The role of non-coding genome in cancer

A thesis submitted to The University of Manchester for the degree of
Doctor of Philosophy
in the Faculty of Medical and Human Sciences

2016

Danish Memon
Faculty of Medical and Human Sciences
School of Medicine

Contents

Total Words = 34,152

List of Figures

Supplementary Figures (Appendix)

List of Tables

Supplementary Tables (Electronic Submission)

Abbreviations

A3SS - Alternative 3′ Splice Sites
A5SS - Alternative 5′ Splice Sites
ADM - Adrenomedullin
ANGPTL4 - Angiopoietin-like 4
ARNT - Arylhydrocarbon Receptor Nuclear Translocator
BH - Benjamini & Hochberg
BLCA - Bladder Urothelial Carcinoma
bp - basepair
CAGE - Cap Analysis of Gene Expression
CDS - Coding Sequence
ChIP – Chromatin Immunoprecipitation
COAD - Colon Adenocarcinoma
CoNCo - Coding to Non-Coding major-isoform
CV - Coefficient of Variation
DABG - Detection Above Background
EBS - Egr1 Binding Site
Egr1 - Early Growth Response 1
ERK - Extracellular signal-Regulated Kinase
ES - Enrichment Score
EST - Expressed Sequence Tag
FDR - False Discovery Rate
FPKM - Fragments Per Kilobase of transcript per Million of mapped fragments
GEO - Gene Expression Ombinus
GO - Gene Ontology
GSEA - Gene Set Enrichment Analysis
HIF - hypoxia-Inducible Factor
HINCR - Hypoxia Induced Non-Coding RNA
HK-lincRNA - Housekeeping lincRNA
HNSC - Head and Neck Squamous Cell Carcinoma
HRE - Hypoxia Response Elements
HRM - Hypoxia Regulated MiRNAs
KICH - Kidney Chromophobe
KIRC - Kidney Renal Clear Cell Carcinoma
KIRP - Kidney Renal Papillary Cell Carcinoma
LIHC - Liver Hepatocellular Carcinoma
LincRNA - Long Intergenic Non-Coding RNA
LncRNA - Long Non-Coding RNA
LUAD - Lung Adenocarcinoma
LUSC - Lung Squamous Cell Carcinoma
MATS - Multivariate Analysis of Transcript Splicing
MFE - Minimum Free Energy
miRNA - microRNA
MM - Mismatch
MMR - Mismatch Repair
MPSS - Massively Parallel Signature Sequencing
mRNA - messenger RNA
MXE - Mutually Exclusive Exons

ncRNA – non-coding RNA
NDRG1 - N-myc Downstream Regulated 1
NHEJ - Non Homologous End Joining
ODD - Oxygen-Dependent Degradation Domain
ORA - Over Representation Analysis
ORF - Open Reading Frame
piRNA - piwi-associated RNA
PM - Perfect Match
PRAD - Prostate Adenocarcinoma
pVHL - von Hippel-Lindau protein
RI - Retained Intron
RIPChIP - RNAimmunoprecipitation followed by ChIP
RNA - Ribo Nucleic Acid
RNP - Ribonucleoprotein
rRNA - Ribosomal RNA
SAGE - Serial Analysis of Gene Expression
SE - Exon Skipping
siRNA - small interfering RNAs
smRNA - small molecular RNA
SNP - Single Nucleotide Polymorphism
snRNA - small nuclear RNAs
snRNPs - small nuclear Ribonuclear Proteins
SNV - Single Nucleotide Variants
SRE - Serum Response Elements
T - Thymine
TCGA - The Cancer Genome Atlas
THCA - Thyroid carcinoma
tRNA - transfer RNA
TS-lincRNA - Tissue-Specific lincRNA
U - Uracil
UCEC - Uterine Corpus Endometrial Carcinoma
UEHG - Ubiquitously Expressed Human Genes
UTR - Untranslated Region

Abstract

The University of Manchester
Danish Memon
Degree of Doctor of Philosophy

The role of non-coding genome in cancer.

January 2016.

Tumour hypoxia is associated with poor patient outcome and resistance to therapy. It impacts upon multiple pathways and causes alterations in the levels of protein encoding transcripts throughout the cell. Next generation sequencing of human datasets have revealed widespread alternative splicing in coding genes and the presence of large numbers of non-coding loci, raising the question as to whether these additional transcripts add additional functional complexity to the genome. The goal of my PhD was to use bioinformatics approaches to identify novel transcriptional events associated with hypoxia, with a particular focus on the role of non-coding RNA expression in regulating the cell's response to changes in oxygenation. In this thesis I describe three studies. The first is a systematic analysis of changes in long non-coding RNA expression in normal (BodyMap RNA-Seq data) and cancer tissue (using data from The Cancer Genome Atlas; TCGA). This revealed a set of long non-coding RNAs that are predicted to participate in housekeeping functions, and are likely to be essential for cell survival. The second is a global transcriptomic analysis of HCT116 colorectal cancer cells following a shift to 1% oxygenation. Analysis of RNA abundance over a hypoxic timecourse identified substantial remodelling of splicing, widespread alterations in the domain structure of many critical protein-coding genes and a global shift towards non-coding isoforms. This transition from coding to non-coding isoforms was recapitulated in a large and independent cohort of colorectal samples taken from TCGA and correlated with patient tumour status at last contact. The third study focused on non-coding RNA expression in the same HCT116 dataset. Expression of one of these loci, HINCR1 (Hypoxia Induced Non Coding RNA 1), was found to be induced in many models of hypoxia, and prognostic of survival in lung cancer patients. Subsequent experimental characterization of HINCR1 revealed it to be a regulator of Egr1 activity, a transcription factor central to the gene expression programs of mitogenesis.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright statement

Acknowledgements

I would like to say a massive thank you to my supervisory team comprising of Dr. Crispin Miller, Dr. Georges Lacaud and Dr. John Brognard who kept me on track during my PhD. I am at loss for words to describe how wonderful and inspirational Crispin has been as a supervisor. There have been phases during my PhD, which have been very challenging and stressful because of my personal circumstances and I will always be grateful to Crispin for his strong support during this period. Thank you Crispin for being a great mentor and for giving me the opportunity to work in your lab.

I would also like to extend my sincere gratitude towards Ms. Keren Dawson who has made major contributions to this thesis and without whom much of the work would have remained as hypothesis. I would also like to thank Dr. Jing Bi for many useful conversations about the projects and for carrying out validation work for one of the projects. The initial stages of the PhD were very challenging due to my inexperience in NGS analysis and I am particularly thankful to Dr. Hui Sun Leong for her patience and extremely useful bioinformatics training at that point. I would like to thank all the members (former and present) of the RNA Biology group for a wonderful time in the lab. A substantial amount of the work in the thesis involved collaborative work utilising the knowledge and resources of many labs including CEP Group (Prof. Caroline Dive, Dr. Ged Brady and Dr. Chris Morrow), Signalling Networks in Cancer Group (Dr. John Brognard, Dr. Shameem Fawdar) and Scientific Computing (Dr. Wei Xing, Dr. Chris Smowton). Much of the bioinformatics work relied upon the use of the linux cluster (Troodon) and I would like to thank Dr. Wei Xing, Mr. ZhiCheng Wang and Mr. Chris Wirth for keeping the system running. A lot of sequencing was performed as part of the projects and I would like to thank members of the molecular biology core facility specially Yvonne Hey and Gillian Newton for their efforts. I would also like to thank the graduate administrator of CRUKMI, Mrs. Julie Edwards, who has been very helpful and supportive throughout the course of my PhD.

This PhD would not have been possible without the support of my family and friends. Life outside the lab was fun due to such wonderful friends including Ajaz, Avinash, Priti, Sree, Nitin, Shubhi, Maria, Hui Sun, Jing, Haoran, Kiran, Thanuja, Sharmin, Bela, Krishna, Eiraj and Mohammad. Special thanks to my batch mate Maria who has been great company throughout the course of PhD and a close friend for lifetime. I was lucky to have such as friendly and cooperative flatmate in Priti who again has become a very close friend. I am extremely grateful to have such a wonderful family both in UK and in India. My Aunt and uncle, Salma Khalajaan and Ismail Uncle, and cousins, Sadia Baji and Abul Bhai, have been great support through the years. Particularly, Khalajaan who treated me like her son and never made me miss home. A big thanks to all my uncles, Nisar Mama, Aslam Mama, Asif Mama and Irfan Mama who have been a great support system. Particularly, Asif Mama who has been a great source of inspiration and positive influence in my life. I would also like to thank my brothers, Yaser Bhai and Ali, sister, Ambarin Baji and brother-in-law, Ather Bhai, for always being there for me. Life has changed considerably for everyone since my dad, Arif, suffered from cancer. Thank you dad for your simplicity, niceness and for always being there. I hope to be there for you while you fight in this tough moment. Thesis writing phase has been one of the most stressful periods of my life and I would like to thank Zainab, for her patience and support.

Finally, none of this could have happened without the unconditional love, sheer determination and constant prayers of my mom, Naseem, who has fought hard her whole life to give me this opportunity.

About The Author

Danish grew up in a beautiful garrison town called Devlali in India. He did his schooling from Barnes School. The school with its scenic location and excellent curriculum was a highly influential phase in his education during these formative years. During the 10th grade in school, he was taught BASIC programming language as part of computer science coursework, which helped in the development of logical thinking which is fundamental to programming skills.

After a move to Mumbai to pursue further education, Danish opted to study Biotechnology from University of Mumbai, India. During his graduation, he developed a lot of interest in Molecular Biology taught by a brilliant and extremely enthusiastic scientist Dr. Haresh Kamdar who nutured scientific curiosity and inspired him to pursue a career in science. During the final year of bachelor's degree, Danish was exposed to the area of Bioinformatics, which brought together both his interests in molecular biology and informatics. He went on to pursue a master's degree in Bioinformatics from University of Pune, India. He then worked as a junior research fellow in Prof. Pramod Wangikar's lab in IIT Bombay, India to perform comparative genomics analyses on the evolution and regulation of gene expression in photosynthetic prokaryotes.

Subsequently, he joined Dr. Crispin Miller's group on a Cancer Research UK (CRUK) funded PhD fellowship at the CRUK Manchester Institute to investigate the role of non-coding RNAs in hypoxia and their relevance in cancer.

Thesis Format

This thesis has been written in alternative format. The thesis starts with a general introduction and is followed by three results chapters. Each of the results chapter is essentially a paper, which describes a self-contained piece of work; together they consider different aspects of the non-coding transcriptome and their relevance to cancer. The contributions of all the authors have been acknowledged at the end of the paper.

The theme of all papers is to use bioinformatics approaches to propose exciting hypothesis on the role of non-coding transcriptome and then perform wet lab experiments to test these hypothesis. Each of these potential papers is an outcome of collaborative effort between bioinformaticians and bench scientists. My efforts have been on the bioinformatics side of the project.

I believe that alternative format is more appropriate to report the work due to the collaborative nature of the project and the potential outcome of three first author papers all based on the core topic of the role of the non-coding transcriptome.

I dedicate this thesis to my parents and my grandmother, Zubeda.

Chapter 1. Introduction

1.1 Aims of the Thesis

The broad objective of this work is to analyse transcriptome data in order to identify novel cancer-associated non-coding RNAs (ncRNAs). Solid tumours tend to be highly hypoxic, leading to many physiological adaptations and therapeutic challenges. In this thesis, the specific goal is to gain a better understanding of transcriptomic changes in response to hypoxia and the role of non-coding RNAs in this process.

1) To assess the overall expression of ncRNAs in normal and cancer tissue.
2) To assess the transcriptomic changes in response to hypoxia with particular focus on alternative splicing and non-coding transcripts.
3) Functional characterization of ncRNAs induced in response to hypoxia.

1.2 An Introduction to RNA

The RNA World Hypothesis proposes that molecules with the ability to self-replicate and catalyse biochemical reactions are likely to be initial precursors of life[1,2]. RNA molecules are self-replicating biomolecules with the ability both to store information and to catalyse biochemical reactions[3,4]. Therefore simpler forms of RNA or "RNA-like polymers" have been hypothesized to be the first biopolymers on Earth[3,4]. This hypothesis underlines the evolutionary and functional importance of RNA for cellular activity.

1.2.1 Transcription In Eukaryotes

The process of transcription involves synthesis of RNA from a DNA molecule in a step-wise manner[3,4]. One DNA strand acts as a template while the other strand acts as a guide. The RNA molecule formed is the reverse complement of the template strand and identical to the guide strand except for the replacement of thymine (T) nucleotides with uracil (U). The overall process of transcription is still an area of intense study, and extremely complex; a brief overview is provided here. Transcription can be sub-divided into 3 majors steps: Initiation, Elongation and Termination. Initiation relies upon RNA polymerase protein acting in combination with multiple protein factors to recognize and bind to promoter regions in the DNA. The initiation complex is able to melt a short 14 base pair (bp) stretch of DNA close to the transcription start site. Once this transcription

bubble is created, the first phosphodiester bond is formed between the ribonucleotides to form a dinucleotide. Initiation is followed by elongation, in which the RNA polymerase moves along the template DNA, simultaneously opening up the double-stranded DNA and extending the emerging ribonucleotide sequence through polymerization. Elongation continues until termination signals are detected downstream of the coding sequence. At this point the synthesized RNA molecule is released and the RNA polymerase dissociates from the DNA.

1.2.2 Steps in RNA Processing

In eukaryotes, an RNA molecules undergoes a number of processing steps both co- and post-transcriptionally. Modifications are made to the 5′ and 3′ ends of the RNA to increase its stability[4]. As soon as the nascent RNA is synthesized, its 5′ end is capped to protect it from enzymatic degradation. At the 3′ end, an endonucleolytic cleavage step is followed by addition of adenylic residues catalyzed by poly(A) polymerases [4]. The poly(A) tail thus formed increases the stability of the RNA molecule and protects it from degradation. The majority of eukaryotic genes contain intronic regions that are spliced out from the primary transcript before the final RNA molecule is produced. Intron start and ends are marked by 5′ and 3′ splice sites[4]. The splicing machinery, which comprises small nuclear ribonuclear proteins (snRNPs) that are able to recognize these splice sites along with a branch site close to the 3′ splice site, mediates two transesterification reactions to remove intron sequence and splice together the adjacent exons[3,5].

1.2.3 RNA – The product of transcription

The final product of successful transcription is a functional RNA molecule. There are many types of RNAs that are produced, depending on the gene undergoing transcription. Protein-coding genes produce messenger RNAs (mRNA) that can be translated to produce a protein. Functional mRNAs have a tripartite structure comprising the 5′ untranslated region (UTR), the protein-coding sequence (CDS) and the 3′ UTR. Many other types of RNA do not encode proteins. These ncRNAs can be sub-divided into structural and regulatory RNA molecules. Structural RNAs (ribosomal RNAs, rRNAs, and transfer RNAs, tRNAs) form the bulk of RNA in the cell and their role in protein synthesis is well established. Recently, other regulatory RNAs have been identified, and their potential to contribute to regulatory pathways within the cell have led them to be the subject of considerable research interest. These regulatory

RNAs can be further sub-divided into three sub-groups based on the length of the mature RNA. ncRNAs with very small RNA molecules (18-25 nt) include microRNAs (miRNAs) that regulate transcript stability and translation through interactions with mRNA molecules, transcription initiation RNAs (tiRNAs) that are expressed close to transcription start site but have unknown function, and small interfering RNAs (siRNAs) that are involved in RNA silencing. The next sub-group also comprises small RNAs (30-300 nt) and includes small nuclear RNAs (snoRNAs), small molecular RNA (smRNAs) and piRNA (piwi-associated RNA). The final sub-group comprises all long, regulatory ncRNAs (200-10,000 nt) expressed in the cell[6,7]. It is this subset of transcripts that forms the basis of the work described in this thesis; their known roles and function are introduced in greater depth below.

1.2.4 Long non-coding RNAs

Several approaches have been developed for the further classification of long non-coding RNAs (lncRNAs) such as those based on genomic location, length of the transcript, annotated protein-coding genes, DNA elements of known function, resemblance to protein-coding RNAs, association with repeats, association with the expression patterns of biochemical pathway components, and sequence and structure conservation[8]. The simplest classification is based on genomic location relative to protein-coding genes. From this perspective, lncRNAs can into five groups: (1) sense overlapping, (2) antisense overlapping, (3) bidirectional, (4) intronic or (5) intergenic. Due to their close association with protein-coding genes, all groups except the 'intergenic' group have been predominantly implicated in cis-regulatory functions affecting neighbouring or overlapping protein-coding genes[8]. In contrast, long intergenic ncRNA (lincRNAs) have been shown to have both cis- and trans- regulatory roles. More than 10,000 mammalian lincRNAs have been reported, and the list continues to grow as more cells, tissue types, and physiological conditions are studied.

An ideal lncRNA classification would include detailed mechanism of action. However, only a subset of lncRNAs have been studied at this level and the function of the vast majority of transcripts is unknown. Where lncRNAs have been characterised this has revealed a diversity of mechanistic roles that implicate lncRNAs at all levels of the cell, where functional interactions with DNA, with other RNAs, and with proteins have all been shown to be important. Many lncRNAs are under the control of canonical transcription factors such as p53 and Sox2[9], and have, conversely, been shown to

regulate protein-coding gene expression both in cis and in trans[10-12]. Mechanism of action can be further segregated through the molecules within which the lncRNA interacts.

1.2.5 Mechanism of Action of lncRNAs

1.2.5.1 RNA-DNA Interactions: Regulation of gene expression

One of the most consistent observations for many nuclear lncRNAs are interactions with chromatin-modifying complexes[3,13], where they are thought to act as a guide to bring the epigenetic factors to genomic sites (Fig 1.1). Since chromatin-modifying complexes lack the ability to bind to the genome, lncRNAs may be the missing link between these proteins and their target sites[13]. Absence of the lncRNA has been shown to disrupt the activity of chromatin-modifying complexes[3]. A number of epigenetic processes are regulated by lncRNAs including X-chromosome inactivation, and determination of cellular differentiation[14]. The role of lncRNA Xist in X chromosome inactivation through recruitment of PRC2 complex is well established. Other lncRNA including Air and Kcnq10t1 have also been shown to interact with PRC2 complex to influence epigenetic regulation[14]. Interestingly, 40% of known lncRNAs were found to be associated with chromatin-remodelling complexes from RNA immunoprecipitation followed by chip (RIPchip) of these proteins[14,15].

Some nuclear lncRNAs may also influence the binding of transcription factors to their target site and therefore have an impact on the downstream transcriptional program[3]. The lncRNA RMST is an example of lncRNA involved in transcription activation and has been shown by Ng *et al.* to be critical for neuronal differentiation[1]. In the absence of RMST, the binding of the SOX2 transcription factor to a subset of its target sites is disrupted and thus prevents activation of the target genes[1,4]. Thus RMST is likely to act as a guide for the SOX2 transcription factor binding on the DNA; the interaction between SOX2 and RMST is possibly mediated by RNA binding protein hnRNPA2/B1[1,4].

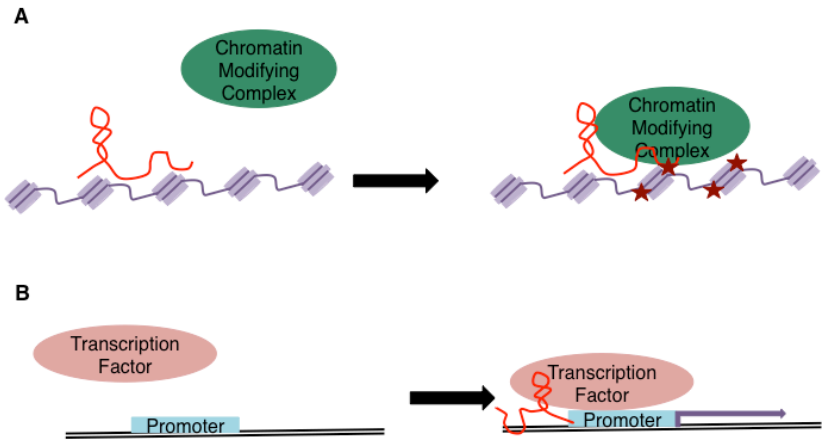

Figure 1.1 A schematic representation of mechanism of action of nuclear lncRNAs in (A) epigenetic modifications and (B) transcriptional regulation. This figure has been adapted from Moran *et al.*, 2012[3].

1.2.5.2 RNA-Protein Interactions: Molecular Bridges and Scaffolds

Certain lncRNAs tend to exist as part of ribonucleoprotein (RNP) complexes in cells and form structural links between proteins within the complex[3,5] (Fig 1.2). For instance HOTAIR has been reported to facilitate interaction between PRC2 and LSD1[3,6,7]. Similarly, XIST also provides the support linking the transcription factor YY1 with the PRC2 complex[3,8]. In some cases lncRNAs can also act as molecular scaffolds to initiate formation of nuclear compartments such as paraspeckles which are sub-nuclear bodies with a potential role in regulation of gene expression[16]. NEAT1 and NEAT2/MALAT1, two of the most highly expressed lncRNAs in the cell, and have been shown to be involved in paraspeckle formation[3].

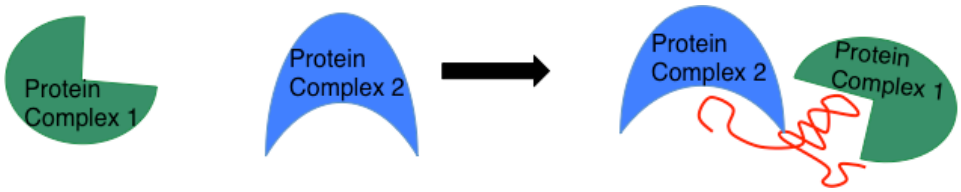

Figure 1.2 A schematic representation of mechanism of action of lncRNA acting as a scaffold facilitating protein-protein interactions. This figure has been adapted from Moran *et al.*, 2012[3].

1.2.5.3 RNA-RNA interactions: Post-transcriptional Regulation

In a landmark study Poliseno *et al.*, 2010 showed that a pseudogene can compete with its protein-coding paralogue to titrate away miRNAs if they share common miRNA binding sites[17]. In the absence of PTENP1, the protein-coding transcript of PTEN will be more effectively inactivated by miRNAs miR-19b and miR-20a. However, in the presence of PTENP1, the effect of the miRNA is abrogated. This regulatory relationship between PTEN and PTENP1 is dependent on common miRNA binding sites and therefore applicable to other coding/non-coding transcript pairs as long as they share high sequence similarity within the regulatory region i.e. common miRNA binding sites (Fig 1.3A). This is supported by a recent study, which has shown that the lncRNA linc-MD1 acts as sponge to abrogate the effect of miRNA miR-133, which would normally target the transcription factors MAML1 and MEF2C involved in muscle differentiation[18].

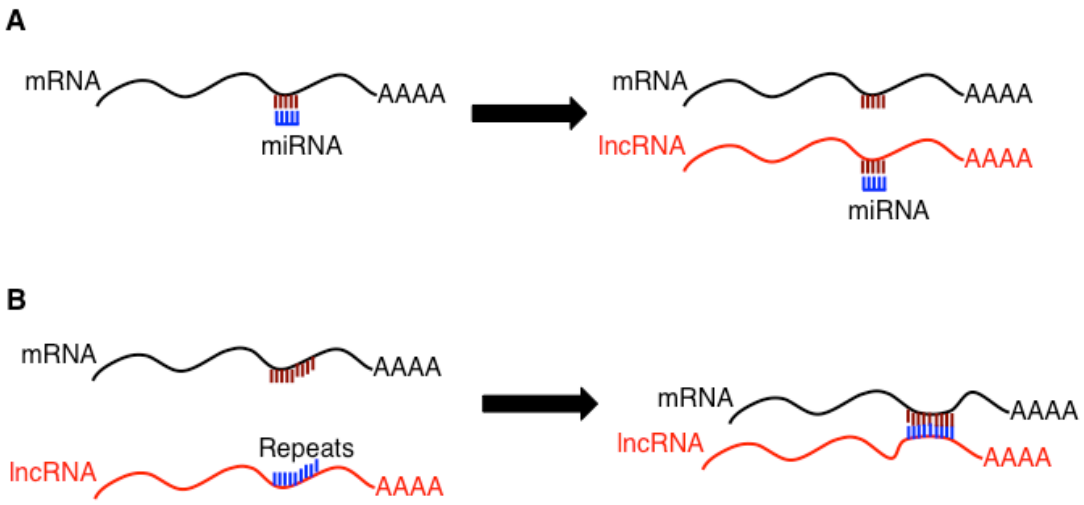

Figure 1.3 A schematic representation of mechanism of action of lncRNA in post-transcriptional regulation through direct interaction with other RNA molecules. (A) LncRNA can act as a 'sponge' to prevent miRNA-dependent down-regulation of mRNA. (B) LncRNA can directly bind to mRNA with complementary repeat elements. This figure has been adapted from Moran *et al.,* 2012[3].

LncRNAs can also exert their function via direct interaction with mRNA. Like microRNAs, they may affect the abundance of coding transcripts through direct interaction, leading to destabilization and/or degradation of the mRNA (Fig 1.3B). Gong and Maquat identified a lncRNA, lncRNA_AF087999, which downregulates the expression of SERPINE1 mRNA through direct interaction between the lncRNA and the mRNA[19]. It was suggested that the direct interaction between the lncRNA and the

target mRNA is facilitated by imperfect yet stable base pairing of repeat-like regions within the lncRNA to repeats with the 3′ UTR of SERPINE1. The double stranded, stem-like RNA structure thus formed can be recognized by Stau1, leading to the degradation of the mRNA. A similar mechanism was reported involving SINE repeats in a rodent lncRNA to influence adipogenesis[20], indicating that RNA-RNA interactions through repeat-like regions may be a general mechanism of lncRNA-based regulation.

1.3 Global RNA Abundance Measurement Techniques

The initial draft sequencing of the human genome took more than a decade with an estimated cost of $3 billion[21]. Since then, the introduction of next generation sequencing has led costs to fall to a point where a human genome can be sequenced for less than $2000[22]. Advances in sequencing technologies have also allowed more accurate identification and quantification of all the transcripts expressed in a cell: the 'transcriptome'. Before genome wide transcriptome sequencing, the major technologies for global gene expression analysis were based either on hybridization-based approaches or low-throughput sequence-based approaches[23].

Hybridization-based approaches estimate gene expression from the strength of signal produced by the incubation of fluorescently labelled cDNA amplified from the initial RNA molecule, with probes designed to hybridize to unique regions of a target transcript or DNA sequence. They are best exemplified by microarrays, which comprise multiple cells, each targeting a different locus, and feature densities are such that a single microarray can individually characterise every known and predicted exon in the human genome. A number of platforms employ probes spanning exon junctions to identify the presence of specific transcript isoforms, while others feature probes tiled along the length a genome in order to improve coverage[23].

By contrast, sequence-based approaches made use of small-scale Sanger sequencing of Expressed Sequence Tag (EST) libraries, produced from partial sequencing of cloned cDNAs[24]. Subsequent modifications to these approaches including serial analysis of gene expression (SAGE), cap analysis of gene expression (CAGE) and massively parallel signature sequencing (MPSS) have allowed more precise quantification of expression levels[23]. Despite the limitations of these approaches, both in terms of accuracy of quantification, and reliance on transcript annotations, these techniques have contributed significantly to our present understanding of the human

transcriptome.

