Title:

Diverse regulatory functions of long non-coding RNAs

Inaugural-Dissertation

to obtain the academic degree

Doctor rerum naturalium (Dr. rer. nat.)

submitted to the Department of Biology, Chemistry and Pharmacy

of Freie Universität Berlin

by

Dubravka Vučićević

From Serbia

Berlin, 2015

The present work was carried out from April 2012 until September 2015 at the Max Planck Institute for Molecular Genetics under the supervision of Dr. Ulf Andersson Ørom.

1st Reviewer: Dr. Ulf Andersson Ørom

2ndReviewer: Prof. Dr. Christian Freund

Date of Defense: 15.12.2015

Мојој мајци

Acknowledgements

First of all, I would like to thank Dr. Ulf Anderson Ørom for making me a part of a stimulating scientific environment and for giving me the opportunity to work on exciting research projects. Also, I would like to thank him for guiding and supporting me, for pushing me out of my comfort zone and investing time in teaching me the art of science. I would also like to thank Prof. Dr. Christian Freund for reviewing this thesis.

I thank the members of Ørom lab, my scientific family: Evgenia Ntini, Anne Musahl, Thomas Conrad, Annita Louloupi, Julia Liz, Alexander Kiefer, Maja Gehre, Antonia Hilbig and Masha Kenda, for amazing scientific discussions and the great times we had in and out of the lab. For their advices, support and true friendship. It was an honor and pleasure to be working with them.

I would also like to thank my collaborators for their contributions: Evgenia Ntini, Ruping Sun, David Meierhofer and Sascha Sauer (Max Planck Institute for Molecular Genetica); Olivia Corradin and Peter C Scacheri (Case Western Reserve University Clevelan); Sonam Dhamija (Medizinische Hochschule Hannover); Lennart Friis-Hansen (Næstved hospital Denmark); and all of the people working at the MPIMG who constructively criticized my work.

Thanks also goes to all of my friends for their support, friendship and for making my doctoral studies enjoyable.

I would also like to thank my parents, especially my mom Aneta. She taught me science at a young age and encouraged me to ask questions and be curious. Without her my scientific career would not be possible. I thank her for the devotion, support and endless love. Воли те Чеда.

Table of Contents

SUMMARY 1

ZUSAMMENFASSUNG 3

1 INTRODUCTION 5

1.1 Human genome 5

1.1.1 Organization of the human genome 5

1.1.2 Genetic regulatory elements 6

1.2 Enhancer predictions 8

1.3 Long non-coding RNAs 12

1.3.1 Characteristics of long ncRNAs 12

1.3.2 Diverse functions of long ncRNAs 14

1. 4 Enhancer derived long ncRNAs 17

1.4.1 Characteristics of enhancer derived long ncRNAs 17

1.4.2 Functional enhancer derived long ncRNAs 18

1.4.3 Mechanism of action of enhancer derived long ncRNAs 19

1.5 MYC oncogene 21

1.6 Aim of the thesis 24

2 MATERIALS AND METHODS 25

2.1 Materials 25

2.1.1 Instrumentation 25

2.1.2. Consumables 26

2.1.3 Chemicals 26

2.1.4 Buffers, Solutions and Media 28

2.1.5 Molecular biology kits 30

2.1.6 Enzymes 30

2.1.7 Plasmids 30

2.1.8 Antibodies 31

2.1.9 Oligonucleotides 31

2.1.10 Cell lines and bacteria 32

2.1.11 Software 32

2.2 Methods 33

2.2.1 *In silico* methods 33

2.2.1.1 Enhancer prediction 33

2.2.1.2 Overlap between long ncRNAs and enhancers 33

2.2.1.3 Pol II association analysis 34

2.2.1.4 Expression analysis of long ncRNAs 34
2.2.1.5 Determining H3K4me1 and H3K4me3 levels at selected enhancers 34
2.2.1.6 RNA-sequencing data analysis 34
2.2.1.7 Mass-spectrometry data analysis 34
2.2.1.8 Gene ontology analysis 35
2.2.1.9 Statistical analysis 35
2.2.2 DNA/RNA molecular biology methods 36
2.2.2.1 Polymerase chain reaction 36
2.2.2.2 Agarose gel electrophoresis 36
2.2.2.3 Extraction of DNA from agarose gels 37
2.2.2.4 Determination of nucleic acid concentration and purity 37
2.2.2.5 Topo TA cloning 37
2.2.2.6 Restrictional digestion 38
2.2.2.7 Dephosphorilation of digested plasmids 38
2.2.2.8 Ligation 38
2.2.2.9 Generation of competent *E. coli* bacteria 38
2.2.2.10 Transformation of *E. coli* bacteria 38
2.2.2.11 Colony PCR 39
2.2.2.12 Plasmid extraction 39
2.2.2.13 Sanger sequencing 39
2.2.2.14 Extraction of RNA from human cells 39
2.2.2.15 DNAse I treatment of RNA 39
2.2.2.16 Reverse transcription 40
2.2.2.17 Quantitative real-time PCR 40
2.2.2.18 RNA sequencing 41
2.2.3 Cell culture methods 41
2.2.3.1 Culturing of human cell lines 41
2.2.3.2 Transfection of human cell lines 41
2.2.2.3 Luciferase assay 42
2.2.3.4 Fractionation of cells 42
2.2.3.5 SILAC labeling of the cells 43
2.2.3.6 Preparation of protein cellular extracts 43
2.2.3.7 Migration assay 43
2.2.3.8 Crystal violet cellular viability assay 43
2.2.3.9 Sucrose gradient fractionation of polysomes- polysome profiling 44
2.2.4 Protein biochemical methods 44
2.2.4.1 Determination of protein concentration 44
2.2.4.2 SDS polyacrylamide gel electrophoresis 45
2.2.4.3 Western blot 45
2.2.4.4 Puromycin translation assay 45

2.2.4.5 *In vitro* translation assay 46
2.2.4.6 Mass-spectrometry 46
2.2.4.7 Caspase3/7 activity assay 47
2.2.4.8 Cell cycle analysis by flow cytometry 47
2.2.4.9 Apoptosis analysis by flow cytometry 47

3 RESULTS **49**

3.1 Long ncRNA expression associates with tissue-specific enhancers 50
3.1.1 PreSTIGE enhancer-target prediction approach 50
3.1.2 Long ncRNAs overlapping PreSTIGE predicted enhancers 52
3.1.3 Comparison of PreSTIGE to other enhancer prediction methods 56
3.1.4 Epigenetic profile of transcribed enhancers 57
3.1.5 Experimental validation of the approached used in the study 59

3.2 The long non-coding RNA PARROT is an upstream regulator of c-Myc and affects proliferation and translation 60
3.2.1 Characterization of the long ncRNAs PARROT 60
3.2.2 Biological functions of the long ncRNA PARROT 64
3.2.3 PARROT's molecular mechanism of action 66
3.2.4 PARROT in senescence and cancer 69

4 DISCUSSION **73**

4.1 Long ncRNA expression associates with tissue-specific enhancers 73
4.1.1 Predicting enhancers and their targets 73
4.1.2 Long ncRNAs overlapping PreSTIGE predicted enhancers 75
4.1.3 PreSTIGE compared to other methods 78
4.1.4 Characteristic epigenetic mark of enhancers transcribed into long ncRNAs 79

4.2 The long non-coding RNA PARROT is an upstream regulator of c-Myc and affects proliferation and translation 80
4.2.1 Characterisation of the long ncRNA PARROT 80
4.2.2. Biological function of the long ncRNA PARROT 82
4.2.3 PARROT in senescence and cancer 83

4.3 Outlook 84
4.3.1 Long ncRNA expression associates with tissue-specific enhancers 84
4.3.2 The long non-coding RNA PARROT is an upstream regulator of c-Myc and affects proliferation and translation 86

REFERENCES **89**

APPENDIX **105**

ABBREVIATIONS **121**

PUBLICATIONS **123**

SUMMARY

Long non-coding RNAs (ncRNAs) control almost every level of the gene expression program adding an unexpected layer of complexity in the regulation of gene expression. They have been shown to control fundamental biological processes such as X chromosome inactivation, imprinting, proliferation, development and differentiation. Furthermore, they are involved in the development of a wide variety of human disorders such as cancer and neurodegenerative disorders.

Recent research revealed that some long ncRNAs are expressed from a subset of enhancers and are required for mediating their function. However, the extent to which long ncRNAs are required for enhancer function is still unknown. Additionally, although enhancers have been studied for decades, there is still no consensus on how to predict tissue-specific enhancers. In this thesis we employed a recently developed methodology- PreSTIGE to predict tissue specific enhancers and their targets based on the tissue specific presence of H3K4me1 marks and tissue specific gene expression. We find that 28 % (2,695) of all ENCODE annotated long ncRNAs overlap tissue-specific enhancers predicted by PreSTIGE. The expression of enhancer overlapping long ncRNAs is significantly higher in a tissue in which an overlapping enhancer is predicted to be active suggesting that some enhancers might require long ncRNAs for their activity. This dependency for long ncRNA expression is not observed at enhancers predicted by a different methodology. Additionally, we find that enhancers expressing long ncRNAs have a lower H3K4me1/H3K4me3 ratio suggesting that they might have a specific epigenetic profile. In summary, we verify the tissue-specific predictive power of PreSTIGE and demonstrate that almost one third of long ncRNAs are expressed from tissue-specific enhancers suggesting that the interplay between long ncRNAs and enhancers is important for regulation of tissue-specific gene expression.

Although we are able to detect thousands of long ncRNAs due to the technological progress identifying functional long ncRNAs and functional characterization of these low abundant transcripts is still challenging. By using functional data for differential expression of long ncRNAs in differentiating keratinocytes and RNA polymerase II

association, we identified PARROT, a functional long ncRNA expressed at a relatively high level in HeLa cells. Genome wide transcriptome and proteome analysis upon depletion of PARROT revealed that PARROT acts as an upstream regulator of c-Myc affecting cellular proliferation, migration and translation. Furthermore, we find that PARROT is down-regulated in senescence and up-regulated in some cancers further suggesting that PARROT has an important role in the regulation of cellular proliferation.

ZUSAMMENFASSUNG

Lange nicht-kodierende RNAs (ncRNAs) kontrollieren nahezu jede Ebene des Genexpressionsprogramms und fügen diesem eine unerwartete Komplexität hinzu. NcRNAs kontrollieren fundamentale biologische Prozesse wie z.B. X-Chromosom-Inaktivierung, Imprinting, Zellteilung, Entwicklung und Differenzierung. Zusätzlich sind sie involviert in der Entstehung einer Vielzahl von menschlichen Krankheiten wie Krebs oder neurodegenerative Erkrankungen.

Jüngste Forschungsergebnisse haben gezeigt, dass einige lange ncRNAs von einer Teilmenge der Enhancerelemente im menschlichen Genom exprimiert werden und notwendig sind um die Enhancerfunktion zu vermitteln. Dennoch sind das Ausmaß zu welchem lange ncRNAs für die Enhancerfunktion benötigt werden noch nicht bekannt. Zusätzlich besteht bisher kein Konsens darüber wie Gewebs-spezifische Enhancer vorher gesagt werden sollten obwohl Enhancer bereits seit Jahrzehnten erforscht werden. In dieser Doktorarbeit verwenden wir eine kürzlich entwickelte Methode – PreSTIGE – um Gewebs-spezifische Enhancer sowie ihre Zielgene vorherzusagen. Diese Methode basiert auf dem Gewebs-spezifischen Vorhandensein von Histonmethylierungen, im speziellen H3K4me1, sowie der Gewebs-spezifischen Genexpression. Dabei finden wir, dass 28% (2.695) aller ENCODE annotierten langen ncRNAs mit Gewebs-spezifischen Enhancern überlappen, die als solche durch PreSTIGE vorhergesagt werden. Die Expression von solchen langen ncRNAs, die mit Enhancern überlappen, ist signifikant höher im jeweiligen Gewebe in dem der entsprechende Enhancer vorhergesagt wird aktiv zu sein. Dies deutet darauf hin, dass einige Enhancer die Expression von langen ncRNAs zur Ausführung ihrer Enhancerfunktion benötigen. Diese Abhängigkeit zur gleichzeitigen Expression von langen ncRNAs wird nicht beobachtet an Enhancern, die mittels einer anderen Methode vorausgesagt werden. Zusätzlich finden wir, dass Enhancer, die eine lange ncRNA exprimieren, ein geringeres Verhältnis von H3K4me1/H3K4me3 Histonmodifikationen aufweisen, was darauf hindeutet, dass sie ein spezifisches epigenetisches Profil aufweisen könnten. Zusammenfassend weisen wir die Gewebs-spezifische Vorhersagekraft von Enhancern durch PreSTIGE nach und zeigen

außerdem, dass fast ein Drittel aller annotierten long ncRNAs von Gewebs-spezifischen Enhancern transkribiert werden. Dies deutet darauf hin, dass das Zusammenspiel von langen ncRNAs und Enhancern wichtig ist für die Regulation von Gewebs-spezifischer Genexpression.

Trotz des technologischen Fortschritts tausende lange ncRNAs zu detektieren, ist die Identifizierung funktionaler langer ncRNAs sowie die funktionelle Charakterisierung dieser gering exprimierten Transkripte anspruchsvoll. Mittels Verwendung funktioneller Daten für die differentielle Expression von langen ncRNAs in differenzierenden Keratinocyten und deren Assoziation mit RNA Polymerase II, identifizieren wir PARROT, eine funktionale lange ncRNA, die zu relativ hohen Leveln in HeLa Zellen exprimiert ist. Genomweite Transkriptom- und Proteomanalyse im Anschluss an den knock-down von PARROT zeigte, dass PARROT als ein vorgeschalteter Regulator von c-Myc fungiert und aufgrund dessen Zellteilung, Migration und Translation in HeLa Zellen beeinflusst. Weiterhin beobachten wir, dass die Expression von PARROT herunterreguliert ist während zellulärer Seneszenz und heraufreguliert ist in einigen Krebsarten. Damit könnte PARROT eine wichtige Rolle in der Regulation der zellulären Proliferation zugeschrieben werden.

1 INTRODUCTION

Mapping the human genome reference sequence was a milestone in modern molecular biology. Soon after the human genome was mapped it was recognized that almost 99% of ~3.3 billion nucleotides that constitute the human genome organized into 23 chromosomes do not code for proteins (Harrow et al., 2012). Additionally, we now know that the human genome is pervasively transcribed into non-coding RNA. Furthermore, genome-wide association studies demonstrated that the majority of trait-associated loci lie outside of protein-coding regions suggesting that the non-coding portion of the human genome carries a wide variety of functionally significant elements (Harrow et al., 2012).

RNA is most likely the primordial molecule of life containing both informational and catalytic functions. It subsequently devolved to the more stable and easily replicable DNA and its catalytic functions shifted to more versatile protein molecules. Although RNA has crucial roles in splicing (the process by which introns are removed from nascent RNA) and translation (the process by which RNA is decoded into an amino acid sequence of a polypeptide chain) it was considered to be a mere intermediate between DNA and proteins embedded in the central dogma: DNA makes RNA through a process called transcription and RNA makes protein through a process called translation (Morris and Mattick, 2014). Nowadays RNA molecules have emerged as the key players in regulation of every level of the gene expression program, wide variety of cellular processes and an increasing number of human disorders (Geisler and Coller, 2013; Harrow et al., 2012; Vucicevic et al., 2014).

1.1 Human genome

1.1.1 Organization of the human genome

The human genome is organized into chromatin (reviewed in (Zhou et al., 2011)). The nucleosome is the fundamental unit of chromatin and consists of 147 base pairs (bp) of DNA wrapped around a histone octamer that consists of histone H2A, H2B, H3 and H4. Histone H1 binds to linker DNA, DNA between the nucleosomes, and is responsible for higher order packing of the chromatin. Nucleosome positioning,

regulated by nucleosome remodeling complexes, can be specific to a certain cell type and determines which genes are accessible for transcription and binding of *trans* acting regulatory factors (Higgs et al., 2007; Zhou et al., 2011). The four histones which constitute the histone octamer can be substituted with alternative variants and can be chemically modified on their protruding N-terminal tails. These chemical modifications include methylation, acetylation, phosphorylation and ubiquitination (Higgs et al., 2007; Zhou et al., 2011). Different histone modifications are responsible for the recruitment of different regulatory factors. Based on histone modifications: open, accessible and transcriptionally active chromatin- euchromatin can be distinguished from compacted, transcriptionally inactive chromatin state- heterochromatin. For instance, transcriptionally inactive heterochromatin regions are marked by trimethylation of histone H3 lysine 9 (H3K9me3). On the other hand, transcriptionally active euchromatin regions are marked with mono/di/trimethylation of lysine 4 of histone H3 (H3K4me1, H3K4me2, H3K4me3) and bodies of actively transcribed genes are marked with trimethylation of lysine 36 of histone H3 (H3K36me3) (Figure 1.1.) (Higgs et al., 2007; Orom, 2010; Zhou et al., 2011). In summary, the establishment and maintenance of gene expression patterns, which determine cellular phenotypes, can be influenced by how DNA is packaged into chromatin which is in turn determined by the modification of the nucleosomes, the presence of histone variants, and the binding of non-histone proteins that associate with chromatin (Higgs et al., 2007).

1.1.2 Genetic regulatory elements

Gene expression is a tightly regulated process and can be achieved at every level of the gene expression program. In humans and in eukaryotes in general, transcription of most genes is performed by RNA polymerase II (Pol II). Those genes typically contain two distinct transcriptional regulatory DNA elements: 1) promoters and 2) distal regulatory elements like enhancers, silencers and insulators. These *cis* regulatory elements contain recognition sites for *trans* binding transcription factors (TFs), which can either enhance or repress transcription (reviewed in (Maston et al., 2006)) (**Figure 1**).

Promoters are located at the 5' ends of genes adjacent to the transcriptional start site (TSS) of a gene and serve as the point of assembly of the transcriptional machinery and

initiation of transcription (**Figure 1**). At the promoter, all regulation directed to a gene is converted into the rate of transcription initiation (Maston et al., 2006). The typical histone modification associated to promoters is H3K4me3 (Heintzman, 2007).

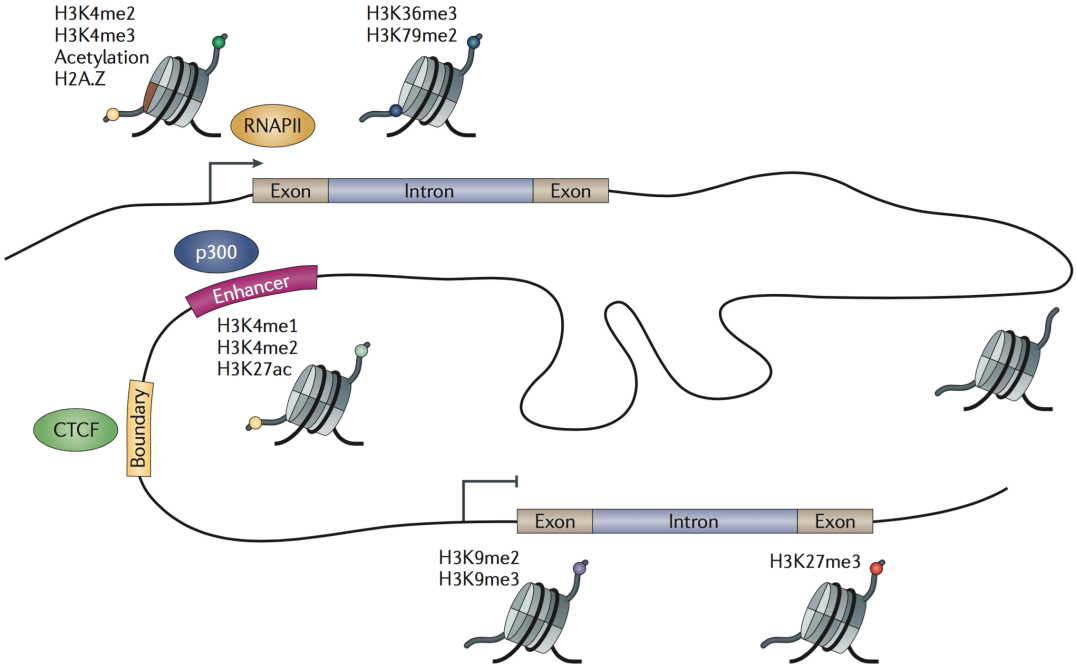

Figure 1. Histone modifications of functional elements in the human genome. Promoters, gene bodies, an enhancer and a boundary element are indicated. Active promoters are commonly marked by histone H3K4me2, H3K4me3, acetylation and H2A.Z histone variant. Actively transcribed regions are enriched for H3K36me3 and H3K79me2. Repressed genes may be located in large domains of H3K9me2 and/or H3K9me3 or H3K27me3. Enhancers are relatively enriched for H3K4me1, H3K4me2, H3K27ac and sometimes the histone acetyltransferase p300. CTCF binds sites that may function as boundary elements, insulators or structural scaffolds (Zhou et al., 2011).

Silencers act in an orientation independent manner to repress the transcription of their target genes. They do so by binding to repressors that can in turn either block the binding of an activator or recruit repressive histone modifiers that form a repressive chromatin structure (Maston et al., 2006). Insulators serve as chromatin barriers that block unwanted chromatin interactions in the genome (**Figure 1**). Therefore, they act in a position dependent and orientation independent manner. In vertebras, the insulating function is mediated by the binding of CTCF (CCCTC-binding factor). The binding of CTCF serves as a barrier between euchromatin and heterochromatin (Maston et al., 2006). Additionally, the binding of this protein is crucial for the three-dimensional architecture of chromatin since its binding can physically separate proximal regions but also bring regions of the chromatin which are far apart into close proximity.

Therefore, it was proposed that this protein organizes global chromatin architecture by mediating intra and inter-chromosomal contacts (Ong and Corces, 2011).

Enhancers are genetic regulatory elements that can activate transcription of their target genes in a temporal and tissue specific manner independent of distance and orientation. They contain binding motifs for both common and cell type specific TFs and lie within histone free, accessible regions (Maston et al., 2006) (**Figure 1**). An estimate is that 400 000 to > 1 million putative enhancers exist in the human genome. The precise activity pattern of specific cohorts of enhancers is crucial for cell type development, cell lineage determination and cellular response to stimuli (Lam et al., 2014). Early studies identified that the p300 transcriptional cofactor binds to tissue-specific enhancers (Visel et al., 2009a). Analysis of histone marks at these enhancers revealed that they are enriched in H3K4me1 and depleted from H3K4me3 (Heintzman, 2007; Heintzman and Ren, 2009). The H3K4me1 histone mark, along with others, is being used to predicted enhancers in multiple cell lines (details in section 1.2). The binding of TFs to enhancers affects the transcription of the target gene by recruiting Pol II and the assembly of the preinitiation complex. The contact between an enhancer and the promoter is achieved through long-range chromatin looping allowing interaction of the necessary co-transcriptional factors (Calo and Wysocka, 2013). For a long time it was thought that enhancers work exclusively on the DNA level. However, recent research has identified long ncRNAs transcribed from enhancers as key players in mediating enhancer function (Lam et al., 2014; Natoli and Andrau, 2012; Orom and Shiekhattar, 2013). The involvement of long ncRNAs in mediating enhancer function is an expanding topic that recently resulted in many publications on the molecular mechanisms of long ncRNA transcription in enhancer function (revived in (Natoli and Andrau, 2012; Orom and Shiekhattar, 2013). Enhancer derived long ncRNAs and their mechanism of action are described in details in section 1.4.

1.2 Enhancer predictions

Based on various enhancer features several approaches have been developed to predict them. Both computational and experimental approaches have been applied in search for enhancers and each of them has its advantages as well as limitations.

The binding of TFs to specific motifs within enhancers is the first enhancer feature that has been used to predict enhancers. This feature allows computational scanning for a TF motif across entire genomes (Arnone and Davidson, 1997). Some studies, aside from enrichment of binding sites also include the conservation of the TF motif (Del Bene, 2007; Kheradpour et al., 2007). Although these approaches identified some functional enhancers they also predict a lot of false positive enhancers. This is for instance, due to the fact that short binding motifs of TFs frequently match to genomic DNA (6 bp motif would be expected to occur every 46bp) and only a small proportion is actually bound *in vivo*. Additionally, TF binding can be context specific and depend on different factors and cofactors. Moreover, even if a motif is conserved, that does not mean that that motif is bound by a TF in a given cell line nor that this sequence will act as an enhancer (Shlyueva et al., 2014).

Since TF binding prediction is a difficult problem which has been shown to be of limited accuracy, several studies have applied genome-wide methods to determine TF binding *in vivo* (Shlyueva et al., 2014). The most commonly used method is chromatin immunoprecipitation followed by deep sequencing (Chip-seq) (**Figure 2A**). In this method, by chemical crosslinking, TFs are covalently linked to their binding sites. The chromatin is then sheared and the DNA fragments bound by the TF are co-precipitated with the TF specific antibodies and determined by sequencing (Johnson et al., 2007). This approach recovers direct binding sites of a TF and has a low false-negative rate. However, TF binding can occur without affecting the expression of any gene. This might be due to general affinity of TFs to bind DNA and the fact that the enhancers are activated by a combination of different transcriptional factors and cofactors and not all of them need to interact directly with the DNA (Spitz and Furlong, 2012). Transcriptional cofactors typically do not interact directly with the DNA. They are recruited by TFs which bind DNA. Once recruited, they perform various enzymatic activities that lead to activation or repression of transcription (Kvon et al., 2012; Moorman, 2006). Based on this, identification of cofactor binding sites has been used to predict enhancers. For instance, binding sites of a transcriptional co-factor histone acetyltransferase p300 are used to predict enhancers. As many as 58-82% of enhancers predicted in this way are reported to function as enhancers (Visel et al., 2009b).

However, this approach identifies only a subset of enhancers that require p300 binding for their activity.

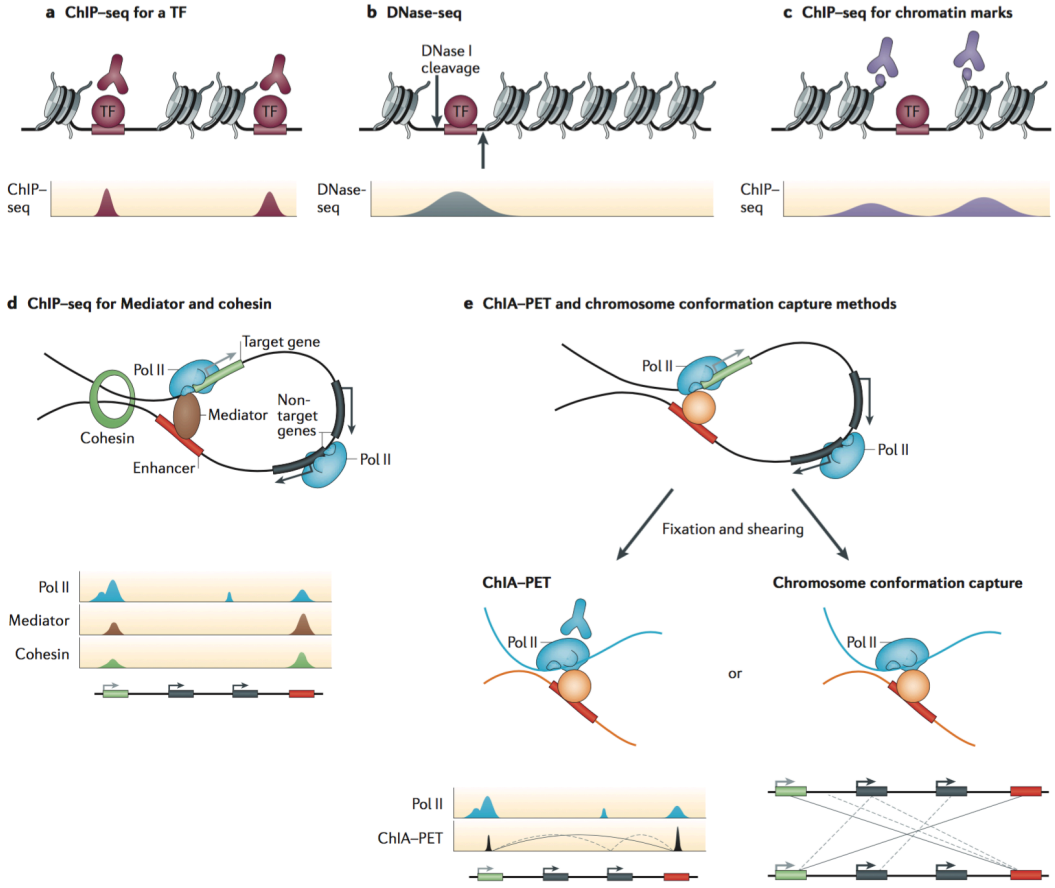

Figure 2. Methods for predicting enhancers through the detection of transcription factor binding, accessible chromatin, chromatin marks or long-range interactions. Modified from (Shlyueva et al., 2014).

Active enhancers tend to reside in regions of open chromatin such that the DNA is accessible. This stereotypical chromatin structure is used for the identification of regulatory regions by, for example, enzymatic cleavage of accessible DNA using DNase I (**Figure 2B**). Of course, not all accessible regions correspond to active enhancers and additional information is needed to predict enhancers (Shlyueva et al., 2014). Additional information can be found in stereotypical biochemical properties of histone proteins in the flanking nucleosomes. For instance, enhancers are typically marked with H3K4me1 and promoters with H3K4me3 while both are decorated with H3K427ac upon activation and can be identified by Chip-seq (**Figure 2C**). Therefore, accessibility and histone modifications are frequently used in combination to identify

active enhancers (Shlyueva et al., 2014).

An even greater challenge in predicting enhancers, is predicting their targets. Traditionally, it was thought that enhancers act on their nearest gene. However, recent evidence from chromosome conformation capture experiments shows that the average distance between an enhancer and its target is 120kb (Sanyal et al., 2012). Therefore, the fact that enhancers are brought into close proximity to the promoters of their target genes has been utilized to predict both enhancers and their targets. These interactions are shown to be mediated via cohesin and the Mediator complex so their ChIP profiles can point out the potential location of the enhancers (Whyte, 2013) (**Figure 2D**). Techniques that can directly assess long range DNA contacts are chromosome conformation capture techniques- 3C, 4C, 5C (reviewed in (van Steensel and Dekker, 2010)). In these assays, formaldehyde allows the crosslinking between spatially close DNA regions after which the DNA is sheared and the proximal DNA fragments are ligated. Long-range spatial DNA contact can be detected via deep sequencing of these chimeric DNA ligation products. Another method, called chromatin interaction analysis with paired end tag sequencing (ChIA-PET) combines ChiP and 3C in search of Pol II dependent intra-chromosomal and inter-chromosomal interactions, predicting enhancers and their targets based on these interactions (Li, 2012) (**Figure 2E**). Although the knowledge of chromatin interactions is very useful in determining enhancers and their target promoters, these methods have a low resolution (order of magnitude 1kb to 10kb) and are not reliable in predicting interactions that are in close proximity to each other (Jin, 2013; Li et al., 2012b). Additionally, not all chromatin interactions are those between enhancers and promoters.

A recently developed approach for tissue-specific enhancer prediction, PreSTIGE (Predicting Specific Tissue Interactions of Genes and Enhancers) uses gene expression and H3K4me1 across a panel of cell lines to identify both tissue-specific enhancers and their targets (Corradin et al., 2014). PreSTIGE predicts enhancers by first finding protein coding genes (PCGs) with tissue-specific increased expression. In the tissue in which the PCG has an increased expression and within the specified domain size surrounding the TSS of the PCG, PreSTIGE predicts enhancers based on the presence of cell type specific H3K4me1 domains (Corradin et al., 2014). This method has an advantage over previously published methods since it takes into account the expression

of the target gene and predicts tissue specific enhancers regardless of the presence of H3K4me3 marks. However, this approach also has its own limitations. It is limited to predicting only tissue specific enhancers and can only predict enhancer-gene pairs within the specified window surrounding the TSS. For its advantages, this method has been selected for use in this thesis and will be discussed in more details in the following chapters.

1.3 Long non-coding RNAs

The development of next generation RNA-sequencing (RNA-seq) technologies shifted the protein centric view to ncRNAs causing a dramatic change in molecular biology. The ENCODE project reported that whilst only a small fraction of the human genome codes for a protein (2-3%), 62-75% of the human genome is transcribed into transcripts that do not code for a protein. The vast majority of these transcripts are referred to as long ncRNAs (Derrien, 2012; Djebali, 2012; Harrow et al., 2012).

1.3.1 Characteristics of long ncRNAs

Long ncRNAs are RNA transcripts arbitrarily defined as being greater that 200 nucleotides in length that do not code for a protein judged by the absence of open reading frames and codon conservation (Derrien, 2012; Ulitsky and Bartel, 2013). Their definition alone indicates how little we know about long ncRNAs. They are defined based on what they are not: they do not code for a protein and are not smaller than 200 nt in order to be distinguished from other RNA species such as micro RNAs. It is, therefore, not a surprise that long ncRNAs are a very heterogeneous group of RNA molecules involved in wide variety of biological processes (Geisler and Coller, 2013; Ulitsky and Bartel, 2013).

The classification of long ncRNAs, comprised of molecules with mostly unknown (if any) function, is challenging (reviewed in (St Laurent et al., 2015)). The ENCODE consortium classifies long ncRNAs based on their localization in the human genome with respect to PCGs into 5 categories: 1) Intergenic long ncRNAs (lincRNAs) -long ncRNAs that do not overlap with PCGs; 2) Exonic antisense long ncRNAs that overlap with exons of PCGs and are transcribed from the opposite strand; 3) Intronic long

ncRNAs that overlap an intron of a PCG on either strand; 4) Overlapping transcripts that contain a PCG within one of their introns; 5) Processed transcripts that can't be classified into any of the above categories. This type of long ncRNA classification results in a final set of 14,880 transcripts originating from 9277 loci (Derrien, 2012).

Long ncRNAs resemble mRNAs in their structure. Their genes are typically shorter than PCG and tend to have only 2-3 exons that are on average slightly longer than those of PCGs (Derrien, 2012). However, we should keep in mind that the first and the last exon in PCG are usually longer so this might not be a difference between long ncRNAs and PCGs. Transcriptional regulation, splicing signals, polyadenylation signals and histone modification patterns at long ncRNAs are indistinguishable from those at PCGs (Derrien, 2012). On the other hand, unlike PCG transcripts, which are spliced co-transcriptionally, long ncRNAs tend to be less efficiently spliced (Tilgner et al., 2012). Contrary to most PCGs, long ncRNAs show a highly tissue specific expression pattern with many long ncRNAs being expressed in the brain and testis (Cabili, 2011; Derrien, 2012). Their expression level is generally low, the median expression level of long ncRNAs is approximately ten times lower than the median mRNA expression. It is still unclear whether this difference is caused by less efficient transcription or faster degradation of long ncRNAs (Derrien, 2012; Sigova et al., 2013; Ulitsky and Bartel, 2013). Recent evidence suggests that this difference is likely due to the lower stability rather than the lower transcription of long ncRNAs (Sun et al., 2015).

Relative to mRNAs long ncRNAs are localized predominantly in the nucleus with a threefold nuclear enrichment. However, since mRNAs are not equally distributed in the nucleus and the cytosol, in fact they are highly enriched in the cytosol, these relative enrichments for long ncRNAs in the nucleus might be a misconception. For instance, if an mRNA is sixfold enriched in the cytosol, then a long ncRNA with a threefold nuclear enrichment would still be two times more abundant in the cytosol (Ulitsky and Bartel, 2013). Specific sites of localization, factors that are involved in exporting and localizing long ncRNAs are still unknown.

In contrast to mRNAs, long ncRNAs have evolved very rapidly and lack known orthologs in species outside of vertebrate. Only ~ 12% of long ncRNAs are conserved

between mouse and human (Cabili, 2011). Within rodents only ~ 60% of long ncRNAs are expressed in both the liver of *Mus musculus* and *Mus castaneus*. Intriguingly, a correlation is observed between the presence of a lineage specific long ncRNA and the expression of the neighboring PCG (Kutter et al., 2012). Although long ncRNAs evolve very rapidly there is a detectable level of natural selection acting on them. Their exons are more conserved than intergenic regions but to a much lesser extent than exons of PCGs (Derrien, 2012). It is possible that long ncRNAs are conserved on the level of their secondary structure but not much is known about their secondary structure to this date. Even though long ncRNAs are not highly conserved a growing body of evidence suggests that there are many functional long ncRNAs.

