universität wien

DISSERTATION

Titel der Dissertation

Mapping and characterization of macro non-protein coding RNAs in human imprinted gene regions

angestrebter akademischer Grad

Doktorin der Naturwissenschaften (Dr. rer.nat.)

Verfasserin / Verfasser:	Irena Vlatkovic
Matrikel-Nummer:	0642621
Dissertationsgebiet (lt. Studienblatt):	Molekulare Biologie
Betreuerin / Betreuer:	Prof. Denise P. Barlow, PhD

Wien, im September 2010

The work presented in this thesis was performed in the lab of Dr. Denise Barlow, CeMM, Research Centre for Molecular Medicine of the Austrian Academy of Sciences, in Dr. Bohr-Gasse 9/4, A-1030 Vienna and in Lazarettgasse 14, AKH BT 25.3, A-1090 Vienna, during the period from October 2006 to September 2010

TABLE OF CONTENTS……………………………………………………………………………..3

Zussamenfassung…………………………………………………………………………………..7

Abstract………………………………………………………………………………………….….8

1. Introduction…………………………………………………………………………….….……9

1. 1. Epigenetics and non-protein coding RNAs……………………………………………..9

1. 1. 1. Definition of epigenetics…………………………………………………………….….9

1. 1. 2. The epigenetic pathway………………………………………………………….…10

1. 1. 3. Epigenetic roles of DNA methylation…………………………………………...…11

1. 1. 4. Epigenetic roles of histone modifications……………………….............................14

1. 1. 5. Epigenetic roles of ncRNAs……………………………………………………….......16

1. 2. Genomic imprinting as a model in ncRNA research……………..........................19

1. 2. 1. Genomic imprinting is an epigenetic phenomenon……………...............................19

1. 2. 2. Discovery of genomic imprinting……………………………….……………………19

1. 2. 3. Evolution of genomic imprinting………………………………………………………..20

1. 2. 4. Key features of genomic imprinting ……………………………………………………22

1. 2. 5. Human and mouse imprinted gene regions………………………...…………………24

1. 2. 5. 1. Six well-studied imprinted gene regions in human and mouse……………………..26

1. 2. 5. 2. Less-studied imprinted gene regions in human and mouse…………………….…30

1. 2. 6. Imprinted protein coding genes have diverse functions………………………………...33

1. 2. 7. Atypical biology of imprinted macro ncRNAs…………………………………………34

1. 2. 8. Imprinted macro ncRNAs are transcriptional regulators………………………………..35

1. 2. 9. Human genomic imprinting and disease…………………………………………………38

1. 2. 9. 1. Selected disorders of genomic imprinting…………………………………...…..38

1. 2. 9. 2. Human genomic imprinting and cancer……………………………………...…..40

1. 3. Human transcriptome research and ncRNAs…………………………………….…41

1. 3. 1. New high-throughput technologies and surprises from transcriptomics……………...41

1. 3. 2. Challenges in defining the non- protein coding transcriptome………………………...42

1. 3. 3. Non-protein coding RNAs and disease……………………………………….……45

1. 3. 4. Non-protein coding RNAs as biomarkers and drug targets…………………….……..46

1. 4. Aims of the study………………………………………………………………………......47

2. Results………………………………………………………………………………………..……49

2. 1. HIRTA (Human Imprinted Region Tiling Array) technology successfully detects macro ncRNAs ……………………………………………………………………..49

2. 1. 1. Selection of regions of interest and design of HIRTA ………………………….……49

2. 1. 2. Normal and cancer samples were hybridized on HIRTA………………………………50

2. 1. 3. Reproducibility of HIRTA……………………………………………………………..52

2. 1. 4. Dynamic range of HIRTA……………………………………………………………54

2. 1. 5. Single cDNA hybridization on HIRTA………………………………………………56

2. 1. 6. Double cDNA hybridization on HIRTA……………………………………………..58

2. 2. Macro ncRNAs in six well-studied human imprinted gene regions……….………59

2. 2. 1. Known human macro ncRNAs from the HIRTA well-studied regions……………….…59
2. 2. 1. 1. Mapping and tissue specific expression of known macro ncRNAs……….……….59
2. 2. 1. 2. Nuclear versus cytoplasmic localization of known macro ncRNAs…………….….66
2. 2. 1. 3. 5'-3' slope as a feature of macro ncRNA expression by HIRTA…………………..67
2. 2. 2. Novel macro ncRNA candidates in well-known imprinted gene regions…………….68
2. 3. Human well-studied XIST region…………………………………………………………72
2. 4. Macro ncRNAs in 26 less-studied imprinted gene regions…………………………..75
2. 4. 1. Mapping of known and novel macro ncRNAs in less-studied regions………………..75
2. 4. 2. Developmental regulation of macro ncRNAs in less-studied regions…………...…78
2. 4. 3. Nuclear versus cytoplasmic localization of novel macro ncRNAs in less-studied regions………………………………………………………………………….……79
2. 4. 4. Tissue specific expression of macro ncRNAs from less-studied regions………......80
2. 4. 5. Examples of characterization and regulation of novel macro ncRNAs ……………....82
2. 4. 5. 1. HIRTA REG TP73 (chr1): *LRRC47down* macro ncRNA characterization…………83
2. 4. 5. 2. HIRTA REG MEST (chr7): *KLF14up3* macro ncRNA characterization……......…..86
2. 4. 5. 3. HIRTA REG KCNK9 (chr8): *PEG13* and *KIAA1126up* ncRNA characterization.....87
2. 4. 5. 4. HIRTA REG BAZ2 (chr11): *PRKCDBP* macro ncRNA characterization……….….91
2. 4. 5. 5. HIRTA REG SLC38A4 (chr12): *SLC38A4down2* macro ncRNA characterization..94
2. 4. 5. 6. HIRTA REG RASGRF1 (chr15): *ADAMTS7down*, *TMED3down* and *KIAA1024up* macro ncRNA characterization………………..…………………………..…99
2. 4. 5. 7. HIRTA REG NNAT (chr20): *BLCAPov* macro ncRNA characterization……..……106
2. 4. 5. 8. Summary of characteristics for ten examined macro ncRNAs……………...….108
2. 5. Each human imprinted gene region express macro ncRNA………………………..108
2. 5. 1. Overview of macro ncRNAs in human imprinted gene regions…………………..….108
2. 5. 2. Overview of tissue specific expression of macro ncRNA candidates………………..110
2. 5. 3. Strand prediction of macro ncRNA candidates…………………………………………112
2. 5. 4. Non-coding potential of macro ncRNA candidates……………………………...….115
2. 6. Macro ncRNAs from imprinted regions are deregulated in cancer………………..119
2. 6. 1. Well-studied macro ncRNAs are deregulated in cancer……………………………119
2. 6. 2. Novel macro ncRNAs are deregulated in cancer…………………………………...122
2. 6. 3. Identification of the human homolog of the mouse *Airn* ncRNA in Wilms' tumors....126
2. 7. Ribosomal RNA-depleted RNA-sequencing detects macro ncRNAs in human fibroblasts………………………………………………………………………….……..131
2. 7. 1. Statistics of rRNA depleted RNA-seq reads in fibroblasts………………….....…..132
2. 7. 2. Known macro ncRNAs by RNA-seq…………………………………………………..135
2. 7. 3. Novel HIRTA macro ncRNAs validation by RNA-seq…………………………………137
2. 7. 4. Novel primary small ncRNA transcripts in the human fibroblasts……………………139
2. 7. 5. Novel *SHOX* protein coding gene variant………………………………………………141
3. Discussion……………………………………………………………………………………143
3. 1. Summary of results…………………………………………………………….…..……143

3. 2. HIRTA and RNA-Seq successfully detects macro ncRNAs…………………………143
3. 2. 1. Macro ncRNAs could be mapped from their expression features
on the tiling array……………………………………………………………………………143
3. 2. 1. 1. "Continuous high HIRTA signal" over macro ncRNA body may
originate from differences in splicing …………………………………………...….……144
3. 2. 1. 2. 5'-3' slope macro ncRNA feature may originate in alternative polyadenylation….145
3. 2. 1. 3. 5'-3' slope macro ncRNA feature as a criteria for the transcript
orientation prediction…………………………………………………………………………147
3. 2. 1. 4. Differences in HIRTA mapping for macro ncRNAs having diverse
positions to the annotated genes……………………………………………………………148
3. 2. 2. Ribosomal RNA depleted RNA-Seq successfully maps macro ncRNAs……………149
3. 3. Pervasive transcription or functional ncRNAs?...150
3. 3. 1. Novel macro ncRNAs that show tissue specific expression
and subcellular localization may be functional…………………………......………………..151
3. 3. 2. Developmentally regulated macro ncRNAs overlapping
olfactory receptor genes may be functional…………………......…………………………152
3. 3. 3. A small number of HIRTA mapped macro ncRNAs overlap
with lincRNAs associated with PRC2 and CoREST complexes……………………..…....153
3. 3. 4. Macro ncRNAs may function independently and as precursors for small ncRNAs..153
3. 4. Developmental upregulation of *XIST* macro ncRNA in HES2 cells………………..155
3. 5. Further support to the lack of conservation between mouse and human
***Tsix*/*TSIX* function……………………………………………………………………………156**
3. 6. Are macro ncRNAs one of the key features in genomic imprinting?....................156
3. 6. 1. RASGRF1 may be novel human imprinted gene region…………………………....159
3. 6. 2. Identification of *PEG13*, the human homologue of the mouse *Peg13* ncRNA………161
3. 7. Conservation of imprinted gene regions between human and mouse?...............161
3. 8. Macro ncRNAs from imprinted regions are deregulated in cancer………………..162
3. 9. Finding function of macro ncRNAs in development and disease……………...…164
4. Conclusion……………………………………………………………………..……167
5. Materials and methods……………………………………………………………………..168
5. 1. Human samples………………………………………………………………………..…168
5. 1. 1. Cell lines……………………………………………………………………………**168**
5. 1. 2. Differentiation of the human embryonic stem cells……………………………...169
5. 1. 3. Collection of human normal and patient tissue samples…………………………...…169
5. 2. DNA isolation …………………………………………………………………......…169
5. 2. 1. DNA isolation from the cell lines……………………………………………...………169
5. 2. 2. DNA isolation from the blood………………………………………………………………170
5. 3. Non-quantitative polymerase chain reaction (PCR)…………………………………..170
5. 3. 1. Dissolving the gel slice…………………………………………………………...…..171
5. 4. Cloning……………………………………………………………………………….……172

5. 5. DNA methylation analysis by the Southern blot……………………..………………..173
5. 5. 1. DNA digestion for the Southern probe preparation……………………….…………173
5. 5. 2. Labeling, cleaning and assessing of Southern probes activity………………………173
5. 5. 3. DNA digestion using the methylation sensitive enzymes…………………………….174
5. 5. 4. DNA Blotting and hybridization with the probe……………………………………...175
5. 6. RNA isolation ………………………………………………………………….……176
5. 6. 1. RNA isolation using TRI Reagent……………………………………………………..176
5. 6. 2. RNA isolation from nuclear and cytoplasmic cell fractions…………………..……..176
5. 7. DNaseI treatment……………………………………………………………….....…177
5. 8. Expression analysis…………………………………………………………………..177
5. 8. 1. Reverse transcription (RT) reaction……………………………………….…………177
5. 8. 2. Non-quantitative RT-PCR……………………………………………………………178
5. 8. 3. Quantitative RT-PCR using SYBR and TaqMan Assays……………………….…178
5. 8. 4. Long range RT-PCR/ Proofreading RT-PCR ……………………………………..180
5. 8. 5. Allelic expression analysis using SNPs……………………………………….……180
5. 8. 6. Northern blotting……………………………………………………………………181
5. 8. 7. Sample preparation for the HIRTA hybridization…………………………….…182
5. 8. 7. 1. First strand reverse transcription using SuperScript II Reverse Transcriptase….182
5. 8. 7. 2. Second strand reverse transcription…………………………………………………182
5. 8. 7. 3. Preparation of sonicated genomic DNA.……………………………………………183
5. 8. 7. 4. Quality control using Agilent DNA 7500 Bioanalyzer…………………………....183
5. 8. 7. 5. Hybridization to the HIRTA chip……………………………………………………183
5. 8. 7. 6. Visualization of the HIRTA hybridization data…………………………….…..184
5. 8. 7. 7. Bioinformatics analysis of the HIRTA reproducibility……………..…………..…184
5. 8. 8. Sample preparation for the rRNA depleted total RNA Sequencing…………….184
5. 8. 8. 1. Ribosomal RNA depletion……………………………………………………………..184
5. 8. 8. 2. Hydrolysis of rRNA depleted total RNA…………………………………….……..185
5. 8. 8. 3. First strand cDNA preparation for RNA Sequencing………………………………185
5. 8. 8. 4. Second strand cDNA preparation for RNA Sequencing…………………………..185
5. 8. 8. 5. Library preparation and Illumina/Solexa sequencing………………………………186
5. 8. 8. 6. Bioinformatics analysis of the RNA Sequencing data…….……………………….186
5. 9. Bioinformatics analysis of macro ncRNA potential……………………..………...186
5. 10. Bioinformatics analysis of direct repeats…………………………………..…...186
5. 11. Appendix table I- Materials……………………………………………………….....187
5. 12. Appendix table II- Primers for testing allelic expression…………………………..188
5. 13. Abbreviations…………………………………………………………………………...190
6. References………………………………………………………………………….....…191
7. CV……………………………………………………………………………………………216
8. Lebenslauf……………………………………………………………………………....218
9. Acknowledgements…………………………………………………………………….220

Zussamenfassung

Genomweite Transkriptomstudien haben unterschiedliche Klassen von Nicht-Protein-kodierenden (nk) RNS offenbart und Fragen nach der Komplexität und Regulation des Genoms aufgeworfen. Genomische Prägung, ein epigenetisches Phänomen das die Exprimierung eines Gens in diploiden Zellen auf ein von zwei elterlichen Chromosomen beschränkt, ist ein Modellsystem um die Funktion der makro oder langen nkRNAs zu studieren. Sechs gut erforschte geprägte Gen- Cluster enthalten jeweils eine makro nkRNS die zwischen Maus und Mensch konserviert ist. Weiters wurde gezeigt, dass die zwei murinen makro nkRNAs *Airn* und *Kcnq1ot1* die Exprimierung aller Protein-kodierenden Gene in den jeweiligen Gen-Clustern *Igf2r* und *Kcnq1* unterdrücken. Beim Menschen exprimieren 8 von 27 bekannten geprägten Gen-Regionen geprägte makro nkRNAs. Um herauszufinden ob makro nkRNAs ein universelles Merkmal von allen menschlichen geprägten Gen-Regionen sind, habe ich individuell entwickelte Human Imprinted Tiling Arrays (HIRTA) und RNA-Sequenzierungstechnologien verwendet. Durch Hybridisierung von cDNA von menschlichen Geweben (20 normale und 23 Krebsgewebe) habe ich gewebespezifische Exprimierungsprofile von menschlichen geprägten Regionen erhalten. Dadurch konnte ich 101 neue Transkripte kartieren von denen 95% durch einen bioinformatische Analyse als macro nkRNS bestätigt werden konnten. Die RNA-Sequenzierung von nicht-ribosomaler RNA einer Fibroblasten Zelllinie ergab 26,2 Millionen einmalig zugeordnete Sequenzfragmente. Damit konnte ich sowohl die Exprimierung bereits bekannter geprägter makro nkRNAs erfolgreich detektieren als auch die Exprimierung von 22/23 neuen makro nkRNAs bestätigen die ich mittels HIRTA gefunden habe. Sieben neue makro nkRNAs sind entwicklungsspezifisch reguliert wie ich mittels eines Differenzierungssystems in embryonalen Stammzellen zeigen konnte. Die weitere Charakterisierung von zehn makro nkRNAs hat gezeigt, dass 4/10 ausschliesslich im Zellkern vorliegen, 6/10 monoallelisch oder bevorzugt monoallelisch exprimiert werden und 2/10 einen CpG Insel Promoter haben der unterschiedlich stark auf den beiden elterlichen Chromosomen methyliert ist (DMR). Zusammengenommen habe ich Beweise für sechs neue geprägte makro nkRNAs. Weiters sind 22 der 101 neu kartierten Transkripte nur in Krebsgeweben exprimiert. Diese könnten einen wertvollen Startpunkt für weiterführende Biomarker Forschung darstellen. Weiters konnte ich zeigen, dass alle menschlichen geprägten Gen-Regionen zumindest je eine makro nkRNA exprimieren welche geprägt sein könnte und diese daher möglicherweise eine Rolle in der Genregulation unter normalen als auch krankhaften Zuständen beim Menschen spielt.

Abstract

Recent genome-wide transcriptome studies revealed diverse classes of non-protein-coding (nc)RNAs and raised questions about the complexity and regulation of the genome. Genomic imprinting (an epigenetic phenomenon that restricts gene expression to one of two parental alleles in diploid cells) is a model system to study the function of an unusual class of macro or long ncRNAs. Six well-studied imprinted gene clusters contain a macro ncRNA that are mainly conserved between mouse and human. Furthermore, the mouse *Airn* and *Kcnq1ot1* macro ncRNAs have been shown to repress all protein-coding genes, respectively in the *Igf2r* and *Kcnq1* imprinted gene clusters. In humans, 8 out of 27 known imprinted gene regions express imprinted macro ncRNAs. To determine if macro ncRNAs are universal features of all human imprinted gene regions I used a custom Human Imprinting Region Tilling Array (HIRTA) and RNA-seq technologies. By applying 20 normal and 23 cancer human samples to HIRTA, I obtained the tissue-specific expression profiles of human imprinted gene regions and based on these profiles I mapped 101 novel transcripts of which about 95% were confirmed as macro ncRNA using a bioinformatics approach. Using ribosomal RNA depleted RNA-seq in a fibroblast cell line, I obtained 26.2 million uniquely mapped reads, successfully detected the known imprinted macro ncRNAs and validated expression of 22/23 novel macro ncRNA transcripts detected by HIRTA. 7 novel macro ncRNAs were developmentally regulated in the human embryonic stem cells differentiation system. Characterization of 10 macro ncRNAs showed that 4/10 were exclusively nuclear localized, 6/10 had monoallelic or expression biased towards one parental allele and 2/10 had CpG island promoters that are differentially methylated regions (DMRs). Thus, I have partial evidence for 6 novel imprinted macro ncRNAs. Furthermore, 22 out of the 101 mapped transcripts were expressed exclusively in cancer samples and may represent a valuable starting point for biomarkers research. In summary, all human imprinted gene regions express at least one macro ncRNA that may be imprinted and potentially play a role in gene regulation in normal or disease conditions in human.

1. Introduction

1. 1. Epigenetics and non-protein coding RNAs

1. 1. 1. Definition of epigenetics

Epigenetics has been defined many times through history. Aristotle (384-322 BC) used the term epigenesis in order to define gradual development of individual organic form from unformed in his book “Generation of Animals”. The term epigenetics (epi-greek: above-genetics) was for the first time defined by Conrad Waddington, 1942, as “the branch of biology which studies the causal interactions between genes and their products, which bring the phenotype into being” (Waddington, 1942). One of the phenomena explained as epigenetic by Waddingtons’ work is cellular differentiation. In 1957, he modeled the cellular differentiation process through his epigenetic landscape concept where a ball representing the cell is able to take different permitted trajectories leading to different cell fates (Figure 1).

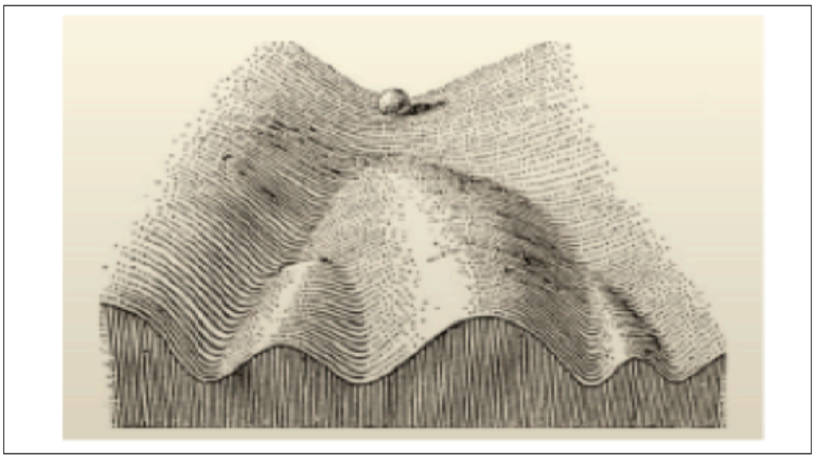


Figure 1. Waddingtons’ epigenetics landscape. Taken unmodified from (*Waddington, 1957*).

One of the most accepted definitions of epigenetics, given by Arthur D. Riggs in 1996, is: “the study of mitotically and/or meiotically heritable changes in genes function that cannot be explained by changes in DNA sequence”. Epigenetics today is a fast growing field. The epigenetic community took a collective effort to define and discuss epigenetics at a meeting hosted by the Banbury Conference Center and Cold Spring Harbor Laboratory which resulted in an operational definition of epigenetics being given by Ali Shilatifard and colleagues in 2009: “An epigenetic trait is a stably heritable phenotype resulting from changes in a chromosome without alterations in the DNA sequence” (Berger et al., 2009). The operational definition will further be discussed in the next section where the epigenetic pathway and its set of operational steps will be dissected. Epigenetics can be defined in different ways but as a conclusion I would point to recent Denise Barlows' quotation: "Epigenetics has always been all the weird and wonderful things that cannot be explained by genetics" (http://epigenome.eu/en/1,1,0).

1. 1. 2. The epigenetic pathway

A recent conference and discussion including a number of leading scientists in the field led to an operational definition of epigenetics based on three categories of signals leading to a stably heritable epigenetic state (Berger et al., 2009). Signals named: "Epigenator", "Epigenetic Initiator" and "Epigenetic Maintainer" are proposed to be parts of the epigenetic pathway (Figure 2).

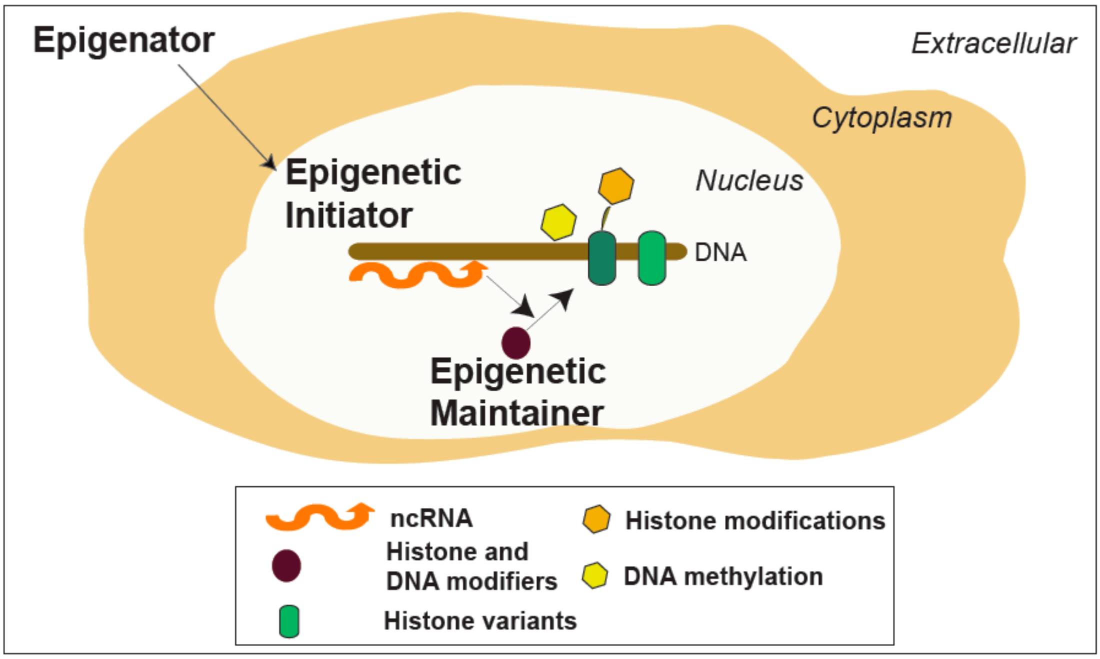


Figure 2. The epigenetic pathway. Includes epigenator that comes from enviroment and "activate" locus specific epigenetic initiator (e.g. ncRNA) leading to establishment of specific chromatin enviroment by action of epigenetic maintainer signals.

The "Epigenator" comes from environment and "activates" the "Epigenetic Initiator". This epigenator signal is transient and could be for example changes in temperature that affect paramutation in plants, where paramutation is a phenomenon in which one allele causes heritable expression change of the homologous allele. The "Epigenetic Initiator" is locus specific. By action of Initiators, such as non- protein coding RNAs (ncRNAs) and DNA binding proteins, a specific chromatin environment is established at a particular location in the genome. The Initiator is not necessarily transient as the "Epigenator" but may persist together with "Epigenetic Maintainer". The "Epigenetic Maintainer" is not locus specific and could operate at any location where it is recruited by an Initiator. Maintainers include a variety of epigenetic signals like DNA methylation, histone modifications, variants of histones and nucleosome positioning. The division of epigenetic pathways into three classes of signals provides a basis for understanding epigenetic pathways and their role in gene regulation through cellular generations. Further investigation of epigenetic pathways, signals and their interconnection is necessary and will provide a better understanding of a whole layer of information “above the genome” implicated in normal development and disease.

1. 1. 3. Epigenetic roles of DNA methylation

DNA methylation is a covalent modification of DNA, in mammals restricted to cytosine residues and almost exclusively found on CpG dinucleotides (Figure 3).

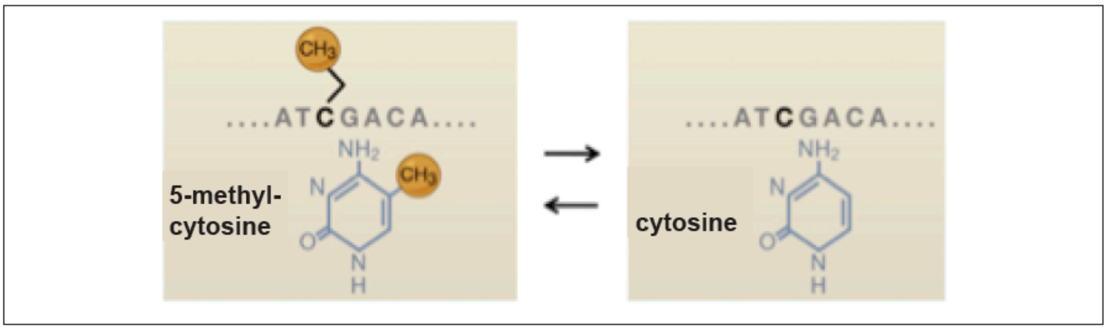


Figure 3. Cytosine methylation. Modified from (*Bernstein et al., 2007*).

CpG dinucleotides in the human genome are present at low density (93%) or concentrated in CpG islands (7%) (Fazzari and Greally, 2004). CpG islands were originally defined as regions of DNA of at least: 200bp length, 55% GC content and 0.6 ratio of observed to expected CpG frequency. Today the CpG island length is usually set to more than 500bp since this lowers the number of false positives (Gardiner-Garden and Frommer, 1987; Takai and Jones, 2002). CpG island promoters are a feature of around 56% of genes in the human genome (Antequera and Bird, 1993). Large-scale studies of cytosine methylation in human showed that global DNA methylation is found throughout the genome: in gene bodies, transposons and intergenic DNA, while CpG islands are mostly unmethylated. No correlation between unmethylated CpG island promoters and expression state of the gene could be found. Exceptionally, methylated CpG islands are involved in X chromosome inactivation and genomic imprinting while their involvement in tissue specific gene silencing is starting to emerge (reviewed in (Suzuki and Bird, 2008)).

DNA methylation is clearly an epigenetic modification as it provides heritable information not encoded by nucleotide sequence. Mouse embryogenesis research has provided insights into the DNA methylation “cycle”. In the preimplantation embryo genome-wide loss of DNA methylation is observed (Monk et al., 1987). In the germline, DNA methylation patterns are set by *de novo* methyltransferases DNMT3A and DNMT3B (Okano et al., 1999). DNMT3A2 isoform and its cofactor DNMT3-like (DNMT3L) introduce DNA methylation to female gametes imprint control elements (ICEs) which are one of the key features of imprinted gene regions (Hata et al., 2002). DNMT3L is not necessary for DNMT3A2 establishment of DNA methylation imprints in sperm (Kaneda et al., 2004). DNMT3B is shown to be required for methylation of pericentromeric repetitive DNA and CpG islands of the inactive X chromosome (Bird, 2002). Propagation of methylation patterns through cell divisions is achieved by involvement of the “maintenance” methyltransferase DNMT1. Specifically hemi- methylated CpG dinucleotides are targeted by this enzyme to methylate newly synthesized DNA strand based on complementary strand methylation (Bird, 2002). However, it was shown that DNMT1 is inefficient at maintaining methylation at some CpG dense regions (Liang et al., 2002), therefore current opinion is that DNMT3A and DNMT3B are also taking part in maintenance of the methylation (Miranda and Jones, 2007).

One of the best-documented roles of DNA methylation is a role in transcriptional gene silencing. Reduction of DNA methylation has been shown to lead to derepression of LINE (Long Interspersed Nuclear Elements) and SINE (Short Interspersed Nuclear Elements) promoters in the human genome (Liu et al., 1994; Woodcock et al., 1997). For example artificial demethylation of promoters of Alu family elements stimulates their expression (Liu et al., 1994). Silencing of repetitive elements by DNA methylation gives support to the genome defense model (Yoder et al., 1997). According to this model DNA methylation has a biological role in silencing repetitive elements and thereby preventing DNA damage due to their transposition (Robertson and Wolffe, 2000).

Long term silencing of transcription by DNA methylation of CpG islands is well established in two forms of non-Mendelian inheritance: X chromosome inactivation and genomic imprinting. During X-inactivation, one female X chromosome is inactivated. CpG island promoters are methylated within the inactive X chromosome by DNA methylation which stabilizes the repressed state of the corresponding genes (Chang et al., 2006). Interestingly, gene-body DNA methylation is two times more abundant on the active X than on the inactive X chromosome, which has been showed by allele-specific DNA methylation of more than 1000 informative loci on the human X chromosome (Hellman and Chess, 2007). Thus, X-inactivation provides evidence of gene-body methylation association with transcriptional activity.

In imprinted gene clusters DNA methylation is present in the form of DMRs (Differentially Methylated Regions) representing CpG islands methylated on one parental allele. DMRs have function as imprint control elements (ICE) if acquired in the primordial germline. The retention of the imprinted gene state through cellular generations is achieved by DNA methylation silencing of ICE that is a *cis*-acting repressor on one parental allele (Koerner and Barlow, 2010).

Different models suggest mechanisms by which DNA methylation could be involved in transcriptional gene silencing. Two common views are that DNA methylation could impede the binding of transcription factors (e.g. CTCF (Bell and Felsenfeld, 2000)) or it could promote binding of MBPs (Methyl-Binding Proteins (Kass et al., 1997)) which by binding to repressors and histone deacetylases could lead to inactive chromatin formation (Harikrishnan et al., 2005).

The epigenetic role of DNA methylation in development and differentiation is still not fully understood. *Weber et al. (2007)* found systematic methylation of germline-specific gene promoters in somatic cells (fibroblasts) while this methylation was absent from mature sperm. This finding implicates DNA methylation as having a role in the silencing of germline specific genes that could contribute to cell differentiation. Homeobox (HOX) and paired box (PAX) genes play important roles in development. DNA methylation has been found on the CpG island promoters of these genes in human somatic tissues implying a role in development (Illingworth et al., 2008). Further, *Lister et al. (2009)* compared human stem cells and fetal fibroblasts at base resolution and identified hundreds of DMRs proximal to genes playing a role in pluripotency and differentiation. Several studies examined DNA methylation in human and showed that DNA methylation of CpG island promoters is more spread in somatic tissues than thought. For example, one group mapped DNA methylation genome-wide profiles in 13 human somatic tissues and proposed that promoter methylation could play role in the context of cell and tissue specific transcriptional programs (Rakyan et al., 2008). Tissue-specific differentially methylated regions (tDMR) are found in both gene-coding and intergenic regions. tDMRs, outside of annotated genes, could potentially play a role as *cis*- regulatory elements involved in gene expression control or as promoters of not yet annotated non-protein coding RNAs (Suzuki and Bird, 2008). The roles of DNA methylation in transcriptional gene silencing, development and differentiation, together with the interplay with other "epigenetic maintainers" such as histone modifications and "epigenetic initiators" such as ncRNAs remain to be fully understood.

1. 1. 4. Epigenetic roles of histone modifications

DNA is wrapped around the histone octamer (two times of each H3, H4, H2A, H2B histone) forming the nucleosome, the basic unit of chromatin. Postranslational modifications can be found on the N-terminal tails or globular domains of histones. There is a debate in the field if histone modifications are epigenetic since it is not clear how and if chromatin modifications are heritable. Still, their roles in epigenetic processes are evident and recently they are recognized as potential "Epigenetic maintainers". Histone modifications have roles in transcription regulation, repair, replication and condensation. Different classes of histone modifications in mammals are shown in Table 1.

Histone modification classes	Residues modified	Histone modifications
Acethylation	K	H3K9, H3K14, H3K18, H3K56, H4K5, H4K8, H4K12, H4K16, H2AK5, H2BK12, H2BK15
Methylation	K, R	H3K4, H3K9, H3K27, H3K36, H3K79, H4K20, H2BK5, H3R2, H3R8, H3R17, H3R26, H4R3
Phosphorylation	S, T	H3S10, H3S28, H4S1, H2BS14, H3T3
Ubiquitylation	K	H2AK119, H2BK120
Sumoylation	K	H4 (Shiio and Eisenman, 2003)
ADP ribosylation	E	H3, H4, H2A, H2B (Burzio et al., 1979)
Deimination	R>C	H3, H2A, H4 (Hagiwara et al., 2002)
Proline Isomerization	P-cis>P-trans	H3P30, H3P38

Table 1. Eight classes of histone modifications are overviewed. Histone amino acid residues modified by different chemical reactions are: K(lysine), R(arginine), S(serine), T(threonine), E(glutamic acid), C(cysteine) and P(proline). Histone modifications found on specific positions of histones are shown (reviewed in (Kouzarides, 2007), (Barski et al., 2007)). Lysine methylation and arginine methylation can show an even higher diversity as mono(me1), di(me2) and tri(me3) methylations are found (where one, two or three methylation groups are subsequently bound) on lysine and mono(me1) or two kinds of di(me2a, me2s) methylations are found on arginine.

Genome wide approaches such as ChIP-chip and ChIP-Seq made it possible map histone modifications in different cell lines and tissues where different modification patterns can be correlated with transcription of the genes. An overview of modifications mapped on specific locations of active/inactive genes, specific regulatory regions and other genetic/epigenetic elements is shown in Table 2 (reviewed in (Wang et al., 2009b)). Use of chromatin maps to map novel genes and regulatory regions is accepted on the basis of the strong correlation between certain histone modifications or their combinations with specific genomic regions. For example H3K4me3 is associated with promoters of expressed genes.

Specific genetic/epigenetic regions	Histone modifications
Promoters of active genes	H3K4me3 (H3K4me2/me1 spreads towards transcribed regions of genes), H3K9ac, H3K14ac, H3K23ac, H2BK5me1, H3K9me1, H3K27me1, H4K20me1, H3K79me1 (H3K79me2/me3 spreads towards the transcribed regions of genes), H4K5/8/12/16ac (spreading through gene body), reviewed in (Wang et al., 2009b)
Active genes body	H3K36me3, H3K79, H3K9me1, H3K27me1, H4K20me1, H3K23ac (decreasing through the gene body), H2BK5me1, H3K9me1, H3K27me1, H3K36me3, H4K5/8/12/16ac, reviewed in (Wang et al., 2009b)
Inactive genes body	H3K9me2/me3, H3K27me2/me3, H4K20me3 reviewed in (Wang et al., 2009b)
Repeats (e. g. LTRs (Long Terminal Repeats))	H3K9me3 + H4K20me3 (Mikkelsen et al., 2007)
Enhancers	H3K4me1 (absence H3K4me3) (Heintzman et al., 2007) H2A.Z + H3K4me3 + H3K4me1 (Barski et al., 2007)
Imprint control elements (ICE)	Modifications depend on the differential methylation status: Methylated ICE: H3K9me3, H4K20me3; Unmethylated ICE: H3K4me3, H3K4me2, H3K9Ac (Regha et al., 2007) Methylated allele: H3K9me3, H4K20me3; Unmethylated allele: H3K4me2, H3ac (Delaval et al., 2007) H3K4me3 + H3K9me3 in ES cells (Mikkelsen et al., 2007)

Table 2. Specific genomic regions enriched by characteristic histone modifications are listed. +; Histone modifications on the same locus on two parental chromosomes.

Histone modifications are dynamic and set or reversed (except methylation of arginine for which demethylating enzyme is still not found) by a number of histone modifying enzymes (reviewed in (Kouzarides, 2007)). The role of histone modifications in DNA condensation and possibly change in higher order chromatin structure is expected and partially shown for acetylation and phosphorylation via their influence on charge changes. For example, Shogren-Knaak at al. (Shogren-Knaak et al., 2006) showed that H4-K16Ac controls chromatin structure. H4 that was homogenously acetylated on K16 by a native chemical ligation strategy after incorporation into nucleosomal arrays impedes formation of one of the levels of chromatin compaction (30-nanometar like fibers). Histone modifications are regulating transcription, repair and replication by recruiting number of other proteins on histones. For example Guccione at al. (Guccione et al., 2006) found that tethering of Myc transcription factor in human is restricted to a stretch of chromatin carrying H3 K4/K79 and H3Ac. This is just one of the examples of proteins recruited by histone modifications functioning in the activation or repression of transcription (reviewed in (Li et al., 2007)). The question about inheritance of histone modifications remains to be answered. One of the models how the memory of specific chromatin states could be transmitted involves RNA as a molecule “carrying” the memory while histone modifications could be the executers of the epigenetic phenomena (Kouzarides, 2007). This model is in agreement with already discussed “Epigenator-Initiator-Maintainer” model of epigenetic pathway.

1. 1. 5. Epigenetic roles of ncRNAs

The central dogma of molecular biology (Francis Crick, 1958, 1970) has been to see RNA as an intermediary molecule into which genomic information from DNA is transcribed and which is further translated into proteins. This dogma is based on the assumption that one gene is coding for one protein has been challenged by the finding that most of the genome is transcribed into non-protein coding RNAs, adding a new layer of complexity (Birney et al., 2007) (Figure 4).

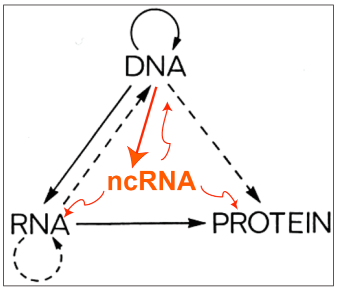


Figure 4. New layer of complexity has been added to Cricks' central dogma of molecular biology. Black; Original figure representing interconnections between DNA, RNA and protein, Orange; Solid arrow shows transcription of ncRNA from DNA, curved arrows represent impact of ncRNAs on DNA, RNA and proteins through different functions of ncRNAs. Modified from Crick (Crick, 1970).

Non-protein coding RNAs (ncRNAs) are not translated into proteins but they are functional RNAs. ncRNAs could be grouped depending on their length (small<200bp and long or macro ncRNAs>200bp), location (nuclear enriched and cytoplasmic RNAs), orientation to protein coding genes (sense or antisense) and function (RNAs functioning *in cis* or *in trans*; housekeeping RNAs and regulatory RNAs). Already in 1961, Jacob and Monod proposed potential interaction of sequence-specific ncRNAs with promoters, which could regulate genes (Jacob and Monod, 1961).

Today it has been shown that a number of long/macro regulatory ncRNAs are gene regulators implicated in wide range of complex epigenetic phenomena such as X-inactivation and genomic imprinting and that they are involved in development. *Xist* (X-Inactive-Specific-Transcript) ncRNA mediates whole chromosome transcriptional silencing during the dose compensation process in mammals. Dosage compensation involves inactivation of n-1 X chromosome in females, where n is the number of X chromosomes. Upregulation and coating of X chromosome by *Xist* ncRNA are the first signs of X-chromosome inactivation (XIC). After the coating numerous changes in chromatin modifications and enrichment in histone variant macroH2A are observed. Observed chromatin modifications on the inactive X chromosome include: loss of “active” modifications such as H3K9Ac and H3K4me3, H4 hypoacetylation and H3K27 hypermethylation, H3K9 hypermethylation, H4K20 monomethylation as well as H2A K119 monoubiquitylation. The modifications lead to establishment of silent chromatin, which keeps X chromosome inactivated throughout the cell cycle (Prasanth and Spector, 2007). *Xist* deletion and transgene analyses showed that *Xist* ncRNA is essential for X-inactivation (Penny et al., 1996; Wutz and Jaenisch, 2000). Beside *Xist* ncRNA, *Tsix*, *Xite* (X-inactivation Intergenic Transcription Elements) and *Jpx/Enox* ncRNAs are involved in mouse dosage compensation. In mice *Tsix* and *Xite* together with part of the *Xist* 3' region are regulating *Xist* ncRNA expression (Heard and Disteche, 2006). The exact mechanism of these RNAs involvement in X-chromosome dosage compensation remains to be understood.

The epigenetic roles of numerous macro ncRNAs involved in genomic imprinting show that imprinting is a valuable model in ncRNA research. An example of a long, non-imprinted RNA that has an epigenetic role and regulates development by transcriptional repression is the *HOTAIR* ncRNA. Human *HOX* genes are organized in 4 clusters (*HOXA-D*) localized on different chromosomes. HOX proteins are transcription factors regulating correct body axis development. Among 231 ncRNAs mapped in human *HOX* clusters in 11 fibroblast cell lines from distinct positions along the body axes, *HOTAIR* ncRNA was found in the *HOXC* cluster (Rinn et al., 2007). Knockdown of *HOTAIR* showed a loss of transcriptional repression, lost of H3K27me3 repressive histone modification and lost of PRC2 (Polycomb Repressor Complex 2) in a 40kb region of the *HOXD* cluster while expression of the *HOXC* cluster was not influenced. These findings by Rinn et al. (2007) and an experiment where PRC2 pull-down showed a specific interaction with *HOTAIR*, are suggesting the epigenetic pathway of the *HOTAIR* ncRNA action in development (Woo and Kingston, 2007).

An example of a ncRNA that regulates development by transcription activation is *Evf-2*. This 3.8kb long ncRNA is found in mouse to form a stable complex with *Dlx-2* (homeodomain protein important for development of neural system) that stabilizes an interaction between the *Dlx-2* and *Dlx5/6* enhancer leading to activation of transcription of target genes (Feng et al., 2006).

Epigenetic roles have also been described for four small regulatory RNAs classes (small nucleolar RNAs (snoRNAs), micro RNAs (miRNAs), short interfering RNAs (siRNAs) and piwi- interacting RNAs (piRNAs)). SnoRNAs guide 2'-O-ribose methylation and pseudouridylation to nucleotides of ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), spliceosomal snRNAs and possiblly mRNAs (Bachellerie et al., 2002). *SNORD115* (*HBII-52*) is a snoRNA with a different function: regulating alternative splicing of the serotonin receptor 2C (Kishore and Stamm, 2006). SnoRNAs can also be processed to small RNAs that can function as miRNAs as for example with the *ACA45* snoRNA (Ender et al., 2008).

MiRNAs have roles in post-transcriptional gene regulation, reviewed in (Malone and Hannon, 2009; Mattick and Makunin, 2005). MiRNAs are 21-25bp in length. They are transcribed by RNA Polymerase II (Borchert et al., 2006; Lee et al., 2004) or RNA Polymerase III (Borchert et al., 2006) as primary RNA (pri-miRNAs) and processed by a complex containing Drosha (RNaseIII) to pre-miRNAs (hairpin structured) and further transported to the cytoplasm where Dicer (RNaseIII) generate a short dsRNA that is unwound leading to a single stranded mature miRNA. MiRNAs are incorporated into miRNPs (micro ribonuclear particles) which in mammals usually target complementary or partially complementary 3'UTRs of protein coding genes mRNAs and can in most cases suppress translation (imperfect match) or cleave target mRNAs (perfect match). Alterations in their activities are associated with cancer and numerous diseases.

SiRNAs are 20-25 nucleotides in length and processed similarly to miRNAs. Precursors of siRNAs are endogenous or exogenous double-stranded RNAs (e.g. viral). Dicer directly cuts these precursors and produces short RNAs that together with Argonaute (AGO) proteins form RISC (RNA-induced silencing complex) where just one strand of the short RNAs is retained (reviewed in (Mattick and Makunin, 2005)). Endogenous siRNAs were found in mammals for the first time in 2008 when Tam et al. (Tam et al., 2008), as well as Watanabe et al. (Watanabe et al., 2008) showed their presence in mouse oocytes. Interestingly, Tam et al., found that endogenous siRNAs are processed from double-stranded RNAs formed when protein-coding transcripts are hybridized to their homologous pseudogene transcripts, or from inverted pseudogene transcripts alone. They showed regulatory activity of these siRNAs in repression of mobile genetic elements. Further, Watanabe et al., found that numerous 21bp endogenous siRNAs found in mouse oocytes correspond to mRNAs or retrotransposons and function in the regulation of gene expression. Synthetic siRNAs are used as an important tool in removal of targeted mRNAs and have potential to be used as therapeutic agents.

PiRNAs were discovered in 2006 (Aravin et al., 2006; Girard et al., 2006; Grivna et al., 2006; Watanabe et al., 2006). These small ncRNAs are associated with germline specific Piwi proteins (Argonaute protein family). Two classes of Piwi proteins (MIWI and MILI) have been found to associate with piRNAs. MIWI associated piRNAs are 30-31 bp long while MILI associated are 26-28bp in length (reviewed in (Thomson and Lin, 2009)). The epigenetic roles of these small RNAs are still unclear. Xu at al. (2008) showed the potential role of piRNAs in repression of transposons since deletion of piRNA cluster in mice leads to increased transposon activity (Xu et al., 2008). Evidence that piRNAs are involved in directing *de novo* DNA methylation by an as yet unknown mechanism are reviewed by Aravin and Bourc’his (Aravin and Bourc'his, 2008).

Current understanding of epigenetic roles of long and small regulatory RNAs has been described. Regulatory RNAs are recently recognized as “Epigenetic Initiators” whose known role in epigenetic pathways are predicted to be the “tip of the iceberg” that remain to be uncovered.

1. 2. Genomic imprinting as a model in ncRNA research

1. 2. 1. Genomic imprinting is an epigenetic phenomenon

Genomic imprinting is an epigenetic phenomenon that restricts expression of a gene to one of two parental chromosomes. While most genes are expressed from both parental alleles in diploid cells, imprinted genes show paternal or maternal monoallelic expression. Different epigenetic mechanisms involving DNA methylation, chromatin modifications and ncRNAs lead to monoallelic expression of imprinted genes from just one of two identical DNA copies. The imprinted status of genes is clearly a heritable epigenetic phenomenon that is passed through cell divisions.

1. 2. 2. Discovery of genomic imprinting

The discovery that the genomes of sperm and egg are different originates from the mid 1970’s. For example, Linder et al. (Linder et al., 1975) showed that the developmental potential of human oocytes depends on the parental genome driving that development. While normal ovarian germ cells have both the mothers’ and fathers’ set of chromosomes, Linder showed that they can give rise to two kinds of tumors: teratomas that are gynogenetic (two mothers’ chromosome sets) and hydatidiform mole that are androgenetic (two fathers’ chromosome sets). The observation that these two tumors are histopathologically very different: teratomas are consisting of all three germinative layers while hydatiform mole contains only trophoblast elements, indicated a difference between the parental genomes. The development of nuclear transfer technology used in experiments examining mouse parthenogenesis in 1980' directly showed that both parental genomes are required for the embryo to develop (McGrath and Solter, 1984; Surani et al., 1984). After removing the male or female pronucleus from one fertilized egg and adding one of two parental pronuclei to the same egg three kinds of diploid embryos were obtained: wildtype with one maternal and one paternal nucleus, gynogenetic with two maternal genomes and androgenetic with two paternal genomes. Both gynogenetic and androgenetic embryos were lethal, whereas the wild type survived. This explained why there is no parthenogenesis in mammals and posed the question of the difference between the parental genomes indicating that this difference could be in the parental specific expression of developmentally important genes. The first imprinted genes were found in 1991, when *Igf2r* (Insulin-like growth factor type 2 receptor) was shown to be maternally expressed (Barlow et al., 1991), *Igf2* was found to be paternally expressed (Insulin-like growth factor type 2) (DeChiara et al., 1991; Ferguson-Smith et al., 1991) and *H19* (Hepatic library clone 19) non-coding RNA was shown to be a maternally expressed imprinted gene in mouse (Bartolomei et al., 1991). These findings led to the discoveries of more imprinted genes and to the establishment of the genomic imprinting field.

1. 2. 3. Evolution of genomic imprinting

Genomic imprinting evolved independently at least three times. This phenomenon is present in angiospermic plants (e.g. *Arabidopsis*), in some Insecta (e.g. *Sciara*, *Coccidae*) and in Therian Mammals (Das et al., 2009). In plants a small subset of genes is imprinted in endosperm by a mechanism involving targeted demethylation leading to activation of the expressed allele (Scott and Spielman, 2006), while in Insecta the whole paternal genome heterochromatization is named genomic imprinting (Khosla et al., 2006).

Genomic imprinting in Therian mammals is complex and a number of hypotheses have attempted to describe: 1) how the imprinting mechanism arose and 2) the evolutionary driving force that could explain why imprinting arose around 125 million years ago and is still present in Therian Mammals. “The host defense” hypothesis is one of the rare hypotheses addressing directly the mechanism of how imprinting arose (Barlow, 1993). This hypothesis raises the possibility that in order to defend the host genome retrotransposons and foreign DNA were DNA methylated and repressed, and this system also methylated imprinted genes. The existence of retrotransposed imprinted genes e.g. *PEG10* could support this theory.

Theories about the potential adaptive advantage gained by imprinting in mammals are specially addressed in order to explain the obvious disadvantage resulting from the nature of imprinting: monoallelic expression of genes cause functional haploidy, thus all mutations of imprinted genes are dominant (imprinted expression leads to increased risk of genetic diseases and cancer). The “Parental conflict” hypothesis (Moore and Haig, 1991) proposes that paternally expressed imprinted genes increase growth leading to the enhancement of fitness of the offspring carrying the paternal genome while maternally expressed imprinted genes suppress fetal growth maximizing reproductive fitness of mother thus enhancing transmission of maternal genome to more offsprings with the potentially different paternal genomes. Imprinting has evolved in response to viviparity and polygamy according to this hypothesis.

“Trophoblast defense” or “ovarian time bomb” hypothesis (Varmuza and Mann, 1994; Weisstein et al., 2002) propose that imprinting evolved in order to protect oocytes and thus female mammals from ovarian teratomas. Imprinted genes “defend” mother from potential malignant trophoblast formation that could arose from partenogenetically activated oocytes developing into ovarian teratomas. Thus selective pressure through evolution would favor females with imprinted genes not developing lethal trophoblast disease. Trophoblast is an invading part of the placenta that mediates implantation of fetus to the mothers’ uterus. Thus silencing of the genes functioning to promote placental development and activation of the genes limiting placental development by imprinting is expected to “defend” the mother from cancer development.

Further, Kono hypothesized that imprinting may evolve as a defense against parthenogenesis to select for sexual reproduction (Kono, 2006). The hypothesis that imprinted expression has evolved in order to regulate development (the complementation hypothesis) is proposed by Kaneko-Ishino (Kaneko-Ishino et al., 2006). Imprinting is proposed to evolve on the basis of dosage compensation required for duplicated genes by Walter and Paulsen (Walter and Paulsen, 2003). Recently, the theory of “coadaptation” links evolution of imprinting with its role in regulation of embryonic development and reproductive behavior (Keverne and Curley, 2008).

As described, theories deciphering the mechanism how imprinting arose and the evolution of imprinting are proposed and are more or less supported with novel findings in imprinting research. Potentially, imprinting could have evolved by different mechanisms at different loci (Renfree et al., 2009), but further analysis of the imprinting mechanism in different species and mammalian lineages is needed to fulfill our understanding of imprinting.

1. 2. 4. Key features of genomic imprinting

Genomic imprinting involves a *cis*-acting mechanism affecting one parental chromosome and leading to differences between paternal and maternal alleles. Genes affected by genomic imprinting are usually clustered. Imprinted gene clusters consists of protein-coding gene mRNAs and often include of macro ncRNAs that show reciprocical imprinted expression. Each set of imprinted genes residing in one cluster has a common regulator known as imprinting control element (ICE). Imprinted gene clusters are of different lengths that are up to 4Mb for the human PWS cluster (UCSC genome browser). Some imprinted genes are “orphans” and belong to micro-imprinted gene clusters (e.g. *Nap1l5* (Evans et al., 2001)).

The presence of an imprint control element (ICE) is one of the key features of genomic imprinting. ICEs are defined by deletion experiments as DNA elements whose epigenetic state controls expression of all imprinted genes in their clusters (Koerner and Barlow, 2010). The ICE is DNA methylated on one allele where its function is repressed and is unmethylated on another allele where it can function as a repressor in three potential ways (Kaneda et al., 2004; Koerner and Barlow, 2010). In the mouse *Igf2* (insulin growth factor 2) imprinted gene cluster, the unmethylated ICE binds the CTCF protein and acts as an insulator since its’ binding blocks influence of enhancers on the *Igf2* gene leading to the absence of *Igf2* expression from the chromosome with the unmethylated ICE (Hark et al., 2000). In mouse the *Igf2r* and *Kcnq1* imprinted gene regions unmethylated ICEs contain promoters for *Airn* and *Kcnq1ot1* macro ncRNAs respectively. These ncRNAs are expressed from unmethylated ICE and are repressing protein-coding genes *in cis*. A third possibility for the function of unmethylated ICE is shown in the mouse *H13* imprinted gene cluster where the unmethylated ICE contains active promoter of the *Mcts2* retrogene causing expression of truncated *H13* transcripts (Wood et al., 2008).

Gametic DMRs (Differentially Methylated Regions) are methylated on only one parental allele, are set in parental gametes (mechanism described in 1.1.3.) and have a function as an ICE. After fertilization, genome-wide reprogramming of DNA methylation takes place, but gametic imprints are not reprogrammed and are maintained on the same parental chromosome in all cells up to the adult stage. Germ cells of embryonic gonads are the only cells where gametic imprints are erased at a stage before the sex of the embryo is determined. With the development of mature gonads parent specific imprints are again established in gametes (Barlow, 2007). 16 gametic DMRs are found in mice and have been shown to be maternally (13) or paternally (3) methylated (Table 3, Table 4, section 1.2.5.). In human 6 gametic DMRs have been identified to be maternally (4) or paternally (2) mathylated (Table 3, Table 4, section 1.2.5.). The low number of known human gametic DMRs may be due to of difficulties in obtaining human oocytes.

Gametic DMRs often contain a series of tandem direct repeats (Hutter et al., 2006; Neumann et al., 1995; Paoloni-Giacobino et al., 2007; Sleutels et al., 2002). The role of tandem direct repeats has been examined for Snurf/Snrpn, Kcnq1 and Igf2r DMRs using transgenic mice. Authors proposed that tandem repeats could have a role in the establishment and maintenance of parental specific methylation (Reinhart et al., 2002; Reinhart et al., 2006).

DMRs show common histone modification signatures named DHMs (Differential Histone Modifications). For example our lab showed that the DNA methylated alleles of DMRs are enriched in H3K9me3 and H4K20me3, whereas the unmethylated alleles have H3K4me3, H3K4me2 and H3K9Ac marks (Regha et al., 2007). Genome-wide chromatin maps in mouse ES cells showed that overlapping H3K4m3 and H3K9me3 are common signatures of an ICE (Mikkelsen et al., 2007) (described in Table 2, 1.1.4).

Second class of DMRs are somatic DMRs that are erased during post-fertilization reprogramming of DNA methylation marks and are set in somatic cells of post-implantation embryos. Numerous human DMRs have been found, yet if they are somatic or gametic remains to be tested.

One of the key features of genomic imprinting is *cis*-regulation of imprinted genes over long genomic distances leading to their parental specific monoallelic expression. Genes showing imprinted expression can be maternally or paternally expressed, from the mother or father’s chromosome respectively. Tissue and developmental stage specific differences in parent-of-origin specific imprinted expression are proposed to exist due different “readers” of imprints. For example, specific transcription factors could influence imprinting expression and lead to expression variation of imprinted genes in tissues and in development. Both protein coding genes and non-protein coding RNAs can be imprinted and show parental specific imprinted expression.

1. 2. 5. Human and mouse imprinted gene regions

Imprinted gene regions could be defined as genomic regions where at least one gene showing imprinted expression has been located, although most regions contain more genes (between 2-15) that form an imprinted gene cluster. 26 imprinted gene regions exist in mouse according to the Harwell web site (http://www.mousebook.org/catalog.php?catalog=imprinting) while 27 imprinted gene regions are found in human (Figure 5). Although, the epigenetic initiator responsible for silencing has been identified in some clusters, the mechanisms of silencing are not so well defined and the borders of many imprinted gene regions are unclear. For example, some imprinted genes considered by one research group as parts of a separate imprinted regions are considered by other groups using different criteria as just border genes of an adjacent region. Thus, 25 to 30 imprinted gene regions in both mouse and human could be provisionally defined.

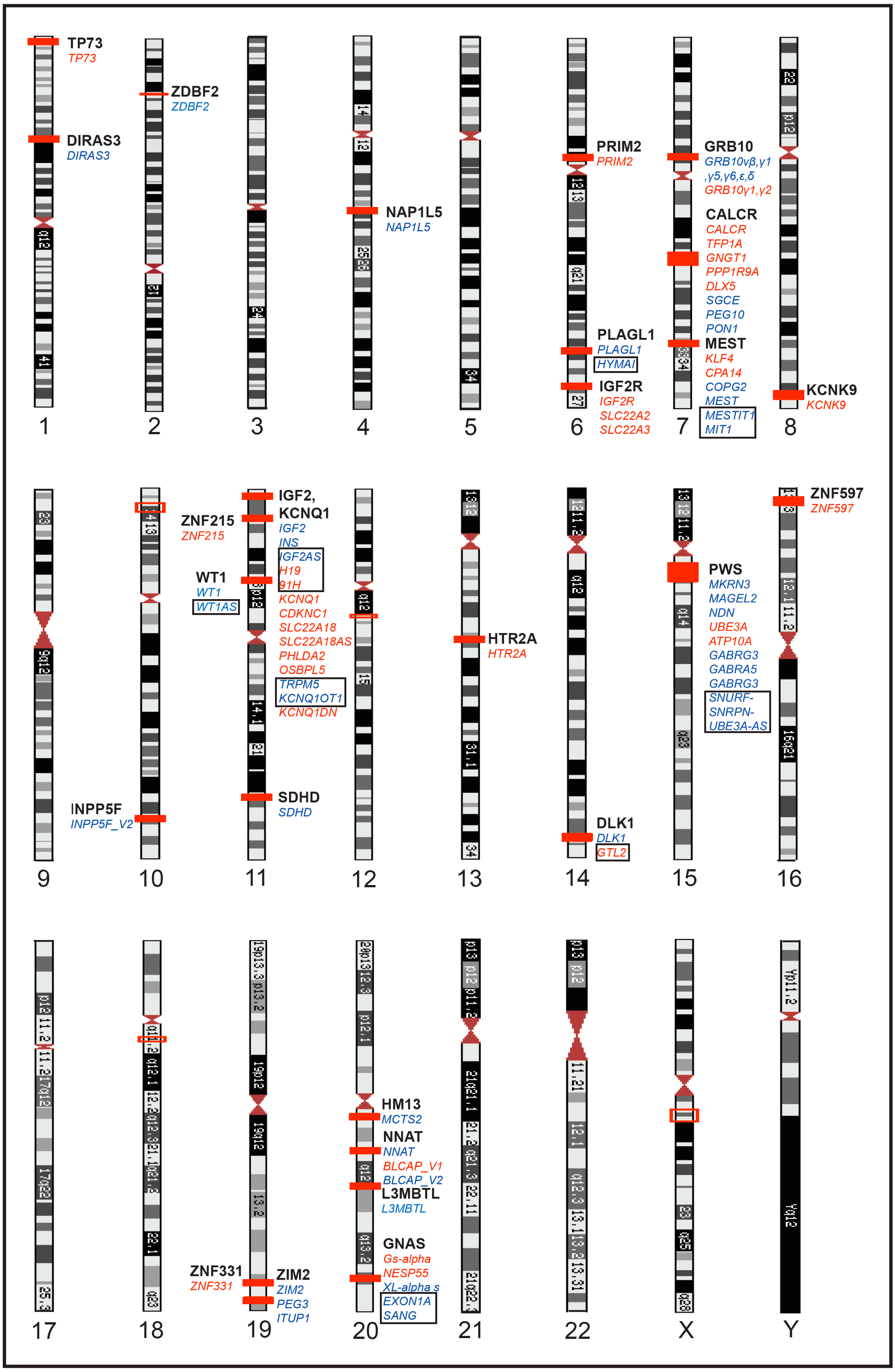


Figure 5. 8 out of 27 human imprinted gene regions express a macro ncRNA. Human imprinted gene regions are located on 14 out of 22 autosomes. Red filled boxes; locations of imprinted gene regions, Bold black letters; human imprinted gene regions named according to the gene with "central" position in the cluster, Blue; paternally methylated, Red; maternally methylated, Black non-filled boxes; Known macro ncRNAs.

1. 2. 5. 1. Six well-studied imprinted gene regions in human and mouse

Six imprinted gene regions (e.g. mouse Igf2r, Igf2, Kcnq1, Dlk1, Gnas and Pws/As) have been well-studied in mouse and human and are used as valuable models for macro ncRNA research since each of them contain at least one macro ncRNA (Koerner et al., 2009). Table 3 shows a comparison between known gametic methylation marks that are imprint control elements (ICEs) in six well-studied imprinted gene regions in mouse and human. Gametic DMR locations in these regions are conserved, as well as, the pattern of parental specific methylation. Gametic DMRs found to be paternally methylated typically lie upstream of ncRNAs promoters, while maternally methylated DMRs contain macro ncRNAs promoters.

Imprinted gene region (chromosome positon)		Gametic DMR (gDMR)	
Mouse	**Human**	**Mouse**	**Human**
Gnas (chr2)	GNAS (chr20)	GNAS-DMR(Williamson et al., 2004)	*EXON1A-DMR(Liu et al., 2000)
Pws/As (chr7)	PWS/AS (chr15)	PWS-IC(Yang et al., 1998)	Bipartite IC: PWS-SRO(Sutcliffe et al., 1994; Zeschnigk et al., 1997), AS-SRO(Buiting et al., 1999)
Igf2 (chr7)	IGF2 (chr11)	H19 DMD(Tremblay et al., 1997)	ICR1(Jinno et al., 1996)
Kcnq1 (chr7)	KCNQ1 (chr11)	KvDMR1(Fitzpatrick et al., 2002)	ICR2(Beatty et al., 2006)
Dlk1 (chr12)	DLK1 (chr14)	IG-DMR(Lin et al., 2003)	IG-DMR(Geuns et al., 2007)
Igf2r (chr17)	IGF2R (chr6)	DMR2(Sleutels et al., 2002)	*CGI-2 (Smrzka et al., 1995)

Table 3. Gametic DMRs and parental origin of methylation are conserved between mouse and human. Names of the imprinted gene regions and chromosome positions have been shown. Gametic DMRs (gDMRs) residing in well-studied imprinted regions, designated with the common literature name are presented. Green; paternally methylated, Orange, maternally methylated, * DMRs not tested if gametic or somatic but predicted from mouse data to be gametic

Our survey of genes showing imprinted expression in mouse and human well-studied imprinted gene regions shows global conservation of imprinted expression in these regions (Table 4). Each well-studied imprinted gene region contains at least one macro ncRNA. Some of these ncRNAs are hosts for small RNAs e.g. miRNAs and snoRNAs, which if tested show the same parental imprinted expression as their precursor macro RNAs.

Imprinted gene region (chromosome position)		Imprinted protein coding genes		Imprinted macro ncRNAs		Small ncRNA host	
Mouse	**Human**	**Mouse**	**Human**	**Mouse**	**Human**	**Mouse**	**Human**

Gnas (chr2)	GNAS (chr20)	Nesp(Peters et al., 1999)
		Gnasxl(Peters et al., 1999)
		Gnas(Williamson et al., 1996)
			XLαS(Hayward et al., 1998a)
			NESP55(Hayward et al., 1998b)
			GS-α(Hayward et al., 2001)
				Nespas(Wroe et al., 2000)
				Exon1A (Li et al., 2000)
					SANG (Hayward and Bonthron, 2000)
					EXON1A (Liu et al., 2000)
						-	hsa-mir-296, hsa-mir-298

Pws/As (chr7)	PWS/ AS (chr15)	Atp10a∗ (Kashiwagi et al., 2003; Kayashima et al., 2003a)
		Ube3a(Albrecht et al., 1997)
		Snrpn(Leff et al., 1992)
		Snurf(Gray et al., 1999)
		Ndn(MacDonald and Wevrick, 1997; Watrin et al., 1997)
		Magel2(Boccaccio et al., 1999)
		Mkrn3(Jong et al., 1999a)
		Peg12 (Chai et al., 2001; Kobayashi et al., 2002)
			ZNF127(Jong et al., 1999b)
			NDN(MacDonald and Wevrick, 1997)
			MAGEL2(Boccaccio et al., 1999)
			SNURF(Gray et al., 1999)
			SNRPN(Glenn et al., 1993)
			UBE3A(Rougeulle et al., 1998)
			ATP10A(Meguro et al., 2001)
			GABRG3∗ (Hogart et al., 2007; Meguro et al., 1997)
			GABRG5∗ (Hogart et al., 2007; Meguro et al., 1997)
				Zfp127as (Jong et al., 1999b)
				AK014392 (Nikaido et al., 2003)
				[Lncat (Landers et al., 2004; Le Meur et al., 2005)
				Ipw(Wevrick and Francke, 1997)
				Ube3a-as(Chamberlain and Brannan, 2001)
				Pec2(Buettner et al., 2005)
				Pec3 (Buettner et al., 2005)]***
					ZNF127AS (Jong et al., 1999b)
					[UBE3A-AS(Rougeulle et al., 1998)
					PWCR1 (de los Santos et al., 2000)
					IPW(Wevrick et al., 1994)
					PAR1(Sutcliffe et al., 1994)
					PAR5(Sutcliffe et al., 1994)
					PAR-SN(Ning et al., 1996))]***
						Snord6 (MBII-13)(Cavaille et al., 2000)
						Snord115(MBII-52)(Cavaille et al., 2000)
						Snord116 (MBII-85)
							HBII-85 (Cavaille et al., 2000)
							HBII-52 (Cavaille et al., 2000)

Igf2 (chr7)	IGF2 (chr11)	Igf2(DeChiara et al., 1991)
		Ins(Deltour et al., 1995; Giddings et al., 1994)
			IGF2(Giannoukakis et al., 1993)
			INS(Moore et al., 2001)
				H19(Bartolomei et al., 1991)
				91H(Berteaux et al., 2008)
				Igf2as(Moore et al., 1997)
					H19(Zhang and Tycko, 1992)
					91H(Berteaux et al., 2008)
					IGF2AS (Okutsu et al., 2000)
						-	miR-675 (Cai and Cullen, 2007)

Kcnq1 (chr7)	KCNQ1 (chr11)	Th(Schulz et al., 2006)
		Ascl2(Guillemot et al., 1995)
		Tspan32(Umlauf et al., 2004)
		Cd81(Umlauf et al., 2004)
		Tssc4(Paulsen et al., 2000)
		Kcnq1(Gould and Pfeifer, 1998; Paulsen et al., 1998)
		Cdkn1c(Hatada and Mukai, 1995)
		Msuit1(Onyango et al., 2000)
		Slc22a18(Dao et al., 1998)
		Phlda2(Qian et al., 1997)
		Nap1l4(Engemann et al., 2000)
		Tnfrsf23(Clark et al., 2002)
		Osbpl15(Engemann et al., 2000)
		Dhcr7(Schulz et al., 2006)
			KCNQ1(Lee et al., 1997)
			KCNQ1DN (Xin et al., 2000)
			CDKN1C(Matsuoka et al., 1996)
			SLC22A18AS (Bajaj et al., 2004)
			SLC22A18 (Dao et al., 1998)
			PHLDA2(Qian et al., 1997)
			OSBPL5(Higashimoto et al., 2002)
			TRPM5(Prawitt et al., 2000)
				Kcnq1ot1 (Lee et al., 1999; Smilinich et al., 1999)	KCNQ1OT1 (Smilinich et al., 1999)	-	-

Dlk1 (chr12)	DLK1 (chr14)	Begain(Tierling et al., 2009)
		Dlk1(Schmidt et al., 2000)
		Mico1(Labialle et al., 2008)
		Mico1os(Labialle et al., 2008)
		Rtl1(Seitz et al., 2003)
		Dio3(Tsai et al., 2002)
			DLK1(Wylie et al., 2000)	[Gtl2(Miyoshi et al., 2000)
				Rtl1as(Seitz et al., 2003)
				Rian(Hatada et al., 2001)
				AK050713 (Hagan et al., 2009)
				AK053394 (Hagan et al., 2009)]***
					GTL2(Miyoshi et al., 2000)	C/D snoRNAs(Cavaille et al., 2002)
						Mirg (Seitz et al., 2004)
						miRNAs(Seitz et al., 2004; Seitz et al., 2003)
							has mir-154(Williams et al., 2007)
							hsa-mir-335
							SNORD 113,
							SNORD 114

Igf2r (chr17)	IGF2R (chr6)	Igf2r(Barlow et al., 1991)
		Slc22a2(Zwart et al., 2001)
		Slc22a3(Zwart et al., 2001)
			IGF2R∗∗(Xu et al., 1993)
			SLC22A2(Monk et al., 2006)
			SLC22A3(Monk et al., 2006)
				Airn (Lyle et al., 2000; Wutz et al., 1997)	AIRN(Yotova et al., 2008)	-	-

Table 4. Each well-studied imprinted gene region expresses at least one macro ncRNA in mouse and human. Imprinted expression of both protein coding genes and ncRNAs is globally conserved. Blue; paternally expressed genes, Red; maternally expressed genes, Black; genes showing imprinted expression in one of the species but not tested in another, ∗; Conflicting data, ∗∗; Polymorphic imprinting; [ ]***; Suggested to represent one long macro ncRNA.

Although most imprinted genes show conserved imprinted expression, the *Igf2r* imprinted gene region is an example that shows differences between human and mouse. In mouse, ubiquitous imprinted expression has been found for both *Igf2r* and *Airn* ncRNA in fetal, extra-embryonic and adult tissues (with the exception of neurons and ES cells), while the *Slc22a2* and *Slc22a3* genes have placental specific imprinted expression (Monk et al., 2006; Yamasaki et al., 2005). In human, the *IGF2R*, *SLC22A2* and *SLC22a3* genes are largely biallelically expressed, but show polymorphic imprinted expression in placenta, early fetal tissue, lymphoblastoid cells, cultured amniotic cells and Wilms' tumors (Monk et al., 2006; Oudejans et al., 2001; Smrzka et al., 1995; Xu et al., 1993). Interestingly, the human *IGF2R* intron 2 CpG island (CGI-2) is maternally methylated, in the same position as the mouse gDMR (gDMR2, Table 3), in all tested fetal and adult tissues (Smrzka et al., 1995). Human *AIRN* ncRNA is expressed from the CGI-2 promoter in Wilms' tumor cell line and 16-40% of Wilms' tumor patients, but has not been tested for imprinted expression (Yotova et al., 2008).

The genomic organization of six well-studied human imprinted gene regions is shown in Figure 6. Differentially methylated regions (DMRs) are present in each region. 11 DMRs including 7 maternally methylated and 4 paternally methylated are shown. 5 are gametic DMRs (including 2 that are the parts of the bipartite IC in the PWS imprinted gene region), 2 are somatic, while for 4 it is still not tested if are gametic or somatic. Each of 6 well-studied imprinted gene regions has at least one macro ncRNA. 5 of these macro ncRNA are paternally expressed, 2 are maternally expressed while for *AIRN* imprinted status is not known.

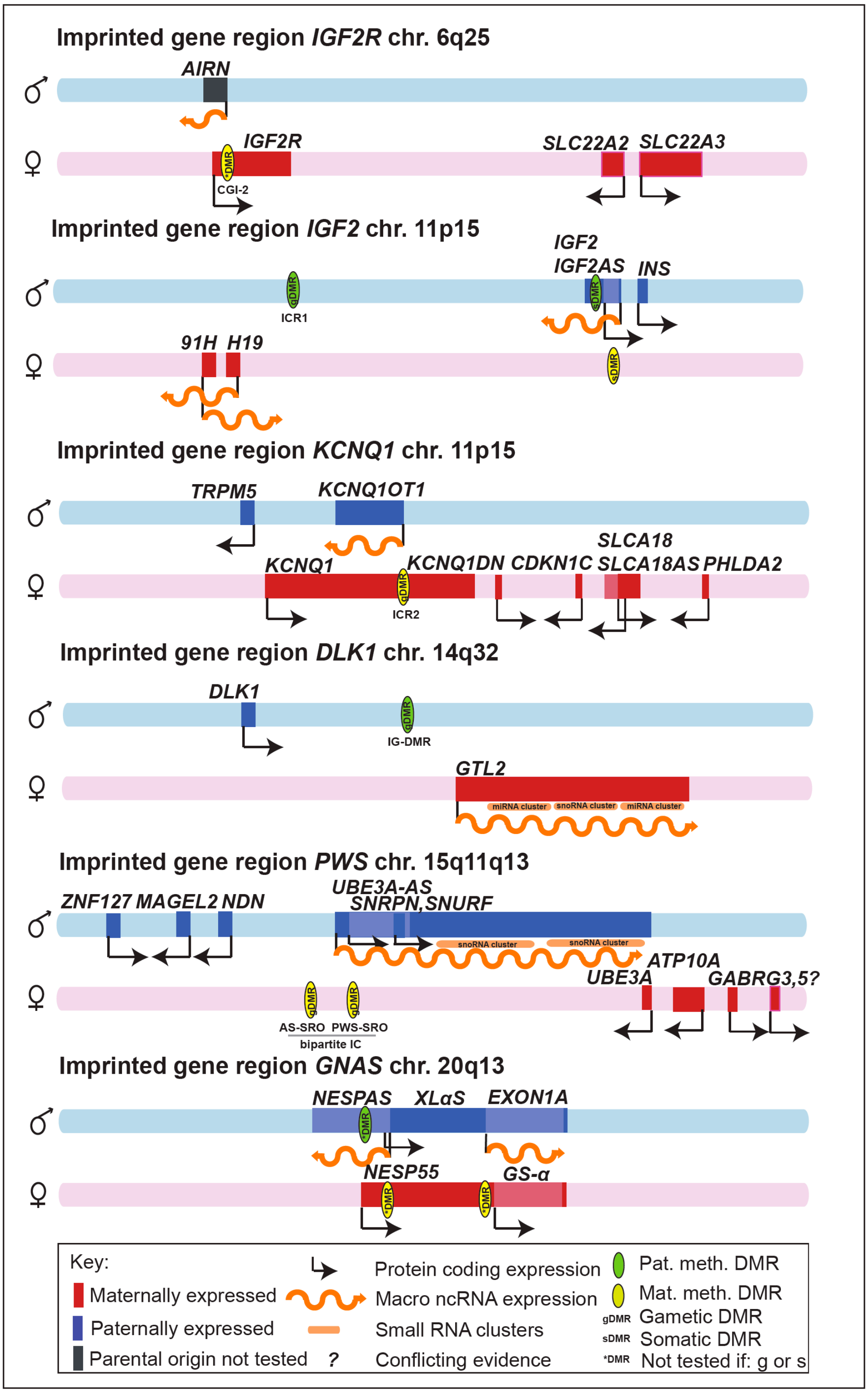


Figure 6. Six well-studied human imprinted gene clusters. Each cluster expresses a macro ncRNA. Differentially methylated regions (DMRs) are present in each cluster. Red; transcription from mothers' chromosome, Blue; transcription from fathers' chromosome, Grey; parental origin not tested, Black arrow; transcription direction of protein coding gene, Wavy orange line; macro ncRNA expression, Pale orange box; small ncRNA (snoRNAs, miRNAs) clusters, Yelow circle; maternally specific methylation, Green circle; paternally specific methylation, gDMR; gametic DMR, sDMR; somatic DMR, *DMR; not tested if DMR is gametic or somatic

1. 2. 5. 2. Less-studied imprinted gene regions in human and mouse

A survey of the literature for gametic DMRs (Table 5) and imprinted gene expression (Table 6A, Table 6B) in less-studied mouse and human imprinted gene regions shows that further study of these regions is necessary in order to understand the mechanisms of genomic imprinting in these regions. While most of less-studied imprinted regions are conserved between mouse and human, some are not or still lack examination to the same extent in both species. In mouse 24 less-studied imprinted gene regions, 9 maternally methylated and 1 paternally methylated DMRs have been previously found and shown to be gametic, while in 21 human regions 8 maternally methylated DMRs have been found (1/8 located in GRB10 region has been shown as gametic while 7 were not tested).

Imprinted gene region (chromosome position)		Gametic DMR (gDMR)	
Mouse	**Human**	**Mouse**	**Human**
Zdbf2 (chr1)	ZDBF2 (chr2)	DMR 10kb upstream Zdbf2(Kobayashi et al., 2009)	-
Sfmbt2 (chr2)	-	-	-
Gatm (chr2)	-	-	-
HM13 (chr2)	HM13 (chr20)	Mcts2 DMR(Wood et al., 2007b; Wood et al., 2008)	-
Nnat (chr2)	NNAT (chr20)	Nnat DMR(Kikyo et al., 1997)	Nnat DMR not tested for gametic(Evans et al., 2001)
Mkrn1-ps1 (chr5)	-	-	-
Calcr-Dlx5 (chr6)	CALCR-DLX5 (chr7)	Exon1 Peg10/SGCE DMR(Ono et al., 2003)	Exon1 SGCE DMR not tested for gametic(Grabowski et al., 2003)
Mest (chr6)	MEST (chr7)	Mest promoter DMR(Lucifero et al., 2002)	5' of MEST DMR not tested for gametic(Riesewijk et al., 1997)
Nap1l5 (chr6)	NAP1L5 (chr4)	Nap1l5 promoter DMR(Smith et al., 2003; Wood et al., 2007b)	-
Zim2 (chr7)	ZIM2 (chr19)	-	-
Ampd3 (chr7)	-	-	-
Inpp5f (chr7)	INPP5F (chr10)	CpG 5'Inpp5f_v2 DMR(Wood et al., 2007b)	-
Rasgrf1 (chr9)	-	30kb 5'Rasgrf1(Pearsall et al., 1999)	-
Plagl1 (chr10)	PLAGL1 (chr6)	Hymai DMR(Arima and Wake, 2006)	HYMAI DMR not tested for gametic(Arima et al., 2001)
Dcn (chr10)	-	-	-
Grb10 (chr11)	GRB10 (chr7)	Grb10 DMR(Arnaud et al., 2003)	CGI2 DMR(Arnaud et al., 2003)
Commd1 (chr11)	-	-	-
Pde4d (chr13)	-	-	-
Htr2a (chr14)	HTR2A (chr13)	-	-
Kcnk9 (chr15)	KCNK9 (chr8)	Peg13 DMR(Ruf et al., 2007)	-
Slc38a4 (chr15)	-	-	-
Impact (chr18)	-	-	-
Tbc1d12 (chr19)	-	-	-
Xist (chrX)	-	-	-
-	TP73 (chr1)	-	-
-	DIRAS3 (chr1)	-	DMRs CpGI,II,III in DIRAS3 gene not tested for gametic(Yu et al., 1999; Yuan et al., 2003)
-	PRIM2 (chr6)	-	-
-	ZNF215 (chr11)	-	-
-	WT1 (chr11)	-	WT1 ARR DMR not tested for gametic(Dallosso et al., 2004)
-	SDHD (chr11)	-	-
-	ZNF597 (chr16)	-	-
-	ZNF331 (chr19)	-	-
-	L3MBTL (chr20)	-	CPG 3, 4 promoter 2 L3MBTL DMR not tested for gametic(Li et al., 2004)

Table 5. Methylation status of gametic DMRs in 24 mouse and 21 human less-studied imprinted gene regions is mostly conserved. Imprinted gene regions are named according to the gene located on the "central" position in the region. Parental methylation status is coserved between mouse and human but human typically lacks the test if observed methylation is set in gametes due to problems in obtaining human oocytes. 10 gametic DMRs

are listed for mouse imprinted gene regions (9 maternally methylated and 1 paternally methylated) while one gametic maternally methylated DMR is confirmed in human GRB10 region and 7 more are found to be maternally expressed. Green; paternally methylated, Orange; maternally methylated, -; not tested.

Imprinted gene region (chromosome position)		Imprinted protein coding genes		Imprinted macro ncRNAs		Imprinted expression of small ncRNAs	
Mouse	**Human**	**Mouse**	**Human**	**Mouse**	**Human**	**Mouse**	**Human**

Zdbf2 (chr1)	ZDBF2 (chr2)	Gpr1(Mishima et al., 1990)
		Zdbf2(Kobayashi et al., 2009)
			ZDBF2(Kobayashi et al., 2009)	-	-	-	-

Sfmbt2 (chr2)	-	Sfmbt2(Kuzmin et al., 2008)	-	-	-	-	-
Gatm (chr2)	-	Gatm(Sandell et al., 2003)	No imprinted expression(Monk et al., 2006)	-	-	-	-
HM13 (chr2)	HM13 (chr20)	HM13a, b, c(Wood et al., 2007b; Wood et al., 2008) HM13d, e(Wood et al., 2008) Mcts2(Wood et al., 2007b)	MCTS2(Wood et al., 2007b)	-	-	-	-

Nnat (chr2)	NNAT (chr20)	Nnat(Kagitani et al., 1997)
		Blcap_v1a(Schulz et al., 2009)
		Blcap_v2a(Schulz et al., 2009)
			NNAT(Evans et al., 2001)
			BLCAP_V1a,b,c (Schulz et al., 2009)
			BLCAP_V2a(Schulz et al., 2009)
				-	-	-	-

Mkrn1-ps1 (chr5)	-	Mkrn1-ps1∗ (Gray et al., 2006; Hirotsune et al., 2003)	-	-	-	-	-

Calcr-Dlx5 (chr6)	CALCR-DLX5 (chr7)	Calcr(Hoshiya et al., 2003)
		Tfpi2(Monk et al., 2008)
		Casd1(Babak et al., 2008)
		Sgce(Monk et al., 2008)
		Peg10(Ono et al., 2001)
		Neurabin(Monk et al., 2008; Ono et al., 2001)
		Pon3(Ono et al., 2001)
		Pon2(Ono et al., 2001)
		Asb4(Mizuno et al., 2002)
		Dlx5∗(Horike et al., 2005; Schule et al., 2007)
			PEG10(Ono et al., 2001)
			GNGT1(Okita et al., 2003)
			CALCR(Okita et al., 2003)
			SGCE(Grabowski et al., 2003)
			PPP1R9A(Nakabayashi et al., 2004)
			PON1(Okita et al., 2003)
			DLX5∗(Okita et al., 2003; Schule et al., 2007)
			APS(Okita et al., 2003)
				-	-	-	-

Mest (chr6)	MEST (chr7)	Mest(Kaneko-Ishino et al., 1995)
		Copg2(Lee et al., 2000)
		Klf14(Parker-Katiraee et al., 2007)
			MEST(Kobayashi et al., 1997)
			COPG2*(Blagitko et al., 1999; Yamasaki et al., 2000)
			CPA4(Kayashima et al., 2003b)
			KLF14(Parker-Katiraee et al., 2007)
				Copg2as1 (Lee et al., 2000)
				Copg2as2 (Lee et al., 2000)
					MESTIT1(Li et al., 2002)
					MIT1 (Yamasaki et al., 2000)
						Mirn-335(Royo and Cavaille, 2008)	-

Nap1l5 (chr6)	NAP1L5 (chr4)	Nap1l5(Smith et al., 2003; Wood et al., 2007b)	NAP1L5(Wood et al., 2007b)	-	-	-	-

Zim2 (chr7)	ZIM2 (chr19)	Zim2(Kim et al., 2004)
		Zim1(Kim et al., 1999)
		Apeg3(Choo et al., 2008)
		Peg3(Kaneko-Ishino et al., 1995)
		Usp29(Kim et al., 2000)
		Zim3(Kim et al., 2001)
		Zfp264(Kim et al., 2001)
			ZIM2(Kim et al., 2004)
			ITUP1(Maegawa et al., 2004)
			PEG3(Murphy et al., 2001)
				-	-	-	-

Ampd3 (chr7)	-	Ampd3(Schulz et al., 2006)	-	-	-	-	-

Inpp5f (chr7)	INPP5F (chr10)	Inpp5f_v2(Choi et al., 2005)
		Inpp5f_v3(Wood et al., 2007a)
			INPP5F_V2(Wood et al., 2007b)	-	-	-	-

Rasgrf1 (chr9)	-	Rasgrf1(Plass et al., 1996)
		As4(Nomura et al., 2008)
			-	A19(de la Puente et al., 2002)	-	Mir184 (Nomura et al., 2008)	-

Plagl1 (chr10)	PLAGL1 (chr6)	Plagl1(Piras et al., 2000; Smith et al., 2002)	PLAGL1(Kamiya et al., 2000)	Hymai(Arima and Wake, 2006)	HYMAI (Inoue et al., 2001)	-	-
Dcn (chr10)	-	Dcn(Mizuno et al., 2002)	No imprinted expression(Monk et al., 2006)	-	-	-	-

Grb10 (chr11)	GRB10 (chr7)	Ddc_exon1a(Menheniott et al., 2008)
		Grb10as(Babak et al., 2008)
		Grb10α,δ(Miyoshi et al., 1998)
		Grb10$\beta1,\beta2$ (Arnaud et al., 2003; Hikichi et al., 2003)
		Cobl(Shiura et al., 2009)
			GRB10$\beta,\gamma1,\gamma5,\gamma6,\varepsilon,\delta$(Blagitko et al., 2000)
			GRB10$\gamma1,\gamma2$ (Blagitko et al., 2000; McCann et al., 2001)
				-	-	-	-

Commd1 (chr11)	-	U2af1-rs1(Hatada et al., 1993; Zhang et al., 2006)
		Commd1(Zhang et al., 2006)
			-	-	-	-	-

Pde4d (chr13)	-	Pde4d(Babak et al., 2008)	-	-	-	-	-
Htr2a (chr14)	HTR2A (chr13)	Htr2a(Kato et al., 1998)	HTR2A∗(Kato et al., 1996; Pastinen et al., 2004)	-	-	-	-

Kcnk9 (chr15)	KCNK9 (chr8)	Kcnk9(Ruf et al., 2007)
		Trappc9(Wang et al., 2008b)
			KCNK9(Ruf et al., 2007)	Peg13(Smith et al., 2003)	-	-	-

Slc38a4 (chr15)	-	Slc38a4(Mizuno et al., 2002; Smith et al., 2003)	-	-	-	-	-

Table 6A. Imprinted expression of protein-coding genes and macro ncRNAs is mostly conserved between human and mouse in less-studied imprinted gene regions. 9 mouse imprinted gene regions express imprinted genes while homologous human regions are not tested or if in two cases tested did not express imprinted genes. Blue; paternally expressed genes, Red; maternally expressed genes, -; not tested, ∗; conflicting data.