1.3.1 The Affymetrix Exon Array Platform

Among the most popular microarray platforms is the Affymetrix microarray, which estimates expression from hybridization intensities of multiple perfect match (PM) and in some cases, mismatch (MM) probes in which the middle residue of the target sequence has been changed[25]. Affymetrix Exon arrays estimate exon-level expression from 1-4 probes each targeting an individual exon and together referred to as a 'probeset'. The Human Exon Array (Human Exon 1.0 array) has high coverage with more than 6.5 million probes designed against RefSeq, EST and computationally predicted transcripts[25], including many that target non-coding genes[26]. Since a large amount of publically available expression data is derived from the Exon Array platform, it forms a valuable and under-utilised resource when studying non-coding RNAs.

There are many steps in the microarray sample preparation, hybridization and imaging protocols that may introduce technical variability between samples[27]. A number of normalization methods have been developed in order to correct for these effects. The majority of these are based on the assumption that the overall distribution of probe intensities is not expected to change across samples[27] and include the popular quantile normalization approach employed by the RMA algorithm[28]. Normalization is followed by aggregation of the individual signals from each probe targeting a given locus into a single value representing a single summary value, in many cases using techniques based on a weighted average[27]. A DABG (Detection above background) score is often then used to filter probesets based on signal to noise ratio using a background value derived from a set of 25,000 background probes on each Exon array[27]. Once reliable, detected, probesets have been identified, normalized signals from these probesets can be used to estimate gene/transcript expression. To do this requires mapping to gene/transcript annotation data from databases such as Ensembl[29]/RefSeq[30]. Subsequently, the signals of all probesets mapped to a particular gene/transcript/exon/etc. can be summarized by taking an average, or in some cases, the maximum signal across each locus. At this point data are ready for further analysis.

1.3.2 Illumina Next Generation Sequencing Platform

Transcriptomics studies often aim to identify all types of transcript (mRNA, ncRNA, other novel RNAs) and determine the accurate start and end site, splice structure and

sequence and structure modifications in order to generate a comprehensive catalogue of the transcript composition of a sample, along with an accurate estimation of abundance for each transcript detected. Microarray platforms have a number of weaknesses including a low dynamic range and poor resolution at the transcript level. These have been overcome to a certain extent by next generation sequencing approaches. Illumina released the Genome Analyzer II a decade ago, and have since followed it with a number of platforms including HiSeq variants, MiSeq, and NextSeq, which together have reduced the cost of sequencing while also offering better resolution of the transcriptome[22]. The main raw data from Illumina sequencers are the read sequences along with quality scores. The following steps are performed in a standard transcriptome analysis pipeline (Fig 1.4).

1.3.2.1 Read Alignment

After the file format conversion from the raw .bcl file output from the sequencer to fastq files, and initial QC of the fastq files, sequenced reads are mapped to a reference genome/transcriptome. A number of tools such as Bowtie[31] and BWA[32] have been developed to perform this step. Transcriptome datasets from higher eukaryotes contain a significant proportion of reads that map to exon-exon junctions and will therefore not map directly to the genome. Mapping to a transcript database generally improves alignment rates but will miss novel junction reads. A number of splice-aware algorithms have been developed to address this problem. The main challenge for these algorithms is the accurate mapping of junction reads by splitting them at the correct splice site. Tools such as MapSplice[33] and TopHat[34] follow a two-step algorithm of 'tag alignment' followed by 'splice inference' to achieve successful mapping of junction reads. Reads that do not map directly to the genome are initially split at potential splice sites and then re-aligned to the genome using a gapped alignment based approach. Therefore, a number of possible alignments are obtained for each junction read. The second step involves filtering out alignments based on spurious splice sites and identifying the most likely alignment for the read. Splice-aware algorithms significantly improve the alignment rates, help identify novel splice sites, and in turn novel transcripts, but are computationally more intensive and risk false positive alignments if run with overly permissive thresholds.

1.3.2.2 *De novo* Transcriptome Assembly

One of the powerful features of next generation sequencing is that it allows the

identification of novel transcripts. To achieve this, algorithms are used to perform *de novo* assembly of the transcriptome i.e. independent of genome annotation databases. Many algorithms do this, although the absence of a gold standard dataset means that it is currently unclear which is the most accurate. One of the popular algorithms, Cufflinks[35], attempts to find the most parsimonious set of transcripts to explain the splicing information at a locus using a graph based approach to consider all reads aligned at that location locus. Cufflinks has been shown to be more conservative in comparison to other algorithms such as Scripture[36], and therefore less likely to produce false positives.

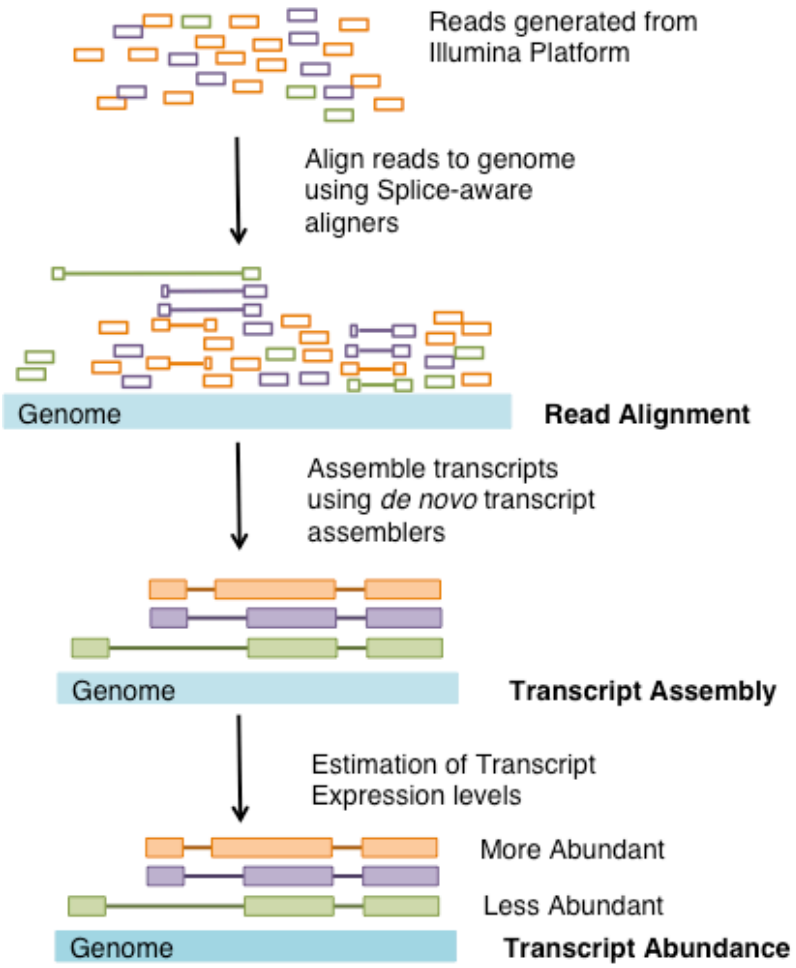

Figure 1.4 A pipeline for transcriptome data analysis for RNA-Seq data generated from Illumina Platform. This figure has been adapted from[38].

1.3.2.3 Estimation of gene/transcript abundance

Following transcriptome assembly, the next step is to estimate abundance. Gene-level counts give an overall estimate of gene expression from a particular locus without attempting to infer the levels of individual (typically overlapping) transcript isoforms from a given gene. The RSubRead[37] package in R offers an efficient method for summarizing read counts across each annotated gene. Unlike gene counts, estimating transcript levels in mammals is much more challenging and requires a more sophisticated approach to predict the likely source of a read from a complex set of overlapping transcripts expressed by a gene. Cufflinks (Cuffquant) attempts to address this by representing the problem as a linear function with a term for each transcript. Each term is assigned a weight corresponding to the level of the transcript and the problem then becomes that of estimating the most likely set of weights, given the input data. This is done using a maximum likelihood approach[35]. The final output from Cufflinks are normalized expression estimates for each transcript. The standard expression normalization takes into account the library size (total number of reads sequenced) and feature (gene/transcript) length. Cufflinks also corrects for sequence-based biases as part of the normalization process. The normalized gene/transcript counts are defined as FPKM (Fragments Per Kilobase of transcript per Million of mapped fragments) values.

1.3.2.4 Differential Expression and Splicing Analysis

Once gene expression measurements have been obtained differentially expressed genes or transcripts can be identified. A widely used tool, edgeR[39] does this using an empirical Bayes estimation and exact tests based on a negative binomial model. EdgeR and other similar tools are applicable only to differential expression analysis over gene level summaries. For transcript level differential expression analysis, the statistical model needs to take into account the unreliability of transcript abundance estimates derived from short read data. Cuffdiff, part of the Cufflinks package, is primarily designed to perform differential testing of isoform abundance estimated by Cufflinks. Cuffdiff uses a one-sided t-test of the Jensen-Shannon Divergence metric to perform these tests.

In some studies, identification of splicing changes is of particular interest. Differential splicing algorithms can be sub-divided into those based on count based models (DEXSeq[40], DSGSeq[41] and MATS[42]) and those based on isoform resolution (Cuffdiff[35]

and DiffSplice[43])[44]. DEXSeq uses transcript annotations for each gene to split them into non-overlapping bins (counting units) and then uses Generalized Linear Models (GLMs) to identify differentially used counting units. One major weakness of DEXSeq is that it does not use the information in junction reads. An alternative to DEXSeq is MATS (Multivariate Analysis of Transcript Splicing), which incorporates junction reads into its statistical model. MATS is based on a Bayesian approach to determine differential alternative splicing events from RNA-Seq data. For each exon, MATS estimates the exon inclusion levels in a pair of samples. This information is then used to model the overall similarity in alternative splicing profiles between the two samples. MATS then uses an MCMC method to calculate the Bayesian posterior probability that the splicing difference is likely to exceed a given threshold. Comparison of the observed posterior probability with simulated posterior probabilities allows the estimation of a p-value and False Discovery Rate (FDR) for each exon. MATS is able to report different alternative splicing events (Fig. 1.5) including Exon Skipping (SE), Intron Retention (RI), Mutually Exclusive Exons (MXE), Alternative 3′ and 5′ Splice Sites (A3SS, A5SS) (Fig 1.5).

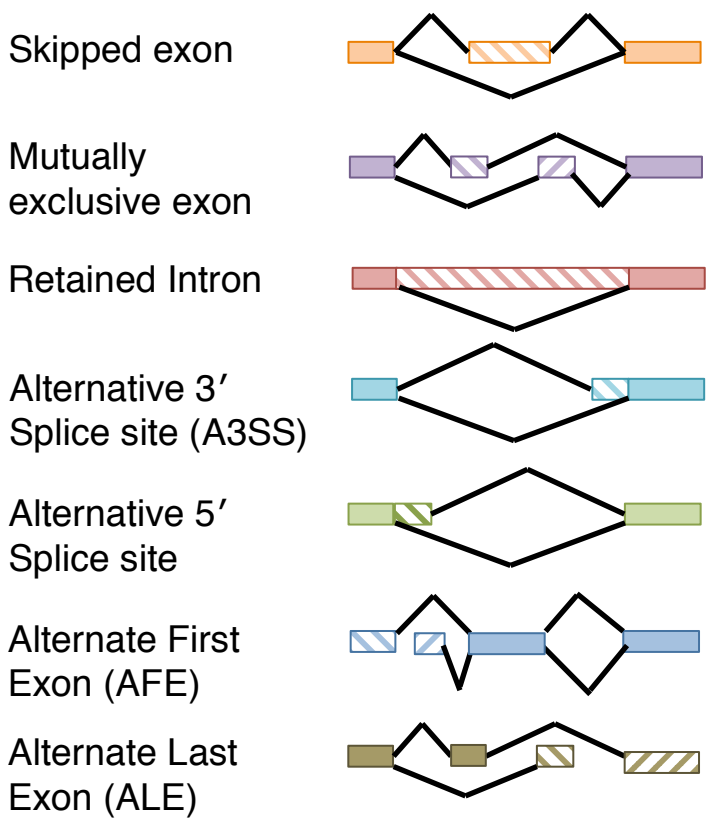

Figure 1.5 Types of alternative splicing events. This figure has been adapted from[42].

An alternative to count-based methods are isoform resolution approaches, which directly compare the relative transcript abundance across conditions. In addition to identifying differentially spliced genes, Cuffdiff also reports differential splicing changes that affect the coding sequence of protein-coding genes as well as differential promoter usage.

1.4 Annotation Databases

A critical aspect of all these pipelines is the fidelity of the genome annotations used to partition reads into different gene/transcript/exon structures. For studies focused on previously annotated genes, transcriptome data can be directly mapped to genome annotations and will not require *de novo* assembly. On the other hand, studies interested in novel transcripts can use existing annotation to guide and assess the quality of a *de novo* transcriptome assembly. There are several genome annotation databases including Ensembl[29] and RefSeq[30]. These databases contain gene and transcript annotations for all types of genes along with the evidence (computational prediction/type of experimental method) used for the annotation. Some databases such as NONCODE[45] and LNCipedia[46] are purely dedicated to ncRNAs and therefore useful for studies focused on these transcripts.

1.4.1 Ensembl Gene Annotations

One of the critical attributes in a genome annotation databases is the class (e.g. miRNA, protein coding, etc.) of a gene or transcript. The Ensembl database does this through a biotype label provided for each entry. In total, there are 57773 annotated genes in human genome in Ensembl (v74)[29], of which 35% have been classed as 'protein-coding', due to the presence of an Open Reading Frame (ORF) in the gene. Gene models lacking an ORF are classified as 'processed_transcript'. Of these, transcripts found to lie between protein coding genes longer than 200 bp are classed as 'lincRNA' while those overlapping with a protein-coding gene model on the opposite strand are classed as 'antisense'. Other gene models with a disrupted ORF, 'pseudogenes', also form a significant proportion of non-coding genes. Pseudogenes are difficult to study, as it is hard to distinguish whether a matching read originated from the protein coding gene or its pseudogene.

The general assumption is that a protein-coding gene will express a protein-coding transcript. However, many protein-coding genes can also express transcripts lacking an ORF, or those that include partial or complete introns within the mature transcript. These could either be a result of mis-splicing or the 'deliberate' expression of a non-coding isoform (generally of indeterminate function). A substantial proportion of the non-coding transcripts in Ensembl (v74) originate from within a protein coding locus. The major non-coding transcripts from protein-coding loci are 'processed_transcript', 'retained_intron' and 'nonsense_mediated_decay'.

1.5 Bioinformatics approaches to predict ncRNA function

As discussed earlier, only a handful of lncRNAs have been fully functionally characterized. Further, our current inability to identify functional elements such as lncRNA 'domains' with similar functional roles across multiple transcripts makes the inference and initial experimental characterisation of lncRNA function very challenging. Therefore, bioinformatic predictions of potential function are particularly useful as an initial lead both for 'candidate' lncRNA selection, as well as helping to identify the most useful experimental approaches to apply to the selected candidate gene[47].

A number of bioinformatics methods have been developed to predict function using expression data. The fundamental basis for all these methods is that genes that are co-expressed are more likely to be co-regulated, and therefore to be involved in the same process or pathway. Since few lncRNAs have had their functions determined, this approach instead relies on seeking associations with better-annotated protein coding genes, typically using correlation-based approaches to identify loci with similar expression profiles. This requires relatively large datasets with a significant degree of variation across the data if reliable correlations are to be found at a reasonable level of statistical significance.

1.5.1 Over Representation Analysis

Over Representation Analysis (ORA) takes a gene list from an experiment and a set of gene lists, each defining different biological signatures (for example, pathways, or biological processes such as differentiation, etc.). It then seeks gene-sets with a disproportionate number of genes present within the experimental gene list. The assumption is that these gene signatures are then likely to be associated with the

experiment[48]. Multiple methods have been used for gene over-representation analysis, including those that take into account the hierarchical structure of Gene Ontology terms and/or the biases introduced by the gene expression measurement technique (microarray/RNA-Seq). ORA is typically done using fold change and p-value cutoffs to define experimental changes observed between treatment and control groups in a microarray experiment. However, in the context of ncRNAs, gene lists could be derived by applying a hard correlation cutoff to genes co-expressed with a lncRNA of interest. This would then provide a principled way of identifying lncRNAs that have an expression profile correlated with a functionally related set of protein coding genes, with significance assessed using a hyper-geometric test followed by multiple testing correction. This approach is used extensively within this thesis.

1.5.2 Gene Set Enrichment Analysis

There are alternatives to applying stringent cutoffs and performing analysis on the list of genes satisfying the cutoff. One of them, Gene Set Enrichment Analysis (GSEA) has proven to be particularly powerful in associating known gene signatures to changes found in an experimental condition[49]. GSEA is based on the concept that in an ordered/ranked list L, the genes S belonging to the associated gene signature will show non-random distribution[49]. Therefore the starting point for GSEA is a ranked gene list. The metric used for ranking genes could be either the fold changes/FDR of genes in a microarray or NGS experiment. The initial step of GSEA is the calculation of Enrichment Score (ES)[49]. The score is derived from a running-sum statistic wherein every time the gene belonging the gene signature is encountered in the ranked gene list, the score is increased, and vice versa. ES score is an indicator of the extent of non-random distribution of genes belonging to the gene signature and therefore a higher ES suggests stronger association between the gene list and the gene signature. The next step is to estimate an empirical p-value as an estimator of the significance of ES score. Finally, the p-value is adjusted for multiple testing to account for the multiple gene signatures tested against the ranked gene list.

Not all genes in a gene set are expected to be strongly associated with a gene list. Therefore, GSEA offers a leading-edge subset analysis wherein the core genes in a gene set with the strongest signal in the gene list can be identified. The GSEA approach has been implemented both in the form of standalone software (GSEA-P) as well as R packages (R-GSEA). GSEA is particularly useful in determining subtle

changes such as small changes in expression of many genes of a metabolic pathway. The genes may not satisfy a fold change cutoff, and therefore the metabolic pathway is more likely to be identified from a GSEA-based approach than by a GOA-based approach. While GSEA is typically applied using differential expression p-values to provide the initial ranking of the gene-list under test, other metrics can also be employed including correlation. Again, this approach is employed within this thesis to identify non-coding RNAs correlated in expression with functionally related sets of protein coding genes.

1.5.3 Gene Co-expression Networks

In some cases the gene signature associated with a gene list may be completely novel and therefore ORA/GSEA based approaches may not be effective. An alternative approach is to focus on genes most strongly associated with the lncRNA by building co-expression networks and analysing these to identify patterns that may not be obvious from previous approaches. For instance, a lncRNA may be a hub in the network i.e. having unusually high number of interactions. Further the most significantly co-expressed genes can also be used as candidates for experimental verification. Although simple correlation (0^{th} order) is a powerful metric in identifying co-expressed genes, it does not imply a causal relationship between the co-expressed genes. In most gene expression datasets the number of variables (genes/transcripts) tends to be much larger than the sample size (experimental conditions). In such cases simple correlation is less effective as calculation of correlation between two variables does not take into account the effect of other variables. A more effective way of identifying co-expressed genes is using partial correlation coefficients, which calculate correlation between two variables after controlling for other variables. In the context of gene expression data, partial correlation coefficients determine the co-expression between two genes after negating the effect of other genes. Hence, partial correlations are able to identify direct relationships in contrast to simple correlations, which cannot distinguish between direct and indirect relationships. Partial correlations have been effectively used in a number of studies to predict potential interactions. A number of tools such as Cytoscape[50] and Gephi [51] have been developed to visualize and analyse co-expression networks.

1.6 Hypoxia

Hypoxia occurs when the oxygenation of a tissue drops below homeostatic levels. There are multiple definitions of hypoxia. Biochemists define it in terms of O_2 limited electron transport while physiologists and clinicians define it as reduced availability, or lack of partial pressure of oxygen[52]. In the context of this thesis, hypoxia is considered from a clinical perspective. The percentage of oxygen in air, 21%, is equivalent to a partial pressure of 159 mm of Hg. Unlike tissue culture, the levels of oxygenation can vary considerably *in vivo* depending on the cell type. For instance, liver O_2 concentration ranges between 4-5% (30-40 mm of Hg)[53] while cerebral cortex O_2 levels are considerably lower (~2.5%; ~20 mm of Hg)[54]. The definition of hypoxia is therefore highly context-dependent, as hypoxic O_2 levels for one cell (or tissue) type may be equivalent to normoxic O_2 levels for another cell type. Furthermore, physiological normoxia corresponds to a considerably lower oxygen concentration, generally in the range of 2-9%, than that of ambient air (21%)[55].

1.6.1 Tumour hypoxia

The tumour microenvironment is highly dynamic and has a significant effect on tumour metastasis, drug response and therapy[56]. It is influenced by a number of factors including pH content, oxygen levels and cellular metabolism[56]. Reduced oxygen levels or hypoxia can confer resistance to solid tumours against ionizing radiation[52,57]. Separately, a large body of clinical evidence has accumulated demonstrating the effect of hypoxia on the pathophysiology of solid tumours (Fig 1.6). More than half of locally advanced solid tumours appear to exhibit heterogeneity in oxygen levels within the tumour[58]. Hypoxia is a direct result of the imbalance between the demand and supply of oxygen[58]. A number of factors may result in tissue hypoxia such as a) oxygen tension caused by low partial pressure in arteries at high altitude, b) reduced capacity of blood to carry oxygen due to anaemia, c) reduced tissue perfusion caused by abnormalities in tumour microvessels and d) loss of diffusion geometry[58]. Of these, a transient change in oxygen levels caused by perfusion-limited oxygen delivery is a major source of *acute hypoxia*[58]. In contrast, low levels of oxygen for a prolonged period of time, caused typically by loss of diffusion geometry, results in *chronic hypoxia*[58]. The existence of both acute and chronic hypoxia in human tumours is characterized by differences with respect to changes in transcription, translation and regulation of the cell cycle. The contradictory behaviour of cells in response to hypoxia, which has been referred to as the “Janus face” of hypoxia, makes it difficult to

understand the underlying mechanism of regulation[58]. At one level, hypoxia can trigger shutdown of protein synthesis, cause cell cycle arrest, restrict proliferation and activate programmed cell death via both p53 and p53-independent mechanisms, in a manner that is dependent on the severity and length of hypoxia [58]. In conjunction, proteins induced during hypoxia enable survival in low nutrient conditions, contributing towards malignancy and promoting tumour progression. Adverse prognostic effects of tumour hypoxia arising from the formation of an aggressive malignant tumour phenotype are well established from clinical studies. Similarly, a hypoxia-induced decrease in DNA repair leading to the accumulation of mutations in the genome is also apparent[56]. Therefore, detection of tumour sub-populations affected by acute and chronic hypoxia and their differences at molecular level is a major challenge, and crucial in devising therapies against the tumours.

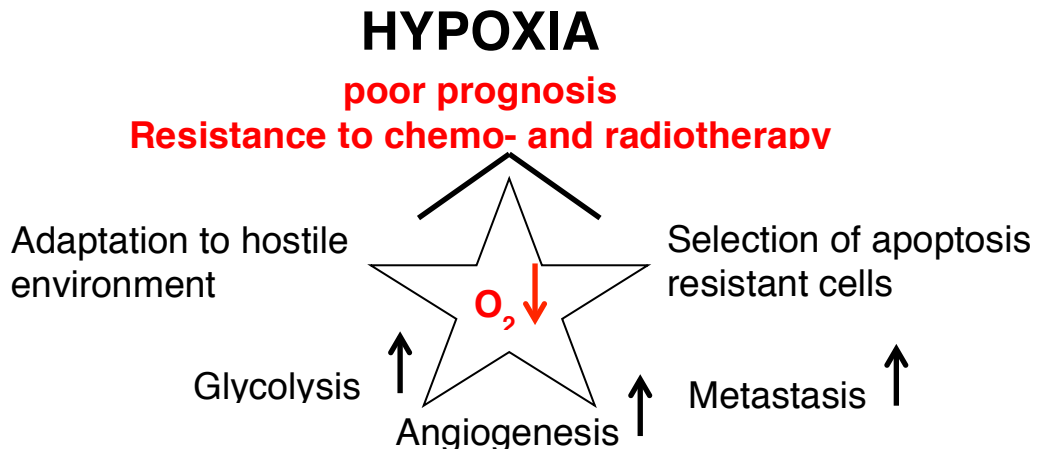

Figure 1.6 A summary of cellular response to hypoxia. This figure has been adapted from http://www.btk.fi/research/research-groups/jaakkola/.

1.6.2 Transcriptional regulation in Hypoxia

One of the most interesting attributes of cellular adaptation to hypoxia is the metabolic effect causing activation of glycolysis pathway due to a shift from aerobic to anaerobic metabolism. Other than forming an energy source, glycolysis can also reduce oxidative stress, thus extending the lifespan of the cells[59]. The regulatory mechanism for activation of the glycolytic pathway relies upon a key transcription factor HIF1A which is able to regulate majority of genes involved in glycolysis and glycolysis associated genes[59] (Fig 1.7). For instance, glucose transporters (GLUT1 and GLUT3), hexokinase, aldolase, glyceraldehyde 3 phosphate dehydrogenase, pyruvate kinase and many

others are regulated by HIF1A[59]. Similarly, *PDK1* and *MXI1*, genes that help reduce mitochondrial function, are also activated by it[59].

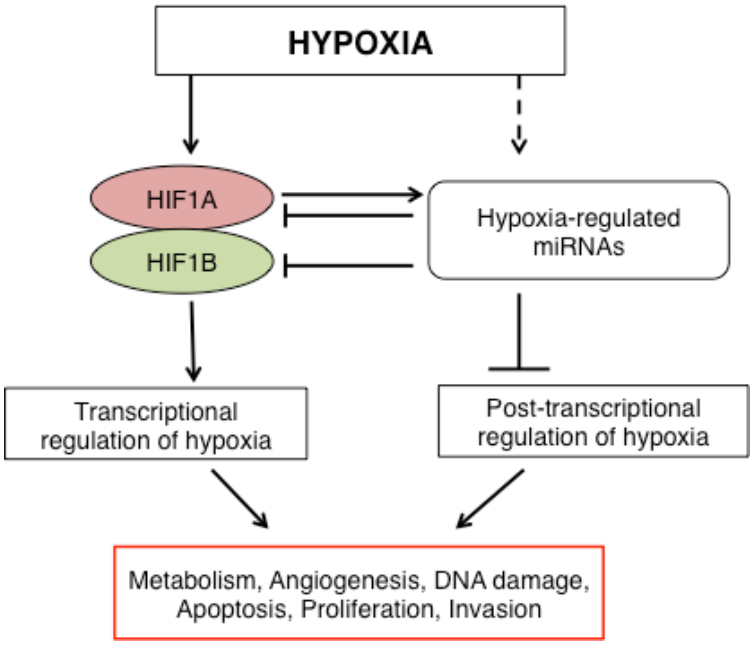

Figure 1.7 Transcriptional regulation in hypoxia. The figure has been adapted from[63].

