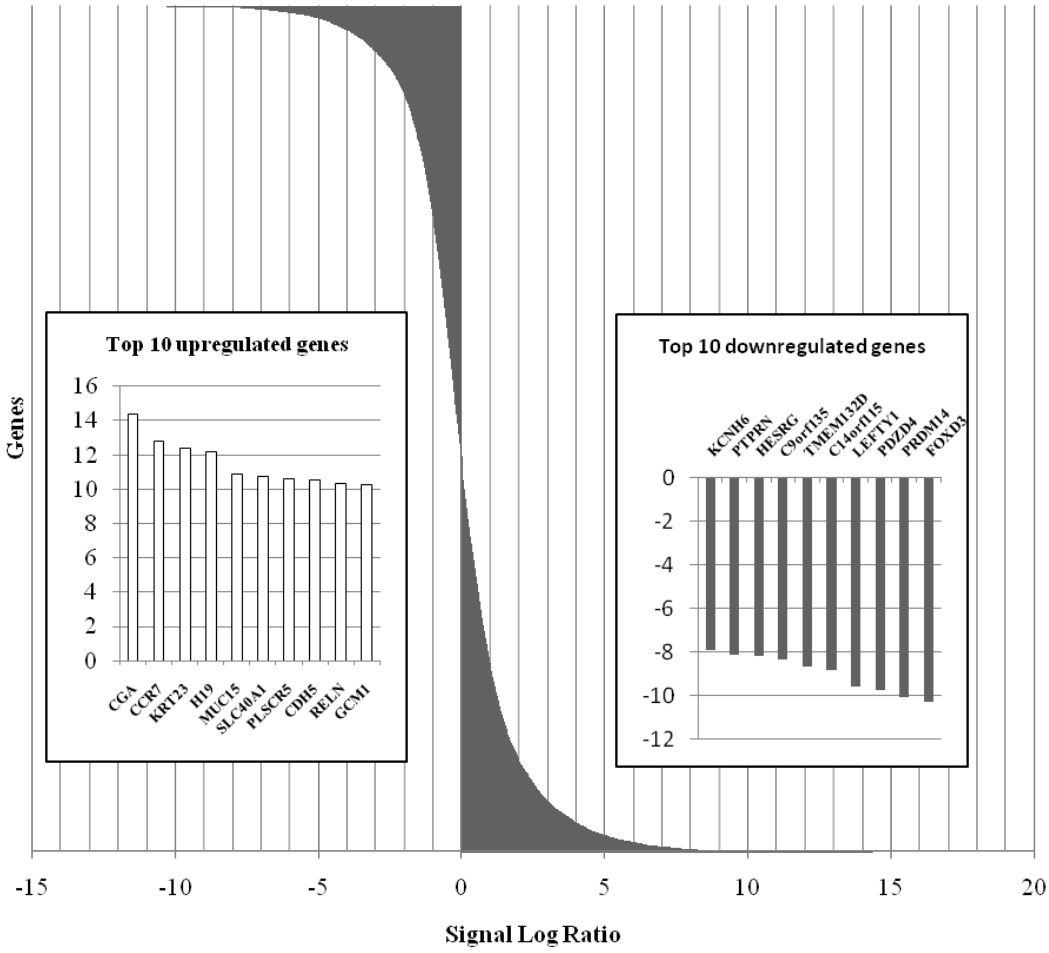


Figure 4.6: Differentially expressed genes during trophoblast differentiation.
This shows the extent of differential expression between day 0 and day 8, during the transition from human ES cells to the trophoblast lineage. Each horizontal bar in the main graph depicts the signal log ratio (SLR) of each of the total 21,296 genes. The top 10 up and down-regulated genes have been magnified and are shown in the two sub charts. The extent of up and down-regulation suggest that the differentiation protocol leads to a profound change in the transcriptome. SLR is defined as the $\log_2$ of the fold change.

4.6 A number of genes which are not expressed in undifferentiated human ES cells gets induced during trophoblast differentiation

Apart from differentially regulated genes, induced genes during the trophoblast differentiation is of particular interest. Here an induced gene is defined as a gene which is not expressed in human embryonic stem cells, but is significantly expressed during the differentiation treatment. There is an increase in the number of induced genes during treatment, and the induction of new genes seems to be correlated with the length of treatment. Even considering a cut off of RPKM > 2, there are 51 genes (shown in table 3) which are induced in day 8.

Treatment	Total induced number of genes (RPKM > 0)	Total induced number of genes (RPKM > 1)	Total induced number of genes (RPKM > 2)
Day 2	679	11	4
Day 4	845	32	16
Day 6	912	59	28
Day 8	1018	78	51

Table 3: Trophoblast differentiation induces a number of genes during the time-course. The number of genes which are induced increase with time.

Gene Symbol
CYP19A1
HOPX
S100P
MRGPRX1
HERV-FRD
DCN
APOA4
FGB
HMGCS2
SMPX
SLN
LOC100129935
C1orf105
LGALS13
SERPINB12
HOXA1
LGR5
BLNK
CCR1
SNORD115-33
SLC22A11
DEFB1
BCAR4
CYP3A7
TFAP2B
PPY
TCL6
ODAM
FBLN7
P11
SNORD115-30
FYB
PLA2G2F
CASP4
ELF5
C6orf155
TREM1
NPR3
CGB8
TLR7
SNORD116-28
SNORD19B
FGA
SNORD115-4
PGC
CGB5
CST4
DAPP1
LY6H
LY6G6C
HOXB3

Table 4: The 51 genes induced during 8 days of treatment which show expression level of more than 2 RPKM units.

In order to identify the functions of the 51 induced genes as a group, a gene ontology study was carried out using panther gene ontology database (www.pantherdb.org).

Panther gene ontology analysis of the 51 induced genes with 8 days of treatment.

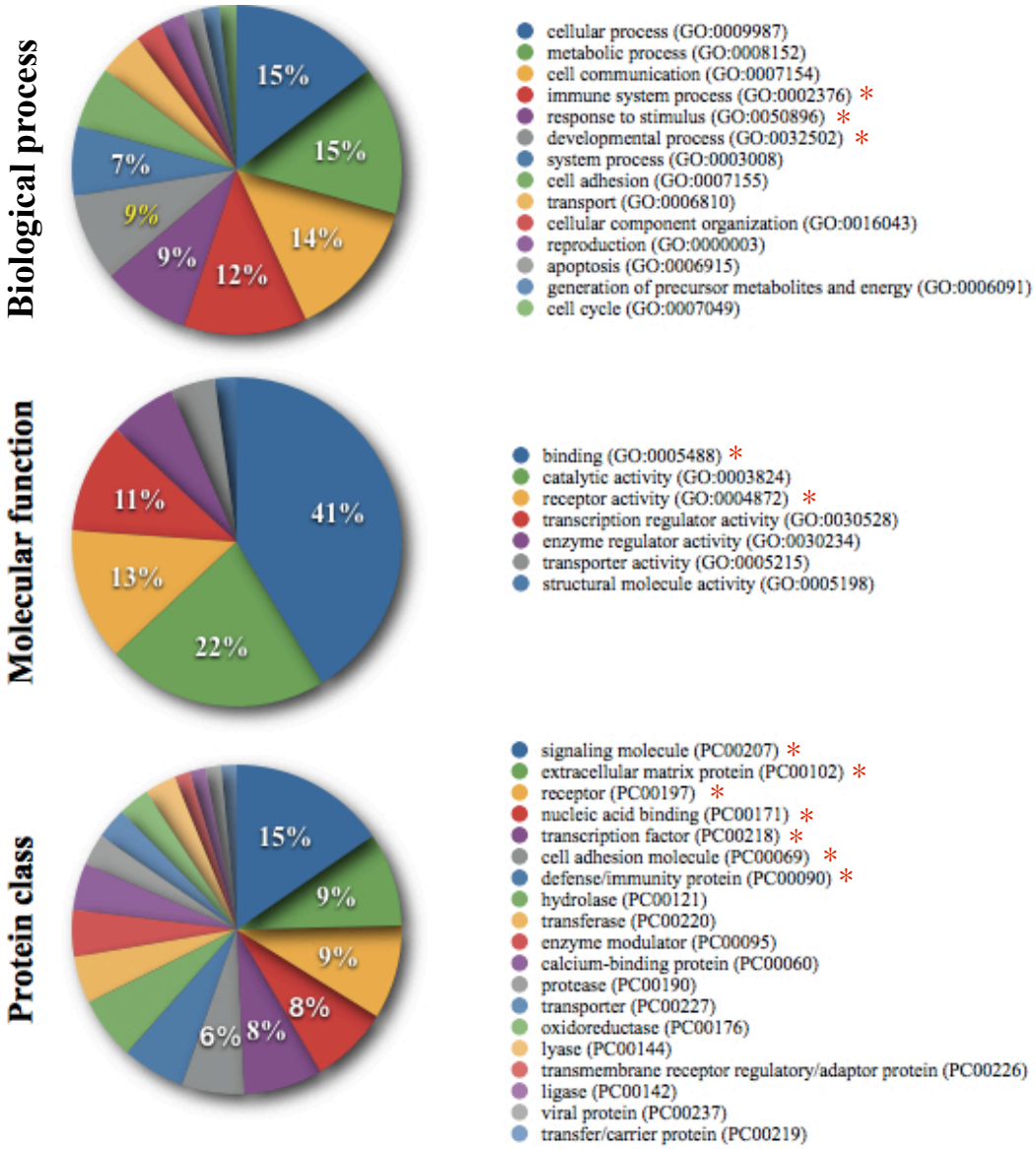


Figure 4.7: Gene ontology results of the 51 highly induced genes during trophoblast differentiation. The ontology terms marked with a red * represents terms significantly enriched by DAVID (http://david.abcc.ncifcrf.gov/) functional gene analysis.

Based on panther gene-ontology analysis, the 51 induced genes at 8 days of treatment seems to be biologically active proteins. The majority of them seems to be involved in regulatory roles which suggests that these might be important effectors of the differentiation process.

4.7 Study of fold change distribution of genes during trophoblast differentiation

Looking at the genes which are up and down regulated during the differentiation protocol, clearly the number of differentially expressed genes gets increased with time.

comparison	Up-regulated (SLR ≥ 2)	Down-regulated (SLR ≤ 2)
day 4 vs day 0	1440	1354
day 6 vs day 0	1642	1498
day 8 vs day 0	1789	1684

Table 5: The total number of up-regulated and down-regulated genes increase with treatment duration.

To broadly evaluate the function of up-regulated genes during the differentiation, a gene ontology analysis similar to the one above was carried out.

4.8 Gene ontology analysis of up regulated genes during trophoblast differentiation

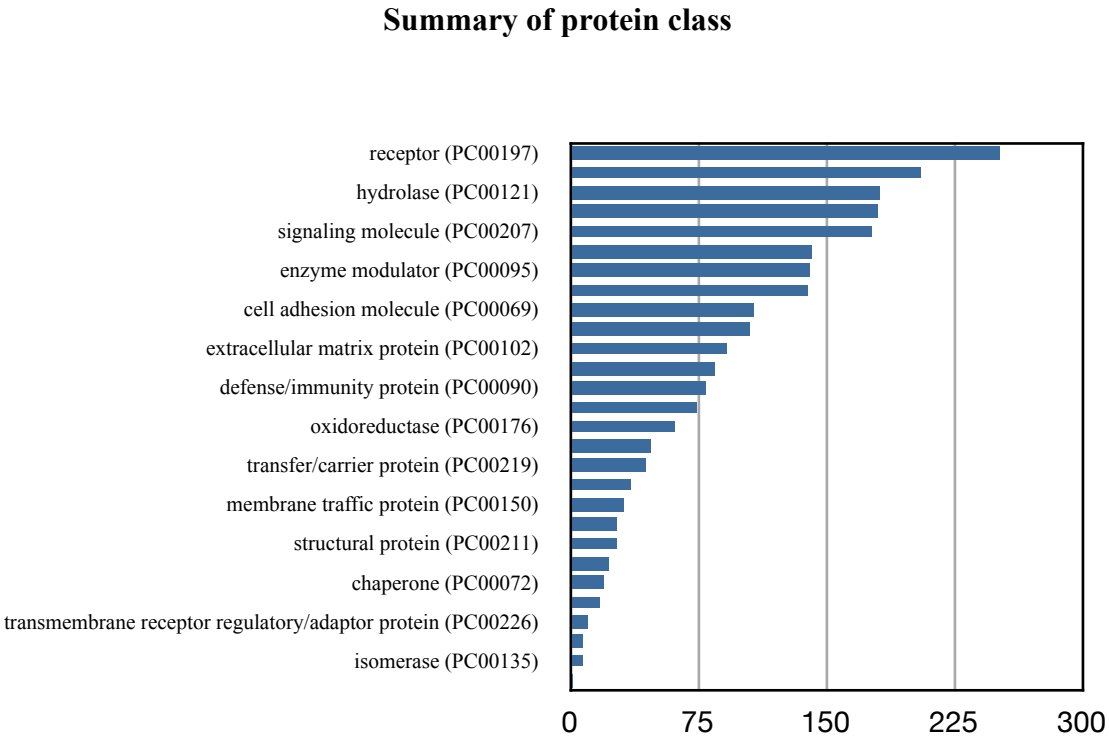


Figure 4.8: Panther protein classes of the up-regulated genes.
Y axis shows the significantly expressed protein class and the x axis shows the number of genes belonging to each class.

Just as induced genes, the up-regulated genes during 8 days of treatment are enriched with transcription factors and signaling proteins. Apart from these two main groups, most other groups of proteins are also represented in the up-regulated genes. These results imply that while transcription factors and signaling molecules are the major effectors of the observed differentiation outcomes, the treatment brings about a profound change which affects almost all the processes of the cell.

The pathways affected by genes which are up-regulated during trophoblast differentiation.

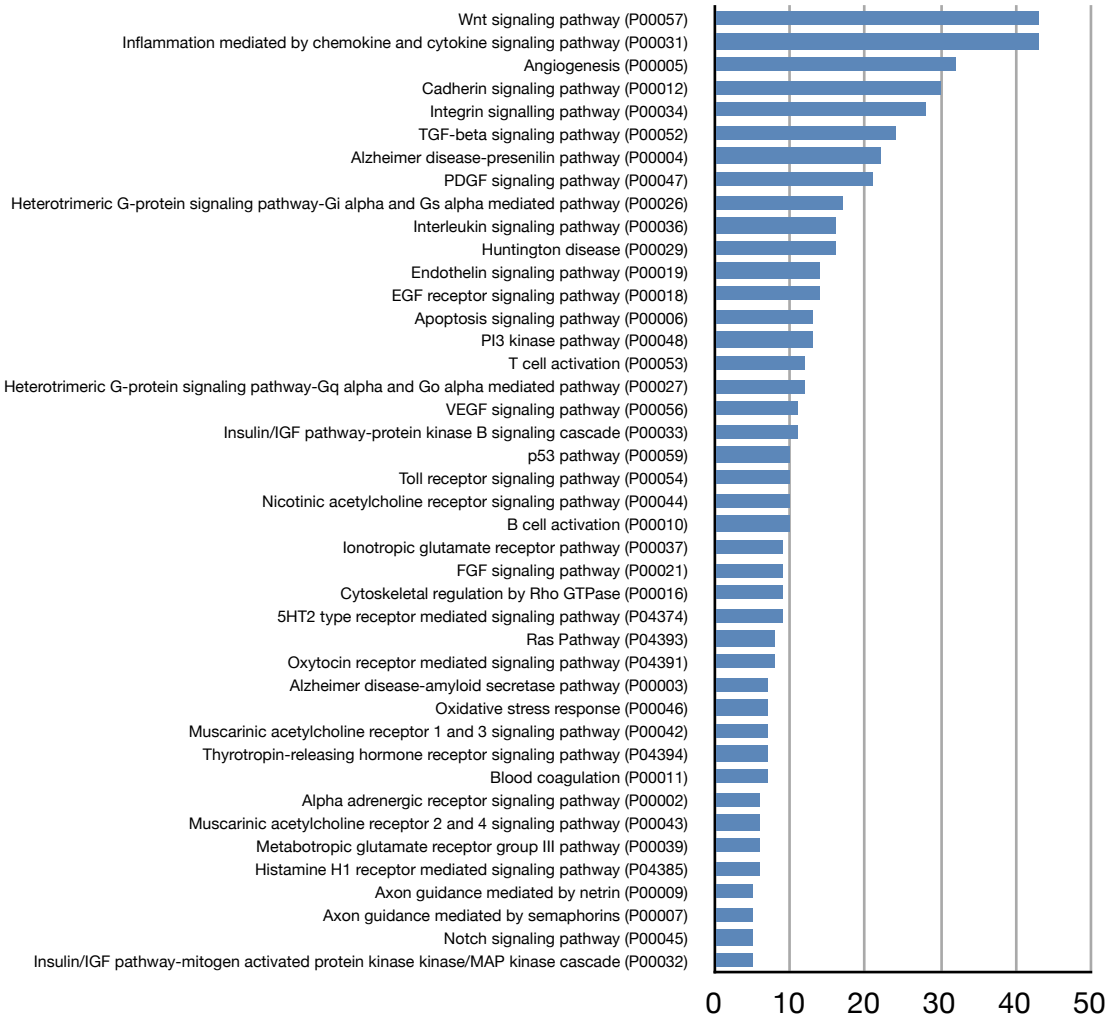


Figure 4.9: Significantly affected pathways from up-regulated genes during trophoblast differentiation.

Pathways with the highest number of up-regulated genes, namely Wnt signaling, Cadhering signaling and TGF-beta signaling are all reported to be involved in mesenchymal to epithelial transition. In addition, Wnt signaling has been implicated in embryo implantation (Mohamed, Jonnaert et al. 2005) .

Top 50 Up-regulated	Expressed in placenta	Top 50 Down-regulated
CGA	Yes	FOXD3
CCR7		PRDM14
KRT23	Yes	PDZD4
H19	Yes	LEFTY1
MUC15		C14orf115
SLC40A1	Yes	TMEM132D
PLSCR5		C9orf135
CDH5	Yes	HESRG
RELN		PTPRN
GCM1	Yes	KCNH6
CSF3R	Yes	ATCAY
ALPK2		MAGEA4
HSD3B1	Yes	ZIC2
PLCXD3		LCK
NTRK2		SNCB
ERP27		CRLF1
HOXB2		PPP1R16B
LUM	Yes	TMEM151B
VGLL1	Yes	OPCML
TBX3	Yes	POU5F1
P2RY6	Yes	POU3F1
DIO2		PTPRZ1
APOA2		ADAMTS8
GDF6		BCAN
NR2F2	Yes	NRIP3
CYSLTR2		C1QL2
GUCY1A3		RTN4RL2
CYP2C18		PRKCB
KIAA0774		FGF19
FLJ45983		SLC7A3
EPAS1	Yes	CTCFL
MBNL3	Yes	KCNQ2
LRP2		SLITRK3
FLRT3		FAM124A
ZNF750	Yes	CRIP3
IL1F5		NPTX2
ST8SIA4		RASGRP4
GATA6		IGSF21
STS	Yes	CPEB1
HTRA4		NMU
TP63		AIF1
MEIS1		OLFM1
NUPR1		NTRK3
HAL		RHBDL3
SLC6A4		PADI3
ZNF503		DOCK2
CLEC1A	Yes	ZIC5
ADCY10		TMEM145
SYTL5		CXCL5
C8orf4	Yes	HRH3

Table 6: Top 50 up and down regulated genes during trophoblast differentiation.

4.9 Comparison of RNA-Seq gene expression levels with published human preimplantation data shows a considerable overlap

There is a paucity of gene expression data available for human pre-implantation development, a result of the scarcity of such tissues combined with the very limiting amount of RNA available in each individual sample. Thus when Zhang et al (Zhang, Zucchelli et al. 2009) reported a microarray study of human pre-implantation development - the first available data set providing global gene expression of significant quality from the human blastocyst - I eagerly compared it with our hESC-derived data. The Zhang dataset enabled the identification of differentially expressed genes during the transition from the 4 - cell stage to the blastocyst. Using the data presented in the paper, it is possible to identify genes up and down-regulated during blastocyst formation, but it is not possible to separate out trophectoderm / ICM specific / enriched genes as the paper does not report the gene expression of ICM or the trophectoderm separately. Nonetheless, the majority of cells in the blastocyst (>70%) would be trophectoderm and the 4-cell embryo would represent a stage at which the trophectoderm has yet to form. Thus a comparison of 4-cell stage to blastocyst should identify genes up-regulated in the human trophoblast lineage. Though many pluripotent genes are thought to be expressed in the 4-cell stage embryo there are likely other genes specifically expressed in the ICM of the human blastocysts that are up-regulated from the 4-cell to the blastocyst stage. My RNA-Seq data of the SU5402+BMP4 hESC differentaitiotn protocol provides a set of genes up / down-regulated during human trophoblast differentiation. Therefore by overlapping the human embryo dataset with the SU5402+BMP4 differentiation, the genes shown to be responsible for trophoblast development by the differentiation protocol can be validated.

However, it should be noted that while the dataset presented by Zhang et al. is impressive, due to the nature of the sample they only use duplicates for microarray runs and that the variability within the samples is high. As a result 1,501 genes identified as significantly up-regulated 2-fold or more is likely an underestimate of the genes truly differentially regulated. When these 1,501 genes are overlapped with the hESC-derived trophoblast differentiation data 542 genes are significantly up-regulated in both datasets. On average the 542 commonly up-regulated genes show an average fold change of 8 fold (maximum = 16,348 fold and minimum 2 fold). In addition, a hierarchical clustering done between the RNA-Seq and preimplantation microarray data showed that the 4 cell microarray sample clustered with the hESc sample of RNA-Seq and the blastocyst sample of microarray clustered with day 8 time point of RNA-Seq. This clearly shows that there is a good correlation between the SU5402+BMP4 differentiation RNA-Seq data and human preimplantation data (Figure 4.10), and more importantly this correlation suggests that the identified 542 genes identified may play a vital role in human trophoblast formation and development highlighting the utility of my RNA-seq data in identifying the transcriptome of early human trophoblast development. These 542 genes were analyzed further.

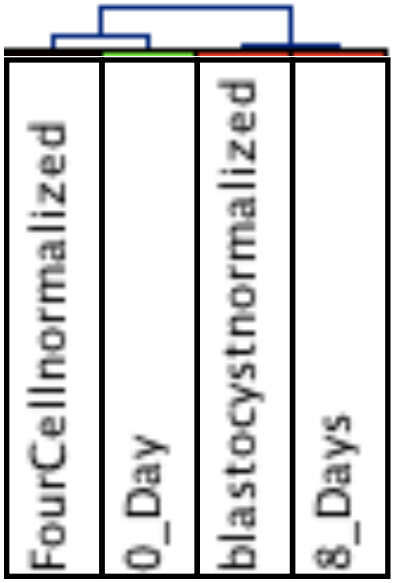


Figure 4.10: Hierarchical clustering of RNA-Seq data with published human preimplantation data.
Human blastocyst sample gets clustered with 8 days differentiation time-point while four cell human embryo sample gets clustered with human ES sample.

geneID	Day 0 RPKM	Day 2 RPKM	Day 4 RPKM	Day 6 RPKM	Day 8 RPKM	SLR	Fold change	Fold change blastocyst vs EightCell
CGA	0.11	0.26	8.09	544.43	2239.42	14.37	21173.91	126.19
S100P	0	0.28	0.51	31.14	83.46	-	-	109.48
GCM1	0.03	0.07	1.23	16.24	33.01	10.26	1226.22	88.05
MUC15	0.03	7.43	25.34	39.22	62.84	10.87	1871.53	80.58
ABCG2	1.55	13.3	94.63	118.81	95.97	5.95	61.82	76.64
ANXA1	11.11	24.35	113.18	481.26	751.31	6.08	67.65	74.84
H19	0.18	15.78	81.45	598.54	808.91	12.16	4576.41	58.78
LRP2	0.15	2.92	23.12	39.26	43.24	8.13	280.14	54.18
KRT19	100	368.9	582.93	630.19	606.56	2.6	6.06	51.01
RCN1	19.12	18.4	26.8	38.87	43.59	1.19	2.28	49.97
CCR7	0.02	0.95	3.35	42.51	122.07	12.8	7131.55	46.9
ZNF750	0.06	27.41	31.55	12.68	15.58	8.07	268.73	41.31
CBLB	10.46	15.66	20.21	34.68	51.22	2.29	4.89	41.1
LYN	12.37	12.57	25.91	54.9	91.76	2.89	7.41	39.84
ENPEP	0.34	18.29	34.39	57.63	54	7.33	160.90	37.57
SGMS1	7.74	12.16	12.98	16.42	16.75	1.11	2.16	36.46
GADD45G	11.56	14	6.93	33.33	82.97	2.84	7.16	35.36
SERPINB9	31.42	233.14	82.11	45.03	73.85	1.23	2.35	34.81
SLC38A1	29.15	47.37	101.27	134.29	128.59	2.14	4.41	33.98
KRT23	0.03	0.49	19.3	96.09	182.97	12.36	5256.91	33.55
SMAD7	5.32	17.88	21.43	21.09	30.92	2.54	5.82	33.21
CCKBR	0.54	1.11	5.27	27.9	32.85	5.92	60.55	32.39
HERV-FRD	0	0	0.14	6.67	32.13	-	-	31.54
RAB31	10.26	30.39	26.51	37.7	34.48	1.75	3.36	29.34
RHOU	1.87	20.23	33.93	68.54	71.29	5.25	38.05	29
KANK4	0.18	0.18	3.12	23.74	29.38	7.33	160.90	27.86
GATA2	0.52	31.45	31.26	45.02	77.57	7.23	150.12	27.52
SLC1A3	2.41	4.07	9.87	38.02	54.55	4.5	22.63	27.19
CEBPA	2.25	11.21	33.38	39.25	47.26	4.39	20.97	26.59
REEP1	0.31	0.45	0.94	5.88	6.61	4.42	21.41	26.53
SLC40A1	0.09	0.65	4.16	82.5	152.16	10.74	1710.26	26.03
ANXA3	14.61	68.63	150.09	322.63	290.26	4.31	19.84	25.74
SDC1	21.99	5.05	5.24	29.7	54.74	1.32	2.50	25.29
KRT18	69.59	230.83	322.61	458.84	410.36	2.56	5.90	23.94
COL21A1	1.19	0.6	1.43	10.82	9.26	2.96	7.78	23.36
TACSTD2	0.9	45.09	48.68	32.7	30.38	5.08	33.82	23.31
AMOTL2	14.86	40.87	53.73	77.06	90.51	2.61	6.11	22.28
GPRC5A	0.57	7.43	10	19.25	31.05	5.76	54.19	21.35
S100A14	1.32	25.65	121.89	72.48	57.72	5.45	43.71	21.16

geneID	Day 0 RPKM	Day 2 RPKM	Day 4 RPKM	Day 6 RPKM	Day 8 RPKM	SLR	Fold change	Fold change blastocyst vs EightCell
TNS3	10.54	55.64	28.83	19.3	26.34	1.32	2.50	21.1
IL1R1	0.12	1.35	2.71	8.16	10.7	6.46	88.03	20.87
TGFBR2	2.02	0.9	2.43	10.67	27.39	3.76	13.55	20.73
FHDC1	6.81	18.21	17.82	24.05	39.22	2.53	5.78	20.32
HOPX	0	0.44	0.16	42.53	98.19	-	-	20.14
SLC7A11	1.37	1.98	4.29	3.07	2.88	1.07	2.10	19.35
CDKN1C	15.65	41.64	46.75	162.39	252.66	4.01	16.11	19.3
C10orf10	3.21	9.83	12.33	28.87	56.97	4.15	17.75	19.28
TPD52L1	5.86	9.31	21.58	35.53	36.09	2.62	6.15	19.13
HSPB8	3.49	17.14	43.68	150.95	388.24	6.8	111.43	19.03

Table 7: The top 50 up-regulated genes (based on human embryo 4-cell to blastocyst fold enrichment) which are also up-regulated in the hESC-based trophoblast differentiation protocol.

4.10 Genes induced / up-regulated during trophoblast differentiation

Table 7 highlights genes that are best induced / up-regulated in human trophectoderm formation (based on preimplantation data) that are also up-regulated in the hESC-based trophoblast system. The RNA-Seq profile of a few of these genes will be described hereafter.

4.10.1 *CGA* (Chorionic gonadotrophin alpha)

CGA codes for the alpha subunit of the human chorionic gonadotropin, the signature hormone of the trophoblast. *CGA* expression is initiated at day 4 and is greatly up-regulated as differentiation progresses.

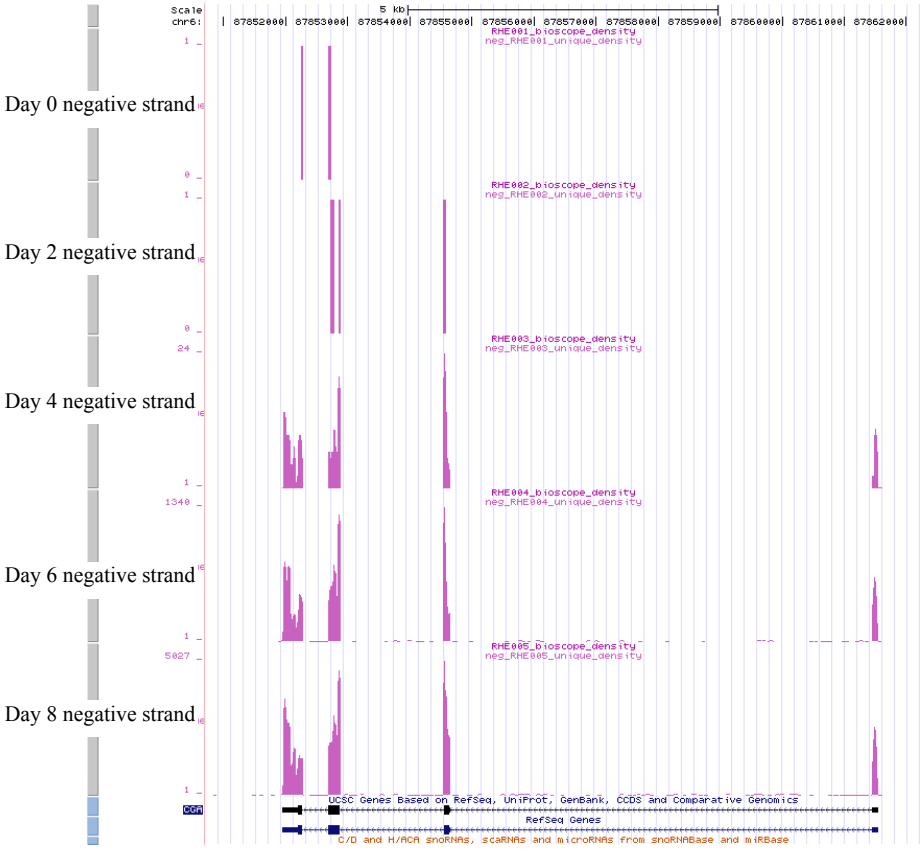


Figure 4.11: RNA-Seq peak profile of *CGA* on the UCSC browser.
The *CGA* expression gets up-regulated from a very low value of 0.11 RPKM at day 0 to 2239.4 at day 8, showing a fold change of 21173.9. Only the negative strand for each of the time point is shown here for clarity.

4.10.2 *CGB* (Chorionic gonadotropin beta)

CGB, which together with CGA produces hCG hormone, has six genes in the human genome with virtually the same sequence. These are a result of extensive gene duplication of the original *LHB* gene with some of these gene duplication occurring since the divergence from the chimpanzee (Hallast, Saarela et al. 2008). The extent of sequence similarity can be seen in the multiple alignment diagram in Figure 4.12. This presents a major issue when it comes to aligning the reads from these regions into the genome. During the counting phase of RNA-Seq, reads which map to two places or more with the same score are discarded and therefore a considerable number of reads originating from *CGB* genes are not taken into account during calculation of expression values. This leads to an underestimation of the level of expression from this locus. This is in contrast to the microarray data which indicates rather robust up-regulation (*CGB5* max. probe intensity of 267, *CGB8* max. probe intensity of 281) as the hybridization signal is not lost. This multi-mapping issue is a weakness of RNA-seq but the sequence data can also be used to an advantage here. By focusing on specific bases that vary between the genes, and identifying whether any unique reads are mapped in these regions, it is possible to identify which of the 6 *CGB* genes are being expressed, something that would be challenging with an array based method of expression detection. From this analysis, RNA-seq clearly identifies expression from *CGB8*, *CGB5* and *CGB7* as showing significant up-regulation during the differentiation. Such information is valuable particularly if one is interested in the transcriptional control of specific *CGB* genes.

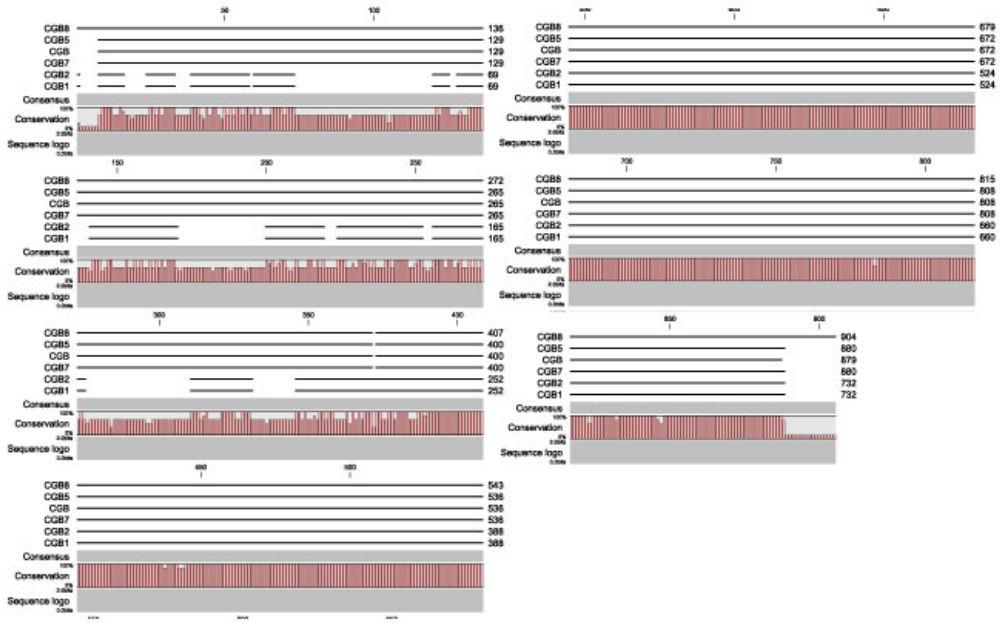


Figure 4.12: Multiple alignment of *CGB8, CGB5, CGB, CGB7, CGB2* and *CGB1*.
The analysis was performed using CLC Main workbench. The lines next to the gene names show consensus sequences and the bar graph shows the conservation %. Note that in most cases the conservation is 100% and that it rarely goes below 50% suggesting extreme sequence similarity.

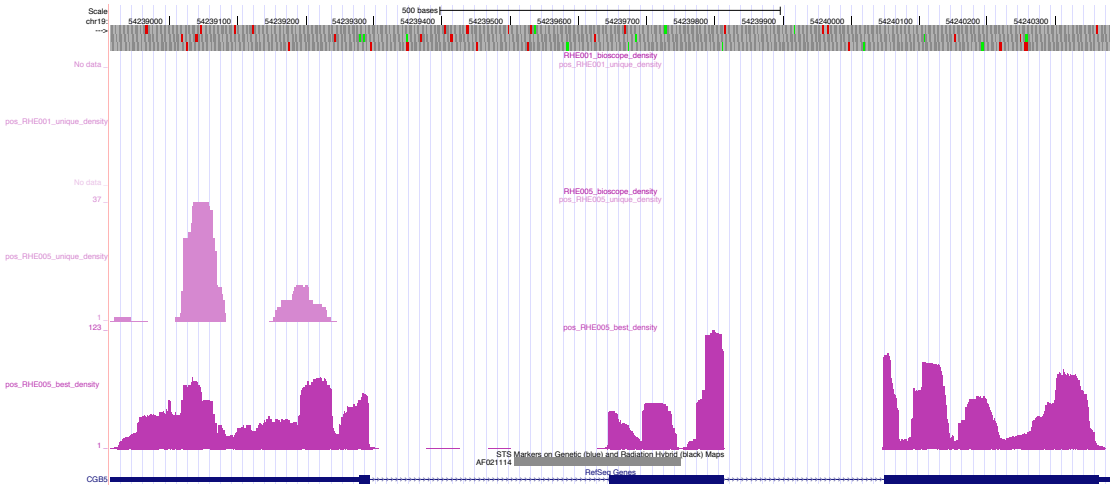


Figure 4.13: The UCSC view for *CGB5*.
Tracks from the top are, (1) Day 0 unique reads - which there are none, (2) Day 8 unique reads - to a maximum height of 37 suggesting that there is robust expression, and (3) Day 8 multi-map reads - which up to a certain extent recovers some of the lost reads due to the high homology of the CGB group of genes. The multimap track therefore show a much higher read count and a 100% coverage of the transcript.

4.10.3 *CCR7* (CC chemokine receptor type 7)

CCR7 is best known for its involvement in the maturation of dendritic cells and thus in adaptive immune response (Sanchez-Sanchez, Riol-Blanco et al. 2006). *CCR7* expression goes from an RPKM value of 0.02 at 0 days to 122 at 8 days and ranks as the eleventh most abundant transcript in the 8 day trophoblast. There are no previous reports of its expression in the trophoblast but an analysis of the Zhang *et al* data clearly indicates its expression in the human blastocyst. Potential functions for CCR7 could either be a measure to protect the fetus from future infections, or a way to modulate inflammatory reactions between the fetal - maternal interface, through the communication between the trophoblasts and dendritic/Treg/NK cells in the endometrium which are known to play a positive role during implantation. Considering the up-regulation of the retroviral elements (discussed later) it is likely that CCR7 and expressed retroviral elements misdirects the maternal immune system providing an immunosuppressive function.

4.10.4 *KRT23* (Keratin type I cytoskeletal 23)

Keratins are classical markers of epithelial cells. In the mouse blastocyst, Krt8 and Krt18 (and Krt7 and 19) are used to mark the early trophoblast lineage, while in the human blastocyst it is KRT7 that is the classical marker of these cells. I see abundant expression and up-regulation of *KRT7* (7.8 fold), *KRT8* (6.8 fold), *KRT18* (5.8 fold), and *KRT19* (6.0 fold). In addition, I also see *KRT23*, a keratin not previously defined as a trophoblast keratin but is highly expressed in the placenta (*KRT23* entry of biogps expression database at http://biogps.gnf.org) and the trophoblast differentiation (5272.3 fold). KRT23 is required for epithelial cells and its up-regulation can be used

as an additional confirmation that the differentiation has produced cells of the epithelial lineage. Since *Krt23* expression is not seen in mouse early development, *KRT23* can be considered as a potential human trophoblast specific gene.

Comparing the RefSeq annotation of *KRT23* gene and the RNA-Seq peak profile, it is clear that the *KRT23* isoform expressed during differentiation has a different transcription start site to that shown in RefSeq. The transcript seems to skip the first exon in the 5' UTR region and start at the next exon which contains the start codon. While this does not affect the structure and therefore the function of the protein since the coding sequence remains the same, it might be differently regulated post transcriptionally, due to the change in UTR (Figure 4.14).

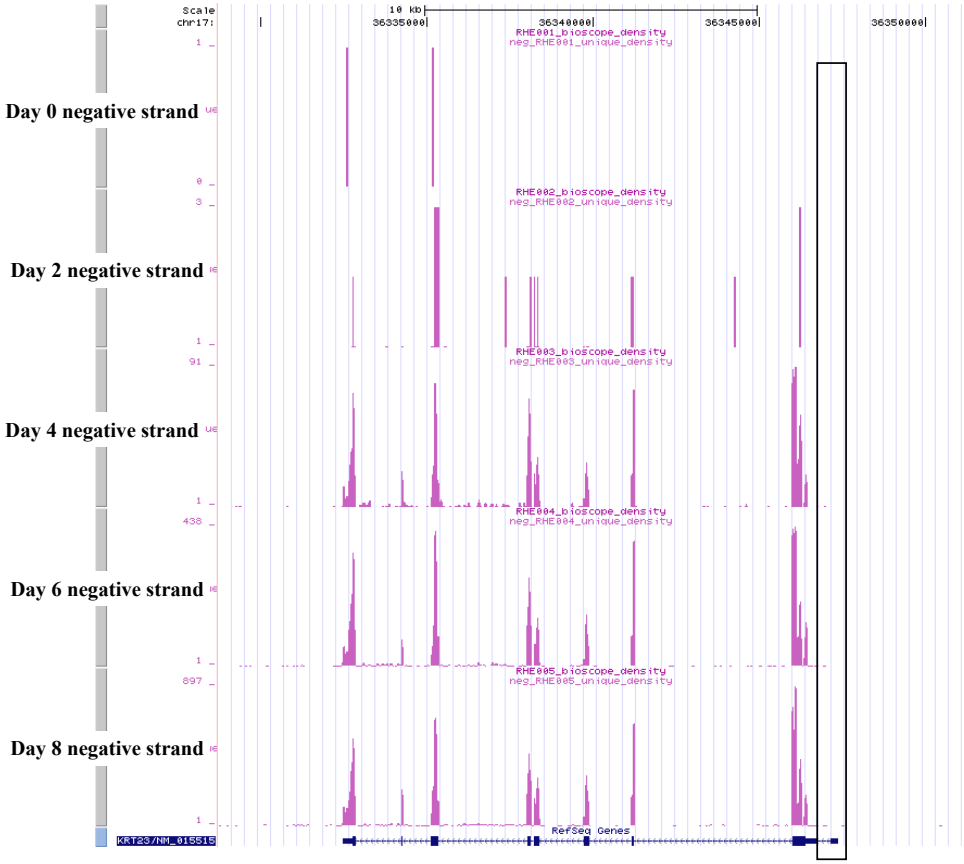


Figure 4.14: The RNA-Seq peak profile of *KRT23* gene.
The first exon in the 5' URT region (shown within the box) is not transcribed during differentiation, giving rise to a new isoform. For clarity only the negative strand is shown.

4.10.5 *H19*

H19 is a long non-coding RNA, which is well known to be highly expressed in the placenta. *H19* is reported to be modulated by steroid hormones including 17-ß-estradiol which is the dominant form of estrogen, in mammary glands and the uterus (Adriaenssens, Lottin et al. 1999).

Considering RNA-Seq data, *H19* expression is induced immediately after treatment and gets highly up-regulated during trophoblast differentiation. This expression pattern clearly shows that *H19* expression in early development is not only limited to maternal tissues, but is also expressed in the embryo and that H19 is involved in trophoblast differentiation. Furthermore miR-675 which originates from the original H19 transcript (Cai and Cullen 2007) is highly up-regulated based on the microRNA RNA-Seq data presented in the latter part of this thesis suggesting a regulatory role for H19.

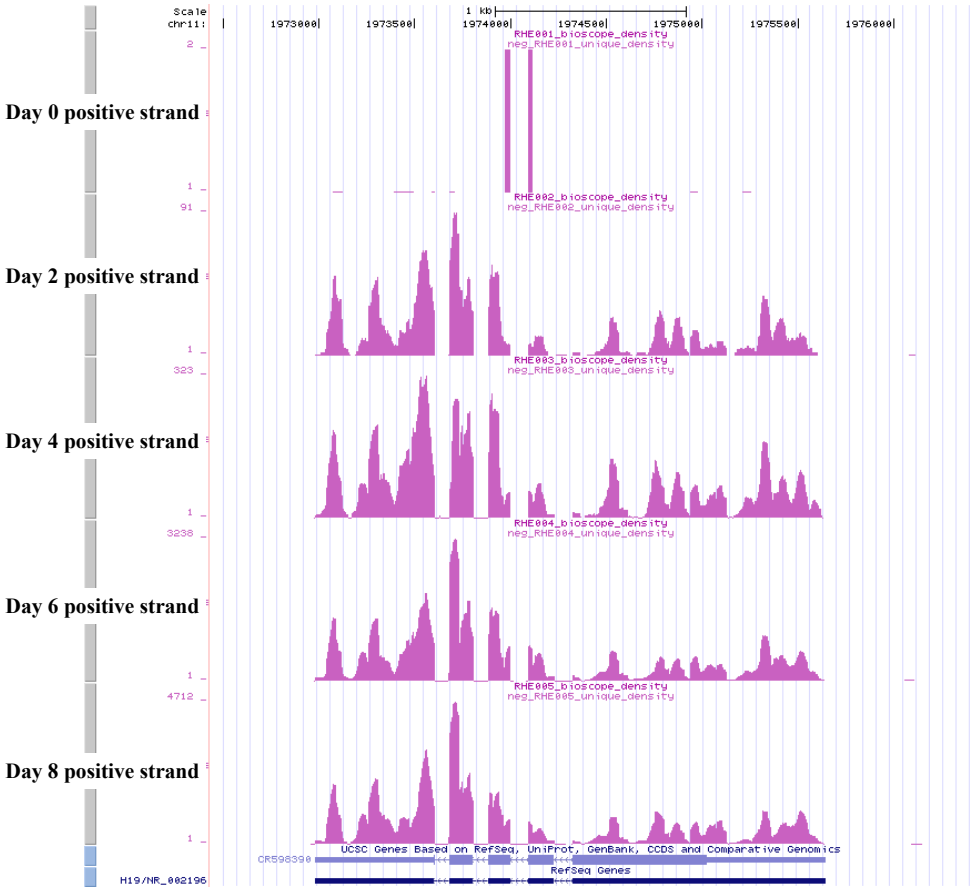


Figure 4.15: RNA-Seq peak profile of H19.
H19 gets up-regulated from day 2 onwards. For clarity only the positive strand is shown.

4.10.6 *MUC15* (Mucin 15)

Mucins are high molecular weight substances secreted by epithelial cells to form a sticky mass. Mucins are secreted by the uterus and is reported to aid implantation (Carson, DeSouza et al. 1998). The function of mucins is not limited to implantation. They are also reported to be expressed in the human placenta and suppress the invasion of trophoblast-like cells in vivo (Shyu, Lin et al. 2007). This suggest that mucins directly or indirectly regulate the migratory properties, first by facilitating implantation and then by regulating invasion and plays an important part during early trophoblast development.

RNA-Seq data shows that *MUC15* expression is induced during trophoblast differentiation and is highly up-regulated (1867.4 fold) at day 8 of differentiation. This suggests that mucins are secreted by the blastocyst and it is confirmed by the fact that *MUC15* is one of the greatest up-regulated genes in the Zhang *et. al.* data during the transition from the 8-cell to the blastocyst stage. In addition to this mucin secretion also indicates the epithelial phenotype is acquired by the differentiated cells. RNA-Seq junction reads show that two out of the three known *MUC15* isoforms are expressed.

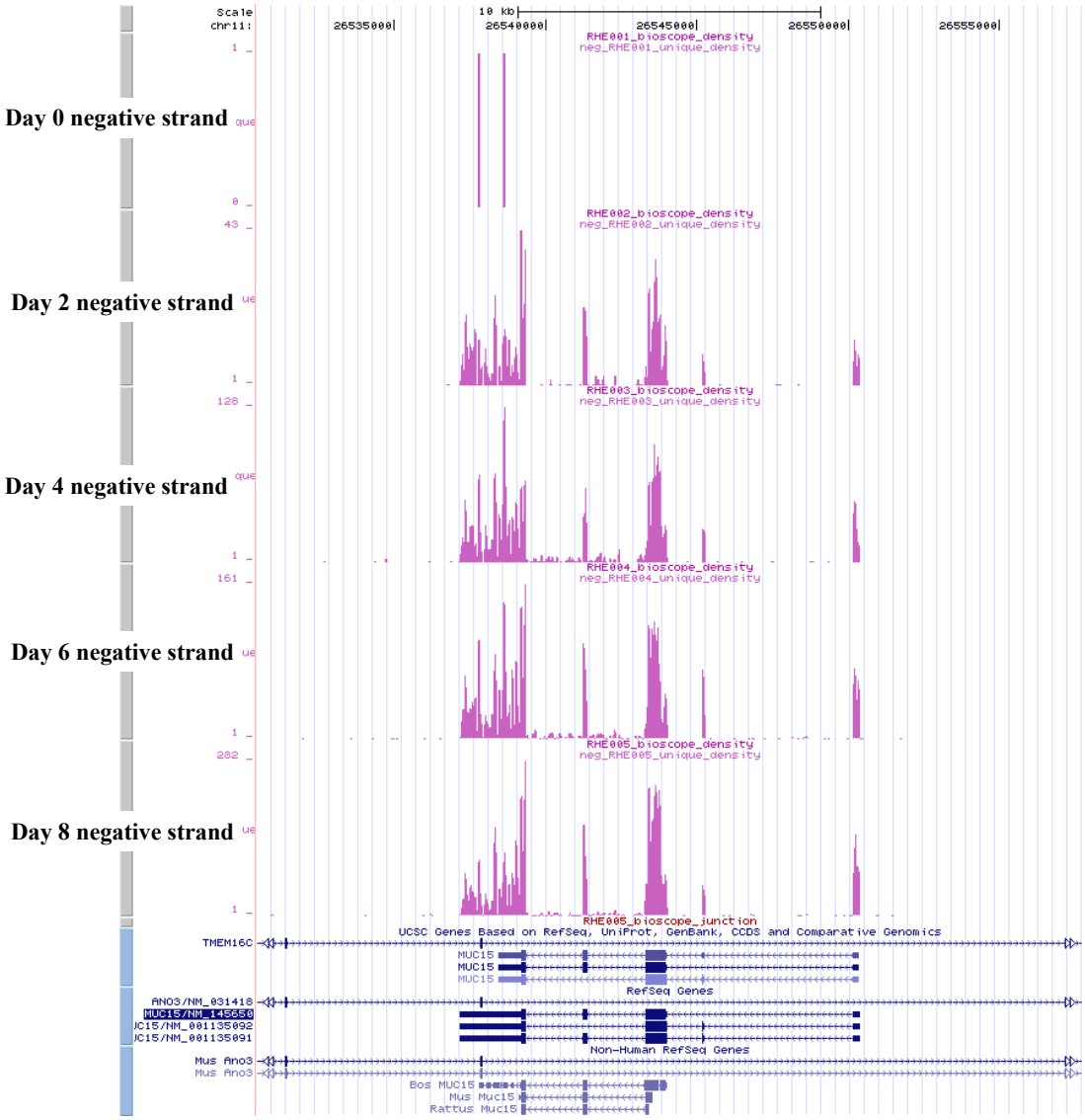


Figure 4.16: RNA-Seq peak profile of *MUC15*.
MUC15 shares its gene footprint with *TMEM16C* which is coded by the opposite (positive strand) strand.