Imprinted gene region (chromosome position)		Imprinted protein coding genes		Imprinted macro ncRNAs		Imprinted expression of small ncRNAs	
Mouse	**Human**	**Mouse**	**Human**	**Mouse**	**Human**	**Mouse**	**Human**
Impact (chr18)	-	*Impact(Hagiwara et al., 1997)*	-	-	-	-	-
Tbc1d12 (chr19)	-	*Tbc1d12(Babak et al., 2008)* *Ins1∗(Deltour et al., 1995; Giddings et al., 1994)*	-	-	-	-	-

Xist (chrX)	-	*Fthl17(Kobayashi et al., 2010)*
		Xlr3b(Raefski and O'Neill, 2005)
		Xlr4b(Raefski and O'Neill, 2005)
		Xlr4c(Raefski and O'Neill, 2005)
		Rhox5(Kobayashi et al., 2006)
			-	*Xist(Kay et al., 1994)*
				Tsix(Lee, 2000)
					No imprinted expression(Migeon et al., 2002)	-	-

-	TP73 (chr1)	-	*TP73(Kaghad et al., 1997)*	-	-	-	-
-	DIRAS3 (chr1)	-	*DIRAS3(Yu et al., 1999)*	-	-	-	-
-	PRIM2 (chr6)	-	*PRIM2(Pant et al., 2006)*		-	-	-
-	ZNF215 (chr11)	-	*ZNF215(Alders et al., 2000)*	-	-	-	-
-	WT1 (chr11)	-	*WT1∗(Mitsuya et al., 1997)* *AWT1(Dallosso et al., 2004)* *WT1∗(Jinno et al., 1994)*	*No imprinted expression(Dallosso et al., 2007)*	*WT1-AS(Dallosso et al., 2004)*	-	-
-	SDHD (chr11)	-	*SDHD(Badenhop et al., 2001)*	-	-	-	-
-	ZNF597 (chr16)	-	*ZNF597(Pant et al., 2006)*	-	-	-	-
-	ZNF331 (chr19)	-	*ZNF331(Pant et al., 2006)*	-	-	-	-
-	L3MBTL (chr20)	*No imprinted expression(Li et al., 2005)*	*L3MBTL(Li et al., 2004)*	-	-	-	-

Table 6B. Genes from imprinted gene regions (3 mouse and 9 human) show imprinted expression in mouse or human. 7 more mouse regions that also show imprinted expression in mouse, but not in human (mostly not tested in human) were already shown in Table 6A. For two paternally expressed human genes (*L3MBTL* and *WT1-AS*) no imprinted expression has been found in mouse, while imprinted mouse *Xist* and *Tsix* ncRNAs do not show imprinted expression in human. Blue; paternally expressed genes, Red; maternally expressed genes, Black; genes showing imprinted expression in one of the species, but not tested in another, ∗; conflicting data, -; not tested.

A census of mammalian imprinting has been published in 2005 by Morison et al. (Morison et al., 2005) when 112 imprinted "functional components" were found (53 human and 96 in mice with 37 overlapping). They found a number of discordances between human and mouse imprinting data. Interestingly a recent review by Frost and Moore (Frost and Moore, 2010) examined conservation of mouse genes imprinted in placenta and found that up to date most of the genes showing imprinted expression in mouse, but not in human are exactly those that are imprinted just in the mouse placenta.

By literature search I found 126 mouse and 88 human imprinted genes or gene variants (Table 4, Table 6A, B). I also observed certain number of discordances, but similarly to the conclusion from Morison et al. (Morison et al., 2005), most of them are due to lack of data in one of the species or a missing the orthologous gene. *ZIM2* is an example of the gene that shows different parental expression in human and mouse. Interestingly, the basis for this difference has been found: the insertion of *Zim1* between *Peg3* and *Zim2* in mouse lead to the different imprinted expression of these genes between the species (Kim et al., 2004).

By looking into parental expression of human imprinted genes I observed that a large majority of genes that are found to be imprinted in both mouse and human show expression from the same parental allele in both species. Thus, human and mouse imprinted genes are mostly conserved considering parental expression status, but still with some exceptions and the necessity for further investigation in number of cases where expression has not been tested in one of the species. Interestingly, a number of human and mouse imprinted gene regions express imprinted macro ncRNAs. Functions of imprinted protein coding genes and imprinted macro ncRNAs will be further described.

1. 2. 6. Imprinted protein coding genes have diverse functions

The evidence for the functions of imprinted genes is gained from knockouts or rare random mutations in mice, while natural deletions and mutations are the basis for finding the function of imprinted genes in human. Imprinted protein coding genes have diverse functions but they can be grouped into genes that affect growth, those with no obvious role in development and those affecting behavior through their role in the nervous system (mouse imprinted genes and their functions are listed on the Harwell web site http://www.har.mrc.ac.uk/research/genomic_imprinting/function.html). Numerous imprinted genes have a function in development by acting as a growth regulators among which those paternally expressed promote growth and those maternally expressed act as growth repressors (e.g. *Igf2* and *Igf2r* respectively) (Barlow, 2007). These genes are in agreement with the “parental conflict” theory of evolution of genomic imprinting (described in 1.2.3.). The group of by-stander genes (genes with no obvious role in development) remain to be understood since they cannot be easily explained by any of the theories about evolution of imprinting (introduced in 1.2.3.). A number of imprinted genes are expressed in brain and associated with cognitive, behavioral and neurological disorders (lists of these genes are available at Cardiff

University web site http://www.bgg.cardiff.ac.uk/imprinted_tables/index.html). Evolution of genomic imprinting, for some genes that affect maternal care (nourishment and protection of a newborn) also could be explained by “parental conflict” theory. An example of an imprinted gene associated with a neurological disorder is *Ub3a*, the maternally expressed gene involved in Angelman syndrome recently found to be required for experience-dependent synapse plasticity in the mouse visual cortex (Sato and Stryker, 2010).

1. 2. 7. Atypical biology of imprinted macro ncRNAs?

Imprinted macro ncRNAs show parent-of-origin specific expression, are unusually large (defined as more than 200bp in length, but typically several hundred thousand nucleotides long) and do not have a continuous open reading frame (ORF). In a few cases it has been showed that macro ncRNAs have a methyl-7-guanosine cap (7mGcap) and that are RNA Polymerase (RNAP) II transcripts with polyA-tail similar to RNAP II mRNA transcripts. Atypical features of macro ncRNAs e.g. reduced splicing potential, relative unstability and nuclear retention are overviewed for mouse macro ncRNAs in a Table 7. In human imprinted macro RNA research, the biology of the majority of ncRNA transcripts remains to be tested.

Mouse imprinted macro ncRNA	Length (kb)	CpG Island promoter (UCSC browser)	Promoter parental specific methylation (Table 3)	Direct repeats	RNAP II	7mGcap	Splicing	Stability (half-life)	Cellular localization
Nespas	Unspliced >3.35 Spliced 1.4-15.8 Genomic size: >30	Yes	Yes	-	-	-	S and U	-	-
Exon1A	Unspliced >1.1kb Spliced >1.4 Genomic size: 19	Yes	Yes	-	-	-	S and U	-	-
Lncat	~1000	No	Yes	Yes (Paoloni-Giacobino et al., 2007)	-	-	S	-	C (Le Meur et al., 2005)
H19	2.2, Genomic size: 2.5	No	-	-	Yes (Brannan et al., 1990)	Yes (Pachnis et al., 1988)	S (low I/E)	-	N, C (Brannan et al., 1990)
Igf2as	4.8, Genomic size: 10.7	Yes	-	-	-	-	S	-	-
Kcnq1ot1	~90	Yes	Yes	Yes (Paoloni-Giacobino et al., 2007)	Yes (Redrup et al., 2009)	-	U	Moderately stable (3.4h)(Redrup et al., 2009)	N (Redrup et al., 2009)
Gtl2	1.9-7 Genomic size: 30.7	No	-	Yes (Dindot et al., 2009)	-	-	U	-	-
Airn	108	Yes	Yes	Yes (Neumann et al., 1995)	Yes (Seidl et al., 2006)	Yes (Seidl et al., 2006)	U 95% Low S 5%	Unstable (1.6-2.1h)(Seidl et al., 2006)	N (Seidl et al., 2006)

Table 7. Macro ncRNAs may have atypical biology. Mouse macro ncRNAs from six well-studied clusters are overwieved. -; Not tested, S; spliced, U; unspliced, I/E; intron/exon ratio,

C; cytoplasmic localization, N; nuclear localization. Length and splicing of macro ncRNAs were reviewed in (Koerner et al., 2009).

The largest gene in the human genome, *DYSTROPHIN* has a genomic size of 2.22Mb, but only 14.069kb is transcribed as a mRNA. The reason for this difference is the very high intronic content of most messenger RNAs (mRNAs), leading to the high intron/exon ratio of these transcripts. In comparison, macro ncRNAs are mostly unspliced or with a low intronic content that leads to the unusual length of their mature transcripts. For example, mouse *Airn* ncRNA is 108kb long while *Lncat* could be the longest mature transcript known to date at around 1000kb in length. Since their mature transcripts are covering large genomic regions macro ncRNAs are rich in transposons that are normally depleted from mature mRNAs (Latos and Barlow, 2009). Unspliced forms of mouse *Airn* and *Xist* imprinted macro ncRNAs are found to be relatively unstable. Seidl et al. (Seidl et al., 2006) for example used Actinomycin D in order to determine the stability of mouse *Airn* ncRNA and *Igf2r* mRNA and found that unspliced *Airn* had a 2.1h half-life while spliced *Airn* had a 15-17h half-life and *Igf2r* mRNA 14.3h half-life. Furthermore, some of these ncRNA transcripts (especially those found to be unspliced) are nuclear localized since they escape nuclear export by still unclear mechanism (Redrup et al., 2009; Seidl et al., 2006). Human *KCNQ1OT1* is similarly to mouse *Kcnq1ot1* found to be unspliced and nuclear localized transcript (Murakami et al., 2007).

Interestingly, mouse *Airn* and *Kcnq1ot1* imprinted macro ncRNAs that are long; transposon rich, unspliced and nuclear localized transcripts have a function in gene regulation. The roles of macro ncRNAs will be the focus of the next section.

1. 2. 8. Imprinted macro ncRNAs are transcriptional regulators

Studies of imprinted macro ncRNAs function have been performed on mouse embryonal stem (mES) cells and knockout mice models. Three types of experiments involving deletions of ICEs, replacements of ncRNAs promoters and deletions or truncations of ncRNA genes have been done for a restricted number of imprinted macro ncRNAs. Those experiments showed clearly *in cis* silencing of protein coding genes by two tested imprinted macro ncRNAs (*Airn* and *Kcnq1ot1*) and in one case (*H19*), *in trans* regulation of at least 16 co-expressed imprinted genes belonging to a recently defined “imprinted gene network” (Gabory et al., 2009; Mancini-Dinardo et al., 2006; Sleutels et al., 2002).

Lost of macro ncRNA expression and de-repression of imprinted protein coding genes in their clusters are found after deletions of the unmethylated ICE containing the promoters of *Airn*, *Kcnq1ot1* and *Nespas* ncRNAs (Fitzpatrick et al., 2002; Shin et al., 2008; Williamson et al., 2006; Wutz et al., 1997). For example, a 3.7kb deletion of the unmethylated ICE expressing *Airn* ncRNA led to loss of imprinted expression of *Igf2*, *Slc22a2* and *Slc22a3* genes (Wutz et al., 2001; Zwart et al., 2001). These experiments show the ICE is important, but they do not distinguish between the role of the ICE as a ncRNA promoter (e.g. KvDMR1 for *Kcnq1ot1* ncRNA) or regulator (e.g. IG-DMR for *GTL2* ncRNA) and other possible functions of the ICE.

The second group of experiments involving promoter replacements showed that high ncRNA expression is necessary for silencing. For example, when the *Airn* promoter was replaced with a strong PGK promoter, these cells gained *Igf2r* promoter DNA methylation in differentiated ES cells and silenced *Igf2r in cis*, while cells where the *Airn* promoter was replaced with a weak TET promoter resulting in low *Airn* expression failed to establish *Igf2r* methylation and silence *Igf2r* (Stricker et al., 2008). These experiments showed that the *Airn* promoter itself has no role in silencing, while the expression level of *Airn* macro ncRNA is a key factor in silencing.

Truncations are the third group of experiments that further examined the function of ncRNAs in silencing. *Airn* ncRNA has been truncated from 108kb to 3kb length by inserting 1.2kb long polyadenylation cassette without changing the ICE. 3kb *Airn* remained paternally expressed and the promoter maternally methylated, while loss of imprinted expression was observed again for all three genes in the *Igf2r* cluster (Sleutels et al., 2002). Similarly, *Kcnq1ot1* ncRNA truncated on the paternal chromosome from about 90kb to 1.5kb led to a derepression of the seven genes (located over 775kb of DNA sequence) from the *Kcnq1* region in ES cells (Mancini-Dinardo et al., 2006). These kinds of experiments showed that the initiation of ncRNAs transcription is not critical for silencing, while transcriptional elongation or the transcript itself are the critical factors for silencing *in cis*.

Interestingly, deletion of *H19* ncRNA gene in the *Igf2* locus showed that this macro ncRNA does not have a role in imprinted gene regulation *in cis* in liver, while it showed small effect in skeletal muscle (Schmidt et al., 1999). Instead this locus is an example of the insulator model of *cis*-acting silencing in imprinted clusters. Deletion of the paternally methylated DMR 2kb upstream from *H19* led to loss of imprinted expression of both *H19* and *Igf2* (Thorvaldsen et al., 1998). This experiment showed that this gametic DMR is the ICE, which has been further shown to have insulator function. The unmethylated ICE located on the maternal chromosome binds CTCF and enhancers located downstream of *H19* activate expression of this ncRNA, while their interaction with *Igf2* and *Ins* is blocked. In contrast, methylation of ICE on the paternal chromosome does not allow binding of CTCF allowing the enhancers activate *Igf2* and *Ins* expression (Bell and Felsenfeld, 2000; Hark et al., 2000).

Models of *cis*-mediated ncRNA imprinted silencing are the *Igf2r* and *Kcnq1* clusters where truncations of macro ncRNAs led to loss of imprinted expression of protein coding genes in the cluster. The exact mechanism how *cis*-mediated ncRNA silencing takes place is still under debate and number of hypothesis that are less or more supported by current evidence are present in the literature (reviewed in (Pauler et al., 2007)). Two main types of hypotheses are that the ncRNA product or ncRNA transcription *per se,* leads to *cis*-imprinted silencing. Recent evidence provide support for the ncRNA itself recruiting Polycomb group (PcG) and G9a (functions as H3K9me2 histone methyltransferase) proteins and targeting them to silence imprinted genes in placenta. For example, interaction of G9a with *Kcnq1ot1* and *Airn* macro ncRNAs has been found and in the case of *Airn* it has been shown that this macro ncRNA targets G9a to chromatin of *Slc22a3* promoter in placenta suggesting this may epigenetically silence transcription of the genes that show imprinted expression only in placenta (genes like *Igf2r* that shows ubiquitous imprinted expression were not affected) (Nagano et al., 2008; Pandey et al., 2008). Further careful examinations of all epigenetic players in imprinting gene regions will be necessary for further understanding of the mechanism how macro ncRNAs silence imprinted genes.

Insulator model and RNA-mediated silencing model showed that macro ncRNAs present in imprinted regions could have role in silencing *in cis* (macro ncRNAs like *Airn*, *Kcnq1ot1*) or do not have this role (macro ncRNAs like *H19*). Interestingly, a role for *H19* regulation of imprinted gene network *in trans* has been found in mice. After deletion of *H19* macro ncRNA in two knock out models, upregulation of five imprinted genes located on different chromosomes was shown (Gabory et al., 2009). The role of macro ncRNAs in regulation of imprinting are clear, but the exact mechanisms of ncRNA mediated epigenetic regulation remain to be fully understood. Important functions of imprinted protein coding genes and imprinted macro ncRNAs imply that if disrupted they may be involved in disease and next section will focus on this subject.

1. 2. 9. Human genomic imprinting and disease

Disruption of genomic imprinting by diverse genetic and epigenetic causes leads to human disease. Uniparental disomies (UPDs), deletions, translocations and point mutations are genetic causes affecting imprinted genes or ICEs in diverse imprinted gene clusters found to be involved in imprinting disorder syndromes. Epigenetic causes leading to disease are found to involve changes in DNA methylation. Interestingly, genetic and epigenetic causes leading to disease can be combined, such as genetic mutations in *MeCP2* (methyl CpG binding protein 2) gene, that lead to a failure to read the DNA methylation imprint and cause changes in histone modifications in Rett syndrome (reviewed in (Arnaud and Feil, 2005)).

Involvement of genomic imprinting in about ten human syndromes and cancer is well documented (e.g. Beckwith Wiedmann, Prader-Willi (PWS), Angelman, Silver-Russel, Transient Neonatal Diabetes Mellitus (TNDM), McCune-Albright syndrome (MAS), Albrights hereditary osteodystrophy (AHO), pseudohypothyroidism type 1a and 1b (PHP1a and PHP1b), reviewed in (Murrell, 2006; Robertson, 2005)). Numerous complex genetic diseases e.g. autism, diabetes, Alzheimer disease and schizophrenia are showing parent-of-origin effects by linkage studies (reviewed in (Das et al., 2009)). The potential role of imprinting in these kinds of genetic disorders remains to be elucidated. There is a controversy in the field about the association between diseases caused by perturbations in imprinting and assisted reproductive technologies (AST). A number of papers showed that there is an increased risk of different imprinting diseases in babies born after AST (reviewed in (Laprise, 2009)), however other studies, like Danish large study involving around 450,000 babies born naturally or by AST, found no risk (Lidegaard et al., 2005).

1. 2. 9. 1. Selected human disorders of genomic imprinting

Human syndromes involving genomic imprinting could be grouped into three main types: those that affect growth and development, neurological and hormonal/metabolic disorders (Arnaud and Feil, 2005). Here I describe three well-studied imprinted clusters (PWS, IGF2 and KCNQ1) where disturbances are documented to cause Prader-Willi (PWS), Angelman (AS) and Beckwith-Wiedemann (BWS) syndrom.

Prader-Willi (PWS) is a neuro-developmental disorder with an incidence of about one in 10,000-20,000 individuals (Butler, 1990) (reviewed in (Horsthemke and Wagstaff, 2008)). The clinical presentation includes infantile hypotonia, childhood obesity, small hands and feet, hypogonadism, behavior problems and mental retardation. Genetic defects like deletion of 15q11-q13 on the paternal chromosome (70% cases) and maternal UPD of chromosome 15 (29% cases) are involved in PWS. An imprinting defect involving DNA methylation change on the ICE in the PWS imprinted cluster is found in ~1% of PWS patients. Interestingly, there is an increased risk for myeloid leukemia in PWS patients (Davies et al., 2003).

Angelman (AS) syndrome is a neurological disorder that also involves the 15q11-q13 chromosome region including the PWS imprinted cluster, and occurs with the same frequency as PWS (reviewed in (Horsthemke and Wagstaff, 2008; Van Buggenhout and Fryns, 2009). This disorder is characterized by mental retardation, microcephaly, ataxia, behavioral problems including hyperactivity, a happy personality, and sleeping problems. Approximately 60-75% of AS patients show maternal deletion or re-arrangement on 15q11-q13; 2-5% show paternal UPD, while around 10% shows mutation in the *UBE3A* imprinted gene. An imprinting centre (IC) defect is found in 3-5% of AS where both alleles show unmethylated *SNRPN*, *NDN* and *MKRN3* imprinted genes promoter regions (Horsthemke and Wagstaff, 2008).

Beckwith-Wiedmann (BWS) is a growth disorder with an incidence of one in 13,700 (Weksberg et al., 2010). Major clinical findings of BWS involve e.g. macroglossia, visceromegaly, renal abnormalities and abdominal wall defects. The molecular basis of 75-80% of BWS involves changes to the 11p15.5 chromosome region including IGF2 and KCNQ1 imprinted clusters. Uniparental disomies, parent-of-origin specific duplications, translocations/inversions, microdeletions and mutations in the *CDKN1C* imprinted gene are genetic alterations found in BWS patients (Weksberg et al., 2010). Epigenetic alterations involve the loss of methylation on ICR2, that involves promoter of *KCNQ1OT1* ncRNA (Table 3, Figure 6, section 1.2.5) in 50% of BWS cases and gain of methylation on ICR1 in 2-7% of cases, while loss of ICR2 methylation happens in 95% of BWS patients born following assisted reproduction. Interestingly, BWS patients have 7.5-9% overall risk for developing embryonic tumors within 5-8 years of age (Murrell, 2006).

Imprinted non-protein coding RNAs, linked with human syndromes involving imprinting, are overviewed in Table 8. Different molecular mechanisms might disturb macro ncRNAs in disease, for example there are some cases of BWS where a microdeletion of the *KCNQ1OT1* ncRNA gene is found (Niemitz et al., 2004) while in other cases *KCNQ1OT1* shows biallelic expression (Lee et al., 1999).

Imprinted macro ncRNA	Imprinting disorder	Reference
H19	Beckwith-Wiedemann Syndrome	(Sparago et al., 2004)
KCNQ1OT1	Beckwith-Wiedemann Syndrome	(Niemitz et al., 2004)
MESTIT1	Russell-Silver syndrome	(Nakabayashi et al., 2002)
MIT1	Russell-Silver syndrome	(Yamasaki et al., 2000)
UBE3A-AS	Angelman syndrome	(Chamberlain and Brannan, 2001)
IPW	Prader-Willi syndrome	(Wevrick et al., 1994)

Table 8. **Imprinted macro ncRNAs are involved in human imprinting disorders.**

1. 2. 9. 2. Human genomic imprinting and cancer

Loss of imprinted expression (LOI), based on the changes in parent-of-origin specific DNA methylation leading to gain of biallelic expression, is common in cancer. Demethylated mouse ES cells and chimeric mice derived from these ES cells showed that LOI may lead to widespread tumorigenesis (Holm et al., 2005). In humans, complete hydatiform moles that are mostly androgenetic (two fathers' chromosome sets) and partial moles that are triploid (usually two paternal and one maternal chromosome sets) have predominant paternal imprints and show the potential to develop into malignant choriocarcinomas. Gynogenotes, on the other hand (two mothers' chromosome sets) form benign ovarian teratomas (reviewed in (Murrell, 2006; Paoloni-Giacobino, 2007)). Interestingly, some imprinting disorders predispose suffers to cancers, e.g. BWS patients are most commonly predisposed to Wilms' tumor and hepatoblastoma.

LOI of several studied genes occurs in high frequency in spontaneous tumors that commonly include the activation of a normally silent growth-promoting gene or silencing of a normally active growth-inhibitory gene. The most common gene exhibiting LOI is the *IGF2* gene coding for growth a factor, that has been found in 100% of chronic myeloid leukemia, 66% of colorectal cancer and 70% of Wilms' tumors (reviewed in (Jelinic and Shaw, 2007)). It has been shown that LOI of *IGF2* occurs in the nephrogenic embryonic cells (potentially stem cells) in Wilms' tumors and that LOI is an early event in cancer progression (Yuan et al., 2005). Interestingly, LOI also shows an increased occurrence in normal tissue surrounding the Wilms' tumor and colorectal cancers (Kaneda and Feinberg, 2005; Nakagawa et al., 2001). These studies support the epigenetic origin of cancer hypothesis according to which epigenetic disruption of progenitor cells is the first step in the cancer progression (Feinberg et al., 2006). The question if LOI has the causal role in cancer progression is still open, but it is clear that understanding how disruptions of genomic imprinting occur will have a further impact on cancer research. Except for the *IGF2* gene, LOI of the human protein coding genes has been suggested for *PLAGL1* in ovarian cancer

(Abdollahi et al., 2003; Kamikihara et al., 2005) and *MEST* in lung, colon and breast cancers (Kohda et al., 2001; Nakanishi et al., 2004; Pedersen et al., 2002; Pedersen et al., 1999).

Certain imprinted macro ncRNAs that show altered expression in cancer suggest a link between aberrant genomic imprinting and cancer. For example, *KCNQ1OT1* ncRNA shows LOI in 40% of colorectal cancer patients (Tanaka et al., 2001) while *WT1-AS* ncRNA disply biallelic expression in Wilms' tumors, but monoallelic expression in normal kidney also suggesting LOI (Malik et al., 2000). Furthermore, overexpression of the imprinted *IGF2AS* macro ncRNA is found in Wilms' tumors (Okutsu et al., 2000), while the imprinted *H19* macro ncRNA is upregulated in carcinogenesis and found to be an oncogene in hepatocellular and bladder carcinoma (Matouk et al., 2007).

1. 3. Human transcriptome research and ncRNAs

1. 3. 1. New high-throughput technologies and surprises from transcriptomics

The human transcriptome (complete set of transcripts of cell under a specific physiological condition) is much more complex than estimated ten years ago. The encyclopedia of DNA Elements (ENCODE) pilot project examined 1% of human genome by three high-throughput approaches: tilling arrays hybridizations, tag sequencing of cap-selected RNAs (CAGE) and integrated annotation of cDNAs and ESTs. ENCODE showed that less than 2% of the genome codes for proteins but more than 90% is transcribed (Birney et al., 2007). The first surprise of human transcriptomics data was the unexpectedly low number of protein coding genes (previously estimated to be 30-70,000 (Harrison et al., 2002) and by this research to be 20,000-25,000) and the high overall transcription with more than 88% of genome transcribed into unannotated transcripts named by ENCODE as 'transcripts of unknown function' (TUF), today recognized as putative ncRNAs with a possible regulatory function (Birney et al., 2007).

Large intervening non-coding RNAs (lincRNAs) have been recently identified genome-wide using a combination of chromatin maps (overlapping H3K4me3 and H3K36me3 domains) and a tiling array approach. In mouse ~1600 lincRNAs have been mapped across 4 cell types with more than 95% of them beeing evolutionarily conserved (Guttman et al., 2009), while in human the same group found ~3300 highly conserved lincRNAs across 6 cell types (Khalil et al., 2009).

By using next generation mRNA-seq by Illumina technology 66% of polyadenylated human transcriptome has been mapped to annotated genes and 34% to unannotated regions, where these unannotated regions again have the potential of representing regulatory ncRNAs (Sultan et al., 2008). The second surprise of the transcriptomics data came with the discovered complexity of both the coding and non-coding portion of the transcriptome. Even less abundant than thought, the protein-coding transcriptome is highly diverse through alternative splicing, alternative initiation of transcription and alternative polyadenylation. For example, by combining mRNA-Seq and EST sequence data it has been estimated that ~95% of multiexon genes in human undergo alternative splicing (Pan et al., 2008) and two or more alternative promoters are used by ~58% of mouse protein coding transcripts (Carninci et al., 2006).

Furthermore, complexity is found in the abundance of overlapping transcripts where it has been estimated for the mouse that an average 7.6 transcripts are overlapping from the same strand and could be grouped into one transcription unit (Carninci et al., 2005), while overlapping antisense transcription has been found in mouse for 72% of all transcriptional units (Katayama et al., 2005). Overlapping transcription is a feature of both coding and non-coding regions and supports a new model of genomic organization in higher Eukaryotes as discussed by Kapranov et al. (2007). They suggest the old ‘collinear’ model based on Jacob and Monod description of lac operon in Bacteria (Jacob and Monod, 1961) needs to be replaced by an ‘interleaved’ model where: “multiple functional elements can overlap in the same genomic space” (Kapranov et al., 2007b). Underestimated complexity has been found also in the non-protein coding transcriptome (Kapranov et al., 2007a). This complexity is not completely understood or defined and challenges in defining this portion of transcriptome will be further described.

1. 3. 2. Challenges in defining the non-protein coding transcriptome

Next generation RNA Sequencing, tiling arrays and full-length cDNA sequencing projects together with bioinformatics has shown that the greatest portion of the mammalian genome is non-protein coding. When the majority of the transcription was revealed to map outside of annotated regions the first question was if all new transcriptional units are novel protein coding genes or if they are non-coding. Transcripts that did not have open reading frame (ORF)>300bp have been classified as non-coding. For example, in one of the first large studies of full-length cDNAs in the mouse genome it was found that 35% of 33,400 of clustered ‘transcriptional units’ were new non-coding transcripts on the basis of the ORF criteria (Okazaki et al., 2002), while even ~70% of 7500 full-length human cDNA sequences were shown to be non-coding by another group (Ota et al., 2004).

The ORF criterion has been useful but there could still be problems with this approach. For example, the *SRA* gene was first characterized as a ncRNA on the basis of ORF criteria, while later it was found that this transcript has multiple isoforms including some protein coding. Today we know that *SRA* has in the same RNA both regulatory and the protein coding functions (Chooniedass-Kothari et al., 2004). The second possible weakness of the ORF definition of ncRNA is based on the finding that in *Drosophila* transcripts with an ORF of 33bp could be translated, so it is not clear if 300bp is a valid number for the cut off of the ORF length for all cases. A third potential obstacle with the ORF criteria is possibility that ORF>300bp present in a transcript could be spurious (ORF that occur just by chance), something that could only be examined further by mutational analysis. Evolution favors synonymous over non-synonymous mutations in protein-coding genes and on that basis true ORF (protein coding gene) can be distinguished from spurious (non-coding RNA) (Lin et al., 2007).

Conservation is a second often used criteria for ncRNA mapping from any kind of the transcription data. NcRNAs are generally not conserved as highly as protein coding genes, although ncRNAs often have shorter stretches of sequence conserved, while whole ORF need to be conserved in protein coding genes (reviewed in (Mercer et al., 2009)). One of the explanations for these different conservation patterns is the finding that ncRNAs evolution is less restricted than for protein coding genes (Pang et al., 2006). The conservation approach in defining ncRNAs can be based on two measurements: reading frame conservation (RFC) and codon substitution frequency (CSF) (Clamp et al., 2007). RFC shows what is the percentage of nucleotides of an open reading frame that is conserved among species, while CSF score shows different patterns in nucleotide substitutions between protein coding genes and non-protein coding genes. For example, mouse linc (long intervening non-coding) RNAs have CSF<0 while known protein coding genes are showing CSF>0 scores (Guttman et al., 2009).

Guttman et al. (2009) developed a method that combines the rate of mutations with the level of constraint, where the calculated Pi LOD (logarithm of the odds) score represents the comparison of sequence evolution to neutrally evolving sequence. By using this method they found that lincRNA sequence conservation score is in the middle between lowly conserved introns and highly conserved protein-coding genes for both mouse (Guttman et al., 2009) and human (Khalil et al., 2009) lincRNAs.

The PhastCons score from the UCSC annotation is another measure of the sequence conservation where number of phastCons elements is counted across each ncRNA candidate region and compared with random regions scores. PhastCons scores>0.8 were used in a computational pipeline where non-coding RNAs were mapped from human low abundance expressed sequence tags (EST) (Xue and Li, 2008).

One group used secondary structure prediction criteria in combination with conservation to identify small ncRNAs (Weile et al., 2007). They used the RNAz program to filter out conserved secondary structured RNAs from multispecies conserved sequences and further they combined these data with their tiling array expression data from human neuroblastoma cell line. By this approach novel non-coding, structured and conserved RNA genes have been found, but what needs to be considered is that this kind of approach is optimized for detecting small ncRNAs (e.g. miRNAs and snoRNAs), while previous “ORF” and conservation approaches have been used also for macro ncRNA prediction. The ideal approach to computationally define a portion of the transcriptome as non-coding is still not available since all mentioned approaches are predictions with possible exceptions and need further experimental validation. Still, these approaches are highly valuable since they have increased our understanding of the transcriptome.

Wide-spread transcription outside of annotated regions has been named pervasive transcription, and if this transcription is functional (reviewed in (Mercer et al., 2009)) or represents transcriptional ‘noise’ (Ebisuya et al., 2008; Struhl, 2007) has been the focus of debates in the field. Today, research of many groups is showing supporting evidence towards functionality of the biggest part of ncRNA transcriptome. Developmental regulation, specific localization in the cell, evolutionary selection and the association with disease of many ncRNAs are supporting a functional role (Wilusz et al., 2009). Roles of the non-coding portion of the genome in disease will be further described.

1. 3. 3. Non-protein coding RNAs and disease

Non-protein coding RNAs have numerous regulatory roles in the cell and have been linked with diseases. The roles of micro ncRNAs in diverse human diseases from cardiovascular and muscular diseases to cancer are well established. Numerous micro ncRNAs are regulating cell growth and differentiation and are specifically involved in cancer where some of them are acting as oncogenes, while some are found to have tumor suppressor roles (reviewed in (Trang et al., 2008)). Involvement of imprinted macro ncRNAs in human syndromes and cancers has been discussed already in sections: 1.2.9.1 and 1.2.9.2, while in this section the main focus will consider non-imprinted macro ncRNAs and their role in disease.

The involvement of macro ncRNAs in disease is mainly based on correlative data showing deregulation of ncRNA expression in numerous diseases and cancer. Examples of macro ncRNAs altered in disease are reviewed in (Prasanth and Spector, 2007). Still, there are just few reports showing a causative role of ncRNA in disease. An interesting example is a report of an individual with α-thalassemia where deletion of *HBA1* and *HBQ1* genes juxtaposes a region normally located 18kb downstream from α-globin *(HBA2)* gene, next to the *HBA2* gene, leading to expression of a newly formed ncRNA transcript. This novel ncRNA formed by this deletion is antisense to *HBA2* and has been shown to cause *de novo* DNA methylation on the CpG island promoter of *HBA2* and silencing of the *HBA2* gene that further cause the disease (Tufarelli et al., 2003). A second interesting example has been found in Lynch syndrome patients where germline deletion of the 3' exons of the *EPCAM* gene cause transcriptional read through that leads to promoter DNA methylation and silencing of the *MSH2* gene (Ligtenberg et al., 2009; Niessen et al., 2009). Germline inactivation of one of the mismatch repair genes (*MLH1*, *MSH2*, *MSH6* and *PMS2*) is the known cause of Lynch syndrome, therefore the newly formed *EPCAM-MSH2* fusion ncRNA that silence *MSH2* has a causative role in Lynch syndrome (Ligtenberg et al., 2009; Niessen et al., 2009). In the past few years, there is more and more evidence supporting the roles of macro ncRNAs in disease and the possibilities of their use in medicine as biomarkers and potentially drug targets will be further described.

1. 3. 4. Non-protein coding RNAs as biomarkers and drug targets

There is evidence about the involvement of different classes of ncRNAs in human disease and especially in cancer. An emerging question is the possibility of their use as diagnostic or prognostic markers. The micro ncRNA field already widely recognized the potential for the use of these small ncRNAs as biomarkers and companies like Rosetta Genomics and Exiquon are focused on revealing novel miRNA biomarkers for various cancers. Macro ncRNAs are emerging as new molecules with a potential in prognostics. Numerous macro ncRNAs have been found up or downregulated in various multigenic diseases and cancer and just some of them will be discussed here.

The *DD3* macro ncRNA is expressed specifically in prostate and over-expressed in prostate cancer (Bussemakers et al., 1999). *DD3* was found to be a sensitive and specific marker for early detection of prostate cancer (de Kok et al., 2002). *aHIF* is a ncRNA found after exposure of cells to hypoxia and over-expressed in renal (Thrash-Bingham and Tartof, 1999) and breast cancer where is established as a strong predictor for poor breast cancer prognosis (Cayre et al., 2003). *MALAT-1* ncRNA is significantly associated with metastasis in non-small cell lung cancer (NSCLC) and has prognostic potential for the survivals of patients in stage I of NSCLC (Ji et al., 2003). The same ncRNA has been found to be over-expressed in endometrial sarcoma (Yamada et al., 2006) and in hepatocellular carcinomas (HCCs) where the possibility of using *MALAT-1* as a marker of neoplastic cells has been proposed (Lin et al., 2007). *P15AS* macro ncRNA is overexpressed in two forms of leukemia where it correlates with antisense silencing, and heterochromatization of the *p15* tumor suppressor gene showing the potential role of this macro ncRNA in cancerogenesis and potential use as a biomarker (Yu et al., 2008). Interestingly, fragments of ncRNAs can be detected from human blood by two approaches used by Semenov et al., which supports the possibility of biomarker ncRNAs usage in medical practice (Semenov et al., 2008).

The growing evidence of the role of macro ncRNAs in gene regulation suggests they could be potential molecular targets for epigenetic therapy. In the past, most of the drugs were directed towards proteins and delivering drugs to target ncRNAs will be a completely new challenge. Macro ncRNAs deregulated in disease have a potential to be targeted by siRNAs, antisense oligonucleotides, ribozymes or small molecules, but still there are many shortcomings in usage of these approaches, especially in patients. β-secretase-1 antisense (*BACE1-AS*) is a ~2kb long macro ncRNA involved in Alzheimer's disease that has the potential to be used as drug target. This ncRNA is upregulated in Alzheimer's patients and importantly its' function in regulating the *BACE1* enzyme crucial for cleavage of amyloid precursor protein (APP) is established (Faghihi et al., 2008). This ncRNA is directly implicated in increased abundance of plaque formation, one of the key features of Alzheimers' disease (Faghihi et al., 2008). A company named CuRNA was founded in 2008 with the goal of identifying of new biomarkers and drugs based on macro ncRNAs deregulated in disease (http://www.curna.com/). Further development of new technologies like tiling arrays and RNA Sequencing will lead to the identification of novel macro ncRNAs deregulated in disease e.g. cancer, potentially resulting in numerous benefits for medicine, in diagnostics, prognostics, and development of new drugs.

1. 4. Aims of the study

Recent development of new technologies: next generation RNA Sequencing and Tilling arrays hybridizations, led to a break through in the transcriptomics field and changed our perception of the complexity of genome. The majority of the transcripts in the human genome are non-protein coding, but still little is known about the diversity and functional potential of these transcripts. In order to give my contribution to the understanding of that diversity, I focused on the identification and characterization of one of the ncRNAs classes: macro ncRNAs. Mapping of macro ncRNAs by new technologies has been in itself the challenge since at the time I started this project the exact methodology for this kind of transcriptome research was not fully established.

My first goal was to establish methodologies and data analysis procedures based on Tiling array and RNA sequencing in order to reliably map novel macro ncRNAs and this part of project has been done in collaboration with Ru Huang and Florian Pauler in the lab. Focus on this specific ncRNAs class came from ten years expertise of our lab in the field of imprinted macro ncRNAs. In this thesis I use genomic imprinting as a model for macro ncRNA research based on the fact that well-studied imprinted gene regions express macro ncRNAs and that for two of these ncRNAs a direct role in gene regulation has been shown. Current understanding of the imprinting phenomenon and the role of macro ncRNAs in imprinted gene regulation is mainly based on 6 out of about 30 imprinted gene clusters. Thus, my second goal was to analyze transciption of all human imprinted gene clusters in different tissues and developmental stages and map and characterize novel macro ncRNA transcripts in these regions in order to address the question if macro ncRNAs are a universal feature of imprinted gene regions in human. Characterization of novel imprinted macro ncRNAs is valuable in broadening our knowledge about the biology of these transcripts and could be helpful in predicting their functions. With the recent development of ncRNA field it became clear that macro ncRNA transcripts are associated with a spectrum of human diseases and in a few cases a causative role of a macro ncRNA in disease development has been found. Numerous macro ncRNAs are biomarkers for different cancer types and some of them are proposed as potential drug targets. Thus, my third goal was to investigate deregulation of novel macro ncRNAs found in imprinted gene regions in different types of cancer. In conclusion, this PhD thesis aims to broaden our knowledge of macro ncRNAs from imprinted gene regions in both normal and disease conditions and will be a valuable resource for further functional studies of human macro ncRNAs.

2. Results

2. 1. HIRTA (Human Imprinted Region Tiling Array) technology successfully detects macro ncRNAs

2. 1. 1. Selection of regions of interest and design of HIRTA

The Human Imprinted Region Tiling Array (HIRTA) was designed with the aim of mapping novel macro ncRNAs in human regions containing imprinted genes. HIRTA is a NimbleGen custom tiling array with one 50bp oligonucleotide (further referred as a tile) per 100bp of single copy sequence from selected gene regions (Figure 7A). HIRTA covers around 2% of the human genome. 4% of all HIRTA probes cover exons of annotated genes, 3% cover intron-exon junctions, 49% cover intronic regions while 43% of the HIRTA cover intergenic regions (Figure 7B). Note that interspersed repeats identified by Repeat Masker (http://www.repeatmasker.org/) were excluded from HIRTA (black bars in Figure 7A).

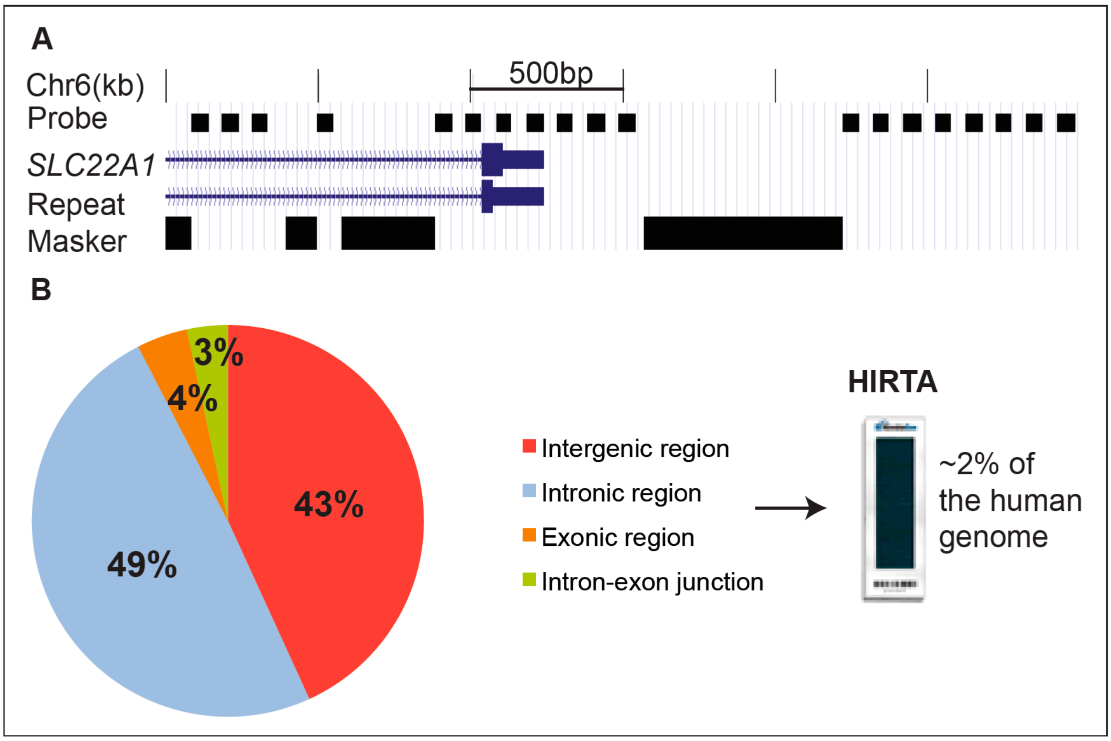


Figure 7. HIRTA covers ~2% of the human genome. A. UCSC snapshot showing a 3kb region of human chromosome 6 and HIRTA oligonucleotide probes 50bp in length (short black bars) that cover each 100bp of a single copy sequence. RefSeq genes UCSC track shows part of *SLC22A1* gene (dark blue) indicating that HIRTA oligonucleotide probes cover both genic and intergenic regions. Repeat Masker track shows that interspersed repeats (black bars) are not covered by HIRTA oligonucleotide probes. **B.** 49% of the sequence covered by HIRTA oligonucleotide probes are positioned in intronic regions, 43% in

intergenic, 4% in intron-exon junctions and 3% in exonic regions of annotated genes. HIRTA is a NimbleGen array that covers ~2% of human genome.

The 2% of human genome covered by HIRTA includes three types of regions containing imprinted genes: 1) well-studied regions containing genes imprinted in both human and mouse, 2) less-studied regions containing genes imprinted in mouse, human or both and 3) the XIST region that is well-studied but imprinted just in mouse extraembryonic tissues (Table 9; genes showing imprinted expression in these regions were described in section 1.2.5.). Regions contained on HIRTA map on 16 human chromosomes and range in length from 1-5Mb (Table 9). The HIRTA regions contain centrally positioned genes known to be imprinted in human and/or mouse and cover the known imprinted gene cluster plus flanking regions containing non-imprinted genes.

HIRTA REGION	Chr.	REGION START	REGION END	LENGTH (Mb)	REGION imprinting status
TP73	1	2,900,000	3,900,000	1	L (H)
DIRAS3	1	67,800,000	68,800,000	1	L (H)
COMMD1	2	61,500,000	62,500,000	1	L (M)
NAP1L5	4	89,300,000	90,300,000	1	L (M, H)
PLAGL1	6	143,830,000	144,830,000	1	L (M, H)
IGF2R	6	160,000,000	161,000,000	1	W (M, H)
GRB10	7	50,200,000	51,200,000	1	L (M, H)
CALCR	7	92,400,000	97,400,000	5	L (M, H)
MEST	7	129,500,000	130,500,000	1	L (M, H)
KCNK9	8	139,600,000	142,600,000	3	L (M, H)
SFMBT2	10	5,800,000	8,800,000	3	L (M)
INPP5F	10	121,100,000	122,100,000	1	L (M, H)
BAZ2	11	6,000,000	8,000,000	2	L (H)
AMPD3	11	10,000,000	11,000,000	1	L (M)
IGF2, KCNQ1	11	1,700,000	3,700,000	2	W (M, H)
WT1	11	31,830,000	32,830,000	1	L (H)
SDHD	11	110,800,000	111,800,000	1	L (H)
SLC38A4	12	45,000,000	46,000,000	1	L (M)
DCN	12	89,600,000	90,600,000	1	L (M)
HTR2A	13	45,800,000	46,800,000	1	L (M, H)
DLK1	14	99,600,000	101,600,000	2	W (M, H)
PWS	15	20,170,000	25,170,000	5	W (M, H)
GATM	15	42,900,000	43,900,000	1	L (M)
RASGRF1	15	76,600,000	77,600,000	1	L (M)
IMPACT	18	19,900,000	20,900,000	1	L (M)
ZIM2	19	61,600,000	62,600,000	1	L (M, H)
HM13	20	29,200,000	30,200,000	1	L (M, H)
NNAT	20	35,080,000	36,080,000	1	L (M, H)
GNAS	20	56,500,000	57,500,000	1	W (M, H)
L3MBTL	20	41,100,000	42,100,000	1	L (H)
XIST	X	70,700,000	75,700,000	5	X (M)

Table 9. HIRTA oligonucleotide probes are covering 32 genomic regions with at least one gene having imprinted expression in human or mouse. Names of the HIRTA regions (given according to the RefSeq name of the gene that shows imprinted expression and is centrally positioned within the region), positions on human chromosomes (Chr.) in the human build NCBI36/hg18, starts and ends of the regions, their length (in mega bases) and grouping according to the imprinting status in human and mouse (into well-studied imprinted (W), less-studied imprinted (L) and XIST (X) region) is shown. M; region cointains genes showing imprinted expression in mouse, H; region contains genes showing imprinted expession in human.

2. 1. 2. Normal and cancer samples were hybridized on HIRTA

HIRTA was hybridized with cDNA from 43 normal or cancer human samples including cells, tissues and patient tissue samples (Figure 8).

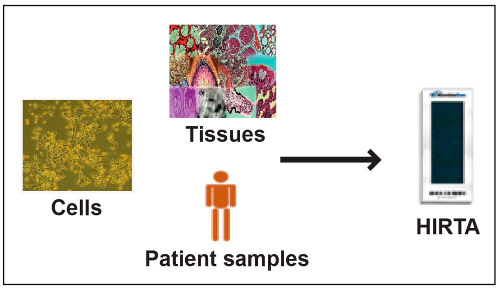


Figure 8. 43 human samples were hybridized to HIRTA. cDNA from three types of samples: cultured cells, tissues and patient samples were hybridized to HIRTA.

The 20 normal cells/tissues hybridized to the HIRTA array are described in Table 10. Undifferentiated human embryonic stem cells (HES2d0) and day 7 differentiated human embryonic stem cells (HES2d7) were hybridized to HIRTA in order to test the developmental regulation of known and novel macro ncRNAs. Normal human fibroblast cell line, samples from 3 fetal human tissues, placenta (extraembryonic tissue) and 13 different adult tissues were hybridized to HIRTA with the aim of testing tissue specific expression of the macro ncRNAs (Table 10). RNA from human tissues was purchased from Clontech and in most cases only pooled tissues were available. While this does not allow individual variation to be assessed, it has the advantage of reducing effects of biological variation at the level of the gene expression that is relatively high in human (Cheung et al., 2003).

NORMAL	
Cell line/tissue	**Description**
Hs27	Normal male human fibroblasts
HES2d0	Undifferentiated female embryonic stem cells
HES2d7	7 days differentiated female embryonic stem cells
Fetal Brain	21 male/female: Caucasian fetuses, ages: 26-40 weeks
Fetal Liver	63 male/female Caucasian fetuses, ages: 22-40 weeks
Fetal Kidney	34 Caucasian male/female fetuses, ages: 12-31 weeks
Placenta	Obtained from 4 female Caucasians, ages: 21-39
Adult Brain	2 male Caucasians, ages: 47-55
Adult Lung	3 male/female Caucasians, ages: 32-61
Adult Uterus	8 female Caucasians, ages: 23-63
Adult Heart	10 male/female Caucasians, ages: 21-51
Adult Kidney	14 male/female Caucasians, ages: 18-59
Skeletal Muscle	7 male/female Caucasians, ages:20-68
Bone Marrow	8 male/female Caucasians, ages: 18-56
Colon	1 female Caucasian, age: 23
Cervix	1 female African American, age: 40
Mammary Gland	1 female Caucasian, age: 27
Adult Liver	1 male Caucasian, age: 51
Whole Blood	1 female Caucasian, age: 30
Testis	39 male Caucasians, ages: 14-69

Table 10. 20 normal human cells/tissues has been hybridized to HIRTA. Names of used samples and their description is presented.

Further, 17 cell lines from 6 different cancer types (cervical, breast, colon, teratocarcinoma, rhabdomyosarcoma and neuroblastoma) and 6 patient samples (4 acute myeloid leukemias and 2 myeloproliferative disorders) were hybridized to HIRTA in order to test regulation of macro ncRNAs in cancer (Table 11).

CANCER	
Cell line/tissue	**Description**
HeLa	Cervical cancer cell line, adenocarcinoma
HT3	Cervical cancer cell line, carcinoma
C4I	Cervical cancer cell line, carcinoma
C4II	Cervical cancer cell line, carcinoma
SiHa	Cervical cancer cell line, squamous cell carcinoma, grade II
C33A	Cervical cancer cell line, carcinoma
DoTc2	Cervical cancer cell line, carcinoma
ME180	Cervical cancer cell line, epidermoid carcinoma
SW756	Cervical cancer cell line, squamous cell carcinoma
HCT116	Colon cancer cell line, colorectal adenocarcinoma
Caco2	Colon cancer cell line, colorectal adenocarcinoma
MCF7	Breast cancer cell line, adenocarcinoma
CAMA-1	Breast cancer cell line, adenocarcinoma
Tera2	Malignant embryonal carcinoma
NCCIT	Teratocarcinoma
A201	Rhabdomyosarcoma
SH-SY-5Y	Neuroblastoma
AML5_BMMC	Acute myeloid leukemia patient, bone marrow mononuclear cells
AML5_PBMC	Acute myeloid leukemia patient, periferal blood mononuclear cells
AML7	Acute myeloid leukemia patient, bone marrow mononuclear cells
AML8	Acute myeloid leukemia patient, bone marrow mononuclear cells
MP_0351B	Myeloproliferative disorder patient, periferal blood
MP_0363	Myeloproliferative disorder patient, periferal blood

Table 11. 23 cancer samples including: cervical, colon, breast, teratocarcinoma, rhabdomyosarcoma, neuroblastoma, acute myeloid leukemia and myeloproliferative disorder cancer types have been hybridized to HIRTA. Names of the cell lines, patient tissue samples and short descriptions are shown.

2. 1. 3. Reproducibility of HIRTA

HIRTA reproducibility was tested by performing five replicates of HIRTA hybridizations. Three biological (Hs27, Adult uterus and HeLa) and 2 technical replicates (HES2d0 and HES2d7) were tested. Scatter plots of the replicates that showed near linear relationships between replicates and Pearson's correlations ranging from r= 0.885 to 0.951 (where +1 is a perfect linear correlation) demonstrated that HIRTA is highly reproducible (Figure 9A). Looking at the single gene level, the *Decorin* (*DCN*) gene was showing no (HES2d0, HES2d7), low (HeLa), medium (Hs27) or high (Adult uterus) expression in both replicates of each of five tested samples (Figure 9B). High reproducibility of highly expressed and saturated *DCN* gene was indicated by its intronic signal that showed similar levels of expression in two biological replicates (*DCN* expression in Adult uterus, Figure 9B). As the HIRTA hybridization technique is highly reproducible for 5 examples, replicates were not performed for other tissues and cell lines.

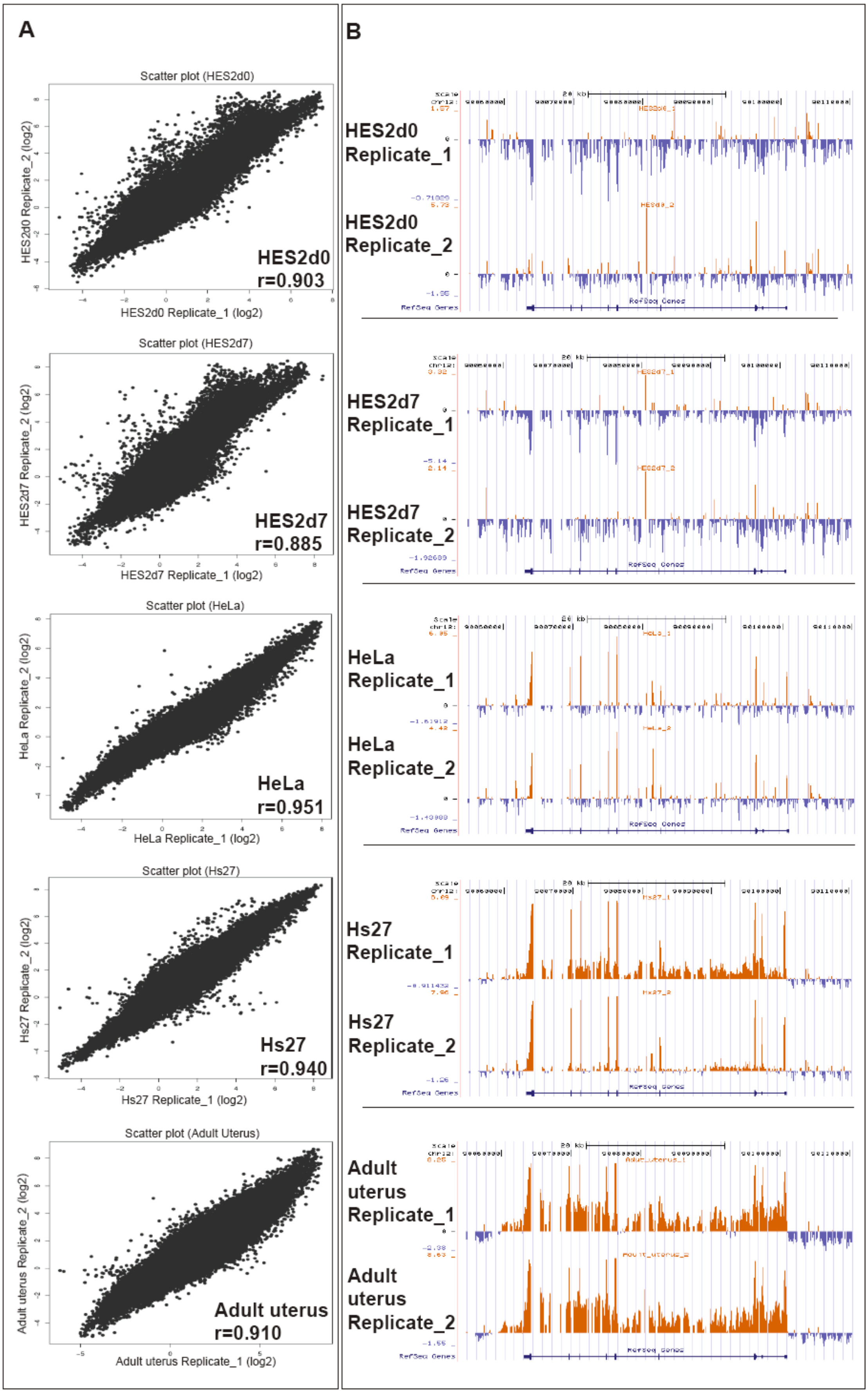

A
Scatter plot (HES2d0)
HES2d0 Replicate_2 (log2)
HES2d0 Replicate_1 (log2)
HES2d0
r=0.903
Scatter plot (HES2d7)
HES2d7 Replicate_2 (log2)
HES2d7 Replicate_1 (log2)
HES2d7
r=0.885
Scatter plot (HeLa)
HeLa Replicate_2 (log2)
HeLa Replicate_1 (log2)
HeLa
r=0.951
Scatter plot (Hs27)
Hs27 Replicate_2 (log2)
Hs27 Replicate_1 (log2)
Hs27
r=0.940
Scatter plot (Adult Uterus)
Adult uterus Replicate_2 (log2)
Adult uterus Replicate_1 (log2)
Adult uterus
r=0.910
B
HES2d0
Replicate_1
HES2d0
Replicate_2
HES2d7
Replicate_1
HES2d7
Replicate_2
HeLa
Replicate_1
HeLa
Replicate_2
Hs27
Replicate_1
Hs27
Replicate_2
Adult
uterus
Replicate_1
Adult
uterus
Replicate_2
Scale
20 kb
chr12:
RefSeq Genes

Figure 9. HIRTA hybridizations show high reproducibility. A. Comparissons of $\log_2$ expression signals from two replicates of Hs27, HES2d0, HES2d7, Adult uterus and HeLa are represented using scater plots. Each black dot represents $\log_2$ expression from one HIRTA tile. Scatter plots show a nearly linear relationship of two replicates for five HIRTA samples, r; Pearson's correlations. **B.** *Decorin* (*DCN*) expression for five tested HIRTA samples in two replicate hybridizations. HIRTA hybridization results were loaded on the UCSC browser in the form of custom track where X-axis represents position in the genome (hg18), while Y-axis is a $\log_2$ ratio of cDNA normalized to genomic DNA. Name of the custom track is shown on the left side of the figure. Publicaly available UCSC track for RefSeq genes is shown, as well as scale and hg18 chromosome positions.

2. 1. 4. Dynamic range of HIRTA

Tiling arrays have been shown to display limited dynamic range with signal saturation of highly expressed genes (Wang et al., 2009a). The dynamic range of HIRTA was tested by comparing expression of *Decorin* gene (*DCN*) from three different cell lines by HIRTA hybridization and RT-qPCR. *DCN* showed high expression in Hs27 (value of *DCN* gene expression in fibroblasts was 8.21, while maximal saturated value of HIRTA expression in fibroblasts was 8.38, observed for *COL1A2* gene), medium expression in HeLa (value of 6.05) and no expression in HESd7 cells (values down to minus 6.14 over the exonic regions) by HIRTA hybridizations (Figure 10A). RT-qPCR data for the same gene using four exonic (DCNEx1-4) and four intronic primer pairs (DCNIn1-4) confirmed the relatively high expression in Hs27 cell line (set to 100) while expression in HeLa was about 50 fold less in exons and 100 fold less in introns. *DCN* was not expressed in HES2d7 by both RT-qPCR and the tiling array (Figure 10B).

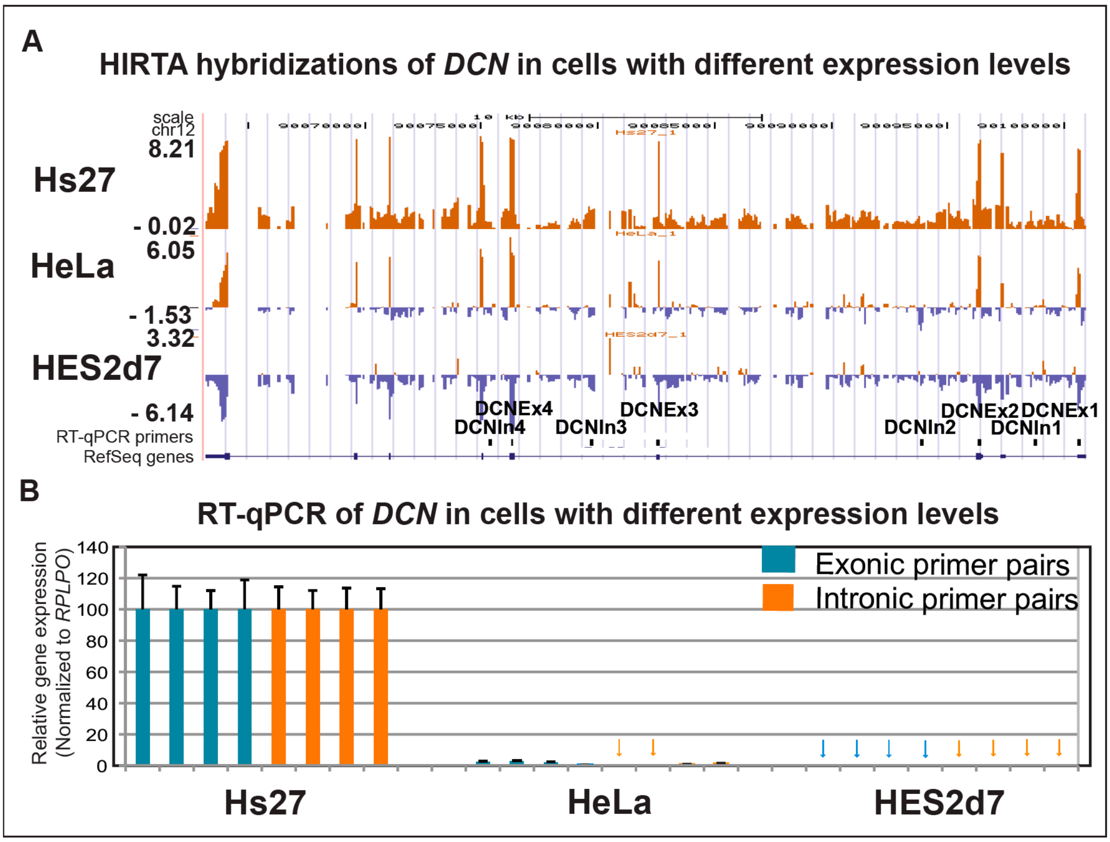


Figure 10. Tiling arrays can distinguish high, medium and no expression but has a low dynamic range and a saturation limit. A. Tissue specific expression of the *Decorin* in Hs27, HeLa and HES2d7 cells. Tracks are displayed on UCSC as on Figure 9. B. RT-qPCR primers positions are also shown. **B.** RT-qPCR validation of tiling array data using the human *Decorin* (*DCN*) gene was done in cell lines with different expression levels. Expression is assayed by eight primer pairs, four of which were located in exons (blue) and four in introns (orange) of the *Decorin* gene. Negative exonic and intronic expression showed by these primer pairs are indicated using blue and orange arrows respectivelly. Each primer pair is set to 100 in Hs27 cells and normalized to expression of the *RPLPO* gene.

The dynamic range of HIRTA was ~250 fold expression difference (Table 12). *DCN* expression values in Hs27 and HeLa, obtained from HIRTA hybridizations, were about a 5 fold different, while RT-qPCR showed a 50 fold difference in *DCN* expression between the cell lines*,* indicating that *DCN* expression on HIRTA was saturated (Table 12).

The dynamic range and saturation of HIRTA Chip based on RT-qPCR and HIRTA expression of *Decorin* gene

	Hs27 expression values	HeLa expression values	Difference in expression values between the cell lines
HIRTA	$\text{Log}_2 n=8.21$ n= 296.11	$\text{Log}_2 n=6.05$ n=66.26	= ~4.5 fold
RT-qPCR	100	~2	= ~50 fold

If HIRTA data range from $\text{Log}_2 n=0$, n=1 to $\text{Log}_2 n=8$, n=256, than by HIRTA we are able to observe **~250 fold** difference in expression (potential background not taken into account) Note that HIRTA hybridizations have different saturation limit but typically they have maximal observed $\log_2$ expression in range of 7.5 to 8.5.

Table 12. Comparisson of HIRTA data and RT-qPCR in different tissues shows that dynamic range of HIRTA is about 250 and that HIRTA has a saturation limit.

2. 1. 5. Single cDNA hybridization on HIRTA

Total RNA from cell lines, tissues or patient samples was converted to double stranded cDNA, labeled with a flourochrome Cy5 and co-hybridized with sonicated DNA from the Hs27 cell line, labeled with a fluorochrome Cy3 (Figure 9A). This type of hybridization was named “single cDNA hybridization”. The raw data was Tukey bi-weight normalized and the HIRTA hybridization results are displayed on the UCSC (University of California Santa Cruz) genome browser using the human Mar. 2006 (NCBI 36/hg18) assembly with the position in the genome on the X-axis and $\log_2$ signal values from HIRTA hybridizations on Y-axis (Figure 9). HIRTA probe intensities were $\log_2$ transformed in order to enable visualization of lowly transcribed genes. Log_2 signal values displayed in orange and positioned above the zero line represent enrichment of cDNA over genomic DNA and indicate RNA expression. The blue signal below the zero line shows that there is more genomic DNA than cDNA and that RNA is not expressed. Zero line represents equal amounts of cDNA and genomic DNA. The results of human fibroblasts hybridizations (Hs27) for a typical imprinted macro ncRNA and a typical protein coding gene show different tiling array patterns (Figure 11B). Protein coding genes (e.g. *DCN*) typically show high signals matching to exons and low to moderate signals over introns. The high exonic signals are often saturated, depending on the expression in examined tissue. This observed signal pattern of protein coding genes is based on their high intron/exon ratio, typical of most mammalian genes. Macro ncRNAs (e.g. *GTL2*) are found to show a HIRTA pattern with most of the tiles strongly positive through whole body of macro ncRNA gene (they have low intron/exon ratio), with the HIRTA signal that is typically not saturated. On the basis of these distinct tilling array patterns macro ncRNA can be differentiated from protein coding genes by visual inspection of tiling array hybridization results displayed on the UCSC genome browser.

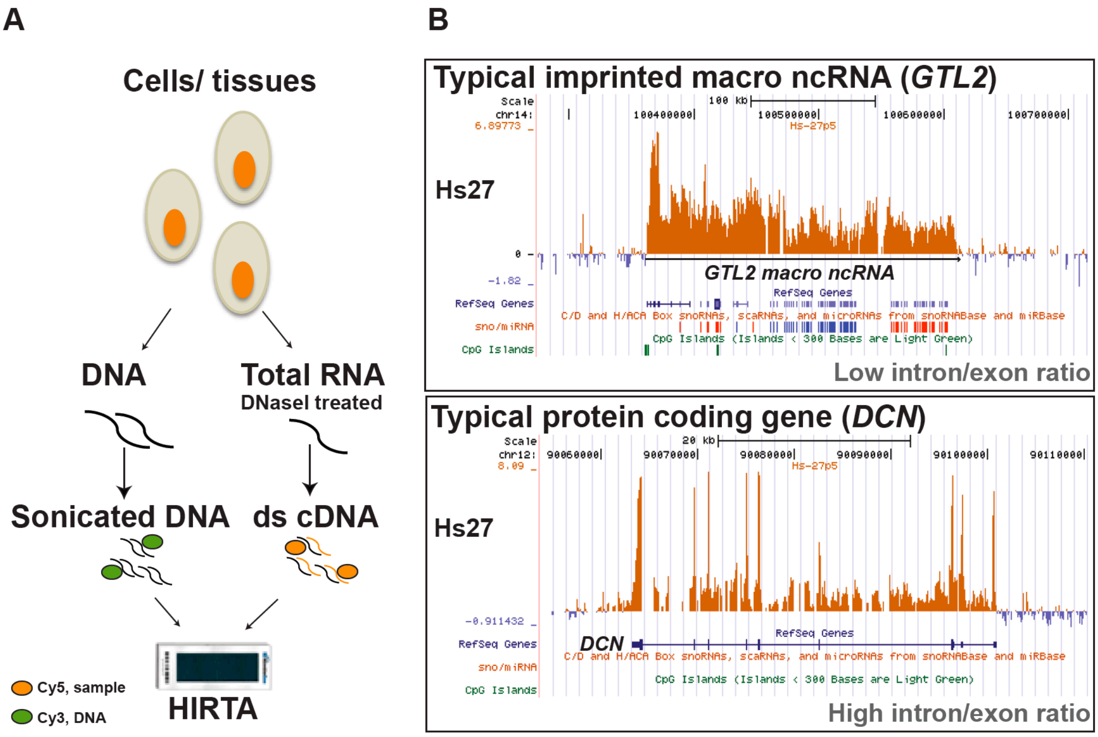


Figure 11. Single cDNA hybridizations on HIRTA can detect macro ncRNAs and distinguish them from protein coding genes. A. Short overview of the sample preparation using “single cDNA hybridization” on HIRTA. ds cDNA; double stranded cDNA **B.** The figure is displayed as in Figure 9B. In addition, publically available UCSC tracks for sno/miRNA and CpG islands are shown. Sno/miRNA UCSC track display miRNAs in red and snoRNAs in blue color. The upper panel shows a typical known macro ncRNA, the *GTL2* transcript expressed from the DLK1 imprinted gene region that is also precursor to miRNAs and snoRNAs. The lower panel shows *Decorin (DCN)* as an example of typical protein coding gene showing high expression signals (orange focal peaks) matching the exons of *DCN* (shown as dark blue blocks on RefSeq gene track) and moderately expressed introns.

The expression patterns of already known macro ncRNAs and their differences to annotated protein coding genes, that resulted from single cDNA hybridizations on HIRTA, were used to develop criteria for mapping novel variants of known or novel macro ncRNAs. HIRTA macro ncRNA mapping by visual inspection criteria include: 1) high coverage of positive expressed probes through the body of the transcript (coverage was used as a relative measure of low intron/exon ratio with typical coverage >90%, and in the case of lowly transcribed transcripts >70% coverage); 2) absence of typical protein coding exons visualized as highly expressed focal signals, that typically consist of 1-2 tiles as the average human exon is <200bp in length, and are often positioned on the distances of about 2-8kb typically resembling introns (Lander et al., 2001; Sakharkar et al., 2004)); 3) more than 1kb in length; 4) Simultaneous analysis in 43 cells/tissues/patients. The first two criteria are based on the already shown difference between macro ncRNAs and protein-coding genes in the HIRTA expression pattern (Figure 11) while the third criteria limits the analysis to very long/macro ncRNAs; potential 200bp to 1kb long ncRNAs were not analysed since transcription of these lengths are more difficult to distinguish from cross-hybridizations, pseudogenes or background. Further, the fourth criteria requiring simultaneous detection from all HIRTA hybridizations resolves three potential problems: a) difficulties in distinguishing macro ncRNAs from protein coding genes showing high transcription rate, which can have introns expressed to the similar extent as exons, b) potential overlapping transcripts c) distinguishing between low transcription (less than $\log_2=1$) and tilling array background, if a transcript is highly expressed in one of the tissues and a low expression from another tissue maps to the same position, then this expression was recognized as a ncRNA. Positions of the first and the last HIRTA oligonucleotide probe giving high expression were defined as the start and the end of the macro ncRNA transcript. Positions of intergenic transcripts were reliably mapped while the positions of overlapping and very lowly expressed transcripts were given provisionally since exact mapping of these transcripts on the basis of tiling array data was not possible.

2. 1. 6. Double cDNA hybridization on HIRTA

In order to test the cellular localization of macro ncRNAs, double hybridizations on HIRTA were performed using two-color labeling of double stranded (ds) cDNA from nuclear and cytoplasm fractions of a normal human fibroblasts cell line (Hs27) (Figure 12A). The raw data was normalized by method implemented by Dr. Florian Pauler (unpublished data). The normalized hybridization results were loaded into UCSC genome browser as the custom track with black bars above the zero line indicating nuclear enriched RNA regions and grey bars below the zero line indicating the cytoplasmic enriched RNA. As expected, exons of protein coding genes were enriched in cytoplasmic fraction (e.g. *DCN* exons, gray bars below the zero line), while a known nuclear *KCNQ1OT1* macro ncRNA (Murakami et al., 2007) was enriched in the nucleus (Figure 12B). Thus, we further used this technique for testing cellular localization of novel macro ncRNAs.

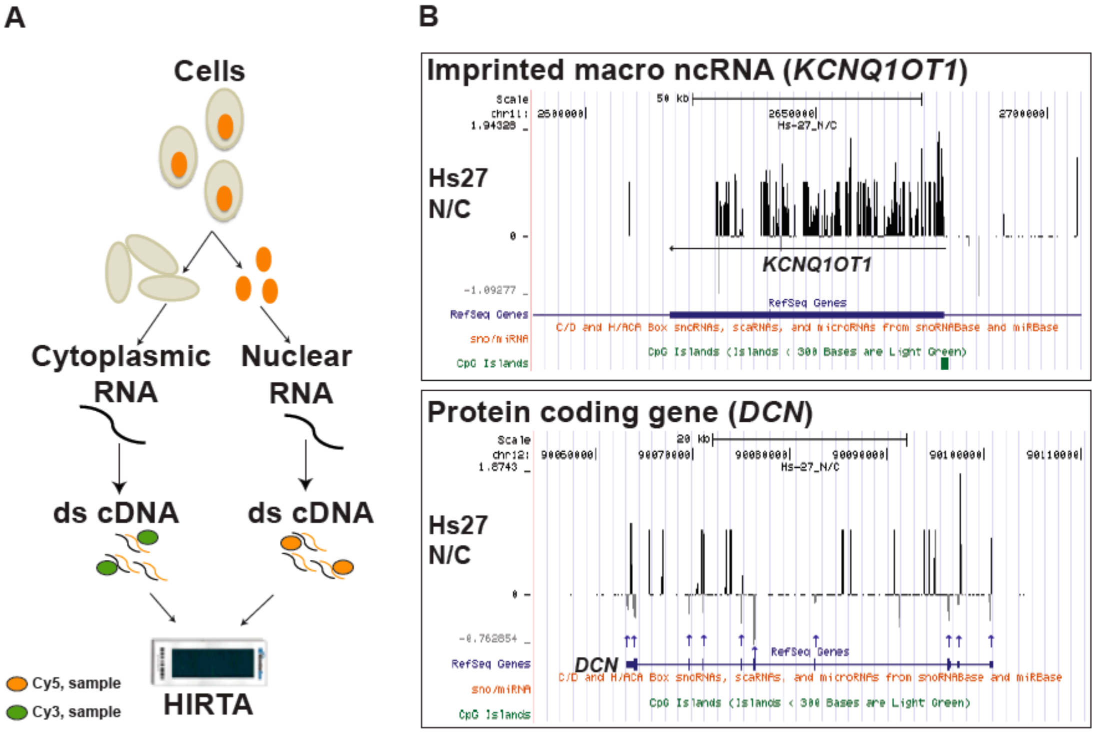


Figure 12. Double cDNA hybridization shows nuclear enrichment of the *KCNQ1OT1* macro ncRNA and cytoplasmatic enrichment of the *Decorin* protein coding gene. A. Short overview of the sample preparation using “double cDNA hybridization” on HIRTA. ds cDNA; double stranded cDNA **B.** The upper panel shows result of double cDNA hybridization on HIRTA for known nuclear localized *KCNQ1OT1* imprinted macro ncRNA indicated with the black arrowed line. The lower panel shows example of cytoplasmic localization of typical protein coding gene. Black signals above the zero line represent nuclearly localized introns while gray signals below the zero line correspond to exons of the *DCN* gene indicated by the blue arrowed lines. Names of the double cDNA hybridization custom tracks are shown on left part of both panels as well as names of displayed UCSC custom tracks (RefSeq Genes, sno/miRNA, CpG islands).

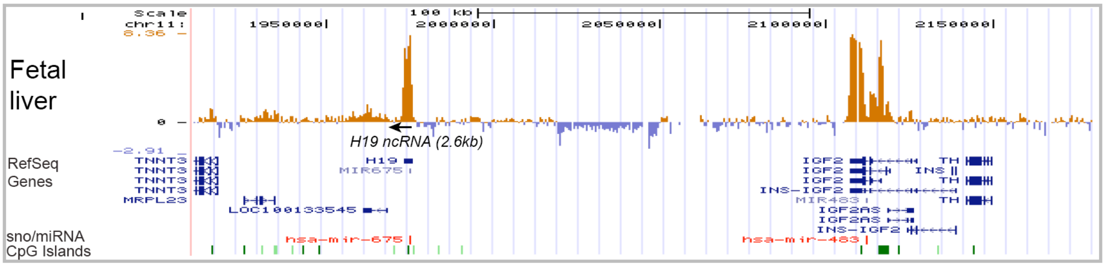
2. 2. Macro ncRNAs in six well-studied human imprinted gene regions

2. 2. 1. Known human macro ncRNAs from the HIRTA well-studied regions

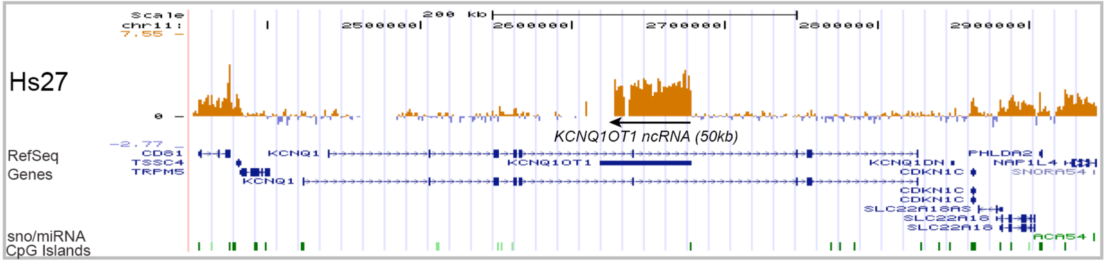
2. 2. 1. 1. Mapping and tissue specific expression of known macro ncRNAs

Six well-studied regions in mouse cointaining imprinted genes are: Igf2, Kcnq1ot1, Igf2r, Pws/As, Gnas and Dlk1 (Koerner et al., 2009). Analysis of the same regions in human, using HIRTA, showed that six known human imprinted macro ncRNAs (*H19*, *KCNQ1OT1*, *GTL2*, *UBE3A-AS*, *NESPAS*, *EXON1A*) were successfully detected in five of well-studied regions (Figure 13). No single tissue expressed all six well-studied macro ncRNAs and the human *AIRN* macro ncRNA (that is described in section 2.6.3.) was not detected in any of tissues studied by HIRTA. The detected macro ncRNAs were used as positive controls for HIRTA and based on their expression pattern, criteria for mapping novel macro ncRNAs were established (described in section 2.1.5.).

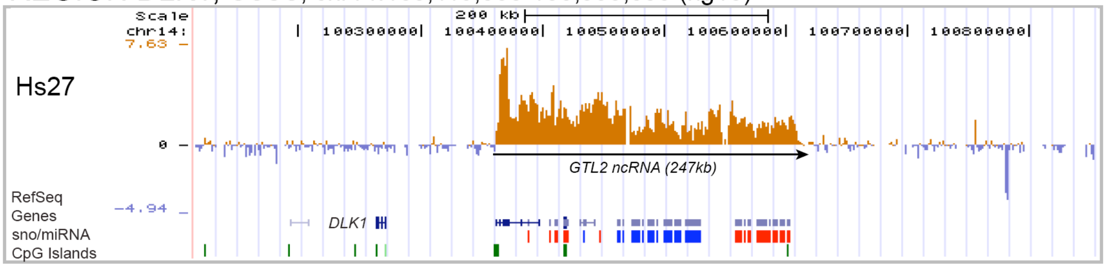
REGION IGF2, UCSC; chr11:1,910,000-2,180,000 (hg18)



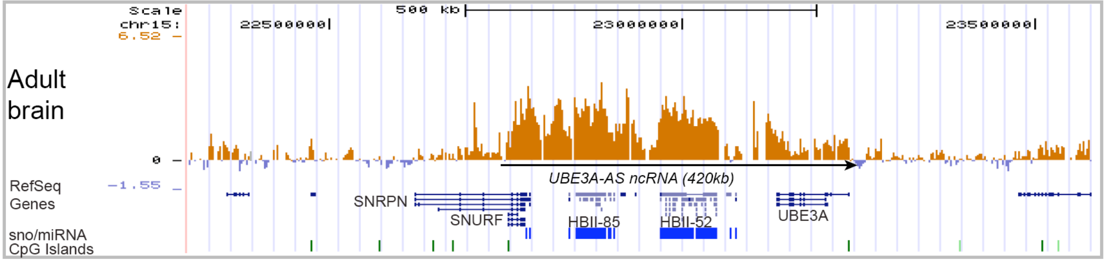
REGION KCNQ1, UCSC; chr11:2,350,000-2,945,000 (hg18)



REGION DLK1, UCSC; chr14:100,115,000-100,855,000 (hg18)



REGION PWS, UCSC; chr15:22,300,000-23,580,000 (hg18)



REGION GNAS, UCSC; chr20:56,790,000-56,960,000 (hg18)

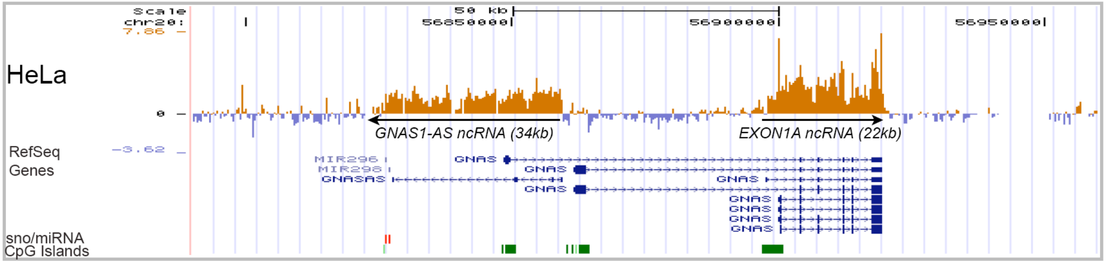


Figure 13. Macro ncRNAs in well studied human imprinted gene regions. Expression of *H19, KCNQ1OT1, GTL2, UBE3A-AS, GNAS1-AS* and *EXON1A* macro ncRNAs is shown by presentation of one HIRTA track from a cell line/tissue that highly expressed each ncRNA. Known macro ncRNAs orientation is depicted with black arrowed lines while their lengths in kilobases (kb) are shown in brackets. Names of HIRTA regions and screenshot positions in

hg18 are shown on the top of each box. Custom HIRTA tracks as well as RefSeq genes, sno/miRNA and CpG islands UCSC tracks are shown.

H19, KCNQ1OT1, GNAS1-AS and EXON1A are macro ncRNAs annotated by RefSeq with gene lengths of 2.3kb, 59.5kb, 32kb and 22kb respectively and the relatively same lengths were mapped by HIRTA except in the case of the *KCNQ1OT1* ncRNA when the repetitive region positioned downstream of ncRNA transcript was not spoted on the HIRTA Chip, therfore the exact end of this ncRNA could not be determined (Table 13).

Mapping of the *GTL2* and *UBE3A-AS* macro ncRNAs in the DLK1 and PWS HIRTA regions respectively showed a complex transcriptional pattern in different tissues that resulted in the mapping of 6 tissue specific transcripts that are novel variants of, or transcripts overlapping *GTL2* (Figure 14), and 4 novel tissue specific transcripts that are variants of, or are overlapping with the *UBE3A-AS* macro ncRNA (Figure 16). *GTL2* (*MEG3*) is by RefSeq Genes as a ~35kb long transcript. *GTL2var1* (Figure 13, 14) was mapped using HIRTA as ~250kb transcript, overlapping known miRNAs and snoRNA clusters. In diverse tissues, different novel variants were mapped using HIRTA: *GTL2var2* overlapping both annotated *MEG3* and miRNA cluster, *GTL2var3* that may be short variant of annotated *MEG3, GTL2var4* coresponding to the second miRNA cluster in the region, *GTL2var5* that is lowly expressed and corresponds to two known snoRNA clusters (14qI and 14qII) and *GTL2var6* that is positioned next to the retrotransposon like-1 (*RTL1*) and may be overlapping this gene (Figure 14). Interestingly, *GTL2* long and short variants may use different promoters in different tissues. The *GTL2* long transcripts (*GTL2var1* and *GTL2var2*) use one promoter in Hs27 and adult brain, while the *GTL2*var3 uses another promoter in adult lung and heart (Figure 15).

HIRTA REGION	Chr.	NAME of macro ncRNA	POSITION (hg18)	Length (kb)	COV (%)	Position to annotated PC genes
IGF2	11	***H19***	chr11:1972982-1975641 (RefSeq)	2.6	100	IG
KCNQ1		***KCNQ1OT1***	chr11:2618344-2677804 (RefSeq)	59.5	99.7	IN
DLK1	14	***GTL2var1***	chr14:100362198-100609108	246.9	99.3	IG
		GTL2var2	chr14:100362198-100473308	111.1	93	IG
		GTL2var3	chr14:100364721-100372481	7.7	98.4	IG
		GTL2var4	chr14:100537556-100609108	71.5	74.9	IG
		GTL2var5	chr14:100459386-100536578	77.2	72.3	IG
		GTL2var6	chr14:100420937-100440719	19.8	85.5	IG
PWS	15	***UBE3A-ASvar1***	chr15:22751163-23171714	420.5	97.3	OV
		UBE3A-ASvar2	chr15:22751163-22960127	209	99.5	OV
		UBE3A-ASvar3	chr15:22751163-22918716	167.5	98.3	OV
		UBE3A-ASvar4	chr15:22954110-23050454	96.4	58.3	IG
GNAS	20	***GNASAS***	chr20:56827368-56859353 (RefSeq)	32	98.1	5'/OV
		EXON1A	chr20:56897575-56919642 (RefSeq)	22	98.5	OV

Table 13. Known macro ncRNAs in five well-studied imprinted gene regions. *H19, KCNQ1OT1, GNASAS* and *EXON1A* are mapped by HIRTA to positions annotated as RefSeq genes on UCSC. *GTL2* and *UBE3A-AS* show a complex transcription patterns and number of tissue specific variants of these macro ncRNA are detected by HIRTA. The length of visualy mapped macro ncRNAs ranges between 2.6 and 420kb. HIRTA regions named in accordance to Table 9, chromosome positions according to hg18, name of macro ncRNAs and their length (in kilobases) were shown. Coverage (COV) represents % of positive probes in a tissue expressing high level of the macro ncRNA candidate. Macro ncRNAs were grouped according to their position in relation to annotated protein coding genes: intergenic (IG)= non-overlapping, 5'/OV= 5' exons or overlapping, 3'/OV= 3'UTRs or overlapping, OV=overlapping, IN=inside of annotated protein-coding genes. PC; protein-coding gene

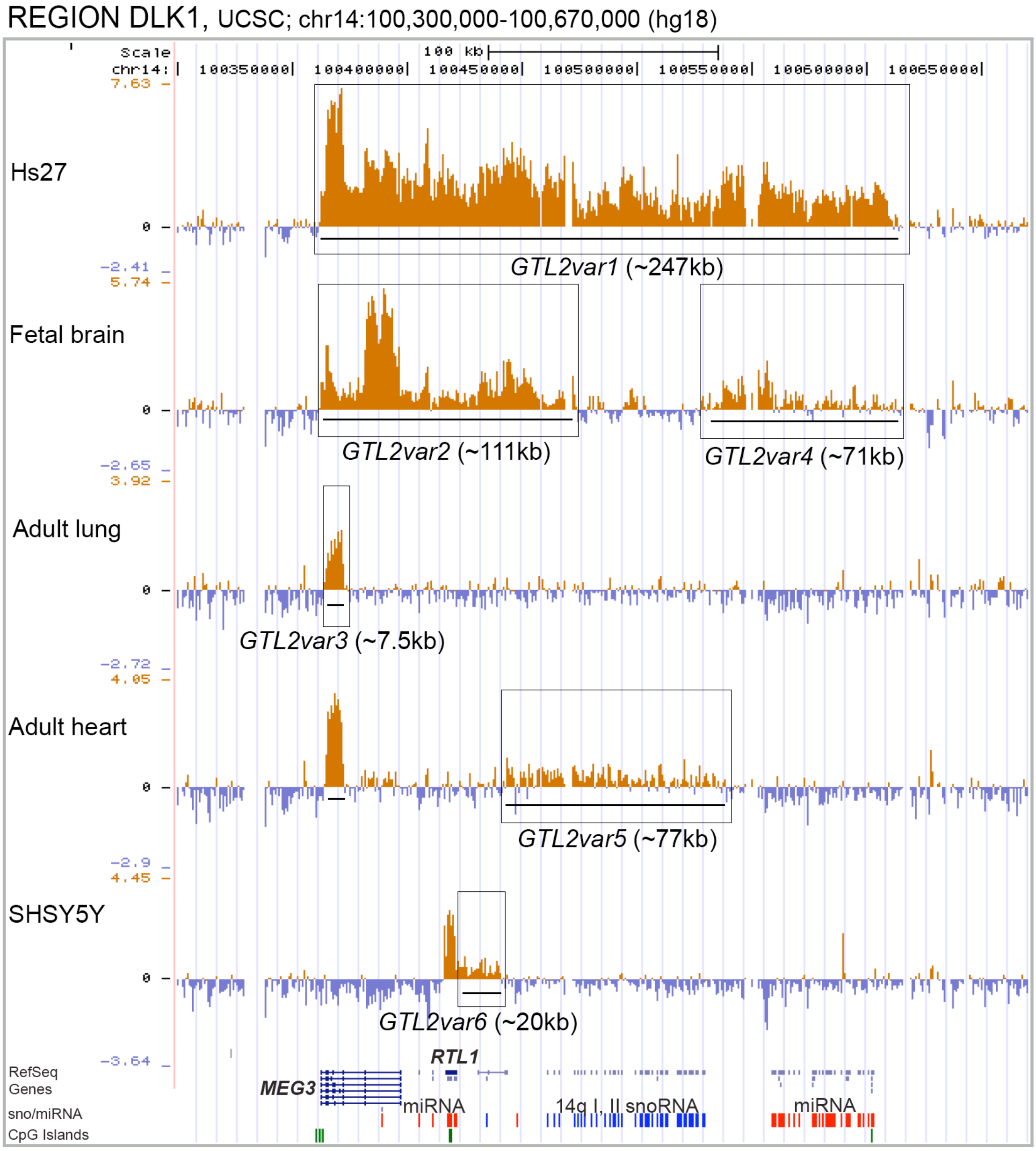

REGION DLK1, UCSC; chr14:100,300,000-100,670,000 (hg18)

Figure 14. Tissue specific expression of the *GTL2* macro ncRNA complex transcription unit. *GTL2* variants or overlapping macro ncRNA transcripts ranging in length from 7.5 to 247kb (marked by the black boxes) show tissue specific expression. Note positions of miRNA (red) and snoRNA clusters (blue). Black lines show expressed macro ncRNAs. Custom HIRTA tracks as well as RefSeq genes, sno/miRNA and CpG islands UCSC tracks are shown.