Hypoxia-inducible factors (HIFs) are heterodimeric proteins with α and β subunits that facilitate cellular adaptation in hypoxia. There are three known HIFs (HIF-1, 2 and 3) with the ability to regulate transcription in hypoxia[60]. Among them HIF1A is ubiquitously expressed and known to play a key role in activation of hypoxia-inducible genes[61]. The HIF1A protein comprises of four degradation domains and two activation domains[62]. The oxygen-dependent degradation domain (ODD) is involved in ubiquitin-proteosomal degradation pathway active during normoxia[62]. More specifically, hydroxylation of proline residues in the ODD domain of HIF1A permits targeting by the von Hippel-Lindau protein (pVHL) and subsequent degradation of HIF1A in normoxic condition[62]. However, in hypoxia, there is interference in hydroxylation of the ODD domain of HIF1A, causing stabilization and translocation to the nucleus, where it dimerises with the β subunit, arylhydrocarbon receptor nuclear translocator (ARNT), and then binds to hypoxia response elements (HREs) in the promoter region of hypoxia-inducible genes[62]. The HIF-1 binding site 5′-(A-G)CGTG-3′ is part of the HRE and allows activation of a large number of genes[63]. HIF-1 dependent programs have been shown to be activated pO2 level falls below 10-15 mm Hg triggering expression of genes

involved in pH maintenance, glucose transport and angiogenesis. Further, a drop in oxygen levels leads to cut down on energy production (ATP synthesis) and decrease in protein synthesis.

1.6.3 Egr1 and Hypoxia

There are several HIF-independent transcription changes in response to hypoxia. For instance, hypoxia triggers induction of c-Fos which heterodimerises to form Jun/c-Fos AP-1 complex, regulating expression of a number of cell growth associated genes[64-67]. Similarly, expression of inflammatory cytokine IL-6 is also regulated by hypoxia-dependent activation of the transcription factor C/EBPβ[68,69]. Recently, several groups have shown that many HIF1A dependent genes such as VEGFA, NDRG1 and TF can also be regulated by other transcription factors such as Egr1. In fact HIF1A levels have also been shown to be modulated by Egr1 in hypoxia. Egr1 (also known as NGFI-A, zif268 and krox24) is an Early Growth Regulator Protein and belongs to a family of transcription factors that are induced immediately in response to various external stimuli. Egr1 was first identified by Sukhatme *et. al* [70] in mouse fibroblasts. The human Egr1 gene is located on the 5q31.2 locus and is expressed as a 3.7 kb mRNA. The promoter region of Egr1 comprises of binding sites for serum response elements (SRE elements), CRE element, NFkB and SP1 [71] (Fig 1.8A). In addition, Egr1 has the ability to bind to its own promoter region via its own Egr1 binding site (EBS)[71]. The Egr1 protein is an 80kD nuclear phosphoprotein with Ser/Thr rich region and Pro/Ser/thr rich region near the N and C-terminus respectively. The Egr1 protein can be sub-divided into three domains: the DNA-binding domain, the activator domain and the Repressor domain[71] (Fig 1.8B). The DNA-binding domain comprises of three canonical C2H2 zinc-finger motifs that facilitate Egr1 binding to sequence specific regions in the major groove of the DNA double helix[71]. The activator domain of Egr1 extends from the N-terminus of the protein and distributed over the Serine/Threonine rich region[71]. The repressor domain of Egr1 is critical for its interaction with NAB1 and NAB2 protein which act as co-repressors of Egr1[71]. Egr1 has very low basal expression across majority of tissues but highly induced in response to growth factors, mitogens, stress and cytokines[71]. Induction of Egr1 expression can be driven by the MAPK/ERK pathway (Fig 1.9). Interaction of growth factors with their receptors triggers a MAPK Signalling cascade initiated via adaptor proteins and cytoplasmic tyrosine kinase to activate Raf (MAPKKK). Egr1 has the ability to regulate the expression of genes by binding to the promoter regions at the EBS, a GC-rich consensus motif, and regulate

diverse sets of genes include growth factors, hormones (LH), cytokines (IL8), lipoproteins and adhesion molecules. Egr1 is able to both activate as well as repress transcription.

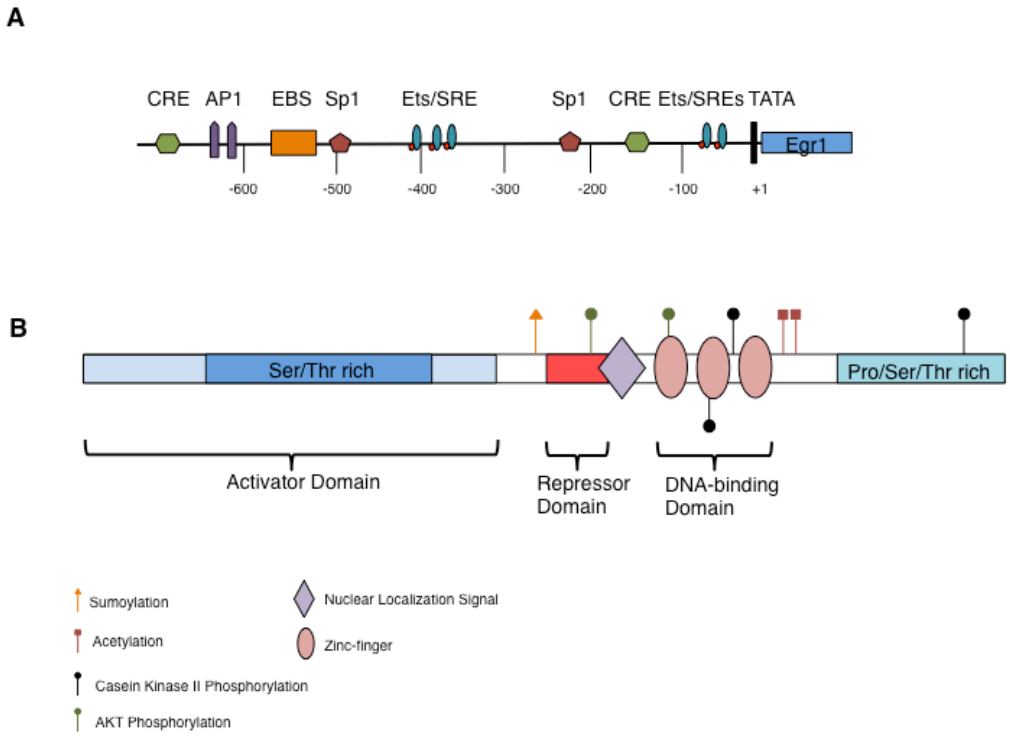

Figure 1.8 A schematic representation of Egr1 (A) promoter region and (B) protein domains. Panel A has been adapted from[72]. Panel B has been adapted from http://atlasgeneticsoncology.org/Genes/EGR1ID496ch5q31.html.

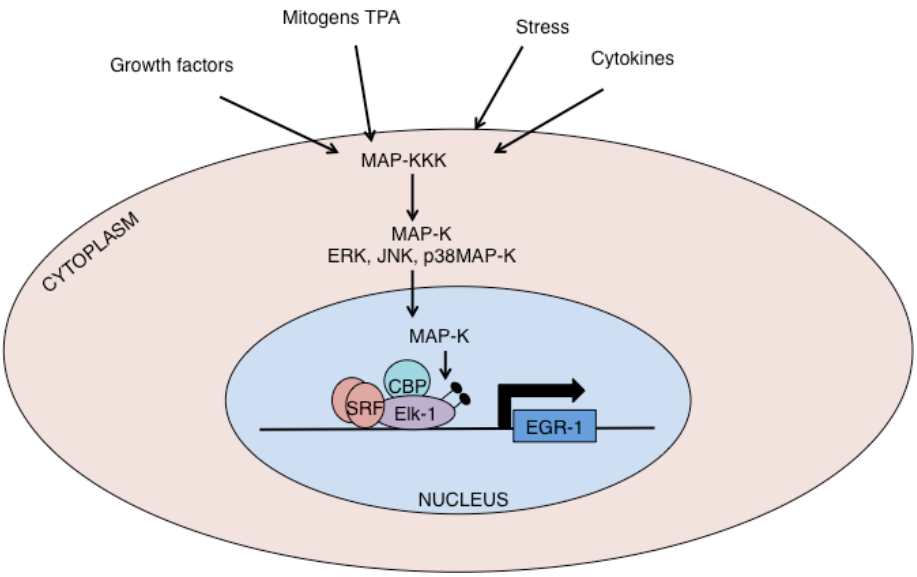

Figure 1.9 Mechanism of activation of Egr1 in response to external stimuli. The figure has been adapted from[73].

1.6.4 Non-coding RNAs in hypoxia

As discussed earlier, ncRNAs have been shown to play a critical role in transcriptional and post-transcriptional regulation, therefore it is reasonable to ask whether the gene expression changes observed in response to hypoxia are influenced by ncRNAs. Among ncRNAs, the role of miRNAs in hypoxia has been extensively studied (Fig 1.7). Several independent studies have carried out miRNA expression profiling across different cancer cell levels at varying oxygen concentrations[74-79] and more than 90 hypoxia regulated miRNAs (HRMs) have been identified[61,62]. The majority of these miRNAs are cell line specific to the extent that miR-210 is the only miRNA that was found to be upregulated across all studies. This miRNA is expressed from an intronic region of a non-coding RNA host gene, *MIR210HG*, also differentially expressed in hypoxia[80]. Investigations into the role of miR-210 have revealed dose-dependent changes in miR-210 abundance in response to changes in oxygen concentration. More interestingly, it appears that the miR-210 activity is hypoxia-specific, as no change was observed due to osmotic stress, changes in pH or growth factors[81,82]. The promoter of miR-210 contains HREs, thus allowing activation of transcription through binding of HIF1A in the promoter region. Therefore, the current model of miR-210 activity involves HIF1A-based upregulation of *MIR210HG* and in resulting in increased levels of miR-210, which then down-regulates a large number of other genes. Recent efforts to identify targets of miR-210 have shown that miR-210 does not upregulate hypoxia-inducible genes, but that its main role is in the repression of a large number of genes active during normoxia[83]. In fact, only one of 50 genes identified as targets of miR-210 was found to be hypoxia-inducible[83].

In contrast to miRNAs, only a handful of hypoxia-related lncRNAs have been identified. These have been extensively reviewed in[8]. Among the hypoxia-dependent lncRNA include NEAT1, linc-ROR, HINCUT1, UCA1, H19, WT1 lncRNA, AK058003, lncRNA LET, lincRNA-P21 and EFNA3. Several report HIF-dependent transcription of ncRNAs in hypoxia. An initial study reported widespread binding of hypoxia-inducible factor (HIF) and RNA polII on non-coding loci in hypoxic cells. Further, HIF was shown to activate transcription via release of pre-bound promoter-paused RNA polII[84,85]. The same study also reported the lncRNA NEAT1 to be induced upon oxygen deprivation in MCF-7 breast cancer cells. Hypoxia-independent induction of HIF also increased NEAT1 levels suggesting the critical role of HIF in activation of this lncRNA. HIF1A has

been implicated to be direct or indirect upstream regulator for several lncRNA including lincRNA-p21, UCA1, HINCUT-1, H19 and EFNA3[8]. The majority of these lncRNAs play a critical role in modulating key processes in hypoxia and loss of these lncRNAs can deregulate cell metabolism (lincRNA-p21), induce tumourigenesis (lincRNA-p21), influence invasion and metastasis (EFNA3, UCA1), and affect cell viability and survival (UCA1, lncRNA-ROR, HINCUT and NEAT1).

As described in section 1.2.5, a number of mechanisms of action for lncRNAs have been reported via interaction with DNA, RNA and/or proteins. In the context of hypoxia, several of these mechanisms have been described. LncRNA driven epigenetic changes have been consistently reported in the literature, including the lncRNA WT1 which modulates methylation levels at the TSS of WT1 mRNA, thus affecting its expression levels[86]. LncRNAs can in some cases act as a sponge to protect the mRNA from miRNA-based regulation. For example, the hypoxia-induced lncRNA linc-RoR prevents the miR-145 mediated down-regulation of p70S6K1 (RPS6KB1)[87]. Since p70S6KB1 is a kinase that activates protein synthesis through phosphorylation of the S6 ribosomal protein, the sponge activity of linc-RoR affects the proteins levels of many genes including HIF1A[8]. Other mechanisms include the interaction of lncRNA with proteins, such as lincRNA-P21. Both HIF1A and lincRNA-p21 interact with the VHL protein at a common site, therefore competing for binding. Since interaction of HIF-1A and VHL triggers the proteasome dependent degradation of HIF-1A, induction of lincRNA-p21 leads to accumulation of HIF1A in hypoxia[8]. In this thesis, we explore the role of hypoxia-induced ncRNAs.

1.7 References

1. Ng, S.-Y., Bogu, G. K., Soh, B. S. & Stanton, L. W. The long noncoding RNA RMST interacts with SOX2 to regulate neurogenesis. *Mol. Cell* **51,** 349–359 (2013).
2. Robertson, M. P. & Joyce, G. F. The origins of the RNA world. *Cold Spring Harb Perspect Biol* **4,** (2012).
3. Moran, V. A., Perera, R. J. & Khalil, A. M. Emerging functional and mechanistic paradigms of mammalian long non-coding RNAs. *Nucleic Acids Research* **40,** 6391–6400 (2012).
4. Lodish, H. *et al. Molecular Cell Biology*. (W. H. Freeman, 2016).
5. Kornblihtt, A. R. *et al.* Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nat. Rev. Mol. Cell Biol.* **14,** 153–165 (2013).
6. Rinn, J. L. & Chang, H. Y. Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* **81,** 145–166 (2012).
7. Pavet, V., Portal, M. M., Moulin, J. C., Herbrecht, R. & Gronemeyer, H. Towards

novel paradigms for cancer therapy. *Oncogene* **30,** 1–20 (2011).
8. Chang, Y.-N. *et al.* Hypoxia-regulated lncRNAs in cancer. *Gene* **575,** 1–8 (2016).
9. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458,** 223–227 (2009).
10. Ponjavic, J., Oliver, P. L., Lunter, G. & Ponting, C. P. Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain. *PLoS Genet* **5,** e1000617 (2009).
11. Bitton, D. A. *et al.* Programmed fluctuations in sense/antisense transcript ratios drive sexual differentiation in S. pombe. *Mol. Syst. Biol.* **7,** 559 (2011).
12. Leong, H. S. *et al.* A global non-coding RNA system modulates fission yeast protein levels in response to stress. *Nature Communications* **5,** 1–10 (2014).
13. Vance, K. W. & Ponting, C. P. Transcriptional regulatory functions of nuclear long noncoding RNAs. *Trends Genet.* **30,** 348–355 (2014).
14. Roberts, T. C., Morris, K. V. & Weinberg, M. S. Perspectives on the mechanism of transcriptional regulation by long non-coding RNAs. *Epigenetics* **9,** 13–20 (2014).
15. Khalil, A. M. *et al.* Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 11667–11672 (2009).
16. Fox, A. H. & Lamond, A. I. Paraspeckles. *Cold Spring Harb Perspect Biol* **2,** a000687 (2010).
17. Poliseno, L. *et al.* A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **465,** 1033–1038 (2010).
18. Cesana, M. *et al.* A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell* **147,** 358–369 (2011).
19. Gong, C. & Maquat, L. E. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3′ UTRs via Alu elements. *Nature* **470,** 284–288 (2011).
20. Wang, J., Gong, C. & Maquat, L. E. Control of myogenesis by rodent SINE-containing lncRNAs. *Genes & Development* **27,** 793–804 (2013).
21. Berg, P. Origins of the human genome project: why sequence the human genome when 96% of it is junk? *Am. J. Hum. Genet.* **79,** 603–605 (2006).
22. Reuter, J. A., Spacek, D. V. & Snyder, M. P. High-throughput sequencing technologies. *Mol. Cell* **58,** 586–597 (2015).
23. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10,** 57–63 (2009).
24. Nagaraj, S. H., Gasser, R. B. & Ranganathan, S. A hitchhiker's guide to expressed sequence tag (EST) analysis. *Brief. Bioinformatics* **8,** 6–21 (2007).
25. Kapur, K., Xing, Y., Ouyang, Z. & Wong, W. H. Exon arrays provide accurate assessments of gene expression. *Genome Biol* **8,** R82 (2007).
26. Du, Z. *et al.* Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. *Nat. Struct. Mol. Biol.* **20,** 908–913 (2013).
27. Okoniewski, M. J. & Miller, C. J. Comprehensive analysis of affymetrix exon arrays using BioConductor. *PLoS Comp Biol* **4,** e6 (2008).
28. Hardiman, G. *Microarray Innovations*. (CRC Press, 2009).
29. Flicek, P. *et al.* Ensembl 2014. *Nucleic Acids Research* **42,** D749–55 (2014).
30. Pruitt, K. D. *et al.* RefSeq: an update on mammalian reference sequences. *Nucleic Acids Research* **42,** D756–63 (2014).
31. Langmead, B. Aligning short sequencing reads with Bowtie. *Curr Protoc Bioinformatics* **Chapter 11,** Unit 11.7 (2010).
32. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
33. Wang, K. *et al.* MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Research* **38,** e178 (2010).
34. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions

with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).
35. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28,** 511–515 (2010).
36. Guttman, M. *et al.* Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* **28,** 503–510 (2010).
37. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30,** 923–930 (2014).
38. Haas, B. J. & Zody, M. C. Advancing RNA-Seq analysis. *Nat. Biotechnol.* **28,** 421–423 (2010).
39. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).
40. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Research* **22,** 2008–2017 (2012).
41. Wang, W., Qin, Z., Feng, Z., Wang, X. & Zhang, X. Identifying differentially spliced genes from two groups of RNA-seq samples. *Gene* **518,** 164–170 (2013).
42. Shen, S. *et al.* MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. *Nucleic Acids Research* **40,** e61 (2012).
43. Hu, Y. *et al.* DiffSplice: the genome-wide detection of differential splicing events with RNA-seq. *Nucleic Acids Research* **41,** e39 (2013).
44. Liu, R., Loraine, A. E. & Dickerson, J. A. Comparisons of computational methods for differential alternative splicing detection using RNA-seq in plant systems. *BMC Bioinformatics* **15,** 364 (2014).
45. Liu, C. *et al.* NONCODE: an integrated knowledge database of non-coding RNAs. *Nucleic Acids Research* **33,** D112–5 (2005).
46. Volders, P.-J. *et al.* LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Research* **41,** D246–51 (2013).
47. Ma, H. *et al.* Molecular mechanisms and function prediction of long noncoding RNA. *ScientificWorldJournal* **2012,** 541786 (2012).
48. Beissbarth, T. & Speed, T. P. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics* **20,** 1464–1465 (2004).
49. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **102,** 15545–15550 (2005).
50. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research* **13,** 2498–2504 (2003).
51. Bastian, M., Heymann, S. & Jacomy, M. Gephi : An Open Source Software for Exploring and Manipulating Networks. 1–2 (2009). at <http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154>
52. Höckel, M. & Vaupel, P. Tumor hypoxia: definitions and current clinical, biologic, and molecular aspects. *J. Natl. Cancer Inst.* **93,** 266–276 (2001).
53. Wölfle, D., Schmidt, H. & Jungermann, K. Short-term modulation of glycogen metabolism, glycolysis and gluconeogenesis by physiological oxygen concentrations in hepatocyte cultures. *Eur. J. Biochem.* **135,** 405–412 (1983).
54. Whalen, W. J., Ganfield, R. & Nair, P. Effects of breathing O 2 or O 2 +CO 2 and of the injection of neurohumors on the PO 2 of cat cerebral cortex. *Stroke* **1,** 194–200 (1970).
55. Simon, M. C. & Keith, B. The role of oxygen availability in embryonic development and stem cell function. *Nat. Rev. Mol. Cell Biol.* **9,** 285–296 (2008).

56. Bristow, R. G. & Hill, R. P. Hypoxia and metabolism. Hypoxia, DNA repair and genetic instability. *Nat. Rev. Cancer* **8,** 180–192 (2008).
57. GRAY, L. H., CONGER, A. D., EBERT, M., HORNSEY, S. & SCOTT, O. C. The concentration of oxygen dissolved in tissues at the time of irradiation as a factor in radiotherapy. *Br J Radiol* **26,** 638–648 (1953).
58. Vaupel, P. & Mayer, A. Hypoxia in cancer: significance and impact on clinical outcome. *Cancer Metastasis Rev.* **26,** 225–239 (2007).
59. Welford, S. M. & Giaccia, A. J. Hypoxia and senescence: the impact of oxygenation on tumor suppression. *Mol Cancer Res* **9,** 538–544 (2011).
60. Lu, X. & Kang, Y. Hypoxia and hypoxia-inducible factors: master regulators of metastasis. *Clin. Cancer Res.* **16,** 5928–5935 (2010).
61. McCormick, R., Buffa, F. M., Ragoussis, J. & Harris, A. L. The role of hypoxia regulated microRNAs in cancer. *Curr. Top. Microbiol. Immunol.* **345,** 47–70 (2010).
62. Rankin, E. B. & Giaccia, A. J. The role of hypoxia-inducible factors in tumorigenesis. *Cell Death Differ.* **15,** 678–685 (2008).
63. Pocock, R. Invited review: decoding the microRNA response to hypoxia. *Pflugers Arch.* **461,** 307–315 (2011).
64. Müller, J. M., Krauss, B., Kaltschmidt, C., Baeuerle, P. A. & Rupec, R. A. Hypoxia induces c-fos transcription via a mitogen-activated protein kinase-dependent pathway. *J. Biol. Chem.* **272,** 23435–23439 (1997).
65. Norris, M. L. & Millhorn, D. E. Hypoxia-induced protein binding to O2-responsive sequences on the tyrosine hydroxylase gene. *J. Biol. Chem.* **270,** 23774–23779 (1995).
66. Yao, K. S., Xanthoudakis, S., Curran, T. & O'Dwyer, P. J. Activation of AP-1 and of a nuclear redox factor, Ref-1, in the response of HT29 colon cancer cells to hypoxia. *Mol. Cell. Biol.* **14,** 5997–6003 (1994).
67. Mishra, R. R., Adhikary, G., Simonson, M. S., Cherniack, N. S. & Prabhakar, N. R. Role of c-fos in hypoxia-induced AP-1 cis-element activity and tyrosine hydroxylase gene expression. *Brain Res. Mol. Brain Res.* **59,** 74–83 (1998).
68. Yan, S. F. *et al.* Induction of interleukin 6 (IL-6) by hypoxia in vascular cells. Central role of the binding site for nuclear factor-IL-6. *J. Biol. Chem.* **270,** 11463–11471 (1995).
69. Yan, S. F. *et al.* Nuclear factor interleukin 6 motifs mediate tissue-specific gene transcription in hypoxia. *J. Biol. Chem.* **272,** 4287–4294 (1997).
70. Sukhatme, V. P. *et al.* A novel early growth response gene rapidly induced by fibroblast, epithelial cell and lymphocyte mitogens. *Oncogene Res.* **1,** 343–355 (1987).
71. Pagel, J.-I. & Deindl, E. Early growth response 1--a transcription factor in the crossfire of signal transduction cascades. *Indian J. Biochem. Biophys.* **48,** 226–235 (2011).
72. Hasan, R. N. & Schafer, A. I. Hemin upregulates Egr-1 expression in vascular smooth muscle cells via reactive oxygen species ERK-1/2-Elk-1 and NF-kappaB. *Circ. Res.* **102,** 42–50 (2008).
73. Ngiam, N., Post, M. & Kavanagh, B. P. Early growth response factor-1 in acute lung injury. *Am. J. Physiol. Lung Cell Mol. Physiol.* **293,** L1089–91 (2007).
74. Kulshreshtha, R. *et al.* Regulation of microRNA expression: the hypoxic component. *Cell Cycle* **6,** 1426–1431 (2007).
75. Hua, Z. *et al.* MiRNA-directed regulation of VEGF and other angiogenic factors under hypoxia. *PLoS ONE* **1,** e116 (2006).
76. Hebert, C., Norris, K., Scheper, M. A., Nikitakis, N. & Sauk, J. J. High mobility group A2 is a target for miRNA-98 in head and neck squamous cell carcinoma. *Mol. Cancer* **6,** 5 (2007).
77. Donker, R. B., Mouillet, J.-F., Nelson, D. M. & Sadovsky, Y. The expression of

Argonaute2 and related microRNA biogenesis proteins in normal and hypoxic trophoblasts. *Mol. Hum. Reprod.* **13,** 273–279 (2007).
78. Guimbellot, J. S. *et al.* Correlation of microRNA levels during hypoxia with predicted target mRNAs through genome-wide microarray analysis. *BMC Med Genomics* **2,** 15 (2009).
79. Pulkkinen, K., Malm, T., Turunen, M., Koistinaho, J. & Ylä-Herttuala, S. Hypoxia induces microRNA miR-210 in vitro and in vivo ephrin-A3 and neuronal pentraxin 1 are potentially regulated by miR-210. *FEBS Lett* **582,** 2397–2401 (2008).
80. Camps, C. *et al.* hsa-miR-210 Is induced by hypoxia and is an independent prognostic factor in breast cancer. *Clin. Cancer Res.* **14,** 1340–1348 (2008).
81. Chan, S. Y. *et al.* MicroRNA-210 controls mitochondrial metabolism during hypoxia by repressing the iron-sulfur cluster assembly proteins ISCU1/2. *Cell Metab.* **10,** 273–284 (2009).
82. Fasanaro, P. *et al.* MicroRNA-210 modulates endothelial cell response to hypoxia and inhibits the receptor tyrosine kinase ligand Ephrin-A3. *J. Biol. Chem.* **283,** 15878–15883 (2008).
83. Huang, X. *et al.* Hypoxia-inducible mir-210 regulates normoxic gene expression involved in tumor initiation. *Mol. Cell* **35,** 856–867 (2009).
84. Choudhry, H. *et al.* Extensive regulation of the non-coding transcriptome by hypoxia: role of HIF in releasing paused RNApol2. *EMBO Rep.* **15,** 70–76 (2014).
85. Choudhry, H. & Mole, D. R. Hypoxic regulation of the noncoding genome and NEAT1. *Brief Funct Genomics* (2015). doi:10.1093/bfgp/elv050
86. McCarty, G. & Loeb, D. M. Hypoxia-sensitive epigenetic regulation of an antisense-oriented lncRNA controls WT1 expression in myeloid leukemia cells. *PLoS ONE* **10,** e0119837 (2015).
87. Zhou, X. *et al.* Linc-RNA-RoR acts as a 'sponge' against mediation of the differentiation of endometrial cancer stem cells by microRNA-145. *Gynecol. Oncol.* **133,** 333–339 (2014).