4.10.7 *SLC40A1* (Solute carrier family 40 (iron-regulated transporter), member 1)

The product for this gene, Ferroportin 1, is essential for iron efflux. It has been identified as expressed in the human placenta, found on the basal surface of the syncytiotrophoblasts (Donovan, Brownlie et al. 2000). In my data, *SLC40A1* is highly up-regulated during 4 days of trophoblast differentiation and by 8 days reaches an RPKM value of 152 (1715.9 fold change). This is one of many examples in my expression data, where up-regulation is seen in a key molecule involved in nutrient supply between mother and the fetus through the trophoblast.

Located on the basal surface (Donovan, Brownlie et al. 2000), SLC40A1 is in a position to secrete iron, out from the trophoblast cytoplasm towards the developing embryo. With respect to iron uptake by the trophoblast from the maternal side, it is interesting to note that *TFRC,* encoding transferrin receptor-1 and functioning in iron uptake, is abundantly expressed (86 RPKM) in the 8 day trophoblast. Thus presumably, both apically positioned transferrin receptor-1 and basally positioned SLC40A1 are able to supply iron to the fetus from the mother. This along with folate, of which the transporter is also expressed (*FOLR1* 25 RPKM at 8 days), adds to the mounting evidence that iron supplementation at preconception and early pregnancy is important for improved pregnancy outcomes (Titaley, Dibley et al. 2010). Relevant to this is the recent finding that *Slc40a1* is essential for mouse neural tube closure (Mao, McKean et al. 2010), which is precisely the role folate supplementation is known to play in early human development.

4.10.8 *GCM1* (Glial cells missing homolog 1)

As described in the introduction, GCM1 is a transcription factor essential for mouse placental function, though its expression in the mouse trophoblast lineage occurs much later than initial trophoblast formation. Based on RNA-Seq data, *GCM1* is not expressed in human ES cells, but is highly up-regulated from day 4 of the trophoblast differentiation (Figure 4.17) to a final fold change of 1226.1.

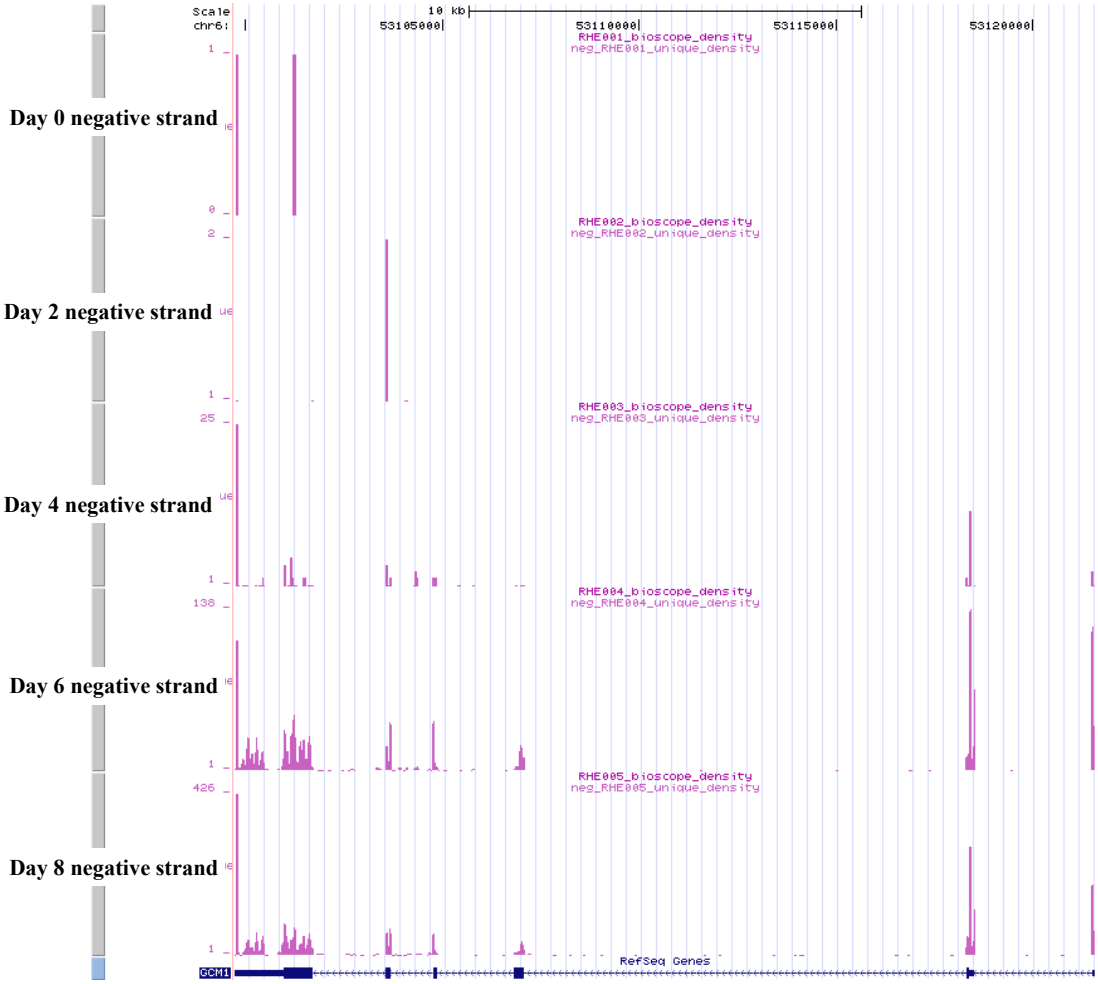


Figure 4.17: RNA-Seq peak profile of *GCM1* gene expression.
There is negligible expression in day 0 (undifferentiated ES cells). The Expression begins at day 4 of differentiation and increases through the time-points.

4.10.8.1 Regulation of GCM1

Regulation of *GCM1* is carried out by the proteins GSK3B, FBW2 and DUSP23 at the post-transcriptional level. GSK3B phosphorilates GCM1, marking it for degradation by FBW2 (Chiang, Liang et al. 2009) while DUSP23 dephosphorilates GCM1 (Lin, Chang et al. 2010) preventing the GCM1 degradation. Interestingly during SB differentiation GSK3B gets up-regulated and DUSP23 gets down-regulated suggesting that even at earlier preimplantation stages the GCM1 regulatory machinery is active.

4.10.9 Placental *BDNF* (Brain-derived neurotropic factor) / *NTRK2* (Neurotropic tyrosine kinase 2) system

It has been reported that in mice, BDNF plays an important role in implantation and placental development (Mayeur, Silhol et al. 2010). In the hESC-trophoblast differentiation protocol *BDNF* is not expressed at very significant levels (max 3.31 RPKM at 8 days) but its receptor - TrkB (*NTRK2*) - is, being up-regulated from 0 to 62 RPKM over the 8 day time course. This would suggest that the trophoblast is responsive to BDNF, perhaps supplied from the maternal endometrium.

TrkB, in mammals has a full length and a truncated isoform (Tapia-Arancibia, Rage et al. 2004). The truncated isoform, while lacking intracellular tyrosine kinase activity, is active and can trigger transduction signals (Tapia-Arancibia, Rage et al. 2004; Skaper 2008). The truncated isoform of *TrkB* (Trkb-T1) is able to regulate Rho A signaling (Ohira, Homma et al. 2006) and Rho A is shown to be predominant in cytotrophoblast cells and is implicated in trophoblast migration (Shiokawa 2002; Mayeur, Silhol et al. 2010).

The RNA-Seq peak profile of both *NTRK2* (Figures 4.18) shows an interesting transcriptomic phenomenon. In the case of *NTRK2*, which based on RefSeq has five different isoforms, only the shortest isoform is expressed during trophoblast differentiation. Based on peak heights, it seems a few of the longer isoforms are expressed at very low levels but the shortest isoform is clearly the highest expressed. It is this isoform (*Trkb* - T1) that has shown to be involved in regulating RhoA signaling (Ohira, Homma et al. 2006).

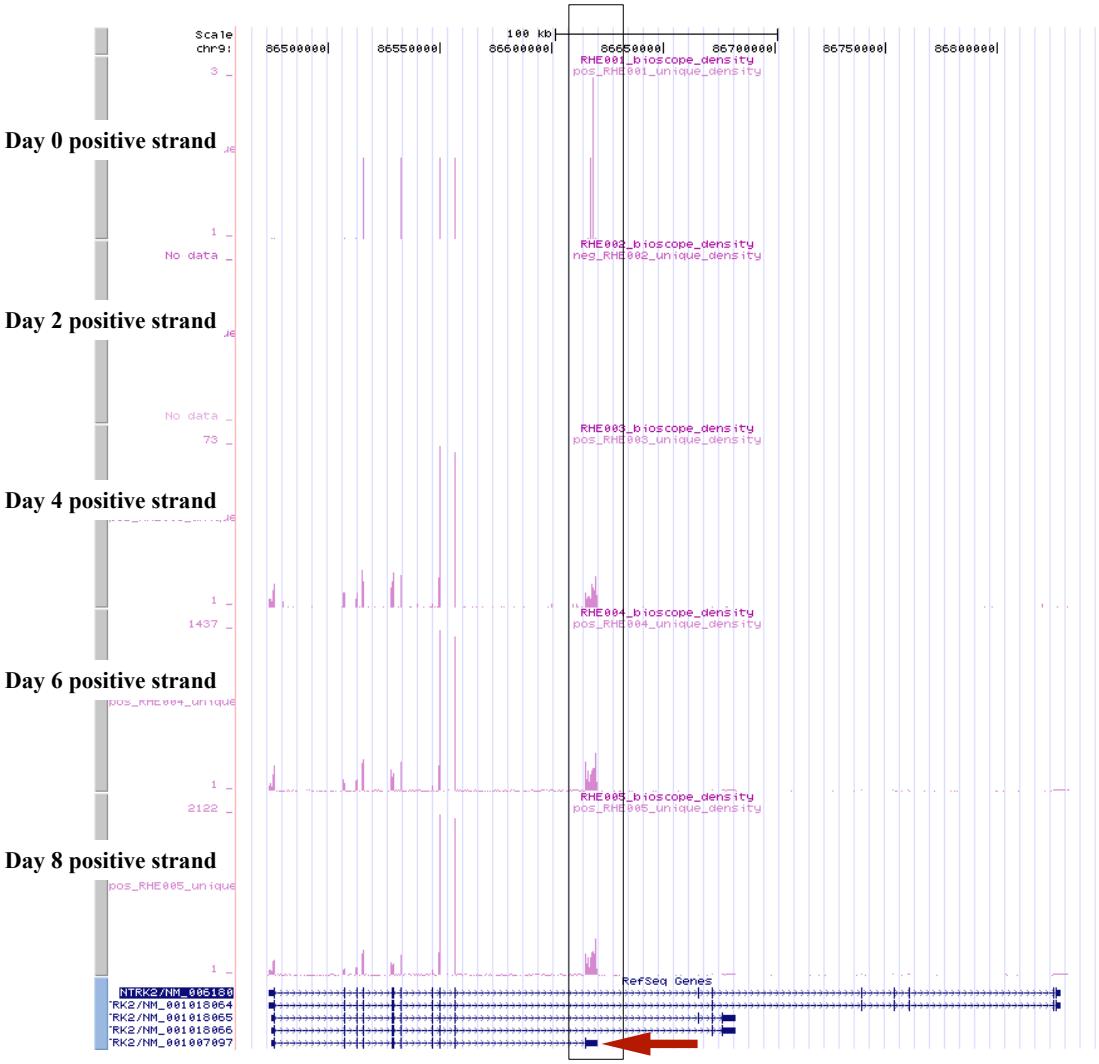


Figure 4.18: The RNA-Seq peak profile of *NTRK2*.
Data on peak distribution proves that shortest isoform, indicated with a red arrow is highly expressed.

4.10.10 *ELF5* (E74-like factor 5)

As a transcription factor essential for mouse trophoblast, much attention has been drawn to *ELF5* in the human trophoblast literature. Indeed, its lack of expression in other hESC-derived trophoblast populations has been used to argue against these cells being true trophoblast (Hemberger, Udayashankar et al. 2010). Thus it is comforting to see from my RNA-Seq data which clearly shows that the SU5402+BMP4 differentiation protocol used in this thesis does indeed induce *ELF5* expression (0 RPKM at day 0 and 2.96 at day 8) thereby providing a more realistic transcriptomics picture of trophoblast differentiation. ELF5 expression is not high, but since its been reported to be methylated in human ES cells, any form of expression indicates that the trophoblast differentiation leads to its de-methylation. Furthermore ELF5 is considered to be a trophoblast stem cell marker and not a marker for the entire trophoblast lineage, which is brought about by the differentiation.

ELF5 has two isoforms, 2a and 2b. It has been reported that *ELF5* - 2b is the major variant found in the placenta (Hemberger, Udayashankar et al. 2010). In agreement with above, the SU+BMP4 differentiation clearly induces *ELF5* - 2b (Figure 4.19).

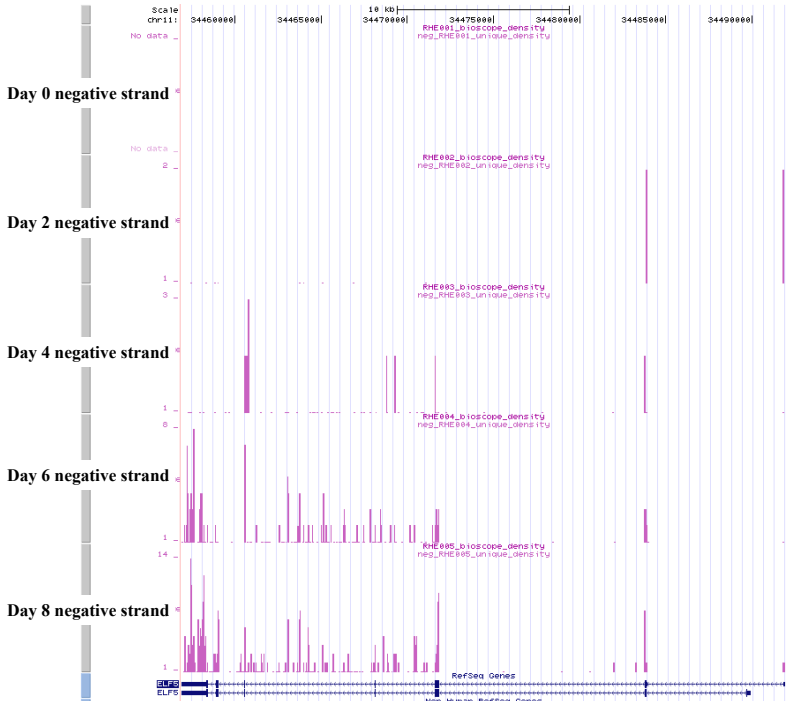


Figure 4.19 : Peak profile of *ELF5*.
ELF5 expression begins at round day 4 - day 6 during the differentiation. ELF5 has two isoforms 2b and 2a. 2a has its first exon within the first intron of 2b. SU+BMP4 induces the ELF5 - 2b isoform, just as in the placenta where it is the major variant.

4.10.11 *ABCG2* (ATP-binding cassette sub-family G member 2)

ABCG2 is highly expressed in the human placenta, and is believed to protect the fetus from xenobiotics transported from the maternal circulation (Kolwankar 2005). It has been shown that knocking down of ABCG2 in BeWo cells, causes the down-regulation of trophoblast markers and reduces cell fusion (Evseenko, Paxton et al. 2007). ABCG2 is directly regulated by estrogen and PPARgamma (Szatmari 2006), both highly expressed during early development and during the trophoblast differentiation.

The RNA-Seq peak profile of *ABCG2* shows an interesting transcriptomics dynamic. *ABCG2* gene which is expressed at low levels in human ES cells get highly up-

regulated during differentiation (61.7 fold). Furthermore, as can be seen in Figure 4.20 panel A, the *ABCG2* changes the starting exon of transcription during differentiation. In addition to this the third exon which is not expressed in human ES cells starts getting expressed.

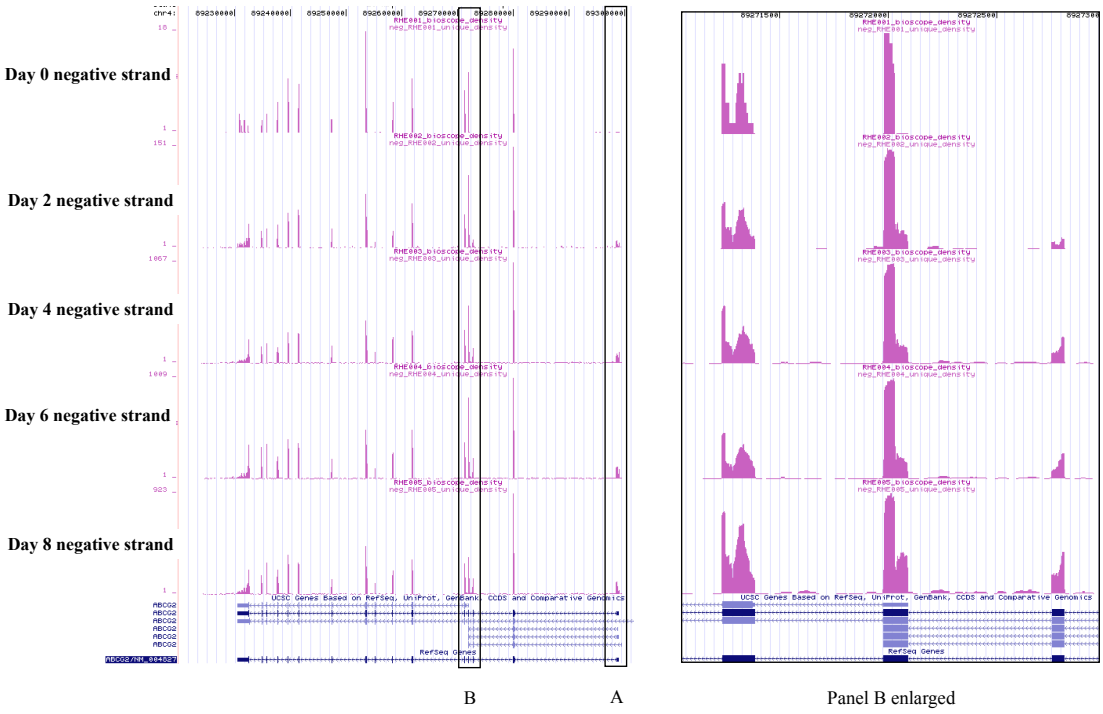


Figure 4.20: The expression and splicing dynamics of *ABCG2*.
The entire RNA-Seq profile is shown on the right panel. The box marked A shows the first exon which is unused in human ES cells and used during differentiation. The box marked B shows the third exon which has a similar expression pattern as the first - see the enlarged view on the right.

4.11 Comparison with mouse pre-implantation data

While the main focus of this thesis is human trophoblast development, it is important to compare the data from the human samples with the mouse model. This comparison enables the validation of existing knowledge as most observations regarding early development originate from the mouse system, and enables the identification of human specific phenomena during trophoblast differentiation.

The mouse pre implantation RNA-Seq dataset presented in this thesis consists of four samples - uncompacted 8-cell, E3.5 blastocyst, E4.5 Blastocyst and E4.5 ICM. (The E4.5 trophoblast cells are difficult to isolate from the embryo without being contaminated by cells in the ICM). Dr. Guo Guoji, a former member of the lab, carried out the collection of the mouse embryos and performed the immunosurgery (Solter and Knowles 1975) to isolate the ICM.

Using the four mouse RNA-Seq samples, trophoblast related gene expression changes can be identified via the comparison of gene expression changes between the E4.5 blastocyst / 8-cell and E4.5 ICM / E4.5 blastocyst. i.e genes which are involved in mouse trophoblast differentiation can be considered as genes which show an up-regulation in both E4.5 Blastocyst / 8 cell stage comparison and E4.5 Blastocyst / E4.5 ICM comparison (i.e. low in 8 cell, low in ICM but high in blastocyst). At the uncompacted 8-cell stage, markers and some key regulators of the trophectoderm are not yet expressed, morphological epithelialization of the outer cells does not occur until the early 32-cell stage. The ICM is clearly distinguishable from the trophectoderm by gene expression as early as the later 32-cell stage (Guo, Huss et al. 2010).

To validate the mouse RNA-Seq data it was compared with the expression levels of the 48 genes presented in Guo et al. There was a good qualitative co-relation between the two datasets. For example *Gata3* and *Cdx2* which are expressed at low levels in the 8 cell stage and becomes TE specific showed a 4.59 and 4.11 fold up-regulation (8 cell stage vs outer cells at 32 cell stage) in Guo et al and a 35.1 and 35.7 fold (8 cell stage vs E4.5 blastocyst) in mouse RNA-Seq data. *Nanog* and *Sox2* which are specific to ICM compared to TE shows a fold enrichment of 17.1 and 354.6 in the ICM (32 cell stage in vs out cells) based on Guo et al data and mouse RNA-Seq data showed an enrichment of 2 and 1.5 fold between E4.5 blastocyst and E4.5 ICM. As can be seen from these data the expression of key genes is qualitatively the same in both datasets. The values of fold change differs as the sample types used for the comparison are different.

When comparing the gene sets related to trophoblast development in both human and mouse systems, the most clear observation is the significant difference of both the expression level and expression pattern with each other. This point is clearly illustrated in Figure 4.21. The figure is a scatterplot of the top 500 up-regulated genes in human during trophoblast differentiation. The genes which have a RPKM value of less than 1 has been reset to one to make the comparison simpler. As can be clearly seen there is no correlation between human and mouse. There are a number of genes which are highly expressed in human but are not significantly expressed in mouse (see the area highlighted in red).

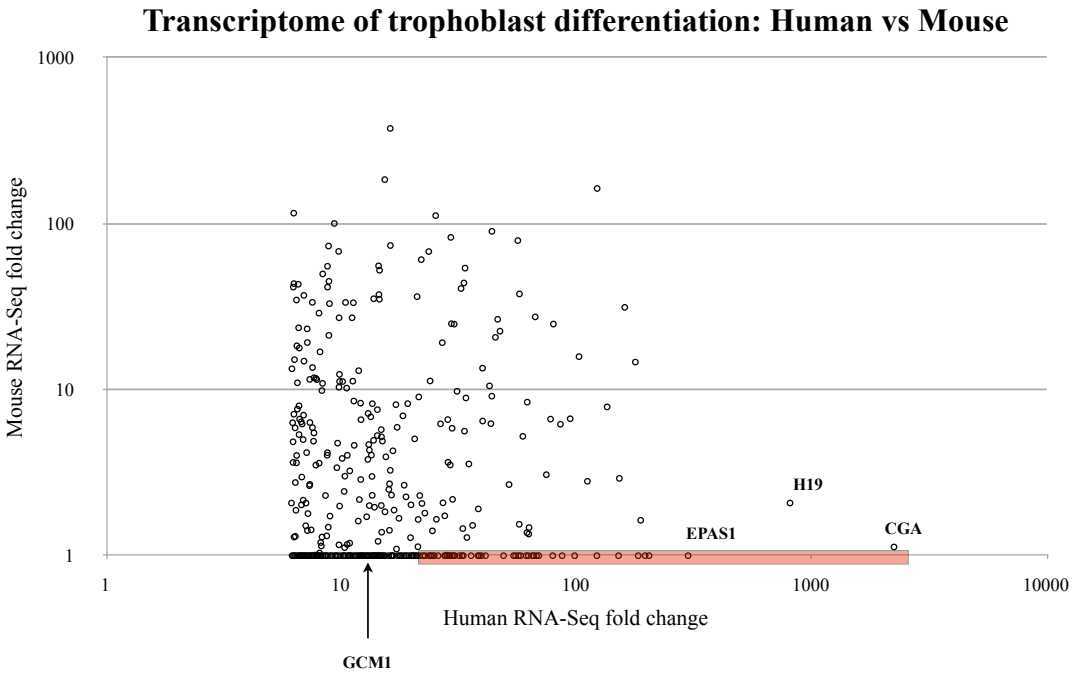


Figure 4.21: Comparison of mouse and human RNA-Seq data during trophoblast differentiation.
The RPKM values less than 1 has been reset to 1. Only the top 500 up-regulated genes in human are shown. The top 35 genes which are highly up-regulated in human and expressed at very low level (less than 1 RPKM) in mouse are highlighted in red.

It should be highlighted that the comparison done in Figure 4.21 is not exact. In-fact considering the lack of clinical samples of human early trophoblast development, a one-to-one comparison is not feasible. Therefore for the comparison human ES cells differentiated into the trophoblast lineage was used. Given the evidence presented here it could be correctly assumed to produce a realistic early trophoblast transcriptome.

Due to the difference in the initially available RNA amounts two different protocols (as outlined in the methods section) was used to process the samples. The protocol used to process the trophoblast differentiation samples produce reads specific for transcripts with a poly A tail while the single cell protocol used to process mouse embryos provides information on all the transcripts - *including* poly A ones. Therefore as far as the poly A genes are concerned the data should be comparable. The usage of

fold change as the primary measure of gene expression also removes sample specific biases.

The purpose of the Figure 4.21 and this entire section, is to emphasize that at the transcriptome level, human and mouse systems, during trophoblast differentiation show significant difference and cellular heterochrony.

4.11.1 *GCM1* expression in SB differentiation, human and mouse early development

The gene expression of *GCM1* in both human and mouse systems during early development is drastically different. Based on the Zhang et. al. paper, raw probe intensity value of *GCM1* during the human 4 cell stage embryo is 50 and it gets increased up to 7117 during the blastocyst stage. The human trophoblast RNA-Seq data has a similar pattern where in human ES cells *GCM1* gets only 2 reads and at day 8 it increases to 2399 reads - a 1226 fold up-regulation based on RPKM value. However in the mouse RNA-Seq system this drastic increase is not seen. In fact the up-regulation of Gcm1 from 8 cell stage to E4.5 blastocyst is just 2.73 fold, where the E4.5 blastocyst sample gets only 45 reads being aligned to the gene. This drastic up-regulation of *GCM1* during human trophoblast differentiation as compared to the mouse system suggests that GCM1 plays a more important role in the human system compared with the mouse. This is further confirmed by the observed up-regulation of *PGF*, Syncytins and Aromatase, which are genes regulated by GCM1, suggesting that GCM1 protein is highly active during trophoblast formation.

4.12 Retroviral expression as a possible explanation for the transcriptomic difference of early development factors in human and mouse

As explained in the Introduction, expression of endogenous retroviral components have the to capacity influence the transcriptome of early development. Existing data shows that they can create new genes and form regulatory regions of existing genes influencing their expression. Since the endogenous retroviral component of the genome changes with evolution, the retroviral elements in mouse and human can be considered to be quite different, and this difference has the potential to bring about the changes in the gene regulatory mechanisms of early development in mouse and human.

4.12.1 Expression of genes originated from retroviral elements during trophoblast differentiation

The fusogenic Syncytin 1 and Syncytin 2 are primate-specific genes, which originated from retroviral elements inserted into the ancestral genome ~25 and 40 million years ago respectively (Cheynet, Ruggieri et al. 2005; Renard, Varela et al. 2005). They are induced during the hESC-to-trophoblast differentiation. The human pre-implantation data also shows a similar expression pattern. Their peak profiles are shown in Figures 4.22 and 4.23.

Expression of Syncytin 1 and 2 starts at day 6 and gets up-regulated at day 8. Considering the distribution of uniquely mapped reads, Syncytin 1 which is the newer gene among the two, has a coverage of around 50% while the older Syncytin 2 has a 100% coverage. Once the multi-mapped reads are used to measure the coverage, Syncytin 1 reaches a 100% coverage (Figure 4.24). This highlights an important point

when trying to identify new retroviral insertions. When the insertion is relatively new, even if it is highly expressed, due to sequence similarity, aligning reads becomes difficult and the coverage goes down. However when the insertion gets "older" and accumulates point mutations, then identification of those regions through RNA-Seq becomes less difficult due to ease of alignment.

ERVWE1 (*Syncytin* 1)

Figure 4.22 : RNA-Seq peak profile of *Syncytin* 1.
Syncytin 1 is not expressed in day 0 and day 2. It starts to get expressed in day 6 and is up-regulated there after. There is only a partial UCSC annotation. Note the new exon on the 3' end. Only the negative strand of each time point is shown for clarity.

***HERV-FRD* (*Syncytin* 2)**

Figure 4.23: RNA-Seq peak profile of *Syncytin* 2.
This has a similar expression pattern as Syncytin 1.

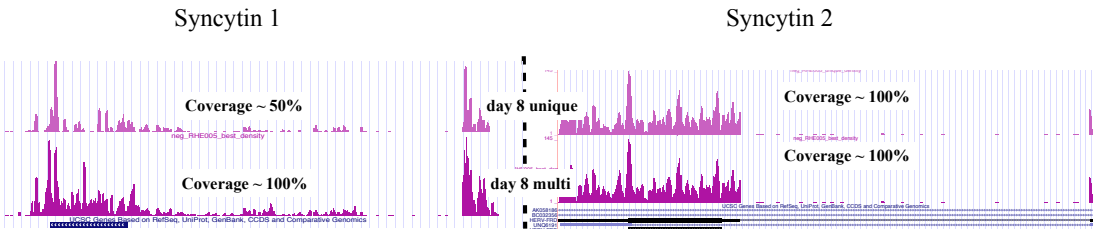


Figure 4.24 : Change of coverage of expressed retroviral elements with insertion time.
The first track shows the standard unique reads whereas the second track shows multi-mapped reads which are reads that map to the genome up to 10 times. *Syncytin* 2 gene which is the older of the two has 100% coverage in both unique and multi-mapped tracks, while *Syncytin* 1 show only around 50% coverage in the unique track.

4.12.2 Expression of genes with retroviral derived regulatory elements during trophoblast differentiation

Trophoblast-specific expression from ERV LTRs has been described for a number of genes (Cohen, Lock et al. 2009) . I first investigated my data set to determine if there was expression derived from these ERV LTRs. Of the 9 placenta-specific ERV LTRs described in Cohen et al. there was evidence for significant expression from *CYP19A1* (RPKM = 102 at day 8), *PTN* (RPKM = 244), *INSL4* (RPKM = 1.97), *PAPPA2* (RPKM = 25), *MID1* (RPKM = 14) and *EDNRB* (RPKM = 60) but no expression from *IL2RB*, *NOS3* and *ENTPD1*.

4.12.3 *CYP19A1* (Cytochrome P450, family 19, subfamily A, polypeptide 1)

Human trophoblast expression of *CYP19A1* is known to be driven by an ERV LTR promoter (Conley and Hinshelwood 2001; Cohen, Lock et al. 2009). In my data set *CYP19A1* is highly up-regulated during SB differentiation (Figure 4.25). Based on RNA-Seq data, the placenta-specific isoform is expressed while the others, driven from different tissue-specific promoters, are not. Based on the UCSC annotation track it seems that another - third isoform is expressed, but its expression is not as nearly as high as the placenta specific one.

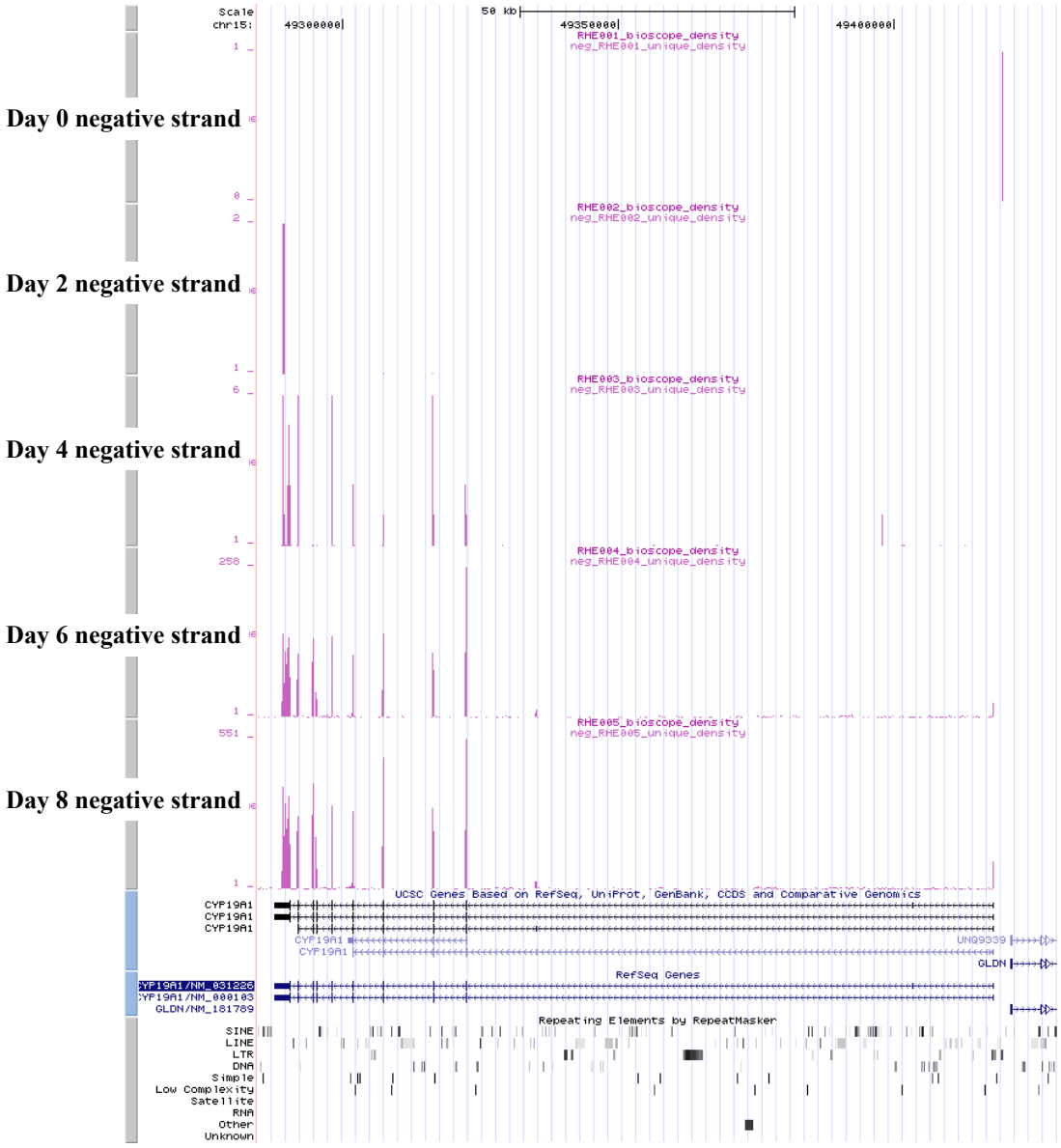


Figure 4.25: The RNA-Seq peak profile of *CYP19A1*.
It is not expressed at day 0 (undifferentiated human ES cells) and gets highly up-regulated during differentiation.

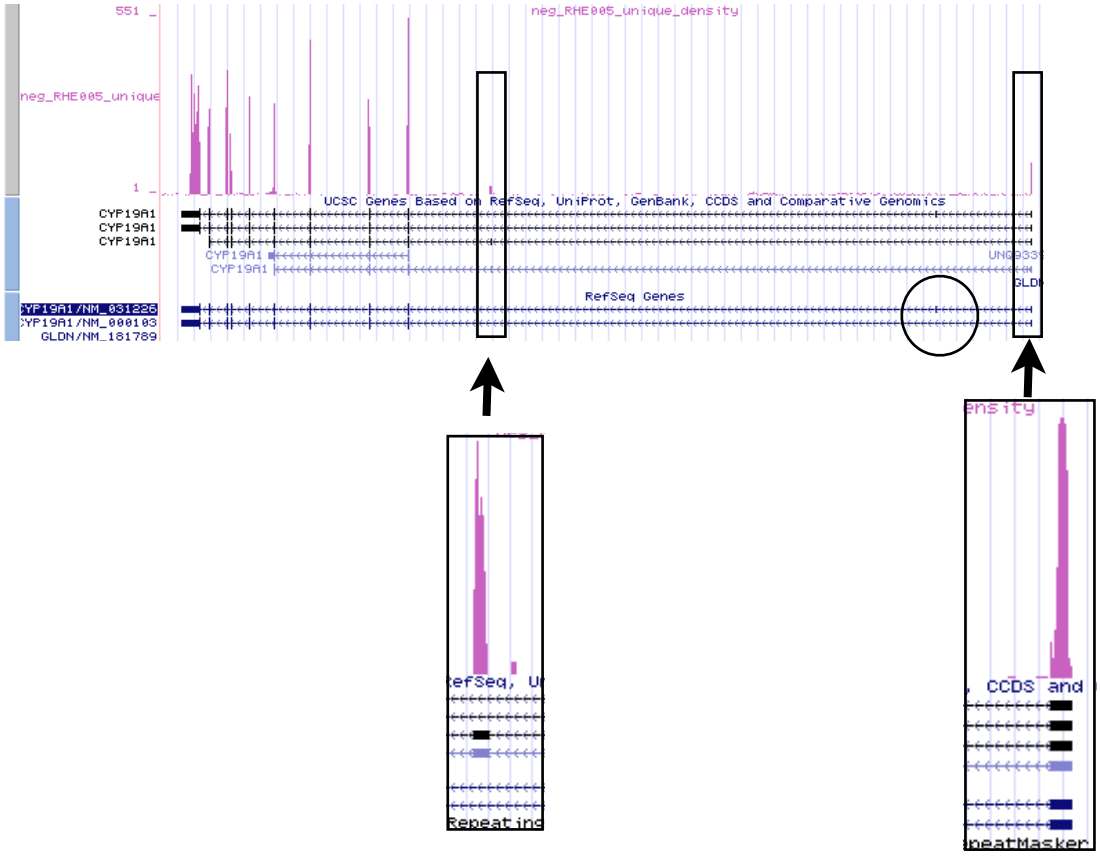


Figure 4.26: The RNA-Seq expression profile of *CYP19A1* at day 8 time point of SB differentiation protocol.

The major expressed isoform incorporates the first exon (enlarged view in the box on the right) and is the one reported to be placenta specific. The other RefSeq isoform incorporating the second exon (shown by the circle) is not expressed. RNA-Seq data also shows the expression of a third isoform (shown enlarged in the box on the right), which is unannotated in RefSeq.

4.12.4 *EDNRB* (Endothelin receptor type B)

Based on RNA-Seq data *EDNRB* is highly up-regulated during SB differentiation. Unlike *CYP19A1,* it is expressed in undifferentiated human ES cells and this expression is further up-regulated throughout the treatment. What has been previously described as the placental-specific isoform, driven from an ERV LTR promoter, starts to be expressed from day 6 onwards of the differentiation process. Expression from this LTR-based promoter only accounts for 10-15% of total expression at the 8 day time-point. Based on RNA-Seq data, a novel third exon is observed between the expressed first exon and the non-expressed second exon. This novel isoform seems to have an expression pattern similar to the placenta specific one (Figure 4.27).

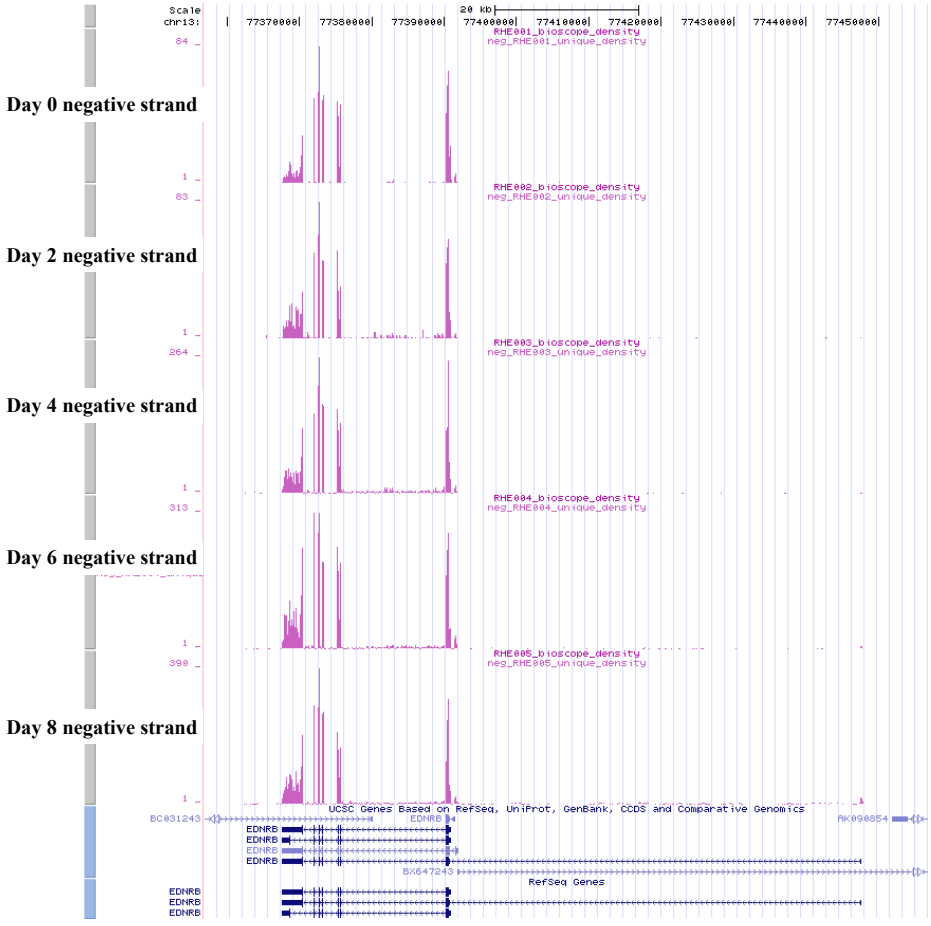


Figure 4.27 : The RNA-Seq peak profile of *EDNRB*.
The gene gets up-regulated during differentiation. While the gene is expressed (albeit at a lower level) in undifferentiated human embryonic stem cells, the placenta specific isoform is only expressed from day 6 - 8 onwards.

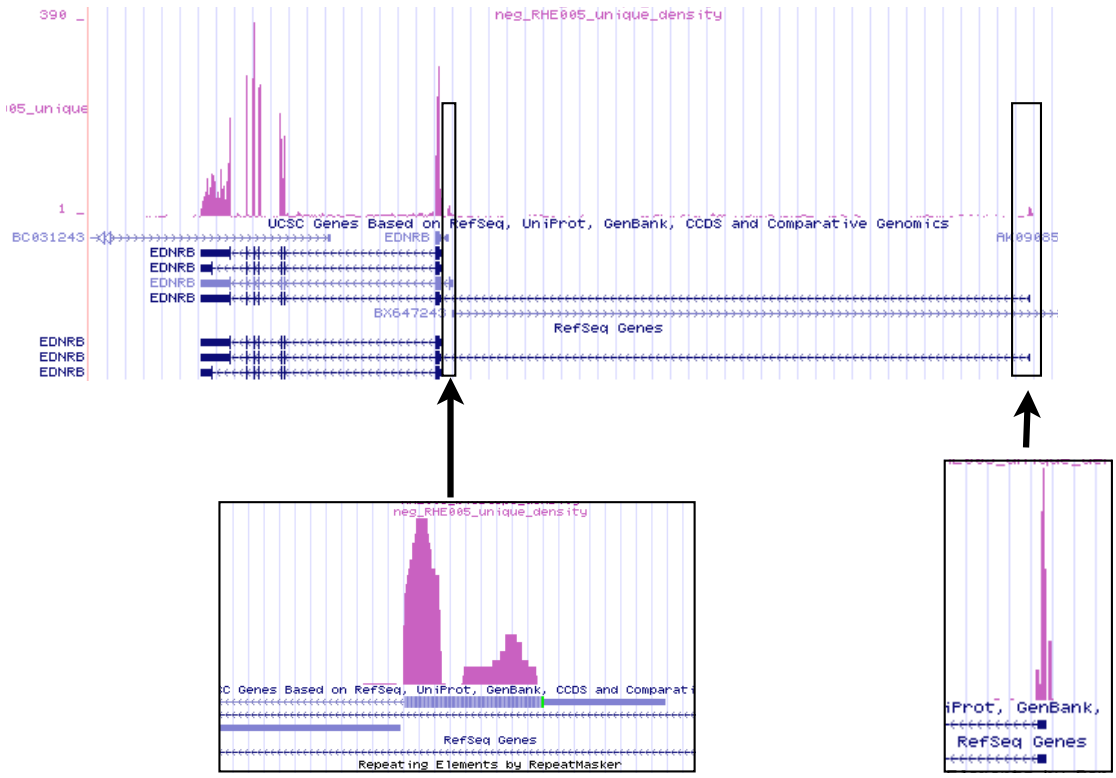


Figure 4.28: An enlarged view of the RNA-Seq expression profile of *EDNRB* gene at day 8 timepoint.

The box on the right shows the exon responsible for the placenta specific isoform under the regulation of the LTR promoter. In accordance with published data this isoform is expressed at around 10 - 15% of the total transcripts (based on peak height). The third isoform (which does not have a RefSeq annotation) shows a similar pattern as the placenta specific one, and its unique exon is shown in the box on the left.

4.12.5 PTN (Pleiotrophin)

PTN gene has an HERV-C family LTR region, which acts as a placenta-specific alternative promoter producing a different isoform (Schulte, Lai et al. 1996) . Based on RNA-Seq, *PTN* is highly up-regulated during differentiation. The placenta-specific isoform is highly expressed at day 8, but there also is expression in undifferentiated human ES cells albeit at a much lower level. The RNA-Seq data also indicates the presence of another novel exon, which is not annotated in either the RefSeq or UCSC tracks. This is absent in day 0 and the early days of differentiation but starts to get expressed at day 6 and onwards (Figure 4.29) and, interestingly, the new exon is actually an expressed LTR - ERV1 element.

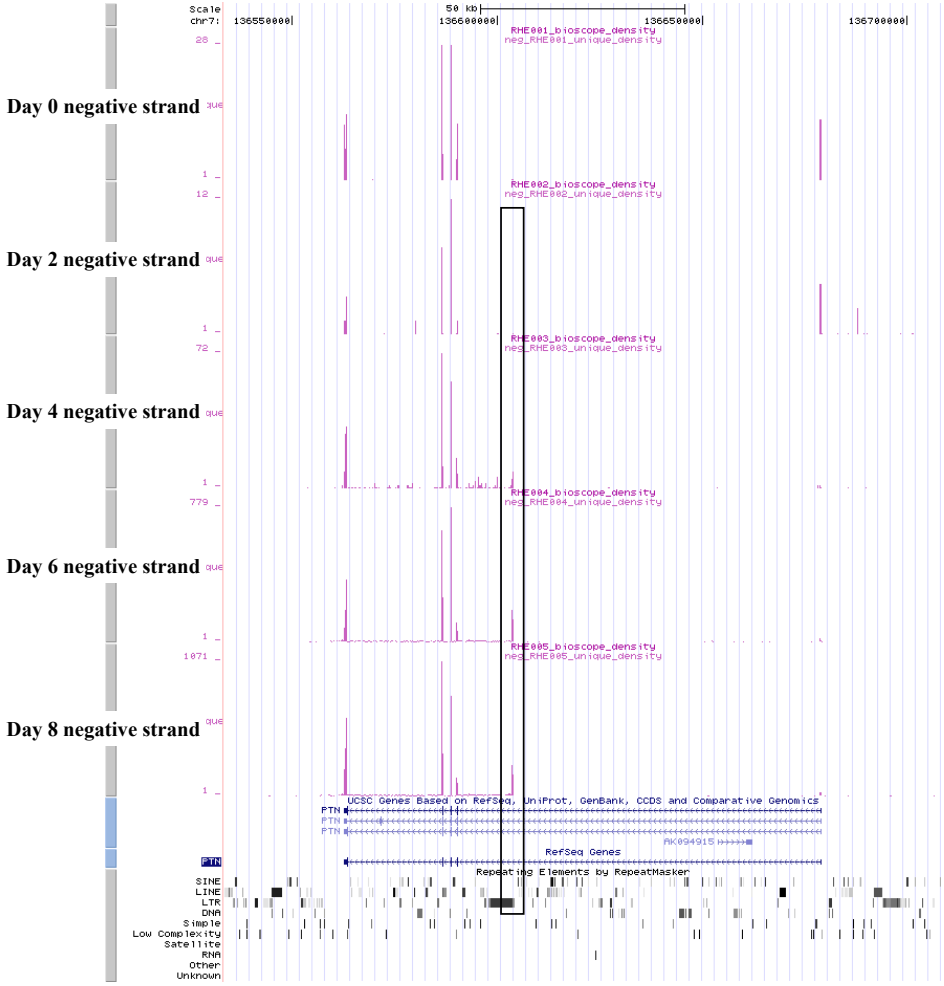


Figure 4.29: The RNA-Seq expression profile of *PTN* gene.
It gets highly up-regulated during differentiation. Novel exon is highlighted.