REGION DLK1, UCSC; chr14:100,350,000-100,380,000 (h18)

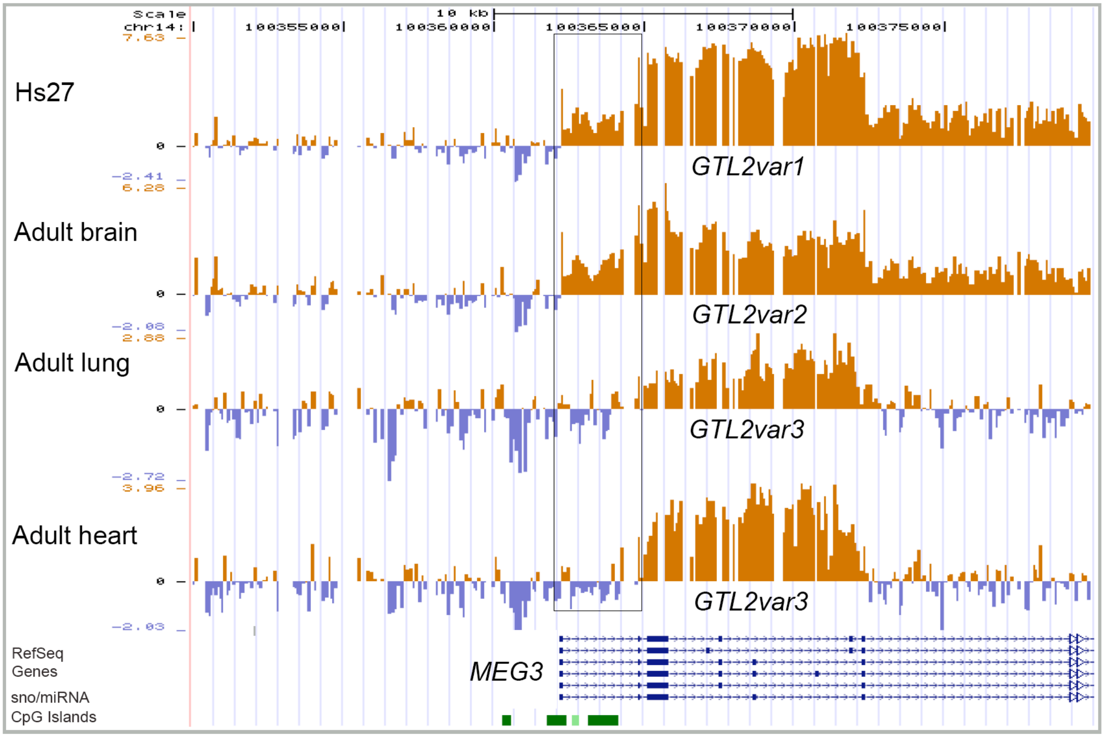


Figure 15. *GTL2* macro ncRNA long and short variants use different promoters. Long *GTL2* variants are present in Hs27 and adult brain, while short lung and heart *GTL2* variants use another promoter. Short *GTL2* transcript partially map to the previously annotated *MEG3 (GTL2)* ncRNA. Custom HIRTA tracks as well as RefSeq genes, sno/miRNA and CpG islands UCSC tracks are shown.

Four variants or transcripts overlapping *UBE3A-AS* macro ncRNA were expressed in representative tissues in lengths ranging from 96 to 420kb (Figure 16). The existence of *UBE3A-ASvar1* has been previously suggested in mouse where *Ube3a-as* has been mapped to ~1000kb (Landers et al., 2004) and here in human was visualized using HIRTA as about 420kb long. Two imprinted snoRNA clusters (*HBAII-85* or *SNORD116* and *HBAII-52* or *SNORD115*) have been mapped in this region previously (Cavaille et al., 2000). Interestingly, *UBE3A-ASvar2* that overlaps the HBII-85 and the 40kb shorter *UBE3A-ASvar3* RNAs, were ubiquitously expressed in all cells/tissues hybridized to HIRTA. The *UBE3A-ASvar4* overlapping to HBAII-52 in contrast had tissue specific expression and showed developmental upregulation (it was not expressed in human undifferentiated embryonic stem cells while it was lowly expressed in day 7 differentiated ES cells). High expression of HBAII-52 was restricted to fetal and adult brains that were also the same two tissues uniquely expressing 420kb *UBE3A-ASvar1*.

One of possible explanations for complexity of both *GTL2* and *UBE3A-AS* macro ncRNA regions could be based on the presence of overlapping micro and snoRNA clusters. Thus some of the mapped variants could represent primary miRNA or snoRNA transcripts.

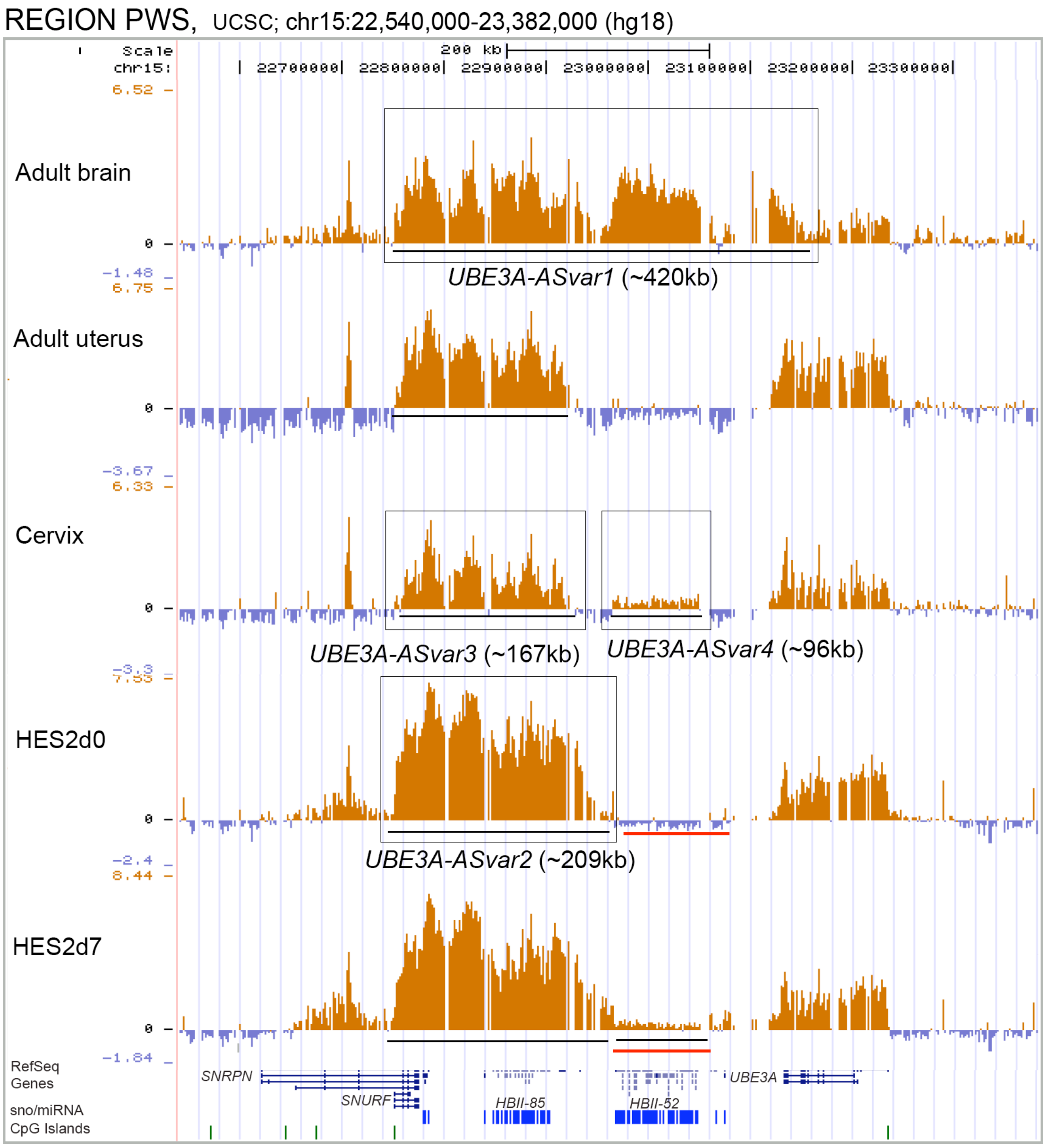


Figure 16. *UBE3A-AS* macro ncRNA tissue specific expression. Five cells/tissues (undifferentiated and differentiated embryonic stem cells, adult brain, uterus and cervix) were chosen in order to depict four different *UBE3-AS* variants or overlapping transcripts that range in length from 96 to 420kb (lengths depicted in brackets). The red line highlights developmental upregulation (in human ES cell system) of *UBE3A-ASvar4* that is corresponding to HBII-52 snoRNA cluster. Black boxes highlight macro ncRNA variants while black lines show expressed macro ncRNAs. Custom HIRTA tracks as well as RefSeq genes, sno/miRNA and CpG islands UCSC tracks are shown.

Tissue specific expression of 6 known and novel macro ncRNAs based on tiling array expression in 43 samples is depicted on Figure 17. Level of macro ncRNAs expression in different tissues was assessed by visual inspection of HIRTA hybridization profiles. Since tiling arrays have potential background and saturation limitations, tissue specific expression of macro ncRNAs is shown as two clear expression states: ON and OFF, where ON represents low, medium or high expression whereas OFF means no expression. The exact level of ON expression of each transcript can be further assessed by loading available .wig tracks for specific tissues showing ON expression into hg18 of UCSC browser and by searching macro ncRNA position.

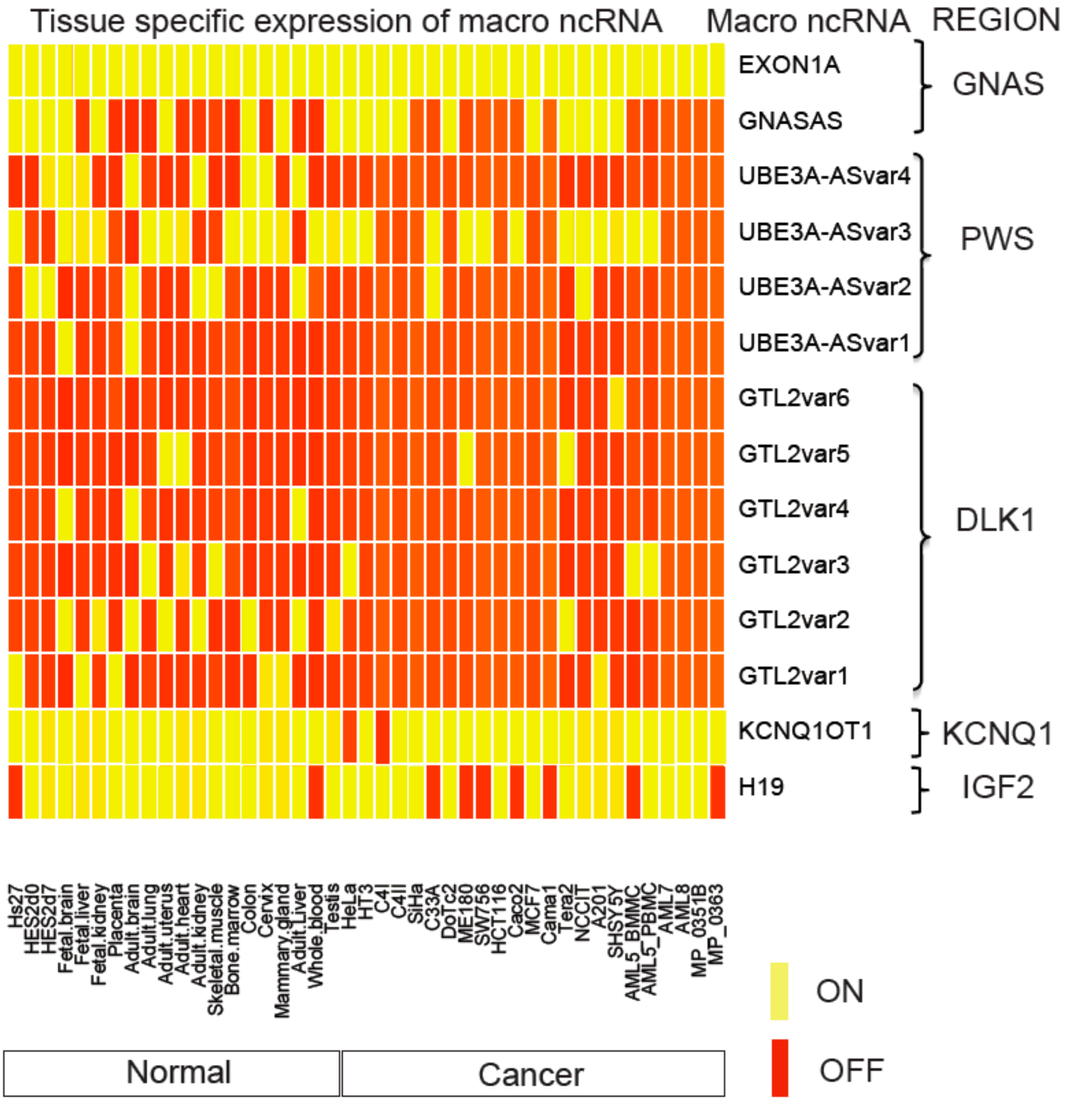


Figure 17. Tissue specific expression of well-studied macro ncRNAs in normal and cancer cells. Expression of known and novel variants macro ncRNAs from well known imprinted regions in 20 normal and 23 cancer samples is shown. Names of the HIRTA regions and macro ncRNAs that are previously defined in Table 9 and Table 13, respectively are shown on the left. 20 normal and 23 cancer cells/tissues/ patients are named below and grouped with corresponding normal and cancer white boxes. Yellow boxes; transcript expressed, Red boxes; transcript is not expressed.

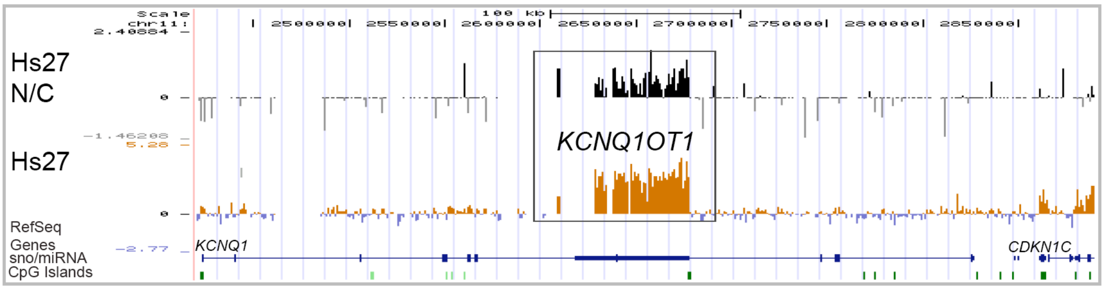
The *EXON1A*, *KCNQ1OT1* and *H19* macro ncRNAs are largely ubiquitously expressed (exceptions: *H19* is not expressed in fibroblasts and whole blood while *KCNQ1OT1* is downregulated in two cervical cancers lines). *GNAS1-AS* and most

GTL2 and *UBE3A-AS* variants have diverse tissue specific patern. *GTL2var6* is specific for neuroblastoma (SHSY5Y) while *UBE3A-AS* long variant (var1) is specifically expressed in brain (both fetal and adult) (Figure 17).

2. 2. 1. 2. Nuclear versus cytoplasmic localization of known macro ncRNAs

Nuclear versus cytoplasmic localization of six well-known macro ncRNAs was assessed by double cDNA hybridization on HIRTA in normal human fibroblasts cell line (Hs27) (Figure 12, section 2.1.6). *KCNQ1OT1* and *GTL2var1* macro ncRNAs were highly expressed in fibroblasts and were clearly nuclearly enriched by double HIRTA hybridization (Figure 18). Nuclear localization of *KCNQ1OT1* was also confirmed by RT-PCR and RT-qPCR and has been used as a positive control in further experiments testing localization of novel macro ncRNA candidates. *GNASAS* was lowly expressed and showed both nuclear and cytoplasmic localization while *EXON1A* localization was difficult to determine since this ncRNA differs just in the first exon (that corresponds to one HIRTA probe) from the overlapping protein coding gene *GNAS*. *H19* and *UBE3A-ASvar1* localizations could not be determined since they showed no expression in fibroblasts. Human tissues were not available for nuclear versus cytoplasmic study of macro ncRNA localization.

A REGION KCNQ1, UCSC; chr11:2,420,000-2,890,000 (hg18)



B REGION DLK1, UCSC; chr14:100,230,000-100,700,000 (hg18)

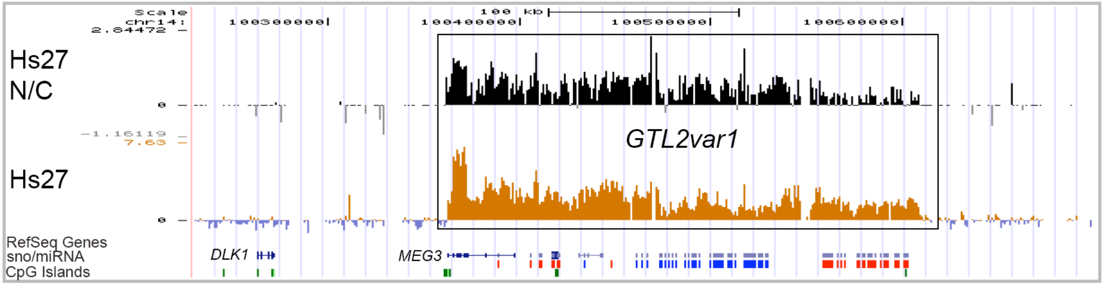


Figure 18. *KCNQ1OT1* and *GTL2var1* macro ncRNAs are nuclear-enriched. A. Expression of *KCNQ1OT1* ncRNA in fibroblasts is shown in orange while its nuclear enrichment is visible as black bars probe (enrichment in double nucleus and cytoplasmic hybridization) **B.** *GTL2var1* ncRNA is nuclearly enriched in fibroblasts. Tracks are displayed on UCSC as on the Figure 12B, section 2.1.6.

2. 2. 1. 3. 5'-3' slope as a feature of macro ncRNA expression by HIRTA

HIRTA mapping of known macro ncRNAs led to the observation of that tile expression intensity decreases through the macro ncRNA body in a 5' to 3' direction. This feature was typically observed for highly expressed and very long macro ncRNAs like *KCNQ1OT1*, *GNASAS*, *GTL2var1* and *UBE3A-ASvar1* (Figure 19) in different tissues and cell lines. SLOPE function (that assess the slope of a linear regression line through the data points, where the slope is defined as vertical distance divided by horizontal distance between any two points on the line showing rate of the change along the regression line) was used to calculate decreases in tile expression intensities through macro ncRNA bodies. Interestingly, positive/negative value of the calculated slope and the known strand orientation of the transcript showed an opposite correlation. If the calculated slope was negative, transcript was expressed from plus (+) DNA strand (strand defined by UCSC) while positive slope indicated transcription from minus (-) DNA strand. Thus, the decrease in the HIRTA probes intensity has a 5' to 3' direction and this feature was named the 5'-3' slope. This shows that the slope has potential in predicting macro ncRNAs transcription orientation and promoter location. Correlation between slope and strand was shown for four known macro ncRNAs showing low to moderate non-saturated expression, while *H19* ncRNA was showing high, saturated level of expression and did not show the slope and *EXON1A* could not be tested since it highly overlaps with the *GNAS* protein coding gene. Wheather this correlation is a universal feature of long ncRNAs will be further tested in the section 2.5.3, where prediction of transcriptional orientation, based on 5'-3' slope and other genetic/epigenetic features, for novel macro ncRNAs will be described.

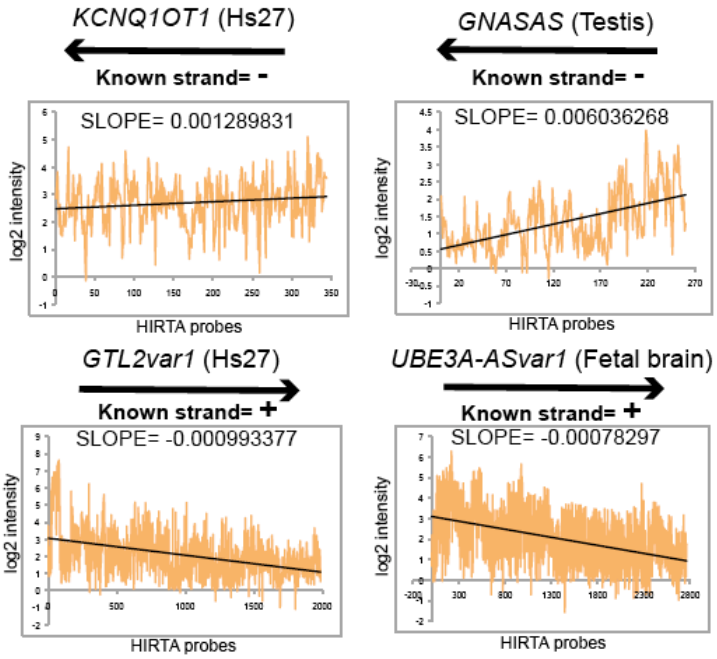


Figure 19. Reverse correlation between 5'-3' slope and transcriptional orientation. Log_2 tile expression intensities through the length of macro ncRNA from cells/tissues highly transcribing macro ncRNAs are ploted to the HIRTA probes (number of tiles through the macro ncRNA region) and shown in orange. Trendline corresponding to signal intensities is depicted in black. *KCNQ1OT1* and *GNASAS* macro ncRNAs known to be transcribed from the - DNA strand show positive SLOPE values while *GTL2var1* and *UBE3A-ASvar1* macro ncRNas are known to be transcribed form the + DNA strand and show negative SLOPE values.

2. 2. 2. Novel macro ncRNA candidates in well-known imprinted gene regions

In addition to the six macro ncRNAs presented on Figure 13, in six well-studied imprinted clusters there is a number of less-studied macro ncRNAs (*AIRN* in IGF2R cluster, *91H* and *IGF2AS* in IGF2 gene cluster, *LOC650368* and *LOC100133545* in KCNQ1OT1 cluster, *NCRNA00239* in DLK1 and *PWRN1* in PWS imprinted gene cluster (described in introduction or annotated by RefSeq genes)). HIRTA also detected expression of less-studied macro ncRNAs except for *AIRN* ncRNA, which was not detected in the tissues and cell lines that were used (data not shown).

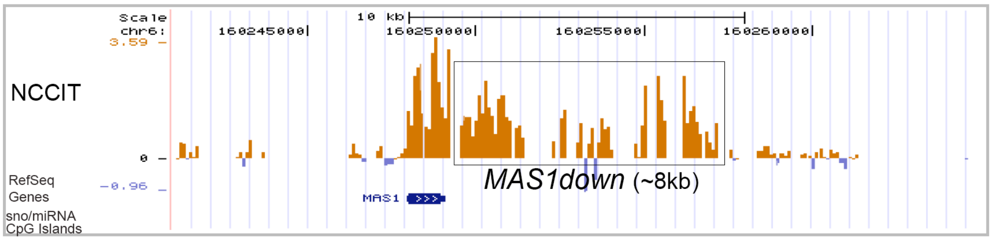
28 novel macro ncRNA candidates that range from 3.9 to 666kb in length (Table 14) were mapped in six well-known human imprinted gene regions based on criteria described in section 2.1.5. Positions of the candidate macro ncRNAs, their lengths, % of positive probes covered and their positions corresponding to annotated genes are shown.

HIRTA REGION	Chr.	NAME of macro ncRNA	POSITION (hg18)	Length (kb)	COV (%)	Position to annotated PC genes
IGF2R	6	***SLC22A2up***	chr6:160616598-160620501	3.9	100	IG
		MAS1down	chr6:160249097-160257182	8	95.9	3'/OV
		MAS1ov	chr6:160171198-160260777	89.6	89.9	OV
IGF2	11	***H19down***	chr11:1930616-1972982	42.4	94	3'/OV
		H19up	chr11:2019588-2048590	29	100	IG
		ASCL2ov	chr11:2167733-2255607	87.8	89.5	OV
		ASCL2up	chr11:2261646-2270383	8.7	100	3'/OV
		TSPAN32down1	chr11:2296006-2355798	59.8	100	IG
KCNQ1	11	***ZNF195down1***	chr11:3225444-3318834	93.4	87.6	IG
		ZNF195down2	chr11:3321897-3329926	8	100	IG
		ZNF195down3	chr11:3311198-3317597	6.4	92.3	IG
		ZNF195down4	chr11:3303912-3308999	5	97	IG
		ZNF195up1	chr11:3379056- 3421678	42.6	86.5	IG
		ZNF195up2	chr11:3463567-3560541	97	89.1	IG
DLK1	14	***BEGAINup***	chr14:100105884-100123177	17.2	100	5'/OV
		PPP2R5Cup1	chr14:101163838-101268901	105	86.6	OV
		PPP2R5Cup2	chr14:101262293-101297952	35.6	100	OV
PWS	15	***NIPA1up1***	chr15:20647107-20666695	19.5	92	IG
		WHAMML1up	chr15:20759798-20797989	38.2	94.7	5'/OV
		WHAMML1up1	chr15:20800834-20804904	4	100	IG
		SNRPNup1	chr15:21621941-21834998	213	87.6	IG
		SNRPNup2	chr15:21887504-22553483	666	98.2	IG
		GABRB3down1	chr15:24040777-24097047	56.3	75	IG
		GABRB3down2	chr15:24191839-24252220	60.4	93	IG
GNAS	20	***APCDD1Lup1***	chr20:56523841-56628352	104.5	84.3	OV
		APCDD1Lup2	chr20:56633435-56655122	21.7	85.4	IG
		ZNF831up1	chr20:57154647-57199470	44.8	99.3	5'/OV
		ZNF831up2	chr20:57070823-57103486	32.6	91	IG

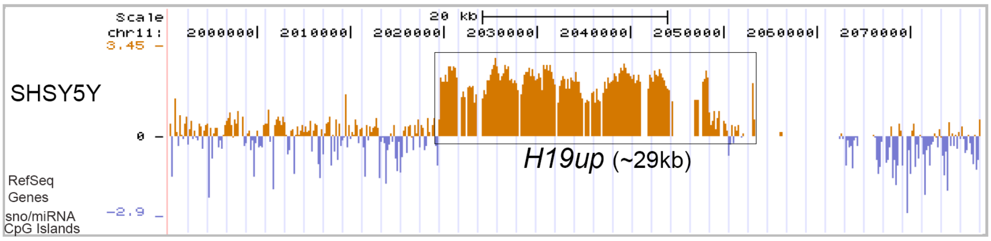
Table 14. 28 novel macro ncRNA candidates are mapped in six well-known imprinted gene clusters in a total of 43 tissues/cell lines. The columns are as described in Table 13., section 2.2.1.1.

To illustate these candidates, six examples of novel macro ncRNAs: *MAS1down* from the IGF2R imprinted gene region, *H19up* from the IGF2 imprinted gene region, *ASCL2up* from the KCNQ1 region, *PPP2R5Cup2* from the DLK1 region, *WHAMML1up* from the PWS region and *ZNF831up2* from GNAS region are shown (Figure 20). Relative to RefSeq and UCSC annotated genes, 17 macro ncRNAs do not overlap, 5 overlap with annotated genes, 3 are located next to the gene from its' 5' end and 3 are located adjacent to the 3'end of a gene. Transcripts located next to the genes could be overlapping ncRNAs, novel 5' exons or 3'UTRs. For example *PPP2R5Cup2* and *WHAMML1up* could be novel 5' exons or overlapping ncRNA while *MAS1down* and *ASCL2up* could be novel 3'UTRs or novel overlapping macro ncRNAs (Figure 20). From 28 mapped transcripts *H19up* is the only that showed very slight upregulation through human embryonic cell differentiation ($\log_2$=1.5 in HESd0 and $\log_2$=2.1 in HES2d7). Eight novel macro ncRNA candidates have CpG islands that may be their promoters. Interestingly, five of the candidates are overlapping with annotated Expression Sequenced Tags (ESTs) that do not have coding potential in UCSC description while *PPP2R5Cup2* that is exclusively expressed in testis partially overlaps *NCRNA00239* ncRNA. Thus, these evidences straighten the hypothesis that these transcripts are indeed non-protein coding RNAs.

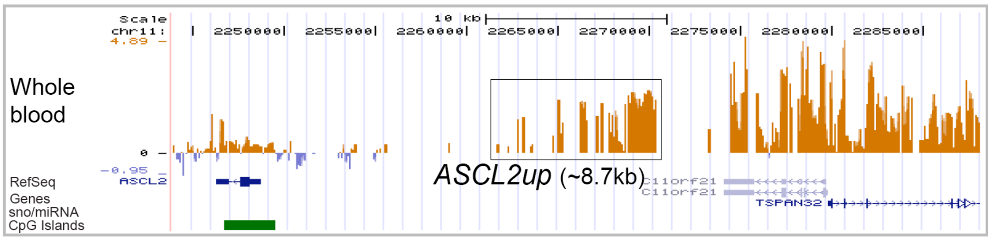
REG IGF2R, UCSC; chr6:160,241,000-160,265,000 (hg18)



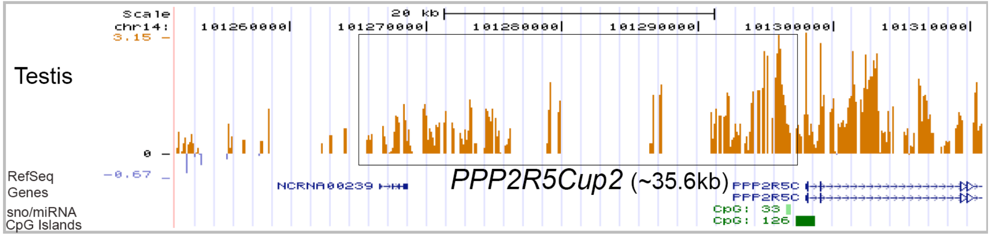
REG IGF2, UCSC; chr11:2019588-2048590 (hg18)



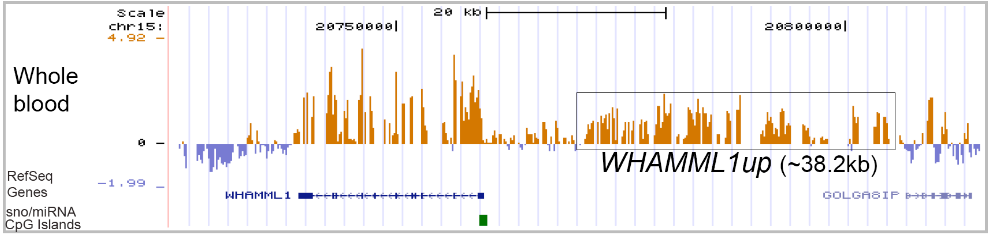
REG KCNQ1, UCSC; chr11:2261646-2270383 (hg18)



REG DLK1, UCSC; chr14:101262293-101297952 (hg18)



REG PWS, UCSC; chr15:20,725,000-20,815,000 (hg18)



REG GNAS, UCSC; chr20:57,038,000-57,120,000 (hg18)

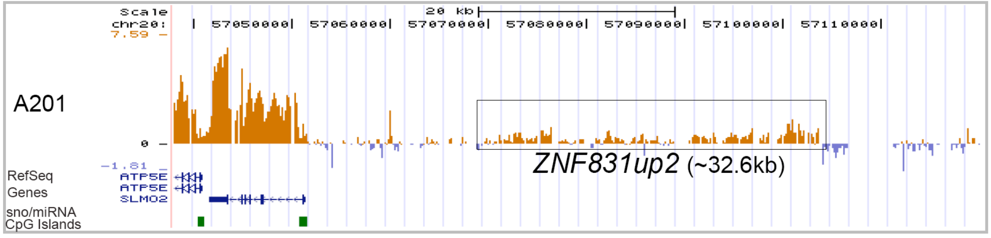


Figure 20. Typical examples of novel macro ncRNA candidates from 6 well-known imprinted gene regions. Diverse examples of macro ncRNA candidates expressed in

different cell lines/tissues from six well-known imprinted regions are noted by black boxes. *PPP2R5Cup2* is an example of a group of macro ncRNA candidates rich in repetitive regions. Name of the HIRTA region and viewed UCSC screen position in hg18 is shown on the top of each gray box. The name and the lengths of the novel transcripts are showed below the boxes. Custom HIRTA tracks as well as RefSeq genes, sno/miRNA and CpG islands UCSC tracks are shown.

Expression of novel candidate macro ncRNAs mapped by HIRTA in 6 well-known imprinted gene regions was analysed using the described ON/OFF system showing if ncRNA candidates are expressed or not expressed in normal and cancer cells or tissues (Figure 21).

Among these 28 mapped novel ncRNAs there were no ubiquitously expressed transcripts. Five transcripts (*MAS1ov*, *PPP2R5Cup2*, *GABRB3down1*, *GABRB3down2* and *ZNF831up2*) were expressed in a single cell line/tissue. 6 transcripts were expressed exclusively from cancer cell lines/ tissues (*MAS1ov*, *ZNF195down1, 2, 3, 4* and *PPP2R5Cup1*). Examples of cancer specific and cancer dowregulated transcripts will be further shown in section 2.6. Fetal liver and placenta do not express any of the 28 macro ncRNA candidates althow they do express other macro ncRNAs (Figure 17, section 2.2.1.1.), which is the same for two cervical cancer cell lines (DoTc2 and ME180).

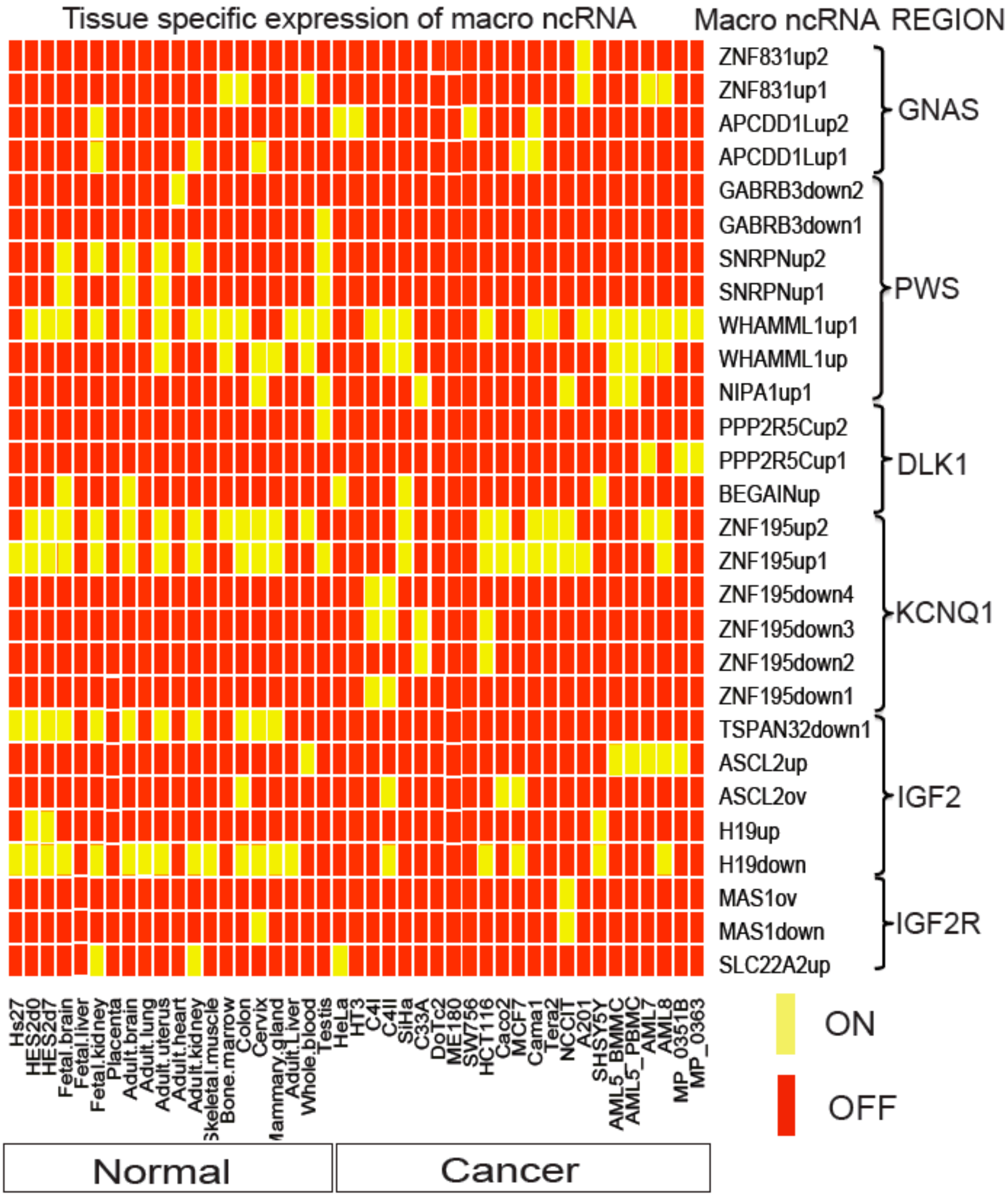
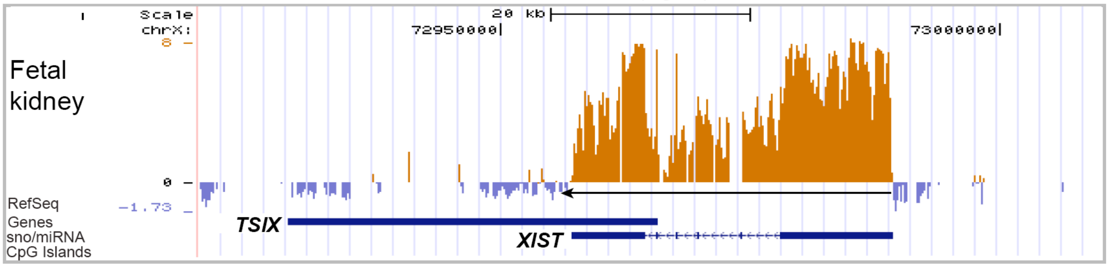


Figure 21. Tissue specific expression of macro ncRNA candidates from the 6 well-known imprinted gene regions. Expression of 28 macro ncRNA candidates in 43 normal and cancer cells/tissues. Description of the figure is as for Figure 17., section 2.2.1.1. Yellow; ON- expressed transcripts, Red; OFF-no expression.

2. 3. Human *XIST* region

The XIC (X Inactivation Centre) region is included here as an example of region containing a wel-studied *Xist*/XIST macro ncRNA. The XIC was genetically identified as a locus on the X chromosome that is required and sufficient for the X chromosome inactivation (Russell, 1963). However the human XIC has several differences compared to mouse XIC. The mouse XIC region contains *Xist* macro ncRNA that is expressed exclusively from inactivated X chromosome in female, shows imprinted expression in extraembryonic tissues and has silencing function in X-inactivation (reviewed in (Wutz and Gribnau, 2007)). The human *XIST* ncRNA is also expressed exclusively from inactivated X chromosome in female cells containing at least two X chromosomes and has a silencing function in X-inactivation but it does not show imprinted expression in extraembryonic tissues (Brown et al., 1991; Migeon, 2002). Expression of the *XIST* macro ncRNA was shown by HIRTA single hybridizations in a number of human cells/tissues (one example of high *XIST* expression is shown on Figure 22A). The undifferentiated human embryonic stem cell line HES2 did not express *XIST* while after seven days of differentiation *XIST* was expressed in the same cell line (Figure 22B).

A REG XIST, UCSC; chrX:72,920,000-73,010,000 (hg18)



B *XIST* is developmentally regulated in HES2 cells

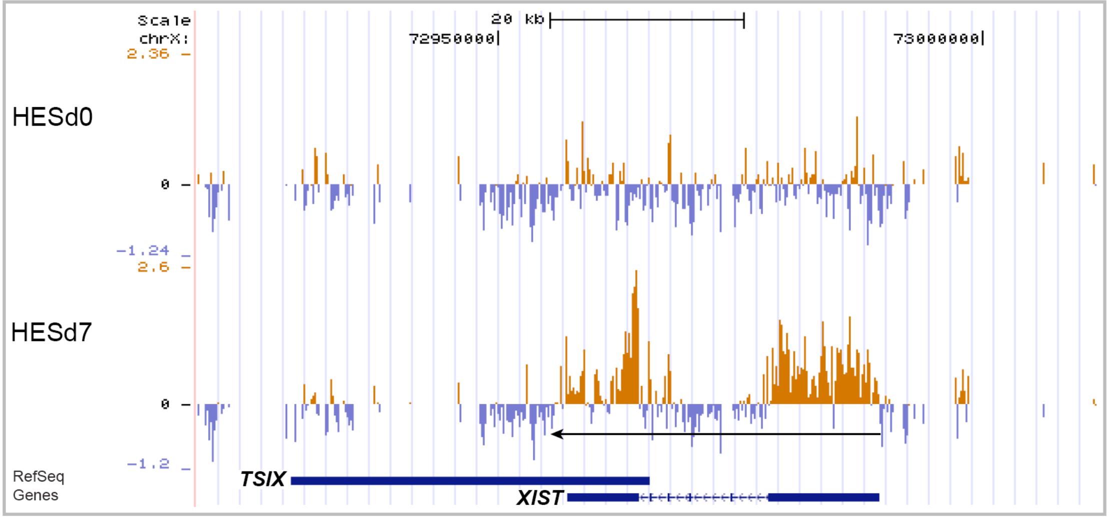


Figure 22. *Xist* macro ncRNA is developmentally regulated in HES2 cell line. A. HIRTA expression of *XIST* macro ncRNA corresponds to the annotated gene. *TSIX* ncRNA is not expressed. **B.** *XIST* macro ncRNA was not expressed in HESd0 cells but showed expression after 7 days of differentiation in HES2d7 cell line. Custom HIRTA tracks as well as RefSeq genes, sno/miRNA and CpG islands UCSC tracks are shown.

Mouse *Tsix* and human *TSIX* macro ncRNA show different expression patterns and possibly different roles between mouse and human (Chang and Brown, 2010; Migeon et al., 2002). The mouse *Tsix* is expressed early in development and silenced upon differentiation (Debrand et al., 1999; Lee and Lu, 1999). The *Tsix* completely overlaps *Xist* ncRNA, and was found to have a role in regulating *Xist* expression and in partitioning chromatin domains within the mouse Xic (Navarro et al., 2009). The human *TSIX* macro ncRNA just partially overlapps *XIST* and is expressed in chorionic villus cells, embryonal bodies and the human embryonal carcinoma N-Tera2D1 cell line (Chow et al., 2003; Migeon et al., 2002). The role of human *TSIX* is not fully elucidated but it was proposed that the *TSIX* is not a regulator of *XIST* in humans (Chang and Brown, 2010). Using HIRTA analysis *TSIX* was not expressed in undifferentiated and differentiated HES2 cells (Figure 22B) or in two tested embryonal carcinoma cell lines (Tera2 and NCCIT) as well as any other tested cells/tissues (data not shown). Surprisingly, the N-Tera2D1 cell line that was previously found to express *TSIX* was derived by cloning the NTERA-2 cell line that was itself established from the nude mouse xenograft of Tera-2 (http://www.lgcstandards-atcc.org/), the cell line that was hybridized to HIRTA and did not show *TSIX* expression.

The *Jpx* that is also known as *Enox* ncRNA is the first gene expressed upstream of the *Xist* ncRNA and it was found to partially escape X-inactivation in mouse (Johnston et al., 2002). Similarly, human *JPX/ENOX* was also found to be expressed from both inactive and active X chromosomes (Chow et al., 2003). In the UCSC RefSeq genes *JPX/ENOX* ncRNA is annotated as *NCRNA00182* that is located about 90kb upstream of *XIST. NCRNA00183 ncRNA* is another ncRNA that overlapps *NCRNA00182* and was recently annotated by UCSC RefSeq genes. Ubiquitous expression in all tested normal and cancer tissues of both overlapping *NCRNA00182* and *NCRNA00183* macro ncRNAs, was found by HIRTA (typical expression pattern in one of the tissues is shown on Figure 23). These ncRNAs were not developmentaly regulated after seven days of differentiation in ES cell system (data not shown). Both *NCRNA00182* and *NCRNA00183* were rich in repeats and *NCRNA00183* overlapped with three annotated miRNAs (Figure 23).

REG XIST, UCSC; chrX:72,910,000-73,450,000 (hg18)

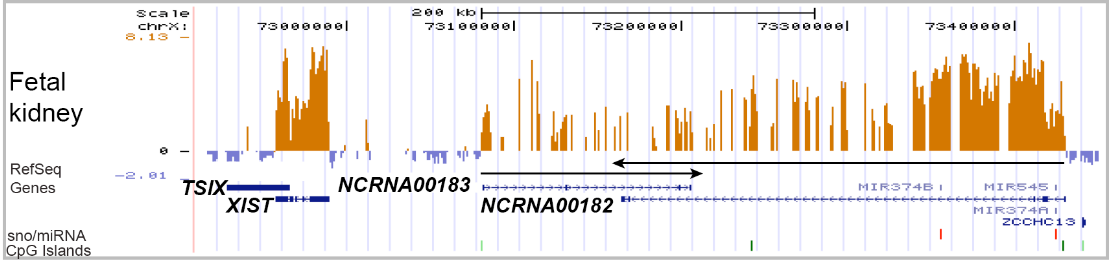


Figure 23. *NCRNA00138* and *NCRNA00182* macro ncRNAs are expressed from XIC region. HIRTA detects two overlapping macro ncRNAs (*NCRNA00182* and *NCRNA00183*) that are positioned more than 90kb upstream of *XIST*. Custom HIRTA tracks as well as RefSeq genes, sno/miRNA and CpG islands UCSC tracks are shown.

2. 4. Macro ncRNAs in 26 less-studied imprinted gene regions

2. 4. 1. Mapping of known and novel macro ncRNAs in less-studied regions

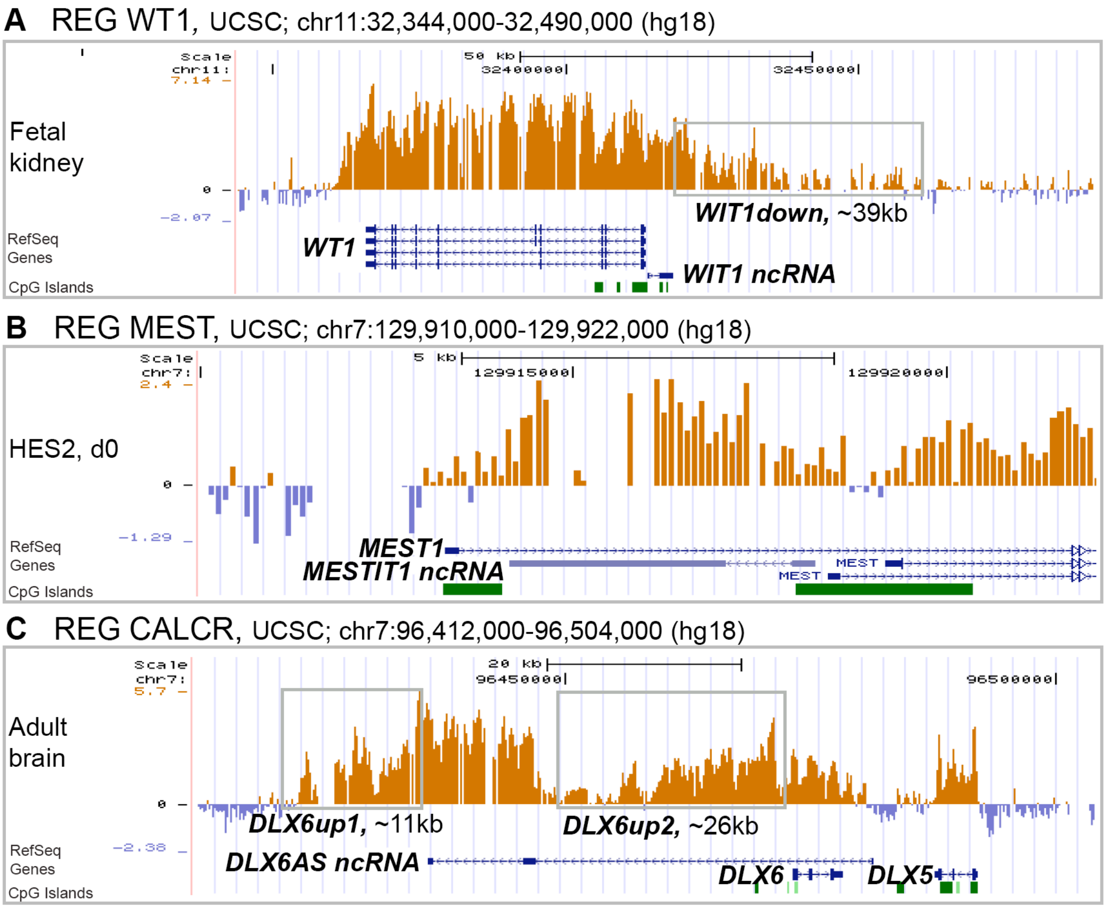
In the previous sections six well-studied imprinted gene clusters and the *XIST* genomic region were shown to express already known or novel macro ncRNAs (sections 2.2.; 2.3.). To examine whether all human imprinted gene clusters express macro ncRNAs, we analysed 26 less-well studied imprinted gene regions spotted on HIRTA (less-studied regions were defined in Table 9, section 2.1.1.). Mapping of novel macro ncRNAs in 43 cells/tissues/patients in 26 less-studied regions was done on the basis of the already presented criteria (section 2.1.5.). 11/26 tested regions had been shown previously to express macro ncRNAs (e.g. WT1 region: *WIT1* ncRNA, MEST region: *MESTIT1* and *MIT1* ncRNAs, PLAGL1: *HYMAI* ncRNA, CALCR region: *DLX6AS* ncRNA) (introduced in Table 6A and Table 6B, section 1.2.5.2.).

HIRTA Chip confirmed expression of the paternally expressed, 2.4kb long *WIT1* ncRNA although the size of this ncRNA could not be precisely mapped since a new ~39kb long transcript overlapping *WIT1* or presenting novel variant of *WIT1*, was found (*WIT1down*). This new 39kb transcript showed specifically high expression in fetal kidney, similar to *WIT1* (Figure 24A).

MESTIT1 ncRNA is 3.2kb long, paternally expressed and overlaps the *MEST* gene (Li et al., 2002). Expresion of this transcript was confirmed by HIRTA and showed tissue specific expression (Figure 24B). *MIT1,* from the same MEST imprinted gene region was previously identified as a ncRNA mapping to intron 20 of *COPG2* and having an antisense orientation to this protein-coding gene (Yamasaki et al., 2000). Since the reference sequence, from the hg18 version of the UCSC human genome that was used for HIRTA design, has a gap of about 80kb including 10 previously mapped *COPG2* exons, *MIT1* expression could not be confirmed.

HYMAI is a paternally expressed 5kb long ncRNA previously shown to overlap the *PLAGL1* protein-coding gene in the sense orientation. HIRTA showed high expression signals over the known HYMAI region, but since the *PLAGL1* gene is simultaneously highly expressed and shows high intronic signals in the region overlapping *HYMAI*, the expression of *HYMAI* itself could not be independently elucidated.

DLX6AS ncRNA expression was detected from the CALCR imprinted gene region by HIRTA. But simultaneously, two novel macro ncRNA candidates could be mapped in the DLX6AS region: the ~11kb long *DLX6up1* positioned upstream of *DLXAS* and the ~26kb long *DLX6up2* ncRNA overlapping *DLX6AS*. These results indicate, DLX6 is a region with potentially highly complex overlapping transcriptional units that are still not fully elucidatet (Figure 24C).


A REG WT1, UCSC; chr11:32,344,000-32,490,000 (hg18)

B REG MEST, UCSC; chr7:129,910,000-129,922,000 (hg18)

C REG CALCR, UCSC; chr7:96,412,000-96,504,000 (hg18)

Figure 24. Examples of known macro ncRNAs from 26 less studied imprinted gene regions. A. 2.4kb long known *WIT1* macro ncRNA positioned downstream of *WT1* protein-coding gene is expressed in fetal kidney. *WIT1down* is a 39kb long macro ncRNA candidate. **B.** In the MEST region the *MESTIT1* macro ncRNA is expressed in undifferentiated ES cells. **C.** The DLX6 region is part of the CALCR imprinted gene region and may contain a number of overlapping transcripts (e.g. *DLX6up1*, *DLX6AS, DLX6up2, DLX6*). Custom HIRTA tracks as well as RefSeq genes, sno/miRNA and CpG islands UCSC tracks are shown.

Mapping of novel macro ncRNAs in 26 less-studied regions revealed 73 novel macro ncRNA candidates in 43 tested cell lines/tissues (Table 15). Overall positions in the human genome, lengths, coverage and positions in accordance to protein coding genes were mapped. Novel macro ncRNA transcripts that were identified from were in between 6 and 460kb in length. Five of these candidates are found to be between 1kb and 10kb, 51 between 10 and 100kb and 17 more than 100kb in length. 19 of these transcripts have potential CpG island promoters on one of the mapped ends, and 13 of them are partially overlapped with ESTs annotated with RefSeq genes where 10 of these ESTs are showing non-protein coding potential according to the UCSC description.

HIRTA REGION	Chr.	NAME of macro ncRNA	POSITION (hg18)	Length (kb)	COV (%)	Position to annotated PC genes
TRP73	1	***LRRC47down***	chr1:3679188- 3686484	7.1	99	IG
DIRAS3	1	***GADD45Aup***	chr1:67842948-67867981	25	92.4	IG
		GPR177up	chr1:68530492-68571923	41.4	97.4	IG
		RPE65ov	chr1:68626430-68712423	85.9	100	OV
		RPE65down1	chr1:68484480-68624258	139.7	96.9	3'/OV
		RPE65down2	chr1:68530492-68571923	41.4	89.6	IG
COMMD1	2	***B3GNT2down1***	chr2:62376964-62498341	121.3	99.2	IG
		B3GNT2down2	chr2:62305370-62373270	67.9	93	3'/OV
		FAM161Adown	chr2:61794031-61845495	51.4	83.1	IG
NAP1L5	4	***TIGD2down***	chr4:90255075-90296601	41.5	91.9	3'/OV
PLAGL1	6	***PHACTR2ov***	chr6:143899928-144040795	140.8	98	OV
GRB10	7	***COBLdown1***	chr7:50948084-51023474	75.4	87.9	IG
		COBLdown2	chr7:50902101-50930499	28.4	98.7	IG
CALCR	7	***DLX6up1***	chr7:96422607-96434058	11.4	100	IG
		DLX6up2	chr7:96446866-96472851	25.9	97.8	IG
		COL1A2up	chr7:93758984-93858857	99.8	97.7	IG
MEST	7	***KLF14up1***	chr7:130121904-130170014	41.8	89.7	IG
		KLF14up2	chr7:130186541-130248421	61.8	99.7	IG
		KLF14up3	chr7:130248421-130279459	31	100	IG
KCNK9	8	***PEG13***	chr8:141173723-141180357	6.6	100	IN
		PTK2up	chr8:142107593-142163814	56.2	92	IG
		KIAA1126up	chr8:142333407-142388337	54.9	99.2	5'/OV
SFMBT2	10	***PFKFB3down***	chr10:6317511-6509111	191.6	99.9	3'/OV
		SFMBT2down1	chr10:6866778-6926405	59.6	99.6	IG
		SFMBT2down2	chr10:6820088-6866778	46.7	99.3	IG
		SFMBT2down3	chr10:7185016-7201477	16.5	87.6	IG
		SFMBT2down4	chr10:6915718-7015793	100	95.9	IG
		SFMBT2down5	chr10:6929387-6964245	34.8	99.4	IG
		GATA3down	chr10:8452227-8487562	35.3	73.8	IG
INPP5F	10	***BAG3down***	chr10:121427002-121475599	48.6	92.6	3'/OV
		SEC23IPdown1	chr10:121922168-122083651	161.5	90.7	IG
		SEC23IPdown2	chr10:122068853-122096845	28	99.3	IG
		SEC23IPdown3	chr10:121922168-122085936	163.8	90.6	IG
BAZ2	11	***OR556B4up1***	chr11:6000641-6087607	87	98.5	OV
		OR56A1down	chr11:6020186-6041034	20.8	99.3	IG
		PRKCDBPup	chr11:6298316-6323682	25.4	100	5'/OV
		PRKCDBPdown	chr11:6272107-6296751	24.6	93.9	3'/OV
		HPXdown	chr11:6400477-6409017	8.5	100	3'/OV
		ZNF215up	chr11:6724246-6904230	180	95.3	OV
		RBMXL2down	chr11:7068955-7116437	47.5	93	3'/OV
		OR5P2ov	chr11:7684517-7876797	192.3	99	OV
AMPD3	11	***AMPD3up***	chr11:10325169-10428800	103.6	78.5	5'/OV
		LYVE1ov	chr11:10519395-10551214	16.6	98.8	OV
		MRVI1up	chr11:10672111-10707462	35.5	83.8	5'/OV
WT1	11	***WIT1down***	chr11:32418196-32457422	39.2	96.3	3'/OV
SDHD	11	***SDHDdown***	chr11:111471727-111514225	42.5	97.9	3'/OV
		PTSdown	chr11:111645888-111756700	120	97.4	3'/OV
SLC38A4	12	***SLC38A4down1***	chr12:45063900-45174483	110.6	83.52	IG
		SLC38A4down2	chr12:45063900-45527065	463.2	94.3	OV
		SLC38A4up	chr12:45563328-45661331	98	98	IG
DCN	12	***EPYCov***	chr12:89855372-89888024	32.6	83.8	OV
		DCNup1	chr12:90400030-90416192	16.1	87.1	IG
		DCNup2	chr12:90469318-90484664	15.3	93.2	IG
		DCNup3	chr12:90538500-90546595	8	92.7	IG
HTR2A	13	***LRCH1up1***	chr13:45925127-45939792	14.6	92.7	IG
		LRCH1up2	chr13:45947398-46015616	68.2	87.6	5'/OV
		HTR2Aup1	chr13:46364433-46472289	107.8	85.5	5'/OV
		HTR2Aup2	chr13:46364433-46405913	41.4	90.8	OV
		HTR2Ain	chr13:46317545-46333910	16.4	94.1	IN
GATM	15	***SQRDLdown***	chr15:43784646-43889432	104.8	74.4	IG
		C15orf43up	chr15:42962890-42998200	35.3	93.4	IG
RASGRF1	15	***ADAMTS7down***	chr15:76830029-76836494	6.4	96.7	IG
		TMED3down	chr15:77402383-77490747	88.4	98.5	3'/OV
		KIAA1024up	chr15:77492611-77511913	19.3	89	5'/OV
		CHRNB4up	chr15:76739929-76826996	87	88.6	5'/OV
IMPACT	18	***ZNF521up***	chr18:20555579-20779652	224	94.4	IG
ZIM	19	***ZNF71down1***	chr19:61824857-61835672	10.8	100	3'/OV
		ZNF71down2	chr19:61835974-61866765	30.8	91.3	3'/OV
HM13	20	***ID1up1***	chr20:29623883-29635995	12.1	91.7	IG
		ID1up2	chr20:29639171-29654648	15.4	95.6	5'/OV
NNAT	20	***BLCAPov***	chr20:35547467-35583161	35.7	99.5	3'/OV
L3MBTL	20	***L3MBTLup***	chr20:41552855-41572354	19.5	100	IG
		TOX2up	chr20:41817556-41913383	95.8	72.7	IG

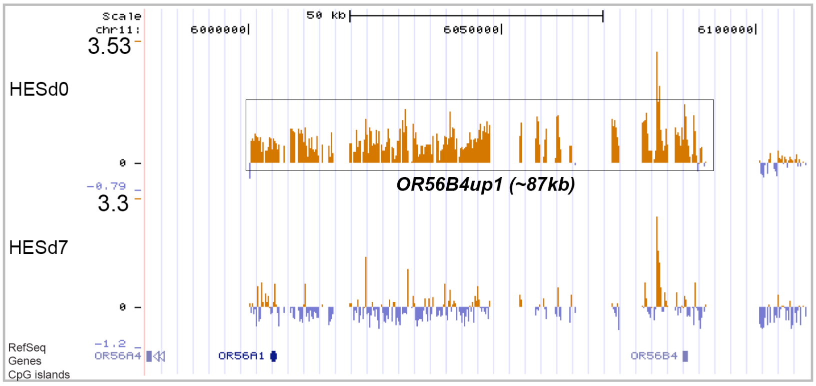
Table 15. 73 novel macro ncRNA candidates have been mapped in 26 less-well studied imprinted gene regions in 43 cells/tissues/patients by HIRTA. The columns are as described in Table 13., section 2.2.1.1.

2. 4. 2. Developmental regulation of macro ncRNAs in less-studied regions

We tested developmental regulation of macro ncRNA candidates by differentiating human ES cells and testing candidate gene expression using HIRTA single hybridizations before and after differentiation. Among 73 mapped macro ncRNA candidates in 26 less-well studied imprinted gene regions, 6/22 that were expressed in ES cells showed some degree of developmental regulation.

The *OR56B4up1* ncRNA candidate was expressed from the BAZ2 region in undifferentiated ES cells while after seven days of differentiation expression of this gene was not detected (Figure 25A). Interestingly, another macro ncRNA candidate that also overlaps olfactory receptor genes, *OR5P2ov* was upregulated after seven days of ES cells differentiation (Figure 25B). *SLC38A4down2* is ~463kb long ncRNA expressed from SLC38A4 region and was downregulated to a lesser extent during ES cells differentiation. Three other candidates (*PKCDBPup, ZNF521up, ID1up1*) that show downregulation during HES2 differentiation were lowly expressed in undifferentiated ES cells and by day 7 of differentiation of HES2 they did not show expression (data not shown).

A REG BAZ2, UCSC; chr11:5,980,000-6,110,000 (hg18)



B REG BAZ2, UCSC; chr11:7,669,000-7,950,000 (hg18)

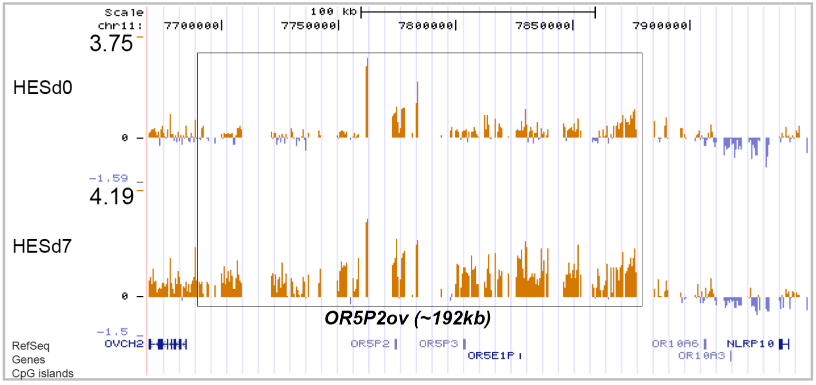


Figure 25. Comparison of novel macro ncRNAs expression before and after day 7 of human embryonal stem cells differentiation. A. The *OR56B4up1* macro ncRNA expressed from the BAZ2 imprinted gene region is downregulated upon differentiation of HES2 cell line. **B.** *OR5P2ov* shows upregulation upon ES cells differentiation. Both *OR56B4up1* and *OR5P2ov* are overlapping olfactory receptor genes. Custom HIRTA tracks as well as RefSeq genes, sno/miRNA and CpG islands UCSC tracks are shown.

2. 4. 3. Nuclear versus cytoplasmic localization of novel macro ncRNAs in less-studied regions

Nuclear versus cytoplasmic localization of macro ncRNA candidates in less-studied imprinted gene regions was first tested by the double cDNA hybridization procedure in normal human fibroblasts cell line (Hs27) (section 2.1.6). 14 out of 73 transcripts showed expression in Hs27 cell line. For 5 macro ncRNAs (*KLF14up2, KLF14up3, OR5P2ov, ADAMTS7down, TMED3down*) double HIRTA hybridization showed nuclear enrichment, 7 (*TIGD2down, PEG13, KIAA1126ov, PRKCDBPup, PTSdown, KIAA1024up, BLCAPov*) were found to be both nuclear and cytoplasmic and one appeared to be cytoplasmic enriched (*LRRC47down*). To illustrate these findings, examples of double cDNA hybridizations on HIRTA for novel macro ncRNAs will be shown under section 2.4.5., where ten novel macro ncRNAs will be characterized using different techniques.

Further, *SLC38A4down2* and *ADAMTS7down* macro ncRNAs nuclear enrichment relative to the *RPLPO* housekeeping gene was shown using quantitative RT-PCR (qRT-PCR) (Figure 25). RNA from nuclear (N) and cytoplasmic (C) fractions of the Hs27 cell line and total Hs27 RNA (T) were tested by primers specific for these macro ncRNA candidates. The *H19* macro ncRNA was used as a control as it was previously found in both nucleus and cytoplasm (Brannan et al., 1990). *H19* expression showed a N/C ratio of 1.5/1 as expected. Similarly, the *GAPDH* housekeeping gene N/C ratio was 0.8/1. Two primer pairs were used to test localization of *KCNQ1OT1*, which was used as a positive control known to be nuclear enriched (Murakami et al., 2007). *KCNQ1OT1* was entirely found in nuclear fraction of the Hs27 fibroblast cell line (with ratios N/C=208/1 and N/C=142/1 depending of the primer pair), while *SLC38A4down2* and *ADAMTS7down* were present in both the nuclear and cytoplasmic fractions, but enriched in nucleus with N/C ratios of 5:1 and 11:1 respectively. Still, it cannot be excluded that cytoplasmic aboundance was under estimated due to presence of rRNA.

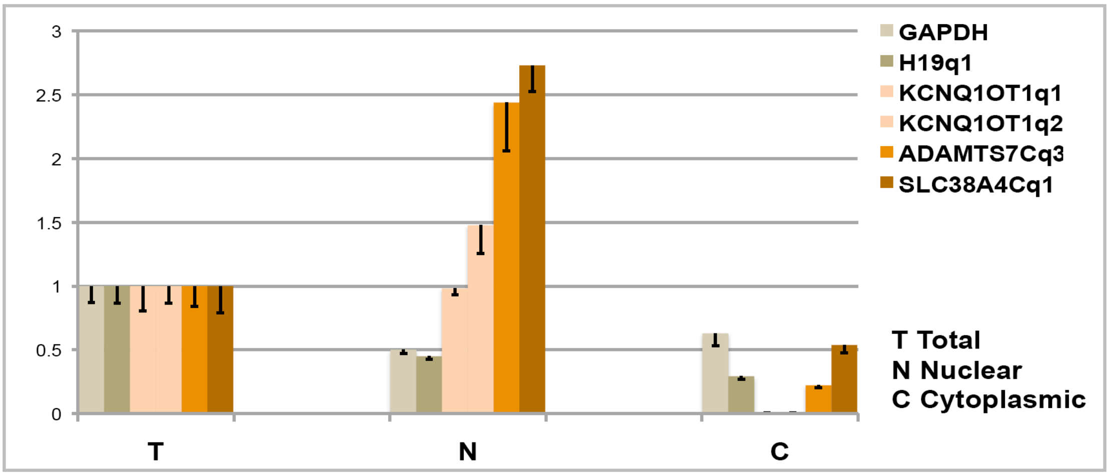


Figure 26. qRT-PCR showed nuclear enrichment of the *ADAMTS7down* and *SLC38A4down2* macro ncRNAs relative to the *RPLPO* housekeeping gene. Hs27 cells were fractionated to nuclear (N) and cytoplasmic (C) fractions. Total (T) RNA from the same cell line was set to 1 and used for comparisson. The *H19* macro ncRNA was found to be partially exported to the cytoplasm as previously published. The *GAPDH* housekeeping gene (GAPDH primer pair) showed the same kind of distribution showing both nuclear and cytoplasmic localization. *KCNQ1OT1* ncRNA was tested by two primer pairs (KCNQ1OT1q1 and KCNQ1OT1q2) and found to be nuclearly localized. *ADAMTS7down* macro ncRNA candidate has been tested with ADAMTS7Cq3 primer pair and found to be nuclearly enriched same as *SLC38A4down2* macro ncRNA tested with SLC38A4Cq1 primer pair. Standard deviation bars represent three technical replicates.

2. 4. 4. Tissue specific expression of macro ncRNAs from less-studied regions

Tissue specific expession of 73 novel macro ncRNA candidates in 43 tested cells/tissues/patients was examined from HIRTA single hybridization data by the ON/OFF system where ON means expressed at any level and OFF no expression of a ncRNA candidate (Figure 27). 2/73 transcripts (*ADAMTS7down* and *KLF14up3*) showed ubiquitous expression in all tested samples, while two other macro ncRNA candidates: *KIAA1126ov* and *BLCAPov* were expressed in most of the cells/tissues. The *SEC23IPdown3* expression was restricted to fetal and adult kidney and *L3MBTLup* restricted to fetal and adult brain tissue. 13/73 transcripts showed expression detected just in one of the 43 tested cells/tissues/patients, and 7 out of these 13 transcripts were testis specific. 56 transcripts showed different levels of expression or lack of expression in different tissues. 16 macro ncRNA candidates were found to be cancer specific (not expressed in any of tested normal tissues). From these candidates 9 were found exclusively in one cancer cell line; for example *SFMBT2down5* was cervical cancer specific and *OR56A1down* was AML5 patient specific. Examples of cancer specific expression of these transcripts will be shown in the section focused on macro ncRNA expression changes in cancer (2.6).

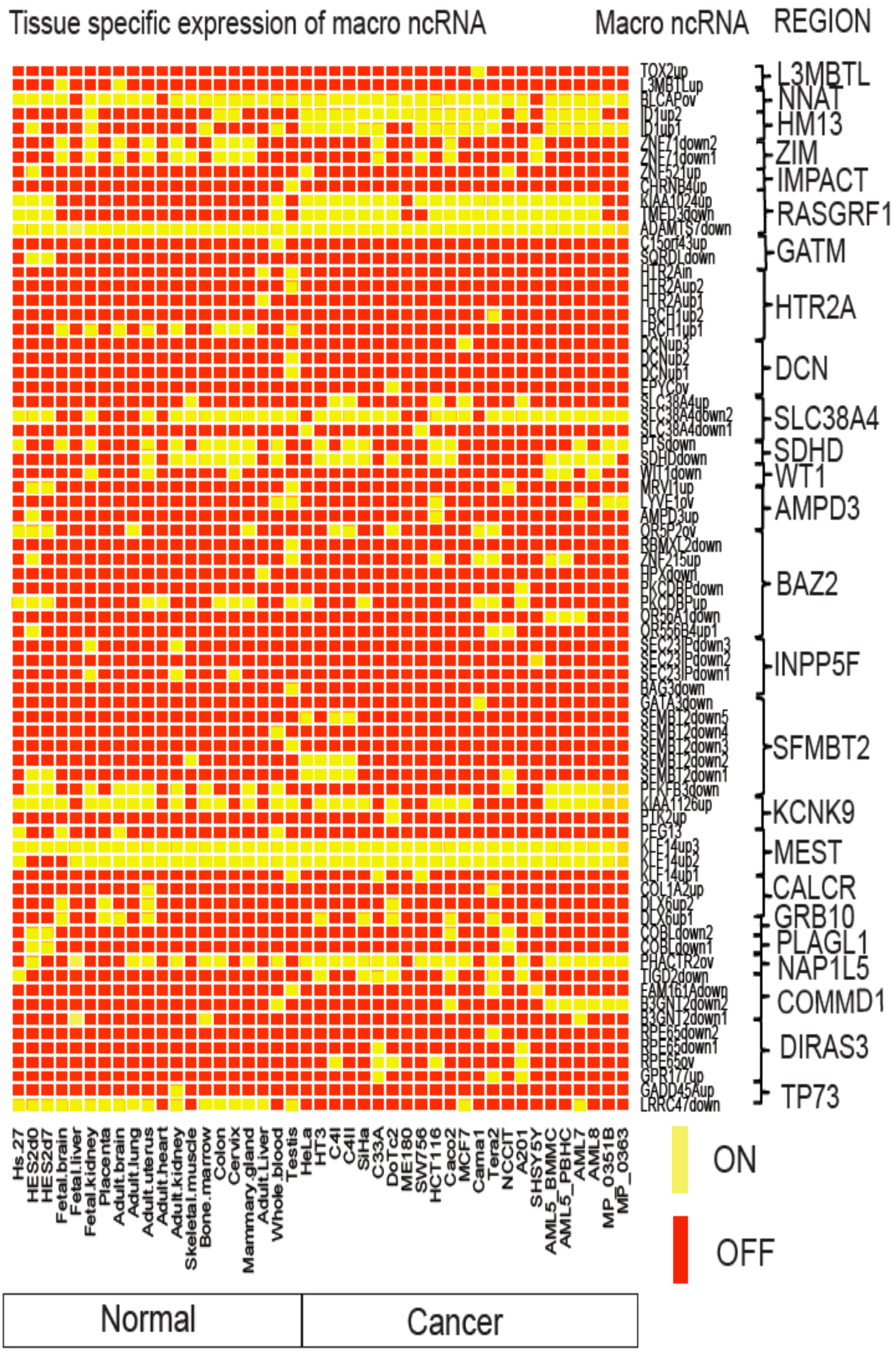


Figure 27. Tissue specific expression of novel macro ncRNAs in normal and cancer cells. Tissue specific expression of 73 macro ncRNA candidates from 26 less-known imprinted gene regions is shown for 43 tested cells/tissues. Description of the figure is as for Figure 17, section 2.2.1.1. Yellow; ON- expressed transcripts, Red; OFF-no expression.

2. 4. 5. Examples of characterization of novel macro ncRNAs

Ten novel macro ncRNA candidates expressed from less-studied imprinted gene regions were selected for further characterization. Characterization included assessment of: 1) Length of macro ncRNA, 2) Tissue specific expression and subcellular localization, 3) Annotation, 4) CpG island features, 5) Monoallelic/Imprinted expression of macro ncRNA.

Length of macro ncRNA transcripts as well as tissue specific expression was assessed by analyzing HIRTA hybridization data, or by Northern blots using diverse cells/tissues and probes specific for macro ncRNA of interest (for details about the method see section 5.8.6.). Mouse *Airn* and *Kcnq1ot1* ncRNAs that have function in gene silencing *in cis* are both nuclearly localized (Redrup et al., 2009; Seidl et al., 2006). Thus, characterization of transcripts included the test of subcellular localization for transcripts expressed in normal human fibroblasts using double cDNA hybridization (section 2.1.6.) and RT-PCRs on nuclear, cytoplasmic and total RNA.

Annotation of transcripts orientation was done using 5’-3’ slope feature, CpG islands, H3K4me3, RNAP II. CpG islands, visualized using CpG island track available on UCSC browser (Gardiner-Garden and Frommer, 1987), overlapping one of the transcript ends are potential promoters of the transcripts. CpG islands and/or borders of macro ncRNA candidate transcription were compared with maps of H3K4me3 histone modification peaks from 9 cell lines (GM12878, H1-hESC, HepG2, HMEC, HSMM, HUVEC, K562, NHEK and NHLF) and with maps of RNAP II peaks from 3 cell lines (HUVEC, K562 and NHEK), visualized using ENCODE Histone Modifications by Broad Institute ChIP-seq track available on UCSC browser, http://genome.ucsc.edu/cgibin/hgTrackUi?hgsid=168678869&c=chr1&g=wgEncodeBroadChipSeq, (Bernstein et al., 2005; Bernstein et al., 2006). The large intergenic non-coding (linc) RNAs (Khalil et al., 2009) were compared with HIRTA mapped transcripts expressed from Hs27 and HeLa cell lines. Since Khalil et al. data also include tracks of lincRNA association with Polycomb repressive complex (PRC)2 and chromatin-modifying protein CoREST complexes, indication of this association was also investigated for three novel HIRTA mapped macro ncRNAs overlapping with linc ncRNAs.

The assessed CpG island features were: test of CpG island promoter methylation status (if the CpG island is a DMR), test for presence of Differential Histone Modifications (DHM) and test for presence of tandem direct repeats in the CpG island. A number of imprinted macro ncRNA (e.g. *KCNQ1OT1*) have promoter CpG islands that are also Differentially Methylated Regions (DMRs) with one unmethylated and one methylated allele depending of parent-of-origin. Methylation status of CpG islands mapping to ends of novel macro ncRNAs was tested using methylation sensitive enzymes in Southern blot method and the Southern probe specific for tested macro ncRNA (for details about the method see section 5.5.). Deletions of DMRs showed previously that they could function as Imprint Control Elements (ICEs), controlling imprinted expression in the entire gene cluster (introduced in section 1.2.4.). Two known features of ICEs are: 1) presence of Differential Histone Modifications (DHM) including both activation H3K4me3 and repressive H3K9me3 chromatin modifications, 2) often direct repeats are present. Presence of DHM was tested comparing HIRTA expression profiles with ChIP-Seq profiles, published for both H3K4me3 and H3K9me3 in T cells by Barski (Barski et al., 2007). Presence of tandem direct repeats in the CpG island promoters of novel macro ncRNAs was assessed using dotmatcher program (EMBOSS).

The imprinted gene expression was assessed by a combination of PCR/Proofreading RT-PCR and sequencing. In the first step, DNA from the cell lines that express genes of interest, was amplified using PCR with primers overlapping with known SNPs from the dbSNP build 129 database (Sherry et al., 2001) and sequenced on the Applied Biosystems 3730xl DNA Analyzer. Results of sequencing were analyzed using Sequencher 4.7. The heterozygous SNPs were visualized as two sequencing peaks at one base pair (bp) position and the same primer pairs were further used in the proofreading RT-PCR reaction using cDNA as a template. If two peaks were overlapping on Sequencher tracks gained from both DNA and cDNA sequencing than biallelic, and if one peak was present on the cDNA than monoallelic expression was found. The blood from one family and three lymphoblastoid cell lines that originated from families genotyped from the international HapMap project (http://hapmap.ncbi.nlm.nih.gov/) were assessed using the same method (for details see section 5.8.5). Macro ncRNA candidates that were assessed for listed characteristics and that will be presented through the section are: *LRRC47down*, *KLF14up3, PEG13, KIAA1126up, PRKCDBPup, SLC38A4down2, ADAMTS7down, TMED3down, KIAA1024up* and *BLCAPov*.

2. 4. 5. 1. HIRTA REG TP73 (chr1): *LRRC47down* macro ncRNA characterization

TP73 is expressed from human chromosome 1 and encodes the p73 transcription factor involved in cellular response to stress and development (Dickman, 1997).

TP73 is the only known imprinted gene in the HIRTA TP73 region and shows expression from maternal chromosome in neuroblastoma cells (Kaghad et al., 1997). No macro ncRNAs and DMRs have been mapped previously to this region.

1) Length of macro ncRNA

The *LRRC47down* macro ncRNA candidate is ~7kb long. It is positioned ~ 40kb downstream of the *TP73* gene and immediately downstream from *LRRC47* (Leucine-rich repeat containing protein 47). *LRRC47down* partially overlaps the hypothetical protein *LOC388588* (Figure 28A).

2) Tissue specific expression and subcellular localization

LRRC47down macro ncRNA candidate was expressed in 16/20 normal cells/tissues while in cancer downregulation was apparent since just 2/23 cancer samples expressed *LRRC47down* (Figure 27, section 2.4.4). Double cDNA hybridization in Hs27 fibroblast cell line indicated potential cytoplasmic enrichment for this candidate (Figure 28A).

3) Annotation

The *LRRC47down* macro ncRNA candidate did not show clear presence of 5’-3’ slope, but it had CpG island 92 located at one of the ends of the transcript that could be a potential promoter of this macro ncRNA candidate, as well as H3K4me3 in 9/9 (shown for H1-hESC) and RNAP II in 1/3 (shown in K562) Broad Histone cell lines (Figure 28A).

4) CpG island features

The CpG island 92 was shown to be unmethylated in fibroblasts (Hs27) and HeLa cells using three methylation sensitive enzymes (BstUI, BssHII and EglI) in Southern blot hybridization with LRRC47SBP probe (Figure 28C). Presence of DHM could not be observed on this CpG island (data not shown). Direct repeats were present in the CpG island 92 (Figure 28D).

5) Monoallelic/Imprinted expression of macro ncRNA

One heterozygous SNP (rs12061931, C/T) was found in fibroblast, Hs27 DNA using sequencing of PCR products from LRRC47CIE1 and LRRC47CIE2 primer pairs. Sequencing of RT-PCR band containing heterozygous SNP showed that *LRRC47down* macro ncRNA candidate was bialelically expressed in fibroblasts Hs27 cell line (Figure 28B).

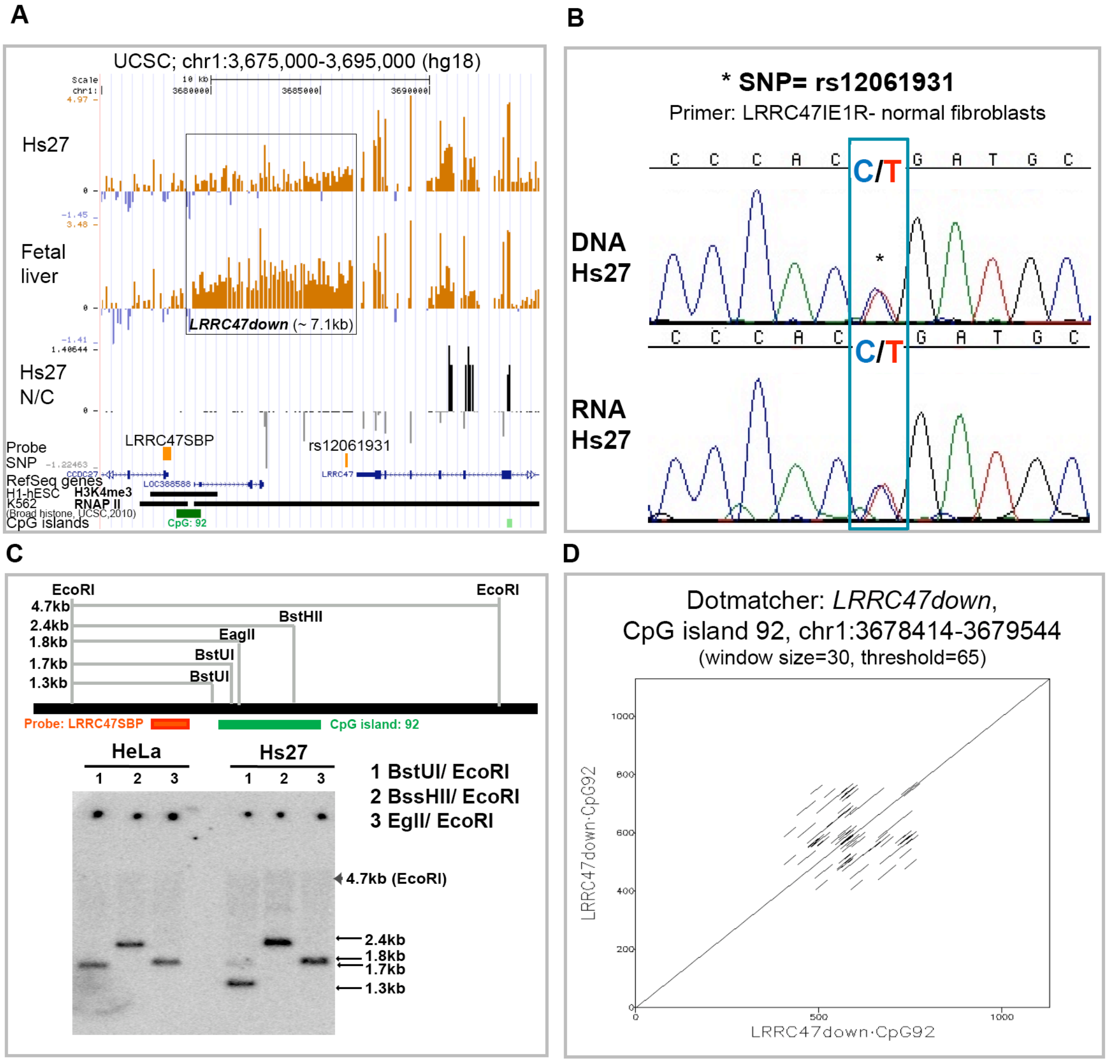


Figure 28. *LRRC47down* macro ncRNA candidate characterization. A. Expression of *LRRC47down* ncRNA candidate from normal fibroblasts (Hs27) and fetal liver is shown. This candidate is ~7kb in length and has a potential CpG island promoter (CpG 92) overlapped with H3K4me3 in H1-hESC cell line and RNAP II in K562 cell line. This candidate is potentially cytoplasmically enriched since gray cytoplasm signals could be observed in the Hs27, N/C track. Positions of the Southern blot probe LRRC47SBP and heterozygous SNP (rs12061931) are shown in the hg18 build on UCSC browser by loading the custom .wig track. UCSC browser position is shown on the top. Tracks presented as on Figure 11B, section 2.1.5. and Figure 12B., section 2.1.6. **B.** Primer pair used for PCRs and RT-PCRs is shown on the top. Products were sequenced and sequencing tracks from Hs27 DNA and RNA showing five nucleotides from both sides of heterozygous SNP are shown. rs12061931 express both C and T alleles in the Hs27 cell line. *; heterozygous SNP, blue box; position of the SNP. **C.** Map of the genomic region surrounding CpG island 92 is shown. Positions where enzymes cut are shown using vertical gray lines while numbers on the left side from vertical lines represent expected lengths (in kilobases) of the DNA fragments digested with corresponding enzymes and recognized with the probe that is shown in red. Southern blot using three methylation sensitive enzymes (BstUI, BssHII and EgII) in combination with EcoRI and LRRC47SBP probe showed that both alleles of CpG 92 are unmethylated in HeLa and Hs27 cell lines, since three methylation senzytive enzymes cutted DNA producing 2.4kb, 1.8kb, 1.7kb and 1.3kb bands while 4.7kb band expected from digestion of EcoRI enzyme alone was not present (indicated with arrow). **D.** CpG island 92 shows presence of direct repeats using dotmatcher program (EMBOSS) with following criteria, window size: 30 and treshold: 65.

In summary, *LRRC47down* was shown to be ~7kb long transcript expressed in most of the tested normal tissues (16/19) and not expressed in most of the tested cancer cell lines (21/23) that had unmethylated CpG island promoter and was biallelically expressed in fibroblasts.

2. 4. 5. 2. HIRTA REG MEST (chr7): *KLF14up3* macro ncRNA characterization

The HIRTA MEST region is positioned on human chromosome 7 and includes four known protein coding genes showing imprinted expression (*MEST* and *COPG2* that are paternally expressed and *CPA4* and *KLF14* that are maternally expressed) and two macro ncRNAs: *MESTIT1* and *MIT1,* both previously found to be paternally expressed imprinted genes (introduced in Table 6A, section 1.2.5.2.).

1) Length of macro ncRNA

The *KLF14up3* macro ncRNA candidate was a part of complex transcription unit potentially containing numerous overlapping transcripts expressed in both transcriptional orientations. The *KLF14up3* (~30kb long) alone was located directly upstream of *Homo sapiens* hypotetical LOC378805 (FLJ43663), transcript variant 1, ncRNA (RefSeq genes, UCSC) and could be part of this transcript (Figure 29A).

2) Tissue specific expression and subcellular localization

KLF14up3 was ubiquitously expressed in all 43 tested cells/tissues (Figure 27, section 2.4.4). HIRTA double cDNA hybridization (Figure 29A) and RT-PCRs on nuclear and cytoplasmic fractions of normal human fibroblasts (Figure 29B) both showed nuclear enrichment of *KLF14up3* macro ncRNA candidate.