1.3.2 Diverse functions of long ncRNAs

Although only a small fraction of long ncRNAs have been well characterized to date we already know that they can regulate every level of a gene expression program (**Figure 3**) (Geisler and Coller, 2013; Ulitsky and Bartel, 2013). They regulate fundamental biological processes such as imprinting, differentiation and X chromosome inactivation and they are involved in the development of many human disorders like cancer and neurological disorders (Geisler and Coller, 2013; Ulitsky and Bartel, 2013; Vucicevic et al., 2015).

Often long ncRNAs associate to proteins to bring about regulatory functions (Bertani et al., 2011; Gong and Maquat, 2011; Lai, 2013; Rinn, 2007; Tripathi, 2010; Wang et al., 2011b) emphasizing one area of intensive research. For a growing number of long ncRNAs the function as a guide to a chromatin remodeling complex has been described (**Figure 3**) (Bertani et al., 2011; Rinn, 2007; Wang et al., 2011a).

For instance, long ncRNA BDNF-AS directly interacts with a subunit of polycomb repressive complex 2 (PRC2) called EZH2 to regulate the expression of its sense protein partner brain derived neurotrophic factor (BDNF). At the BDNF promoter, depletion of BDNF-AS lead to a reduction in the occupancy of EZH2 as well as H3K27me3, a mark deposited by PRC2 complex to repress the expression of a gene (Vashishtha et al., 2013). Similarly, long ncRNA ANRIL binds to the CBX7 subunit of the polycomb repressive complex 1 (PRC1) and at the same time interacts with PRC2 acting as a scaffold for those two protein complexes. Through these interactions,

ANRIL mediates the silencing of the INK4a/ARF/INK4a locus that encodes for tumor-suppressor genes responsible for regulating cell cycle and senescence (Pasmant et al., 2011). Reduced proliferation of the cells is observed upon depletion of this long ncRNAs. ANRIL is also up-regulated in prostate cancer and leukemia suggesting that this long ncRNA can alone act as an oncogene (Huarte and Rinn, 2010; Pasmant et al., 2011). Long ncRNAs can also act as decoys for transcription factors to prevent the transcription of certain genes. Gas5 is one such long ncRNA. Upon starvation its expression is induced and by mimicking a glucocorticoid receptor binding motif it decoys it away from its DNA binding sites preventing the transcription of metabolic genes (Kino et al., 2010). Similarly, long ncRNA PANDA associates with a transcriptional factor NF-YA preventing p53 mediated apoptosis (Hung et al., 2011). Long ncRNAs can also act in an enhancer like manner to activate the expression of their target genes. Since this thesis is focusing on them, they will be introduced in details in the following section.

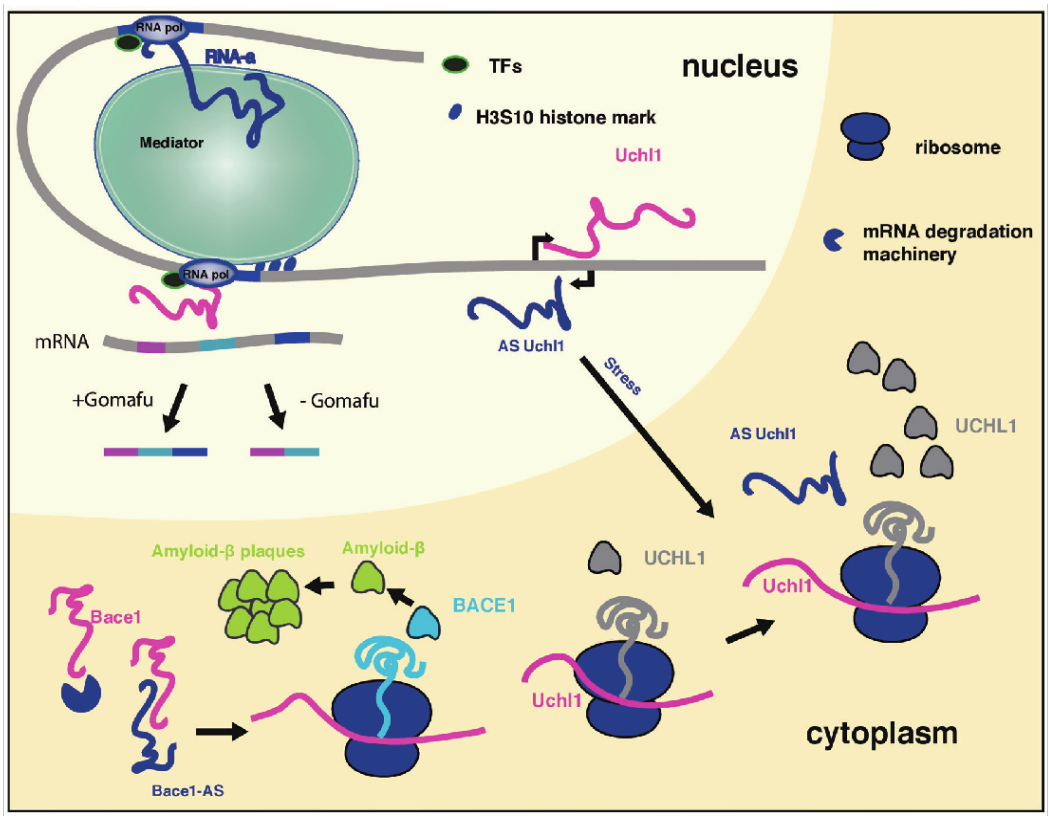

Figure 3. Long ncRNAs can regulate every level of gene expression program. Shown is a summary of selected long ncRNA functions that act in different cellular compartments (Vucicevic et al., 2014).

In addition to regulating gene expression on a transcriptional level, long ncRNAs can also regulate gene expression post-transcriptionally. They can regulate mRNA processing, mRNA stability, and translation (Geisler and Coller, 2013; Ulitsky and Bartel, 2013; Vucicevic et al., 2015).

An example of a long ncRNA regulating mRNA splicing is a long ncRNA called Gomafu (also called MIAT and RNBR2) (**Figure 3**). This ncRNA has a distinct feature, it contains a tandem repeat sequence UACUAAC that is a conserved intron branch point that binds to the SF1 splicing factor. Gomafu also has the ability to bind two other splicing factors, QKI and SRSF1 (Albertson et al., 2006; Barry, 2014). Dysregulation of this long ncRNA leads to alternative splicing patterns of DISC1 and ERBB4 which are similar to splicing patterns observed in schizophrenic disorder. Additionally, Gomafu is dysregulated in the cortex of schizophrenic subject and upregulated in the region of the brain involved in behavior and addiction to cocaine and heroin, suggesting that Gomafu has a role in development of behavioral abnormalities (Albertson et al., 2006; Barry, 2014).

The long ncRNA BACE1-AS is transcribed antisense to β-secretase-1 protein (BACE1) and regulates mRNA stability of BACE1, an enzyme that generates amyloid-β. Amyloid-β clusters in amyloid plaques that are a histological hallmark of Alzheimer's disease (Faghihi et al., 2008; Kim et al., 2013). In stress conditions BACE1-AS is upregulated and increases BACE1 mRNA stability by duplexing with BACE1 mRNA (**Figure 3**) leading to the generation of additional BACE1 enzyme and amyloid-β (Faghihi et al., 2008). In mouse brain, the reduction of BACE1 protein levels, reduction of amyloid-β synthesis and its aggregation in the brain is observed upon depletion of BACE1-AS signifying the importance of this long ncRNA for the development of Alzheimer's disease (Modarresi et al., 2011). BACE1-AS is an interesting example of a long ncRNA acting without the help of a protein partner suggesting that there might be many more long ncRNAs that don't require protein partners to mediate their activity.

A long ncRNA transcribed antisense of the mouse ubiquitin carboxy-terminal hydrolase L1 (*Uchl1*) gene can induce the translation of Uchl1 without affecting its mRNA levels. Uchl1 mRNA localizes predominantly in the cytoplasm whereas the antisense ncRNA is enriched in the nucleus of dopaminergic neurons. Upon treatment of dopaminergic

cells with an mTOR inhibitor, antisense Uchl1 relocalizes to the cytoplasm, triggers the binding of Uchl1 mRNA to polysomes and leads to an increase in UCHL1 protein level (**Figure 3**). (Carrieri, 2012).

In summary, long ncRNAs can regulate every level of the gene expression program. They regulate many biological processes and are involved in the development of a wide variety of disorders.

1. 4 Enhancer derived long ncRNAs

Although there were few examples of enhancers being transcribed (Ashe et al., 1997; Collis et al., 1990; Tuan et al., 1992) the discovery that enhancers are being pervasively transcribed into ncRNAs came as a surprise adding an unexpected layer of complexity to the regulation of gene expression (reviewed in (Lam et al., 2014; Natoli and Andrau, 2012; Orom and Shiekhattar, 2013)). The question that is under intense investigation is: what is the function of these enhancer derived long ncRNAs? These enhancer-derived ncRNAs can be merely just transcriptional noise, a consequence of being in open chromatin. It was also proposed that the act of transcription *per se* and not the transcript itself is important for the enhancer function. Finally, a growing body of evidence is showing that the enhancer derived transcript itself is responsible for mediating the activity of the enhancer it is derived from (Lam et al., 2014; Natoli and Andrau, 2012; Orom and Shiekhattar, 2013).

1.4.1 Characteristics of enhancer derived long ncRNAs

Enhancer derived long ncRNAs represent a very heterogeneous group of RNA species. Initially discovered enhancer derived long ncRNAs are spliced and polyadenylated resembling mRNAs in their structure and belong to the group of intergenic or intervening long ncRNAs. They are often referred to as unidirectional (1D-eRNAs) since they are transcribed in only one direction (Lam et al., 2014; Natoli and Andrau, 2012; Orom and Shiekhattar, 2013). A great number of these ncRNAs are transcribed upon stimuli or are developmentally regulated (Onoguchi et al., 2012; Orom, 2010; Wang et al., 2011b). More often bidirectionally transcribed long ncRNAs are observed at enhancers and are therefore called 2D-eRNAs. These enhancer derived bidirectional transcripts are often shorter than classical long ncRNAs, they are not spliced nor

polyadenilated. Since 2D-eRNAs have a short half-life, they can only be found associated to chromatin so this feature can additionally distinguish 1D-eRNAs from 2D-eRNAs (Lam et al., 2014; Natoli and Andrau, 2012). Upon certain stimuli both 1D-eRNAs and 2D-eRNAs are produced from enhancers. Adding an additional layer of complexity is the observation that intragenic enhancers can act as alternative TSSs generating both multiexonic polyadenilated and short nonpolyadenilated transcripts simultaneously (Kowalczyk et al., 2012). Even though the overlap between 1D-eRNAs and 2D-eRNAs is not yet clear and why their biogenesis differs, a growing body of evidence demonstrates that both of these enhancer derived RNA classes exert their function *via* affecting the expression of their neighboring genes (Bhatt et al., 2012; Lam et al., 2014; Orom and Shiekhattar, 2013).

1.4.2 Functional enhancer derived long ncRNAs

Several studies have shown that the expression of long ncRNAs from enhancers correlates with the expression of their neighboring PCGs (De Santa, 2010; Hah et al., 2013; Kim, 2010; Koch and Andrau, 2011). A pioneering study by Kim and colleagues identified thousands of bidirectionally transcribed nonpolyadenylated ncRNAs in mouse neurons. These transcripts are induced upon stimuli from enhancers defined by the presence of H3K4me1 mark and the binding of p300 (Kim, 2010). Another study of Pol II association to enhancers in macrophages also identified many enhancers that produce nonspliced, polyadenylated long noncoding RNAs. The transcription of some of these enhancer transcripts was induced upon stimulation of macrophages with the endotoxin. Their inducible activation preceded endotoxin-induced transcription of their neighboring PCGs. These results imply that these ncRNAs might be responsible for the activation of the neighboring PCGs (De Santa, 2010). Similarly, in MCF7 cells treated with estradiol, transcription of a majority of long ncRNAs preceded the activation of their target gene (Hah et al., 2013).

A pioneering study by Ørom and colleagues provided for the first time evidence that long ncRNAs transcribed from enhancers have a crucial role in mediating enhancer activity in human cells (Orom, 2010). In this study, a knock-down approach demonstrated that the expression of long ncRNAs is necessary for the expression of the neighboring PCGs. In the absence of the long ncRNA called activating RNA (RNA-a) the expression of the target PCGs was diminished. Furthermore, they were

able to confirm this in a classical enhancer reporter assay. A similar approach was applied in several other studies identifying individual enhancer derived long ncRNAs mediating enhancer activity. For instance: long ncRNA HOTTIP causes transcriptional activation of HOXA genes (Wang et al., 2011b) and long ncRNA *utNgn1* is required for transcriptional activation of neurogenin 1 during development (Onoguchi et al., 2012). Additionally, *in vivo* experiments in mice demonstrated that long ncRNA NeST activates the transcription of adjacent interferon-γ coding gene (Gomez et al., 2013).

Investigation of how transcription factors mediate their effects gave further insight into the role of enhancer derived long ncRNAs in transcriptional activation. Recent studies showed that many transcription factors exert their effect by affecting the expression of long ncRNAs at enhancers. The oestrogen receptor, upon stimulation by estradiol, preferentially binds to enhancers that produce long ncRNAs. Knock-down of these long ncRNAs led to a depletion in expression of surrounding genes (Li, 2013). Similarly, p53 also exerts its effect by regulating the expression of enhancer-derived long ncRNAs upon whose depletion a decrease in expression of surrounding p53 target genes is observed (Melo et al., 2013). Contrastingly, nuclear receptor Rev-Erb regulates the expression of its target genes by binding and suppressing the expression of enhancer derived long ncRNAs in macrophages. These authors also show that most distal Rev-Erb binding sites expressing long ncRNAs show enhancer features such as the presence of H3K4me1 marks. (Lam et al., 2013).

1.4.3 Mechanism of action of enhancer derived long ncRNAs

Several different mechanisms of action have been described for enhancer derived long ncRNAs. Most of them involve a direct interaction with a protein partner that helps them mediate their effect. They also often physically loop to the promoters of their target genes. In fact, it is demonstrated that enhancers that loop to the promoters of their target genes express RNAs at higher levels than those that don't loop (Lin et al., 2012; Sanyal et al., 2012).

A study, in a search for the mechanism of action of RNA-a, identified the Mediator complex as a protein partner of these long ncRNAs (Lai, 2013; Orom, 2010). This interaction regulates the localization and kinase activity towards histone H3 serine 10 of the Mediator complex. The localization and the kinase activity promotes

transcription of the target gene. Intriguingly, Mediator carrying a disease associated mutation displays a diminished ability to associate with RNA-a. Furthermore, chromosome conformation capture demonstrated the presence of DNA looping between the RNA-a loci and its targets. This interaction is reduced in the absence of either RNA-a or Mediator (Lai, 2013).

Similarly, eRNAs induced by oestrogen also contribute to the looping between an enhancer they are derived from and the target gene. A reduction in the looping between an enhancer and a promoter and a consequent reduction of expression of the target genes was observed upon knock-down of these eRNAs at NR1P1 and GREB1 loci demonstrating the importance of these eRNAs in mediating enhancer activity (Li et al., 2013). The authors suggest that the role of eRNAs in chromosomal looping might be mediated *via* their interaction with the components of complex known to control the enhancer- promoter looping - the cohesin complex (Li, 2013).

The long ncRNA HOTTIP directly interacts with WDR5, a key component of the mixed lineage leukemia (MLL) complex (Wang et al., 2011a). Through chromatin looping, HOTTIP is brought into the proximity of multiple genes in the HOXA cluster. By targeting WDR5 to the HOXA locus HOTTIP causes histone H3 trimethylation at lysine 4 (H3K4me3) leading to transcriptional activation of HOXA genes. *In vivo,* HOTTIP mediated activation of genes in the HOXA locus leads to developmental defects of chicken limbs demonstrating the importance of this long ncRNA (Wang et al., 2011b).

In mouse embryonic stem cells, the long ncRNA Mistral also interacts with the MLL complex and recruits it to activate the transcription of HOXA6 and HOXA7 genes. This process leads to dynamic changes in chromatin conformation and activation of genes involved in germ-layer specification during mouse embryonic stem cell differentiation (Bertani et al., 2011).

In accordance with their low level of expression and nuclear localization most described enhancer derived long ncRNAs act in *cis* - in the proximity of the region from which they are transcribed (Orom and Shiekhattar, 2013). There are also examples in which long ncRNAs can activate the transcription of their targets in *trans* - over a great distance. One such example is Jpx, a long ncRNA that activates the

expression of another long ncRNA. By evicting CTCF from its promoter, long ncRNA Jpx activates in *trans* the expression of long ncRNA Xist, the long ncRNA that is crucially involved in mammalian X chromosome inactivation (Sun et al., 2013; Tian et al., 2010). Another example is a long ncRNA *NeST* that activates the expression of INF-ɣ. In mice, the functional consequence of the loss of *NeST* can be restored through overexpression of this long ncRNA in a transgenic mice demonstrating that the site of transcription of this long ncRNA and the context in which it is produced is not important, clearly demonstrating that this long ncRNA acts in *trans* (Gomez et al., 2013).

In summary, enhancer derived long ncRNAs are a heterogeneous group of ncRNAs comprised of both short bidirectional unspliced transcripts and unidirectional processed transcript that regulate the expression of their target genes in a positive manner. Most of them act in *cis* and are brought in the close proximity to their target genes through chromatin looping. Enhancer derived long ncRNAs add an unexpected layer of complexity into the regulation of gene expression and further examination of the enhancer derived long ncRNAs can shed light onto the complex gene expression regulatory network.

1.5 MYC oncogene

The members of the MYC family of proteins are pivotal regulators in tumorigenesis. This family consists of three members c-Myc, n-Myc and l-Myc (Dang, 2012; Deng et al., 2014). All three members function in a similar manner but their expression differs in different types of cancers. The c-Myc oncogene was the first one to be discovered as an oncogene with a translocation into the immunoglobulin heavy chain locus in human Burkitt lymphoma. C-Myc is amplified in both blood-borne and solid tumors such as breast cancer, prostate cancer and colon cancer. N-Myc is expressed mostly in neural tumors and can be used as a prognostic indicator for neuroblastoma whereas l-Myc is overexpressed in small cell lung carcinomas (Chen et al., 2014; Dang, 2012; Dang et al., 2009). Overall, many human cancers show increased expression of one of the MYC family members. All three members belong to a group of basic helix-loop-helix-leucine zipper transcription factors that act to regulate transcription. They contain a b-HLH-LZ motif that allows MYC family members to

form heterodimers with MYC- associated protein X (MAX). The MYC-MAX heterodimer binds to E-box (CACGTG) DNA recognition sequence through which they regulate the expression of their target genes. With the help of MAX, c-Myc regulates the transcription of at least 15% of the genes in the human genome (van Riggelen et al., 2010). C-Myc regulates many biological processes such as cell cycle, differentiation, apoptosis and angiogenesis by modulating the expression of genes involved in these processes. Additionally, c-Myc regulates protein synthesis through regulation of expression of the components of the ribosome complex (Meyer et al., 2008; van Riggelen et al., 2010). Since c-Myc regulates cellular processes which are crucial for malignant transformation and is upregulated in many human cancers, it is a subject of intense investigation for cancer treatment (Chen et al., 2014; Dang et al., 2009).

Examples of long ncRNAs acting either upstream or downstream of oncogenes and tumor suppressors have been described (Huarte et al., 2010; Melo et al., 2013; Musahl et al., 2015; Sanchez et al., 2014). The long ncRNAs lincRNA-p21 (Huarte et al., 2010), Pint; PR-lncRNA-1 and PR-lncRNA-10 (Sanchez et al., 2014) are regulated by the tumor suppressor p53 and are involved in mediating its effects (Leveille et al., 2015; Melo et al., 2013). Using a different mechanism, the long ncRNA-RB1 is co-expressed with RB1 tumor suppressor from a bidirectional promoter. This ncRNA associates RB1 transcription to the transcription of another tumor suppressor, Calreticulin, through which it affects immunogenic cell death (Musahl et al., 2015).

Recent research has shown that there are long ncRNAs that act either upstream or downstream of c-Myc. The long ncRNA H19 is imprinted at H19 insulin-like growth factor 2 locus and is expressed only from the maternal allele. Allele specific ChIP experiments in breast and lung cancer cell lines showed that c-Myc directly binds to the promoter of H19 activating the expression of H19 by recruiting a histone acetyltransferase. A reduction in clonogenicity and anchorage independent growth was observed upon depletion of this long ncRNA in both breast and lung cancer cells. These results suggest that H19 acts downstream of c-Myc to promote tumorigenesis in breast and lung cancer cells (Barsyte-Lovejoy et al., 2006).

Similarly, in gastric cancer c-Myc was shown to activate the expression of colon cancer associated transcript (CCAT1), a long ncRNA that was first discovered to be associated to colon cancer. Overexpression of this long ncRNA leads to an increase in proliferation and migration of gastric cancer cells (Yang et al., 2013). A longer isoform of the same transcript called CCAT1-L was shown to regulate the expression of MYC in colon cancer. It is proposed that CCAT1-L enables the interaction between the enhancer and the MYC promoter through modulation of CTCF concentration thereby promoting tumorigenesis (Xiang et al., 2014). Another long ncRNA associated to colon cancer called CCAT2 can enhance the expression of MYC. It does so by increasing the transcriptional activity of TCF7L2 leading to abnormal activation of Wnt signaling that leads to activation of MYC. This ncRNA alone is also capable of promoting growth and tumorigenesis of colon cancer cells (Ling et al., 2013).

A highly expressed long RNA in gastric cancer called gastric carcinoma high expressed transcript 1 (GHET1) also regulates c-Myc. This ncRNA enhances the stability and the expression of c-Myc by cooperating with insulin-like growth factor 2 mRNA binding protein 1 (IGF2BP1) leading to an enforced physical interaction between IGF2BP1 and c-Myc RNA. Depletion of c-Myc reduces the ability of GHET1 to promote proliferation of cancer cells (Yang et al., 2014).

In cervical cancer long ncRNA XLOC_010588 can decrease the expression of c-Myc. The expression of this long ncRNA is lower in cancer cells than in the healthy surrounding tissue suggesting that this long ncRNA might act as a tumor suppressor (Liao et al., 2014). Additionally, long ncRNA PCGEM1 can activate the transcription of c-Myc (Hung et al., 2014) whereas long ncRNA GAS5 can inhibit the translation of this oncogene by directly interacting with both c-Myc mRNA and eIF4E translation initiation factor (Hu et al., 2014).

Further investigation of the interplay between c-Myc and long ncRNAs can potentially shed light onto its role in cancer and the understanding of the molecular mechanisms involved.

1.6 Aim of the thesis

In this thesis, I aimed at functionally characterizing long ncRNAs. Two different approaches were used: 1) a computational approach for a genome-wide search for long ncRNAs involved in enhancer function in multiple cell lines and 2) an experimental approach to gain insight into the function and mechanism of action of a long ncRNA with an unknown function.

It has recently been demonstrated that long ncRNAs are expressed from a subset of enhancers. Furthermore, it is shown that some of these long ncRNAs are required for the enhancer activity and are responsible for the activation of the enhancer target gene. Based on this evidence, in this thesis I aimed at investigating the interplay between long ncRNAs genome wide across a panel of 11 cell lines to determine how many of the annotated long ncRNAs can potentially act to mediate enhancer activity. To this end I aimed at predicting tissue specific enhancers and their targets based on the presence of the tissue-specific H3K4me1 enhancer mark and tissue specific increased gene expression. Correlation in expression between long ncRNAs overlapping enhancers and predicted enhancer activity were also examined. Additionally, to gain further insight, the epigenetic profile of enhancers overlapping long ncRNAs were examined.

Long ncRNAs have been shown to control almost every level of the gene expression program. However, studying long ncRNAs is challenging due to their low expression level and association to chromatin. To bypass the challenges involved in studying lowly expressed transcripts, in the second project I aimed at identifying a functional, highly expressed long ncRNA transcribed by Pol II. The potential function of a long ncRNA identified in this manner has been addressed by the loss of function approach, i.e. knock-down experiments. Functional assays like the translation assay, cellular viability and migration assays were used to address the function of the selected ncRNA. Additionally, to gain insight into the mechanism of action of a selected long ncRNA, genome wide approaches such as RNA-seq and mass-spectrometry were employed. Furthermore, the aim was to examine the potential role of this long ncRNA in biological processes such as senescence and malignant transformation.

2 MATERIALS AND METHODS

2.1 Materials

2.1.1 Instrumentation

All devices used in this work are listed in Table 1.

Table 1: Devices.

Device	Name, Manufacturer, Country
Analytical balance	BP 61, Sartorius, Germany
Blotting System	Mini Trans-Blot Cell, Bio-Rad Laboratories, USA
Centrifuge	FRESCO 17, Heraeus, Germany
Chemiluminescence imager	FUSION-SL Advance 4.2 MP, PeqLab, Germany
Flow cytometer	Cyan ADP, Beckman Coulter, USA
Fluid aspiration system	BioChem-VacuuCenter BVC 21, Vacuubrand, Germany
Freezer	Comfort, Premium NoFrost, Liebherr, Switzerland
Heating Block	Thermomixer5436, Eppendorf, Germany
Horizontal electrophoresis system	Mini-Sub Cell GT Cell, Bio-Rad Laboratories, USA
Ice machine	AF30, Scotsman Ice Systems, USA
Incubator for bacteria	Heraeus-Brutschrank B 504, Heraeus, Germany
Incubator for cell culture	Heracell CO2, Heraeus, Germany
Liquid chromatographer	Dionex Ultimate 3000, Thermo Scientific, USA
Magnetic stirrer	TK22, Kartell Labware, Australia
Mass spectrometer	Q-Exactive Plus Orbitrap,Thermo Scientific, USA
Microplate luminometer	LUMIstar Omega, BMG Labtech, Germany
Microscope	Axiovert 40 CFL, Zeiss, Germany
Microwave	SEVERIN 900&Grill, Severin, Germany
Mini Centrifuge	5430, 5810 R, MiniSpin, Eppendorf, Germany
Multi-pipette	Multipette Xstream, Eppendorf, Germany
pH meter	HI 221, Hanna Instruments, Canada
Photometer	Ultrospec 10 Cell Density Meter, Amersham Biosciences, UK
Pipettes	PIPETMAN P2, P20, P200, P1000, Gilson, USA
Pipettor	VacuuHandControl, Vacuubrand, Germany
Power supply	Power Pac 300, Bio-Rad Laboratories, USA
qRT-PCR cycler	ABI (PRISM 7900 HT), Life Technologies, USA
Refrigerator	ProfiLine, Liebherr, Switzerland
Rocking platform	ST5, Ingenieurbüro CAT, Germany
Shaker for culturing bacteria	Innova 440, New Brunswick Scientific, Germany
Sonicator	W375, Heat Systems, USA
Spectrophotometer	NanoDrop 2000, PeqLab, Germany
Sterile bench	HERAsafe HSF 12, Heraeus, Germany
SW40Ti rotor	Beckman, Palo Alto, CA
Thermocycler	Peqstar 2x gradient, PeqLab, Germany
UV transilluminator	Gel iX20 Imager, Intas, Deutschland
Vertical electrophoresis system	XCell SureLock Mini-Cell, Life Technologies, USA
Water bath	WNE, Memmert, Germany
Water purification system	Purelab Chorus, Elga Labwater, Germany

2.1.2. Consumables

Table 2 lists the items routinely used in this thesis.

Table2: Consumables.

Product, Manufacturer, Country
4-12% NuPAGE Bis-Tris Precast Gels, Life Technologies, USA
96-well Black/Clear Imaging Plates, BD Biosciences, USA
96-well white plates (LumiNunc), Thermo Fisher Scientific, USA
Bottle top filter (Steritop-GV, 0.22 μm), Merck Millipore, Germany
Cell culture plates (10 cm, 6-, 24-, 96-well), TPP, Switzerland
Cell scraper, Sarstedt, Germany
Combitips advanced (0.1 ml, 0.5 ml), Eppendorf, Germany
Eppendorf safe-lock micro test tubes (1.5 ml, 2 ml), Eppendorf, Germany
Falcon Tubes (15 ml and 50 ml), Greiner-Bio-One, Germany
Gloves, VWR International, Germany
MicroAmp Clear Adhesive Film, Life Technologies, USA
Microscope slides, Thermo Fisher Scientific, USA
Needles, BD Biosciences, USA
Optical well plates for qPCR, Life Technologies, USA
Pasteur-Pipettes, VWR International, Germany
Petri dishes, Greiner-Bio-One, Germany
Pipette tips, DeckWorks, Corning, USA
Precision Plus Protein Dual Color Marker, Bio-Rad Laboratories, USA
PVDF membrane, Bio-Rad Laboratories, USA
Serological pipettes, Sarstedt, Germany
Surgical blades, B. Braun Melsungen AG, Germany
Syringes, BD Biosciences, USA
TiO_2 columns, GL Sciences, Japan
Weighting dishes, Roth, Germany
Whatman Gel-Blotting Paper, 1.4 mm, Thermo Fisher Scientific, USA

2.1.3 Chemicals

All chemicals used in this study are listed in Table 3.

Table 3: Chemicals.

Chemical	Manufacturer, Country
1,4-Dithiothreitol (DTT)	Biomol, Germany
2-Propanol	Merck, Germany
3,3', 5-tri-iodo-L-thyronine (T3)	Sigma-Aldrich, USA
HEPES pH 7.5, 1 M	AppliChem, Germany
4,6-diamidino-2-phenylindole (DAPI)	Sigma-Aldrich, USA
Acetic acid	Merck, Germany
Agar, Bacto	BD Biosciences, USA
AlbuMAX	Invitrogen, USA
Ampicillin, sodium salt	AppliChem, Germany
b-Estradiol (E2)	Sigma-Aldrich, USA
Bovine serum albumin	Sigma-Aldrich, USA
Cholera toxin	Sigma-Aldrich, USA
Complete, EDTA free, protease inhibitor cocktail tablets	Roche, Switzerland

Table 3 continued

Chemical	**Manufacturer, Country**
Cycloheximide	Sigma-Aldrich, USA
Deoxyadenosine triphosphate (dATP)	Life Technologies, USA
Deoxycytidine triphosphate (dCTP)	Life Technologies, USA
Deoxyguanosine triphosphate (dGTP)	Life Technologies, USA
Deoxythymidine triphosphate (dTTP)	Life Technologies, USA
Diethyldicarbonat (DEPC)	Sigma-Aldrich, USA
Epidermal Growth Factor	Upstate Biotechnology, USA
Ethylenediaminetetraacetic acid (EDTA), 500 mM	AppliChem, Germany
Fetal bovine serum (FBS)	EuroClone, Italy
Fluorescent mounting medium	Dako, Germany
Formaldehyde (37%)	AppliChem, Germany
GeneRuler 100bp Plus DNA Ladder	Thermo Fisher Scientific, USA
GeneRuler 1kb DNA Ladder	Thermo Fisher Scientific, USA
Glycerol	Merck Millipore, Germany
Glycerol, BioUltra	Sigma-Aldrich, USA
Hydrocortisone	Sigma-Aldrich, USA
Insulin	Sigma-Aldrich, USA
Isoproterenol	Sigma-Aldrich, USA
L-glutamine 200 mM	Invitrogen, USA
Magnesium chloride (MgCl2)	Sigma-Aldrich, USA
Methanol	Merck, Germany
Milk powder	Biomol, Germany
NuPAGE MOPS SDS Running Buffer	Life Technologies, USA
Orange G	Sigma-Aldrich, USA
Oxytocin	Bachem, USA
Paraformaldehyde	Sigma-Aldrich, USA
Penicillin-Streptomycin	Life Technologies, USA
Phosphate-buffered saline (PBS), 10x	Life Technologies, USA
Ponceau S	Sigma-Aldrich, USA
Precision Plus Protein Dual Color Marker	Bio-Rad Laboratories, USA
RNASEZAP	Sigma-Aldrich, USA
RotiLoad	Roth, Germany
Sodium chloride (NaCl)	AppliChem, Germany
Sodium deoxycholate	Sigma-Aldrich, USA
Sodium dodecyl sulfate (SDS)	Promega, USA
Sodium hydroxide (NaOH)	Sigma-Aldrich, USA
Stripping buffer (for western blots)	Thermo Fisher Scientific, USA
Sucrose	Sigma-Aldrich, USA
SUPERase In RNase Inhibitor	Life Technologies, USA
SuperSignal West DURA Extended Duration 100	Thermo Fisher Scientific, USA
SYBR Safe DNA Gel Stain	Life Technologies, USA
Tris-HCl, 1M soln., pH 7.4, RNase free	Alfa Aesar, USA
Tris/Borate/EDTA (TBE) buffer (10x)	AppliChem, Germany
Triton X-100	Sigma-Aldrich, USA
Trizma base	Sigma-Aldrich, USA
TRIzol Reagent	Life Technologies, USA
Tryptone, Bacto	BD Biosciences, USA
Tween-20	VWR International, Germany
Yeast extract, Bacto	BD Biosciences, USA

2.1.4 Buffers, Solutions and Media

All solutions were prepared with Milli-Q water or RNAse free water (for RNA). The buffers, solutions and media are listed in Table 4 and 5. DMEM/F12 was supplemented to make MCDB170 medium as listed in Table 6. MEGM Mammary Epithelial Cell Growth Medium was supplemented as listed in Table 7 to make MM4 medium. Ingredients of SILAC heavy and light medium are listed in Table 8 and 9.

Table 4: Solutions.

Solution	Composition
Ampicillin (1000x)	100 mg ampicillin in 1 ml water
Blocking solution	5% (w/v) milk powder, 1x PBS, 0.1% (v/v) Tween-20
DNA loading buffer (6x)	30% (v/v) glycerol, 0.25% (w/v) orange G
Extraction buffer	20 mM Tris-HCl pH 8.0, 140 mM KCl, 0.5 mMDTT, 5 mM MgCl2, 0.5% Nonidet-P40, 0.1 mg/ml cycloheximide, 0.5 mg/ml heparin
Glycerol buffer	20 mM Tris (pH 7.5, RNase-free), 75 mM NaCl, 0.5 mM EDTA, 50% (v/v) glycerol
LB agar	10 g tryptone, 5 g yeast extract, 10 g NaCl, 15 g agar in 1 l water (pH 7.0, adjusted
LB medium	10 g tryptone, 5 g yeast extract, 10 g NaCl in 1 l water (pH 7.0, adjusted with NaOH)
Lysis buffer	10 mM Tris-HCl (pH 7.5, RNase-free), 150 mM NaCl, 0.15% (v/v) IGEPAL CA-630
Nuclear lysis buffer	10 mM HEPES (pH 7.6), 7.5 mM MgCl2, 0.2 mM EDTA, 0.3 M NaCl
PBST	1x PBS, 0.1% (v/v) Tween-20
Ponceau S	1g Ponceau S, 50 ml acetic acid, up to 1l water
Ripa buffer	25 mM Tris (pH 7.6), 150 mM NaCl, 1% (v/v) IGEPAL CA-630, 1% SDS
SDS Lysis buffer	4% SDS, 0.1M DTT, 0.1M Tris ph8
Sucrose buffer	10 mM Tris-HCl (pH 7.5, RNase-free), 150 mM NaCl, 24% (w/v) sucrose
Transfer buffer	3.03 g Trizma base, 14.4 g glycine, 140 ml methanol in 1 l water

Table 5: Media.

Media	**Manufacturer, Country**
DMEM, High Glucose, Pyruvate	Life Technologies, USA
DMEM/F12	Life Technologies, USA
MEGM Mammary Epithelial Cell Growth Medium	Lonza, USA
Opti-MEM Reduced Serum Medium	Life Technologies, USA

Table 6: MCDB170 medium.

Supplement	**Final conc.**
Insulin	5.0 μg/ml

Table 6 continued

Supplement	**Final conc.**
Hydrocortisone	0.5 μg/ml
EGF	5.0 ng/ml
Transferrin	5.0 μg/ml
Isoproterenol (IP)	10^{-5} M
Glutamine	2.0 mM

Table 7: MM4 medium.

Supplement	**Final conc.**
Insulin	10 μg/ml
Tri-iodothyronine	10 nM
b-estradiol	1.0 nM
Hydrocortisone	0.1 μg/ml
Fetal calf serum	0.50%
EGF	5 ng/ml
Glutamine	2 mM

Table 8: Light SILAC medium.