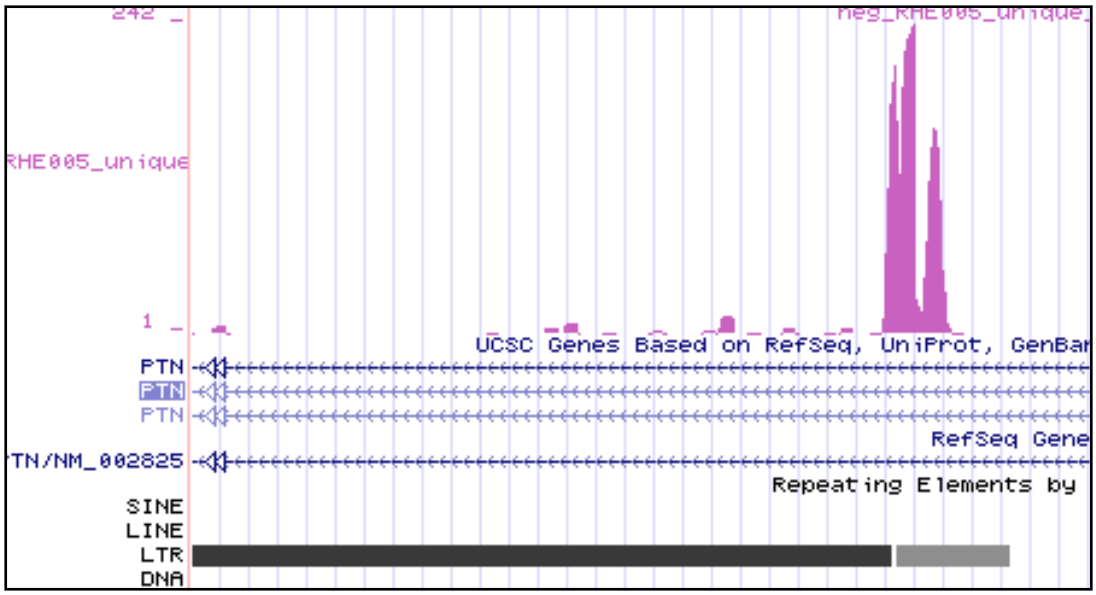


Figure 4.30 : A magnified view of the novel exon, with an LTR footprint of *PTN* gene found by RNA-Seq.

4.13 Novel transcribed regions (NTRs) active during trophoblast differentiation

One of the advantages of RNA-seq over other gene expression detection methods is that it provides an unbiased view of the transcriptome and thus has the opportunity to identify transcripts not defined previously. This feature is particularly important in the context of the cell type I was analyzing as the early human trophoblast has not been fully transcriptionally explored before. For this reason I spent some effort in trying to identify and characterize novel transcribed regions (NTRs). NTRs are defined as expressed regions in the genome that do not have any valid RefSeq annotations (see Methods for details). NTR detection was done for all five samples.

Time point	Total NTRs	Total mapped reads to the genome.	NTRs per million mapped reads
Day 0	556,207	42,845,342	12,981.74
Day 2	1,074,476	42,203,140	25,459.62
Day 4	1,144,784	40,421,804	28,320.95
Day 6	1,007,134	40,218,029	25,041.85
Day 8	975,546	40,174,214	24,282.89

Table 8: Total Novel Transcribed Regions (NTRs) identified from each sample. Showing NTRs per million mapped reads normalizes the total NTRs to the sequencing depth.

Table 8 shows the total NTR counts for each sample, and the total NTRs per million mapped reads to normalize for sequencing depth. The main pattern which stands out from the above dataset is the increase of total NTRs from day 0 to day 2. This increase is maintained throughout the time-course. The increase of total NTRs from day 0 to day 2 is 196% while the increase from day 0 to day 8 is 187%. Almost doubled increase in NTRs in the differentiated cells fits with my hypothesis that

unbiased transcriptomic analysis of the trophoblast lineage would uncover greater novelty over the more extensively explored embryonic stem cell transcriptome.

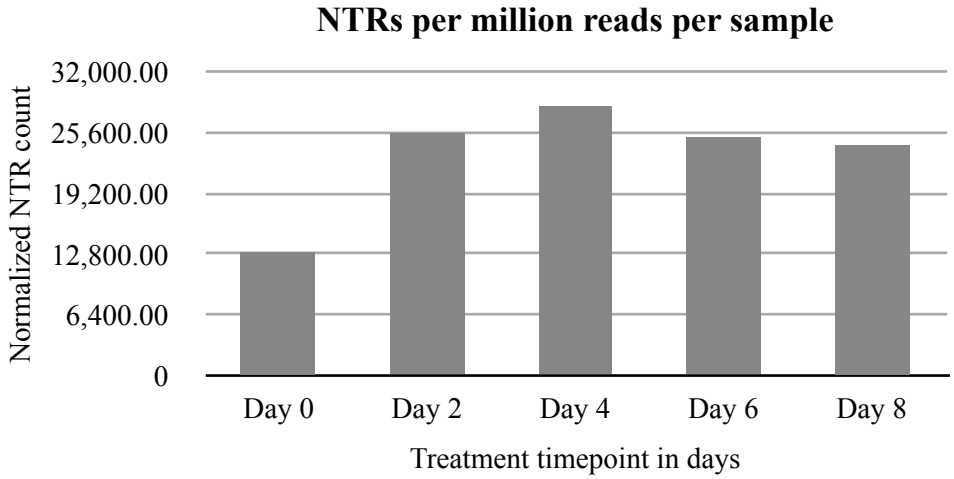


Figure 4.31: Distribution of NTRs per million reads during treatment.
There is a marked increase in NTRs during the initial stages of treatment (day 0 to day 8) and the total NTR number remains at elevated levels throughout the treatment.

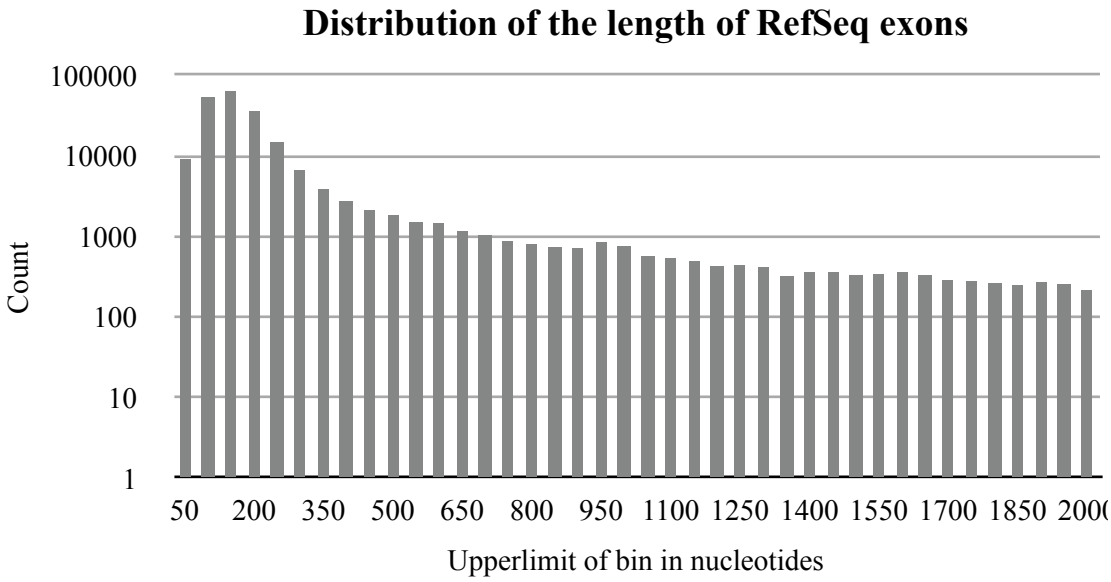


Figure 4.32: Distribution of the size of known exon from RefSeq.
The distribution is skewed to the left but the exon counts are maintained even beyond 1000 nucleotides. Note that the logarithmic scale is used for the Y axis and only exons less than 2000 nucleotides are represented in the histogram for clarity.

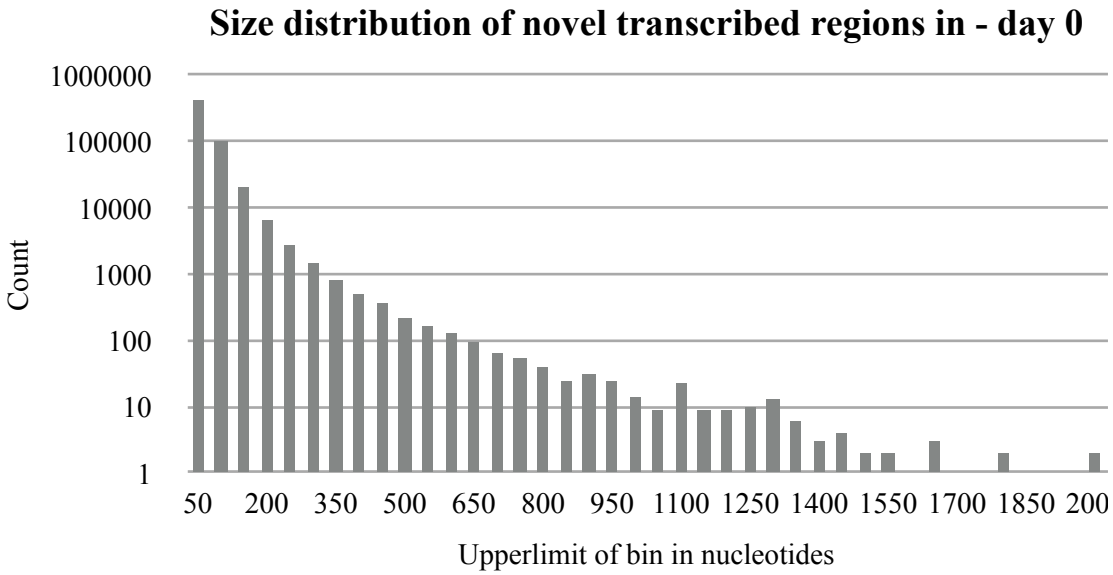


Figure 4.33: Distribution of size in novel transcribed regions in day 0.
The size distribution follows a logarithmic decrease as the size of the NTR increases. Note that the logarithmic scale is used for the Y axis. Clearly the NTR distribution differs from that of RefSeq exons.

Since NTRs could be assumed to be potential unannotated exons, one would assume that the NTR distribution would follow the same distribution as that of RefSeq exons. Therefore I first studied the distribution of RefSeq exon size (Figure 4.32), which peaks at around 150 - 250 nucleotide. While the RefSeq distribution is skewed to the left, there are a significant number of exons which are larger than 1000 nucleotides. This is significantly different from the length distribution of NTRs which peaks at the smallest size bin (Figure 4.33). The highest number of NTRs belong to the very short 0 to 50 nucleotide range and counts gets exponentially reduced as the NTR length increases. In contrast to the RefSeq known exon distribution which contains a considerable number of exons more than 1000 nucleotides long, there seem to be very few NTRs of that size. And there seems to be quite a high number of NTRs which are of smaller size (100 nucleotides or less).

Distribution pattern of NTR length remains unchanged during treatment, all time points share the same histogram shape. However, the rapid increase in NTRs due to treatment is reflected in the read counts, where all size bins show a considerable increase between day 0 and day 2.

Range	day 0	day 8	Fold Change (8D / 0D)	Increase %
0 - 50	419367	709193	1.69	169
50 - 100	102981	202784	1.97	197
100 - 150	20424	39536	1.94	194
150 - 200	6562	12033	1.83	183
200 - 250	2761	5109	1.85	185
250 - 300	1456	2498	1.72	172
300 - 350	827	1415	1.71	171
350 - 400	502	841	1.68	168
400 - 450	368	577	1.57	157
450 - 500	215	384	1.79	179
500 - 550	164	263	1.60	160
550 - 600	130	198	1.52	152
600 - 650	95	149	1.57	157
650 - 700	64	114	1.78	178
700 - 750	54	77	1.43	143
750 - 800	40	69	1.73	173
800 - 850	24	53	2.21	221
850 - 900	31	39	1.26	126
900 - 950	24	27	1.13	113
950 - 1000	14	22	1.57	157

Table 9: Comparison of NTRs of difference sizes between day 0 and day 8.
There is a clear up-regulation of NTRs on all size bands. Note that only NTRs of less than 1000 nucleotides are shown for clarity.

NTR Range	day 0	day 2	day 4	day 6	day 8
0 - 50	419367	777948	820404	720453	709193
50 - 100	102981	225709	243633	218154	202784
100 - 150	20424	44070	49552	42624	39536
150 - 200	6562	13412	15516	13019	12033
200 - 250	2761	5633	6427	5464	5109
250 - 300	1456	2799	3372	2756	2498
300 - 350	827	1561	1917	1537	1415
350 - 400	502	984	1156	921	841
400 - 450	368	602	739	602	577
450 - 500	215	418	507	388	384
500 - 550	164	318	350	266	263
550 - 600	130	231	264	221	198
600 - 650	95	166	200	153	149
650 - 700	64	115	152	107	114
700 - 750	54	89	100	92	77
750 - 800	40	84	98	58	69
800 - 850	24	56	71	61	53
850 - 900	31	44	51	41	39
900 - 950	24	38	36	38	27
950 - 1000	14	33	38	22	22
1000 - 1050	9	21	31	27	22
1050 - 1100	23	25	21	24	25
1100 - 1150	9	13	22	18	16
1150 - 1200	9	11	26	15	20
1200 - 1250	10	12	13	9	12
1250 - 1300	13	10	14	11	10
1300 - 1350	6	12	8	9	13
1350 - 1400	3	6	6	8	11
1400 - 1450	4	4	5	7	11
1450 - 1500	2	11	11	1	5
1500 - 1550	2	5	8	5	4
1550 - 1600	0	3	4	4	0
1600 - 1650	3	7	5	1	1
1650 - 1700	0	2	2	1	0
1700 - 1750	1	3	3	0	2
1750 - 1800	2	3	1	2	2
1800 - 1850	1	0	2	1	1
1850 - 1900	0	2	3	1	2
1900 - 1950	1	1	0	0	0
1950 - 2000	2	3	0	2	2

Table 10: NTR counts of all the treatments divided into size bands of 50 nucleotides.

For clarity only NTRs less than 2000 nucleotides are shown. As observed in the total NTR counts, there is a marked increase in NTR in day 2 compared to day 0.

During the discovery of novel transcribed regions, it was believed that most NTRs would be either new exons of known genes, or exons from totally novel genes. While NTR counts and the distribution of length of NTRs support the potential existence of new transcripts and new exons, the presence of large numbers of very small NTRs appeared to be a mystery.

4.14 Identification of Novel transcripts

The potential for identification of novel transcripts, presumably found in a subset of the NTRs, was one of the reasons RNA-Seq was applied in this study. The novel nature of the cell type caused by the SU5402+BMP4 differentiation creates the possibility of identifying new genes / exons which have not been described previously.

The strategy I used to identify novel transcripts from NTR data is fully described in the materials and methods section. Briefly, to be identified as a cluster of exons contributing to a new gene, the NTRs had to be significantly expressed (on average 5 reads per base) and exist away from any know exon / gene footprint but significantly close with each other (less than 10,000 nucleotides). Samples from Day 0 representing human embryonic stem cells and Day 8 representing the most differentiated time-point were used for the novel transcript discovery.

The novel transcript discovery pipeline identified 741 potential novel transcripts in Day 0 and 701 potential novel transcripts in Day 8. Out of these 367 were present in both the undifferentiated hESC and the 8 day differentiated trophoblast.

To further study theses novel transcripts their distribution of average exon length was observed. It became evident that the majority of potentially novel transcripts had an average exon length less than the 110 - 120 nucleotides found to be the average size RefSeq exons. This potential novel transcript exon size distribution was similar to the length distribution in the total NTR set where the majority of NTRs were less than 100 nucleotides.

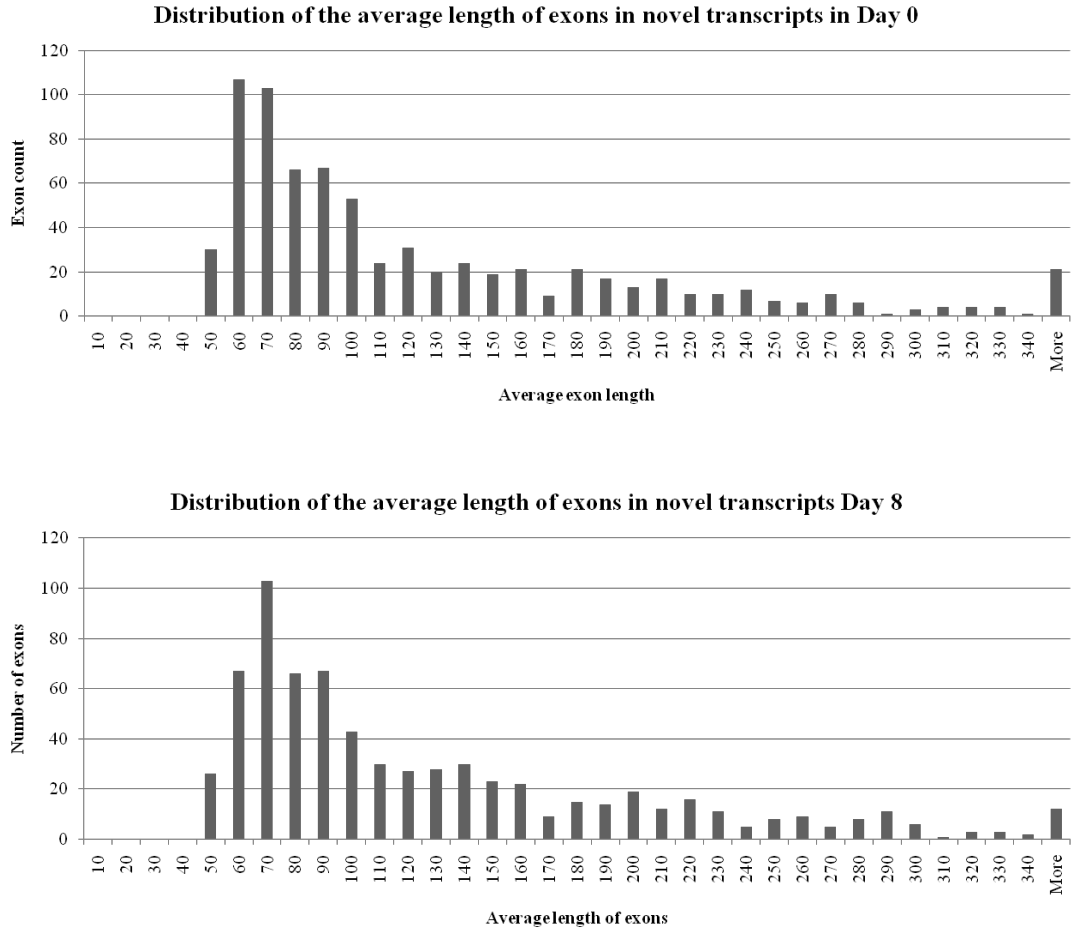
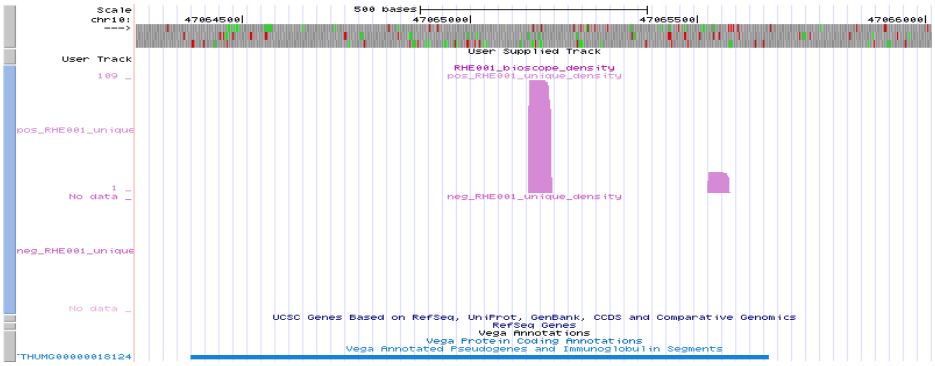


Figure 4.34: The distribution of the average exon length of the potential novel transcripts.
The distribution is skewed showing a bias towards exon lengths less than 100.

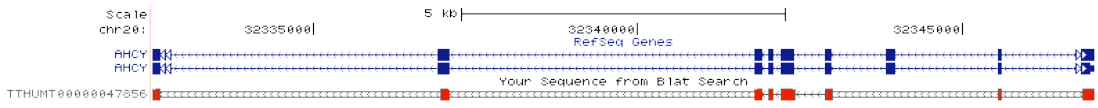
4.14.1 Interference of the novel transcript discovery by processed pseudogenes

While looking at the footprints of the potentially novel transcripts, it was observed that a considerable number of them had overlaps with footprints with processed pseudogenes. Though one hypothesis could be that these represented expressed processed pseudogenes, this should not be the case as the definition of a processed pseudogene is that it contains no introns. Thus there should not be multiple peaks but only a single peak detected from a processed pseudogene; I detected multiple, small (50-60 base window) peaks. A sequence search indicated that the footprint of these peaks were the same as the exon - exon junctions of the parental gene from where the pseudogene originates. This implies that these reads are actually from the parental gene transcript but they get mapped to the processed pseudogene than the actual exon - exon junction as the aligner favors alignment without gaps like in the intron between the exons. This leads to the creation of small peaks outside the footprints of known genes and located in pseudogene regions, which are then (incorrectly) identified as NTRs. This explains the unexpected high number of short NTRs observed. Since the processed pseudogene contains sequences to all the exon - exon junctions, these small RNAs exists as groups representing all the exon - exon junctions of the active transcript, thereby falsely showing as a new transcript. For an example please see the explanation below.

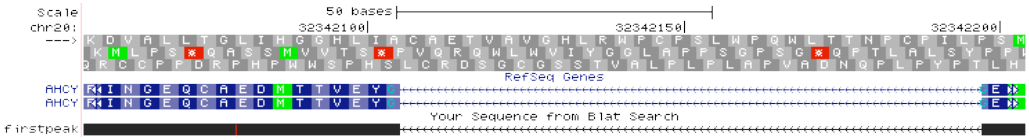
1. Observation: Presence of short peaks (less than ~100 nucleotides), often in groups (mostly two but can be more). In most cases (such as in this) they are on an annotated processed pseudogene footprint shown in blue.



2. When a sequence search of the annotated processed pseudogene (in blue) is done using a UCSC blat search, the second hit is the gene *AHCY*, which is the active counterpart of the pseudogene. (The first hit is the processed gene itself). Visualization of the blat result is shown below. The first track is the RefSeq annotation and the blat alignment of the processed pseudogene is shown as the second track. Note the alignment with only the exons - a characteristic of a processed pseudogene.



3. When the sequence of the footprint of the first peak is used to search the human genome (same as above) it results in the following location. The sequence match is shown in black. Note that the sequence of the small peak footprint is identical to a junction region of the *AHCY* gene. Therefore the conclusion is that the small peaks in clusters originate from the junction reads of active genes.



This observation clearly explains the high number of short NTRs which were then removed from the dataset and the novel transcript identification was repeated.

It has been reported that LINE-1 repeat elements have the ability to create pseudogenes (Esnault, Maestre et al. 2000) and that they are active in human ES cells (Garcia-Perez, Marchetto et al. 2007). Therefore the 'noise' created by pseudogenes in RNA-Seq experiments should be monitored and removed. This observation would be quite useful for the RNA-Seq community as it introduces a source of false positives in an RNA-Seq experiment and because it leads to an under estimation of read counts of genes which have pseudogenes, by taking away their junction reads.

Even after removing the small NTRs, all the patterns of NTR such as the marked up-regulation of NTRs during the start of the differentiation remains the same.

4.14.2 Novel transcripts discovered from RNA-Seq data after removing interferences by pseudogenes

Identification of the above mentioned phenomenon which created false positive peaks due to exon - exon junctions was a disappointment as it brought down the total number of novel transcribed regions and thereby novel transcripts in each sample. Despite this, after removing NTRs which are less than 120 nucleotides in length (the ones which are most likely be mapped to exon-exon junctions of pseudogenes) and re-running the novel transcript discovery pipeline, 260 potentially novel transcripts from day 0 and 272 transcripts from Day 8 were identified. A subset of these were validated by PCR, cloned and sequenced. Some of the examples are described below.

4.15 Examples of identified and validated novel transcripts

The sequences of the novel transcripts 1 - 8, which were obtained by PCR and cloning are given in appendix II.

4.15.1 Novel transcript 1 (chr1:63,559,143 - 63, 560, 695)

As can be seen in Figure 4.35, there is a novel multi-exonic gene which overlaps *FOXD3* gene, and is coded by the opposite strand. There are no RefSeq or UCSC annotations describing it. However to support the above observation there are valid split ESTs (one originating from ES cells) with a shared footprint. FOXD3, being a major pluripotency factor, is inhibited immediately upon treatment, and interestingly the novel gene has exactly the same expression pattern, suggesting that there may be a functional relationship and co-regulation, between the two, potentially through a bidirectional promoter.

This novel transcript does not have a valid open reading frame starting form AUG (but does have a 405 nucleotide coding sequence beginning from UUG). Therefore it is most likely a non-coding transcript. Novel transcript 1 was validated by PCR, cloning and sequencing.

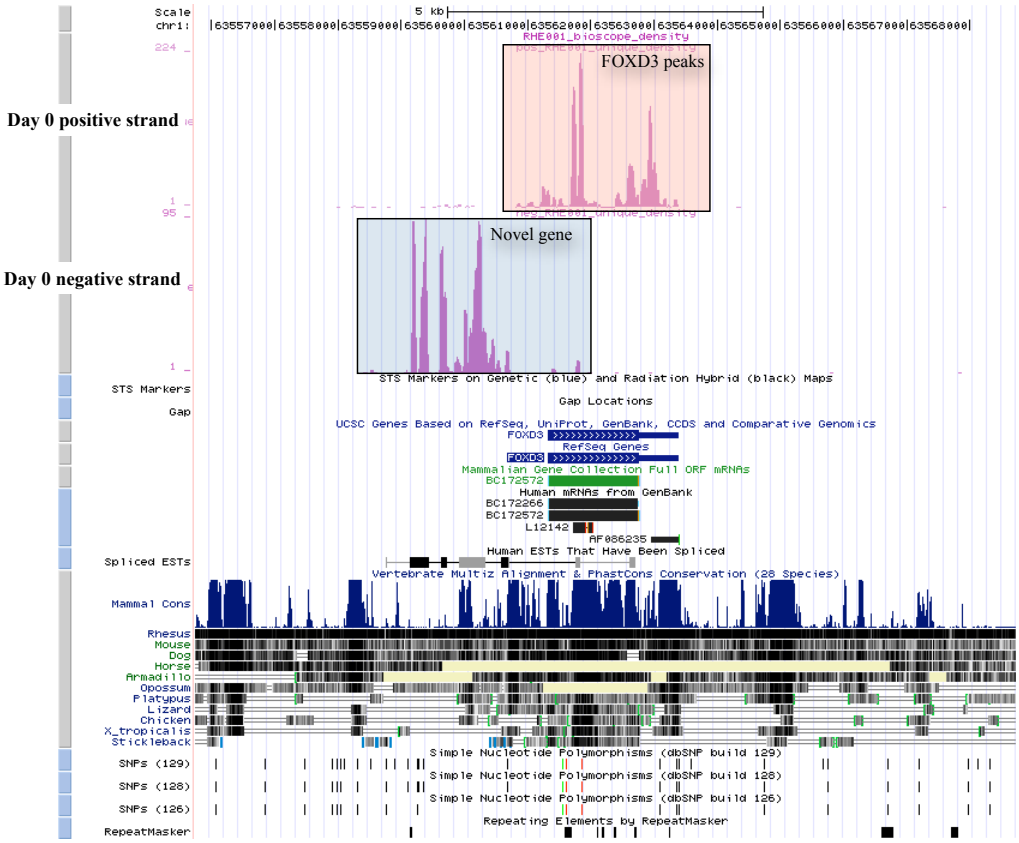


Figure 4.35: The novel gene next to *FOXD3*.
The peaks belonging to FOXD3 is shown highlighted in red while the peaks of the novel gene is shown in blue.

4.15.2 Novel transcript 2 (chr7:100,729,591-100,731,304)

This new transcript was identified in the undifferentiated hESC sample. It is down-regulated immediately upon differentiation (maximum peak height of 105 in day 0 goes down to 2 in day 2). There is a LINE element which has an overlap with this transcript thus suggesting that this transcript originated from a LINE insertion. This has an open reading frame of 447 nucleotides. Novel transcript 2 was validated using PCR, cloning and sequencing.

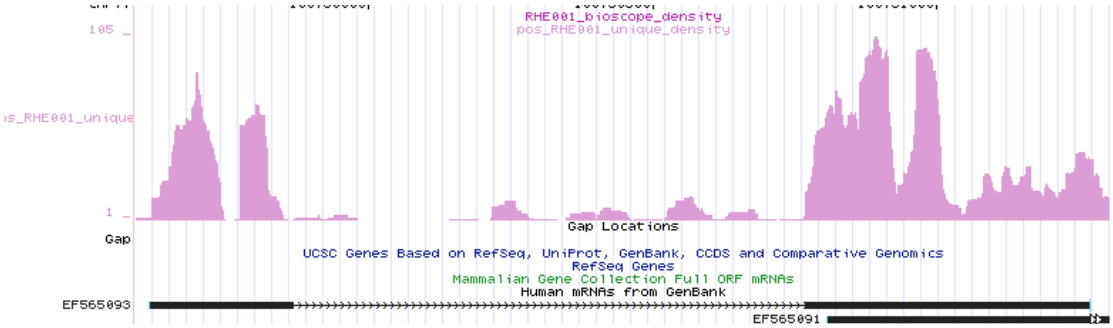


Figure 4.36: RNA-Seq peak profile of the novel transcript 2.
Only the positive strand of Day 0 is shown for clarity.

4.15.3 Novel transcript 3 (chr7:100,738,332-100,740,838)

This novel transcript shows a down-regulation during the trophoblast differentiation but does not show a rapid suppression upon treatment compared to novel transcript 2. It has a maximum peak height of 213 in day 0, which goes down to 14 in day 8. This does not have a RefSeq annotation, and UCSC only has a putative annotation, and reports that it is from an IMAGE clone. It should be noted that based on RNA-Seq data there appears to be an additional exon on the 3' side of the transcript. The transcript does have an open reading frame and appears to be originating from a LINE insertion based on the LINE sequences found on the base of both exons. This novel transcript has been validated by PCR, cloning and sequencing.

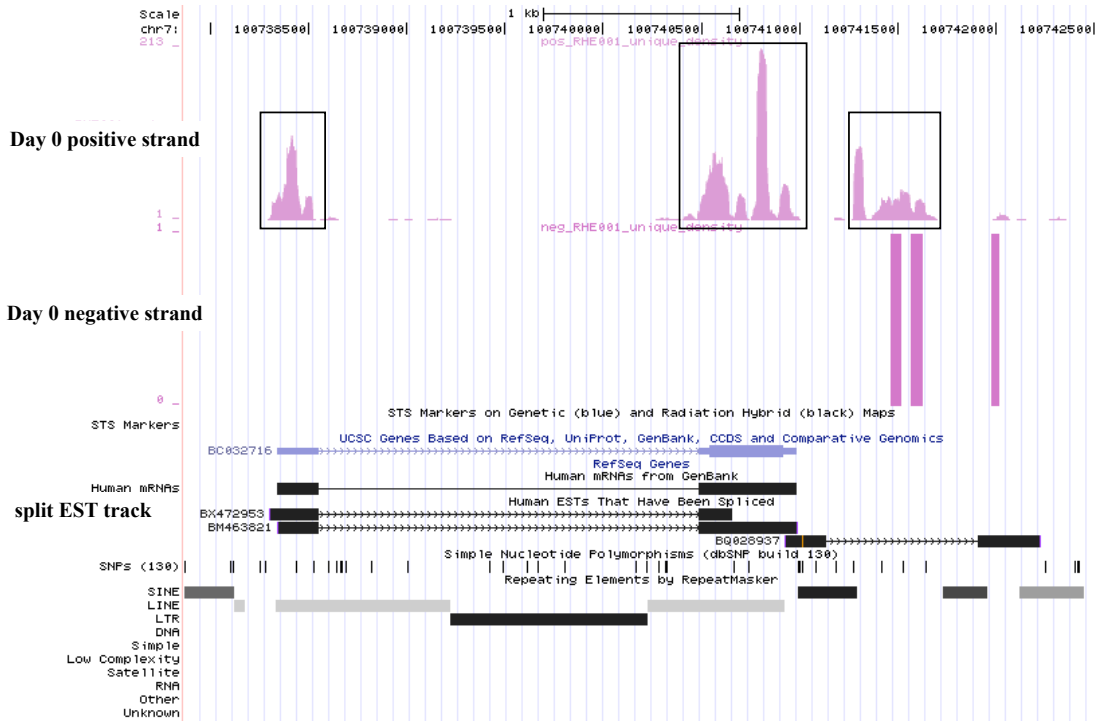


Figure 4.37: RNA-Seq profile of the novel transcript 3.
The three potential exons are shown in boxes. There is no RefSeq transcript for this, but there is an incomplete UCSC annotation which excludes the last exon on the 3' end.

4.15.4 Novel transcript 4 (chr17:34,456,005-34,462,831)

This novel transcript is human ES specific. It does not have a RefSeq annotation, and has only an incorrect UCSC annotation. The footprint of this transcript is supported by split ESTs. The 5' exon appears to be originating from an LTR region.

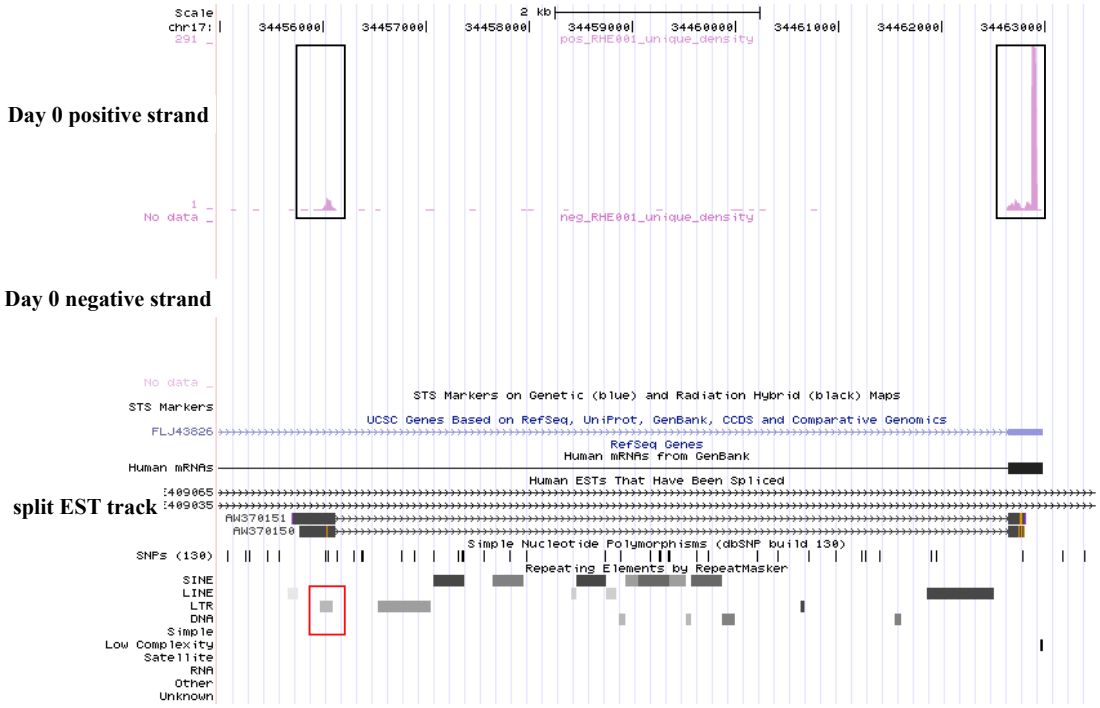


Figure 4.38: RNA-Seq peak profile of the novel transcript 4.
The two exons are enclosed in black boxes and the LTR region which overlaps the first exon is shown within the red box.

4.15.5 Novel transcript 5 (chr19:44,838,393-44,843,124)

This transcript originally did not have a RefSeq annotation or a UCSC annotation. Its existence is supported by split - human ESTs with placental origins. This transcript is expressed only at later time points (day 6 and day 8) during differentiation implying that it might be important in trophoblast differentiation. PCR validation and sequencing proved the existence of the four exons. The latest version of UCSC browser shows this gene as LGALS16 supported by a publication (Than, Romero et al. 2009) which reports that its placenta specific. While this takes away the novelty of this transcript, this proves the effectivity of the differentiation protocol for inducing this transcript and the transcript detection pipeline for identifying it.

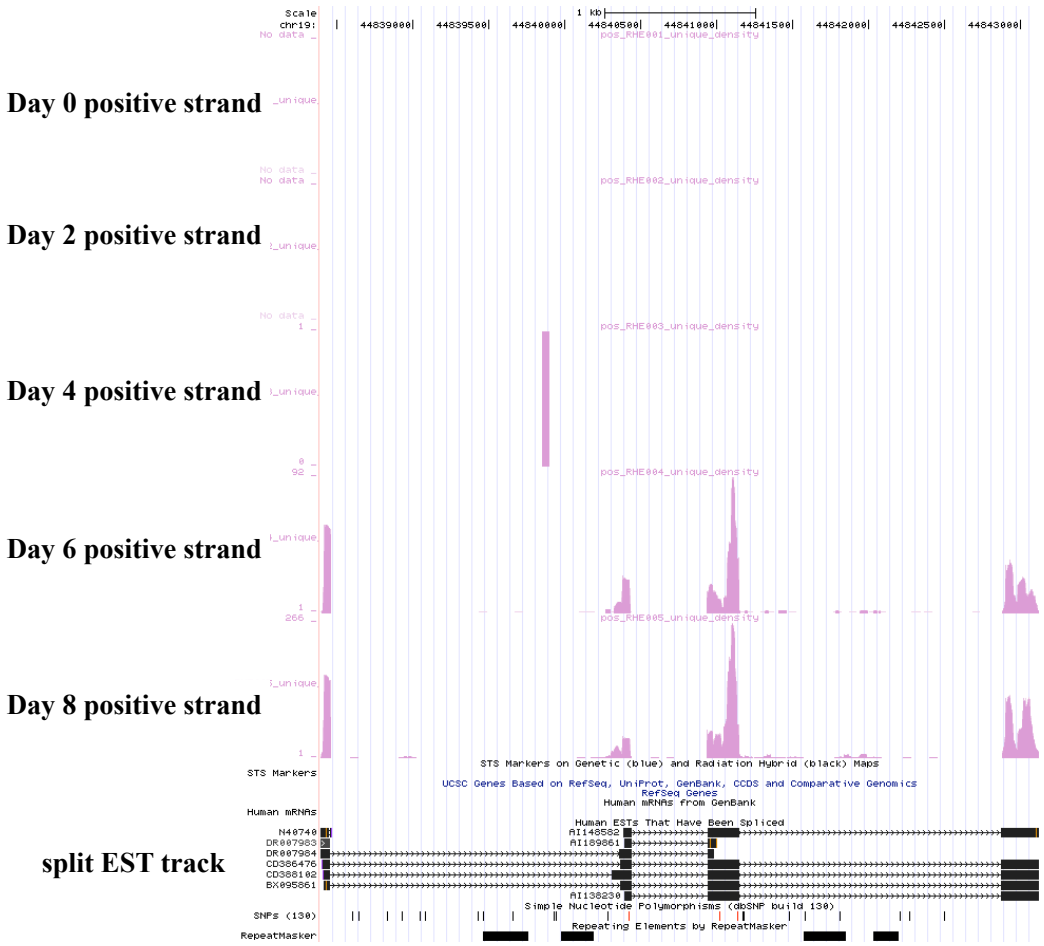


Figure 4.39: The UCSC view of the novel transcript 5.
The expression begins late at day 6. There are split ESTs supporting it.

4.15.6 Novel transcript 6 (chr13:99,536,264-99,539,117)

This transcript is expressed during the entire differentiation but shows a significant down-regulation through the course of differentiation. It does not have RefSeq or UCSC annotations, but is supported by split ESTs.

The gene *PCCA* (Propionyl CoA carboxylase, alpha polypeptide) is just next to this transcript and is coded by the opposite strand, thereby suggesting that both transcripts could be regulated by a bi-directional promoter. *PCCA* has a similar expression pattern to this novel transcript. Novel transcript 6 was validated using PCR, cloning and sequencing.

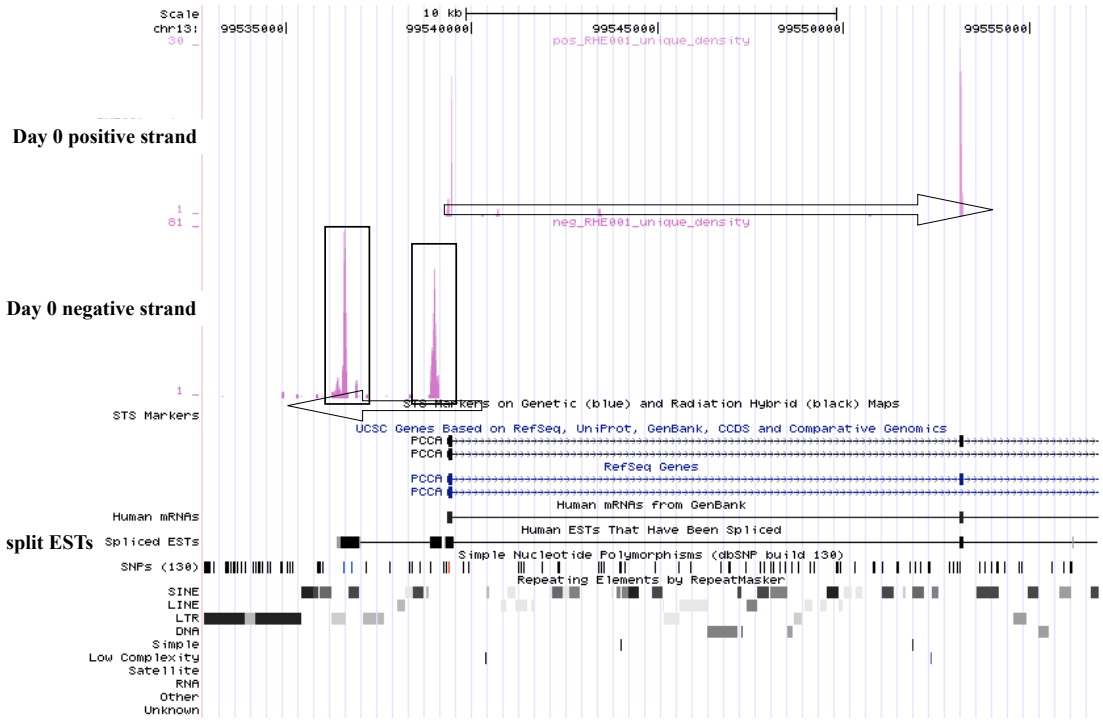


Figure 4.40: UCSC view of the novel transcript 6.
Note that the novel transcript is expressed from the negative strand (right to left), while the *PCCA* gene is expressed from the positive strand (left to right). The proximity of these two transcripts, common expression pattern and their orientation suggests the regulation through a bi-directional promoter.

4.15.7 Novel transcript 7 (chr13:90,577,939-90,644,334)

This transcript is specific for the day 8 time-point of trophoblast development. Its footprint is quite long, and does not have RefSeq or UCSC annotations. However it is supported by EST data including one originating from embryonic trophoblast. Expression of this transcript begins at day 4 at a maximum peak height of 34 and increase up to 140 in day 8. Novel transcript 7 was validated using PCR, cloning and sequencing.

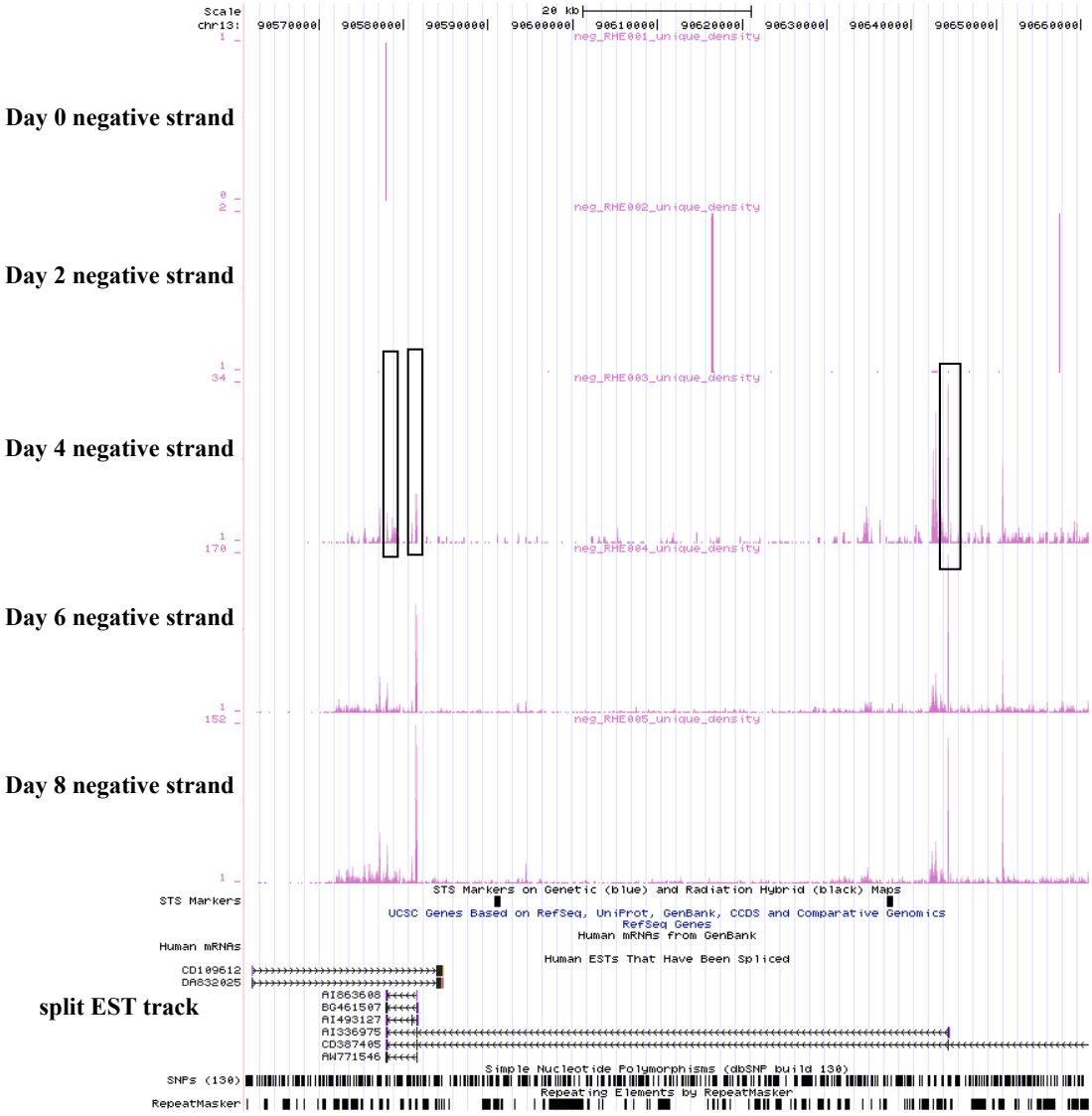


Figure 4.41: RNA-Seq peak profile of the novel transcript 8.
This transcript gets up-regulated during trophoblast differentiation. Three exons which make up the transcript as identified by PCR and sequencing is shown within the boxes. Additional peaks in the surrounding suggest that there could be additional transcripts originating from this locus.

4.15.8 Novel transcript 8 (chr10:54,432,626-54,459,840)

This transcript is composed of three exons and has no RefSeq or UCSC annotations. Based on the RNA-Seq data it is up-regulated throughout the differentiation. Its existence is supported by a human EST which has a fetal origin. Expression of this transcript starts at day 2 and gets up-regulated during the course of the differentiation. There is no clear open reading frame. Novel transcript 8 was validated using PCR, cloning and sequencing.