3) Annotation

The presence of H3K4me3 in GM12878 lymphoblastoid cell line (and in 7 more cell lines mapped by Broad Histone, UCSC track, section 2.4.5.) at the distal end of the *KLF14up3* transcript, argues for plus (+) strand transcription that is opposite to transcription orientation of FLJ43663 ncRNA transcript. *KLF14up3* has been previously mapped as linc ncRNA, associated with PRC2 and CoREST complexes (Figure 29A).

4) CpG island features

KLF14up3 did not show presence of a CpG island promoter (Figure 29A).

5) Monoallelic/Imprinted expression of macro ncRNA

Despite the nuclear localization and genome position close to known imprinted genes, *KLF14up3* was biallelically expressed in fibroblasts on the basis of three heterozygous SNPs (rs17165272, rs205712 and rs13230391) (data shown for rs17165272, Figure 29B).

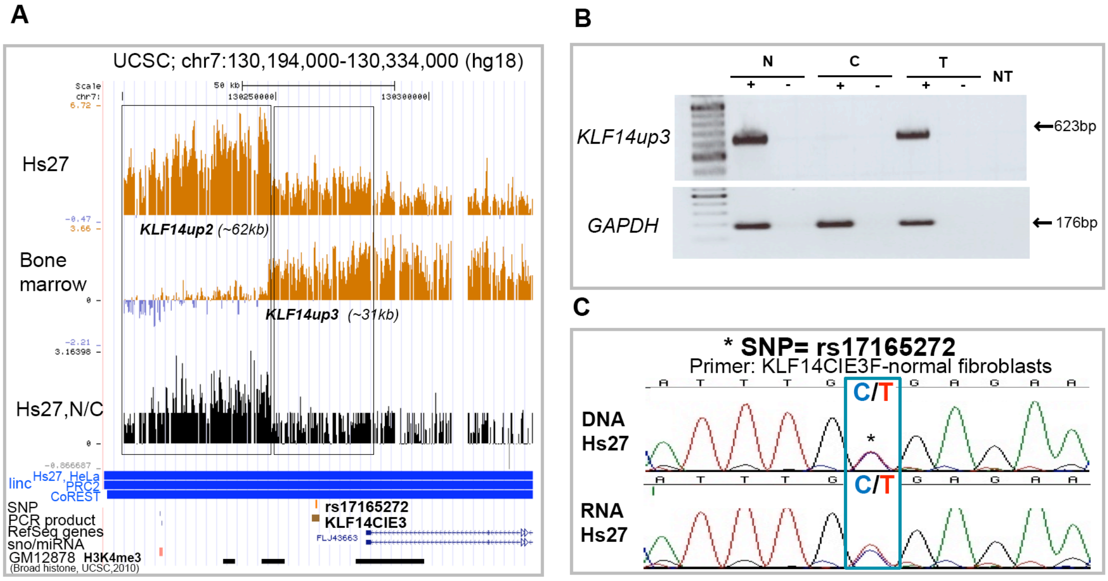


Figure 29. *KLF14up3* (~31kb) is nuclear enriched biallelic macro ncRNA. A. Expression of *KLF14up2* (~52kb) and *KLFup3* (~31kb) in Hs27 cells and bone marrow from single HIRTA hybridizations is shown. This macro ncRNA is enriched in nucleus by double cDNA hybridization and overlapps to previousy found linc RNAs that are associated with PRC2 and CoREST complexes. Positions of heterozygous SNP (rs17165272) and PCR product (KLF14CIE3) used for sequencing of SNPs and for testing cellular localization are shown in the hg18 build on UCSC browser by loading the custom .wig track. Tracks presented as on Figure 11B, section 2.1.5. and Figure 12B., section 2.1.6. **B.** *KLF13up3* nuclear localization has been confirmed using RT-PCR on nuclear (N), cytoplasmic (C) and total (T) fibroblasts RNA. *GAPDH* is used as a loading control. +; +RT reaction, -; -RT reaction, NT; no template **C.** *KLF14up3* is biallelically expressed in fibroblasts based on rs17165272 SNP. For detailed description look at the Figure 28B, section 2.4.5.1.

In summary, *KLF14up3* was ~30kb long, ubiquitous, nuclear, biallelically expressed macro ncRNA that is associated with PRC2 and CoREST complexes.

2. 4. 5. 3. HIRTA REG KCNK9 (chr8): *PEG13* and *KIAA1126up* ncRNA characterization

KCNK9 (potassium chanel, subfamily K, member 9) is the only known gene showing imprinted expression (maternally expressed (Ruf et al., 2007)) in HIRTA region KCNK9. In mouse this imprinted gene region contains two imprinted protein coding: maternally expressed genes *Kcnk9* and *Trappc9* (Ruf et al., 2007; Wang et al., 2008b) and one paternally expressed non-protein coding RNA, *Peg13* (Smith et al., 2003). The CpG island promoter of mouse *Peg13* is a gametic DMR (Ruf et al., 2007). Previously human *PEG13* was not identified, despite the in silico mapping of the orthologous human region of the *Peg13* DMR, to intron 17 of the *TRAPPC9* gene (Ruf et al., 2007).

1) Length of macro ncRNAs

HIRTA single hybridizations showed expression of the human ~ 6.6kb long transcript from intron 17 of *TRAPPC9*, overlapping the CpG island corresponding to mouse *Peg13* DMR and with specifically high expression in fetal and adult brain (Figure 30A). Thus, I named this transcript *PEG13*. The second macro ncRNA candidate identified in the KCNK9 region, *KIAA1128up* was ~55kb long macro ncRNA, located upstream of *KIAA1126*, a putative uncharacterized protein (that is annotated by UCSC gene track with eight annotated exons, but not present in RefSeq genes) with no reports about its imprinted expression in both human or mouse (Figure 31A).

2) Tissue specific expression and subcellular localization

Normal human fibroblasts (Hs27) showed low expression of *PEG13*. In the same cell line this ncRNA candidate was present in both nuclear and cytoplasm fractions by the double cDNA hybridization (Figure 30A). Except from fibroblasts, *PEG13* was expressed from the whole blood, and it showed especially high expression from fetal and adult brain (Figure 27, 2.4.4.). The *KIAA1126up*, candidate was highly expressed from the whole blood, while fibroblasts showed low expression (Figure 31A). The KIAA1126up was expressed from 15/20 of normal and from 14/23 cancer cell lines (Figure 27, section 2.4.4.). This macro ncRNA candidate was present in both nuclear and cytoplasm fractions of fibroblasts, based on the double cDNA HIRTA hybridizations (Figure 31A).

3) Annotation

The CpG island that overlapps *PEG13* (CpG: 210, UCSC) showed presence of H3K4me3 in 7/9 Broad Histone Modification tracks (presence of H3K4me3 in H1-ESC is shown on Figure 30A). The *KIAA1026up* had a potential CpG island promoter (CpG: 77, UCSC) that was overlapped by H3K4me3 in 8/9 cell lines on the Broad UCSC track.

4) CpG island features

A Southern blot using methylation sensitive enzyme BssHII in combination with EcoRI or EcoRI alone as a control, and specific probe (PEG13SBP) revealed that the CpG island 210 is a DMR in the Hs27 cell line where the methylated and unmethylated alleles respectively were visually present as bands of similar intensities indicating a 50%: 50% ratio (feature of a DMR) (Figure 30C). In the same time this CpG island shows hypermethylation in HeLa cells (Figure 30C). A dotplot of CpG 210 showed the presence of numerous direct repeats (Figure 30D). *KIAA1026up* CpG island was not tested for presence of methylation.

5) Monoallelic/Imprinted expression of macro ncRNA

To test potential monoallelic expression of *PEG13,* I used 4 primer pairs (P13CIE1-4) and were able to map one known heterozygous SNP (rs35257944) (dbSNP build 130, UCSC) and one novel heterozygous SNP on position chr8: 141176581-141176581 in Hs27 DNA by sequencing PCR products with both forward and reverse primers. Both SNPs showed a strong bias towards one allele (preferential monoallelic expression) being expressed in the fibroblast cell line (Figure 30B, rs35257944 presented). To test allelic expression of *KIAA1126up* transcript, I used five primer pairs (KIAA1126CIE1-5) positioned through the body of the *KIAA1126up* gene and mapped three heterozygous SNPs in the Hs27 cell line (rs62524186, rs67307958 and rs72681595). rs62524186 was heterozygous in Hs27 and had A nucleotide present on one parental chromosome (allele) and G nucleotide present on the same position on another parental chromosome. By using the KIAA1126CIE4 primer pair and sequencing RT-PCR products, biallelic expression of *KIAA1126up* was shown since both A and G alleles were expressed at the rs62524186 SNP (Figure 31B).

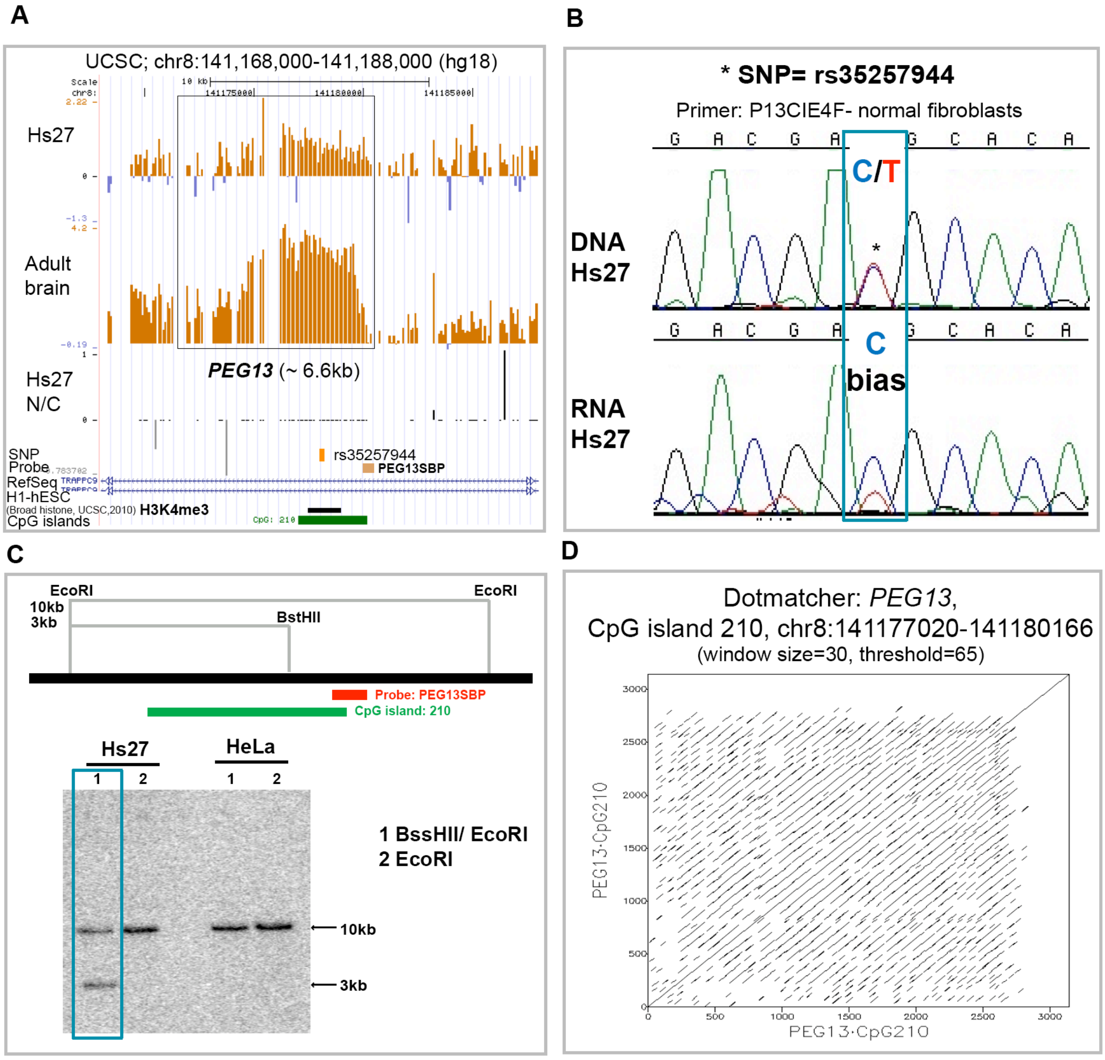


Figure 30. *PEG13* is a human homolog of mouse *Peg13* gene that shows preferential expression from one allele and the presence of Differentially Methylated Region (DMR) in normal human fibroblasts. A. Expression and nucleus vs. cytoplasmic localization of *PEG13* (~6.6kb long transcript) in fibroblasts (Hs27) and expression in adult brain is shown. CpG island 210 (UCSC, CpG island track) position and H3K4me3 presence in H1-ESC is shown (UCSC, Broad Histone Modifications). Positions of heterozygous SNP (rs35257944) and Southern blot probe (PEG13SBP) are presented using orange and brown boxes. Tracks presented as on Figure 11B, section 2.1.5. and Figure 12B., section 2.1.6. **B.** CpG 210 is a DMR in Hs27 fibroblasts while in HeLa cells this CpG island is hypermethylated. Methylation was tested by Southern blot using BssHII methylation sensitive enzyme in combination with EcoRI and EcoRI alone as a control while hybridization was done with specific probe (PEG13SBP). 10kb and 3.3kb bands represent methylated and unmethylated alleles respectively. **D.** Preferential monoallelic expression of *PEG13* ncRNA was found in Hs27 fibroblasts using the P13CIE4 primer pair flanking the rs35257944 SNP. For detailed description look at the Figure 28B, section 2.4.5.1. **C.** Numerous direct repeats are present in CpG 210 using Dotmatcher with criteria: window size=30 and threshold=65.

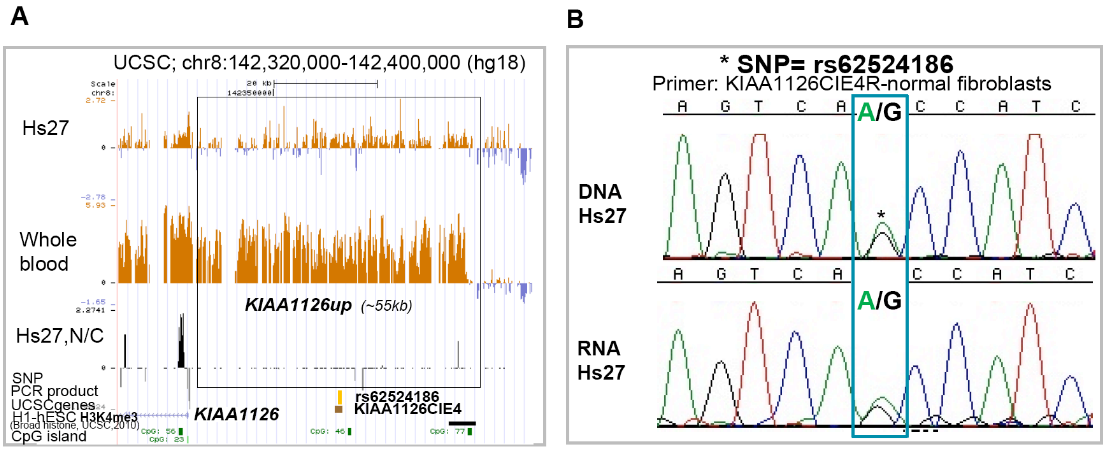


Figure 31. *KIAA1126up* macro ncRNA candidate is ~55kb long, both nuclear and cytoplasmic biallelically expressed transcript. A. Expresssion profiles in Hs27 and whole blood, double cDNA hybridization (Hs27, N/C) and presence of CpG: 77 island mapping to the distal end of the candidate and marked with H3K4me3 from H1-hESC CHIP-seq is shown. Tracks presented as on Figure 11B, section 2.1.5. and Figure 12B., section 2.1.6. **B.** *KIAA1126up* is biallelically expressed from Hs27 cell line. For detailed description look at the Figure 28 B, section 2.4.5.1.

In sumary, *PEG13* was a novel ~6.6kb long, nuclear macro ncRNA that was highly expressed from brain, showed biased expression towards one parental allele in fibroblasts and had CpG island that was a DMR. The *KIAA1126up* was a novel ~55kb long ncRNA candidate that is expressed in most of the tested tissues and cell lines, had both nuclear and cytoplasmic localization, had a CpG island promoter and was biallelically expressed in fibroblasts.

2. 4. 5. 4. HIRTA REG BAZ2 (chr11): *PRKCDBPup* macro ncRNA characterization

The HIRTA region named BAZ2 contains the *ZNF215* gene showing maternal imprinted expression in human while in the same region there is no known imprinted genes in mouse (Alders et al., 2000). Eight novel macro ncRNA candidates were mapped in this region. Macro ncRNA candidate *PRKCDBPup* will be presented in this section.

1) Length of macro ncRNA

PRKCDBPup was ~25kb in length and could represent novel 5' exons of *PRKCDBP* protein coding gene or macro ncRNA that overlapps this gene in the same direction.

2) Tissue specific expression and subcellular localization

The *PRKCDBPup* showed tissue specific expression and was expressed in 9/20 normal cells/tissues and 5/23 cancer samples (Figure 27, section 2.4.4.). The transcript showed high expression in adult uterus and it was very lowly expressed in fibroblasts where it showed both nuclear and cytoplasmic localization (data not shown).

3) Annotation

The CpG island 108 island could be a bidirectional promoter for both *PRKCDBPup* and *PRKCDBP* (Protein kinase C, delta binding protein). This CpG island was marked with H3K4me3 in 7/9 Broad Histone Modifications tracks with the example of H1-hESC shown in Figure 32A.

4) CpG island features

A Southern blot using BssHII methylation sensitive enzyme in combination with EcoRI and HindIII enzymes and the PRKCDBPSBP probe showed that CpG 101 is unmethylated in both the HeLa and Hs27 cell lines (Figure 32D), and therefore is not a DMR, since 14.8kb band and 4.3kb bands predicted to be observed from the methylated allele were not observed. The CpG island 108 showed presence of a low number of direct repeats using previously described criteria for the dotmatcher program (Figure 32E).

5) Monoallelic/Imprinted expression of macro ncRNA

By using both forward and reverse sequencing of PCR bands with two primer pairs (PRKCDBPCIE1 and 2) one known heterozygous SNP (C/T, rs12807110)(dbSNP build 130, UCSC) and one novel heterozygous SNP (T/A, chr11: 6314801- 6314801) were mapped in HeLa cells. Interestingly, these two SNPs were positioned with three base pairs in between. Both SNPs showed slight bias towards one allele in HeLa cells (Figure 31B). In order to test if these two SNPs are showing imprinted expression (both are expressed always from one parental allele) or they show random monoallelic expression (situation when some cells in the population transcribe one allele while other cells transcribe other allele, thus expression of two SNPs could be: both from the first allele, both from the second allele or one from the first and another SNP from the second allele (Krueger and Morison, 2008)), RT-PCR bands from HeLa cells using the PRKCDBPCIE2 primer pair were isolated and cloned into pGEM-T-Easy vector and sequenced from the T7 promoter. If candidate macro ncRNA shows expression from one allele uniformly (and thus could be imprinted) linkage of C and A alleles and linkage of T and T alleles (CA/TT SNPs combination) or linkage of C and T alleles and linkage of T and A alleles (CT/TA SNPs combination) would be expected in each tested case, while if the macro ncRNA has random monoallelic expression mixture of possible SNPs combinations (CT/TA and CA/TT SNPs combinations) would be expected (Figure 32C). Nine tested clones all showed CA/TT SNPs combinations, showing that these two SNPs are linked and thus indicating imprinted expression of *PRKCDBPup* (Figure 32C).

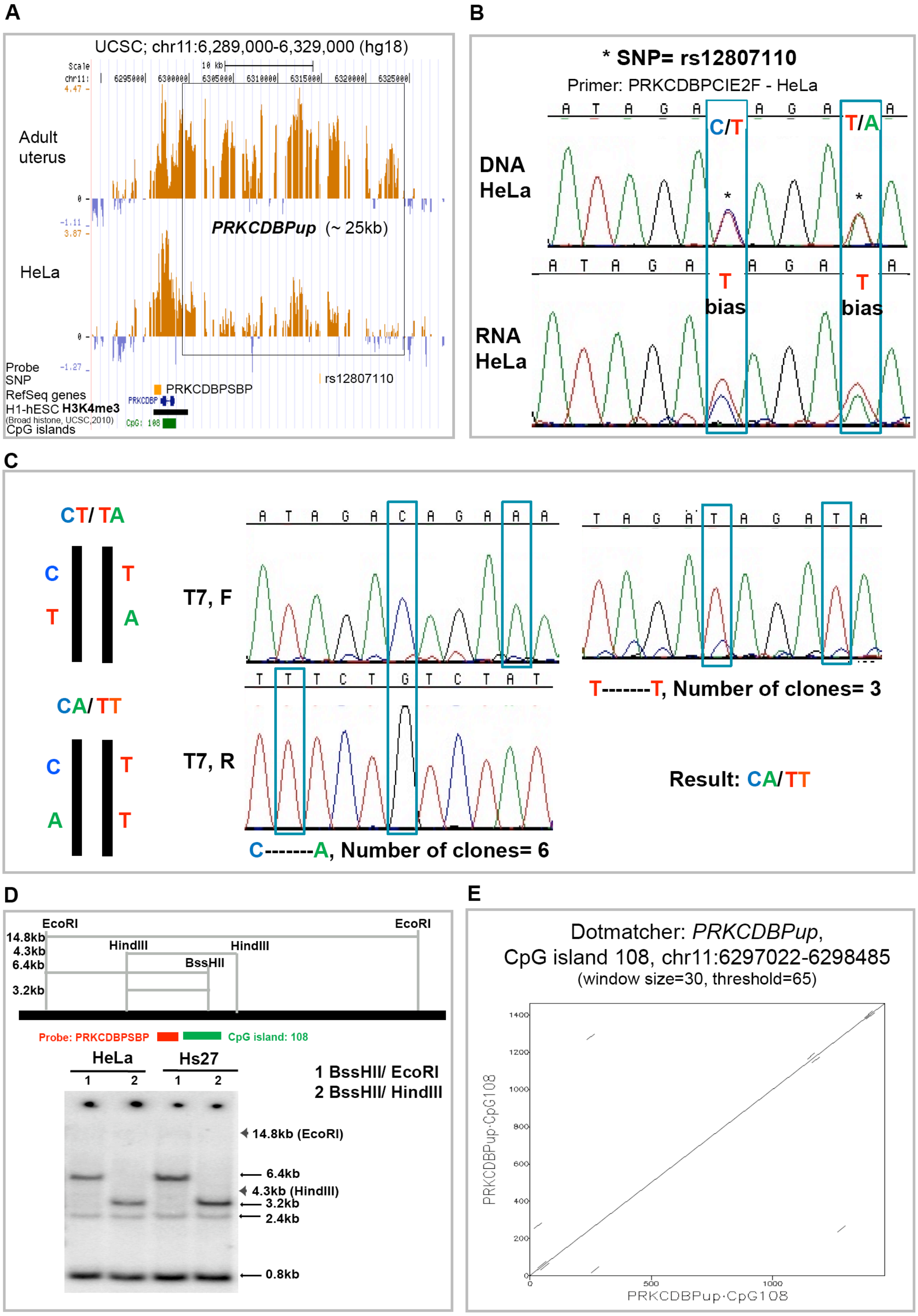


Figure 32. *PRKCDBPup* is a ~25kb long macro ncRNA candidate showing slight biased, but non-random expression towards one allele. A. Expression of *PRKCDBPup* in the adult uterus and in HeLa cells. CpG island 108 is marked with H3K4me3 in H1-hESC (Broad, UCSC). Positions of the Southern probe PRKCDBPSBP and rs12807110 SNP are shown as previously described. **B.** C/T known SNP and novel T/A SNP are mapped in HeLa DNA. Both SNPs show slight biased expression towards one allele in HeLa cells. **C.** On the left side possible combinations of expression from two close SNPs are shown. Sequencing from T7 promoter is shown. Nine clones show that two heterozygous SNPs are linked on the same chromosome (CA/TT SNPs combination) and argue against random expression of *PRKCDBPup* ncRNA candidate. **D.** Southern blot using BssHII methylation sensitive enzyme in combination with EcoRI and HindIII enzymes and PRKCDBPSBP probe in Hs27 and Hela cells. **E.** Dotplot of CpG island 108 is shown as previously.

In summary, *PRKCDBPup* was ~25kb long macro ncRNA candidate showing biased imprinted expression in HeLa cells and potentially having unmethylated bidirectional CpG island promoter.

2. 4. 5. 5. HIRTA REG SLC38A4 (chr12): *SLC38A4down2* macro ncRNA characterization

Imprinted expression of human genes residing in the SLC38A4 HIRTA region is unknown while mouse *Slc38a4* (Solute carrier family 38, member 4) has paternal imprinted expression (Smith et al., 2003).

1) Length of macro ncRNA

The *SLC38A4down2* macro ncRNA candidate was ~463kb long and the candidate was located about 10kb downstream of the human *SLC38A4* gene (Figure 33A).

2) Tissue specific expression and subcellular localization

The *SLC38A4down2* transcript was expressed in 14/20 HIRTA hybridizations of normal cells/tissues and in 19/23 cancer cells/patients (Figure 27, section 2.4.4.). Relative high expression of *SLC38A4down2* in HES2d0 was shown by both HIRTA single hybridizations and by qRT-PCR (normalized to *RPLPO* gene) using two primer pairs, SLC38A4Cq1 and SLC38A4Cq3 (Figure 33A). Fetal liver and Hs27 showed very low but no negative expression signals (blue signals) by HIRTA. This was in agreement with qRT-PCR that also showed low expression in both tissues. The *SLC38A4down2* showed 5:1 nuclear enrichment in Hs27 cells using qRT-PCR (shown on Figure 26, section 2.4.3.)

3) Annotation

Part of *SLC38A4down2* transcript was a lincRNA associated with PRC2 and CoREST complexes (Figure 33A). Since both lincRNA data and HIRTA data did not show strand specific information the strand of *SLC38A4down2* candidate was predicted as plus (+) since a CpG island 106 was observed on one end of the mapped transcript and the slope of the HIRTA expression through the body of the novel ncRNA was negative (reverse correlation between slope and transcriptional orientation was shown previously on Figure 19, section 2.2.1.) (Figure 33B). In order to further check the reliability of the predicted transcript strand I looked into published chromatin modification maps and RNAP II maps at the CpG island 106 (Figure 34A). In both examples shown in Figure 34A: T cells (examined by Chip-seq method (Barski et al., 2007)) and Normal Human Keratinocytes (NHEK) cell line from Broad Histone track (UCSC), enrichment of H3K4me3 and RNAP II was found on the CpG island 106 position supporting its function as a promoter of *SLC38A4down2*. Similar H3K4me3 methylation enrichment of the CpG island position was present in a range of cell lines on the Broad Histone track, UCSC.

4) CpG island features

The CpG island 106 adjacent to *SLC38A4down2* does not show the features of a DHM since no H3K9me3 was present on this island (Figure 34A), but this CpG island showed presence of direct repeats when dotmatcher program using following criteria: window=30 and threshold=65, was used. In order to experimentally test if CpG island 106 is a DMR, Southern blot using two methylation sensitive enzymes (BstUI and BssHII) and specific probe (SLC38A4SBP) located upstream of CpG island was performed in two cell lines (Hs27 and HeLa). Methylation sensitive enzymes were used in combination with the EcoRI enzyme where a 15kb band recognized with specific probe would be expected if allele is methylated (when BstUI and BssHII can not cut) while 3.1 or 3.3kb bands would be expected from unmethylated alleles (when BstUI and BssHII are able to cut DNA). If a DMR is present both 15kb and 3.1kb bands or 15kb and 3.3kb bands are expected showing presence of unmethylated and methylated allele. Both alleles of the CpG island promoter of *SLC38A4down2* were unmethylated by two methylation senzytive enzymes in both Hs27 and HeLa cell lines (Figure 34C). Thus, CpG island promoter of *SLC38A4 down2* is not a DMR and cannot have a function as an ICE in this region.

5) Monoallelic/Imprinted expression of macro ncRNA

In order to test if *SLC38A4down2* macro ncRNA is biallelic or shows imprinted expression 23 primer pairs for PCR/RT-PCR were designed. By doing PCRs with these primers on Hs27 cell line DNA and sequencing products, I was not able to map any heterozygous Single Nucleotide Polymorphism (SNP) in those tested 23 regions that range in length from 130 to 990bp, of the *SLC38A4down2* gene body. Therefore, I further tested these primers on the blood from a volunteer whose mother also volunteered blood. After sequencing, 7 heterozygous SNPs from 4 primer pairs were found. Expression of 4 heterozygous SNPs (rs12814298, rs11183486, rs2131371, rs2408497) was further tested using forvard plus reverse primers (SLC38A4CIE1 and SLC38A4CIE12) to sequence the SNPs and in all eight cases preferential expression of one allele was found (Figure 33D). In order to reveal parent-of-origin of *SLC38A4down2* expression we tested the Mothers' blood and on the basis of rs2408497 SNP, a bias towards the Fathers' G allele was found. Thus, preferential paternal expression of *SLC38A4down2* was detected.

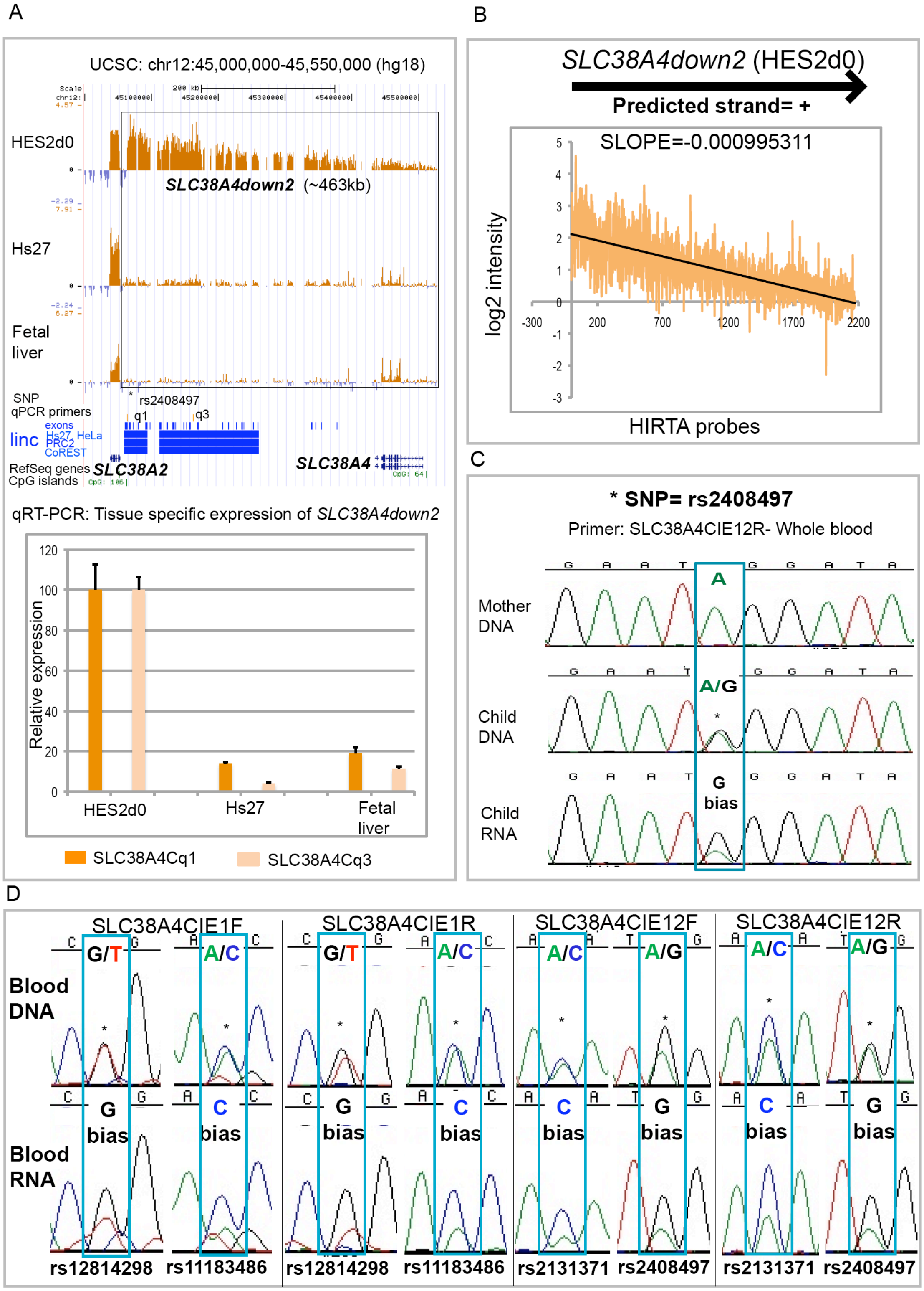


Figure 33. *SLC38A4down2* is macro ncRNA candidate predicted to be expressed in plus (+) transcriptional orientation and showing preferentially paternal imprinted expression. A. Expression of *SLC38A4down2* (~463kb long) macro ncRNA candidate is shown using HIRTA and qRT-PCR (primer pairs: SLC38A4Cq1 and SLC38A4Cq3) in fibroblasts, HES2, d0 cells and fetal liver. qRT-PCR expression relative to *RPLPO* gene. Linc RNAs exons, Hs27/HeLa expression, PRC2 and CoREST association tracks as well as custom SNP and qPCR primers tracks are shown. Tracks presented as on Figure 11B, section 2.1.5. **B.** Slope of *SLC38A4down2* in HES, d0 cells predicts + strand expression of this candidate. Details as on Figure 19, section 2.2.1. **C.** *SLC38A4down2* contains SNP (rs2408497) showing bias towards paternal allele in one family whole blood. For detailed

description look at the Figure 28B, section 2.4.5.1. **D.** Forward and reverse sequencing from SLC38A4CIE1 and SLC38A4CIE12 primers revealed 4 heterozygous SNPs (rs12814298, rs11183486, rs2131371 and rs2408497) in blood DNA and showed biased expression from blood RNA.

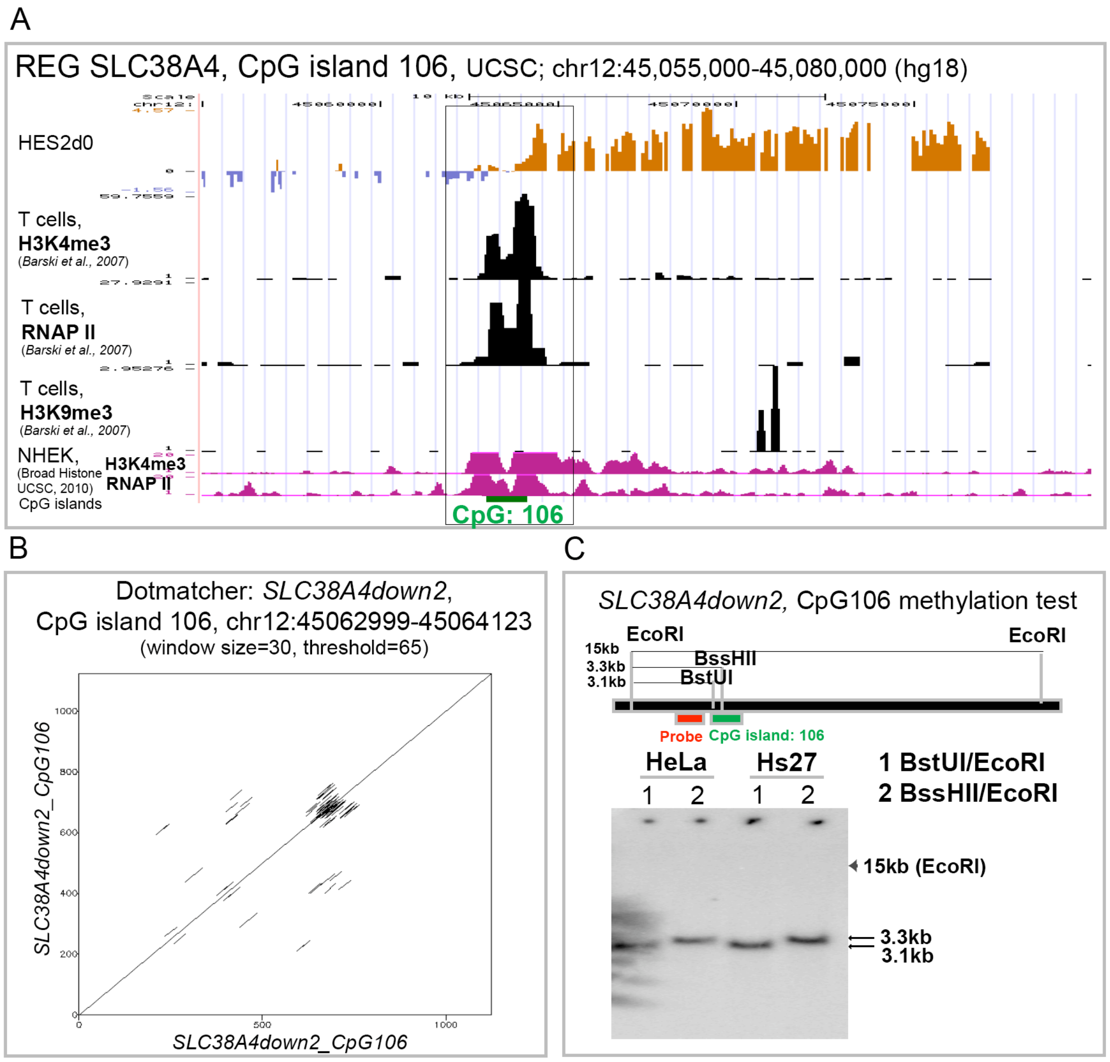


Figure 34. The CpG island 106 matches to the proximal end of *SLC38A4down2* HIRTA expression, is enriched for H3K4me3 and RNAP II, direct repeats and is unmethylated. A. CpG island 106 region on chromosome 12 shows enrichment of H3K4me3 in T cells and NHEK cells and RNAP II peaks over the same region. Peak of H3K9me3 repressive mark is found upstream, but not matching the CpG island. **B.** Dotplot of CpG island 106 is shown as previously. **C.** Southern blot using BstUI and BssHII in combination with EcoRI and SLC38A4SBP in HeLa and Hs27 cells.

In sumary, *SLC38A4down2* was ~463kb long macro ncRNA that was predicted to have plus (+) strand transcriptional orientation, partially overlapped to lincRNAs associated with PRC2 and CoREST complexes, showed paternal biased imprinted expression and had unmethylated CpG island promoter.

2. 4. 5. 6. HIRTA REG RASGRF1 (chr15): *ADAMTS7down*, *TMED3down* and *KIAA1024up* macro ncRNA characterization

Rasgrf1 and RASGRF1 regions are located on mouse chromosome 9 and human chromosome 15 respectively. Previous studies showed that mouse Rasgrf1 region contains four genes showing paternal imprinted expression: *Rasgrf1*, *As4*, *A19* and *miR-184* (de la Puente et al., 2002; Nomura et al., 2008; Plass et al., 1996). Imprinting status of genes residing in human *RASGRF1* region is not known. I characterized three macro ncRNA candidates (*ADAMTS7down*, *TMED3down* and *KIAA1024up*), found in the human RASGRF1 region by HIRTA single cDNA hybridizations.

1) Length of macro ncRNA

ADAMTS7down macro ncRNA was ~6kb long and located downstream of *ADAMTS7* protein coding gene (ADAM metallopeptidase with thrombospondin type 1, motif, 7) (Figure 35A). The length of the candidate mapped from the HIRTA hybridization data, was further confirmed in fibroblasts using Northern blot with a specific ADAMTS7downNOR probe. ß-Actin was used as a loading control for the Northern blot (Figure 35C). The Northern blot was repeated using fibroblasts and HeLa and same ~6kb band was found in both Hs27 and HeLa (data not shown). In addition to *ADAMTS7down*, I mapped two more candidates in human RASGRF1 region: *TMED3down* (~88kb in length) and *KIAA1024up* (~19kb in length) (Figure 39A).

2) Tissue specific expression and subcellular localization

The *ADAMTS7down* was ubiquitously expressed in all cells/tissues hybridized to HIRTA, including skeletal muscle that was the only tissue showing very low expression of this candidate (Figure 27, section 2.4.4. and Figure 35A). Expression of *ADAMTS7down* was also assessed by qRT-PCR using the ADAMTS7Cq3 primer pair in three different tissues and showed the expected profile matching to HIRTA hybridizations of the same tissues (Figure 35A, B). Cellular localization of *ADAMTS7down* was tested and its preferential nuclear localization using double HIRTA hybridization (Figure 35A), RT-PCR (Figure 35D) and qRT-PCR (11:1 ratio of nuclear to cytoplasmic) (Figure 25, section 2.4.1.) was shown.

TMED3down and *KIAA1024up* candidates were expressed in fibroblasts, blood and human embryonic stem cells but were not expressed in any other tested normal cells/tissues (Figure 27, 2.4.1). Both candidates showed upregulation in different cancer types (Figure 27, 2.4.1). To test nuclear versus cytoplasmic localization of these candidates double cDNA hybridization and RT-PCRs were used. *TMED3down* was enriched in the nuclear compartment by both techniques. *KIAA1024up* also showed nuclear enrichment, but in this case I was also able to detect substantial amount of this transcript in cytoplasmic fraction by RT-PCR, showing both nuclear and cytoplasmic localizations (Figure 39B).

3) Annotation

The CpG island 17 maps to the *ADAMTS7down* proximal position. The same CpG island was marked with the H3K4me3 histone in 6/9 cell lines from the Broad Histone track (Figure 36A). However, the H3K4me3 activation mark of promoters was not only found on the CpG island 17 position in *ADAMTS7down* region, but also on a distal position where 3/9 cell lines mapped by Broad Histone track showed its presence (Figure 36A). *TMED3down* and *KIAA1026up* showed no presence of CpG island. H3K4me3 were not found in any of 9 Broad cell lines on both ends of *TMED3down* while 7/9 cell lines showed peaks of H3K4me3 coresponding to the minus (-) strand transcriptional orientation of *KIAA1026up*.

4) CpG island features

In mouse, the 252bp long DMR located about 30kb upstream of *Rasgrf1* is paternally methylated in both monoallelic and biallelic tissues while in human there is no data about the presence of a DMR in the *RASGRF1* gene region that contains *ADAMTS7* and the novel *ADAMTS7down* transcripts. To test the methylation status of CpG island 17, which is potential promoter of *ADAMTS7down,* Southern blot with the BstUI methylation sensitive enzyme in combination with EcoRI and specific ADAMTS7downSBP probe, was used. The CpG 17 island was a DMR in both fibroblasts and HeLa cells since from BstUI/EcoRI digestion were gained both a 10.6kb band representing the methylated allele and a 1.6kb band representing the unmethylated allele (Figure 36B). By using EcoRI alone as a control, a 10.6kb band coming from the methylated allele and a 1.9kb band originating from the unmethylated allele were observed (Figure 36B). The CpG island 17 did not show any existence of direct repeats using dotmatcher program with standard criteria (data not shown).

5) Monoallelic/Imprinted expression of macro ncRNA

To determine if *ADAMTS7down* macro ncRNA candidate shows monoallelic expression I tested 11 primer pairs (ADAMTS7CIE1-11) on DNA from the Hs27 cell line and found two heterozygous SNPs (rs7174572 and rs34019568). Both SNPs showed monoallelic expression of *ADAMTS7down* in fibroblasts (Figure 37B). To see whether *ADAMTS7down* shows imprinted expression, same 11 primer pairs were tested on DNA from whole blood of one family consisting of the child and its mother (family described in section 2.4.5.5.). Two heterozygous SNPs were found in the child's DNA (rs71211234 and rs12908299) and after testing its RNA I observed monoallelic expression of *ADAMTS7down* in blood. The child's RNA showed expression of the G allele in both SNPs and the mothers DNA showed a bias towards the expression of A allele in both cases, indicating paternal expression of *ADAMTS7down* candidate on the basis of these two SNPs in blood (Figure 37C). To confirm imprinted expression of this candidate two more families were tested using RNA from lymphoblastoid cell lines (GM108054 and GM108046). Genotypes of the Utah families (father, mother and children) have been published by the international Haplotype Map of the Human Genome (HapMap) project (http://snp.cshl.org/) (Frazer et al., 2007). Two heterozygous SNPs in HapMap data from the children (SNP_A-8391868 and SNP_A-8407002) were found and RNA from lymphoblastoid cells originated from the same children was tested. Sequencing from both the forward and reverse primers ADAMTS7CIE11 in the GM108054 cell line showed C allele expressed on the position of SNP_A-8391868. This expression confirmed monoallelic expression of *ADAMTS7down* in lymphoblastoid cells and by assessing HapMap genotyping data of parents, paternal expression of *ADAMTS7down* was found.

However, in contrast to this, sequencing from both forward and reverse primers ADAMTS7CIE10 showed only the A allele expressed on the position of SNP_A-8407002, which according to the HapMap genotyping data indicated maternal expression of *ADAMTS7down* (Figure 37D). The same heterozygous SNPs were tested on RNA from the second lymphoblastoid cell line (GM108046) originating from another family. In this case SNP_A-8391868 showed maternal while SNP_A-8407002 showed paternal expression of *ADAMTS7down* (Figure 37E). In summary my results show clear monoallelic expression of *ADAMTS7down* from fibroblasts, blood and lymphoblastoid cell lines. Concerning the imprinted expression we tested three families and found paternal expression on the basis of four SNPs and maternal expression on the basis of two SNPs. Possible expleantions for these findings could be monoallelic random expression of *ADAMTS7down* macro ncRNA or potential artefacts in HapMap whole genome genotyping data (discussed in section 3.6.1.).

To test if the *ADAMTS7* protein coding gene positioned ~2kb from *ADAMTS7down* also shows monoallelic expression four primer pairs (ADAMTS7GIE1-4) were tested and one heterozygous SNP (rs7173267) positioned in exon2 of *ADAMTS7* protein coding gene in fibroblasts (Hs27) was found. The *ADAMTS7* protein-coding gene was biallelically expressed in fibroblasts (Figure 38).

To examine monoallelic expression of *TMED3down* and *KIAA1024up* ncRNA candidates I tested seven primer pairs (TMED3CIE1-7) and six primer pairs (KIAA1024CIE1-6) on normal fibroblasts cell lines, respectively. By using TMED3CIE4 primer pair on Hs27 DNA the rs1532968 heterozygous SNP (T/G) that showed preferential expression of one allele (T bias) in fibroblasts was mapped (Figure 39C). Further, by using the KIAA1024CIE3 primer pair the heterozygous rs769770 (A/G) SNP that also showed preferential expression from one allele (G bias) in fibroblasts was mapped (Figure 39C). In summary, both *TMED3down* and *KIAA1024up* were expressed preferentially from one allele in fibroblasts and therefore may show imprinted gene expression.

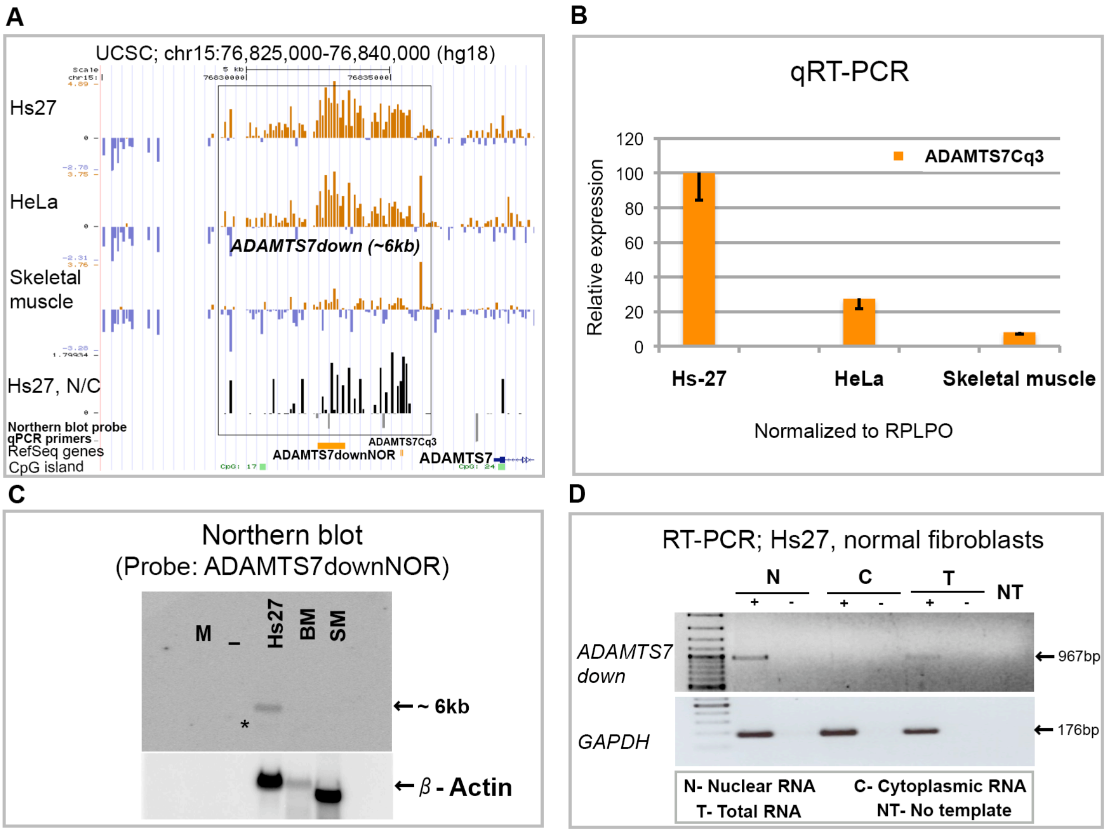


Figure 35. *ADAMTS7down* is a ~6kb long preferentially nuclearly localized macro ncRNA candidate. A. *ADAMTS7down* expression is high in fibroblasts (Hs27), medium in HeLa cells and low in skeletal muscle by HIRTA single cDNA hybridizations. Double cDNA hybridization of nuclear versus cytoplasmic fractions of Hs27 (Hs27, N/C) shows preferential nuclear localization. Positions of the Northern blot probe (ADAMTS7downNOR), qPCR

primers (ADAMTS7Cq3) and CpG island 17 and 24 are shown together with the RefSeq genes UCSC track in hg18 human build. **B.** qRT-PCR using ADAMTS7Cq3 primer pair and normalized to the *RPLPO* gene in three tissues confirmed findings from HIRTA single hybridizations. **C.** Northern blot on fibroblasts (Hs27), bone marrow (BM) and skeletal muscle (SM) RNA using ADAMTS7downNOR probe, detects a strong band ~6kb in length in fibroblasts and very faint (*) ~4kb long band that is likely an artefact of 28S rRNA. Northern blot using a ß-Actin specific probe was used as a loading control. **D.** RT-PCR of nuclear, cytoplasmic and total fibroblasts RNA using ADAMTS7downNOR primer pair (967bp band) showed nuclear enrichment of *ADAMTS7down* macro ncRNA candidate. *GAPDH* was used as a loading control (176bp band).

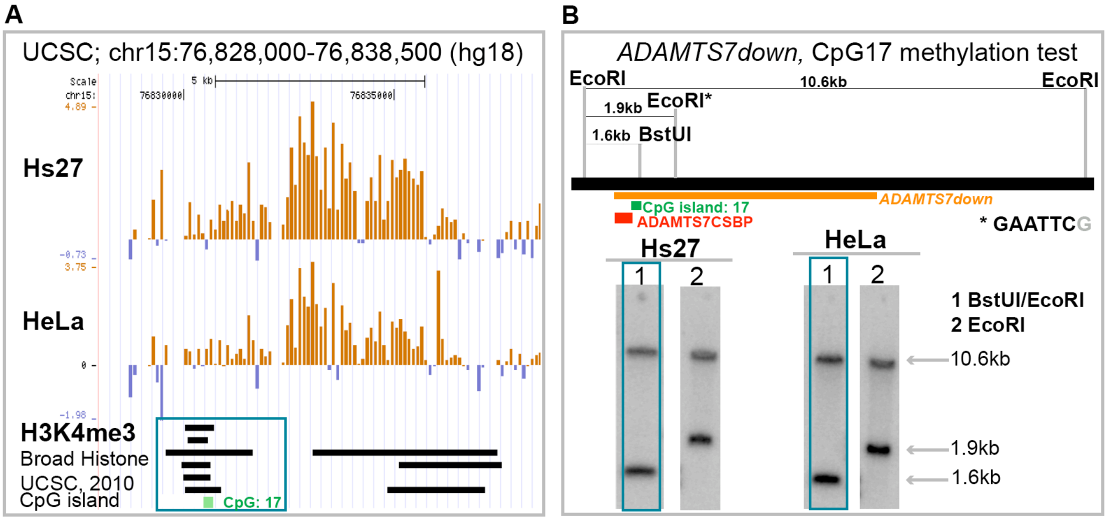


Figure 36. CpG island promoter of *ADAMTS7down* is a differentially methylated region (DMR) in fibroblasts and HeLa cells. A. HIRTA expression of *ADAMTS7down* in fibroblasts and HeLa cells. The proximal end of *ADAMTS7down* expression overlaps with CpG island 17 and H3K4me3 from 6/9 cell lines displayed by Broad Histone, UCSC track. The distal-end of *ADAMTS7down* HIRTA expression overlaps with H3K4me3 in 3/9 cases. Tracks presented as on Figure 11B, section 2.1.5. **B.** Southern blot using BstUI methylation sensitive enzyme in combination with EcoRI and EcoRI alone as a control and ADAMTS7CSBP specific probe, showed that CpG island 17 is a DMR in Hs27 and HeLa cells. The 10.6kb band represents the methylated allele, the 1.6kb band the unmethylated allele and a 1.9kb band is a result of an unusual EcoRI digestion: when G is present after GAATTC EcoRI cutting site this enzyme is methylation sensitive and cuts just the unmethylated allele.

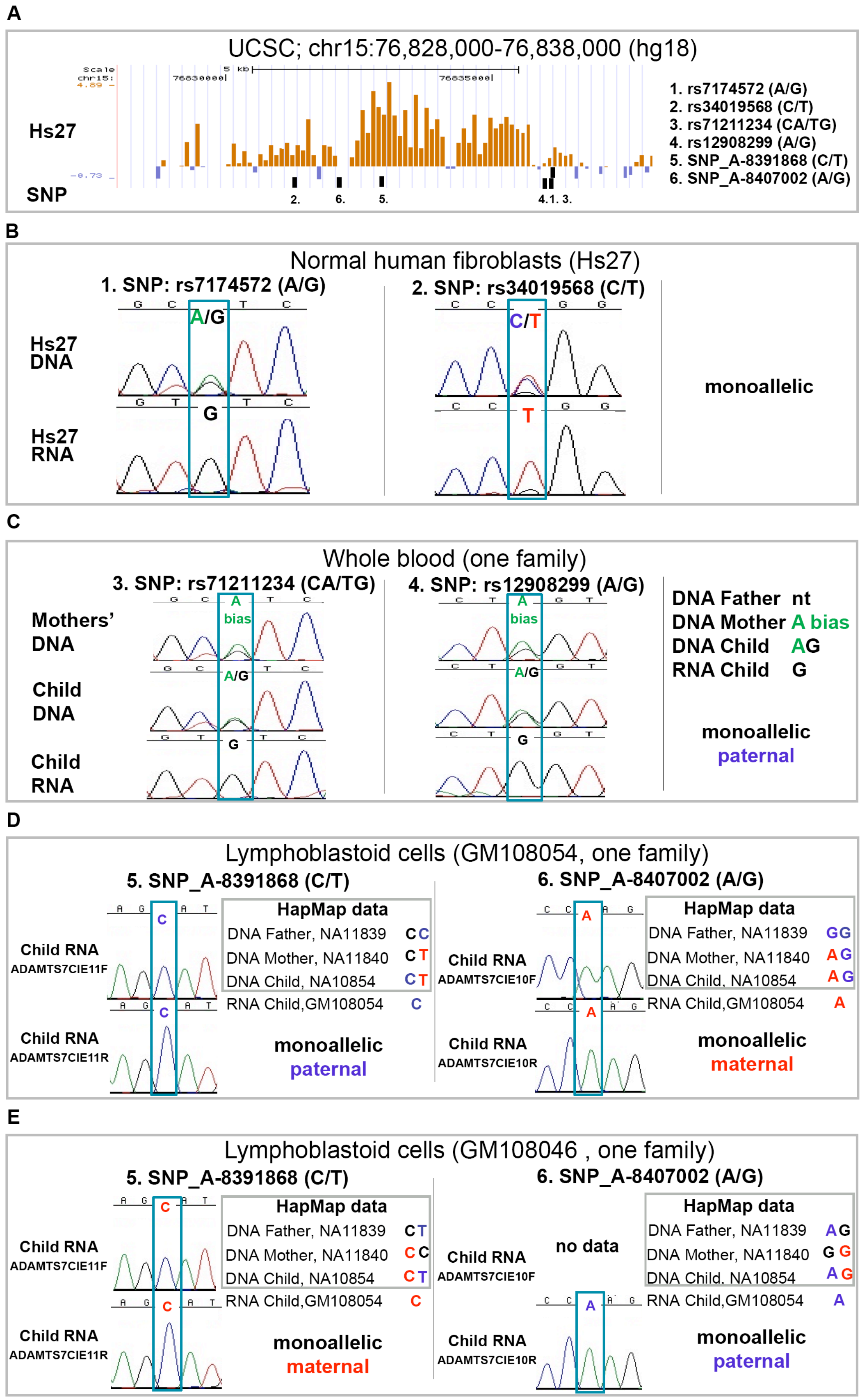

A
UCSC; chr15:76,828,000-76,838,000 (hg18)
Scale
chr15:
5 kb
76830000
76835000
4.89
-0.73
Hs27
SNP
2.
6.
5.
4.1. 3.
1. rs7174572 (A/G)
2. rs34019568 (C/T)
3. rs71211234 (CA/TG)
4. rs12908299 (A/G)
5. SNP_A-8391868 (C/T)
6. SNP_A-8407002 (A/G)
B
Normal human fibroblasts (Hs27)
1. SNP: rs7174572 (A/G)
2. SNP: rs34019568 (C/T)
Hs27 DNA
Hs27 RNA
A/G
G
C/T
T
monoallelic
C
Whole blood (one family)
3. SNP: rs71211234 (CA/TG)
4. SNP: rs12908299 (A/G)
Mothers' DNA
Child DNA
Child RNA
A bias
A/G
G
DNA Father nt
DNA Mother A bias
DNA Child AG
RNA Child G
monoallelic
paternal
D
Lymphoblastoid cells (GM108054, one family)
5. SNP_A-8391868 (C/T)
6. SNP_A-8407002 (A/G)
Child RNA ADAMTS7CIE11F
Child RNA ADAMTS7CIE11R
Child RNA ADAMTS7CIE10F
Child RNA ADAMTS7CIE10R
C
A
HapMap data
DNA Father, NA11839 CC
DNA Mother, NA11840 CT
DNA Child, NA10854 CT
RNA Child,GM108054 C
monoallelic
paternal
HapMap data
DNA Father, NA11839 GG
DNA Mother, NA11840 AG
DNA Child, NA10854 AG
RNA Child,GM108054 A
monoallelic
maternal
E
Lymphoblastoid cells (GM108046 , one family)
5. SNP_A-8391868 (C/T)
6. SNP_A-8407002 (A/G)
Child RNA ADAMTS7CIE11F
Child RNA ADAMTS7CIE11R
Child RNA ADAMTS7CIE10F
Child RNA ADAMTS7CIE10R
C
no data
A
HapMap data
DNA Father, NA11839 CT
DNA Mother, NA11840 CC
DNA Child, NA10854 CT
RNA Child,GM108054 C
monoallelic
maternal
HapMap data
DNA Father, NA11839 AG
DNA Mother, NA11840 GG
DNA Child, NA10854 AG
RNA Child,GM108054 A
monoallelic
paternal

Figure 37. The *ADAMTS7down* macro ncRNA candidate region is monoallelically expressed in all tested cells/tissues. A. Positions of 6 heterozygous SNPs in hg18 assembly coresponding to HIRTA expression of *ADAMTS7down* in fibroblasts (Hs27). **B.** Sequencing tracks showed as previously. Monoallelic expression in normal human fibroblasts shown using two SNPs (rs7174572 and rs34019). **C.** *ADAMTS7down* mapped region shows paternal expression based on two SNPs (rs71211234 and rs12908299) from the whole blood of one family. **D.** *ADAMTS7down* mapped region shows paternal expression based on SNP_A-8391868 when lymphoblastoid cell GM108054 RNA from the child was sequenced and compared to the parents' genotypes from the HapMap project. The same region shows maternal expressin based on SNP_A-8407002. **E.** *ADAMTS7down* mapped region shows maternal expression based on SNP_A-8391868 when lymphoblastoid cell GM108046 RNA from the child was tested and compared to the parents' genotypes from published the HapMap consortium. The same region shows paternal expression based on SNP_A-8407002.

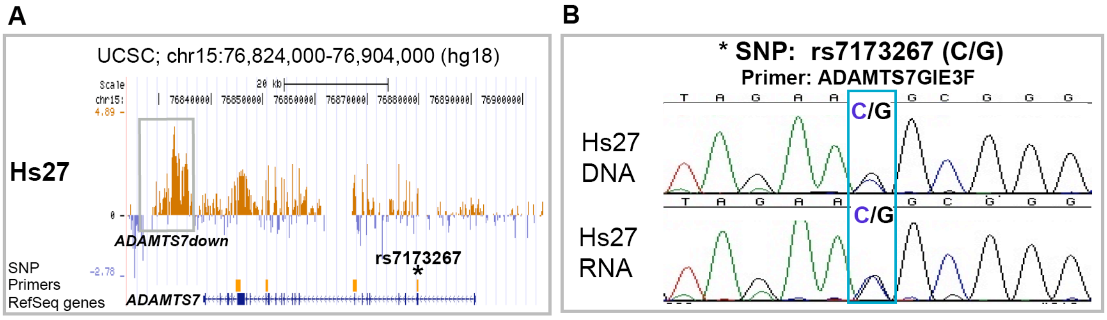


Figure 38. *ADAMTS7* protein coding gene is biallelically expressed in fibroblasts. A. HIRTA expression in fibroblasts in *ADAMTS7* gene region. The position of *ADAMTS7down* is marked by gray box. The positions of four ADAMTS7GIE primers are shown by orange boxes. Position of heterozygous rs7173267 SNP is marked by a star. **B.** Sequencing of PCR and RT-PCR products from primer ADAMTS7GIE3F showed biallelic expression of *ADAMTS7* in fibroblasts on the basis of rs7173267 SNP.

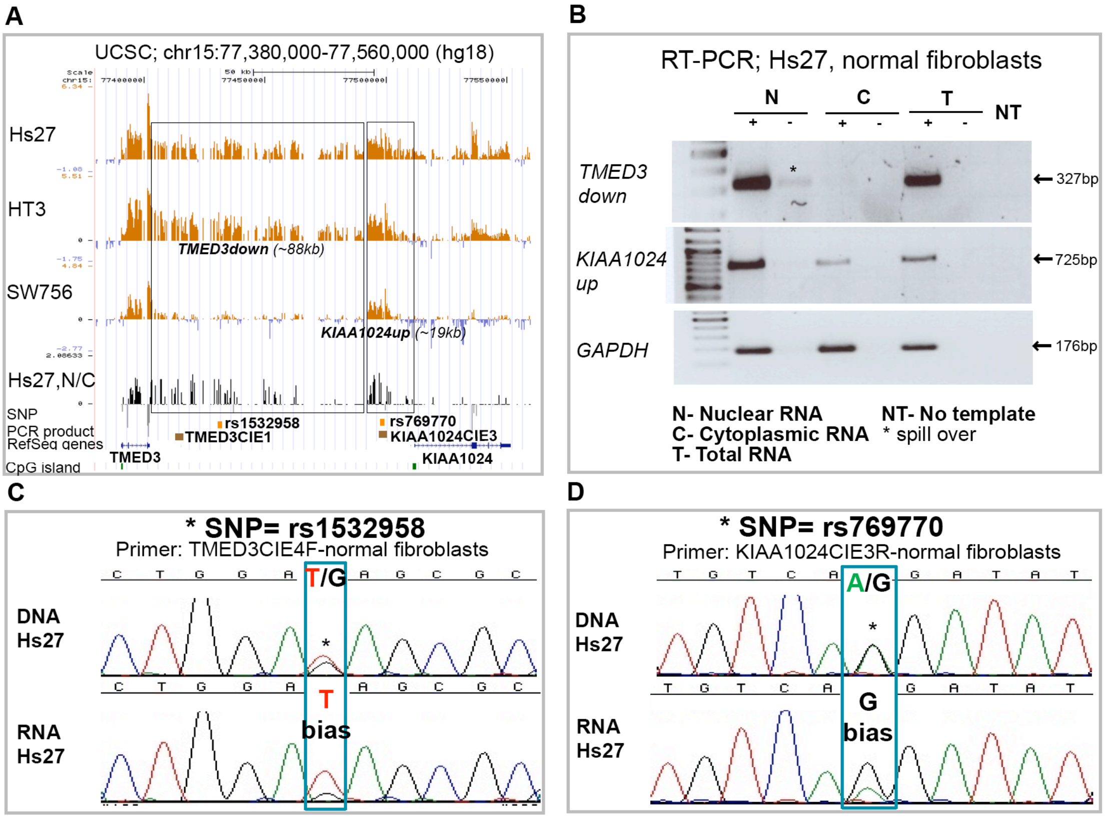


Figure 39. *TMED3down* and *KIAA1024up* are nuclearly enriched macro ncRNA candidates that show preferential expression from one allele in fibroblasts. A. Expression of *TMED3down* (~88kb in length) and *KIAA1024up* (~19kb in length) in fibroblasts and two cervical cancer cell lines and nuclear enrichment using double HIRTA hybridization is shown. Positions of mapped heterozygous SNPs and PCR products of primer pairs used for testing nuclear vs. cytoplasmic fractionations are shown. Tracks presented as on Figure 11B, section 2.1.5. and Figure 12B., section 2.1.6.**B.** RT-PCR with primer pairs TMED3CIE1 and KIAA1024CIE3 on nuclear, cytoplasmic and total RNA from fibroblasts shows nuclear enrichment of *TMED3down,* while *KIAA1024up* is both nuclear and cytoplasm enriched. *GAPDH* has been used as a loading control. *; Spill over. **C.** *TMED3down* shows expression bias towards one allele in fibroblasts using the rs1532958 heterozygous SNP. For detailed description look at the Figure 28B, section 2.4.5.1. **D.** *KIAA1024up* is preferentially expressed from one allele in fibroblasts using the rs769770 heterozygous SNP.

2. 4. 5. 7. HIRTA REG NNAT (chr20): *BLCAPov* macro ncRNA characterization

Human HIRTA region NNAT, paternally expresses *NNAT* (Neuronatin) and *BLCAP* v2a (Bladder cancer-associated) protein-coding genes while other variants of *BLCAP*: v1a, b and c are maternally expressed (Evans et al., 2001; Schulz et al., 2009).

1) Length of macro ncRNA

The only macro ncRNA candidate I mapped in the NNAT region was the ~36kb long *BLCAPov* transcript that was potentially overlapping and mapping also downstream of *BLCAP* protein coding gene.

2) Tissue specific expression and subcellular localization

The *BLCAPov* was expressed in 40/43 tested samples (Figure 27, 2.4.1). Double cDNA hybridization of nuclear versus cytoplasmic Hs27 RNA together with RT-PCR (using primer pair BLCAPCIE3) on fractionated fibroblasts cells showed both nuclear and cytoplasmic localization of *BLCAPov* (Figure 40A, B). Double cDNA hybridization showed cytoplasmic enrichment of three consecutive probes (typical for exons), thus we cannot exclude that *BLCAPov* represents part of the *BLCAP* gene.

3) Annotation

Taking into consideration the 5’-3’ slope through its’ gene body and 6/9 cell Broad lines that showed enrichment of H3K4me3, *BLCAPov* was predicted to be transcribed from minus (-) strand (same as *BLCAP*) (Figure 40A). However, a linc RNA expressed in Hs27 and HeLa partially maped to *BLCAPov* and indicated that this region is a complex transcriptional unit consisting of a number of overlapping protein coding and non-coding transcripts (Figure 40A).

4) CpG island features

BLCAPov did not show presence of potential CpG island promoter (Figure 40A).

5) Monoallelic/Imprinted expression of macro ncRNA

10 tested primer pairs (BLCAPCIE1-10) failed to identify a heterozygous SNP in the Hs27 fibroblasts cell line. Thus, I used the HapMap data (Frazer et al., 2007) from GM108046 lymphoblastoid cell line where the authors found both the A and G alleles present in DNA at the position of the A-2275664 SNP. I used primer pair BLCAPCIE11 covering this SNP and by sequencing of the RT-PCR product from both forward and reverse primers I showed *BLCAPov* biallelic expression in the lymphoblastoid cell line.

In summary, *BLCAPov* was ~36kb long, both nuclear and cytoplasm enriched transcript showing biallelic expression in a lymphoblastoid cell line.

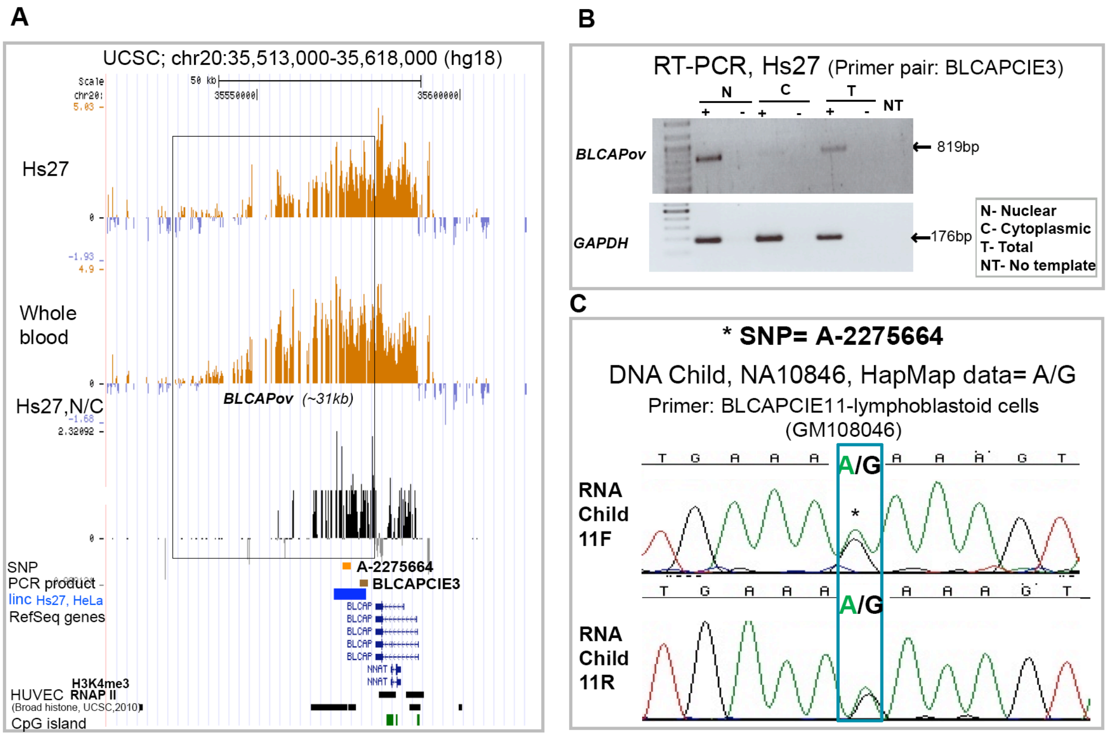


Figure 40. *BLCAPov* is a nuclear localized biallelic macro ncRNA candidate. A. Expression of *BLCAPov* in fibroblasts and whole blood using single HIRTA hybridizations and enrichment in fibroblasts nuclear fraction using double HIRTA hybridization is shown. *BLCAPov* partially overlaps a lincRNA expressed in Hs27 and HeLa (blue colored linc Hs27, HeLa track). H3K4me3 and RNAP II enrichments from HUVEC cell line (Broad Histoane modifications, UCSC track) in this region are shown. Positions of a SNP (A-2275664) heterozygous in GM108046 lymphoblastoid cell line and PCR primer product used for testing of nuclear versus cytoplasmic localization (BLCAPCIE3) are shown. Tracks presented as on Figure 11B, section 2.1.5. and Figure 12B., section 2.1.6. **B.** RT-PCR confirms nuclear enrichment of *BLCAPov* already shown using HIRTA double cDNA hybridization. 819bp product from primer pair BLCAPCIE3 shows nuclear enrichment while very faint band (*) could also be detected in cytoplasm. *GAPDH* has been used as a loading control where the expected 176bp band was observed in both fractions and total RNA from Hs27. **C.** *BLCAPov*

is biallelic since A and G alleles are both present in heterozygous SNP (A-2275664) from the HapMap data of Child DNA (NA10846) and in RNA from lymphoblastoid cell line originating from the same Child (GM108046). Sequencing from both forward and reverse primer BLCAPCIE11 is shown.

2. 4. 5. 8. Summary of characteristics for ten examined macro ncRNAs

Ten macro ncRNAs from 7 gene regions containing imprinted genes were characterized. These transcripts had lengths in a range from 6 to 463kb. 2/10 samples showed ubiquitous expression in all 43 tested cells/tissues while 8 were tissue specifically expressed. From 10 tested candidates that were all expressed in Hs27, 4 were enriched in the nuclear fraction, 5 were present in both nuclear and cytoplasmic and one was cytoplasmically localized by RT-PCRs or HIRTA double nuclear versus cytoplasm cDNA hybridizations. Three out of ten selected macro ncRNA candidates were mapping to previously published linc (large intergenic non-coding) RNAs (Khalil et al., 2009).

6/10 candidates had CpG islands, while 9/10 showed presence of H3K4me3 peaks on one of the transcript ends. The methylation of CpG islands promoters was tested on 5 candidates and 3 CpG island promoters were found to be unmethylated while two (*PEG13* and *ADAMTS7down*) showed the presence of Differentially Methylated Region (DMR) in tested cell lines. Direct repeats were present in 4/5 tested CpG island promoters of macro ncRNA candidates.

Characterization of 10 macro ncRNA candidates showed that *ADAMTS7down* was exclusively monoallelically expressed while 5 other candidates showed biased expression and 4 were found to be biallelically expressed from tested cells and/or tissues. *SLC38A4down2* macro ncRNA candidate was found to be paternally expressed from the blood.

2. 5. Each human imprinted gene region express macro ncRNA

2. 5. 1. Overview of macro ncRNAs in human imprinted gene regions

Previous studies have shown that all human well-studied imprinted gene regions contain macro ncRNAs (Table 4 and Figure 6, section 1.2.5.1.). To see whether macro ncRNAs could be a universal feature of all human imprinted gene regions 43 different cells/tissues/patients were hybridized and novel macro ncRNAs in 32 HIRTA regions were mapped. I showed that each of the 32 studied HIRTA regions contains at least one known or candidate macro ncRNA in human normal and cancer tissues

(Figure 41). 18/32 HIRTA regions that were previously shown to contain a macro ncRNAs were grouped into well-studied (*H19, KCNQ1OT1, GTL2, UBE3A-AS, GNASAS, EXON1A, XIST and TSIX*) and other known (*HYMAI, AIRN, LOC100129427, DLX6AS, MIT1, MESTIT1, FLJ4663, 91H, IGF2AS, LOC650368, LOC100133545, WT1AS, LOC100233209, NCRNA00239, PWRN1, LOC145663, PEG3AS, NCRNA00028, NCRNA00182 and NCRNA00183*) macro ncRNAs. *GTL2* and *UBE3A-AS* are highly complex transcriptional units that contain a number of known overlapping ncRNAs of different length that were previously extensively studied and were not considered separately in this analysis. 27/30 known macro ncRNAs were detected by HIRTA while 3 were not (*TSIX, HYMAI and AIRN*). 101 novel macro ncRNA candidates in 32 HIRTA regions from 43 tested cells/tissues/patients were mapped. The results illustrate that regions with a high number of macro ncRNAs are: KCNQ1 and PWS with 9 macro ncRNAs each, and SFMBT2, BAZ2 and IGF2, with 8 ncRNAs each. Regions with only one mapped macro ncRNA candidate were: TRP73, NAP1L5, IMPACT and NNAT. In total, 131 known or candidate macro ncRNAs were expressed from 32 HIRTA regions.

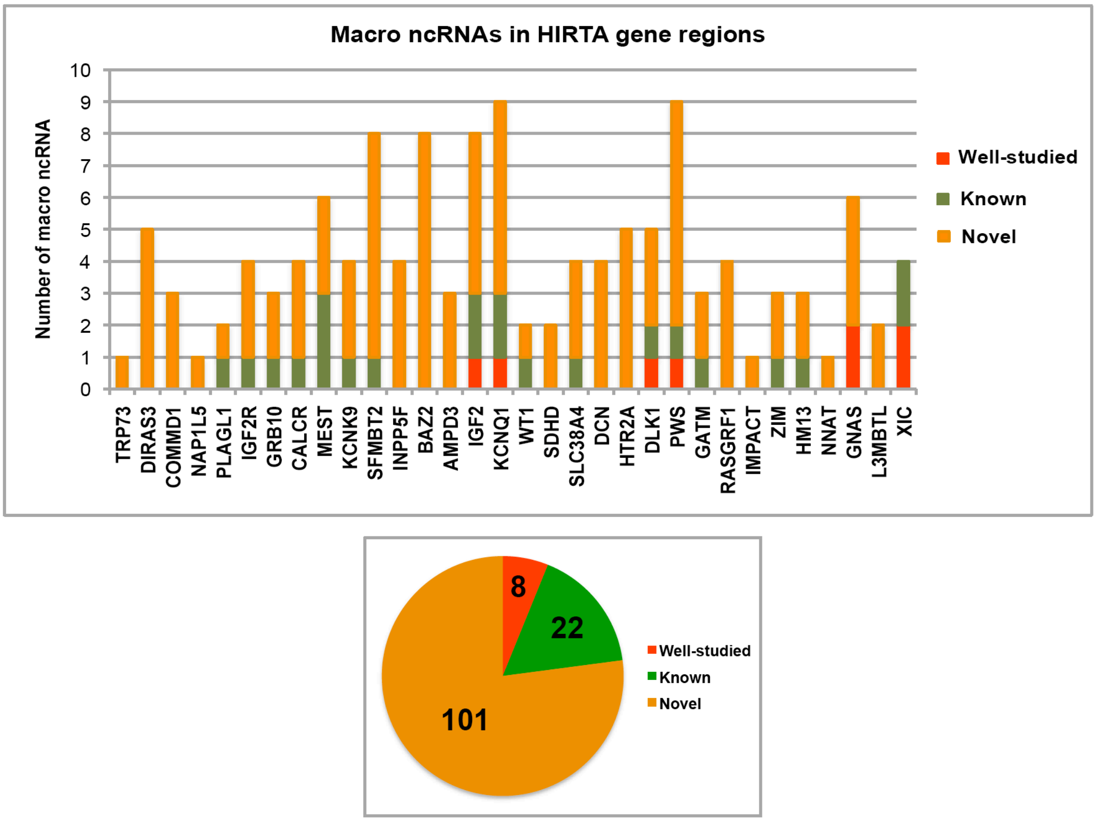


Figure 41. Distribution of known and novel macro ncRNAs from 32 HIRTA regions. Number of novel macro ncRNA candidates per HIRTA region (TRP73 to XIC region) is shown in orange (101 candidates), while we grouped known macro ncRNAs in two groups (well-studied; red color (8), and known; green color (22)).

Novel macro ncRNA candidates were grouped according to their position in accordance with RefSeq and UCSC annotated genes. 55/101 novel transcripts were positioned intergenically which means they did not overlap any annotated gene, 12/101 were positioned 5' to an annotated gene and had the potential to be novel 5'exons of annotated genes or distinct upstream ncRNAs that are overlapping these transcripts, 18/101 were mapped 3' to annotated transcripts and could represent novel 3'UTRs or macro ncRNAs that overlaped these transcripts. Further, 2/101 were positioned inside of an intron of an annotated gene, while 14/101 overlapped one or more short genes, for example long transcripts overlapping olfactory receptor genes (Figure 42).

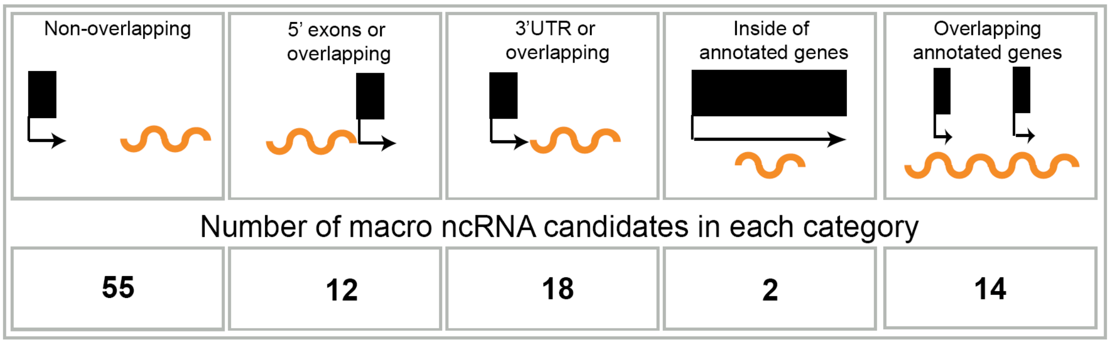

Non-overlapping
5' exons or overlapping
3'UTR or overlapping
Inside of annotated genes
Overlapping annotated genes
Number of macro ncRNA candidates in each category
55
12
18
2
14

Figure 42. Grouping of novel macro ncRNA candidates in accordance with positions of annotated genes. 5 groups of positioning could be distinguished with a certain number of candidates mapping to these categories: non-overlapping (55/101), 5'exons or overlapping (12/101), 3'UTR or overlapping (18/101), inside of annotated genes (2/101) and overlapping annotated genes (14/101).

The *GNASAS* macro ncRNA expressed from the GNAS region, *GTL2var1* from the DLK1 region and *UBE3A-ASvar1* from PWS imprinted gene regions were overlapped by a large intergenic noncoding (linc) RNA published by Khalil et al. (Khalil et al., 2009). Comparison of 101 macro ncRNA candidates mapped from HIRTA hybridizations with large intergenic noncoding (linc) RNA showed that 9/101 transcripts were previously mapped in fibroblasts and HeLa cells. From these transcripts, two were fully overlapped with linc RNAs while 7 others were partially overlapped. 5/9 transcripts that overlap between our study and previously published linc RNAs were associated with PRC2 and CoREST complexes (Khalil et al., 2009).

2. 5. 2. Overview of tissue specific expression of macro ncRNA candidates

20 normal cells/tissues were hybridized to the HIRTA Chip using the single cDNA hybridization technique. These samples included undifferentiated and differentiated embryonic stem cells (HES2), 3 fetal tissues, placenta as extra embryonic tissue and 13 adult tissues. With 29 novel mapped macro ncRNAs in 32 HIRTA regions, testis

was a tissue with the highest number of macro ncRNA candidates (Figure 43A). Further, undifferentiated ES cells had 28 and adult uterus, 27 candidates. Tissues showing 10 or less macro ncRNA candidates were: fetal liver, placenta, adult lung and adult heart.

HIRTA was hybridized with 23 cancer samples including 9 cervical, 2 colon, 2 breast, 1 teratocarcinoma, 1 rhabdomyosarcoma, 1 neuroblastoma cell line and 6 patients (4 AML (Acute lymphoid leukemia) and 2 MPD (Myeloproliferative disorder)). Overall, 20/23 samples expressed more than 15 novel macro ncRNA candidates with the highest number in C4II, one of the cervical cancer cell lines (25) (Figure 43B). On average, 18.65 candidates were expressed per normal and 19 candidates per tested cancer sample showing no overall difference in number of expressed macro ncRNA candidates in imprinted gene regions in human between normal and cancer samples.

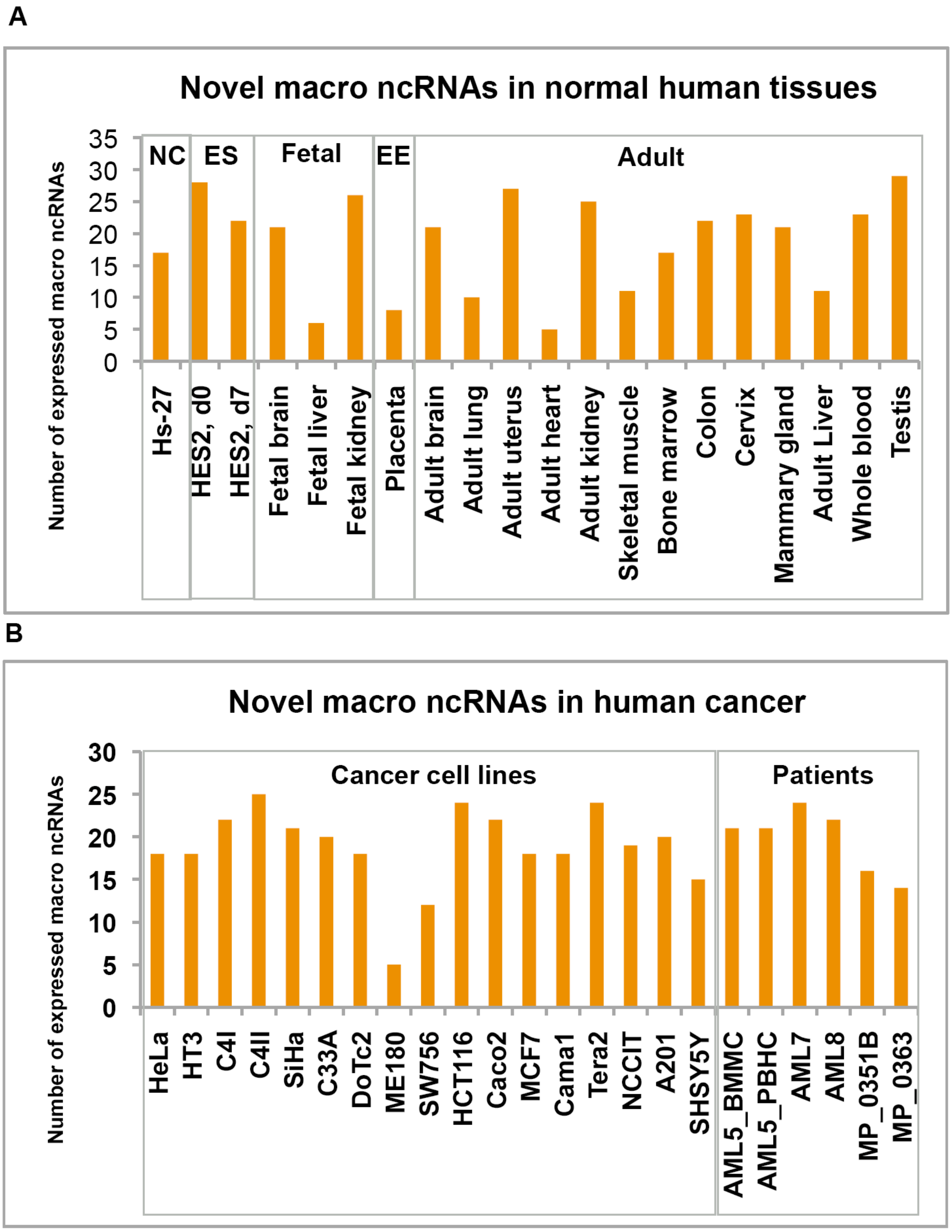


Figure 43. Distribution of novel macro ncRNAs in normal and cancer cells/tissues. A. Number of macro ncRNA candidates in 20 normal human cells/tissues are shown. NC; normal cell line, ES; embryonic stem cells, EE; extraembryonic tissue **B.** Number of macro ncRNA candidates in 23 human cancer cells/patients is presented.

2. 5. 3. Strand prediction of macro ncRNA candidates

To determine the expression of known and novel transcripts from imprinted gene regions I used single cDNA hybridizations to the HIRTA tiling array. We co-hybridized double-stranded cDNA and sonicated DNA, to the HIRTA Chips and mapped 101 novel macro ncRNA candidates in 43 tested samples. As this approach did not allow the orientation of transcription to be determined for these candidates I developed a novel approach in order to predict the strand of the novel macro ncRNA candidates. This approach was based on the usage of 4 features (Tabe 16). The CpG islands (CpG island track, UCSC (Gardiner-Garden and Frommer, 1987)), H3K4me3 and RNAP II enrichments (Broad Histone tracks, UCSC; ENCODE Histone Modifications by Broad Institute ChIP-Seq, Peaks from 9 cell lines for H3K4me3 and 3 cell lines for RNAP II) were used as indicating of promoters. To these three criteria feature based on the reverse correlation between the SLOPE and the strand, was added (Figure 19, 2.2.1.) and to each criteria a relative value was assigned (Table 16).