Light ^{12}C medium (550 ml)
500.0 ml DMEM
50.0 ml dialyzed FBS
5.0 ml Pen-Strep
46,2 mg ^{12}C Arginin –HCl (0.398mM)
80 mg ^{12}C Lysine –HCl (0.798mM)
110 mg ^{12}C Proline (1.74mM)
2mM glutamine

Table 9: Heavy SILAC medium.

Heavy ^{13}C medium (550 ml)
500.0 ml DMEM
50.0 ml dialyzed FBS
5.0 ml Penicillin Streptomycin
46,2 mg ^{13}C Arginin –HCl (0.398mM)
80 mg ^{13}C Lysine –HCl (0.798mM)
110 mg ^{12}C Proline (1.74mM)
2mM glutamine

2.1.5 Molecular biology kits

All molecular biology kits used in this thesis are listed in Table 9.

Table 9: Kits.

Kit	Manufacturer, Country
BD Annexin V FITC	BD Biosciences, USA
Bicinchoninic Acid Kit for Protein Determination	Sigma-Aldrich, USA
Caspase-Glo3/7 Assay Kit	Promega, USA
Dual-Glo Luciferase Assay System	Promega, USA
Fast SYBR Green Master Mix	Life Technologies, USA
High-Capacity RNA-to-cDNA Kit	Life Technologies, USA
Mix & Go E. coli Transformation Kit & Buffer Set	Zymo Research, USA
QIAGEN Plasmid Mini Kit	Qiagen, Germany
QIAprep Spin Miniprep Kit	Qiagen, Germany
QIAquick Gel Extraction Kit	Qiagen, Germany
QIAquick PCR Purification Kit	Qiagen, Germany
SuperSignal West Dura Chemiluminescent Substrate	Thermo Fisher Scientific, USA
TNT Quick Coupled Transcription/Translation Systems	Promega, USA
TOPO TA Cloning Kit for Subcloning	Life Technologies, USA

2.1.6 Enzymes

All enzymes used in this work are listed in Table 10.

Table 10: Enzymes.

Enzyme	Manufacturer, Country
Antarctic Phosphatase BamHI	New England Biolabs, USA
BglII	New England Biolabs, USA
DNase I	New England Biolabs, USA
HindII	New England Biolabs, USA
MluI	New England Biolabs, USA
Phusion High-Fidelity DNA Polymerase	Thermo Fisher Scientific, USA
T4 DNA Ligase	New England Biolabs, USA
Taq DNA Polymerase	Life Technologies, USA

2.1.7 Plasmids

All plasmids used in this work are listed in Table 11.

Table 11: Plasmids.

Plasmid	Manufacturer, Country
pcDNA3	Invitrogen, USA
pCR2.1-TOPO	Life Technologies, USA
pGL3-Basic Vector	Promega, USA
pRL-TK Vector	Promega, USA

2.1.8 Antibodies

All antibodies used in this thesis are listed in Table 12.

Table 12: Antibodies.

Antibody	Manufacturer, Country
anti-puromycin antibody	Kerafast, USA
c-Myc (sc-40)	Santa Cruz Biotechnology, USA
p16 (sc-468)	Santa Cruz Biotechnology, USA
phospho-c-Myc (ab32029)	Abcam, USA

2.1.9 Oligonucleotides

All primers and dsiRNAs used in this work are listed in Table 13-15.

Table 13: Cloning primers.

Name	Sequence 5'-3'
PARROT_prom_fw_166bp	TGAAGATCTAAGCTCCCAGAAATGTCAGC
PARROT_prom_fw_223bp	TGAAGATCTCCACTCAGTTATTTTTGTCTCTCA
PARROT_prom_fw_298bp	TGAAGATCTAGGAAATTGGTCAAGGTTGC
PARROT_prom_fw_404bp	TGAAGATCTCTGACCACCTGTTGAGCTGT
PARROT_prom_rv	TGAAAGCTTCCACCCAGAGTCAAGGAGAC

Table 14: qPCR primers.

Name	Forward (5'-3')	Reverse (5'-3')
7SL	GTCAAAACTCCCGTGCTGAT	GCTGGAGTGCAGTGGCTATT
c-Myc	GCTGCTTAGACGCTGGATTT	CCTCCTCGTCGCAGTAGAAA
EFHD2	CATGATCAAGGAGGTGGATG	CGCTGTCCTCCTGAAGCTC
GAPDH	GCTCTCTGCTCCTCCTGT	ACGACCAAATCCGTTGACTC
PARROT	CAGAACAGAGCCACCTCCAG	GCACCGTCTGTTGTTCATTC
pre-GAPDH	CAATGACCCCTTCATTGACC	GGCTCACCATGTAGCACTCA
β-actin	CGACAGGATGCAGAAGGAG	GTACTTGCGCTCAGGAGGAG

Table 15: dsiRNA oligos.

Name	Sense (5'-3')	Antisense (5'-3')
siRNA1 PARROT	GAAUGAAAGCACAGCACCAUCCU GGAA	CCAGGAUGGUGCUGUGCUUUC AUTC
siRNA2 PARROT	GCUGAAUCAAGAUGCUGACUUCA GCAC	GCUGAAGUCAGCAUCUUGAUU CAGC

2.1.10 Cell lines and bacteria

All cell lines and the bacterial strain used in this thesis are listed in Table 16 and Table 17.

Table 16: Cell lines.

Cell Line	**Origin**
A549	Human lung carcinoma cell line
GM12878	Lymphoblastoid cell line (LCL)
H1ES	Human embryonic stem cells
HEK293	Human embryonic kidney
HeLa	Cervical cancer cell line
HepG2	Liver carcinoma
HMEC	Human mammary epithelial cells
HSMM	Human skeletal muscle myoblasts
HUVEC	Human umbilical vein enodothelial
K562	Erythrocytic leukemia
MCF7	Breast cancer cell line
NHEK	Normal epidermal keratinocytes
NHLF	Normal human lung fibroblasts

Table 17: Bacteria strains.

Bacterial strain	**Manufacturer, Country**
E. coli Zymo DH5-alpha	Zymo Research, USA

2.1.11 Software

The software used in this thesis is listed in Table 18.

Table 18: Software.

Software	Source
BEDTools	(Quinlan and Hall, 2010)
IPA	www.qiagen.com/ingenuity
Primer3	(Untergasser et al., 2012)
SDS Software 2.2	Applied Biosystems, USA
UCSC genome browser	(Kent et al., 2002)

2.2 Methods

2.2.1 *In silico* methods

2.2.1.1 Enhancer prediction

PreSTIGE methodology was developed as described previously (Corradin et al., 2014) with modifications. PreSTIGE predicts enhancers and their targets by first finding PCGs with tissue-specific increased expression and assumes that these are the targets of tissue-specific enhancers. In a cell line in which a PCG has an increased expression, it searches for H3K4me1 domains within the specified domain size (+/-100kb) surrounding the TSS. Normalized H3K4me1–enhancer signal had to be high above background (>10) for an interaction to be predicted in a given cell line. Specificity of both the enhancer and the target gene were determined by calculating the Shannon entropy Q scores. The original version of this method takes into account CTCF binding sites and expands the domain until the first CTCF binding site if no CTCF binding site is found within the search domain. In this thesis PreSTIGE was modified to address the association of long ncRNAs to enhancers. These modifications are: 1) CTCF domains are not considered 2) the domain size is expanded to 200 kb surrounding the TSS of the PCG and 3) all enhancers overlapping a long ncRNA predicted to target that long ncRNA were filtered out from the enhancer dataset used in the subsequent analysis.

2.2.1.2 Overlap between long ncRNAs and enhancers

Annotated long ncRNAs were obtained from (Derrien, 2012). They were further filtered for long ncRNAs that do not overlap PCGs by using UCSC genome browser BEDTools (Quinlan and Hall, 2010). Intersection of long ncRNAs with cell type

specific PreSTIGE enhancers, and vice versa, was performed by using BEDTools (Quinlan and Hall, 2010). Minimal overlap for this analysis was set at one nucleotide.

Previously predicted HeLa enhancers were obtained from (Heintzman, 2009). Overlap of these enhancers with HeLa enhancers predicted by PreSTIGE, and the intersection of long ncRNAs with these enhancers was performed using BEDTools (Quinlan and Hall, 2010). Minimal overlap for this analysis was set at one nucleotide.

2.2.1.3 Pol II association analysis

Differentially expressed long ncRNAs in keratinocyte differentiation were obtained from (Orom, 2010). ENCODE Chip-seq data for Pol II was used. Pol II signal and long ncRNAs were intersected using BEDTools (Quinlan and Hall, 2010).

2.2.1.4 Expression analysis of long ncRNAs

RNA-seq data for long ncRNAs was obtained from (Derrien, 2012). Expression value-RPKM (reads per kilobase of transcript per million reads mapped) of each gene was normalized to the average RPKM across the 11 cell lines used in this study.

2.2.1.5 Determining H3K4me1 and H3K4me3 levels at selected enhancers

ENCODE HeLa broad.peaks were used to determine the H3K4me1 and H3K4me3 levels for each enhancer predicted by either PreSTIGE or (Heintzman, 2009). Enhancers that have no histone mark peak were assigned a pseudo signal value: in the case of H3K4me3 pseudo value was set at 0.5 and in the case of H3K4me1 pseudo value was set at 1. The mean signal value was calculated in cases in which one enhancer overlapped more than one peak. Evgenia Ntini performed this analysis.

2.2.1.6 RNA-sequencing data analysis

RNA-seq data was analyzed with the TRUP pipeline (Fernandez-Cuesta et al., 2015) and differential expression analysis was performed in edgeR Bioconductor package (Robinson et al., 2010). The cutoff was set at FDR<0.05 for differentially expressed genes. This analysis was performed by Ruping Sun.

2.2.1.7 Mass-spectrometry data analysis

Raw MS data were processed with MaxQuant v1.4.1.2 and searched against the UniProtKB human proteome database. FDR of 0.01 for proteins and peptides and a minimum peptide length of 7 amino acids were required. A maximum of two missed

cleavages was allowed for the tryptic digest. Following SILAC modifications were used: $^{13}C_6{}^{15}N_4$-arginine and $^{13}C_6{}^{15}N_2$-lysine. Cysteine carbamidomethylation was set as fixed modification, while N-terminal acetylation and methionine oxidation were set as variable modifications in all runs, phosphorylation of serine, threonine and tyrosine were set as variable modifications in the according enriched fractions. David Meierhofer performed this analysis.

The cutoff for the ratio of light (knock-down treatment) to heavy (control) was set to be >1.2 for upregulated and <0.8 for downregulated to filter for differentially regulated proteins or phosphoproteins. Additionally, we also required that the ratio of knock-down/control to be >1.2 or <0.8 compared to the ratio of control knock-down(NK)/control.

2.2.1.8 Gene ontology analysis

The Gene Ontology (GO) provides a system for classifying gene products into categories by integrating all functions of gene products described in the literature. The terms are grouped into three categories: molecular function, biological process and cellular process. GO can be used to functionally profile selected set of genes to determine which GO terms appear more frequently than would be expected by chance. In this thesis QIAGEN's Ingenuity Pathway Analysis (IPA) gene ontology software was used. In addition to finding enriched GO terms in the dataset, IPA can predict an upstream regulator, an affected protein that can potentially be responsible for differential abundance of proteins in the dataset based on published regulatory relationships between proteins.

PreSTIGE predicted targets of enhancers that overlap long ncRNAs were subjected to IPA analysis. Also, all transcripts, proteins and phosphoproteins affected by the depletion of PARROT were subjected to IPA analysis. Upstream pathway analysis was also performed by IPA. Networks shown in **Figure 14A** were generated by the use of IPA.

2.2.1.9 Statistical analysis

Statistical analysis was performed using two-tailed Student's T-test (for **Figure 8B-C**, **Figure 11**, **Figure 12A-C, Figure 15C,E and G**) or using paired Mann-Whitney-Wilcoxon test (in **Figure 5**, **Figure 8D-E**) comparing expression of the predicted long

ncRNA overlapping enhancers to expression of all long ncRNAs prior to normalization of RNA sequencing data.

2.2.2 DNA/RNA molecular biology methods

2.2.2.1 Polymerase chain reaction

Polymerase chain reaction (PCR) was used to amplify genomic regions of interest. PCR reactions were prepared according to the Phusion High-Fidelity DNA Polymerase manufacturer's protocol (listed in Table 19). The amplification was performed using the PCR program as in table 20.

Table 19. Composition of PCR reaction.

Component	**20 µl Reaction**
Nuclease-free water	to 20 µl
5X Phusion HF or GC Buffer	4 µl
10 mM dNTPs	0.4 µl
10 µM Forward Primer	1 µl
10 µM Reverse Primer	1 µl
Template DNA	50-200 ng
DMSO (optional)	0.6 µl
Phusion DNA Polymerase	0.2 µl

Table 20. PCR settings.

STEP	**TEMP**	**TIME**
Initial Denaturation	98°C	30 seconds
	98°C	5-10 seconds
25-35 Cycles	45-72°C	10-30 seconds
	72°C	15-30 seconds per kb
Final Extension	72°C	5-10 minutes

2.2.2.2 Agarose gel electrophoresis

In order to determine the amount of PCR product and the specificity of primers used for PCR or restriction digestion efficiency, PCR products or plasmids were separated by agarose gel electrophoresis. For that, each DNA sample was mixed with 6x DNA loading buffer and loaded onto a 1% agarose gel supplemented with SybrSafe for visualization. Commercial DNA ladder was loaded in parallel as a size control.

Electrophoresis was performed in 1x TBE buffer at 110V and DNA was visualized by the use of UV illuminator (254 nm).

2.2.2.3 Extraction of DNA from agarose gels

Surgical blade was used to excise from agarose gels DNA of interest while agarose gel was exposed to minimum UV illumination. QIAquick Gel Extraction Kit was used to extract DNA from the gel following the manufacturer's instructions and was eluted in sterile water.

2.2.2.4 Determination of nucleic acid concentration and purity

Concentration of DNA or RNA was determined via photometrical measurements (extinction at 260 nm) using a Nanodrop spectrophotometer. The ratio of absorbance measured at 260 nm and 280 nm was used to determine the purity of the sample. For a pure DNA sample this ratio is ~ 2 and for a pure RNA sample the ratio is ~ 1.8.

2.2.2.5 Topo TA cloning

Full length DNA sequence of PARROT was cloned into pCR2.1-TOPO vector. TOPO TA Cloning Kit allows the ligation of A-tailed PCR products into the pCR2.1-TOPO vector. This vector is linearized and harbors single 3'-T overhangs and has a covalently bound topoisomerase allowing the ligation of PCR product of interest in a single reaction in 10 minutes.

To enable cloning into the pCR2.1-TOPO vector PCR products amplified by Phusion DNA polymerase were A-tailed. Taq polymerase catalyzes the addition of an adenine nucleotide to the 3' end of DNA and was therefore used for A-tailing. Mixture (Table 21) was incubated at 72◦C for 20 min and TOPO TA cloning was performed according to manufacturer's instructions.

Table 21. PolyA tailing of PCR products

Compound	Amount
PCR buffer (10x)	2µl
MgCl2 (50 mM)	0.6 µl
dATP (1 mM)	2µl
Taq DNA polymerase (5 U/µl)	0.5 µl
dd H2O	to 20µl

2.2.2.6 Restrictional digestion

Purified PCR products with added restriction sites or pGL3 vector were digested with BglII and HindIII restriction enzymes to allow directional cloning and to produce compatible ends for ligation. Buffers, components and the incubation temperature used for the reaction were selected based on manufacturer's recommendation. All restriction enzyme digestion reactions were performed for 20 minutes at 37°C for optimal enzyme activity. Subsequently, the enzymes were inactivated at 65°C for 20 minutes. Following digestion, PCR products were purified from restriction enzymes and small DNA fragments using the QIAquick PCR Purification Kit according to the manufacturer's instructions and plasmids were purified by agarose gel electrophoresis followed by gel extraction with a QIAquick Gel Extraction Kit.

2.2.2.7 Dephosphorilation of digested plasmids

To prevent the re-ligation of cohesive ends produced by restriction digestion of a plasmid, a dephosphorylation reaction was performed right after the purification of the digested plasmid. This was achieved by the use of Antarctic phosphatase according to the manufacturer's instructions. 20 µl reaction containing 17 µl of vector, 2 µl of 10x Antarctic phosphatase buffer and 1µl of Antarctic phosphatase was incubated at 37°C for 20 min after which the enzyme was heat-inactivate by a 10 min incubation at 65°C.

2.2.2.8 Ligation

The presence of cohesive ends in both the PCR product and the plasmid allow directional ligation. Typically, 3:1 molar ratio of insert to plasmid was used for a ligation reaction. According to the manufacturer's instruction, 10 µl ligation reaction containing 1µl of T4 ligase (400 U), 1µl of T4 ligase buffer, vector and insert was incubated at room-temperature for 3 h or at 16°C overnight. Ligation reaction was used for transformation of *E. coli* cells.

2.2.2.9 Generation of competent *E. coli* bacteria

Mix & Go E. coli Transformation Kit was used to generate competent *E. coli* bacteria according to the manufacturer's instruction.

2.2.2.10 Transformation of *E. coli* bacteria

Chemically competent *E. coli* bacteria were thawed on ice for 5 min after which 5 µl of ligation reaction (Section 2.2.2.6) was gently added. The mixture was incubated on

ice for 10 min after which bacteria were spread on warmed agar plates containing 100 µg/ml of ampicillin and incubated at 37°C overnight.

2.2.2.11 Colony PCR

Colonies of transformed *E. coli* bacteria were tested for the presence of desired inserts by colony PCR. To this end, in a typical PCR reaction containing the primer pair used to amplify the insert or primers that annealed in the vector backbone (section 2.2.2.1). A small fraction of the *E. coli* colony was added as DNA template.

2.2.2.12 Plasmid extraction

E. coli colonies containing the insert of insert determined by colony PCR, were cultured in the LB medium containing ampicillin (100 µg/ml) at 37°C with shaking at 300 rpm overnight. QIAGEN Plasmid Mini Kit was used for the extraction and the purification of plasmids from bacterial cultures according to the manufacturer's protocol. The QIAprep Spin Miniprep Kit was used for plasmid extraction if the plasmids were subsequently used for sequencing or further cloning.

2.2.2.13 Sanger sequencing

Sanger sequencing was employed for verification of successful cloning of the desired insert. 10µl reaction contained 80-100 ng/µl of plasmid DNA, 5µM primer annealing to the plasmid backbone and water. GATC Biotech AG performed the sequencing.

2.2.2.14 Extraction of RNA from human cells

Trizol reagent was used for extracting RNA from cultured human cells. First, growth medium was removed and cells were rinsed with 1x PBS and Trizol reagent was added directly to the cells (the amount was adjusted to the number of cells) for cellular lysis. RNA was extracted according to the manufacturer's protocol. GlycoBlue was added in the precipitation step for visualization of RNA pellet. RNA pellet was washed, re-suspended in RNase free water and stored at -20°C (short term) or -80°C (long term).

2.2.2.15 DNAse I treatment of RNA

To remove the traces of co-purified DNA from RNA isolated with Trizol reagent, RNA was treated with DNase I. 2 µg of RNA was incubated with 2 µl of DNase I Reaction Buffer (10x), 1 μl of DNase I and water at 37°C for 15 min. DNase I enzyme was inactivated by incubating the sample for 10 min at 75°C with 5 mM EDTA.

2.2.2.16 Reverse transcription

High-Capacity RNA-to-cDNA Kit was used to reverse transcribe RNA to complementary DNA (cDNA). 2 µg of RNA or DNase I treated RNA was reverse transcribed according to the manufacturer's protocol. cDNA was diluted 4 times prior to downstream application.

2.2.2.17 Quantitative real-time PCR

By using sequence-specific primers, the number of copies of RNA reverse transcribed into cDNA can be determined by quantitative real-time PCR (qPCR). qPCR involves the use of fluorescence to detect the threshold cycle (Ct). Ct is the cycle number at which the fluorescent signal of the reaction crosses the threshold and is used for quantification of the amplicon of interest. SYBR Green is a dye that intercalates in the double-stranded DNA. A qPCR machine detects this fluorescence and appropriate software calculates the Ct values from the intensity of the fluorescence.

The Fast SYBR Green Master Mix was used to perform qPCR. The kit contains: AmpliTaq Fast DNA Polymerase designed to allow instant hot start, SYBR Green I dye to enable detection of DNA, deoxynucleotides (dNTPs) as building blocks, uracil-DNA Glycosylase (UDG) to reduce carryover contamination and ROX dye as passive internal reference. Primers were designed in Primer3 software so that the resulting PCR product were 80-120 bp in length and their melting temperature was ~60 °C. In a final 10µl reaction, 5 µl of SYBR Green Master Mix was mixed with 2 µl of cDNA (from 2.2.1.11) 0.5 μl of primers (FW and RV, each 10 μM) and 2.5 μl of water. qPCR was performed in 384-well plates in triplicates. RNA that was not reverse transcribed was used as a contamination control. The amplification was performed using the program listed in Table 22. on a PRISM 7900 HT (ABI) cycler using the standard setting and including a melting curve profile.

Table 22. qPCR settings.

Step	Temperature (°C)	Time (sec)	Cycles
Initial denaturation	95	20	1
Denaturation	95	1	
Annealing & Extension	60	20	40
Melting	95	15	
Melting	60	15	1

The SDS2.2 software was used to determine the Ct for each primer pair. The fold-difference in expression for each gene of interest was determined using the comparative Ct method ($2^{-\Delta\Delta Ct}$ method). The relative expression was calculated by normalizing to the Actin or GAPDH expression level as control housekeeping genes.

2.2.2.18 RNA sequencing

RNA-seq libraries were prepared with the TruSeq RNA Sample Preparation Kit v2. Sequencing was performed on a HiSeq 2000 instrument using paired-end sequencing (2 × 50 bp) at the MPIMG sequencing facility.

2.2.3 Cell culture methods

2.2.3.1 Culturing of human cell lines

HeLa, HEK293 and A549 human cell lines were cultured in complete DMEM medium containing 10% fetal bovine serum (FBS) and 50µg/ml of penicillin–streptomycin. Trypsin- EDTA (0.25%) was used to detach the cells from the cell culture dish. HMEC cells were cultured in a 1:1 mixture of MM4 (Table 7) and MCDB170 (Table 6) medium supplemented with 5μg/ml of cholera toxin, 5% AlbuMAX and 1μM oxytocin as described previously (Garbe et al., 2009). Cells were routinely diluted for maintenance. All cells were cultured at 37 °C with 5% CO_2.

2.2.3.2 Transfection of human cell lines

Depending on the number of cells needed for further analysis, cells were transfected in 6-well plates (for q-PCR analysis and RNA-seq) or in 10 cm dishes (mass-spectrometry and polysome profiling). In a 10 cm dish $1.1x10^6$ cells in 5ml DMEM+FCS (with ought antibiotic/antimycotic) were seeded 6 hours prior to transfection. In tube 1, 550 µl of Opti-MEM was added to 13 µl of the 20µM siRNA stock (50nM final concentration upon transfection). In tube 2, 20 µl/ml of lipofectamine 2000 was added to the desired amount of Opti-MEM and incubated for 5 minutes. Once incubation was finished, 550 µl of the content of tube 2 (lipofectamine+Opti-MEM) was added to tube 1 (siRNA+Opti-MEM), mixed and incubated for 20min. 1 ml of this mixture was added to the cells after 1ml of the medium was remove from the cells. 24h later, 5ml of the medium with antibiotic/antimycotic was added to the cells. Cells were collected 24 (for

q-PCR analysis and RNA-seq), 48 (for polysome profiling) or 72 (for q-PCR analysis and mass-spectrometry) hours after the transfection.

2.2.2.3 Luciferase assay

10000 cells/well (HeLa or HEK293) were plated in 200 µl complete growth medium in 96-well white plates 24 h before transfection. Cells were co-transfected with 0.1 µg pGL3-basic-promoter constructs and 0.02 µg of pRL-TK vector as a control using lipofectamine 2000 (as described in 2.2.3.2) in 50 µl of antibiotics-free growth medium. Luciferase activity was determined by using the Dual-Luciferase Reporter Assay System kit. 24 h after the transfection the medium was replaced with 25 µl of 1x PBS and 25 µl of Dual-Glo reagent was added and incubated for 10 min after which Firefly luciferase activity was determined by using a microplate luminometer. 25 µl of Stop & Glo reagent was added, incubated for 10 min and Renilla reniformis luciferase activity was determined. Firefly luciferase activity was normalized to that of Renilla luciferase. All transfections were carried out in triplicates at least three times.

2.2.3.4 Fractionation of cells

Cells grown on 10 cm dishes were trypsinized, resuspended in complete growth medium and centrifuged for 5 min at 200 g. Then, cells were washed with PBS. Cell pellets were lysed in 400 µl lysis buffer, incubated on ice for 5 min and pipetted up and down 3-4 times before layered on top of 1 ml sucrose buffer followed by a centrifugation for 10 min at 3,000 g at 4°C. The supernatant was taken into new tube representing the cytoplasmic fraction. The residual pellet was gently rinsed with PBS-EDTA (1 mM) before resuspended in 500 µl of glycerol buffer. An equal volume of nuclear lysis buffer was added, the sample was vortexed and incubated on ice for 2 min followed by a centrifugation for 2 min at 14,000 g at 4°C. The supernatant was transferred into a new tube representing the nucleoplasmic fraction. The residual pellet represented the chromatin fraction and was rinsed with PBS-EDTA (1 mM) before resuspended in 1 ml of Trizol using a syringe and needle. For RNA isolation 200 µl of cytoplasmic and nucleoplasmic fractions were resuspended in 1 ml Trizol reagent. All buffers used for fractionation were ice-cold and freshly supplemented with 1 mM DTT, proteinase inhibitors and RNase inhibitor.

2.2.3.5 SILAC labeling of the cells

Untreated cells were labeled with heavy isotopes (Table 9) whereas the cells in which PARROT was knocked-down with siRNAs were grown in medium with light isotopes (Table 8). Three days after the knock-down was performed, protein extracts were collected and mixed in equal ratios.

2.2.3.6 Preparation of protein cellular extracts

Cells were placed on ice, washed twice with ice-cold PBS and RIPA buffer supplemented with proteinase inhibitors was added to the cells (for western blotting) or SDS lysis buffer (for mass-spectrometry). Cells were scraped off the culture dish surface and lysates were transferred into tubes for sonication (5 bursts, minimum intensity). Lysates were centrifuged for 5 min at 14,000 g at 4°C and the supernatant was transferred into fresh reaction tubes. Samples were kept on ice during the procedure, extracts were snap frozen in liquid nitrogen and stored at -20°C.

2.2.3.7 Migration assay

Some cancer cells migrate to the source of nutrients like the blood vessel leading to metastasis. In this thesis, the ability of A549 cells to migrate was assessed in a migration assay by the use of migration chambers.

A549 cells were seeded in a 6 well plate (100000cells/well) 4 h before transfection with siRNAs against PARROT, SnaiI or control siRNAs. 48 hours later cells were starved for 24h in a DMEM medium without FCS. 50000 cells were added in DMEM medium without FCS to a migration chamber placed in a well of 24 well plate containing 750 µl of DMEM with FCS. 14 hours later cells where washed with PBS and cells that did not migrate were removed with a cotton swiper. Migrated cells were fixed with methanol for 15min, washed with water and stained with 0.1% crystal violet. 20 minutes later cells were washed with water, dried and migrating cells were recorded.

2.2.3.8 Crystal violet cellular viability assay

Crystal violet carries a positively charged ammonium ion that binds to the negatively charged phosphate backbone of DNA allowing the determination of the amount of DNA in the cellular population corresponding to the number of viable cells.

HeLa or HMEC cells (10000cells/well) were plated in 96-well plates and transfected with siRNAs or negative control. 72h after the transfection cells where washed with

PBS and fixed/permeabilized with methanol for 15min. Cells were washed with water and stained with 0.1% crystal violet. After 20min cells were washed with water and dried. 50 µl/well of 33% acetic acid that dissolves crystal violet was added and the absorbance at 590nm was measured. All transfections were carried out in triplicates at least three times.

2.2.3.9 Sucrose gradient fractionation of polysomes- polysome profiling

Polysome profiling is a method that allows the separation of polysome-RNA complexes based on their molecular mass. Cycloheximide is used to stall the 80S ribosomes on mRNA in a polysomal state. Centrifugation of cytoplasmic extracts through a linear sucrose gradient allows velocity separation of translation complexes. By quantifying RNA of interest from fractionated velocity separated complexes the translation efficiency can be monitored for an RNA of interest.

Cytoplasmic lysates of HeLa cells were prepared and subjected to centrifugation through linear sucrose gradients (10–50% sucrose) as described previously (Dhamija et al., 2010). SiRNA transfected cells (were rinsed and scraped in ice-cold PBS containing 100μg/mL cycloheximide. All subsequent steps were performed in ice-cold conditions. Cells were pelleted and resuspended in extraction buffer, incubated 10 min on ice and centrifuged for 10 min at 12000g. Approximately 400μL of supernatant was layered onto a 12-ml linear sucrose gradient (10–50% sucrose (w/v) in detergent free extraction buffer) and centrifuged at 4 °C in an SW40Ti rotor at 35,000 rpm for 120 min. The gradients were collected into 12 fractions (1ml each), and absorbance profiles at 260 nm were recorded (ISCO, UA-6 detector). 0.1 volume of 3 M sodium acetate (pH 5.2) and 1 volume of isopropyl alcohol were added to the probes for overnight precipitation at −20 °C. RNA was purified, RNA concentration was determined, and the samples were stored at −80 °C. This experiment was performed by Sonam Dhamija.

2.2.4 Protein biochemical methods

2.2.4.1 Determination of protein concentration

Protein concentrations of protein cellular extracts were determined by the use of Bicinchoninic Acid Kit for Protein Determination according to the manufacturer's instructions. This assay is based on the reduction of Cu2+ to Cu1+ within some amino

acids and peptide bonds in alkaline conditions. Bicinchoninic acid forms a stable purple-blue complex with Cu1+ in alkaline conditions enabling monitoring of the protein concentration. For each experiment, a standard curve was generated by measuring the concentration of a protein with a known concentration, bovine serum albumin (BSA). Each sample was pipetted in duplicates and at two different dilutions. Absorbance was measured at 562 nm using a microplate reader. Absorbance measured for BSA was used to create a linear regression curve and calculate protein concentrations in the samples.

2.2.4.2 SDS polyacrylamide gel electrophoresis

Cellular protein extracts were separated based on their molecular weight by SDS Polyacrylamide gel electrophoresis (SDS-PAGE). Prior to loading onto 4-12% Bis-Tris gels, protein extracts were mixed with Roti Load buffer and incubated at 95°C for 5min to denature proteins. Protein ladder was also loaded in the gel as the size marker. Electrophoresis was performed in 1x MOPS buffer at 180 V for 50 min.

2.2.4.3 Western blot

Following SDS-PAGE, gel separated proteins were transferred to the polyvinylidene difluoride (PVDF) membranes for 90 min at 120 V and 4°C in a transfer buffer by using a wet transfer system. The membranes were blocked in blocking solution for 1 h with agitation. After washing, membranes were incubated for 1 h at room temperature or overnight at 4°C with primary antibodies diluted 1:1000 (or in case of tubulin 1:2000) in a blocking solution. The membrane was washed vigorously 3 times with PBST and incubated with an appropriate secondary horseradish peroxidase-conjugated antibody diluted 1:10000 in a blocking solution for 1h. Membrane was washed 3 times with PBST and the HRP signal was developed using ECL reagent (Pierce). The protein of interest was visualized using a chemiluminescence imager.

2.2.4.4 Puromycin translation assay

Puromycin is an antibiotic that causes premature chain termination during translation. It resembles the 3' end of the aminoacylated tRNA and is incorporated in the nascent peptide chain causing the formation of a puromycylated nascent chain and premature peptide chain release. This feature is being used to detect the efficiency of translation based on the amount of incorporated puromycin in the nascent peptide chain. The

amount of puromycin incorporated in the newly synthesized chain can be detected with an Anti-puromycin antibody.

Three days after the knockdown 1µM puromycin was added to the cells for 30 min. Cells were harvested in RIPA buffer supplemented with protease inhibitors, sonicated and boiled for 5min at 95°C in Laemmli buffer. 15µg of protein was loaded onto 4-12% Bis-Tris gradient gel and transferred onto a PVDF membrane which was blocked in 5% milk and incubated overnight at 4°C with 1µg/ml anti-puromycin antibody. The signal was detected as in a standard western blot procedure (section 2.2.3.3). The membrane was stained with Ponceau S protein dye prior to blocking as the loading control.

2.2.4.5 *In vitro* translation assay

The TNT Quick Coupled Transcription/Translation Systems was used for *in vitro* translation. This kit allows single-tube, coupled transcription/translation reactions for eukaryotic *in vitro* translation by combining the RNA polymerase, nucleotides, salts and Recombinant RNasin Ribonuclease Inhibitor with the reticulocyte lysate solution. According to the manufacturer's instructions 0.2µg of circular plasmid DNA containing a T7 promoter, was added to an aliquot of the TNT Quick Master Mix and incubated in a 10µl reaction volume for 75 minutes at 30°C. Additionally, unlabeled methionine and biotinylated lysine was added to the translation reaction as a precharged ε-labeled biotinylated lysine-tRNA complex which incorporates into nascent proteins during translation enabling non-radioactive detection. The synthesized proteins were analyzed by SDS-polyacrylamide gel electrophoresis (SDS-PAGE) and detected colorimetrically with a Transcend Colorimetric kit according to the manufacturer's instructions. As a control luciferase-encoding control plasmid was used.

2.2.4.6 Mass-spectrometry

HeLa cells were grown in SILAC medium (see section 2.2.3.4). All steps were performed as previously reported (Gielisch and Meierhofer, 2015). Cells were lysed and digested by trypsin under denaturing conditions. Peptides were purified with C18 columns and fractionated by a strong cation exchange column. 5% of resulting peptides in each of the 11 fractions were used for proteome profiling. The remaining peptides were further enriched for phosphor-peptides by the use of TiO_2 columns. Each sample

fraction was dissolved in 3 μL of 5% ACN and 2% FA and analyzed by liquid chromatography-tandem mass spectrometry (LC-MS/MS).LC–MS/MS was carried out by nanoflow reverse phase liquid chromatography coupled to a Q-Exactive Plus Orbitrap mass spectrometer. This was performed by MPIMG mass-spectrometry facility.

2.2.4.7 Caspase3/7 activity assay

Caspase3/7 activity assay was performed by the use of Caspase-Glo3/7 Assay Kit according to manufacturer's instructions. Caspase-3 and caspase-7 are proteases that have key effector roles in apoptosis in mammalian cells. The assay provides a luminogenic caspase-3/7 substrate. Addition of the kit reagent leads to cell lysis, followed by caspase cleavage of the luminogenic substrate and generation of a luminescent signal that can be detected by using a microplate luminometer. This signal corresponds to the number of apoptotic cells.

2.2.4.8 Cell cycle analysis by flow cytometry

By using a florescent DNA binding due- propidium iodide (PI), the distribution of cells in cell cycle phases can be determined by flow cytometry. Following desired treatment of cells, they were washed with 1X PBS, fixed and permiabilized with 100% ethanol to allow the entrance of PI in the cells. Cells were re-washed and resuspended in PI solution. 15 minutes later they were analyzed on a flow cytometer.

2.2.4.9 Apoptosis analysis by flow cytometry

One of the earlier events of programmed cell death-apoptosis, is translocation of membrane phosphatidylserine (PS) from the inner side of the plasma membrane to the outer side of the plasma membrane. Annexin V has a high affinity for PS, and fluorochrome-labeled Annexin V such as FITCH - Annexin V can be used for the detection of the exposed PS using flow cytometry. To allow for discrimination between the cells in apoptosis and cells in necrosis, cells are additionally stained with PI. This analysis was performed by the use of BD Annexin V FITC Assay kit according to the manufacturer's instruction.

3 RESULTS

Statement of contributions:

The work presented here is the result of collaborative projects. All experiments were performed by myself unless otherwise stated.

Olivia Corradin and Peter C Scacheri developed the PreSTIGE method and modified enhancer prediction criteria to accommodate the needs of this work.

Evgenia Ntini analyzed the occupancy of H3K4me3 and the ratio between H3K4me1 and H3K4me3 mark on selected enhancers.