Chapter 2. The mammalian ncRNAome

In silico analysis of whole body gene expression data reveals non-coding RNA regulators of the cell cycle

Danish Memon, Jing Bi, Crispin J Miller*

RNA Biology Group, CRUK Manchester Institute, Manchester M20 4BX UK
*corresponding author: crispin.miller@cruk.manchester.ac.uk

2.1 Introduction

The relatively recent emergence of non-coding RNAs as functional molecules in the cell means that few have been functionally characterised and existing classifications have been based solely on relatively arbitrary criteria such as gene size. Selecting 'important' long intergenic ncRNAs (lincRNAs) for functional characterisation is a critical challenge, and the identification of essential and disease-associated lincRNA genes would be of great utility. Previous studies have shown that 'core' genes essential for cell viability are over-represented among ubiquitously expressed genes, while tissue-specific genes are frequently perturbed in human disease. We used the Human BodyMap RNA-Seq dataset to identify ubiquitously expressed and tissue-specific lincRNAs (HK/TS-lincRNAs). We identified a small yet significant subset of HK-lincRNAs, including *Neat1, Malat1 and JPX*, that are ubiquitously expressed in the majority of human tissue types. HK-lincRNAs tend to be mono-exonic, localized in the neighbourhood of housekeeper protein-coding genes, show higher base-level conservation, lower mutation rates and higher editing rates. Exons of HK-lincRNAs are long and enriched for SINE and LINE elements that are predicted to contribute to stable secondary structures. The majority of HK-lincRNAs are ubiquitously expressed in other mammalian transcriptomes. Correlative analyses using independent data from TCGA and Tumourscape implicated the majority of HK-lincRNAs in core 'housekeeping' functions including a significant subset with a predicted role in the cell cycle. Core essential protein-coding genes are less likely to be down-regulated or deleted across tumour datasets derived from TCGA and Tumourscape in comparison with other genes. By applying the same logic to HK-lincRNAs we identified a set

of lincRNAs with stable expression in tumours thus likely to be essential for 'survival'.

The catalogue of known genes has expanded considerably since publication of the first draft of the human genome in 2001[1] in part through the detection of additional protein-coding loci[2-4] but also as a result of the identification of thousands of novel non-coding genes – to the extent that they are now thought to outnumber proteins by a factor of at least two to one[5]. The majority of these comprise a highly heterogenous set of transcripts termed 'long non-coding RNAs' (lncRNAs), a definition based purely on their length (> 200 bp) and inability to code for a protein. LncRNAs regulate a large number of developmental and biological pathways including gene imprinting, cell differentiation, the cell cycle and apoptosis. Many lncRNAs are under the control of canonical transcription factors such as p53 and Sox2[6], and have, conversely, been shown to regulate protein-coding gene expression both in *cis* and in *trans*[7,8,9].

Given their relatively recent discovery, it is not surprising that despite rapid progress in the field, the majority of lncRNAs have yet to be the subject of detailed investigation. Where they have been characterised two common themes have emerged: First, the ability of lncRNAs to hybridise through sequence complementarity, often via repeat sequences[10,11], allows precise targeting of a lncRNA to a given DNA or RNA locus. Second, the formation of double stranded RNA (dsRNA) within a molecule supports the establishment of stable structures that lend specificity to interactions with particular proteins[12,13]. Together, these properties allow lncRNAs to perform a diversity of scaffolding and targeting roles throughout the cell.

However, unlike proteins, for which function is determined by the complex biochemical properties of a set of interacting amino acids, the primacy of sequence in driving lncRNA function allows them to evolve more rapidly; a substitution of one base can often be compensated for by a complementary substitution at its binding partner. Thus, while negative correlation between evolutionary rate and lncRNA expression levels has been demonstrated[14], the majority of lncRNAs are less well-conserved than those of proteins, and undergo only weak positive or neutral selection at the sequence level[11]. A major consequence of this is that sequences typically evolve too rapidly to allow evolutionary lineages to be traced using phylogenetic approaches. This, when combined with the relative paucity of annotated data points in lncRNA-space, means

that attempts to establish a comprehensive functional taxonomy of lncRNAs have so far been unsuccessful.

A number of studies have sought to identify protein-coding genes that are essential for growth and viability[15-25]. These provide significant insights into the potential role of different protein coding genes and an important step towards more detailed classification. Currently, no such catalogue exists for lncRNAs. Here we provide an initial list of candidate essential lncRNA genes, and use expression-based approaches to infer potential roles for these loci.

2.2. Results

2.2.1 Identification of housekeeping lncRNAs

Transcript profiles across 16 human tissues: adipose, adrenal, brain, breast, colon, heart, kidney, liver, lung, lymph node, ovary, prostate, skeletal muscles, testes, thyroid and white blood cells were generated from the Illumina Human BodyMap 2.0 RNA-Seq dataset. We first reanalysed the data using Cufflinks to generate *de novo* transcript assemblies and mappings to Ensembl (v74). A total of 107651 known transcripts (28660 genes) were detected, including 15637 protein-coding loci and 4770 lncRNA genes (2343 antisense and 2427 lincRNAs). Since the BodyMap data do not preserve strand information, expression measurements for antisense transcripts were less reliable. We therefore discarded these, and considered only long intergenic non-coding RNAs (lincRNAs) > 1kb from the nearest protein coding gene. As expected[11], lincRNA levels were substantially lower than those at protein coding loci (Supplementary Figure 2.1). Normalised transcript-levels (Coefficient of Variance; CV) segregated into a clear bimodal distribution irrespective of gene type (Fig 2.1A), with the majority of protein-coding genes exhibiting low CV (> 60% with CV < 1.5). In contrast, lincRNAs were considerably more variable (< 25% with CV < 1.5; Fig 2.1A). We classified lincRNAs as 'housekeeping' (HK-lincRNAs) if detected across more than 75% of tissues with a CV < 1.5, and tissue-specific (TS-lincRNAs) if detected in less than 25% of tissues with a CV > 3.5 (Fig 2.1B). In total, 55 HK-lincRNA and 721 TS-lincRNA remained following this stringent classification (Fig 2.1C; Supplementary Table 2.1). Both NEAT1 and NEAT2/MALAT1 were classified as HK-lincRNAs by this strategy, in keeping with their critical role in paraspeckle structures in the nucleus[26,27]. It is tempting to speculate that their ubiquity of expression across tissue types is indicative of a role in basic cellular function. While HK-lincRNAs are detected across the tissue panel, many are expressed

at higher levels in tissues derived from the ovary (median expression ~4 FPKM) and at lower levels in the liver (median expression = ~2 FPKM). The majority of TS-lincRNAs are specific to testis.

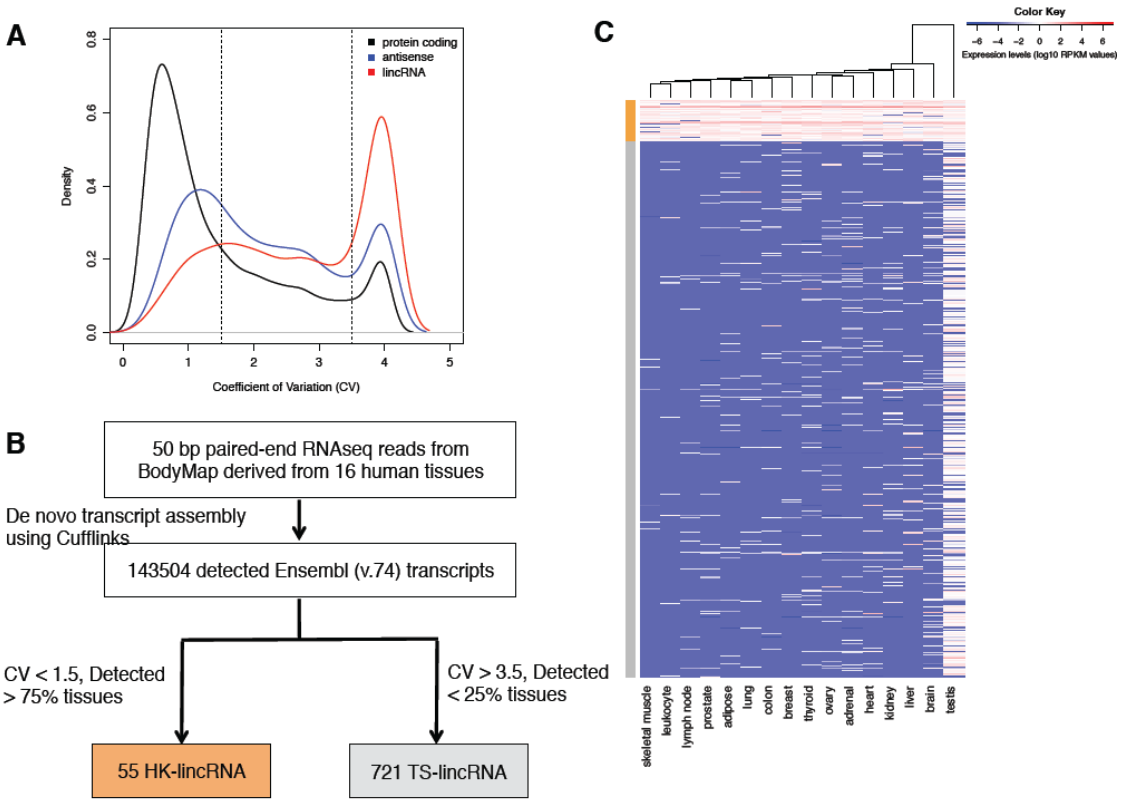

Figure 2.1 Identification of HK-lincRNA and TS-lincRNA. (A) Distribution of coefficient of variation (CV) for transcripts belonging to different Ensembl biotypes. (B) A flowchart of the steps followed in the selection of HK-LincRNA and TS-LincRNA. (C) Expression profile of HK-LincRNA and TS-LincRNA across 16 tissues in the BodyMap data.

HK-lincRNA	Symbol	Minimum Free Energy (MFE)	Differential Expression in Tumours	Copy Number Alterations in Tumours
ENST00000416769	LINC00339	-341.6	-	-
ENST00000602412	RP11-343N15.5	-270.17	Unaffected	-
ENST00000540383	CROCCP2	-722.7	Unaffected	-
ENST00000606641	RP4-758J24.5	-487.64	Unaffected	-
ENST00000435649	RP4-665J23.1	-259.53	-	-
ENST00000442526	RP11-517P14.2	-835.5	Down	-
ENST00000605920	RP11-182L21.6	-791.76	Unaffected	Amplified
ENST00000499732	NEAT1	-678.74	Unaffected	-
ENST00000534336	MALAT1	-2627.97	Unaffected	-
ENST00000508564	RP11-834C11.4	-527.19	Down	-
ENST00000551450	RP3-462E2.3	-3238.35	-	-
ENST00000609803	LINC00938	-687.19	Unaffected	-
ENST00000546580	RP11-620J15.3	-188.31	Up/Down	-
ENST00000552780	RP11-2H8.2	-353.26	-	-
ENST00000554921	RP11-1112J20.2	-244.16	Up	-
ENST00000602330	CTD-2576F9.2	-163.7	-	-
ENST00000582940	RP11-160O5.1	-219.3	Up	-
ENST00000566986	RP13-516M14.1	-438	-	-
ENST00000602353	RP11-78O7.2	-170.2	Down	Amplified
ENST00000585086	RP11-690G19.3	-473.31	Unaffected	-
ENST00000582866	RP11-498C9.15	-805.25	-	-
ENST00000581471	LINC00667	-714.54	Down	-
ENST00000600047	CTC-444N24.8	-375.18	Unaffected	-
ENST00000587762	MIR24-2	-753.4	Down	-
ENST00000590677	LINC00662	-269.22	Unaffected	Deleted
ENST00000602458	RP11-95D17.1	-1053.25	Unaffected	Deleted
ENST00000567540	RP11-254F7.2	-508.41	Unaffected	Deleted
ENST00000426713	LINC00116	-171.5	Unaffected	-
ENST00000409569	MIR4435-1HG	-164.8	Up	-
ENST00000435844	LINC00493	-140.7	Unaffected	-
ENST00000565493	LINC00657	-1738.33	Up	-
ENST00000602901	LINC00478	-80.8	-	-
ENST00000460407	RP11-38P22.2	-501.95	-	-
ENST00000609183	RP11-434H6.7	-82.51	Down	-
ENST00000307533	AC093323.3	-947.4	Unaffected	-
ENST00000466692	RP11-1398P2.1	-216.9	Unaffected	-
ENST00000513179	RP11-539L10.3	-102.63	Unaffected	-
ENST00000514608	RP11-21I10.2	-259.31	-	-
ENST00000502001	MIR4458HG	-317	Unaffected	-
ENST00000607056	RP11-53O19.3	-642.72	Up	-
ENST00000501937	LINC00847	-715.94	Unaffected	Amplified/Deleted
ENST00000606482	CTD-2081C10.7	-178.76	-	-
ENST00000411553	HCG11	-1545.92	-	-
ENST00000567732	CTA-14H9.5	-77.9	Down	-
ENST00000564837	RP11-611L7.1	-700.02	Up	-

ENST00000564834	LINC01003	-827.38	Unaffected	-
ENST00000580458	SNHG15	-279.14	Up	-
ENST00000439105	AC074183.4	-366.5	-	-
ENST00000610021	RP4-813F11.4	-659.41	-	-
ENST00000606064	RP11-722E23.2	-159.62	Unaffected	-
ENST00000606963	CTD-3025N20.3	-128.22	Unaffected	Deleted
ENST00000593237	RP11-220I1.1	-1070.94	-	Deleted
ENST00000444125	RP11-65J3.1	-296.82	Up/Down	-
ENST00000603385	RP11-258C19.7	-443.56	Unaffected	-
ENST00000602985	JPX	-108.57	Unaffected	-

Table 2.1 HK-lincRNAs identified from the BodyMap RNA-Seq dataset. The table also lists information about estimated minimum free energy, differential expression in tumours and overlap with regions in the genome showing copy number alterations.

2.2.2 qRT-PCR based validation of HK-lincRNAs

To verify the predictions we performed qRT-PCR of two of the HK-lincRNAs, AC093323.3 and RP11-220I1.1, across 20 tissue types. Both lincRNAs were detected in all tissues, despite their low expression, confirming the predictions made from the BodyMap data (Fig 2.2) and indicating that other predicted HK-lincRNAs are also likely to be ubiquitously expressed.

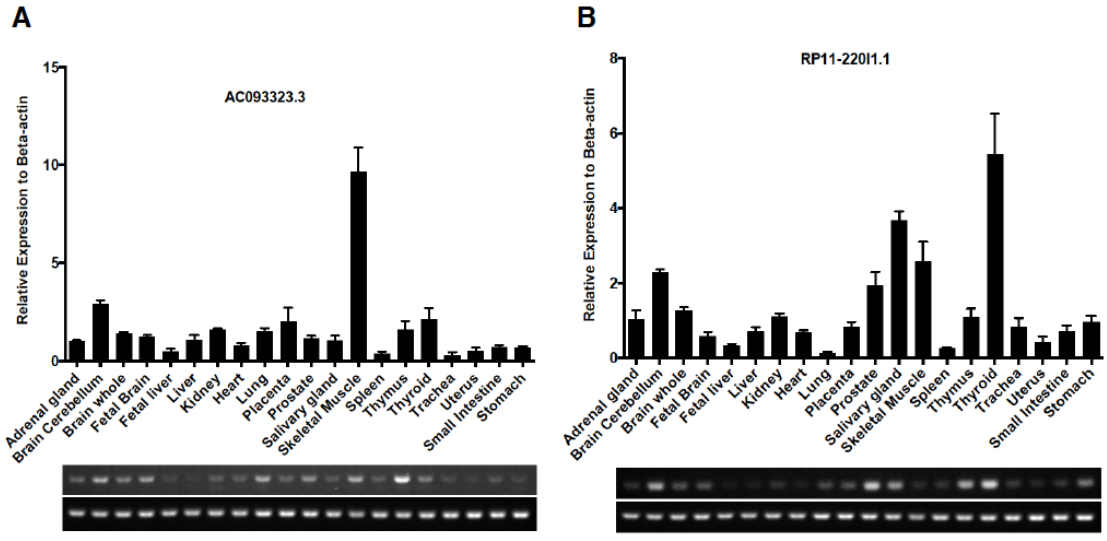

Figure 2.2 qRT-PCR of AC093323.3 and RP11-220I1.1. Relative expression levels of HK-lincRNAs: (A) AC093323.3 and (B) RP11-220I1.1. Expression levels of the lincRNAs have been normalized relative to B-Actin.

2.2.3 HK-lincRNAs are frequently single exon and proximal to a protein coding locus

HK-lincRNAs were enriched for single exon transcripts (51%: 28/55) with respect to TS-lincRNA (6.5%: 47/721; *p*-value < 0.01) or randomly sampled lncRNAs (mean 22%: 12/55; *p*-value < 0.01), and tend to comprise longer exons (median exon length for HK-lincRNAs: *L* = 332 bp; TS-lincRNAs: *L* = 152 bp; *p*-value < 0.01). A substantial proportion of HK-lincRNAs (19/55; ~35%) are within 10kb of protein-coding genes (median inter-gene distance: 18224 bp), while the majority of TS-lincRNAs (669/721; ~93%) are further than 10Kb, placing them in apparent 'gene deserts'. HK-lincRNA are pre-dominantly in tandem configuration with respect to their neighbouring gene (70% in tandem, 18% in convergent, 12% in divergent) and show moderate positive correlation with their nearest protein-coding neighbour (median Pearson correlation of gene expression in the BodyMap data: 0.35).

2.2.4 HK-lincRNAs are more conserved, contain fewer SNPs, are more frequently edited than TS-lincRNAs and have more stable secondary structure

We compared the Phylop conservation score[28] of lincRNA exons derived from a 46-way alignment of mammalian genomes for HK- and TS-lincRNAs. Nucleotide level conservation was significantly higher for the HK set (mean conservation score 0.135) than the TS-lincRNAs (mean conservation score 0.065; *p*-value < 0.01; Figure 2.3A). Mutation rates were also marginally lower for HK vs TS-lincRNA exons (18.5 vs. 19.7 SNPs per gene per kb; *p*-value < 0.05; Figure 2.3B). Previous studies have shown that housekeeping protein-coding genes are marked by presence of CpG islands in their promoter region[18,23]. 45% (25/55) of HK-lincRNAs (including MALAT1) contain CpG islands in their 1kb 5′ upstream proximal region in contrast to 15% (107/721) of TS-lincRNAs. Since lincRNAs have been reported to be enriched for repeat elements[10,29], we compared the repeat distribution of HK-lincRNAs and TS-lincRNAs. We found SINE/Alu elements to constitute ~7.7% of nucleotides in HK-lincRNA exons in contrast to ~2.4% of nucleotides in TS-lincRNA exons (Figure 2.4A). This observation is consistent with previous reports of a positive correlation between number of SINE elements and higher expression[10]. In addition, we observed a strong negative correlation between the proportion of SINE elements and the distance from the nearest protein-coding neighbour. Since, HK-lincRNAs tend to be more localized in the neighbourhood of protein-coding genes (median distance from closest protein-coding gene less than 10 kb), the enrichment of SINE elements may be driven by genome

architecture. Repeat elements are subject to high levels of editing in the nucleus[30]. We compared the RNA editing rates for HK-lincRNA and TS-lincRNAs in publically available datasets using the RADAR database[31]. The mean edit rate for HK-lincRNA was 1.3 per kb, significantly higher than the mean edit rate for TS-lincRNAs (0.18 per kb; *p*-value < 0.01). The secondary structure of ncRNA plays a critical role in ncRNA activity[32] and regulatory non-coding RNA sequences (such as pre-miRNAs) tend to have folding free energies indicative of a tendency to form stable secondary structure[33]. We used minimum free energy (MFE) estimated using randfold[33] as a parameter to assess the stability of lincRNA secondary structure. Overall, HK-lincRNA sequences exhibit significantly lower MFE values (median MFE = -375.18 kcal/mol) than TS-lincRNA (median MFE = -140.93 kcal/mol; *p*-value < $2.9e^{-16}$). In order to negate the effect of sequence length on MFE values, we compared the MFE value of each lincRNA with a distribution of MFE values generated from shuffling the lincRNA sequence while maintaining the dinucleotide composition. 22% (12/55) of HK-lincRNAs were found to have significantly lower MFE values when compared to randomly shuffled sequences (FDR < 0.05; Fig 2.4B). No TS-lincRNA was found to have significantly low MFE values with this approach.

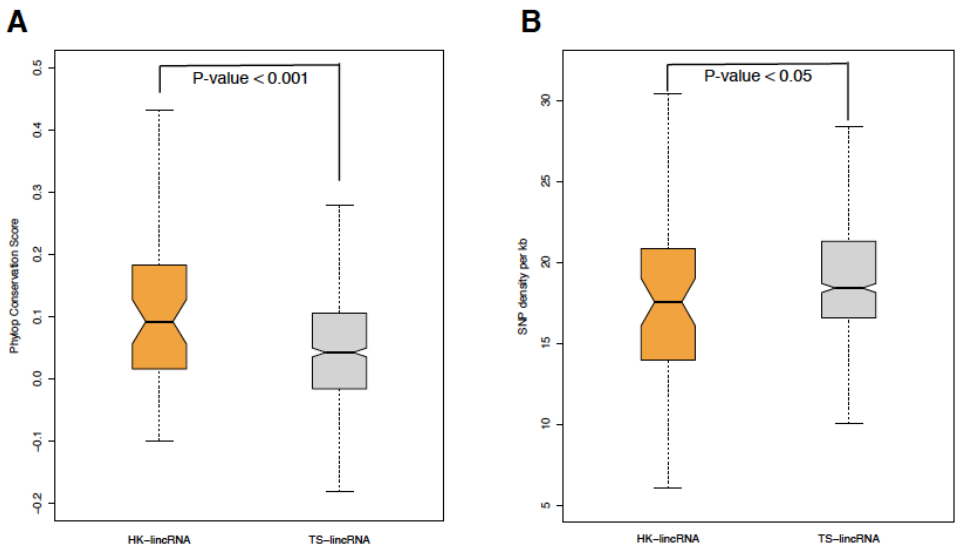

Figure 2.3 Properties of HK-lincRNA and TS-lincRNA. Comparison of (A) nucleotide-level conservation score and (B) SNP density of exons of HK-lincRNAs and TS-lincRNAs.

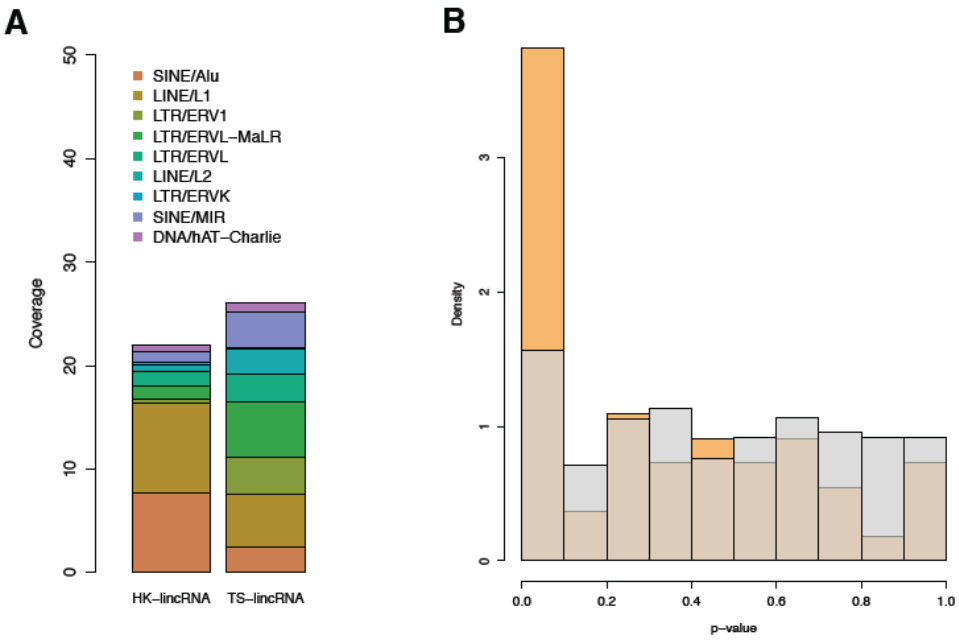

Figure 2.4 Repeat composition and secondary structure stability of HK-lincRNAs and TS-lincRNAs. Comparison of (A) repeat composition and (B) stability of secondary structure of HK-lincRNAs and TS-lincRNAs. Minimum free energy, calculated using Randfold software [33], was used as an indicator of secondary structure stability. An empirical p-value was estimated for each lincRNA after randomly shuffling the sequence and re-calculating the MFE for the random sequences. Each lincRNA sequence was shuffled a 1000 times and the dinucleotide composition of the original sequence was retained. The p-values of HK-lincRNAs are indicated in orange while the p-values of TS-lincRNAs are indicated in grey.

2.2.5 HK-lincRNAs are ubiquitously expressed in other mammals

Housekeeping protein coding genes have been reported to be highly conserved across other species. We therefore used BLAST tool to infer homologues of HK-lincRNAs in *de novo* assembled transcriptomes from five other eukaryotes (Rhesus, cow, rat, mouse, chicken), generated using the same pipeline used to reannotate the BodyMap data. The number of predicted homologues decreased progressively with phylogenetic distance. We found 95.5% (50/55) of HK-lincRNAs to be expressed in the transcriptome of at least one other mammalian species. In contrast, only 63.9% (462/721) of TS-lincRNA were expressed in other species (Fig 2.5A). Further, 89% (48/54) HK-lincRNA were also ubiquitously expressed in other species (CV < 1.5), with a low overall CV (median CV =1; Fig 2.5B). In contrast, 63% (283/447) TS-lincRNAs showed tissue-specific expression profile in other species with high CV (median CV = 2). Together, these data indicate that human TS- and HK-lincRNA are not only conserved, but display similar expression patterns in other species.