Presence of Overlapping Feature	Given Value (Sum= -4 to +4)
CpG island inside of +/- 1kb from transcript start or end	+1 or -1
SLOPE value negative or positive	+1 or -1
H3K4me3 peak inside of +/- 1kb from transcript start or end (9 cell lines)	+1 (for each cell line positive on the start add +0.11) or -1 (for each cell line positive on the end add –0.11)
RNAP II peak inside of +/- 1kb from transcript start or end (3 cell lines)	+1 (for each cell line positive on the start add +0.33) or -1 (for each cell line positive on the end add –0.33)

Table 16. Four criteria used in prediction of macro ncRNA candidates orientation and their assigned values. Each feature has value +1 if present on proximal or -1 if present on distal end of the novel transcript.

The starting hypothesis was that each criterion has the same value (+1 if present on the position +/- 1kb from the proximal end of the transcript and -1 if present on the position +/- 1kb from the distal end of the transcript) and thus showed equal prediction importance. Thus, the maximal sum from the criteria values was +4 if the transcript showed plus (+) strand orientation or -4 if the transcript showed minus (-) strand orientation. In order to calculate the Score depicting likelihood of the transcript orientation prediction I used the value of: (sum of criteria values)/4. Thus +1 was the score showing highest probability that transcript is plus (+) strand while -1 is the highest probability that the transcript had minus (-) strand orientation based on used criteria. Further, the scores were grouped into: high probability score (+/- 0.5 to +/- 1), medium probability (+/- 0.25 to +/- 0.5) and low probability (- 0.25 to + 0.25). This approach does not allow potential novel overlapping transcripts transcribed in the opposite directions to be distinguished since for these transcripts low probability for each strand is expected. Thus, a separate group of transcripts was formed (marked with *) that were showing a low/medium probability prediction for certain transcription orientation, but at the same time were overlapping at least one tested feature on each end. I first tested this approach on 6 well-studied macro ncRNAs from imprinted gene regions and found that prediction based on the four used criteria's matched the known strands of all these transcripts (Table 17).

Well-known imprinted Macro ncRNAs	CpG island	Reverse SLOPE	H3K4me3, proximal	RNAP II, proximal	H3K4me3, distal	RNAP II, distal	SUM	Score = SUM/4	Predicted orientation	Known orientation
H19	-1	-1	0.11	0	-0.33	0	-2.22	-0.555	-	-
KCNQ1OT1	-1	-1	0	0	-0.99	-0.66	-3.65	-0.912	-	-
GTL2var1	1	1	0.55	0.33	-0.11	-0.33	2.44	0.61	+	+
UBE3A-ASvar1	1	1	0.88	0.66	0	0	3.54	0.885	+	+
GNASAS	-1	-1	0	0	-0.77	0	-2.77	-0.692	-	-
EXON1A	1	-1	0.99	0.99	0	-0.99	0.99	0.247	+	+

Table 17. Predicted orientations of 6 well-known macro ncRNA are matching to known orientations using approach based on four epigenetic/genomic features that associate with promoters. Calculations are done according to the scheme described in Table 16.

Interestingly, a low probability prediction also matched to the known transcription orientation as shown on the example of the well-known *EXON1A* macro ncRNA. Further, we predicted the strands for 101 novel macro ncRNA candidates (Table 18) where distribution of candidates in each category is shown on Figure 44.

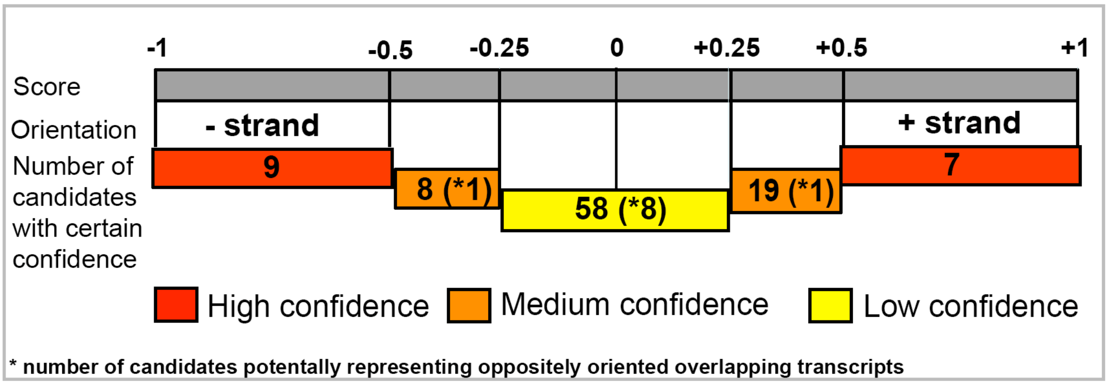

-1
-0.5
-0.25
0
+0.25
+0.5
+1
Score
Orientation
- strand
+ strand
Number of candidates with certain confidence
9
8 (*1)
58 (*8)
19 (*1)
7
High confidence
Medium confidence
Low confidence
* number of candidates potentally representing oppositely oriented overlapping transcripts

Figure 44. Orientation is confidently predicted for 43 macro ncRNA candidates. Number of candidates per each of three prediction categories is shown. Nine candidates that are predicted to represent overlapping transcripts of opposite orientations are marked with star (*).

Macro ncRNA candidate	Score	C	P. S.	Macro ncRNA candidate	Score	C	P. S.	Macro ncRNA candidate	Score	C	P. S.
SLC22A2up	0.222	L	+	*B3GNT2down1*	0.387	M	+	*OR5P2ov*	-0.25	L	-
MAS1down	0.25	L	+	*B3GNT2down2*	0.195	L	+	*AMPD3up*	-0.302	M	-
MAS1ov	0.222	L	+	*FAM161Adown*	-0.995	H	-	*LYVE1ov*	0.495	M	+
H19down	0.582	H	+	*TIGD2down*	0.305	M	+	*MRVI1up*	-0.5	M	-
H19up	0.25	L	+	*PHACTR2ov*	-0.025	L	-*	*WIT1down*	0.555	H	+
ASCL2ov	-0.25	L	-	*COBLdown1*	0.25	L	+	*SDHDdown*	-0.167	L	-
ASCL2up	-0.222	L	-	*COBLdown2*	-0.25	L	-	*PTSdown*	0.36	M	+
TSPAN32down1	-0.385	M	-	*DLX6up1*	-0.25	L	-	*SLC38A4down1*	0.385	M	+
ZNF195down1	0.222	L	+	*DLX6up2*	-0.692	H	-	*SLC38A4down2*	0.995	H	+
ZNF195down2	-0.222	L	-	*COL1A2up*	-0.305	M	-	*SLC38A4up*	-0.25	L	-
ZNF195down3	0.222	L	+	*KLF14up1*	0.25	L	+	*EPYCov*	0.277	M	+
ZNF195down4	0.25	L	+	*KLF14up2*	-0.607	H	-	*DCNup1*	0.25	L	+
ZNF195up1	-0.167	L	-	*KLF14up3*	-0.222	L	-*	*DCNup2*	0.25	L	+
ZNF195up2	0.14	L	+	*PEG13*	-0.082	L	-	*DCNup3*	-0.222	L	-
BEGAINup	0.085	L	+*	*PTK2up*	0.332	M	+	*LRCH1up1*	-0.25	L	-
PPP2R5Cup1	-0.275	M	-*	*KIAA1126ov*	-0.08	L	-*	*LRCH1up2*	0.25	L	+
PPP2R5Cup2	-0.967	H	-	*PFKFB3down*	0.332	M	+	*HTR2Aup1*	0.387	M	+
NIPA1up1	-0.137	L	-*	*SFMBT2down1*	0.222	L	+	*HTR2Aup2*	0.387	M	+
WHAMML1up	0.995	H	+	*SFMBT2down2*	0.415	M	+	*HTR2Ain*	0.25	L	+
WHAMML1up1	0.25	L	+	*SFMBT2down3*	-0.277	M	-	*SQRDLdown*	0.912	H	+
SNRPNup1	-0.25	L	-	*SFMBT2down4*	-0.25	L	-	*C15orf43up*	-0.14	L	-
SNRPNup2	0.25	L	+	*SFMBT2down5*	0.25	L	+	*ADAMTS7down*	0.472	M	+*
GABRB3down1	-0.25	L	-	*GATA3down*	0.222	L	+	*TMED3down*	0.332	M	+
GABRB3down2	-0.25	L	-	*BAG3down*	-0.72	H	-	*KIAA1024up*	-0.22	L	-*
APCDD1Lup1	0.39	M	+	*SEC23IPdown1*	-0.25	L	-	*CHRNB4up*	0.25	L	+
APCDD1Lup2	0.25	L	+	*SEC23IPdown2*	0.25	L	+	*ZNF521up*	0.332	M	+
ZNF831up1	-0.582	H	-	*SEC23IPdown3*	-0.25	L	-	*ZNF71down1*	0.5	M	+
ZNF831up2	-0.14	L	-	*OR556B4up1*	0.25	L	+	*ZNF71down2*	0	-	No
LRRC47down	0.247	L	+*	*OR56A1down*	0.25	L	+	*ID1up1*	0.967	H	+*
GADD45Aup	0.222	L	+	*PKCDBPup*	0.885	H	+	*ID1up2*	-0.247	L	-
GPR177up	-0.112	L	-	*PKCDBPdown*	-0.885	H	-	*BLCAPov*	-0.747	H	-
RPE65ov	0.277	M	+	*HPXdown*	-0.167	L	-	*L3MBTLup*	0.305	M	+
RPE65down1	-0.497	M	-	*ZNF215up*	-0.692	H	-	*TOX2up*	-0.277	M	-
RPE65down2	0.387	M	+	*RBMXL2down*	0.25	L	+				

Table 18. Strand of 100 macro ncRNA candidates is predicted based on four epigenetic/genomic features. Score for each candidate and predicted strand (P. S.) is shown. Strand of *ZNF1down2* could not be predicted since the score was equal to zero. Candidates that could represent two overlapping transcripts expressed in opposite directions are marked by *. C; confidence, H; High, M; Medium, L; Low.

2. 5. 4. Non-coding potential of macro ncRNA candidates

To further examine the non-coding potential of macro ncRNA candidates we used RNAcode (Washietl et al., 2010, unpublished data) that predicts protein-coding regions based on evolutionary signatures. Jan Engelhart, a visiting MSc student from Peter Stadlers' laboratory, Institut für Informatik, Universität Leipzig, conducted RNAcode over 44 vertebrate species using Multiz Align from UCSC and obtained whole genome predictions of protein-coding regions. RNAcode was found to be highly reliable in mapping known protein coding exons, while there was no RNA code predictions mapping through the body of known macro ncRNAs from imprinted gene regions in human (in some cases 1-2 "orphan" RNAcode predicted exons were found). The "orphan" RNAcode predictions were short, positioned centrally in the body of ncRNA, alone or paired with one more exon (Lander et al., 2001). The "orphan" prediction could represent exon of overlapping protein-coding gene (this exon could splice to another distant exon, if non-annotated gene has introns longer than average), but it is unlikely to represent exons of tested macro ncRNA candidate.

Examples of typical protein coding genes: *DCN* and *KCNQ1* (Figure 45A, B), and typical macro ncRNAs: *KCNQ1OT1* and *H19*, are shown (Figure 45B, C).

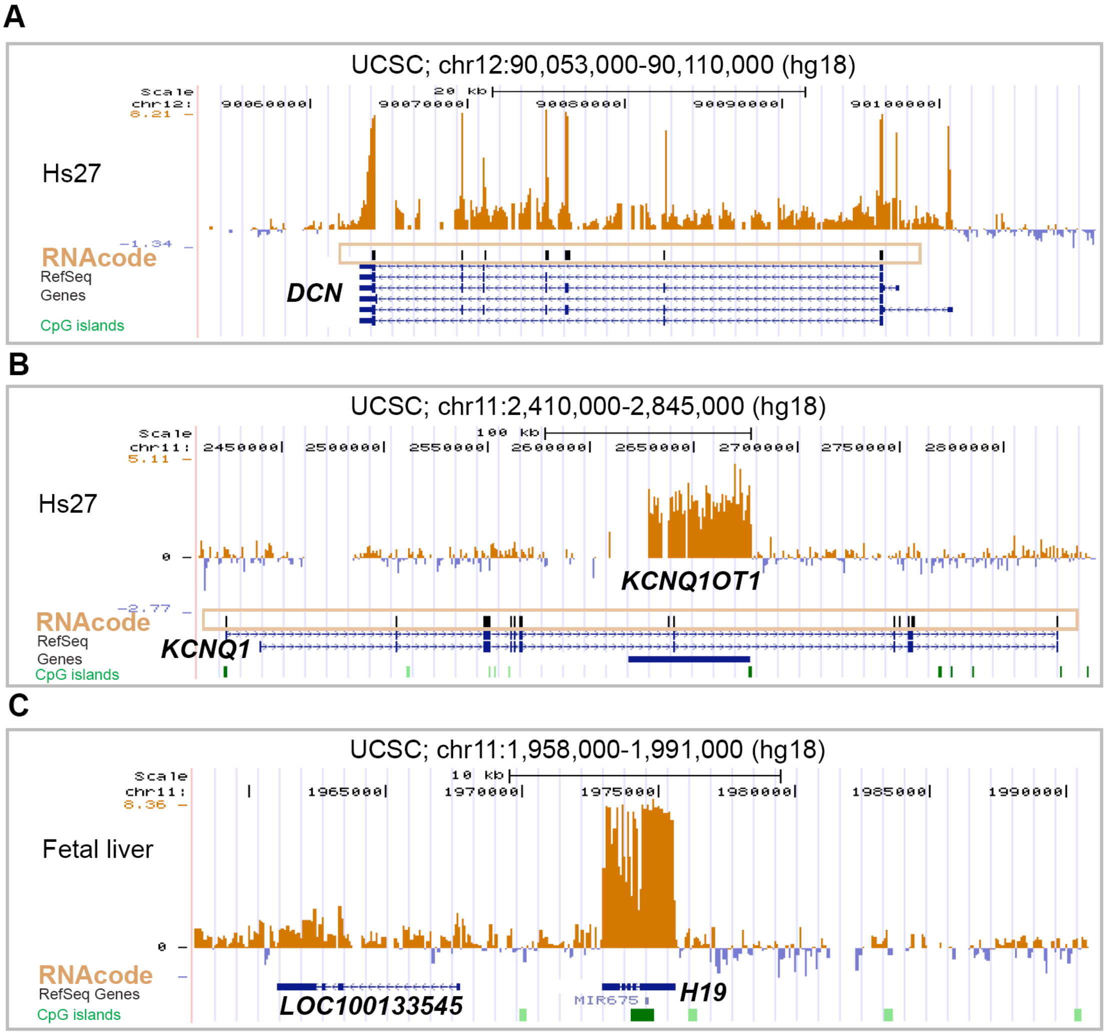


Figure 45. RNAcode protein-coding predictions match to exons of known protein coding genes and do not find any protein coding potential in known non-coding RNAs. A. RNAcode mapped 7/9 known *DCN* exons. **B.** RNA code mapped 10/10 known and two novel exons for the *KCNQ1* protein-coding gene, while no exons mapping to the start and end of the *KCNQ1OT1* ncRNA were mapped. **C.** RNAcode did not find any protein-coding region in *H19* and *LOC100133545* ncRNAs. RNA code predictions are presented as black filled boxes and are highlited with light brown boxes.

I further analyzed results of the RNAcode predictions over 101 macro ncRNA candidates from imprinted gene regions in human. I found that 89/101 macro ncRNA transcripts mapped from HIRTA expression data were predicted to be non-protein coding since no RNAcode output could be found through the body of mapped genes. From these candidates 63 had no overlap with any of RNA code predictions (group 1) while for 27 others, part overlapped RNAcode exons matching known overlapping protein coding genes (group 2- macro ncRNA candidates overlapping annotated genes, Figure 42, 2.5.1) or 1-2 short exons ("orphans") were detected located in the central part of a mapped ncRNA gene (group 3) (examples of each of the three groups are shown in Figure 46).

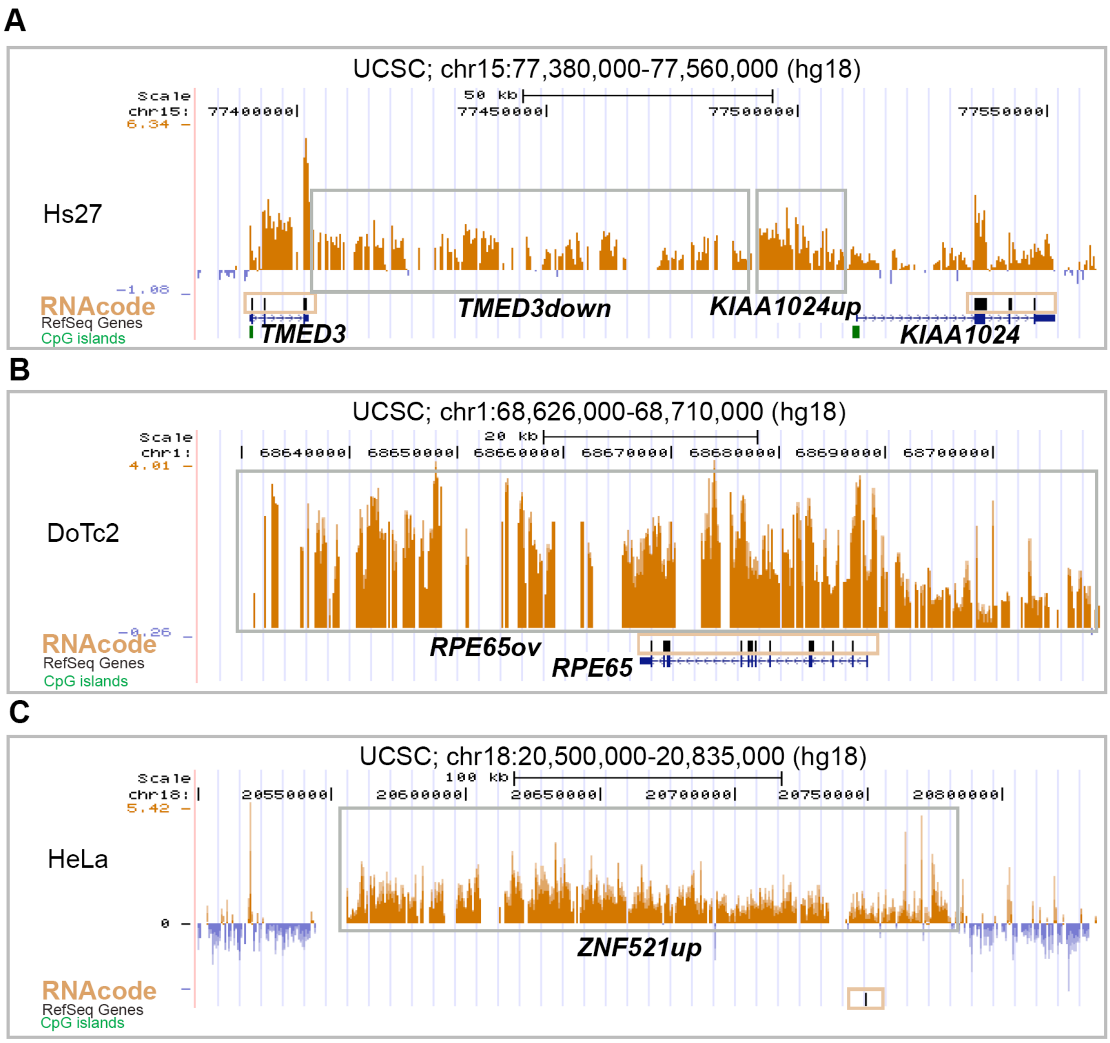


Figure 46. Examples of HIRTA macro ncRNA candidates confirmed to be non- protein coding using RNAcode. A. *TMED3down* and *KIAA1024* are 2/63 macro ncRNAs that do not overlap any RNAcode protein-coding region. **B.** RNAcode finds exons of annotated *RPE65* protein-coding gene while does not find any other region mapping to *RNA65ov* macro ncRNA. **C.** *ZNF521up* is an example of ncRNA that overlaps with 1 “orphan” RNAcode region.

The 12 macro ncRNAs that were found to have protein-coding potential could be also further subgrouped. The first subgroup included those where potentially both a non-protein coding and coding RNA could be present in the same time in the mapped HIRTA region (7 candidates: *OR556B4up1, OR56A1down, OR5P2ov, AMPD3up, SLC38A4down2, SQRDLdown, CHRNB4up*). The second subgroup of 5 candidates: *ZNF195up1* (RNAcode=9), *ZNF195up2* (RNAcode=4), *SNRPNup2* (RNAcode=7), *RPE65down1* (RNAcode=5) and *SLC38A4up* (RNAcode=3) had protein-coding potential with RNAcode regions positioned at distances resembling typical introns (Lander et al., 2001) and could potentially present novel exons of novel genes (number of RNAcode regions mapping to candidates shown in brackets) (examples are shown on Figure 47). In summary, 96/101 mapped HIRTA candidates are macro ncRNAs that are intergenic or overlap known or potentially novel protein-coding genes, while 5/101 candidates have protein-coding potential. Thus, RNAcode indicated that the HIRTA expression macro ncRNA mapping approach had 4.5% false positives while 95.5% of the HIRTA mapped macro ncRNAs were predicted to be non-coding.

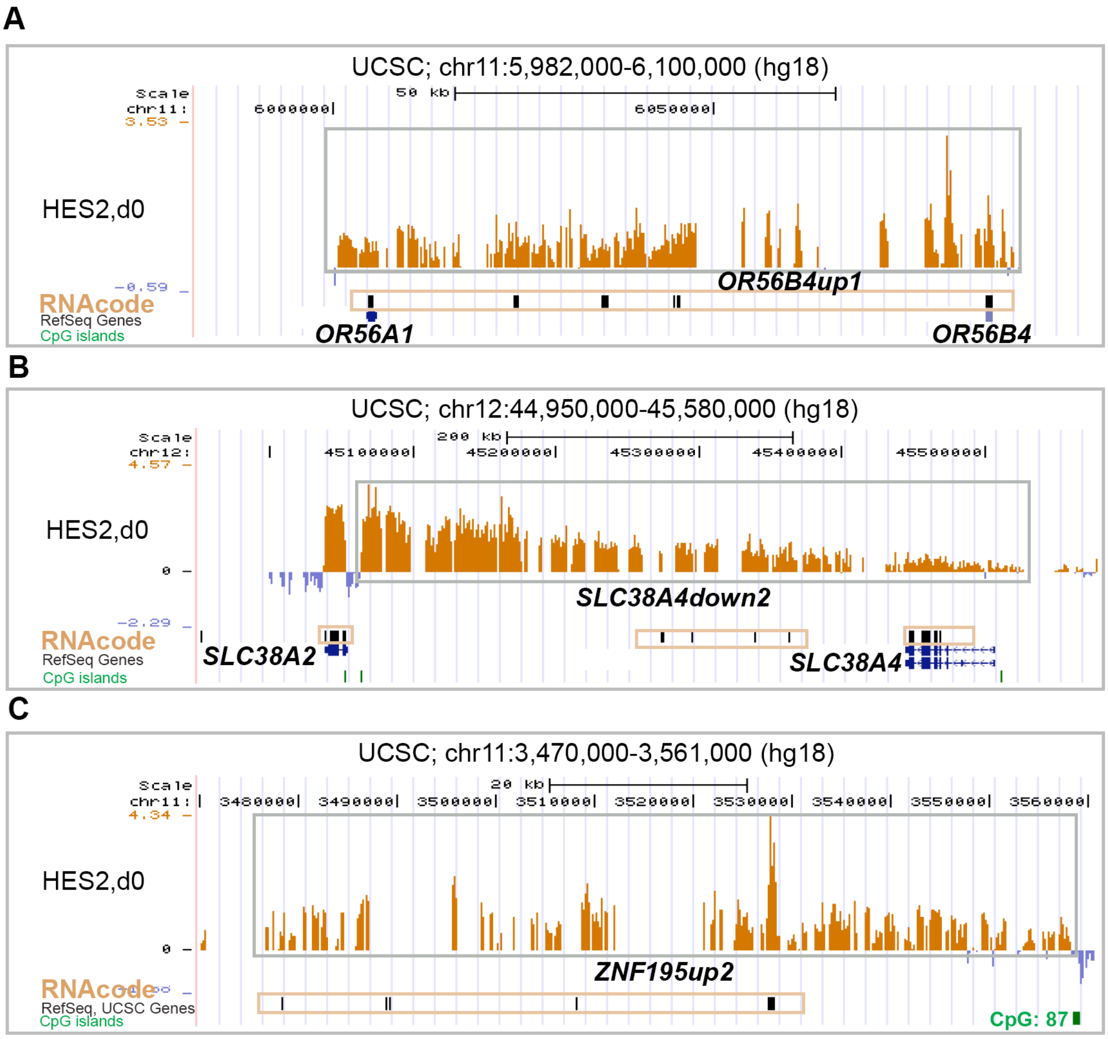


Figure 47. Examples of 3/12 HIRTA candidates with both non-coding and coding, or protein-coding potential alone. A. RNAcode predicts two known olfactory genes and three new protein-coding regions that could to also represent olfactory receptor genes, while *ORF56B4up1* macro ncRNA positioned in this region could be a precursor or it may overlap short olfactory receptor genes. **B.** RNAcode finds known exons of *SLC38A4* and *SLC38A2* protein-coding genes and 4 regions overlapping *SLC38A4down2* showing that part of this macro ncRNA could be overlapped with a novel protein coding gene. **C.** RNAcode regions are potential novel exons of *ZNF195up2* transcript showing the protein-coding potential of this transcript.

2. 6. Macro ncRNAs from imprinted regions are deregulated in cancer

Macro ncRNA expression can be deregulated in cancer through genetic and epigenetic mechanisms as introduced in sections 1.2.9.2 and 1.3.3. Macro ncRNA deregulation in cancer could be valuable for prognostics since macro ncRNAs have the potential to be used as biomarkers (introduced in section 1.3.4). Among 20 normal and 23 tumor samples, cancer cell lines/patients matching normal tissues (from which the cancer originated), were distinguished. Thus, it was possible to correlate expression in mammary gland and two breast cancer cell lines (MCF7 and Cama1), cervix and 9 cervical cancer cell lines, colon and two colon cancer lines (HCT116 and Caco2), skeletal muscle and rhabdomyosarcoma (A201) line, and whole blood or bone marrow and 6 AML/MPD patients.

2. 6. 1. Well-studied macro ncRNAs are deregulated in cancer

Imprinted macro ncRNAs can be deregulated in cancer through loss of imprinted expression (LOI). Examples of *H19* upregulation in bladder and hepatocellular carcinoma and *IGF2AS* overexpression in Wilms' tumor were introduced in section 1.2.9.2. Analysis of expression of 6 well-known macro ncRNAs (*H19, KCNQ1OT1, GTL2, UBE3A-AS, GNAS* and *EXON1A*) in pairs of normal and the corresponding cancer cell lines/patients was performed.

An overview of expression of 6 well-known macro ncRNAs is shown in the Figure 17, section 2.2.1.1. and here details of the observed deregulation of the ncRNAs are presented. In the IGF2 imprinted gene region, the *H19* macro ncRNA was expressed in cervix, colon and mammary gland while 3/9 cervical cancer cell lines, 1/2 colon cancer and 1/2 breast cancer cell lines did not show expression of *H19*. We found *KCNQ1OT1* macro ncRNA to be a widely expressed ncRNA that was downregulated in 2/9 cervical cancer cell lines (Figure 48).

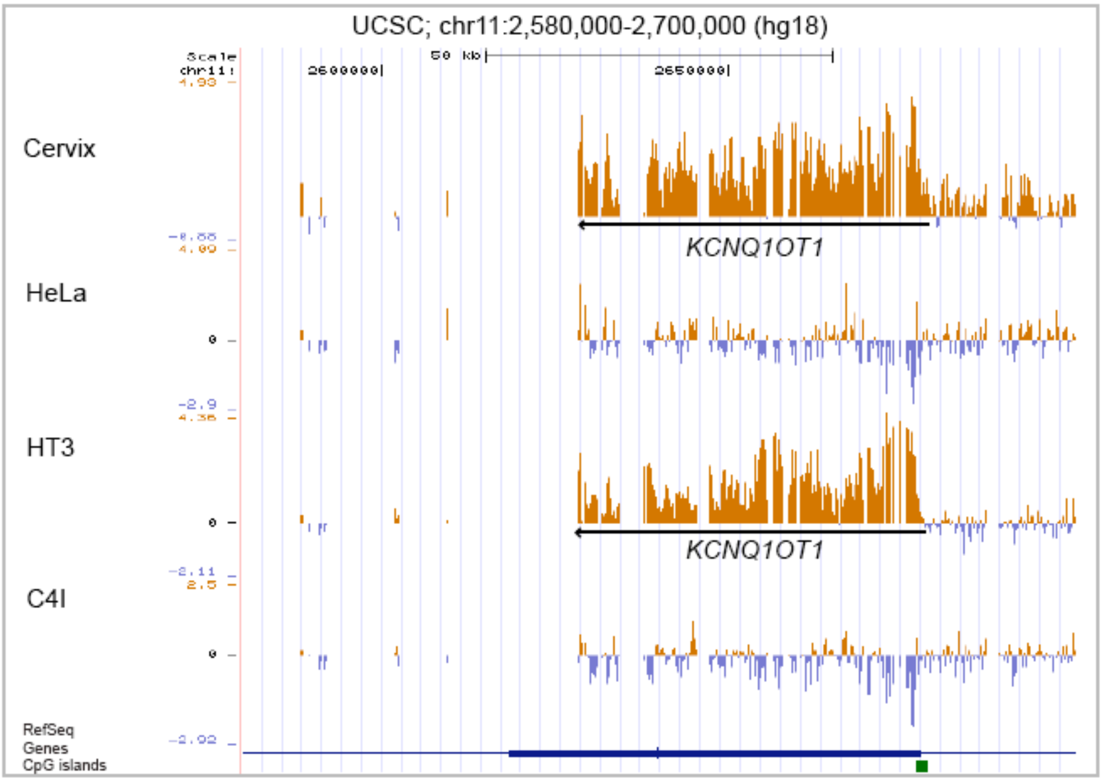


Figure 48. *KCNQ1OT1* ncRNA is downregulated in HeLa and the C4I cervical cancer cell line and expressed from normal cervix. HT3 and 6 more cervical cancer cell lines that are not shown express *KCNQ1OT1*. UCSC position shown on the top. Custom HIRTA tracks as well as RefSeq genes, sno/miRNA and CpG islands UCSC tracks are shown.

In the DLK1 HIRTA region, the *GTL2var1* macro ncRNA is expressed from cervix and mammary gland while in 9/9 cervical cancer cell lines and 2/2 breast cancer cell lines hybridized to HIRTA this macro ncRNA was fully downregulated (examples in Figure 49A). Interestingly, the same ncRNA was not expressed in skeletal muscle, but expressed in a A201 cell line (Figure 49B). One of the *GTL2* variants (*GTL2var6*) was characteristic for neuroblastoma cell line (Figure 17, section 2.2.1.1.).

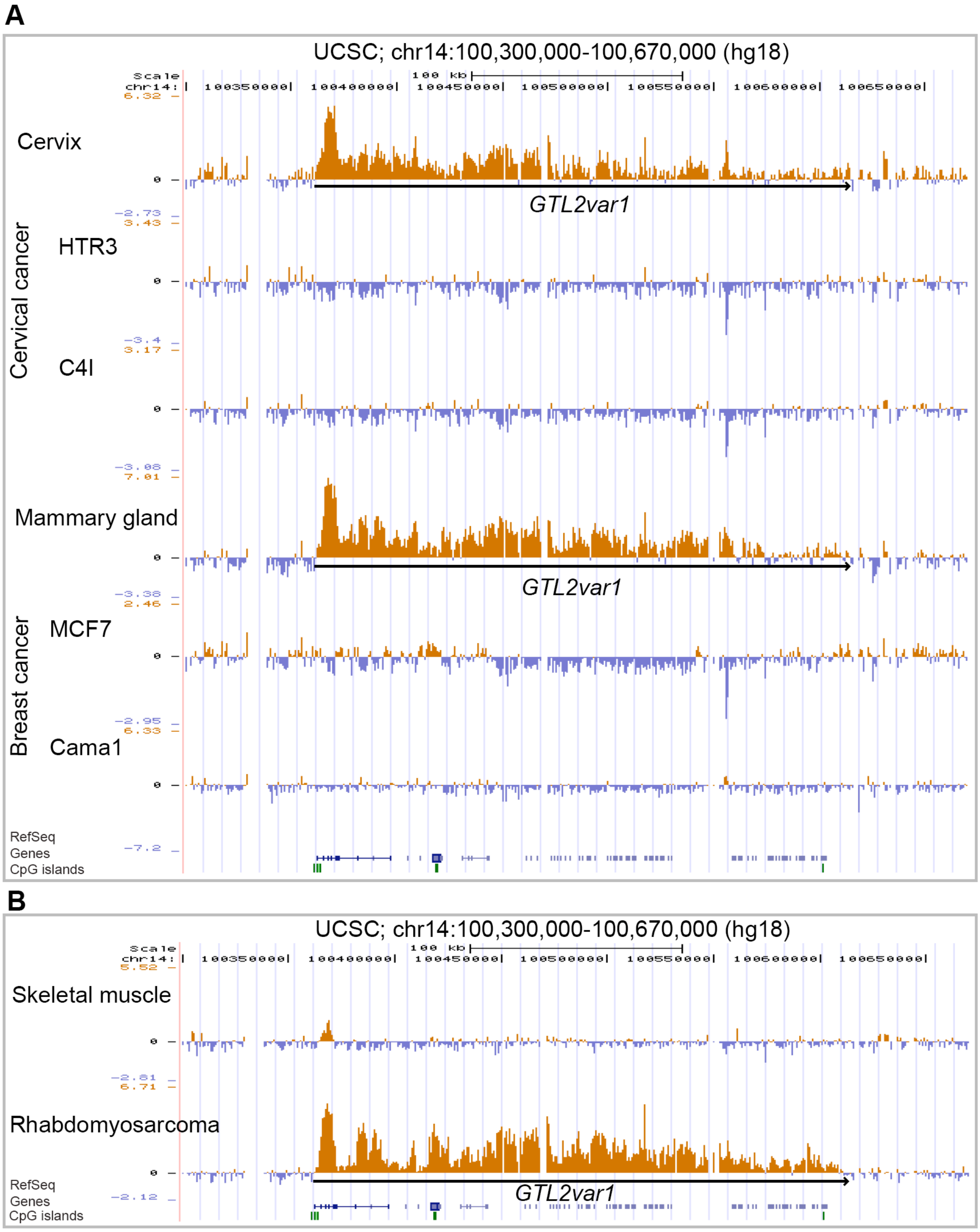


Figure 49. *GTL2var1* macro ncRNA is deregulated in cancer. A. *GTL2var1* expressed in normal cervix and mammary gland tissues is downregulated in cervical and breast cancers. **B.** *GTL2var1* shows upregulation in rhabdomyosarcoma (A201) compared to the skeletal muscle sample. UCSC position shown on the top. Custom HIRTA tracks as well as RefSeq genes, sno/miRNA and CpG islands UCSC tracks are shown.

UBE3-ASvar4 ncRNA from PWS imprinted gene region was expressed in diverse normal tissues, but not expressed in any of tested cancer cells. In the GNAS imprinted gene region, *GNASAS* ncRNA was not expressed in cervix but showed upregulation in 6/9 cervical cancer cell lines (Figure 50A). The same ncRNA was expressed in normal colon, but showed downregulation in 2/2 colon cancer cell lines

(Figure 50B). *EXON1A* ncRNA was ubiquitously expressed in all tested normal and cancer samples and did not show cancer deregulation. In summary, 5/6 well-known macro ncRNAs were deregulated in cancer.

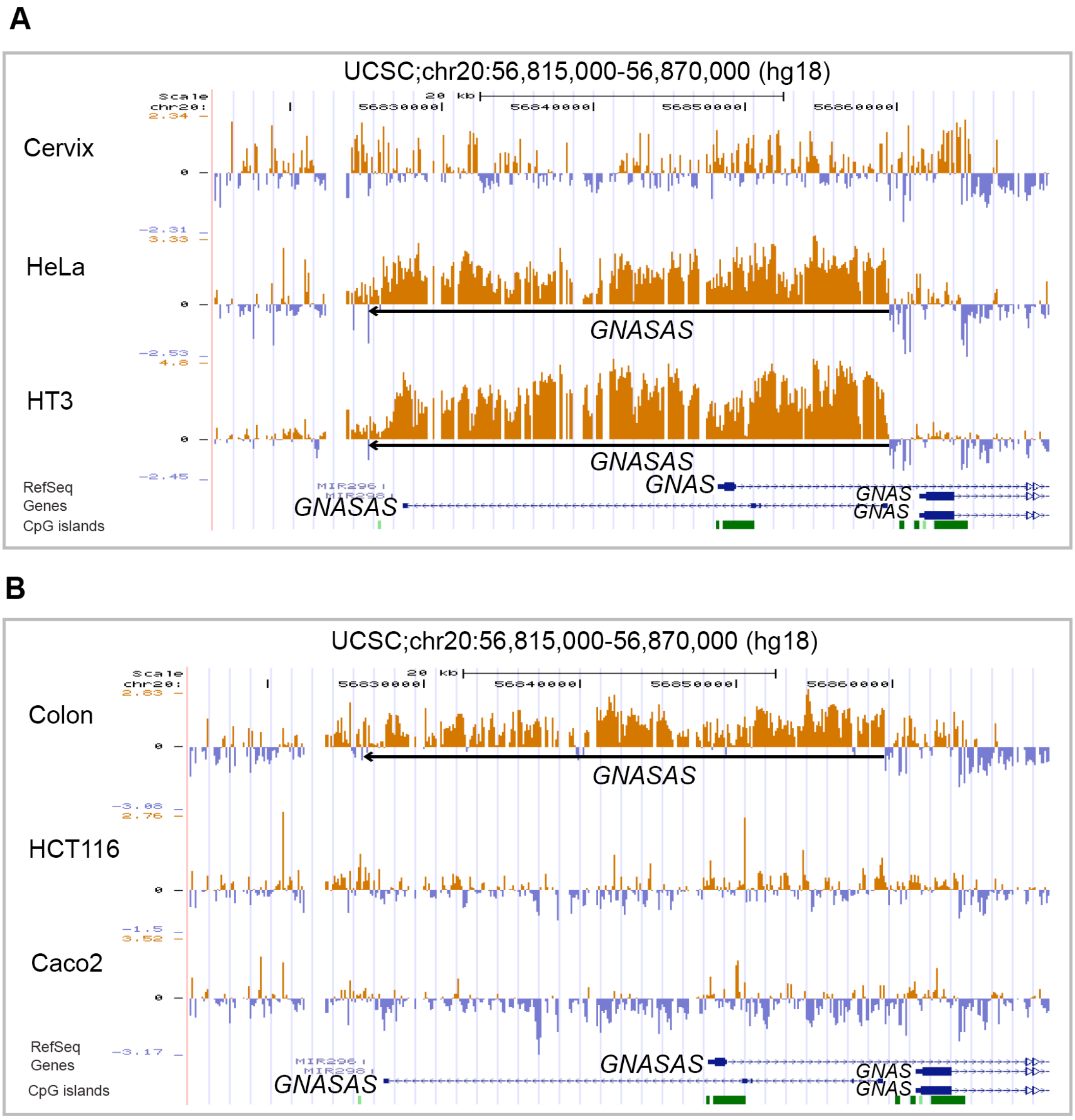


Figure 50. *GNASAS* ncRNA is deregulated in cervical and colon cancers. A. *GNASAS* is not expressed in cervix but it is upregulated in HeLa and HT3 (and 4 more cervical not shown cancer cell lines). **B.** *GNASAS* is expressed from colon and downregulated in 2/2 colon cancer cell lines hybridized to HIRTA. UCSC position is shown on the top. Tracks presented as on Figure 11B, section 2.1.5.

2. 6. 2. Novel macro ncRNAs are deregulated in cancer

101 novel macro ncRNA candidates were found to be expressed using HIRTA in 43 cell lines/tissues/patients. I showed previously that there was no overall difference in number of macro ncRNAs in imprinted gene regions between normal and cancer samples since on average ~19 novel macro ncRNA candidates are expressed per both normal or cancer sample (section 2.5.2.), but I found a number of samples

expressed exclusively in cancer cells or in normal cells as well as candidates showing different levels of expression between a normal tissue and the corresponding cancer of that tissue.

22 macro ncRNAs were expressed exclusively in cancers (Table 19). From these macro ncRNAs 10 were expressed in only one tested cancer cell line. 16 of these transcripts were characteristic for one type of cancer while 6 others were expressed from more than one type of cancer but not expressed from any normal tissue. In 11/22 cases macro ncRNAs were not expressed in normal cervix (or any other tested normal tissue), but expressed from 1-3 cervical cancer cell lines. Cancer deregulation of *RPE56ov* and *RPE65down1* is presented on Figure 51. *OR56A1down* and *PPP2R5Cup1* (low expression) were expressed in acute myeloid leukemia patients, but not expressed in whole blood and bone marrow.

Cancer specific macro ncRNA candidates	Cell line expressing macro ncRNA	Cancer type
GPR177up	C33A, Tera2, A201	cervical, teratocarcinoma, rhabdomyosarcoma
RPE65ov	C4I, C33A, DoTc2, HCT116, A201	cervical, colon, rhabdomyosarcoma
RPE65down1	C33A, A201	cervical, rhabdomyosarcoma
RPE65down2	Tera2	teratocarcinoma
MAS1ov	NCCIT	teratocarcinoma
PTK2up	DoTc2	cervical
SFMBT2down5	HeLa, C4I, C4II	cervical
GATA3down	Cama1	breast
SEC23IPdown2	SHSY5Y	neuroblastoma
OR56A1down	AML5_BMMC, AML5_PBMC, AML7	acute lymphoid leukemia
PKCDBPdown	A201	rhabdomyosarcoma
ZNF195down1	C4I, C4II	cervical
ZNF195down2	HCT116, C33A	colon, cervical
ZNF195down3	C4I, C4II, C33A, HCT116	cervical, colon
ZNF195down4	C4I, C4II	cervical
SLC38A4down1	Hela, SW756	cervical
EPYCov	DoTc2	cervical
DCNup3	MCF7	breast
LRCH1up2	Tera2	teratocarcinoma
PPP2R5Cup1	AML7, MP_0351B, MP_0363	acute lymphoid leukemia, myeloproliferative disorder
ZNF831up2	A201	rhabdomyosarcoma
TOX2up	Cama1	breast

Table 19. 22 novel macro ncRNAs are cancer specific. Cancer specific macro ncRNAs are not expressed in any of 20 tested normal tissues, but are expressed in presented cell lines or patients of a specific cancer type.

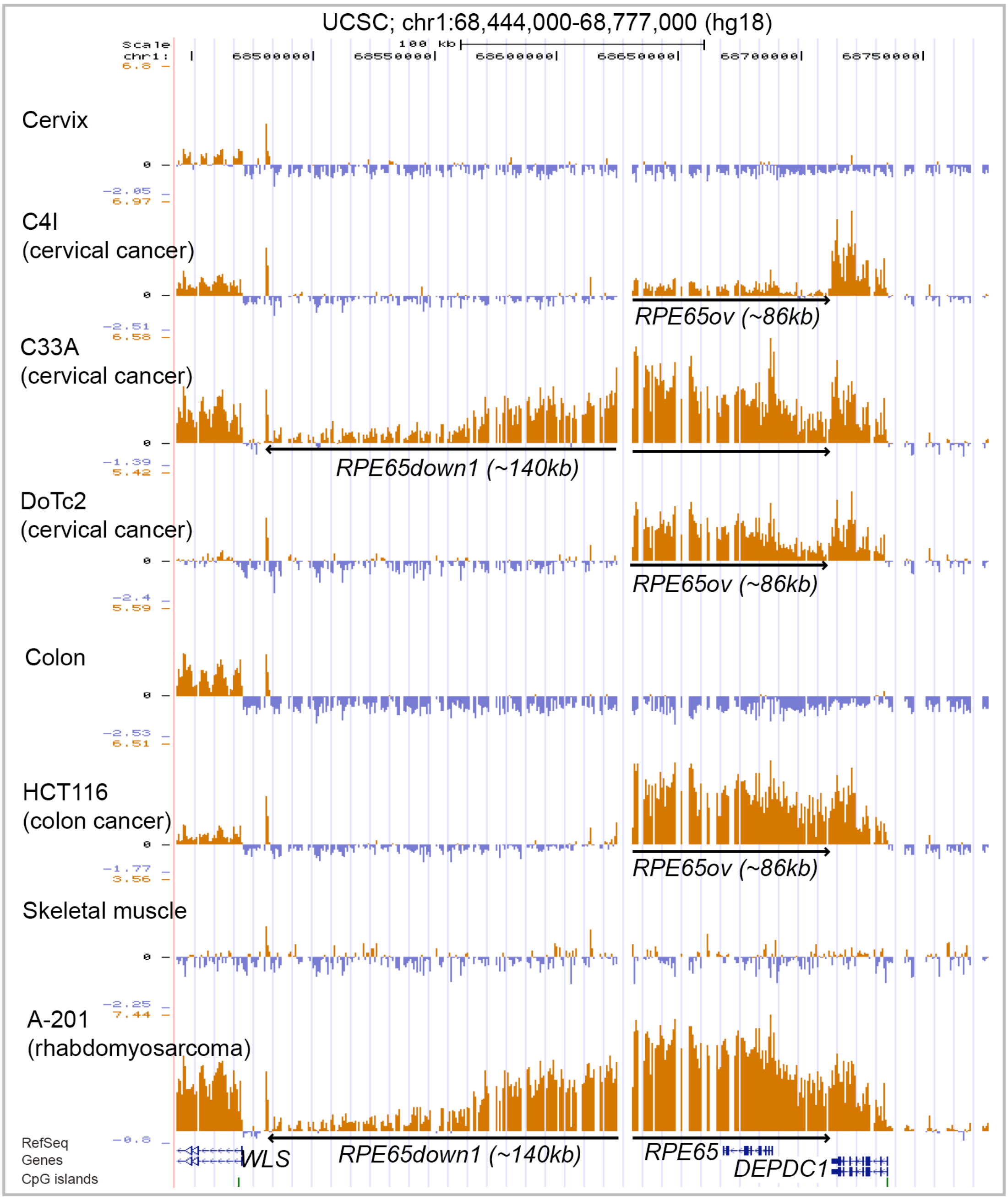


Figure 51. *RPE65ov* and *RPE65down1* macro ncRNAs are not expressed in any of 20 tested normal tissues, but show cancer specific expression. *RPE65ov* is a ~86kb long macro ncRNA not expressed in normal cervix, colon and skeletal muscle, but expressed in corresponding cancers (3/9 cervical, 1/2 colon and 1/1 rhabdomyosarcoma cell lines). *RPE65down1* is specifically expressed in C33A and A201 cell lines. Predicted orientation of macro ncRNAs is shown with black arrows. UCSC position is shown on the top. Custom HIRTA tracks as well as RefSeq genes, sno/miRNA and CpG islands UCSC tracks are shown.

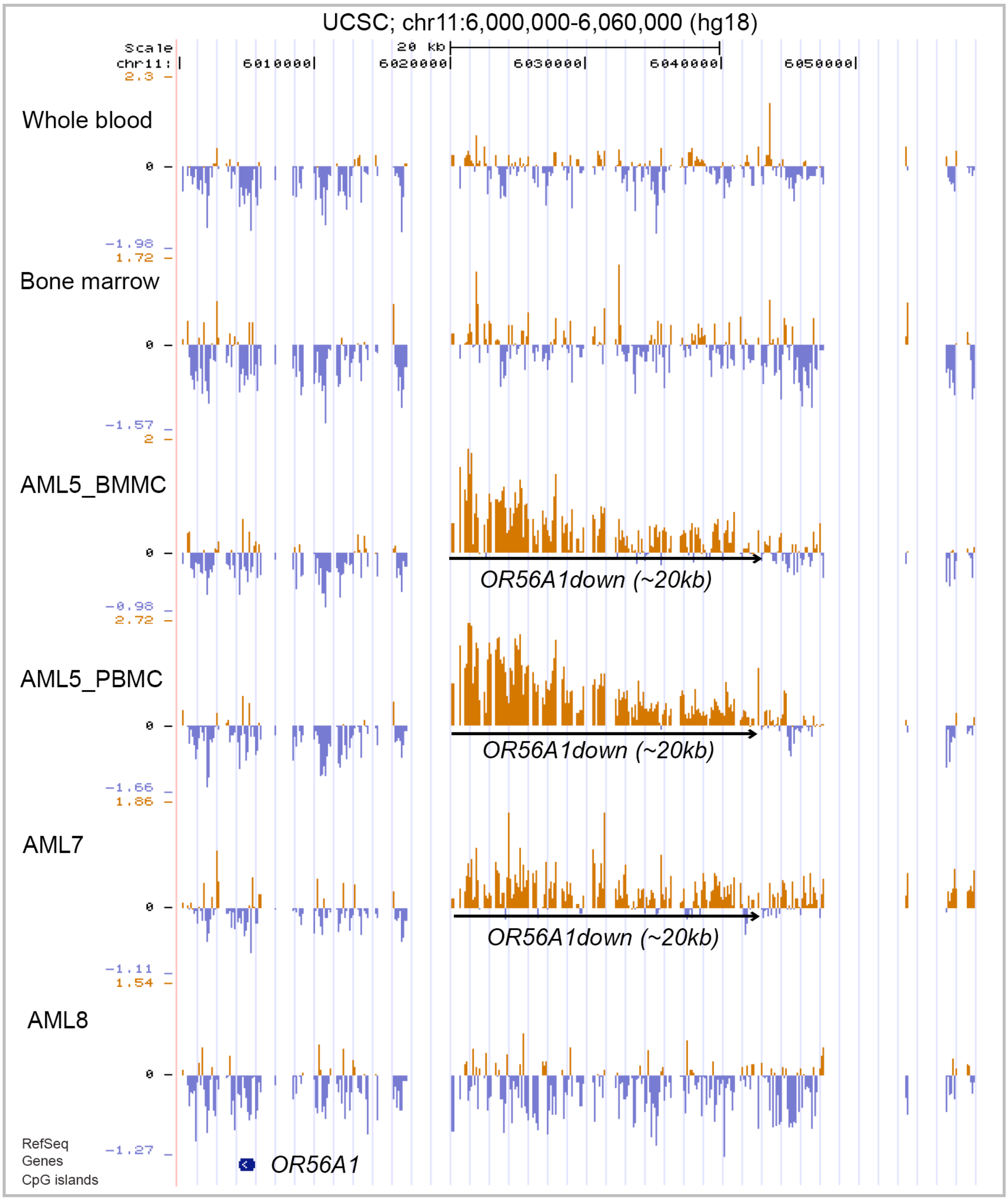
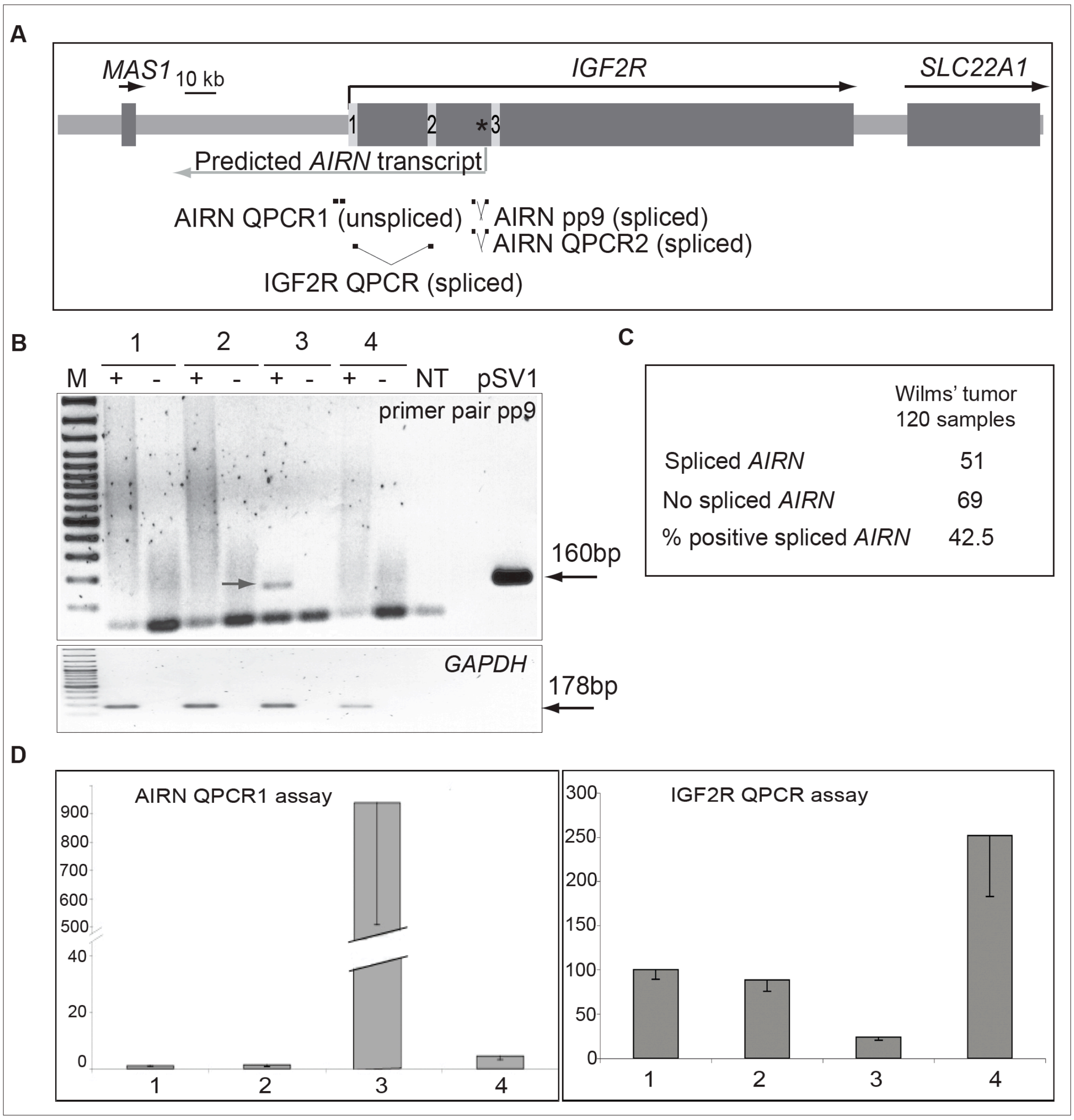


Figure 52. *OR56A1down* macro ncRNA lacks expression in normal blood and bone marrow (as well as any of tested 20 normal tissues), but shows upregulation in 3/4 acute myeloid leukemia patient samples. This ncRNA is ~20kb in length, predicted to be expressed from + strand (black arrow) and expressed from ZNF215 HIRTA region on chromosome 11. UCSC position is shown on the top. Custom HIRTA tracks as well as RefSeq genes, sno/miRNA and CpG islands UCSC tracks are shown.

A number of macro ncRNAs were expressed in both normal and cancer samples, but showed deregulation in cancer. For example, *SFMBT2down1* was not expressed in cervix, but expressed in 4/9 cervical cancer cell lines, or *B3GNT2down2* was not expressed in colon, but showed upregulation in 1/2 colon cancer cell lines. Further, *SLC38A4up* was not expressed in cervix, mammary gland and colon but is upregulated in 2/9 cervical cancer cell lines, 1/2 breast and 1/2 colon cancer cell lines. In contrast, *LRCH1up* macro ncRNA was expressed in 9/20 normal samples, but not expressed in any of examined cancers while *C15orf43up* showed expression in normal blood, but was downregulated in 1/3 AML and 2/2 MPD patients. Since tiling arrays have limitations (in background and dynamic range) numerous subtle changes in expression were not examined but rather I focused on candidates fully silenced or fully upregulated in cancers. Therefore, I expect that the expression change of presented transcripts is at least equal to the dynamic range of HIRTA and possibly higher, meaning that the expected fold difference in transcript expression between normal and corresponding cancer tissue was <= 200 fold (based on previous RT-qPCR experiments, Table 10, section 2.1.4.)

2. 6. 3. Identification of the human homolog of the mouse *Airn* ncRNA in Wilms' tumors

Previously it was shown that the CpG island in intron 2 of *IGF2R* (CGI-2), similar to the mouse CpG island located on the same position of *Igf2r*, carries a maternal-specific DNA methylation imprint (Smrzka et al., 1995). Since this CpG island is associated with the transcription start site of paternally expressed *Airn* ncRNA in mouse we tested if the human CpG island could also act as a promoter for a macro ncRNA (Figure 53A). While mouse *Igf2r* shows ubiquitous imprinted expression in fetal, extra-embryonic and adult tissues (except post-mitotic neurons) (Yamasaki et al., 2005), human *IGF2R* shows polymorphic imprinted expression in early fetal tissue, amniotic and lymphoblastoid cells, placenta and Wilms' tumors (Oudejans et al., 2001; Smrzka et al., 1995; Xu et al., 1993; Xu et al., 1997). No evidence of a human *AIRN* has been previously found in placenta (Oudejans et al., 2001). I focused on Wilms' tumors that were not previously tested for *AIRN* expression, but had polymorphic imprinted expression of human *IGF2R* gene.



	Wilms' tumor 120 samples
Spliced *AIRN*	51
No spliced *AIRN*	69
% positive spliced *AIRN*	42.5

Figure 53. *AIRN* is expressed from a Wilms' tumor. A. Map of human chromosome 6q27 region showing *IGF2R* and flanking genes. Transcription orientation of protein coding genes and predicted *AIRN* transcripts is indicated by arrows. Exons of *IGF2R* are marked by numbers (1-3) and position of CGI-2 is marked by an asterisk. Positions of RT-PCR (pp9) and qRT-PCR assays (IGF2R QPCR and AIRN QPCR1 and 2) are indicated underneath. **B.** Expression of spliced *AIRN* in Wilms' tumor cell lines using primer pair pp9 amplifying a 160bp product (1: Sk-Nep1, 2: G-401, 3: STA-WT3ab). The Hs27 normal human fibroblasts cell line (4) was used as a negative control and plasmid SV1 containing the spliced human *AIRN* product from mouse transgenes known to express *AIRN* as a positive control. Human *GAPDH* was used as a loading control (178bp band). NT; no template control. **C.** Summary of expression of *AIRN* by RT-PCR using pp9 in 120 Wilms' tumor samples from the German SIOP/GPOH 93-01 study (see Figure 54). **D.** Quantitative RT-PCR of unspliced *AIRN* (AIRN QPCR1 assay) and *IGF2R* (IGF2R QPCR assay) expression in human cell lines (1: Sk-Nep1, 2: G-401, 3: STA-WT3ab, 4: Hs27) using assays shown in A. Values average three technical replicates normalized to *GAPDH*. (Modified from Yotova IY, Vlatkovic IM et al., 2008 (Yotova et al., 2008))

AIRN was expressed from 1 in 3 tested Wilms' tumor cell lines (Figure 53B) by using RT-PCR and primer pair pp9 that amplified the expected 160bp band. Further, using same assay I tested 120 Wilms' tumor samples and found that 42.5% Wilms' tumor patients expressed spliced *AIRN* (Figure 53C, Figure 54A). pp9 assay examining spliced *AIRN* expression in these samples was used since 65% of the samples were DNA contaminated (data not shown) and thus could not be tested for unspliced *AIRN* expression (to examine DNA contamination, primer pair CON located intergenically between *IGF2R* and *SLC22A1* and amplifying 509bp band by RT-PCR was used).

Quantitative RT-PCR in cell lines confirmed *AIRN* expression from the STA-WT3ab Wilms' tumor cell line using the AIRN QPCR1 assay for unspliced *AIRN* (Figure 53D*)*. Since expression of mouse *Airn* correlates with silencing of *Igf2r* on one parental allele I further tested if in human similar correlation was observed. qRT-PCR using IGF2R QPCR primer pair showed the lowest expression of *IGF2R* in the STA-WT3ab cell line that expressed *AIRN* compared with other 3 tested cell lines that did not express *AIRN* (Figure 53D). When we applied the AIRN QPCR2 assay (corresponding to pp9 primer pair) for spliced *AIRN* and the IGF2R QPCR assay, on 20 Wilms' tumor patient samples and the Hs27 cell line that expressed no *AIRN* or showed medium or high *AIRN* expression from RT-PCR data, no correlation between no or medium expressed *AIRN* and *IGF2R* in same samples was found, while 3/6 samples with high *AIRN* showed reduced *IGF2R* (Figure 54B). The data did not show a clear correlation between human *AIRN* and *IGF2R*, however, since Wilms' tumor patient samples most likely contain mixtures of cells expressing, or not expressing *AIRN* the data does not exclude a correlation and thus potential silencing of *IGF2R* by *AIRN in cis*.

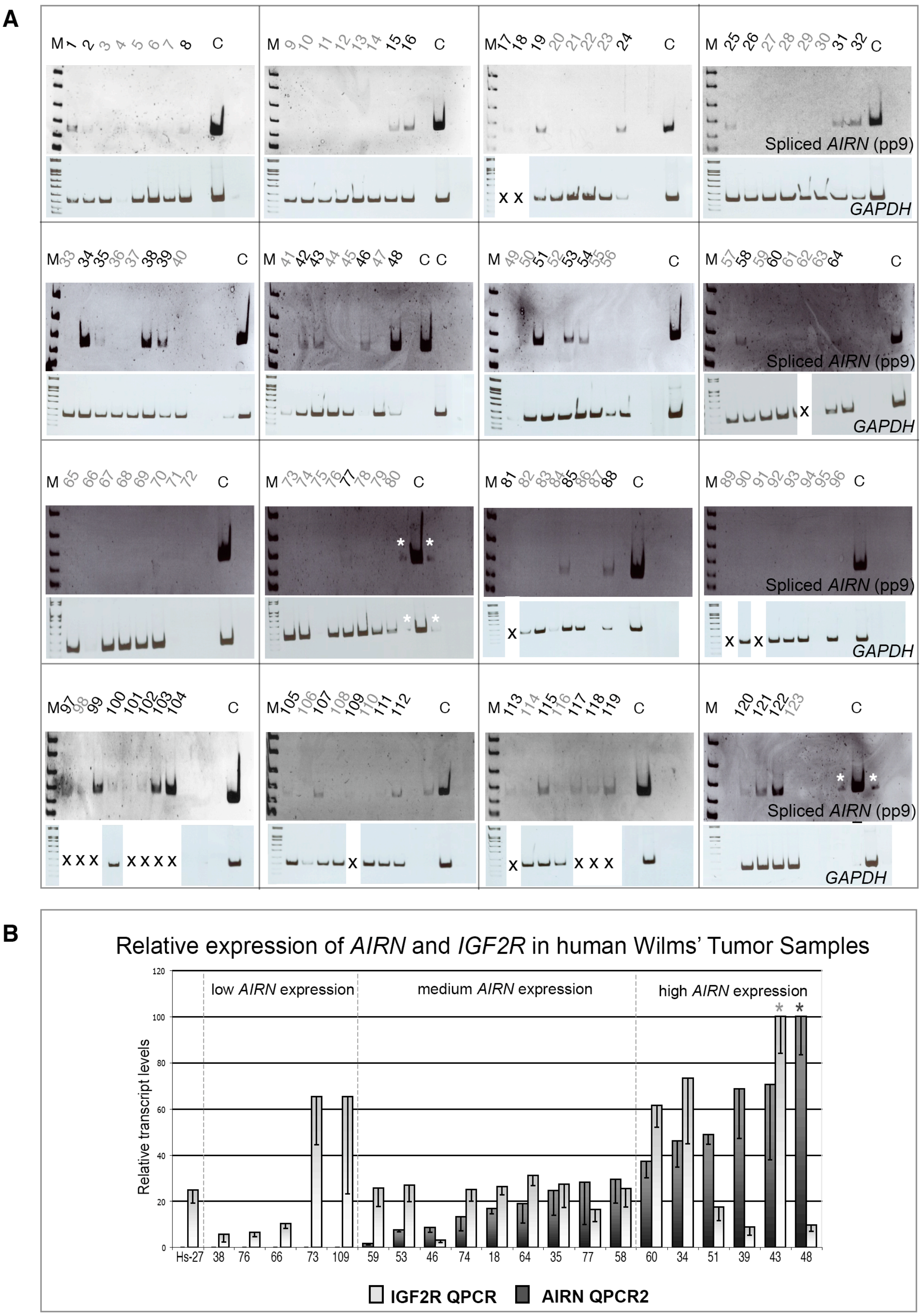


Figure 54. *AIRN* expressed in Wilms tumor patients do not show clear correlation to *IGF2R* in same patients. A. RT-PCR assay of spliced *AIRN* in 123 human Wilms' tumor samples using primer pp9. *GAPDH* was used as a loading control. M: Size markers, C; the plasmid pSpIF containing the SV1 *AIRN* spliced variant was used a positive control for the pp9 primers and HS27 cDNA used as the *GAPDH* primer control. 51 out of 120 samples (excluding 3 samples negative for both *GAPDH* and *AIRN*) were positive for spliced *AIRN* (black type). X, insufficient material to assay *GAPDH*, *spill over from control lane. **B.** Quantitative RT-PCR assay for human *IGF2R* and *AIRN* in 20 Wilms' tumor samples selected for no, medium or high *AIRN* expression. AIRN QPCR2 assay amplified spliced *AIRN* from

exons 1-2 (these *AIRN* exons were also amplified by the primer pair pp9 shown above), and assay IGF2R QPCR amplified *IGF2R* from exons 1-2. Hs27 cells were used as a negative control for *AIRN* expression. qRT-PCR data was normalized to 18S rRNA. Sample 43 with maximal *IGF2R* expression and sample 48 with maximal *AIRN* expression, were each set to 100 (*). Error bars indicate the standard deviation of three technical replicates. Three independent experiments (each with 3 technical replicates) were performed with the same sample set. The measurements of *AIRN* abundance from the non-quantitative and quantitative assays were in general agreement with some exceptions (e.g. sample 38). (Modified from Yotova IY, Vlatkovic IM et al., 2008 (Yotova et al., 2008))

Normal human tissues showed the presence of maternally methylated CGI-2 that we found to be the promoter for human *AIRN* macro ncRNA in all tested fetal and adult tissues (Riesewijk et al., 1996; Smrzka et al., 1995). To determine if *AIRN* expression is specific for Wilms' tumors or could be also found in normal tissues, I tested 20 normal human tissues (placenta; fetal brain and liver; adult: brain, liver, heart, lung, uterus, thyreoidea, bone marrow, skeletal muscle, salivary gland, adrenal gland, prostate, testis, spinal cord, trachea, thymus, kidney and cerebellum) using the pp9 primer pair for spliced *AIRN*. I could not detect expression of *AIRN* in any of 20 tested normal tissues (data not shown). Further, I tested both undifferentiated (d0) and differentiated (d7) human embryonic stem cells (HES2) for expression of spliced *AIRN* and found that spliced *AIRN* was not expressed in these cells (data not shown).

To determine if *AIRN* could be present in other cancers we focused on embryonic cancers expected to show polymorphic expression of *IGF2R*. I tested four cancer cell lines: Tera2, NCCIT, PA1 and A201, for the presence of differentially methylated region on the CpG island in intron 2 of *IGF2R* (CGI-2) shown to be the promoter of *AIRN* in Wilms' tumors. By using Southern blot with methylation sensitive enzyme Not1 combined with EcoRI digestion and the specific Bx probe, both alleles of the CpG island were methylated in the four tested cell lines (Figure 55). The CpG island in intron 2 of *IGF2R* in Tera2 cell line was confirmed to be methylated using three more methylation sensitive enzymes (BstUI, XhoI and HpaII) in Southern blot with the same Bx probe. Additionally I confirmed existence of both methylated alleles of the same CpG island in the A201 cell line, using BstUI, HpaII, Eco52I and PauI methylation sensitive enzymes (data not shown). It was previously shown that DNA methylation is a repressor of transcription if found on promoter regions (reviewed in (Koerner and Barlow, 2010)) , thus it is expected that *AIRN* would not be expressed from tested cancer cell lines. In summary, *AIRN* could be a cancer and particularly Wilms' tumor specific macro ncRNA. Partially this work has been published in Yotova IY, Vlatkovic IM et al., 2008 (Yotova et al., 2008).

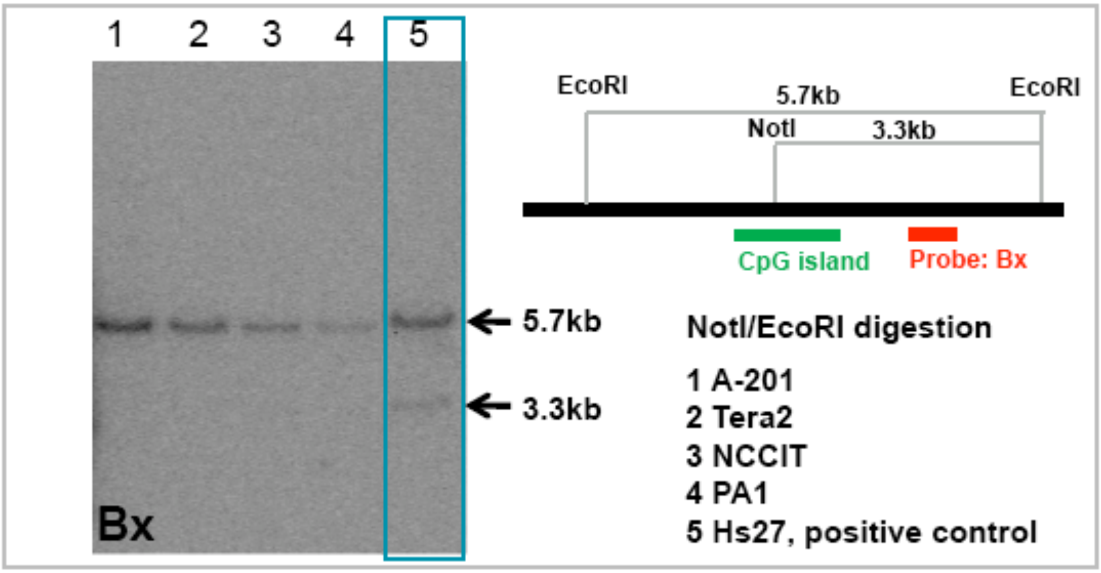


Figure 55. Lack of DMR2 in four teratocarcinoma and rhabdomyosarcoma cell lines. Southern blot analysis using NotI methylation sensitive enzyme in combination with EcoRI using specific Bx probe showed the methylation of both alleles (presence of 5.7kb band) of CpG island in intron 2 of *IGF2R* in four tested cancer cell lines. Hs27 normal human fibroblasts cell line was used as a positive control and showed a 5.7kb methylated and 3.3kb unmethylated bands (presence of CGI-2 DMR) as expected.

2. 7. Ribosomal RNA-depleted RNA-sequencing detects macro ncRNAs in human fibroblasts

Transcriptome sequencing (RNA-seq) using next-generation technologies e.g. illumina has a great potential in whole genome research. I examined part of the human RNA-seq data from diverse tissues published during 2008, in order to compare them to our HIRTA tiling arrays expression results. First I examined if published RNA-seq data could detect known imprinted macro ncRNAs. Interestingly, published RNA-seq data poorly detected expression of well-known macro ncRNAs (shown in Figure 56) even in tissues and cells that are known to express these transcripts. For example in Pan et al. (Pan et al., 2008) where a mRNA-seq dataset consisting of 17-32 million 32bp reads that examined whole brain, cerebral cortex, liver, lung, skeletal muscle and heart was published, I was not able to detect known macro ncRNAs. For example, expression of *KCNQ1OT1* and *UBE3A-AS* macro ncRNAs was clearly detectable by HIRTA, but poorly detectable using RNA-seq (Figure 56A, B). The only known imprinted macro ncRNA that could be clearly detected in Pan et al., 2008*,* was the 2.6kb *H19* macro ncRNA that differs from *KCNQ1OT1* in that it has small introns and a cytoplasm localization. Similarly, I found almost no expression of known imprinted macro ncRNAs in RNA-seq data from Sultan et al. (Sultan et al., 2008) where poly (A) RNA was extracted from human embryonic kidney (HEK cell line) 293T and B cells (RAMOS cell line) and produced

8.6 million and 7.6 million of 27bp sequencing reads respectively. This analysis of public data indicated that RNA-seq using poly A selected RNA does not allow optimal macro ncRNA detection and I tested if total RNA could potentially improve macro ncRNA detection using normal human fibroblasts, a cell line that express most of the known macro ncRNAs. Ru Huang, a PhD student in the lab developed a protocol for RNA-seq of the mouse cells using total RNA depleted of ribosomal RNA (rRNA) and hydrolyzed prior to double stranded (ds) cDNA preparation. I performed the rRNA depletion RNA-seq protocol, on normal Hs27 human fibroblast sample and performed double stranded (ds) cDNA. The sample was further processed (see section 5.8.8.5.) by Andy Sommer, GENAU consortium, applied on 4 flowcells and runned on Illumina Genome Analyser obtaining in total 58.8 milion 36bp sequence reads (13.1-15.5 million reads per flowcell).

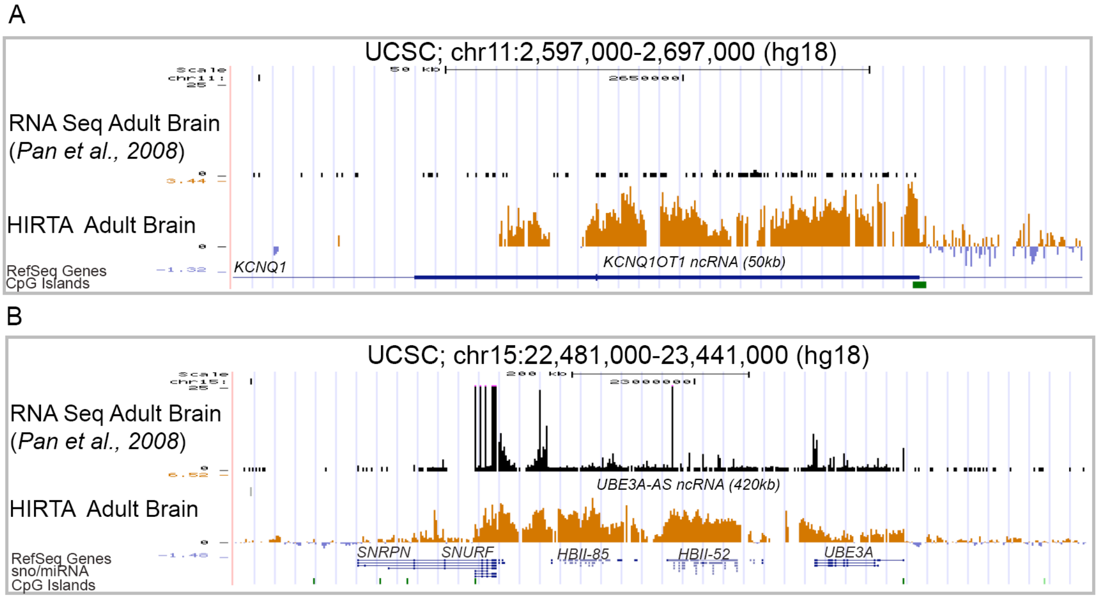


Figure 56. Imprinted macro ncRNAs were not detectable in mRNA-Seq data published prior to our RNA-Seq experiment. A. *KCNQ1OT1* macro ncRNA is clearly expressed from adult brain hybridization on HIRTA while it is poorly detected from the same tissue in mRNA-Seq from *Pan et al., 2008* data. **B.** *UBE3A-AS* is highly expressed in adult brain by HIRTA, but shows very low, not continuous expression in mRNA-seq data from *Pan et al., 2008*. UCSC view of RNA-seq (black) and HIRTA (orange) expression data on specified UCSC position (hg18). HIRTA tracks are presented as previously (Figure 11B, section 2.1.5.). Black: RNA-seq track; RNA-seq with cut off of 25 reads.

2. 7. 1. Statistics of rRNA depleted RNA-seq reads in fibroblasts

From 58.8 milion reads, 44.6% (26.2 million reads) were uniquely mapped on the NCBI36/hg18 using the Bowtie alignment program (the alignment was done by Ido Tamir, GENAU consortium). Further, 49.95% of unique mapped reads were mapping to rRNA genes after the depletion. This showed that depletion was not complete, but considering that rRNA consists ~95% of the cell RNA, depletion was enriched for non-rRNA.

To examine the distribution of the reads through different categories of genomic regions the TopHat (Trapnell et al., 2009) program that aligns reads from the RNA-Seq to the reference genome (hg18) without relying on known splice sites, was used. This program finds potential exons in Bowtie aligned data and further it builds a database of possible splice junctions that the program further confirms by mapping the experimentally gained reads to the junctions. This program was used to find known and novel splice junctions (across “GT-AG” introns) and to determine the distribution of the spliced and by extracting from the starting Bowtie data, the unspliced portion of the human genome.

The TopHat algorithm aligned ~11 million reads to the hg18, consisting of reads that are complete (they contiguously align to the hg18) and that are spliced (discontinuous alignment). Approximately 8.3 million complete reads and 0.5 million of spliced reads that map to genic regions (annotated exons and introns of the RefSeq genes), and 2.1 million of complete and 0.13 million of spliced reads that map to intergenic regions (Figure 57A), were found. As expected, most of spliced reads partially map to exons and introns. Reads found to map to intergenic regions represented potential novel spliced transcripts in the human genome. This analysis showed that there are about 20% more spliced transcripts than annotated by RefSeq genes genes, and about 25% of novel exons mapping to introns (new spliced variants of known genes).

The results of RNA-Seq were visualized on the UCSC browser (NCBI36/hg18) with the X-axis representing position in the genome and Y-axes representing number of reads. Representation of the Bowtie and the TopHat outputs are shown with the example of the *DCN* protein coding gene where 7/9 known exons have been found using exon-exon mapping (Figure 57B). As expected, the number of reads matching to exons of protein coding genes was very high in the RNA-Seq data, for example for *DCN*, 553 reads were mapped to the exon 5 showing highest expression (note that on Figure 57B a cut off of 25 reads was used).

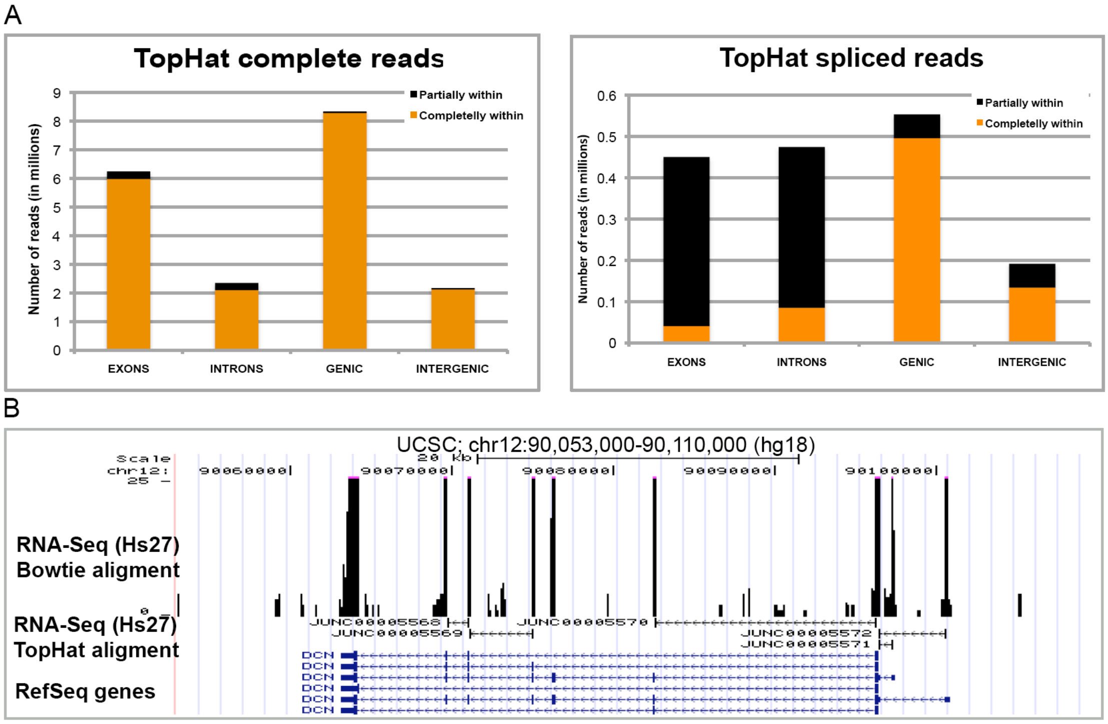


Figure 57. TopHat alignment of fibroblasts RNA-seq finds novel splice-junctions in intergenic and intronic regions. A. Comparison of TopHat complete and spliced reads with RefSeq genes (reads that are completely or partially within exons, introns, genic and intergenic regions are shown). **B.** UCSC screenshot of Bowtie and TopHat aligned RNA-seq fibroblasts data. TopHat alignment is shown in the form of splice junctions where exons are represented as vertical lines. 7/9 exons of *DCN* were found using TopHat. Black RNA-seq track; RNA-seq reads with cut off of 25 reads.

To examine the number of Hs27 RNA-seq reads that are represented in the unspliced fraction of the genome, I extracted the TopHat spliced fraction (based on Bowtie alignment) from uniquely mapped reads mapped with Bowtie and found that 57.4% (~15 millions) of reads were unspliced (Figure 58). This rather large fraction potentially contains exons that are spliced in the tested cell line, but were not detected by TopHat (false negative rate), known unspliced transcripts (e.g. *Kcnq1ot1* macro ncRNA lacks any spliced products (Pandey et al., 2008)) and novel unspliced transcripts e.g. macro ncRNAs.

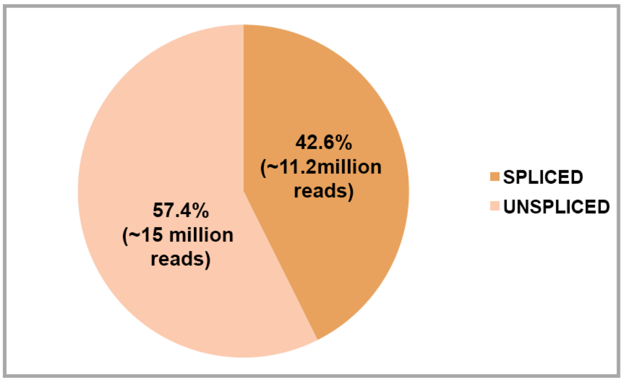


Figure 58. Using different aligners mapping spliced and total unique reads, 42.6% of the fibroblasts transcriptome has been found to be spliced while 57.4% is unspliced. Note that unspliced fraction contains TopHat false negatives, but also novel macro ncRNAs.

2. 7. 2. Known macro ncRNAs by RNA-seq

I further examined expression of known imprinted macro ncRNAs by RNA-seq and found that *KCNQ1OT1*, *GTL2var1*, *UBE3A-ASvar2* and *EXON1A* that were expressed in fibroblasts by HIRTA hybridization, also showed expression in the Hs27 RNA-seq data (examples in Figure 59). Previously, exact mapping of *KCNQ1OT1* was not possible with HIRTA mapping since the end of *KCNQ1OT1* macro ncRNA is highly repetitive and was not spotted on the Chip. Using RNA-seq, I found the end of the human *KCNQ1OT1*, that is annotated RefSeq transcript of 59.5kb, and mapped the *KCNQ1OT1* transcript as a 100kb long. The *H19* and *GNASAS* macro ncRNAs that were very lowly expressed in fibroblasts using HIRTA compared to other cells/tissues e.g. liver for *H19* and fetal kidney for *GNASAS*, could not be mapped using RNA-seq (these macro ncRNAs were covered with a very low number of reads). Interestingly, I observed that using RNA-seq, all known imprinted macro ncRNAs were lowly expressed (below 20 uniquely mapped reads) compared to protein coding genes.

Further, expression of known non-imprinted macro ncRNAs expressed from fibroblasts was analysed and it was found that RNA-seq clearly detected the *HOTAIR* and *MALAT1* macro ncRNAs (Figure 60A, B). Interestingly, *MALAT1* was among the most highly expressed genes excluding rRNA genes, having the expression maximum of 5052 mapped reads (Figure 60B).

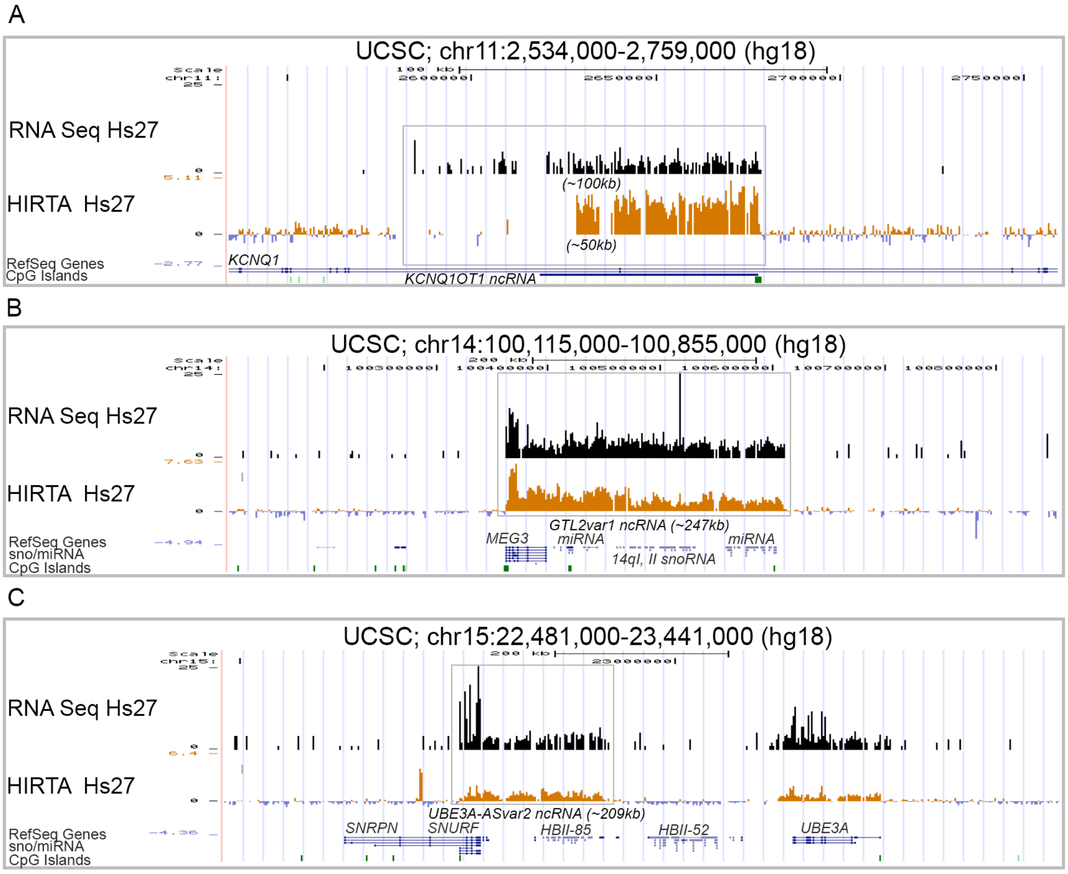


Figure 59. Known imprinted macro ncRNAs are detected by rRNA depleted RNA-seq. A. Both RNA-seq and HIRTA detect expression of *KCNQ1OT1* with RNA-seq mapping a longer, 100kb *KCNQ1OT1* variant. **B.** *GTL2var1* ncRNA is detected using RNA-seq. **C.** *UBE3A-ASvar2* is expressed in Hs27 using HIRTA and RNA-seq. UCSC view of RNA-seq (black) and HIRTA (orange) expression data on specified UCSC position (hg18). HIRTA tracks are presented as previously (Figure 11B, section 2.1.5.) Black: RNA-seq track; RNA-seq reads with cut off of 25 reads.