Maja Gehre cloned the PARROT promoter constructs used in this study.

Sonam Dhamija performed polysome profiling.

Lennart Friis-Hansenprovided tumor samples and planned stomach cancer RNA sequencing experiments.

Deep sequencing of PARROT depleted RNA samples was performed by the MPIMG sequencing facility. Ruping Sun mapped and filtered deep sequencing data and performed differential expression analysis.

Mass-spectrometry analysis was performed by MPIMG Mass Spectrometry facility and David Meierhofer and Sascha Sauer contributed to the data analysis.

Ulf Andersson Ørom conceived the experiments.

3.1 Long ncRNA expression associates with tissue-specific enhancers

3.1.1 PreSTIGE enhancer-target prediction approach

We sought to investigate the interplay between long ncRNAs and enhancers. To this end we used modified PreSTIGE enhancer prediction methodology to identify tissue-specific enhancers and their targets in 11 cell lines (Corradin et al., 2014). The original version of PreSTIGE predicts enhancers by first finding PCGs with tissue-specific increased expression and assumes that these are the targets of enhancers. In the cell line in which a PCG shows increased expression, PreSTIGE looks for the H3K4me1 domains within the 100 kb domain surrounding the TSS of the PCG and defines them as enhancers of those PCG. Thus, tissue-specific enhancers are predicted based on the presence of tissue specific H3K4me1 domains in proximity to their targets - genes with tissue-specific elevated expression. PreSTIGE also takes into account the presence of conserved CTCF binding sites and expands the 100kb search domain surrounding the TSS until the first CTCF binding site is found if there is no CTCF binding within the 100 kb domain (Corradin et al., 2014).

In order to address the interplay between long ncRNAs and tissue specific enhancers, we introduced several modifications into the PreSTIGE methodology (**Figure 4A**). First, we eliminated the CTCF binding sites from the prediction. The main reason for that is the recent research that showed that CTCF does not only act to block an interaction between an enhancer and its target, which was traditionally thought, but organizes global chromatin architecture by mediating intra- and inter-chromosomal contacts enabling interactions between distant enhancers and its targets (Ong and Corces, 2011). Additionally, CTCF binding sites occur very frequently in the human genome and are very frequently found within 200 bp surrounding a TSS (Corradin et al., 2014). Therefore, the addition of this factor would not add a contribution to the prediction. Second, we eliminated from the prediction all ncRNAs as targets of enhancers that overlap with those ncRNAs. This was done in order to avoid biases that would potentially arise from having a long ncRNA transcribed from an enhancer predicted to target that long ncRNA. Third, we expanded the domain size in which we look for a tissue specific enhancer to 200 kb surrounding the TSS of a PCG. We did

that to include recent evidence from chromosome conformation capture techniques showing that promoter-enhancer interaction occur on distances larger than 100 kb (120 kb on average) and also to take into account the length that long ncRNAs occupy in the genome. This modified PreSTIGE enhancer prediction methodology, predicts in total 131,917 enhancer in 11 cell lines (**Figure 4C**).

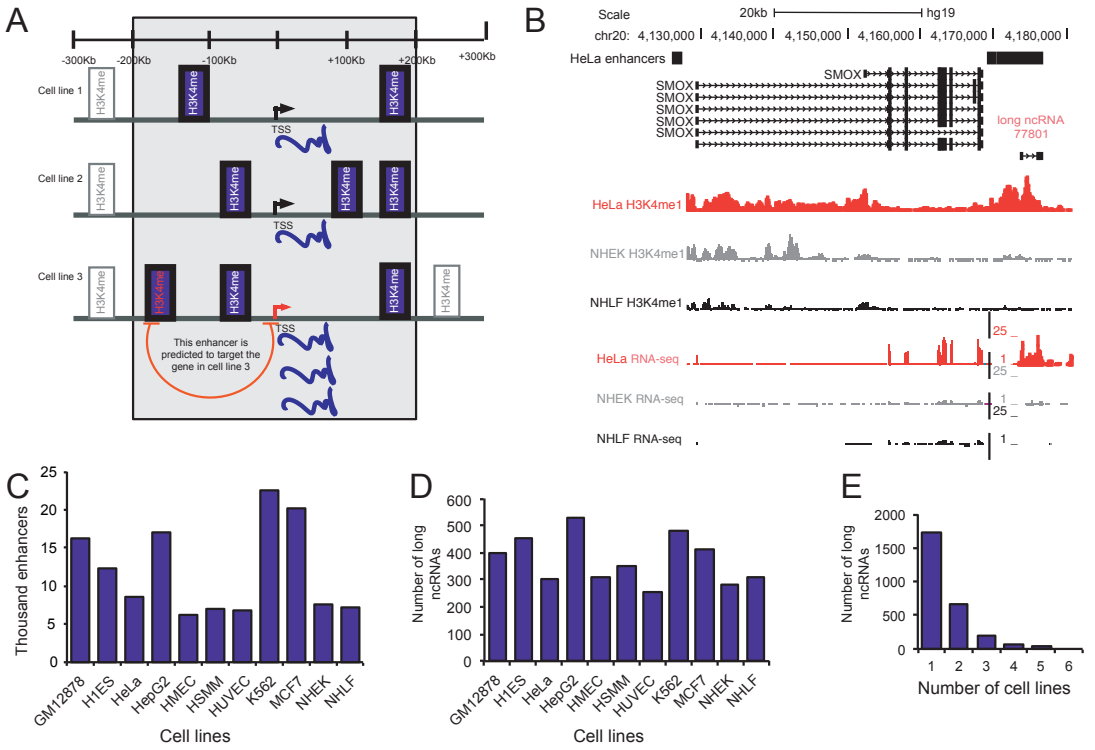

Figure 4. Overlap of long ncRNAs with PreSTIGE predicted tissue-specific enhancers. (**A**) Illustration of the modified PreSTIGE enhancer prediction methodology used in this thesis. PreSTIGE predicts enhancers by finding PCGs with tissue-specific increased expression and associating enhancers to them based on the presence of tissue-specific H3K4me1 domains in the vicinity (200kb surrounding the TSS). (**B**) Example of a long ncRNA overlapping an enhancer predicted in HeLa. This ncRNA is expressed exclusively in Hela in accordance with the predicted activity of an enhancer it overlaps. HeLa specific enhancers are depicted as black bars. ENCODE annotated isoforms of long ncRNAs and PCGs are also shown. ENCODE data for selected representative cell lines are shown for H3K4me1 and RNA sequencing data deposited in the UCSC genome browser. (**C**) The number of cell type specific enhancers in 11 different cell lines predicted by PreSTIGE. (**D**) The number of annotated long ncRNAs that are overlapping enhancer predicted by PreSTIGE. (**E**) The number of cell lines in which a long ncRNA is overlapping a PreSTIGE predicted enhancer.

An example of a long ncRNA overlapping an enhancer predicted by PreSTIGE in HeLa cells is shown in **figure 4B**. We can notice in this figure the cell-type specific presence of the H3K4me1 domain based on which the enhancer was predicted in HeLa. Additionally, judged from the available RNA-seq data, both the long ncRNA that overlaps this enhancer and a predicted target of this enhancer-SMOX are expressed only in the cell line in which the enhancer is predicted to be active. SMOX gene, according to PreSTIGE, is targeted by five enhancers based on the presence of HeLa specific H3K4me1 domains within the 200kb domain surrounding the SMOX TSS. Only one of these five enhancers expresses an annotated long ncRNA and only this one enhancer is taken into account in further analysis.

3.1.2 Long ncRNAs overlapping PreSTIGE predicted enhancers

In order to investigate how many long ncRNAs are expressed from these tissue specific enhancers, I examined the overlap between PreSTIGE predicted enhancers and 9,505 ENCODE annotated long ncRNAs (Derrien, 2012). This analysis revealed that 28% (2,695) of all annotated long ncRNAs overlap with at least 1 nt a tissue specific enhancer in any of 11 cell lines (**Figure 4D**). List of all enhancers that overlap with long ncRNAs that includes the predicted targets and all long ncRNAs overlapping enhancers in each of the 11 cell lines can be found in the (Vucicevic et al., 2015) supplementary material (they are not listed here due to their size). 1736 of these long ncRNAs transcribed from 937 genomic regions overlap an enhancer in only one cell line. To a much lesser extent long ncRNAs overlap an enhancer in more than one cell line (**Figure 4E**).

Of all ENCODE annotated long ncRNAs, 17% are not expressed in any of the 11 cell lines used to predict the enhancers and 80% of the long ncRNAs overlapping an enhancer are expressed in at least one of the 11 cell lines used in this study. To assess whether the expression of long ncRNAs from enhancers is in accordance with the predicted enhancer activity I have established the relative expression of long ncRNAs at each enhancers compared to the average across 11 cell lines used in the study by using RPKM values from (Derrien, 2012). This examination revealed a higher median expression of long ncRNAs associated with tissue-specific predicted enhancers for all cell lines (**Figure 5**, **Figure 6** and **Figure 8D**). *E.g.* in a cell line in which a long

ncRNA overlaps an enhancer, that long ncRNA is expressed at a much higher level in that cell line than in any other cell line. In 5 cell lines (GM12878, H1ES, HSMM, NHEK and HeLa) Wilcoxon rank test shows a significantly higher expression of long ncRNAs in a cell line in which they overlap an enhancer compared to the expression of all long ncRNAs in the particular cell line (**Figure 5, Figure 8D**).

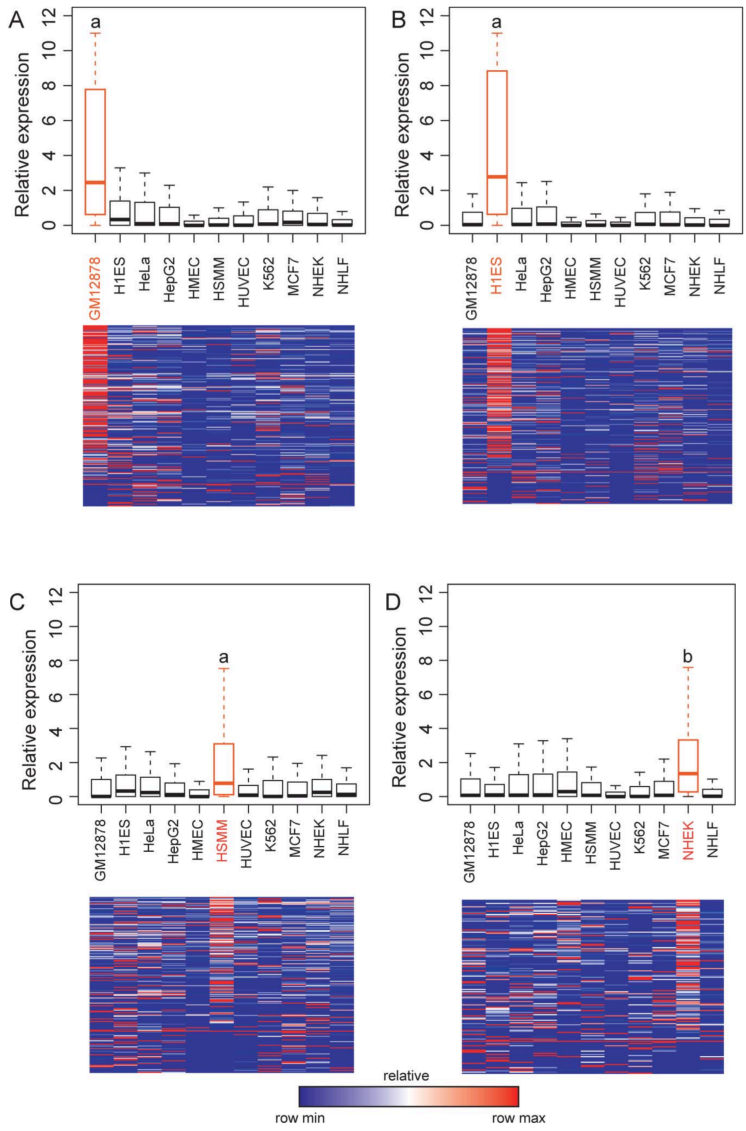

Figure 5. Expression of long ncRNA correlates with the predicted activity of tissue-specific enhancers. In (**A–D**) expression values for all long ncRNAs overlapping a predicted enhancer in (**A**) GM12878, (**B**) H1ES, (**C**) HSMM, and (**D**) NHEK are shown. As bar-plots shown are relative expression values for each long ncRNA normalized to the average expression across all cell lines. As heat maps shown are relative expression values for each long ncRNA normalized for each transcript such that red shows the highest expression and blue shows the lowest expression. Wilcoxon test is used for statistical analysis. a *P*-value < 2.2e-16, b *P*-value 7.9e-15.

For other cell lines, expression of long ncRNAs is higher in a cell line in which they overlap an enhancer. However, their expression, though significant, is not as highly significant as the expression of enhancer derived long ncRNAs in the selected 5 cell lines (**Figure 6**). For an overview of the expression of each long ncRNA that overlaps an enhancer shown are also heat maps. In these heat maps relative expression values for each long ncRNA overlapping an enhancer in each cell line is shown (**Figure 5, Figure 6** and **Figure 8D**).

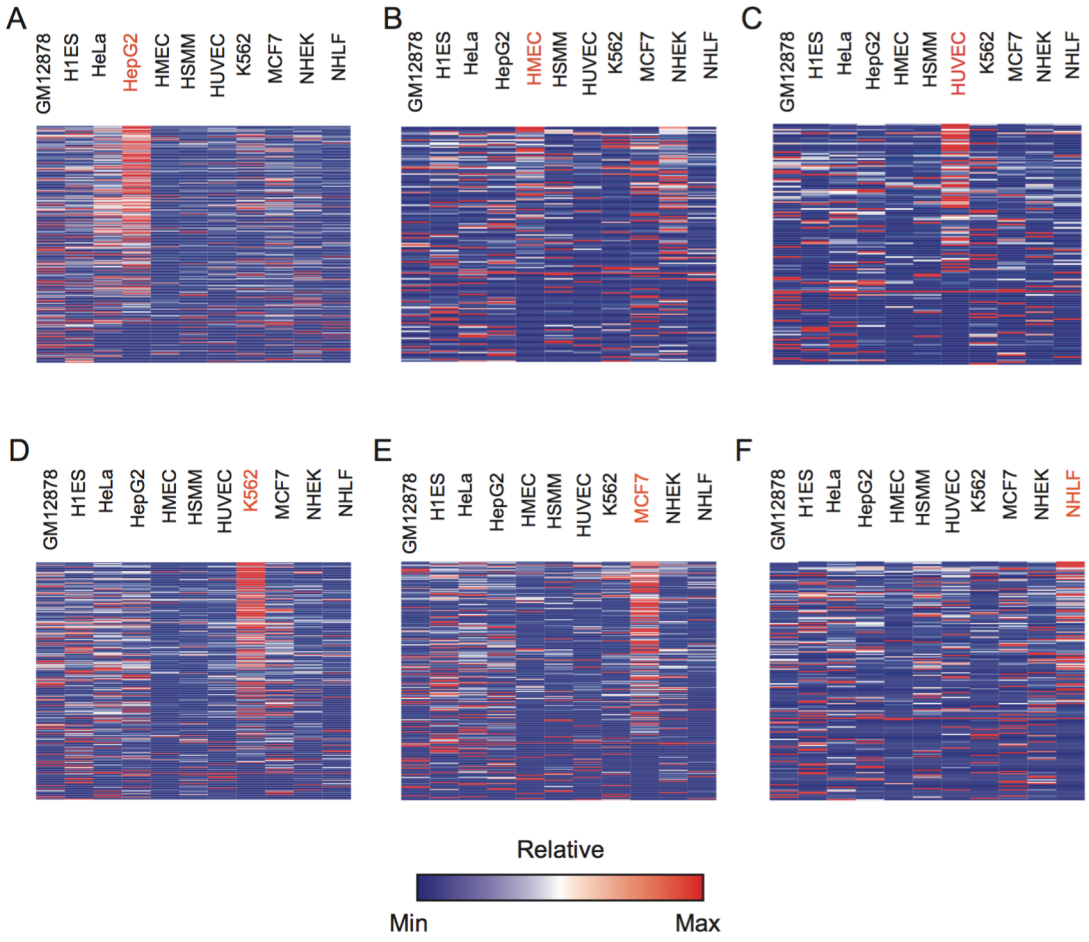

Figure 6. Tissue-specific expression of long ncRNAs overlapping PreSTIGE enhancers in other cell lines. In A-F are shown expression values for all long ncRNAs overlapping a predicted enhancer in (A)HepG2, (B) HMEC, (C) HUVEC, (D) K562, (E) MCF7, and (F) NHLF. As heat maps shown are relative expression values for each long ncRNA normalized for each transcript such that red shows the highest expression and blue shows the lowest expression.

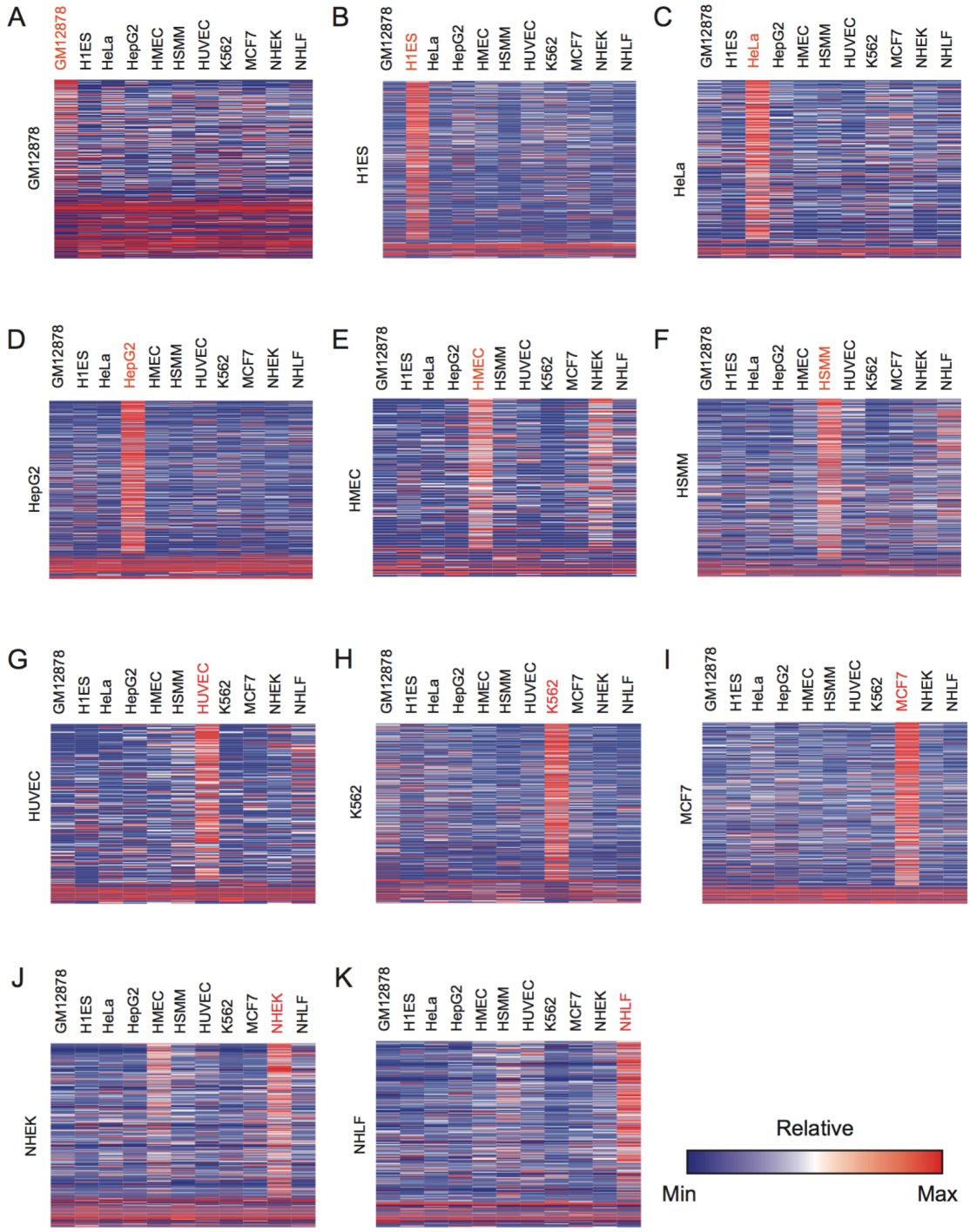

Figure 7. Expression of PreSTIGE predicted protein-coding gene targets of enhancers that overlap long ncRNAs is tissue-specific. In A-K are shown expression values as heatmaps for all PCGs with an enhancer predicted by PreSTIGE that overlaps a long ncRNA. (A) GM12878, (B) H1ES, (C) HeLa, (D) HepG2, (E) HMEC, (F) HSMM, (G) HUVEC, (H) K562, (I) MCF7, (J) NHEK, and (K) NHLF. Heatmaps are normalized for each transcript such that blue shows the lowest expression and red shows the highest expression.

In accordance with the enhancer prediction methodology used, we observed that predicted target genes of enhancers that produce long ncRNAs are also expressed in a tissue specific manner (**Figure 7**). Additionally, predicted PCG targets of enhancers that produce long ncRNAs are enriched in cell-type specific gene ontology terms for GM12878 (Cell-To-Cell Signaling and Interaction, Hematological System Development and Function); H1ES (Embryonic Development, Developmental Disorder); HeLa (Reproductive System Disease); HepG2 (Carbohydrate Metabolism, Lipid Metabolism); K562 (Cardiovascular System Development and Function, Cell-mediated Immune Response); MCF7 (Breast or Ovarian Cancer) and NHEK (Dermatological Diseases and Conditions), further demonstrating the importance of tissue-specific enhancer predictions.

In order to investigate whether these long ncRNAs transcribed from enhancer are potentially functional I examined the correlation between the expression of enhancer derived long ncRNA and the predicted PCG targets. This was achieved by calculating the Pearson correlation for each ncRNA-target gene pair across all 11 cell lines. For 37.2% of gene pairs (long ncRNA~PCG) across all 11 cell lines a significant correlation in expression ($p < 0.05$) is found. The results obtained with this analysis further supports the finding presented here that the subset of tissue specific enhancer expresses long ncRNAs in an activity dependent manner and suggests that these long ncRNAs might mediate their activity.

3.1.3 Comparison of PreSTIGE to other enhancer prediction methods

Many different enhancer prediction methodologies exist and still there is no consensus on how to predict enhancers (Shlyueva et al., 2014). To test the predictive power of PreSTIGE we compared it to a pioneering study by Heintzman and colleagues (Heintzman, 2009). This study investigates the epigenetic landscape of p300 binding sites and based on this investigation they associate H3K4me1 and H3K427ac mark to enhancers and predict enhancers based on the presence of H3K4me1 mark. Additionally, they define promoters based on the presence of H3K4me3 and consider those two marks to be mutually exclusive (Heintzman, 2009). The intersection of PreSTIGE predicted enhancers with enhancers predicted by (Heintzman, 2009) revealed that only 2,125 (24%) PreSTIGE enhancers overlap with 36,552 enhancers

predicted by (Heintzman et al., 2009) in HeLa cells (**Figure 8A**). Additionally, I compared the expression of long ncRNAs that overlap enhancer predicted by those two methodologies in HeLa cells. As in most cell lines used in the study, long ncRNAs that overlap enhancers predicted in HeLa by PreSTIGE are expressed significantly higher in HeLa than in other cell lines (**Figure 8D**). On the other hand, although 1,217 long ncRNAs overlap with (Heintzman, 2009) predicted enhancers, they are expressed at a much lower level and less significantly than the ones overlapping PreSTIGE predicted enhancers (*P*-value 9.496e-07 vs *P*-value < 2.2e-16). Additionally, long ncRNAs overlapping enhancers predicted by (Heintzman, 2009) are not only expressed at a low level in HeLa cells but also in HSMM, NHEK and NHLF (**Figure 8E**). These results clearly show that these two methodologies predict two different subsets of enhancers.

The differences we observe between enhancers predicted in HeLa by PreSTIGE and by Heintzman are possibly a consequence of difference in the enhancer prediction approaches. Unlike PreSTIGE, Heintzman and colleagues exclude all the enhancers that have an H3K4me3 mark present. In this way they potentially eliminate most enhancers that express long ncRNAs since long ncRNA promoters, just like promoters of PCG, are marked with H3K4me3. To test this, we examined the presence of H3K4me3 mark on enhancers predicted by Heintzman and colleagues. In support of this idea, we find a significantly higher H3K4me3 signal at enhancer predicted by (Heintzman, 2009) that overlap with PreSTIGE predicted enhancers than the ones that do not. (**Figure 8B**).

3.1.4 Epigenetic profile of transcribed enhancers

Several studies proposed that the ratio between H3K4me1 and H3K4me3 is being used to discriminate between enhancers and promoters. If the H3K4me1/H3K4me3 ratio is in favor of H3K4me1 mark than an enhancer is predicted and if the ration is in favor of H3K4me3 than a promoter is predicted (Djebali, 2012; Heintzman, 2007; Koch and Andrau, 2011). In order to further investigate PreSTIGE enhancers that produce long ncRNAs we examined the H3K4me1/H3K4me3 ratio on the ones that overlap long ncRNAs and the ones that don't. We find that, on average, enhancers that overlap long ncRNAs display a lower H3K4me1/H3K4me3 ratio than the enhancer that do not overlap with long ncRNAs (**Figure 8C**). These results demonstrate that enhancers that

overlap long ncRNAs have a higher H3K4me3 signal and suggest that these enhancers might have a characteristic epigenetic mark.

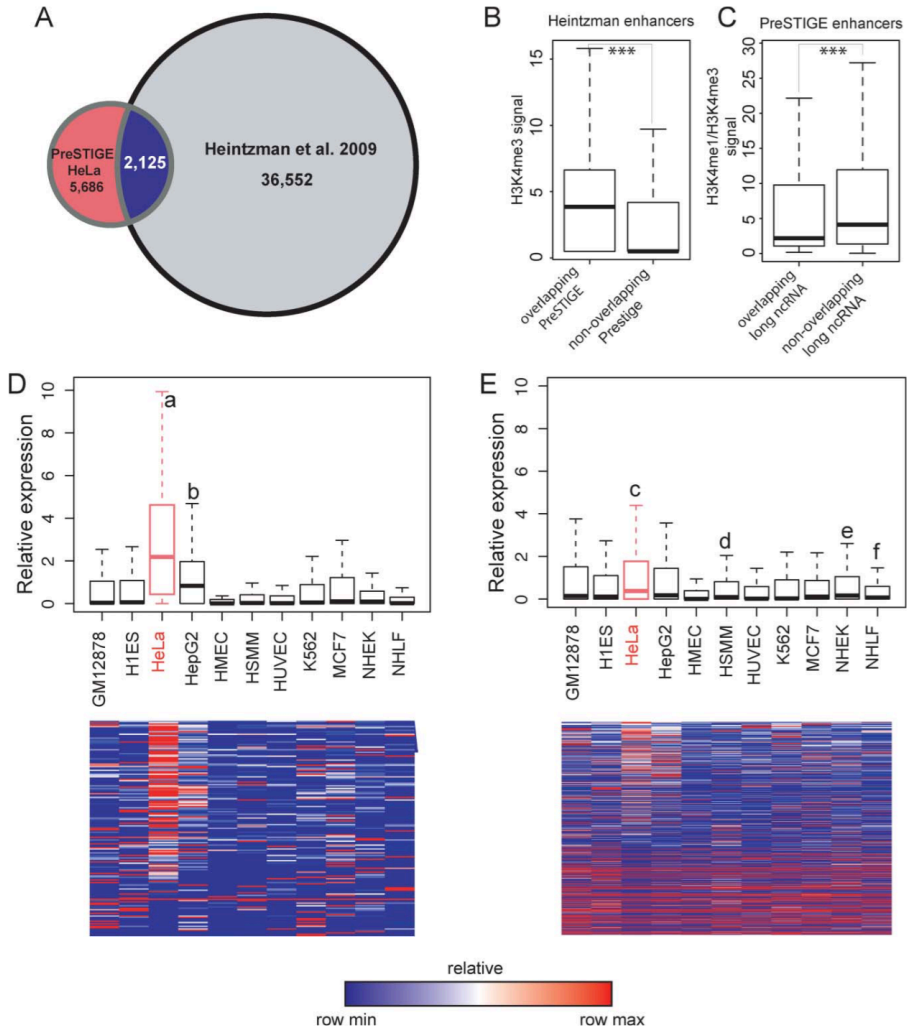

Figure 8. PreSTIGE has a stronger predictive power over previous methods in identifying enhancers that are actively transcribed. (A) Venn diagram showing the overlap between enhancers predicted by(Heintzman, 2009) and by PreSTIGE in HeLa. (Heintzman, 2009) predicts enhancers based on the H3K4me1 domains and excludes the ones that contain H3K4me3 marks whereas PreSTIGE predicts enhancers based on the tissue-specific presence of H3K4me1 mark. (B) Quantification of H3K4me3 signal at enhancers predicted by (Heintzman, 2009) that overlap or do not overlap PreSTIGE predicted enhancers in HeLa, respectively. *** *P*-value < 2.2e-16. (C) H3K4me1 to H3K4me3 ratio for PreSTIGE predicted enhancers that overlap or do not overlap a long ncRNA, respectively. *** *P*-value 0.00039. (D and E) As bar-plots shown are relative expression values for each long ncRNA normalized to the average expression across all cell lines (upper panel). As heat maps shown are relative expression values for each long ncRNA normalized for each transcript such that red shows the highest expression and blue shows the lowest expression (lower panels). (D) Expression of long ncRNAs overlapping enhancers predicted by the modified PreSTIGE method. a *P*-value < 2.2e-16, b *P*-value 9.0e-10. (E) Expression of long ncRNAs overlapping enhancers

figure legend continued from the previous page

predicted by Heintzman et al., 2009. c *P*-value 9.5e-07, d *P*-value 3.9e-05, e *P*-value 6.4e-06 and f *P*-value 1.6e-06. Statistical analysis were done using Mann-Whitney-Wilcoxon test.

3.1.5 Experimental validation of the approached used in the study

A growing number of studies show the involvement of long ncRNAs in mediating enhancer function (Lai, 2013; Lam et al., 2014; Melo et al., 2013; Orom and Shiekhattar, 2013). For instance activating long ncRNAs (RNA-a) have been shown to activate the transcription of their neighboring genes (Orom, 2010). Two of them, long ncRNA called RNA-a3 and long ncRNA-a4 regulating in a positive manner the expression of TAL1 and CMPK1 gene respectively, overlap a PreSTIGE predicted enhancer in K562 (**Figure 9**). The TAL1 gene is indeed predicted to be the targeted by this enhancer since it is expressed at a higher level in this cell line compared to the other 10 cell lines used in the study within the 200kb window, validating the PreSTIGE enhancer-target prediction methodology. The experimental evidences from Pol II ChIA-PET, reporter assays and knock-down experiments demonstrate that expression of CMPK1 is regulated by ncRNA-a4 transcribed from a K562 specific enhancer. However, CMPK1 is not predicted to be the target of PreSTIGE since it is not highly expressed only in K562 cell line but also in other cell lines. This example illustrates how challenging the prediction of long-range regulatory relationships can be and suggests that an even larger fraction of long ncRNAs could be associated to enhancers.

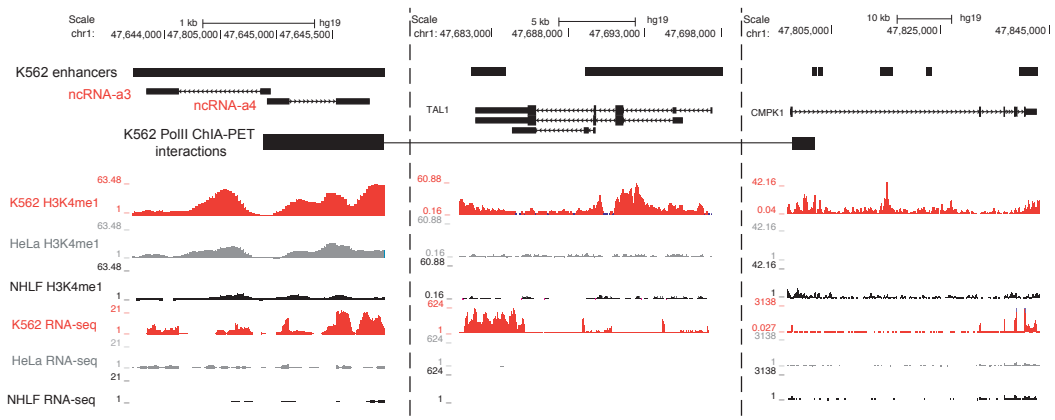

Figure 9. Example of activating ncRNAs overlapping a PreSTIGE predicted enhancer. Activating ncRNAs ncRNA-a3 and ncRNA-a4 shown to activate TAL1 and CMPK1, respectively (Orom, 2010), overlap an enhancer predicted by PreSTIGE in K562 cell line. K562 predicted enhancers are depicted as black boxes. ENCODE

figure legend continued from the previous page

transcript annotation is shown discontinuously (see coordinates and scale bars). Physical interaction experimentally determined by K562 PolII ChIA-PET from Li et al.(Tsuiji et al., 2011) as deposited in the UCSC genome browser is shown. RNA-seq and H3K4me1 ENCODE data as deposited at UCSC genome are shown for three selected cell lines.

3.2 The long non-coding RNA PARROT is an upstream regulator of c-Myc and affects proliferation and translation

3.2.1 Characterization of the long ncRNAs PARROT

Although we are able to detect thousands of long ncRNAs due to the progress in next generation sequencing technologies (Derrien, 2012), finding functional long ncRNAs and functional characterization of low abundant transcripts is still a challenge. To bypass these problems we used a set of 98 long ncRNAs previously reported to be differentially expressed in human primary keratinocytes upon induction of differentiation by 12-O-tetradecanoylphorbol-13-acetate (TPA) (Orom, 2010). Additionally, we performed a Pol II association analysis on these 98 long ncRNAs in HeLa and HEK293 cell lines. This analysis showed that among 98 long ncRNAs responding to TPA treatment 6 are bound by Pol II in both cell lines, 9 are bound by Pol II only in HeLa cells, and 6 are bound by Pol II in HEK293 cell line only (**Figure 10A**). Although several long ncRNAs are bound by Pol II in one cell line only, further examination of the Pol II binding showed that some of them are clearly bound by high levels of Pol II in one cell line only. One of them is a long ncRNA named for its function- Proliferation Associated RNA and Regulator Of Translation (PARROT) (**Figure 10B**). High cell-type specific Pol II association to PARROT is recapitulated in its expression levels determined by qPCR - PARROT is highly expressed in HeLa cells and it is undetectable in HEK293 cells (**Figure 10C**).

To gain insights into the transcriptional mechanism that result in differential expression patterns of PARROT transcript, we cloned the upstream promoter region of PARROT ncRNA in a luciferase reporter vector and examined the luciferase activity (**Figure 10D**). Whereas in HeLa cell line we observe high induction of luciferase expression,

we detect no induction of luciferase in HEK293 cell line. To further characterize the promoter of PARROT we cloned 166bp, 222bp, 298bp and 404 bp sequences located upstream of the PARROT's TSS. This experiment identified a short promoter region responsible for mediating cell-type specific PARROT expression. Intriguingly, this short promoter region, when cloned downstream of the luciferase, can also act as an enhancer in a typical enhancer luciferase assay (not shown).