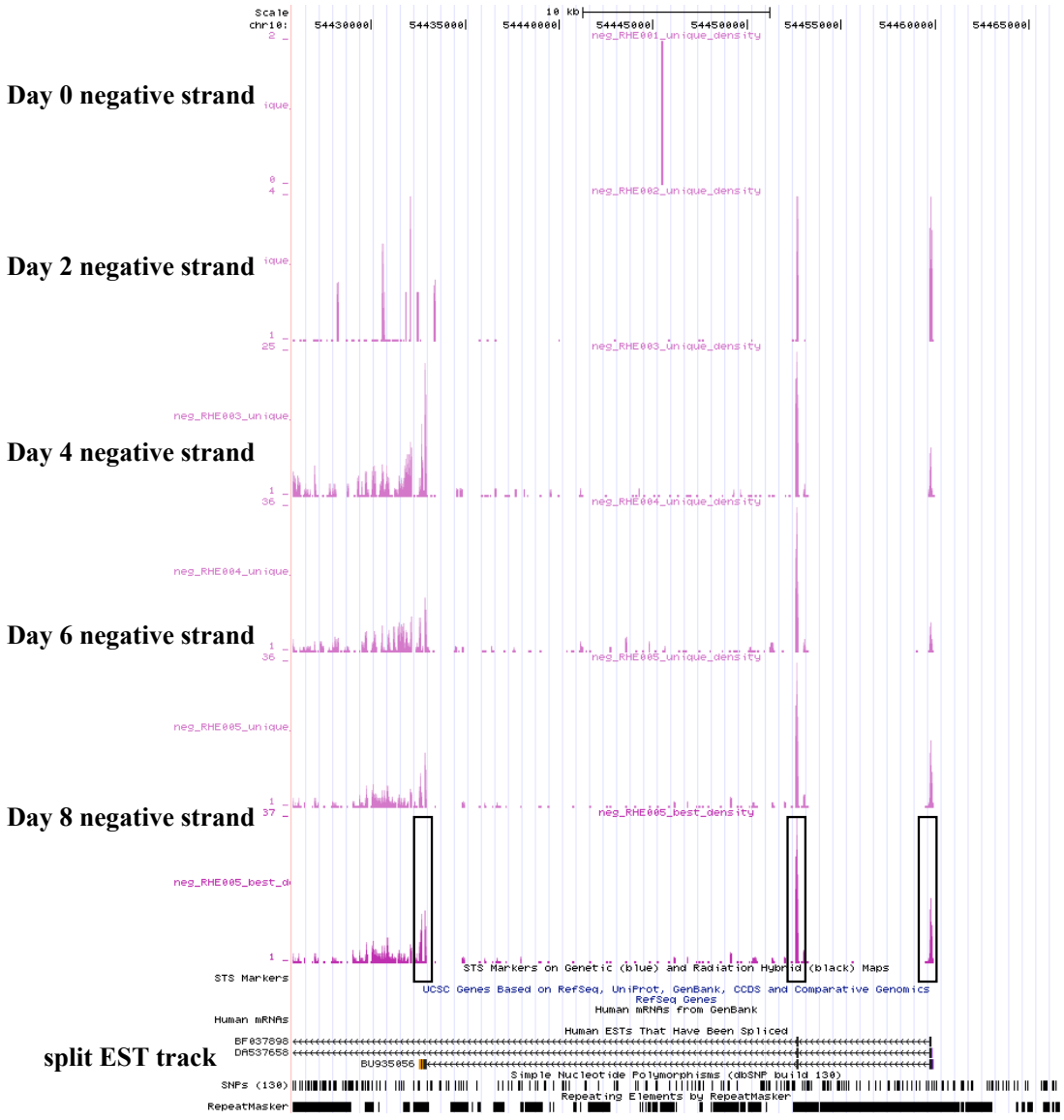


Figure 4.42: RNA-Seq peak profile of novel transcript 8.

4.15.9 A cluster of new transcripts (chr7:100,728,243 - 100,742,923)

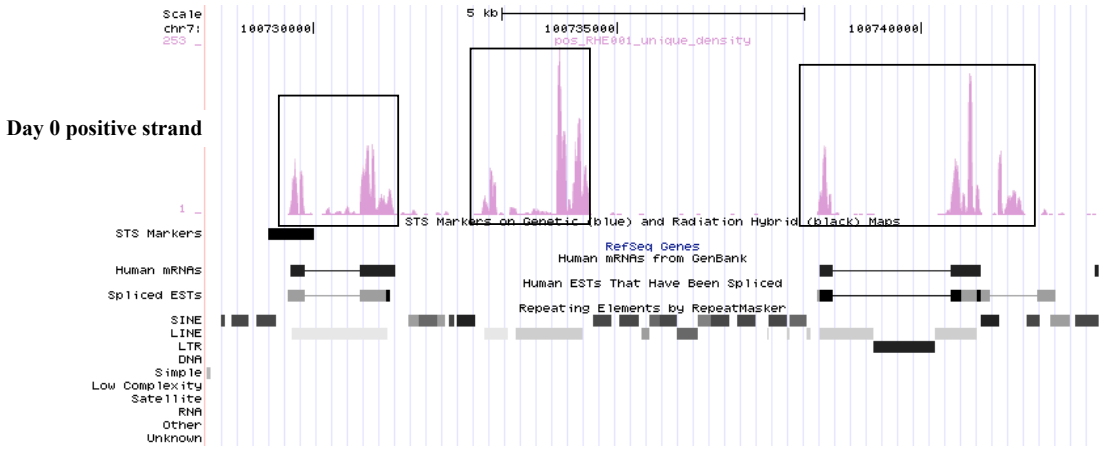


Figure 4.43: A cluster of novel transcripts identified by RNA-Seq.

A cluster of three novel transcripts expressed at day 0. They do not have RefSeq annotations, and only the transcript on the extreme left has a partial UCSC annotation. This cluster was validated using PCR, cloning and sequencing.

4. 16 Expression of retroviral related elements in the genome during trophoblast differentiation

Even after the removal of small NTRs formed by junction reads of active genes (i.e. exon exon junctions of processed pseudogenes), the number of NTRs present in the sample were still considerable. When NTRs longer than 150 nucleotides are considered, there are 6,562 NTRs in day 0 and 12,033 NTRs in day 8. Among these, 3,151 NTRs in day 0 and 2,976 NTRs in day 8 contribute to potential novel transcripts. This leaves 3,411 NTRs in day 0 and 9,057 in day 8 unaccounted for. The significant increase (almost threefold) of NTRs from day 0 to day 8 suggest that these NTRs might serve a biological purpose. To study this, the locations of these NTRs were analyzed. From an initial manual analysis of some of these NTRs it became apparent that many of these were derived from short interspersed nuclear elements (SINEs), long interspersed nuclear elements (LINEs), and long terminal repeat (LTR) elements of the genome, in other words, the "dark matter" of the genome. Indeed, it turns out that a majority (Table 11) of these NTRs were derived from these repetitive elements.

To study the expression of these elements during differentiation, all SINE, LINE and LTR elements which do not fall within any known RefSeq footprint was identified and their expression based on RNA-Seq read counts were analyzed. The following table shows the summary of the counts. Based on the read counts it is clear that both SINE and LINE elements show an increase during differentiation.

Repeat Type	Day 0	Day 2	Day 4	Day 6	Day 8
SINE	315	674(684)	734(777)	656(701)	557(593)
LTR	269	269(284)	261(276)	247(263)	237(252)
LINE	241	459(465)	558(590)	449(478)	417(444)

Table 11: The number of SINE / LINE / LTR elements which show expression during day 0 - day 8 based on uniquely mapping reads.
An element is considered expressed only if on average it has a read count of 4 per base. The number enclosed in brackets are values normalized for the sequencing depth.

Since SINE, LINE and LTR elements all have different subcategories, their expression dynamics were further analyzed to see if there was any sub-type specific expression. The tables containing the number of expressed elements belonging to a particular category and graphs showing their expression pattern are shown in Figures 4.44, 4.45 and 4.46. Overall, this shows that the trophoblast differentiation causes a clear increase in all sub categories of SINE and LINE elements and the highest increase is between day 0 and day 2 - the start of the treatment. As far as the LTR elements are concerned, despite the reduction in total expressed elements, three out of four subcategories - namely ERVL, ERVK and MaLR show a distinct increase in expressed elements at the start of differentiation, and the increase is maintained throughout the differentiation.

4.16.1 Specificity of reads mapping to the repeat elements

The notion that all the repeat regions in the genome have similar sequences and therefore are unable to provide unique read mapping surfaces is untrue. Detection by hybridization is problematic but sequence-based detection is possible. While the repeat regions originally inherits a particular primary structure based on it's type and

class, it gets rapidly altered due to point mutations. And since different repeat regions acquire different mutations their sequences become unique. however it should be kept in mind that this process requires time and that the most recently integrated repeats would not show a sequence diversity as shown by the more mature ones.

RNA-Seq has been designed from ground up to identify all expressed regions including ones that arise from repeat regions while preventing non-specific binding. Firstly the reads which map to more than one location with the same score are discarded and not used in counting. These discarded reads could come from expressed repeat regions which have not yet accumulated enough point mutations to become truly unique. Secondly a read is considered to be uniquely aligned only if the difference between it's best alignment score and the second best is more than four. This too prevents non specific binding.

All the data on the expressed repeat regions reported in this thesis have been obtained by the same alignment criteria used for the rest of the genome (as described in the methods section and as highlighted above). This results in the rejection of large number of reads arising from expressed regions as seen by the huge increase of reads in the multi mapped track compared to that of the uniquely mapped. Therefore the extensive expression of repeat regions reported in this thesis is not a result of mis aligned reads and in fact the reported repeat expression is an underestimate of what actually is due to the large number of reads lost due to the stringent alignment process.

To quantitatively show the above mentioned point the reads mapping to introns and coding sequences were compared with the reads mapping to repeat regions. The data is summarizes in table 12.

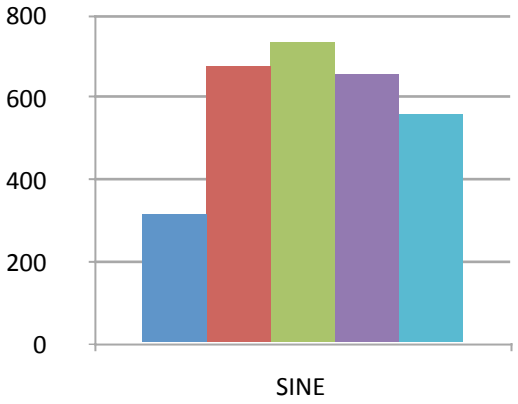
Description	# of regions	Average reads per base
Introns of genes with one known isoform	118432	**0.07**
SINE , LINE , LTR elements with at least 4 reads per base	1211	**10.4**
All RefSeq exons with at least 4 reads per base	4079	**10.1**
NTRs with 5 reads per base with 60 bases or more overlap	1530	**16.0**

Table 12: Statistics of reads mapping to repeat elements, introns and exons of day 8 sample.

As can be seen in the table the average reads mapping per base in known eons and the repeat regions are almost the same while the reads mapping to the introns is negligible. Looking at the novel transcribed regions which arise from a repeat region and go beyond it's footprint, the average reads per base number goes higher even than that of known exons to 16.

This clearly demonstrates that the reported expressed regions are indeed real and that the cutoff used to identify the expressed ones is comparable with that of known exons.

Expression of SINE elements during trophoblast differentiation



Repeat Typ	Day 0	Day 2	Day 4	Day 6	Day 8
MIR	265	586	647	572	12
Alu	50	88	87	84	12

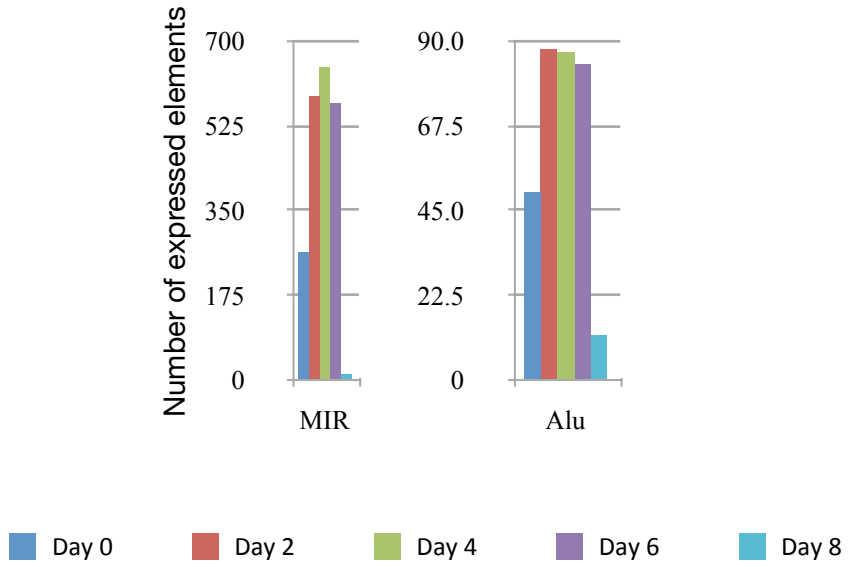
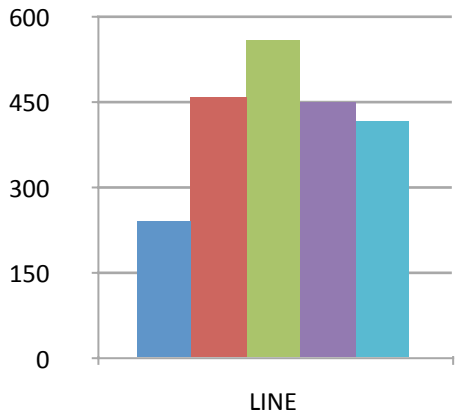


Figure 4.44: Number of expressed SINE elements during trophoblast differentiation.

Expression of LINE elements during trophoblast differentiation



Repeat Type	Day 0	Day 2	Day 4	Day 6	Day 8
L2	128	250	317	230	225
L1	85	151	187	158	134
CR1	21	41	37	45	44
RTE	7	17	17	16	14

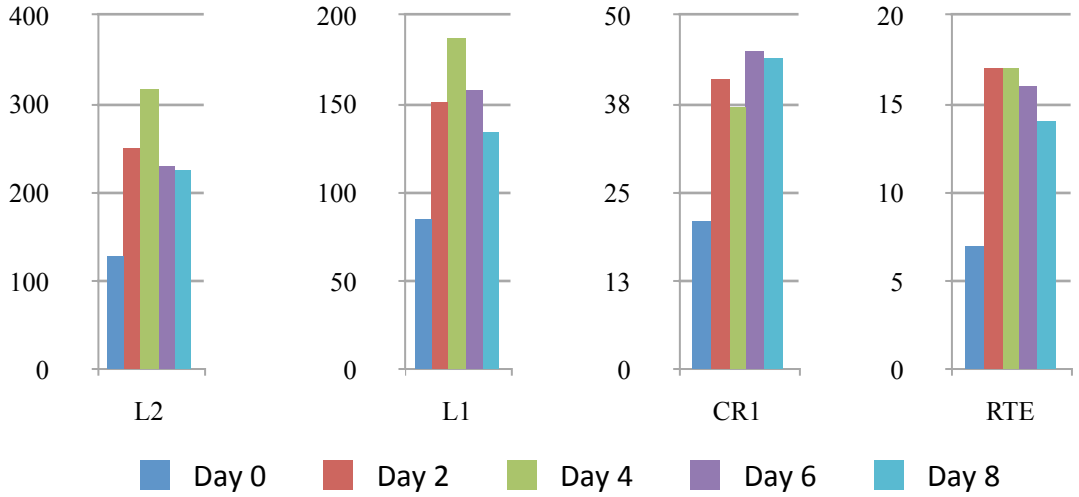
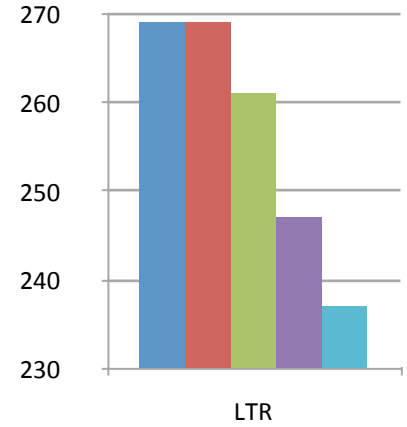


Figure 4.45: Number of LINE elements expressed during trophoblast differentiation.

Expression of LTR elements during trophoblast differentiation



Repeat Type	Day 0	Day 2	Day 4	Day 6	Day 8
ERV1	171	100	95	104	107
MaLR	66	103	98	84	69
ERVL	29	48	47	37	43
ERVK	3	17	20	21	17
ERV	0	1	1	1	1

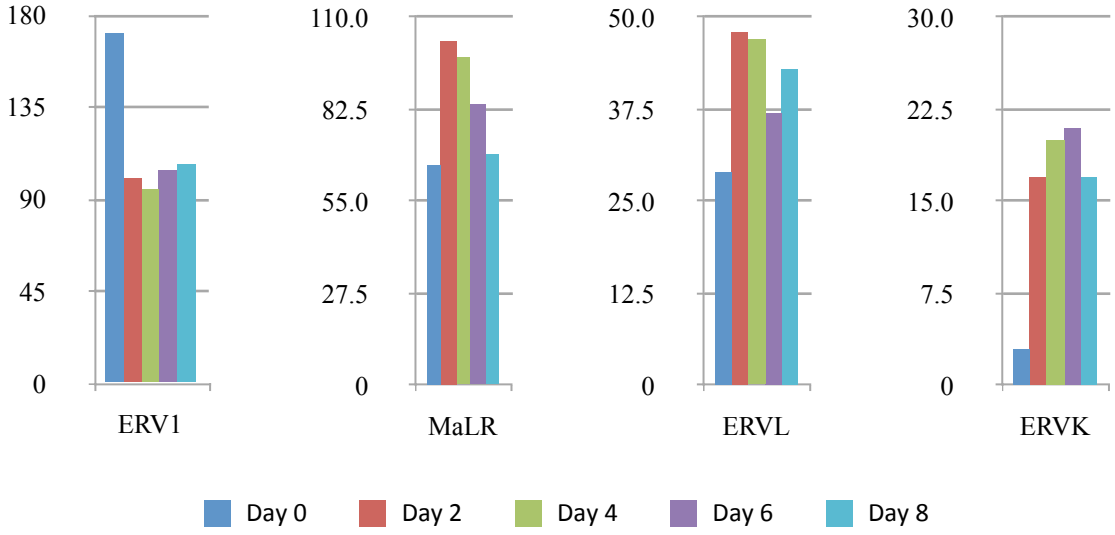


Figure 4.46: Number of LTR regions expressed during trophoblast differentiation.

4.17 Distribution of expressed repeat regions in the genome

Once it was observed that the SINE, LINE and LTR elements are expressed throughout differentiation the location of these expressed elements were studied to see if they are spread throughout the genome or localized to a particular area. A Circos diagram (http://mkweb.bcgsc.ca/circos/) (Krzywinski, Schein et al. 2009) (Figure 4.47) shows that the expressed SINE, LINE and LTR elements are found everywhere in the genome and that they show considerable differential expression.

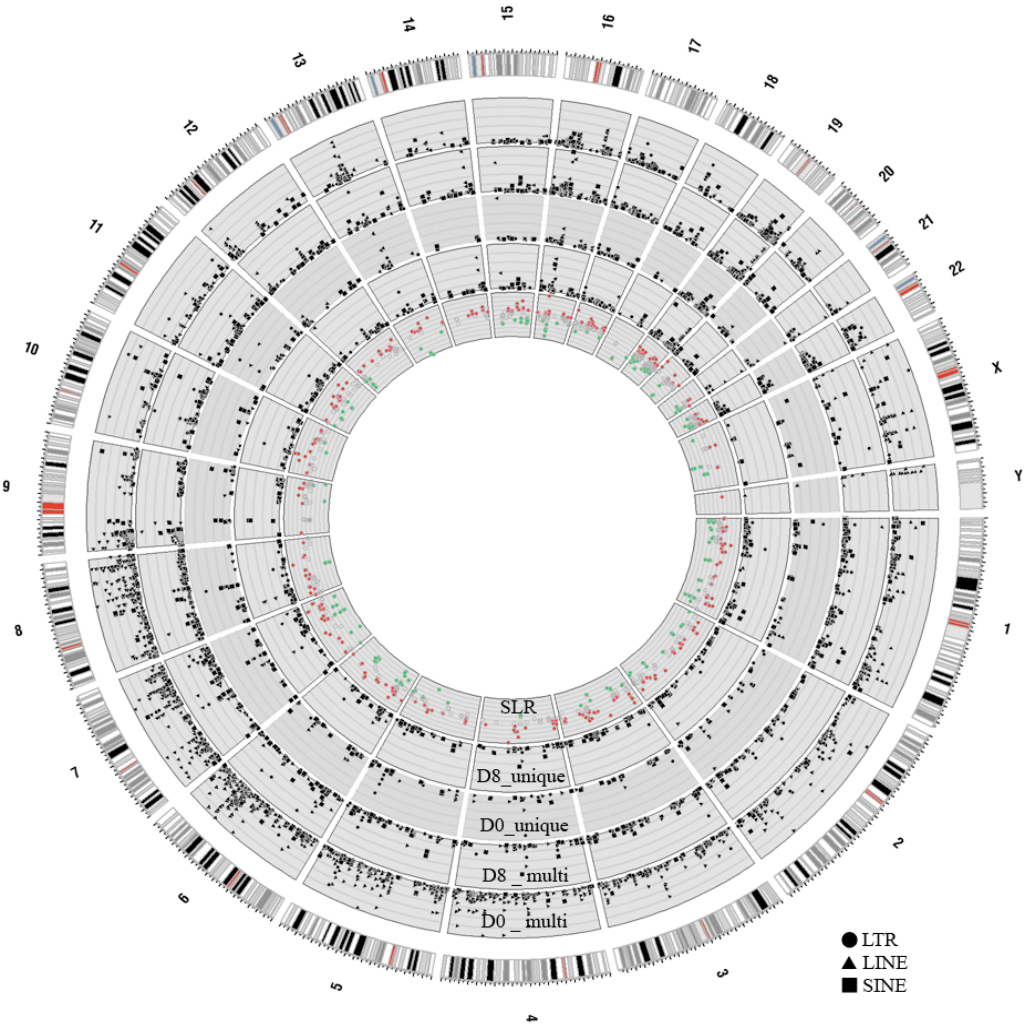


Figure 4.47: A circular chromosomal image (generated by circos software) showing the expression of LINE, SINE and LTR elements.
The chromosomes of the genome is shown in each of the segments of the outermost circle. Each concentric circle is a scatterplot showing the expression level of the particular element in the y axis. The two outer most rings / charts show the expression based on the multi map read counts, while the next two tracks show the expression based on the unique read counts. The innermost track shows the differential expression based on the unique counts (Red - Up-regulated, Green - Down-regulated). The objective of this diagram is to show that the expression of SINE, LINE and LTR elements are widespread throughout the genome.

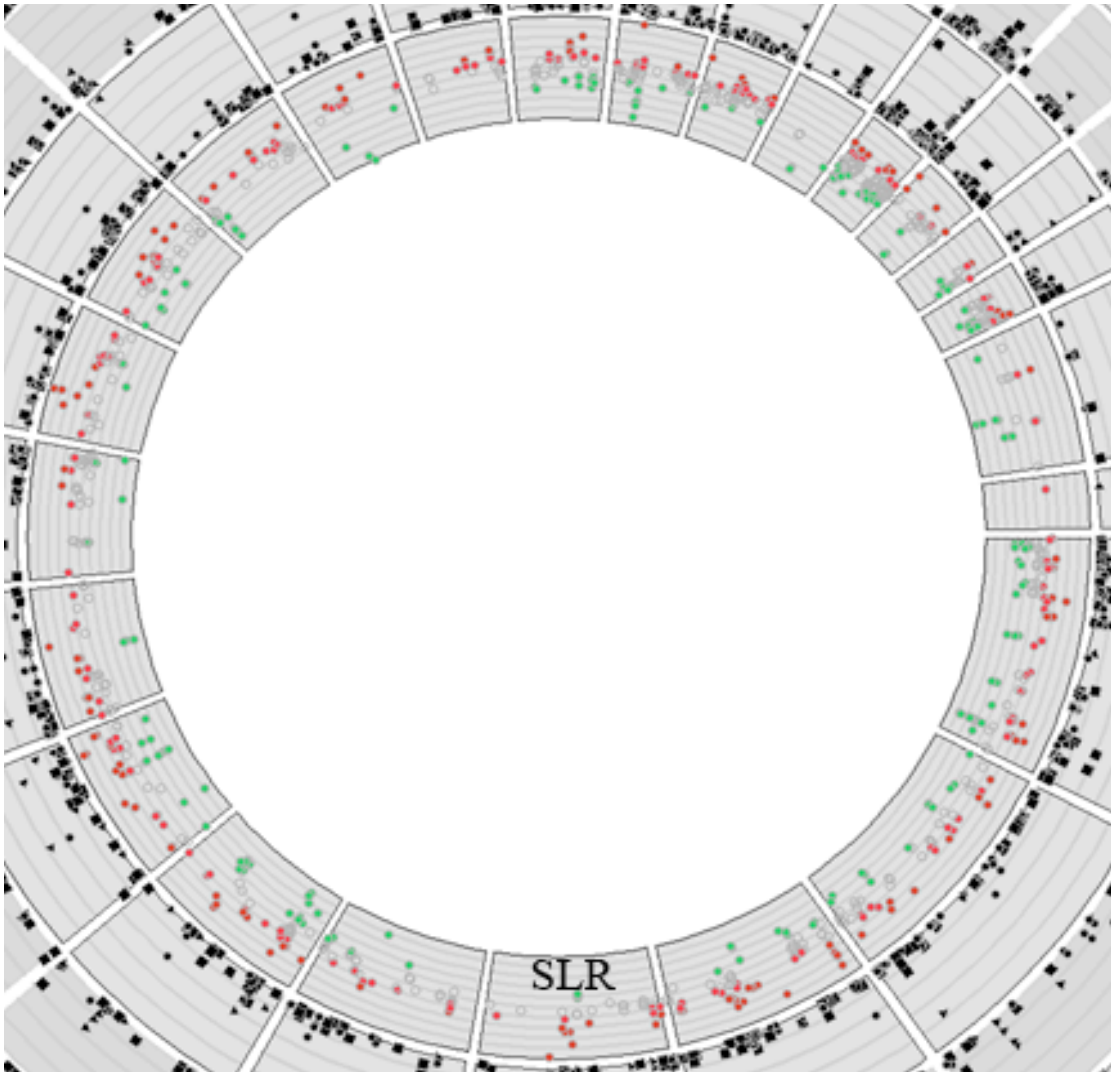


Figure 4.48: The track showing differential expression of the repeat elements (day 0 vs day 8).
This is the SLR (Signal log ratio, which is the $\log_2$ of fold change) track from the previous figure. The red markers showing up-regulation and green showing down-regulation shows clearly that the repeat elements are differentially expressed during differentiation.

4.18 Retroviral elements acting as new exons of known transcripts during trophoblast differentiation

To gauge the influence of retroviral elements in gene expression, a proximity study of the expressed retroviral elements to known genes was carried out. A retroviral element was considered as a potential new exon of a gene if it was found to be within 10, 000 bases of a gene footprint. There were 86 such cases in day 0 and 259 cases in day 8. This threefold increase from day 0 to day 8 further suggests that LTR elements do have a significant biological role in trophoblast differentiation.

4.18.1 *CLDN4* (Claudin 4)

CLDN4 shows a novel exon on its 5' side which has an annotated LTR footprint from the ERV1 subfamily. The new exon is shown in panel 1 of Figure 4.49 and is enlarged in panel 2 to show its LTR footprint. The additional exon is not expressed in day 0, even though the *CLDN4* gene itself is expressed. The new exon is induced by the differentiation, reaching its maxima in day 8. Another novel exon just next to the *CLDN4* extension can also be seen. The novel exon and the extension has been validated by PCR, cloning and sequencing.

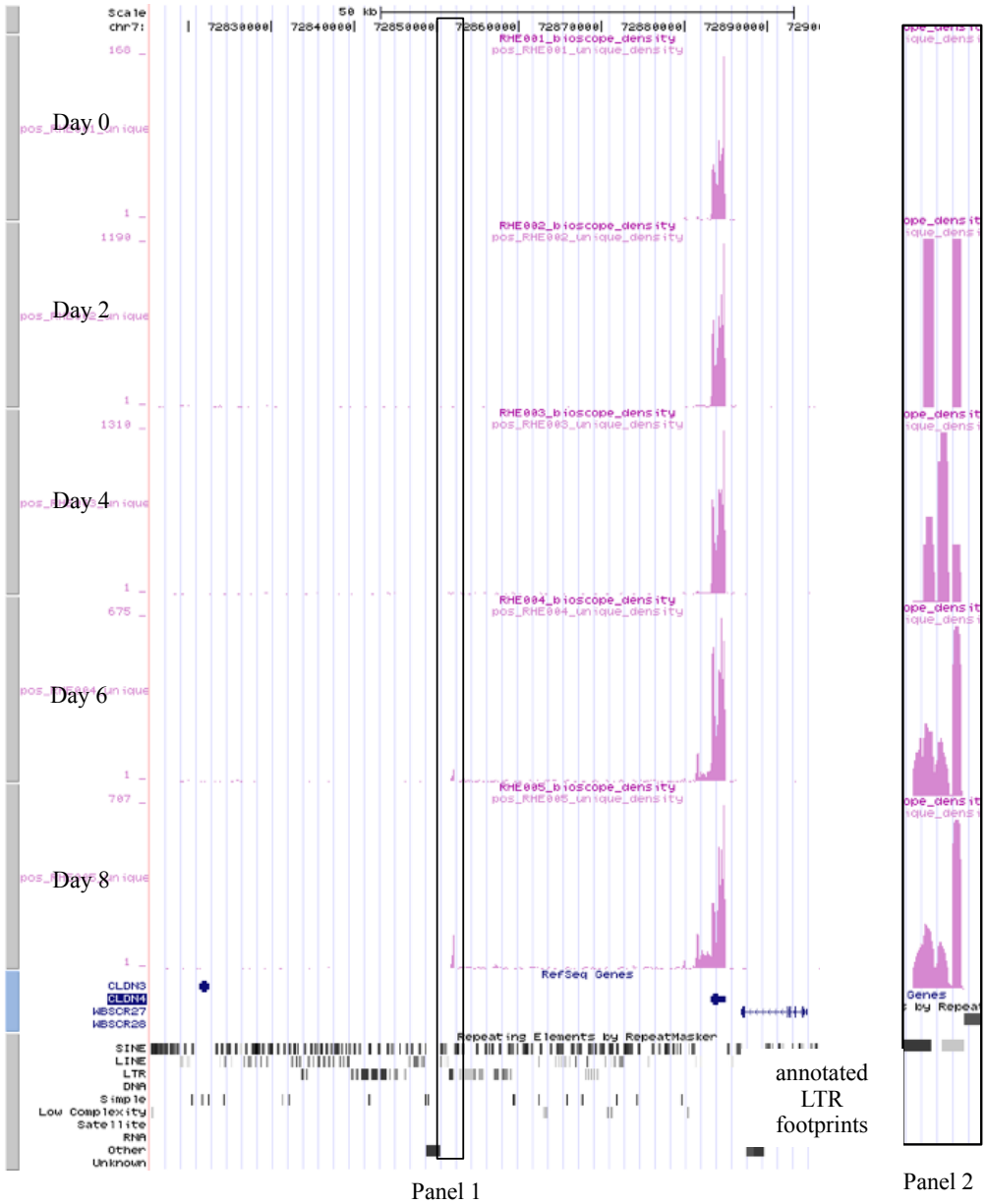


Figure 4.49: RNA-Seq peak profile of *CLDN4* and its novel exons as identified by RNA-Seq.
Panel 1 shows the first novel exon and panel 2 shows an enlarged view of that peak showing its LTR footprint.

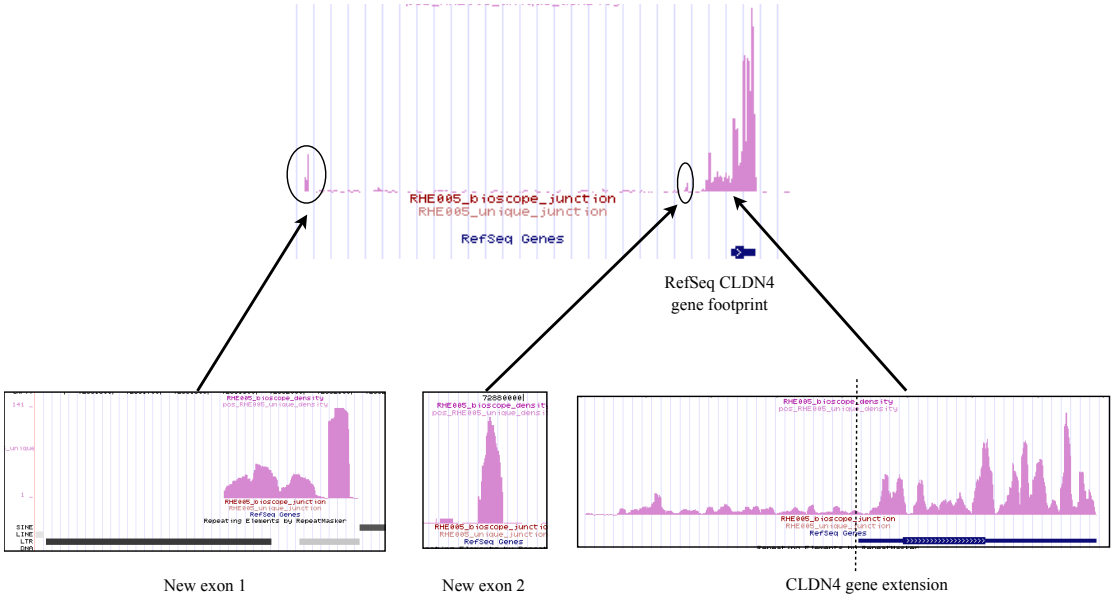


Figure 4.50: Enlarged view of the novel exons of *CLDN4* identified using RNA-Seq.
The extension to the original CLDN4 can also be seen. Here only the peaks of day 8 time point are shown.

4.18.2 *DHX32* (DEAH (Asp-Glu-Ala-His) box polypeptide 32)

The gene *DHX32* gets up-regulated during trophoblast differentiation. Apart from the known exons, the gene seems to have an additional one in the 5' side and this novel exon has an overlap with an LTR element. This observation was validated by PCR.

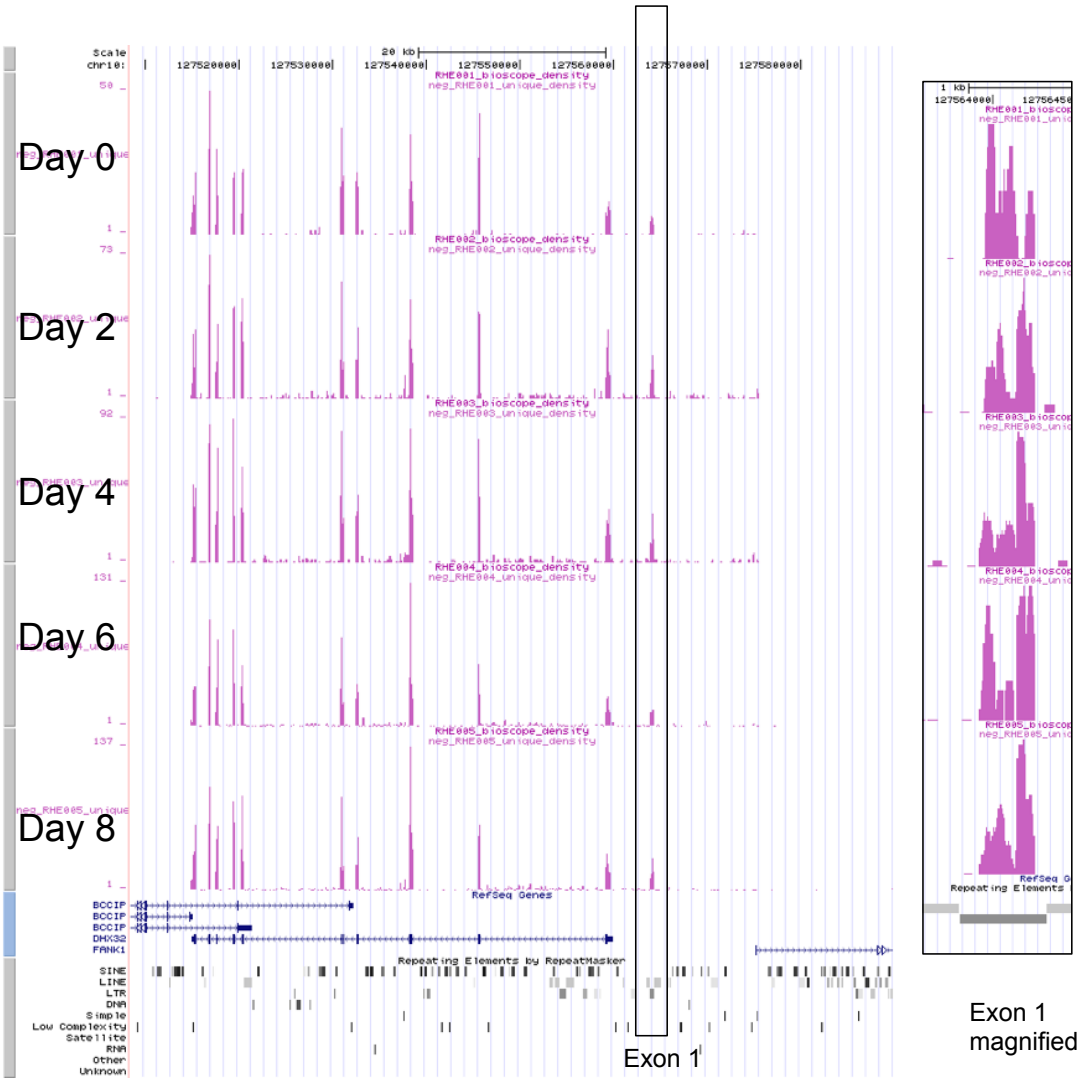


Figure 4.51: The gene *DHX32* has a novel exon on its 5' end (exon 1).
The panel on the right (exon 1 magnified) shows an enlarged view of the same exon with the LTR annotations. The peak falls fully on an LTR element which is flanked by two LINE elements.

4.18.3 *MYCT1* (MYC target 1)

Based on RefSeq annotation the gene *MYCT1* has two exons. Its expression begins at day 2 and continues throughout the differentiation. RNA-Seq data indicates that the first exon of *MYCT1* is not expressed, however there are two novel transcribed regions which are on the 5' side of the transcript, and they share footprints with LTRs. The similar expression pattern between *MYCT1* and the two NTRs suggests that they could be two new exons of *MYCT1*.

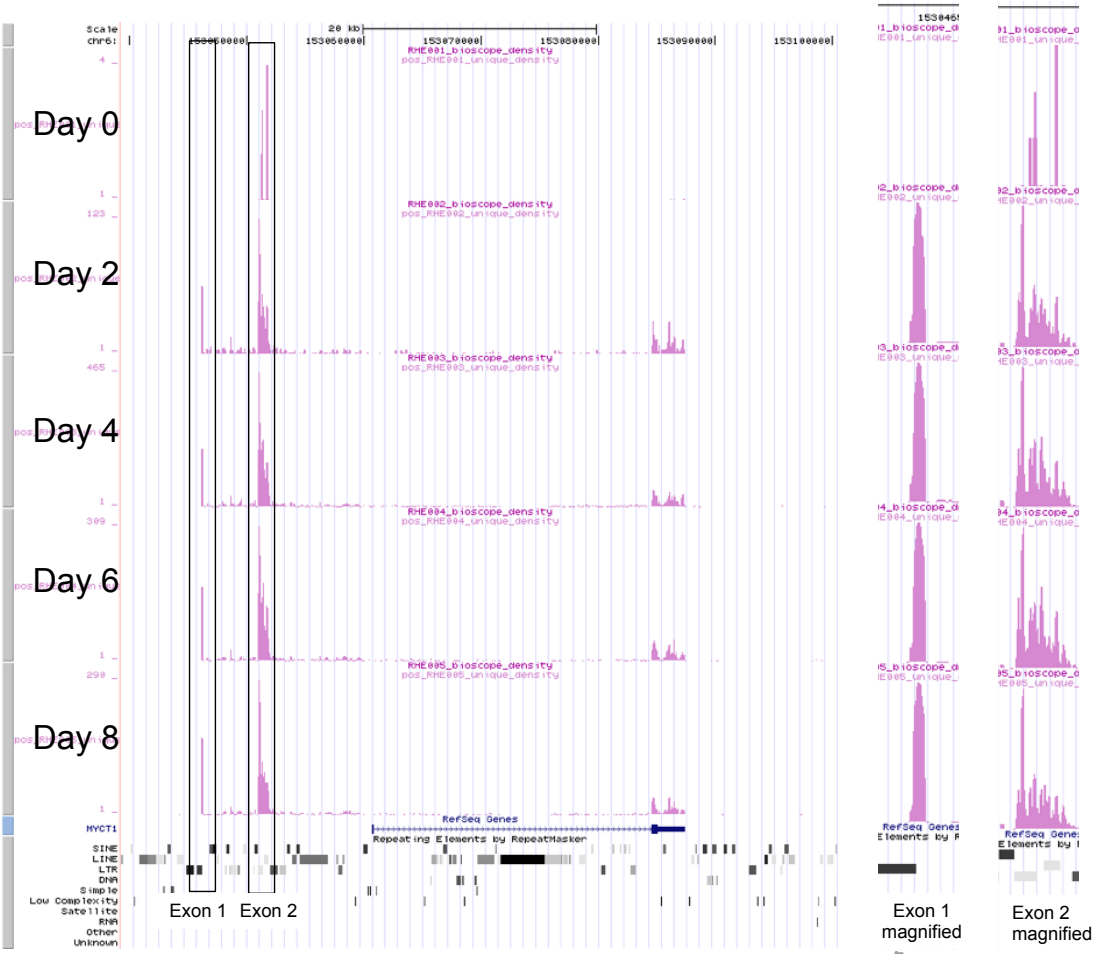


Figure 4.52: RNA-Seq profile of *MYCT1* and its two novel exons identified by RNA-Seq.
There are two new 'exons' on the 5' end of the MYCT1 gene. Both peaks have overlaps with LTRs (shown on the left portion of the Figure).

4.18.4 *ZBTB3* (Zinc finger and BTB domain containing 3)

The gene *ZBTB3* is down-regulated during trophoblast differentiation. Looking at the RNA-Seq data (Figure 4.53) it is clear that there is an additional exon on the 5' side of the transcript. This transcript is only present in day 0 and therefore can be considered as stem cell specific. The footprint of the new exon is derived from an LTR element. This has been validated using PCR.

Interestingly this LTR has been reported to recruit NANOG in ES cells (Kunarso, Chia et al. 2010) which hints at a co-regulation mechanism between the expressions of this novel exon and NANOG.

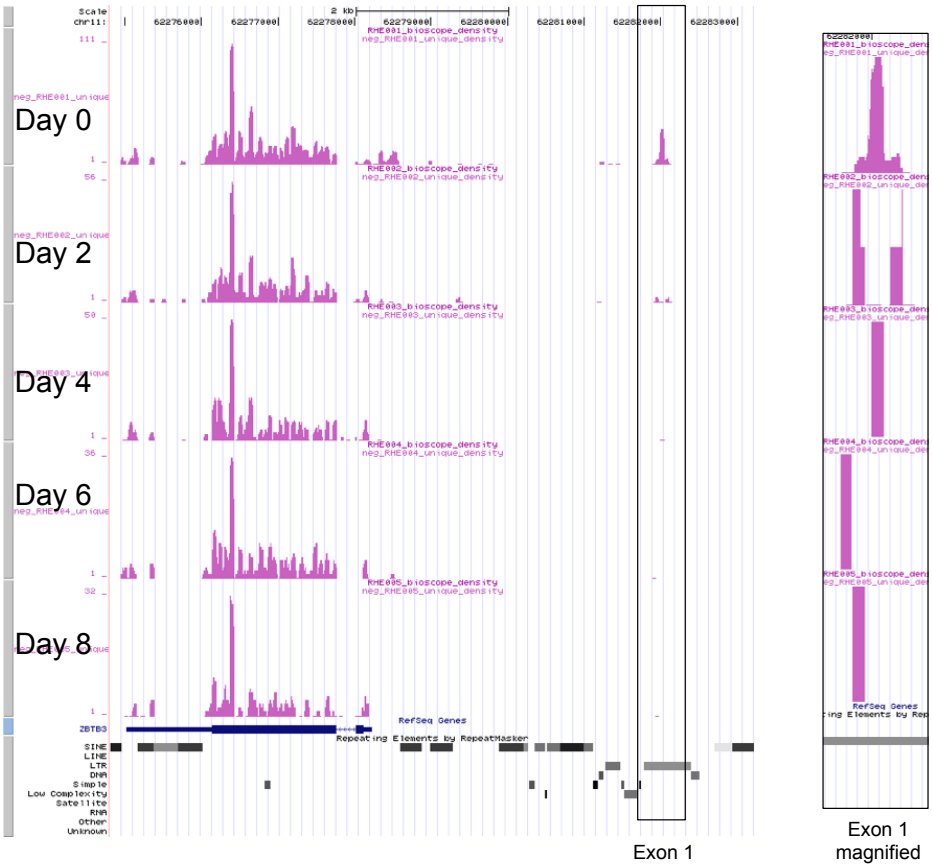


Figure 4.53 : RNA-Seq peak profile of ZBTB3 and its ES specific novel exon.
The gene ZBTB3 has a novel exon on its 5' end which has a complete overlap with an LTR element (ERV1 subfamily)

4.18.5 *SCGB3A2* (Secretoglobin, family 3A, member 2)

SCGB3A2 is down-regulated during trophoblast differentiation. RNA-Seq indicates that the first exon of the transcript is hardly transcribed (has low number of reads aligned to it) and that there is a novel exon, which has originated from an LTR region and expressed at a similar level to the gene. Expression of the novel exon and the skipping of the RefSeq 1st exon which contains the original start codon, suggest that the *SCGB3A2* protein structure could be affected. PCR results showed that the NTR is indeed a novel exon of *SCGB3A2*.

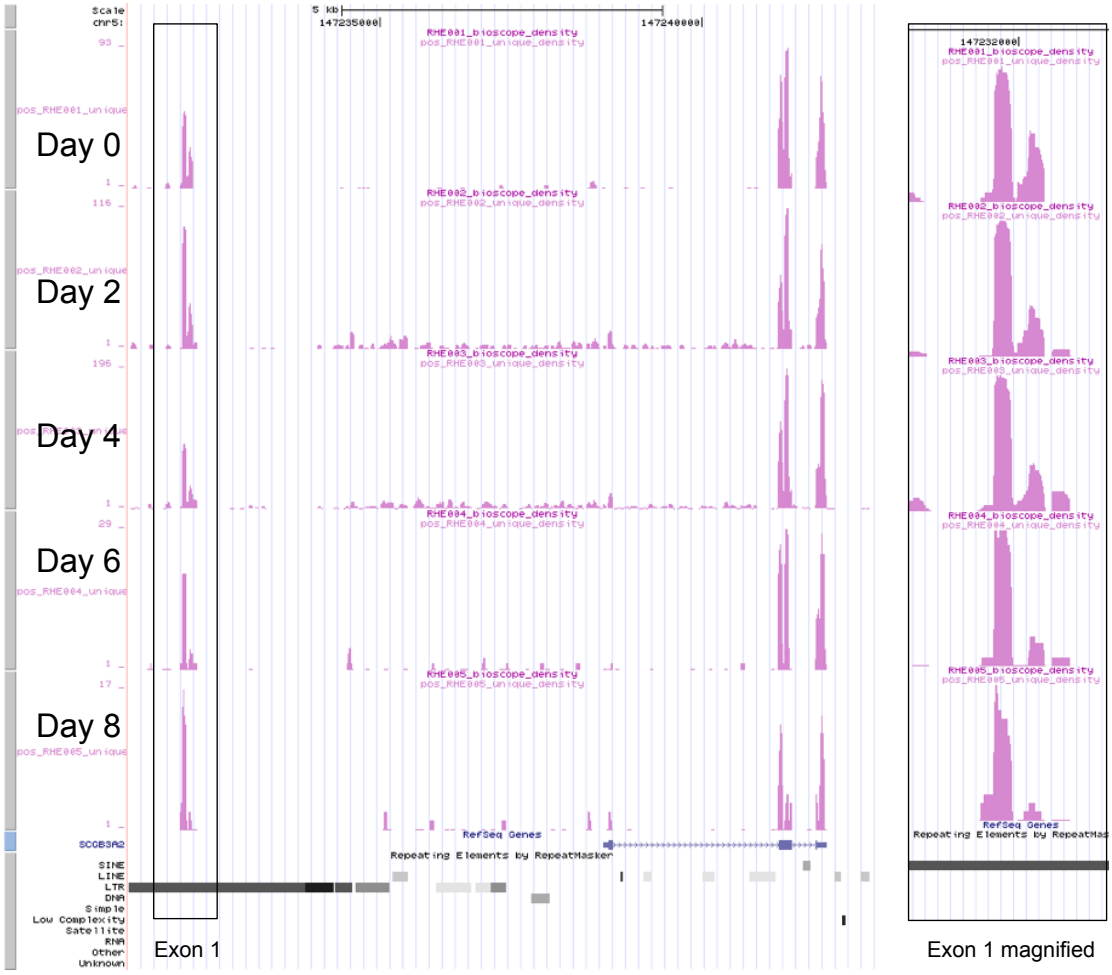


Figure 4.54 :RNA-Seq peak profile of *SCGB3A2* and its novel exon identified by RNA-Seq.
SCGB3A2 gene has one new exon on its 5' side which has an LTR footprint.

4.19 Genes which show a change in their splicing profile during trophoblast differentiation

Transcriptomic dynamics of any system cannot be exclusively described based on the expression level of genes. To sufficiently study a transcriptome, the alternate splicing events needs to be described together with gene expression. Alternative splicing events are quite common in the transcriptome (Wang, Sandberg et al. 2008) and they are reported to be important in regulating developmental processes (Kanadia and Cepko 2010).

If the splicing occurs in a protein coding region then it could influence the biological function of the protein. On the other hand, if the splicing is restricted to an untranslated region then the postranscriptional regulation of the transcript could potentially be affected.