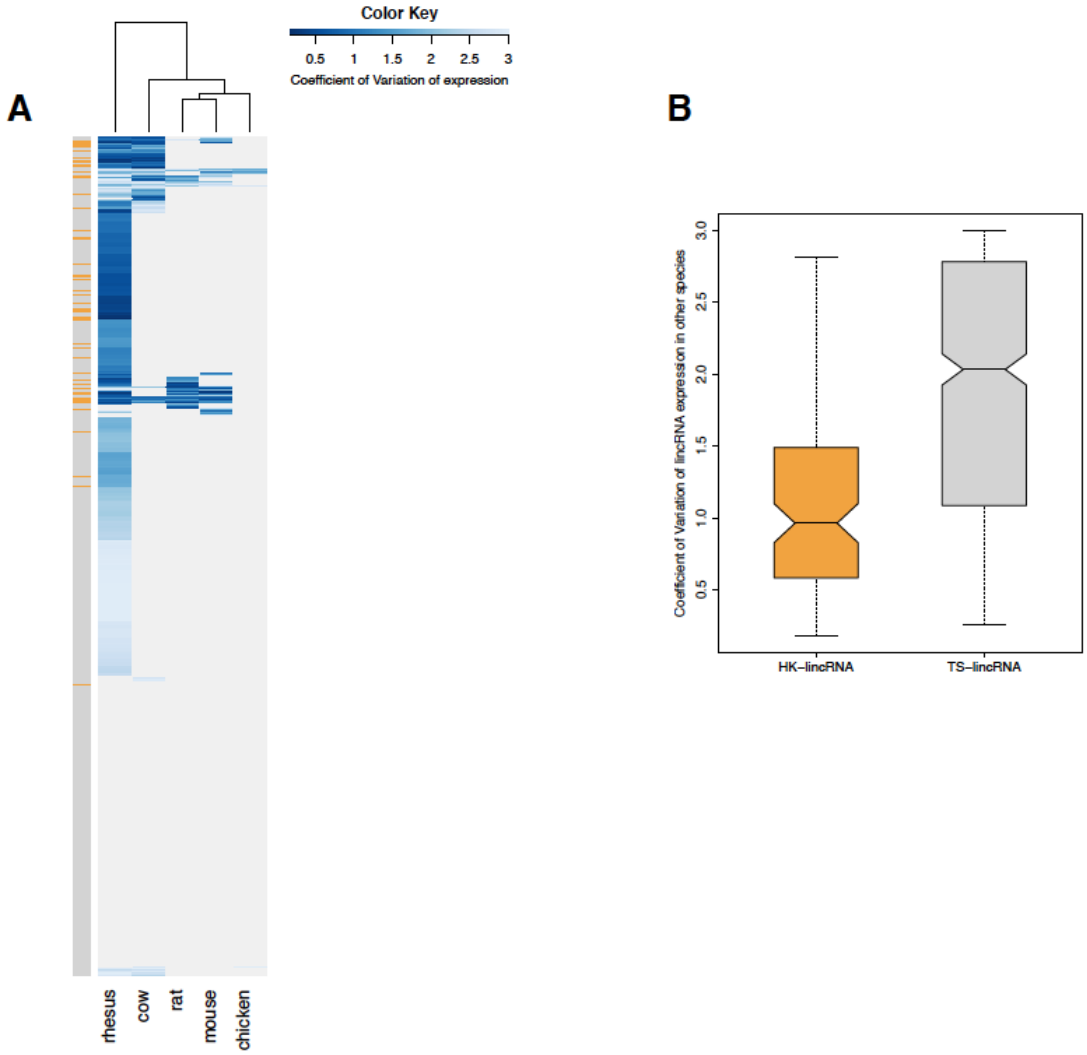

Figure 2.5 Conservation of expression of HK-lincRNA in other species. (A) Coefficient of variation of expression of HK-lincRNAs and TS-lincRNA homologues in other species. (B) Comparison of coefficient of variation of expression between HK-lincRNA and TS-lincRNA. For each of the five species, the expression data were obtained for eight organs from Merkin *et. al.*[36] and was used to calculate the coefficient of variation of lincRNA expression.

2.2.6 HK-lincRNAs are involved in key housekeeping functions

Many lncRNA genes bear strong resemblance to canonical protein-coding loci, with similar chromatin marks[6], PolII mediated transcription, well-defined intron-exon structures, and similar downstream processing including splicing, 5′-capping and 3′ polyadenylation[11]. These clear patterns suggest that lncRNAs are under active regulatory control. We therefore hypothesised that lncRNAs with highly correlated expression profiles to functionally related sets of proteins might therefore be governed by common regulatory mechanisms, and therefore involved in similar processes. We used a large Affymetrix Exon array dataset of 182 Encode cell lines (Tier 1, Tier 2 and Tier 3) derived from a diverse set of normal and tumour tissues to calculate gene expression correlations between protein-coding genes and HK-lincRNAs[34,35]. 50/55 HK-lincRNAs were supported by one or more reliable Exon array probeset, allowing function prediction to be performed for the majority of HK-lincRNAs. As expected, HK-lincRNA had significantly higher expression levels with respect to the TS-lincRNA in the same dataset (Supplementary Figure 2.2). Proteins with significant positive or negative correlations to HK-lincRNAs were subjected to Over-Representation Analysis (ORA) to identify Gene Ontology Biological Processes with strong statistical associations to each HK-lincRNA (Fig 2.6). We were able to detect at least one significantly associated biological process for 40/55 HK-lincRNAs. In an unsupervised analysis, HK-lincRNAs clustered into two major sub-groups comprising 30 and 10 HK-lincRNAs, respectively. In total, 24 were significantly associated with *cell cycle* while other fundamental processes including RNA splicing, chromatin organization, oxidative phosphorylation, protein folding and protein targeting were represented in the larger cluster. The smaller cluster of 10 HK-lincRNA included many with significant positive association to tissue-level processes including the regulation of cell migration, cell adhesion, regulation of signal transduction and angiogenesis.

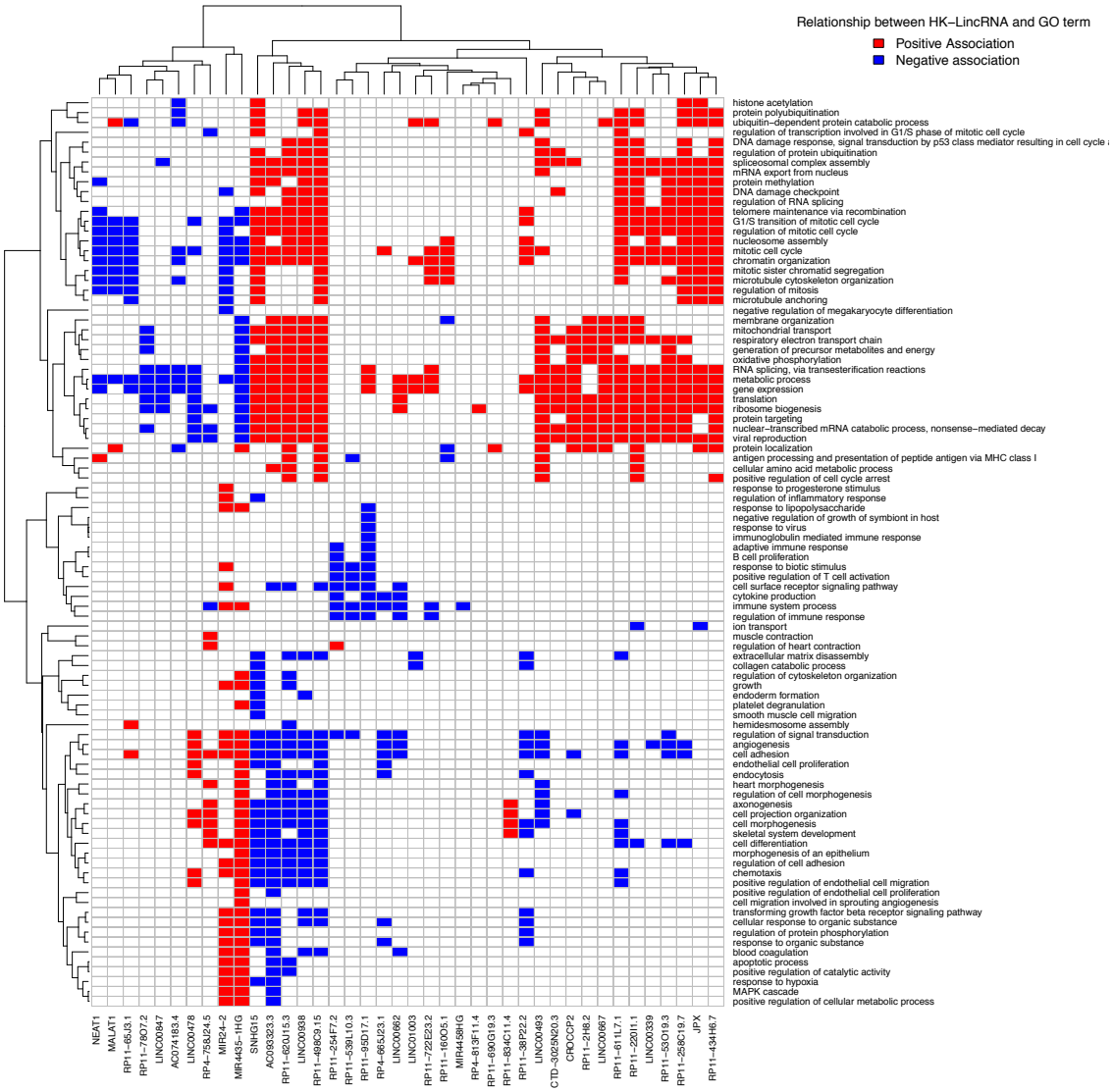

Figure 2.6 Function prediction of HK-lincRNA. HK-lincRNAs are significantly associated with key housekeeping biological processes critical for cellular homeostasis. Rows represent GO 'Biological Processes' terms and columns represent HK-lincRNAs. Cells are coloured as red or blue based on significant positive or negative association between a HK-lincRNA and a biological process. Significant association was defined by strong pearson correlation of genes expression with FDR < 0.05. Correlations were calculated using the publicly available Exon Array dataset of 182 Encode cell lines.

2.2.7 Core essential genes are rarely down-regulated in tumours

We next hypothesised that essential genes would be unlikely to be deleted or down-regulated in human tumours, since loss would be expected to lead to cell death. We first tested this hypothesis using a previously reported set of 291 'core' essential protein-coding genes identified through RNAi knockdown experiments. This is indeed

the case: Differential expression analysis was performed between matched normal and tumour samples for 13 different tumour types obtained from TCGA. In total, 90.7% (264/291) of the 'core' essential genes were detected in normal tissues, consistent with the expectation of ubiquitous expression. Importantly, only 6.1% (16/264) of these were downregulated, while 33.3% (88/264) were upregulated in at least one tumour (differential expression > two-fold change and q-value < 0.05).

We therefore applied a similar approach to identify HK-lincRNAs that were rarely downregulated in TCGA tumour data. Only 2.3% (212/9142) of lncRNAs were detected across all sample types. These included 72.7% (40/55) of the HK-lincRNAs identified using the BodyMap data. Of these, only 9 were downregulated in one or more tumours (RP11-517P14.2, RP11-620J15.3, RP11-834C11.4, RP11-78O7.2, LINC00667, MIR24-2, RP11-434H6.7, CTA-14H9.5, RP11-65J3.1; FC > 2, q-value < 0.05), and a further 9, upregulated (RP11-620J15.3, RP11-1112J20.2, RP11-160O5.1, MIR4435-1HG, LINC00657, RP11-53O19.3, SNHG15, RP11-611L7.1, RP11-65J3.1). The remaining 24 HK-lincRNAs remained unaffected (Table 1, Fig 2.7). These include JPX, loss of which has been previously shown be to lethal in females[37].

2.2.8 Core essential genes are rarely deleted in tumours

Beroukhim *et al.*[38] previously described 76 focal amplifications and 82 focal deletions identified from pooled analysis of copy number alterations across 12 different tumour types. We compared these loci to the core essential protein coding gene set described above and found that essential protein coding genes were under-represented in focal-deletion loci, while no such enrichment was observed in focal-amplifications. A total of 24 essential genes (0.27%) were found to be part of focal deletion regions and 25 essential genes (0.46%) overlapped with focal amplification regions. This is in keeping with the hypothesis that deletion of these genes would be deleterious to cell survival. We therefore applied a similar strategy to HK-lincRNA loci. While 7.6% (545/7109) and 15% (1117/7109) of all lncRNAs were found to fall within focal amplifications and deletions, respectively, only 9 HK-lincRNAs (9.6% of total length) fell in these loci (Table 1). In contrast, 175 TS-lincRNA (20.5% of total length) mapped to these regions of frequent loss or gain. These data demonstrate that HK-lincRNAs are less likely to be amplified or deleted in tumours, while TS-lincRNAs are frequently perturbed, suggesting that they provide a rich source of novel disease-related genes.

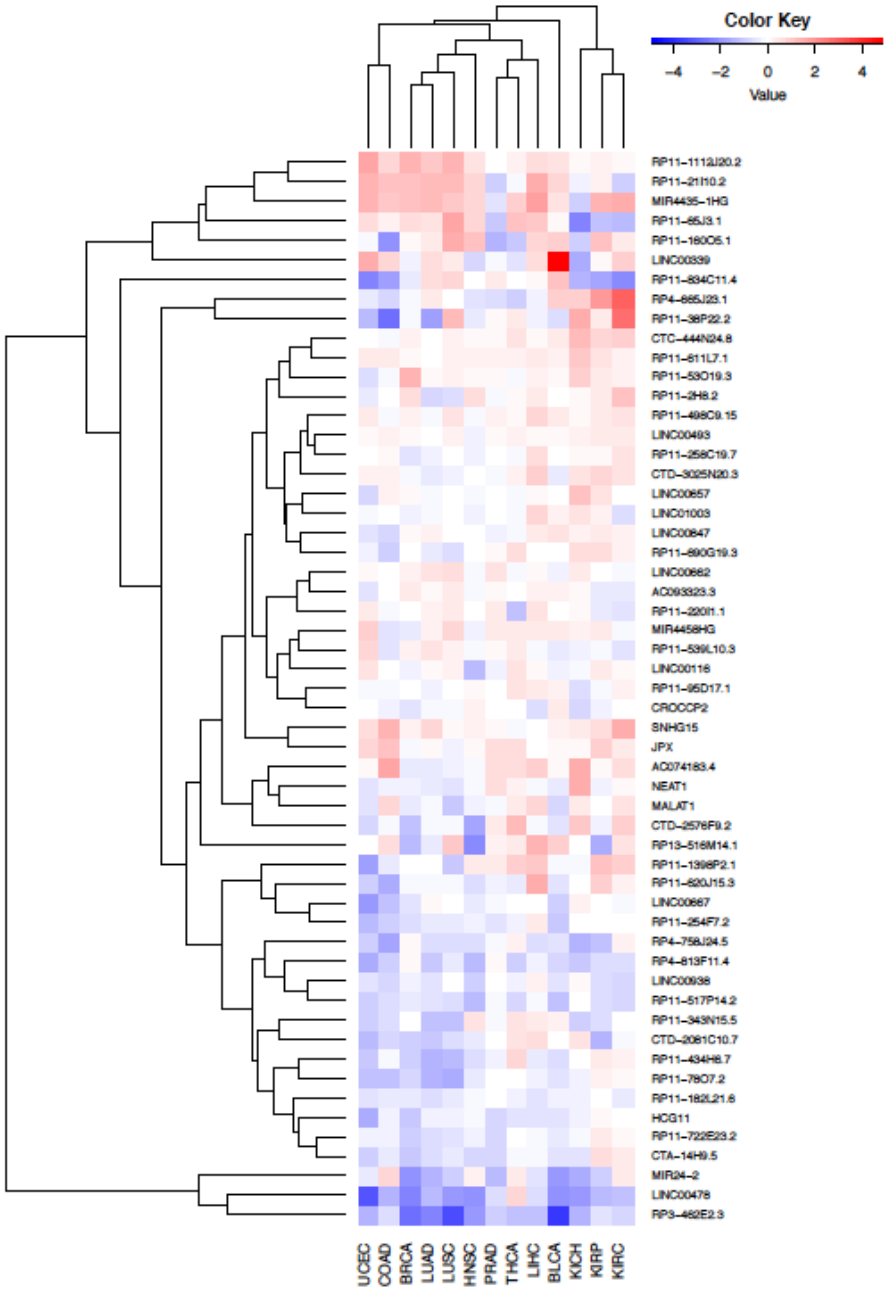

Figure 2.7 Dysregulation of HK-lincRNA in tumours. Expression change of HK-lincRNAs represented as fold changes in tumour samples relative to matched normal samples. The RNA-Seq data for each tumour type was obtained from TCGA.

2.3 Discussion

Studies of Ubiquitously Expressed (protein coding) Human Genes (UEHGs), or 'housekeepers' have been extensive. Although the catalogue of housekeeping genes varies considerably across different studies, with a poor overlap among gene lists reported by different groups[17] arising as a consequence of the different technologies

used to generate expression data, differences in the precise definition of a 'housekeeper', and the considerable changes in genome annotation that have occurred over the past decade, distinct structural, evolutionary and promoter features[17-23] have been described, and mean expression levels are higher than those of tissue-specific loci[18,24]. While coding sequence is slower to evolve in housekeepers relative to tissue-specific genes[18,25], promoters at housekeeping loci show lower sequence conservation[23]. Although housekeeping genes have been the focus of considerable attention, previous work has concentrated on protein coding loci. Here we present the first systematic study of housekeeping lncRNAs and demonstrate that housekeeping and tissue-specific lncRNAs behave in similar ways to their protein-coding counterparts.

Given their ubiquity, it is not surprising that deregulation of lncRNA expression is linked to many diseases including cancer[39], however, despite the rapid progress of the field, a fundamental question that remains unclear is how many are 'essential' for the growth and viability of the organism. Knockout studies in mouse have been able to define an extensive set of essential protein-coding genes, many of which exhibit conserved function in humans[40]. While essential protein coding genes are mechanistically diverse, bioinformatics studies have revealed distinct properties such as slow rates of evolution and a tendency to form highly connected hubs in protein-protein interaction networks[41]. They are often ubiquitously expressed[15] and frequently involved in basic cellular functions and the maintenance of cellular homeostasis[16,18]. Hart *et al.* used an RNAi based approach to distinguish 'core' essential genes necessary for survival in all cell lines in their study from context-dependent essential genes required only in certain cell types [42]. We applied the same concept to lncRNAs arguing that a subset of ubiquitously expressed lncRNA are likely to be 'core' essential lncRNA.

We identified hundreds of loci with stable expression not only in normal tissues but also in the chaotic environment of a tumour cell. Candidate housekeeping lncRNAs are rarely down-regulated or deleted in multiple tumour types, and functional enrichment analysis predicts molecular roles for these loci in multiple core processes including the cell cycle, chromatin organisation and protein biogenesis, as well as critical tissue-level programmes that include migration, adhesion, cell signalling and angiogenesis. Finally, we describe a complementary set of tissue-specific lncRNAs that are frequently perturbed in tumours. Through correlative analyses we were able to significantly associate many of these transcripts with functionally related sets of protein coding

genes, suggesting that they are under similar patterns of regulation within the cell, and thus involved in broadly similar processes. Together these data provide a catalogue of candidate lncRNAs that may act as novel therapeutic targets and biomarkers in human disease, including cancer.

2.4 Experimental Procedures

2.4.1 Quantitative RT-PCR Analysis

Total RNA from cell lines were extracted using the RNeasy mini kit (QIAGEN, 74104) and reverse transcribed using the M-MLV reverse transcriptase (Promega, M1701). Human total RNA from 20 different tissue sites was purchased from Clonetech, 636643. Gene and LincRNAs expression were quantified by qPCR (Fast start SYBR green, Roche, 04673484001) and shown as normalized expression relative to beta-actin. Error bars represent the SDs of the average expression based on three experimental replicates.

Primers for qRT-PCR (AC093323.3): GCCTGCGTTTTCTCCACATT (forward), GCAGCAGCGTACGTACTGTA (reverse);

Primers for qRT-PCR (RP11-220I1.1): AGCAGTACTGGGGACTTACA (forward), GCAAGACTCCACTGCCAAAA (reverse);

2.4.2 Dataset description

LncRNA expression measurements were obtained from the publically available illumina Human Body Map RNA-seq set generated from the Human BodyMap 2.0 Project. This dataset comprises of RNA-seq data obtained from 16 human tissues: adipose, adrenal, brain, breast, colon, heart, kidney, liver, lung, lymph node, ovary, prostate, skeletal muscles, testes, thyroid and white blood cells with an average of 160 million reads sequenced from each tissue. High read depth is critical for non-coding RNAs, which tend to be more lowly expressed as compared to their coding counterparts. The comprehensive nature of the dataset has facilitated a number of bioinformatics studies on various RNA species, their regulation and their relationship with each other.

2.4.3 Processing of BAM files

The BAM files of 50mer paired reads aligned to the human genome (Hg19) using TopHat (v2) were downloaded from the UCSC Browser (link). Transcript models were derived for each sample independently using Cufflinks (v2.2.0; with default parameters). Resultant models were then merged using Cuffmerge to provide a global model and to classify transcripts as novel, or known, when they mapped to ENSEMBL (v74). For each gene, we identified the most abundant (highest mean expression) 'known' transcript and thus ended up with only 28660 transcripts.

2.4.4 LncRNA Conservation, Mutation and Secondary Structure

The nucleotide-level conservation scores for human (hg19) were obtained from the 'phyloP46wayPrimates' track in the UCSC database (http://genome.ucsc.edu/index.html). Mutation data was obtained from dbSNP database (http://www.ncbi.nlm.nih.gov/snp). Both the conservation and mutation data was intersected with lincRNA exon annotations in Ensembl using bedtools (http://code.google.com/p/bedtools/), which allowed us to make inferences about differences between HK-lincRNA and TS-lincRNA in terms of conservation rates and mutation density. Calculation of minimum free energy from lncRNA transcript sequences and estimation of *p*-value from MFE distribution was performed using randfold software [33]. For lncRNA conservation in other species, aligned expression data from eight organs of five species [36] was downloaded and subjected to *de novo* transcript assembly and quantification using cufflinks and the corresponding Ensembl genome annotation as guide.

2.4.5 Gene Over-representation Analysis

The GOA analysis was performed using a publically available Exon Array dataset (GSE19090) comprising of expression measurements from 182 Encode cell lines (tier 1, tier 2 and tier 3 cell types). The Affymetrix GeneChip Human Exon 1.0 ST Array had reliable probesets targeting 50 out of 55 HK-lincRNAs. Reliable probesets were then mapped to the ENSEMBL human genome annotation (v74) using the annmap Bioconductor package [43] and expression for each gene was obtained by calculating median expression levels of all probesets mapped to the gene. For each HK-lincRNA, we identified the most significantly correlated (positive and negative) protein-coding genes (significant pearson correlation with FDR cutoff < 0.01). The significant associations were then subjected to Gene Ontology Enrichment Analysis (GOA) using

the topGO [44] tool. The gene ontology terms list was subject to a number of filtration steps. Gene ontology terms with less than 5 or greater than 1000 genes were filtered out. For ease of analysis, the significant gene ontology terms list was further cut-down to retain only non-redundant by selecting one among many similar GO terms. GO term similarity was calculated using GOSemSim package in R.

2.4.6 Analysis of TCGA data

Aligned expression data from patients with matched normal and tumour samples were obtained from TCGA for 13 different tumour types (BLCA[45] – Bladder Urothelial Carcinoma, BRCA[46] – Breast invasive carcinoma, COAD[47] – Colon Adenocarcinoma, HNSC[48] – Head and Neck squamous cell carcinoma, KICH[49] – Kidney Chromophobe, KIRC[50] – Kidney renal clear cell carcinoma, KIRP[51] – Kidney renal papillary cell carcinoma, LIHC – Liver Hepatocellular Carcinoma, LUAD[52] – Lung adenocarcinoma, LUSC[53] – Lung squamous cell carcinoma, PRAD[54] - Prostate Adenocarcinoma, THCA[55] – Thyroid carcinoma, UCEC[56] - Uterine Corpus Endometrial Carcinoma). The data was used to estimate gene and transcript abundance based on human genome annotations in Ensembl (v74). Differential expression analysis was performed using Cuffdiff. Genes/transcripts were called differentially expressed if they showed 2-fold change in expression between the normal and tumour samples and q-value was less than 0.05.

Acknowledgements

This work was funded by Cancer Research UK (Grant number: C5759/A12328).

Author Contributions

DM performed the data analysis with contributions from CSS. JB performed the bench experiments. DM and CJM wrote the manuscript with assistance from all the authors.

Data Access

Permission was obtained to use the raw data from TCGA under the project #8211: "Identification of non-coding tumour suppressors and oncogenes".

Conflict of Interest

The authors declare that they have no conflict of interest.

2.5 References

1. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
2. Bitton, D. A., Smith, D. L., Connolly, Y., Scutt, P. J. & Miller, C. J. An integrated mass-spectrometry pipeline identifies novel protein coding-regions in the human genome. *PLoS ONE* **5,** e8949 (2010).
3. Gascoigne, D. K. *et al.* Pinstripe: a suite of programs for integrating transcriptomic and proteomic datasets identifies novel proteins and improves differentiation of protein-coding and non-coding genes. *Bioinformatics* **28,** 3042–3050 (2012).
4. Flicek, P. *et al.* Ensembl 2014. *Nucleic Acids Research* **42,** D749–55 (2014).
5. Managadze, D. *et al.* The vast, conserved mammalian lincRNome. *PLoS Comp Biol* **9,** e1002917 (2013).
6. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458,** 223–227 (2009).
7. Ponjavic, J., Oliver, P. L., Lunter, G. & Ponting, C. P. Genomic and Transcriptional Co-Localization of Protein-Coding and Long Non-Coding RNA Pairs in the Developing Brain. *PLoS Genet* **5,** e1000617 (2009).
8. Leong, H. S. *et al.* A global non-coding RNA system modulates fission yeast protein levels in response to stress. *Nature Communications* **5,** 1–10 (2014).
9. Bitton, D. A. *et al.* Programmed fluctuations in sense/antisense transcript ratios drive sexual differentiation in S. pombe. *Mol. Syst. Biol.* **7,** 559 (2011).
10. Kelley, D. & Rinn, J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol* **13,** R107 (2012).
11. Ulitsky, I. & Bartel, D. P. lincRNAs: genomics, evolution, and mechanisms. *Cell* **154,** 26–46 (2013).
12. Gong, C. & Maquat, L. E. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3′ UTRs via Alu elements. *Nature* **470,** 284–288 (2011).
13. Wang, J., Gong, C. & Maquat, L. E. Control of myogenesis by rodent SINE-containing lncRNAs. *Genes & Development* **27,** 793–804 (2013).
14. Managadze, D., Rogozin, I. B., Chernikova, D., Shabalina, S. A. & Koonin, E. V. Negative correlation between expression level and evolutionary rate of long intergenic noncoding RNAs. *Genome Biology and Evolution* **3,** 1390–1404 (2011).
15. Georgi, B., Voight, B. F. & Bućan, M. From mouse to human: evolutionary genomics analysis of human orthologs of essential genes. *PLoS Genet* **9,** e1003484 (2013).
16. Tu, Z. *et al.* Further understanding human disease genes by comparing with housekeeping genes and other genes. *BMC Genomics* **7,** 31 (2006).
17. Chang, C.-W. *et al.* Identification of human housekeeping genes and tissue-selective genes by microarray meta-analysis. *PLoS ONE* **6,** e22859 (2011).
18. Zhu, J. *et al.* On the nature of human housekeeping genes. *Trends Genet.* **24,** 481–484 (2008).
19. Lercher, M. J., Urrutia, A. O. & Hurst, L. D. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat Genet* **31,** 180–183 (2002).
20. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends*

Genet. **29,** 569–574 (2013).
21. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes are compact. *Trends Genet.* **19,** 362–365 (2003).
22. De Ferrari, L., De Ferrari, L., Aitken, S. & Aitken, S. Mining housekeeping genes with a Naive Bayes classifier. *BMC Genomics* **7,** 277 (2006).
23. Farré, D. *et al.* Housekeeping genes tend to show reduced upstream sequence conservation. *Genome Biol* **8,** R140 (2007).
24. Vinogradov, A. E. Compactness of human housekeeping genes: selection for economy or genomic design? *Trends Genet.* **20,** 248–253 (2004).
25. Zhang, L., Zhang, L., Li, W.-H. & Li, W.-H. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Molecular Biology and Evolution* **21,** 236–239 (2004).
26. Bond, C. S. & Fox, A. H. Paraspeckles: nuclear bodies built on long noncoding RNA. *J. Cell Biol.* **186,** 637–644 (2009).
27. Naganuma, T. & Hirose, T. Paraspeckle formation during the biogenesis of long non-coding RNAs. *RNA Biol* **10,** 456–461 (2013).
28. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research* **20,** 110–121 (2010).
29. Kapusta, A. *et al.* Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet* **9,** e1003470 (2013).
30. Athanasiadis, A., Rich, A. & Maas, S. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. *PLoS Biol.* **2,** e391 (2004).
31. Ramaswami, G. & Li, J. B. RADAR: a rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Research* **42,** D109–13 (2014).
32. Mercer, T. R. & Mattick, J. S. Structure and function of long noncoding RNAs in epigenetic regulation. *Nat. Struct. Mol. Biol.* **20,** 300–307 (2013).
33. Bonnet, E. *et al.* Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics* **20,** 2911–2917 (2004).
34. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489,** 75–82 (2012).
35. Hansen, R. S. *et al.* Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl. Acad. Sci. U.S.A.* **107,** 139–144 (2010).
36. Merkin, J., Russell, C., Chen, P. & Burge, C. B. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science* **338,** 1593–1599 (2012).
37. Tian, D., Sun, S. & Lee, J. T. The Long Noncoding RNA, Jpx, Is a Molecular Switch for X Chromosome Inactivation. *Cell* **143,** 390–403 (2010).
38. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463,** 899–905 (2010).
39. Ning, S. *et al.* Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers. *Nucleic Acids Research* (2015). doi:10.1093/nar/gkv1094
40. Park, D., Park, J., Park, S. G., Park, T. & Choi, S. S. Analysis of human disease genes in the context of gene essentiality. *Genomics* **92,** 414–418 (2008).
41. Khuri, S. & Wuchty, S. Essentiality and centrality in protein interaction networks revisited. *BMC Bioinformatics* **16,** 109 (2015).
42. Hart, T. *et al.* Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. *Mol. Syst. Biol.* **10,** 733–733 (2014).
43. Yates, T., Okoniewski, M. J. & Miller, C. J. X:Map: annotation and visualization of genome structure for Affymetrix exon array analysis. *Nucleic Acids Research* **36,** D780–6 (2008).