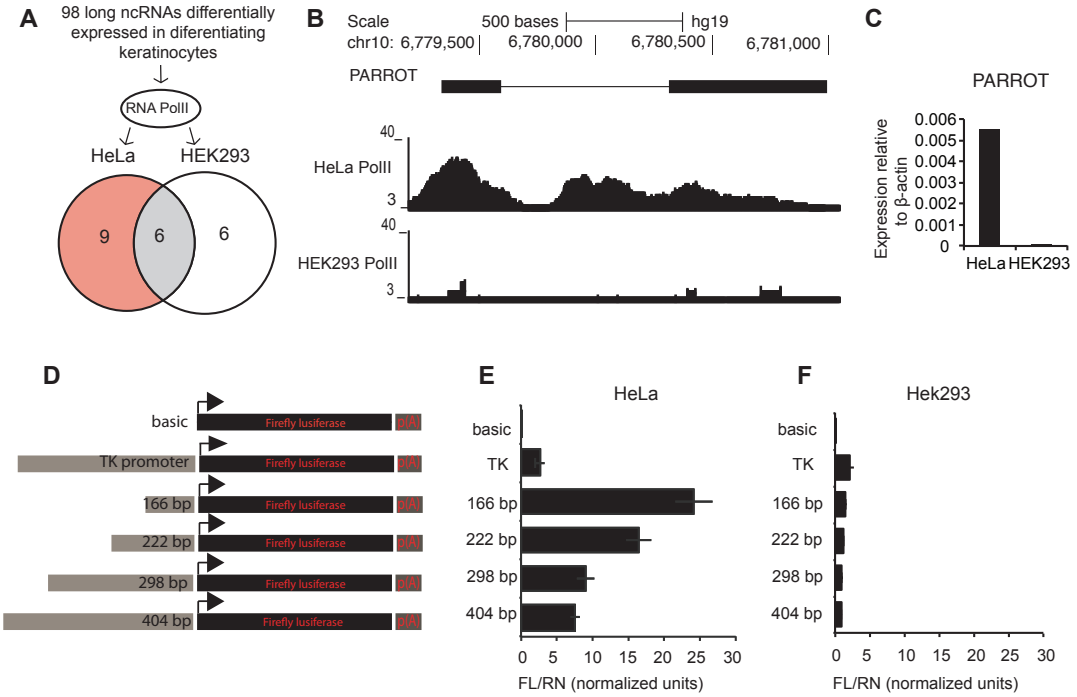

Figure 10. PARROT's differential expression is a consequence of cell type specific promoter activity. A) Association of Pol II in HeLa and HEK293 cell to ncRNAs differentially expressed upon keratinocyte differentiation. B) Pol II association to PARROT in HeLa and HEK293 cell line shown as a genome browser snap shot. C) Expression of PARROT determined by qPCR in HeLa and HEK293 cell line normalized to β-actin. (D-F) Promoter upstream regain of PARROT was cloned in pGL3 vector upstream of the Luciferase gene. D) Graphical illustration of the cloned inserts. (E-F) Luciferase reporter assay. Renilla luciferase vector (pRL-TK) was cotransfected with the Firefly luciferase vectors for transfection control in HeLa (D) and HEK293 cells (E).

Examination of available RNA-seq data shows that in the human body PARROT is expressed in different tissues: Immune (lymph nodes), muscle (heart and skeletal muscle); Internal (kidney and liver); Secretory (thyroid, adrenal gland and breast); Reproductive (ovary, prostate, testis) (Harrow et al., 2012). Additionally, PARROT is

expressed in different human cancer derived cell lines such as HeLa, HepG2 and A549 as well as primary cell lines such as H1ES (human embryonic stem cells), NHEK (normal epidermal keratinocytes) and HMEC (human mammary epithelial cells) (Derrien, 2012; Harrow et al., 2012) (**Figure 15A**).

In order to determine the cellular distribution of PARROT I fractionated HeLa cells into cytoplasmic, nuclear and chromatin fraction extracted the RNA from each fraction and performed a qPCR. Levels of 7SL and pre-GAPDH mRNA as markers of cytoplasm and chromatin, respectively, were also determined to control the purity of cellular fractionation. This experiment revealed that PARROT, unlike most long ncRNAs reported so far, is localized predominantly in the cytoplasm (**Figure 11A**).

To further evaluate this transcript I examined the transcriptional profile of PARROT obtained in following RNA-seq experiments: 1) RNA-seq of chromatin-associated RNA using random primers and 2) RNA-seq from whole cells using oligod(T) primers (Conrad et al., 2014). This examination revealed that PARROT is transcribed as a long chromatin associated primary transcript of which only a short 5'end is further processed by splicing into a mature, cytoplasm associated form (**Figure 11B**). These observations suggest that PARROT might act both in the nucleus in the form of its primary transcript and in the cytoplasm as a processed transcript.

It was reported that some long ncRNAs code for short peptides (Kim et al., 2014). Since PARROT is predominantly localized in the cytosol we tested whether the annotation of PARROT as a ncRNA is correct. To this end, I cloned the full-length processed PARROT DNA sequence behind the T7 promoter and performed an *in vitro* translation assay. As a positive control, *in vitro* translation of a luciferase gene was performed in parallel. This experiment did not reveal the presence of any peptides arising from the PARROT sequence whereas the luciferase protein was translated successfully (**Figure 11C**) verifying that the experiment was performed successfully and suggesting that PARROT does not code for a protein.

We further tested the coding potential of PARROT by examining the distribution of PARROT in polysome profiles by performing qPCR in individual polysome fractions. Compared to a PCG - GAPDH that is translated at a high rate judged from its association with high molecular weight polysome fractions, PARROT is associated only to low molecular weight polysome fractions suggesting that PARROT is not

translated (**Figure 11D**). To account for the differences in length that can skew the interpretation of polysome profiling results, the distribution of a PCG EFHD2 that has the same length like PARROT in polysome fractions has been used as a control. Compared to PARROT, EFHD2 is associated to higher molecular weight polysome fractions further confirming that PARROT does not code for a protein (not shown). Furthermore, no peptides corresponding to the sequence of PARROT are detected in proteogenomic studies (Lipovich et al., 2012) and PARROT has not been detected in ribosome profiling studies in HeLa (Ingolia et al., 2013).

Conservation of PARROT has also been examined. Like most long ncRNAs, PARROT is conserved in primates but not in other vertebrates (not shown).

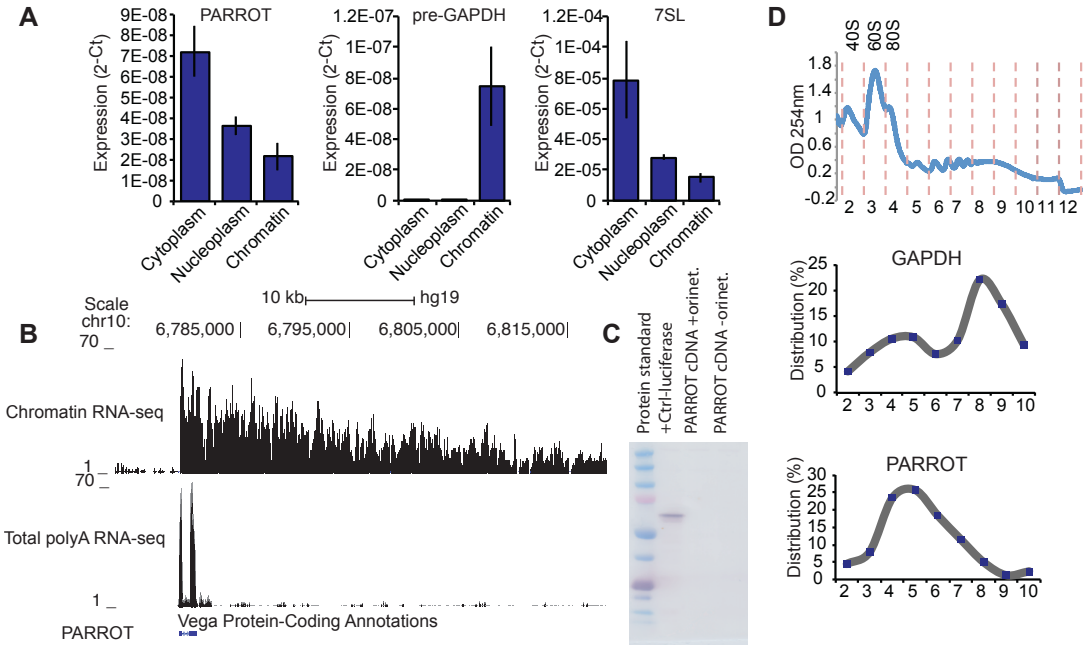

Figure 11. PARROT is a long non-coding RNA predominately localized in the cytoplasm. A) Distribution of PARROT, in cellular fractions. Pre-gapdh and 7SL are shown as controls for the cellular fractionation. Shown are the average of 2^{-Ct} ± s.d., n=3. B) PARROT in chromatin RNA-seq and polyA RNA-seq shown as a snapshot from the genome browser. C) *In vitro* translation assay. D) HeLa polysome profile. Polysomal distribution of GAPDH and PARROT RNAs in polysome fractions was determined by isolating the RNA from each fraction collected from a 10–50% sucrose gradient and performing a qPCR. Percentage of mRNA level in each fraction is shown.

3.2.2 Biological functions of the long ncRNA PARROT

To investigate the function of PARROT, I depleted it in HeLa cells with siRNAs and observed its effect on the cellular phenotype.

In order to determine the optimal concentration of siRNA oligos, HeLa cells were transfected with increasing concentrations of the oligos ranging from 2 nM to 100 nM (data not shown) and the effect was observed by qPCR. Depletion of PARROT was most efficient when I used 50 nM oligo concentration, as efficient as when 100 nM of oligo was used, therefore this oligo concentration was chosen for further analysis (**Figure 12A**). Two different siRNAs that match two different targeting sites along PARROT sequence were used to account for the siRNA off target effects.

Upon depletion of PARROT in HeLa we observe 39.5%-44% reduction in the number of viable cells with a crystal-violet staining (**Figure 12B**). Several different methods were employed in an attempt to determine whether cells lacking PARROT were apoptotic or arrested in a certain phase of the cell cycle: luciferase caspase3/7 activity assay, annexin V fluorescent staining and propidium iodide fluorescent staining in different time points upon depletion of PARROT (data not shown). However, no evidence of apoptosis or cell cycle arrest was obtained from these experiments. We observe a marked reduction in the number of cells upon knock-down of PARROT not only in HeLa but also in A549 human lung adenocarcinoma cell line (data not shown) and primary human mammary epithelial cells (HMECs) (**Figure 15E**).

Since PARROT is predominantly localized in the cytosol and is associated to low molecular weight polysome fractions I tested whether it can potentially regulate translation. To address this question, I carried out a puromycin based translation assay in HeLa cells. Upon knock-down of PARROT we can observe a marked reduction in global translation (37.3% to 47.8%) determined by puromycin incorporation and western blot of the newly synthesized proteins (**Figure 12C**). Ponceau S staining of the same membrane was performed to control for equal protein loading. The quantification of three independent puromycin translation assays is also shown (**Figure 12C**).

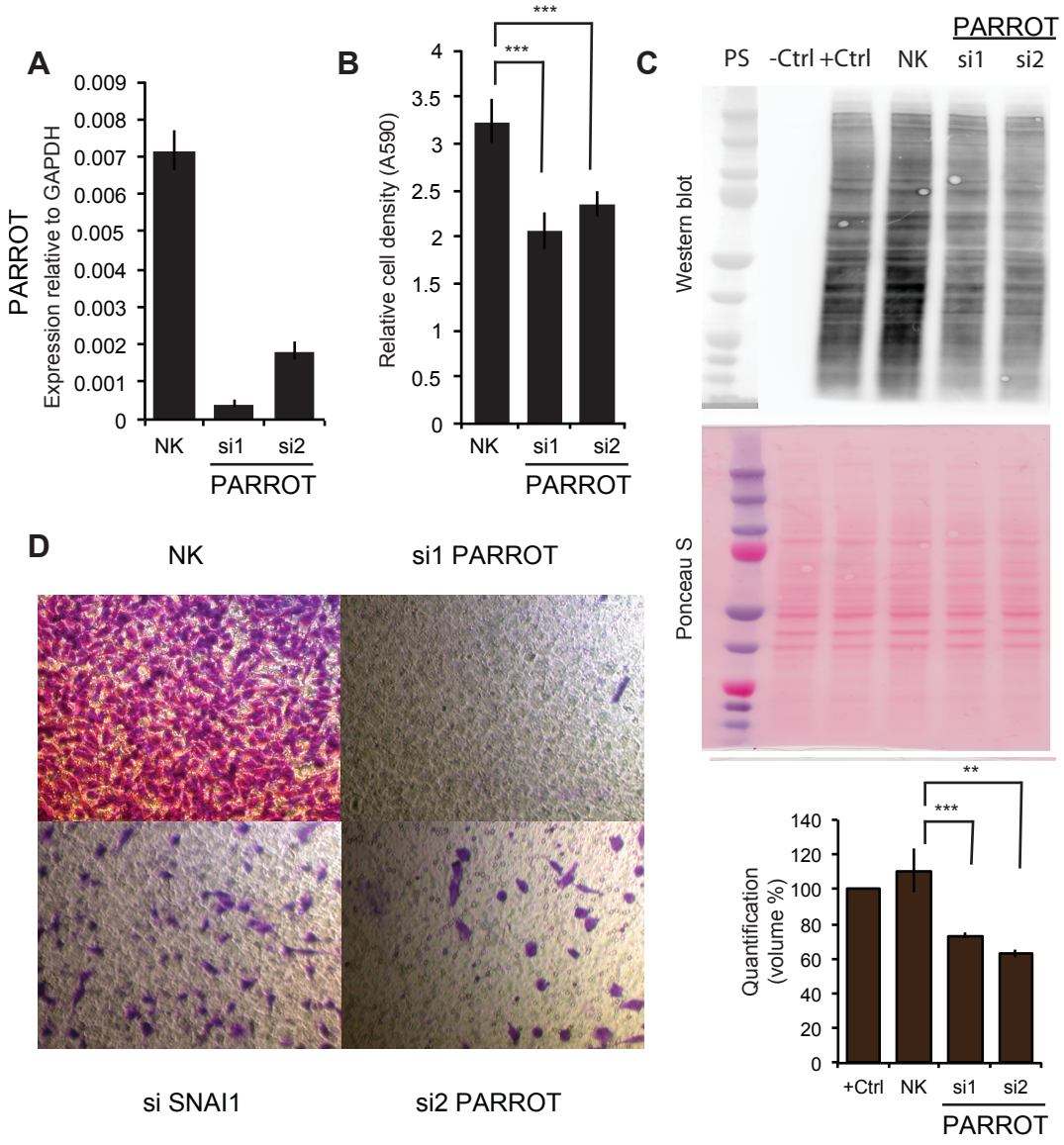

Figure 12. PARROT is a regulator of growth and translation. A) Depletion of PARROT with two different siRNAs in HeLa cells. HeLa cells were also transfected with a non-targeting control siRNA (NK). B) Crystal violet viability assay in HeLa cells treated with either the control siRNA (NK) or depleted of PARROT for 72h. The average ± s.d. are shown, n=5 (si1 p= 4.87581E-06, si2 p= 1.58272E-05). C) Puromycin translation assay. Shown are: Protein marker, protein extracts from control untreated HeLa (-ctrl), control HeLa cells treated with puromycin (+ctrl), and cells treated with puromycin and transfected with either control siRNA (NK) or two different siRNAs against PARROT for 72h. Top panel: western blot with puromycin antibody, middle panel: Ponceau S staining shown as the loading control and the bottom panel: quantification of puromycin incorporation in three independent experiments (The average ± s.d., is shown; si1 p= 0.001054504, si2 p= 0.017090534). D) Migration assay of A549 cells treated with either the control siRNA (NK), two different siRNAs against PARROT or SNAI1 (positive control).

Furthermore, migration assay revealed that A549 cells lose the ability to migrate in the absence of PARROT (**Figure 12D**). This effect on the cellular migration is comparable to the one caused by the knock-down of Snail Family Zinc Finger 1 (SNAI1), a well-known regulator of cellular migration (Barrallo-Gimeno and Nieto, 2005).

Overexpression of PARROT has also been employed to further examine and confirm the impact of this long ncRNA. To this end, I cloned the full length PARROT sequence in a pcDNA3.1 vector under the control of the CMV promoter, and the vector was transfected in HeLa and HEK293 cells. Although a ~ 10 fold increase in the levels of processed transcript was achieved, reaching levels of expression as high as for highly expressed PCG GAPDH, no effect was observed on either expression of target genes (see below) nor cellular proliferation (data not shown). This suggests that PARROT might act in a context dependent manner.

3.2.3 PARROT's molecular mechanism of action

To determine the molecular mechanism by which PARROT exerts its effect on the cellular phenotype, we examined the effect on the whole cell transcriptome, proteome and phosphoproteom upon PARROT depletion in HeLa cells.

To get insight into changes that occur on a transcriptional level in cells lacking PARROT we conducted RNA-seq of polyadenilated transcripts in HeLa cells treated with siRNAs against PARROT. This analysis revealed that the depletion of PARROT affects the expression of 331 genes of which 150 are downregulated and 181 are upregulated (**Appendix 1**). Several targets were validated by qPCR (not shown). In accordance with the observed phenotype, gene ontology analysis shows that PARROT affects the expression of genes involved in cell cycle regulation, cellular growth and proliferation and cellular movement (**Figure 13A**).

In order to further explore the effect of PARROT on translation SILAC labeling was performed and both the abundance and the phosphorylation status of proteins in cells depleted of PARROT was observed by MS. In this experiment untreated cells were labeled with heavy isotopes whereas the cells in which PARROT was knocked-down with siRNAs were grown in medium with light isotopes. Three days after the knock-

down was performed, protein extracts were collected and mixed in equal ratios. Intensities of both heavy and light labeled peptides were determined by MS. Analysis of light/heavy ratios determined by MS revealed that PARROT affects the levels of 73 proteins of which 51 are downregulated and 22 are upregulated (**Appendix 2**). Additionally, we further purified for phosphoproteins to identify differentially phosphorylated proteins by MS in cells depleted of PARROT. Analysis of the phosphoproteome revealed that the depletion of PARROT influences the phosphorylation status of 264 proteins: 98 were hyper-phosphorylated and 166 were hypo-phosphorylated upon depletion of PARROT (**Appendix 3**). All affected proteins were subjected to a gene ontology analysis. In agreement with RNA-seq results and the phenotype observed this analysis shows that PARROT affects proteins involved in cell cycle regulation, cellular growth and proliferation as well as cellular movement on the level of protein synthesis as well as phosphorylation (**Figure 13B**).

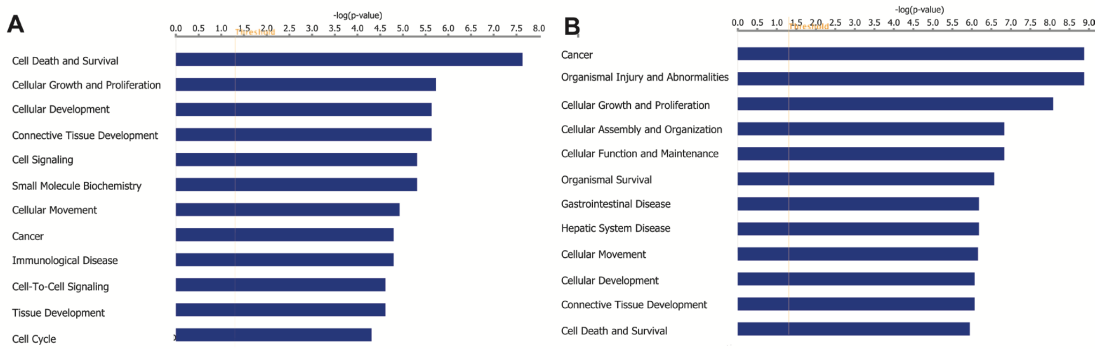

Figure 13. Gene ontology analysis of PARROT affected genes confirms that PARROT affects cellular growth, proliferation and migration. Gene ontology analysis of A) genes differentially expressed upon depletion of PARROT. B) differentially phosphorylated proteins upon depletion of PARROT.

To further examine the mechanism by which PARROT exerts its effect RNA-seq, proteome and phosphoproteome data from HeLa cells depleted of PARROT were subjected to pathway analysis -IPA. Based on published regulatory relationships between proteins, IPA can predict an upstream regulator, an affected protein that can potentially be responsible for differential abundance of proteins in the dataset. For proteins affected by the depletion of PARROT either on the transcriptional, protein or phosphoprotein level, IPA predicts c-Myc as a significant ($p=3.09x10^{-5}$) upstream regulator (**Figure 14A**).

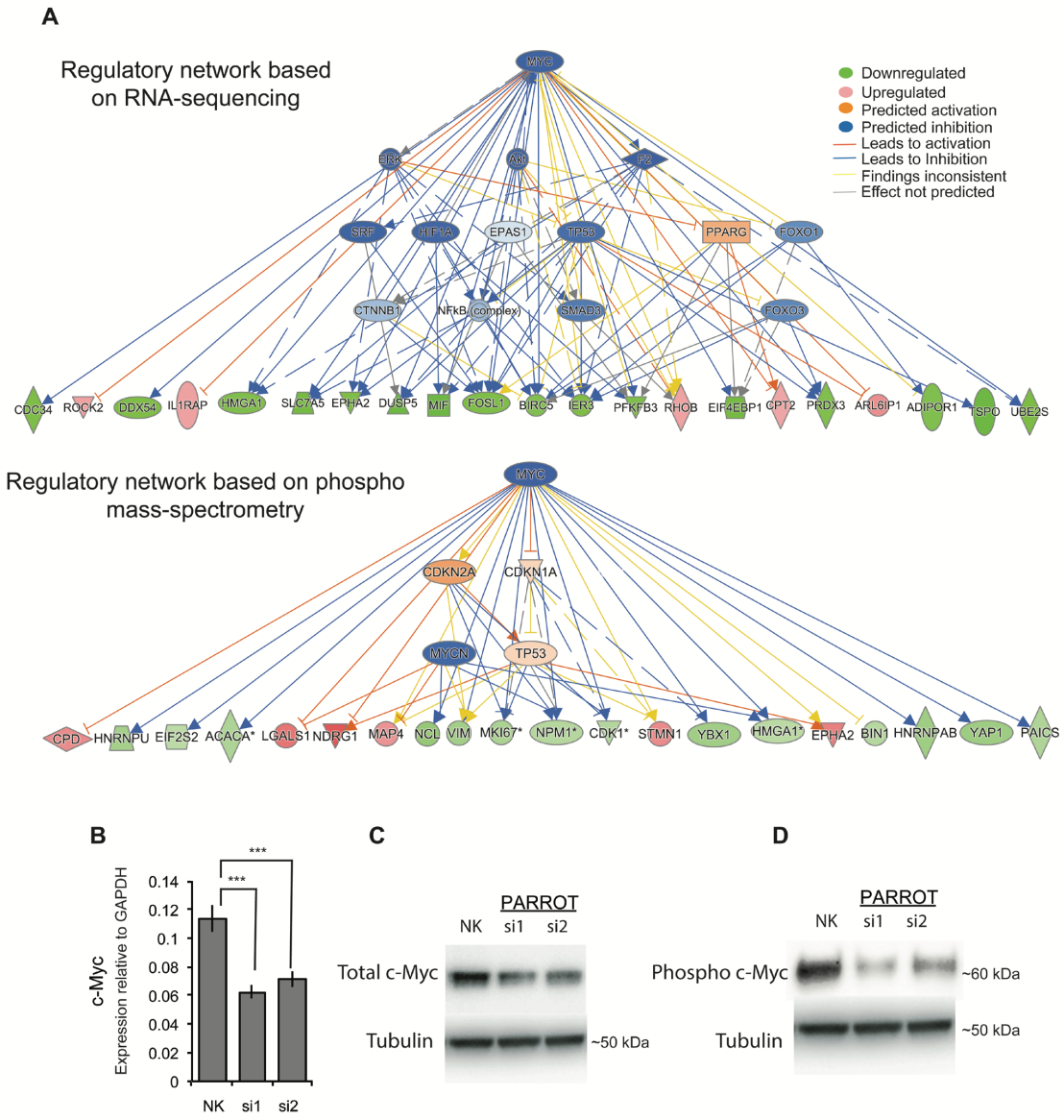

Figure 14. PARROT regulates c-Myc oncogene. A) IPA network analysis showing c-Myc as the upstream regulator of genes affected by depletion of PARROT based on RNA-seq data (upper panel) or phospho mass-spectrometry data (lower panel) B) Levels of processed c-Myc transcript in HeLa cells treated with either the control siRNA (NK) or depleted of PARROT determined by qPCR. The average ± s.d. are shown, n=3 (si1 p= 0.000541914, si2 p= 0.001295592). (C-D) Western blot of total c-Myc or phospho-form of c-Myc (D) from protein extracts of Hela cells transfected with either the control siRNA (NK) or depleted of PARROT with two different siRNAs.

Known c-Myc targets such as CDC34 (involved in cell death) (Blank et al., 2013); FOSL1 (involved in proliferation, migration, motility, invasion, cell cycle progression, differentiation, apoptosis, etc) (Adiseshaiah et al., 2007; Galvagni et al., 2013); ROCK2 (involved in cellular adhesion, migration, apoptosis and differentiation) (Fukata et al.,

1999; Li et al., 2012a; Piazzolla et al., 2005); SLC7A5 (involved in growth and proliferation) (Wang et al., 2011b); EPHA2 (involved in migration, proliferation, adhesion, growth) (Surawska et al., 2004); RHOB (Ras family member involved in apoptosis, growth, proliferation, transformation, migration, adhesion, etc.) (Prendergast, 2001); MIF (involved in proliferation, apoptosis, migration, aging, cell death, etc.) (Nishihira, 2000); DUSP5 (negatively regulates members of MAP kinase superfamily which are associated with cellular proliferation and differentiation) (Nunes-Xavier et al., 2010), are differentially expressed and/or phosphorylated upon depletion of PARROT (**Figure 14A**).

In order to validate c-Myc as the target of PARROT, I examined the direct effect of PARROT depletion on c-Myc at the RNA and protein level. Levels of c-Myc transcript, determined by qPCR, are indeed decreased upon knock-down of PARROT (**Figure 14B**). In accordance with these results, the levels of both total c-Myc (**Figure 14C**) and phosphorylated c-Myc are decreased (**Figure 14D**) upon depletion of PARROT. Since c-Myc is a regulator of cellular proliferation and growth as well as migration (Meyer et al., 2008; van Riggelen et al., 2010) the regulation of c-Myc by PARROT is in agreement with the phenotypes observed.

3.2.4 PARROT in senescence and cancer

PARROT has a profound effect on cellular growth, judged from the effect on translation, and proliferation. Based on this evidence, I examined the expression of PARROT in biological systems in which proliferation is altered, like senescence and cancer.

Examination of available RNA-seq data revealed that PARROT shows the highest expression in human mammary epithelial cells (HMECs) (**Figure 15A**). HMECs are healthy human mammary breast cells that undergo replicative senescence, an irreversible arrest of cellular growth, after approximately 14 passages (Bertram and Hass, 2009). Expression of PARROT was examined by RNA-seq in young, proliferative (passage 2) and senescent (passage 14) HMECs. Whereas PARROT is highly expressed in proliferative (passage 2) HMECs, its expression is greatly reduced in senescent HMECs (passage 14) (**Figure 15B**).

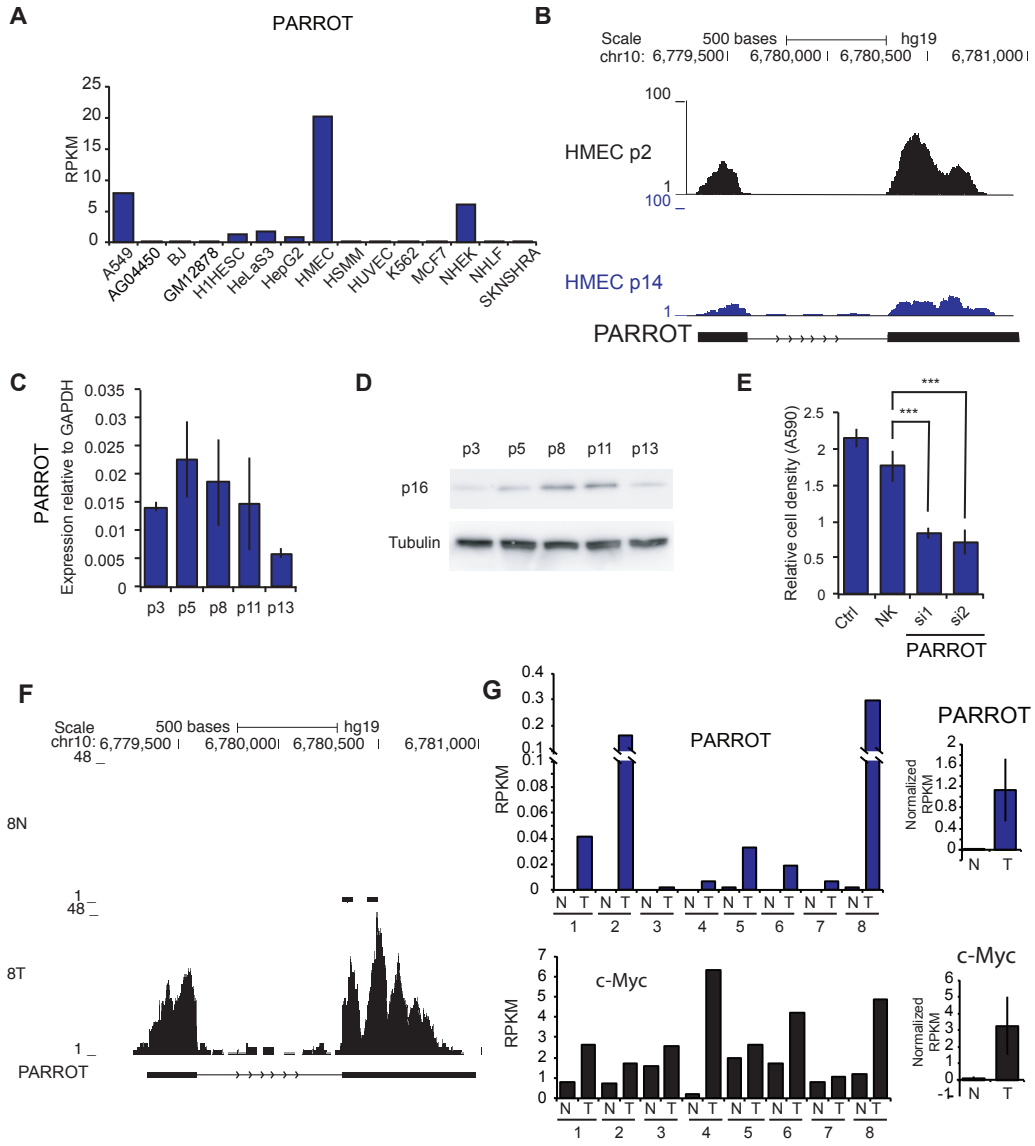

Figure 15. In senescence and stomach cancer PARROT is differentially expressed. A) Expression of PARROT in different cell lines determined by RNA-seq (Derrien, 2012). B) RNA-seq data of HMECs cells in passage 2 and in passage 14 shown as a genome browser snapshot. C) Expression of PARROT in ageing HMECs determined by qPCR. D) p16 western-blot of protein extracts from ageing HMECs. E) Crystal violet viability assay of HMECs cells of untreated cells (ctrl) or cells treated with either control siRNA (NK) or depleted of PARROT with two different siRNAs for 72h in passage 5. The average ± s.d. are shown, n=5 (si1 p=1.09588E-12, si2 p=2.86182E-12).F) RNA-seq in control (N) or tumor (T) stomach cancer sample depicted as a genome browser snapshot. (G) Expression of PARROT and c-Myc in control (N) and tumor (T) stomach cancer samples determined by RNA-seq.

In order to validate this differential expression of PARROT, I cultured HMECs until they reached senescence. As a confirmation that the cells did reach senescence in my hands, I monitored the level of a marker of senescence -p16 protein, a tumor suppressor

whose enhanced expression leads to senescence (Bertram and Hass, 2009). Indeed, as the HMECs age through passaging the expression of p16 increases (**Figure 15D**). In accordance with the RNA-seq data, qPCR analysis showed that as HMECs cells grow older with passaging from third to thirteenth passage the expression of PARROT gradually decreases (**Figure 15B** and **Figure 15C**). Furthermore, as in HeLa, knock-down of PARROT in proliferative HMECs (passage 5) led to a 43-48% decrees in the number of viable cells (**Figure 15E**). This suggests that PARROT is not only involved in expansion of cancer-derived cell lines but also for the expansion of primary mammary cells.

Contrary to HMECs that stop proliferating, transformed cancer cells proliferate at a much higher rate than the untransformed cells of the same origin. In order to examine this proliferation status, we examined the expression of PARROT in stomach cancer by RNA-seq. The analysis of 8 paired normal and tumor samples from stomach cancer patients reveled that whereas PARROT is not expressed in healthy stomach cells of stomach cancer patients, it is highly expressed in the stomach cancer cells of the same patients (**Figure 15F**). Furthermore, when the expression of c-Myc in these patients was examined, the same trend can be observed, c-Myc is expressed at a significantly higher level in tumor that in the control cells (**Figure 15G**). This observation supports our data and further suggests that PARROT has an oncogenic potential that is exerted *via* regulation of the c-Myc oncogene.

4 DISCUSSION

Long ncRNAs regulate almost every level of the gene expression program adding an unexpected layer of complexity in the regulation of gene expression (Geisler and Coller, 2013; Ulitsky and Bartel, 2013). Furthermore, long ncRNAs are crucially involved in development of various diseases such as cancer and neurological disorders (Geisler and Coller, 2013; Ulitsky and Bartel, 2013; Vucicevic et al., 2015).

In this thesis we demonstrate that as much as 28% of long ncRNAs are expressed from tissue specific enhancers and are potentially involved in their function. Furthermore, we characterize a long ncRNA named PARROT and demonstrate its function in regulating cellular growth and proliferation through regulation of c-Myc oncogene.

4.1 Long ncRNA expression associates with tissue-specific enhancers

4.1.1 Predicting enhancers and their targets

Several approaches to predict enhancers have been developed based on various enhancer features and their combinations, however to date there is no consensus on how to predict enhancers (Shlyueva et al., 2014; Smallwood and Ren, 2013). Initial prediction methodologies used the binding TFs, either based on motif search or *in vivo* binding of TFs to genomic loci (Del Bene, 2007; Johnson et al., 2007; Kheradpour et al., 2007; Shlyueva et al., 2014). Although these approaches identified some active enhancers, the FDR of this approach is high since TFs have a general affinity to bind DNA and the binding of a TF does not necessarily affect the expression of any gene. The tendency of enhancers to reside within the accessible chromatin regions sensitive to DNaseI digestion, corresponding to the DNaseI hypersensitivity sites, has been used in search for enhancers (Boyle, 2008; Thurman, 2013). However, not all chromatin accessible regions correspond to enhancers, leading to very high FDR when this approach is applied to predict enhancers.

Binding of a co-transcriptional factor p300 has been used to predict enhancers (Visel et al., 2009a). Although most enhancers called in this way (up to 82%) have been reported to act as enhancers, this approach identifies only enhancers that acquire p300 binding for their activity. Examination of the epigenetic landscape of distal p300 binding sites revealed that these enhancers are enriched in H3K4me1 and H3K27ac mark and depleted of H3K4me3 mark. H3K4me3 mark has been associated to promoters (Heintzman, 2007, 2009). H3K4me1 and H3K4me3 marks and the ratio between them have been used to discriminate between enhancers and promoters (Djebali, 2012; Heintzman, 2007; Koch and Andrau, 2011). It was also shown that, whereas chromatin signatures at promoters are highly invariant across cells lines, histone modification patterns at enhancers, in particular H3K4me1 marks, are highly cell type specific (Heintzman, 2009). These results suggest that enhancers are the most variable genetic elements and are of prime importance in driving cell type specific gene expression patterns (Heintzman, 2009; Heinz et al., 2015). In line with this evidence, PreSTIGE enhancer prediction methodology defines enhancers based on the presence of tissue specific H3K4me1 domains in proximity to genes with high tissue specific expression representing an improvement over previous methods. Although PreSTIGE does not take into account the presence of H3K27ac for enhancer prediction, enhancers predicted in this way are enriched in the presence of H3K27ac, thus validating the approach used (Corradin et al., 2014).

It was traditionally thought that enhancers regulate the expression of the nearest gene on the linear DNA. However, experimental evidence demonstrated that enhancers can act over great distances and often do not target the closest gene (Sanyal et al., 2012; Smallwood and Ren, 2013). Only 40% of PreSTIGE predictions involve an enhancer and the nearest gene. Therefore, 60% of the PreSTIGE predictions would be undetected by the nearest gene method. Additionally, compared to the nearest gene method PreSTIGE has a much lower FDR, ~ 30% lower than the FDR in the nearest gene method (Corradin et al., 2014). This further demonstrates the power and the improvement of PreSTIGE over previously published methods.