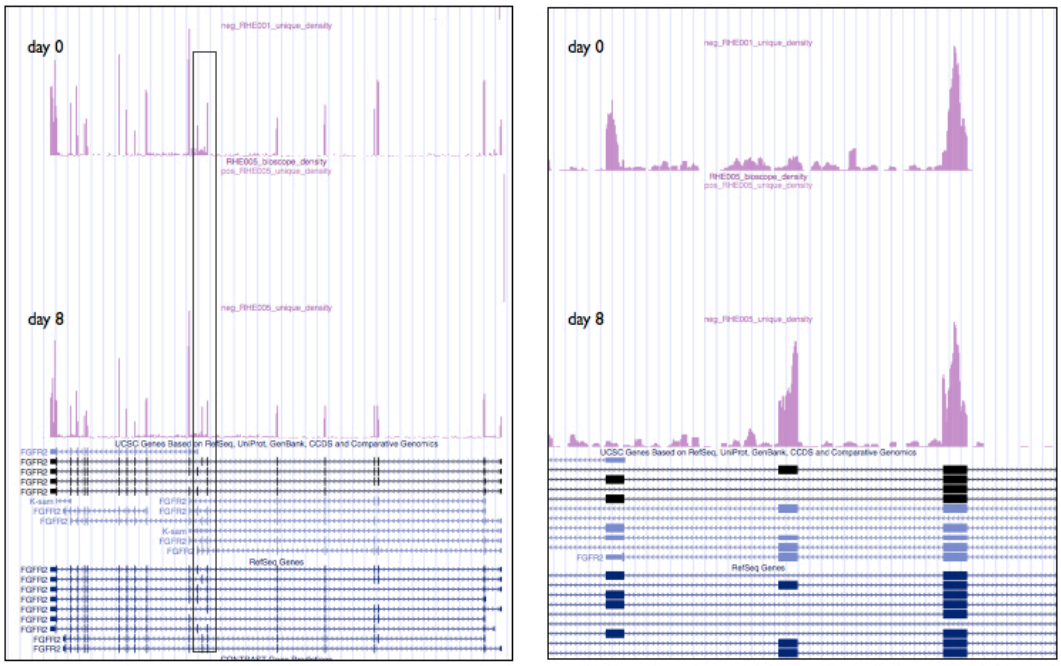
A comprehensive study on the alternative splicing events during early development has not been done before, mainly due to the limitations of microarray technology. Therefore the alternative splicing detection workflow (as described in Results 1 section) was written to identify alternative splicing events leveraging on RNA-Seq data which provides expression information of the entire gene.

Based on the alternative splicing detection workflow, 385 genes which show a change in their alternative splicing profiles were identified. The criteria used were that both exons showing the splicing should have a read count of more than 10 and that they should show a three fold or more differential change in the two exons (i.e. fairly

stringent criteria). Based on results, the workflow identified standard alternative splicing events, mutual exon splicing events and also alternative start / stop sites.

4.19.1 Mutual exclusive splicing of Fibroblast growth factor receptor 2 (*FGFR2*)

One of the most striking examples of alternative splicing was found in the fibroblast growth factor receptor 2 (*FGFR2*) which happens to be mutually exclusive. Data shows that there is a shift of expression from the 8th exon of FGFR2 transcript variant 1 (NM000141.4) to the 8th exon of FGFR2 transcript variant 2 (NM022970.3), when comparing day 0 and day 8 samples (i.e. a shift from exon IIIc to IIIb). It has been reported that the *FGFR2* - IIIb isoform is specific to epithelial cells and the IIIc isoform is specific to mesenchymal lineage (Orr-Urtreger, Bedford et al. 1993). Therefore this observation clearly shows the transformation of stem cells into an epithelial lineage. Furthermore, ESRP1 and ESRP2, which are epithelial cell type-specific splicing regulators of *FGFR2* (Warzecha, Sato et al. 2009), are up-regulated during the early stages of differentiation (day 2 and day 4), which are the time points where the flip in the mutual exclusive isoforms takes place.



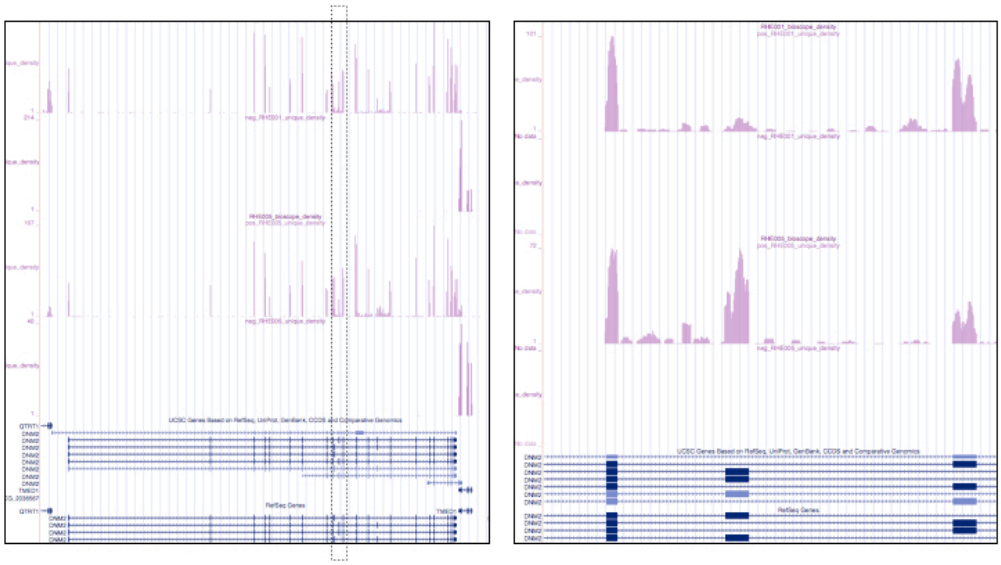
Panel A : The full FGFR2 gene

Panel B : The two mutually exclusive exons

Figure 4.55: Mutual exclusion of *FGFR2* exons.

4.19.2 Mutual exclusion splicing of dynamin 2 (*DNM2*)

Similar to that of *FGFR2*, Dynamin 2, also shows a mutual alternative splicing between its 10th exon of isoform NM_001005361.1 and the 10th exon of isoform NM_004945.2.



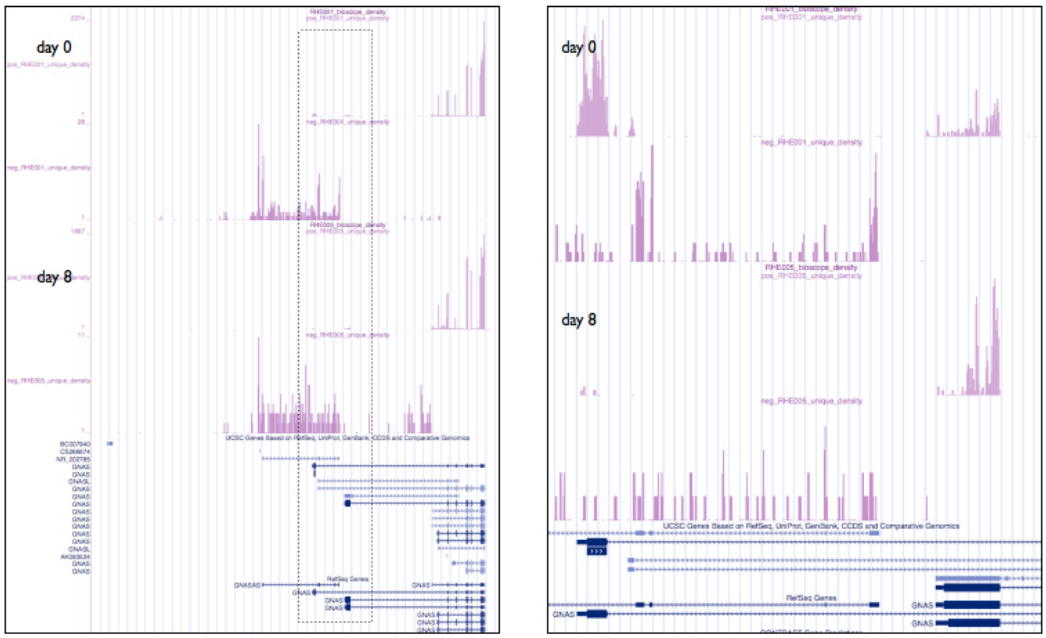
Panel A : The full DNM2 gene

Panel B : The two mutually expressed exons

Figure 4.56: RNA-Seq peak profile of *DNM2*.

4.19.3 Alternative start exon in guanine nucleotide binding protein (G protein), alpha stimulating activity polypeptide (*GNAS*)

The alternative splicing detection workflow also picks up genes which change their transcription start sites by 'dropping' exons. *GNAS* shows such a change in expression where it shifts the expression from transcript variant 4 (NM_080425.2) to transcript variant 2 (NM_016592.2). It should be noted that based on RNA-Seq data, the opposite strand also show some expression at day-0.



Panel A : The full GNAS gene

Panel B : The two alternative transcription start sites

Figure 4.57: RNA-Seq peak profile of *GNAS*.
Panel A shows the profile of the full gene and Panel B shows the enlarged footprints of the two start exons. While the expression of one start exon goes down, the other start exon's expression level goes up suggesting a change in isoforms and start exon usage.

4.19.4 GATA binding protein 2 (*GATA2)*

GATA2 is reported as an important regulator of trophoblast specific gene expression and placental function (Ma, Roth et al. 1997). *GATA2* is one of the highest up-regulated genes in both the RNA-Seq trophoblast differentiation (150 fold) and in human blastocyst development (Zhang, Zucchelli et al. 2009) . Up-regulation of *GATA2* is immediate upon differentiation by SU5402 + BMP4 as can be seen by the RNA-Seq peak profile in Figure 4.58. This immediate up-regulation can, in part, be explained by the fact that this gene is a known BMP4 target.

Based on RefSeq annotation, *GATA2* has three isoforms that differ from each other by use of different transcription start sites leading to three unique first exons. The translation start site resides in exon 2. This alternative promoter use has been conserved between the chick and human (Nony, Hannon et al. 1998; Pan, Minegishi et al. 2000), at least for the most distal and proximal promoters. Analysis of my RNA-Seq data indicates that during differentiation, the most proximal promoter (producing transcript NM 001145662.1) is expressed first, identifiable in the day 2 data, and at around day 6, expression is evident from both the proximal and distal promoter (transcript NM 032638.4) (Figure 4.58). There is a Smad responsive element immediately adjacent to the proximal promoter of *GATA2* (Karaulanov, Knöchel et al. 2004). This could explain early activation of the *GATA2* isoform corresponding to the proximal promoter as Smad7 is expressed and up-regulated throughout the differentiation. The presence of a *GATA2* binding site at the distal promoter might explain the expression of the alternative isoform.

This observation suggests of an instance where a gene is regulated by two promoters, one inducing the expression while the other maintaining it. This dynamic switch in *GATA2* promoter use has not been previously described. This example represents the power of combining a comprehensive transcriptomic analysis (i.e. RNA-seq) with a developmental time-course to provide insight into developmental mechanism.

In addition, RNA-Seq data indicates that there is a novel transcript which is transcribed by the opposite strand but which overlaps with the 5' portion of the *GATA2* gene (Figure 4.58), and that it has a similar expression pattern to *GATA2* (i.e. not expressed in day 0, and gets induced during differentiation).

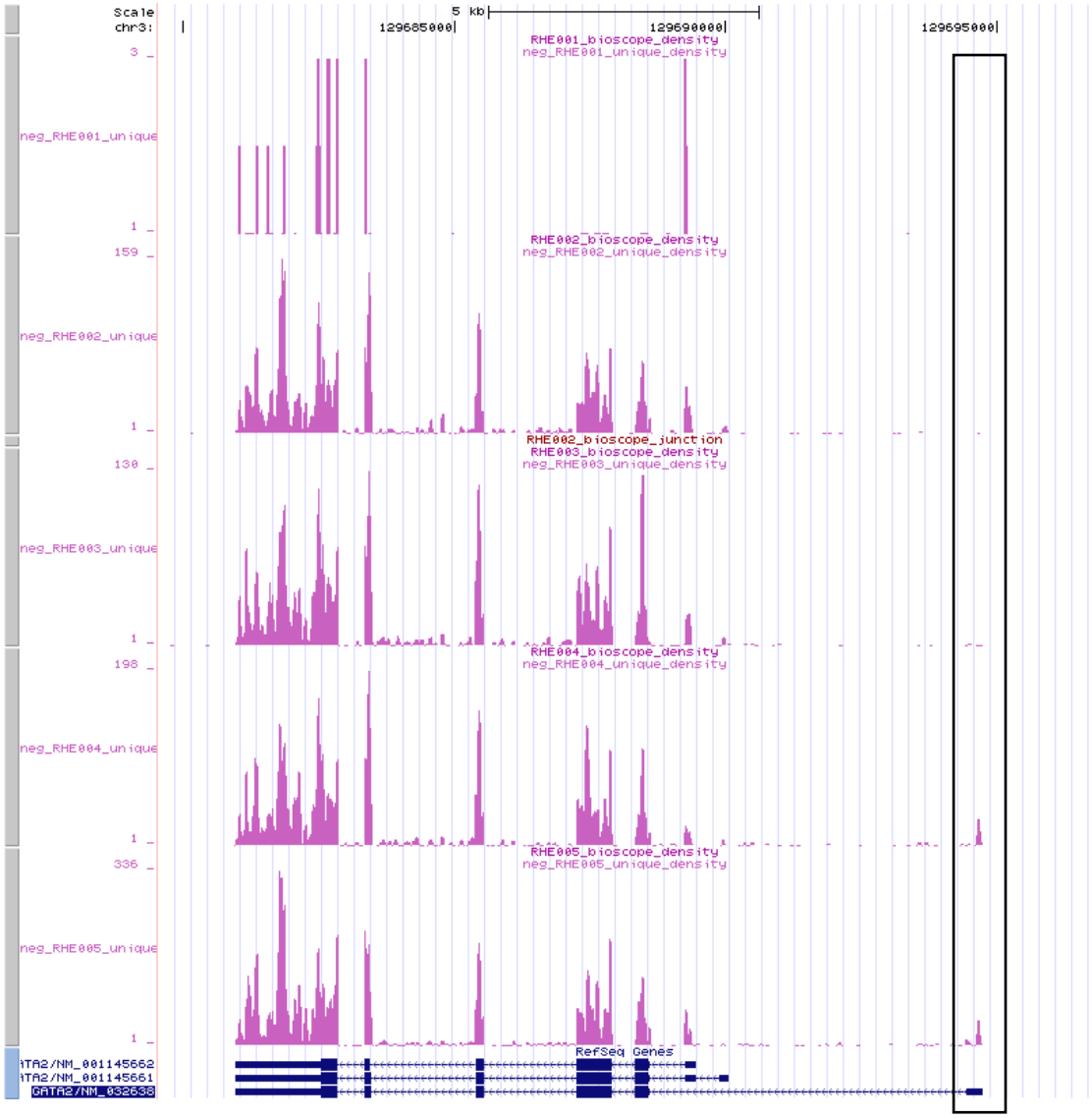


Figure 4.58: RNA-Seq peak profile of *GATA2*.
It gets induced during day 2 and keeps on being up-regulated. Initially a short transcript is expressed and around day 6 a longer transcript is transcribed.

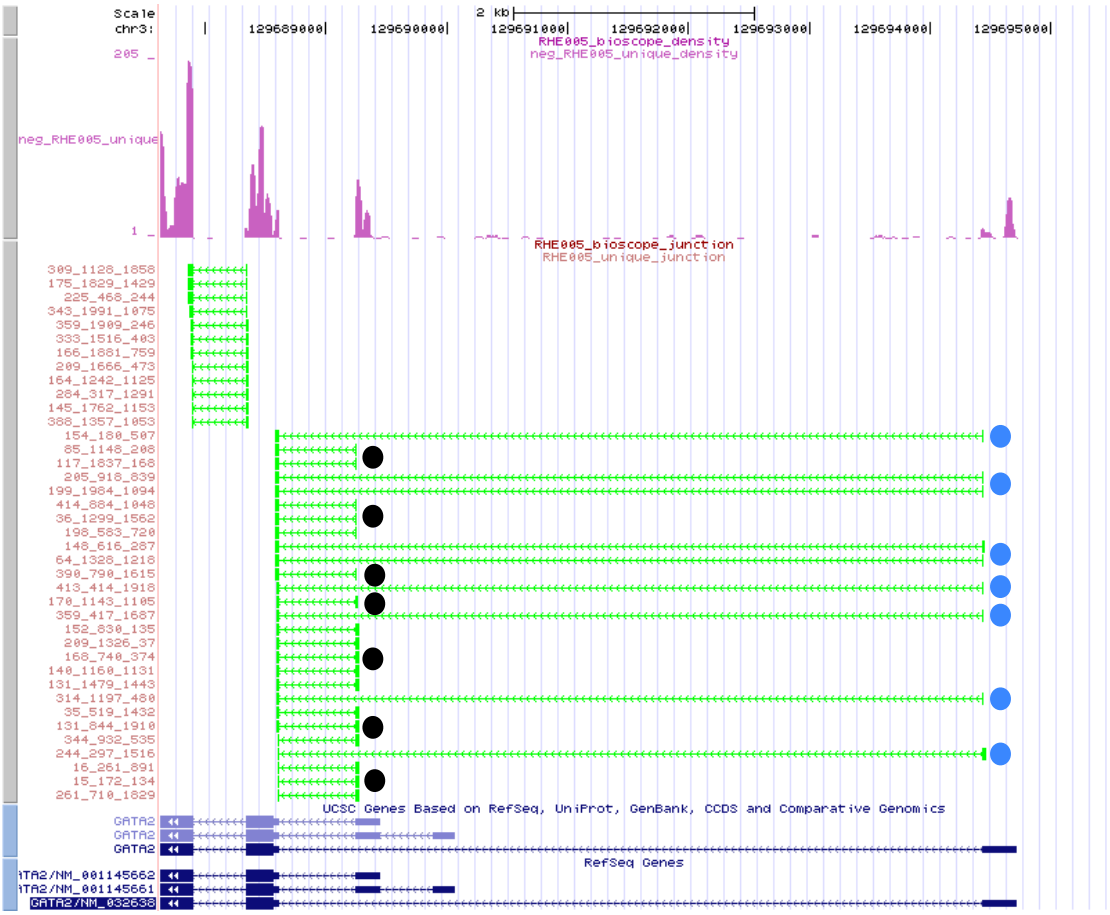


Figure 4.59: Different isoforms of *GATA2* expressed at day 8.
At day 8, the shortest and the longest isoform of the GATA2 gene is expressed, and the third isoform is not. The above Figure shows the junction reads (in green), which joins the exons together. The junctions marked with a blue dot represent the exon - exon connections of the longer isoform, while the junctions marked with the black dot originates from the shortest isoform. In day 2, only the junction reads from the shorter isoform can be seen.

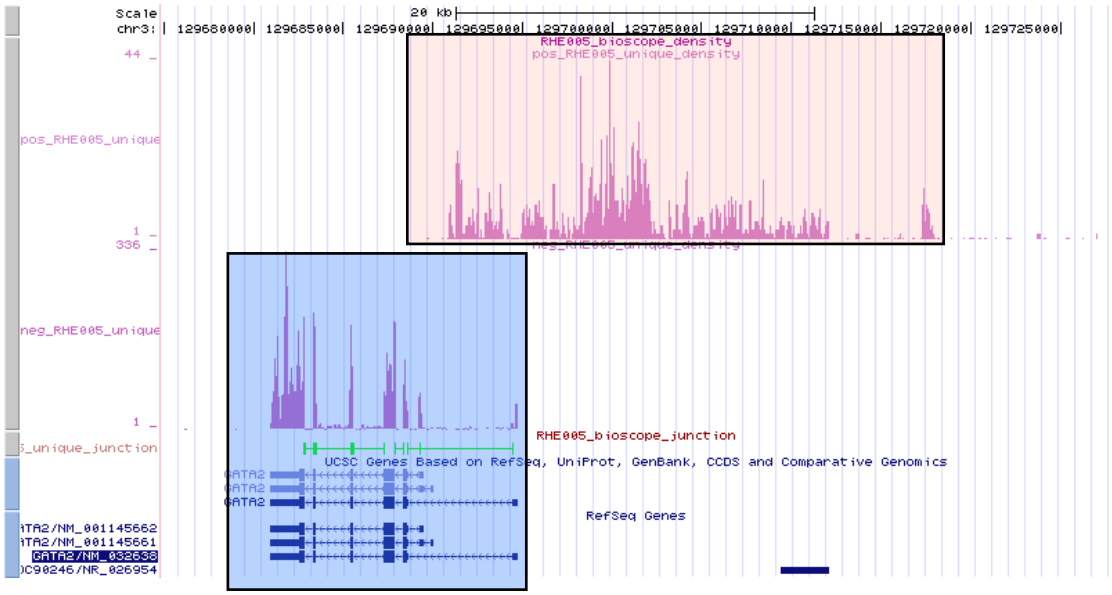


Figure 4.60: The novel transcribed region next to *GATA2*.
The *GATA2* peaks are highlighted in blue and the novel transcribed region / transcript is shown in red. There is considerable overlap between the two (but coded by different strands) and the new transcript has a similar expression pattern to that of *GATA2*.

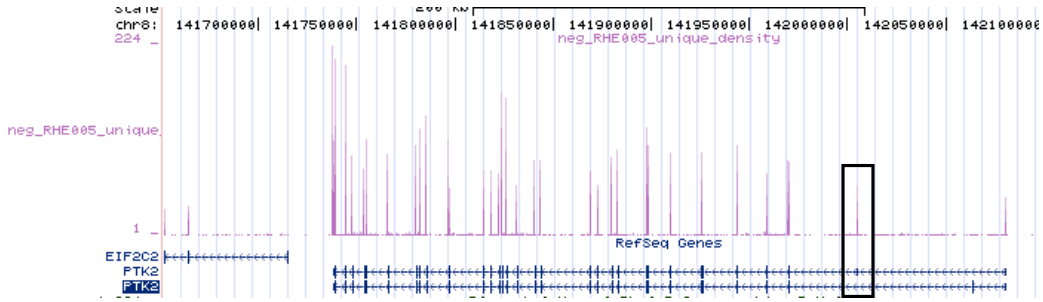
4.20 Identification of novel exon - exon junctions based on RNA-Seq data

As described in the methods section, the mapreads aligner identified reads mapping to exon - exon junctions of known transcripts. This is done through aligning reads (which have not been aligned during the genomic alignment phase) to a sequence database which represents all the exon - exon junctions of RefSeq transcripts including the novel exon - exon junctions.

The reads which map to exon-exon junctions are important as they act as markers indicating that the two exons which makes up a particular junction are connected (i.e. spliced in) with each other. Therefore if in a given gene, there is a significant number of reads mapping to a novel exon - exon junction (i.e the splice junction is not described in RefSeq but the exons which contribute to it are) then it could be used as an indicator to show that there is a new isoform of that particular gene.

I wrote a pipeline to exploit this dataset which predicts novel exon-exon junctions of all the genes in RefSeq. To increase the accuracy of the method, only the reads which are highly specific for the novel junction were used. Based on the results of this workflow which used data from all time points, there were 6,205 genes which showed 12615 potentially novel exon - exon junctions. Among these, there were 253 junctions which had at-least 10 reads mapping to it. Even though on the face of it, 10 reads per junction appears to be too stringent it should be noted that the reads used for this step of analysis showed the best possible unique alignment and that junction reads have a low chance of being aligned on the first place due to the short exon-exon junction foot print.

The main importance of identifying novel exon - exons junctions is that a novel junction could completely change the function of the protein. This is demonstrated by using the novel exon - exon junctions identified in *PTK2* and *PAPOLA* genes (Figures 4.61 and 4.62).



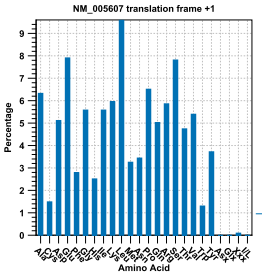
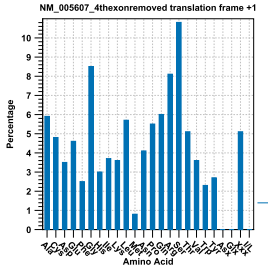
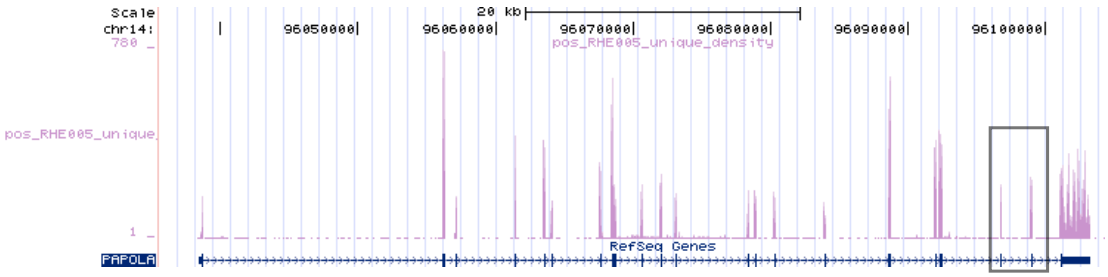
	NM_005607 (PTK2) - original	NM_005607 with 4th exon spliced out
Length	1075 a.a	999a.a
Weight	121kDa	111kDa
Isoelc. point	6.42	9.28
Aliphatic Index	81.16	56.04
Amino acid composition		
Pfam domains	protein kinase domain alpha / beta hydrolase fold cystathionine beta synthase domain aldehyde dehydrogenase	negative regulatory factor

Figure 4.61: The novel exon-exon junction of *PTK2* identified by RNA-Seq and its influence on the protein product.
For clarity only the negative strand is shown in the day 8 time point. Note that the changes brought about by the novel junction ultimately leads to changing the domains contained within the protein.



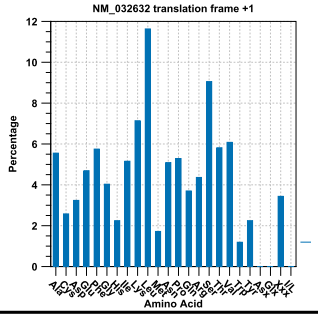
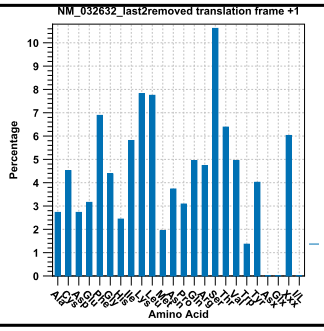
	NM_032632 (PAPOLA) - original	NM_032632 with 4th exon spliced out
Length	1513 a.a	1394 a.a
Weight	170 kDa	161 kDa
Isoelc. point	9.52	9.65
Aliphatic Index	90.66	73.48
Amino acid composition		
Pfam domains	acetyltransferase family aminotransferase class I and II type II intron maturase	ring finger domain HAMP domain

Figure 4.62: The novel exon-exon junction of *PAPOLA* identified by RNA-Seq and its influence on the protein product.

GENE	Chr	Novel Junction (hg 18)	
		Start	End
USP28	chr11	113180979	113184230
PTOV1	chr19	55049613	55052973
FBLN1	chr22	44307709	44321629
ZFP42	chr4	189153989	189157757
PALLD	chr4	170082805	170083939
PBX1	chr1	163048011	163057398
FUS	chr16	31101487	31103761
FMR1	chrX	146826812	146829787
MYO6	chr6	76675065	76680500
GPBP1L1	chr1	45899485	45924671
COL6A2	chr21	46364920	46367217
YAP1	chr11	101582016	101599563
OSBPL8	chr12	75368900	75405421
SYNGAP1	chr6	33517515	33518644
COBLL1	chr2	165265513	165269200
AP1B1	chr22	28054885	28056367
LTA4H	chr12	94920974	94924223
ATP1A1	chr1	116737869	116743565
EPB41L3	chr18	5396968	5400565
RNF213	chr17	75942562	75977568
HISPPD1	chr5	102548345	102554442
DGUOK	chr2	74007688	74027354
PTK2	chr8	141943681	141969824
SIN3B	chr19	16834371	16835491
PRKCSH	chr19	11419434	11419508
CCT2	chr12	68267664	68279910
HMG20A	chr15	75500402	75537801
EPB41L3	chr18	5387426	5396776
COL1A2	chr7	93881524	93894256
PLD3	chr19	45546516	45563300
NTRK2	chr9	86474159	86475112
EPB41L3	chr18	5383478	5385066
EMID1	chr22	27941620	27957009
ACTN4	chr19	43888612	43904018
BAT3	chr6	31714983	31715955
PPP4C	chr16	29995298	30001306
ZNF664	chr12	123038639	123061881
DMKN	chr19	40688729	40692926

Table 13: Genes which have a novel exon - exon junction with more than 20 reads mapping to it.

4.21 Extensions to existing annotations at 3' and 5'

One of the main advantages of RNA-Seq data is that it is not dependent on existing annotation. This becomes a great asset in identifying novel transcriptomic phenomena. One such example is the improvement of existing annotations. The alignment pattern of the sequenced reads to a particular annotated region in the genome can be used to validate the existing annotations and make alterations if required. In certain cases the RNA-Seq peaks tends to 'extend' beyond the known RefSeq annotation boundary, and a workflow was developed to identify these extensions in a genome-wide manner.

The pipeline found that there were 1708 internal exons (all exons except 3' and 5' UTRs) which had an 'extension' of 100 base pairs or more on 5' and / or 3' side in day 0. There were 322 genes which had more than 1000bp extensions on their 3' UTR while 20 genes had more than 1000bp 5'UTR extensions. Based on distribution of extensions, it is clear that the 3' UTR extensions have a longer average length (Figure 4.63).

The majority of target sites of microRNA lies in the 3' UTR region of the transcripts. For this reason the proper annotation of specially the 3' UTR is vital. This study identifies a total of 1,575 kb extensions of 3' UTR regions with respect to RefSeq annotation for the entire genome. This extension region is vital for accurate identification of microRNA-target transcript pairs. For example a microRNA - target prediction performed using the 3' UTR extension regions identified here through miRANDA prediction algorithm (Enright, John et al. 2003) predicted 1000+ additional microRNAs which could bind to *NANOG*.

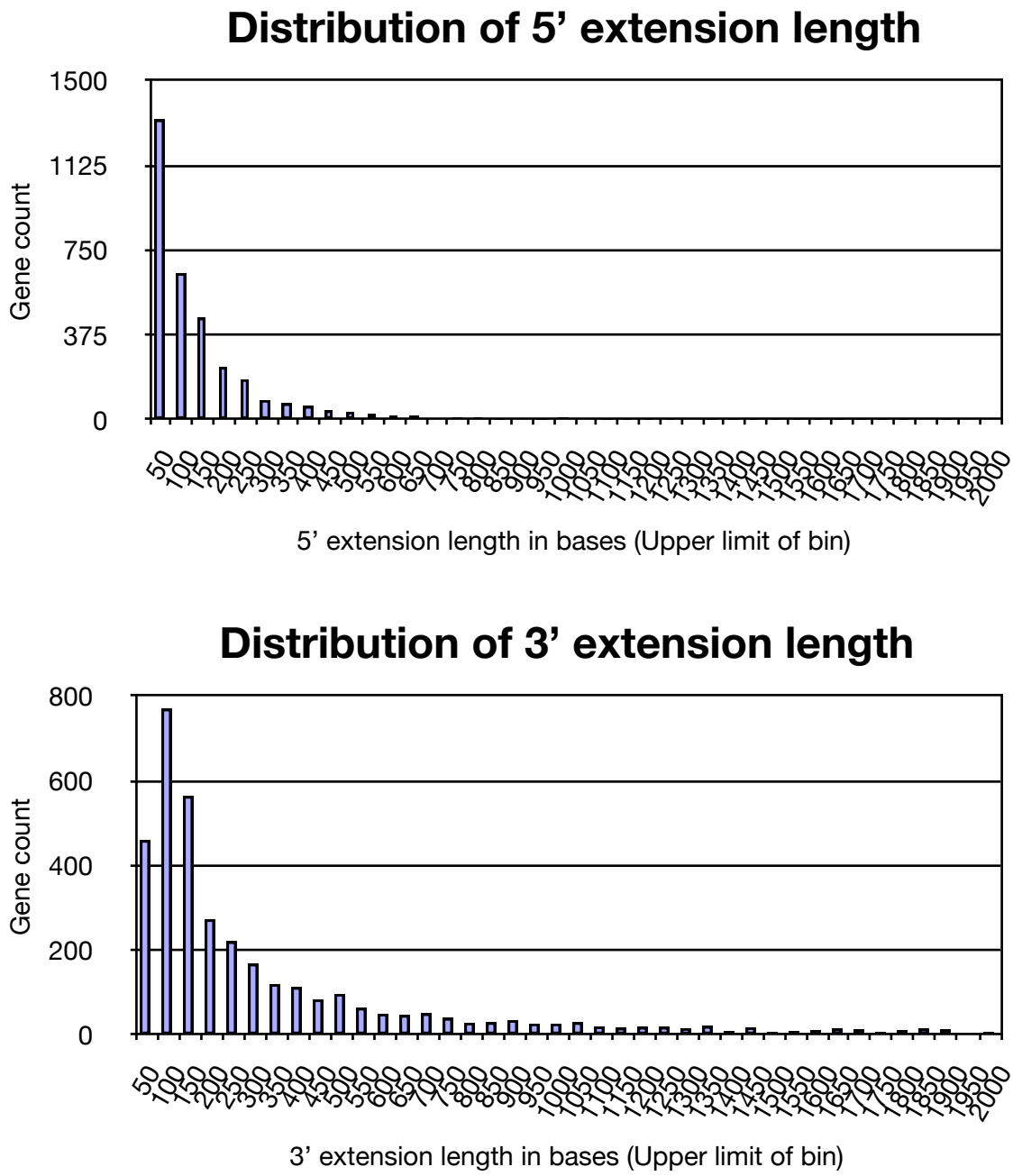


Figure 4.63: Distribution of 3’ and 5’ UTR extensions based on RNA-Seq data.

4.21.1 A few examples of UTR extension

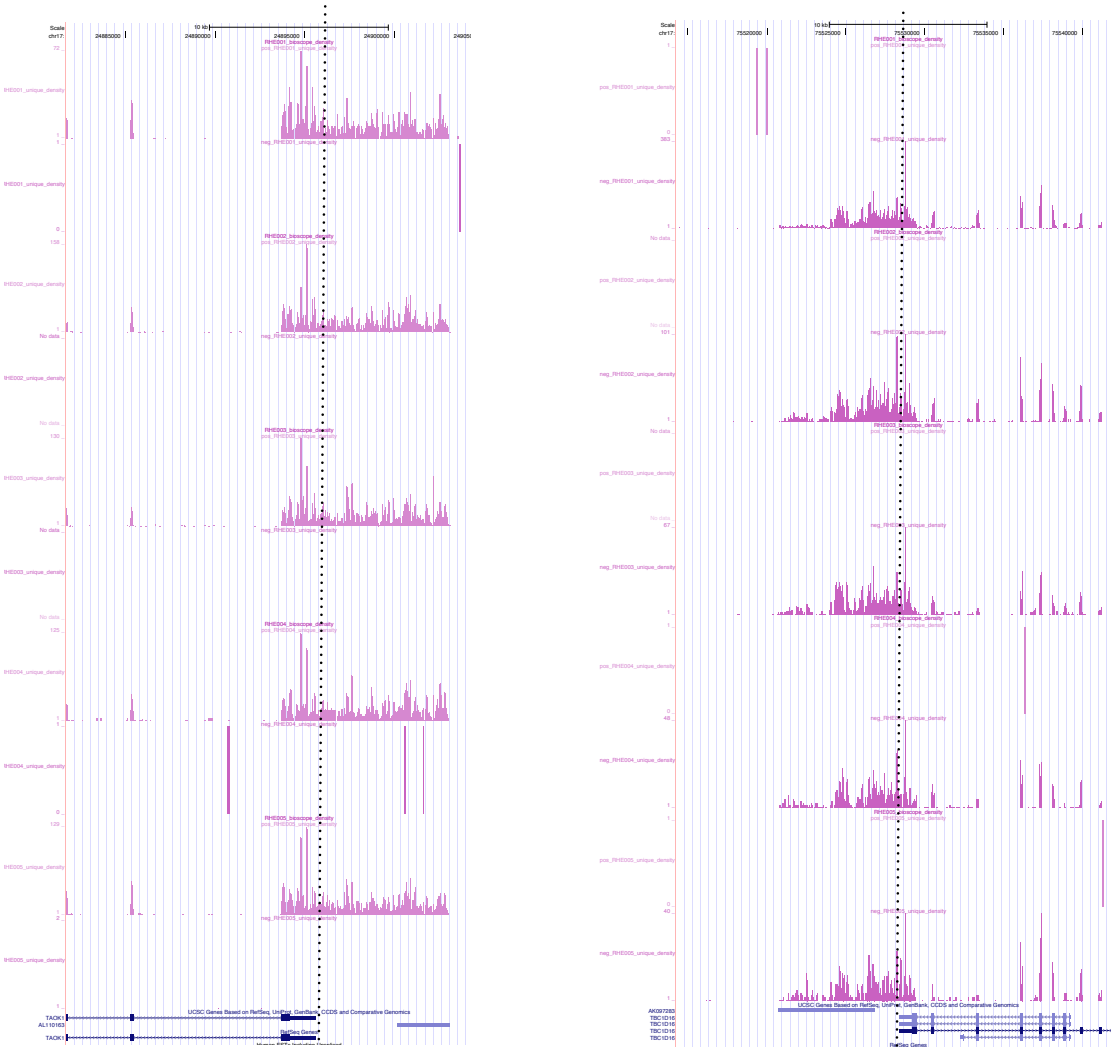


Figure 4.64: UTR extensions of TAOK1 and TBC1D16 two genes with the highest extended 3' UTRs.

These extensions are much longer than the original length of the transcript itself.

To further show the superior nature of RNA-Seq and the extent of 3' UTR extensions as compared to RefSeq annotation data EST BC042436 can be used as an example. This EST does not have a corresponding annotation in RefSeq and it does not have an Illumina probe associated to it. An associated Affymatrix probe (212444_at) has no associated RefSeq gene annotated to it. However RNA-Seq data clearly indicates that this is in fact a result of the 3' extension of the gene *GPRC5A* (Figure 4.65).

In the preimplantation dataset both *GPRC5A* and the probe 212444_at gets up-regulated during blastocyst formation. The co-regulation of these two - seen by the similar expression pattern, further validates that they are indeed from the same transcript.

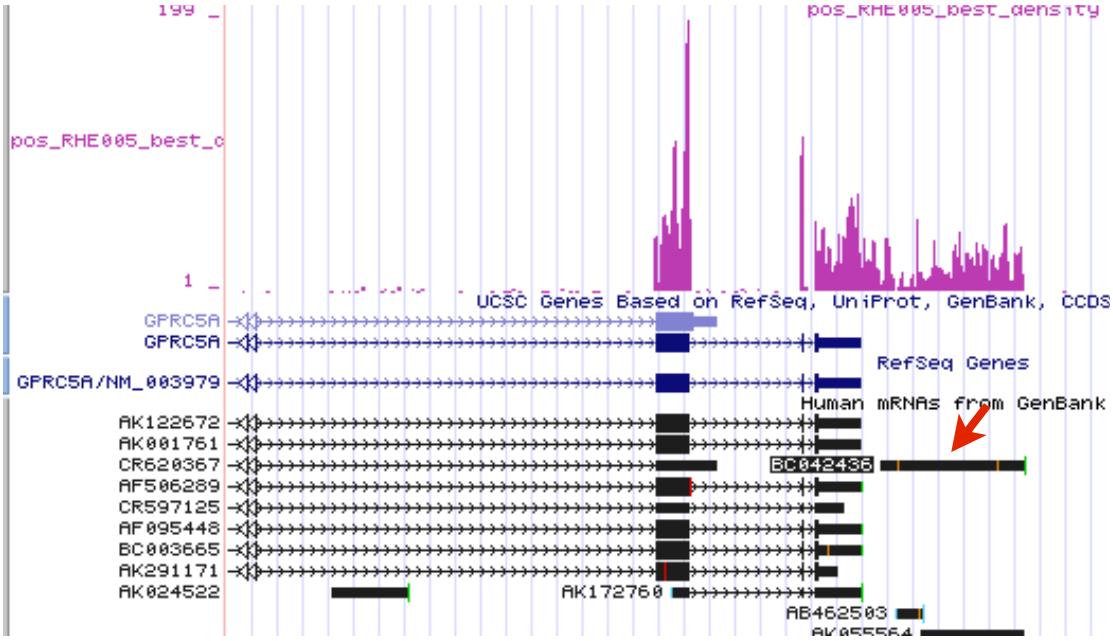


Figure 4.65: Extension region of *GPRC5A*.
The EST BC042436 does not have a RefSeq annotation, and is annotated by GenBank as an independent mRNA with a partial coding sequence. RNA - seq data clearly shows that it is actually an extension at the 3' UTR of the gene *GPRC5A*.

GeneID	Chr	UTRStop	Strand	Extension Length (Nucleotides)
HMBOX1	chr8	28966159	+	11453
SNORD108	chr15	22783233	+	10611
EIF2C2	chr8	141611433	-	10308
POU2F1	chr1	165651950	+	10050
RAB3B	chr1	52158374	-	10016
HELLS	chr10	96351846	+	9930
WDR86	chr7	150709765	-	9508
ZNF704	chr8	81716267	-	8461
HIST1H2AC	chr6	26232897	+	7691
SESN3	chr11	94546153	-	7436
MBNL3	chrX	131341386	-	7251
DYNLL2	chr17	53522617	+	7153
HEBP2	chr6	138776275	+	7127
TBC1D16	chr17	75529501	-	7106
TAOK1	chr17	24895628	+	7099
FAM40A	chr1	110398786	+	7097
PPM1L	chr3	162271511	+	6752
TRIM71	chr3	32908775	+	6710
MPRIP	chr17	17029598	+	6708
PANK3	chr5	167917204	-	6695
RGP1	chr9	35742871	+	6552
PTPN14	chr1	212598016	-	6469
FAM160A1	chr4	152804234	+	6241
C6orf186	chr6	110674156	-	6149
TMED8	chr14	76878084	-	6090
GRPR	chrX	16081562	+	6086
CDS2	chr20	5119989	+	6060
LOC729082	chr15	39379086	+	6016

Table 14: Genes which show a maximum 3' UTR extension of more than 6000 nucleotides beyond the current RefSeq annotation.

4.22 smallRNA data analysis

The focus on mRNA dynamics and technical issues such as difficulty in designing primers for their small footprint and lack of sequence information has resulted in the low number of high throughput genome-wide studies on small RNA, specially during early development.

Small RNA species include microRNA, which are regulators of mRNA. It has been shown that during trophoectoderm formation in mouse, a massive removal of non-TE-specific transcription factors takes place and that this removal is greater than the increase of TE specific transcription factors (Guo, Huss et al. 2010). Therefore, it is logical to assume that microRNAs could play a major part in the reduction of non - TE specific transcription factors. Thus, study of microRNA expression during trophoectoderm formation could be beneficial in describing mRNA transcriptomic events that take place during this time.

Placental microRNA has been detected in maternal blood during pregnancy (Chim, Shing et al. 2008; Enquobahrie, Abetew et al. 2010). So it is vital to know the microRNA component which is involved in trophoblast formation so that they can serve as markers for placental / trophoblast function.

Due to the above mentioned reasons it was decided to perform a small RNA-seq experiment for the samples of the human trophoblast differentiation protocol. Day 0 (undifferentiated human ES cells), Day 2 and Day 4 time-points of the differentiation were used for this experiment and all RNA less than 200 nucleotides were studied. Since microRNAs are the most active among the small RNA, the emphasis of the

analysis was to study the differential expression of known microRNA and to identify novel microRNA.

4.22.1 Differential expression of microRNA

The miRBASE annotation, containing a total of 1048 microRNAs, was used to obtain the footprints of known microRNA. A microRNA was considered as expressed at significant level when it had more than 20 reads mapping to it. Based on these criteria day 0 time-point (undifferentiated human ES cells) showed 350 microRNAs as being significantly expressed while day 2 and day 4 showed the significant expression of 371 and 365 microRNA respectively.

As for differential expression (day 4 vs day 0), 138 microRNA were up-regulated 2 fold or more and 110 microRNA were down-regulated. (Figure 4.66)

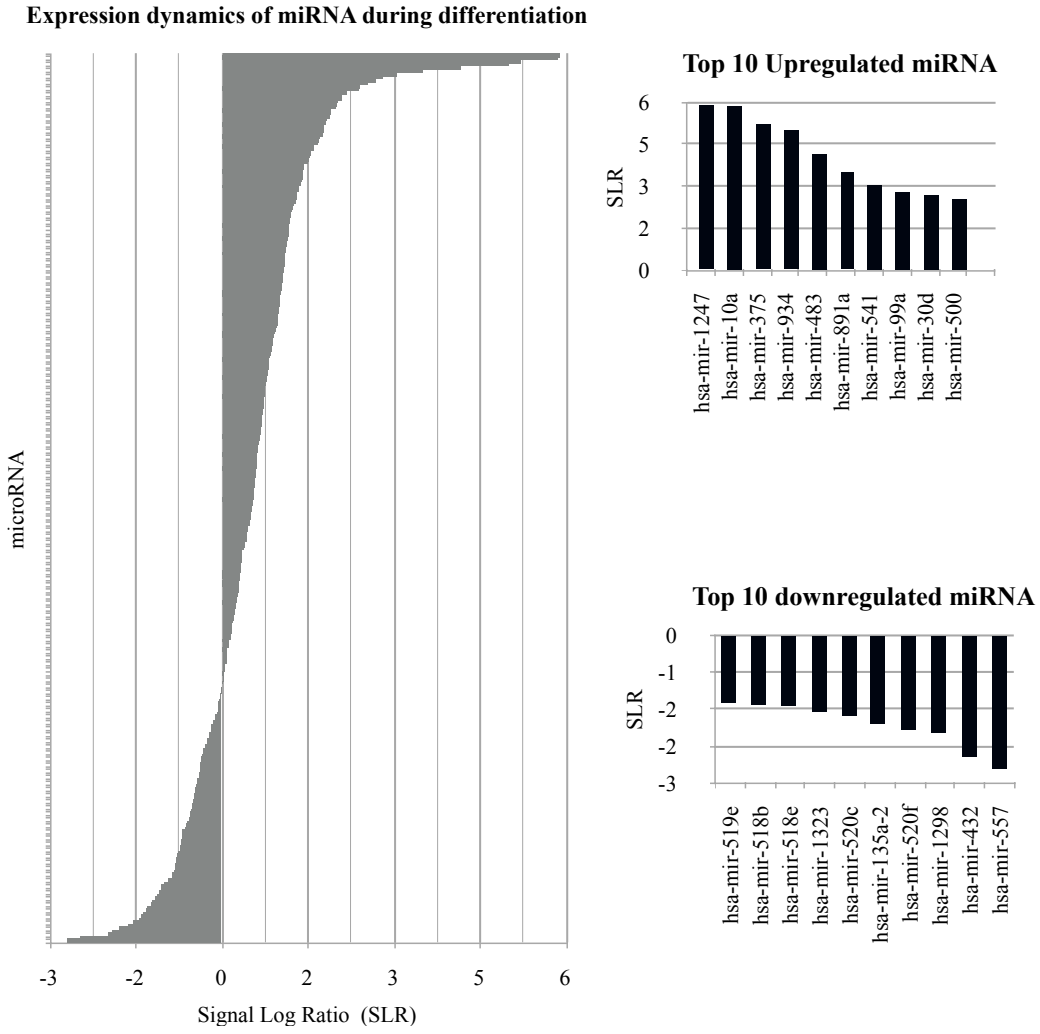


Figure 4.66: The differential expression of microRNA during the trophoblast differentiation.
Each point / bar of the graph on the left shows the expression of a single microRNA. As can be seen, the microRNA component of the transcriptome is highly dynamic during the differentiation, having members which are highly up and down regulated. The top 10 highly up and down-regulated genes and their expression levels are shown on the left.

Presence of a significant number of up and down-regulated microRNA during differentiation suggest that the SU5402 + BMP4 differentiation brings about a considerable change in the small RNA transcriptome. In order to see that this change leads to trophoblast like phenotype / biology seen in the differentiated cells, up-regulated microRNAs were compared with microRNA previously reported to be highly expressed in the placenta and a considerable overlap was observed (Terauchi, Koi et al. 2003; Chim, Shing et al. 2008; Gilad, Meiri et al. 2008; Enquobahrie,

Abetew et al. 2010). Figure 4.67 shows the up-regulated microRNAs during the differentiation protocol together with their expression values. The bars in red shows the microRNAs which, based on literature, are highly expressed in the placenta. The majority of microRNA reported to be expressed in the placenta are up-regulated in the trophoblast differentiation. The microRNAs which are up-regulated during trophoblast differentiation but are not reported to be present in the placenta could be involved in the early stages of placental development or simply unidentified placental microRNA .