44. Alexa, A. *et al.* topGO: Enrichment analysis for Gene Ontology. *Bioconductor package version 2.6.0. http://www.bioconductor.org/packages/release/bioc/html/topGO.html* (2010).
45. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* **507,** 315–322 (2014).
46. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).
47. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487,** 330–337 (2012).
48. Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517,** 576–582 (2015).
49. Davis, C. F. *et al.* The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer Cell* **26,** 319–330 (2014).
50. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499,** 43–49 (2013).
51. Linehan, W. M. *et al.* Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma. *N. Engl. J. Med.* (2015). doi:10.1056/NEJMoa1505917
52. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511,** 543–550 (2014).
53. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489,** 519–525 (2012).
54. Cancer Genome Atlas Research Network. Electronic address: schultz@cbio.mskcc.orgCancer Genome Atlas Research Network. The Molecular Taxonomy of Primary Prostate Cancer. *Cell* **163,** 1011–1025 (2015).
55. Cancer Genome Atlas Research Network. Integrated genomic characterization of papillary thyroid carcinoma. *Cell* **159,** 676–690 (2014).
56. Cancer Genome Atlas Research Network *et al.* Integrated genomic characterization of endometrial carcinoma. *Nature* **497,** 67–73 (2013).

Chapter 3. Global transcriptomic changes in response to hypoxia

Hypoxia driven splicing into non-coding isoforms regulates the DNA damage response

[1]Danish Memon, [1]Keren Dawson, [2]Christopher SF Smowton, [2]Wei Xing, [3]Caroline Dive, [1]Crispin J Miller*

[1]RNA Biology Group, [2]Scientific Computing Team, [3]Clinical and Experimental Pharmacology Group, CRUK Manchester Institute, Manchester M20 4BX UK
*corresponding author: crispin.miller@cruk.manchester.ac.uk

3.1 Introduction

Tumour hypoxia is associated with poor patient outcome and resistance to therapy. It is also associated with a rapid decline in protein production mediated through changes in gene expression. By performing sample specific *de novo* annotation of RNA sequencing data generated in a timecourse of reduced oxygenation, we were able to identify hundreds of novel splicing events occurring in response to hypoxia. ~350 genes switched between coding and non-coding isoforms, including multiple components of the DNA damage response pathway. Notably, HDAC6, a master regulator of the cytotoxic response, and TP53BP1, which sits at the nexus of the double strand break repair pathway, both underwent a marked transition towards an intron-retention pattern with a concomitant decline in protein levels. These transitions from coding to non-coding isoforms were recapitulated in a large cohort of 499 colorectal samples taken from The Cancer Genome Atlas (TCGA). The set of altered genes was enriched for multiple components of the Fanconi Anemia, nucleotide excision and double strand break repair pathways, together forming a strong signature of tumour status at last contact. Together these data demonstrate a new role for hypoxia-driven alternative splicing in regulating DNA damage response.

Hypoxia occurs within the majority of solid tumours and is associated with poor patient outcome and chemo- and radioresistance[1,2]. Hypoxia arises both because disorganization within tumour microvasculature lengthens intracapillary distances

beyond the diffusion range of oxygen and because transient disruptions to blood flow provoke periods of acute oxygen starvation. Hypoxia has multiple impacts on tumour biology including selection of altered cell signaling, angiogenesis, vasculogenesis, changes in central metabolism, suppression of immune reactivity, enhanced receptor tyrosine kinase signaling and down regulation of DNA repair pathways, promotion of pro-survival phenotypes and increased proclivity for invasion and metastasis[3,4]; extensively reviewed in[5,6]. Many of these hypoxia responses are characterized by widespread alterations in transcription profiles driven largely (but not exclusively) by stabilization of the transcription factor subunit hypoxia inducible factor 1α (HIF1A)[7,8]. Hypoxia mediated transcriptional regulation is also controlled by other factors including HIF2A[9] and HIF3A[10]. In addition, signaling through both the growth factor receptor pathways (phosphatidylinositol 3-kinase; PI3K, ERK) and energy depletion pathways (5-AMP-activated protein kinase; AMPK) converge on the tuberous sclerosis complex (TSC1/2) leading to complex patterns of spatial and temporal regulation in response to stress. These signals feed in to mechanistic target of rapamycin (mTOR), which, in the context of hypoxia, leads to the rapid suppression of protein synthesis[11], presumably in order to conserve energy[12]. Levels of hypoxia vary between and within tumours, correlate with patient outcomes, and can lead to differences in response to therapy[13]. A better understanding of heterogeneity in hypoxia-driven changes in gene expression will therefore inform strategies for precision medicine[14], raising the need for reliable biomarkers of tumour hypoxia. To this end, a number of groups have developed multiplex gene expression signatures with the intention of better reflecting the multiplicity of pathways involved in the hypoxic response, e.g.[13,15,16].

In recent years, advances in expression profiling have revealed substantial levels of alternative splicing within the human genome, such that the majority of protein-coding genes are now known to express multiple isoforms (median isoform count per protein coding gene in ENSEMBL 74: 5), of which 44% are annotated as non-coding (63,816: “non-coding”; 81,715: “protein-coding”). Despite their prevalence, the majority of these transcripts have yet to be characterized, raising the question of how much of this “dark matter” is functional, and how much is simply a consequence of aberrant splicing and a passive by-product of gene expression.

Given the widespread alterations in transcript expression that arise in response to changes in oxygen levels, we speculated that similar systematic alterations in splicing might add further levels of transcriptional control. We therefore exploited the increased

precision offered by RNA sequencing to investigate how hypoxia affects alternative splicing, since earlier studies using 3′ IVT arrays were not able to characterize the transcriptome at this level of precision. We used *de novo* sample specific annotation strategies to investigate changes in exon structure alongside multiple events including intron retention and alterations to the 5′ and 3′ boundaries of many genes. We applied these novel annotation approaches both to a time-course of colorectal cancer cells in reduced oxygen and to re-analyze a large cohort of colorectal samples from TCGA.

3.2 Results

3.2.1 Systematic re-splicing of the transcriptome in response to hypoxia

RNA deep sequencing over a time-course of enforced hypoxia was used to identify changes in coding and non-coding RNA levels in HCT116 cells. Cells were harvested at 0, 1, 2, 24 hours in reduced oxygen (1% O_2) and poly(A) RNA sequenced using 100mer paired end strand specific Illumina sequencing. Data were aligned and exon structure determined using a multistage pipeline including DEXSeq[17], Mapsplice[18] and Cufflinks[19,20] to identify significant events such as exon skipping and intron retention. This pipeline generated an augmented catalogue of transcripts in which novel isoforms identified by Cufflinks were amalgamated with existing annotations from ENSEMBL, before splicing changes were identified through Multivariate Analysis of Transcript Splicing (MATS)[21].

Of the 53,936 transcripts identified, 52,733 mapped to 15,334 genes in ENSEMBL (v74)[22] while 1,203 (1,155 genes) were unannotated (Supplementary Table 3.1; Supplementary Figure 3.1). In keeping with previous reports[15], 12% of protein-coding genes exhibited gene-level changes within the timecourse (N=2,387), and included multiple genes associated with hypoxia, glycolysis and MAPK signaling (Supplementary Table 3.2). At the isoform level, 9,222 (60%) of genes were present in at least 2 isoforms, of which 4,982 (54%) were novel. Multiple changes to exons within the coding sequence were identified using Cuffdiff. These included multiple kinases (CLK4, MARK4, ACVR2B, MAP3K3, MAP3K8, STK32C and SGK494) and 2 phosphatases (PPP5C, PPP3CB).

Splicing events were frequent and diverse. 29% (N=869/2,949) of spliced genes were predicted by MATS to exhibit multiple concurrent modifications (Fig 3.1A), including a substantial increase in the number of retained introns, increased exon skipping,

increased usage of non-canonical 5′ and 3′ splice sites (Fig 3.1B; Supplementary Table 3.3), and a global shift towards expression of retrained intron transcripts (Fig 3.1C). Genes involved in exon skipping and inclusion exhibited high enrichment for *Cell cycle* (GO:0007049) and *RNA splicing* (GO:0008380). The *RNA splicing* pathway was also enriched for genes that exhibited increased intron retention in hypoxia, suggestive of a feedback loop in which components of the alternative splicing machinery are themselves regulated in part by changes to splicing patterns. Importantly, genes associated with *Response to DNA Damage Stimulus* and *DNA repair pathway* also exhibited increased intron retention (Benjamini & Hochberg (BH) corrected *p*-value < 0.05).

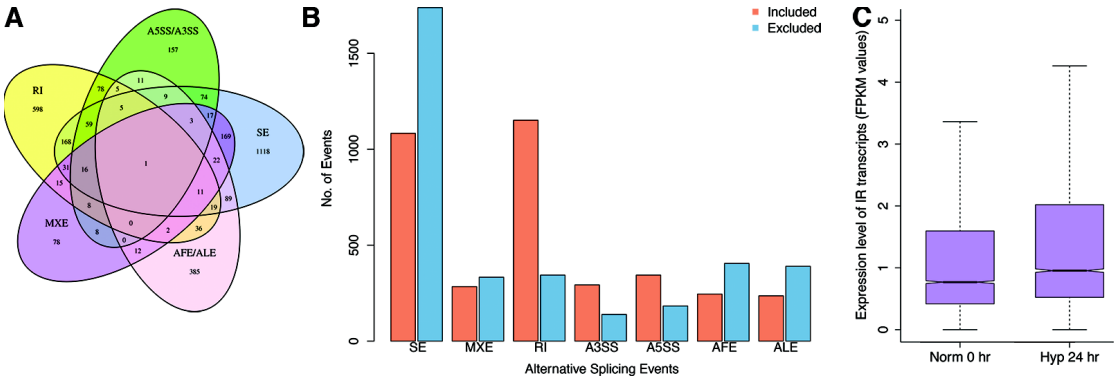

Figure 3.1 Isoform switching in hypoxia increases the abundance of unproductive transcripts. (A) Multiple splicing events affect the majority of alternatively spliced genes. SE: Spliced Exon; MXE: mutually exclusive exons, RI: retained intron; A3SS: alternative 3′ start site; A5SS alternative 5′ start site; AFE alternative first exon; ALE alternative last exon. (B) Inclusion or exclusion of exons in response to hypoxia, detected by MATS. Categories as A. (C) Overall expression levels of genes that switch between 'protein-coding' and 'retained-intron' major isoform between normoxia and hypoxia.

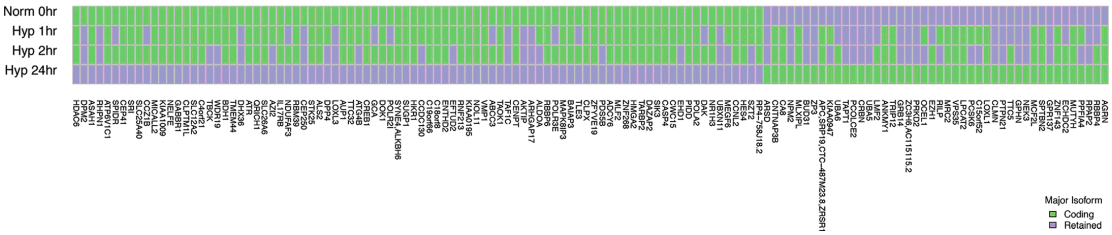

Figure 3.2 Genes switching to or from a retained intron major isoform in response to hypoxia.

3.2.2 Confirmation that retained intron expression regulates expression of HDAC6 and TP53BP1 protein levels

In addition to changes between protein coding isoforms (Supplementary Figure 3.2), 29% of genes switched from a protein-coding to non-coding major-isoform, with the majority of events (199/343) leading to the expression of a non-coding isoform as the primary transcript under hypoxia (Fig 3.2). Of these, most changes were driven by intron retention (88/132) and were consistent with an overall increase in the expression of retained-intron transcripts after 24hrs in hypoxia (p-value < 0.001; Fig 3.1B, C).

We hypothesized that this might provide a previously unreported mechanism by which cells could modulate protein levels in response to hypoxia, and therefore sought to confirm that a switch towards a non-coding isoform was indeed associated with altered protein levels. Two notable genes with substantial changes in intron retention were HDAC6 and TP53BP1 (Fig 3.3A,D). HDAC6 is a class IIb histone deacetylase with an unusually diverse set of substrates that include multiple cytosolic proteins such as HSP90. HDAC6 inhibition has been associated with the processing of protein aggregates, the misfolded protein stress response pathway, and more generally as a master regulator of cytotoxic stress[27,28], including its involvement in the ubiquitination and deacetylation of the mismatch repair (MMR) protein MutS protein homolog 2 (MSH2)[23]. Levels of intron expression at the locus increased in hypoxic conditions and were accompanied by a significant decline in protein levels (Fig 3.3B,C). TP53BP1 also exhibited an increase in intron retention and an associated decline in protein levels (Fig 3.3E,F). It sits at the nexus of the Double Strand Break (DSB) repair pathway where it performs a number of roles, including binding to P53, leading to enhanced transactivation and increased levels of P21[24], interacting with chromatin to promote DNA repair[25], and recognizing H4K20me2 and H2AK15ub histone marks arising from DSB signaling[26]. TP53BP1 has also been shown to cooperate with RIF1 and MAD2L2 (Rev7)[27] to modulate Non Homologous End Joining (NHEJ) and genetic stability. These data are therefore particularly interesting in the context of a ‘mutator phenotype’ in which a shift to increased genetic instability is postulated to promote clonal diversity within a tumour[28].

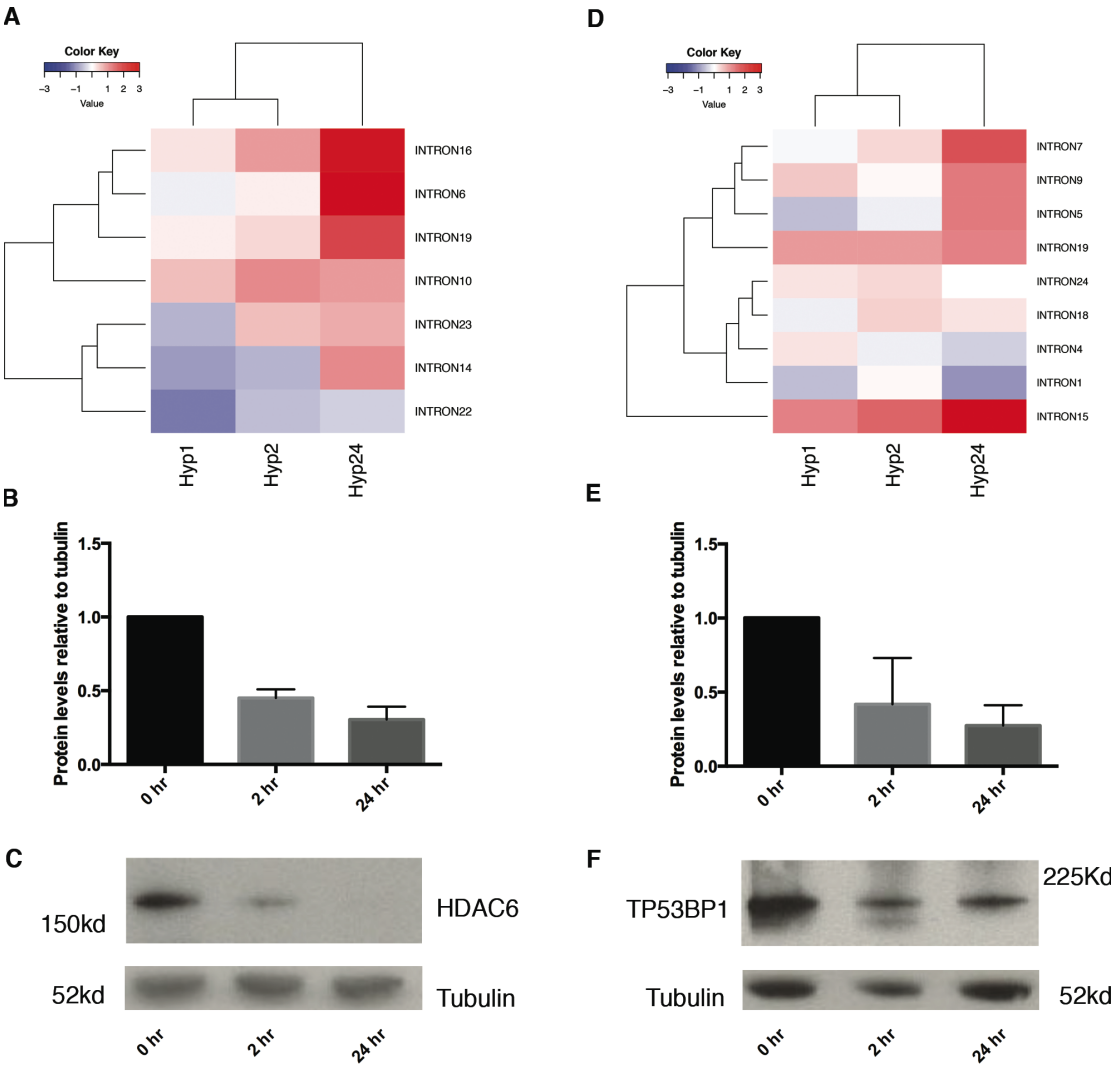

Figure 3.3 Changes in HDAC6 and TP53BP1 transcript and protein levels in response to hypoxia. (A) Expression across introns in HDAC6 at 0, 2, 24 hours following a shift to reduced oxygen (1%). Colour represents $\log_2$ fold change. (B) HDAC6 protein levels as determined by Western blot (C). (D-F) TP53BP1 exhibits similar patterns of expression.

3.2.3 Colorectal tumours express high proportion of unproductive transcripts

Having identified a new role for alternative splicing in regulating protein levels in response to hypoxia, we asked whether similar changes were observed in human tumours. We analyzed RNA-seq data derived from 458 colorectal tumour samples and 41 normal tissues obtained from The Cancer Genome Atlas (TCGA)[29]. The samples included AJCC pathologic staging (Stage I - Stage IVB), metastatic staging (M0, M1 and MX) and therefore included both patients with early-localized disease as well as metastasis. No information was available about previous treatment with radiotherapy or chemotherapy for the majority of patients. In total 4303 genes were found to switch

between coding and non-coding major isoforms in at least 5% of samples (Fig 3.4A; Supplementary Table 3.4). Unsupervised clustering of these data by major isoform coding/non-coding status revealed three major clusters (Fig 3.4A). The proportion of non-coding major-isoforms correlated well with available relapse data, with tumours from patients who were subsequently defined as “tumor free” at last contact expressing fewer non-coding major-isoforms than those defined as “with tumor” (tumor-free: 916; with-tumor: 1945; *p*-value < 0.01; Fig 3.4B). As expected, many of these also switched between coding and non-coding isoforms in the HCT116 cell line data (*p-value* < 0.01; hypergeometric test), suggesting that a significant proportion of these changes are driven by changes in oxygenation status within the tumours.

750 genes changed in major-isoform status from coding to non-coding between these two classes (Fig 3.4C; Supplementary Table 3.4), however switching to a non-coding isoform was not significantly associated with a global change in overall transcription levels associated with these loci (Fig 3.4D). These data highlight the importance of considering the consequences of splicing when seeking expression signatures using transcription data.

Finally we asked whether these splicing changes in tumour samples were associated with specific pathways, and found significant enrichment for DNA damage and repair pathways, as well as alternative splicing (“Response to DNA Damage Stimulus”, “DNA Repair” and “RNA Splicing” Gene Ontology (GO) categories; BH corrected *p*-value < 0.01; Fig 3.4E). A significant proportion of these loci were also known downstream targets of TP53 (Ingenuity; IPA; *p*-value of overlap: $< 10^{-6}$).

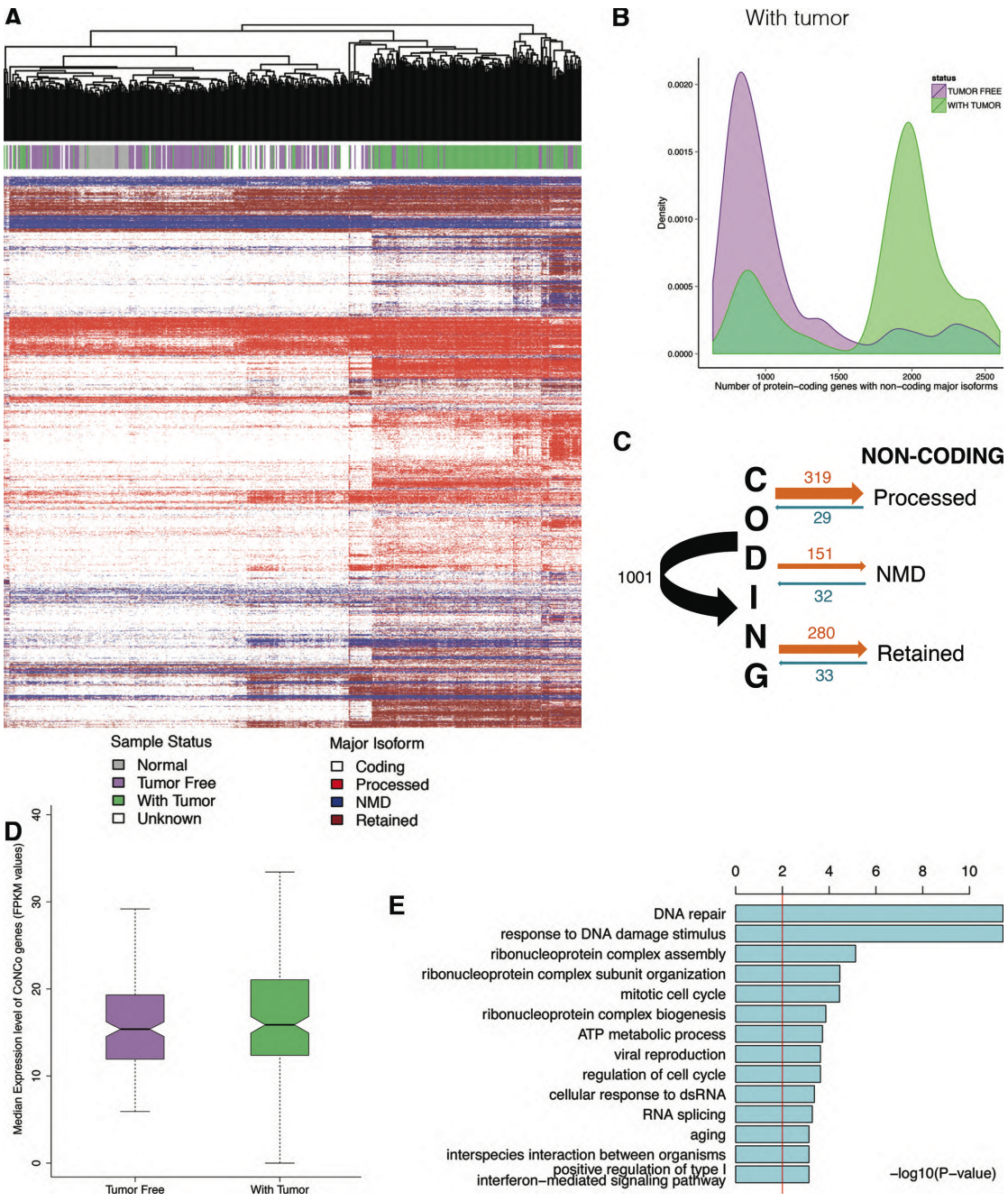

Figure 3.4 Switch from coding to non-coding transcripts is a signature of colorectal tumours. (A) Change in major isoform class between 'tumor-free' and 'with-tumor' samples in TCGA colorectal carcinomas; Orange: enriched in 'with-tumor' samples. (B) Unsupervised clustering of colorectal tumour samples based on major isoform type: Processed – "processed transcript", NMD – "nonsense mediated decay", RI – "Retained Intron"). (C) Proportion of non-coding major isoforms detected across colorectal tumor samples stratified by patient status at last contact: "tumor free" (purple) and "with tumor" (green). Data from TCGA. (D) Average expression level of genes changing from coding to non-coding major-isoform (CoNCo) stratified by patient status on last contact. (E) Gene ontology terms found enriched among CoNCo genes.

3.3 Discussion

Although hypoxia-dependent splicing changes have been reported for individual loci[30-32], there has been no evidence on how widely applicable the phenomenon may be[33]. Here we identified a comprehensive remodeling of transcript structures in response to hypoxia, encompassing significant changes in the levels of 12% of all protein-coding transcripts, and a switch to a different major isoform at 7% of all protein-coding loci. We observed similar changes in colorectal carcinomas and a widespread switch to non-coding isoforms that correlated strongly with tumour status at last contact. Many of these changes involved expression of novel transcripts not present in the ENSEMBL reference annotation database.