The knowledge that enhancers are brought into close proximity to their targets has also been used to predict enhancers and their targets thanks to the development of chromosome conformation capture techniques like 3C, 4C, 5C (van Steensel and

Dekker, 2010) and ChIA-PET (Li et al., 2012b). PreSTIGE predicts 16 out of 17 enhancer-gene interactions previously identified by 3C or other experimental methods validating the approach used in this thesis. PreSTIGE predictions in K562 and MCF-7 cells are significantly enriched among ChIA-PET-identified interactions. Additionally, predictions made in other cell lines are depleted from the ChIA-PET identified K562 and MCF-7 interactions. This reflects the high cell type-specificity of enhancer-gene interactions and suggests that the FDR for PreSTIGE is low (Corradin et al., 2014). When comparing PreSTIGE, the closest gene and the closest expressed gene method to chromatin infarctions detected by 5C, PreSTIGE outperforms the two other methods with enrichment ratios up to sevenfold higher (Corradin et al., 2014). This evidence further demonstrates that PreSTIGE has a greater predictive power over methods previously used to predict enhancers and their targets.

Recent evidence demonstrating that long ncRNAs are transcribed from enhancers adds complexity to the enhancer prediction (Kim, 2010; Lam et al., 2013; Lam et al., 2014; Li, 2013; Melo et al., 2013; Natoli and Andrau, 2012; Orom and Shiekhattar, 2013). Since many factors involved in both enhancer and long ncRNA function are not yet characterized, it is challenging to study long ncRNAs with enhancer-like function. In order to address the interplay between long ncRNAs and enhancers PreSTIGE methodology was chosen because it is an improvement over previously published methods and does not exclude the possibility that an enhancer is being transcribed. PreSTIGE was further modified to better suit the examination of long ncRNAs derived from enhancers. Although this slightly increased the FDR, we expanded the domain in which we search for an enhancer to 200 kb surrounding the TSS, considering the length that long ncRNAs occupy in the genome and recent evidence that shows that enhancer-promoter interactions typically occur over distances larger than 100 kb (average of 120 kb) (Sanyal et al., 2012).

4.1.2 Long ncRNAs overlapping PreSTIGE predicted enhancers

In this thesis I find that 28% (2,695) of the analyzed ENCODE annotated long ncRNAs overlap a predicted cell-type specific active enhancer in all 11 cell lines suggesting that tissue specific enhancer activity could orchestrate the tissue-specific expression of

long ncRNA. In most cell lines expression of enhancer derived long ncRNAs is in accordance with the predicted enhancer activity but not in all 11 cell lines. For instance, in HMECs cells, we do not observe that, like in other cell lines, long ncRNAs are expressed at a higher level in the cell line in which they intersect an active enhancer. The reason for that might be the surgical origin of HMECs. Although these cells in theory should be mammary epithelial cells, they are a mixture of epithelial cells, fibroblasts, adipocytes and other cells present in the mammary tissue since they cannot be fully eliminated. Therefore, it is not surprising that the expression of long ncRNAs in such a mixture of cell types is not cell type specific.

Additionally, in this thesis I observe a strong correlation between the expression levels of long ncRNAs overlapping cell type-specific enhancers and PCGs that are targeted by those enhancers, suggesting a functional link between them. This is in line with the growing body of evidence that demonstrates that the expression of long ncRNAs from enhancers correlates with the expression of their neighboring PCGs in wide variety of biological conditions (De Santa, 2010; Hah et al., 2013; Kim, 2010; Koch and Andrau, 2011; Orom, 2010). For instance, stimulation of mouse neurons (Kim, 2010) or activation of macrophages (De Santa, 2010) leads to transcription of thousands of long ncRNA from enhancers. Transcription of long ncRNAs from enhancers was shown to precede the activation of their neighboring PCG suggesting that they are responsible for the increased expression of the PCG in proximity (De Santa, 2010; Hah et al., 2013) and supporting the idea that long ncRNAs identified here function to regulate their neighbors.

Several recent studies showed that some TFs regulate the expression of their target genes by modulating the expression of enhancer-derived long ncRNAs. TFs can either activate the expression of enhancer-derived long ncRNAs as in the case of p53 (Melo et al., 2013) and oestrogen receptor (Li et al., 2013) or repress their transcription as in the case of Rev-Erb (Lam et al., 2013) to modulate the expression of their target genes. These studies suggest that, in addition to long ncRNAs transcribed from enhancers identified in this study, there might be many more long ncRNAs transcribed from enhancers that are induced only upon certain stimuli like the binding of a TF.

In this thesis, it is shown that the target PCGs of enhancers that are transcribed into long ncRNAs are enriched for ontological terms that are relevant for the cell type specific biological functions. This further suggests that this approach identifies functionally relevant enhancers. Furthermore, this implies that *via* long ncRNAs, enhancers can regulate cellular fate, differentiation and cell type specific gene expression program. This is plausible since here I find that most long ncRNAs overlap an enhancer in one cell line only (**Figure 4E**) which is in line with the fact that long ncRNAs, just like enhancers, are cell type specific (Derrien, 2012).

Several enhancers can regulate the expression of a single gene to fine-tune its expression in response to different cellular stimuli (Heinz et al., 2015). An example of one such gene is SMOX **(Figure 4B)**. PreSTIGE predicts five enhancers for this gene in K562 cell line of which only one is transcribed into an annotated long ncRNA. This example suggests that both enhancers that mediate their activity *via* ncRNAs and the ones that potentially act on the DNA level can together regulate the expression of a single gene. Since it is demonstrated that many enhancers produce short lived exosome sensitive transcripts (Andersson et al., 2014), it is also possible that some of these remaining four enhancers produce short lived long ncRNAs that cannot be detected by steady-state RNA-seq. However, since no function has been attributed to these short lived exosome sensitive transcripts, it is very likely that they are not functional and are probably just transcriptional noise. Therefore, these short lived transcripts were not included in our analysis. Only ENCODE annotated, detectable by RNA-seq, long ncRNAs were used in a search for long ncRNAs derived from enhancers since they are more likely to be functional and mediate enhancer activity due to their stability.

The PreSTIGE approach used in this study to find enhancer derived long ncRNAs has its limitations. This approach is not able to identify enhancer-promoter pairs outside of the 200 kb window surrounding the TSS since for those interactions FDR would increase as the domain increases. Additionally, although most enhancers are cell type specific (Heintzman, 2009; Lam et al., 2014), some enhancers can act in several cell lines and those would not be identified with this approach. Furthermore, not all enhancers require the presence of H3K4me1 mark (Shlyueva et al., 2014) and those enhancers would also be missed with this approach. Therefore, albeit some of the predicted enhancers that express long ncRNAs might be false-positives, due to its

limitations PreSTIGE potentially does not include a great number of true-positives (Vucicevic et al., 2015).

4.1.3 PreSTIGE compared to other methods

Comparing PreSTIGE predicted enhancers to enhancers predicted by (Heintzman, 2009) in HeLa revealed that only 27% enhancers predicted by PreSTIGE overlap the ones predicted by (Heintzman, 2009). This exemplifies how different enhancer prediction methodologies predict different pools of enhancers. Both prediction methodologies predict enhancers based on the presence of H3K4me1 mark. However, PreSTIGE predicts a region as an enhancer only if the covering H3K4me1 mark is cell type specific, whereas (Heintzman, 2009) excludes all enhancers that have H3K4me3 signal present since this mark is considered to be a mark of promotes. In this way (Heintzman, 2009) minimize the possibility that an enhancer is being transcribed in agreement with what was thought for decades that enhancers act exclusively on the DNA level. However, in this study and other recent studies it is demonstrated that many enhancers are being transcribed into long ncRNAs (Andersson et al., 2014; Lam et al., 2013; Melo et al., 2013; Vucicevic et al., 2015). Therefore, the transcriptionally active promoters of these transcribed long ncRNAs carrying H3K4me3 marks can also be present at these enhancers. Indeed, in support of this idea, we observe a higher H3K4me3 signal at enhancers predicted by (Heintzman, 2009) that overlap PreSTIGE predicted enhancer than at enhancers that are only predicted by (Heintzman, 2009; Vucicevic et al., 2015) (**Figure 8B**). In agreements with this is the expression of long ncRNAs that overlap these two sets of enhancers; long ncRNAs that overlap enhancers predicted by (Heintzman, 2009) are expressed at a much lower level and in several cell lines than the ones that overlap PreSTIGE predicted enhancers (**Figure 8D-E**). Therefore, although unlike study by (Heintzman, 2009), PreSTIGE does not include in the prediction enhancers active in several cell lines, when it comes to finding enhancers that are being transcribed, PreSTIGE has a much stronger predictive power than the study by (Heintzman, 2009).

4.1.4 Characteristic epigenetic mark of enhancers transcribed into long ncRNAs

The more we examine enhancers and promoters distinguishing between the two is becoming increasingly challenging. For instance, a study showed that if the promoter of the gene is absent, an intragenic enhancer can act as an alternative tissue-specific promoter (Kowalczyk et al., 2012). Several studies have proposed that although enhancers are being actively transcribed, H3K4me3 levels at enhancers are low compared to promoters. Thus, a higher ratio of H3K4me1 to H3K4me3 is being used to distinguish between enhancers and promoters (Calo and Wysocka, 2013; Djebali, 2012; Heintzman, 2009; Koch and Andrau, 2011). H3K4me3 mark is associated with the initiating form of Pol II (Ser5 phosphorylation of the C-terminal domain). Evidence of widespread binding of Pol II to enhancers and the transcription of enhancers raises a question of why most enhancers do not acquire higher levels of H3K4me3. One possible explanation could be that the levels of Pol II are generally much lower at enhancers than at promoters (Calo and Wysocka, 2013). Indeed, the expression of long ncRNAs is in general at lower levels compared to PCG (Derrien, 2012; Ulitsky and Bartel, 2013). In support of this idea are experiments that showed that ectopically H3K4me3 modified enhancers are associated with aberrantly elevated level of transcription both from the enhancers and from the nearby promoters (Clouaire et al., 2012). In this study we show that on average, enhancers overlapping an annotated long ncRNA show significantly lower H3K4me1/H3K4me3 ratio than the enhancers that do not overlap annotated long ncRNAs. (**Figure 8C**). This result reveals that enhancers overlapping annotated long ncRNAs have a higher H3K4me3 signal compared to H3K4me1 signal, which is in line with the evidence that long ncRNAs are expressed from these enhancers. Furthermore, these results suggest that enhancers expressing long ncRNAs could have a characteristic histone mark profile (Vucicevic et al., 2015) that can potentially be used in search for more transcribed enhancers.

In conclusion, in this thesis we validate the tissue-specific predictive power of enhancer prediction methodology- PreSTIGE. Furthermore, we show that thousands of long ncRNAs are expressed from these enhancers dependent on their activity. Data presented here suggests that there is functional link between enhancer activity and long ncRNAs expressed from these enhancers in determining tissue-specific gene expression (Vucicevic et al., 2015).

4.2 The long non-coding RNA PARROT is an upstream regulator of c-Myc and affects proliferation and translation

4.2.1 Characterization of the long ncRNA PARROT

The development of next generation sequencing technologies enabled the detection of low abundant transcripts revealing that the human genome is pervasively transcribed into thousands of long ncRNAs (Derrien, 2012). However, although we are aware of their existence, studying low abundant transcripts is challenging. Additionally, albeit we know that some long ncRNAs are functional, we still have no evidence on how many long ncRNAs are functional and how many are just merely transcriptional noise (Ulitsky and Bartel, 2013). Furthermore, long ncRNAs are expressed at a highly tissue specific manner, adding a further challenge in studying individual long ncRNAs (Derrien, 2012). To bypass these problems in this study I used functional data for differential expression of long ncRNAs in keratinocytes following differentiation and RNA Pol II association in HeLa and HEK293 cells. We reasoned that if a long ncRNA is differentially expressed, regulated in certain condition like keratinocyte differentiation, it is much more likely to be functional. Additionally, RNA Pol II association analysis enabled me to find a long ncRNA that is expressed at a high level compared to other long ncRNAs, since in most cases RNA Pol II cannot be detected at low abundant transcripts. Furthermore, I performed RNA Pol II analysis in two cell lines, HeLa and HEK293 that are easy to grow and manipulate. With this approach I was able to identify a highly and differentially expressed functional long ncRNA-PARROT.

It is thought that the low expression of long ncRNAs is a consequence of rapid degradation owing to the lack of open reading frames within those transcripts, rendering them prone to decay (Ulitsky and Bartel, 2013). By using a promoter reporter we show that the high expression of long ncRNA PARROT in HeLa cells is a consequence of promoter activity (**Figure 10E,F**). The absence of its expression in HEK293 cells is not a consequence of degradation of this transcript in HEK293 cells but rather shows that the promoter is not active in HEK293 cells.

Long ncRNA promoters are thought to be very similar to the promoters of PCGs (Ulitsky and Bartel, 2013). Indeed, I was able to identify a short core promoter of PARROT with properties characteristic for promoters of PCGs. Interestingly, the longer the upstream promoter region was cloned the lower the activity of the reporter assay was (**Figure 10E,F**). This implies that just like for PCGs the expression of PARROT might be fine-tuned by binding of activators and repressors to its promoter.

It is thought that long ncRNAs are in general predominantly localized in the nucleus (Derrien, 2012). Compared to mRNAs, long ncRNAs are approximately threefold enriched in the nuclear fraction than in the cytoplasmic fraction of the cell. However, this is possibly a misconception since a typical mRNA is approximately sixfold enriched in the cytoplasm, therefore with a 3 fold relative nuclear enrichment a long ncRNA would still be two times more abundant in the cytoplasm (Ulitsky and Bartel, 2013). Although most research has focused on long ncRNAs regulating transcription, examples of long ncRNAs regulating mRNA processing (Albertson et al., 2006; Barry, 2014), mRNA stability (Faghihi et al., 2008) and translation (Carrieri, 2012) have emerged. For a long ncRNA PARROT I observe that its two-exon processed transcript is highly enriched in the cytoplasm suggesting that this long ncRNA might perform its function in the cytoplasm.

Intriguingly, PARROT is transcribed as a very long primary transcript and only a small portion of the 5' end is further processed suggesting that PARROT could also perform its function in the nucleus as a long primary transcript. This observation points out to significant differences that are observed between the way that PCG and long ncRNAs are processed. Whereas PCG transcripts are spliced co-transcriptionally and their primary transcripts can barely be detected in chromatin RNA-seq experiments, long ncRNA primary transcripts are highly detectable in chromatin RNA-seq experiments. This suggests that although they carry the same splice site signals, processing of long ncRNAs does not occur as efficiently as for the mRNAs (Derrien et al., 2012). This is an interesting observation that should be addressed to help us better understand the biology of long ncRNAs.

Only a very small fraction of long ncRNAs is conserved from mouse to humans (Cabili, 2011). When examining the conservation of long ncRNAs we should bear in

mind that long ncRNAs don't have the constraints that a PCGs have and that we are looking at conservation from a perspective of a PCG. For instance, a deletion or an insertion of a single base possibly does not make a difference for a long ncRNA whereas for a PCG it would cause a frameshift and a nonsense protein would be translated. It is possible that long ncRNAs are conserved on the level of their secondary structure but not much is known about their secondary structure. A growing body of evidence suggests that there are many functional long ncRNAs regulating every level of the gene expression program even though long ncRNAs are not highly conserved. An example examined here, PARROT, although conserved only in primates, has a profound effect on the cellular phenotype and regulates a conserved oncogene - c-Myc. Thus, although high evolutionary constrain does imply function, the lack of evolutionary constrain does not mean that the transcript is not functional. Evolutionary new transcripts, among which are long ncRNAs, can be driving evolutionary novelties like the development of the human brain. In support of this idea is the fact that the number of PCGs does not increase with the complexity of the organism (so called C-value paradox) but the number of non-coding genes does increase with the organism complexity (Kung et al., 2013). Therefore, although long ncRNAs are not highly conserved, they can be important evolutionary players and should be studied extensively.

4.2.2. Biological function of the long ncRNA PARROT

One of the key features of malignant transformation is the immortality of cancers cells and their ability to proliferate uncontrollably. This abnormal proliferation and growth requires pervasive translation of cellular building blocks. In order to obtain the energy required for their growth, cancer cells can migrate to the source of energy, like for instance the blood vessel through which they can spread to other organs and cause metastasis (Meyer and Penn, 2008). In this study, it is demonstrated that long ncRNA PARROT affects these key biological processes in a cancer cell line, proliferation, translation and migration, stressing out the impact that a single long ncRNA can have on cellular biology and homeostasis.

Upstream pathway analysis of genes affected by the depletion of PARROT identified c-Myc as the main upstream regulator of these genes. Upon depletion of PARROT we

can observe a decrease of both mRNA and total protein/phosphoprotein levels of c-Myc supporting the idea that PARROT exerts its effect in HeLa via regulating c-Myc. The c-Myc oncogene is one of the first discovered and one of the most studied proteins in molecular biology (Dang, 2012; Deng et al., 2014; Meyer and Penn, 2008). This oncogene is involved in regulation of many biological processes such as cell cycle, differentiation, and protein synthesis through regulation of expression of at least 15% of the genes in the human genome (van Riggelen et al., 2010). C-Myc regulates cellular growth by modulating the expression of the components of the ribosome complex (van Riggelen et al., 2010). The fact that PARROT regulates c-Myc provides an explanation to how a single long ncRNA can cause such a dramatic effect on the cellular growth and translation observed in this thesis. Our data is in agreement with previously published studies in which the knock-down of c-Myc in HeLa cells leads to a reduction in the number of viable cells and reduced ability of cells to migrate (Cappellen et al., 2007). Just like in this thesis, in the case of PARROT, Cappellen and colleagues who studied the effect of c-Myc depletion in HeLa, were also unable to determine whether the reduction of the number of cells is a consequence of cellular death or cell cycle arrest (Cappellen et al., 2007). A possible explanation for that could be that HeLa cells are not suitable for cell cycle analysis due to their error-filled genomes, with one or more extra copies of many chromosomes (Landry et al., 2013). Nevertheless, PARROT seems to be an indispensable transcript in HeLa cells.

4.2.3 PARROT in senescence and cancer

The effect on translation and the association to proliferation together with the regulation of c-Myc by PARROT inspired us to examine biological systems in which proliferation is changed, like senescence and cancer.

Examination of RNA-seq data from proliferative and senescent HMECs revealed that PARROT is downregulated in senescent cells and is required for the proliferation of young HMECs cells further supporting the idea that PARROT is a long ncRNA required for cellular proliferation. How is this effect mediated in untransformed cells remains to be determined since c-Myc does not seem to play an important role in proliferation of young HMECs. These results imply that PARROT might be able to regulate cellular proliferation through mechanisms that do not involve c-Myc in this

type of cells making PARROT an even more interesting candidate for further investigation.

Furthermore, I show that in stomach cancer samples, from eight patients analyzed by RNA-seq, PARROT and c-Myc are expressed at a similar manner. They are both highly expressed in tumor samples and expressed at a much lower level (or absent) in the healthy control tissue implying that PARROT might have a role in the development of this type of cancer via modulation of c-Myc expression. Additionally, PARROT has previously been found to be associated with renal and lung adenocarcinoma (Iyer et al., 2015) and could be an important regulatory factor of the c-Myc pathway in several types of cancer.

Recent research has reported a significant association between long ncRNAs and c-Myc in a wide range of tumors. Some reports revealed that c-Myc can activate some and repress other long ncRNAs. These long ncRNAs act to regulate c-Myc targets and act as oncogenes themselves promoting cellular proliferation and migration. C-Myc itself can be regulated by long ncRNAs (Deng et al., 2014). The findings presented here, that PARROT can regulates a group of genes regulated by c-Myc identifies an additional factor in c-Myc regulatory network and gives further insight into how c-Myc is regulated in cancer.

4.3 Outlook

4.3.1 Long ncRNA expression associates with tissue-specific enhancers

The results presented here suggest that long ncRNA transcription is functionally linked to tissue-specific enhancers. However, they do not tell us about their mechanism of action. Some enhancers derived long ncRNAs have been shown to be crucially involved in mediating the enhancer activity whereas other studies showed that the transcription at the enhancers is a mere consequence of enhancer activity (Natoli and Andrau, 2012). Based on data presented in this thesis, we cannot conclude how many of these enhancer derived long ncRNAs are mediating the enhancer activity and how many are expressed as a consequence of enhancer activation but this can be addressed in future experiments.

The integration of additional genomic information like RNA Pol II dependent chromatin looping from ChIA-PET experiments could improve the PreSTIGE prediction. ChIA-PET data is available for only three cell lines to this date. Hopefully in the future Pol II dependent interactions will be available for more cells lines allowing the integration of these interactions in the PreSTIGE prediction methodology. The combination of the two would allow the prediction of long-range interactions outside of the 200 kb PreSTIGE window that can be predicted from the Pol II ChIA-PET data. On the other hand, PreSTIGE approach can more reliably predict interactions within the 200 kb window since the Pol II ChIA-PET has a high FDR for predicting short range interactions.

Experimental validation of enhancer derived long ncRNAs would shed light on to whether a long ncRNA transcript is functional and is responsible for activating the predicted target PCG or it is just a consequence of enhancer activity and has no function in mediating the enhancer activity. For instance, a siRNA screen could be performed in which hundreds of enhancer derived long ncRNA are depleted. The effect on the expression of their predicted targets could be monitored with qPCR. However, these types of experiments are very laborious, expensive and would require robotic assistance. Additionally, siRNA mediated depletion of chromatin associated transcripts is not very efficient. Therefore, further optimization of depletion of chromatin associated transcripts is necessary for this type of screening.

Recently discovered prokaryotic immune system called clustered regularly interspaced short palindromic repeats (CRISPR) has emerged as a powerful tool for genome editing. A component isolated from this system is called Cas9 nuclease. Cas9 is targeted to DNA by a short guide RNA (sgRNA). Recently discovered Cas9 mutant called nuclease-deficient Sp.Cas9 mutant (dCas9) allows the incorporation of additional RNA cargo within the sgRNA enabling tethering of long ncRNAs of interest to the locus of interest (Shechner et al., 2015). In principle, with this approach whole genome library can potentially be screened for enhancer-like long ncRNAs. This can be done by cloning the whole genome library in the sgRNA sequence. These genomic sequences can be designed to target the site of interest, for instance, promoter of green fluorescent protein (GFP). Both the plasmid carrying the CRISPR/dCas9 components with genome library and the plasmid carrying the GFP can be co-transfected. By using

FACS cells with increased GFP signal can be selected and the enhancer-like long ncRNAs can be determined by sequencing. However, executing this screen is very challenging and would most probably require extensive optimization and validation.

4.3.2 The long non-coding RNA PARROT is an upstream regulator of c-Myc and affects proliferation and translation

Further investigation is needed to shed light onto the mechanism by which PARROT regulates c-Myc, which could reveal important mechanistic insight into how c-Myc is regulated in cancer.

Development of CRISPR/Cas9 technologies now allows an investigator to edit a locus of interest. Deletion of the gene encoding PARROT with this approach would allow us to verify that the effects we observe upon depletion of PARROT are mediated by it and would allow much easier further experimental manipulation. This was attempted but the deletion was either not successful or PARROT is an indispensable transcript and cells cannot survive without it. An interesting experiment would also be to activate the transcription of PARROT by the use of CRISPR/Cas9. Recent publication showed that this could be achieved with the use of special dCas9 mutant fused to VP64 activator co-transfected with MS2-p65-HSF1 fusion protein that improves the activation of the target (Konermann et al., 2015). This experiment would allow us to observe the effect that the endogenously expressed PARROT has in a cell line in which it is not expressed.

To further examine long ncRNAs- c-MYC interplay RNA immunoprecipitation with a c-Myc specific antibody followed by RNA-seq of the associated RNAs could be performed. This experiment could potentially reveal direct ncRNA binding partners of c-Myc that can be tested for the involvement in c-Myc network and effect on tumorigenesis. Potentially PARROT could be one of the long ncRNAs interacting directly with c-Myc *in vivo*.

Several methods for determining DNA binding sites and protein partners of long ncRNA of interest have been described: ChIRP (chromatin isolation by RNA purification), CHART (capture hybridization analysis of RNA targets), RAP (RNA antisense purification), ChIRP-MS (ChIRP optimized for protein identification by

mass spectrometry) and RAP-MS (RAP optimized for protein identification by mass spectrometry). By the use of biotinylated antisense oligonucleotides these methods allow a pull down of the long ncRNA of interest and since all interactions are chemically crosslinked, it allows the determination of DNA binding sites of the ncRNA through DNA sequencing and detection of proteins bound to the ncRNA by mass spectrometry (Chu et al., 2011; Chu et al., 2015; Engreitz et al., 2013; McHugh et al., 2015; Simon et al., 2011). Although these methods are very useful and informative they can only be applied to very long and highly expressed long ncRNAs. For instance, DNA binding sites and protein partners of a long ncRNA Xist have been determined by these methods. Long ncRNA Xist is expressed at levels comparable to that of a housekeeping gene GAPDH and even with this long ncRNA these approaches with subtle differences in the protocol yielded in different results due to high background that cannot be distinguished clearly from RNA specific signal (Chu et al., 2015; McHugh et al., 2015). Further development and improvement of this type of methods as well as the improvement of method used for the detection of ncRNA interactors, for instance improvement of the sensitivity of mass spectrometry instruments, will allow researchers to examine many long ncRNAs like PARROT in more details. The application of one of these methods on long ncRNA PARROT would shed light on its mechanism of action and identify direct DNA and protein interactors.

The involvement of long ncRNA PARROT in the development of cancer could also be further examined. A comprehensive examination of PARROT in wide variety of tumor samples could be performed. Additionally, its tumorigenic potential could be addressed *in vivo,* in mice. Nude mice, mice that lack immune system, can be injected with human cells lines, like HeLa, and the effect on tumor development in different organs can be monitored in the presence or absence of long ncRNA PARROT. This experiment could shed light on to how the oncogenic potential of long ncRNA PARROT is exerted and does it have an effect *in vivo*. This insight might point to PARROT as a biomarker of certain types of cancers, like stomach cancer. Furthermore, there is a novel trend in cancer therapy in which ncRNAs are targeted by the means of siRNAs, RNA or DNA antisense oligonucleotides. Therefore, plausibly, since PARROT is almost completely absent in the non-transformed stomach tissue and

highly expressed in transformed stomach tissue it can serve as an ideal target for therapeutics that can be used to treat patients suffering from this type of cancer.

REFERENCES

Adiseshaiah, P., Lindner, D.J., Kalvakolanu, D.V., and Reddy, S.P. (2007). FRA-1 proto-oncogene induces lung epithelial cell invasion and anchorage-independent growth in vitro, but is insufficient to promote tumor growth in vivo. Cancer Res *67*, 6204-6211.

Albertson, D.N., Schmidt, C.J., Kapatos, G., and Bannon, M.J. (2006). Distinctive profiles of gene expression in the human nucleus accumbens associated with cocaine and heroin abuse. Neuropsychopharmacology *31*, 2304-2312.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T.*, et al.* (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455-461.

Arnone, M.I., and Davidson, E.H. (1997). The hardwiring of development: organization and function of genomic regulatory systems. Development *124*, 1851-1864.

Ashe, H.L., Monks, J., Wijgerde, M., Fraser, P., and Proudfoot, N.J. (1997). Intergenic transcription and transinduction of the human beta-globin locus. Genes Dev *11*, 2494-2509.

Barrallo-Gimeno, A., and Nieto, M.A. (2005). The Snail genes as inducers of cell movement and survival: implications in development and cancer. Development *132*, 3151-3161.

Barry, G. (2014). The long non-coding RNA Gomafu is acutely regulated in response to neuronal activation and involved in schizophrenia-associated alternative splicing. Mol Psychiatry *19*, 486-494.

Barsyte-Lovejoy, D., Lau, S.K., Boutros, P.C., Khosravi, F., Jurisica, I., Andrulis, I.L., Tsao, M.S., and Penn, L.Z. (2006). The c-Myc oncogene directly induces the H19 noncoding RNA by allele-specific binding to potentiate tumorigenesis. Cancer Res *66*, 5330-5337.

Bertani, S., Sauer, S., Bolotin, E., and Sauer, F. (2011). The noncoding RNA Mistral activates Hoxa6 and Hoxa7 expression and stem cell differentiation by recruiting MLL1 to chromatin. Mol Cell *43*, 1040-1046.

Bertram, C., and Hass, R. (2009). Cellular senescence of human mammary epithelial cells (HMEC) is associated with an altered MMP-7/HB-EGF signaling and increased formation of elastin-like structures. Mech Ageing Dev *130*, 657-669.

Bhatt, D.M., Pandya-Jones, A., Tong, A.J., Barozzi, I., Lissner, M.M., Natoli, G., Black, D.L., and Smale, S.T. (2012). Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. Cell *150*, 279-290.

Blank, J.L., Liu, X.J., Cosmopoulos, K., Bouck, D.C., Garcia, K., Bernard, H., Tayber, O., Hather, G., Liu, R., Narayanan, U.*, et al.* (2013). Novel DNA damage checkpoints mediating cell death induced by the NEDD8-activating enzyme inhibitor MLN4924. Cancer Res *73*, 225-234.

Boyle, A.P. (2008). High-resolution mapping and characterization of open chromatin across the genome. Cell *132*, 311-322.

Cabili, M.N. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev *25*, 1915-1927.

Calo, E., and Wysocka, J. (2013). Modification of enhancer chromatin: what, how, and why? Mol Cell *49*, 825-837.

Cappellen, D., Schlange, T., Bauer, M., Maurer, F., and Hynes, N.E. (2007). Novel c-MYC target genes mediate differential effects on cell proliferation and migration. EMBO Rep *8*, 70-76.

Carrieri, C. (2012). Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. Nature *491*, 454-457.

Chen, B.J., Wu, Y.L., Tanaka, Y., and Zhang, W. (2014). Small molecules targeting c-Myc oncogene: promising anti-cancer therapeutics. Int J Biol Sci *10*, 1084-1096.

Chu, C., Qu, K., Zhong, F.L., Artandi, S.E., and Chang, H.Y. (2011). Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. Mol Cell *44*, 667-678.

Chu, C., Spitale, R.C., and Chang, H.Y. (2015). Technologies to probe functions and mechanisms of long noncoding RNAs. Nature structural & molecular biology *22*, 29-35.

Clouaire, T., Webb, S., Skene, P., Illingworth, R., Kerr, A., Andrews, R., Lee, J.H., Skalnik, D., and Bird, A. (2012). Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells. Genes Dev *26*, 1714-1728.

Collis, P., Antoniou, M., and Grosveld, F. (1990). Definition of the minimal requirements within the human beta-globin gene and the dominant control region for high level expression. Embo J *9*, 233-240.

Conrad, T., Marsico, A., Gehre, M., and Orom, U.A. (2014). Microprocessor activity controls differential miRNA biogenesis In Vivo. Cell Rep *9*, 542-554.

Corradin, O., Saiakhova, A., Akhtar-Zaidi, B., Myeroff, L., Willis, J., Cowper-Sal lari, R., Lupien, M., Markowitz, S., and Scacheri, P.C. (2014). Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. Genome Res *24*, 1-13.

Dang, C.V. (2012). MYC on the path to cancer. Cell *149*, 22-35.

Dang, C.V., Le, A., and Gao, P. (2009). MYC-induced cancer cell energy metabolism and therapeutic opportunities. Clin Cancer Res *15*, 6479-6483.

De Santa, F. (2010). A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PLoS Biol *8*, e1000384.

Del Bene, F. (2007). In vivo validation of a computationally predicted conserved Ath5 target gene set. PLoS Genet *3*, 1661-1671.

Deng, K., Guo, X., Wang, H., and Xia, J. (2014). The lncRNA-MYC regulatory network in cancer. Tumour Biol *35*, 9497-9503.

Derrien, T. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res *22*, 1775-1789.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G.*, et al.* (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res *22*, 1775-1789.

Dhamija, S., Doerrie, A., Winzen, R., Dittrich-Breiholz, O., Taghipour, A., Kuehne, N., Kracht, M., and Holtmann, H. (2010). IL-1-induced post-transcriptional mechanisms target overlapping translational silencing and destabilizing elements in IkappaBzeta mRNA. J Biol Chem *285*, 29165-29178.

Djebali, S. (2012). Landscape of transcription in human cells. Nature *489*, 101-108.

Engreitz, J.M., Pandya-Jones, A., McDonel, P., Shishkin, A., Sirokman, K., Surka, C., Kadri, S., Xing, J., Goren, A., Lander, E.S.*, et al.* (2013). The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. Science *341*, 1237973.

Faghihi, M.A., Modarresi, F., Khalil, A.M., Wood, D.E., Sahagan, B.G., Morgan, T.E., Finch, C.E., St. Laurent Iii, G., 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.

Fernandez-Cuesta, L., Sun, R., Menon, R., George, J., Lorenz, S., Meza-Zepeda, L.A., Peifer, M., Plenker, D., Heuckmann, J.M., Leenders, F.*, et al.* (2015). Identification of novel fusion genes in lung cancer using breakpoint assembly of transcriptome sequencing data. Genome Biol *16*, 7.

Fukata, Y., Oshiro, N., Kinoshita, N., Kawano, Y., Matsuoka, Y., Bennett, V., Matsuura, Y., and Kaibuchi, K. (1999). Phosphorylation of adducin by Rho-kinase plays a crucial role in cell motility. J Cell Biol *145*, 347-361.

Galvagni, F., Orlandini, M., and Oliviero, S. (2013). Role of the AP-1 transcription factor FOSL1 in endothelial cells adhesion and migration. Cell Adh Migr *7*, 408-411.

Garbe, J.C., Bhattacharya, S., Merchant, B., Bassett, E., Swisshelm, K., Feiler, H.S., Wyrobek, A.J., and Stampfer, M.R. (2009). Molecular distinctions between stasis and telomere attrition senescence barriers shown by long-term culture of normal human mammary epithelial cells. Cancer Res *69*, 7557-7568.

Geisler, S., and Coller, J. (2013). RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. Nat Rev Mol Cell Biol *14*, 699-712.

Gielisch, I., and Meierhofer, D. (2015). Metabolome and proteome profiling of complex I deficiency induced by rotenone. J Proteome Res *14*, 224-235.

Gomez, J.A., Wapinski, O.L., Yang, Y.W., Bureau, J.F., Gopinath, S., Monack, D.M., Chang, H.Y., Brahic, M., and Kirkegaard, K. (2013). The NeST long ncRNA controls microbial susceptibility and epigenetic activation of the interferon-gamma locus. Cell *152*, 743-754.

Gong, C., and Maquat, L.E. (2011). lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. Nature *470*, 284-288.

Hah, N., Murakami, S., Nagari, A., Danko, C.G., and Kraus, W.L. (2013). Enhancer transcripts mark active estrogen receptor binding sites. Genome Res *23*, 1210-1223.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S.*, et al.* (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res *22*, 1760-1774.

Heintzman, N.D. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature Genet *39*, 311-318.

Heintzman, N.D. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108-112.

Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W.*, et al.* (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108-112.

Heintzman, N.D., and Ren, B. (2009). Finding distal regulatory elements in the human genome. Current opinion in genetics & development *19*, 541-549.

Heinz, S., Romanoski, C.E., Benner, C., and Glass, C.K. (2015). The selection and function of cell type-specific enhancers. Nat Rev Mol Cell Biol *16*, 144-154.

Higgs, D.R., Vernimmen, D., Hughes, J., and Gibbons, R. (2007). Using genomics to study how chromatin influences gene expression. Annu Rev Genomics Hum Genet *8*, 299-325.

Hu, G., Lou, Z., and Gupta, M. (2014). The long non-coding RNA GAS5 cooperates with the eukaryotic translation initiation factor 4E to regulate c-Myc translation. PLoS One *9*, e107016.

Huarte, M., Guttman, M., Feldser, D., Garber, M., Koziol, M.J., Kenzelmann-Broz, D., Khalil, A.M., Zuk, O., Amit, I., Rabani, M.*, et al.* (2010). A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. Cell *142*, 409-419.

Huarte, M., and Rinn, J.L. (2010). Large non-coding RNAs: missing links in cancer? Hum Mol Genet *19*, R152-161.

Hung, C.L., Wang, L.Y., Yu, Y.L., Chen, H.W., Srivastava, S., Petrovics, G., and Kung, H.J. (2014). A long noncoding RNA connects c-Myc to tumor metabolism. Proc Natl Acad Sci U S A *111*, 18697-18702.

Hung, T., Wang, Y., Lin, M.F., Koegel, A.K., Kotake, Y., Grant, G.D., Horlings, H.M., Shah, N., Umbricht, C., Wang, P.*, et al.* (2011). Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. Nature genetics *43*, 621-629.