Upregulated microRNA during differentiation into trophoblast lineage

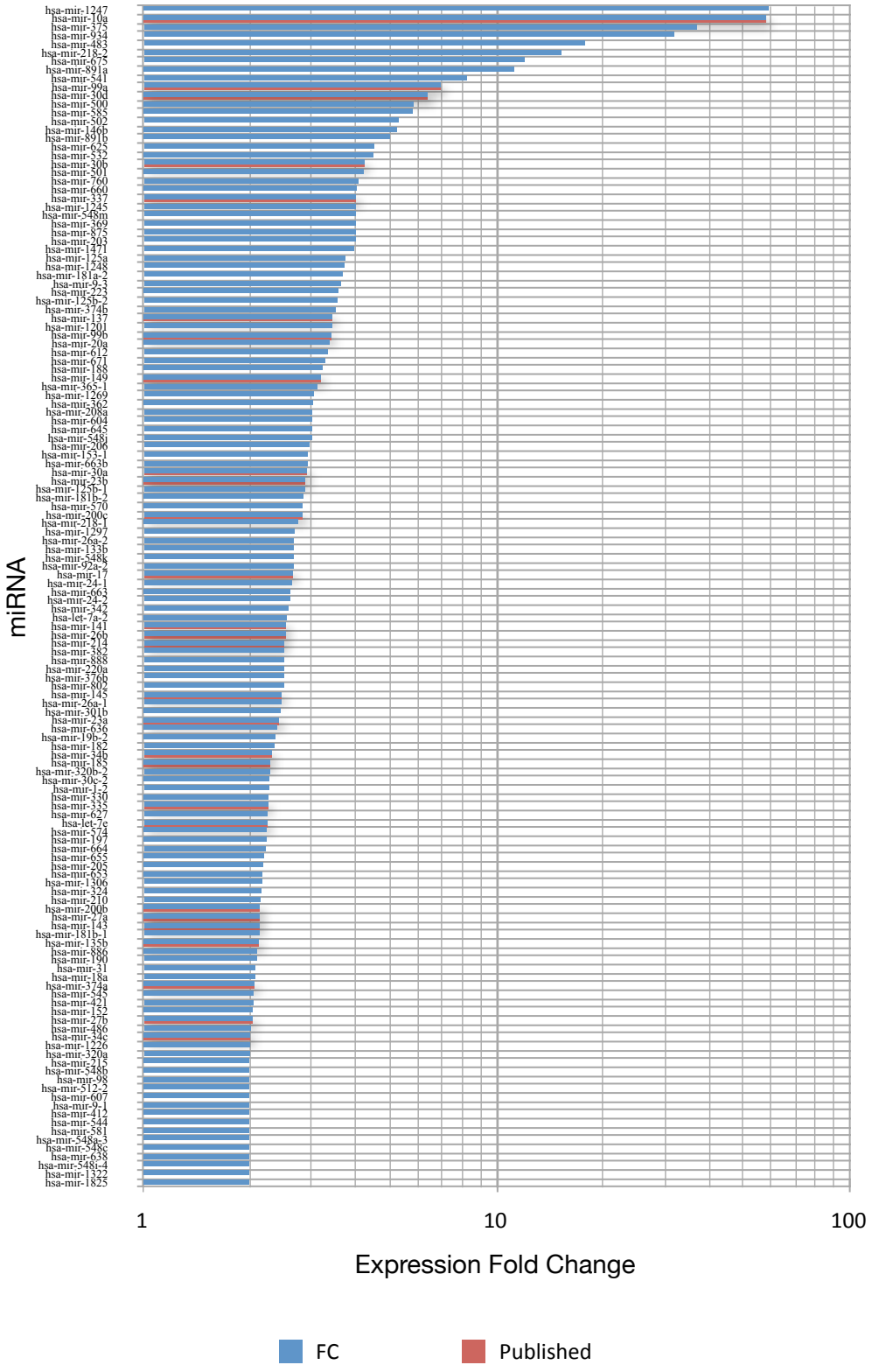


Figure 4.67: The expression level of all the up-regulated microRNA based on RNA - Seq data.

The microRNAs with the red bar are the ones reported in literature to be abundant in the placenta. The ones which are in blue could simply be unidentified new microRNAs of the placenta, or microRNAs involved in the early development of the placenta which once their role is done, gets down-regulated during the mature stages of the placenta.

4.22.2 microRNAs involved in the regulation trophoblast lineage

Based on smallRNA - Seq data, there are 348 microRNA which are up-regulated during the SU5402+BMP4 differentiation. Expression levels of the highest expressed microRNA are shown in Figure 4.67. Since this is the first time that the microRNA component of the early stages of human trophoblast differentiation has been studied, there are no directly comparable data available. The closest available datasets come from placental samples. Chim et al (2008) analyzed third trimester placenta for the expression of 157 microRNA by TaqMan analysis.

Out of the 17 highest expressed microRNAs in placenta based on (Chim, Shing et al. 2008), 12 are up-regulated during the trophoblast differentiation. These include hsa-miR-373, hsa-mir-371, hsa-mir-372, hsa-mir-149, hsa-miR-34c, hsa-miR-34b, hsa-miR-135b, hsa-miR-141, hsa-miR-200b, hsa-miR-137, hsa-miR-184 and hsa-miR-337.

4.22.3 Stem cell related microRNA

Next generation sequencing has been used to study the microRNA component of human embryonic stem cells (Bar, Wyman et al. 2008; Morin, O'Connor et al. 2008). The sequencing depth used in these studies are much smaller than the amount used in my small RNA-Seq sequencing. Therefore it is expected that my dataset would yield additional information on the smallRNA transcriptome.

Using 20 reads per microRNA as cutoff, only 186 microRNAs are significantly expressed according to the dataset in Morin et al, while 331 microRNAs are

significantly expressed in mine. The 145 microRNAs which were detected exclusively in my dataset can be explained by its high sequencing depth which increases sensitivity. Despite this, the two datasets are highly comparable with 87% (164 microRNAs out of 186) expressed in Morin et. being significantly expressed in mine. The difference of sequencing depth also could be a reason for the much higher number of novel microRNA detected using my dataset compared to that of Morin *et al*.

4.23 Identification of novel small RNA

Just like in the standard RNA-Seq analysis, the small RNA-Seq dataset was used to look for novel transcribed regions (NTRs). In this case the footprints of microRNAs in the miRBASE database was used to de-mark the known expressed regions. 12,404 NTR regions were identified in day 0 and 15,145 NTRs were found in day 8. Again just like in the case of the standard RNA-Seq dataset the number of NTR regions showed a significant increase during differentiation.

Since most of the smallRNAs can be classified based on the size, the size distribution of all the identified NTRs were studied. The distribution shown in Figure 4.70, while having a large footprint from 16 nucleotides to beyond 200, shows a clear, very strong maxima of 22 nucleotides. Recall that my small RNA-seq library was created for any RNA that was 200 bases or smaller. Since the average length of microRNAs is around 22 nucleotides, this suggests that identified NTRs from the small RNA dataset are highly enriched in microRNA. Due to this observation the subsequent analysis focussed on identification of novel microRNAs.

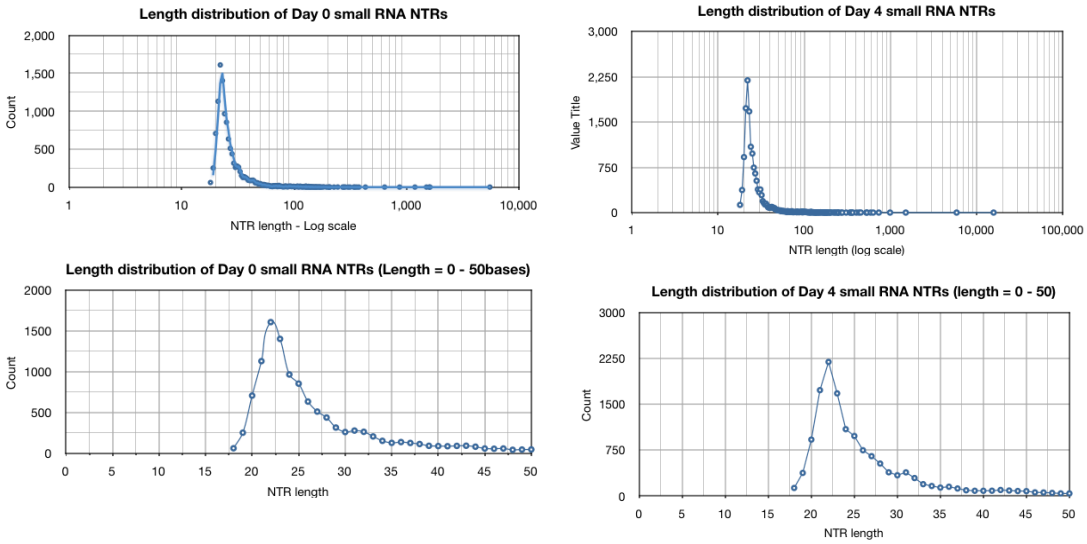


Figure 4.70 : The size distribution of small RNA NTRs expressed in day 0 and day 4.
While the distribution has a broad footprint the maxima is at 22 nucleotides - the average length of microRNAs. The bottom two graphs are shows the enlarged view of the microRNA size peak.

4.23.1 Potentially novel microRNAs

To get a reliable list of potential microRNA from all the NTRs, NTRs which map to known repeatmasker regions from UCSC genome browser, were removed. Repeatmasker contains annotations for rRNA, tRNA as well as LINEs, SINEs and LTRs described previously. By this step most of the tRNA and small ribosomal RNA are removed. Subsequently, known snoRNAs were removed from the list. To narrow down the list further, the secondary structure of these potential microRNA was analyzed using RNAfold. As mentioned in the methods section microRNAs have a unique stem loop like secondary structure with quantifiable criteria such the as number of complementary bases in the stem loop and a free energy cutoff. These criteria enabled the identification of potential microRNA with a stable stem loop secondary structure. Finally 2,360 potential microRNA from day 0 and 2,924 from the day 4 data set were identified. Among these 150 and 180, in day 0 and 4 respectively,

originated from highly conserved regions in placental mammals seen as a 80% or more overlap with mammal conservation track of UCSC. This suggest the possibility that most of the new microRNA are less conserved, meaning that they could be either primate- or human- specific.

When comparing the expression pattern of these novel microRNAs, 927 microRNAs (including highly conserved and non-conserved) showed up-regulation (2-fold or more) and 473 showed a down-regulation. This suggest that these microRNA are affected by the differentiation treatment. As for the highly conserved microRNA they too show significant differential expression. On day 0 out of the 150 highly conserved novel microRNA, 64 show a 2 fold or more up-regulation while only 8 shows down regulation of more than 2 fold.

Description	Day 0	Day 4
Total peaks (min height 1, min reads per base 100)	12,404	15,145
no overlap with RepeatMasker	6,546	8,022
no overlap with RepeatMasker + 80% or more Overlap with Mammalian highly conser	460	589
no overlap with RepeatMasker + no overlap with snRNA	6,369	7,818
no overlap with RepeatMasker + no overlap with snRNA + miRNA criteria pass	2,360	2,924
no overlap with RepeatMasker + no overlap with snRNA + miRNA criteria pass + 80%	150	180

Table 15: Novel microRNA statistics.
Note the clear increase of novel microRNAs brought about by the differentiation.

Description	Count
Up-regulated novel miRNA during differentiation more than 1SLR	927
Up-regulated novel miRNA during differentiation more than 2SLR	307
Down-regulated novel miRNA during differentiation more than 1SLR	473
Down-regulated novel miRNA during differentiation more than 2SLR	132

Table 16: Differential expression of novel microRNA.
The differential expression is measured between day 4 and day 0 during trophoblast differentiation.

4.23.2 A typical view of a known microRNA together with its folded structure

Figure 4.71 shows the UCSC view of a known microRNA and the stem loop structure it produces. Note that the peak profile contains a taller peak and a shorter one. The taller peak represents the mature microRNA sequence while the shorter peak represents the star sequence.

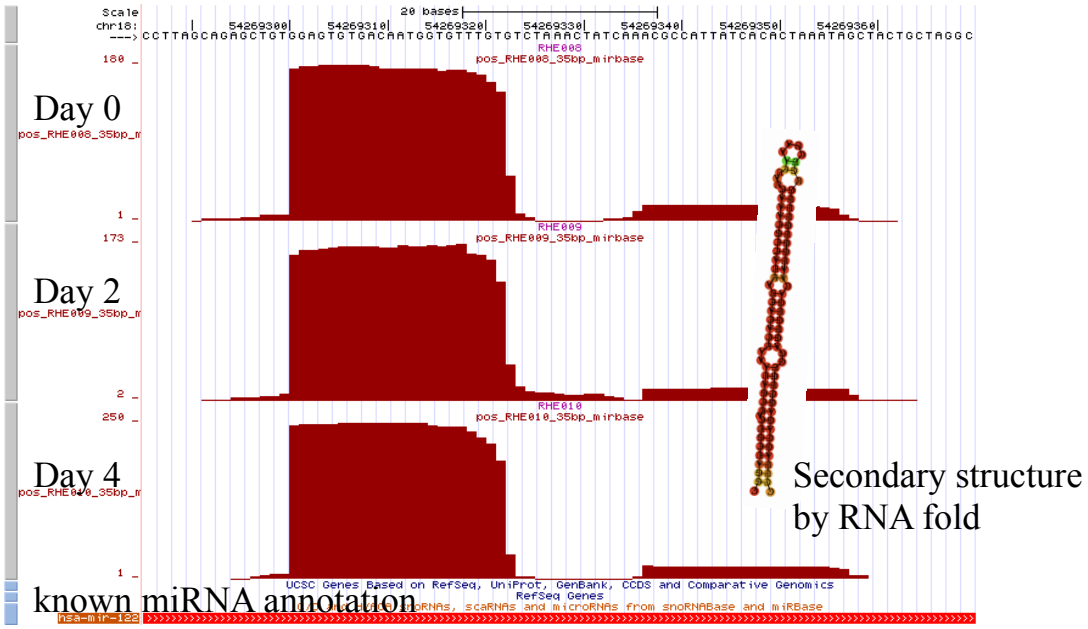


Figure 4.71: A typical UCSC view of the RNA-Seq small RNA dataset.
Here a footprint of a known microRNA is shown. The data tracks from day 0, day 2 and day 4 respectively are arranged from top to bottom. In the insert, next to the peaks, the secondary structure of of the microRNA - which shows the characteristic stem loop structure is shown. This is the format of all the small RNA-Seq related screen shots displayed in the thesis unless stated otherwise.

4.23.3 Examples of novel microRNA

As the alignment process discriminates between a known microRNA and a novel one, and since it is easier to align small RNA-Seq reads to the known microRNA footprints than to entire genome, the small peak representing the star sequence is not seen in novel microRNA. The peak seen in novel microRNA profiles represents the mature microRNA sequence.

The following section will contain examples of potentially novel microRNA which originate from introns, intergenic regions and opposite strand of a known gene. In each of the cases the predicted stem-loop structure is also shown.

4.23.4 Examples of novel miRNA which originate from the opposite strand of a known exon

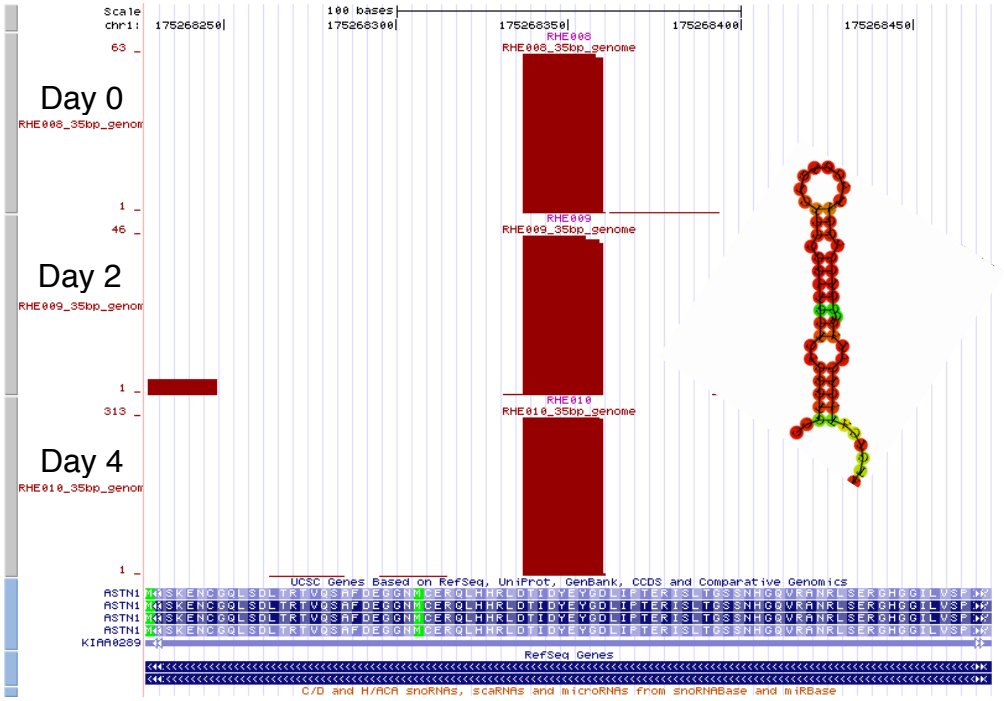


Figure 4.72: novel microRNA which originates from the opposite strand of ASTN1 gene.
This gets up-regulated during differentiation and has a stable stem loop.

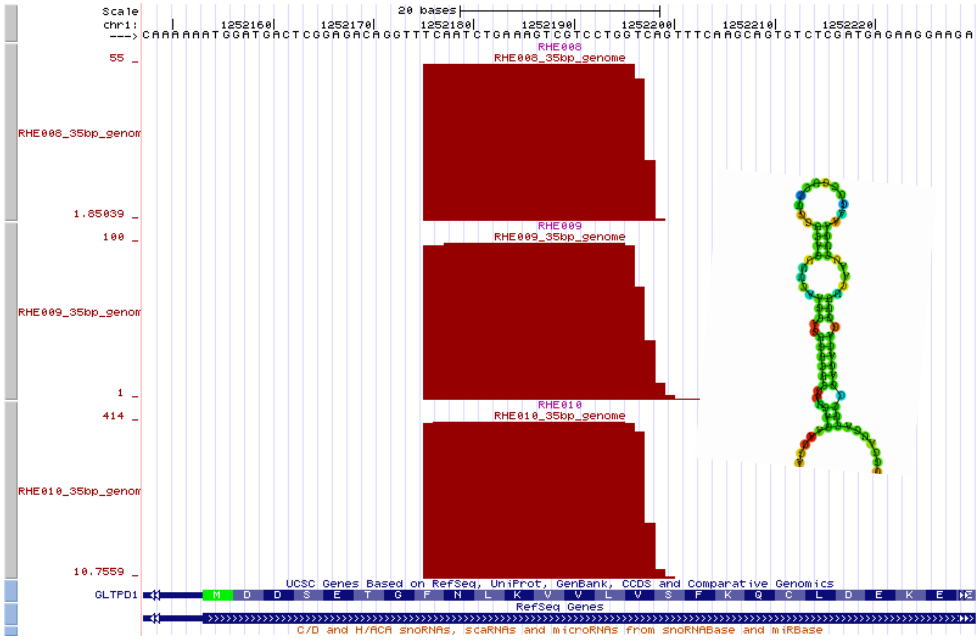


Figure 4.73: A novel microRNA coded by the opposite strand of GLTPD1.
This forms a stable stem loop and gets up-regulated during trophoblast differentiation.

4.23.5 Examples of novel miRNA which originate from an intronic region.

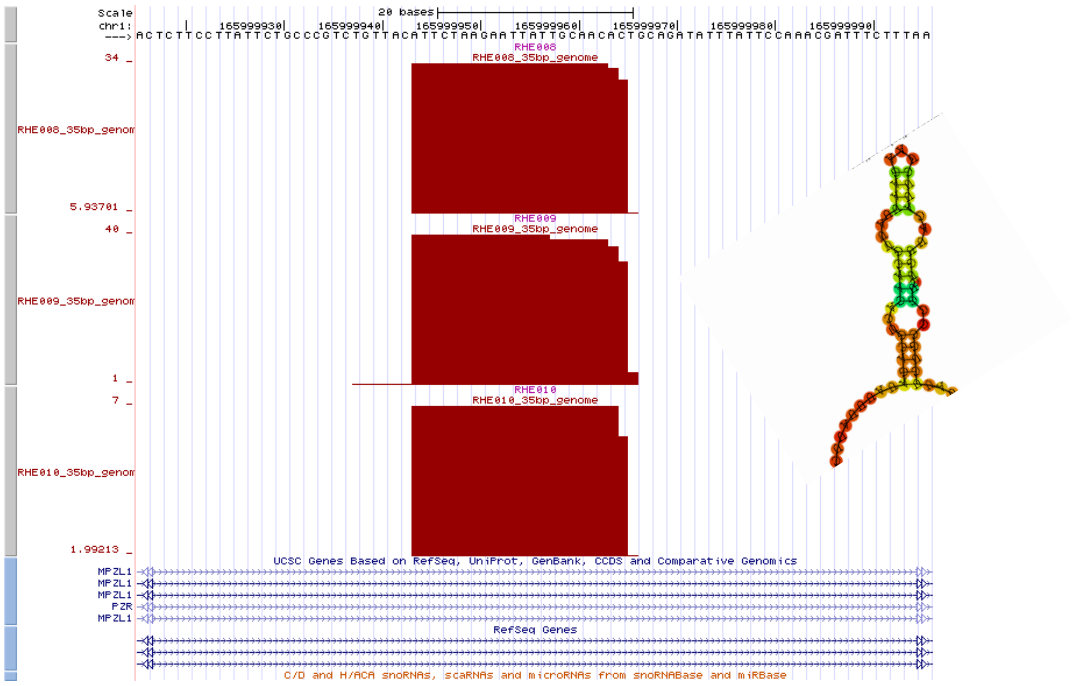


Figure 4.74: A novel microRNA coded by an intron of MPZL1 and PZR genes.
This novel microRNA gets down-regulated during differentiation.

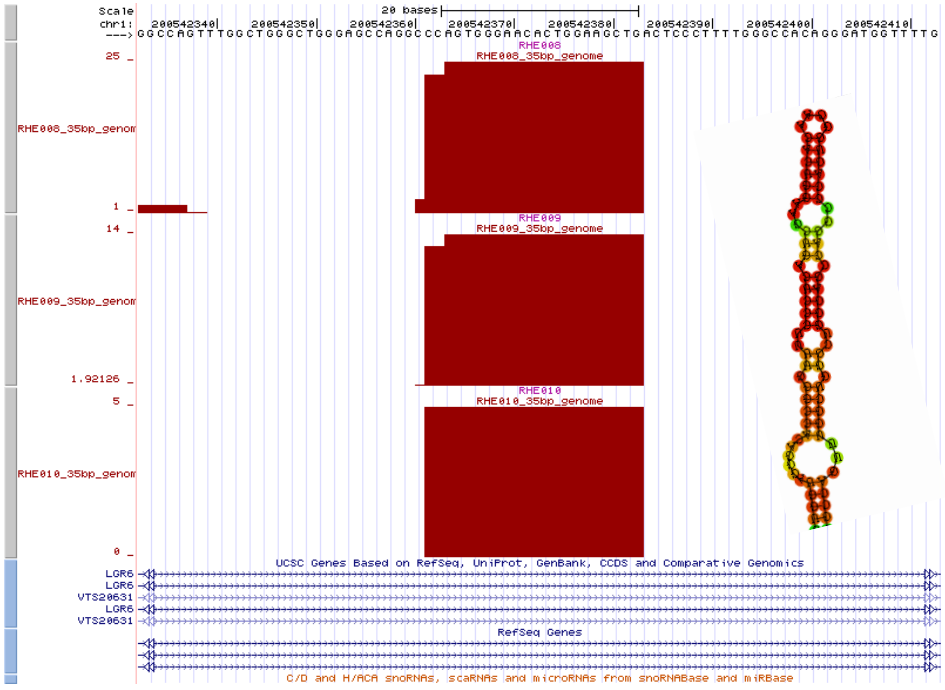


Figure 4.75: A novel microRNA which is coded by the intron sequence of LGR6 and VTS20631.
This gets down-regulated with treatment.

4.23.6 Examples of novel miRNA which originate from an intergenic region of the genome.

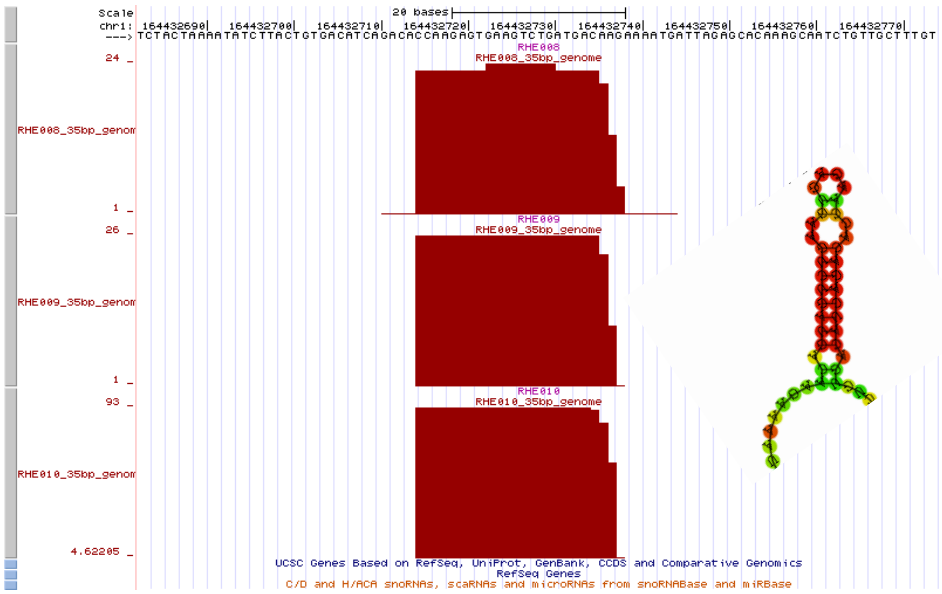


Figure 4.76: novel microRNA coded by an intergenic region.
This gets up-regulated during differentiation.

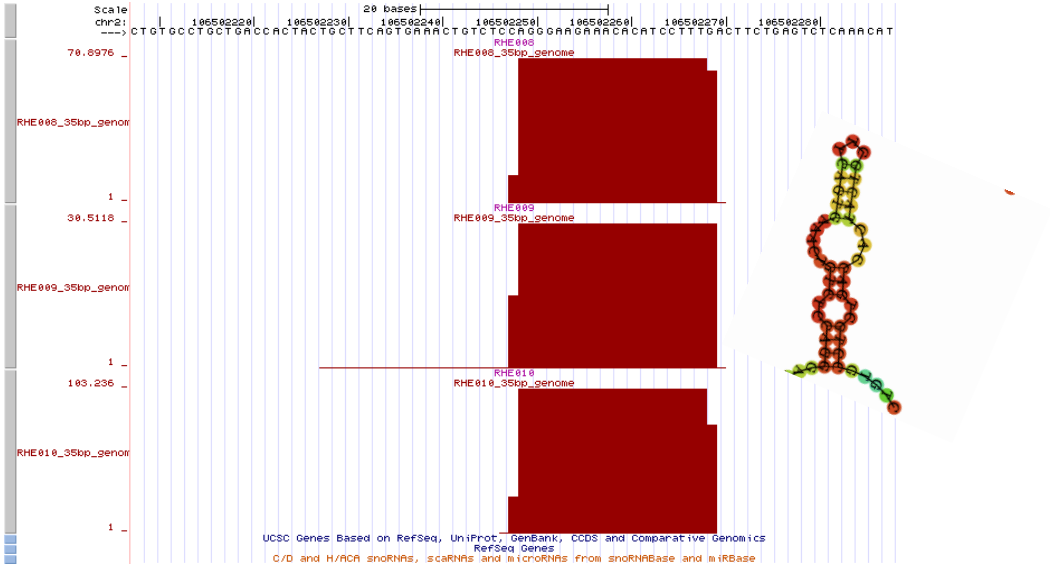


Figure 4.77: Another novel microRNA coded by an intergenic region.

4.23.7 Examples of novel miRNA which show an up-regulation during differentiation.

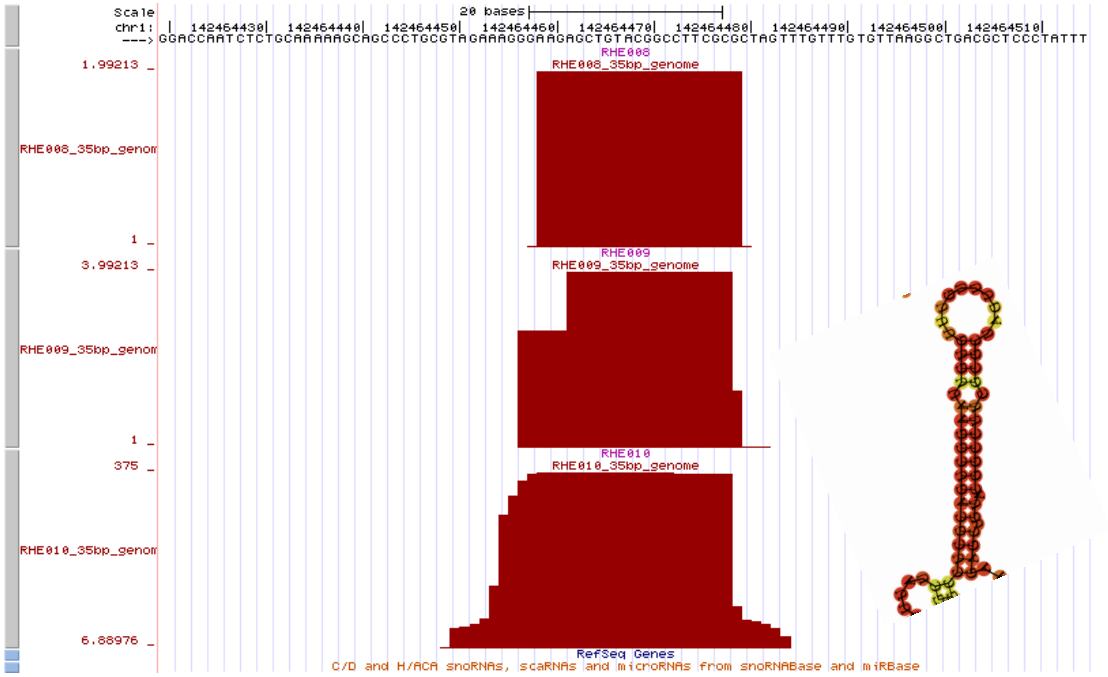


Figure 4.78: A highly up-regulated novel microRNA (~180 fold).

4.23.8 Examples of novel miRNA which show an down-regulation during differentiation.

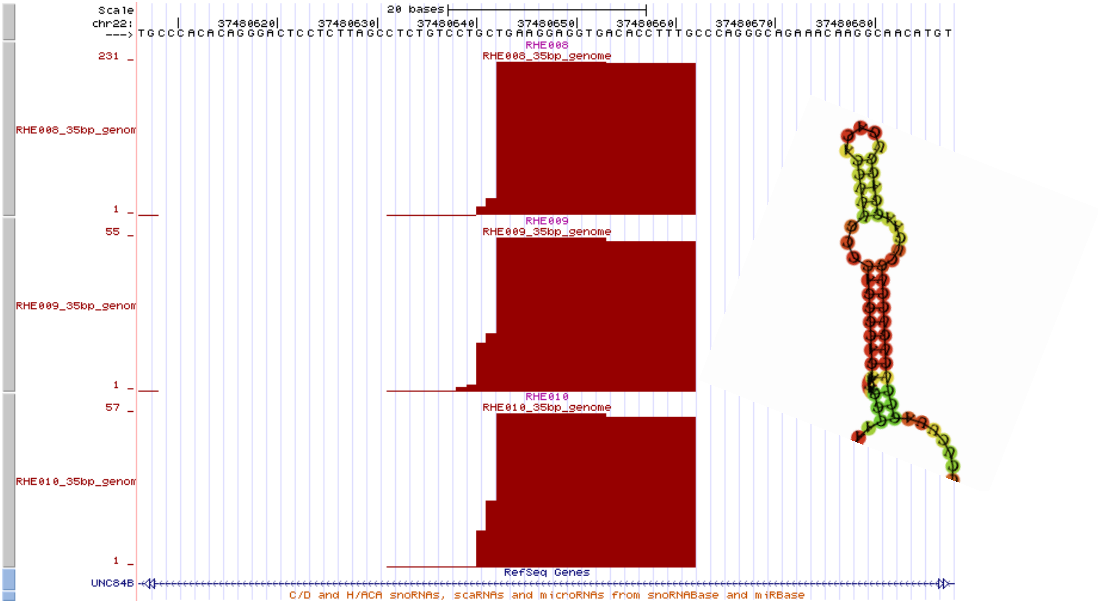


Figure 4.79: A significantly down-regulated microRNA (~ 4 fold).

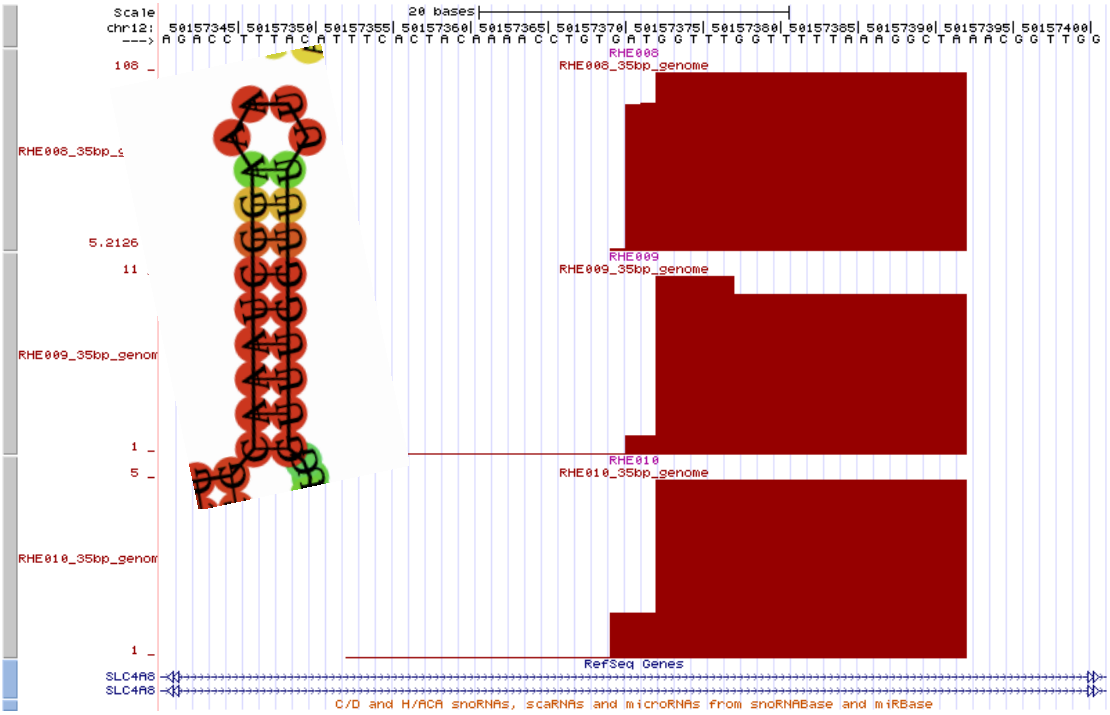


Figure 4.80: A 20 fold down-regulated novel microRNA.

4.23.9 Novel microRNA cluster

A study on the proximity of novel microRNA was carried out to identify microRNAs which exist as clusters. Only one such example was found containing 3 or more microRNA. Figure 4.81 shows the UCSC view of a known microRNA and Figure 4.82 shows the novel microRNA cluster as identified by RNA-Seq.

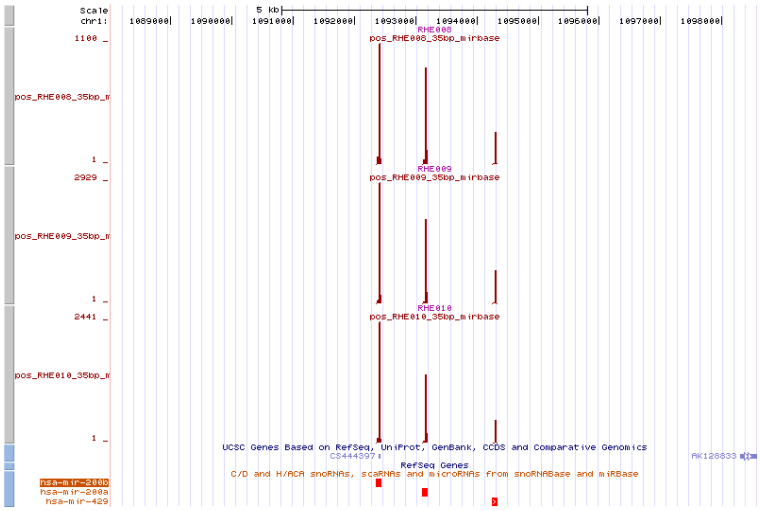


Figure 4.81: A known microRNA cluster

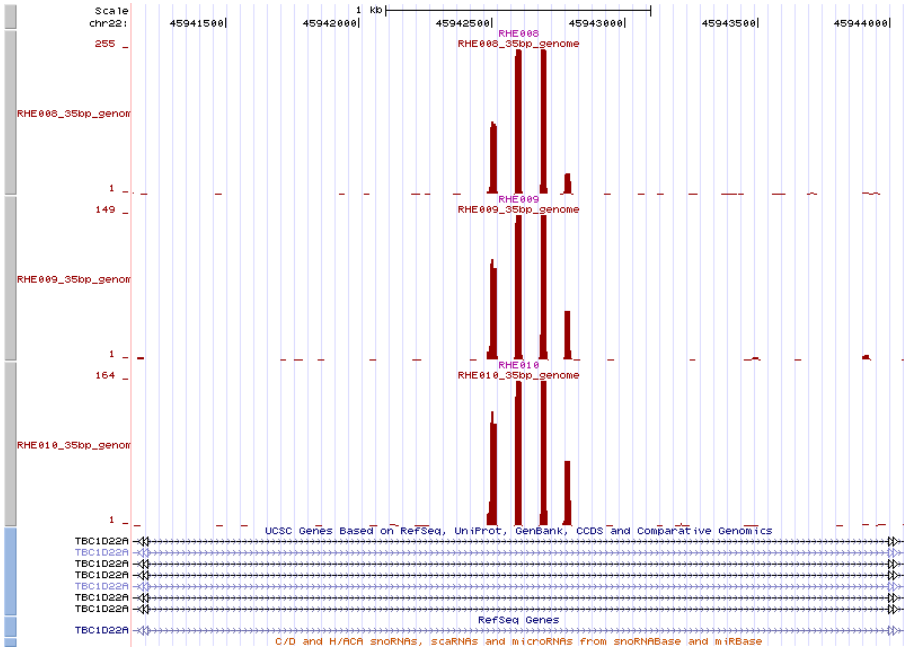


Figure 4.82: The novel microRNA cluster identified by RNA-Seq.

5.0 Discussion

Study of early human development, especially the development of the trophoblast lineage, is important not only from a fundamental biological point of view but also from a clinical perspective.

Formation of the trophectoderm is a significant biological step, as it marks the first lineage commitment and the origin of the first epithelial cell type of the new organism. Furthermore, with the help of maternal tissues, trophoblast cells go on to produce the placenta, through a complex and unique differentiation sequence.

This thesis is an attempt to describe and understand the intricate dynamics of the transcriptome during the establishment and development of the human trophoblast lineage. A scarcity or non-availability of samples, ethical issues and the lack of differentiation protocols that can provide realistic results, have limited the detailed study of this fascinating differentiation program in humans.

In this thesis, using an improved differentiation protocol which induces human embryonic stem cells to assume characteristics of the trophoblast lineage, I have attempted to create a comprehensive record of the transcriptomic dynamics during trophoblast differentiation. To study the transcriptome, I have used RNA-Seq technology to look at both poly-adenylated RNA and small RNA dynamics. The poly(A) RNA data set provides information on the mRNA portion of the transcriptome while the RNA-Seq dataset of small RNA provides insights mainly on microRNA expression.

RNA-Seq technology is superior to other traditional techniques employed for transcriptomic studies. It allows the analysis of the entire transcriptome and is extremely sensitive and accurate since it is based on sequencing. Furthermore, RNA-Seq data is not based on existing annotation, which allows the identification of novel transcriptomic phenomena.

At the time of the analysis there were no proper software to analyze an RNA-Seq data set in an in-depth manner. Therefore I developed a set of workflows / scripts which enabled the extraction of useful information from an RNA-Seq dataset. These workflows identify alternative splicing events, mutual exclusion events, extensions for existing annotations, novel transcribed regions, novel transcripts and novel microRNAs.

To calculate the RNA-Seq expression levels, I first used RPKM values, which are the counts of reads mapping to individual genes, normalized to the gene length and the sequencing depth. To check the validity of the RNA-Seq experiment I compared the RNA-Seq expression levels with microarray data for the same sample. The two datasets had a very good correlation with a coefficient of determination (R^2) value of 0.8.

Next I compared the gene expression dataset of the five time points of the SU5402+BMP4 treatment (our novel trophoblast differentiation protocol) with a large group of human tissues and cell lines, to identify the organ / tissue system which has the closest transcriptional similarity using hierarchical clustering. Reassuringly, it

turned out that the closest organ to the outcome of the differentiation protocol is the human placenta and the closest cell types to it were the cytotrophoblast and syncytiotrophoblast cells.

To further validate the data from the differentiated trophoblast transcriptome it was compared with a published microarray dataset of early human development (Zhang, Zucchelli et al. 2009). This dataset only contains information of the pre-blastocyst and blastocyst samples with no direct trophoectoderm sample. However by overlapping up-regulated genes of blastocyst formation with those of trophoblast differentiation, I was able to identify genes which are exclusively involved in trophoblast formation.

Hierarchal clustering was again used to compare the trophoblast differentiated cell transcriptome with the above mentioned published human blastocyst microarray data. Here the day 8 time point of the differentiation protocol clustered closest with the blastocyst sample and the day 0 (undifferentiated ES cells) was clustered with the 4 cell stage embryo.

Then I looked at the expression dynamics of individual genes to identify the ones which are involved in the initiation and the maintenance of the trophoblast lineage. I looked at significantly up-regulated genes in day 2, 4, 6 and 8 time points compared to day 0. These genes either were induced meaning that they were not expressed in the human embryonic stem cells (day 0), but expressed at significant levels during differentiation or were already being expressed at day 0, but were significantly up-regulated during differentiation.

Looking at the number of expressed genes in each time point, it was clear that the differentiation caused an increase in the total number of expressed genes. The increase of expressed genes was highest in day 2. On the other hand at day 2 some of the pluripotency factors are still expressed albeit at a lower level than undifferentiated ES cells. Therefore day 2 time-point represents a transition phase where pluripotency machinery are being suppressed while trophoblast inducing mechanisms are made active.

In general, the significantly up-regulated genes during trophoblast differentiation include a mix of pregnancy related hormones, placenta specific genes, genes associated with retroviral elements and genes which indicate mesenchymal to epithelial transition.

A considerable number of the genes that were induced / up-regulated during trophoblast differentiation have already been reported to be involved in placenta formation. However in most of the cases these observations have been made in other model systems, trophoblast cell lines, mature placenta or related samples (e.g the uterus). RNA-Seq data confirms the fact that these genes are involved in the early phases of the trophoblast differentiation.

The highest up-regulated gene during trophoblast differentiation is *CGA* which codes for one of the two subunits of human chorionic gonadropin - the “pregnancy hormone” and a hallmark of the trophoblast lineage. *CGA* is initially detected at day 4 and by day 8 has extremely high levels of expression (2,239 RPKM).

Syncytin 1 and Syncytin 2 are fusogenic proteins implicated in the formation of the syncytiotrophoblast. They have originated from endogenous retroviral insertions and are induced during differentiation.

The keratin genes - *KRT19*, *KRT23*, *KRT18*, which are characteristic of epithelial cells, are significantly induced around day 4 and their up-regulation is maintained throughout the differentiation.

The Mucin gene MUC15 is induced immediately during differentiation and is highly expressed up to day 8. Mucins are believed to play a vital part during implantation, creating a sticky surface for the blastocyst to attach to the uterus.

GCM1 is an essential transcription factor for placenta formation. *GCM1* expression is induced during day 4 and keeps on increasing. Interestingly the transcriptomic data suggest that the GCM1 regulatory machinery (which includes the genes *GSK3B* and *DUSP23*), evolved to keep the GCM1 levels in check, is also active.

Induction of genes during differentiation is not only limited to coding genes. The gene *H19* which codes for a long non-coding gene which is modulated by oestrogen, is immediately induced during differentiation and its expression keeps on increasing.

All the genes mentioned here and the majority of the genes being significantly up-regulated during trophoblast treatment are either placenta specific, or highly expressed in the placenta.

RNA-Seq data also identifies genes which are up to now not reported to be involved in trophoblast development. For example *CCR7* (Chemokine receptor type 7) is a gene involved in adaptive immune response. Based on transcriptomics data, it is the second most highly up-regulated gene in the differentiation protocol. This is unexpected because it is thought that the trophoblast has mechanisms to suppress the immune reaction from the mother for a successful pregnancy. However the up-regulation of *CCR7* indicates that the trophoblast is actively producing proteins which has the potential to induce an immune reaction. Published data on the human preimplantation development of actual human embryos, also show an up-regulation of *CCR7* in the human blastocyst compared to the 8 cell embryo. This suggest that *CCR7* is indeed a relevant gene for trophoblast function, and not a side effect of the differentiation. One potential explanation for this is that the human trophoblast secretes CCR7 to put the mothers adaptive immune systems to "overdrive" and therefore reducing its effectiveness. This sounds feasible, specially considering the extremely high number of endogenous retrovirus related transcripts seen to be expressed during the differentiation (discussed later), thereby providing an enormous amount of antigens - most of which are highly dynamic and not critical for the functions of the trophoblast.

VTCN1 (V-set domain containing T cell activation inhibitor 1), which is an inhibitor of innate immunity (Yi and Chen 2009). During the trophoblast differentiation VTCN1 is up-regulated 191 fold. This together with the CCR7 example could shed light on the immunosuppressive processed during early development.

RNA-Seq of small RNA indicates that the smallRNA expression is significantly influenced by the trophoblast differentiation. However the number of induced microRNA during differentiation is extremely limited. This suggests that in the case of microRNA regulation of trophoblast differentiation, changes in microRNA expression level is more important than the induction of new microRNAs.

A considerable number of microRNA up-regulated during trophoblast differentiation, has shown to be expressed in the placenta suggesting that they have a sustained functional role, from the initiation of the trophoblast formation to its later stages of development. In addition to this, there is also another set of microRNA which gets up-regulated during the differentiation protocol and have not been reported yet in literature to be expressed in the placenta. Since the samples used in literature are mostly term placentas, this subset of microRNA can be considered to be specific for the earlier stages of trophoblast differentiation.

A subset of the microRNA which gets up-regulated during trophoblast differentiation has shown to be present in the maternal serum. This opens up the possibility of using these microRNA as biomarkers to monitor the development progress of the trophoblast lineage, and therefore, to an extent, the health of the fetus.

Main issue with the currently available data on trophoblast differentiation is that most of them have been originally discovered in the mouse model and have been extrapolated into the human model. Considering the scarcity of clinical samples and ethical issues, this is understandable. However given the evolutionary difference between mouse and human, and the resulting differences in basic mechanisms of early

development, such an extrapolation could potentially lead to misinterpretations. As an example, the genes for hCG, the main marker of trophoblast lineage is absent in the mouse model. To understand the difference between early development transcriptomics of human and mouse, and to isolate transcriptomic events specific to human, I generated a RNA-Seq dataset of early mouse differentiation which included samples representing mouse 8-cell stage, E3.5 blastocyst, E4.5 blastocyst and E 4.5 inner cell mass. The inclusion of an E4.5 inner cell mass sample enables the isolation of mouse trophoblast specific transcriptomic events. This dataset was then compared to the RNA-Seq data from the human trophoblast differentiation.

Major observation of this comparison was the apparent difference in the expression levels of genes between the human and mouse systems. Based on the expression levels, it is evident that at least as far as the trophoblast lineage is concerned different molecular mechanisms participate in each of the two species.

For example the "classical" trophoblast related genes - *Gata3* and *Cdx2* are up-regulated at extremely high levels in the mouse trophoblast. In comparison, while *GATA3* is expressed in the human (RNA-Seq and pre-implantation) system and *CDX2* is up-regulated during trophoblast formation, the expression levels are much lower compared to that of the mouse system. On the other hand transcription factors such as *GCM1* which are highly expressed in the human system is only faintly expressed in mouse.

The species dependent divergence of trophoblast related biological mechanisms is a characteristic of the hourglass model of development. Briefly explained, the model

suggests that organisms of the same animal phylum have a particular stage, termed the phylotypic stage, where they look morphologically similar to each other. Beyond this stage and before this stage, development is dissimilar just like the ends of an hourglass. The major criticism leveled against this model has been that the observation is based on morphological similarity alone. However two recent papers, characterizing the conservation of gene expression pattern within fish and fly species across developmental time, have shown that the phylotypic similarities are observable at the molecular level as well (Domazet-Lošo and Tautz 2010; Kalinka, Varga et al. 2010). The phylotypic stage is considered to occur following gastrulation at approximately the early somite stage. Therefore the human and mouse transcriptomic data presented here belong to a time-point before the phylotypic stage. This explains the significant difference in the transcriptome of human and mouse seen by RNA-Seq. The area of focus of this thesis falls on a very narrow region of the hourglass model much earlier than the phylotypic stage. Concerning that particular region, the human and mouse transcriptomic differences is in agreement with the hourglass model of development. That being said, due to the lack of information of our dataset on the conserved time points of the model, the data presented in this thesis cannot be used to support the rest of the hourglass model.

Since the differences between the human and mouse shown in this thesis is in alignment with the hourglass model of development, at the developmental stage the samples belong to, it is essential to find the cause of the divergence. In other words it is important to identify what factor / factors contribute to the developmental differences at the molecular level between the mouse and the human system.

This thesis presents ample examples which show that primate-specific retroviral elements in the human genome have an important function in trophoblast development. The number of expressed retroviral elements - namely the trophoblast-specific components of SINES, LINES and HERV-K elements of LTRs - are highly increased during differentiation. This increase is consistent with the increase of overall gene expression which peaks at day 2 during trophoblast differentiation. The expression of retroviral elements is widespread throughout the genome.

Syncytin 1 and Syncytin 2, both of which have origins in retroviral genes, and involved in placenta formation, get induced during differentiation and are among the most highly up-regulated.

The genes *CYP19A1*, *EDNRB* and *PTN* are known to have promoters which have originated from retroviral insertions resulting in production of placenta specific / enriched isoforms. These genes are highly expressed during human trophoblast differentiation and the major isoform in all these cases is the one under the regulation of the retroviral promoter. This suggests that the retroviral elements have an important regulatory role in trophoblast differentiation.

One of the novel observations made during RNA-Seq analysis was where, a new unannotated exon, with origins of retroviral sequences, initiates expression during trophoblast differentiation. This is seen in a number of genes including *CLDN4*, *DHX32* and *ZBTB3*, *SCGB3A2*. In the cases where the gene is a transcription factor the effects of retroviral expression will be amplified.

The expression of retroviral elements in early development has also seen in other species including mouse. Therefore the expression of the retrotransposon elements are by no means a human specific event. However since the retrotransposon complement - both location and sequence - of each species is unique it is clear that the retrotransposon transcriptome is significantly different from each other. Because of this difference and the fact that most of them are expressed, they have the capacity to regulate and create species specific transcriptomic events.