While splicing to remodel the proteome is well understood to be a critical process through which cells achieve increased protein diversity[34], we have identified a novel role for splicing in which switching between non-coding isoforms modulates overall protein output from a locus. Hypoxia-dependent changes were biased towards loss of the protein-coding isoform under hypoxia (Figures 3.2D, 3.3) is consistent with the rapid decline in protein synthesis that accompanies a shift to hypoxic conditions. Importantly, loci that changed in this way were significantly enriched for specific pathways, including those that respond to DNA damage. The effects observed are, therefore, not simply the result of overall loss of fidelity in the splicing machinery, but rather a consequence of its coordinated reprogramming. Together, these data indicate the presence of a tightly regulated pathway, and are therefore important in the context of recent reports that somatic Single Nucleotide Variants (SNVs) can result in aberrant splicing patterns, including intron retention, that deactivate tumour suppressor genes[35]. Our data indicate that these splicing patterns may occur naturally as part of normal regulatory processes, and that SNVs provide a mechanism by which cancer cells can hijack these pathways to subvert normal mechanisms of control. This transformation of the coding competence of the transcriptome occurs in parallel with changes to other regulatory pathways including epigenetic modifiers and multiple kinases, thus providing further opportunities to rewire cancer-signaling pathways[36].

Both HDAC6 and TP53BP1 exhibited a significant decline in protein level concomitant with increased intron retention. These data, together with coding-noncoding splicing changes to multiple key components of the Fanconi Anemia pathway (MU31, FANCB, FANCG, FANCM, ATRIP, POLG, RPA1), the Nucleotide excision repair pathway (AQR, PLD1, ERCC2, LIG1, RPA1, CCNH, RNF11, ACTL6A, RFC1, REV1, ACRT8,

GTF2H2), and the double strand break repair pathway (RAD9, WRN, SMARCA5, UIMC1, RFC1, MUS81, POLD1, ATRIP, CHEK1, CLSPN, BABM1, TP53BP1, RPA1, REV1, CDK2, RAD52) together reveal a hitherto unanticipated role for alternative splicing in modulating the DNA damage response through action both at critical regulators (e.g. HDAC6, TP53BP1; Figure 3.3 and the G2-M checkpoint kinase CHEK1) and by modulating the level of proteins throughout the pathway.

Hypoxia-dependent increases in genetic instability have previously been suggested as a driver of a 'mutator phenotype'[37], in which elevated mutation rates increase clonal diversity, leading to a greater likelihood of the expression of a clone with genetic changes that confer a proliferative advantage[28]. Taken together, our data suggest that a shift to non-coding expression in response to hypoxia may provide a novel, important, and unanticipated mechanisms by which these processes are mediated.

Our reanalysis of TCGA data also revealed a striking signature of outcome in which a switch to a non-coding major isoform at multiple protein coding loci is associated with patients annotated as 'with tumor at last contact'. Importantly, these changes are not detected using gene-level expression summaries since overall RNA levels from these loci do not change substantially. Our findings have clear implications for the analysis of expression data and the development of RNA-based signatures and biomarkers in both diagnostics and personalized medicine.

3.4 Experimental Procedures

3.4.1 Cell culture

HCT116 were cultured in RPMI-1640 media (Life Technologies) supplemented with 10% FBS (Biowest). All cells were maintained in a humidified atmosphere at 37°C and 5% CO_2. For hypoxia treatment, the HCT116 cell line was cultured in 1% O_2 in an Invivo$_2$ hypoxia workstation 4000 (Biotrace, Fred Baker Ltd.) for the given time course 24 hours after plating.

3.4.2 Protein extraction and western blotting

Protein was extracted by washing cells in ice cold PBS and scraping cells in ice cold cell lysis buffer (9803s New England Biolabs) supplemented with PMSF (Sigma 93482)

and protease inhibitors (Roche Diagnostics complete edta free 11 873 580 001). The sample was centrifuged at 4^0C 13,000rpm for 10mins and supernatent kept. 50mg of total protein per sample was resolved by SDS-PAGE 10 % NuPage gels (Invitrogen) and transferred electrophoretically to Immobilon-P™ (Millipore). The membrane was blocked in 5%milk PBS-T for 30 mins and blotted overnight with HDAC6 antibody (1/1000 NEB 7558S), TP53BP1 (1/1000 AT4311a Generon mouse monoclonal antibody), or tubulin antibody (1/5000 Sigma T6199). Detection was performed using a peroxidase-conjugated anti-rabbit or anti-mouse IgG (Amersham Biosciences Pharmacia) and chemiluminescence visualization (ECL+, Amersham Biosciences) was used according to the manufacturer's instructions. Quantification of Western blot signals was performed using the Chemi Genius Bioimaging system (Syngene) and the Chemi genius gel documentation and analysis system.

3.4.3 RNA extraction and construction of sequencing libraries

RNA was extracted using the Qiagen Qiashredder kit (79654) and the Qiagen RNeasy Mini Kit (74104) as per the manufacturer's instructions. The RNA was DNase treated following the protocol in the RNeasy Mini Kit with Qiagen RNase-free DNase I (79254). Indexed PolyA libraries were prepared using 1ug of Total RNA and 13 cycles of amplification in the NEB Next Ultra Directional RNA Library Prep Kit for Illumina (New England Biolabs Inc. Cat No: E7420S). Libraries were quantified by qPCR using a Kapa Library Quantification Kit for Illumina sequencing platforms (Kapa Biosystems Inc. Cat No: KK4835). Pooled libraries were clustered at 15pM on the cBot and 2 x 100bp sequencing was carried out using the High Throughput mode of a HiSeq 2500 using TruSeq SBS Kit v3 chemistry (Illumina inc.)

3.4.4 Data analysis

All statistical analysis including t-tests and Wilcoxon's tests, were performed in R. Time-course data: 100mer paired reads were aligned to the human genome (hg19) using Mapsplice (v2.1.4)[18]. An average of 43.8M (27.5M-58.2M) read-pairs per sample mapped to the genome in the correct orientation and appropriately spaced, corresponding to ~90% of the total reads sequenced. Transcript models were derived for each sample independently using Cufflinks (v2.2.0; with default parameters, except to specify strand specificity). Resultant models were then merged using Cuffmerge to provide a global model and to classify transcripts as novel, or known, when they mapped to ENSEMBL (v74). Resultant gene models were then filtered to keep

transcripts where an exon junction was supported by at least 2 reads in at least two samples, and Fragments Per Kilobase Mapped (FPKM) greater than 0.5 in at least three samples. These data were classified according to transcript type and provided in the Supplementary Table 1. The remaining (53,936) transcripts (15,334 genes) were used to obtain gene level counts using the RsubRead package[38] in R and supplied to edgeR[39] to call differential expression (absolute fold-change > 2 relative to 0 hours; False Discovery Rate (FDR) < 1%) at each time point. Annotation was supplied by the Bioconductor package annmap[40]. For each gene, the major isoform was defined as the transcript with the highest median expression across replicates.

3.4.5 Gene Ontology Enrichment Analysis

Functional enrichment analysis was performed using the Goseq package[41] in R to identify statistically enriched gene ontology (GO) terms[42] (Hypergeometric test: BH corrected *P* - value < 0.01). Non-redundant GO terms were obtained by retaining only one representative term from GO with high semantic similarity, derived using GOSemSim package[43] in R.

3.4.6 Detection of Alternative Splicing

Alternative splicing was detected at both transcript level and exon level. Differential splicing at the exon level was performed for 24 hour samples (in hypoxia) versus 0 hours using the DEXSeq package in R. DEXSeq uses transcript annotations of each gene to split them non-overlapping bins (exon units) and then uses Generalized Linear Models (GLMs) to identify differentially used exon units. Only exon units that were at least 50 bp long, with mean exonic counts greater than 1 and satisfied a FDR cutoff of 5% were considered as differentially used. Differential splicing at the transcript level was performed for 24-hour samples (in hypoxia) versus 0 hours using the Cuffdiff program within Cufflinks. In addition to detecting differential splicing, Cuffdiff also reports statistically significant differential promoter usage and differential coding sequence usage. Statistically significant changes were detected at a FDR cutoff of 5%. Following this, the alternative splicing events were annotated into 5 major categories, Exon Skipping (SE), Mutually Exclusive Exons (MXE), Intron Retention (RI), Alternative 3′ Splice Sites (A3SS) and Alternative 5′ Splice Sites (A5SS), using MATS (3.0.8). The splicing events from MATS were detected using both read information and exon junction data and filtered for statistically significant events (FDR < 1%).

3.4.7 Analysis of TCGA Data

Splice-aware aligned RNA-Seq data (BAM files) of Colorectal cancer cohort (COAD)[29] comprising of 458 tumour samples and 41 normal samples along with corresponding clinical data was obtained from TCGA. Gene and transcripts quantification and normalization was performed using standard Cufflinks pipeline and the annotations in Ensembl (v74).

Acknowledgements

We thank the CRUK MI MCBF for processing the RNA sequencing samples and the Scientific Computing and Computational Biology support teams for maintaining upstream data analysis pipelines and hardware. This work was funded by Cancer Research UK (Grant number: C5759/A12328).

Author Contributions

DM performed the data analysis with contributions from CSS, WX. KD performed the experiments. DM and CJM wrote the manuscript with assistance from all the authors.

Data Access

HCT116 cell line sequencing data have been deposited in the Gene Expression Ombinus (GEO). Permission was obtained to use the raw data from TCGA under the project #8211: "Identification of non-coding tumour suppressors and oncogenes".

Conflict of Interest

The authors declare that they have no conflict of interest.

3.5 References

1. Moulder, J. E. & Rockwell, S. Tumor hypoxia: its impact on cancer therapy. **5,** 313–341 (1987).
2. Teicher, B. A. Hypoxia and drug resistance. **13,** 139–168 (1994).
3. Carmeliet, P. *et al.* Role of HIF-1alpha in hypoxia-mediated apoptosis, cell proliferation and tumour angiogenesis. **394,** 485–490 (1998).
4. Maxwell, P. H. *et al.* Hypoxia-inducible factor-1 modulates gene expression in

solid tumors and influences both angiogenesis and tumor growth. **94,** 8104–8109 (1997).
5. Wilson, W. R. & Hay, M. P. Targeting hypoxia in cancer therapy. **11,** 393–410 (2011).
6. McKeown, S. R. Defining normoxia, physoxia and hypoxia in tumours-implications for treatment response. **87,** 20130676 (2014).
7. Huang, L. E., Gu, J., Schau, M. & Bunn, H. F. Regulation of hypoxia-inducible factor 1alpha is mediated by an O2-dependent degradation domain via the ubiquitin-proteasome pathway. **95,** 7987–7992 (1998).
8. Gordan, J. D. & Simon, M. C. Hypoxia-inducible factors: central regulators of the tumor phenotype. **17,** 71–77 (2007).
9. Makino, Y. *et al.* Inhibitory PAS domain protein is a negative regulator of hypoxia-inducible gene expression. **414,** 550–554 (2001).
10. Tian, H., McKnight, S. L. & Russell, D. W. Endothelial PAS domain protein 1 (EPAS1), a transcription factor selectively expressed in endothelial cells. **11,** 72–82 (1997).
11. Pouysségur, J., Dayan, F. & Mazure, N. M. Hypoxia signalling in cancer and approaches to enforce tumour regression. **441,** 437–443 (2006).
12. Bristow, R. G. & Hill, R. P. Hypoxia and metabolism. Hypoxia, DNA repair and genetic instability. **8,** 180–192 (2008).
13. Eustace, A. *et al.* A 26-gene hypoxia signature predicts benefit from hypoxia-modifying therapy in laryngeal cancer but not bladder cancer. *Clin. Cancer Res.* **19,** 4879–4888 (2013).
14. Bristow, R. G., Berlin, A. & Dal Pra, A. An arranged marriage for precision medicine: hypoxia and genomic assays in localized prostate cancer radiotherapy. **87,** 20130753 (2014).
15. Winter, S. C. *et al.* Relation of a hypoxia metagene derived from head and neck cancer to prognosis of multiple cancers. *Cancer Res.* **67,** 3441–3449 (2007).
16. van Malenstein, H. *et al.* A seven-gene set associated with chronic hypoxia of prognostic importance in hepatocellular carcinoma. **16,** 4278–4288 (2010).
17. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. **22,** 2008–2017 (2012).
18. Wang, K. *et al.* MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. **38,** e178 (2010).
19. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. **28,** 511–515 (2010).
20. Roberts, A., Pimentel, H., Trapnell, C. & Pachter, L. Identification of novel transcripts in annotated genomes using RNA-Seq. **27,** 2325–2329 (2011).
21. Shen, S. *et al.* MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. **40,** e61 (2012).
22. Flicek, P. *et al.* Ensembl 2014. **42,** D749–55 (2014).
23. Zhang, M. *et al.* HDAC6 deacetylates and ubiquitinates MSH2 to maintain proper levels of MutSα. *Mol. Cell* **55,** 31–46 (2014).
24. Iwabuchi, K., Bartel, P. L., Li, B., Marraccino, R. & Fields, S. Two cellular proteins that bind to wild-type but not mutant p53. *Proc. Natl. Acad. Sci. U.S.A.* **91,** 6098–6102 (1994).
25. Dimitrova, N., Chen, Y.-C. M., Spector, D. L. & de Lange, T. 53BP1 promotes non-homologous end joining of telomeres by increasing chromatin mobility. *Nature* **456,** 524–528 (2008).
26. Fradet-Turcotte, A. *et al.* 53BP1 is a reader of the DNA-damage-induced H2A Lys 15 ubiquitin mark. *Nature* **499,** 50–54 (2013).
27. Boersma, V. *et al.* MAD2L2 controls DNA repair at telomeres and DNA breaks by inhibiting 5′ end resection. *Nature* **521,** 537–540 (2015).

28. Loeb, L. A. Human cancers express mutator phenotypes: origin, consequences and targeting. *Nat. Rev. Cancer* **11,** 450–457 (2011).
29. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. **487,** 330–337 (2012).
30. Hirschfeld, M., Hausen, zur, A., Bettendorf, H., Jaeger, M. & Stickeler, E. Alternative Splicing of Cyr61 Is Regulated by Hypoxia and Significantly Changed in Breast Cancer. **69,** 2082–2090 (2009).
31. Sena, J. A., Wang, L., Pawlus, M. R. & Hu, C.-J. HIFs Enhance the Transcriptional Activation and Splicing of Adrenomedullin. **12,** 728–741 (2014).
32. Kemmerer, K. & Weigand, J. E. Hypoxia reduces MAX expression in endothelial cells by unproductive splicing. **588,** 4784–4790 (2014).
33. Weigand, J. E., Boeckel, J.-N., Gellert, P. & Dimmeler, S. Hypoxia-induced alternative splicing in endothelial cells. **7,** e42697–e42697 (2012).
34. Kelemen, O. *et al.* Function of alternative splicing. **514,** 1–30 (2013).
35. Jung, H. *et al.* Intron retention is a widespread mechanism of tumor-suppressor inactivation. **47,** 1242–1248 (2015).
36. Furney, S. J. *et al.* SF3B1 mutations are associated with alternative splicing in uveal melanoma. **3,** 1122–1129 (2013).
37. Luoto, K. R., Kumareswaran, R. & Bristow, R. G. Tumor hypoxia as a driving force in genetic instability. *Genome Integr* **4,** 5 (2013).
38. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. **30,** 923–930 (2014).
39. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. **26,** 139–140 (2010).
40. Yates, T., Okoniewski, M. J. & Miller, C. J. X:Map: annotation and visualization of genome structure for Affymetrix exon array analysis. *Nucleic Acids Res.* **36,** D780–D786 (2008).
41. Young, M. D., Wakefield, M. J., Smyth, G. K. & Oshlack, A. Gene ontology analysis for RNA-seq: accounting for selection bias. **11,** R14 (2010).
42. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. **25,** 25–29 (2000).
43. Yu, G. *et al.* GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. **26,** 976–978 (2010).

Chapter 4. The role of long noncoding RNAs in hypoxia

Comprehensive analysis of the hypoxic transcriptome identifies noncoding RNA HINCR1 as a critical regulator of the hypoxic response

[1]Danish Memon[#], [1]Keren Dawson[#], [2]Yaoyong Li, [3]Shameem Fawdar, [2]Christopher Wirth, [2]Hui Sun Leong, [3]John Brognard, [4]Christopher Morrow, [4]Caroline Dive, [1,2]Crispin J Miller*

[1]RNA Biology Group, CRUK Manchester Institute, Manchester M20 4BX UK

[2]Computational Biology Support Team, CRUK Manchester Institute, Manchester M20 4BX UK

[3]Signalling Networks in Cancer Group, CRUK Manchester Institute, Manchester M20 4BX UK

[4]Clinical and Experimental Pharmacology Group, CRUK Manchester Institute, Manchester M20 4BX UK

[#]equal contribution

*corresponding author: crispin.miller@cruk.manchester.ac.uk

4.1 Introduction

Tumour hypoxia is associated with poor patient outcome and resistance to therapy. It impacts upon multiple pathways and causes alterations in the levels of protein encoding transcripts throughout the cell. Here we show that many noncoding loci are also differentially expressed. One of these loci, HINCR1, is induced within 2 hours in response to hypoxia. HINCR1 was predicted to correlate strongly with hypoxia signatures across a large independent cohort (N=248) of patient derived tumour samples encompassing multiple cancer types. HINCR1 is induced in multiple tumour types and high levels of HINCR1 were found to be prognostic of poor disease-specific survival in lung adenocarcinomas. siRNA mediated knockdown of HINCR1 modulated hypoxia-dependent changes at specific loci throughout the genome, including the critical mitogenic transcription factor Egr1 and a significant number of genes harbouring its binding sequence in their promoter. Widespread binding of HINCR1 on the genome was observed in hypoxia, with significant enrichment at

EGR1-bound sites. Knockdown of HINCR1 led to altered Egr1 binding, correlated with changes in expression at these loci and resulted in attenuation of hypoxia-dependent changes to transcript levels expressed from numerous genes throughout the genome, including Egr1 and its downstream targets. Given the central role of Egr1 in modulating the gene expression programs of mitogenesis and differentiation, these data reveal HINCR1 as a potential therapeutic target in cancer.

While only approximately 1.2% of the human genome encodes amino acids, recent reports suggest that between 70-90% of all nucleotides may be transcribed, resulting in the expression of large numbers of noncoding RNAs (ncRNAs) that are never translated into proteins[1,2]. Although relatively few of these ncRNAs have so far been characterised, a growing subset has been shown to be functional (extensively reviewed in [3,4]). ncRNAs can influence a wide range of processes that regulate gene expression including chromatin modulation and regulation of transcription[5,6], post-transcriptional regulation of transcript stability[7], splicing[8], and, through direct interactions, the modulation of protein function[9]. Many ncRNAs are alternatively spliced, and since many protein-coding genes are expressed in both coding and noncoding isoforms, ncRNAs are also frequently expressed from within protein-coding loci, potentially further increasing the repertoire of functional RNAs. Since the noncoding genome can be subject to the same mutation and selection pressures as coding loci, it may constitute a considerable but largely untapped resource of novel tumour suppressors and oncogenes. We sought to identify novel cancer-associated noncoding RNAs by integrating RNA sequencing data derived from a timecourse of HCT116 cells following a shift to hypoxic conditions with public domain expression data taken from a large cohort of (N=248) human tumours.

Hypoxia occurs within the majority of solid tumours and is associated with poor patient outcome and chemo- and radioresistance[10,11]. It has multiple impacts on tumour biology including selection of pro-survival phenotypes, altered cell signalling, angiogenesis, vasculogenesis, changes in central metabolism, increased proclivity for invasion and metastasis, suppression of immune reactivity, enhanced receptor tyrosine kinase signalling and down regulation of DNA repair pathways[12,13]; extensively reviewed in [14,15]. Levels of hypoxia vary between and within tumours, correlate with patient outcomes, and can lead to differences in response to therapy[16]. These responses to changes in oxygen levels are associated with genome-wide alterations in transcription

profiles driven largely (but not exclusively) by stabilisation of the transcription factor subunit hypoxia inducible factor 1A (HIF1A)[17,18], HIF2A[19], HIF3A[20], and the extracellular signal-regulated kinase (ERK) induced transcription factor early growth response 1 (Egr1)[21].

Given the central role played by transcription in regulating tumour hypoxia, we asked whether noncoding RNA expression might contribute to these alterations in gene expression programmes.

4.2 Results

4.2.1 HINCR1 is a novel hypoxia responsive noncoding RNA

We recently used a novel annotation pipeline to perform *de novo* annotation and analysis of a hypoxia timecourse RNA-Seq data (0, 1, 2, 24 hours) derived from HCT116 cell line. In addition to changes in expression and splicing at protein-coding loci[1], we identified multiple noncoding transcripts, of which 3294 matched existing annotations in ENSEMBL (v74; 'lincRNAs', 'pseudogenes', 'processed transcripts' and 'antisense'). Also, we identified a further 1155 novel noncoding genes at unannotated loci, expressing 1203 transcripts, of which 139 (100 genes) were multi-exonic. A small subset of these (25/100) were predicted to have high coding potential, raising the possibility that they may express novel proteins (Supplementary Table 4.1). Although unannotated, 23 of the 100 novel genes were also detected in ENCODE Caltech RNA-seq data for the same cell line (HCT116)[22], but not in the 13 other cell lines in this set, indicating that many of these transcripts are hypoxia-, tissue-, or cell-line specific (Supplementary Figure 4.1).

685/3294 noncoding genes were found to have altered levels (absolute Fold Change > 2; False Discovery Rate (FDR) < 1%). While the majority of have no prior association with hypoxia, both H19 and UCA1 were induced as expected, confirming previous reports[9,22-25]. In total 80 lincRNAs were differentially expressed in hypoxia, of which 59 were induced (Fig 4.1A). We refer to these as HINCRs (Hypoxia Induced Non-Coding RNAs). Only three of these transcripts (MIR210HG, HINCR1, RMRP) were significantly upregulated at early timepoints (Supplementary Table 4.2), mirroring changes at protein-coding loci, which were also at their most extensive after 24 hours.

[1] Chapter 3 in this thesis

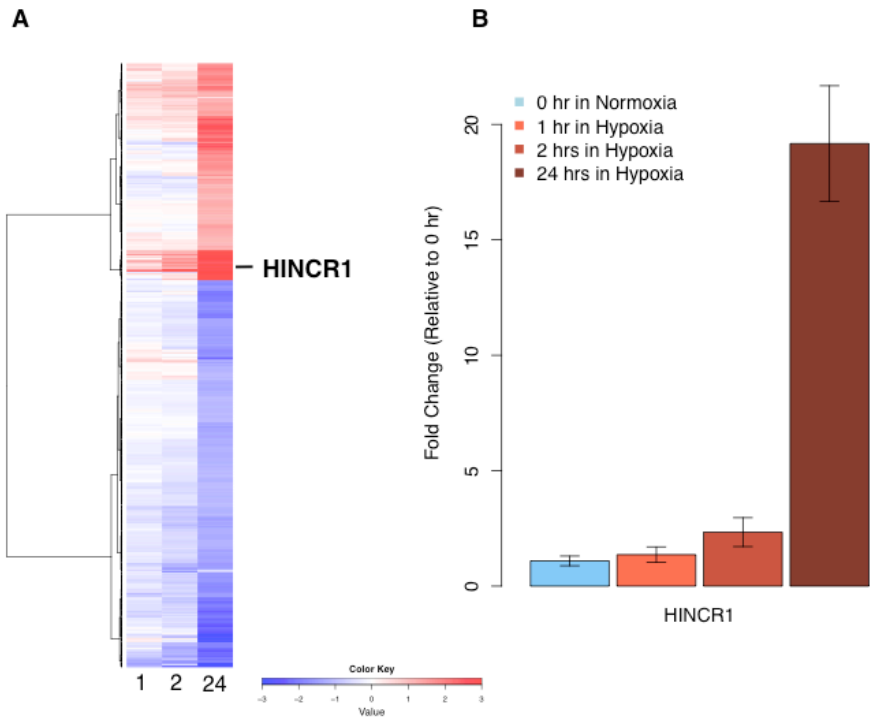

Figure 4.1 Long noncoding RNA HINCR1 is induced in response to hypoxia. (A) Hypoxia-dependent differentially expressed long noncoding RNAs at 1, 2, or 24 hrs relative to the 0 hr timepoint. Rows represent noncoding RNAs, columns ordered by timepoint. Cells are coloured by the RPKM *z*-score. See also Supplementary Table 4.2. (B) Change in HINCR1 levels at 1, 2, and 24 hrs in response to hypoxia. Data normalized relative to 0 hr timepoint.

We next adopted a bioinformatics strategy to infer potential roles for these differentially expressed lincRNAs. We reasoned that transcripts with similar expression profiles might be under similar patterns of regulatory control, and that the broad function of a noncoding RNA might therefore be inferred from the set of protein-coding transcripts with which it shares highly correlated expression profiles. Since hypoxia is a feature of the majority of solid tumours, but present at different levels in individual tumours, we assembled a large independent expression dataset comprising 248 published microarray samples encompassing Breast, Colorectal, Lung and Glioblastoma cancers[26-28]. To do this, we exploited the coverage of Affymetrix Exon Arrays, which feature a significant number of probesets targeting less well-characterised regions of the genome, including many noncoding loci[29]. Together, this cohort provided many more data points from which to calculate correlations between expression profiles than the original timecourse study. Following batch normalisation using ComBat[30], we used these data to derive a co-expression network based on partial correlation coefficients

(0.22% strongest correlations; Fig 4.2A). Of the 7,081 nodes in this network, 2,744 were noncoding, of which 224 were represented in our data.