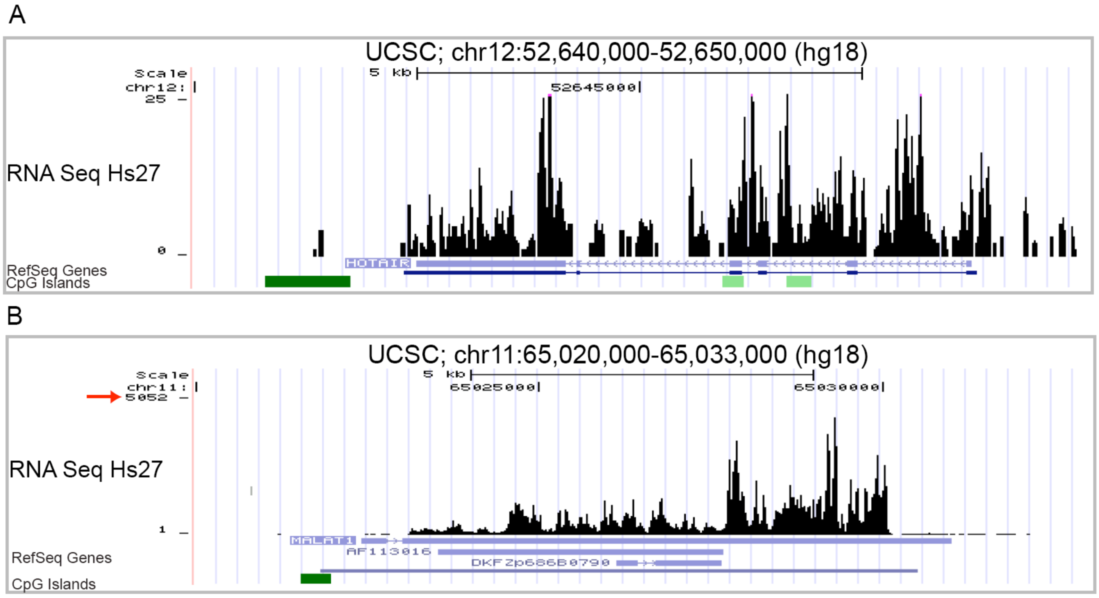


Figure 60. rRNA depleted RNA-seq in normal human fibroblasts detects known non-imprinted macro ncRNAs. A. *HOTAIR* macro ncRNA maps to annotated RefSeq genes

track. **B.** *MALAT1* macro ncRNA maps to annotated ncRNA gene and shows very high expression with maximum of 5052 uniquely mapped reads. UCSC view of RNA-seq (black) expression data on specified UCSC position (hg18). Black RNA-seq track; RNA-seq reads

2. 7. 3. Novel HIRTA macro ncRNAs validation by RNA-seq

24 novel macro ncRNAs that were found by HIRTA in the fibroblast cell line were compared to rRNA depleted RNA-seq of the same cell line (Table 20). 6/24 transcripts found by HIRTA were also clearly expressed by RNA-seq, 16 were very lowly expressed (with few sequencing reads mapping to regions annotated by HIRTA expression) and expression of two macro ncRNA was not found using RNA sequencing (Table 20). Three examples of novel macro ncRNA transcripts RNA-seq expression are shown in Figure 61, where *KLF14up2* and *KLF14up3* (Figure 61A) were clearly expressed and *SLC38A4down2* example of lowly expressed transcript (Figure 61B).

Since previously published RNA-seq data (shown in 2.7) could not detect expression from known imprinted macro ncRNAs, I looked into RNA-seq data recently integrated by UCSC (Burge RNA-seq (Wang et al., 2008a), CSHL Long RNA-seq, GIS RNA-seq, Caltech RNA-seq and Helicos RNA-seq). These RNA-seq data examined expression in brain, breast, heart, lymph node, lymphoblastoid cells and K562 myelogenous leukemia cell line, but there were no available fibroblasts sequencing data to which I could compare our Hs27 RNA-seq. Therefore, RNA-seq was compared to global run-on sequencing (GRO-seq) data (Core et al., 2008) which shows the positions of transcriptionally-engaged RNA polymerases in IMR90 lung fibroblasts. In total, 18/24 macro ncRNAs expressed from HIRTA were also transcribed by GRO-seq and similarly to RNA-seq of Hs27, they showed very low transcription of macro ncRNAs. GRO-seq data are strand specific and I found that 7/18 GRO-seq transcripts showed transcription from both plus (+) and minus (-) UCSC strands indicating transcriptional orientation of probably overlapping transcripts, while 11 were transcribed from one of the UCSC strands (Table 20). When GRO-seq expression data was compared with the strand prediction data (section 2.5.3), overall agreement between these datasets was found, with exception of *SDHDdown* transcript that was predicted as expressed from - strand (with low probability score of -0.167), but found by GRO-seq data to be lowly expressed from + strand. Three more examples: *PHACTR2ov*, *KIAA1024up* and *ADAMTS7down* were similarly predicted as one strand transcripts and found as opposite strand by GRO-seq, but they were all in the group of transcripts marked with a star (Table 18, section 2.5.3.) representing probable overlapping transcripts of opposite orientations.

Macro ncRNAs expressed in fibroblasts (Hs27) by HIRTA	RNA-seq in fibroblasts (Hs27)	GRO-seq + (*Core et al., 2008*) in fibroblasts (IMR90)	GRO-seq - (*Core et al., 2008*) in fibroblasts (IMR90)
H19down	+ (low)	+ (low)	-
TSPAN32down1	+ (low)	-	-
ZNF195up1	+ (low)	-	-
ZNF195up2	+ (low)	-	-
SNRPNup2	+ (low)	-	-
APCDD1Lup1	+ (low)	+ (low)	+ (low)
LRRC47down	+ (low)	+ (low)	+ (low)
TIGD2down	+ (low)	+ (low)	-
PHACTR2ov	+ (low)	+ (low)	-
COL1A2up	+ (low)	+ (low)	+ (low)
KLF14up2	+	-	+
KLF14up3	+	+	+ (low)
PEG13	+ (low)	-	+ (low)
KIAA1126up	-	-	+ (low)
ZNF215up	+ (low)	-	-
OR5P2ov	+	-	-
SDHDdown	+ (low)	+ (low)	-
PTSdown	+ (low)	+ (low)	+ (low)
SLC38A4down1	+ (low)	+ (low)	+ (low)
SLC38A4down2	+ (low)	+ (low)	+ (low)
ADAMTS7down	-	-	+ (low)
TMED3down	+	+ (low)	-
KIAA1024up	+	+ (low)	-
BLCAPov	+	-	+

Table 20. RNA-seq in fibroblasts validated 22/24 macro ncRNAs mapped by HIRTA. 16 transcripts show low expression (<= 5 reads and usually low coverage that could not allow us mapping of transcript borders) while 6 show low expression, but with >5 reads and coverage that could be visually distinguished as a transcript. GRO-seq data confirmed 18/24 transcripts in IMR90. GRO-seq shows that most of the transcripts are lowly transcribed (<= 5 reads and usually low coverage). Transcriptional orientation of 9 transcripts could be determined from GRO-seq data. RNAs that are transcribed from both strands (columns GRO-seq + and GRO-seq -) represent probably overlapping transcripts of opposite transcriptional orientation.

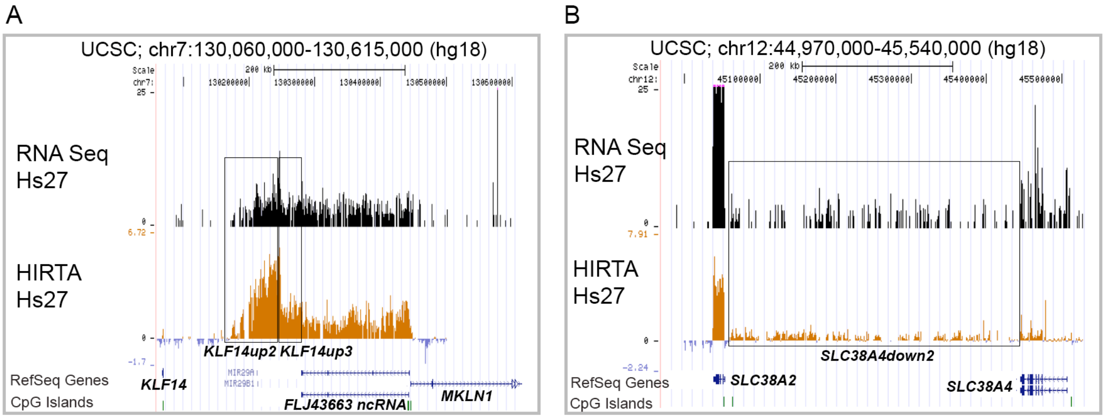


Figure 61. Novel macro ncRNAs are detected by RNA-seq. A. Region upstream of *KLF14* protein coding gene is complex transcriptional unit consisting of annotated *FLJ43663* ncRNA and *KLF14up2* and *KLFup3* macro ncRNAs expression detected by both HIRTA and RNA-seq. RNA-seq expression is shown in black while tiling array expression is orange. **B.** *SLC38A4down2* is a macro ncRNA expressed between *SLC38A2* and *SLC38A4* protein coding genes. This ncRNA shows low expression using HIRTA, but is still found to be expressed using RNA-seq with 26.2 millions of uniquely mapped reads. UCSC view of RNA-seq (black) and HIRTA (orange) expression data on specified UCSC position (hg18). HIRTA tracks are presented as previously (Figure 11B, section 2.1.5.) Black: RNA-seq track; RNA-seq reads with cut off of 25 reads.

I analyzed 24 macro ncRNA candidates expressed in fibroblasts by HIRTA with TopHat alignment in order to test if this program could find exon-exon junctions in our candidates directly from our experimental RNA-seq fibroblasts data. 23/24 candidates did not show any presence of the splice junctions while the only exception was found in the complex transcription unit on chromosome 7 where 2 splice junctions were found in the *KLF14up3* macro ncRNA. These splice junctions were just partially overlapping *KLFup3* and thus potentially represent a novel overlapping protein-coding gene in this region.

2. 7. 4. Novel primary small ncRNA transcripts in the human fibroblasts

Macro ncRNAs can be precursors for different classes of small RNAs. Among six well-known imprinted macro ncRNAs four are known precursors of small ncRNAs, for example: *H19* hosts hsa-mir675, *GNASAS* hosts hsa-mir296, hsa-mir298, *GTL2var1* hosts a cluster of miRNA and the *14I* and *14II* snoRNAs, and, *UBE3A-ASvar1* hosts clusters of *SNORD115* and *SNORD116* snoRNAs. RNA preparation used for the RNA-Seq did not allow detection of RNAs smaller than 200bp, but the precursors of small RNAs should be detectable. To find novel primary micro (pri-mi) and primary snoRNA (pre-sno) transcripts I used miRBase Release 13 (March 2009) and mapped transcription overlapping these micro and snoRNAs on ten human chromosomes (chr1- chr10) in normal human fibroblasts from RNA-seq data.

282 miRNAs and 121 snoRNAs have been annotated from miRBase on the first 10 human chromosomes and I found 26 transcripts overlapping these small RNAs on the first 9 human chromosomes while chromosome 10 was negative (Table 21; names of transcripts will be shown as chromosome number, number of candidate from proximal end of chromosome). 25 were novel while one (chr 9-3) was overlapping the recently validated RefSeq gene *LOC554202* ncRNA. 25 were miRNA precursors (overlapping 1-3 miRNAs) and one pre-snoRNA was also mapped. Lengths of these transcripts were 0.5 to 884kb (for those that are lowly expressed the exact positions of start and end was only provisionally determined). 16/26 transcripts had a CpG island on one of their ends. I used RNAcode prediction to determine non-coding versus coding status of these transcripts and found 20 to be non-coding, 4 protein coding and 2 possibly overlapping both coding and non-coding transcripts. According to the position of novel transcripts to the annotated genes (RefSeq) in the region 16 were intergenic, 5 potential novel 5'exons or overlapping, 3 potential 3'UTRs or overlapping transcripts and 2 intronic transcripts.

Chr.	No. miRNA, No. snoRNA (miRBase, Rel 13)	No / Name		Positions (hg18)	Length (kb)	miRNA/ snoRNA overlapping	CpG island	Position to the annotated gene	RNA code
1	46, 40	6	1	chr1:9131097-9185835	54.7	hsa-mir-34a	yes	IG	NC
			2	chr1:65271069-65305978	34.9	has-mir-101-1	yes	IG	NC
			3	chr1:206037598-206108847	71.2	hsa-mir-29c, hsa-mir-29b-2	yes	IG	NC
			4	chr1:197084516-197173082	88.6	hsa-mir-181a-1, hsa-mir-181b-1	no	IG	NC
			5	chr1:170374221-170380725	6.5	hsa-mir-214, hsa-mir-199a-2	no	IN	NC
			6	chr1:98159044-98283926	124.9	hsa-mir-137	yes	5'/OV	NC (1)
2	24, 14	1	1	chr2:132727100-132755758	28.6	hsa-mir-663b	yes	5'/OV	PC (11)
3	31, 15	3	1	chr3:44016047- 44141886	125.8	hsa-mir-138-1	yes	IG	NC (1)
			2	chr3:115516049-115540802	24.7	has-mir-568	no	3'/OV	NC
			3	chr3:196898055-196932703	34.6	has-mir-570	no	5'/OV	NC (1)
4	25, 5	1	1	chr4:24118111-24138849	20.7	hsa-mir-573	no	3'/OV	NC
5	30, 9	1	1	chr5:148766576-148793784	27.2	hsa-mir-143, hsa-mir-145	no	IG	NC
6	15, 13	4	1	chr6:72136483- 72187456	51	hsa-mir-30c-2, hsa-mir-30a	yes	IG	NC (1)
			2	chr6:126753203-127481933	728.3	hsa-mir-588	yes	IG	PC (12)
			3	chr6:155876033-156759776	883.7	hsa-mir-1202	yes	IG	NC/PC (2)
			4	chr6:18636712-18860589	223.9	hsa-mir-548a-1	no	IG	NC (1)
7	34, 8	4	1	chr7:22859996-22876600	16.6	HBII-336	yes	IG	PC (3)
			2	chr7:27176401-27173373	3	hsa-mir-196b	yes	IG	NC
			3	chr7:32733372- 32851180	117.8	hsa-mir-550-2	yes	IG	PC (8)
			4	chr7:130207753-130252350	44.6	hsa-mir-29a, hsa-mir-29b-1	no	IG	NC/PC (2)
8	29, 4	4	1	chr8:135866372-135914313	47.9	hsa-mir-30b, hsa-mir-30d	yes	IG	NC
			2	chr8:105552966-105570604	17.6	hsa-mir-548a-3	no	3'/OV	NC
			3	chr8:22158068- 22158541	0.5	hsa-mir-320a	yes	5'/OV	NC
9	26, 11	3	1	chr9: 95968398- 96001224	32.8	hsa-let-7a-1, hsa-let-7f-1, hsa-let-7d	yes	IG	NC
			2	chr9:130045825-130047020	1.2	hsa-mir-199b	no	IN	NC
			3*	chr9:21444267-21549697	105.43	hsa-mir-31	yes	5'/OV	NC (1)
10	22, 3	-	-	-	-	-	-	-	-

Table 21. 26 novel primary mi/sno transcripts are mapped on first 10 human chromosomes in normal human fibroblasts. Total number of mi/snoRNAs per human chromosome is shown (using miRBase, Release13, March 2009). Positions (in hg18) and lengths (in kb) of novel transcripts named according to number starting from proximal end of the chromosome they are transcribed from are shown. Overlapping small RNAs, CpG island located on one of the end of the mapped transcript, position according to annotated genes (RefSeq) and RNAcode prediction for these transcripts is presented. * *LOC554202* ncRNA

The example of the novel *LET-7-pri-miRNA* (chr 9-1) overlapping three miRNA (*hsa-let-7a-1, hsa-let-7f-1, hsa-let-7d*) and of novel *SNORD93-pre-snoRNA* (chr 7-1) overlapping *HBII-336* are showed on Figure 62A, B. Both of those have a CpG island at one end putatively identifying the transcriptional start site. Positions of 26 mapped transcripts on 10 human chromosomes are depicted on Figure 62C.

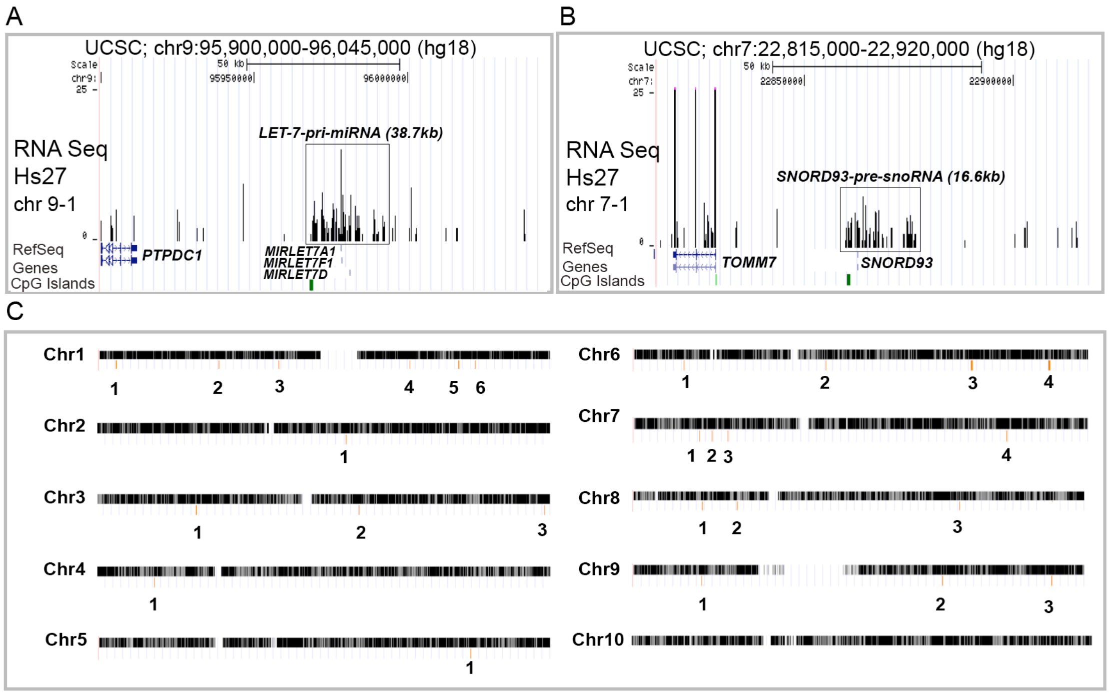


Figure 62. Positions of novel primary small RNA transcripts on 10 human chromosomes and examples of *SNORD93pre-snoRNA* (chr 7-1) and *LET-7-pri-miRNA* (chr 9-1). A. *LET-7-pri-miRNA* in length of 38.7kb is found by RNA-seq in Hs27, maps to three miRNAs and has a CpG island positioned on the proximal end of the transcript. **B.** RNA-seq in fibroblast cell line shows 16.6kb transcription overlapping *SNORD93*. Novel transcript has a CpG island promoter (green). **C.** Positions of novel transcripts expressed in fibroblasts that map to ten human chromosomes (transcripts are numbered from the proximal end of the chromosome).

2. 7. 5. Novel *SHOX* protein coding gene variants

The *SHOX* gene consists of 7 exons. This gene is located on the pseudoautosomal region 1 on chromosomes X and Y and encodes a transcription factor. In collaboration with Claudia Durand from the lab of Gudrun Rappold (Heidelberg, Germany) who had performed a systematic screen of 41 human tissues and identified novel *SHOX* gene variants including four exon variants named 2a, 7-1, 7-2 and 7-3, I verified these findings using RNA-seq. Since the Hs27 normal human fibroblasts cell line expressed some novel *SHOX* variants in levels detectable by RT-PCR, I analyzed rRNA depleted RNA-seq whole genome data from Hs27. I confirmed novel *SHOX* exons in the *SHOX* gene region on chromosome X in Hs27 cell line by RNA-seq (Figure 63A). Applying a cut-off of 5 unique mapped reads, expression of novel *SHOX* exon 2a and variants of the known exon 7: 7-1, 7-2 and 7-3, were detected by RNA-seq (Figure 63B). In summary, novel *SHOX* exon 2a and three variants of exon 7 were validated by RNA-seq technique in the Hs27 cell line. This work has been submitted as: Durand C, Roth R, Vlatkovic I, Dweeph H, Decker E, Schneider K, Rappold G, Alternative splicing and nonsense-mediated RNA decay contribute to the regulation of *SHOX* expression. The analysis shows usefulness of RNA-seq to identify novel spliced variants.

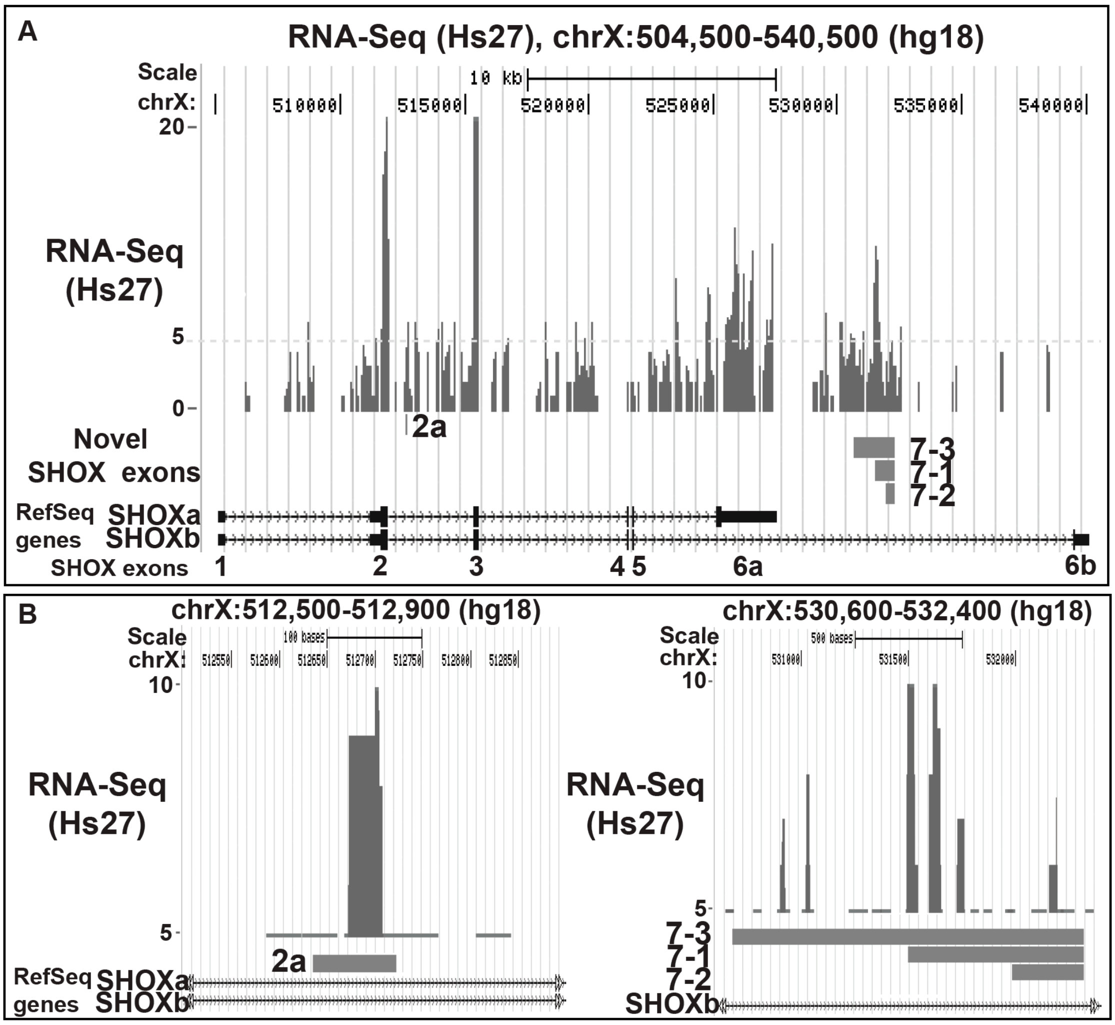


Figure 63. RNA-seq of diploid human fibroblasts (Hs27) confirms novel 2a, 7-1, 7-2 and 7-3 exons of *SHOX* gene. **A.** UCSC browser screenshot of Hs27 RNA-seq on the chromosome X genomic region covering *SHOXa* and *SHOXb* genes (RefSeq UCSC track) shows expression of known and novel *SHOX* exons. The X-axis represents the number of mapped reads; the Y-axis shows the position in the NCBI36/hg18 release of the human genome. Known and novel exons of *SHOX* variants are numbered below. **B.** UCSC browser screenshot focused on novel exons 2a and 7-1, 7-2 and 7-3, shows that more than 5 unique reads could be mapped on the genome positions of these exons confirming their expression in the Hs27 cell line.
(Modified from Durand C, Roth R, Vlatkovic I, et al., submitted)

3. Discussion

3. 1. Summary of results

I successfully used tilling array (HIRTA) and RNA-seq approaches to map novel macro ncRNAs in human imprinted gene regions. From 20 normal and 23 cancer human samples hybridized to HIRTA, tissue specific profiles of human imprinted gene regions were obtained and 101 novel macro ncRNA candidates were identified. Furthermore, rRNA depleted RNA-seq of a fibroblast cell line validated the expression of 22/24 transcripts detected by HIRTA in fibroblasts. 7 novel macro ncRNAs were developmentally regulated in a human embryonic stem cell differentiation system. Characterization of 10 novel macro ncRNAs showed that 6 had preferentially or exclusively monoallelic expression, 5 were exclusively nuclear localized indicating their potential role in gene silencing and 2 (*ADAMTS7down* and *PEG13)* had CpG island promoters that were found to be differentially methylated regions (DMRs). Furthermore, 22 out of 101 mapped transcripts were cancer specific. These results indicate that macro ncRNAs could be a universal feature of human imprinted gene regions since each human imprinted gene region analyzed expressed at least one macro ncRNA.

3. 2. HIRTA and RNA-Seq successfully detects macro ncRNAs

3. 2. 1. Macro ncRNAs could be mapped from their expression features on the tiling array

Tiling arrays have been used for obtaining expression profiles of limited number of cell lines from selected genomic regions (e.g. Cheng (Cheng et al., 2005)) or in whole genome expression studies (e.g. Bertone and colleagues used liver polyA RNA (Bertone et al., 2004) and Kapranov polyA RNA from HeLa and HepG2 cell lines (Kapranov et al., 2007a)). These studies with some exceptions utilized Affymetrix technology and polyA RNA cell fractions and showed that the non-annotated portion of the human genome is still substantial and that high proportion of transcription consists of non-protein coding RNAs. These studies found numerous novel non-annotated transcripts from a small number of tested cell lines, but they did not examine if any novel transcripts were macro ncRNAs of the type found in imprinted gene clusters.

In this thesis, I applied total RNA to a custom NimbleGen tilling array (HIRTA) and showed that known imprinted macro ncRNAs are successfully detected using “single cDNA hybridization on HIRTA". Focusing on imprinted gene regions (HIRTA contains 2% of the human genome) allowed a higher probability of finding novel imprinted macro ncRNAs and also lowered the cost of applying higher number of samples than if the whole genome tiling arrays were used. The expression profile of macro ncRNAs from HIRTA hybridizations showed specific characteristics: a continuous high signal through the whole ncRNA gene body, a 5'-3' slope and a non-saturated transcription level. These features distinguish macro ncRNAs from protein-coding mRNA genes that typically showed a peaked pattern with a high, saturated signal over exons and a low non-saturated signal over introns. I discuss below how these features of macro ncRNA and protein coding mRNA genes may originate from differences in macro ncRNA and pre-mRNA biology and how they were used to develop a strategy for mapping novel macro ncRNAs.

3. 2. 1. 1. "Continuous high HIRTA signal" over macro ncRNA body may originate from differences in splicing

The first macro ncRNA feature observed from the six well-studied imprinted gene clusters were continuous non-saturated signals from typically more than 90% of HIRTA tiles, through the whole length of macro ncRNA gene. In contrast, most pre-mRNAs showed focal saturated signals matching exons and typically low to moderate non-saturated intronic signals. Continuous high signals from macro ncRNAs most probably results from their low level of splicing (e.g. *Airn* is mostly unspliced and *Kcnq1ot1* lacks any spliced products (Pandey et al., 2008; Seidl et al., 2006)).

Nonsense-mediated decay (NMD) is the mechanism that selectively eliminates aberrant RNAs that have premature transcription termination codons in eukaryotic cells. Splicing is coupled with NMD since specific protein complexes that mark premature codon deposited to the RNA during splicing had been found (reviewed in (Chang et al., 2007)). Thus, previously it was proposed that macro ncRNAs could loose their capacity to be spliced during evolution, to escape the nonsense mediated decay pathway (Ponting et al., 2009).

Continuous high signals may also originate from the low intron/exon ratio that had been previously shown to characterize some known macro ncRNAs, e.g. *H19* and *Xist* are spliced but have short introns relative to exon length (Brannan et al., 1990; Brown et al., 1991). These short introns were more difficult to visualize in comparison to the typical long protein-coding introns (protein coding genes are showing high intron/exon ratio (Lander et al., 2001)). Thus when expressed, short introns could not be distinguished from "continuous high signals" appearance of macro ncRNAs, while protein-coding genes that typically splice out long introns had a clearly different high-exonic/low-intronic HIRTA expression pattern.

Potential obstacles in differentiating between ncRNAs and protein coding genes based on the "continuous high signal" criterion were: 1) protein coding genes that consist of one exon (12% of human annotated genes are single-exon genes (Sakharkar et al., 2005), e.g. *NDN* located on chr15) and 2), protein coding genes that show high intronic signals. Based on the HIRTA profile alone I could not exclude that some of the novel mapped transcripts are single-exon protein coding genes and thus I used the RNAcode algorithm (Washietl et al., unpublished data) that predicts exons based on evolutionary signatures, to show that no single-exon genes were among novel mapped transcripts. Protein coding genes that show high intronic expression on the tiling array may represent transcripts with a high transcriptional rate (Ebisuya et al., 2008). I excluded these potential false positives from the mapped macro ncRNA candidates since I compared all 43 hybridized samples simultaneously. This approach maximizes possibility to exclude protein-coding genes with high transcriptional rate since these genes are expected not to have high intronic expression, but typical high-exonic/low-intronic HIRTA expression pattern in most of the tissues.

3. 2. 1. 2. 5'-3' slope macro ncRNA feature may originate in alternative polyadenylation

By visual examination of the HIRTA macro ncRNA expression profiles, as a second feature, the decreases of the HIRTA expression through the body of known macro ncRNAs were observed. These decreases showed opposite correlation with the transcription orientations of the known ncRNA transcripts and thus were named as 5'-3' slope. In this section I will discuss how 5'-3' slope macro ncRNA feature may originate in alternative polyadenilation.

The biology of macro ncRNAs is still poorly characterized. For a few tested imprinted macro ncRNAs it has been shown that they are RNA Polymerase (RNAP) II transcripts (e.g. mouse *H19*, *Airn*, *Kcnq1ot1* (Brannan et al., 1990; Redrup et al., 2009; Seidl et al., 2006)) and that have low level of splicing (e.g. *Airn*) or that they are unspliced (e.g. *Kcnq1ot1*) (Pandey et al., 2008; Seidl et al., 2006).

Mammalian pre-mRNA transcription termination is directed by sequence elements on the pre-mRNA (Mandel et al., 2008). Primary sequence elements consist of polyadenylation signal sequence, the cleavage site and the G/U rich downstream element (Zhao et al., 1999). The auxiliary elements were found downstream of the RNAP II cleavage site or upstream of the polyadenylation sequence (Zhao et al., 1999). Downstream auxiliary elements are generally G-rich while upstream auxiliary elements are usually U-rich (UUUU, UGUA or UAUA) (Bagga et al., 1995; Hu et al., 2005). Interestingly upstream auxiliary elements that enhance efficiency of cleavage and polyadenilation of primary transcript RNAs by binding auxiliary polyadenylation factors are mostly found together with enhanced transcription of intronless genes (Le Hir et al., 2003; Moreira et al., 1995). Thus, we speculate that auxiliary elements may be potentially more often found in intronless macro ncRNAs than in pre-mRNAs containing long introns, enhancing their transcription and promoting their polyadenylation.

If macro ncRNAs are RNA polymerase (RNAP) II transcripts, three non-mutually exclusive hypothesis could explain the 5'-3' slope of macro ncRNAs: 1) very long, macro ncRNAs may have a number of polyA signals that are usually not functional through the genomic region, but could potentially be more often used by RNAP II when the coupling with the splicing machinery does not take place, 2) macro ncRNAs may have a higher number of additional termination elements (auxiliary elements) in 3' regions that will advance RNAP II termination on different polyA sites comparing to pre-mRNAs 3), macro ncRNAs may be characterized by a higher level of alternative polyadenylation than pre-mRNAs and thus different cells will produce macro ncRNA transcripts of different lengths that could be seen as 5'-3' slope in the RNA pool detected by the HIRTA array (Figure 64).

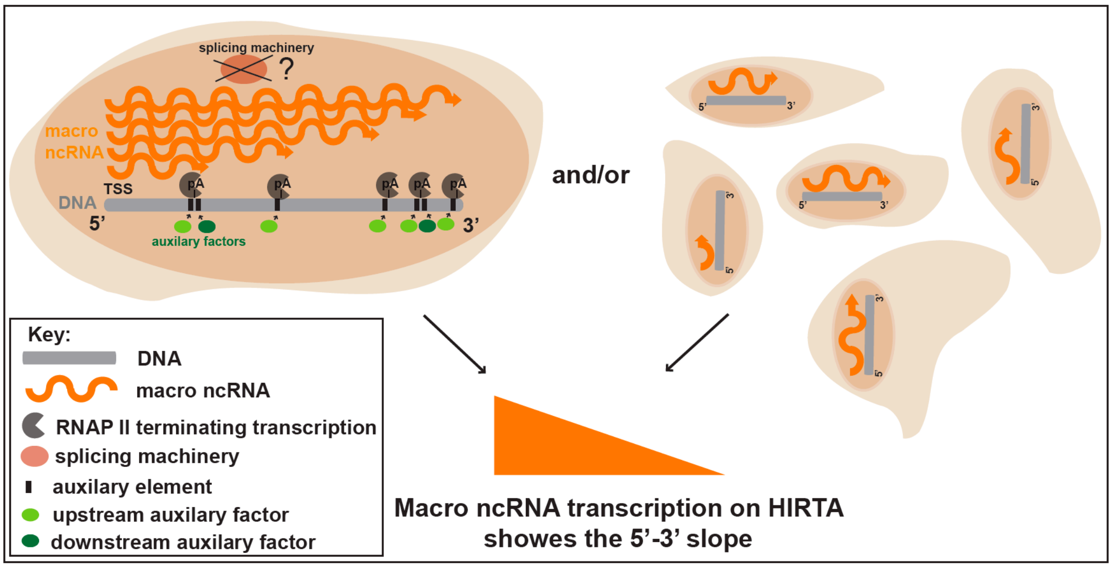


Figure 64. Model showing how 5'-3' HIRTA feature of macro ncRNAs may originate in macro ncRNA biology. A number of macro ncRNA variants may be transcribed from one cell by RNAP II, when different polyadenylation sites are more efficiently recognized with the help of auxiliary factors and in the potential absence of splicing machinery each cell of the population may transcribe macro ncRNA variant of different length by using different polyadenylation and cleavage site for transcription termination. TSS; transcriptional start site, pA; polyadenylation site. See key for further details.

These hypothesis remain to be tested, for example by using RNA FISH with different probes for potential diverse 3' ends, it could be tested if each cell transcribe just one alternatively polyadenylated macro ncRNA transcript or each cell transcribes a pool of macro ncRNA transcripts of different lengths. Further, sequences of high number of macro ncRNAs and pre-mRNAs could be tested for statistical difference in number of G-rich and U-rich auxiliary elements.

3. 2. 1. 3. 5'-3' slope macro ncRNA feature as a criteria for the transcript orientation prediction

HIRTA hybridizations did not allow strand specific information to be obtained, since double stranded cDNAs were applied to the HIRTA chips. In order to indirectly gain this information, I used 5'-3' slope expression feature as one of the criteria in prediction of the macro ncRNA strand. In addition to the slope criteria, I based prediction of the transcript orientation on three more published genomic/epigenomic features: CpG island, H3K4me3 enrichment and RNAP II binding site. I used these features since they are known marks of promoters (described in section 1.) and thus could predict the 5' ends of the transcripts. I found overall agreement between transcript orientation predictions of the novel macro ncRNAs based on these four features and the GRO-Seq data that mapped the amount, position and orientation of transcriptionally engaged RNA polymerases in the IMR90 fibroblast cell line (Core et al., 2008). Further, I found that around 50% of the macro ncRNAs that were overlapping with GRO-seq data showed presence of RNA polymerases in both directions. Thus, a certain percentage of novel macro ncRNAs that I mapped may represent overlapping transcripts supporting the “interleaved” model (extensive overlap of transcriptional units and regulatory regions on the same genomic space) proposed by Kapranov and colleagues (Kapranov et al., 2007b).

3. 2. 1. 4. Differences in HIRTA mapping for macro ncRNAs having diverse positions to the annotated genes

101 novel macro ncRNAs that were grouped as: 1) intergenic, 2) intronic, 3) adjacent to the 5’ end of annotated genes (5’/OV), and 4) adjacent to the 3’ end of annotated genes (3’/OV), were mapped using the described strategy (section 2.1.5.) based on the HIRTA expression patterns. The intergenic transcripts were mapped reliably since borders of these transcripts could be easily visualized. The borders of potential intronic transcripts could not be assigned in most of the cases when high expression was observed through the whole length of the host gene intronic regions. The intronic macro ncRNAs I was able to distinguish from the host gene expression were *PEG13* where I knew the orthologous position from the mouse, and *HTR2Ain* where the *HTR2A* host protein-coding gene was not expressed in number of tissues. Human introns have been found as a source of regulatory RNAs (Mattick and Gagen, 2001), and among imprinted macro ncRNAs some are intronic e.g. *MESTIT1* macro ncRNA (Figure 24, section 2.4.1.). Thus, there is a high potential for existence of a more intronic macro ncRNAs than already known or mapped by HIRTA in imprinted regions and for study of these transcripts strand specific mapping technology in numerous tissues would be necessary.

Many known macro ncRNAs overlap protein-coding genes (e.g. *KCNQ1OT1, UBE3A-AS, GNAS1-AS, EXON1A*), thus I also mapped transcription longer than 1kb positioned adjacent to 5’ and 3’ ends of known genes. This transcription could represent macro ncRNAs that overlap annotated genes or alternatively novel 5’UTRs or 3’ UTRs of annotated genes. Further examination of each specific example would be necessary to reveal the origin of mapped transcription. However, I showed that 29 transcripts that are positioned 5’ and 3’ to the known protein-coding gene are predicted to be non-coding by using the RNAcode algorithm (Washietl et al., unpublished data) where no protein coding potential could be found in these regions based on evolutionary signatures in 44 species.

3. 2. 2. Ribosomal RNA depleted RNA-seq successfully maps macro ncRNAs

Previous RNA-seq studies from poly (A) selected RNAs have been focused on alternative splicing of protein coding genes while the macro non-protein coding portion of the genome was not directly assessed (Pan et al., 2008; Sultan et al., 2008). Visual inspection of mRNA-seq data from these studies showed that imprinted macro ncRNAs were poorly detected showing the necessity for employment of modified RNA-seq sample preparation in order to improve the macro ncRNAs detection. Thus, RNA-seq using the Illumina/Solexa technology on total rRNA depleted RNA from fibroblasts was performed, and successfully detected the known macro ncRNAs such as *KCNQ1OT1*, *GTL2var1*, *UBE3A-ASvar2*, *HOTAIR* and *MALAT1* and 22/24 of the novel macro ncRNAs that I detected by tiling array.

With some exceptions, most of the known and novel macro ncRNAs were relatively lowly expressed in comparison to annotated protein coding genes by rRNA depleted RNA-seq. Low expression of macro ncRNAs was also shown by HIRTA hybridizations indicating that low macro ncRNA expression reflects their biology and not technical problems. Both HIRTA and RNA-seq detect the steady-state level of gene expression. The steady-state level is expected to be a result of the combined rate of transcription and degradation (stability) of the transcript. The low expression of macro ncRNAs could be based on the low stability of macro ncRNAs measured by their relatively short half-life (e.g. mouse unspliced *Airn* half-life was 2.1h, whereas *Igf2r* mRNA had half-life of 14.3h (Seidl et al., 2006)), while in the same time macro ncRNA transcription may be high. To examine if low expression of novel mapped macro ncRNAs results from combined rate of transcription and degradation, measure of transcription rate using nuclear-run on assays in diverse cells/tissues and measure of macro ncRNA stability for example by using actinomicin D for each novel lowly expressed macro ncRNA would be necessary.

In the RNA-seq data, the read coverage of the 16/24 lowly expressed novel macro ncRNAs was also often low. Due to this low coverage in 66% of the novel transcripts, if HIRTA mapping data would not be available, RNA-seq alone would not allow us to map 5’ and 3’ macro ncRNA boundaries. This indicates that the 26.2 million of uniquely mapped reads provided by the RNA-seq data was not deep enough to completely cover lowly expressed macro ncRNAs and this could be further improved by more efficient rRNA depletion of total RNA prior to the RNA-seq procedure.

The ~45% of the total read number was uniquely mapped in the Hs27 cell line by rRNA depleted RNA-seq. This number was in overall agreement with other RNA-seq data. For example Sultan et al. (Sultan et al., 2008) had 50% uniquely mapped reads using mRNA-seq. In the rRNA depleted RNA-seq from the Hs27 cell line, 20% more spliced transcripts than annotated by RefSeq genes and 25% novel exons mapping to introns of known genes, were found by TopHat alignment. This result was in concordance with previous findings that alternative splicing is one of the key factors on which complexity of higher eukaryotes is based (Matlin et al., 2005). For example, Pan et al. detected novel splice junctions in about 20% of multiexon genes from six human tissues (Pan et al., 2008). Novel spliced transcripts and exons that were detected in Hs27 cells and described in section 2.7.1., further add to the human alternative splicing complexity.

The ~60% of the unspliced reads found in Hs27 cell line after extracting TopHat spliced transcripts from Bowtie aligned uniquely mapped reads, contained known unspliced transcripts, potentially TopHat false negative transcripts, but also showed that high percentage of uniquely mapped reads may represent novel unspliced macro ncRNAs. The finding of a high proportion of unspliced transcripts in the human genome from RNA-seq data, was in the overall agreement with HIRTA study based on 2% of the genome, where 101 novel transcripts, that showed “continuous high signal” patterns and thus were potentially unspliced macro ncRNAs, were mapped. RNA-seq using rRNA depleted total RNA successfully detected macro ncRNAs but further advancement in detection is expected from more efficient rRNA depletion, paired-end sequencing, strand-specific sequencing and from further development of technology towards the longer reads lengths.

3. 3. Pervasive transcription or functional ncRNAs?

Rapid development of technologies such as tiling arrays and next-generation RNA-sequencing in the last years led to novel insights into the genome complexity. These findings led to the use of the term “pervasive transcription” that refers to widespread unannotated transcription arising over most of the human genome (reviewed in (Kapranov et al., 2007b)). In the literature the term “pervasive transcription” was not uniquely defined and fully understood and become a matter of debate in the field (for publications see: http://pervasivetranscription.com/). For example, Jacquier et al., (Jacquier, 2009) referred to pervasive transcription as transcription not restricted to well-defined functional features, such as genes, and Ponting et al. (Ponting et al., 2009) noted that pervasive transcription does not necessarily imply an abundance of functional RNAs. Oppositely, Beretta and Morillon predicted a crucial role of pervasive transcription in controlling gene expression and genomic plasticity while Mattick also suggests a functionality of transcripts (ncRNAs), derived from pervasive transcription (Berretta and Morillon, 2009; Dinger et al., 2009; Mattick and Makunin, 2006). The overall picture is that there are different possibilities of what pervasive transcription could represent: novel transcripts (coding or non-coding) that have currently unknown functional roles, spurious products of transcription (junk RNAs) that do not have a function and represent side products of functional processes in the cell, novel 5' and 3' UTRs of annotated genes, transcriptional read through or even experimental artifacts (reviewed in (Huttenhofer et al., 2005; Johnson et al., 2005; Mendes Soares and Valcarcel, 2006; Wilusz et al., 2009)).

Global transcription resembling pervasive transcription, but found specifically in undifferentiated mouse ES cells on the whole genome level, was proposed to be the key mechanism that could keep the pluripotent state of the cells. The same authors found that after 7 days of ES cell differentiation total and mRNA levels were ~2 fold lower compared to undifferentiated ES cells (Efroni et al., 2008). Global transcription of undifferentiated human ES cells in imprinted gene regions was not observed in my HIRTA studies since transcribed and silenced genomic regions were present at similar level to the differentiated ES cells.

In this study 101 unannotated transcripts were mapped in cell lines and tissues with 95% of them showing non protein coding potential as assessed by RNAcode (Washietl et al., unpublished data). It could not be fully excluded that some of these transcripts are UTRs of annotated genes, transcriptional read through, junk RNAs or experimental artifacts (originating from cross-hybridization on the tiling arrays or being background signals). However, newly mapped macro ncRNA transcripts were found in gene regions containing imprinted genes, with a number of them showing different indices of functionality that will be discussed through this section.

3. 3. 1. Novel macro ncRNAs that show tissue specific expression and subcellular localization may be functional

Tissue specific expression and a specific subcellular localization of macro ncRNAs are indices of functionality of macro ncRNAs (reviewed in (Mattick, 2009)). I tested expression of 20 normal and 23 cancer cells/tissues on HIRTA and found differential tissue specific expression of 99/101 novel macro ncRNA candidates in imprinted gene regions. Most of the macro ncRNAs (29%) were expressed in testis. This observation was in agreement with work of Sasaki where they found that 29% of their tested mRNA-like ncRNAs were expressed in testis where the testis macro ncRNA fraction was the largest among 11 human tissues. Together with similar mouse data this implies that macro ncRNAs could potentially have specific roles in complex regulatory mechanisms in testis (Ravasi et al., 2006; Sasaki et al., 2007).

I used a "double cDNA HIRTA hybridization" (section 2.1.6.) approach to identify nuclear-enriched macro ncRNAs. I confirmed the subcellular localization of the *KCNQ1OT1* macro ncRNA and identified the subcellular localizations of 15 other known and novel macro ncRNAs using this technique. I observed a limitation in the reliable interpretation of subcellular localization for very lowly expressed transcripts since correct interpretation highly depends on the position of the base line and the used normalization, however those with higher expression macro ncRNAs were reliably assigned to the nuclear or cytoplasmic compartment. A second difficulty was the reliable interpretation of cytoplasm (grey) signals that only spanned one tile and were located in non-annotated regions. These signals may represent novel exons of cytoplasm-enriched transcripts. Additionally, it cannot be excluded that one tile cytoplasmic signals observed by double cDNA HIRTA hybridization may originate from cross-hybridizations with partially complementary transcripts mapping to other genomic regions. Overcoming the limitations in the data interpretation, double cDNA HIRTA hybridization reliably detected that 8/15 transcripts were localized in both cytoplasm and nucleus of the Hs27 cells while 6 transcripts were exclusively nuclear indicating their potential role in gene regulation.

3. 3. 2. Developmentally regulated macro ncRNAs overlapping olfactory receptor genes may be functional

HIRTA expression profiling of undifferentiated and day 7 differentiated human ES cells showed that 7 novel transcripts were developmentally regulated and thus may have regulatory functions. Two of these transcripts (*OR56B4up1* and *OR5P2ov*) overlapped olfactory receptor genes (Figure 25, section 2.4.2.). Olfactory receptor genes are known as typical example of randomly monoallelically expressed genes and in mouse it has been previously shown that monoallelic expression of olfactory receptors is regulated by enhancer element H that is methylated on one chromosome and that can act *in cis* or *in trans* (Chess et al., 1994; Lomvardas et al., 2006). Further it was proposed that in mouse this enhancer element interacts just with one olfactory receptor (OR) gene per nucleus leading to expression of just this one OR gene out of 1300 OR genes, in one olfactory neuron (Serizawa et al., 2003). The mapping of two macro ncRNAs that overlap a number of OR genes raises the question if these macro ncRNAs contribute to random monoallelic expression in a similar manner as found for imprinted macro ncRNA in inducing imprinted monoallelic expression? It would be interesting to test if the macro ncRNAs overlapping OR genes expressed in human ES cells are also present in olfactory neurons, and if they have the same transcription profiles in olfactory neurons expressing different olfactory genes. One of the possibilities how macro ncRNAs may regulate OR gene monoallelic expression is by transcription through enhancer elements where macro ncRNAs could potentially set a methylation mark on enhancer of one parental chromosome, that would further activate OR gene by known mechanism involving intra or interchromosomal interaction with promoter of OR gene.

3. 3. 3. A small number of HIRTA mapped macro ncRNAs overlap with lincRNAs associated with PRC2 and CoREST complexes

Large intergenic transcripts (lincRNAs) have been mapped in human fibroblasts and HeLa cells by approach based on finding of evolutionary conserved domains where H3K4me3 and H3K36me3 overlap (Khalil et al., 2009). Only 9 out of the 101 HIRTA mapped transcripts overlapped with lincRNAs. Interestingly, 5 out of these 9 transcripts have been found to associate with PRC2 and CoREST complexes in the lincRNA study (Khalil et al., 2009). PRC2 is a methyltransferase that trimethylates H3K27 and has a function in repressing transcription of specific genes (Bracken et al., 2006). Recently, it has been shown that PRC2 interacts in mouse placenta with the imprinted *Kcnq1ot1* ncRNA (Pandey et al., 2008) that is known to have a function in silencing imprinted genes in the *Kcnq1* mouse imprinted cluster (described in section 1). The *HOTAIR* macro ncRNA that is transcribed from the HOXC cluster has been shown to bind PRC2 and to repress genes *in trans* in the HOXD cluster (Rinn et al., 2007). The CoREST complex is another chromatin-modifying complex known as a repressor of neuronal genes (Andres et al., 1999). The 5 transcripts found by both HIRTA and lincRNAs study that are physically associated with PRC2 and CoREST complexes may have a function as epigenetic initiators and target histone methylation at specific genomic location (epigenetic pathway model was described in 1.1.2. (Berger et al., 2009)).

3. 3. 4. Macro ncRNAs may function independently and as precursors for small ncRNAs

Macro ncRNAs have different roles in the cell, but one function is to act as precursors for different classes of small ncRNAs, since macro ncRNAs as well as protein-coding genes (especially their introns) can be post-transcriptionally processed into small RNAs (Fejes-Toth, 2009). Imprinted macro ncRNAs are known precursors for miRNAs and snoRNAs (e.g *H19*, *GNASAS*, *UBE3A-AS*, *GTL2,* described in section 2.7.4.), as well as some non-imprinted macro ncRNAs e.g. *MALAT1*, ~7kb nuclear macro ncRNA known to generate mascRNAs (*MALAT1* associated small cytoplasmic RNA). It is becoming apparent that macro ncRNAs and their processed small ncRNAs can have different biology (e.g. cellular localization, half-life) and possibly different functions (Wilusz et al., 2008; Wilusz et al., 2009).

The only large snoRNA clusters found in the human genome up to now are those embedded into the imprinted *UBE3A-AS* and *GTL2* macro ncRNAs. The *UBE3A-AS* macro ncRNA that is expressed from PWS imprinted gene cluster contains two large snoRNA clusters: *HBAII-85* and *HBAII-52*. The *HBAII-85* has a significant role in the Prader-willi Syndrome (PWS) (Duker et al., 2010) and the HBAII-52 could be involved in the editing and/or alternative splicing of the pre-mRNA of 5-hydroxytryptamine 2C (5-HT_{2C}) receptor that plays a role in serotonergic signal transduction (such a regulatory role has been found for corresponding mouse *MBAII-52* snoRNA cluster) (Kishore and Stamm, 2006; Nicholls and Knepper, 2001). In this thesis *UBE3A-ASvar4* macro ncRNA overlapping exclusively *HBAII-52* snoRNA cluster (Figure 16, section 2.2.1.1.) that could be a potential precursor for *HBAII-52* was found in a small number of tissues including fetal and adult brain and developmentally upregulated in hES cell system correlating with a previously shown functional role in brain.

In the DLK1 imprinted gene region, *GTL2var5* overlapping the 14qI and 14qII snoRNA clusters (Figure 14, section 2.2.1.1.) was expressed only in adult heart and uterus and could also represent precursor for those snoRNA clusters. The question for the future work is if these macro ncRNAs that overlap small RNAs function exclusively as a precursors of small RNAs or they could have in the same time other roles in the cell, presumably in gene regulation *in cis* (Figure 65).

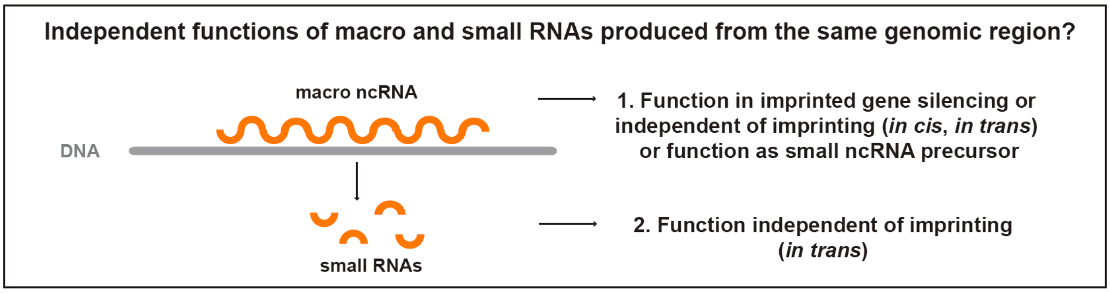

Independent functions of macro and small RNAs produced from the same genomic region?
macro ncRNA
DNA
small RNAs
1. Function in imprinted gene silencing or independent of imprinting (*in cis*, *in trans*) or function as small ncRNA precursor
2. Function independent of imprinting (*in trans*)

Figure 65. Macro ncRNAs can be precursors of small RNAs. Macro ncRNAs could function as small ncRNA precursors but may also have independent functions.

The methodology of sample preparation for rRNA depleted RNA-seq did not preserve small RNA species (less than 200bp long), but by examining 10 human chromosomes (chr1 to chr10), 25 potential primary miRNAs transcripts (novel transcripts found to overlap known miRNAs) and 1 potential precursor for snoRNA (novel transcript that overlaps known snoRNA) were mapped, of which 20 were predicted to be macro ncRNAs, 4 predicted to be novel protein-coding genes and 2 possibly overlapping both non-coding and coding transcripts.

The number of novel precursors for small RNAs that were mapped was dependent on already known small RNAs from the miRBase, thus it could be predicted that more unknown precursors and small ncRNAs may be mapped in normal human fibroblasts. An interesting question will be if these precursors always have the same expression profiles as the small RNAs processed from them in different tissues, or they could be also transcribed in the tissue specific manner without processing into the small ncRNAs. For resolving of this question RNA-seq expression profiles of both small and precursor RNAs in number of tissues should be compared.

Mapping of novel macro ncRNAs in human and indirect evidence implying their function are a valuable resource and the beginning point for direct testing of function of mapped macro ncRNAs, that is certainly the important challenge of future work (further discussed in section 3.8.).

3. 4. Developmental upregulation of *XIST* macro ncRNA in HES2 cells

X-chromosome inactivation takes place in female mammals and requires the X Inactive Specific Transcript (*Xist*/*XIST*) (Lee et al., 1996; Penny et al., 1996). Not all human ES cell lines have the ability to show developmental regulation of *XIST* ncRNA and to recapitulate X-inactivation in an *in vitro* system (Silva et al., 2008). The human *XIST* macro ncRNA was not expressed in the female undifferentiated human embryonic stem HES2 cell line by HIRTA, while was developmentally upregulated and showed relatively low expression after 7 days of HES2 differentiation (section 2.3.). Interestingly, Silva et al. could not detect any expression of the *XIST* in the same HES2 cell line in any time point before or after differentiation, using the RNA FISH technique (Silva et al., 2008). The same study classified HES2 cells to be “class III” human embryonic cell lines that have lost the capacity to express *XIST* and that do not have a counterpart in mouse ES cells (all mouse ES cells express *Xist* after differentiation). Silva et al., hypothesized that class III cells are prone to lose *XIST* expression once X inactivation is initiated in the culture, since they found that class III cells even if *Xist* was lost, still showed inactivation of X chromosome. A possible explanation for the differences in *XIST* expression through HES2 differentiation observed between HIRTA and Silva et al. studies is that RNA FISH technique used in the Silva study was not sensitive enough to detect the low levels of *XIST* expression in differentiated HES2 cells. Thus, other human embryonic stem cell lines that were explored for the *XIST* expression by the Silva study and grouped as class III cell lines, should be also examined by techniques that detect RNA expression other than RNA FISH. This could potentially show that all human embryonic cell lines have a property to show developmental regulation of *XIST* ncRNA as found for mouse ES cells (Chaumeil et al., 2002). Potentially, human HES2 cells showing upregulation of *XIST* and X-inactivation may provide a model for *in vitro* studies of epigenetic changes during early human development.

3. 5. Further support to the lack of conservation between mouse and human *Tsix*/*TSIX* function

In the mouse the *Tsix* macro ncRNA is transcribed antisense to and regulates *Xist* macro ncRNA during both random and imprinted X-inactivation (Lee, 2000; Sado et al., 2001). Human *TSIX* transcripts have been found to be co-expressed with *XIST* only from the inactive X-chromosome in human fetal cells using RNA FISH technique indicating that *TSIX* does not have the same function as mouse *Tsix* in repressing *Xist* (Migeon et al., 2002). Further, Chang and Brown found lack of conservation of regulatory elements for human *TSIX* and thus also showed that *TSIX* is most probably not a regulator of *XIST* in humans (Chang and Brown, 2010). Previously *TSIX* was found expressed from small number of cells types: chorionic villus cells, embryonal bodies and one human embryonal carcinoma cell line (Chow et al., 2003; Migeon et al., 2002). In this thesis *TSIX* was found not to be expressed in undifferentiated and differentiated human embryonic stem cells (HES2), in three fetal tissues and in all other tested cancer cell lines including teratocarcinoma and malignant embryonal carcinoma. This finding further suggests that human *TSIX* lacks a conserved function as a regulator of *XIST* macro ncRNA and if functional, *TSIX* may have function specific for the cell type where it is expressed.

3. 6. Are macro ncRNAs one of the key features in genomic imprinting?

Genomic imprinting is an epigenetic phenomenon that utilizes different mechanisms of gene regulation and thus represents a valuable model system enabling deeper understanding of how genes and genomic regions may be regulated. Genome regions that contain imprinted genes also contain different numbers of usually clustered imprinted genes, and show certain key features. The well-established key features of imprinted gene regions in human and mouse are: the presence of a short element known as the Imprint Control Element (ICE) that carries a parental specific methylation mark set in the germline (gametic Differentialy Methylated Region (DMR)) and *cis*-regulation over long distances that leads to monoallelic expression of clusters of flanking imprinted genes. In last years, it has been hypothesized that macro ncRNAs could be one of the key features of imprinted gene regions since they are expressed from the six-well characterized imprinted gene clusters and two, the mouse *Airn* and *Kcnq1ot1* have been shown to play a role in silencing of protein coding genes *in cis* (introduced in section 1.2.5.1. and 1.2.8.). Thus, we tested known human gene regions containing imprinted genes, by using tiling array (HIRTA) and found that each of the 32 tested regions containing imprinted genes also contain macro ncRNAs. Since it was previously shown that non-coding portion of the human genome is substantial (introduced in section 1.3.) it could be that macro ncRNAs are not specifically enriched in the imprinted gene regions when compared to other randomly selected genomic regions. This could be directly tested by statistically comparing macro ncRNAs expression in 32 imprinted and the same number of randomly selected genomic regions of the same length. The question that remains is if at least one of the mapped macro ncRNAs per imprinted gene region shows imprinted expression. Thus, while HIRTA mapping of macro ncRNAs that showed that macro ncRNAs are expressed from each of 32 tested imprinted gene regions to some extent supports the idea that imprinted macro ncRNAs may be universal, key feature of imprinted gene regions in human, the additional evidence will come from further testing of imprinting expression status of novel macro ncRNAs that were mapped in human regions containing imprinted genes (Figure 66).

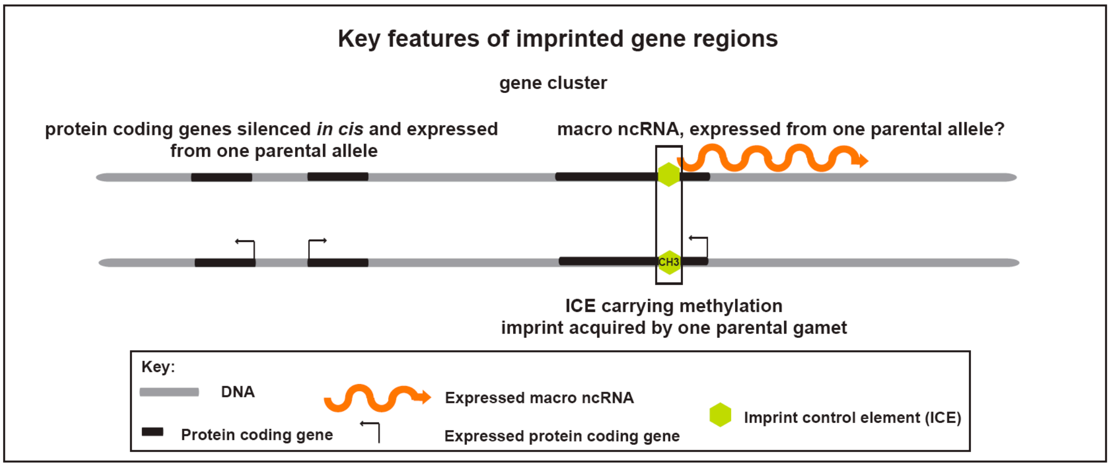

Key features of imprinted gene regions
gene cluster
protein coding genes silenced in cis and expressed from one parental allele
macro ncRNA, expressed from one parental allele?
CH3
ICE carrying methylation imprint acquired by one parental gamet
Key:
DNA
Protein coding gene
Expressed macro ncRNA
Expressed protein coding gene
Imprint control element (ICE)

Figure 66. Macro ncRNAs may be a universal feature of human imprinted gene regions. Hypothetical imprinted gene region contains ICE that carries parental specific methylation mark set in the germline and monoallelically expressed clustered imprinted genes of which at least one could be a macro ncRNA. Black box points ICE. See key for further details.

The two novel Differentially Methylated Regions (DMRs) corresponding to *ADAMTS7down2* and *PEG13* macro ncRNA CpG island promoters were found and could represent ICEs or somatic DMRs in corresponding RASGRF1 and KCNK9 regions. For testing if this DNA methylation has germline origin and was set in mother's or father's germ lines, analysis of oocytes and sperm cells would be necessary.

Monoallelic expression has been found in X-inactivation, genomic imprinting and as random monoallelic expression (reviewed in (Krueger and Morison, 2008)). Monoallelic imprinted expression that originates from a maternal or a paternal allele is one of the key features of imprinted genes. Imprinted genes can show ubiquitous or tissue specific monoallelic expression. In humans, monoallelic gene expression is usually tested from blood samples from families while testing of other tissues is highly restricted by the availability of the human samples. 6 out of 10 tested HIRTA mapped macro ncRNAs were found to show monoallelic or biased expression towards one allele based on expression of one to six Single Nucleotide Polymorphisms (SNPs), in fibroblasts and/or blood samples (section 2.4.5.). Further testing using higher number of SNPs, in more families and in other diverse tissues would further clarify imprinted status of novel HIRTA mapped macro ncRNAs. Two examples of human regions that may possess all the key features of imprinted gene regions will be further discussed below.

3. 6. 1. RASGRF1 may be novel human imprinted gene region

The Mouse *Rasgrf1* imprinted gene region (chr9) contains four paternally expressed imprinted genes: *Rasgrf1* (monoallelic in brain and liver and biallelically in lung, thymus, kidney and stomach), *As4*, *A19* ncRNA (that is monoallelically expressed in brain but biallelic in testis) and *miRNA184* (described in section 2.4.5.6.). A paternally methylated DMR located 30kb upstream of *Rasgrf1* that was shown to binds CTCF when unmethylated, functions *in cis* to silence maternal *Rasgrf1* allele by blocking enhancers located further upstream from the DMR (section 2.4.5.6., (Yoon et al., 2005)). Human genes from syntenic RASGRF1 region on chromosome 15 were not previously tested for imprinted expression and in this thesis no orthologous *A19* ncRNA expression could be seen in any of the 43 tested human samples (section 2.4.1.). However, I could detect biased expressed of the *TMED3down* and *KIAA1024up* macro ncRNAs, monoallelic expression of the *ADAMTS7down* macro ncRNA and finally, I identified a DMR located on the *ADAMTS7 down* CpG island. Thus I show, for the first time that the human RASGRF1 region is an imprinted gene region (a comparison of mouse and human Rasgrf1/RASGRF1 imprinted gene regions is shown in Figure 67).

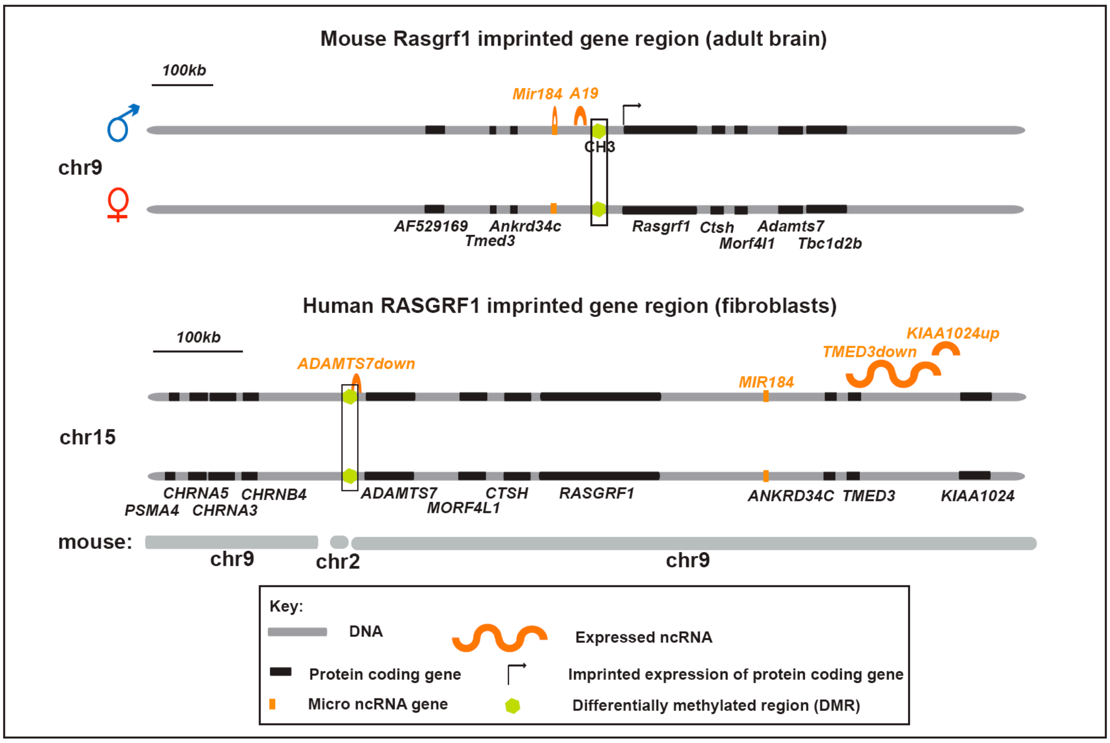


Figure 67. Mouse Rasgrf1 and human RASGRF1 region both contain imprinted macro ncRNAs. In mouse brain *Rasgrf1*, *A19* ncRNA and *Mir184* show paternal expression while in human fibroblasts *ADAMTS7down*, *TMED3down* and *KIAA1024up* show biased or monoallelic expression. Paternally methylated DMR has been found in mouse about 30kb upstream of *Rasgrf1* gene while in human CpG island promoter of *ADAMTS7down* ncRNA is a DMR. This DMR is positioned 500bp downstream of a chromosome break point in the

mouse syntenic region. Both mouse and human chromosomes are presented as on UCSC, placing chromosomal centromere always on the right side of the figure and thus inversion in the gene positions could be seen between mouse and human. Black boxes show differentially methylated region. See key for further details.

Surprisingly, expression of the *ADAMTS7down* macro ncRNA showed clear monoallelic expression in fibroblasts and blood, but paternal expression was seen based on 4 SNPs and maternal expression was seen based on 2 SNPs, by testing cells and blood of 3 families. The first possibility that could explain these findings, was random monoallelic expression combined with clonal expansion from a single cell (Krueger and Morison, 2008). Since we used polyclonal lymphoblastoid cells for RNA isolation but exclusively found monoallelic expression, this disagreement of parental origin is not expected to be due to the random monoallelic expression. The second possibility that could explain finding of both maternal and paternal imprinted expression could be due to the technical artifacts originating from the whole genome HapMap genotyping data (described in 2.4.5.6.). While finding that 4 SNPs mapping to *ADAMTSdown* were paternally expressed based on both direct examining DNA and RNA from blood of one family and from examination of HapMap data, the finding that 2 SNPs were maternally expressed came solely from examination of parental genotypes available from the HapMap data. Thus, there may be a possibility that HapMap data could misinterpret small percentage of genotypes and thus also by chance these 2 maternally expressed SNPs even if it was shown that HapMap genotyping data had been passed very strict quality control filters (2005; Frazer et al., 2007). The quality score of the each nucleotide assignment by the HapMap data to *ADAMTS7down* SNP positions could be further assessed by reanalyzing the HapMap raw data using the Genome Console, Affymetrix.

When comparing mouse and human Rasgrf1/RASGRF1 regions, it can be observed that mouse syntenic region of chromosome 15 RASGRF1 human imprinted gene region, consists mostly of chromosome 9 where mouse Rasgrf1 imprinted gene region maps, but have a break point where 25kb of the chromosome 2 was inserted (Figure 67). Similar break point was observed through other mammalian genomes (data not shown) but was not present in Primates. This break point maps to the very distal end of human *ADAMTS7down* macro RNA and thus there is the possibility that this transcript and/or its regulation could be exclusively feature of Primates. Further testing of imprinted expression of genes from this region in human brain and in mouse fibroblasts could show if this imprinted gene region is conserved between human and mouse.

3. 6. 2. Identification of *PEG13*, the human homologue of the mouse *Peg13* ncRNA

The mouse Kcnk9 imprinted gene region contains the maternally expressed *Kcnk9*, *Trappc9* genes and the paternally expressed *Peg13* ncRNA found to be ~4kb long, highly expressed in brain and unspliced (section 1.2.5.2., (Smith et al., 2003)). In human, KCNK9 imprinted gene region contains the maternally expressed KCNK9 (Ruf et al., 2007). In the study of Ruf et al., methylation from both parental alleles has been detected in human blood and brain on the CpG island orthologous to the mouse germline maternally methylated Peg13 DMR (Ruf et al., 2007).

In this thesis, for the first time a *PEG13* macro ncRNA of ~6.6kb in length, consistent with mouse *Peg13,* expressed highly in brain, but also highly expressed in blood and fibroblasts, was found. In human fibroblasts, the DMR corresponding to the mouse Peg13-DMR was found while in HeLa cells only a methylated allele was detected (Figure 30, section 2.4.5.3.). Finding of only a methylated allele in HeLa may be due to Uniparental disomy (UPD) that could retain two copies of the chromosome carrying methylated allele. Mapping of UPDs in HeLa cells would be necessary to further investigate the origin of both chromosome methylation on the examined *PEG13* CpG island.

The *PEG13* CpG island promoter was differentially methylated in a ratio that indicated allele specific methylation in fibroblasts and the same cell line also showed biased monoallelic expression of *PEG13*. Expression of *PEG13* in blood from families and human brain needs to be examined to test if the parental origin of human *PEG13* expression is same as found for mouse and to address if *PEG13* macro ncRNA has ubiquitous or a tissue-specific imprinted expression.

In summary, I show that macro ncRNAs are expressed from all tested human regions containing imprinted genes and suggest this may represent one of the key features of imprinted gene regions. Further testing of the imprinted expression and methylation status and direct testing of macro ncRNA function will lead to a better understanding of the imprinting process and its role in mammalian biology.

3. 7. Conservation of imprinted gene regions between human and mouse?

The question of conservation of imprinted gene regions in mammals is still the matter of debate in the field since it is necessarily based on provisionally set cut off of how many of genes could show differences in imprinting status between the species.

Different laboratories have different opinions if these differences indicate a general lack of conservation or just sporadic exceptions from the overall conservation of imprinted gene regions (Frost and Moore, 2010; Luedi et al., 2007; Morison et al., 2005). An interesting recent suggestion is that most of the differences between imprinted genes in mouse and human may have their origin in differences in their reproductive biology, e.g. multiple offspring compared to single offspring, that leads to a larger number of genes with imprinted expression in mouse placenta compared to human placenta (Frost and Moore, 2010). The question of conservation further gains on complexity when not only imprinted expression but also differential methylation is taken into account when comparing species. It is evident that there are differences between human and mouse in imprinted expression of certain genes but as an overall picture imprinted expression of these species may be conserved (introduced in section 1.2.5.1. and 1.2.5.2.), especially since the human imprinting field still suffers from the lack of tissue specific imprinted expression reports due to the difficulties in obtaining normal tissues from families. Thus, regions selected for the HIRTA array were not just those containing known human imprinted genes but also those whose imprinted status was only known in mouse. The identification here of novel human *PEG13* ncRNA homologue that is expressed preferentially from one allele and has a DMR, also supports the conservation between mouse and human imprinted regions.

3. 8. Macro ncRNAs from imprinted regions are deregulated in cancer

The role of macro ncRNAs in disease started to come into view with findings of a few macro ncRNAs that showed deregulation correlating with disease (introduced in section 1.3.3.). Deregulation of macro ncRNAs were usually reported as sporadic cases but few studies showed this process on the large scale. For example one study showed that number of non-coding ultraconserved expressed regions that were longer than 200bp (UCRs) were deregulated in chronic lymphocytic leukemia's, colorectal and hepatocellular carcinomas when compared with their normal counterparts, indicating their possible involvement in tumorigenesis (Calin et al., 2007). Perez et al. also showed that 15 ncRNAs that were >400bp in length, were deregulated in breast and ovarian cancers (Perez et al., 2008). In this thesis (in section 2.6.) finding of 22 macro ncRNAs expressed exclusively in cancers and a number of known and novel ncRNAs showing complete upregulation or downregulation when normal and cancer tissues were compared, provided additional evidences that macro ncRNAs may be involved in tumorigenesis of tested rhabdomyosarcoma, AML, MPD, cervical, breast and colon cancers.

The observed lack of expression of macro ncRNAs in cancer compared to normal counterpart tissues could be caused by both genetic (e.g. Uniparental disomies, deletions) and epigenetic (e.g. DNA methylation/ repressive histone modifications on regulatory regions) changes. For example my results show the imprinted *GTL2var1* macro ncRNA was not expressed in 9/9 cervical cancer and 2/2 breast cancer lines, but was expressed in normal colon and breast tissues. This raises the possibility that *GTL2var1* expression or the genomic region, from which it was expressed, may be changed in cancers comparing to the tested normal tissues. Further, this change could happen early in tumorigenesis since all tested cancer cell lines resembling different stages and/or cell types of the cervical and breast cancers showed lack of *GTL2var1* expression. In order to resolve the potential cause of observed lack of macro ncRNA expression further genetic and methylation tests of the examined cell lines and patients would be necessary. In particular since *GTL2var1* is only expressed from the maternal chromosome it would be necessary to investigate if cells still retained this parental chromosome.

Macro ncRNAs from imprinted gene regions found to be upregulated in cancers by HIRTA are valuable resource for further examination involving patients having different stages of disease, in order to study their potential usage as biomarkers for specific grades of different cancer types. Most valuable candidates for this type of research are the 16 macro ncRNAs that were found to be upregulated uniquely in specific cancer types.

AIRN macro ncRNA was shown to be expressed in the STA-WT3ab Wilms' tumor cell line and in 42.5% of Wilms' tumor patient samples, but it was not expressed in two other cancer cell lines nor in 20 tested normal tissues. Mouse *Airn* has a function in silencing of *Igf2r in cis* leading to its' imprinted expression. *IGF2R* loss of function genetic mutations accompanied by loss of heterozygosity are often seen in human tumors, suggesting its role as a tumor suppressor gene (De Souza et al., 1997; Sleutels et al., 2002). However, no clear correlation between high *AIRN* and low *IGF2R* expression in Wilms' tumors could be seen (Figure 54, section 2.6.3., (Yotova et al., 2008)) and, to date there is no evidence that *IGF2R* could undergo epigenetic inactivation during tumorigenesis.

Expression changes in cancer of 5/6 well-studied imprinted macro ncRNAs were identified. Imprinted macro ncRNAs could be also subjected to loss of imprinted expression (LOI) (introduced in section 1.2.9.2.). Imprinted macro ncRNAs that show

LOI have biallelic expression and thus their expression is expected to be two-fold upregulated in comparison to their monoallelic imprinted expression. Macro ncRNAs expression detected by HIRTA showed numerous subtle changes between normal and cancer cells but these changes remain to be quantified using for example qRT-PCR that can reliably detect the two fold changes. In order to examine if LOI and not upregulation from the same allele that monoallelically express macro ncRNA occurred RNA FISH could be used.

Myeloproliferative disease (MPD) can evolve into acute myeloid leukemia (AML). The higher number of macro ncRNAs found in all AML samples comparing to all MPD samples may indicate a globally upregulation during progression of MPD towards AML. Since only 6 patient samples were tested, number of the patients should be expanded to gain further evidence for potential global upregulation of macro ncRNAs through leukemia development. The *OR56A1down* macro ncRNA that was found exclusively in 2 AML patients and the *PPP2R5Cup1* macro ncRNA found in 1/3 AML and 2/2 MPD patients could be potential biomarkers of these diseases since they were not found in any of normal and cancer tissues. Further investigation of these macro ncRNA would be necessary to reveal more about their potential function in tumorigenesis.

Further detailed examination of each of 22 specific macro ncRNAs candidates, using a number of cancer patients as well as over-expression studies and testing their potential function in proliferation and/or apoptosis cell-based assays, may enlarge our knowledge about the role of macro ncRNAs in gene regulation and disease and show if these macro ncRNAs may have potential as novel biomarkers or drug targets.

3. 9. Finding function of macro ncRNAs in development and disease

In this thesis 101 novel macro ncRNAs were mapped in 43 tested cells/tissues and a number of them showed indices of functionality that were discussed through section 3.3. These findings are a valuable resource for further functional tests of mapped macro ncRNAs. Functional tests could involve the macro ncRNA knock down. RNA interference (RNAi) knock down of macro ncRNAs using small interfering RNAs (siRNAs) has been found restricted to cytoplasm in human cells and thus not useful in degradation of nuclear macro ncRNAs (Zeng and Cullen, 2002). One of the recent possibilities for targeting macro ncRNA in human cells involves custom designed zinc-finger nucleases, but still this method is not widely used because of high costs (Hockemeyer et al., 2009). A second possibility for the knock down of macro ncRNAs and examination of their function may be chemically modified chimeric antisense oligonucleotides (ASO) that have been found efficient in degradation of nuclear ncRNAs in mammalian cells in culture (Ideue et al., 2009).

Some of the novel mapped macro ncRNAs identified in this thesis may be involved in imprinting where they could silence genes *in cis* as shown for the mouse *Airn* and *Kcnq1ot1* ncRNAs, by mechanisms such as transcriptional interference (TI) or RNA-mediated targeting (reviewed in (Koerner et al., 2009; Pauler et al., 2007)). Although expressed from regions containing imprinted genes, macro ncRNAs may also have functions independent to imprinting. Examples of roles played by non-coding RNAs include: 1) as regulatory RNAs regulating gene expression *in trans* through mechanisms as targeting proteins to specific genomic loci to control transcription, 2) modulating activity of proteins that are bound to macro ncRNA, 3) modulating alternative splicing, 4) altering protein localization, 5) novel housekeeping RNAs fulfilling structural roles (reviewed in (Wilusz et al., 2009)). In numerous human imprinted gene regions more than one macro ncRNA was mapped per cluster and thus it may be possible that while one of the mapped macro ncRNA has a function in silencing *in cis* (on the parental chromosome from which the macro ncRNA is expressed) leading to imprinted expression of all protein coding genes in the imprinted gene cluster, other macro ncRNAs from the same cluster may have diverse other functions that are independent of imprinting (Figure 68).

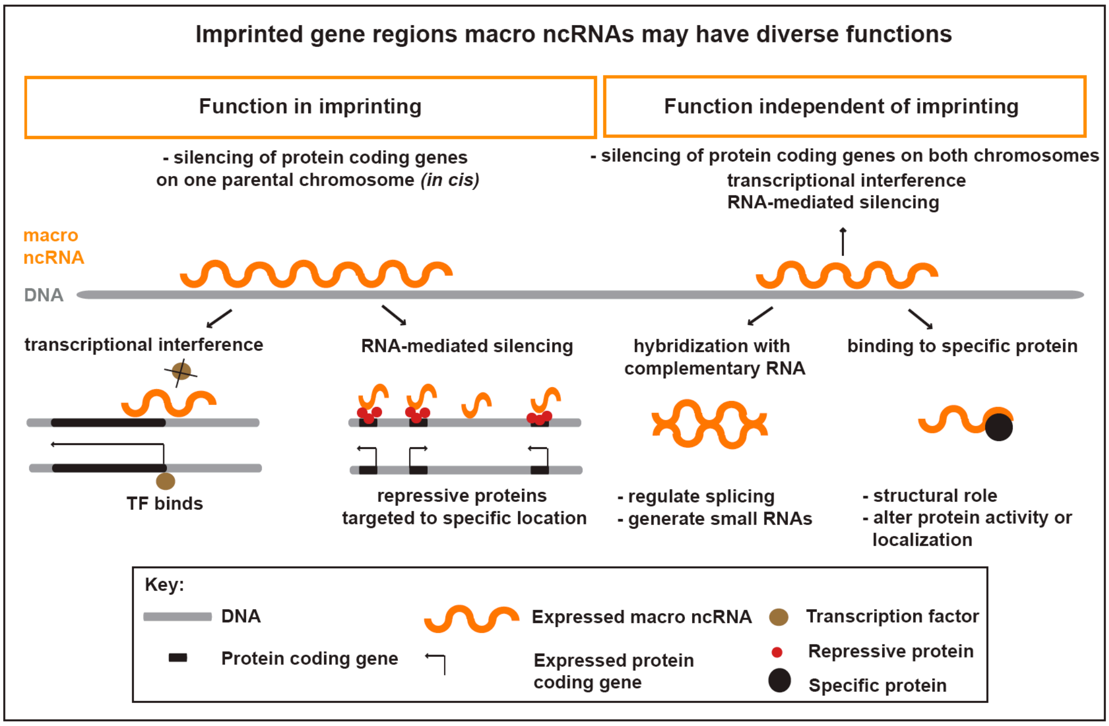

Imprinted gene regions macro ncRNAs may have diverse functions
Function in imprinting
Function independent of imprinting
- silencing of protein coding genes on one parental chromosome (in cis)
- silencing of protein coding genes on both chromosomes
transcriptional interference
RNA-mediated silencing
macro ncRNA
DNA
transcriptional interference
RNA-mediated silencing
hybridization with complementary RNA
binding to specific protein
TF binds
repressive proteins targeted to specific location
- regulate splicing
- generate small RNAs
- structural role
- alter protein activity or localization
Key:
DNA
Protein coding gene
Expressed macro ncRNA
Expressed protein coding gene
Transcription factor
Repressive protein
Specific protein

Figure 68. Models showing how macro ncRNAs from imprinted gene regions may function in or independently of imprinting. Transcriptional interference and RNA-mediated silencing are shown as most common models for macro ncRNA imprinted gene silencing *in cis*. In transcriptional interference transcription of a macro ncRNA through the promoter or an enhancer leads to the silencing of a protein coding gene by interfering with binding of necessary transcription factors. In RNA-mediated silencing the macro ncRNA acts as a locus specific epigenetic initiator inducing recruitment of repressive proteins and inducing chromatin remodeling and histone modifications. Some of the potential macro ncRNA functions independent of imprinting induced by hybridization to complementary RNA or binding to specific proteins are shown. Further, macro ncRNAs may regulate gene expression on both parental chromosomes *in cis,* but also on other chromosomes *in trans*. See key for further details.

Mechanisms how and when during cancer progression macro ncRNAs may be involved in tumorigenesis are still far from understood, and different models could be proposed how these transcripts may have a role in cancer progression (Figure 69). Imprinted macro ncRNAs can lose their imprinted expression (LOI) and become biallelic in cancer (e.g. *KCNQ1OT1* in colorectal cancer and *WT1-AS* in Wilms' tumor (Malik et al., 2000; Tanaka et al., 2001)) while in normal tissues biallelic macro ncRNAs potentially could gain imprinted expression (GIE) in cancer, further adding to the complexity of potential roles of macro ncRNAs in cancer (Figure 69). Cancer is a genetic disease generally based on the changes involving tumor suppressors and oncogenes. Up to date two examples of a macro ncRNAs correlating with silencing of tumor suppressor genes were found: *P15AS* that silence *P15* in leukemia (Yu et al., 2008) and *EPCAM-MSH2* fusion ncRNA that silence *MSH2* in Lynch syndrome (Ligtenberg et al., 2009; Niessen et al., 2009), while no example of activation of oncogenes by macro ncRNAs was found even if macro ncRNAs may act as transcriptional activators, as shown for mouse *Evf-2* macro ncRNA (Feng et al., 2006). Further investigation of novel macro ncRNAs expressed form human gene regions containing imprinted genes, that were deregulated in cancer may provide novel biomarkers and/or potential drug targets (see section 1.3.4.).

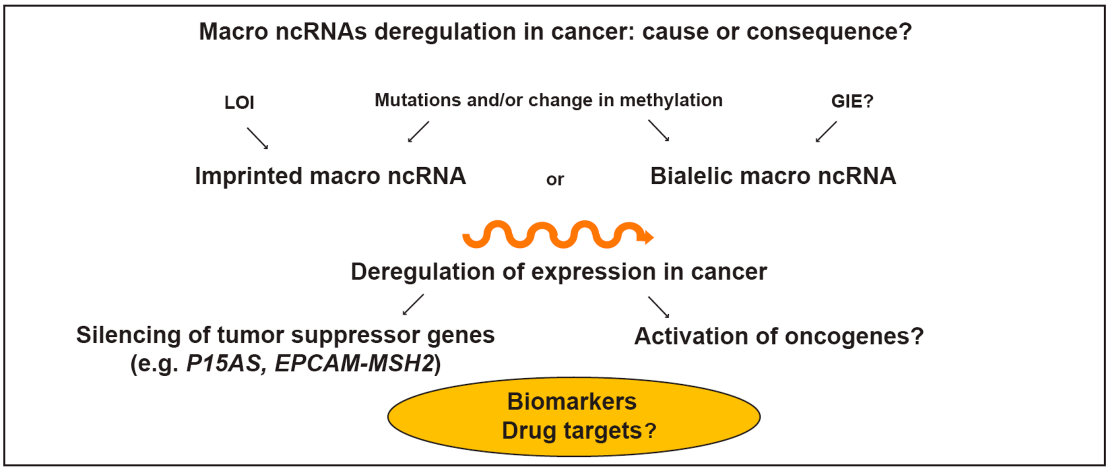

Macro ncRNAs deregulation in cancer: cause or consequence?
LOI
Mutations and/or change in methylation
GIE?
Imprinted macro ncRNA
or
Bialelic macro ncRNA
Deregulation of expression in cancer
Silencing of tumor suppressor genes
(e.g. *P15AS, EPCAM-MSH2*)
Activation of oncogenes?
Biomarkers
Drug targets?

Figure 69. Macro ncRNAs are deregulated in cancer and their role in cancer progression remains to be tested. Genetic and methylation changes, loss or gain of imprinted expression (LOI or GIE) may deregulate macro ncRNA expression and in combination with other genetic and epigenetic factors lead to silencing of tumor suppressor genes or potentially activation of oncogenes. Deregulated macro ncRNAs may potentially be used as biomarkers and/or drug targets.

4. Conclusion

In this thesis I showed that novel macro ncRNAs could be successfully mapped from their expression features on a tiling array and by using rRNA depleted RNA-seq technology. Mapping of 101 novel macro ncRNAs that showed different indications of functionality and were located in human regions containing imprinted genes further supports the idea that imprinted macro ncRNAs may be a universal feature of imprinted gene regions in human. The 22 novel macro ncRNAs that were expressed exclusively in cancer cells/tissues may provide the starting point for biomarker research. The data presented in this thesis overall broadens the knowledge about macro ncRNAs from human imprinted gene regions in both normal and disease conditions and will be an valuable resource for further studies of the function of macro ncRNAs in regulating imprinting and potentially in other cellular processes.

5. Materials and methods

The overview of used chemicals, other materials and kits, with information about manufacturers, is shown in Appendix table I. Primers used for the allelic expression tests are listed in the Appendix table II, at the end of the section.