In summary, Focusing exclusively on the transcriptome of early development, mainly in human and to a limited extent in the mouse I have catalogued the dynamics of known genes and also described novel transcriptomic phenomena. I also provide evidence for the hourglass model of development in human and mouse at the molecular level during trophoblast differentiation, and suggest that the expression of retroviral elements might be the driving force for species specific transcriptomic events.

Be it human, mouse or any other species, early development is one of the most important and biologically fascinating field of study. Considering its importance and complexity, it will take quite some time for science to be able to fully describe it. I believe that the data and information presented in this thesis will be beneficial in this regard.

6.0 Bibliography

Adriaenssens, E., S. Lottin, et al. (1999). "Steroid hormones modulate H19 gene expression in both mammary gland and uterus." Oncogene **18**(31): 4460.

Alsat, E., J. Guibourdenche, et al. (1998). "Physiological role of human placental growth hormone." Molecular and cellular endocrinology **140**(1-2): 121-127.

Anson-Cartwright, L., K. Dawson, et al. (2000). "The glial cells missing-1 protein is essential for branching morphogenesis in the chorioallantoic placenta." Nature genetics **25**(3): 311-4.

Aplin, J. and S. Kimber (2004). "Trophoblast-uterine interactions at implantation." Reprod Biol Endocrinol **2**: 48.

Arnold, S., U. Hofmann, et al. (2008). "Pivotal roles for eomesodermin during axis formation, epithelium-to-mesenchyme transition and endoderm specification in the mouse." Development **135**(3): 501.

Baczyk, D., S. Drewlo, et al. (2009). "Glial cell missing-1 transcription factor is required for the differentiation of the human trophoblast." Cell Death Differ **16**(5): 719-27.

Bar, M., S. Wyman, et al. (2008). "MicroRNA discovery and profiling in human embryonic stem cells by deep sequencing of small RNA libraries." Stem Cells **26**(10): 2496-2505.

Bartel, D. (2004). "MicroRNAs:: Genomics, Biogenesis, Mechanism, and Function." Cell **116**(2): 281-297.

Beanan, M. J. and T. D. Sargent (2000). "Regulation and function of Dlx3 in vertebrate development." Developmental Dynamics **218**(4): 545-553.

Benirschke, K., P. Kaufmann, et al. (2006). "Early Development of the Human Placenta." 42-49.

Bilban, M., P. Haslinger, et al. (2009). "Identification of novel trophoblast invasion-related genes: heme oxygenase-1 controls motility via peroxisome proliferator-activated receptor gamma." Endocrinology **150**(2): 1000-13.

Bilban, M., S. Tauber, et al. (2010). "Trophoblast invasion: Assessment of cellular models using gene expression signatures." Placenta **31**(11): 989-996.

Bischof, P., J. Aplin, et al. (2006). "Implantation of the human embryo: research lines and models." Gynecol Obstet Invest **62**(4): 206-216.

Black, S. G., F. Arnaud, et al. (2010). "Endogenous retroviruses in trophoblast differentiation and placental development." Am J Reprod Immunol **64**(4): 255-64.

Blaise, S., N. de Parseval, et al. (2003). "Genomewide screening for fusogenic human endogenous retrovirus envelopes identifies syncytin 2, a gene conserved on primate evolution." Proc Natl Acad Sci USA **100**(22): 13013-8.

Blond, J., D. Lavillette, et al. (2000). "An envelope glycoprotein of the human endogenous retrovirus HERV-W is expressed in the human placenta and fuses cells expressing the type D mammalian retrovirus receptor." Journal of Virology **74**(7): 3321.

Blond, J. L., F. Besème, et al. (1999). "Molecular characterization and placental expression of HERV-W, a new human endogenous retrovirus family." J Virol **73**(2): 1175-85.

Boyd, M., C. Bax, et al. (1993). "The human endogenous retrovirus ERV-3 is upregulated in differentiating placental trophoblast cells." Virology **196**(2): 905-909.

Burleigh, D. W., C. M. Kendziorski, et al. (2007). "Microarray analysis of BeWo and JEG3 trophoblast cell lines: identification of differentially expressed transcripts." Placenta **28**(5-6): 383-9.

Byrne, M. and C. Warner (2008). "MicroRNA expression in preimplantation mouse embryos from Ped gene positive compared to Ped gene negative mice." J Assist Reprod Genet **25**(5): 205-214.

Cai, X. and B. R. Cullen (2007). "The imprinted H19 noncoding RNA is a primary microRNA precursor." RNA **13**(3): 313-6.

Cameo, P., S. Srisuparp, et al. (2004). "Chorionic gonadotropin and uterine dialogue in the primate." Reprod Biol Endocrinol **2**(1): 50.

Carson, D. D., M. M. DeSouza, et al. (1998). "Mucin and proteoglycan functions in embryo implantation." Bioessays **20**(7): 577-83.

Carter, A. (2007). "Animal models of human placentation-a review." Placenta **28**: S41-S47.

Chang, M., D. Mukherjea, et al. (2008). "Glial Cell Missing 1 Regulates Placental Growth Factor (PGF) Gene Transcription in Human Trophoblast." Biology of Reproduction **78**(5): 841-851.

Cheynet, V., A. Ruggieri, et al. (2005). "Synthesis, assembly, and processing of the Env ERVWE1/syncytin human endogenous retroviral envelope." J Virol **79** (9): 5585-93.

Chiang, M., F. Liang, et al. (2009). "Mechanism of Hypoxia-induced GCM1 Degradation." Journal of Biological Chemistry **284**(26): 17411.

Chim, S., T. Shing, et al. (2008). "Detection and characterization of placental microRNAs in maternal plasma." Clinical chemistry **54**(3): 482.

Chim, S. S. C., T. K. F. Shing, et al. (2008). "Detection and characterization of placental microRNAs in maternal plasma." Clin Chem **54**(3): 482-90.

Ciruna, B. and J. Rossant (2001). "FGF signaling regulates mesoderm cell fate specification and morphogenetic movement at the primitive streak." Developmental Cell **1**(1): 37-49.

Cohen, C. J., W. M. Lock, et al. (2009). "Endogenous retroviral LTRs as promoters for human genes: a critical assessment." Gene **448**(2): 105-14.

Comijn, J., G. Berx, et al. (2001). "The two-handed E box binding zinc finger protein SIP1 downregulates E-cadherin and induces invasion." Molecular cell **7**(6): 1267-1278.

Conley, A. and M. Hinshelwood (2001). "Mammalian aromatases." Reproduction **121** (5): 685-95.

Cross, J., D. Baczyk, et al. (2003). "Genes, development and evolution of the placenta." Placenta **24**(2-3): 123-130.

De Parseval, N., V. Lazar, et al. (2003). "Survey of human genes of retroviral origin: identification and transcriptome of the genes with coding capacity for complete envelope proteins." Journal of virology **77**(19): 10414.

Domazet-Lošo, T. and D. Tautz (2010). "A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns." Nature **468**(7325): 815-818.

Donovan, A., A. Brownlie, et al. (2000). "Positional cloning of zebrafish ferroportin1 identifies a conserved vertebrate iron exporter." Nature **403**(6771): 776-81.

Eger, A., K. Aigner, et al. (2005). "DeltaEF1 is a transcriptional repressor of E-cadherin and regulates epithelial plasticity in breast cancer cells." Oncogene **24**(14): 2375-2385.

Enders, A. (2000). "Trophoblast-uterine interactions in the first days of implantation: models for the study of implantation events in the human." Semin Reprod Med **18**(3): 255-264.

Enquobahrie, D., D. Abetew, et al. (2010). "Placental microRNA expression in pregnancies complicated by preeclampsia." American Journal of ….

Enright, A. J., B. John, et al. (2003). "MicroRNA targets in Drosophila." Genome Biol **5**(1): R1.

Esnault, C., J. Maestre, et al. (2000). "Human LINE retrotransposons generate processed pseudogenes." Nature genetics **24**(4): 363-367.

Evseenko, D., J. Paxton, et al. (2007). "The xenobiotic transporter ABCG2 plays a novel role in differentiation of trophoblast-like BeWo cells." Placenta **28**: S116-S120.

Foshay, K. and G. Gallicano (2009). "miR-17 family miRNAs are expressed during early mammalian development and regulate stem cell differentiation." Developmental Biology **326**(2): 431-443.

Fürbass, R., R. Selimyan, et al. (2007). "DNA methylation and chromatin accessibility of the proximalCyp19 promoter region 1.5/2 correlate with expression levels in sheep placentomes." Mol. Reprod. Dev. **75**(1): 1-7.

Garcia-Perez, J., M. Marchetto, et al. (2007). "LINE-1 retrotransposition in human embryonic stem cells." Human molecular genetics **16**(13): 1569.

Ge, X., S. Yamamoto, et al. (2005). "Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues." Genomics **86** (2): 127-41.

Genbacev, O. and R. Miller (2000). "Post-implantation differentiation and proliferation of cytotrophoblast cells: In vitro models: A review." Placenta **21**: S45-S49.

Georgiades, P. and J. Rossant (2006). "Ets2 is necessary in trophoblast for normal embryonic anteroposterior axis development." Development **133**(6): 1059.

Gerami-Naini, B., O. V. Dovzhenko, et al. (2004). "Trophoblast differentiation in embryoid bodies derived from human embryonic stem cells." Endocrinology **145**(4): 1517-24.

Gilad, S., E. Meiri, et al. (2008). "Serum microRNAs are promising novel biomarkers." PLoS ONE **3**(9): e3148.

Goldenberg, R. and W. Andrews (1996). "Intrauterine infection and why preterm prevention programs have failed." American Journal of Public Health **86**(6): 781.

Golos, T. G., M. Giakoumopoulos, et al. (2010). "Embryonic stem cells as models of trophoblast differentiation: progress, opportunities, and limitations." Reproduction **140**(1): 3-9.

Golos, T. G., L. M. Pollastrini, et al. (2006). "Human embryonic stem cells as a model for trophoblast differentiation." Semin Reprod Med **24**(5): 314-21.

Goodier, J. L. and H. H. Kazazian (2008). "Retrotransposons revisited: the restraint and rehabilitation of parasites." Cell **135**(1): 23-35.

Gregory, P. A., A. G. Bert, et al. (2008). "The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1." Nat Cell Biol **10**(5): 593-601.

Griffiths, D. (2001). "Endogenous retroviruses in the human genome sequence." Genome Biol **2**(6): 1017.1—1017.5.

Guo, G., M. Huss, et al. (2010). "Resolution of Cell Fate Decisions Revealed by Single-Cell Gene Expression Analysis from Zygote to Blastocyst." Developmental Cell **18**(4): 675-685.

Hallast, P., J. Saarela, et al. (2008). "High divergence in primate-specific duplicated regions: Human and chimpanzee Chorionic Gonadotropin Beta genes." BMC Evol Biol **8**(1): 195.

Hanahan, D. and R. Weinberg (2000). "The hallmarks of cancer." Cell **100**(1): 57-70.

Handwerger, S. (2009). "New insights into the regulation of human cytotrophoblast cell differentiation." Molecular and Cellular Endocrinology.

Hay, D. C., L. Sutherland, et al. (2004). "Oct-4 knockdown induces similar patterns of endoderm and trophoblast differentiation markers in human and mouse embryonic stem cells." Stem Cells **22**(2): 225-35.

Hemberger, M., R. Udayashankar, et al. (2010). "ELF5-enforced transcriptional networks define an epigenetically regulated trophoblast stem cell compartment in the human placenta." Human Molecular Genetics **19**(12): 2456.

Hofacker, I. L. (2003). "Vienna RNA secondary structure server." Nucleic Acids Research **31**(13): 3429.

Home, P., S. Ray, et al. (2009). "GATA3 Is Selectively Expressed in the Trophectoderm of Peri-implantation Embryo and Directly Regulates Cdx2 Gene Expression." Journal of Biological Chemistry **284**(42): 28729-28737.

Hough, S. R., I. Clements, et al. (2006). "Differentiation of Mouse Embryonic Stem Cells after RNA Interference-Mediated Silencing of OCT4 and Nanog." Stem Cells **24**(6): 1467-1475.

Hurteau, G. J., J. A. Carlson, et al. (2007). "Overexpression of the MicroRNA hsa-miR-200c Leads to Reduced Expression of Transcription Factor 8 and Increased Expression of E-Cadherin." Cancer Research **67**(17): 7972-7976.

Hustin, J. and J. P. Schaaps (1987). "Echographic [corrected] and anatomic studies of the maternotrophoblastic border during the first trimester of pregnancy." Am J Obstet Gynecol **157**(1): 162-8.

Ivanova, N., R. Dobrin, et al. (2006). "Dissecting self-renewal in stem cells with RNA interference." Nature **442**(7102): 533-538.

Kalinka, A., K. Varga, et al. (2010). "Gene expression divergence recapitulates the developmental hourglass model." Nature.

Kalluri, R. (2009). "EMT: when epithelial cells decide to become mesenchymal-like cells." J Clin Invest **119**(6): 1417-9.

Kalluri, R. and R. A. Weinberg (2009). "The basics of epithelial-mesenchymal transition." J Clin Invest **119**(6): 1420-8.

Kamat, A. and C. Mendelson (2001). "Identification of the regulatory regions of the human aromatase P450 (CYP19) gene involved in placenta-specific expression* 1." The Journal of Steroid Biochemistry and Molecular Biology **79**(1-5): 173-180.

Kanadia, R. N. and C. L. Cepko (2010). "Alternative splicing produces high levels of noncoding isoforms of bHLH transcription factors during development." Genes Dev **24**(3): 229-34.

Karaulanov, E., W. Knöchel, et al. (2004). "Transcriptional regulation of BMP4 synexpression in transgenic Xenopus." EMBO J **23**(4): 844-856.

Kent, W. J., C. W. Sugnet, et al. (2002). "The human genome browser at UCSC." Genome Research **12**(6): 996.

Khoo, N., J. Bechberger, et al. (1998). "SV40 Tag transformation of the normal invasive trophoblast results in a premalignant phenotype. I. Mechanisms responsible for hyperinvasivess and resistance to anti-invasive action of TGF&bgr." International Journal of Cancer **77**(3): 429-439.

Kim, K., M. Kugler, et al. (2006). "Alveolar epithelial cell mesenchymal transition develops in vivo during pulmonary fibrosis and is regulated by the extracellular matrix." Proceedings of the National Academy of Sciences **103** (35): 13180.

King, A., L. Thomas, et al. (2000). "Cell culture models of trophoblast II: trophoblast cell lines--a workshop report." Placenta **21 Suppl A**: S113-9.

Kolwankar, D. (2005). "EXPRESSION AND FUNCTION OF ABCB1 AND ABCG2 IN HUMAN PLACENTAL TISSUE." Drug Metabolism and Disposition **33** (4): 524-529.

Krzywinski, M., J. Schein, et al. (2009). "Circos: an information aesthetic for comparative genomics." Genome Research **19**(9): 1639-45.

Kuckenberg, P., S. Buhl, et al. (2010). "The Transcription Factor TCFAP2C/AP-2 Cooperates with CDX2 To Maintain Trophectoderm Formation." Molecular and Cellular Biology **30**(13): 3310-3320.

Kunarso, G., N.-Y. Chia, et al. (2010). "Transposable elements have rewired the core regulatory network of human embryonic stem cells." Nat Genet **42**(7): 631-4.

Lee, Y., C. Ahn, et al. (2003). "The nuclear RNase III Drosha initiates microRNA processing." Nature **425**(6956): 415-419.

Lewis, M. A. and K. P. Steel (2010). "MicroRNAs in mouse development and disease." Seminars in Cell and Developmental Biology **21**(7): 774-780.

Lin, F., C. Chang, et al. (2010). "Dual-specificity phosphatase 23 mediates GCM1 dephosphorylation and activation." Nucleic Acids Research.

Lin, L., B. Xu, et al. (2000). "The cellular mechanism by which the human endogenous retrovirus ERV-3 env gene affects proliferation and differentiation in a human placental trophoblast model, BeWo." Placenta **21**(1): 73-78.

Lin, L., B. Xu, et al. (1999). "Expression of endogenous retrovirus ERV-3 induces differentiation in BeWo, a choriocarcinoma model of human placental trophoblast." Placenta **20**(1): 109-18.

Lindsley, R., J. Gill, et al. (2008). "Mesp1 coordinately regulates cardiovascular fate restriction and epithelial-mesenchymal transition in differentiating ESCs." Cell Stem Cell **3**(1): 55-68.

Liu, P., M. Wakamiya, et al. (1999). "Requirement for Wnt3 in vertebrate axis formation." Nature genetics **22**(4): 361-365.

Liu, Y., O. Dovzhenko, et al. (2004). "Maintenance of pluripotency in human embryonic stem cells stably over-expressing enhanced green fluorescent protein." Stem Cells and Development **13**(6): 636-645.

Loh, Y., Q. Wu, et al. (2006). "The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells." Nature genetics **38**(4): 431-440.
Ma, G., M. Roth, et al. (1997). "GATA-2 and GATA-3 regulate trophoblast-specific gene expression in vivo." Development **124**(4): 907.
Macleod , J. N., I. Worsley, et al. (1991). "Human Growth Hormone-Variant Is a Biologically Active Somatogen and Lactogen." Endocrinology **128**(3): 1298-1302.
Mangeney, M., M. Renard, et al. (2007). "Placental syncytins: genetic disjunction between the fusogenic and immunosuppressive activity of retroviral envelope proteins." Proceedings of the National Academy of Sciences **104**(51): 20534.
Mao, J., D. M. McKean, et al. (2010). "The iron exporter ferroportin 1 is essential for development of the mouse embryo, forebrain patterning and neural tube closure." Development **137**(18): 3079.
Mayeur, S., M. Silhol, et al. (2010). "Placental BDNF/TrkB Signaling System is Modulated by Fetal Growth Disturbances in Rat and Human." Placenta **31**(9): 785-791.
McConnell, J., L. Petrie, et al. (2005). "Eomesodermin is expressed in mouse oocytes and pre-implantation embryos." Mol Reprod Dev **71**(4): 399-404.
Medstrand, P., J. Landry, et al. (2001). "Long terminal repeats are used as alternative promoters for the endothelin B receptor and apolipoprotein CI genes in humans." Journal of Biological Chemistry **276**(3): 1896.
Mi, S., X. Lee, et al. (2000). "Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis." Nature **403**(6771): 785-9.
Mohamed, O. A., M. Jonnaert, et al. (2005). "Uterine Wnt/beta-catenin signaling is required for implantation." Proceedings of the National Academy of Sciences of the United States of America **102**(24): 8579-84.
Morin, R., M. O'Connor, et al. (2008). "Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells." Genome research **18**(4): 610.
Mortazavi, A., B. A. Williams, et al. (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq." Nature Methods **5**(7): 621-628.
Murchison, E., P. Stein, et al. (2007). "Critical roles for Dicer in the female germline." Genes & Development **21**(6): 682.
Nanna, R. (2007). "miR-200b mediates post-transcriptional repression of ZFHX1B." RNA **13**(8): 1172.
Nelson, P. N., P. R. Carnegie, et al. (2003). "Demystified. Human endogenous retroviruses." Mol Pathol **56**(1): 11-8.
Ng, R., W. Dean, et al. (2008). "Epigenetic restriction of embryonic cell lineage fate by methylation of Elf5." Nature cell biology **10**(11): 1280-1290.
Nieto, M. (2002). "The snail superfamily of zinc-finger transcription factors." Nat. Rev. Mol. Cell Biol. **3**(3): 155-166.
Nilson, J. H., J. A. Bokar, et al. (1991). "Different combinations of regulatory elements may explain why placenta-specific expression of the glycoprotein hormone alpha-subunit gene occurs only in primates and horses." Biol Reprod **44**(2): 231-7.
Nishioka, N., K.-i. Inoue, et al. (2010). "The Hippo Signaling Pathway Components Lats and Yap Pattern Tead4 Activity to Distinguish Mouse Trophectoderm from Inner Cell Mass." Developmental Cell **16**(3): 398-410.

Nishioka, N., S. Yamamoto, et al. (2008). "Tead4 is required for specification of trophectoderm in pre-implantation mouse embryos." Mechanisms of Development **125**(3-4): 270-283.

Niwa, H., J. Miyazaki, et al. (2000). "Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells." Nature genetics **24**(4): 372-6.

Niwa, H., Y. Toyooka, et al. (2005). "Interaction between Oct3/4 and Cdx2 determines trophectoderm differentiation." Cell **123**(5): 917-929.

Nony, P., R. Hannon, et al. (1998). "Alternate promoters and developmental modulation of expression of the chicken GATA-2 gene in hematopoietic progenitor cells." Journal of Biological Chemistry **273**(49): 32910.

Ohira, K., K. J. Homma, et al. (2006). "TrkB-T1 regulates the RhoA signaling and actin cytoskeleton in glioma cells." Biochem Biophys Res Commun **342**(3): 867-74.

Orr-Urtreger, A., M. T. Bedford, et al. (1993). "Developmental localization of the splicing alternatives of fibroblast growth factor receptor-2 (FGFR2)." Developmental Biology **158**(2): 475-86.

Pan, X., N. Minegishi, et al. (2000). "Identification of human GATA-2 gene distal IS exon and its expression in hematopoietic stem cell fractions." Journal of Biochemistry **127**(1): 105.

Pera, M. F., J. Andrade, et al. (2004). "Regulation of human embryonic stem cell differentiation by BMP-2 and its antagonist noggin." J Cell Sci **117**(Pt 7): 1269-80.

Perea-Gomez, A., F. Vella, et al. (2002). "Nodal antagonists in the anterior visceral endoderm prevent the formation of multiple primitive streaks." Developmental Cell **3**(5): 745-756.

Pierce, J. G. and T. F. Parsons (1981). "Glycoprotein hormones: structure and function." Annu Rev Biochem **50**: 465-95.

Popperl, H., C. Schmidt, et al. (1997). "Misexpression of Cwnt8C in the mouse induces an ectopic embryonic axis and causes a truncation of the anterior neuroectoderm." Development **124**(15): 2997.

Potenta, S., E. Zeisberg, et al. (2008). "The role of endothelial-to-mesenchymal transition in cancer progression." Br J Cancer **99**(9): 1375-1379.

Ralston, A., B. Cox, et al. (2010). "Gata3 regulates trophoblast development downstream of Tead4 and in parallel to Cdx2." Development **137**(3): 395.

Ralston, A. and J. Rossant (2008). "Cdx2 acts downstream of cell polarization to cell-autonomously promote trophectoderm fate in the early mouse embryo." Developmental Biology **313**(2): 614-629.

Rasmussen, H. B. and J. Clausen (1998). "Large number of polymorphic nucleotides and a termination codon in the env gene of the endogenous human retrovirus ERV3." Disease Markers **14**(3): 127-133.

Rawn, S. M. and J. C. Cross (2008). "The Evolution, Regulation, and Function of Placenta-Specific Genes." Cell and Developmental Biology **24**(1): 159-181.

Renard, M., P. F. Varela, et al. (2005). "Crystal structure of a pivotal domain of human syncytin-2, a 40 million years old endogenous retrovirus fusogenic envelope gene captured by primates." J Mol Biol **352**(5): 1029-34.

Reubinoff, B. E., M. F. Pera, et al. (2000). "Embryonic stem cell lines from human blastocysts: somatic differentiation in vitro." Nat Biotechnol **18**(4): 399-404.

Riley, P., L. Anaon-Cartwight, et al. (1998). "The Hand1 bHLH transcription factor is essential for placentation and cardiac morphogenesis." Nature genetics **18**(3): 271-275.

Rote, N. S., S. Chakrabarti, et al. (2004). "The role of human endogenous retroviruses in trophoblast differentiation and placental development." Placenta **25**(8-9): 673-83.

Russ, A., S. Wattler, et al. (2000). "Eomesodermin is required for mouse trophoblast development and mesoderm formation." Nature **404**(6773): 95-99.

Sakurai, T., M. Yanagisawa, et al. (1990). "Cloning of a cDNA encoding a non-isopeptide-selective subtype of the endothelin receptor."

Sanchez-Sanchez, N., L. Riol-Blanco, et al. (2006). "The multiple personalities of the chemokine receptor CCR7 in dendritic cells." The Journal of Immunology **176** (9): 5153.

Schulte, A., S. Lai, et al. (1996). "Human trophoblast and choriocarcinoma expression of the growth factor pleiotrophin attributable to germ-line insertion of an endogenous retrovirus." Proceedings of the National Academy of Sciences of the United States of America **93**(25): 14759.

Schulz, L. C., T. Ezashi, et al. (2008). "Human embryonic stem cells as models for trophoblast differentiation." Placenta **29 Suppl A**: S10-6.

Senner, C. E. and M. Hemberger (2010). "Regulation of early trophoblast differentiation - Lessons from the mouse." Placenta.

Shiokawa, S. (2002). "Small Guanosine Triphospatase RhoA and Rho-Associated Kinase as Regulators of Trophoblast Migration." Journal of Clinical Endocrinology & Metabolism **87**(12): 5808-5816.

Shiverick, K. T., A. King, et al. (2001). "Cell culture models of human trophoblast II: trophoblast cell lines--a workshop report." Placenta **22 Suppl A**: S104-6.

Shyu, M.-K., M.-C. Lin, et al. (2007). "Mucin 15 is expressed in human placenta and suppresses invasion of trophoblast-like cells in vitro." Human Reproduction **22**(10): 2723-2732.

Simpson, E. R., M. S. Mahendroo, et al. (1994). "Aromatase Cytochrome P450, The Enzyme Responsible for Estrogen Biosynthesis." Endocr Rev **15**(3): 342-355.

Skaper, S. D. (2008). "The biology of neurotrophins, signalling pathways, and functional peptide mimetics of neurotrophins and their receptors." CNS Neurol Disord Drug Targets **7**(1): 46-62.

Skromne, I. and C. Stern (2001). "Interactions between Wnt and Vg1 signalling pathways initiate primitive streak formation in the chick embryo." Development **128**(15): 2915.

Solter, D. and B. B. Knowles (1975). "Immunosurgery of mouse blastocyst." Proceedings of the National Academy of Sciences of the United States of America **72**(12): 5099.

Strumpf, D., C. Mao, et al. (2005). "Cdx2 is required for correct cell fate specification and differentiation of trophectoderm in the mouse blastocyst." Development **132**(9): 2093.

Sturn, A., J. Quackenbush, et al. (2002). "Genesis: cluster analysis of microarray data." Bioinformatics **18**(1): 207-8.

Su, Y., S. Liebhaber, et al. (2000). "The human growth hormone gene cluster locus control region supports position-independent pituitary-and placenta-specific

expression in the transgenic mouse." Journal of Biological Chemistry **275**(11): 7902.

Szatmari, I. (2006). "Peroxisome Proliferator-activated Receptor -regulated ABCG2 Expression Confers Cytoprotection to Human Dendritic Cells." Journal of Biological Chemistry **281**(33): 23812-23823.

Takeda, K., K. Noguchi, et al. (1997). "Targeted disruption of the mouse Stat3 gene leads to early embryonic lethality." Proceedings of the National Academy of Sciences of the United States of America **94**(8): 3801.

Tang, F., C. Barbacioru, et al. (2010). "Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis." Cell Stem Cell **6** (5): 468-78.

Tang, F., M. Kaneda, et al. (2007). "Maternal microRNAs are essential for mouse zygotic development." Genes & Development **21**(6): 644.

Tapia-Arancibia, L., F. Rage, et al. (2004). "Physiology of BDNF: focus on hypothalamic function." Front Neuroendocrinol **25**(2): 77-107.

Terauchi, M., H. Koi, et al. (2003). "Placental extravillous cytotrophoblasts persistently express class I major histocompatibility complex molecules after human cytomegalovirus infection." Journal of virology **77**(15): 8187.

Than, N. G., R. Romero, et al. (2009). "A primate subfamily of galectins expressed at the maternal–fetal interface that promote immune cell death." Proceedings of the National Academy of Sciences **106**(24): 9731.

Thiery, J. (2002). "Epithelial–mesenchymal transitions in tumour progression." Nat Rev Cancer **2**(6): 442-454.

Thomson, J., J. Itskovitz-Eldor, et al. (1998). "Embryonic stem cell lines derived from human blastocysts." Science **282**(5391): 1145.

Thomson, J. A., J. Kalishman, et al. (1995). "Isolation of a primate embryonic stem cell line." Proc Natl Acad Sci U S A **92**(17): 7844-8.

Titaley, C. R., M. J. Dibley, et al. (2010). "Iron and folic acid supplements and reduced early neonatal deaths in Indonesia." Bull. World Health Organ. **88**(7): 500-8.

van de Lagemaat, L., J. Landry, et al. (2003). "Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions." Trends in Genetics **19**(10): 530-536.

Velkey, J. M. and K. S. O'Shea (2003). "Oct4 RNA interference induces trophectoderm differentiation in mouse embryonic stem cells." Genesis **37**(1): 18-24.

Venables, P., S. Brookes, et al. (1995). "Abundance of an endogenous retroviral envelope protein in placental trophoblasts suggests a biological function." Virology **211**(2): 589-592.

Wang, E. T., R. Sandberg, et al. (2008). "Alternative isoform regulation in human tissue transcriptomes." Nature **456**(7221): 470-6.

Warzecha, C. C., T. K. Sato, et al. (2009). "ESRP1 and ESRP2 Are Epithelial Cell-Type-Specific Regulators of FGFR2 Splicing." Molecular Cell **33**(5): 591-601.

Wen, F., J. Tynan, et al. (2007). "Ets2 is required for trophoblast stem cell self-renewal." Developmental Biology **312**(1): 284-299.

Wenlong, L. (2008). "A Human Embryonic Stem Cell–based Model of Trophoblast Formation." Thesis, National University of Singapore.

Xu, N., T. Papagiannakopoulos, et al. (2009). "MicroRNA-145 regulates OCT4, SOX2, and KLF4 and represses pluripotency in human embryonic stem cells." Cell **137**(4): 647-58.

Xu, R., X. Chen, et al. (2002). "BMP4 initiates human embryonic stem cell differentiation to trophoblast." Nat Biotechnol **20**(12): 1261-1264.

Xu, X., M. Weinstein, et al. (1998). "Fibroblast growth factor receptor 2 (FGFR2)-mediated reciprocal regulation loop between FGF8 and FGF10 is essential for limb induction." Development **125**(4): 753.

Yagi, R., M. Kohn, et al. (2007). "Transcription factor TEAD4 specifies the trophectoderm lineage at the beginning of mammalian development." Development **134**(21): 3827.

Yamada, K., H. Ogawa, et al. (1999). "A GCM motif protein is involved in placenta-specific expression of human aromatase gene." Journal of Biological Chemistry **274**(45): 32279.

Yi, K. H. and L. Chen (2009). "Fine tuning the immune response through B7-H3 and B7-H4." Immunol Rev **229**(1): 145-51.

Yu, C. (2002). "GCMa Regulates the Syncytin-mediated Trophoblastic Fusion." Journal of Biological Chemistry **277**(51): 50062-50068.

Zeisberg, E. M., O. Tarnavski, et al. (2007). "Endothelial-to-mesenchymal transition contributes to cardiac fibrosis." Nat Med **13**(8): 952-61.

Zeisberg, M. and E. G. Neilson (2009). "Biomarkers for epithelial-mesenchymal transitions." J Clin Invest **119**(6): 1429-37.

Zhang, P., M. Zucchelli, et al. (2009). "Transcriptome Profiling of Human Pre-Implantation Development." PloS one. 11.e784

Appendix I : Python code for workflows

RPKM calculation

Input file:
The default exon counts file produced by ABI bioscope pipeline

Command:
python counts2RPKM.py <bioscope counts file> <sequencing depth>

Code:
counts2RPKM.py

```
#this script takes in a counts file and produces RPKM values for individual genes.
from __future__ import division
import sys

fName = sys.argv[1]
seqDepth = int(sys.argv[2])
countsDic = {}
lengthDic = {}

print seqDepth
for line in open(fName,'r'):
    temp = line.strip(`).split('\t')
    if temp[2] == 'exon':
        chr = temp[1]
        start = int(temp[3])
        stop = int(temp[4])
        length = stop - start
        hits = int(temp[5])
        geneID = temp[8].split('"')[1].split('"')[0]
        if geneID[-1:]=='P':
            geneID = geneID[:-1]
        #print line.split()
        #print chr,start,stop,hits,geneID

        if geneID in countsDic:
                    #geneId has already been added


            countsDic[geneID] = countsDic[geneID]+hits
            lengthDic[geneID] = lengthDic[geneID]+length
        else:
            countsDic[geneID] = hits
            lengthDic[geneID] = length
k = countsDic.keys()
k.sort()

for gene in k:
    totalCounts = countsDic[gene]
    totalLength = lengthDic[gene]/2
    if totalCounts == 0:
        print gene, totalCounts,totalLength,0
    else:
        print gene, totalCounts,totalLength, (totalCounts/(totalLength/1000))/seqDepth
* 1000000
```

Identification of novel transcribed regions

Inputs:

1) A folder containing all the wig files for each chromosome-strand pair.(The naming convention should be as follows - chr<#>.<pos/neg>.wig - eg. chr1.pos.wig)
2) The default counts file containing RefSeq counts from the ABI bioscope output. (The scripts can be modified to use others.)

Output:

The final output will be a folder containing identified NTRs.

There will be a seperate file for NTRs for all chromosome / strand combinations.

The NTR files will be tab delimited and will have the following columns describing the identified NTR peak.

[<chromosome><strand><start><stop><min height><max height><length><ucsc notation><total area>

Command:

The NTR identification process includes the following steps...

1)Processing the counts file

2)Creating a gap file from the counts file.

3)Identification of NTRs.

1)Processing the counts file.

The biosocpe counts file shows counts for each exon / cds / start codon of refSeq.
First remove CDS and start_codon counts.
Then sort the counts file in the following order

i)chromosome ASC
ii)strand ASC
iii)start position ASC

Name the counts file as [sampleID].exons.sorted.txt, and use it for the next step.

2)Creating a gap file from the counts file.

python gaps.py [counts file] > [gapfile.txt]
Recommended gapfile name - [sampleID].exon.gap.txt

3)Identification of NTRs.

Once the above script is run type...
python peakcall_iterate_bsub.py [wiggle file folder] [resuls folder] [gap file] [minthreshhold] [min peak size]

Code:

Gaps.py

```
#gaps.py
#finds gaps in a list of annotated regions.
#Input - tab delimited file of strand specific annotated regions (e.g refseq)
#Input format - chr,start,stop,strand,Info - Sorted Ascending in the following order -
chr,start,strand # corrction 18/8/2010 chr,strand,start
#Output - list of ranges of the 'gaps' between the given annotations.

import sys

try:
    annotfile = sys.argv[1] #The input file
except:
    print 'Incorrect input.'
    print ' python gaps.py <Inputfile>'



#Initialization of variables which describe the range in the line immediately before.
prevstop = 0
prevchr = 'a'
prevstrand = '~'


for line in open(annotfile,'r'): #reads the annotation file line by line

    #stripping data from the current line


    temp = line.split('\t')
    chr = temp[0].strip()
    strand = temp[3].strip()
    starts = int(temp[1])
    ends = int(temp[2])
    #info = temp[8].split('\n')[0] #this field in unique for the refseq annotation



    #This begins the comparing process

    if prevstop > starts:
 #usually this cannont be the case, if this is true then either there is an overlap
between two annotations or the annotation range is on a different chromosome / strand

        if (prevchr == chr and prevstrand ==strand):
#checks to see if its the same chr/strand combination; if this is the case then its an
overlap.

            if prevstop < ends:
# Checks if the current seqments if fully immersed in the previous one. i.e. a 100!
overlap.

                prevstop = ends
# if ends is greater then shifts the prevstop position and does not print anything as
there is not 'gap'

        else:
            #its a change in either the chromosme or the strand
            prevstop = -1
#resets the prev position and prints the gap from start to the currenet position
            print chr,'\t',prevstop+1,'\t',starts-1,'\t',strand

    else:
            a=1
            #A normal sequential annotation.
            print chr,'\t',prevstop+1,'\t',starts-1,'\t',strand



    prevstop = ends
    prevchr = chr
    prevstrand = strand
```

peakcall.py

```
#peakcall.py
#(c)  Genome Institute of Singapore.
#Input - tab delimited file; Format  - chr,start,stop,strand

try:
    import psyco
    psyco.full()
except:
    pass

import sys



chromosome = sys.argv[1]
inputstrand = sys.argv[2]
wigfile = sys.argv[3]
minthreshhold = int(sys.argv[4])
minpeaksize = int(sys.argv[5])
gapfile = sys.argv[6]



dic = {} #this is the dictionary which will contain all the data from the wig file

#first the script adds the entire wig file to a dictionary. WARNING - you need a
considerable amount of ram. This mac has 4Gig and it seems sufficient.

for line in open(wigfile,'r'): #loops through the file reading each line and adds it
to the dictionary - dic
    temp = line.split('\t')
    pos = temp[0]

    if pos.isdigit(): #prevents the header giving an error.
        pos = int(pos)
        count = int(temp[1].split()[0])

        dic[pos] = count #dic is the dictionary




line = '' #reusing line and temp variables
temp = ''
prevplace = 0


peak={} # this will act as a temporary dictionary to contain the data of a peak

for line in open(gapfile,'r'): #reads the file containing the gap positions line by
line
    temp = line.split('\t')
    chr = temp[0].split()[0]
    strand = temp[3].split()[0]

    if chr==chromosome and strand == inputstrand: #This limits the search for only the
given chromosome and strand.
        starts = int(temp[1])
        stops = int(temp[2]) #start and stop coordinates of the gap

        prevplace = starts-1 #this holds the previous stop+1

        for i in range(starts,stops+1): #why use stops+1?

            try:
                d =  dic[i] # checks if the position contains a value in the
dictionary. if not there will be an error.
                #ithere IS a count value for the given position


                if d > minthreshhold: #provides a threshhold value for the peaks
                    #itpiht there is a count for the position and it is more than the
threshhold

                    if (i-prevplace) == 1: #checks if the current position is adjecent
to the  previous. i.e checks for continuity.

                        peak[i] = int(d)
                        d=0

                    else:
                        #itpiht there is gap i.e the peak has stopped! So showld
display the peak.
                        q=2



                prevplace = i


            except :
                q=2 #The peak has ended
                if len(peak) > minpeaksize:
```

```
                    peakpoints = peak.values()
                    peakkeys = peak.keys()
                    #print 'max', max(peakpoints)
                    #print 'min',min(peakpoints)
                    #print 'length',len(peakpoints)
                    c = len(peakpoints)-1
                    tot = 0
                    for i in peakpoints:
                        tot = tot+int(i)
                    print chr,'\t',strand,'\t',min(peakkeys),'\t',max
(peakkeys),'\t',max(peakpoints),'\t',min(peakpoints),'\t',len
(peakpoints),'\t',chr,':',min(peakkeys),'-',max(peakkeys),'\t',tot


                    peak.clear() #resetting the peak dictionary
                else:
                    peak.clear()

    #to run after the loop ends (this is for peaks which end with the range). i.e the
peaks which continues throught the annotations, i.e. the extensions?

    if len(peak) > minpeaksize:
                    peakpoints = peak.values()
                    peakkeys = peak.keys()


                    c = len(peakpoints)-1
                    tot = 0
                    for i in peakpoints:
                        tot = tot+ int(i)
                    print chr,'\t',strand,'\t',min(peakkeys),'\t',max
(peakkeys),'\t',max(peakpoints),'\t',min(peakpoints),'\t',len
(peakpoints),'\t',chr,':',min(peakkeys),'-',max(peakkeys),'\t',tot


                    peak.clear() #resetting the peak dictionary
    else :
        peak.clear()


    #print '------------------------------'
    peak.clear()
```

Identification of novel transcripts

Steps:

Input:
1) A concatenated NTR file
2) File containing gene footprints

Command:
1) To identify NTRs which are close to a known gene
python getCloseNTRs.py <NTR file> <Genefootprint file> <max gap>

2) To get a list of NTRs which are not close to any gene
python compare.py <output of 1> <NTR file>

3) To identify transcripts
python idenitfyTranscripts.py <output of 2> <max gap between exons>

Compare.py

```
import sys

list = []

for line in open(sys.argv[1],'r'):
    list.append(line.strip())

b = 0

for line in open(sys.argv[2],'r'):
    temp = line.strip().split('\t')
    chr = temp[0].strip()
    start = int(temp[2].strip())
    stop = int(temp[3].strip())
    strand = temp[1].strip()
    #print chr,start,stop,strand
    for i in list:
        temp2 = i.strip().split('\t')
        chr1 = temp2[6].strip()
        start1 = int(temp2[8].strip())
        stop1 = int(temp2[9].strip())

        strand1 = temp2[7].strip()
        #print chr1,start1,stop1,strand1
        if chr == chr1 and start == start1 and stop == stop1 and strand == strand1 :
            #print line.strip()
            b = 1

    if b == 0:
        print line.strip()
    if b == 1:
        b = 0
```

getCloseNTRs.py

```
import sys
NTRFile = sys.argv[1]
geneFootPrintFile = sys.argv[2]
n = int(sys.argv[3])
#addint the NTRfile to a dictionary...
dic = {}
for line in open(NTRFile,'r'):
    temp = line.strip().split('\t')
    tID = temp[0]+temp[1]+temp[2]+temp[3]
    #print temp[4],temp[12],temp[13],temp[14],temp[15]
    dic[tID] = line.strip()

for line in open(geneFootPrintFile,'r'):
    temp = line.strip().split('\t')
    print line
    #print temp[0],temp[1],temp[2],temp[3],temp[4]
    geneID  =temp[0].strip()
    chr = temp[1].strip()
    start = int(temp[2])
    stop = int(temp[3])
    strand = temp[4].strip()

    setGene = set(range(start,stop))
    setGeneExtended = set(range(start-n,stop+n))

    for ntr in dic:
        t = dic[ntr].strip().split('\t')
        nchr = t[0].strip()
        nstrand = t[1].strip()
        if nchr == chr and nstrand == strand :
            nstart = int(t[2])
            nstop = int(t[3])
            setNTR = set(range(nstart,nstop))

            L1 = len(setGene & setNTR)
            L2 = len(setGeneExtended & setNTR)

            if L1 == len(setNTR) or L2 == len(setNTR):
                'L3','\t',line.strip(),'\t',dic[ntr].strip()
            else:
                if L1> 0:
                    print 'L1','\t',line,'\t',dic[ntr],'\t',L1
                #print geneID,chr,start,stop,strand,nchr,nstrand,nstart,nstop,'L1',L1
                if L2 > 0:
                    print 'L2','\t',line.strip(),'\t',dic[ntr].strip(),'\t',L2

                #print geneID,chr,start,stop,strand,nchr,nstrand,nstart,nstop,'L2',L2
```

identifyTranscripts.py

```
from __future__ import division
import sys
n = int(sys.argv[2])
prevStop = 0
tempList = []
prevChr = ''
prevStrand = '~'

for line in open(sys.argv[1],'r'):
    temp = line.strip().split('\t')
    chr = temp[0].strip()
    strand = temp[1].strip()
    start = int(temp[2].strip())
    stop = int(temp[3].strip())

    if chr == prevChr and prevStrand == strand and (start - prevStop) <=  n:
        print '@@@@@@@', start - prevStop, start,prevStop
        tempList.append(line.strip())
    else:
        if len(tempList)>= 2:
            j = 0.00
            for i in tempList:
                print i.strip()
                j = j + float(i.strip().split()[-1])

            print j / len(tempList)
('\t')[2],'-',tempList[0].strip().split('\t')[3],tempList[0].strip().split('\t')[0],':',tempList[0].strip().split
            tempList = []
            tempList.append(line.strip())
            print '======================'
        else:
            tempList = []
            tempList.append(line.strip())


    prevStop = stop
    prevChr = chr
    prevStrand = strand
    #print chr,strand,start,stop
```

Finding 3‘UTR extensions.

Input:
3UTR - bed file containing 3’ UTRs
allexonfile - bed file containing all exons
wigfile - original ABI file

Command:
python <3UTR> <chr> <strand> <wigfile> <allexonfile>

Code:

correctEnds.3prime.+.py

```
#this script is only for + strand.
import sys
exonFile = sys.argv[1] #exon file containing only annotated 3' exons
chr = sys.argv[2] #in the chrn format
strand = sys.argv[3].strip()
wigFile = sys.argv[4].strip() #corresponding wig file
originalExonFile = sys.argv[5] #the exon file containing all exons

oriExon = []
#loading the originalExonFile
for line in open(originalExonFile,'r'):
    temp = line.strip()
    oriExon.append(temp)

#loading the wiggle file into the dictionary
wigDic = {}

for line in open(wigFile,'r'):
    temp = line.strip().split('\t')
    #print temp[0],temp[1]
    wigDic[int(temp[0])] = int(temp[1]) #wigDic[pos]=count



for line in open(exonFile,'r'): #goes through all the 3' UTR exons one by one
    flag = 0
    temp = line.strip().split('\t')
    echr = temp[1].strip()
    estrand = temp[4].strip()
    #print echr,estrand,chr,strand
    if chr == echr and strand == estrand: #looks if the 3' UTR exon is withing the
given chr and strand combination

        NM = temp[0]
        #print NM
        start = int(temp[2]) #start and stop of the 3' UTR
        stop = int(temp[3])
        geneID = temp[5]
        coverage = 0
        total = 0
        avgheight = 0
        #finding the next anootation after this so that the gap between the current 3'
UTR and the next gene
        for exon in oriExon:
            t = exon.strip().split('\t')
            oriChr = t[0]
            oriStart = int(t[1])
            oriStop = int(t[2])
            oriStrand = t[3]
            oriGene = t[4]
```

```
            if chr == oriChr and oriStart == start and oriStop == stop and oriStrand
== strand: #looking for the 3' UTR from the list containing all exons
                oriIndex = oriExon.index(exon) #save the index of it
                tempgeneID = oriGene
                while tempgeneID.strip() == oriGene.strip(): #passes the other 3' UTRs
of the same gene! ??
                    oriIndex = oriIndex +1
                    tempgeneID = oriExon[oriIndex].strip().split('\t')[4].strip()
                    flag = flag + 1
                   # print tempgeneID,oriGene

                #print chr,strand,start,stop,geneID

                #print oriExon[oriIndex], 'is the gene next to it'
                nextStart = int(oriExon[oriIndex].strip().split('\t')[1].strip()) #
this marks then begining of the next gene #nextStart is the start of the next gene
(i.e. the 5' UTR of the next gene)




        tempstart = stop
        coverage = 0
        total = 0
        gaps = 0
        lastPeakStop = 0
        if nextStart - tempstart > 20000:#the maximum gap the scrpt looks for is
10000bp
            nextStart = tempstart+20000
        #print tempstart,nextStart
        if nextStart < tempstart:
            continue #bypasses cases where the next genefootprint starts before the
end of the 3'UTR
        gapCount = 0
        while gaps < 200 and  tempstart < nextStart -1 : #max gap allowed is 100
            #print tempstart,nextStart
            #print gaps, gaps < 200
            if tempstart in wigDic:
                coverage = coverage +1
                total = total + wigDic[tempstart]
                lastPeakStop = tempstart
                #if gaps >  0:
                    #gapCount = gapCount+1
                    #gaps = 0
                gaps = 0

            else:
                gaps = gaps +1
                gapCount = gapCount+1
            tempstart = tempstart + 1


        if coverage > 10: #coverage limit

            #print '========================================='
            print
chr,start,stop,strand,geneID,coverage,total,lastPeakStop,start,gapCount,tempgeneID,fla
g
            #print '========================================='


        ## for i in range(stop,stop+1001):
##              if i in wigDic:
##                  coverage = coverage +1
##                  total = total + wigDic[i]
##         if coverage > 10:
##             print chr,start,stop,geneID,coverage,total
```

Identification of genes which show a change in their splicing profile during differentiation.