Ingolia, N.T., Brar, G.A., Rouskin, S., McGeachy, A.M., and Weissman, J.S. (2013). Genome-wide annotation and quantitation of translation by ribosome profiling. Curr Protoc Mol Biol *Chapter 4*, Unit 4 18.

Iyer, M.K., Niknafs, Y.S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T.R., Prensner, J.R., Evans, J.R., Zhao, S.*, et al.* (2015). The landscape of long noncoding RNAs in the human transcriptome. Nature genetics *47*, 199-208.

Jin, F. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature *503*, 290-294.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. Science *316*, 1497-1502.

Kheradpour, P., Stark, A., Roy, S., and Kellis, M. (2007). Reliable prediction of regulator targets using 12 Drosophila genomes. Genome Res *17*, 1919-1931.

Kim, M.S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S.*, et al.* (2014). A draft map of the human proteome. Nature *509*, 575-581.

Kim, T., Vidal, G.S., Djurisic, M., William, C.M., Birnbaum, M.E., Garcia, K.C., Hyman, B.T., and Shatz, C.J. (2013). Human LilrB2 is a beta-amyloid receptor and its murine homolog PirB regulates synaptic plasticity in an Alzheimer's model. Science *341*, 1399-1404.

Kim, T.K. (2010). Widespread transcription at neuronal activity-regulated enhancers. Nature *465*, 182-187.

Kino, T., Hurt, D.E., Ichijo, T., Nader, N., and Chrousos, G.P. (2010). Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor. Sci Signal *3*, ra8.

Koch, F., and Andrau, J.C. (2011). Initiating RNA polymerase II and TIPs as hallmarks of enhancer activity and tissue-specificity. Transcription *2*, 263-268.

Konermann, S., Brigham, M.D., Trevino, A.E., Joung, J., Abudayyeh, O.O., Barcena, C., Hsu, P.D., Habib, N., Gootenberg, J.S., Nishimasu, H.*, et al.* (2015). Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. Nature *517*, 583-588.

Kowalczyk, M.S., Hughes, J.R., Garrick, D., Lynch, M.D., Sharpe, J.A., Sloane-Stanley, J.A., McGowan, S.J., De Gobbi, M., Hosseini, M., Vernimmen, D.*, et al.* (2012). Intragenic enhancers act as alternative promoters. Mol Cell *45*, 447-458.

Kung, J.T., Colognori, D., and Lee, J.T. (2013). Long noncoding RNAs: past, present, and future. Genetics *193*, 651-669.

Kutter, C., Watt, S., Stefflova, K., Wilson, M.D., Goncalves, A., Ponting, C.P., Odom, D.T., and Marques, A.C. (2012). Rapid turnover of long noncoding RNAs and the evolution of gene expression. PLoS Genet *8*, e1002841.

Kvon, E.Z., Stampfel, G., Yanez-Cuna, J.O., Dickson, B.J., and Stark, A. (2012). HOT regions function as patterned developmental enhancers and have a distinct cis-regulatory signature. Genes Dev *26*, 908-913.

Lai, F. (2013). Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. Nature *494*, 497-501.

Lam, M.T., Cho, H., Lesch, H.P., Gosselin, D., Heinz, S., Tanaka-Oishi, Y., Benner, C., Kaikkonen, M.U., Kim, A.S., Kosaka, M.*, et al.* (2013). Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription. Nature *498*, 511-515.

Lam, M.T., Li, W., Rosenfeld, M.G., and Glass, C.K. (2014). Enhancer RNAs and regulated transcriptional programs. Trends Biochem Sci *39*, 170-182.

Landry, J.J., Pyl, P.T., Rausch, T., Zichner, T., Tekkedil, M.M., Stutz, A.M., Jauch, A., Aiyar, R.S., Pau, G., Delhomme, N.*, et al.* (2013). The genomic and transcriptomic landscape of a HeLa cell line. G3 (Bethesda) *3*, 1213-1224.

Leveille, N., Melo, C.A., Rooijers, K., Diaz-Lagares, A., Melo, S.A., Korkmaz, G., Lopes, R., Akbari Moqadam, F., Maia, A.R., Wijchers, P.J.*, et al.* (2015). Genome-wide profiling of p53-regulated enhancer RNAs uncovers a subset of enhancers controlled by a lncRNA. Nat Commun *6*, 6520.

Li, B., Antonyak, M.A., Zhang, J., and Cerione, R.A. (2012a). RhoA triggers a specific signaling pathway that generates transforming microvesicles in cancer cells. Oncogene *31*, 4740-4749.

Li, G. (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell *148*, 84-98.

Li, G., Ruan, X., Auerbach, R.K., Sandhu, K.S., Zheng, M., Wang, P., Poh, H.M., Goh, Y., Lim, J., Zhang, J.*, et al.* (2012b). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell *148*, 84-98.

Li, W. (2013). Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. Nature *498*, 516-520.

Li, W., Notani, D., Ma, Q., Tanasa, B., Nunez, E., Chen, A.Y., Merkurjev, D., Zhang, J., Ohgi, K., Song, X.*, et al.* (2013). Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. Nature *498*, 516-520.

Liao, L.M., Sun, X.Y., Liu, A.W., Wu, J.B., Cheng, X.L., Lin, J.X., Zheng, M., and Huang, L. (2014). Low expression of long noncoding XLOC_010588 indicates a poor prognosis and promotes proliferation through upregulation of c-Myc in cervical cancer. Gynecol Oncol *133*, 616-623.

Lin, Y.C., Benner, C., Mansson, R., Heinz, S., Miyazaki, K., Miyazaki, M., Chandra, V., Bossen, C., Glass, C.K., and Murre, C. (2012). Global changes in the nuclear positioning of genes and intra- and interdomain genomic interactions that orchestrate B cell fate. Nat Immunol *13*, 1196-1204.

Ling, H., Spizzo, R., Atlasi, Y., Nicoloso, M., Shimizu, M., Redis, R.S., Nishida, N., Gafa, R., Song, J., Guo, Z.*, et al.* (2013). CCAT2, a novel noncoding RNA mapping to 8q24, underlies metastatic progression and chromosomal instability in colon cancer. Genome Res *23*, 1446-1461.

Lipovich, L., Dachet, F., Cai, J., Bagla, S., Balan, K., Jia, H., and Loeb, J.A. (2012). Activity-dependent human brain coding/noncoding gene regulatory networks. Genetics *192*, 1133-1148.

Maston, G.A., Evans, S.K., and Green, M.R. (2006). Transcriptional regulatory elements in the human genome. Annu Rev Genomics Hum Genet *7*, 29-59.

McHugh, C.A., Chen, C.K., Chow, A., Surka, C.F., Tran, C., McDonel, P., Pandya-Jones, A., Blanco, M., Burghard, C., Moradian, A.*, et al.* (2015). The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. Nature *521*, 232-236.

Melo, C.A., Drost, J., Wijchers, P.J., van de Werken, H., de Wit, E., Oude Vrielink, J.A., Elkon, R., Melo, S.A., Leveille, N., Kalluri, R.*, et al.* (2013). eRNAs are required for p53-dependent enhancer activity and gene transcription. Mol Cell *49*, 524-535.

Meyer, K.D., Donner, A.J., Knuesel, M.T., York, A.G., Espinosa, J.M., and Taatjes, D.J. (2008). Cooperative activity of cdk8 and GCN5L within Mediator directs tandem phosphoacetylation of histone H3. Embo J *27*, 1447-1457.

Meyer, N., and Penn, L.Z. (2008). Reflecting on 25 years with MYC. Nat Rev Cancer *8*, 976-990.

Modarresi, F., Faghihi, M.A., Patel, N.S., Sahagan, B.G., Wahlestedt, C., and Lopez-Toledano, M.A. (2011). Knockdown of BACE1-AS Nonprotein-Coding Transcript Modulates Beta-Amyloid-Related Hippocampal Neurogenesis. Int J Alzheimers Dis *2011*, 929042.

Moorman, C. (2006). Hotspots of transcription factor colocalization in the genome of Drosophila melanogaster. Proc Natl Acad Sci USA *103*, 12027-12032.

Morris, K.V., and Mattick, J.S. (2014). The rise of regulatory RNA. Nat Rev Genet *15*, 423-437.

Musahl, A.S., Huang, X., Rusakiewicz, S., Ntini, E., Marsico, A., Kroemer, G., Kepp, O., and Orom, U.A. (2015). A long non-coding RNA links calreticulin-mediated immunogenic cell removal to RB1 transcription. Oncogene.

Natoli, G., and Andrau, J.C. (2012). Noncoding transcription at enhancers: general principles and functional models. Annu Rev Genet *46*, 1-19.

Nishihira, J. (2000). Macrophage migration inhibitory factor (MIF): its essential role in the immune system and cell growth. J Interferon Cytokine Res *20*, 751-762.

Nunes-Xavier, C.E., Tarrega, C., Cejudo-Marin, R., Frijhoff, J., Sandin, A., Ostman, A., and Pulido, R. (2010). Differential up-regulation of MAP kinase phosphatases MKP3/DUSP6 and DUSP5 by Ets2 and c-Jun converge in the control of the growth arrest versus proliferation response of MCF-7 breast cancer cells to phorbol ester. J Biol Chem *285*, 26417-26430.

Ong, C.T., and Corces, V.G. (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. Nat Rev Genet *12*, 283-293.

Onoguchi, M., Hirabayashi, Y., Koseki, H., and Gotoh, Y. (2012). A noncoding RNA regulates the neurogenin1 gene locus during mouse neocortical development. Proc Natl Acad Sci U S A *109*, 16939-16944.

Orom, U.A. (2010). Long noncoding RNAs with enhancer-like function in human cells. Cell *143*, 46-58.

Orom, U.A., and Shiekhattar, R. (2013). Long noncoding RNAs usher in a new era in the biology of enhancers. Cell *154*, 1190-1193.

Pasmant, E., Sabbagh, A., Vidaud, M., and Bieche, I. (2011). ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS. Faseb J *25*, 444-448.

Piazzolla, D., Meissl, K., Kucerova, L., Rubiolo, C., and Baccarini, M. (2005). Raf-1 sets the threshold of Fas sensitivity by modulating Rok-alpha signaling. J Cell Biol *171*, 1013-1022.

Prendergast, G.C. (2001). Actin' up: RhoB in cancer and apoptosis. Nat Rev Cancer *1*, 162-168.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

Rinn, J.L. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell *129*, 1311-1323.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139-140.

Sanchez, Y., Segura, V., Marin-Bejar, O., Athie, A., Marchese, F.P., Gonzalez, J., Bujanda, L., Guo, S., Matheu, A., and Huarte, M. (2014). Genome-wide analysis of the human p53 transcriptional network unveils a lncRNA tumour suppressor signature. Nat Commun *5*, 5812.

Sanyal, A., Lajoie, B.R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. Nature *489*, 109-113.

Shechner, D.M., Hacisuleyman, E., Younger, S.T., and Rinn, J.L. (2015). Multiplexable, locus-specific targeting of long RNAs with CRISPR-Display. Nature Methods *12*, 664-670.

Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet *15*, 272-286.

Sigova, A.A., Mullen, A.C., Molinie, B., Gupta, S., Orlando, D.A., Guenther, M.G., Almada, A.E., Lin, C., Sharp, P.A., Giallourakis, C.C.*, et al.* (2013). Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. Proc Natl Acad Sci U S A *110*, 2876-2881.

Simon, M.D., Wang, C.I., Kharchenko, P.V., West, J.A., Chapman, B.A., Alekseyenko, A.A., Borowsky, M.L., Kuroda, M.I., and Kingston, R.E. (2011). The genomic binding sites of a noncoding RNA. Proc Natl Acad Sci U S A *108*, 20497-20502.

Smallwood, A., and Ren, B. (2013). Genome organization and long-range regulation of gene expression by enhancers. Curr Opin Cell Biol *25*, 387-394.

Spitz, F., and Furlong, E.E. (2012). Transcription factors: from enhancer binding to developmental control. Nature Rev Genet *13*, 613-626.

St Laurent, G., Wahlestedt, C., and Kapranov, P. (2015). The Landscape of long noncoding RNA classification. Trends Genet *31*, 239-251.

Sun, M., Gadad, S.S., Kim, D.S., and Kraus, W.L. (2015). Discovery, Annotation, and Functional Analysis of Long Noncoding RNAs Controlling Cell-Cycle Gene Expression and Proliferation in Breast Cancer Cells. Mol Cell *59*, 698-711.

Sun, S., Del Rosario, B.C., Szanto, A., Ogawa, Y., Jeon, Y., and Lee, J.T. (2013). Jpx RNA activates Xist by evicting CTCF. Cell *153*, 1537-1551.

Surawska, H., Ma, P.C., and Salgia, R. (2004). The role of ephrins and Eph receptors in cancer. Cytokine Growth Factor Rev *15*, 419-433.

Thurman, R.E. (2013). The accessible chromatin landscape of the human genome. Nature *488*, 75-82.

Tian, D., Sun, S., and Lee, J.T. (2010). The long noncoding RNA, Jpx, is a molecular switch for X chromosome inactivation. Cell *143*, 390-403.

Tilgner, H., Knowles, D.G., Johnson, R., Davis, C.A., Chakrabortty, S., Djebali, S., Curado, J., Snyder, M., Gingeras, T.R., and Guigo, R. (2012). Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. Genome Res *22*, 1616-1625.

Tripathi, V. (2010). The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. Mol Cell *39*, 925-938.

Tsuiji, H., Yoshimoto, R., Hasegawa, Y., Furuno, M., Yoshida, M., and Nakagawa, S. (2011). Competition between a noncoding exon and introns: Gomafu contains tandem UACUAAC repeats and associates with splicing factor-1. Genes Cells *16*, 479-490.

Tuan, D., Kong, S., and Hu, K. (1992). Transcription of the hypersensitive site HS2 enhancer in erythroid cells. Proc Natl Acad Sci U S A *89*, 11219-11223.

Ulitsky, I., and Bartel, D.P. (2013). lincRNAs: genomics, evolution, and mechanisms. Cell *154*, 26-46.

van Riggelen, J., Yetil, A., and Felsher, D.W. (2010). MYC as a regulator of ribosome biogenesis and protein synthesis. Nat Rev Cancer *10*, 301-309.

van Steensel, B., and Dekker, J. (2010). Genomics tools for unraveling chromosome architecture. Nature Biotech *28*, 1089-1095.

Vashishtha, M., Ng, C.W., Yildirim, F., Gipson, T.A., Kratter, I.H., Bodai, L., Song, W., Lau, A., Labadorf, A., Vogel-Ciernia, A.*, et al.* (2013). Targeting H3K4 trimethylation in Huntington disease. Proc Natl Acad Sci U S A *110*, E3027-3036.

Visel, A., Blow, M.J., Li, Z., Zhang, T., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F.*, et al.* (2009a). ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature *457*, 854-858.

Visel, A., Rubin, E.M., and Pennacchio, L.A. (2009b). Genomic views of distant-acting enhancers. Nature *461*, 199-205.

Vucicevic, D., Corradin, O., Ntini, E., Scacheri, P.C., and Orom, U.A. (2015). Long ncRNA expression associates with tissue-specific enhancers. Cell Cycle *14*, 253-260.

Vucicevic, D., Schrewe, H., and Orom, U.A. (2014). Molecular mechanisms of long ncRNAs in neurological disorders. Front Genet *5*, 48.

Wang, K.C., Yang, Y.W., Liu, B., Sanyal, A., Corces-Zimmerman, R., Chen, Y., Lajoie, B.R., Protacio, A., Flynn, R.A., Gupta, R.A.*, et al.* (2011a). A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature *472*, 120-124.

Wang, Q., Bailey, C.G., Ng, C., Tiffen, J., Thoeng, A., Minhas, V., Lehman, M.L., Hendy, S.C., Buchanan, G., Nelson, C.C.*, et al.* (2011b). Androgen receptor and nutrient signaling pathways coordinate the demand for increased amino acid transport during prostate cancer progression. Cancer Res *71*, 7525-7536.

Whyte, W.A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell *153*, 307-319.

Xiang, J.F., Yin, Q.F., Chen, T., Zhang, Y., Zhang, X.O., Wu, Z., Zhang, S., Wang, H.B., Ge, J., Lu, X.*, et al.* (2014). Human colorectal cancer-specific CCAT1-L lncRNA regulates long-range chromatin interactions at the MYC locus. Cell Res *24*, 513-531.

Yang, F., Xue, X., Bi, J., Zheng, L., Zhi, K., Gu, Y., and Fang, G. (2013). Long noncoding RNA CCAT1, which could be activated by c-Myc, promotes the progression of gastric carcinoma. J Cancer Res Clin Oncol *139*, 437-445.

Yang, F., Xue, X., Zheng, L., Bi, J., Zhou, Y., Zhi, K., Gu, Y., and Fang, G. (2014). Long non-coding RNA GHET1 promotes gastric carcinoma cell proliferation by increasing c-Myc mRNA stability. Febs J *281*, 802-813.

Zhou, V.W., Goren, A., and Bernstein, B.E. (2011). Charting histone modifications and the functional organization of mammalian genomes. Nat Rev Genet *12*, 7-18.

APPENDIX

Appendix 1. Differentially expressed genes upon depletion of PARROT as determined by RNA-seq (ens-ID -ensemble gene ID; FC-fold change; CPM- counts-per-million; FDR-false discovery rate; DE-differential expression where 1 indicates upregulation and -1 indicates downregulation).

ens-id	logFC	logCPM	P-value	FDR	DE
ENSG00000108468	0.85	6.65	0.00	0.00	1
ENSG00000162368	0.80	7.14	0.00	0.00	1
ENSG00000170540	0.66	8.12	0.00	0.00	1
ENSG00000108960	0.62	4.37	0.00	0.00	1
ENSG00000171314	0.72	5.44	0.00	0.00	1
ENSG00000144228	0.64	4.96	0.00	0.00	1
ENSG00000146842	0.55	5.54	0.00	0.00	1
ENSG00000072786	0.48	5.27	0.00	0.00	1
ENSG00000172667	0.66	4.60	0.00	0.00	1
ENSG00000184949	0.55	3.42	0.00	0.00	1
ENSG00000198146	0.55	6.71	0.00	0.00	1
ENSG00000145293	0.64	5.87	0.00	0.00	1
ENSG00000100580	0.45	6.04	0.00	0.00	1
ENSG00000072571	0.51	6.35	0.00	0.00	1
ENSG00000006576	0.47	5.55	0.00	0.00	1
ENSG00000137992	0.52	4.72	0.00	0.00	1
ENSG00000212768	0.68	4.95	0.00	0.00	1
ENSG00000151414	0.99	5.59	0.00	0.00	1
ENSG00000157617	0.42	5.87	0.00	0.00	1
ENSG00000163597	0.40	6.64	0.00	0.00	1
ENSG00000165449	0.52	2.64	0.00	0.00	1
ENSG00000126821	0.64	3.83	0.00	0.00	1
ENSG00000135069	0.67	8.91	0.00	0.00	1
ENSG00000142166	0.43	5.58	0.00	0.00	1
ENSG00000152642	0.46	6.48	0.00	0.00	1
ENSG00000118307	1.31	-1.00	0.00	0.00	1
ENSG00000127314	0.51	5.45	0.00	0.00	1
ENSG00000168026	0.52	2.22	0.00	0.00	1
ENSG00000134698	0.45	3.66	0.00	0.00	1
ENSG00000122367	1.24	-1.57	0.00	0.00	1
ENSG00000171241	0.51	5.95	0.00	0.00	1
ENSG00000180329	0.40	5.70	0.00	0.00	1
ENSG00000080371	0.48	5.04	0.00	0.00	1
ENSG00000077514	0.37	5.63	0.00	0.00	1
ENSG00000237714	0.65	1.51	0.00	0.00	1
ENSG00000143033	0.44	5.73	0.00	0.00	1
ENSG00000165244	0.53	4.95	0.00	0.00	1
ENSG00000112893	0.41	6.45	0.00	0.00	1
ENSG00000100583	1.10	-0.47	0.00	0.00	1
ENSG00000115414	0.39	4.73	0.00	0.00	1
ENSG00000167065	0.54	1.74	0.00	0.00	1
ENSG00000151553	0.44	4.50	0.00	0.00	1
ENSG00000062582	0.38	5.93	0.00	0.00	1
ENSG00000134602	0.53	6.58	0.00	0.00	1
ENSG00000153707	0.44	4.27	0.00	0.00	1
ENSG00000241772	0.70	1.44	0.00	0.00	1
ENSG00000165898	0.50	3.75	0.00	0.00	1

Appendix 1 continued

ens-id	logFC	logCPM	P-value	FDR	DE
ENSG00000134318	0.39	6.12	0.00	0.00	1
ENSG00000197603	0.41	4.61	0.00	0.00	1
ENSG00000139433	0.39	3.94	0.00	0.00	1
ENSG00000159082	0.41	3.90	0.00	0.00	1
ENSG00000101844	0.51	3.05	0.00	0.00	1
ENSG00000118276	0.69	0.92	0.00	0.00	1
ENSG00000167984	0.81	-0.11	0.00	0.00	1
ENSG00000251018	0.55	1.99	0.00	0.00	1
ENSG00000115267	0.85	2.24	0.00	0.00	1
ENSG00000110911	0.32	6.80	0.00	0.00	1
ENSG00000161570	3.54	-0.36	0.00	0.00	1
ENSG00000100320	0.33	7.98	0.00	0.00	1
ENSG00000070214	0.46	5.82	0.00	0.00	1
ENSG00000167525	0.62	1.56	0.00	0.00	1
ENSG00000144043	0.35	5.83	0.00	0.00	1
ENSG00000089723	0.56	3.42	0.00	0.00	1
ENSG00000109436	0.37	4.06	0.00	0.00	1
ENSG00000165322	0.36	5.63	0.00	0.00	1
ENSG00000087502	0.39	6.46	0.00	0.00	1
ENSG00000137965	1.94	0.82	0.00	0.00	1
ENSG00000205464	0.49	2.65	0.00	0.00	1
ENSG00000135114	1.90	3.16	0.00	0.01	1
ENSG00000151151	0.57	3.88	0.00	0.01	1
ENSG00000109929	0.48	5.40	0.00	0.01	1
ENSG00000162599	0.44	4.85	0.00	0.01	1
ENSG00000137941	0.44	3.70	0.00	0.01	1
ENSG00000157578	0.56	1.64	0.00	0.01	1
ENSG00000104093	0.36	5.29	0.00	0.01	1
ENSG00000138642	0.72	1.28	0.00	0.01	1
ENSG00000116459	0.31	7.97	0.00	0.01	1
ENSG00000030419	0.56	1.88	0.00	0.01	1
ENSG00000047346	0.44	2.92	0.00	0.01	1
ENSG00000100228	0.61	0.61	0.00	0.01	1
ENSG00000118762	0.39	4.32	0.00	0.01	1
ENSG00000151575	0.58	0.93	0.00	0.01	1
ENSG00000134313	0.30	5.84	0.00	0.01	1
ENSG00000027697	0.39	4.69	0.00	0.01	1
ENSG00000096968	0.49	3.00	0.00	0.01	1
ENSG00000116574	0.45	2.83	0.00	0.01	1
ENSG00000135052	0.32	6.33	0.00	0.01	1
ENSG00000128284	2.77	-2.23	0.00	0.01	1
ENSG00000131437	0.40	4.40	0.00	0.01	1
ENSG00000095015	0.33	4.67	0.00	0.01	1
ENSG00000180998	0.99	-0.44	0.00	0.01	1
ENSG00000119042	0.32	4.70	0.00	0.01	1
ENSG00000074621	0.40	2.68	0.00	0.01	1
ENSG00000134265	0.33	5.22	0.00	0.01	1
ENSG00000197168	0.67	-0.15	0.00	0.01	1
ENSG00000257026	0.44	3.05	0.00	0.01	1
ENSG00000266801	1.60	-2.74	0.00	0.01	1
ENSG00000172318	0.72	-0.20	0.00	0.01	1
ENSG00000137710	0.46	7.02	0.00	0.01	1
ENSG00000164402	0.29	5.69	0.00	0.01	1
ENSG00000138760	0.33	6.17	0.00	0.01	1
ENSG00000174684	0.65	3.53	0.00	0.01	1
ENSG00000184486	0.84	-0.26	0.00	0.01	1
ENSG00000197568	0.40	3.38	0.00	0.01	1
ENSG00000119917	1.67	3.45	0.00	0.01	1
ENSG00000118997	0.64	-0.14	0.00	0.01	1
ENSG00000127989	0.38	3.52	0.00	0.02	1
ENSG00000107164	0.29	5.97	0.00	0.02	1

Appendix 1 continued

ens-id	logFC	logCPM	P-value	FDR	DE
ENSG00000072415	0.30	6.57	0.00	0.02	1
ENSG00000256223	0.44	2.15	0.00	0.02	1
ENSG00000133678	0.46	2.89	0.00	0.02	1
ENSG00000103160	0.40	5.41	0.00	0.02	1
ENSG00000166106	0.50	3.65	0.00	0.02	1
ENSG00000238261	0.58	0.50	0.00	0.02	1
ENSG00000149948	0.34	5.26	0.00	0.02	1
ENSG00000188706	0.29	5.49	0.00	0.02	1
ENSG00000124098	0.36	3.74	0.00	0.02	1
ENSG00000181143	0.34	4.33	0.00	0.02	1
ENSG00000080546	0.53	2.90	0.00	0.02	1
ENSG00000105875	0.43	2.44	0.00	0.02	1
ENSG00000167680	0.54	2.55	0.00	0.02	1
ENSG00000145012	0.32	6.91	0.00	0.02	1
ENSG00000062725	0.34	5.23	0.00	0.02	1
ENSG00000145354	0.34	4.81	0.00	0.02	1
ENSG00000132964	0.32	5.18	0.00	0.02	1
ENSG00000178695	0.78	0.99	0.00	0.02	1
ENSG00000197417	0.36	4.26	0.00	0.02	1
ENSG00000196247	0.39	3.85	0.00	0.02	1
ENSG00000196083	0.38	3.71	0.00	0.02	1
ENSG00000133640	0.43	1.87	0.00	0.02	1
ENSG00000120162	0.55	0.50	0.00	0.02	1
ENSG00000113356	0.44	5.46	0.00	0.02	1
ENSG00000119862	0.36	4.55	0.00	0.02	1
ENSG00000122591	0.35	5.79	0.00	0.02	1
ENSG00000197857	0.37	3.15	0.00	0.02	1
ENSG00000111727	0.45	2.48	0.00	0.02	1
ENSG00000197442	0.33	4.61	0.00	0.02	1
ENSG00000162878	0.48	0.95	0.00	0.02	1
ENSG00000241790	0.52	0.65	0.00	0.02	1
ENSG00000128581	0.34	4.15	0.00	0.02	1
ENSG00000096093	0.33	4.46	0.00	0.02	1
ENSG00000134755	0.37	4.50	0.00	0.03	1
ENSG00000004799	0.33	6.00	0.00	0.03	1
ENSG00000157184	0.31	4.77	0.00	0.03	1
ENSG00000236723	0.38	2.60	0.00	0.03	1
ENSG00000128609	0.40	6.05	0.00	0.03	1
ENSG00000151470	0.35	3.19	0.00	0.03	1
ENSG00000107890	0.31	4.52	0.00	0.03	1
ENSG00000106004	0.43	3.96	0.00	0.03	1
ENSG00000240571	0.44	1.88	0.00	0.03	1
ENSG00000084676	0.35	3.74	0.00	0.03	1
ENSG00000254894	0.45	1.86	0.00	0.03	1
ENSG00000166167	0.30	4.51	0.00	0.03	1
ENSG00000064393	0.46	4.07	0.00	0.03	1
ENSG00000166483	0.29	6.51	0.00	0.03	1
ENSG00000260597	0.57	0.17	0.00	0.03	1
ENSG00000006468	0.33	4.76	0.00	0.03	1
ENSG00000163521	0.44	1.43	0.00	0.03	1
ENSG00000168216	0.36	4.58	0.00	0.03	1
ENSG00000171503	0.29	4.91	0.00	0.03	1
ENSG00000149488	1.15	-2.37	0.00	0.03	1
ENSG00000134901	0.30	4.70	0.00	0.03	1
ENSG00000102287	0.27	6.18	0.00	0.03	1
ENSG00000114354	0.27	7.06	0.00	0.03	1
ENSG00000155530	1.05	-1.92	0.00	0.03	1
ENSG00000148481	0.30	4.75	0.00	0.03	1
ENSG00000111554	0.35	3.38	0.00	0.03	1
ENSG00000168772	0.63	0.53	0.00	0.03	1
ENSG00000089091	0.51	0.71	0.00	0.03	1

Appendix 1 continued

ens-id	logFC	logCPM	P-value	FDR	DE
ENSG00000182253	0.40	3.61	0.00	0.03	1
ENSG00000166436	0.27	5.85	0.00	0.03	1
ENSG00000260236	0.56	0.42	0.00	0.03	1
ENSG00000115525	0.40	4.63	0.00	0.03	1
ENSG00000169548	1.14	-1.97	0.00	0.03	1
ENSG00000139505	0.32	4.72	0.00	0.03	1
ENSG00000228140	1.09	-2.10	0.00	0.03	1
ENSG00000260686	0.95	-1.52	0.00	0.04	1
ENSG00000163328	0.60	0.95	0.00	0.04	1
ENSG00000112182	0.39	2.39	0.00	0.04	1
ENSG00000152749	0.35	5.56	0.00	0.04	1
ENSG00000110931	0.26	5.38	0.00	0.04	1
ENSG00000185621	0.31	4.55	0.00	0.04	1
ENSG00000255038	0.85	-1.63	0.00	0.04	1
ENSG00000100342	1.45	-0.54	0.00	0.04	1
ENSG00000122417	0.32	5.09	0.00	0.04	1
ENSG00000266456	0.48	0.61	0.00	0.04	1
ENSG00000260465	0.33	5.24	0.00	0.04	1
ENSG00000197020	0.41	2.15	0.00	0.04	1
ENSG00000185567	0.33	7.05	0.00	0.04	1
ENSG00000163214	0.27	5.07	0.00	0.04	1
ENSG00000105939	0.37	6.21	0.00	0.04	1
ENSG00000152022	0.32	5.87	0.00	0.04	1
ENSG00000142687	0.25	5.95	0.00	0.04	1
ENSG00000145569	0.51	3.81	0.00	0.04	1
ENSG00000137628	0.90	2.23	0.00	0.04	1
ENSG00000138646	0.69	3.35	0.00	0.04	1
ENSG00000087253	0.31	5.26	0.00	0.04	1
ENSG00000163482	0.27	5.17	0.00	0.04	1
ENSG00000157106	0.27	8.56	0.00	0.04	1
ENSG00000198363	0.38	10.86	0.00	0.04	1
ENSG00000111911	0.40	4.48	0.00	0.04	1
ENSG00000123384	0.37	4.25	0.00	0.04	1
ENSG00000058091	0.35	3.09	0.00	0.04	1
ENSG00000186638	0.33	3.98	0.00	0.04	1
ENSG00000143878	0.36	3.28	0.00	0.04	1
ENSG00000112210	0.35	4.08	0.00	0.04	1
ENSG00000148019	0.35	5.14	0.00	0.04	1
ENSG00000136100	0.26	6.08	0.00	0.04	1
ENSG00000120992	0.48	6.94	0.00	0.04	1
ENSG00000009335	0.27	7.32	0.00	0.04	1
ENSG00000144357	0.30	5.28	0.00	0.04	1
ENSG00000113369	0.29	4.89	0.00	0.04	1
ENSG00000204052	0.89	-1.78	0.00	0.04	1
ENSG00000261389	0.66	-0.91	0.00	0.04	1
ENSG00000184995	0.47	0.51	0.00	0.05	1
ENSG00000171467	0.28	5.38	0.00	0.05	1
ENSG00000148468	0.31	3.25	0.00	0.05	1
ENSG00000163378	0.33	3.91	0.00	0.05	1
ENSG00000261485	0.44	1.06	0.00	0.05	1
ENSG00000203880	0.30	5.42	0.00	0.05	1
ENSG00000259895	0.89	-1.36	0.00	0.05	1
ENSG00000165288	0.27	6.06	0.00	0.05	1
ENSG00000106031	0.63	0.03	0.00	0.05	1
ENSG00000263859	0.54	-0.20	0.00	0.05	1
ENSG00000119636	0.37	1.89	0.00	0.05	1
ENSG00000164049	0.88	-1.87	0.00	0.05	1
ENSG00000111711	0.36	6.29	0.00	0.05	1
ENSG00000088854	0.30	3.78	0.00	0.05	1
ENSG00000235261	0.34	2.32	0.00	0.05	1
ENSG00000128596	0.81	-1.32	0.00	0.05	1

Appendix 1 continued

ens-id	logFC	logCPM	P-value	FDR	DE
ENSG00000188629	0.40	1.77	0.00	0.05	1
ENSG00000223784	-1.53	2.56	0.00	0.00	-1
ENSG00000100300	-0.69	6.41	0.00	0.00	-1
ENSG00000254010	-0.71	1.49	0.00	0.00	-1
ENSG00000230882	-0.75	3.22	0.00	0.00	-1
ENSG00000241553	-0.43	6.72	0.00	0.00	-1
ENSG00000169100	-0.80	1.72	0.00	0.00	-1
ENSG00000181649	-0.60	3.99	0.00	0.00	-1
ENSG00000256167	-1.03	0.10	0.00	0.00	-1
ENSG00000224281	-0.57	4.32	0.00	0.00	-1
ENSG00000235852	-0.66	1.76	0.00	0.00	-1
ENSG00000178550	-0.87	1.91	0.00	0.00	-1
ENSG00000143549	-0.40	8.83	0.00	0.00	-1
ENSG00000166847	-0.41	6.61	0.00	0.00	-1
ENSG00000172366	-0.58	3.90	0.00	0.00	-1
ENSG00000107404	-0.53	6.23	0.00	0.00	-1
ENSG00000118523	-0.57	3.16	0.00	0.00	-1
ENSG00000187840	-0.49	5.75	0.00	0.00	-1
ENSG00000140859	-0.45	5.94	0.00	0.00	-1
ENSG00000236552	-0.67	3.27	0.00	0.00	-1
ENSG00000218537	-0.66	5.64	0.00	0.00	-1
ENSG00000240972	-0.68	5.73	0.00	0.00	-1
ENSG00000224126	-1.38	-1.82	0.00	0.00	-1
ENSG00000167987	-0.47	4.33	0.00	0.00	-1
ENSG00000232888	-0.67	0.37	0.00	0.00	-1
ENSG00000103257	-0.50	9.64	0.00	0.00	-1
ENSG00000170545	-0.42	4.67	0.00	0.00	-1
ENSG00000260466	-0.54	4.96	0.00	0.00	-1
ENSG00000175602	-0.66	2.35	0.00	0.00	-1
ENSG00000142871	-0.37	5.44	0.00	0.00	-1
ENSG00000224858	-0.66	1.75	0.00	0.00	-1
ENSG00000176978	-0.45	5.12	0.00	0.00	-1
ENSG00000213442	-0.61	2.08	0.00	0.00	-1
ENSG00000229344	-0.53	4.81	0.00	0.00	-1
ENSG00000213225	-1.17	-1.71	0.00	0.00	-1
ENSG00000258088	-0.76	-0.08	0.00	0.00	-1
ENSG00000133101	-0.37	6.26	0.00	0.00	-1
ENSG00000198517	-0.45	4.68	0.00	0.00	-1
ENSG00000248527	-0.44	5.83	0.00	0.00	-1
ENSG00000225159	-0.53	2.20	0.00	0.00	-1
ENSG00000229251	-0.75	0.20	0.00	0.00	-1
ENSG00000100292	-0.37	5.04	0.00	0.00	-1
ENSG00000170638	-0.44	5.27	0.00	0.00	-1
ENSG00000262323	-0.62	3.27	0.00	0.00	-1
ENSG00000233762	-0.64	1.77	0.00	0.00	-1
ENSG00000172409	-0.46	4.09	0.00	0.00	-1
ENSG00000228502	-0.62	4.74	0.00	0.01	-1
ENSG00000174886	-0.42	5.33	0.00	0.01	-1
ENSG00000166166	-0.51	4.01	0.00	0.01	-1
ENSG00000130383	-0.42	5.05	0.00	0.01	-1
ENSG00000228981	-0.73	1.23	0.00	0.01	-1
ENSG00000175592	-0.33	6.87	0.00	0.01	-1
ENSG00000167962	-0.48	5.89	0.00	0.01	-1
ENSG00000237039	-0.76	0.23	0.00	0.01	-1
ENSG00000223529	-0.67	1.16	0.00	0.01	-1
ENSG00000112658	-0.38	5.93	0.00	0.01	-1
ENSG00000140961	-0.61	3.61	0.00	0.01	-1
ENSG00000224094	-0.60	1.08	0.00	0.01	-1
ENSG00000243679	-0.64	0.96	0.00	0.01	-1
ENSG00000158615	-0.36	7.03	0.00	0.01	-1
ENSG00000182795	-0.58	1.57	0.00	0.01	-1