5. 1. Human samples

5. 1. 1. Cell lines

Cell lines were purchased from the ATCC, http://www.lgcstandards-atcc.org/ (Table 22), Coriell Cell Repositories, http://ccr.coriell.org (Table 23), or obtained from Peter Ambros's laboratory, St. Anna Kinderspital, Vienna (STA-WT3ab human Wilms' tumor cell line). The ATCC purchased cell lines were grown as adherent cultures under cell culture conditions using media, 10%-15% FBS, gentamicine (50µg/ml) and supplements recommended by repository (used media are listed in Tables 22 and 23). The adherent STA-WT3ab cells were cultured in RPMI-1640, 25ml FBS (12.5%), supplemented with 5ml of HEPES (25mM) and 2.5ml of sodium pyruvate (1mM) per 200ml of media. Cells were subcultured by trypsinization before they reach a 80% to 90% confluence and media was renewed twice a week.

ATCC purchased					
Designations	**ATCC Number**	**Media**	**Designations**	**ATCC Number**	**Media**
Hs27	CRL-1634	DMEM	HT-3*	HTB-32	McCoy's 5a
HeLa	CCL-2	DMEM	C-4 I*	CRL-1594	Waymouth's 752/1
HCT116	CCL-247	McCoy's 5a	C-4 II*	CRL-1595	Waymouth's 752/1
Caco-2	HTB-37	MEM	SiHa*	HTB-35	MEM
MCF7	HTB-22	MEM	C-33 A*	HTB-31	MEM
CAMA-1	HTB-21	MEM	DoTc2*	CRL-7920	MEM
Tera2	HTB-106	McCoy's 5a	ME-180*	HTB-33	McCoy's 5a
NCCIT	CRL-2073	RPMI-1640	SW756*	CRL-10302	Leibovitz's L15
A-204	HTB-82	McCoy's 5a	SK-NEP1*	HTB-48	McCoy's 5a
SH-SY5Y	CRL-2266	MEM	G-401*	CRL-1441	McCoy's 5a
PA-1	CRL-1572	MEM			

Table 22. Designations and ATCC number of 21 ATCC purchased cell lines are shown.
* Cell lines that were grown by Dr. Iveta Yotova, Medical University Vienna

The Corriel purchased lymphoblastoid cell lines were grown as suspension cultures in RPMI-1640, 15%FBS media. Cultures were typically seeded at a concentration of about 250,000 viable cells/ml (concentration of viable cells was measured using Casy Cell Counter, Scharfe System) and split after 2-4 days before the cultures reached confluence.

Coriel purchased			
Designations	**Catalog ID**	**CEPH/UTAH pedigree**	**Media**
Lymphoblastoid	GM12878	1463	RPMI-1640
Lymphoblastoid	GM10854	1349	RPMI-1640
Lymphoblastoid	GM10846	1334	RPMI-1640

Table 23. Lymphoblastoid cell lines were purchased from the Coriell Cell Repositories. Listed cell lines originated from the three pedigrees of the CEPH/UTAH collection used in the international HapMap project.

5. 1. 2. Differentiation of the human embryonic stem cells

Human embryonic stem cells HES-2 (NIH code ESO2) were grown as undifferentiated for one day on ES cell media supplemented with basic fibroblast growth factor (bFGF). Further, cells were differentiated on gelatinized dishes in ES cell media without bFGF. Human embryonic stem cells were differentiated by Dr. Helia Berrit Schonthaler from laboratory of Erwin Wagner, IMP, Vienna (present address: CNIO, Madrid).

5. 1. 3. Collection of human normal and patient tissue samples

Whole blood was obtained from two healthy volunteers (mother and child) in the AKH, Vienna. Blood was collected in EDTA anticoagulant tubes to prevent clotting and processed for DNA and RNA isolation on the same day. The 24 normal tissue RNAs used for preparation of samples for HIRTA hybridizations and RT-PCR testing of *AIRN* expression were obtained from Clontech as a Human Total RNA Master Panel II or as separate Total RNAs.

The peripheral blood and bone marrow tissue samples from Acute Myeloid Leukemia patients and Myeloproliferative disorder patients (see description in Table 11, section 2.1.2.) were obtained from the Robert Kralovics laboratory, CeMM, Vienna and further processed for RNA isolation.

The 123 Wilms' tumor patient cDNA samples that were screened for the expression of *AIRN* ncRNA, were obtained from the laboratory of Prof. Martin Gessler, TBI, Wuerzburg. The Wilms' tumor samples were part of the German SIOP/GPOH 93-01 study and the patients were predominantly with preoperative chemotherapy as mandated by the European protocol.

5. 2. DNA isolation

5. 2. 1. DNA isolation from the cell lines

The cells (presented in sections 5.1.1. and 5.1.2.) were washed twice with 1XPBS and lysed using 1ml of DNA Lysis buffer per T75 flask. Lysates were incubated overnight at 55ºC. 300µl of saturated (>5M) NaCl was added per 700µl of the lysed cells and mixed by inversion. The samples were spun for 10min using 14000rpm at room temperature. The supernatant was added to 0.6V of 4ºC isopropanol and shaker by inverting to precipitate the DNA. The sample was spun 10min using 14000rpm at 4ºC and the supernatant was discarded. In order to free the pellet of salt, samples were washed with 1ml of 70% ethanol for 15min at room temperature. The sample was spun for 10min using 14000rpm at 4ºC, the supernatant was removed, and the spin was repeated for 1min. After complete removal of the supernatant, the sample was resuspended in 200µl of TE buffer and further incubated overnight at 55ºC to dissolve the DNA. Concentration of the samples was measured using the NanoDrop ND-1000 Spectrophotometer, Peqlab.

Solutions:
DNA Lysis buffer
1xTEN pH 9.0 (50mM Tris pH 9.0, 20mM EDTA pH8.0, 40mM NaCl in MQ H_20)
1% SDS
0.5mg/ml Proteinase K

TE buffer
10mM Tris-Cl pH8.0
1mM EDTA

5. 2. 2. DNA isolation from blood

DNA was isolated from blood samples according to the manufacturer instructions using Wizard Genomic DNA Purification Kit, Promega. Briefly, 3ml of the blood from the anticoagulant EDTA tubes was mixed, red blood cells were lysed using 9ml of Cell Lysis Solution for 10min at room temperature and samples were centrifuged using 2000g for 10min at room temperature. For the RNA isolation the white pellet was resuspended in TRI Reagent (further processed as in the section 5.6.1.). For the DNA isolation, the white pellet, consisting mostly of the white blood cells, was resuspended in 3ml of the Nuclei Lysis solution supplemented with RNase Solution (15µl per 3ml sample volume) and the sample was incubated for 15min at 37ºC. After incubation, Protein Precipitation Solution was added (1ml per 3ml sample volume), vortexed for 20sec and centrifuged using 2000g for 10min at room temperature. RNA was further extracted using isopropanol (3ml of isopropanol per 3ml of the sample volume) and washed with 75% ethanol. DNA was dissolved in DNA Rehydratation Solution (250µl per 3ml sample volume) according to the manufacturer recommendations. DNA was stored at 4ºC.

5. 3. Non-quantitative polymerase chain reaction (PCR)

DNA isolated from fibroblasts was used as a template in Polymerase Chain Reaction (PCR). Non-quantitative PCR was used for the preparation of Southern blot probes for testing methylation status of CpG island promoters of macro ncRNAs. Primers were designed using Primer3 (http://frodo.wi.mit.edu/primer3/) or using Primer-BLAST (http://www.ncbi.nlm.nih.gov/tools/primer-blast/) and synthesized by VBC-Biotech Service GmbH (http://www.vbc-biotech.at/) by Standard DNA-Oligonucleotide Synthesis protocol. Primers for the Southern probes were specifically placed in the genomic region of interest with the goal to amplify single copy sequences in length from 300bp to 1kb (probes and primers listed in Table 24). In a typical PCR reaction: 5µl 5X GoTaqPol buffer, 1µl 25mM $MgCl_2$, 0.5µl 10mM dNTP, 1µl 10pmol/µl forward primer, 1µl 10pmol/µl reverse primer, 2µl 5M Betaine, 0.125µl 5U/µl GoTaqPol and H_2O, were used in a 25µl reaction. The PCR cycle conditions were: 95ºC, 3min; 35 cycles of 95ºC, 30sec; 59ºC, 30sec; 72ºC, 45sec and final extension of 72ºC, 7min (with optimization of the annealling temperature and the extension time depending on the primer pair). PCR was performed using the Peltier Thermal Cycler PTC-200.

Name of the probe	Primer Name	Sequence	Length of the probe (bp)
LRRC47SBP	LRRC47SBPF	AAGCCTCTCTGGAGGAGGAG	381
	LRRC47SBPR	AGAAAAAGGTGGGACAGTGC	
PEG13SBP	P13SBP1F	TGCATTCAGGCTCACGCGCT	451
	P13SBP1R	GCTGCCTGGCCAAAAGATGGCT	
PRKCDBPSBP	PRKCDBPSBPF	AGGCAGCGGCTGTATTAGAA	801
	PRKCDBPSBPR	CTTGCGCTCACCATCAATAA	
SLC38A4SBP	SLC38ASBPF	CCTTTTCATTTGACCCTGGA	865
	SLC38ASBPR	ACTCAAAGGGGGTTGTTGTG	
ADAMTS7CSBP	ADAMTS7SBPF	ATCCGTATATCCCCTGGACC	384
	ADAMTS7SBPR	CCCAGTACAGAACTGAGGGC	

Table 24. Primers for the Southern blot probes

5. 3. 1. Dissolving the gel slice

PCR reactions and marker (GeneRuler 100bp Plus DNA Ladder) were loaded into a 2% agarose gel in 1xTAE, electrophoresed and stained using ethidium bromide solution (1mg/l). Fragments corresponding to expected probes lengths were excised from the gel under UV light and DNA was purified from the gel slice using Wizard SV Gel and PCR Clean-Up System, Promega according to manufacturers' protocol.

Solutions:
TAE Buffer
40mM Tris
0.1142% glacial acetic acid
1mM EDTA pH8.0

5. 4. Cloning

Cloning of the Southern and Northern blot probes and of the PRKCDBPCIE2 band (section 2.4.5.4.) was done using the pGEM-T Easy Vector System I, Promega. In the first step, PCR products isolated from the agarose gel were ligated into the pGEM T-Easy vector as recommended by manufacturer. 1:1 and 1:3 ratios of vector to fragment were used, and 10µl ligation reactions were assembled using 2X Rapid Ligation buffer, T4 DNA Ligase, pGEM T-Easy vector and the fragment of interest. Ligation was performed at 16ºC overnight.

In the second step, transformation of *E.coli* (JM109 competent cells) with ligated plasmid containing insert of interest, by heat shock, was performed. Bacteria were mixed with plasmids and incubated on ice for 30min. Heat shock was performed at 42ºC for 1min. The reaction was placed on ice for 2min and 250µl of LB media per tube was added. Reactions were shaked at 37ºC for 1h. The reactions were plated on Ampicilin plates containing IPTG and X-Gal and growned on 37ºC overnight. Colonies containing inserts of interest were selected as white colored colonies, while negative colonies were blue colored.

The positive white colonies were inoculated in 3ml LB with Ampicilin cultures and grown on 37ºC for 16h. The plasmid minipreps were prepared using Alkaline Lysis Protocol, Sambrook and Russel, Molecular Cloning, 2001. 1.5ml from each 3ml overnight culture were centrifuged using 14000rpm for 1min at room temperature. The supernatant was removed and the cells were resuspended on ice in 225µl of Alk-1, than 450µl of Alk-2 was added, mixed by inverting and after 5min 340µl of Alk-3 was added. The reaction stayed 5min on ice prior to centrifugation at 14000rpm for 7min at room temperature. The DNA pellet was washed in 1ml of 75% ethanol for 5min at room temperature and centrifuged again using 14000rpm for 7min at room temperature. The plasmid DNA pellet was resuspended in 100µl of TE buffer and dissolved at 55ºC overnight. Concentrations were measured using the Nano Drop. Plasmids were sent to Agowa (http://www.agowa.de/) for sequencing from the T7 promoter and the sequences of all DNA fragments that were further used as Southern blot probes, were confirmed.

Solutions:

Alk-1
50mM glucose
25mM Tris pH 8.0
10mM EDTA pH 8.0

Alk-2

0.2M NaOH
1% SDS

Alk-3
3M KAc
11.5% glacial acetic acid

5. 5. DNA methylation analysis by the Southern blot

5. 5. 1. DNA digestion for the Southern probe preparation

The DNA fragments for the Southern probes were cloned into pGEM T-Easy vector (probes LRRC47SBP, PEG13SBP, PRKCDBPSBP, SLC38A4SBP and ADAMTS7CSBP), or were already available in the laboratory as a part of pE3Up vector (pBlueskript II SK with 5.7kb IGF2R region containing CGI2-DMR, X83701) (probe Bx). The probes were digested from the plasmids using the EcoRI enzyme for pGEM T-Easy cloned probes or the BstXI enzyme for excision of the 600bp long Bx probe from the pE3Up vector. Digestion reactions using recommended buffers (EcoRI buffer or O buffer for BstXI) in total volume of 20μl or 40μl were set up and digestion of DNA took place at 37ºC (with EcoRI) or 55ºC (with BstXI) overnight. DNA digestions were separated on 2% agarose gels together with a marker (GeneRuler 100bp Plus DNA Ladder) and DNA probes were isolated from the bands of expected sizes (as described in the section 5.3.1.).

5. 5. 2. Labeling, cleaning and assessing of Southern probes activity

Labeling of the Southern probes was done using ^{32}P radioactive isotope. 20ng of the DNA probe was made up to 14μl with MQ H_2O, denatured for 5min at 100ºC and to the labeling reaction was added: 20μl LS, 6μl CTG mix, 1μl Klenow fragment and 2μl α^{32}P dATP (total reaction volume=43μl). The labeling mixture was incubated overnight (not more than 18h) at room temperature. Cleaning of the probe to remove unincorporated nucleotides, salts and contaminants was done using G50-Sephadex columns prepared as 1ml syringes stuffed with glass wool and filled with Sephadex glass beads dissolved in TE by centrifugation using 3000rpm for 3min. Labeled probes were diluted with 60μl TE, loaded into the column and respun at 3000rpm for 3min. A 1:100 dilution of the cleaned probe in TE was measured by Liquid Scintillation Analyzer 1600 TR and probes showing 20,000 to 100,000 cpm were further used.

Solutions:

LS

25ml 1M Hepes pH 6.6
25ml of OL (25ml 1M Tris pH 8.0 + 10ml 25mM $MgCl_2$+ 350µl ß-merkaptoethanol)
1ml of TM (50units of pd$(N)_6$ Random Hexamer Primer in 1.6ml TE, pH 8.0)

CTG mix

100µM dTTP
100µM dCTP
100µM dGTP
2mg/ml BSA

5. 5. 3. DNA digestion using the methylation sensitive enzymes

DNA methylation analysis was done using methylation sensitive enzymes (BstUI, BssHII, EglI, NotI) and non-methylation sensitive enzymes (EcoRI, HindIII) in single or double enzyme digestions. The 20µg of genomic DNA from Hs27, HeLa, A-201, Tera2, NCCIT and PA-1 cells, were digested with 2µl of the enzyme in 40µl reaction using buffers recommended for single or double digestions with chosen enzymes (http://www.fermentas.com/en/tools/doubledigest). Each digestion reaction was separated on a 0.8% agarose gel in 1XTBE. The gels were stained using ethidium bromide and photographed with the ruler. DNA methylation assays are presented in Table 25.

Region tested	Enzymes	Name of the probe	Tested cell lines
LRRC47down, CpG island: 92	BstUI/EcoRI BssHII/EcoRI EglI/EcoRI	LRRC47SBP	Hs27, HeLa
PEG13, CpG island: 210	BssHII/EcoRI	PEG13SBP	Hs27, HeLa
PRKCDBPup, CpG island: 108	BssHII/EcoRI BssHII/HindIII	PRKCDBPSBP	Hs27, HeLa
SLC38A4down2, CpG island: 106	BstUI/EcoRI BssHII/EcoRI	SLC38A4SBP	Hs27, HeLa
ADAMTS7down, CpG island: 17	BstUI/EcoRI	ADAMTS7CSBP	Hs27, HeLa
AIRN, CGI2-DMR	NotI/EcoRI	Bx	A-201, Tera2, NCCIT, PA-1, Hs27

Table 25. DNA methylation assays. CpG islands that are potential promoters of macro ncRNAs were tested for methylation status using double digestions with methylation sensitive and non-methylation sensitive enzymes. Names of the probes used for Southern blot and positions of the probes according to UCSC (hg18) are shown.

Solutions:

TBE buffer

89.1 mM Tris
89 mM Boric acid
1mM EDTA pH8.0

5. 5. 4. DNA Blotting and hybridization with the probe

The digested DNA from section 5.5.3. was electroforesed and photographed under UV light and the DNA were further denaturated into single stranded DNAs by shaking agarose gels twice for 30min covered with Denaturing solution. First three sheets of 3MM Whatman paper pre-wet in Denaturing solution were placed on the glass plate with the overhanging ends dipped into the solution. Than the gel was placed upside down on the top of the Whatman papers and the edges of the gel were covered with plastic strips in order to prevent buffer short circuits. The Hybond XL membrane that was prewet in MQ H_2O and than in denaturing solution was placed on the gel and two Whatman papers, a stack of the paper towels, glass plate and a weight were placed on the top. DNA was transferred from the gel to the membrane by capillary transfer that take place over at least 18h.

The blots were disassembled and neutralized with 200ml of 20mM Na_2HPO_4 for 2min. The blots were pre-hybridized with Church buffer for 1 to 3h at 65ºC in hybridization tubes. After removing the pre-hybridization solution, Hybridization buffer (Church buffer containing denatured, cleaned and labeled Southern blot probe) was placed into the hybridization tube containg the membrane, for at least 18h. The membrane was washed twice with Wash buffer at 65ºC for 30min. The sealed membrane was exposed to the Phosphoimager screen overnight. The membrane was scanned using Typhoon Scanner 5600, Amersham.

Solutions:

Denaturing solution
0.5M NaOH
1.5M NaCl

Church buffer
500ml 0.5M Na_2HPO_4
350ml 20% SDS
2ml 0.5M EDTA
Filled up to 1l with MQ H_2O

Wash buffer
40ml 0.5M Na_2HPO_4
50ml 20% SDS
Filled up to 1l with MQ H_2O

5. 6. RNA isolation

5. 6. 1. RNA isolation using TRI Reagent

RNA work was done under strict RNA-free conditions. RNA was isolated from samples listed in section 5.1, using TRI Reagent, according to the manufacturers' protocol. Briefly, cells and tissues were lysed in TRI Reagent, for 5min on the room temperature. 1ml of TRI Reagent per $10cm^2$ of culture dish area for adherent or per $5\text{-}10x10^6$ viable cells for suspension cells was used. 0.1ml of 1-bromo-3-chloropropane (BCP) was added per 1ml of TRI Reagent, shaked and left 15min at room temperature. The mixture was centrifuged at 12,000g for 15min at 4ºC and RNA was separated as colorless upper aqueous phase. The RNA was isolated from the aqueous phase using isopropanol (0.5ml per 1ml TRI Reagent), the vortexed sample was incubated for 10min at room temperature and than centrifuged using 12,000g for 10min at 4ºC. The RNA pellet was washed using 75% ethanol, air dryed 3-5min and dissolved in RNA Storage Solution. RNA was precipitated using 2.5X 96% ethanol and 0.1 volumes 3M NaAc, and stored at -20ºC.

5. 6. 2. RNA isolation from nuclear and cytoplasmic cell fractions

Normal human fibroblast cells (Hs27) were cultured under the standard conditions. Nucleus versus cytoplasmic fractionation was adapted from Sambrook and Russel, Molecular Cloning, 2001. The cells (4x T75 flasks) were washed three times with ice cold PBS on ice and cell suspensions were centrifuged using 2,000g for 5min at 4ºC. The cells were lysed using N/C Lysis buffer, underlayed with an equal volume of the N/C Lysis Buffer containing sucrose (24% w/v) and NP-40 (1%) and centrifuged using 10,000g for 20min at 4ºC. 2xProteinase K Buffer and proteinase K (200µg/ml) was added to the cytoplasmic fraction that was recovered as upper turbid layer and incubated at 37ºC for 30min. The nuclear fraction that was in the form of the pellet was sheared by squirting through a needle and further proteinase K treated as above. Proteins were removed by phenol/chloroform extraction and further RNA was extracted using Acid Phenol/chloroform. Nucleus and cytoplasm RNA fractions were recovered as upper, aqueous phases after centrifugation using 13,000g for 10min at 4ºC and were precipitated (by adding 2.5 volumes of 96% ethanol and 0.1 volumes of 3M NaAc pH 5.5). RNA was recovered by centrifugation, subject to a 75% ethanol wash and resuspended in RNA storage solution. RNA preparation from nuclear and cytoplasmic cell fractions was done two times, once by Federica Santoro, PhD student in the laboratory and once by myself.

Solutions:

<u>N/C Lysis Buffer</u>
0.14M NaCl
1.5mM MgCl2
10mM Tris HCl pH 8.6
0.5% NP-40
10mM Vanadyl-Ribonucleoside Complex

<u>Proteinase K Buffer</u>
0.2M Tris-HCl pH7.5
25mM EDTA pH8.0
0.3M NaCl
2% SDS

5. 7. DNaseI treatment

DNaseI treatments were done for all RNAs prior to Reverse transcriptase (RT) reactions using the DNA-free Kit, Ambion. Typically DNAseI treatment was done in 50µl or in 300µl reactions. For a routine 50µl reaction, 10µg of the RNA, 10X DNaseI Buffer and 1µl of DNaseI was incubated at 37° for 30min and a further 5µl of DNase Inactivation Reagent was added and mixed for 2min at room temperature. Reaction was centrifuged using 10,000g for 1.5min at room temperature. The supernatant containing RNA was precipitated and stored at -20ºC. In 300µl reactions rigorous DNaseI treatments were done, with 250µl of sample RNA, 30µl 10x Buffer, 10µl DNaseI and 10µl H_2O. These reactions were inactivated using 60µl of DNase Inactivation Reagent. Specifically, nuclear fractions of Hs27 were DNaseI treated two times (the first time in a 50µl reaction with 1µl of DNaseI for 30min at 37ºC and the second time 1µl of DNaseI was again added to the same reaction and incubated for another 30 min at 37ºC). DNaseI treated RNAs were precipitated and recovered prior to reverse transcription reactions.

5. 8. Expression analysis

5. 8. 1. Reverse transcription (RT) reaction

DNaseI treated RNAs were processed into cDNAs using the RevertAid First Strand cDNA Synthesis Kit, Fermentas. Reactions were performed according to the “Synthesis of First Strand cDNA Suitable for Second Strand Synthesis” protocol. Briefly, about 1µg of total RNA was mixed with 1µl of random hexamer primers (0.2µg/µl) and DEPC-treated H_2O, in 12µl reactions, and incubated at 70ºC for 5min. The following components were added to the reaction: 4µl of 5x reaction buffer, 1µl of 20 u/µl of RiboLock Ribonuclease Inhibitor, 2µl of 10mM dNTP mix. Reactions were incubated at 25ºC for 5min. 1µl of 200u/µl of RevertAid M-MuLV Reverse Transcriptase was added to +RT reactions while 1µl of DEPC-treated H_2O was added to control –RT reactions and incubated on 25ºC for 10min, 42ºC for 60min and finally reaction was stopped by heating on 70ºC for 10min. Reactions were placed on ice and stored in -20ºC prior to RT-PCRs, proofreading RT-PCRs or qRT-PCRs.

5. 8. 2. Non-quantitative RT-PCR

Reverse transcription polymerase chain reactions (RT-PCRs) were done similarly as PCR reactions (section 5.3.), with the exception that as a template, cDNAs were used. These reactions were done to test expression of *KLF14up3*, *ADAMTS7down*, *TMED3down*, *KIAA1024up* and *BLCAPov* macro ncRNAs in nuclear and cytoplasmic fractions of Hs27 cell line, where *GAPDH* expression was used as a loading control and expression of *KCNQ1OT1* macro ncRNA as a positive control (section 2.4.5.). Further, *AIRN* expression was tested by RT-PCRs in three Wilms' tumor and Hs27 cell lines and in 123 Wilms' tumor patients (Figure 53 and 54 in section 2.6.3.).

RT-PCR results were typically separated on 2% agarose gels in 1xTAE. An exception to this was for the 123 Wilms' tumor samples, where a 160bp band was expected using the pp9 primer pair. There the separation was done using 12% polyacrylamide gel electrophoresis (PAGE) in 1xTAE. The RT-PCR results were visualized by ethidium bromide staining.

5. 8. 3. Quantitative RT-PCR using SYBR and TaqMan Assays

Quantitative RT-PCRs (qRT-PCR) were done using SYBR® Assays and TaqMan® Assays. Primer pairs were designed using PrimerExpress. The SYBR® Assays were performed with 100mM primers (listed in the Table 26) and MESA Green qPCR MasterMix Plus for SYBR Green assays. The reactions were amplified on the ABI PRISM 7000 Sequence Detection System, Applied Biosystems with cycling conditions: 95ºC for 5min, 40 cycles of 95ºC for 15sec, 60ºC for 1min. The standard curve method of analysis using serial dilutions of cDNA was used as a basis for the quantification of RNA. Data was normalized to *RPLPO* gene expression.

qRT-PCR Assay	Primer Name	Sequence
DCNEx1	DCNEx1F	CCACAGGAGCCCTCAAAG
	DCNEx1R	CTACCCCCTCCTCCTTTCCA
DCNEx3	DCNEx2F	GTCGCGGTCATCAGGAACTT
	DCNEx2R	ACAGAGAGGCTTATTTGACTTTATGCT
DCNEx3	DCNEx3F	CAGGTTCTTAAAGTCTCCATCTTTGA
	DCNEx3R	GGATCTTCCCCCTGACACAA
DCNEx4	DCNEx4F	TCGCACTTTGGTGATCTCATT
	DCNEx4R	GCCAGAAAAAATGCCCAAAA
DCNIn1	DCNIn1F	TGTATCTGTTTCCCATTAAAAATGCA
	DCNIn1R	GCAATGACTTTGCTTCATTTTTCTT
DCNIn2	DCNIn2F	TCTTCATCTGTTACTGCATATAATCATCA
	DCNIn2R	TCTTAGAAATCCTATTAATCGTGTGAGGTA
DCNIn3	DCNIn3F	AAAACTCCTTCCTCGCATATTCTC
	DCNIn3R	GTAGTGAGTGTTATGGACTTAAAGTAAAAGAAA
DCNIn4	DCNIn4F	GAAAAAGACTATTAGTGAAAGCAATACCAA
	DCNIn4R	AAGATGGGAATTGTAAACTTGCTTTAG
GAPDH	GAPDHF	TGAAGGTCGGAGTCAACGG
	GAPDHR	ACCAGAGTTAAAAGCAGCCCT
H19q1	H19q1F	GTGTGACGGCGAGGACAGA
	H19q1R	TCCGTGGAGGAAGTAAAGAAACA
KCNQ1OT1q1	KCNQ1OT1q1F	ATTCCTCAAGTGTTGACCATTTTG
	KCNQ1OT1q1R	TGGTCCTGTGGGCTCCATT
KCNQ1OT1q2	KCNQ1OT1q2F	CTGCCTTCTCAGGTTATGGTCAT
	KCNQ1OT1q2R	GCTGGGCCTCCTTTGGA
ADAMTS7q3	ADAMTS7Cq3F	TGTTATGTGTGTGACTCCCTTGTG
	ADAMTS7Cq3R	GGGCCAGAGGGAAAAGCA
SLC38A4q1	SLC38A4Cq1F	AAGGCTCTAGGAGCTGTCAGATTAA
	SLC38A4Cq1R	CCACGCTCACCGAAGCTT
SLC38A4q3	SLC38A4Cq3F	TTTTCTATTCTCAGCCCCACTAAAG
	SLC38A4Cq3R	CTGAGATGAGCACTTGGATCCA
RPLPO1	RPLPO1F	CCACGCTGCTGAACATGCT
	RPLPO1R	TCGAACACCTGCTGGATGAC

Table 26. Primers used in the SYBR® Assays.

The TaqMan® Assays were performed using the primers and TaqMan probes listed in the Table 27. Primer pairs and TaqMan probes were designed using PrimerExpress. Primers were synthesized by VBC-Biotech Service GmbH, and TaqMan probes by Eurogentec (http://www.eurogentec.com/). The TaqMan qRT-PCR reaction was set with 900mM primers and 200nM probe using qPCR MasterMix Plus, Eurogentec. The reaction was amplified on a ABI PRISM 7000 Sequence Detection System using following cycle conditions: 50ºC for 2min, 40 cycles of 95ºC for 15sec, 60ºC for 1min. Quantification was done using the standard curve method. RNA expression was normalized to GAPDH or 18S rRNA gene expression. To evaluate GAPDH expression the same primer pair as for the SYBR® Assay was used, with a GAPDH TaqMan probe (Table 27). For assessment of expression of 18S rRNA, 18S rRNA Control kit (FAM-TAMRA), Eurogentec was used according to the manufacturers' recommendations.

qRT-PCR Assay	Primer/Probe Name	Sequence
AIRN QPCR1	H2354F	TCAGATGCAGGAAGATTGGGT
	H2354R	AGGCTTGGCATCCAGGTG
	H2354	CTCACAACAGGGCGGTGGTTGGA
AIRN QPCR2	P4HuSplF	GGCTCAGCCAAAAGGACACA
	P4HuSplR	TCGAACTGGGAGCCATGG
	P4HuSpl	AAGCCCTGCAGAGGCTCTGAAACCAA
IGF2R QPCR	P8HuF	CACGCAGGCCCAGGC
	P8HuR	TGGTATCAACAGCTTCCCATGT
	P8Hu	CCCGTTCCCCGAGCTGTGCA

Table 27. Primers and probes used in the TaqMan® Assays.

5. 8. 4. Long range RT-PCR/ Proofreading RT-PCR

Non-quantitative RT-PCRs and PCRs were done using the Long PCR enzyme Mix, Fermentas, when further allelic expression using Single Nucleotide Polymorphisms (SNPs) was tested. Long PCR enzyme Mix contains highly processive Taq DNA polymerase and second thermo stabile polymerase that exhibits 3' to 5' exonuclease activity. The proofreading activity was important to have greater protection against depurination and nicking and give good quality template for further sequencing through the genomic region containing SNPs. In the standard reaction: 5µl of 10x Long PCR Buffer with $MgCl_2$, 1µl of 10mM dNTPs, 2.5µl of forward and reverse primers (10pmol/µl), 1µl DMSO, 0.25µl of Long PCR Enzyme Mix (1.25u) and 2µl of template cDNA were used per 50µl reaction. The reaction was amplified in the Peltier Thermal Cycler PTC-200, under the cycling conditions: initial denaturation on 94ºC for 1.5min, 35 cycles of 94ºC for 15sec, 59ºC for 30sec, 68ºC for 45sec and final elongation at 68ºC for 7min. Reactions were separated on a 2% agarose gel and stained using the ethidium bromide solution. The bands of expected size were further cleaned as in section 5.3.1.

5. 8. 5. Allelic expression analysis using SNPs

In order to test gene expression from parental alleles, a combination of PCR/Proofreading RT-PCR and sequencing was done. In the first step, specific primers for the known and novel macro ncRNAs and the *ADAMTS7* protein coding gene were designed using Primer-BLAST on the genomic regions overlapping with known SNPs from the dbSNP build 129 database (Sherry et al., 2001) (primers are listed in Appendix Table II). In the second step, DNA from the cell lines (Hs27, HeLa) that were known to express genes of interest, were PCR amplified with specific primers; bands of expected size were cleaned (section 5.3.1.) and sent to the AGOWA, GmbH, for sequencing on the Applied Biosystems 3730xl DNA Analyzer. Results of the sequencing were further analyzed using Sequencher 4.7 (http://www.genecodes.com/). The heterozygous SNPs were visualized as two overlapping sequencing peaks at one base pair (bp) position. If a heterozygous SNP was observed for the tested genomic locus, the same primer pairs were used in the proofreading RT-PCR reaction (section 5.8.4) using cDNA as a template. Sequencing tracks were further visualized on Sequencher enabling analysis of expression at the genomic position where heterozygous SNP was previously detected in the DNA. If two peaks were overlapping on both Sequencher tracks gained from DNA and cDNA sequencing then biallelic expression was found. If one peak was present after cDNA sequencing on the position where two peaks were present in the DNA, monoallelic expression was observed. The same analysis was done in one family blood and in three lymphoblastoid cell lines that originated from families genotyped from the international HapMap project (http://hapmap.ncbi.nlm.nih.gov/).

5. 8. 6. Northern blotting

Northern blots were done on the Hs27, bone marrow, skeletal muscle and HeLa, cells and tissues using ADAMTS7downNOR and ß-Actin NOR probes (loading control). The probes were amplified using RT-PCR from Hs27 cDNA with specific primers (listed in Table 28). The probes were cloned into the pGEM T-Easy vector, confirmed by sequencing and excised from the plasmid by digestion prior to labeling with α^{32}P dATP (as previously described in the sections 5.4. and 5.5.1).

Name of the probe	Primer Name	Sequence	Length of the probe (bp)
ADAMTS7downNOR	ADAMTSNORF1	GCCGCTGATTCTCTTGTCTC	967
	ADAMTSNORR1	ACAGAGCAGCCCAGTGATCT	
ß-Actin NOR	ß actinNORF	CAGGCACCAGGGCGTGATGG	994
	ß actinNORR	GATGGAGGGGCCGGACTCGT	

Table 28. Primers for the Northern blot probes.

The Northern blotting was done under strict RNase-free conditions. The agarose gel was prepared using 2g of Agarose-LE with 20ml of NorthernMax 10x Denaturing Gel Buffer and 180ml of DEPC-treated H_20. 20µg of the RNA sample was dissolved in 6µl of RNA Storage Solution and 3 volumes of Formaldehyde loading dye and 1µl of ethidium bromide were added to the mixture. Millennium marker RNA was similarly prepared. The RNA and marker samples were denatured for 15min at 65ºC and than put on ice briefly until loading. The denaturing gel was loaded with samples and electrophoresed at 100V for 2h, in MOPS 1xRunning Buffer.

The northern blot was assembled similar as described for Southern blots in section 5.5.4. with the exception that as the blotting buffer: 50mM Na_2HP0_4 freshly treated with DEPC (1:1000) was used. After the capillary transfer of RNA to the Hybond XL membrane took place, the membrane was dried for 15min in the 55ºC oven. The RNA was cross linked to the membrane by AUTO cross linking in the UV Stratalinker 1800. The membrane was pre-hybridized with Church Buffer at 65ºC for 3h, hybridized at 65ºC for 18h with Church buffer containing the denatured labeled probe and washed using the Washing buffer (as described in section 5.5.4.). The membrane was exposed to the Phosphoimager screen and after 4 days scanned by the Typhoon Scanner 5600, Amersham.

5. 8. 7. Sample preparation for the HIRTA hybridization

The dscDNA samples used for the HIRTA hybridization, were prepared from the cell lines and tissues listed in Table 10 and Table 11 (section 2.1.2.) by first and second strand reverse transcription. The genomic DNA was sonicated and hybridized together with the dscDNA to HIRTA in order to enable normalization.

5. 8. 7. 1. First strand reverse transcription using SuperScript II Reverse Transcriptase

Typically 6µg of total DNaseI treated RNA was dissolved in 16µl of DEPC-treated H_2O and incubated together with 2µl of pd(N)6 Random Hexamer 5'Phosphate, Sodium Salt prepared to 2.5µg/µl, at 70ºC for 10min. The mixture was placed on ice for 5min. To the mixture: 7µl of 5X First Strand cDNA Buffer, 3.5µl of 0.1M DTT, 3µl 10mM dNTP, 1µl of 40u/µl RNase Inhibitor and 2µl of SuperScript II Reverse Transcriptase were added and filled with DEPC-treated H_2O up to a total volume of 35µl. The reaction was incubated at 42ºC for 1h and placed on ice prior to second strand synthesis.

5. 8. 7. 2. Second strand reverse transcription

20 µl of cDNA prepared using SuperScript II reverse Transcriptase was further mixed with: 30µl of 5X Second Strand Buffer, 10mM dNTP, 4µl DNA Polymerase I, 1µl E.coli DNA Ligase, 1µl RNase H and filled with DEPC-treated H_2O up to a 150µl total reaction volume. The mixture was incubated at 16ºC for 2h. Further 2µl of 5u/µl T4 DNA Polymerase was added and incubated at 16ºC for 10min and further on 70ºC for 10min to stop the reaction.

The dscDNA was cleaned using QIAquick PCR Purification Kit, Qiagen using the procedure recommended by manufacturer. Briefly, 600µl of PBI Buffer was added per reaction, sample was applied onto QIAquick column and centrifuged using

13,000rpm for 1min. Reactions were washed for two times using 700µl of PE Buffer and centrifuged using 13,000rpm for 1min. Each sample was eluted with two times 50µl of EB buffer and precipitated according to standard procedures.

5. 8. 7. 3. Preparation of sonicated genomic DNA

100µg of DNA isolated from Hs27 cells was first RNaseA treated (final concentration 25µg/ml) and than incubated at 37ºC overnight. Further, the DNA was Proteinase K treated (to final concentration of 200µg/ml) at 55ºC for 3h and phenol-chloroform extracted by adding the same volume of phenol-chlorophorm to the sample and centrifuging at 14,000rpm for 15min at 4ºC. Collected supernatants were chloroform extracted, centrifuged 14,000rpm for 10min on 4ºC and filled with MQ H2O up to 1ml. Samples were sonicated on ice. 18 sonication cycles at 20sec "on", 1min "off"; with the Power set on 40% and Cycle set at 90% were performed. Sonicated genomic DNA from the Hs27 cell line was cleaned according to standard procedures using QIAquick PCR Purification Kit, Qiagen, with 10µg of sonicated DNA per QIAquick column.

5. 8. 7. 4. Quality control using Agilent DNA 7500 Bioanalyzer

The quality of dscDNA and sonicated genomic DNA were tested using the Agilent DNA 7500 Kit, Agilent Technologies. Agilent DNA Chip 7500 was prepared according to the manufacturers recommendations and 1µl of the sample was loaded onto the Chip. The chip was run in the Agilent 2100 bioanalyzer and analyzed using 2100 Expert Software for the DNA 7500 Assay. Sonicated genomic Hs27 DNA in a range from 100 to 800bp was used for the HIRTA hybridization.

5. 8. 7. 5. Hybridization to the HIRTA chip

The dscDNA samples and sonicated genomic DNA samples dissolved in nuclease free water with concentrations of 100 to 500ng/µl and OD260/280 ratios>= 1.7 (measured using NanoDrop), that showed no signs of degradation through visual inspection of Agilent Bioanalyzer gel results and the 2% agarose gel electrophoresis of the sample, were send to ImaGenes (http://www.imagenes-bio.de). The dscDNA samples were labeled using Cy5 and genomic DNA using Cy3 fluorochromes. Samples were hybridized in Cy5/Cy3 pairs to HIRTA (as presented on Figure 11, section 2.1.5.). The HIRTA Chip was further washed and scanned by ImaGenes.

5. 8. 7. 6. Visualization of the HIRTA hybridization data

The raw data from HIRTA hybridizations were normalized by Tuckey bi-weight normalization and prepared as .gff files by Ido Tamir, GENAU consortium. The data were further transformed into .wig files and visualized on the UCSC genome browser.

5. 8. 7. 7. Bioinformatics analysis of the HIRTA reproducibility

Assessment of HIRTA reproducibility was performed by comparison of HIRTA hybridization replicates, visualized by scatter plots of two numerical columns. Each column was a normalized biological or technical replicate HIRTA hybridization data. Pearsons' correlation was the statistical test used to show correlation coefficients between the columns. Display of the data and statistics were done using Galaxy (http://main.g2.bx.psu.edu/).

5. 8. 8. Sample preparation for the rRNA depleted total RNA Sequencing

RNA was isolated from Hs27 fibroblasts using TRI Reagent. Quality of RNA was assessed using the Agilent RNA 6000 Nano Kit, Agilent Technologies, according to the manufacturers' instructions. The Hs27 RNA sample had a RNA Integration Number (RIN)>9. The Hs27 RNA was DNaseI treated in a routine 50µl reaction and RNA quality was again assessed on the Agilent 2100 Bioanalyzer, showing RIN>9. DNaseI treated Hs27 RNA was than further depleted of ribosomal RNA.

5. 8. 8. 1. Ribosomal RNA depletion

Ribosomal RNA was depleted from the high quality Hs27 RNA sample using RiboMinus Transcriptome Isolation kit (Human/Mouse), Invitrogen according to the manufacturers' recommendations. Briefly, 10µg of RNA was hybridized with the RiboMinus Human/Mouse Probe, RNA was denatured by heating at 70ºC for 5min and further cooled down to 37ºC during 30min in the water bath. The prepared RiboMinus Magnetic Beads were mixed with the RNA/RiboMinus probe sample and incubated at 37ºC for 15min. The supernatant containing RiboMinus RNA fraction was separated using a magnetic stand. The RiboMinus RNA fraction was concentrated using RiboMinus Concentration module, Invitrogen by a procedure recommended by manufacturer. Depletion efficiency was assessed using Agilent RNA 6000 Pico kit and Eukaryote Total RNA Pico Series II program on the 2100 Agilent bioanalyzer.

5. 8. 8. 2. Hydrolysis of rRNA depleted total RNA

100ng of the rRNA depleted Hs27 total RNA in nuclease free H_20 was hydrolyzed using 5xHydrolysis buffer (200mM Tris pH8.2, 500mM KAc and 150mM $MgAc_2$). The reaction mixture was heated at 94ºC for 3, 3.5 and 4min. The reaction was cleaned using RNasy MinElute Kit, Qiagen according to the manufacturers' protocol. Cleaned RNA was diluted from the RNase Spin column in two times 50µl of nuclease free H_20, precipitated according to standard procedures and further recovered in 10µl of nuclease free H_20. The hydrolyses of rRNA depleted total Hs27 RNA was assessed using Agilent RNA 6000 Pico kit and the Eukaryote Total RNA Pico Series II program on the 2100 Agilent bioanalyzer. RNA that was hydrolyzed at 94ºC for 4min had the peak between 200 and 500bp and was further used for cDNA preparation.

5. 8. 8. 3. First strand cDNA preparation for RNA Sequencing

100ng of the hydrolyzed, rRNA depleted Hs27 sample was used in a 14µl reaction where 1.67µl of the pd(N)6 Random Hexamers 5'-Phosphate, Sodium Salt, dissolved to 3µg/µl and 1µl of 10mM dNTPs were added. The mixture was heated for 5min at 65ºC. In the second step, 5µl of 5xFirst Strand Buffer, 2.5µl of 0.1M DTT, 1µl of 40u/µl RNaseOUT and 1.5µl of nuclease free H20 were added and the mixture was heated for 2min on 25ºC. Finally, 1µl (200u) of SuperScript II Reverse Transcriptase was added and mixture was heated for 10min at 25ºC and further for 1h at 42ºC, to synthesize the Hs27 cDNA.

5. 8. 8. 4. Second strand cDNA preparation for RNA Sequencing

To the first strand cDNA preparation, 90.5µl of nuclease free H20 was added together with: 1µl of DNA Polymerase I, 0.5µl of RNase H, 30µl of the 5xSecond strand buffer and 3µl of 10mM dNTPs, and than the mixture was heated for 2h at 16ºC. 1µl of T4 DNA Polymerase was further added to the mixture. The mixture was heated for 10min at 16ºC and further for 10min at 70ºC. The dscDNA prepared from the hydrolyzed rRNA depleted Hs27 total RNA was cleaned using MinElute Reaction Cleanup kit, Qiagen according to the manufacturers' protocol. The dscDNA was eluted from the MinElute column using two times 10µl of EB buffer.

The concentration of the Hs27 dscDNA was examined using PicoGreen® Assay for dscDNA. The assay was performed according to the manufacturers' protocol and analyzed on a NanoDrop ND-3300 Fluorospectrometer.

5. 8. 8. 5. Library preparation and Illumina/Solexa sequencing

Library preparation and Illumina/Solexa RNA Sequencing were performed by Andreas Sommer, GENAU consortium. Library preparation was done using ChIP-Seq DNA Sample Prep Kit, Illumina according to the manufacturers' protocol. The library was loaded on four lines of the flowcell and RNA Sequencing was performed by the Illumina Genome Analyzer II. Image analysis and base calling were performed using Illumina Pipeline and 36bp single-reads were obtained.

5. 8. 8. 6. Bioinformatics analysis of the RNA Sequencing data

rRNA depleted total RNA Sequencing 36bp single-reads were aligned to the NCBI36/hg18 genome build using ELAND (http://bioit.dbi.udel.edu/howto/eland), Bowtie (http://bowtie-bio.sourceforge.net/) and TopHat (http://tophat.cbcb.umd.edu/) allowing two mismatches. Alignments were performed by Andreas Sommer and Ido Tamir from the GENAU Bioinformatics team.

5. 9. Bioinformatics analysis of macro ncRNA potential

The RNAcode, program that predicts regions with protein coding potential based on evolutionary signatures (Washietl et al., 2010, unpublished data) was performed over 44 vertebrate species using Multiz Align from UCSC, by Jan Engelhart, a visiting MSc student from Peter Stadlers' laboratory, Institut für Informatik, Leipzig. RNAcode predictions were in the form of a .wig track that was loaded into UCSC browser and compared with visually mapped macro ncRNAs expressed by HIRTA.

5. 10. Bioinformatics analysis of direct repeats

Presence of direct repeats in the CpG island promoters of macro ncRNAs was assessed using Dotmatcher program (http://emboss.bioinformatics.nl/cgi-bin/emboss/dotmatcher). All CpG islands were tested using the same criteria: window size=30 and Threshold=65 and presented in the form of dotplots where dots from regions of similarity align to form diagonal lines and repeats are visualized as parallel diagonal lines.

5. 11. Appendix table I- Materials

Chemicals, enzymes and other materials	Company
$\alpha\ ^{32}$PdATP	PerkinElmer
ß-merkaptoethanol	Sigma
Acid Phenol Chlorophorm	Ambion
Agar	AppliChem
Agarose	Biozym
Agarose-LE	Ambion
Ampicilin	Roche
BCP	MRC
Betaine	Fermentas
Boric Acid	AppliChem
Bovine Serum Albumine (BSA)	Sigma
BstXI	Fermentas
BstUI	Fermentas
BssHII	Fermentas
Chlorophorm	Merck
Diethyl pyrocarbonate (DEPC)	Sigma
Dimethyl sulfoxide (DMSO)	Sigma
Dithiothreitol (DTT)	Invitrogen
DNA polymerase I	Invitrogen
dCTP	Bioron
dGTP	Bioron
dNTP mix	Fermentas
dTTP	Bioron
Dulbecco’s Modified Eagle Medium (DMEM)	Gibco
E.coli DNA Ligase	Invitrogen
EcoRI	Fermentas
EglI	Fermentas
Ethanol	Merck
Ethidium Bromide	Merck
Ethylendiaminetetraacetic acid (EDTA)	Merck
Fetal bovine serum (FBS)	Gibco
First strand cDNA buffer, 5x	Invitrogen
Formaldehyde loading dye	Ambion
GeneRuler 100bp Plus DNA Ladder	Fermentas
Gentamicine	Gibco
Glacial Acetic Acid	VWR
Glucose	Gibco
GoTaq DNA Polymerase	Promega
HCl	Merck
Hepes	Roth
Hind III	Fermentas
Hybond XL	Amersham
Isopropanol	Merck
Isopropyl-ß-D-thiogalactopyraniside (IPTG)	AppliChem
KAc	Sigma
Klenow Fragment	Fermentas
L-Glutamin	Gibco
LB Medium	Bio101
McCoys’ media	Gibco
Minimum Essential Medium Eagle (MEM)	Gibco
Mesa Green qPCR MasterMix Plus	Eurogentec
MgAc$_2$	Sigma
MgCl$_2$	Sigma
MgCl$_2$ (25mM)	Fermentas
NaAc, 3M, pH5.5	Ambion
NaCl	Neo Lab
Na$_2$HPO$_4$	Sigma
NaOH	AppliChem
NorthernMax^{TM} 10x Denaturing gel buffer	Ambion
NorthernMax^{TM} 10x MOPS gel running buffer	Ambion
NotI	Fermentas
NP-40	Calbiochem
Nuclease free wather	Ambion
pd(N)$_6$ Random Hexamer 5’ Phosphate, Sodium salt	GE Healthcare
Phosphate buffered saline (PBS)	Qbiogene
Proteinase K	Qbiogene

qPCR MasterMix Plus	Eurogentec
Random hexamer primers (for RT-PCR)	Invitrogen
Random hexamer primers (for radioactive probe labeling)	Pharmacia
RiboLock™ RNase Inhibitor	Fermentas
Ribonucleoside Vanadyl Complex	New England Biolabs
RNA Millenium Marker	Ambion
RNA Storage Solution	Ambion
Rotiphorese Gel 30	Roth
RNaseA	Fermentas
RNaseH	Invitrogen
RNaseOUT	Invitrogen
RPMI 1640 Media	Gibco
Second strand cDNA Buffer, 5x	Invitrogen
Sephadex™ G-50	Amersham
Sodium Dodecyl Sulfate (SDS)	AppliChem
Sodium Pyruvate	Sigma
SuperScript II Reverse Transcriptase	Invitrogen
TEMED	Roth
T4 DNA Ligase	Fermentas
T4 DNA Polymerase	Invitrogen
TRI Reagent	Sigma
Tris	AppliChem
Tris-Cl	QBiogene
Trypsin-EDTA	Gibco
X-Gal	Roth

Kits	Company
18S rRNA Control Kit (FAM-TAMRA)	Eurogentec
Agilent DNA 7500 Kit	Agilent Technologies
Agilent RNA 6000 Nano Kit	Agilent Technologies
Agilent RNA 6000 Pico Kit	Agilent Technologies
DNA-free Kit	Ambion
ChIP-Seq DNA Sample Preparation Kit	Illumina
Long PCR enzyme mix	Fermentas
MinElute Reaction Cleanup Kit	Qiagen
pGEM T-Easy Vector System I	Promega
PicoGreen Assay	NanoDrop technologies
RevertAid First Strand cDNA Synthesis Kit	Fermentas
RiboMinus Concentration Module	Invitrogen
RiboMinus Transcriptome Isolation Kit (Human/Mouse)	Invitrogen
RNasy MinElute Kit	Qiagen
QIAquick PCR Purification Kit	Qiagen
Wizard Genomic DNA Purification Kit	Promega
Wizard SV Gel and PCR Clean-Up System	Promega

5. 12. Appendix table II- Primers for testing allelic expression

Candidate ncRNA	F_Primer_Name	F_Primer_Sequence	R_Primer _Name	R_Primer_Sequence	Product Length (bp)	Product position (hg18)
LRRC47down	LRRC47CIE1F	CACCCAGATCGTAAGGCAGT	LRRC47CIE1R	CATATCATGCCAGGATGCAG	634	chr1 3685834 3686467
	LRRC47CIE2F	ATCTGTCTGGCTGCATCTGG	LRRC47CIE2R	GTTCGTGAGTCTCTGGGAGG	341	chr1 3684319 3684659
KLF14up3	KLF14CIE1F	GACGTGCGCAGACACAGGCT	KLF14CIE1R	TCATGGGGCCCGAGCTCTCC	743	chr7 130232070 130232812
	KLF14CIE2F	GGTAGAGCACCTCCGGGCCA	KLF14CIE2R	GCCTCACTCCTGGGCAGCCT	473	chr7 130239881 130240353
	KLF14CIE3F	CGCCTGGGTCCCGACCTGTA	KLF14CIE3R	CAGTGGGCCCCCAGCAAAGG	623	chr7 130262960 130263582
KIAA11264up (SLC45A4down)	SLC45A4CIE1F	ACGCCTGCTTTTGGGGTCCT	SLC45A4CIE1R	ATGCGTGCTGCAGCTTTGGC	849	chr8 142384577 142385425
	SLC45A4CIE2F	TGCCACTTGCCGCCTGGTAA	SLC45A4CIE2R	ACAGTGAACTGCTGGCGTGC	429	chr8 142371976 142372404
	SLC45A4CIE3F	AAACCAGCCAGGTGCGCGAT	SLC45A4CIE3R	AGTTGGCTCTTGGTGGGCTCCT	787	chr8 142368363 142369148
	SLC45A4CIE4F	TGGCCAGCTGCTCTGAACGA	SLC45A4CIE4R	TGCAGCTTTGCCGCAGGTCT	641	chr8 142361627 142362267
	SLC45A4CIE5F	GCACAGCACTGCACTCACAGGT	SLC45A4CIE5R	CGGCGTGTGTGGCTTTGCAT	529	chr8 142355919 142356447
PEG13	P13CIE1F	TCAGCGCGCAGCTTCAGCAT	P13CIE1R	TTAGCGCTGACGCCTCCGAA	931	chr8 141179240 141180170
	P13CIE2F	AGCTGCAGAATTGCCGCCGT	P13CIE2R	GCTGAAGCCGCGCTTGAAGA	534	chr8 141177184 141177717
	P13CIE3F	GCACGGGGCAGGGCAAAGAT	P13CIE3R	GGCACAGCTGCGACTGCGTA	764	chr8 141176319 141177082
	P13CIE4F	TAAACACGCCCACCGGGGTT	P13CIE4R	ATGCTGAAGCCGCGCACTGA	726	chr8 141178345 141179070
	P13CIE5F	GGGAGGGCGCACATTCCACC	P13CIE5R	GCTGCAGGAGGAGCTGCGAG	908	chr8 141175322 141176229
	P13CIE6F	CGGTGCCCTGGCAAGCAAGA	P13CIE6R	TCCCCTTCCTGCAGCCTCCC	827	chr8 141173975

						141174801
PRCDBPup	PRKCDBPCIE1F	TATCCCATTACCCAGGATGC	PKCDBPCIE1R	GTAGAGACAGAGGGCCACCC	395	chr11 6303553 6303947
	PRKCDBPCIE2F	AGAGTATGTTCAGTGCGGAGC	PKCDBPCIE2R	CCAATTGACAGTGGATTTTTGA	330	chr11 6314515 6314844
SLC38A4down2	SLC38CIE1F	GGCATGGCGTAGGTGTAGAT	SLC38CIE1R	CAACAGGTGGAGCCAGGTAT	708	chr12 45123429 45124137
	SLC38CIE2F	ATCATTCCCTTGATGATGCC	SLC38CIE2R	CCCCAAAAATGACCATTCAC	714	chr12 45306921 45307634
	SLC38CIE3F	TGGGAAGAGAAATTGAACGG	SLC38CIE3R	TTTCGTTTGTGCTGCACTTC	678	chr12 45141873 45142551
	SLC38AIE4F	TACACCATTGCCAATTCCAG	SLC38AIE4R	ACATCGACGTTCTCCCTTGA	255	chr12 45292045 45292299
	SLC38AIE5F	ATGGGGTACCAAAAAGGGAC	SLC38AIE5R	TTGCTAAACCCCAAAAGAACA	424	chr12 45198803 45199226
	SLC38AIE6F	ACAGAAGGCCAGTGTTCAGG	SLC38AIE6R	AAGTCCCCCAGGGTTGTAAG	386	chr12 45339237 45339622
	SLC38AIE7F	GGATTTTCCCCCTCGTTTTA	SLC38AIE7R	CCTGGAGGATTCACCATTTG	258	chr12 45065055 45065312
	SLC38AIE8F	TTCAAATTGGAGGGAGCTTG	SLC38AIE8R	GGCTAAAGGCATAGCTGACG	312	chr12 45067607 45067918
	SLC38AIE9F	GATGTGGCCTAAATTGGGAA	SLC38AIE9R	AGCTGGGGTTCAAGGTTCTC	230	chr12 45072960 45073187
	SLC38AIE10F	TTACCCTGGCAACACCATTT	SLC38AIE10R	AGTGATGGGGACACAACCAT	433	chr12 45075519 45075951
	SLC38AIE11F	TTCTGGGCTAATCCTCTCCA	SLC38AIE11R	TCCAAAAGGGAACAAAGCAG	196	chr12 45081679 45081874
	SLC38AIE12F	ATGGGCTTGTTTGGAGACTG	SLC38AIE12R	ATGGTGTGGGGACTGTGTTT	491	chr12 45082671 45083161
	SLC38AIE13F	TTCCCCCAGCACCTCCACCC	SLC38AIE13R	GGGTGGGGGCACGCATCTTC	678	chr12 45086582 45087259
	SLC38AIE14F	TCCCTGCATCCTTCCTGGCT	SLC38AIE14R	TGTCGGAATGTGGCACTTGGCAG	850	chr12 45134077 45134926
	SLC38AIE15F	ACAGCAGGTTTGCTGGCCCC	SLC38AIE15R	AGCCAGGACTGAGGGATGGCA	931	chr12 45253825 45254755
	SLC38ACIE16F	TGTGCAATTCCCCACCCTGTGC	SLC38ACIE16R	ACCAGCTCAGGCTCCATGTTCCT	839	chr12 45091399 45092237
	SLC38ACIE17F	GCCCGGGGAATTGCCTCTCC	SLC38ACIE17R	ACAGCACCATTGCTCCTGCGG	593	chr12 45153862 45154454
	SLC38ACIE18F	AGGCAGGGTGTGGCTGGAGA	SLC38ACIE18R	CCATCACAGGGCACCGCAGG	864	chr12 45156579 45157442
	SLC38ACIE19F	TCCCTTCCTTCCCTGCCTTGCT	SLC38ACIE19R	GGGGGTGAGAAGGGTGGTCCT	509	chr12 45172401 45172906
	SLC38ACIE20F	CAGGAGGGCCTGTGGAGCCT	SLC38ACIE20R	CTGCCAGCAGCAGGGTGGTC	975	chr12 45203656 45204630
	SLC38ACIE21F	TGGCAGGGTGGGAGGTGGAA	SLC38ACIE21R	TTGCCCTTTCTGGCTGGCCT	876	chr12 45226838 45227713
	SLC38ACIE22F	ACCCTCCAGGCACTCCCACC	SLC38ACIE22R	TGCCGGTCCCTCTCTGCTGG	991	chr12 45332357 45333347
	SLC38A4CIE23F	TGAGCAAAGGTTGCGGGCGT	SLC38A4CIE23R	GCCACCAGGGTCCACAGGTC	900	chr12 45216764 45217663
ADAMTS7down	ADAMTS7CIE1F	CCCTGCAGCTTGCTTTAGAA	ADAMTS7CIE1R	ATTTGGGAGACGATGCTCAG	325	chr15 76830664 76830988
	ADAMTS7CIE2F	CACCAGCTGCCCTTAGGTT	ADAMTS7CIE2R	GAGGGAGGGGAAGTGTTAGC	419	chr15 76834538 76834957
	ADAMTS7IE3F	TTGGGCGTTCTCTGTTCTCT	ADAMTS7IE3R	GACGTCTGTGTCCCAGGATT	723	chr15 76831349 76832071
	ADAMTS7IE4F	CAGTGCTCCTGGTGTCTCCT	ADAMTS7IE4R	CTAAAGACACGAAGCCGGAG	549	chr15 76833701 76834249
	ADAMTS7IE5F	CTGACCATGGGACACCTTCT	ADAMTS7IE5R	GCGTGTGGATTTCTCAGGTT	342	chr15 76830990 76831347
	ADAMTS7IE6F	TCGTGGAAATCATTCACCAA	ADAMTS7IE6R	AGCTGTCCAAGACCGTTCAC	175	chr15 76833499 76833673
	ADAMTS7IE7F	CTCCGGCTTCGTGTCTTTAG	ADAMTS7IE7R	AAAACCTAAGGGCAGCTGGT	329	chr15 76834230 76834558
	ADAMTS7IE8F	CACCAGCTGCCCTTAGGTT	ADAMTS7IE8R	CTGAGGCTGGCACAGATACA	859	chr15 76834538 76835396
	ADAMTS7IE9F	GGAGACGGTTCTGGTTTCAA	ADAMTS7IE9R	TCTGGGAATCCTTGATGGAG	885	chr15 76835542 76836425
	ADAMTS7IE10F	AGGACTGCCTGGGCCTGTGT	ADAMTS7IE10R	TCAGCGGCGCTCCAGAGAGT	464	chr15 76832016 76832479
	ADAMTS7CIE11F	CCTGCTGTGCCTGTGAGGGC	ADAMTS7CIE11R	AGGCAGCAACCCTCTGGCCT	743	chr15 76832429 76833171
TMED3down	TMED3CIE1F	CGCAGCCAGAACCCTCAGCC	TMED3CIE1R	TCGGCCTCCTGGGATTGGCA	327	chr15 77433842 77434168
	TMED3CIE2F	CCCAGGCTGAGAGGGCAGGT	TMED3CIE2R	AGCCTGGGGAAGGCCAGAGG	992	chr15 77460995 77461986
	TMED3CIE3F	AGTGGCTGTCAAATGCAGTG	TMED3CIE3R	CCGTGATCCTGGTCTTGAAT	443	chr15 77413645 77414087
	TMED3CIE4F	TCGGCTTCTCTGTGAGCATA	TMED3CIE4R	TCACAGGGTTGTGAGACAGC	249	chr15 77437400 77437648
	TMED3CIE5F	AAGCTAAGCTCGTGGTGGCAGC	TMED3CIE5R	AGCAGCTGGGGTCAGGAGAA	382	chr15 77446103 77446484
	TMED3CIE6F	ATCCTGAGTCCCCCAGCCCC	TMED3CIE6F	AGGCCTCTGCACTGGGGACA	345	chr15 77397874 77398218
	TMED3CIE7F	TGCCTGCTTGCAGTGGGACA	TMED3CIE7R	GGGCCTCAGCCACCAGGAGT	696	chr15 77412460 77413155
KIAA1024up	KIA1024CIE1F	GCAAACGCACACAGCAGGGC	KIA1024CIE1R	TGCCCCCAGCTTTTGCCACC	979	chr15 77497123 77498101
	KIA1024CIE2F	GGGCTCTGTCCCTGGGGAGG	KIA1024CIE2R	AGGCCAGGGCCAAGTGGTCA	518	chr15 77510515 77511032
	KIA1024CIE3F	TGTGCTGGGGTTTCCTGCCT	KIA1024CIE3R	AGGGTGCTGGTTTGTACCCGT	725	chr15 77493565 77494289
	KIA1024CIE4F	AGGGTGCATTCAGGCCAGGT	KIA1024CIE4R	ACCAGCTGCCAATTCAGGGCT	657	chr15 77495061 77495717
	KIA1024CIE5F	ACAGCTGCCAGGCGTCACAT	KIA1024CIE5R	TTGGCCCTGCTGTGTGCGTT	731	chr15 77496415 77497145
	KIAA10246F	GGCTCCCTGGGACCAGCCTT	KIAA10246R	GCAGCCCAGTGCTCACAAAGGG	180	chr15 77506036 77506215
BLCAPov	BLCAPCIE1F	GGCCTTAGCACCTGCCTGCC	BLCAPCIE1R	CAGCGCTTACGCCCTGCCTT	836	chr20 35575145 35575980
	BLCAPCIE2F	CAGGCAGCCCTCCACCTCCT	BLCAPCIE2R	TTGGGCCTCCCGATCCCTGG	960	chr20 35572583 35573552
	BLCAPCIE3F	ACTGACAATGCAGGCCCCCT	BLCAPCIE3R	TCCTGTGGCGACACCTGGATGA	819	chr20 35577256 35578074
	BLCAPCIE4F	TCAGCCAGCTCCTGCCAAGT	BLCAPCIE4R	AGGAAGGCCAGGAACCACCCTT	777	chr20 35570967 35571743
	BLCAPCIE5F	AAAGGGCCTGCCGAGCATCT	BLCAPCIE5R	AGCATGGTGGCCCTGCTGAT	957	chr20 35563865 35564821
	BLCAPCIE6F	ACTGGGTCATGGGAGCCCTCT	BLCAPCIE6R	TTTGGCCACCCACTGCCACC	549	chr20 35558111 35558659
	BLCAPCIE7F	TCCTGCAAGCTCACCTGCCT	BLCAPCIE7R	TGGGCCCAATCCCTTGGCTT	576	chr20 35574201 35574776

	BLCAPCIE8F	TTGGGTGGGTGGGTCGTTGC	BLCAPCIE8R	AGGAGGTGGAGGGCTGCCTG	391	chr20 35572212 35572602
	BLCAPCIE9F	TTGTTGCAGTAGCCGGGCTTT	BLCAPCIE9R	ACCACTGCCACTCCATTCTGCT	696	chr20 35554657 35555352
	BLCAPCIE10F	TGGGCAGCACCAGCATTGGA	BLCAPCIE10R	ACATTTCCAGCCCTCCAGCCCT	786	chr20 35547819 35548604
	BLCAPCIE11F	GCTGCTGCCACAGCGAGGAT	BLCAPCIE11R	CTCCAAGGAGGGGGCAGGCA	766	chr20 35574415 35575180
Protein coding gene	**F_Primer_Name**	**F_Primer_Sequence**	**R_Primer _Name**	**R_Primer_Sequence**	**Product Length (bp)**	**Product position (hg18)**
ADAMTS7	ADAMTS7GIE1F	GACCGTCCCCACTGCACAGC	ADAMTS7GIE1R	TGCAGACACCTGCCACCCCT	984	chr15 76844774 76845757
	ADAMTS7GIE2F	GGCTGTGCCTGCCCCACTTC	ADAMTS7GIE2R	CACATACGCACGCAGGGGCA	520	chr15 76850535 76851054
	ADAMTS7GIE3F	AGGGTCCTGCACCTCGCCAA	ADAMTS7GIE3R	CTCTCCCTGCAGGACGTGCAAC	323	chr15 76879634 76879956
	ADAMTS7GIE4F	CCTTGCTCAGGGTTCCGCCG	ADAMTS7GIE4R	GGCCAGTGAGCTTGCAGGGG	847	chr15 76867241 76868114

5. 13. Abbreviations

bp	basepairs
cpm	counts per minute
g	gram
h	hour
kb	kilobasepairs
rpm	rounds per minute
l	litre
mg	milligram
min	minute
ml	mililitre
µg	microgram
µl	microlitre
MQ	milliq
ng	nanogram
u	units
sec	seconds
V	volts
w/v	weight/volume percentage

6. References

(2005). A haplotype map of the human genome. Nature *437*, 1299-1320.

Abdollahi, A., Pisarcik, D., Roberts, D., Weinstein, J., Cairns, P., and Hamilton, T.C. (2003). LOT1 (PLAGL1/ZAC1), the candidate tumor suppressor gene at chromosome 6q24-25, is epigenetically regulated in cancer. J Biol Chem *278*, 6041-6049.

Albrecht, U., Sutcliffe, J.S., Cattanach, B.M., Beechey, C.V., Armstrong, D., Eichele, G., and Beaudet, A.L. (1997). Imprinted expression of the murine Angelman syndrome gene, Ube3a, in hippocampal and Purkinje neurons. Nat Genet *17*, 75-78.

Alders, M., Ryan, A., Hodges, M., Bliek, J., Feinberg, A.P., Privitera, O., Westerveld, A., Little, P.F., and Mannens, M. (2000). Disruption of a novel imprinted zinc-finger gene, ZNF215, in Beckwith-Wiedemann syndrome. Am J Hum Genet *66*, 1473-1484.

Andres, M.E., Burger, C., Peral-Rubio, M.J., Battaglioli, E., Anderson, M.E., Grimes, J., Dallman, J., Ballas, N., and Mandel, G. (1999). CoREST: a functional corepressor required for regulation of neural-specific gene expression. Proc Natl Acad Sci U S A *96*, 9873-9878.

Antequera, F., and Bird, A. (1993). Number of CpG islands and genes in human and mouse. Proc Natl Acad Sci U S A *90*, 11995-11999.

Aravin, A., Gaidatzis, D., Pfeffer, S., Lagos-Quintana, M., Landgraf, P., Iovino, N., Morris, P., Brownstein, M.J., Kuramochi-Miyagawa, S., Nakano, T.*, et al.* (2006). A novel class of small RNAs bind to MILI protein in mouse testes. Nature *442*, 203-207.

Aravin, A.A., and Bourc'his, D. (2008). Small RNA guides for de novo DNA methylation in mammalian germ cells. Genes Dev *22*, 970-975.

Arima, T., Drewell, R.A., Arney, K.L., Inoue, J., Makita, Y., Hata, A., Oshimura, M., Wake, N., and Surani, M.A. (2001). A conserved imprinting control region at the HYMAI/ZAC domain is implicated in transient neonatal diabetes mellitus. Hum Mol Genet *10*, 1475-1483.

Arima, T., and Wake, N. (2006). Establishment of the primary imprint of the HYMAI/PLAGL1 imprint control region during oogenesis. Cytogenet Genome Res *113*, 247-252.

Arnaud, P., and Feil, R. (2005). Epigenetic deregulation of genomic imprinting in human disorders and following assisted reproduction. Birth Defects Res C Embryo Today *75*, 81-97.

Arnaud, P., Monk, D., Hitchins, M., Gordon, E., Dean, W., Beechey, C.V., Peters, J., Craigen, W., Preece, M., Stanier, P.*, et al.* (2003). Conserved methylation imprints in the human and mouse GRB10 genes with divergent allelic expression suggests differential reading of the same mark. Hum Mol Genet *12*, 1005-1019.

Babak, T., Deveale, B., Armour, C., Raymond, C., Cleary, M.A., van der Kooy, D., Johnson, J.M., and Lim, L.P. (2008). Global survey of genomic imprinting by transcriptome sequencing. Curr Biol *18*, 1735-1741.

Bachellerie, J.P., Cavaille, J., and Huttenhofer, A. (2002). The expanding snoRNA world. Biochimie *84*, 775-790.

Badenhop, R.F., Cherian, S., Lord, R.S., Baysal, B.E., Taschner, P.E., and Schofield, P.R. (2001). Novel mutations in the SDHD gene in pedigrees with familial carotid body paraganglioma and sensorineural hearing loss. Genes Chromosomes Cancer *31*, 255-263.

Bagga, P.S., Ford, L.P., Chen, F., and Wilusz, J. (1995). The G-rich auxiliary downstream element has distinct sequence and position requirements and mediates efficient 3' end pre-mRNA processing through a trans-acting factor. Nucleic Acids Res *23*, 1625-1631.

Bajaj, V., Markandaya, M., Krishna, L., and Kumar, A. (2004). Paternal imprinting of the SLC22A1LS gene located in the human chromosome segment 11p15.5. BMC Genet *5*, 13.

Barlow, D., Bartolomei, M. (2007). Genomic Imprinting in Mammals (Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press).

Barlow, D.P. (1993). Methylation and imprinting: from host defense to gene regulation? Science *260*, 309-310.

Barlow, D.P., Stoger, R., Herrmann, B.G., Saito, K., and Schweifer, N. (1991). The mouse insulin-like growth factor type-2 receptor is imprinted and closely linked to the Tme locus. Nature *349*, 84-87.

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823-837.

Bartolomei, M.S., Zemel, S., and Tilghman, S.M. (1991). Parental imprinting of the mouse H19 gene. Nature *351*, 153-155.

Beatty, L., Weksberg, R., and Sadowski, P.D. (2006). Detailed analysis of the methylation patterns of the KvDMR1 imprinting control region of human chromosome 11. Genomics *87*, 46-56.

Bell, A.C., and Felsenfeld, G. (2000). Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. Nature *405*, 482-485.

Berger, S.L., Kouzarides, T., Shiekhattar, R., and Shilatifard, A. (2009). An operational definition of epigenetics. Genes Dev *23*, 781-783.

Bernstein, B.E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D.K., Huebert, D.J., McMahon, S., Karlsson, E.K., Kulbokas, E.J., 3rd, Gingeras, T.R.*, et al.* (2005). Genomic maps and comparative analysis of histone modifications in human and mouse. Cell *120*, 169-181.

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

Berretta, J., and Morillon, A. (2009). Pervasive transcription constitutes a new level of eukaryotic genome regulation. EMBO Rep *10*, 973-982.

Berteaux, N., Aptel, N., Cathala, G., Genton, C., Coll, J., Daccache, A., Spruyt, N., Hondermarck, H., Dugimont, T., Curgy, J.J.*, et al.* (2008). A novel H19 antisense RNA overexpressed in breast cancer contributes to paternal IGF2 expression. Mol Cell Biol *28*, 6731-6745.

Bertone, P., Stolc, V., Royce, T.E., Rozowsky, J.S., Urban, A.E., Zhu, X., Rinn, J.L., Tongprasit, W., Samanta, M., Weissman, S.*, et al.* (2004). Global identification of human transcribed sequences with genome tiling arrays. Science *306*, 2242-2246.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. Genes Dev *16*, 6-21.

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E.*, et al.* (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799-816.

Blagitko, N., Mergenthaler, S., Schulz, U., Wollmann, H.A., Craigen, W., Eggermann, T., Ropers, H.H., and Kalscheuer, V.M. (2000). Human GRB10 is imprinted and expressed from the paternal and maternal allele in a highly tissue- and isoform-specific fashion. Hum Mol Genet *9*, 1587-1595.

Blagitko, N., Schulz, U., Schinzel, A.A., Ropers, H.H., and Kalscheuer, V.M. (1999). gamma2-COP, a novel imprinted gene on chromosome 7q32, defines a new imprinting cluster in the human genome. Hum Mol Genet *8*, 2387-2396.

Boccaccio, I., Glatt-Deeley, H., Watrin, F., Roeckel, N., Lalande, M., and Muscatelli, F. (1999). The human MAGEL2 gene and its mouse homologue are paternally expressed and mapped to the Prader-Willi region. Hum Mol Genet *8*, 2497-2505.

Borchert, G.M., Lanier, W., and Davidson, B.L. (2006). RNA polymerase III transcribes human microRNAs. Nat Struct Mol Biol *13*, 1097-1101.

Bracken, A.P., Dietrich, N., Pasini, D., Hansen, K.H., and Helin, K. (2006). Genome-wide mapping of Polycomb target genes unravels their roles in cell fate transitions. Genes Dev *20*, 1123-1136.

Brannan, C.I., Dees, E.C., Ingram, R.S., and Tilghman, S.M. (1990). The product of the H19 gene may function as an RNA. Mol Cell Biol *10*, 28-36.

Brown, C.J., Ballabio, A., Rupert, J.L., Lafreniere, R.G., Grompe, M., Tonlorenzi, R., and Willard, H.F. (1991). A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. Nature *349*, 38-44.

Buettner, V.L., Walker, A.M., and Singer-Sam, J. (2005). Novel paternally expressed intergenic transcripts at the mouse Prader-Willi/Angelman Syndrome locus. Mamm Genome *16*, 219-227.

Buiting, K., Lich, C., Cottrell, S., Barnicoat, A., and Horsthemke, B. (1999). A 5-kb imprinting center deletion in a family with Angelman syndrome reduces the shortest region of deletion overlap to 880 bp. Hum Genet *105*, 665-666.

Burzio, L.O., Riquelme, P.T., and Koide, S.S. (1979). ADP ribosylation of rat liver nucleosomal core histones. J Biol Chem *254*, 3029-3037.

Bussemakers, M.J., van Bokhoven, A., Verhaegh, G.W., Smit, F.P., Karthaus, H.F., Schalken, J.A., Debruyne, F.M., Ru, N., and Isaacs, W.B. (1999). DD3: a new prostate-specific gene, highly overexpressed in prostate cancer. Cancer Res *59*, 5975-5979.

Butler, M.G. (1990). Prader-Willi syndrome: current understanding of cause and diagnosis. Am J Med Genet *35*, 319-332.

Cai, X., and Cullen, B.R. (2007). The imprinted H19 noncoding RNA is a primary microRNA precursor. RNA *13*, 313-316.

Calin, G.A., Liu, C.G., Ferracin, M., Hyslop, T., Spizzo, R., Sevignani, C., Fabbri, M., Cimmino, A., Lee, E.J., Wojcik, S.E.*, et al.* (2007). Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. Cancer Cell *12*, 215-229.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C.*, et al.* (2005). The transcriptional landscape of the mammalian genome. Science *309*, 1559-1563.

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C.A., Taylor, M.S., Engstrom, P.G., Frith, M.C.*, et al.* (2006). Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet *38*, 626-635.

Cavaille, J., Buiting, K., Kiefmann, M., Lalande, M., Brannan, C.I., Horsthemke, B., Bachellerie, J.P., Brosius, J., and Huttenhofer, A. (2000). Identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization. Proc Natl Acad Sci U S A *97*, 14311-14316.

Cavaille, J., Seitz, H., Paulsen, M., Ferguson-Smith, A.C., and Bachellerie, J.P. (2002). Identification of tandemly-repeated C/D snoRNA genes at the imprinted human 14q32 domain reminiscent of those at the Prader-Willi/Angelman syndrome region. Hum Mol Genet *11*, 1527-1538.

Cayre, A., Rossignol, F., Clottes, E., and Penault-Llorca, F. (2003). aHIF but not HIF-1alpha transcript is a poor prognostic marker in human breast cancer. Breast Cancer Res *5*, R223-230.

Chai, J.H., Locke, D.P., Ohta, T., Greally, J.M., and Nicholls, R.D. (2001). Retrotransposed genes such as Frat3 in the mouse Chromosome 7C Prader-

Willi syndrome region acquire the imprinted status of their insertion site. Mamm Genome *12*, 813-821.

Chamberlain, S.J., and Brannan, C.I. (2001). The Prader-Willi syndrome imprinting center activates the paternally expressed murine Ube3a antisense transcript but represses paternal Ube3a. Genomics *73*, 316-322.

Chang, S.C., and Brown, C.J. (2010). Identification of regulatory elements flanking human XIST reveals species differences. BMC Mol Biol *11*, 20.

Chang, S.C., Tucker, T., Thorogood, N.P., and Brown, C.J. (2006). Mechanisms of X-chromosome inactivation. Front Biosci *11*, 852-866.

Chang, Y.F., Imam, J.S., and Wilkinson, M.F. (2007). The nonsense-mediated decay RNA surveillance pathway. Annu Rev Biochem *76*, 51-74.

Chaumeil, J., Okamoto, I., Guggiari, M., and Heard, E. (2002). Integrated kinetics of X chromosome inactivation in differentiating embryonic stem cells. Cytogenet Genome Res *99*, 75-84.

Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G.*, et al.* (2005). Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science *308*, 1149-1154.

Chess, A., Simon, I., Cedar, H., and Axel, R. (1994). Allelic inactivation regulates olfactory receptor gene expression. Cell *78*, 823-834.

Cheung, V.G., Conlin, L.K., Weber, T.M., Arcaro, M., Jen, K.Y., Morley, M., and Spielman, R.S. (2003). Natural variation in human gene expression assessed in lymphoblastoid cells. Nat Genet *33*, 422-425.

Choi, J.D., Underkoffler, L.A., Wood, A.J., Collins, J.N., Williams, P.T., Golden, J.A., Schuster, E.F., Jr., Loomes, K.M., and Oakey, R.J. (2005). A novel variant of Inpp5f is imprinted in brain, and its expression is correlated with differential methylation of an internal CpG island. Mol Cell Biol *25*, 5514-5522.

Choo, J.H., Kim, J.D., and Kim, J. (2008). Imprinting of an evolutionarily conserved antisense transcript gene APeg3. Gene *409*, 28-33.

Chooniedass-Kothari, S., Emberley, E., Hamedani, M.K., Troup, S., Wang, X., Czosnek, A., Hube, F., Mutawe, M., Watson, P.H., and Leygue, E. (2004). The steroid receptor RNA activator is the first functional RNA encoding a protein. FEBS Lett *566*, 43-47.

Chow, J.C., Hall, L.L., Clemson, C.M., Lawrence, J.B., and Brown, C.J. (2003). Characterization of expression at the human XIST locus in somatic, embryonal carcinoma, and transgenic cell lines. Genomics *82*, 309-322.

Clamp, M., Fry, B., Kamal, M., Xie, X., Cuff, J., Lin, M.F., Kellis, M., Lindblad-Toh, K., and Lander, E.S. (2007). Distinguishing protein-coding and noncoding genes in the human genome. Proc Natl Acad Sci U S A *104*, 19428-19433.

Clark, L., Wei, M., Cattoretti, G., Mendelsohn, C., and Tycko, B. (2002). The Tnfrh1 (Tnfrsf23) gene is weakly imprinted in several organs and expressed at the trophoblast-decidua interface. BMC Genet *3*, 11.

Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science *322*, 1845-1848.

Crick, F. (1970). Central dogma of molecular biology. Nature *227*, 561-563.

Dallosso, A.R., Hancock, A.L., Brown, K.W., Williams, A.C., Jackson, S., and Malik, K. (2004). Genomic imprinting at the WT1 gene involves a novel coding transcript (AWT1) that shows deregulation in Wilms' tumours. Hum Mol Genet *13*, 405-415.

Dallosso, A.R., Hancock, A.L., Malik, S., Salpekar, A., King-Underwood, L., Pritchard-Jones, K., Peters, J., Moorwood, K., Ward, A., Malik, K.T., and Brown, K.W. (2007). Alternately spliced WT1 antisense transcripts interact with WT1 sense RNA and show epigenetic and splicing defects in cancer. RNA *13*, 2287-2299.

Dao, D., Frank, D., Qian, N., O'Keefe, D., Vosatka, R.J., Walsh, C.P., and Tycko, B. (1998). IMPT1, an imprinted gene similar to polyspecific transporter and multi-drug resistance genes. Hum Mol Genet *7*, 597-608.

Das, R., Hampton, D.D., and Jirtle, R.L. (2009). Imprinting evolution and human health. Mamm Genome *20*, 563-572.

Davies, H.D., Leusink, G.L., McConnell, A., Deyell, M., Cassidy, S.B., Fick, G.H., and Coppes, M.J. (2003). Myeloid leukemia in Prader-Willi syndrome. J Pediatr *142*, 174-178.

de Kok, J.B., Verhaegh, G.W., Roelofs, R.W., Hessels, D., Kiemeney, L.A., Aalders, T.W., Swinkels, D.W., and Schalken, J.A. (2002). DD3(PCA3), a very sensitive and specific marker to detect prostate tumors. Cancer Res *62*, 2695-2698.

de la Puente, A., Hall, J., Wu, Y.Z., Leone, G., Peters, J., Yoon, B.J., Soloway, P., and Plass, C. (2002). Structural characterization of Rasgrf1 and a novel linked imprinted locus. Gene *291*, 287-297.

de los Santos, T., Schweizer, J., Rees, C.A., and Francke, U. (2000). Small evolutionarily conserved RNA, resembling C/D box small nucleolar RNA, is transcribed from PWCR1, a novel imprinted gene in the Prader-Willi deletion region, which Is highly expressed in brain. Am J Hum Genet *67*, 1067-1082.

De Souza, A.T., Yamada, T., Mills, J.J., and Jirtle, R.L. (1997). Imprinted genes in liver carcinogenesis. FASEB J *11*, 60-67.

Debrand, E., Chureau, C., Arnaud, D., Avner, P., and Heard, E. (1999). Functional analysis of the DXPas34 locus, a 3' regulator of Xist expression. Mol Cell Biol *19*, 8513-8525.

DeChiara, T.M., Robertson, E.J., and Efstratiadis, A. (1991). Parental imprinting of the mouse insulin-like growth factor II gene. Cell *64*, 849-859.

Delaval, K., Govin, J., Cerqueira, F., Rousseaux, S., Khochbin, S., and Feil, R. (2007). Differential histone modifications mark mouse imprinting control regions during spermatogenesis. EMBO J *26*, 720-729.

Deltour, L., Montagutelli, X., Guenet, J.L., Jami, J., and Paldi, A. (1995). Tissue- and developmental stage-specific imprinting of the mouse proinsulin gene, Ins2. Dev Biol *168*, 686-688.

Dickman, S. (1997). First p53 relative may be a new tumor suppressor. Science *277*, 1605-1606.

Dindot, S.V., Person, R., Strivens, M., Garcia, R., and Beaudet, A.L. (2009). Epigenetic profiling at mouse imprinted gene clusters reveals novel epigenetic and genetic features at differentially methylated regions. Genome Res *19*, 1374-1383.

Dinger, M.E., Amaral, P.P., Mercer, T.R., and Mattick, J.S. (2009). Pervasive transcription of the eukaryotic genome: functional indices and conceptual implications. Brief Funct Genomic Proteomic *8*, 407-423.

Duker, A.L., Ballif, B.C., Bawle, E.V., Person, R.E., Mahadevan, S., Alliman, S., Thompson, R., Traylor, R., Bejjani, B.A., Shaffer, L.G., *et al.* (2010). Paternally inherited microdeletion at 15q11.2 confirms a significant role for the SNORD116 C/D box snoRNA cluster in Prader-Willi syndrome. Eur J Hum Genet.

Ebisuya, M., Yamamoto, T., Nakajima, M., and Nishida, E. (2008). Ripples from neighbouring transcription. Nat Cell Biol *10*, 1106-1113.

Efroni, S., Duttagupta, R., Cheng, J., Dehghani, H., Hoeppner, D.J., Dash, C., Bazett-Jones, D.P., Le Grice, S., McKay, R.D., Buetow, K.H., *et al.* (2008). Global transcription in pluripotent embryonic stem cells. Cell Stem Cell *2*, 437-447.

Ender, C., Krek, A., Friedlander, M.R., Beitzinger, M., Weinmann, L., Chen, W., Pfeffer, S., Rajewsky, N., and Meister, G. (2008). A human snoRNA with microRNA-like functions. Mol Cell *32*, 519-528.

Engemann, S., Strodicke, M., Paulsen, M., Franck, O., Reinhardt, R., Lane, N., Reik, W., and Walter, J. (2000). Sequence and functional comparison in the Beckwith-Wiedemann region: implications for a novel imprinting centre and extended imprinting. Hum Mol Genet *9*, 2691-2706.

Evans, H.K., Wylie, A.A., Murphy, S.K., and Jirtle, R.L. (2001). The neuronatin gene resides in a "micro-imprinted" domain on human chromosome 20q11.2. Genomics *77*, 99-104.

Faghihi, M.A., Modarresi, F., Khalil, A.M., Wood, D.E., Sahagan, B.G., Morgan, T.E., Finch, C.E., St Laurent, G., 3rd, Kenny, P.J., and Wahlestedt, C. (2008). Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. Nat Med *14*, 723-730.

Fazzari, M.J., and Greally, J.M. (2004). Epigenomics: beyond CpG islands. Nat Rev Genet *5*, 446-455.

Feinberg, A.P., Ohlsson, R., and Henikoff, S. (2006). The epigenetic progenitor origin of human cancer. Nat Rev Genet *7*, 21-33.

Fejes-Toth (2009). Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. Nature *457*, 1028-1032.

Feng, J., Bi, C., Clark, B.S., Mady, R., Shah, P., and Kohtz, J.D. (2006). The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. Genes Dev *20*, 1470-1484.

Ferguson-Smith, A.C., Cattanach, B.M., Barton, S.C., Beechey, C.V., and Surani, M.A. (1991). Embryological and molecular investigations of parental imprinting on mouse chromosome 7. Nature *351*, 667-670.

Fitzpatrick, G.V., Soloway, P.D., and Higgins, M.J. (2002). Regional loss of imprinting and growth deficiency in mice with a targeted deletion of KvDMR1. Nat Genet *32*, 426-431.

Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M.*, et al.* (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851-861.

Frost, J.M., and Moore, G.E. (2010). The importance of imprinting in the human placenta. PLoS Genet *6*, e1001015.

Gabory, A., Ripoche, M.A., Le Digarcher, A., Watrin, F., Ziyyat, A., Forne, T., Jammes, H., Ainscough, J.F., Surani, M.A., Journot, L., and Dandolo, L. (2009). H19 acts as a trans regulator of the imprinted gene network controlling growth in mice. Development *136*, 3413-3421.

Gardiner-Garden, M., and Frommer, M. (1987). CpG islands in vertebrate genomes. J Mol Biol *196*, 261-282.

Geuns, E., De Temmerman, N., Hilven, P., Van Steirteghem, A., Liebaers, I., and De Rycke, M. (2007). Methylation analysis of the intergenic differentially methylated region of DLK1-GTL2 in human. Eur J Hum Genet *15*, 352-361.

Giannoukakis, N., Deal, C., Paquette, J., Goodyer, C.G., and Polychronakos, C. (1993). Parental genomic imprinting of the human IGF2 gene. Nat Genet *4*, 98-101.

Giddings, S.J., King, C.D., Harman, K.W., Flood, J.F., and Carnaghi, L.R. (1994). Allele specific inactivation of insulin 1 and 2, in the mouse yolk sac, indicates imprinting. Nat Genet *6*, 310-313.

Girard, A., Sachidanandam, R., Hannon, G.J., and Carmell, M.A. (2006). A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature *442*, 199-202.