Input:

datafile - file containing exon read counts of all samples arranged in different columns

Command:

print 'input -> python altsplice.py <datafile> <column number of first sample> <column number of second sample> <minimum_adjoining peak size> <min peak size> <ratio of ratios>'

To identify genes with mutually exclusive exons substitute altsplice.py with mutualExclusive.py

Code:

altsplice.py

```
#Given a list of combined exon counts sorted according the exon and start position,
this script identifies transcripts which show potential alt splicing.
#It assumes that the ratio of expression level of individual exons whithin a single
transcript at a single time point is the same.


#input format:
#e.g sortedQuery, tab delimited, no header, optional [ID, chromosome,start, stop,
strand, exon, strand] level for each sample,

#import statements
from __future__ import division # this line makes python division behave like normal
i.e with decimals
#from statlib import stats #statlib package used to do statistical calculations
import sys


#assigning command line arguments to variables
try:
    fname = sys.argv[1] # data file

    d0place = int(sys.argv[2]) #position of the first sample
    d8place = int(sys.argv[3]) #position of the second sample

    minAdj = float(sys.argv[4]) #min size of adjoining peaks
    minPeak = float(sys.argv[5]) #min size of peak under study
    mul = float(sys.argv[6]) # multiplication factor


except:

    print 'input -> python altsplice.py <datafile>  <column number of first sample>
<column number of second sample> <minimum_adjoining peak size> <min peak size> <ratio
of ratios>'
    exit(1)
    # Typical input python altsplice.py sortedQuery.txt 10 5 9


#variable initiation
lines = [] #array of lines belonging to the same gene and having a non zero exon count
pgenename = '' #holds the name of the previous gene
ratio = [] #the ration between the two exon counts
```

```
d0 = []
d8 = []
exonStart = []
exonStop = []
exonStartStop = []
r = 0.01
dic= {} #dictionary of lines and the start position, user for sorting
count = 0 #counts the number of lines
dir = ''
prevExonLength = 0
nextExonLength = 0
nowExonLength = 0


print
'pgenename','\t','exon','\t','pratio','\t','nratio','\t','d0prev','\t','d0now','\t','d
0next','\t','d8prev','\t','d8now','\t','d8next'

for line in open(fname,'r'): #sortedQuery.txt contains the combination of all the exon
counts sorted according to the gene
    temp = line.strip().split('\t')
    #genename = temp[0].split(';')[0].split()[1].split('"')[1].split('"')[0]
    genename = temp[9].strip()
    #print genename
    if pgenename == genename: #The objective here is to make a list (lines) of all the
exons belonging to the same gene.
        lines.append(line)
    else:
        #once the lines list is filled it gets processed here

        #sorting the lines in the exon order
        for l in lines:
            dic[int(l.strip().split('\t')[2])] = l #dictionary format startposition :
line

        #sorting dic
        dk = dic.keys()
        dk.sort() #get the list of keys i.e. startpositions and then sort that.


        for l in dk:
            t = dic[l].strip().split('\t')
            d0i = int(t[d0place])
            d8i = int(t[d8place])
            d0.append(d0i)
            d8.append(d8i)
            exonStartStop.append( str(t[0]).strip()+':'+str(t[1]).strip()+'-'+str(t
[2]).strip())
            exonStart.append(int(t[1].strip()))
            exonStop.append(int(t[2].strip()))

            #print t[1],t[2]


        #now that the two lists are filled....
        #print 'sdfad'
        #print exonStart
        #print exonStop
        prevExonLength = 0
        nextExonLength = 0
        nowExonLength = 0

        for i in range(1,len(d0)-2):



            if len(nowSet & prevSet) == 0:
                d0prev = d0[i-1]
                d8prev = d8[i-1]
```

```
            else:
                #print 'wishva',pgenename
                d0prev = d0[i-2]
                d8prev = d8[i-2]

            if len(nowSet & nextSet) == 0:

                d0next = d0[i+1]
                d8next = d8[i+1]
            else:
                #print 'wishva',pgenename
                d0next = d0[i+2]
                d8next = d8[i+2]


            #start code for finding exons which goes completely AWOL

            if d0now < minPeak and d0now > 0 and d8now >= minPeak and d8prev > 0 and
d8next > 0 and d8now / d0now >= mul:
                if (d0prev / d8prev  >=  mul/2 or d0next / d8next >= mul/2):
                #potential appearance of a peak
                    if d0prev >= minAdj and d0next>= minAdj and d8prev >= minAdj and
d8next >= minAdj:
                        print pgenename,'\t', exonStartStop[i],'\t',pratio,'\t',
nratio,'\t',d0prev,'\t',d0now,'\t',d0next,'\t',d8prev,'\t',d8now,'\t',d8next,'\t',
pratio >= mul\
     and nratio>= mul, int(exonStartStop[i].split('-')[1]) - int(exonStartStop
[i].split(':')[1].split('-')[0]),"Extream UP"


            if d8now < minPeak and d8now > 0 and d0now >= minPeak and d0prev > 0 and
d0next > 0 and d0now / d8now >= mul :
                if (d8prev / d0prev  >=  mul/2 or d8next / d0next >= mul/2) :
                #potential dissappearance of a peak

                    if d0prev >= minAdj and d0next>= minAdj and d8prev >= minAdj and
d8next>= minAdj:
                        print pgenename,'\t', exonStartStop[i],'\t',pratio,'\t',
nratio,'\t',d0prev,'\t',d0now,'\t',d0next,'\t',d8prev,'\t',d8now,'\t',d8next,'\t',
pratio >= mul\
     and nratio>= mul, int(exonStartStop[i].split('-')[1]) - int(exonStartStop
[i].split(':')[1].split('-')[0]),"Extream DOWN"


              #      print 'vertical','\t',pgenename,'\t', exonStartStop
[i],'\t',pROR,'\t',
nROR,'\t',d0prev,'\t',d0now,'\t',d0next,'\t',d8prev,'\t',d8now,'\t',d8next,'\t',
pratio >= mul and nratio>= mul, int(exonStartStop[i].split('-')[1]) - int
(exonStartStop[i].split(':')[1].split('-')[0])







            #end code for vertical
comparison-------------------------------------------------------------------------
---------------------------------
            if d0prev >= minAdj and d0next >= minAdj  and d8prev >= minAdj and d8next
>= minAdj:

                if d0now >= minPeak and d8now >= minPeak:




                    d0prevratio = d0now/d0prev
```

```
                    d0nextratio = d0now/d0next

                    d8prevratio = d8now/d8prev
                    d8nextratio = d8now/d8next



                    #previous ratios
                    if d0prevratio >= d8prevratio:
                        pratio = d0prevratio/d8prevratio

                    else:
                        pratio = d8prevratio / d0prevratio


                    #next ratios
                    if d0nextratio >= d8nextratio:
                        nratio = d0nextratio/d8nextratio
                    else:
                        nratio = d8nextratio / d0nextratio



                    if pratio >= mul or nratio>= mul  : # or?
                        print pgenename,'\t', exonStartStop[i],'\t',pratio,'\t',
nratio,'\t',d0prev,'\t',d0now,'\t',d0next,'\t',d8prev,'\t',d8now,'\t',d8next,'\t',
pratio >= mul and nratio>= mul, int(exonStartStop[i].split('-')[1]) - int
(exonStartStop[i].split(':')[1].split('-')[0]),'\t',"Normal"
                        dir = ''
                        count = count+1



        d0 = []
        d8 = []
        exonStartStop = []
        ratio = []
        dic = {}



        lines = []
        lines.append(line)
    pgenename = genename

print count
```

mutualExclusive.py

```
#Given a list of combined exon counts arranged according to the exon this script
identifies transcripts which show potential alt splicing.
#The combined counts file MUST be sorted based on geneID

#Wishva Herath, Robson Lab, Genome Institute of Singapore, Singapore
#(C) Jan 2010.



#import statements
from __future__ import division # this line makes python division behave like normal
i.e with decimals
#from statlib import stats #statlib package used to do statistical calculations
import sys


#assigning command line arguments to variables
try:
    fname = sys.argv[1] # combined counts file
    d0place = int(sys.argv[2]) #The index of the first sample
    d8place = int(sys.argv[3]) #The index of the second sample
    geneIDpos = int(sys.argv[4]) #The index of the geneID
    gT = float(sys.argv[5]) # this is the min of first+second / exonlength
    ratio = float(sys.argv[6]) #the ratio between the first and the second
    #minAdj = int(sys.argv[4])
    #minPeak = int(sys.argv[5])
    #mul = float(sys.argv[6]) # multiplication factor of the standard deviation


except:

    print "Input error"
    print "python findMutualSpliced.py <combined exoncounts file.txt> <d0> <d8>
<geneIDpos> <normTotal> <ratio>"

    exit(1)
    # Typical input python altsplice.py sortedQuery.txt 10 5 9


#variable initiation
lines = [] #array of lines belonging to the same gene and having a non zero exon count
pgenename = '' #holds the name of the previous gene

d0 = []
d8 = []
exonStartStop = []

dic= {} #dictionary of lines and the start position, user for sorting



for line in open(fname,'r'): #sortedQuery.txt contains the combination of all the exon
counts sorted according to the gene
    temp = line.strip().split('\t')
    #genename = temp[0].split(';')[0].split()[1].split('"')[1].split('"')[0] #This is
for the combined gap files produced by accsess.
    genename = temp[geneIDpos].strip()
    #print genename
    if pgenename == genename: #The objective here is to make a list (lines) of all the
exons belonging to the same gene.
        lines.append(line)
    else:
        #once the lines list gets filled with exon data of a particular gene it gets
here.
```

```
        #sorting the lines in the exon order
        for l in lines:
            dic[int(l.strip().split('\t')[1])] = l #dictionary format startposition :
line

        #sorting dic
        dk = dic.keys()
        dk.sort() #get the list of keys i.e. startpositions and then sort that.


        for l in dk:
            t = dic[l].strip().split('\t')
            d0i = int(t[d0place])
            d8i = int(t[d8place])
            d0.append(d0i)
            d8.append(d8i)
            exonStartStop.append( str(t[0]).strip()+':'+str(t[1]).strip()+'-'+str(t
[2]).strip())#contains chr:start-stop

        FC = []
        normGap = []
        rr = []
        rr.append(1.234)
        rr.remove(1.234)
        for i in range(0,len(d0)):
            first = int(d0[i])
            second = int(d8[i])
            gap = int(exonStartStop[i].split('-')[1]) - int(exonStartStop[i].split
(':')[1].split('-')[0])

            normGap.append((first+second)/gap)

            if ((first+second)/gap)  > gT :
                if first == 0 and  second == 0 :
                    #print 'zero'
                    rr.append(float(0))
                    FC.append('z')
                elif first == 0 and second > 1:
                    FC.append('Mu')
                    rr.append(float(0))
                elif second == 0 and first > 1:
                    FC.append('Md')
                    rr.append(float(0))
                elif first == second:
                    FC.append('e')
                    rr.append(float(1))
                elif first / second  > ratio:
                    rr.append(first/second)
                    FC.append('d')
                elif second / first > ratio:
                    rr.append(float(second/first))
                    FC.append('u')
                else:
                    FC.append('n')
                    if first > second:
                        rr.append(float(first / second))
                    else:
                       # rr = second/first
                        rr.append(float(second/first))
            else:
                FC.append('l')
                if first <> 0 and second <> 0:
                    if first > second:
                        rr.append(float(first / second))
                    else:
                        rr.append(float(second / first))
                else:
                    rr.append(float(0))
```

```
    #print FC

    if 'u' in FC and 'd' in FC:
        print '-----------------------------------------------'
        print 'GENEID=',pgenename


        for j in range(0,len(d0)):
            #print "Complete Mutual Exclusion!"
            print FC[j],d0[j],d8[j],exonStartStop[j],normGap[j],rr[j]
            #pass
    if 'Mu' in FC and 'Md' in FC:
        print "complete mutual exclusivitiy"
        for j in range(0,len(d0)):
            print FC[j],d0[j],d8[j],exonStartStop[j],normGap[j],rr[j]



    d0 = []
    d8 = []
    exonStartStop = []

    dic = {}



    lines = []
    lines.append(line)
pgenename = genename
```

Appendix II: Sequences of novel transcripts

Novel transcripts were validated by designing primers to amplify the transcript, running PCR, performing gel purification, cloning the amplicon into a top10 vector and sequencing the vector. The entire sequence of the sequenced vector is given. The primer sequence is shown in bold and underlined text.

Novel transcript 1 (chr1:63,559,143-63,560,695)

Forward strand sequence:

GGTCATTCACAAAGACTCACTATAGGGCGAATTGGGCCCTCTAGATGCATGCTCGAAGCGG
CCGCCAGTGTGATGGATATCTGCAGAATTCGCCCTT**GATTTTTAAATTTTTATTTTTATTTT**
ATTGAATTATTTTTGGTGTGTCAAGGCCAAGGAAAGAGGAGATCGTGGGTGGGGAAACAG
ACAGAGGGAATCAGAAGCACCACTGTCCATCCGGAATTAAATCCACATCCCAGCATCTTCT
GCAAATATTTCACTAATTATTTCCTCTCGGAACTCCTCCCCTCGTGCTCCTTCCTCTGGTGAG
GCCGGCGCTCCCCTCCCAGGCCGCAGCGGACAGACAGGGATTGGGTTCCGTGTGCCTGCC
ACACCAGGCAGGCTCTTGCGGCTCCCAACTAGGCGGCCTTGCACTCCGCGTGCATTGGCCA
CACATCCTCGCCTCCTCCACCCGCTCCGCCGCCGGTTTTCTTGGAAGTTAAATCTTGGAGGA
TTTGTCCACACCCGCTCCCTGGGCCCCAGGGCCCGGATCCAGCCTGGGTGGGGGGGTCTCC
GGGCGGGCCGCAGCGCCCTCCGTGCCCCGGGGATGCTGGCGCACAGTGCGGAGCGGAGTT
GCGCGTCTCTCAAGGGCGAATTCCAGCACACTGGCGGCCGTTACTAGTGGATCCGAGCTCG
GTACCAAGCTTGGCGTAATCATGGTCATAGCTGTTTCCTGTGTGAAATTGTTATCCGCTCAC
AATTCCACACAACATACGAGCCGGAAGCATAAAGTGTAAAGCCTGGGGTGCCTAATGAGTG
AGCTAACTCACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTG
CCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCTCT
TCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCGAGCGGTATCAG
CTCACTCAAAGGCGGTAATACGGTTATCCCCAGAATCAGGGGATAACGCAGGAAAAAACAT
GTGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAGGCCCCGTTGCTGGGGTTTTT
CCATAGGTTCGG

Reverse strand sequence:

CAACGATAATGATACGCCAAGCTTGGTACCGAGCTCGGATCCACTAGTAACGGCCG
CCAGTGTGCTGGAATTCGCCCTT**GAGAGACGCGCAACTCCG**CTCCGCACTGTGC
GCCAGCATCCCCGGGGCACGGAGGGCGCTGCGGCCCGCCCGGAGACCCCCCCAC
CCAGGCTGGATCCGGGCCCTGGGGCCCAGGGAGCGGGTGTGGACAAATCCTCCA
AGATTTAACTTCCAAGAAAACCGGCGGCGGAGCGGGTGGAGGAGGCGAGGATGT
GTGGCCAATGCACGCGGAGTGCAAGGCCGCCTAGTTGGGAGCCGCAAGAGCCTG
CCTGGTGTGGCAGGCACACGGAACCCAATCCCTGTCTGTCCGCTGCGGCCTGGGA
GGGGAGCGCCGGCCTCACCAGAGGAAGGAGCACGAGGGGAGGAGTTCCGAGAG
GAAATAATTAGTGAAATATTTGCAGAAGATGCTGGGATGTGGATTTAATTCCGGATG
GACAGTGGTGCTTCTGATTCCCTCTGTCTGTTTCCCCACCCACGATCTCCTCTTTCC
TTGGCCTTGACACACCAAAAATAATTCAATAAAATAAAAATAAAAATTTAAAAATC
AAGGGCGAATTCTGCAGATATCCATCACACTGGCGGCCGCTCGAGCATGCATCTAG
AGGGCCCAATTCGCCCTATAGTGAGTCGTATTACAATTCACTGGCCGTCGTTTTACA
ACGTCGTGACTGGGAAAACCCTGGCGTTACCCAACTTAATCGCCTTGCAGCACAT
CCCCCTTTCGCCAGCTGGCGTAATAGCGAAGAGGCCCGCACCGATCGCCCTTCCC
AACAGTTGCGCAGCCTGAATGGCGAATGGACGCGCCCTGTAGCGGCGCATTAAGC
GCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAG
CGCCCGCTCCTTTCCCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCC
GTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTTACGGCAC
CTCGACCCCAAAAAACTTGATTAGGGGGATGGTTCACGTAATGGGCCATCCCCC

Novel transcript 2 (chr7:100,729,591-100,731,304)

Forward strand:

GGGGGCGTAGAATCGACTCCTATAGGGCGAATTGGGCCCTCTAGATGCATGCTCGAGCGGC
CGCCAGTGTGATGGATATCTGCAGAATTCGCCCTT**ACCCCAGAATAAGAACTAACC**CAGC
AGGTTTTGTTCACATTAATGGCCAGAGTACAACCCATGAATCTTTAAAGTGAGAAAAATGCT
TCAGATTTTATGCCCTGAGAGTTTTTCTTCTATTTGTTCTAGTCCATCATTGAAGCTTTTGAGT
GTACTTTGTATTTCATCCAATGATTTCTTCAGTTCCAGAAATACTCTTCTGTTGAATTTTAAA
ACCTCTTTCTGGCAGGTAATCATCTCATATATCTCCCCACCTCTTTTCCTCATTTCTTTATGTT
GTTGTTCAAAAGACTCTTCCATGTTGATGACCTTCTTTACAATCAGTCCGGTCAATTCTCTG
CTCAATTCTCTGTTTGAGCTTTCCTGAATTTCTTTTGGATGGGGATCTGTTGCTGGAGAATTA
CTGGTTTCTTCTGAAGGCGTCACATTTCCTTGCTTTTTCATGTCTCCTGGGTCCTTCCGTTGA
GATCTGCGCATCCTGTGGAACAGTCACTTCTTCCGTTTGTGAATTCGCTTTGTAGGGGAGGA
CTTCTTCCCTTTCCCCGCGCGCGGGGGCTGCAGGCCGTCCCAGCCGATCCGATTTCTCCGC
GCAGGCTGCCTGGGTGCCTCTGCTCCACTCCTCGGGCTCCCAGTGGCTTCTCGGCTGAATC
CCGCGTTCTCTTAGCGGATCTGCTGCAAGTGTGAAAACCTACCGGCTCCTTCTGTTCCCGTC
GGAGGAGAGGCGCGTGCCGGCTGCATCTACCAGGCCATCTTGAGCCTGTGGTCCTGGGTG
GAAACGGGGTCCCAAGTGCGGGAGTTACTCAAGGGCGTTCAAGGGCGAATTCCAGCACAC
TGGCGGCCGTTACTAGTGGATCCGAGCTCGGTACCAAGCTTGGCGTAATCATGGTCATAGCT
GTTTCCTGTGTGAAATTGTTATCCGCTCACAATTCCACACAACATACGAGCCGGAAGCATAA
AGTGTAAAGCCTGGGGGGCCTAATGAGTGAGCTAACTCACATTAATTGGGTTGCCCTCCCT
GC

Reverse strand:

AACGACACGTTATACGCCAAGCTTTGGTACGAGCTCGGATCCACTAGTAACGGCCGCCAGT
GTGCTGGAATTCGCCCTT**GAACGCCCTTGAGTAACT**CCCGCACTTGGGACCCCGTTTCCA
CCCAGGACCACAGGCTCAAGATGGCCTGGTAGATGCAGCCGGCACGCGCCTCTCCTCCGA
CGGGAACAGAAGGAGCCGGTAGGTTTTCACACTTGCAGCAGATCCGCTAAGAGAACGCGG
GATTCAGCCGAGAAGCCACTGGGAGCCCGAGGAGTGGAGCAGAGGCACCCAGGCAGCCT
GCGCGGAGAAATCGGATCGGCTGGGACGGCCTGCAGCCCCCGCGCGCGGGGAAAGGGAA
GAAGTCCTCCCCTACAAAGCGAATTCACAAACGGAAGAAGTGACTGTTCCACAGGATGCG
CAGATCTCAACGGAAGGACCCAGGAGACATGAAAAAGCAAGGAAATGTGACGCCTTCAG
AAGAAACCAGTAATTCTCCAGCAACAGATCCCCATCCAAAAGAAATTCAGGAAAGCTCAA
ACAGAGAATTGAGCAGAGAATTGACCGGACTGATTGTAAAGAAGGTCATCAACATGGAAG
AGTCTTTTGAACAACAACATAAAGAAATGAGGAAAAGAGGTGGGGAGATATATGAGATGAT
TACCTGCCAGAAAGAGGTTTTAAAATTCAACAGAAGAGTATTTCTGGAACTGAAGAAATCA
TTGGATGAAATACAAAGTACACTCAAAAGCTTCAATGATGGACTAGAACAAATAGAAGAAA
AACTCTCAGGGCATAAAATCTGAAGCATTTTTCTCACTTTAAAGATTCATGGGTTGTACTCT
GGCCATTAATGTGAACAAAACCTGCTGGGTTAGTTCTTATTCTGGGGTAAGGGCGAATTCTG
CAGATATCCATCACACTGGCGGCCGCTCGAGCATGCATCTAGAGGGCCCAATTCGCCCTATA
GTGAGTCGTATTACAATTCACTGGCCGTCGTTTTACAACGTCGTGACTGGGAAAACCCTGG
CGTTACCCAACTTAATCGCCTTGCAGCACATCCCCCTTTCGCCAGCTGGCGTAATAGCGAAA
AGGCCCCCCCCGATCGCCCTTTCCAAAAGTG

Novel Transcript 3 (chr7:100,738,332-100,740,838)

Forward strand:

GGGTCGTTTGACATCGAATCACTAAAAGGGCGAATTGGGCCCTTCTAGATGCATGCTCGAG
CGGCCGCCAGTGTGATGGATATCTGCAGAATTCGCCCTT**AAGATGGCCTGGTAGATG**CAG
CCGGCACGCACCTCTCCTCTGACGGGAACCGAAGGAGCCGGTAGGTTTTCACACTTGCAG
CAGATCCGCTAAGAGAACGCGGGATTCAGCCGAGAAGCCACGGGGAGCCCGAGGAGCGG
AGCAGAGGCACCCAGGCAGCCTGCGCGGAGAAATTGGATCGGCGGGGACGACCTGCAGC
TCCCGCGCGCGGGGAAAGGGAAGAAGTCCTCCCCTACAAAGCAAATTCACAAACTTGGAA
GAAGCAATTTACACAGGATGTGCAGATCTCAATGGAAGGACACGGGAAACGTGAAAAAGC
AAGGAAGTGGGACGCCTCCAAAGGAACCCAGTAATTCTCCAGCAACAGATCCCCATCCAA
AAGAAATTCAAGAAATGTCATATAGAGAATTGTGGAAACTGATTTTAACCAAGATTAGAGG
GATTCAAGAGACTTCTGAAAAAGAAAGTAAGGAAATGTCAACAGCAATTCTGGATATGGTT
GAGGTATTTACCAACCAGATACAGAGTTTTCCAGAGCACATGGCAAATGTGGAACTGAAGA
AATCACTGGAAAGGGCGAATTCCAGCACACTGGCGGCCGTTACTAGTGGATCCGAGCTCGG
TACCAAGCTTGGCGTAATCATGGTCATAGCTGTTTCCTGTGTGAAATTGTTATCCGCTCACA
ATTCCACACAACATACGAGCCGGAAGCATAAAGTGTAAAGCCTGGGGTGCCTAATGAGTGA
GCTAACTCACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGC
CAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCTCTT
CCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCGAGCGGTATCAGC
TCACTCAAACGCGGTAATACGGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAACATG
TGAGCAAAAGCCCAGCAAAAGGCCAGGAACC

Reverse strand:

AGATTTCAAAGTATTCGCCAAGCTTGGTACCGAGCTCGGATCCACTAGTAACGGCCGCCAG
TGTGCTGGAATTCGCCCTT**TCCAGTGATTTCTTCAGTTCC**ACATTTGCCATGTGCTCTGGA
AAACTCTGTATCTGGTTGGTAAATACCTCAACCATATCCAGAATTGCTGTTGACATTTCCTTA
CTTTCTTTTTCAGAAGTCTCTTGAATCCCTCTAATCTTGGTTAAAATCAGTTTCCACAATTCT
CTATATGACATTTCTTGAATTTCTTTTGGATGGGGATCTGTTGCTGGAGAATTACTGGGTTCC
TTTGGAGGCGTCCCACTTCCTTGCTTTTTCACGTTTCCCGTGTCCTTCCATTGAGATCTGCAC
ATCCTGTGTAAATTGCTTCTTCCAAGTTTGTGAATTTGCTTTGTAGGGGAGGACTTCTTCCCT
TTCCCCGCGCGCGGGAGCTGCAGGTCGTCCCCGCCGATCCAATTTCTCCGCGCAGGCTGCC
TGGGTGCCTCTGCTCCGCTCCTCGGGCTCCCCGTGGCTTCTCGGCTGAATCCCGCGTTCTCT
TAGCGGATCTGCTGCAAGTGTGAAAACCTACCGGCTCCTTCGGTTCCCGTCAGAGGAGAGG
TGCGTGCCGGCTGCATCTACCAGGCCATCTTAAGGGCGAATTCTGCAGATATCCATGACACT
GGCGGCCGCTCGAGCATGCATCTAGAGGGCCCAATTCGCCCTATAGTGAGTCGTATTACAAT
TCACTGGCCGTCGTTTTACAACGTCGTGACTGGGAAAACCCTGGCGTTACCCAACTTAATC
GCCTTGCAGCACATCCCCCTTTCGCCAGCTGGCGTAATAGCGAAGAGGCCCGCACCGATCG
CCCTTCCCAACAGTTGCGCAGCCTGAATGGCGAATGGACGCGCCCTGTAGCGGCGCATTAA
GCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGC
CCGCTCCTTTCGCTTTCTTCCCTTCCTTTTCTCGCCACGTTTCGCCGGCTTTCCCCCGTCTGA
GCTCTAAATCGGGGGGGCTCCCTTTAGAGGGGTTCCGAATTTAAGGGCTTTACGGGCAACC
CCCAACCCCAAAAAAAAAACTTTAT

Novel transcript 4 (chr17:34,456,005-34,462,831)

Forward strand:

GGGCTATTATCGACTCCTATAGGGCGAATTGGGCCCTCTAGATGCATGCTCGAGCGGCCGCC
AGTGTGATGGATATCTGCAGAATTCGCCCTT**TGTAAATGGGAAGGGAAGAA**AATAACATG
AAGTGGAGGCAATAGGAAGAAGAAATGAAGAATCCCTGAGTGAGAACAGGAGTCTTGGA
CTGACTCCGTGGTGCACACACACCCTGTTTCATCTCGGGCAGCATCCTGTCAGCCAGTAGG
AGAGTGGCCGGCCCGAATAGTGCAACCTCCATTCTACCCGCTTGCCATGGTTTCGTTGTGG
GTGGAGGATACTTTCTTGCCCCGGCTTCAGACTTGCCCATGTGGCTTGCTTTGGCCATGGAA
TGAAGCAGAAATGAAAGCCTACCAGTTCCAAAGGGCGAATTCCAGCACACTGGCGGCCGT
TACTAGTGGATCCGAGCTCGGTACCAAGCTTGGCGTAATCATGGTCATAGCTGTTTCCTGTG
TGAAATTGTTATCCGCTCACAATTCCACACAACATACGAGCCGGAAGCATAAAGTGTAAAG
CCTGGGGTGCCTAATGAGTGAGCTAACTCACATTAATTGCGTTGCGCTCACTGCCCGCTTTC
CAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGC
GGTTTGCGTATTGGGCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCG
GCTGCGGCGAGCGGTATCAGCTCACTCAAAGGCGGTAATACGGTTATCCACAGAATCAGGG
GATAACGCAGGAAAGAACATGTGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAA
GGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGA
CGCTCAAGTCAGAGGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTG
GAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTT
CTCCCTTCGGGAAGCGTGGCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGTGTA
GGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAACCCCCCGTTTCAGCCCGACCGCTGGCG
GCCTTA

Reverse strand:

AAGTACAGATTACGCCAGCTTGGTACCGAGCTCGGATCCACTAGTAACGGCCGCCAGTGTG
CTGGAATTCGCCCT**TTGGAACTGGTAGGCTTT**CATTTCTGCTTCATTCCATGGCCAAAGCA
AGCCACATGGGCAAGTCTGAAGCCGGGGCAAGAAAGTATCCTCCACCCACAACGAAACCA
TGGCAAGCGGGTAGAATGGAGGTTGCACTATTCGGGCCGGCCACTCTCCTACTGGCTGACA
GGATGCTGCCCGAGATGAAACAGGGTGTGTGTGCACCACGGAGTCAGTCCAAGACTCCTG
TTCTCACTCAGGGATTCTTCATTTCTTCTTCCTATTGCCTCCACTTCATGTTATTTTCTTCCCT
TCCCATTTACAAAGGGCGAATTCTGCAGATATCCATCACACTGGCGGCCGCTCGAGCATGCA
TCTAGAGGGCCCAATTCGCCCTATAGTGAGTCGTATTACAATTCACTGGCCGTCGTTTTACA
ACGTCGTGACTGGGAAAACCCTGGCGTTACCCAACTTAATCGCCTTGCAGCACATCCCCCT
TTCGCCAGCTGGCGTAATAGCGAAGAGGCCCGCACCGATCGCCCTTCCCAACAGTTGCGCA
GCCTGAATGGCGAATGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTT
ACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCC
CTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTA
GGGTTCCGATTTAGTGCTTTACGGCACCTCGACCCCAAAAAACTTGATTAGGGTGATGGTTC
ACGTAGTGGGCCATCGCCCTGATAGACGGTTTTTCGCCCTTTGACGTTGGAGTCCACGTTCT
TTAATAGTGGACTCTTGTTCCAAACTGGAACAACACTCAACCCTATCTCGGTCTATTCTTTT
GATTTATAAGGGATTTTGCCGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAACAAAA
ATTTAACGCGAATTTTAACAAAATTCAGGGCGCAAGGGCTGCTAAAGGAAGCGGAACACG
TAAAAAGCCAGTCCGCAAAAACGGGTGCTGACCCCGGATG

Novel transcript 5 (chr19:44,838,393-44,843,124)

Forward strand:

GGCGTAGTATCGACTCACTATAGGGCGAATTGGGCCCTCTAGATGCATGCTCGAGCGGCCGC
CAGTGTGATGGATATCTGCAGAATTCGCCCTT**ACTCAGAAGACTGGACAC**AATTCCGAAG
GTCGCCCAGAAGGAGAGGACAATGTCATTTCTAACTGTGCCATACAAACTGCCTGTGTCTT
TGTCTGTTGGTTCCTGCGTGATAATCAAAGGGACACTGATCGACTCTTCTATCAGCGAACCA
CAGCTGCAGGTGGATTTCTACACTGAGATGAATGAGGACTCAGAAATTGCCTTCCATTTGC
GAGTGCACTTAGGCCGTCGTGTGGTCGTGAACAGTCGTGAGTTTGGGATATGGATGTTGGA
GGAGAATTTACACTATGTGCCCTTTGAGGATGGCAAACCATTTGACTTGCGCATCTACGTGT
GTCACAATGAGTATGAGGTAAAGGTAAATGGTGAATACATTTATGCCTTTGTCCATCGAATC
CCGCCATCATATGTGAAGATGATTCAAGTGTGGAGAGATGTCTCCCTGGACTCAGTGCTTGT
CAACAATGGACGGAGATGATCACACTCCTCATTGTTGAGGAAACCCTCTTTCTACCTGACC
ATGGGATTCCTAGAGCCTGCCAACAGAATAATCCCTCCTCAACCCCTTCCCCTACACTTGGT
CATTAAAACAGCACCAAACCAAGGGCGAATTCCAGCACACTGGCGGCCGTTACTAGTGGAT
CCGAGCTCGGTACCAAGCTTGGCGTAATCATGGTCATAGCTGTTTCCTGTGTGAAATTGTTA
TCCGCTCACAATTCCACACAACATACGAGCCGGAAGCATAAAGTGTAAAGCCTGGGGTGCC
TAATGAGTGAGCTAACTCACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAA
CCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATT
GGGGCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCGA
GCGGTTTCAGCCCCCTCAAAAGGCGGTAAAACGGTTTTCCACAAAAATC

Reverse strand:

ACAGAATGATTCGCCAAGCTTGGTACCGAGCTCGGATCCACTAGTAACGGCCGCCAGTGTG
CTGGAATTCGCCCTT**GGTTTGGTGCTGTTTTAATG**ACCAAGTGTAGGGGAAGGGGTTGAG
GAGGGATTATTCTGTTGGCAGGCTCTAGGAATCCCATGGTCAGGTAGAAAGAGGGTTTCCT
CAACAATGAGGAGTGTGATCATCTCCGTCCATTGTTGACAAGCACTGAGTCCAGGGAGACA
TCTCTCCACACTTGAATCATCTTCACATATGATGGCGGGATTCGATGGACAAAGGCATAAAT
GTATTCACCATTTACCTTTACCTCATACTCATTGTGACACACGTAGATGCGCAAGTCAAATGG
TTTGCCATCCTCAAAGGGCACATAGTGTAAATTCTCCTCCAACATCCATATCCCAAACTCAC
GACTGTTCACGACCACACGACGGCCTAAGTGCACTCGCAAATGGAAGGCAATTTCTGAGTC
CTCATTCATCTCAGTGTAGAAATCCACCTGCAGCTGTGGTTCGCTGATAGAAGAGTCGATCA
GTGTCCCTTTGATTATCACGCAGGAACCAACAGACAAAGACACAGGCAGTTTGTATGGCAC
AGTTAGAAATGACATTGTCCTCTCCTTCTGGGCGACCTTCGGAATTGTGTCCAGTCTTCTGA
GTAAGGGCGAATTCTGCAGATATCCATCACACTGGCGGCCGCTCGAGCATGCATCTAGAGG
GCCCAATTCGCCCTATAGTGAGTCGTATTACAATTCACTGGCCGTCGTTTTACAACGTCGTG
ACTGGGAAAACCCTGGCGTTACCCAACTTAATCGCCTTGCAGCACATCCCCCTTTCGCCAG
CTGGCGTAATAGCGAAGAGGCCCGCACCGATCGCCCTTCCCAACAGTTGCGCAGCCTGAAT
GGCGAATGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCA
GCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTT
CTCGCCCCGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGG

Novel transcript 6 (chr13:99,536,264-99,539,11)

Forward strand:

GGGGCGTTTTATCGACTCACTATAGGGCGAATTGGGCCCTCTAGATGCATGCTCGAGCGGCC
GCCAGTGTGATGGATATCTGCAGAATTCGCCCTT**CAGGAGTGAGAGAGACAG**GGGCGTGC
GTGAGTTCCGGCGGCCTGCACCGGGCAAACCCCGTACCTTCCCAGCATCGGCTCAGCAACC
CACGTGCATCCAGGCCGGTCAATGTCATTGAGTCACCTCCGCGCCTTGGCCACCCTGGAGT
CCCGAGAATCCGAAGTTCCGGACAAATGCCCAAACTACATTCCTGCATGTTCGAAAGCGTA
AATTGCAAAGCACAAATCCAGTTGTAGATTGTGGCCGGGAGCAGTGGCTCACGCCTATAAT
CCCAGCACTTTGGGAGGCCGAGGCGGACAGATCACGAGGTCAGCACTTCGAGACCAGCAT
GGCCAACATGGTGAAGCCCCATCTCTACTAAAAATACAAACATTAGCCAGGCATGGTGGTA
GGCCCCTGTAATCCCAGCTACTCGGTAGGCTGAGGCAGGAGAATCACTTGAACCCGGGAG
GCAGAGGTTTCAGTGAGCTGAGACTGCACCATTGCACTCCAGCCTGGGCGACAGACCAAG
ACTCCATCTCAAAAAAAAAAAAAAAGTTATAGATTGTAAGGAAAATACCCCCAAGGAAGTT
GAGGACACAGCAGACTTGGACTGTCTCCAAACCTGTTCATTCTTCTGAGTGCACTGCTCGG
AGCATCCTATTGGGCAGCATATCCTGGCCTCCTTTCCAGTTCGATGTGGTACTATACCTGATT
TCTGGCCAATAAAATATGAGAGGACATGAAATTCAACAGCCCCTGGGCTTGGCTTAAAAAA
CCTTCCATGCAATCCTCCACACCTTTTCACCTCTCAGCTGCTGAGACTTTTCTAGGCACTAC
CCAAAAACCAGCCTGGGTCTCCAGTAACAAAGGAAAACAAATTCCTTCCCACTATCCTGCA
CCGAACTGTGACTGGGCAAAAAATAAAGTTTTTATTTTCTTACCCCCTTGAAATTTGAGGAT
GCTAAAACAATTATTTCTTTCCACCCTAATACCGTTTTTTTAAGGACAATTTGGCCCAACACG
TTCTTATGCCCTTATATCCCCCCCTTTA

Reverse strand:

CGCAGACAGATACGCCAGCTTGGTACCGAGCTCGGATCCACTAGTAACGGCCGCCAGTGTG
CTGGAATTCGCCCTT**CCTGGCCATTCTTGATATT**GTTAATGTATGTTTTTATTCTGAAAAAG
GTTTTTTGTTGTTTATTTATTTATTTTTGAGACAGAGTCTCACTCTGTTGCCCAGGCTGGAGT
GCACTGGTACAATCTTGGTTCACTGCAACCTCCGCCTCCTGGCTTGAAGATTTTCCTGCCTC
AGCTTCCTGAGTAGCTGGGATCACAGGCACGTGCCACCATGTTTGGCTAATTTTCGTATTTT
TTGTAGAGCTGGGGTTTCACCATGTCGGCCAGGCTGGTCTCAAACTCCTGACATCAGGTGA
TCTGCCCACCTCGGCCTTATAAACTGCTGTGATTATAGGCATAAGCTACTGTGTCTGGCCTAA
TTGTCCTCTACTAATACTGTATTAGGCTGGGAAGACTAACTGCTCTAGCAATCCTCAAATCTC
AATGGTGTAAGAAAATAAAAACTTTATTTTTTGCTCATGTCACAGTTCGGTGCAGGATAGTG
AGAAGGATCTGCTCTCCTTTGCTACTGGAGACCCAGGCTGCTTCTGTGTAGTGCCTAGAAA
AGTCTCAGCAGCTGAGAGGTGAAAAGGTGTGGAGGATTGCATGGAAGGTTTTATAAGCCA
AGCCCAGGGGCTGTTGAATTTCATGTCCTCTCATATTTTATTGGCCAGAAATCAGGTATAGTA
CCACATCGAACTGGAAAGGAGGCCAGGATATGCTGCCCAATAGGATGCTCCGAGCAGTGCA
CTCAGAAGAATGAACAGGTTTGGAGACAGTCCAAGTCTGCTGTGTCCTCAACTTCCTTGGG
GGTATTTTCCTTACAATCTATAACTTTTTTTTTTTTTTGAAATGGAGTCTTGGTCTGTCGCCC
AGGCTGGAGTGCAATGGTGCAGTCTCAGCTCACTGAAACCTCTGCCTCCCGGGTTCAAGTG
ATTCTCCTGCCTCAGCCTACCGAGTAGCTGGGATTACAGGGGCCCTACCACCATGCCTGGCT
AATGTTTGTATTTTTTAGTAAAAATGGGGCTTCACCATGTTGGCCCATGCTGGTCTCGAAGT
GCTGACCTC

Novel transcript 7 (chr13:90,577,939-90,644,334)

Forward strand:

GGGAGGATGTTTCGACTCACTATAGGGCGAATTGGGCCCTCTAGATGCATGCTCGAGCGGC
CGCCAGTGTGATGGATATCTGCAGAATTCGCCCTT**ATTTCTAGGTGCAGACGA**GGCATTTG
GGGCATAGAAGATCACACTCTTCTTCCGCCATGTCTTAAGATATTACTTTATAGTAATTTATCT
TAGTCCAGGTGCAGTGCCTCACACCTGTAATCCCAGCACTTTGGAAGGCTGAGCTGGGAGG
ATTGCTTGAGGCTGGCAGTTCAAGACTTCTTCCATTTTAAGGGGCCTTGTGATTACAGCTGG
TCCATCTGGATATTCTAAGATATTCTCCCTATTTTAAGGAGAGAAATCTGAAATCTGAGGTGC
AGTTGATTTGAATCCAGGAAATCTAAGGAAAAAGTTCAAGCTCTTAATCATTTCATACCCTT
CTTGTTGCTAACTTAAACTTTTTTTTTAAAAAAGTTAATCTTGTCTATGAAGCATGAATCTAT
AATACTAGGGAAAAAACTGGCTAACAAGGGCGAATTCCAGCACACTGGCGGCCGTTACTA
GTGGATCCGAGCTCGGTACCAAGCTTGGCGTAATCATGGTCATAGCTGTTTCCTGTGTGAAA
TTGTTATCCGCTCACAATTCCACACAACATACGAGCCGGAAGCATAAAGTGTAAAGCCTGG
GGTGCCTAATGAGTGAGCTAACTCACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTC
GGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTT
GCGTATTGGGCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGC
GGCGAGCGGTATCAGCTCACTCAAAGGCGGTAATACGGTTATCCACAGAATCAGGGGATAA
CGCAGGAAAGAACATGTGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAGGCCG
CGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGCTC
AAGTCAGAGGTGGCGAAACCCGACAGGACTTTAAAGATACCAGGCGTTTCCCCCTGGAAG
CTCCCTCGGTGCGCTCTCCTGTTCCGACCCTGCCCCTTACCGG

Reverse strand:

NNNNNNNNNNAANNNNNNNAACGCCAAGCTNGGGTACCGAAGCTTCGGANNNACTAGTA
ACGGCCGCCAGTGTGCTGGAATTCGCCCTT**GTTAGCCAGTTTTTTCCC**TAGTATTATAGAT
TCATGCTTCATAGACAAGATTAACTTTTTTAAAAAAAAAGTTTAAGTTAGCAACAAGAAGG
GTATGAAATGATTAAGAGCTTGAACTTTTTCCTTAGATTTCCTGGATTCAAATCAACTGCACC
TCAGATTTCAGATTTCTCTCCTTAAAATAGGGAGAATATCTTAGAATATCCAGATGGACCAGC
TGTAATCACAAGGCCCCTTAAAATGGAAGAAGTCTTGAACTGCCAGCCTCAAGCAATCCTC
CCAGCTCAGCCTTCCAAAGTGCTGGGATTACAGGTGTGAGGCACTGCACCTGGACTAAGAT
AAATTACTATAAAGTAATATCTTAAGACATGGCGGAAGAAGAGTGTGATCTTCTATGCCCCA
AATGCCTCGTCTGCACCTAGAAATAAGGGCGAATTCTGCAGATATCCATCACACTGGCGGCC
GCTCGAGCATGCATCTAGAGGGCCCAATTCGCCCTATAGTGAGTCGTATTACAATTCACTGG
CCGTCGTTTTACAACGTCGTGACTGGGAAAACCCTGGCGTTACCCAACTTAATCGCCTTGC
AGCACATCCCCCTTTCGCCAGCTGGCGTAATAGCGAAGAGGCCCGCACCGATCGCCCTTCC
CAACAGTTGCGCAGCCTGAATGGCGAATGGACGCGCCCTGTAGCGGCGCATTAAGCGCGG
CGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCC
TTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCG
GGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTTACGGCACCTCGACCCCAAAAAACTTGAT
TAGGGTGATGGNTNCACGTANNGGGCCATCGCCCTGATAGACGGTTTTTCGCCCTTTGACN
TTNGGAGTCCACNNNNNNNANNGGNNTNTNNNCNAACTNNNNNNACTCANNCCNNTCTC
NNNCNNNTNNTTNNNNNNNNNNNNNTNNNGAATTCGNNNCTATTTGNNNNNN

Novel transcript 8 (chr10:54,432,626-54,459,840)

Forward strand:

GGGAGGTTAGTAATCGACTCACTATAGGGGCGAATTGGGCCCTCTAGATGCATGCTCGAGC
GGCCGCCATGTGATGGATATCTGCAGAATTCGCCCTT**CTTTGTACTTTCACTCTGCT**CAAT
AAAGCCTGCAGCTTTTTCTCACTCTCAGTCCATGTCTCTTTCACTCACTGTGGTCAGCTTCC
ACACCATTTCTTTGGTGTGGCTTGGCAAGAACCTCAGGTGTTACATCTTGGCGAGCCAGAC
AGGAGACTCCAGAAAAGGGGTGATTTTCCTGTACCAGTCCAATGCCTCCAGAGGAAGATC
ATACATTTGCCATTTTACTGCTTAGTACGCATGCTTGAGCCTGCTCGCCCAACTGCTGAGATC
TTATTCAGAAACTGCTGATCACCAACTCCAGCGTCAAATGCTGAGAACCCAGTGGAGGAGT
CCAAGACCTTAGGGGATTGTGGAGCCGCTTGTCCAACACACAGCCCATGGGCCACATGTGG
CTCAGGATTGCTTTGAATGCAGCCCAACACAAATTCACAAACTTTCTTAAAACATTATGAGT
TTTTTTGTGATTTTTTTTTTTAGTAGCTCATAAGCTATGGTTAGTGGTAGTGTATTTTATGTGT
GTGTCCCAAGACAATTCTTCTTCCAGTGTGGCACAGGGAAGCCAAAAGATTGTACACCCAT
GAATTAGAAAGAACAAGCATCAGGAAGGGCGAATTCCAGCACACTGGCGGCCGTTACTAG
TGGATCCGAGCTCGGTACCAAGCTTGGCGTAATCATGGTCATAGCTGTTTCCTGTGTGAAAT
TGTTATCCGCTCACAATTCCACACAACATACGAGCCGGAAGCATAAAGTGTAAAGCCTGGG
GTGCCTAATGAGTGAGCTAACTCACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCG
GGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTG
CGTATTGGGCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGCG
GCGAGCGGTATCAGCTCACTCAAAGGCGGTAATACGGTTATCCACAGAATCAGGGGGATAA
CGCAGGAAAGAACATGTGAGCAAAAAGGCCAGCAAAAAGGCCAGGAACCG

Reverse strand:

CAACGAGACGTCATTCGCCAGCTTGGTACCGAGCTCGGATCCACTAGTAACGGCCGCCAGT
GTGCTGGAATTCGCCCTT**CCTGATGCTTGTTCTTTCT**AATTCATGGGTGTACAATCTTTTG
GCTTCCCTGTGCCACACTGGAAGAAGAATTGTCTTGGGACACACACATAAAATACACTACC
ACTAACCATAGCTTATGAGCTACTAAAAAAAAAAATCACAAAAAAACTCATAATGTTTTAAG
AAAGTTTGTGAATTTGTGTTGGGCTGCATTCAAAGCAATCCTGAGCCACATGTGGCCCATG
GGCTGTGTGTTGGACAAGCGGCTCCACAATCCCCTAAGGTCTTGGACTCCTCCACTGGGTT
CTCAGCATTTGACGCTGGAGTTGGTGATCAGCAGTTTCTGAATAAGATCTCAGCAGTTGGG
CGAGCAGGCTCAAGCATGCGTACTAAGCAGTAAAATGGCAAATGTATGATCTTCCTCTGGA
GGCATTGGACTGGTACAGGAAAATCACCCCTTTTCTGGAGTCTCCTGTCTGGCTCGCCAAG
ATGTAACACCTGAGGTTCTTGCCAAGCCACACCAAAGAAATGGTGTGGAAGCTGACCACA
GTGAGTGAAAGAGACATGGACTGAGAGTGAGAAAAAGCTGCAGGCTTTATTGAGCAGAGT
GAAAGTACAAAGAAGGGCGAATTCTGCAGATATCCATCACACTGGCGGCCGCTCGAGCATG
CATCTAGAGGGCCCAATTCGCCCTATAGTGAGTCGTATTACAATTCACTGGCCGTCGTTTTAC
AACGTCGTGACTGGGAAAACCCTGGCGTTACCCAACTTAATCGCCTTGCAGCACATCCCCC
TTTCGCCAGCTGGCGTAATAGCGAAGAGGCCCGCACCGATCGCCCTTCCCAACAGTTGCGC
AGCCTGAATGGCGAATGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTTGTGGTG
GTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTT
CCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTT
TAGGGTTCCGATTTAGTGCTTTACGGCACCTCGACCCCA

Appendix III: Sequencing statistics

Day 0				
Counts:				
Reads mapped:	42845342	(100.0%)		
Reads filtered:	5866117	(13.7%)		
Reads with too many mappings (N > 10):	3261829	(7.6%)		
Reads with number of mappings in proper range (N <= 10):	39583513	(92.4%)		
Reads uniquely aligned (score.clear.zone = 4):	26955116	(62.9%)		
Reads mapped to NTR:				1703606
Day 2				
Counts:				
Reads mapped:	42203140	(100.0%)		
Reads filtered:	4238302	(10.0%)		
Reads with too many mappings (N > 10):	3363120	(8.0%)		
Reads with number of mappings in proper range (N <= 10):	38840020	(92.0%)		
Reads uniquely aligned (score.clear.zone = 4):	27790743	(65.8%)		
Reads mapped to NTR:				3055869
Day 4				
Counts:				
Reads mapped:	40421804	(100.0%)		
Reads filtered:	4105623	(10.2%)		
Reads with too many mappings (N > 10):	3378454	(8.4%)		
Reads with number of mappings in proper range (N <= 10):	37043350	(91.6%)		
Reads uniquely aligned (score.clear.zone = 4):	26832589	(66.4%)		
Reads mapped to NTR:				3283724
Day 6				
Counts:				
Reads mapped:	40218029	(100.0%)		
Reads filtered:	5020898	(12.5%)		
Reads with too many mappings (N > 10):	3065324	(7.6%)		
Reads with number of mappings in proper range (N <= 10):	37152705	(92.4%)		
Reads uniquely aligned (score.clear.zone = 4):	26237091	(65.2%)		
Reads mapped to NTR:				2814868
Day 8				
Counts:				
Reads mapped:	40174214	(100.0%)		
Reads filtered:	4056325	(10.1%)		
Reads with too many mappings (N > 10):	2866193	(7.1%)		
Reads with number of mappings in proper range (N <= 10):	37308021	(92.9%)		
Reads uniquely aligned (score.clear.zone = 4):	26370337	(65.6%)		
Reads mapped to NTR:	2642161			

3.5BL				
Reads mapped:	63348717	(100.0%)		
Reads filtered:	4357607	(6.9%)		
Reads with too many mappings (N > 10):	8221261	(13.0%)		
Reads with number of mappings in proper range (N <= 10):	55127456	(87.0%)		
Reads uniquely aligned (score.clear.zone = 4):	38227773	(60.3%)		
E4.5BL				
Reads mapped:	69274851	(100.0%)		
Reads filtered:	4162828	(6.0%)		
Reads with too many mappings (N > 10):	9230754	(13.3%)		
Reads with number of mappings in proper range (N <= 10):	60044097	(86.7%)		
Reads uniquely aligned (score.clear.zone = 4):	45011450	(65.0%)		
E4.5 ICM				
Reads mapped:	65735255	(100.0%)		
Reads filtered:	2739158	(4.2%)		
Reads with too many mappings (N > 10):	12190267	(18.5%)		
Reads with number of mappings in proper range (N <= 10):	53544988	(81.5%)		
Reads uniquely aligned (score.clear.zone = 4):	37920430	(57.7%)		
8 cell				
Reads mapped:	66423689	(100.0%)		
Reads filtered:	2155598	(3.2%)		
Reads with too many mappings (N > 10):	10926984	(16.5%)		
Reads with number of mappings in proper range (N <= 10):	55496705	(83.5%)		
Reads uniquely aligned (score.clear.zone = 4):	37635809	(56.7%)		