Appendix 1 continued

ens-id	logFC	logCPM	P-value	FDR	DE
ENSG00000173621	-0.48	3.76	0.00	0.01	-1
ENSG00000175920	-0.79	0.26	0.00	0.01	-1
ENSG00000229659	-0.61	0.34	0.00	0.01	-1
ENSG00000244265	-0.55	0.96	0.00	0.01	-1
ENSG00000258048	-0.99	-0.27	0.00	0.01	-1
ENSG00000108106	-0.38	5.49	0.00	0.01	-1
ENSG00000260879	-0.59	0.86	0.00	0.01	-1
ENSG00000214708	-0.60	0.27	0.00	0.01	-1
ENSG00000241429	-0.82	1.01	0.00	0.01	-1
ENSG00000262136	-0.34	5.56	0.00	0.01	-1
ENSG00000214391	-0.56	2.47	0.00	0.01	-1
ENSG00000266163	-0.98	-0.41	0.00	0.01	-1
ENSG00000242299	-0.53	2.81	0.00	0.01	-1
ENSG00000102753	-0.31	6.51	0.00	0.01	-1
ENSG00000218175	-0.53	1.96	0.00	0.01	-1
ENSG00000177169	-0.46	4.84	0.00	0.01	-1
ENSG00000204196	-0.53	2.40	0.00	0.01	-1
ENSG00000226360	-0.55	1.73	0.00	0.01	-1
ENSG00000163814	-0.34	7.12	0.00	0.01	-1
ENSG00000159346	-0.28	6.85	0.00	0.01	-1
ENSG00000099795	-0.41	4.46	0.00	0.01	-1
ENSG00000139644	-0.31	9.30	0.00	0.01	-1
ENSG00000164117	-0.38	2.93	0.00	0.01	-1
ENSG00000170525	-0.30	5.94	0.00	0.01	-1
ENSG00000104904	-0.32	8.44	0.00	0.01	-1
ENSG00000070423	-0.48	4.02	0.00	0.01	-1
ENSG00000242372	-0.35	7.52	0.00	0.01	-1
ENSG00000148426	-0.53	2.73	0.00	0.01	-1
ENSG00000234851	-0.63	2.55	0.00	0.01	-1
ENSG00000166002	-0.46	1.92	0.00	0.01	-1
ENSG00000109736	-0.46	4.66	0.00	0.01	-1
ENSG00000220937	-1.12	-1.25	0.00	0.01	-1
ENSG00000157510	-0.40	3.99	0.00	0.01	-1
ENSG00000164823	-0.34	5.43	0.00	0.01	-1
ENSG00000256005	-0.54	3.64	0.00	0.01	-1
ENSG00000264281	-0.71	0.75	0.00	0.01	-1
ENSG00000169297	-0.37	3.69	0.00	0.02	-1
ENSG00000229119	-0.68	1.01	0.00	0.02	-1
ENSG00000254741	-0.34	6.11	0.00	0.02	-1
ENSG00000234648	-0.85	-0.60	0.00	0.02	-1
ENSG00000198355	-0.38	5.16	0.00	0.02	-1
ENSG00000152104	-0.43	5.69	0.00	0.02	-1
ENSG00000236334	-0.77	-0.19	0.00	0.02	-1
ENSG00000250151	-0.30	6.08	0.00	0.02	-1
ENSG00000251357	-0.53	3.57	0.00	0.02	-1
ENSG00000198744	-0.47	4.27	0.00	0.02	-1
ENSG00000213553	-0.52	3.41	0.00	0.02	-1
ENSG00000249264	-0.60	3.69	0.00	0.02	-1
ENSG00000184990	-0.43	4.67	0.00	0.02	-1
ENSG00000225630	-0.57	4.38	0.00	0.02	-1
ENSG00000188910	-0.43	2.42	0.00	0.02	-1
ENSG00000138166	-0.68	5.84	0.00	0.02	-1
ENSG00000230979	-0.78	0.83	0.00	0.02	-1
ENSG00000100439	-0.37	4.65	0.00	0.02	-1
ENSG00000153443	-0.47	3.18	0.00	0.02	-1
ENSG00000105538	-0.32	4.88	0.00	0.02	-1
ENSG00000240567	-0.74	1.03	0.00	0.02	-1
ENSG00000103363	-0.36	6.62	0.00	0.02	-1
ENSG00000071655	-0.44	4.47	0.00	0.02	-1
ENSG00000120337	-1.05	2.64	0.00	0.02	-1
ENSG00000167671	-0.35	5.29	0.00	0.02	-1

Appendix 1 continued

ens-id	logFC	logCPM	P-value	FDR	DE
ENSG00000218426	-0.53	2.74	0.00	0.02	-1
ENSG00000137331	-0.34	7.34	0.00	0.02	-1
ENSG00000237433	-0.82	-0.67	0.00	0.02	-1
ENSG00000187653	-0.44	2.58	0.00	0.02	-1
ENSG00000182154	-0.39	3.68	0.00	0.02	-1
ENSG00000122490	-0.37	4.14	0.00	0.02	-1
ENSG00000137309	-0.31	10.40	0.00	0.02	-1
ENSG00000259936	-0.51	1.92	0.00	0.02	-1
ENSG00000118785	-0.29	8.51	0.00	0.02	-1
ENSG00000259952	-0.37	3.73	0.00	0.02	-1
ENSG00000233913	-0.67	1.49	0.00	0.02	-1
ENSG00000130733	-0.38	4.22	0.00	0.02	-1
ENSG00000069011	-0.38	3.71	0.00	0.02	-1
ENSG00000214199	-0.57	3.64	0.00	0.02	-1
ENSG00000104880	-0.34	5.21	0.00	0.02	-1
ENSG00000142541	-0.31	9.30	0.00	0.02	-1
ENSG00000185838	-0.44	2.60	0.00	0.02	-1
ENSG00000164379	-0.36	4.37	0.00	0.02	-1
ENSG00000143322	-0.30	6.67	0.00	0.02	-1
ENSG00000205609	-0.55	3.15	0.00	0.02	-1
ENSG00000215030	-0.47	2.59	0.00	0.02	-1
ENSG00000162032	-0.36	4.09	0.00	0.02	-1
ENSG00000152082	-0.50	4.40	0.00	0.02	-1
ENSG00000224864	-0.84	-0.01	0.00	0.02	-1
ENSG00000132768	-0.34	5.67	0.00	0.02	-1
ENSG00000167513	-0.36	4.54	0.00	0.02	-1
ENSG00000251196	-0.44	2.73	0.00	0.02	-1
ENSG00000077348	-0.37	4.60	0.00	0.02	-1
ENSG00000099998	-0.48	4.15	0.00	0.02	-1
ENSG00000245768	-0.86	-1.19	0.00	0.03	-1
ENSG00000241624	-0.72	-0.54	0.00	0.03	-1
ENSG00000253102	-0.50	2.56	0.00	0.03	-1
ENSG00000142546	-0.32	5.51	0.00	0.03	-1
ENSG00000232499	-0.50	0.94	0.00	0.03	-1
ENSG00000226084	-0.65	2.59	0.00	0.03	-1
ENSG00000078902	-0.39	3.40	0.00	0.03	-1
ENSG00000226015	-0.55	0.42	0.00	0.03	-1
ENSG00000233493	-1.14	-2.02	0.00	0.03	-1
ENSG00000219747	-0.68	-0.48	0.00	0.03	-1
ENSG00000230698	-0.72	-0.14	0.00	0.03	-1
ENSG00000262413	-0.42	4.75	0.00	0.03	-1
ENSG00000074416	-0.39	2.31	0.00	0.03	-1
ENSG00000165802	-0.32	6.31	0.00	0.03	-1
ENSG00000213370	-0.88	-1.45	0.00	0.03	-1
ENSG00000099624	-0.38	4.88	0.00	0.03	-1
ENSG00000149806	-0.28	7.53	0.00	0.03	-1
ENSG00000167701	-0.65	-0.46	0.00	0.03	-1
ENSG00000108107	-0.33	7.64	0.00	0.03	-1
ENSG00000255769	-1.07	-1.63	0.00	0.03	-1
ENSG00000213988	-0.79	0.09	0.00	0.03	-1
ENSG00000173156	-0.37	3.69	0.00	0.03	-1
ENSG00000197258	-0.48	2.13	0.00	0.03	-1
ENSG00000184779	-0.65	2.87	0.00	0.03	-1
ENSG00000251279	-0.50	1.70	0.00	0.03	-1
ENSG00000204308	-0.34	5.97	0.00	0.03	-1
ENSG00000183298	-0.38	3.27	0.00	0.03	-1
ENSG00000125657	-0.28	5.93	0.00	0.03	-1
ENSG00000103152	-0.37	4.53	0.00	0.03	-1
ENSG00000212123	-0.66	0.10	0.00	0.03	-1
ENSG00000178715	-0.82	-1.37	0.00	0.03	-1
ENSG00000103222	-0.31	6.99	0.00	0.03	-1

Appendix 1 continued

ens-id	logFC	logCPM	P-value	FDR	DE
ENSG00000233406	-0.71	1.20	0.00	0.03	-1
ENSG00000177494	-0.58	0.77	0.00	0.03	-1
ENSG00000116604	-0.33	5.13	0.00	0.03	-1
ENSG00000226221	-0.50	2.24	0.00	0.03	-1
ENSG00000250182	-0.51	3.89	0.00	0.03	-1
ENSG00000176946	-0.28	5.55	0.00	0.03	-1
ENSG00000137880	-0.35	3.64	0.00	0.03	-1
ENSG00000144136	-0.29	8.73	0.00	0.03	-1
ENSG00000169218	-0.50	1.34	0.00	0.03	-1
ENSG00000171223	-0.42	2.74	0.00	0.03	-1
ENSG00000148362	-0.34	3.81	0.00	0.03	-1
ENSG00000164236	-0.49	1.04	0.00	0.03	-1
ENSG00000062822	-0.32	5.07	0.00	0.03	-1
ENSG00000133315	-0.37	4.49	0.00	0.03	-1
ENSG00000104825	-0.41	4.12	0.00	0.04	-1
ENSG00000123064	-0.34	6.26	0.00	0.04	-1
ENSG00000186792	-0.33	4.05	0.00	0.04	-1
ENSG00000242735	-0.61	1.56	0.00	0.04	-1
ENSG00000225663	-0.44	3.11	0.00	0.04	-1
ENSG00000230580	-0.76	-0.31	0.00	0.04	-1
ENSG00000099804	-0.33	5.20	0.00	0.04	-1
ENSG00000165672	-0.30	8.33	0.00	0.04	-1
ENSG00000145287	-0.32	6.56	0.00	0.04	-1
ENSG00000142751	-0.38	3.64	0.00	0.04	-1
ENSG00000170265	-0.35	4.70	0.00	0.04	-1
ENSG00000198911	-0.27	8.35	0.00	0.04	-1
ENSG00000177700	-0.37	5.56	0.00	0.04	-1
ENSG00000250251	-0.40	3.68	0.00	0.04	-1
ENSG00000261582	-0.31	6.12	0.00	0.04	-1
ENSG00000169692	-0.36	3.59	0.00	0.04	-1
ENSG00000233328	-0.53	1.15	0.00	0.04	-1
ENSG00000204922	-0.35	3.56	0.00	0.04	-1
ENSG00000213236	-0.77	-0.94	0.00	0.04	-1
ENSG00000171222	-0.44	2.28	0.00	0.04	-1
ENSG00000233016	-0.32	4.89	0.00	0.04	-1
ENSG00000232856	-0.42	4.01	0.00	0.04	-1
ENSG00000198113	-0.32	4.34	0.00	0.04	-1
ENSG00000228218	-0.73	0.22	0.00	0.04	-1
ENSG00000241781	-0.93	-1.55	0.00	0.04	-1
ENSG00000255351	-0.70	-0.52	0.00	0.04	-1
ENSG00000100994	-0.35	8.19	0.00	0.04	-1
ENSG00000167394	-0.38	2.11	0.00	0.04	-1
ENSG00000251593	-0.85	-1.14	0.00	0.04	-1
ENSG00000183458	-0.43	3.39	0.00	0.04	-1
ENSG00000266066	-0.49	0.83	0.00	0.04	-1
ENSG00000237846	-0.59	1.28	0.00	0.04	-1
ENSG00000135535	-0.32	7.50	0.00	0.04	-1
ENSG00000103326	-0.38	4.94	0.00	0.04	-1
ENSG00000142634	-0.29	6.08	0.00	0.04	-1
ENSG00000105701	-0.36	6.94	0.00	0.04	-1
ENSG00000225461	-0.90	-0.48	0.00	0.04	-1
ENSG00000079313	-0.38	4.09	0.00	0.04	-1
ENSG00000110046	-0.32	4.71	0.00	0.04	-1
ENSG00000231240	-0.83	-1.48	0.00	0.04	-1
ENSG00000229638	-0.40	3.84	0.00	0.04	-1
ENSG00000130255	-0.30	8.11	0.00	0.04	-1
ENSG00000237550	-0.97	-0.83	0.00	0.04	-1
ENSG00000226608	-0.53	0.45	0.00	0.04	-1
ENSG00000115268	-0.39	7.48	0.00	0.04	-1
ENSG00000234397	-0.87	-1.42	0.00	0.04	-1
ENSG00000170909	-1.19	-0.08	0.00	0.05	-1

Appendix 1 continued

ens-id	logFC	logCPM	P-value	FDR	DE
ENSG00000167775	-0.30	5.88	0.00	0.05	-1
ENSG00000240489	-0.45	1.50	0.00	0.05	-1
ENSG00000136997	-0.27	8.86	0.00	0.05	-1
ENSG00000104529	-0.29	7.99	0.00	0.05	-1
ENSG00000159202	-0.27	6.64	0.00	0.05	-1
ENSG00000152492	-0.40	6.40	0.00	0.05	-1
ENSG00000234975	-0.71	3.75	0.00	0.05	-1
ENSG00000227097	-0.31	6.00	0.00	0.05	-1
ENSG00000142627	-0.33	7.78	0.00	0.05	-1
ENSG00000105671	-0.37	4.84	0.00	0.05	-1
ENSG00000089685	-0.28	6.83	0.00	0.05	-1
ENSG00000175793	-0.30	7.22	0.00	0.05	-1
ENSG00000255823	-0.47	0.90	0.00	0.05	-1
ENSG00000206228	-0.46	0.56	0.00	0.05	-1
ENSG00000183199	-0.42	2.91	0.00	0.05	-1
ENSG00000176101	-0.35	4.15	0.00	0.05	-1

Appendix 2. Differentially expressed proteins upon depletion of PARROT as determined by MS (L/H- light to heavy ratio; DE-differential expression where 1 indicates upregulation and -1 indicates downregulation).

Gene names	ctrl L/H	siRNA1 L/H	siRNA2 L/H	DE
APOB	3.34	4.45	5.14	1
THBS1	1.55	4.24	2.28	1
ITGA5	2.30	3.75	3.08	1
NT5E	2.41	3.66	4.36	1
MVP	1.74	2.28	2.49	1
ERAP1	1.11	2.03	2.09	1
HEXB	1.38	2.01	2.11	1
MVK	1.20	1.73	1.77	1
CTSZ	1.12	1.68	1.69	1
GNS	1.13	1.64	1.66	1
AKT1S1	0.82	1.63	1.38	1
IDH1	1.03	1.61	1.66	1
CAPZA2	1.09	1.51	1.38	1
CSPG4	0.88	1.45	1.35	1
MARS2	1.04	1.42	1.69	1
TBCB	1.09	1.41	1.51	1
RRBP1	0.72	1.40	1.47	1
BCL10	1.03	1.39	1.30	1
PRKAR1A	0.87	1.36	1.37	1
SH3BGRL3	0.85	1.33	1.50	1
OSTF1	1.00	1.28	1.27	1
NOP16	1.00	0.83	0.75	-1
MTX2	0.99	0.82	0.79	-1
RBM7	1.09	0.82	0.82	-1
RPRD2	1.14	0.81	0.67	-1
CTBP2	1.01	0.81	0.66	-1
KPNA3	0.99	0.81	0.73	-1
CD3EAP	1.00	0.80	0.82	-1

Appendix 2 *continued*

Gene names	ctrl L/H	siRNA1 L/H	siRNA2 L/H	DE
DCAF13	1.07	0.77	0.76	-1
ALDH16A1	0.95	0.77	0.73	-1
MSI2	0.98	0.77	0.71	-1
FDPS	1.01	0.74	0.82	-1
NOA1	0.94	0.74	0.75	-1
RNF20	0.96	0.73	0.78	-1
METTL7A	0.93	0.72	0.72	-1
PACSIN2	0.88	0.71	0.72	-1
SELI;EPT1	1.06	0.70	0.71	-1
FASN	1.07	0.70	0.81	-1
PCID2	0.91	0.69	0.20	-1
EFTUD2	0.85	0.64	0.70	-1
METAP1	1.15	0.61	0.74	-1
DIMT1	0.88	0.61	0.73	-1
RPRD1B	0.92	0.61	0.74	-1
GLO1	1.04	0.58	0.70	-1
MNF1	1.15	0.57	0.78	-1
NOB1	1.09	0.55	0.73	-1
MBLAC2	1.03	0.51	0.75	-1
PPP1R12C	0.77	0.44	0.38	-1
SELENBP1	0.50	0.38	0.38	-1

Appendix 3. Differentially phosphorylated proteins upon depletion of PARROT as determined by MS (L/H- light to heavy ratio; DE-differential expression where 1 indicates upregulation and -1 indicates downregulation).

Gene names	ctrl L/H	siRNA1 L/H	siRNA2 L/H	DE
PSRC1	1.74	6.55	3.17	1
NDRG1	2.88	6.43	7.10	1
ITGA5	4.06	6.33	5.86	1
MVP	2.38	4.13	6.88	1
HMGA2	1.77	2.77	3.00	1
CDK18	1.40	2.74	1.89	1
SIPA1L2	1.79	2.72	2.70	1
PI4KA	1.32	2.42	2.56	1
EPHA2	0.97	2.37	1.28	1
LGALS1	1.08	2.33	1.74	1
MPDZ	1.40	2.32	2.54	1
PPP1R18	1.24	2.30	1.72	1
TNIK	1.74	2.22	2.60	1
FLNB	0.97	2.19	2.02	1
ZNF608	1.21	2.16	1.78	1
STK10	1.40	2.14	1.95	1
DOS	1.07	2.12	1.82	1

Appendix 3 continued

Gene names	ctrl L/H	siRNA1 L/H	siRNA2 L/H	DE
BRSK2	0.99	2.02	2.61	1
MLF2	1.00	1.96	1.52	1
HECW2	0.93	1.94	1.52	1
CALD1	1.39	1.94	1.94	1
DNAJC2	1.10	1.92	1.37	1
TSC2	1.44	1.91	1.95	1
RFC1	0.72	1.91	1.27	1
ABCC2	1.31	1.90	2.05	1
ERC1	1.14	1.82	1.53	1
MBD1	1.01	1.82	1.32	1
SRRM2	1.31	1.81	1.98	1
SNTB2	1.40	1.78	1.92	1
PPP1R18	1.17	1.77	1.49	1
AAK1	1.14	1.77	1.55	1
SPRED2	1.39	1.76	1.79	1
GOLGA4	1.03	1.76	1.61	1
AAK1	1.26	1.75	1.63	1
ZNF532	1.26	1.74	1.60	1
AAK1	1.18	1.70	1.52	1
FSCN1	1.07	1.66	1.36	1
SLK	1.27	1.64	1.75	1
MAST2	1.01	1.64	1.28	1
CHD9	0.99	1.63	1.26	1
TJAP1	0.89	1.63	1.40	1
MKI67	1.09	1.63	1.55	1
CPD	1.09	1.62	1.68	1
MAST2	0.86	1.61	1.37	1
RAB11FIP5	1.11	1.60	1.72	1
EML4	1.28	1.60	1.72	1
MAP4K5	0.88	1.60	1.37	1
C2CD2	0.90	1.60	1.36	1
TNFAIP2	0.66	1.60	1.46	1
STMN1	1.03	1.59	1.55	1
CSPP1	0.98	1.59	1.34	1
ARHGEF12	1.07	1.58	1.46	1
MAP4	1.05	1.58	1.32	1
MKL2	0.81	1.54	1.28	1
STIM1	1.09	1.52	1.49	1
AHNAK	1.03	1.52	1.44	1
DISP1	1.08	1.51	1.58	1
IWS1	0.91	1.50	1.32	1
SH3BP4	0.96	1.49	1.86	1
PHLDB1	0.84	1.49	1.29	1
KDM3A	1.16	1.49	1.55	1

Appendix 3 *continued*

Gene names	ctrl L/H	siRNA1 L/H	siRNA2 L/H	DE
MKI67	1.08	1.49	1.60	1
ETV6	1.11	1.48	1.40	1
USP24	1.15	1.46	1.46	1
IQSEC1	1.00	1.45	1.31	1
CLCN2	0.88	1.44	1.82	1
NES	1.08	1.44	1.54	1
RRBP1	0.50	1.43	1.47	1
BCL7C	1.09	1.43	1.59	1
DOCK7	1.05	1.41	1.73	1
COPB2	0.78	1.39	1.30	1
PHLDB1	0.83	1.38	1.49	1
CHD8	1.09	1.38	1.40	1
DDX41	1.07	1.36	1.51	1
STIM1	1.08	1.36	1.41	1
HOMEZ	1.00	1.35	1.37	1
PHKB	1.01	1.35	1.51	1
PLEKHA4	0.70	1.35	1.27	1
SPAG9	0.94	1.35	1.66	1
UBXN2B	1.00	1.35	1.29	1
ZDHHC5	0.40	1.34	1.36	1
UBXN2B	1.00	1.32	1.46	1
BAG3	1.01	1.32	1.32	1
SP100	0.94	1.31	1.74	1
STX10	0.99	1.31	1.43	1
ASXL2	0.94	1.30	1.44	1
RRBP1	0.61	1.28	1.55	1
PRKAR1A	0.94	1.28	1.58	1
CLN3	0.98	1.27	1.34	1
UBA5	0.99	1.27	1.65	1
KIAA1671	0.94	1.27	1.35	1
STIM1	0.98	1.26	1.34	1
NES	0.76	1.26	1.31	1
EIF3B	1.04	0.83	0.80	-1
DYNC1LI1	1.01	0.82	0.73	-1
LIMCH1	1.04	0.82	0.81	-1
DCUN1D5	1.00	0.80	0.81	-1
LEMD3	0.98	0.80	0.68	-1
ERICH1	1.14	0.80	0.72	-1
XRCC6	1.10	0.80	0.72	-1
YTHDF1	1.02	0.80	0.73	-1
RPS2	1.06	0.79	0.79	-1
ABI2	0.96	0.79	0.69	-1
LMNA	1.00	0.79	0.63	-1
DUT	1.07	0.79	0.80	-1
NUP153	0.95	0.79	0.77	-1

Appendix 3 continued

Gene names	ctrl L/H	siRNA1 L/H	siRNA2 L/H	DE
MGME1	1.06	0.79	0.79	-1
HDAC4	1.01	0.78	0.64	-1
PHF14	4.92	0.78	0.81	-1
RIC8A	1.01	0.78	0.80	-1
RAPGEF3	2.67	0.78	0.71	-1
ZYX	0.96	0.78	0.60	-1
BOLA1	1.11	0.78	0.65	-1
NOP2	0.94	0.78	0.75	-1
MAP7D1	0.98	0.78	0.76	-1
HMGA1	0.95	0.78	0.70	-1
C19orf21	0.94	0.77	0.49	-1
TRIM28	1.01	0.77	0.78	-1
ACLY	1.02	0.77	0.74	-1
KRT18	1.19	0.77	0.57	-1
PDLIM1	0.99	0.77	0.83	-1
MEPCE	1.03	0.77	0.80	-1
NUPL1	1.19	0.76	0.70	-1
RPL29	1.17	0.76	0.75	-1
NAP1L4	0.98	0.76	0.69	-1
ZNF687	1.08	0.76	0.77	-1
TSNAX	0.95	0.76	0.79	-1
EIF2S2	0.98	0.76	0.78	-1
YAP1	0.92	0.76	0.37	-1
KLC1	1.20	0.76	0.81	-1
MAPK1	1.49	0.76	0.67	-1
LRRFIP1	1.30	0.75	0.40	-1
PTRF	1.06	0.75	0.72	-1
PLEKHG4	0.91	0.75	0.68	-1
TRIM28	1.07	0.75	0.78	-1
EPB41L2	1.11	0.75	0.75	-1
HN1L	1.03	0.75	0.82	-1
WDR70	0.92	0.75	0.74	-1
GTF3C2	0.91	0.74	0.75	-1
EIF2A	1.16	0.74	0.64	-1
TOR1AIP1	0.97	0.74	0.76	-1
YBX1	1.12	0.74	0.53	-1
PTRF	0.93	0.74	0.57	-1
NUP133	1.11	0.73	0.77	-1
CDK17	1.00	0.73	0.67	-1
CDK16	1.00	0.73	0.67	-1
VIM	1.20	0.73	0.62	-1
PPP1R14B	1.02	0.73	0.82	-1
SF3B1	1.21	0.73	0.83	-1
IKBKB	1.04	0.73	0.55	-1
NAP1L4	0.97	0.73	0.67	-1

Appendix 3 *continued*

Gene names	ctrl L/H	siRNA1 L/H	siRNA2 L/H	DE
RALGPS2	1.16	0.72	0.82	-1
HMGA1	1.00	0.72	0.81	-1
INTS5	0.93	0.72	0.75	-1
NUMA1	0.87	0.72	0.46	-1
BPTF	0.99	0.72	0.79	-1
TMCO1	1.05	0.72	0.76	-1
UTP14A	0.89	0.71	0.67	-1
ITPR3	1.16	0.71	0.35	-1
FLNB	0.92	0.71	0.70	-1
RPL4	1.42	0.71	0.73	-1
TACC3	0.88	0.71	0.60	-1
SCEL	1.33	0.71	0.58	-1
HNRNPL	0.95	0.71	0.65	-1
MTFR1	1.13	0.71	0.38	-1
UNG	1.07	0.71	0.68	-1
YEATS2	1.19	0.70	0.71	-1
EIF3E	1.02	0.70	0.77	-1
FAM120A	1.21	0.70	0.67	-1
EI24	0.95	0.70	0.77	-1
MKI67IP	0.97	0.70	0.34	-1
BCAR1	1.71	0.70	0.79	-1
NUP107	0.89	0.69	0.61	-1
NUP107	0.89	0.69	0.61	-1
NUP107	0.89	0.69	0.61	-1
MKI67	1.09	0.69	0.67	-1
NPM1	0.92	0.69	0.75	-1
CDK1	0.92	0.69	0.71	-1
NPM1	0.91	0.69	0.65	-1
TRIM28	1.13	0.68	0.69	-1
CALD1	1.18	0.68	0.71	-1
HNRNPAB	1.18	0.68	0.60	-1
NCAPH	0.82	0.67	0.59	-1
CAAP1	2.82	0.67	0.73	-1
HNRNPU	1.28	0.67	0.80	-1
ZKSCAN1	0.82	0.67	0.66	-1
LASP1	0.89	0.67	0.64	-1
SLC19A1	0.95	0.67	0.78	-1
ABL2	1.15	0.67	0.74	-1
WDR43	1.10	0.67	0.83	-1
TMX1	1.09	0.66	0.48	-1
BIN1	0.86	0.66	0.59	-1
AMD1	0.97	0.66	0.81	-1
TNS3	0.81	0.66	0.54	-1
ACACA	0.97	0.65	0.81	-1
HOXA3	0.87	0.65	0.64	-1

Appendix 3 continued

Gene names	ctrl L/H	siRNA1 L/H	siRNA2 L/H	DE
RPL3	1.18	0.65	0.80	-1
USP8	0.82	0.65	0.65	-1
CALD1	0.86	0.65	0.70	-1
WIBG	0.90	0.64	0.69	-1
WIBG	0.90	0.64	0.69	-1
NPM1	0.95	0.64	0.72	-1
FOXK1	0.85	0.64	0.57	-1
GTF2F1	0.82	0.63	0.62	-1
GTF2F1	0.82	0.63	0.62	-1
GTF2F1	0.82	0.63	0.62	-1
RSL1D1	1.26	0.63	0.47	-1
RSL1D1	0.76	0.63	0.63	-1
MYBBP1A	0.93	0.62	0.65	-1
DUT	0.76	0.62	0.57	-1
CBX5	0.98	0.62	0.77	-1
ETV3	0.79	0.62	0.41	-1
PRKAR2A	0.95	0.61	0.33	-1
PRKAR2A	0.95	0.61	0.33	-1
EEF1E1	0.90	0.61	0.69	-1
GTPBP4	0.81	0.61	0.64	-1
GTF2F1	1.13	0.60	0.65	-1
TSC22D4	0.87	0.60	0.50	-1
MEF2D	0.84	0.60	0.68	-1
FLNB	1.35	0.60	0.82	-1
CDK1	0.78	0.60	0.51	-1
ILF3	0.84	0.60	0.62	-1
EI24	0.94	0.60	0.74	-1
NCAPD2	0.85	0.59	0.43	-1
NCL	1.30	0.59	0.64	-1
CD3EAP	0.83	0.58	0.62	-1
PAICS	0.95	0.58	0.69	-1
ACACA	0.95	0.58	0.64	-1
CCDC86	0.94	0.58	0.63	-1
HOXA5	0.71	0.57	0.43	-1
HNRNPA3	1.03	0.57	0.57	-1
NCAPH	0.72	0.57	0.58	-1
OTUD4	0.88	0.56	0.32	-1
NCOA7	0.81	0.56	0.63	-1
DARS2	1.35	0.56	0.57	-1
DYNC1LI2	0.87	0.56	0.66	-1
RIF1	0.82	0.56	0.67	-1
LASP1	0.73	0.56	0.54	-1
NOP2	1.09	0.55	0.64	-1
PRRC2C	0.95	0.55	0.55	-1
BAIAP2	0.66	0.55	0.32	-1

Appendix 3 continued

Gene names	ctrl L/H	siRNA1 L/H	siRNA2 L/H	DE
STAT5B	0.89	0.55	0.65	-1
DPYSL3	0.80	0.54	0.46	-1
CHAF1A	0.74	0.53	0.53	-1
CHAF1A	0.74	0.53	0.53	-1
PRPF31	0.80	0.50	0.39	-1
XRCC1	1.01	0.50	0.82	-1
JUP	0.67	0.49	0.53	-1
KRT18	1.19	0.48	0.83	-1
HAUS6	1.38	0.48	0.66	-1
SLC35F2	1.15	0.47	0.81	-1
PLEKHA1	1.15	0.47	0.67	-1
ANXA2	0.90	0.47	0.62	-1
FTSJ3	0.79	0.46	0.52	-1
RALGPS2	1.26	0.45	0.81	-1
PPP1R12A	0.91	0.42	0.43	-1
SVIL	0.55	0.41	0.32	-1
EPB41L1	0.49	0.41	0.40	-1
NOVA1	0.57	0.34	0.39	-1
EPB41L4A	0.48	0.28	0.28	-1
BAZ2B	0.49	0.28	0.40	-1
UBASH3B	0.30	0.24	0.25	-1

ABBREVIATIONS

1D-eRNAs	Unidirectional enhancer RNAs
2D-eRNAs	Bidirectional enhancer RNAs
A549	Human lung carcinoma cell line
BACE1	β-secretase-1 protein
BDNF	Brain derived neurotrophic factor
bp	Base pare
CCAT1	Colon cancer associated transcript
cDNA	Complementary DNA
ChIA-PET	Chromatin interaction analysis with paired end tag sequencing
Chip-seq	Chromatin immunoprecipitation followed by deep sequencing
CRISPR	Clustered regularly interspaced short palindromic repeats
Ct	Threshold cycle
CTCF	CCCTC-binding factor
dATP	Deoxyadenosine triphosphate
dCTP	Deoxycytidine triphosphate
DEPC	Diethyldicarbonat
dGTP	Deoxyguanosine triphosphate
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
dsiRNA	Dicer-substrate RNA
DTT	1,4-Dithiothreitol
dTTP	Deoxythymidine triphosphate
EDTA	Ethylenediaminetetraacetic acid
ENCODE	Encyclopedia of DNA Elements
EZH2	Enhancer of zeste homolog 2
FBS	Fetal bovine serum
FDR	False discovery rate
GFP	Green fluorescent protein
GHET1	Gastric carcinoma high expressed transcript 1
GM12878	Lymphoblastoid cell line (LCL)
GO	Gene ontology
H1ES	Human embryonic stem cells
H3K27ac	Histone 3 lysine 27 acetylation
H3K27me3	Trimethylation of lysine 27 of histone H3
H3K36me3	Trimethylation of lysine 36 of histone H3
H3K4me1, H3K4me2, H3K4me3	Mono/di/trimethylation of lysine 4 of histone H3
H3K9me3	Trimethylation of histone H3 lysine 9
HEK293	Human embryonic kidney
HeLa	Cervical cancer cell line
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HepG2	Liver carcinoma
HMEC	Human mammary epithelial cells
HSMM	Human skeletal muscle myoblasts cell line
HUVEC	Human umbilical vein enodothelial cell line
IGF2BP1	Insulin-like growth factor 2 mRNA binding protein 1
IPA	Ingenuity Pathway Analysis
K562	Erythrocytic leukemia cell line
kb	Kilobase pair
lincRNAs	Intergenic long ncRNAs
MAX	MYC- associated protein X
MCF7	Breast cancer cell line
MLL	Mixed lineage leukemia complex

NaOH	Sodium hydroxide
ncRNAs	Non-coding RNAs
NHEK	Normal epidermal keratinocytes
NHLF	Normal human lung fibroblasts
p300	E1A binding protein p300
PARROT	Proliferation Associated RNA and Regulator Of Translation
PBS	Phosphate-buffered saline
PCGs	Protein coding genes
PCR	Polymerase chain reaction
PI	Propidium iodide
Pol II	RNA polymerase II
PRC1	Polycomb repressive complex 1
PRC2	Polycomb repressive complex 2
PreSTIGE	Predicting Specific Tissue Interactions of Genes and Enhancers
PS	Phosphatidylserine
PVDF	Polyvinylidene difluoride
qPCR	Quantitative real-time PCR
RNA	Ribonucleic acid
RNA-a	Activating RNA
RNA-seq	RNA-sequencing
RPKM	Reads per kilobase per million read mapped
SDS	Sodium dodecylsulfate
SDS-PAGE	SDS-polyacrylamide gel electrophoresis
sgRNA	Short guide RNA
SILAC	Stable isotope labeling by amino acids in cell culture
TFs	Transcription factors
TPA	12-O-tetradecanoylphorbol-13-acetate
TSS	Transcriptional start site
Uchl1	Ubiquitin carboxy-terminal hydrolase L1

PUBLICATIONS

2016

Vucicevic D, Gehre M, Dhamija S, Friis-Hansen L, Meierhofer D, Sauer S, Ørom UA. The long non-coding RNA PARROT is an upstream regulator of MYC and affects proliferation and translation. Oncotarget 2016; 7(23):33934-47.

2015

Vucicevic D, Corradin O, Ntini E, Scacheri PC, Orom UA. Long ncRNA expression associates with tissue-specific enhancers. Cell Cycle 2015; 14:253-60.

2014

Vucicevic D, Schrewe H, Orom UA. Molecular mechanisms of long ncRNAs in neurological disorders. Front Genet 2014; 5:48.

2013

Marsico A, Huska MR, Lasserre J, Hu H, **Vucicevic D,** Musahl A, Orom U, Vingron M. PROmiRNA: a new miRNA promoter recognition method uncovers the complex regulation of intronic miRNAs. Genome Biol 2013; 14:R84.