Glenn, C.C., Porter, K.A., Jong, M.T., Nicholls, R.D., and Driscoll, D.J. (1993). Functional imprinting and epigenetic modification of the human SNRPN gene. Hum Mol Genet *2*, 2001-2005.

Gould, T.D., and Pfeifer, K. (1998). Imprinting of mouse Kvlqt1 is developmentally regulated. Hum Mol Genet *7*, 483-487.

Grabowski, M., Zimprich, A., Lorenz-Depiereux, B., Kalscheuer, V., Asmus, F., Gasser, T., Meitinger, T., and Strom, T.M. (2003). The epsilon-sarcoglycan gene (SGCE), mutated in myoclonus-dystonia syndrome, is maternally imprinted. Eur J Hum Genet *11*, 138-144.

Gray, T.A., Saitoh, S., and Nicholls, R.D. (1999). An imprinted, mammalian bicistronic transcript encodes two independent proteins. Proc Natl Acad Sci U S A *96*, 5616-5621.

Gray, T.A., Wilson, A., Fortin, P.J., and Nicholls, R.D. (2006). The putatively functional Mkrn1-p1 pseudogene is neither expressed nor imprinted, nor does it regulate its source gene in trans. Proc Natl Acad Sci U S A *103*, 12039-12044.

Grivna, S.T., Beyret, E., Wang, Z., and Lin, H. (2006). A novel class of small RNAs in mouse spermatogenic cells. Genes Dev *20*, 1709-1714.

Guccione, E., Martinato, F., Finocchiaro, G., Luzi, L., Tizzoni, L., Dall' Olio, V., Zardo, G., Nervi, C., Bernard, L., and Amati, B. (2006). Myc-binding-site recognition in the human genome is determined by chromatin context. Nat Cell Biol *8*, 764-770.

Guillemot, F., Caspary, T., Tilghman, S.M., Copeland, N.G., Gilbert, D.J., Jenkins, N.A., Anderson, D.J., Joyner, A.L., Rossant, J., and Nagy, A. (1995). Genomic imprinting of Mash2, a mouse gene required for trophoblast development. Nat Genet *9*, 235-242.

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

Hagan, J.P., O'Neill, B.L., Stewart, C.L., Kozlov, S.V., and Croce, C.M. (2009). At least ten genes define the imprinted Dlk1-Dio3 cluster on mouse chromosome 12qF1. PLoS One *4*, e4352.

Hagiwara, T., Nakashima, K., Hirano, H., Senshu, T., and Yamada, M. (2002). Deimination of arginine residues in nucleophosmin/B23 and histones in HL-60 granulocytes. Biochem Biophys Res Commun *290*, 979-983.

Hagiwara, Y., Hirai, M., Nishiyama, K., Kanazawa, I., Ueda, T., Sakaki, Y., and Ito, T. (1997). Screening for imprinted genes by allelic message display: identification of a paternally expressed gene impact on mouse chromosome 18. Proc Natl Acad Sci U S A *94*, 9249-9254.

Harikrishnan, K.N., Chow, M.Z., Baker, E.K., Pal, S., Bassal, S., Brasacchio, D., Wang, L., Craig, J.M., Jones, P.L., Sif, S., and El-Osta, A. (2005). Brahma links the SWI/SNF chromatin-remodeling complex with MeCP2-dependent transcriptional silencing. Nat Genet *37*, 254-264.

Hark, A.T., Schoenherr, C.J., Katz, D.J., Ingram, R.S., Levorse, J.M., and Tilghman, S.M. (2000). CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. Nature *405*, 486-489.

Harrison, P.M., Kumar, A., Lang, N., Snyder, M., and Gerstein, M. (2002). A question of size: the eukaryotic proteome and the problems in defining it. Nucleic Acids Res *30*, 1083-1090.

Hata, K., Okano, M., Lei, H., and Li, E. (2002). Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice. Development *129*, 1983-1993.

Hatada, I., Morita, S., Obata, Y., Sotomaru, Y., Shimoda, M., and Kono, T. (2001). Identification of a new imprinted gene, Rian, on mouse chromosome 12 by fluorescent differential display screening. J Biochem *130*, 187-190.

Hatada, I., and Mukai, T. (1995). Genomic imprinting of p57KIP2, a cyclin-dependent kinase inhibitor, in mouse. Nat Genet *11*, 204-206.

Hatada, I., Sugama, T., and Mukai, T. (1993). A new imprinted gene cloned by a methylation-sensitive genome scanning method. Nucleic Acids Res *21*, 5577-5582.

Hayward, B.E., Barlier, A., Korbonits, M., Grossman, A.B., Jacquet, P., Enjalbert, A., and Bonthron, D.T. (2001). Imprinting of the G(s)alpha gene GNAS1 in the pathogenesis of acromegaly. J Clin Invest *107*, R31-36.

Hayward, B.E., and Bonthron, D.T. (2000). An imprinted antisense transcript at the human GNAS1 locus. Hum Mol Genet *9*, 835-841.

Hayward, B.E., Kamiya, M., Strain, L., Moran, V., Campbell, R., Hayashizaki, Y., and Bonthron, D.T. (1998a). The human GNAS1 gene is imprinted and encodes distinct paternally and biallelically expressed G proteins. Proc Natl Acad Sci U S A *95*, 10038-10043.

Hayward, B.E., Moran, V., Strain, L., and Bonthron, D.T. (1998b). Bidirectional imprinting of a single gene: GNAS1 encodes maternally, paternally, and biallelically derived proteins. Proc Natl Acad Sci U S A *95*, 15475-15480.

Heard, E., and Disteche, C.M. (2006). Dosage compensation in mammals: fine-tuning the expression of the X chromosome. Genes Dev *20*, 1848-1867.

Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A.*, et al.* (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet *39*, 311-318.

Hellman, A., and Chess, A. (2007). Gene body-specific methylation on the active X chromosome. Science *315*, 1141-1143.

Higashimoto, K., Soejima, H., Yatsuki, H., Joh, K., Uchiyama, M., Obata, Y., Ono, R., Wang, Y., Xin, Z., Zhu, X.*, et al.* (2002). Characterization and imprinting status of OBPH1/Obph1 gene: implications for an extended imprinting domain in human and mouse. Genomics *80*, 575-584.

Hikichi, T., Kohda, T., Kaneko-Ishino, T., and Ishino, F. (2003). Imprinting regulation of the murine Meg1/Grb10 and human GRB10 genes; roles of brain-specific promoters and mouse-specific CTCF-binding sites. Nucleic Acids Res *31*, 1398-1406.

Hirotsune, S., Yoshida, N., Chen, A., Garrett, L., Sugiyama, F., Takahashi, S., Yagami, K., Wynshaw-Boris, A., and Yoshiki, A. (2003). An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. Nature *423*, 91-96.

Hockemeyer, D., Soldner, F., Beard, C., Gao, Q., Mitalipova, M., DeKelver, R.C., Katibah, G.E., Amora, R., Boydston, E.A., Zeitler, B.*, et al.* (2009). Efficient targeting of expressed and silent genes in human ESCs and iPSCs using zinc-finger nucleases. Nat Biotechnol *27*, 851-857.

Hogart, A., Nagarajan, R.P., Patzel, K.A., Yasui, D.H., and Lasalle, J.M. (2007). 15q11-13 GABAA receptor genes are normally biallelically expressed in brain yet are subject to epigenetic dysregulation in autism-spectrum disorders. Hum Mol Genet *16*, 691-703.

Holm, T.M., Jackson-Grusby, L., Brambrink, T., Yamada, Y., Rideout, W.M., 3rd, and Jaenisch, R. (2005). Global loss of imprinting leads to widespread tumorigenesis in adult mice. Cancer Cell *8*, 275-285.

Horike, S., Cai, S., Miyano, M., Cheng, J.F., and Kohwi-Shigematsu, T. (2005). Loss of silent-chromatin looping and impaired imprinting of DLX5 in Rett syndrome. Nat Genet *37*, 31-40.

Horsthemke, B., and Wagstaff, J. (2008). Mechanisms of imprinting of the Prader-Willi/Angelman region. Am J Med Genet A *146A*, 2041-2052.

Hoshiya, H., Meguro, M., Kashiwagi, A., Okita, C., and Oshimura, M. (2003). Calcr, a brain-specific imprinted mouse calcitonin receptor gene in the imprinted cluster of the proximal region of chromosome 6. J Hum Genet *48*, 208-211.

Hu, J., Lutz, C.S., Wilusz, J., and Tian, B. (2005). Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation. RNA *11*, 1485-1493.

Huttenhofer, A., Schattner, P., and Polacek, N. (2005). Non-coding RNAs: hope or hype? Trends Genet *21*, 289-297.

Hutter, B., Helms, V., and Paulsen, M. (2006). Tandem repeats in the CpG islands of imprinted genes. Genomics *88*, 323-332.

Ideue, T., Hino, K., Kitao, S., Yokoi, T., and Hirose, T. (2009). Efficient oligonucleotide-mediated degradation of nuclear noncoding RNAs in mammalian cultured cells. RNA *15*, 1578-1587.

Illingworth, R., Kerr, A., Desousa, D., Jorgensen, H., Ellis, P., Stalker, J., Jackson, D., Clee, C., Plumb, R., Rogers, J.*, et al.* (2008). A novel CpG island set identifies tissue-specific methylation at developmental gene loci. PLoS Biol *6*, e22.

Inoue, J., Mitsuya, K., Maegawa, S., Kugoh, H., Kadota, M., Okamura, D., Shinohara, T., Nishihara, S., Takehara, S., Yamauchi, K.*, et al.* (2001). Construction of 700 human/mouse A9 monochromosomal hybrids and analysis of imprinted genes on human chromosome 6. J Hum Genet *46*, 137-145.

Jacob, F., and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol *3*, 318-356.

Jacquier, A. (2009). The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. Nat Rev Genet *10*, 833-844.

Jelinic, P., and Shaw, P. (2007). Loss of imprinting and cancer. J Pathol *211*, 261-268.

Ji, P., Diederichs, S., Wang, W., Boing, S., Metzger, R., Schneider, P.M., Tidow, N., Brandt, B., Buerger, H., Bulk, E.*, et al.* (2003). MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. Oncogene *22*, 8031-8041.

Jinno, Y., Sengoku, K., Nakao, M., Tamate, K., Miyamoto, T., Matsuzaka, T., Sutcliffe, J.S., Anan, T., Takuma, N., Nishiwaki, K.*, et al.* (1996). Mouse/human sequence divergence in a region with a paternal-specific methylation imprint at the human H19 locus. Hum Mol Genet *5*, 1155-1161.

Jinno, Y., Yun, K., Nishiwaki, K., Kubota, T., Ogawa, O., Reeve, A.E., and Niikawa, N. (1994). Mosaic and polymorphic imprinting of the WT1 gene in humans. Nat Genet *6*, 305-309.

Johnson, J.M., Edwards, S., Shoemaker, D., and Schadt, E.E. (2005). Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. Trends Genet *21*, 93-102.

Johnston, C.M., Newall, A.E., Brockdorff, N., and Nesterova, T.B. (2002). Enox, a novel gene that maps 10 kb upstream of Xist and partially escapes X inactivation. Genomics *80*, 236-244.

Jong, M.T., Carey, A.H., Caldwell, K.A., Lau, M.H., Handel, M.A., Driscoll, D.J., Stewart, C.L., Rinchik, E.M., and Nicholls, R.D. (1999a). Imprinting of a RING zinc-finger encoding gene in the mouse chromosome region homologous to the Prader-Willi syndrome genetic region. Hum Mol Genet *8*, 795-803.

Jong, M.T., Gray, T.A., Ji, Y., Glenn, C.C., Saitoh, S., Driscoll, D.J., and Nicholls, R.D. (1999b). A novel imprinted gene, encoding a RING zinc-finger protein, and overlapping antisense transcript in the Prader-Willi syndrome critical region. Hum Mol Genet *8*, 783-793.

Kaghad, M., Bonnet, H., Yang, A., Creancier, L., Biscan, J.C., Valent, A., Minty, A., Chalon, P., Lelias, J.M., Dumont, X.*, et al.* (1997). Monoallelically expressed gene related to p53 at 1p36, a region frequently deleted in neuroblastoma and other human cancers. Cell *90*, 809-819.

Kagitani, F., Kuroiwa, Y., Wakana, S., Shiroishi, T., Miyoshi, N., Kobayashi, S., Nishida, M., Kohda, T., Kaneko-Ishino, T., and Ishino, F. (1997). Peg5/Neuronatin is an imprinted gene located on sub-distal chromosome 2 in the mouse. Nucleic Acids Res *25*, 3428-3432.

Kamikihara, T., Arima, T., Kato, K., Matsuda, T., Kato, H., Douchi, T., Nagata, Y., Nakao, M., and Wake, N. (2005). Epigenetic silencing of the imprinted gene ZAC by DNA methylation is an early event in the progression of human ovarian cancer. Int J Cancer *115*, 690-700.

Kamiya, M., Judson, H., Okazaki, Y., Kusakabe, M., Muramatsu, M., Takada, S., Takagi, N., Arima, T., Wake, N., Kamimura, K.*, et al.* (2000). The cell cycle control gene ZAC/PLAGL1 is imprinted--a strong candidate gene for transient neonatal diabetes. Hum Mol Genet *9*, 453-460.

Kaneda, A., and Feinberg, A.P. (2005). Loss of imprinting of IGF2: a common epigenetic modifier of intestinal tumor risk. Cancer Res *65*, 11236-11240.

Kaneda, M., Okano, M., Hata, K., Sado, T., Tsujimoto, N., Li, E., and Sasaki, H. (2004). Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. Nature *429*, 900-903.

Kaneko-Ishino, T., Kohda, T., Ono, R., and Ishino, F. (2006). Complementation hypothesis: the necessity of a monoallelic gene expression mechanism in mammalian development. Cytogenet Genome Res *113*, 24-30.

Kaneko-Ishino, T., Kuroiwa, Y., Miyoshi, N., Kohda, T., Suzuki, R., Yokoyama, M., Viville, S., Barton, S.C., Ishino, F., and Surani, M.A. (1995). Peg1/Mest imprinted gene on chromosome 6 identified by cDNA subtraction hybridization. Nat Genet *11*, 52-59.

Kapranov, P., Cheng, J., Dike, S., Nix, D.A., Duttagupta, R., Willingham, A.T., Stadler, P.F., Hertel, J., Hackermuller, J., Hofacker, I.L.*, et al.* (2007a). RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science *316*, 1484-1488.

Kapranov, P., Willingham, A.T., and Gingeras, T.R. (2007b). Genome-wide transcription and the implications for genomic organization. Nat Rev Genet *8*, 413-423.

Kashiwagi, A., Meguro, M., Hoshiya, H., Haruta, M., Ishino, F., Shibahara, T., and Oshimura, M. (2003). Predominant maternal expression of the mouse Atp10c in hippocampus and olfactory bulb. J Hum Genet *48*, 194-198.

Kass, S.U., Landsberger, N., and Wolffe, A.P. (1997). DNA methylation directs a time-dependent repression of transcription initiation. Curr Biol *7*, 157-165.

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap, C.C., Suzuki, M., Kawai, J.*, et al.* (2005). Antisense transcription in the mammalian transcriptome. Science *309*, 1564-1566.

Kato, M.V., Ikawa, Y., Hayashizaki, Y., and Shibata, H. (1998). Paternal imprinting of mouse serotonin receptor 2A gene Htr2 in embryonic eye: a conserved imprinting regulation on the RB/Rb locus. Genomics *47*, 146-148.

Kato, M.V., Shimizu, T., Nagayoshi, M., Kaneko, A., Sasaki, M.S., and Ikawa, Y. (1996). Genomic imprinting of the human serotonin-receptor (HTR2) gene involved in development of retinoblastoma. Am J Hum Genet *59*, 1084-1090.

Kay, G.F., Barton, S.C., Surani, M.A., and Rastan, S. (1994). Imprinting and X chromosome counting mechanisms determine Xist expression in early mouse development. Cell *77*, 639-650.

Kayashima, T., Yamasaki, K., Joh, K., Yamada, T., Ohta, T., Yoshiura, K., Matsumoto, N., Nakane, Y., Mukai, T., Niikawa, N., and Kishino, T. (2003a). Atp10a, the mouse ortholog of the human imprinted ATP10A gene, escapes genomic imprinting. Genomics *81*, 644-647.

Kayashima, T., Yamasaki, K., Yamada, T., Sakai, H., Miwa, N., Ohta, T., Yoshiura, K., Matsumoto, N., Nakane, Y., Kanetake, H.*, et al.* (2003b). The novel imprinted carboxypeptidase A4 gene (CPA4) in the 7q32 imprinting domain. Hum Genet *112*, 220-226.

Keverne, E.B., and Curley, J.P. (2008). Epigenetics, brain evolution and behaviour. Front Neuroendocrinol *29*, 398-412.

Khalil, A.M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., Thomas, K., Presser, A., Bernstein, B.E., van Oudenaarden, A.*, et al.* (2009). Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci U S A *106*, 11667-11672.

Khosla, S., Mendiratta, G., and Brahmachari, V. (2006). Genomic imprinting in the mealybugs. Cytogenet Genome Res *113*, 41-52.

Kikyo, N., Williamson, C.M., John, R.M., Barton, S.C., Beechey, C.V., Ball, S.T., Cattanach, B.M., Surani, M.A., and Peters, J. (1997). Genetic and functional analysis of neuronatin in mice with maternal or paternal duplication of distal Chr 2. Dev Biol *190*, 66-77.

Kim, J., Bergmann, A., Lucas, S., Stone, R., and Stubbs, L. (2004). Lineage-specific imprinting and evolution of the zinc-finger gene ZIM2. Genomics *84*, 47-58.

Kim, J., Bergmann, A., Wehri, E., Lu, X., and Stubbs, L. (2001). Imprinting and evolution of two Kruppel-type zinc-finger genes, ZIM3 and ZNF264, located in the PEG3/USP29 imprinted domain. Genomics *77*, 91-98.

Kim, J., Lu, X., and Stubbs, L. (1999). Zim1, a maternally expressed mouse Kruppel-type zinc-finger gene located in proximal chromosome 7. Hum Mol Genet *8*, 847-854.

Kim, J., Noskov, V.N., Lu, X., Bergmann, A., Ren, X., Warth, T., Richardson, P., Kouprina, N., and Stubbs, L. (2000). Discovery of a novel, paternally expressed ubiquitin-specific processing protease gene through comparative analysis of an imprinted region of mouse chromosome 7 and human chromosome 19q13.4. Genome Res *10*, 1138-1147.

Kishore, S., and Stamm, S. (2006). The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. Science *311*, 230-232.

Kobayashi, H., Yamada, K., Morita, S., Hiura, H., Fukuda, A., Kagami, M., Ogata, T., Hata, K., Sotomaru, Y., and Kono, T. (2009). Identification of the mouse paternally expressed imprinted gene Zdbf2 on chromosome 1 and its imprinted human homolog ZDBF2 on chromosome 2. Genomics *93*, 461-472.

Kobayashi, S., Fujihara, Y., Mise, N., Kaseda, K., Abe, K., Ishino, F., and Okabe, M. (2010). The X-linked imprinted gene family Fthl17 shows predominantly female expression following the two-cell stage in mouse embryos. Nucleic Acids Res.

Kobayashi, S., Isotani, A., Mise, N., Yamamoto, M., Fujihara, Y., Kaseda, K., Nakanishi, T., Ikawa, M., Hamada, H., Abe, K., and Okabe, M. (2006). Comparison of gene expression in male and female mouse blastocysts revealed imprinting of the X-linked gene, Rhox5/Pem, at preimplantation stages. Curr Biol *16*, 166-172.

Kobayashi, S., Kohda, T., Ichikawa, H., Ogura, A., Ohki, M., Kaneko-Ishino, T., and Ishino, F. (2002). Paternal expression of a novel imprinted gene, Peg12/Frat3, in the mouse 7C region homologous to the Prader-Willi syndrome region. Biochem Biophys Res Commun *290*, 403-408.

Kobayashi, S., Kohda, T., Miyoshi, N., Kuroiwa, Y., Aisaka, K., Tsutsumi, O., Kaneko-Ishino, T., and Ishino, F. (1997). Human PEG1/MEST, an imprinted gene on chromosome 7. Hum Mol Genet *6*, 781-786.

Koerner, M.V., and Barlow, D.P. (2010). Genomic imprinting-an epigenetic gene-regulatory model. Curr Opin Genet Dev.

Koerner, M.V., Pauler, F.M., Huang, R., and Barlow, D.P. (2009). The function of non-coding RNAs in genomic imprinting. Development *136*, 1771-1783.

Kohda, M., Hoshiya, H., Katoh, M., Tanaka, I., Masuda, R., Takemura, T., Fujiwara, M., and Oshimura, M. (2001). Frequent loss of imprinting of IGF2 and MEST in lung adenocarcinoma. Mol Carcinog *31*, 184-191.

Kono, T. (2006). Genomic imprinting is a barrier to parthenogenesis in mammals. Cytogenet Genome Res *113*, 31-35.

Kouzarides, T. (2007). Chromatin modifications and their function. Cell *128*, 693-705.

Krueger, C., and Morison, I.M. (2008). Random monoallelic expression: making a choice. Trends Genet *24*, 257-259.

Kuzmin, A., Han, Z., Golding, M.C., Mann, M.R., Latham, K.E., and Varmuza, S. (2008). The PcG gene Sfmbt2 is paternally expressed in extraembryonic tissues. Gene Expr Patterns *8*, 107-116.

Labialle, S., Yang, L., Ruan, X., Villemain, A., Schmidt, J.V., Hernandez, A., Wiltshire, T., Cermakian, N., and Naumova, A.K. (2008). Coordinated diurnal regulation of genes from the Dlk1-Dio3 imprinted domain: implications for regulation of clusters of non-paralogous genes. Hum Mol Genet *17*, 15-26.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W.*, et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Landers, M., Bancescu, D.L., Le Meur, E., Rougeulle, C., Glatt-Deeley, H., Brannan, C., Muscatelli, F., and Lalande, M. (2004). Regulation of the large (approximately 1000 kb) imprinted murine Ube3a antisense transcript by alternative exons upstream of Snurf/Snrpn. Nucleic Acids Res *32*, 3480-3492.

Laprise, S.L. (2009). Implications of epigenetics and genomic imprinting in assisted reproductive technologies. Mol Reprod Dev *76*, 1006-1018.

Latos, P.A., and Barlow, D.P. (2009). Regulation of imprinted expression by macro non-coding RNAs. RNA Biol *6*, 100-106.

Le Hir, H., Nott, A., and Moore, M.J. (2003). How introns influence and enhance eukaryotic gene expression. Trends Biochem Sci *28*, 215-220.

Le Meur, E., Watrin, F., Landers, M., Sturny, R., Lalande, M., and Muscatelli, F. (2005). Dynamic developmental regulation of the large non-coding RNA associated with the mouse 7C imprinted chromosomal region. Dev Biol *286*, 587-600.

Lee, J.T. (2000). Disruption of imprinted X inactivation by parent-of-origin effects at Tsix. Cell *103*, 17-27.

Lee, J.T., and Lu, N. (1999). Targeted mutagenesis of Tsix leads to nonrandom X inactivation. Cell *99*, 47-57.

Lee, J.T., Strauss, W.M., Dausman, J.A., and Jaenisch, R. (1996). A 450 kb transgene displays properties of the mammalian X-inactivation center. Cell *86*, 83-94.

Lee, M.P., DeBaun, M.R., Mitsuya, K., Galonek, H.L., Brandenburg, S., Oshimura, M., and Feinberg, A.P. (1999). Loss of imprinting of a paternally expressed transcript, with antisense orientation to KVLQT1, occurs frequently in Beckwith-Wiedemann syndrome and is independent of insulin-like growth factor II imprinting. Proc Natl Acad Sci U S A *96*, 5203-5208.

Lee, M.P., Hu, R.J., Johnson, L.A., and Feinberg, A.P. (1997). Human KVLQT1 gene shows tissue-specific imprinting and encompasses Beckwith-Wiedemann syndrome chromosomal rearrangements. Nat Genet *15*, 181-185.

Lee, Y., Kim, M., Han, J., Yeom, K.H., Lee, S., Baek, S.H., and Kim, V.N. (2004). MicroRNA genes are transcribed by RNA polymerase II. EMBO J *23*, 4051-4060.

Lee, Y.J., Park, C.W., Hahn, Y., Park, J., Lee, J., Yun, J.H., Hyun, B., and Chung, J.H. (2000). Mit1/Lb9 and Copg2, new members of mouse imprinted genes closely linked to Peg1/Mest(1). FEBS Lett *472*, 230-234.

Leff, S.E., Brannan, C.I., Reed, M.L., Ozcelik, T., Francke, U., Copeland, N.G., and Jenkins, N.A. (1992). Maternal imprinting of the mouse Snrpn gene and conserved linkage homology with the human Prader-Willi syndrome region. Nat Genet *2*, 259-264.

Li, B., Carey, M., and Workman, J.L. (2007). The role of chromatin during transcription. Cell *128*, 707-719.

Li, J., Bench, A.J., Piltz, S., Vassiliou, G., Baxter, E.J., Ferguson-Smith, A.C., and Green, A.R. (2005). L3mbtl, the mouse orthologue of the imprinted L3MBTL, displays a complex pattern of alternative splicing and escapes genomic imprinting. Genomics *86*, 489-494.

Li, J., Bench, A.J., Vassiliou, G.S., Fourouclas, N., Ferguson-Smith, A.C., and Green, A.R. (2004). Imprinting of the human L3MBTL gene, a polycomb family member located in a region of chromosome 20 deleted in human myeloid malignancies. Proc Natl Acad Sci U S A *101*, 7341-7346.

Li, T., Vu, T.H., Lee, K.O., Yang, Y., Nguyen, C.V., Bui, H.Q., Zeng, Z.L., Nguyen, B.T., Hu, J.F., Murphy, S.K., *et al.* (2002). An imprinted PEG1/MEST antisense expressed predominantly in human testis and in mature spermatozoa. J Biol Chem *277*, 13518-13527.

Li, T., Vu, T.H., Zeng, Z.L., Nguyen, B.T., Hayward, B.E., Bonthron, D.T., Hu, J.F., and Hoffman, A.R. (2000). Tissue-specific expression of antisense and sense transcripts at the imprinted Gnas locus. Genomics *69*, 295-304.

Liang, G., Chan, M.F., Tomigahara, Y., Tsai, Y.C., Gonzales, F.A., Li, E., Laird, P.W., and Jones, P.A. (2002). Cooperativity between DNA methyltransferases in the maintenance methylation of repetitive elements. Mol Cell Biol *22*, 480-491.

Lidegaard, O., Pinborg, A., and Andersen, A.N. (2005). Imprinting diseases and IVF: Danish National IVF cohort study. Hum Reprod *20*, 950-954.

Ligtenberg, M.J., Kuiper, R.P., Chan, T.L., Goossens, M., Hebeda, K.M., Voorendt, M., Lee, T.Y., Bodmer, D., Hoenselaar, E., Hendriks-Cornelissen, S.J., *et al.* (2009). Heritable somatic methylation and inactivation of MSH2 in families with Lynch syndrome due to deletion of the 3' exons of TACSTD1. Nat Genet *41*, 112-117.

Lin, R., Maeda, S., Liu, C., Karin, M., and Edgington, T.S. (2007). A large noncoding RNA is a marker for murine hepatocellular carcinomas and a spectrum of human carcinomas. Oncogene *26*, 851-858.

Lin, S.P., Youngson, N., Takada, S., Seitz, H., Reik, W., Paulsen, M., Cavaille, J., and Ferguson-Smith, A.C. (2003). Asymmetric regulation of imprinting on the maternal and paternal chromosomes at the Dlk1-Gtl2 imprinted cluster on mouse chromosome 12. Nat Genet *35*, 97-102.

Linder, D., McCaw, B.K., and Hecht, F. (1975). Parthenogenic origin of benign ovarian teratomas. N Engl J Med *292*, 63-66.

Liu, J., Litman, D., Rosenberg, M.J., Yu, S., Biesecker, L.G., and Weinstein, L.S. (2000). A GNAS1 imprinting defect in pseudohypoparathyroidism type IB. J Clin Invest *106*, 1167-1174.

Liu, W.M., Maraia, R.J., Rubin, C.M., and Schmid, C.W. (1994). Alu transcripts: cytoplasmic localisation and regulation by DNA methylation. Nucleic Acids Res *22*, 1087-1095.

Lomvardas, S., Barnea, G., Pisapia, D.J., Mendelsohn, M., Kirkland, J., and Axel, R. (2006). Interchromosomal interactions and olfactory receptor choice. Cell *126*, 403-413.

Lucifero, D., Mertineit, C., Clarke, H.J., Bestor, T.H., and Trasler, J.M. (2002). Methylation dynamics of imprinted genes in mouse germ cells. Genomics *79*, 530-538.

Luedi, P.P., Dietrich, F.S., Weidman, J.R., Bosko, J.M., Jirtle, R.L., and Hartemink, A.J. (2007). Computational and experimental identification of novel human imprinted genes. Genome Res *17*, 1723-1730.

Lyle, R., Watanabe, D., te Vruchte, D., Lerchner, W., Smrzka, O.W., Wutz, A., Schageman, J., Hahner, L., Davies, C., and Barlow, D.P. (2000). The imprinted antisense RNA at the Igf2r locus overlaps but does not imprint Mas1. Nat Genet *25*, 19-21.

MacDonald, H.R., and Wevrick, R. (1997). The necdin gene is deleted in Prader-Willi syndrome and is imprinted in human and mouse. Hum Mol Genet *6*, 1873-1878.

Maegawa, S., Itaba, N., Otsuka, S., Kamitani, H., Watanabe, T., Tahimic, C.G., Nanba, E., and Oshimura, M. (2004). Coordinate downregulation of a novel imprinted transcript ITUP1 with PEG3 in glioma cell lines. DNA Res *11*, 37-49.

Malik, K., Salpekar, A., Hancock, A., Moorwood, K., Jackson, S., Charles, A., and Brown, K.W. (2000). Identification of differential methylation of the WT1 antisense regulatory region and relaxation of imprinting in Wilms' tumor. Cancer Res *60*, 2356-2360.

Malone, C.D., and Hannon, G.J. (2009). Small RNAs as guardians of the genome. Cell *136*, 656-668.

Mancini-Dinardo, D., Steele, S.J., Levorse, J.M., Ingram, R.S., and Tilghman, S.M. (2006). Elongation of the Kcnq1ot1 transcript is required for genomic imprinting of neighboring genes. Genes Dev *20*, 1268-1282.

Mandel, C.R., Bai, Y., and Tong, L. (2008). Protein factors in pre-mRNA 3'-end processing. Cell Mol Life Sci *65*, 1099-1122.

Matlin, A.J., Clark, F., and Smith, C.W. (2005). Understanding alternative splicing: towards a cellular code. Nat Rev Mol Cell Biol *6*, 386-398.

Matouk, I.J., DeGroot, N., Mezan, S., Ayesh, S., Abu-lail, R., Hochberg, A., and Galun, E. (2007). The H19 non-coding RNA is essential for human tumor growth. PLoS One *2*, e845.

Matsuoka, S., Thompson, J.S., Edwards, M.C., Bartletta, J.M., Grundy, P., Kalikin, L.M., Harper, J.W., Elledge, S.J., and Feinberg, A.P. (1996). Imprinting of the gene encoding a human cyclin-dependent kinase inhibitor, p57KIP2, on chromosome 11p15. Proc Natl Acad Sci U S A *93*, 3026-3030.

Mattick, J.S. (2009). The genetic signatures of noncoding RNAs. PLoS Genet *5*, e1000459.

Mattick, J.S., and Gagen, M.J. (2001). The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNAs in the development of complex organisms. Mol Biol Evol *18*, 1611-1630.

Mattick, J.S., and Makunin, I.V. (2005). Small regulatory RNAs in mammals. Hum Mol Genet *14 Spec No 1*, R121-132.

Mattick, J.S., and Makunin, I.V. (2006). Non-coding RNA. Hum Mol Genet *15 Spec No 1*, R17-29.

McCann, J.A., Zheng, H., Islam, A., Goodyer, C.G., and Polychronakos, C. (2001). Evidence against GRB10 as the gene responsible for Silver-Russell syndrome. Biochem Biophys Res Commun *286*, 943-948.

McGrath, J., and Solter, D. (1984). Completion of mouse embryogenesis requires both the maternal and paternal genomes. Cell *37*, 179-183.

Meguro, M., Kashiwagi, A., Mitsuya, K., Nakao, M., Kondo, I., Saitoh, S., and Oshimura, M. (2001). A novel maternally expressed gene, ATP10C, encodes a putative aminophospholipid translocase associated with Angelman syndrome. Nat Genet *28*, 19-20.

Meguro, M., Mitsuya, K., Sui, H., Shigenami, K., Kugoh, H., Nakao, M., and Oshimura, M. (1997). Evidence for uniparental, paternal expression of the human GABAA receptor subunit genes, using microcell-mediated chromosome transfer. Hum Mol Genet *6*, 2127-2133.

Mendes Soares, L.M., and Valcarcel, J. (2006). The expanding transcriptome: the genome as the 'Book of Sand'. EMBO J *25*, 923-931.

Menheniott, T.R., Woodfine, K., Schulz, R., Wood, A.J., Monk, D., Giraud, A.S., Baldwin, H.S., Moore, G.E., and Oakey, R.J. (2008). Genomic imprinting of Dopa decarboxylase in heart and reciprocal allelic expression with neighboring Grb10. Mol Cell Biol *28*, 386-396.

Mercer, T.R., Dinger, M.E., and Mattick, J.S. (2009). Long non-coding RNAs: insights into functions. Nat Rev Genet *10*, 155-159.

Migeon, B.R. (2002). X chromosome inactivation: theme and variations. Cytogenet Genome Res *99*, 8-16.

Migeon, B.R., Lee, C.H., Chowdhury, A.K., and Carpenter, H. (2002). Species differences in TSIX/Tsix reveal the roles of these genes in X-chromosome inactivation. Am J Hum Genet *71*, 286-293.

Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.K., Koche, R.P.*, et al.* (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553-560.

Miranda, T.B., and Jones, P.A. (2007). DNA methylation: the nuts and bolts of repression. J Cell Physiol *213*, 384-390.

Mishima, K., Watanabe, T., Sasaki, T., Saito, I., and Takakura, K. (1990). [An infected partially thrombosed giant aneurysm of the azygos anterior cerebral artery]. No Shinkei Geka *18*, 475-481.

Mitsuya, K., Sui, H., Meguro, M., Kugoh, H., Jinno, Y., Niikawa, N., and Oshimura, M. (1997). Paternal expression of WT1 in human fibroblasts and lymphocytes. Hum Mol Genet *6*, 2243-2246.

Miyoshi, N., Kuroiwa, Y., Kohda, T., Shitara, H., Yonekawa, H., Kawabe, T., Hasegawa, H., Barton, S.C., Surani, M.A., Kaneko-Ishino, T., and Ishino, F. (1998). Identification of the Meg1/Grb10 imprinted gene on mouse proximal chromosome 11, a candidate for the Silver-Russell syndrome gene. Proc Natl Acad Sci U S A *95*, 1102-1107.

Miyoshi, N., Wagatsuma, H., Wakana, S., Shiroishi, T., Nomura, M., Aisaka, K., Kohda, T., Surani, M.A., Kaneko-Ishino, T., and Ishino, F. (2000). Identification of an imprinted gene, Meg3/Gtl2 and its human homologue MEG3, first mapped on mouse distal chromosome 12 and human chromosome 14q. Genes Cells *5*, 211-220.

Mizuno, Y., Sotomaru, Y., Katsuzawa, Y., Kono, T., Meguro, M., Oshimura, M., Kawai, J., Tomaru, Y., Kiyosawa, H., Nikaido, I.*, et al.* (2002). Asb4, Ata3, and Dcn are novel imprinted genes identified by high-throughput screening using RIKEN cDNA microarray. Biochem Biophys Res Commun *290*, 1499-1505.

Monk, D., Arnaud, P., Apostolidou, S., Hills, F.A., Kelsey, G., Stanier, P., Feil, R., and Moore, G.E. (2006). Limited evolutionary conservation of imprinting in the human placenta. Proc Natl Acad Sci U S A *103*, 6623-6628.

Monk, D., Wagschal, A., Arnaud, P., Muller, P.S., Parker-Katiraee, L., Bourc'his, D., Scherer, S.W., Feil, R., Stanier, P., and Moore, G.E. (2008). Comparative analysis of human chromosome 7q21 and mouse proximal chromosome 6 reveals a placental-specific imprinted gene, TFPI2/Tfpi2, which requires EHMT2 and EED for allelic-silencing. Genome Res *18*, 1270-1281.

Monk, M., Boubelik, M., and Lehnert, S. (1987). Temporal and regional changes in DNA methylation in the embryonic, extraembryonic and germ cell lineages during mouse embryo development. Development *99*, 371-382.

Moore, G.E., Abu-Amero, S.N., Bell, G., Wakeling, E.L., Kingsnorth, A., Stanier, P., Jauniaux, E., and Bennett, S.T. (2001). Evidence that insulin is imprinted in the human yolk sac. Diabetes *50*, 199-203.

Moore, T., Constancia, M., Zubair, M., Bailleul, B., Feil, R., Sasaki, H., and Reik, W. (1997). Multiple imprinted sense and antisense transcripts, differential methylation and tandem repeats in a putative imprinting control region upstream of mouse Igf2. Proc Natl Acad Sci U S A *94*, 12509-12514.

Moore, T., and Haig, D. (1991). Genomic imprinting in mammalian development: a parental tug-of-war. Trends Genet *7*, 45-49.

Moreira, A., Wollerton, M., Monks, J., and Proudfoot, N.J. (1995). Upstream sequence elements enhance poly(A) site efficiency of the C2 complement gene and are phylogenetically conserved. EMBO J *14*, 3809-3819.

Morison, I.M., Ramsay, J.P., and Spencer, H.G. (2005). A census of mammalian imprinting. Trends Genet *21*, 457-465.

Murakami, K., Oshimura, M., and Kugoh, H. (2007). Suggestive evidence for chromosomal localization of non-coding RNA from imprinted LIT1. J Hum Genet *52*, 926-933.

Murphy, S.K., Wylie, A.A., and Jirtle, R.L. (2001). Imprinting of PEG3, the human homologue of a mouse gene involved in nurturing behavior. Genomics *71*, 110-117.

Murrell, A. (2006). Genomic imprinting and cancer: from primordial germ cells to somatic cells. ScientificWorldJournal *6*, 1888-1910.

Nagano, T., Mitchell, J.A., Sanz, L.A., Pauler, F.M., Ferguson-Smith, A.C., Feil, R., and Fraser, P. (2008). The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. Science *322*, 1717-1720.

Nakabayashi, K., Bentley, L., Hitchins, M.P., Mitsuya, K., Meguro, M., Minagawa, S., Bamforth, J.S., Stanier, P., Preece, M., Weksberg, R.*, et al.* (2002). Identification and characterization of an imprinted antisense RNA (MESTIT1) in the human MEST locus on chromosome 7q32. Hum Mol Genet *11*, 1743-1756.

Nakabayashi, K., Makino, S., Minagawa, S., Smith, A.C., Bamforth, J.S., Stanier, P., Preece, M., Parker-Katiraee, L., Paton, T., Oshimura, M.*, et al.* (2004). Genomic imprinting of PPP1R9A encoding neurabin I in skeletal muscle and extra-embryonic tissues. J Med Genet *41*, 601-608.

Nakagawa, H., Chadwick, R.B., Peltomaki, P., Plass, C., Nakamura, Y., and de La Chapelle, A. (2001). Loss of imprinting of the insulin-like growth factor II gene occurs by biallelic methylation in a core region of H19-associated CTCF-binding sites in colorectal cancer. Proc Natl Acad Sci U S A *98*, 591-596.

Nakanishi, H., Suda, T., Katoh, M., Watanabe, A., Igishi, T., Kodani, M., Matsumoto, S., Nakamoto, M., Shigeoka, Y., Okabe, T.*, et al.* (2004). Loss of imprinting of PEG1/MEST in lung cancer cell lines. Oncol Rep *12*, 1273-1278.

Navarro, P., Chantalat, S., Foglio, M., Chureau, C., Vigneau, S., Clerc, P., Avner, P., and Rougeulle, C. (2009). A role for non-coding Tsix transcription in partitioning chromatin domains within the mouse X-inactivation centre. Epigenetics Chromatin *2*, 8.

Neumann, B., Kubicka, P., and Barlow, D.P. (1995). Characteristics of imprinted genes. Nat Genet *9*, 12-13.

Nicholls, R.D., and Knepper, J.L. (2001). Genome organization, function, and imprinting in Prader-Willi and Angelman syndromes. Annu Rev Genomics Hum Genet *2*, 153-175.

Niemitz, E.L., DeBaun, M.R., Fallon, J., Murakami, K., Kugoh, H., Oshimura, M., and Feinberg, A.P. (2004). Microdeletion of LIT1 in familial Beckwith-Wiedemann syndrome. Am J Hum Genet *75*, 844-849.

Niessen, R.C., Hofstra, R.M., Westers, H., Ligtenberg, M.J., Kooi, K., Jager, P.O., de Groote, M.L., Dijkhuizen, T., Olderode-Berends, M.J., Hollema, H.*, et al.* (2009). Germline hypermethylation of MLH1 and EPCAM deletions are a frequent cause of Lynch syndrome. Genes Chromosomes Cancer *48*, 737-744.

Nikaido, I., Saito, C., Mizuno, Y., Meguro, M., Bono, H., Kadomura, M., Kono, T., Morris, G.A., Lyons, P.A., Oshimura, M.*, et al.* (2003). Discovery of imprinted transcripts in the mouse transcriptome using large-scale expression profiling. Genome Res *13*, 1402-1409.

Ning, Y., Roschke, A., Christian, S.L., Lesser, J., Sutcliffe, J.S., and Ledbetter, D.H. (1996). Identification of a novel paternally expressed transcript adjacent to snRPN in the Prader-Willi syndrome critical region. Genome Res *6*, 742-746.

Nomura, T., Kimura, M., Horii, T., Morita, S., Soejima, H., Kudo, S., and Hatada, I. (2008). MeCP2-dependent repression of an imprinted miR-184 released by depolarization. Hum Mol Genet *17*, 1192-1199.

Okano, M., Bell, D.W., Haber, D.A., and Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell *99*, 247-257.

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H.*, et al.* (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature *420*, 563-573.

Okita, C., Meguro, M., Hoshiya, H., Haruta, M., Sakamoto, Y.K., and Oshimura, M. (2003). A new imprinted cluster on the human chromosome 7q21-q31, identified by human-mouse monochromosomal hybrids. Genomics *81*, 556-559.

Okutsu, T., Kuroiwa, Y., Kagitani, F., Kai, M., Aisaka, K., Tsutsumi, O., Kaneko, Y., Yokomori, K., Surani, M.A., Kohda, T.*, et al.* (2000). Expression and imprinting status of human PEG8/IGF2AS, a paternally expressed antisense transcript from the IGF2 locus, in Wilms' tumors. J Biochem *127*, 475-483.

Ono, R., Kobayashi, S., Wagatsuma, H., Aisaka, K., Kohda, T., Kaneko-Ishino, T., and Ishino, F. (2001). A retrotransposon-derived gene, PEG10, is a novel imprinted gene located on human chromosome 7q21. Genomics *73*, 232-237.

Ono, R., Shiura, H., Aburatani, H., Kohda, T., Kaneko-Ishino, T., and Ishino, F. (2003). Identification of a large novel imprinted gene cluster on mouse proximal chromosome 6. Genome Res *13*, 1696-1705.

Onyango, P., Miller, W., Lehoczky, J., Leung, C.T., Birren, B., Wheelan, S., Dewar, K., and Feinberg, A.P. (2000). Sequence and comparative analysis of the mouse 1-megabase region orthologous to the human 11p15 imprinted domain. Genome Res *10*, 1697-1710.

Ota, T., Suzuki, Y., Nishikawa, T., Otsuki, T., Sugiyama, T., Irie, R., Wakamatsu, A., Hayashi, K., Sato, H., Nagai, K.*, et al.* (2004). Complete sequencing and characterization of 21,243 full-length human cDNAs. Nat Genet *36*, 40-45.

Oudejans, C.B., Westerman, B., Wouters, D., Gooyer, S., Leegwater, P.A., van Wijk, I.J., and Sleutels, F. (2001). Allelic IGF2R repression does not correlate with expression of antisense RNA in human extraembryonic tissues. Genomics *73*, 331-337.

Pachnis, V., Brannan, C.I., and Tilghman, S.M. (1988). The structure and expression of a novel gene activated in early mouse embryogenesis. EMBO J *7*, 673-681.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet *40*, 1413-1415.

Pandey, R.R., Mondal, T., Mohammad, F., Enroth, S., Redrup, L., Komorowski, J., Nagano, T., Mancini-Dinardo, D., and Kanduri, C. (2008). Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. Mol Cell *32*, 232-246.

Pang, K.C., Frith, M.C., and Mattick, J.S. (2006). Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. Trends Genet *22*, 1-5.

Pant, P.V., Tao, H., Beilharz, E.J., Ballinger, D.G., Cox, D.R., and Frazer, K.A. (2006). Analysis of allelic differential expression in human white blood cells. Genome Res *16*, 331-339.

Paoloni-Giacobino, A. (2007). Epigenetics in reproductive medicine. Pediatr Res *61*, 51R-57R.

Paoloni-Giacobino, A., D'Aiuto, L., Cirio, M.C., Reinhart, B., and Chaillet, J.R. (2007). Conserved features of imprinted differentially methylated domains. Gene *399*, 33-45.

Parker-Katiraee, L., Carson, A.R., Yamada, T., Arnaud, P., Feil, R., Abu-Amero, S.N., Moore, G.E., Kaneda, M., Perry, G.H., Stone, A.C.*, et al.* (2007). Identification of the imprinted KLF14 transcription factor undergoing human-specific accelerated evolution. PLoS Genet *3*, e65.

Pastinen, T., Sladek, R., Gurd, S., Sammak, A., Ge, B., Lepage, P., Lavergne, K., Villeneuve, A., Gaudin, T., Brandstrom, H.*, et al.* (2004). A survey of genetic and epigenetic variation affecting human gene expression. Physiol Genomics *16*, 184-193.

Pauler, F.M., Koerner, M.V., and Barlow, D.P. (2007). Silencing by imprinted noncoding RNAs: is transcription the answer? Trends Genet *23*, 284-292.

Paulsen, M., Davies, K.R., Bowden, L.M., Villar, A.J., Franck, O., Fuermann, M., Dean, W.L., Moore, T.F., Rodrigues, N., Davies, K.E.*, et al.* (1998). Syntenic organization of the mouse distal chromosome 7 imprinting cluster and the Beckwith-Wiedemann syndrome region in chromosome 11p15.5. Hum Mol Genet *7*, 1149-1159.

Paulsen, M., El-Maarri, O., Engemann, S., Strodicke, M., Franck, O., Davies, K., Reinhardt, R., Reik, W., and Walter, J. (2000). Sequence conservation and variability of imprinting in the Beckwith-Wiedemann syndrome gene cluster in human and mouse. Hum Mol Genet *9*, 1829-1841.

Pearsall, R.S., Plass, C., Romano, M.A., Garrick, M.D., Shibata, H., Hayashizaki, Y., and Held, W.A. (1999). A direct repeat sequence at the Rasgrf1 locus and imprinted expression. Genomics *55*, 194-201.

Pedersen, I.S., Dervan, P., McGoldrick, A., Harrison, M., Ponchel, F., Speirs, V., Isaacs, J.D., Gorey, T., and McCann, A. (2002). Promoter switch: a novel mechanism causing biallelic PEG1/MEST expression in invasive breast cancer. Hum Mol Genet *11*, 1449-1453.

Pedersen, I.S., Dervan, P.A., Broderick, D., Harrison, M., Miller, N., Delany, E., O'Shea, D., Costello, P., McGoldrick, A., Keating, G.*, et al.* (1999). Frequent loss of imprinting of PEG1/MEST in invasive breast cancer. Cancer Res *59*, 5449-5451.

Penny, G.D., Kay, G.F., Sheardown, S.A., Rastan, S., and Brockdorff, N. (1996). Requirement for Xist in X chromosome inactivation. Nature *379*, 131-137.

Perez, D.S., Hoage, T.R., Pritchett, J.R., Ducharme-Smith, A.L., Halling, M.L., Ganapathiraju, S.C., Streng, P.S., and Smith, D.I. (2008). Long, abundantly expressed non-coding transcripts are altered in cancer. Hum Mol Genet *17*, 642-655.

Peters, J., Wroe, S.F., Wells, C.A., Miller, H.J., Bodle, D., Beechey, C.V., Williamson, C.M., and Kelsey, G. (1999). A cluster of oppositely imprinted transcripts at the Gnas locus in the distal imprinting region of mouse chromosome 2. Proc Natl Acad Sci U S A *96*, 3830-3835.

Piras, G., El Kharroubi, A., Kozlov, S., Escalante-Alcalde, D., Hernandez, L., Copeland, N.G., Gilbert, D.J., Jenkins, N.A., and Stewart, C.L. (2000). Zac1 (Lot1), a potential tumor suppressor gene, and the gene for epsilon-sarcoglycan are maternally imprinted genes: identification by a subtractive screen of novel uniparental fibroblast lines. Mol Cell Biol *20*, 3308-3315.

Plass, C., Shibata, H., Kalcheva, I., Mullins, L., Kotelevtseva, N., Mullins, J., Kato, R., Sasaki, H., Hirotsune, S., Okazaki, Y.*, et al.* (1996). Identification of Grf1 on mouse chromosome 9 as an imprinted gene by RLGS-M. Nat Genet *14*, 106-109.

Ponting, C.P., Oliver, P.L., and Reik, W. (2009). Evolution and functions of long noncoding RNAs. Cell *136*, 629-641.

Prasanth, K.V., and Spector, D.L. (2007). Eukaryotic regulatory RNAs: an answer to the 'genome complexity' conundrum. Genes Dev *21*, 11-42.

Prawitt, D., Enklaar, T., Klemm, G., Gartner, B., Spangenberg, C., Winterpacht, A., Higgins, M., Pelletier, J., and Zabel, B. (2000). Identification and characterization of MTR1, a novel gene with homology to melastatin (MLSN1) and the trp gene family located in the BWS-WT2 critical region on chromosome 11p15.5 and showing allele-specific expression. Hum Mol Genet *9*, 203-216.

Qian, N., Frank, D., O'Keefe, D., Dao, D., Zhao, L., Yuan, L., Wang, Q., Keating, M., Walsh, C., and Tycko, B. (1997). The IPL gene on chromosome 11p15.5 is imprinted in humans and mice and is similar to TDAG51, implicated in Fas expression and apoptosis. Hum Mol Genet *6*, 2021-2029.

Raefski, A.S., and O'Neill, M.J. (2005). Identification of a cluster of X-linked imprinted genes in mice. Nat Genet *37*, 620-624.

Rakyan, V.K., Down, T.A., Thorne, N.P., Flicek, P., Kulesha, E., Graf, S., Tomazou, E.M., Backdahl, L., Johnson, N., Herberth, M.*, et al.* (2008). An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs). Genome Res *18*, 1518-1529.

Ravasi, T., Suzuki, H., Pang, K.C., Katayama, S., Furuno, M., Okunishi, R., Fukuda, S., Ru, K., Frith, M.C., Gongora, M.M.*, et al.* (2006). Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. Genome Res *16*, 11-19.

Redrup, L., Branco, M.R., Perdeaux, E.R., Krueger, C., Lewis, A., Santos, F., Nagano, T., Cobb, B.S., Fraser, P., and Reik, W. (2009). The long noncoding RNA Kcnq1ot1 organises a lineage-specific nuclear domain for epigenetic gene silencing. Development *136*, 525-530.

Regha, K., Sloane, M.A., Huang, R., Pauler, F.M., Warczok, K.E., Melikant, B., Radolf, M., Martens, J.H., Schotta, G., Jenuwein, T., and Barlow, D.P. (2007). Active and repressive chromatin are interspersed without spreading in an imprinted gene cluster in the mammalian genome. Mol Cell *27*, 353-366.

Reinhart, B., Eljanne, M., and Chaillet, J.R. (2002). Shared role for differentially methylated domains of imprinted genes. Mol Cell Biol *22*, 2089-2098.

Reinhart, B., Paoloni-Giacobino, A., and Chaillet, J.R. (2006). Specific differentially methylated domain sequences direct the maintenance of methylation at imprinted genes. Mol Cell Biol *26*, 8347-8356.

Renfree, M.B., Hore, T.A., Shaw, G., Graves, J.A., and Pask, A.J. (2009). Evolution of genomic imprinting: insights from marsupials and monotremes. Annu Rev Genomics Hum Genet *10*, 241-262.

Riesewijk, A.M., Hu, L., Schulz, U., Tariverdian, G., Hoglund, P., Kere, J., Ropers, H.H., and Kalscheuer, V.M. (1997). Monoallelic expression of human PEG1/MEST is paralleled by parent-specific methylation in fetuses. Genomics *42*, 236-244.

Riesewijk, A.M., Schepens, M.T., Welch, T.R., van den Berg-Loonen, E.M., Mariman, E.M., Ropers, H.H., and Kalscheuer, V.M. (1996). Maternal-specific methylation of the human IGF2R gene is not accompanied by allele-specific transcription. Genomics *31*, 158-166.

Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E., and Chang, H.Y. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell *129*, 1311-1323.

Robertson, K.D. (2005). DNA methylation and human disease. Nat Rev Genet *6*, 597-610.

Robertson, K.D., and Wolffe, A.P. (2000). DNA methylation in health and disease. Nat Rev Genet *1*, 11-19.

Rougeulle, C., Cardoso, C., Fontes, M., Colleaux, L., and Lalande, M. (1998). An imprinted antisense RNA overlaps UBE3A and a second maternally expressed transcript. Nat Genet *19*, 15-16.

Royo, H., and Cavaille, J. (2008). Non-coding RNAs in imprinted gene clusters. Biol Cell *100*, 149-166.

Ruf, N., Bahring, S., Galetzka, D., Pliushch, G., Luft, F.C., Nurnberg, P., Haaf, T., Kelsey, G., and Zechner, U. (2007). Sequence-based bioinformatic prediction and QUASEP identify genomic imprinting of the KCNK9 potassium channel gene in mouse and human. Hum Mol Genet *16*, 2591-2599.

Russell, L.B. (1963). Mammalian X-chromosome action: inactivation limited in spread and region of origin. Science *140*, 976-978.

Sado, T., Wang, Z., Sasaki, H., and Li, E. (2001). Regulation of imprinted X-chromosome inactivation in mice by Tsix. Development *128*, 1275-1286.

Sakharkar, M.K., Chow, V.T., Ghosh, K., Chaturvedi, I., Lee, P.C., Bagavathi, S.P., Shapshak, P., Subbiah, S., and Kangueane, P. (2005). Computational prediction of SEG (single exon gene) function in humans. Front Biosci *10*, 1382-1395.

Sakharkar, M.K., Chow, V.T., and Kangueane, P. (2004). Distributions of exons and introns in the human genome. In Silico Biol *4*, 387-393.

Sandell, L.L., Guan, X.J., Ingram, R., and Tilghman, S.M. (2003). Gatm, a creatine synthesis enzyme, is imprinted in mouse placenta. Proc Natl Acad Sci U S A *100*, 4622-4627.

Sasaki, Y.T., Sano, M., Ideue, T., Kin, T., Asai, K., and Hirose, T. (2007). Identification and characterization of human non-coding RNAs with tissue-specific expression. Biochem Biophys Res Commun *357*, 991-996.

Sato, M., and Stryker, M.P. (2010). Genomic imprinting of experience-dependent cortical plasticity by the ubiquitin ligase gene Ube3a. Proc Natl Acad Sci U S A *107*, 5611-5616.

Schmidt, J.V., Levorse, J.M., and Tilghman, S.M. (1999). Enhancer competition between H19 and Igf2 does not mediate their imprinting. Proc Natl Acad Sci U S A *96*, 9733-9738.

Schmidt, J.V., Matteson, P.G., Jones, B.K., Guan, X.J., and Tilghman, S.M. (2000). The Dlk1 and Gtl2 genes are linked and reciprocally imprinted. Genes Dev *14*, 1997-2002.

Schule, B., Li, H.H., Fisch-Kohl, C., Purmann, C., and Francke, U. (2007). DLX5 and DLX6 expression is biallelic and not modulated by MeCP2 deficiency. Am J Hum Genet *81*, 492-506.

Schulz, R., McCole, R.B., Woodfine, K., Wood, A.J., Chahal, M., Monk, D., Moore, G.E., and Oakey, R.J. (2009). Transcript- and tissue-specific imprinting of a tumour suppressor gene. Hum Mol Genet *18*, 118-127.

Schulz, R., Menheniott, T.R., Woodfine, K., Wood, A.J., Choi, J.D., and Oakey, R.J. (2006). Chromosome-wide identification of novel imprinted genes using microarrays and uniparental disomies. Nucleic Acids Res *34*, e88.

Scott, R.J., and Spielman, M. (2006). Deeper into the maize: new insights into genomic imprinting in plants. Bioessays *28*, 1167-1171.

Seidl, C.I., Stricker, S.H., and Barlow, D.P. (2006). The imprinted Air ncRNA is an atypical RNAPII transcript that evades splicing and escapes nuclear export. EMBO J *25*, 3565-3575.

Seitz, H., Royo, H., Bortolin, M.L., Lin, S.P., Ferguson-Smith, A.C., and Cavaille, J. (2004). A large imprinted microRNA gene cluster at the mouse Dlk1-Gtl2 domain. Genome Res *14*, 1741-1748.

Seitz, H., Youngson, N., Lin, S.P., Dalbert, S., Paulsen, M., Bachellerie, J.P., Ferguson-Smith, A.C., and Cavaille, J. (2003). Imprinted microRNA genes transcribed antisense to a reciprocally imprinted retrotransposon-like gene. Nat Genet *34*, 261-262.

Semenov, D.V., Baryakin, D.N., Kamynina, T.P., Kuligina, E.V., and Richter, V.A. (2008). Fragments of noncoding RNA in plasma of human blood. Ann N Y Acad Sci *1137*, 130-134.

Serizawa, S., Miyamichi, K., Nakatani, H., Suzuki, M., Saito, M., Yoshihara, Y., and Sakano, H. (2003). Negative feedback regulation ensures the one receptor-one olfactory neuron rule in mouse. Science *302*, 2088-2094.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res *29*, 308-311.

Shiio, Y., and Eisenman, R.N. (2003). Histone sumoylation is associated with transcriptional repression. Proc Natl Acad Sci U S A *100*, 13225-13230.

Shin, J.Y., Fitzpatrick, G.V., and Higgins, M.J. (2008). Two distinct mechanisms of silencing by the KvDMR1 imprinting control region. EMBO J *27*, 168-178.

Shiura, H., Nakamura, K., Hikichi, T., Hino, T., Oda, K., Suzuki-Migishima, R., Kohda, T., Kaneko-ishino, T., and Ishino, F. (2009). Paternal deletion of Meg1/Grb10 DMR causes maternalization of the Meg1/Grb10 cluster in mouse proximal Chromosome 11 leading to severe pre- and postnatal growth retardation. Hum Mol Genet *18*, 1424-1438.

Shogren-Knaak, M., Ishii, H., Sun, J.M., Pazin, M.J., Davie, J.R., and Peterson, C.L. (2006). Histone H4-K16 acetylation controls chromatin structure and protein interactions. Science *311*, 844-847.

Silva, S.S., Rowntree, R.K., Mekhoubad, S., and Lee, J.T. (2008). X-chromosome inactivation and epigenetic fluidity in human embryonic stem cells. Proc Natl Acad Sci U S A *105*, 4820-4825.

Sleutels, F., Zwart, R., and Barlow, D.P. (2002). The non-coding Air RNA is required for silencing autosomal imprinted genes. Nature *415*, 810-813.

Smilinich, N.J., Day, C.D., Fitzpatrick, G.V., Caldwell, G.M., Lossie, A.C., Cooper, P.R., Smallwood, A.C., Joyce, J.A., Schofield, P.N., Reik, W., *et al.* (1999). A maternally methylated CpG island in KvLQT1 is associated with an antisense paternal transcript and loss of imprinting in Beckwith-Wiedemann syndrome. Proc Natl Acad Sci U S A *96*, 8064-8069.

Smith, R.J., Arnaud, P., Konfortova, G., Dean, W.L., Beechey, C.V., and Kelsey, G. (2002). The mouse Zac1 locus: basis for imprinting and comparison with human ZAC. Gene *292*, 101-112.

Smith, R.J., Dean, W., Konfortova, G., and Kelsey, G. (2003). Identification of novel imprinted genes in a genome-wide screen for maternal methylation. Genome Res *13*, 558-569.

Smrzka, O.W., Fae, I., Stoger, R., Kurzbauer, R., Fischer, G.F., Henn, T., Weith, A., and Barlow, D.P. (1995). Conservation of a maternal-specific methylation signal at the human IGF2R locus. Hum Mol Genet *4*, 1945-1952.

Sparago, A., Cerrato, F., Vernucci, M., Ferrero, G.B., Silengo, M.C., and Riccio, A. (2004). Microdeletions in the human H19 DMR result in loss of IGF2 imprinting and Beckwith-Wiedemann syndrome. Nat Genet *36*, 958-960.

Stricker, S.H., Steenpass, L., Pauler, F.M., Santoro, F., Latos, P.A., Huang, R., Koerner, M.V., Sloane, M.A., Warczok, K.E., and Barlow, D.P. (2008). Silencing and transcriptional properties of the imprinted Airn ncRNA are independent of the endogenous promoter. EMBO J *27*, 3116-3128.

Struhl, K. (2007). Transcriptional noise and the fidelity of initiation by RNA polymerase II. Nat Struct Mol Biol *14*, 103-105.

Sultan, M., Schulz, M.H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., *et al.* (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science *321*, 956-960.

Surani, M.A., Barton, S.C., and Norris, M.L. (1984). Development of reconstituted mouse eggs suggests imprinting of the genome during gametogenesis. Nature *308*, 548-550.

Sutcliffe, J.S., Nakao, M., Christian, S., Orstavik, K.H., Tommerup, N., Ledbetter, D.H., and Beaudet, A.L. (1994). Deletions of a differentially methylated CpG island at the SNRPN gene define a putative imprinting control region. Nat Genet *8*, 52-58.

Suzuki, M.M., and Bird, A. (2008). DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet *9*, 465-476.

Takai, D., and Jones, P.A. (2002). Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc Natl Acad Sci U S A *99*, 3740-3745.

Tam, O.H., Aravin, A.A., Stein, P., Girard, A., Murchison, E.P., Cheloufi, S., Hodges, E., Anger, M., Sachidanandam, R., Schultz, R.M., and Hannon, G.J. (2008).

Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. Nature *453*, 534-538.

Tanaka, K., Shiota, G., Meguro, M., Mitsuya, K., Oshimura, M., and Kawasaki, H. (2001). Loss of imprinting of long QT intronic transcript 1 in colorectal cancer. Oncology *60*, 268-273.

Thomson, T., and Lin, H. (2009). The biogenesis and function of PIWI proteins and piRNAs: progress and prospect. Annu Rev Cell Dev Biol *25*, 355-376.

Thorvaldsen, J.L., Duran, K.L., and Bartolomei, M.S. (1998). Deletion of the H19 differentially methylated domain results in loss of imprinted expression of H19 and Igf2. Genes Dev *12*, 3693-3702.

Thrash-Bingham, C.A., and Tartof, K.D. (1999). aHIF: a natural antisense transcript overexpressed in human renal cancer and during hypoxia. J Natl Cancer Inst *91*, 143-151.

Tierling, S., Gasparoni, G., Youngson, N., and Paulsen, M. (2009). The Begain gene marks the centromeric boundary of the imprinted region on mouse chromosome 12. Mamm Genome *20*, 699-710.

Trang, P., Weidhaas, J.B., and Slack, F.J. (2008). MicroRNAs as potential cancer therapeutics. Oncogene *27 Suppl 2*, S52-57.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.

Tremblay, K.D., Duran, K.L., and Bartolomei, M.S. (1997). A 5' 2-kilobase-pair region of the imprinted mouse H19 gene exhibits exclusive paternal methylation throughout development. Mol Cell Biol *17*, 4322-4329.

Tsai, C.E., Lin, S.P., Ito, M., Takagi, N., Takada, S., and Ferguson-Smith, A.C. (2002). Genomic imprinting contributes to thyroid hormone metabolism in the mouse embryo. Curr Biol *12*, 1221-1226.

Tufarelli, C., Stanley, J.A., Garrick, D., Sharpe, J.A., Ayyub, H., Wood, W.G., and Higgs, D.R. (2003). Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. Nat Genet *34*, 157-165.

Umlauf, D., Goto, Y., Cao, R., Cerqueira, F., Wagschal, A., Zhang, Y., and Feil, R. (2004). Imprinting along the Kcnq1 domain on mouse chromosome 7 involves repressive histone methylation and recruitment of Polycomb group complexes. Nat Genet *36*, 1296-1300.

Van Buggenhout, G., and Fryns, J.P. (2009). Angelman syndrome (AS, MIM 105830). Eur J Hum Genet *17*, 1367-1373.

Varmuza, S., and Mann, M. (1994). Genomic imprinting--defusing the ovarian time bomb. Trends Genet *10*, 118-123.

Waddington, C.H. (1942). Endeavour, Vol 1.

Walter, J., and Paulsen, M. (2003). The potential role of gene duplications in the evolution of imprinting mechanisms. Hum Mol Genet *12 Spec No 2*, R215-220.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008a). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470-476.

Wang, X., Sun, Q., McGrath, S.D., Mardis, E.R., Soloway, P.D., and Clark, A.G. (2008b). Transcriptome-wide identification of novel imprinted genes in neonatal mouse brain. PLoS One *3*, e3839.

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

Wang, Z., Schones, D.E., and Zhao, K. (2009b). Characterization of human epigenomes. Curr Opin Genet Dev *19*, 127-134.

Watanabe, T., Takeda, A., Tsukiyama, T., Mise, K., Okuno, T., Sasaki, H., Minami, N., and Imai, H. (2006). Identification and characterization of two novel classes of small RNAs in the mouse germline: retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes. Genes Dev *20*, 1732-1743.

Watanabe, T., Totoki, Y., Toyoda, A., Kaneda, M., Kuramochi-Miyagawa, S., Obata, Y., Chiba, H., Kohara, Y., Kono, T., Nakano, T.*, et al.* (2008). Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. Nature *453*, 539-543.

Watrin, F., Roeckel, N., Lacroix, L., Mignon, C., Mattei, M.G., Disteche, C., and Muscatelli, F. (1997). The mouse Necdin gene is expressed from the paternal allele only and lies in the 7C region of the mouse chromosome 7, a region of conserved synteny to the human Prader-Willi syndrome region. Eur J Hum Genet *5*, 324-332.

Weile, C., Gardner, P.P., Hedegaard, M.M., and Vinther, J. (2007). Use of tiling array data and RNA secondary structure predictions to identify noncoding RNA genes. BMC Genomics *8*, 244.

Weisstein, A.E., Feldman, M.W., and Spencer, H.G. (2002). Evolutionary genetic models of the ovarian time bomb hypothesis for the evolution of genomic imprinting. Genetics *162*, 425-439.

Weksberg, R., Shuman, C., and Beckwith, J.B. (2010). Beckwith-Wiedemann syndrome. Eur J Hum Genet *18*, 8-14.

Wevrick, R., and Francke, U. (1997). An imprinted mouse transcript homologous to the human imprinted in Prader-Willi syndrome (IPW) gene. Hum Mol Genet *6*, 325-332.

Wevrick, R., Kerns, J.A., and Francke, U. (1994). Identification of a novel paternally expressed gene in the Prader-Willi syndrome region. Hum Mol Genet *3*, 1877-1882.

Williams, A.E., Moschos, S.A., Perry, M.M., Barnes, P.J., and Lindsay, M.A. (2007). Maternally imprinted microRNAs are differentially expressed during mouse and human lung development. Dev Dyn *236*, 572-580.

Williamson, C.M., Ball, S.T., Nottingham, W.T., Skinner, J.A., Plagge, A., Turner, M.D., Powles, N., Hough, T., Papworth, D., Fraser, W.D.*, et al.* (2004). A cis-acting control region is required exclusively for the tissue-specific imprinting of Gnas. Nat Genet *36*, 894-899.

Williamson, C.M., Schofield, J., Dutton, E.R., Seymour, A., Beechey, C.V., Edwards, Y.H., and Peters, J. (1996). Glomerular-specific imprinting of the mouse gsalpha gene: how does this relate to hormone resistance in albright hereditary osteodystrophy? Genomics *36*, 280-287.

Williamson, C.M., Turner, M.D., Ball, S.T., Nottingham, W.T., Glenister, P., Fray, M., Tymowska-Lalanne, Z., Plagge, A., Powles-Glover, N., Kelsey, G.*, et al.* (2006). Identification of an imprinting control region affecting the expression of all transcripts in the Gnas cluster. Nat Genet *38*, 350-355.

Wilusz, J.E., Freier, S.M., and Spector, D.L. (2008). 3' end processing of a long nuclear-retained noncoding RNA yields a tRNA-like cytoplasmic RNA. Cell *135*, 919-932.

Wilusz, J.E., Sunwoo, H., and Spector, D.L. (2009). Long noncoding RNAs: functional surprises from the RNA world. Genes Dev *23*, 1494-1504.

Woo, C.J., and Kingston, R.E. (2007). HOTAIR lifts noncoding RNAs to new levels. Cell *129*, 1257-1259.

Wood, A.J., Bourc'his, D., Bestor, T.H., and Oakey, R.J. (2007a). Allele-specific demethylation at an imprinted mammalian promoter. Nucleic Acids Res *35*, 7031-7039.

Wood, A.J., Roberts, R.G., Monk, D., Moore, G.E., Schulz, R., and Oakey, R.J. (2007b). A screen for retrotransposed imprinted genes reveals an association between X chromosome homology and maternal germ-line methylation. PLoS Genet *3*, e20.

Wood, A.J., Schulz, R., Woodfine, K., Koltowska, K., Beechey, C.V., Peters, J., Bourc'his, D., and Oakey, R.J. (2008). Regulation of alternative polyadenylation by genomic imprinting. Genes Dev *22*, 1141-1146.

Woodcock, D.M., Lawler, C.B., Linsenmeyer, M.E., Doherty, J.P., and Warren, W.D. (1997). Asymmetric methylation in the hypermethylated CpG promoter region of the human L1 retrotransposon. J Biol Chem *272*, 7810-7816.

Wroe, S.F., Kelsey, G., Skinner, J.A., Bodle, D., Ball, S.T., Beechey, C.V., Peters, J., and Williamson, C.M. (2000). An imprinted transcript, antisense to Nesp, adds complexity to the cluster of imprinted genes at the mouse Gnas locus. Proc Natl Acad Sci U S A *97*, 3342-3346.

Wutz, A., and Gribnau, J. (2007). X inactivation Xplained. Curr Opin Genet Dev *17*, 387-393.

Wutz, A., and Jaenisch, R. (2000). A shift from reversible to irreversible X inactivation is triggered during ES cell differentiation. Mol Cell *5*, 695-705.

Wutz, A., Smrzka, O.W., Schweifer, N., Schellander, K., Wagner, E.F., and Barlow, D.P. (1997). Imprinted expression of the Igf2r gene depends on an intronic CpG island. Nature *389*, 745-749.

Wutz, A., Theussl, H.C., Dausman, J., Jaenisch, R., Barlow, D.P., and Wagner, E.F. (2001). Non-imprinted Igf2r expression decreases growth and rescues the Tme mutation in mice. Development *128*, 1881-1887.

Wylie, A.A., Murphy, S.K., Orton, T.C., and Jirtle, R.L. (2000). Novel imprinted DLK1/GTL2 domain on human chromosome 14 contains motifs that mimic those implicated in IGF2/H19 regulation. Genome Res *10*, 1711-1718.

Xin, Z., Soejima, H., Higashimoto, K., Yatsuki, H., Zhu, X., Satoh, Y., Masaki, Z., Kaneko, Y., Jinno, Y., Fukuzawa, R.*, et al.* (2000). A novel imprinted gene, KCNQ1DN, within the WT2 critical region of human chromosome 11p15.5 and its reduced expression in Wilms' tumors. J Biochem *128*, 847-853.

Xu, M., You, Y., Hunsicker, P., Hori, T., Small, C., Griswold, M.D., and Hecht, N.B. (2008). Mice deficient for a small cluster of Piwi-interacting RNAs implicate Piwi-interacting RNAs in transposon control. Biol Reprod *79*, 51-57.

Xu, Y., Goodyer, C.G., Deal, C., and Polychronakos, C. (1993). Functional polymorphism in the parental imprinting of the human IGF2R gene. Biochem Biophys Res Commun *197*, 747-754.

Xu, Y.Q., Grundy, P., and Polychronakos, C. (1997). Aberrant imprinting of the insulin-like growth factor II receptor gene in Wilms' tumor. Oncogene *14*, 1041-1046.

Xue, C., and Li, F. (2008). Finding noncoding RNA transcripts from low abundance expressed sequence tags. Cell Res *18*, 695-700.

Yamada, K., Kano, J., Tsunoda, H., Yoshikawa, H., Okubo, C., Ishiyama, T., and Noguchi, M. (2006). Phenotypic characterization of endometrial stromal sarcoma of the uterus. Cancer Sci *97*, 106-112.

Yamasaki, K., Hayashida, S., Miura, K., Masuzaki, H., Ishimaru, T., Niikawa, N., and Kishino, T. (2000). The novel gene, gamma2-COP (COPG2), in the 7q32 imprinted domain escapes genomic imprinting. Genomics *68*, 330-335.

Yamasaki, Y., Kayashima, T., Soejima, H., Kinoshita, A., Yoshiura, K., Matsumoto, N., Ohta, T., Urano, T., Masuzaki, H., Ishimaru, T.*, et al.* (2005). Neuron-specific relaxation of Igf2r imprinting is associated with neuron-specific histone modifications and lack of its antisense transcript Air. Hum Mol Genet *14*, 2511-2520.

Yang, T., Adamson, T.E., Resnick, J.L., Leff, S., Wevrick, R., Francke, U., Jenkins, N.A., Copeland, N.G., and Brannan, C.I. (1998). A mouse model for Prader-Willi syndrome imprinting-centre mutations. Nat Genet *19*, 25-31.

Yoder, J.A., Walsh, C.P., and Bestor, T.H. (1997). Cytosine methylation and the ecology of intragenomic parasites. Trends Genet *13*, 335-340.

Yoon, B., Herman, H., Hu, B., Park, Y.J., Lindroth, A., Bell, A., West, A.G., Chang, Y., Stablewski, A., Piel, J.C.*, et al.* (2005). Rasgrf1 imprinting is regulated by a CTCF-dependent methylation-sensitive enhancer blocker. Mol Cell Biol *25*, 11184-11190.

Yotova, I.Y., Vlatkovic, I.M., Pauler, F.M., Warczok, K.E., Ambros, P.F., Oshimura, M., Theussl, H.C., Gessler, M., Wagner, E.F., and Barlow, D.P. (2008). Identification of the human homolog of the imprinted mouse Air non-coding RNA. Genomics *92*, 464-473.

Yu, W., Gius, D., Onyango, P., Muldoon-Jacobs, K., Karp, J., Feinberg, A.P., and Cui, H. (2008). Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA. Nature *451*, 202-206.

Yu, Y., Xu, F., Peng, H., Fang, X., Zhao, S., Li, Y., Cuevas, B., Kuo, W.L., Gray, J.W., Siciliano, M.*, et al.* (1999). NOEY2 (ARHI), an imprinted putative tumor suppressor gene in ovarian and breast carcinomas. Proc Natl Acad Sci U S A *96*, 214-219.

Yuan, E., Li, C.M., Yamashiro, D.J., Kandel, J., Thaker, H., Murty, V.V., and Tycko, B. (2005). Genomic profiling maps loss of heterozygosity and defines the timing and stage dependence of epigenetic and genetic events in Wilms' tumors. Mol Cancer Res *3*, 493-502.

Yuan, J., Luo, R.Z., Fujii, S., Wang, L., Hu, W., Andreeff, M., Pan, Y., Kadota, M., Oshimura, M., Sahin, A.A.*, et al.* (2003). Aberrant methylation and silencing of ARHI, an imprinted tumor suppressor gene in which the function is lost in breast cancers. Cancer Res *63*, 4174-4180.

Zeng, Y., and Cullen, B.R. (2002). RNA interference in human cells is restricted to the cytoplasm. RNA *8*, 855-860.

Zeschnigk, M., Schmitz, B., Dittrich, B., Buiting, K., Horsthemke, B., and Doerfler, W. (1997). Imprinted segments in the human genome: different DNA methylation patterns in the Prader-Willi/Angelman syndrome region as determined by the genomic sequencing method. Hum Mol Genet *6*, 387-395.

Zhang, Y., and Tycko, B. (1992). Monoallelic expression of the human H19 gene. Nat Genet *1*, 40-44.

Zhang, Z., Joh, K., Yatsuki, H., Wang, Y., Arai, Y., Soejima, H., Higashimoto, K., Iwasaka, T., and Mukai, T. (2006). Comparative analyses of genomic imprinting and CpG island-methylation in mouse Murr1 and human MURR1 loci revealed a putative imprinting control region in mice. Gene *366*, 77-86.

Zhao, J., Hyman, L., and Moore, C. (1999). Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. Microbiol Mol Biol Rev *63*, 405-445.

Zwart, R., Sleutels, F., Wutz, A., Schinkel, A.H., and Barlow, D.P. (2001). Bidirectional action of the Igf2r imprint control element on upstream and downstream imprinted genes. Genes Dev *15*, 2361-2366.

7. Curriculum vitae

Name:	Irena Mihaila Vlatkovic
Maiden name:	Selakovic
Address:	Juchgasse 40/25, 1030 Vienna, Austria
Permanent address:	4/10 Cerova St., 11000 Belgrade, Serbia
E-Mail:	irena.vlatkovic@univie.ac.at ivlatkovic@cemm.oeaw.ac.at
Date of Birth:	26.08.1978
Place of Birth:	Valjevo, Serbia, Yugoslavia
Citizenship:	Serbian

Education

High school: 1993-1997 Specialization to nature sciences and mathematics, Valjevo Grammar School, Serbia

Summer project**:** "Frequency of various structural types of inversions in *Drosophila subobscura*", Supervisor: Dr. Tatjana Terzic-Savic, Institute for Biological Research "Sinisa Stankovic", Belgrade, Serbia, 1995.

University: 1997-2006 Diploma degree in Molecular Biology and Physiology, Faculty of Biology, University of Belgrade, Serbia

Summer project: "Towards elucidating the function of the carboxyl terminus tail of ErbB-3 receptor tyrosine kinase", Supervisor: Yosef Yarden Ph.D., Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel, 1997.

Diploma thesis: "Analysis of genomic instability in human lung tumours by DNA profiling", Supervisor: Dr. Sabera Ruzdijic, Department of Neurobiology, Institute for Biological Research "Sinisa Stankovic", Belgrade, Serbia, 2005-2006.

Doctoral study: 2006-2010 Department of Molecular Biology, Doctoral Studies of Natural Sciences, Vienna University, Austria

Employment

2006-2010 PhD student in the lab of Denise Barlow, CeMM, Research Center for Molecular Medicine of the Austrian Academy of Sciences, from 2006 to Jun 2010 in Dr. Bohr-Gasse 9/4, A-1030 Vienna and from Jun 2010 in Lazarettgasse 14, AKH BT 25.3, A-1090 Vienna, http://www.cemm.oeaw.ac.at/

Meetings

EHA Scientific Workshop: The role of Epigenetics in Hematological Malignancies, February 2007, Mandieleu, France

The Epigenome Network of Excellence 3rd Annual Meeting, June 2007, Stockholm, Sweden (Poster presentation)

EMBO Molecular Medicine Workshop: Drug Action& Chemical Biology, August 2007, Vienna, Austria

RNA 2008, August 2008, Berlin, Germany (Poster presentation)

Keystone Symposia: Epigenetics, Development and Human Disease, January 2009, Breckenridge, Colorado, USA (Poster presentation)

EMBO Workshop: New functions of Regulatory RNAs in Pro- and Eukaryotes, January 2009, Vienna, Austria

European Human Genetics Conference, May 2009, Vienna, Austria (Poster presentation)

The EMBO meeting, August 2009, Amsterdam, Netherlands (Poster presentation)

Epigenome NoE meeting: Epigenetic Regulation in Cell Fate & Disease, March 2010, Vienna, Austria (Poster presentation)

Honours, awards

3rd place on Republic Biology Competition, Gornji Milanovac, Serbia, 1995.
1st place on International Young's Bio-Olympiad, Sankt-Petersburg, Russia, 1996.
2nd place on Republic Biology Competition, Uzice, Serbia, 1997.

Publications

Vlatkovic IM, Pauler FM, Huang R, Santoro F, Guenzl P, Tamir I, Sommer A, Barlow D. (in preparation). Identification of macro ncRNAs in human imprinted gene regions by RNA-Chip and RNA-seq.

Durand C, Roth R, Vlatkovic I, Dweeph H, Decker E, Schneider K, Rappold G, (submitted, August 2010). Alternative splicing and nonsense-mediated RNA decay contribute to the regulation of *SHOX* expression.

Yotova I, Vlatkovic IM, Pauler FM, Warczok KE, Ambros PF, Oshimura M, Theussl HC, Gessler M, Wagner EF, Barlow DP. 2008, Identification of the human homologue of the imprinted mouse Air non-coding RNA, Genomics, 92, 464-473

Selakovic I. Treiber N. 1997, Towards Elucidating the Function of the Carboxyl-terminus Tail of ErbB-3 Receptor Tyrosine Kinase, Scientific Works. Dr. Bessie Lawrence 29th International Summer Science Institute. The Weizmann Institute of Science, Rehovot, Israel

8. Lebenslauf

Name:	Irena Mihaila Vlatkovic
Geburtsname:	Selakovic
Adresse:	Juchgasse 40/25, 1030 Wien, Österreich
Permanente Adresse:	4/10 Cerova St., 11000 Belgrad, Serbien
E-Mail:	irena.vlatkovic@univie.ac.at ivlatkovic@cemm.oeaw.ac.at
Geburtsdatum:	26.08.1978
Geburtsort:	Valjevo, Serbien, Jugoslawien
Nationalität:	Serbisch

Ausbildung

Mittelschule: 1993-1997 mit Spezialisierung in Naturwissenschaften & Mathematik Valjevo Gymnasium, Serbien
Sommerprojekt**:** “Frequency of various structural types of inversions in *Drosophila subobscura*”, Betreuer: Dr. Tatjana Terzic-Savic, Institute for Biological Research “Sinisa Stankovic”, Belgrad, Serbien, 1995.

Universität: 1997-2006 Diplom in Molekularbiologie und -physiologie, Faculty of Biology, Universität Belgrad, Serbien
Sommerprojekt: “Towards elucidating the function of the carboxyl terminus tail of ErbB-3 receptor tyrosine kinase”, Betreuer: Yosef Yarden Ph.D., Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel, 1997.
Diplomarbeit: “Analysis of genomic instability in human lung tumours by DNA profiling”, Betreuer: Dr. Sabera Ruzdijic, Department of Neurobiology, Institute for Biological Research “Sinisa Stankovic”, Belgrad, Serbien, 2005-2006.

Doktoratsstudium: 2006-2010 Department für Molekularbiologie, Doktoratsstudium der Naturwissenschaften, Universität Wien, Österreich

Anstellungen

2006-2010 PhD Student im Labor von Prof. Denise Barlow, CeMM, Forschungszentrum für molekulare Medizin der österreichischen Akademie der Wissenschaften, von 2006 bis Juni 2010 in Dr. Bohr-Gasse 9/4, A-1030 Wien und ab Juni 2010 in Lazarettgasse 14, AKH-BT25.3, A-1090 Wien

Meetings

EHA Scientific Workshop: The role of Epigenetics in Hematological Malignancies, Februar 2007, Mandieleu, France

The Epigenome Network of Excellence 3rd Annual Meeting, Juni 2007, Stockholm, Schweden (Posterpräsentation)

EMBO Molecular Medicine Workshop: Drug Action & Chemical Biology, August 2007, Wien, Österreich

RNA 2008, August 2008, Berlin, Deutschland (Posterpräsentation)

Keystone Symposia: Epigenetics, Development and Human Disease, Jänner 2009, Breckenridge, Colorado, USA (Posterpräsentation)

EMBO Workshop: New functions of Regulatory RNAs in Pro- and Eukaryotes, Jänner 2009, Wien, Österreich

European Human Genetics Conference, Mai 2009, Wien, Österreich (Posterpräsentation)

The EMBO meeting, August 2009, Amsterdam, Niederlande (Posterpräsentation)

Epigenome NoE meeting: Epigenetic Regulation in Cell Fate & Disease, März 2010, Wien, Österreich (Posterpräsentation)

Ehrungen & Awards

3. Platz beim Republik Biologie Wettbewerb, Gornji Milanovac, Serbien, 1995.

1. Platz bei der Internationalen Nachwuchs Bio-Olympiade, Sankt-Petersburg, Russland, 1996.

2. Platz beim Republik Biologie Wettbewerb, Uzice, Serbien, 1997.

Publikationen

Vlatkovic IM, Pauler FM, Huang R, Santoro F, Guenzl P, Tamir I, Sommer A, Barlow D. (in preparation). Identification of macro ncRNAs in human imprinted gene regions by RNA-Chip and RNA-seq

Durand C, Roth R, Vlatkovic I, Dweeph H, Decker E, Schneider K, Rappold G, (submitted, August 2010). Alternative splicing and nonsense-mediated RNA decay contribute to the regulation of *SHOX* expression.

Yotova I, Vlatkovic IM, Pauler FM, Warczok KE, Ambros PF, Oshimura M, Theussl HC, Gessler M, Wagner EF, Barlow DP. 2008, Identification of the human homologue of the imprinted mouse Air non-coding RNA, Genomics, 92, 464-473

Selakovic I. Treiber N. 1997, Towards Elucidating the Function of the Carboxyl-terminus Tail of ErbB-3 Receptor Tyrosine Kinase, Scientific Works. Dr. Bessie Lawrence 29th International Summer Science Institute. The Weizmann Institute of Science, Rehovot, Israel

9. Acknowledgements

The data presented in this thesis could not have been obtained without the help from a team of people that I would like to acknowledge. First I would like to tank to Dr. Denise Barlow, my supervisor that gave me the constant scientific support and great resources for work. With her superb enthusiasm about the science and talent for lab organization she was and she is a great boss.

These four years in the Barlow lab were really fulfilling with the gain of knowledge through number of seminars and meetings that we attended, a lot of experiments, excitements and numerous great people that I met. I would like especially to thank Dr. Iveta Yotova with whom I worked for the first few months of my PhD and published the *AIRN* story. Iveta taught me really a lot and always had a time to answer my questions. I would also like to thank a lot to Ru Huang and Florian Pauler that worked with me on parts of the macro ncRNAs project presented in this thesis. With Ru I spent so much time in discussing about differences in imprinting between human and mouse and why some experiment worked or did not work… Thank you Ru! With Florian every computer problem did find it’s solution and many ideas were challenged… Thank you Florian, and hopefully we will collaborate more to finish NORBERT... Federica Santoro performed a nuclear vs. cytoplasmic fractionation and was really great all of this time. Quanah Hudson did an absolutely great job by reading and re-reading this thesis and Philipp Günzl was great help for the German parts. Also, I worked with Philipp on other imprinting and cancer stories and really enjoyed the time we spent together. I thank as well Tomek, Martha, Romana, Kasia, Ruth, Simon as well as previous lab members: Basak, Mathew, Stephan, Paulina and Lukas for ideas, helping me with techniques, for doing many basic lab stuff, or “just” for being friends. I thank to Angelika and Xenia that made my administrative life much easier. Thank to CeMM, since the spirit of the institute made me fell like a part of something bigger. Big thanks to Renping Qiao for the enormous friend support...

I express my gratitude to the bioinformatics guys: Ido Tamir and Andy Sommer from GENAU and Jan Engelhart from Stadlers’ lab for doing a lot for HIRTA and rRNA-seq analysis and especially for teaching me some informatics stuff. I got a great help from Ashot Harutyunyan, Kralovics’ lab who did a great job by teaching me “all about SNPs and Affymetrix”. I would also thank to Helia Schonthaler from Wagners’ lab for human ES cell differentiation and Claudia Durand for collaboration on the *SHOX* gene.

My PhD committee: Dr. Robert Kralovics, CeMM and Dr. Silke Dorner, MFPL did a great job supporting me on every step, from ideas about the project, samples and materials for the work, to the stories about the future.

I thank to my parents for their maximal support and constant interest in how the life and the project goes. The biggest special thanks goes to my husband who had such patience, constant understanding and who maximally supported me during these four years.