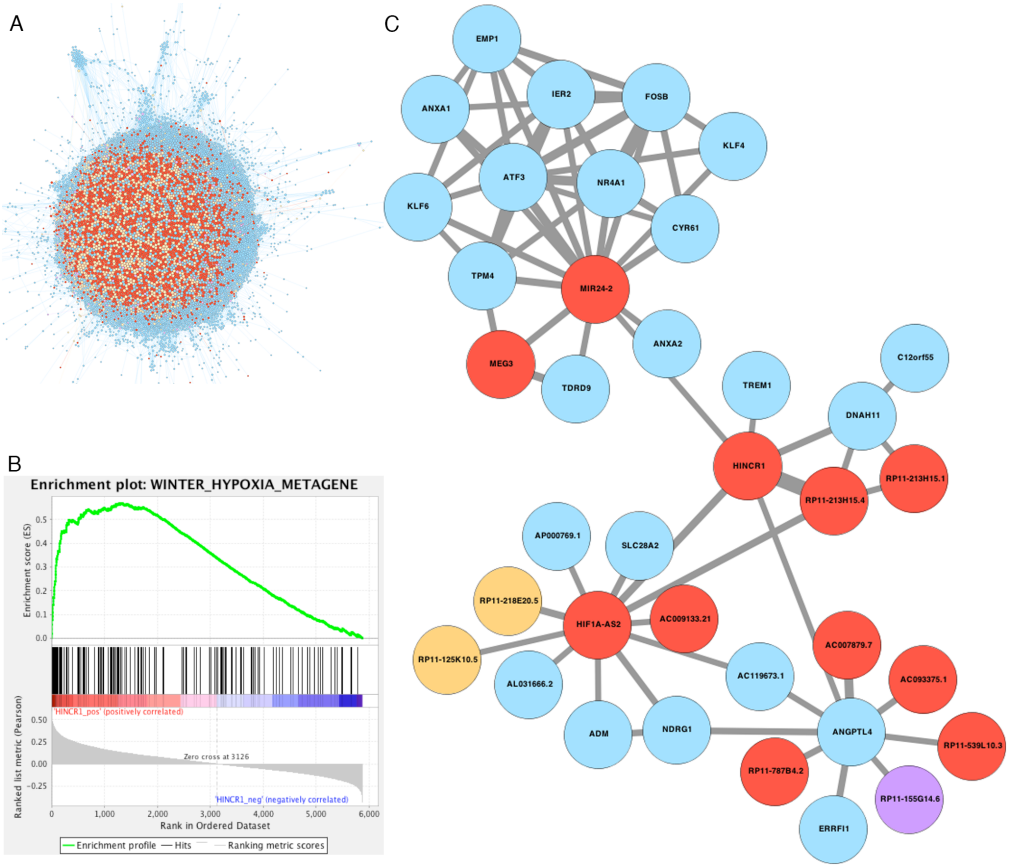

Figure 4.2 HINCR1 transcript levels correlate with known hypoxia-regulated genes. (A) Partial correlation network. Nodes represent genes. Blue: protein-coding. Red: lincRNA. Orange: antisense. Purple: processed transcript. Nodes sized according to weighted degree. Arcs represent 0.22% of all correlations, weighted by strength of correlation. Figure generated using Gephi[31]. (B) Gene-set enrichment analysis of correlations between HINCR1 and protein-coding transcripts, showing strong association with members of a hypoxia gene signature[32] in an independent cohort of tumour Exon array datasets complied from public domain experiments. (C) A sub-graph of (A) ≤ 2 nodes away from HINCR1, coloured as in A. Edges connecting nodes indicate strong absolute expression correlation between the gene products.

We then sought to determine if HINCRs upregulated in our *in vitro* dataset were also changing in the independent tumour cohort, and used Gene Set Enrichment Analysis (GSEA) to seek significant statistical associations between HINCRs and functionally related sets of protein-coding genes. Transcript levels at one of these loci, HINCR1 (LUCAT1/RP11-213H15.3), were significantly elevated at 2 hrs in hypoxia, with further induction at later timepoints (Fig 4.1B). HINCR1 expression patterns were highly correlated with a set of protein-coding genes previously reported by Winter *et al.*, 2007

as a signature of hypoxia[32] (Fig 4.2B,C). Genes from this set (referred to here as 'Winter signature') included angiopoietin-like 4 (ANGPTL4), N-myc downstream regulated 1 (NDRG1) and adrenomedullin (ADM). A total of 63 protein-coding genes are significantly induced at the 2 hr timepoint in hypoxia. These early responders were also highly correlated to HINCR1 in the tumour expression data (p-value < 0.001). In addition, reanalysis of previously published expression data from HUVEC cells[33,34] found HINCR1 to be induced both in response hypoxia and to chemical induction of HIF by $CoCl_2$ (Supplementary Figure 4.2).

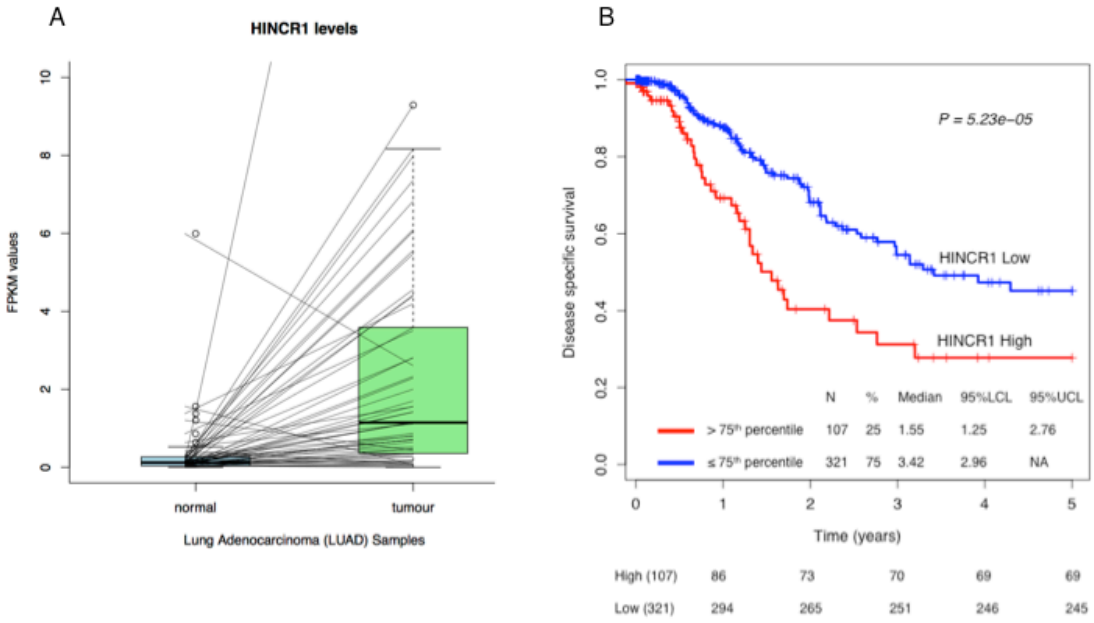

Figure 4.3 Survival analysis of HINCR1. (A) Comparison of HINCR1 expression levels between matched normal and tumour samples from lung adenocarcinoma (LUAD) patients. (B) Kaplan-Meier plot of disease-specific survival among lung adenocarcinoma samples (N=428) using TCGA data. Data stratified on HINCR1 gene expression levels around the 75th quantile value for the dataset.

4.2.2 HINCR1 expression is predictive of survival

HINCR1 is located at 5q14.3. Deletions encompassing this locus have been reported in multiple cancers: 5q11-5q23 was found deleted in 17% of breast basal cancer patients[35]. Loss of 5q has been reported in 43% of bladder tumour cell lines[23], gastric cancer[36], and the frequent loss of 5q14.3 has been reported in colorectal flat adenomas[37]. Differential expression analysis of HINCR1 levels between matched normal and tumour samples obtained from different tumour types from TCGA found

HINCR1 to be significantly induced in 7 out of 13 tumour types including cancers of lung, colon, head and neck, kidney and liver (Supplementary Figure 4.3). Particularly high levels of HINCR1 were observed in tumour samples of lung adenocarcinoma (LUAD) patients[38] (Fig 4.3A). Elevated expression levels of HINCR1 were also indicative of poor disease-specific survival in 428 lung adenocarcinoma patients (Fig 4.3B).

4.2.3 HINCR1 is integral to the hypoxic response

We therefore selected HINCR1 for further characterization. We designed a siRNA pool targeting exons common to all three predicted isoforms, using our annotation data as a guide (Fig 4.4A). Knockdown of the locus led to significant alterations in levels of a subset of the transcripts differentially expressed in hypoxia (Fig 4.4B; Supplementary Table 4.3). A systematic effect was observed in which changes in the levels of a specific subset of transcripts were attenuated relative to the scrambled control, indicating a role for HINCR1 in mediating their response to hypoxia. These genes were enriched for particular pathways (Fig 4.4C): GO:0006695 – *Cholesterol biosynthesis*, GO:0006984 – *ER-nucleus signalling* pathways.

We then asked whether the set of genes that responded to the HINCR1 knockdown shared similar expression profiles to HINCR1 in the independent clinical cohort used to generate Fig 4.2A. This is indeed the case: HINCR1 dependent transcripts with altered levels in the knockdown are significantly more correlated to HINCR1 than would be expected by chance (Wilcoxon's test p-value $< 10^{-6}$; Supplementary Figure 4.4). These include the Winter signature genes ANGPTL4, NDRG1, and ADM, identified in the initial *in silico* analysis (Fig 4.2C). Knockdown of HINCR1 led to altered levels of more than half of all early responders to hypoxia (N=33/63), thus identifying HINCR1 as a novel regulator of immediate early genes in hypoxia.

Importantly, the close correspondence between the genes that exhibit changes in expression level following siRNA mediated depletion of HINCR1 with those that are correlated with HINCR1 in the diverse and entirely independent set of tumour samples used to generate Fig 4.2A indicate both the clinical relevance of these data and their potential to generalise across a wide variety of tumour- and tissue types.

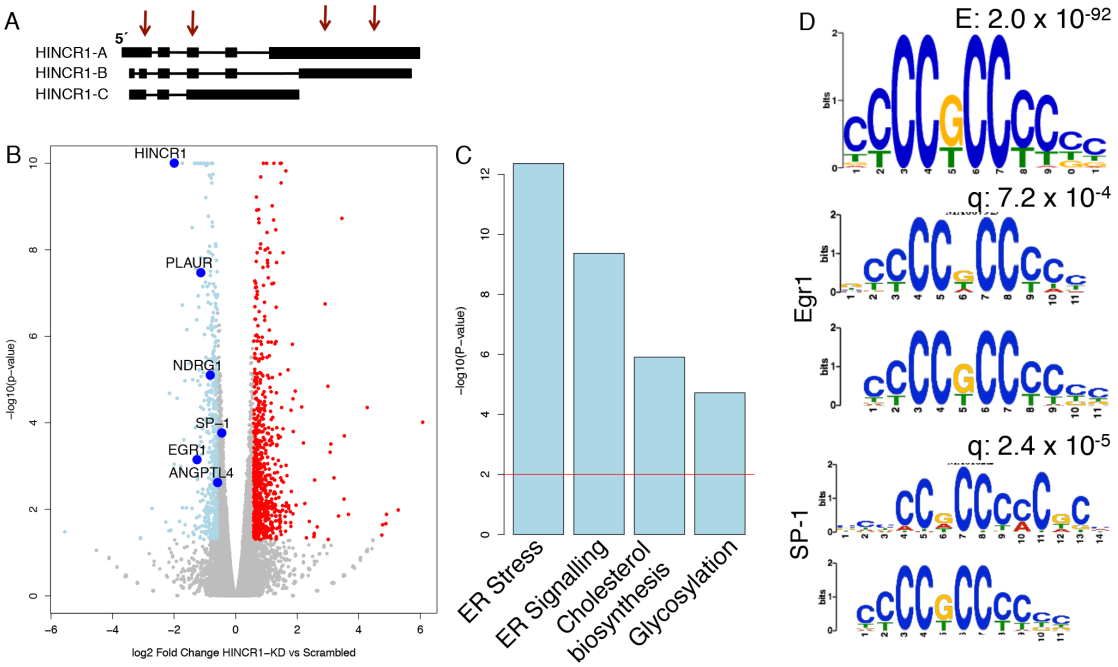

Figure 4.4 Knockdown of HINCR1 in hypoxia prevents upregulation of hypoxia-induced pathways. (A) A pool of siRNAs targeting the consensus region of the three predicted isoforms (A,B and C) of HINCR1. Arrows indicate region targeted by individual siRNAs. (B) Log_2 Fold Change in gene expression vs. $-log_{10}$(*p*-value) of HINCR1-KD HCT116 cells 24 hrs at 1% oxygen, relative to non-target controls. Significantly upregulated genes and downregulated genes are indicated in red and blue respectively. Genes with fold change > 1.5 and corrected *p*-value < 0.05 are coloured. Key downregulated genes are highlighted. (C) Significantly enriched GO biological processes are listed (Only non-redundant GO terms are represented). *ER stress*: GO:0034976; *ER Signaling*: GO:0006984; *Cholesterol biosynthesis*: GO:0006695; *Glycosylation*: GO:0006487) (D) DNA motif enriched in upstream region (1kb) of downregulated genes on knockdown of HINCR1. Also shown are Egr1 and SP-1 transcription factor binding site motifs, which show significant sequence similarity with the *de novo* detected motif. See also Supplementary Table 4.3.

4.2.4 HINCR1 modulates targets of Egr1

Motif analysis of the upstream regions of the loci dysregulated in the knockdown identified significant enrichment for an Egr1/SP-1 binding motif (Fig 4.4D). Egr1 is an immediate early response transcription factor that is induced within 1-2 hrs following a shift to hypoxic conditions but with levels that decline subsequently such that the protein is virtually undetectable at 24 hrs (Fig 4.5A). Knockdown of HINCR1 enhanced hypoxia responsive changes to Egr1 leading to further upregulation of transcript and protein levels at 2 hrs followed by an increased decline in transcript levels at 24 hrs (Fig 4.5A) HINCR1 knockdown in the lung cancer cell line A549 also led to

dysregulation of many genes (including Egr1) dysregulated in HCT116 (Fisher exact test *p*-value < 10^{-5}; Supplementary Figure 4.5; Supplementary Table 4.3). Together these data identify a novel and robust regulatory interaction between the noncoding RNA HINCR1 and Egr1.

4.2.5 HINCR1 binds to the promoter region of key hypoxia-regulated genes

Together these data, combined with its predominant nuclear localisation in A549 cells (data not shown) led us to hypothesise that HINCR1 might interact with Egr1 binding sites. We therefore performed pulldown and sequencing of DNA bound to HINCR1 at 0, 2 and 24 hrs in hypoxia. A total of 23499, 53473 and 63988 HINCR1 binding peaks were observed at 0, 2 and 24 hrs in hypoxia (Supplementary Table 4.4), correlating with the increased expression of HINCR1 and indicating more widespread binding of HINCR1 to the genome in response to hypoxia. Although 48% of the HINCR1 bound peaks were in intergenic regions, HINCR1 binding was enriched in promoter regions (2-fold enrichment) at the 2 hr timepoint. *De novo* analysis of HINCR1 binding peaks against a library of 264 reliable motifs using Homer, found Egr1 and Egr2 motifs among the most enriched motifs at both the 2 hr and 24 hr timepoints (*q*-value = 0). No enrichment of the Egr1 motif was found within the HINCR1 peaks detected at the 0 hr timepoint, confirming the tendency of HINCR1 to bind to Egr1 binding sites in hypoxia.

We then performed ChIP-Seq of Egr1 in the presence or absence of HINCR1. A total of 49658 significant Egr1 binding peaks were observed at 2 hrs in hypoxia (Supplementary Table 4.5), nearly 50% of which were found in promoter regions and exhibited high enrichment for the Egr1 binding motif, as expected (*p*-value < 1e-1437). Multiple Egr1 binding sites were observed at the HINCR1 locus suggesting a feedback loop exists between HINCR1 and Egr1 (Supplementary Figure 4.6). Excitingly, 1735 Egr1 peaks overlapped by at least 100bp with a HINCR1 binding peak at 2 hours. 728 (41.9%) of these were within 1000 bp of the start site of a gene. Following knockdown of HINCR1 binding of Egr1 to these sites increased (*p*-value < 10^{-3}), relative to non-overlapping distal Egr1 binding sites (>10 kb away from a HINCR1 peak) (Fig 4.5B). Further, genes nearest to HINCR1/Egr1 overlapping peaks were significantly more likely to be dysregulated by HINCR1 knockdown than genes lacking these peaks (Fisher exact test *p*-value < 0.05) and included VEGFA (Fig 4.5C) and INSIG1.

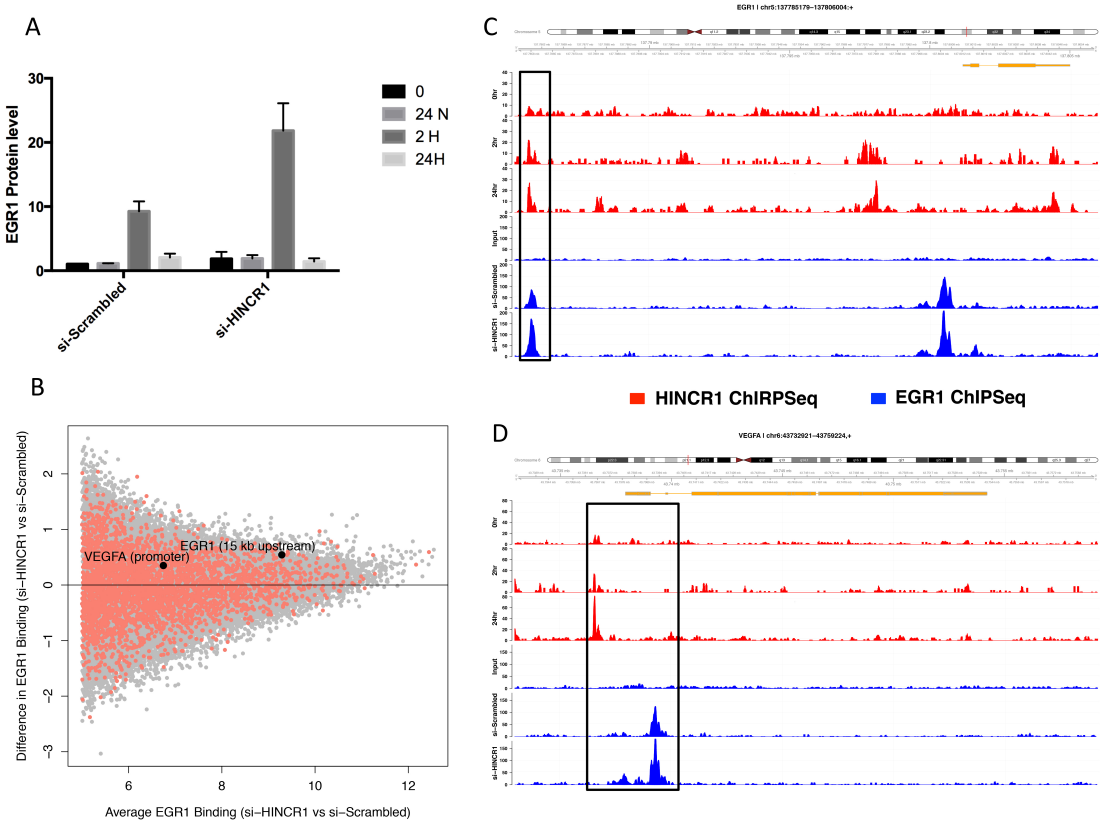

Figure 4.5 HINCR1 regulates Egr1 binding to its target sites. (A) Egr1 protein levels in normoxia (0 hr, 24 hr) and hypoxia (2 hr, 24 hr) in HCT116 cells in the presence and absence of HINCR1. (B) Differential binding of EGR1 in response to HINCR1 knockdown. All EGR1 binding sites have been plotted (in grey). The sites which have a HINCR1 binding peak within 0.5 kb distance are coloured pale pink. (C) HINCR1 binding peaks (in red) and EGR1 binding peaks (in blue) in the upstream region (15 kb) of EGR1. (D) Binding peaks same as in C, in the upstream region (1 kb) of VEGFA.

Although no HINCR1 binding sites were observed in the immediate promoter region of Egr1, it was found to bind to an intergenic region 15 kb upstream of Egr1 (Fig 4.5D) and overlapping with an Egr1 binding site. Loss of HINCR1 led to a significant increase (FC=1.45; *p*-value < 1e-10) in Egr1 binding at this site.

Taken together these data indicate that HINCR1 modulates Egr1 activity by altering the binding specificity of the transcription factor to its target sites.

4.3 Discussion

We found HINCRs with strong statistical associations to proteins in pathways known to respond to hypoxia. Knockdown of one locus, HINCR1, led to an attenuation of the gene expression changes that occur in hypoxia at a substantial number of loci across the genome. The same locus has recently been reported as being upregulated in the presence of cigarette smoke extract, and placed downstream of the oxidative stress-responsive transcription factor, nuclear factor erythroid 2-related factor (NRF2)[39]. Genes affected by HINCR1 knockdown were closer to HINCR1 in the (independent) correlation network than would be expected by chance, confirming HINCR1 as a *bona fide* regulator of hypoxia responses, while the presence of common Egr1 and SP-1 motifs in the promoter of many of these genes suggests that it acts preferentially on the expression of Egr1/SP-1 targets. HINCR1 was found to regulate Egr1 and preferentially bind to Egr1 target sites. Prevention of hypoxia-mediated increase in HINCR1 levels led to significant increase in Egr1 binding to its target confirming the role of HINCR1 as a novel noncoding immediate early response gene that acts upstream of the transcript factor Egr1. These data therefore identify for the first time a critical role for a noncoding RNA in coordinating the levels of key hypoxia responsive transcripts. The presence of multiple HINCRs suggests that many other noncoding RNAs will be revealed to exert similar controls across the genome.

4.4 Experimental Procedures

4.4.1 Cell culture

HCT116 were cultured in RPMI-1640 media (Life Technologies) supplemented with 10% FBS (Biowest). All cells were maintained in a humidified atmosphere at 37°C and 5% CO_2. For hypoxia treatment, the HCT116 cell line was cultured in 1% O_2 in an Invivo$_2$ hypoxia workstation 4000 (Biotrace, Fred Baker Ltd.) for the given time course 24 hrs after plating.

4.4.2 siRNA transfection

Cells were transfected with 20 μM siRNAs targeted to HINCR1 and nontargeting control siRNA (Thermo Scientific) using Dharmafect2 (Thermo Scientific) according to the manufacturer's instructions. After 48 hrs siRNA-treated cells were cultured in hypoxia for 24 hrs as above.

Sequence of siRNA oligos for HINCR1: 5′-UGUAUUUCUCUCACGUUAA-3′, 5′-UUUGGAAGGAUGAGACUUA-3′, 5′-GGAAAGAGACGAAGAGAAA-3′, 5′-GGTCAGTGAGTGAAGAGGA-3′

Non-target siRNA: D-001810-01 20 (Thermo Scientific).

RNA was extracted using the Qiagen Qiashredder kit (79654) and the Qiagen RNeasy Mini Kit (74104) as per the manufacturer's instructions. The RNA was DNase treated following the protocol in the RNeasy Mini Kit with Qiagen RNase-free DNase I (79254).

4.4.3 Protein extraction and western blotting

Protein was extracted by washing cells in ice cold PBS and scraping cells in ice cold cell lysis buffer (9803s New England Biolabs) supplemented with PMSF (Sigma 93482) and protease inhibitors (Roche Diagnostics complete EDTA free 11 873 580 001). The sample was centrifuged at 4^0C 13,000rpm for 10mins and supernatent kept. 50mg of total protein per sample was resolved by SDS-PAGE 10 % NuPage gels (Invitrogen) and transferred electrophoretically to Immobilon-P™ (Millipore). The membrane was blocked in 5%milk PBS-T for 30 mins and blotted overnight with Egr1 antibody (Fisher 11594971), or tubulin antibody (1/5000 Sigma T6199). Detection was performed using a peroxidase-conjugated anti-rabbit or anti-mouse IgG (Amersham Biosciences Pharmacia) and chemiluminescence visualization (ECL+, Amersham Biosciences) was used according to the manufacturer's instructions. Quantification of Western blot signals was performed using the Chemi Genius Bioimaging system (Syngene) and the Chemi genius gel documentation and analysis system.

4.4.4 Data analysis

All statistical analysis including t-tests and Wilcoxon's tests, were performed in R. Timecourse data: *De novo* annotation data combined with Ensembl (v74) from previous work (submitted) were used to obtain gene level counts using the RsubRead package in R and supplied to edgeR[40] to call differential expression (absolute fold-change > 2 relative to 0 hr; False Discovery Rate (FDR) < 1%) at each timepoint.

HINCR1 knockdown data: 100mer paired end strand specific Illumina sequencing was performed as before for the HINCR1-knockdown and Scrambled siRNA treated samples. Gene-level counts for these samples were obtained using the gene models

established in the timecourse analysis and edgeR used to perform differential expression in the HINCR1 knockdown samples relative to the siRNA treated samples. Genes were classified as differentially expressed if they had an absolute fold change > 1.5 and FDR < 5%. Genes unaffected by HINCR1 knockdown were those with a FDR equal to 100%.

4.4.5 Gene Ontology Enrichment Analysis

Functional enrichment analysis was performed using the Goseq[41] package in R to identify statistically enriched gene ontology (GO) terms (Hypergeometric test: Benjamini & Hochberg corrected *p*-value < 0.01). Non-redundant GO terms were obtained by retaining only one representative term from GO with high semantic similarity, derived using GOSemSim[42] package in R.

4.4.6 Coexpression Network

The majority of publically available large-scale expression data has been generated using microarrays. Unlike the majority of other microarray platforms, Affymetrix GeneChip Human Exon 1.0 ST Arrays feature many reliable probesets targeting a substantial number of lincRNAs annotated in ENSEMBL. We therefore established a large cohort of expression data by integrating three large tumour Exon array datasets (GEO accessions: GSE16534, GSE12236 and GSE9385). Batch effects were addressed using ComBat and the 248-sample dataset was then subjected to RMA normalization using default parameters. Reliable probesets were then mapped to the ENSEMBL human genome annotation (v74) using the annmap[43] Bioconductor package and expression for each gene was obtained by calculating median expression levels of all probesets mapped to the gene. Genes were ordered according to the expression range across the 248 samples. For co-expression network construction, only the top 25% most varying genes (highest range) belonging to the four biotypes, 'protein-coding', 'lincRNA', 'processed_transcript' and 'antisense', were used. The network was constructed using Partial Correlation Coefficients estimated for all gene pairs by the GeneNet[44] package in R. Only the top 0.22% most significantly correlated gene pairs with probability > 0.90 (local FDR < 0.1) were connected in the network. Gene Set Enrichment Analysis (GSEA) was performed on the expression data using the javaGSEA[45] tool along with chemical and genetic perturbation annotation data (c2.cgp.v4.0.symbols.gmt).

4.4.7 Analysis of TCGA Data

Aligned expression data from patients with matched normal and tumour samples were obtained from TCGA for 13 different tumour types (BLCA[46] – Bladder Urothelial Carcinoma, BRCA[47] – Breast invasive carcinoma, COAD[48] – Colon Adenocarcinoma, HNSC[49] – Head and Neck squamous cell carcinoma, KICH[50] – Kidney Chromophobe, KIRC[51] – Kidney renal clear cell carcinoma, KIRP[52] – Kidney renal papillary cell carcinoma, LIHC – Liver Hepatocellular Carcinoma, LUAD[38] – Lung adenocarcinoma, LUSC[53] – Lung squamous cell carcinoma, PRAD[54] - Prostate Adenocarcinoma, THCA[55] – Thyroid carcinoma, UCEC[56] - Uterine Corpus Endometrial Carcinoma). Expression levels were estimated for gene-level annotations from Ensembl (v74) for each sample using Cufflinks and expression profiles were normalized using Cuffnorm. For survival analysis, a larger cohort of 428 lung adenocarcinoma samples[38] and corresponding clinical annotations were obtained from TCGA and expression data was processed as before. Kaplan-Meier analysis was performed using the survival[57] package in R.

4.4.8 Motif detection in upstream regions

The MEME[58] package was used to detect motifs in the upstream region (1 kb) differentially expressed genes in response to HINCR1 knockdown. Initially a psp-gen model was built to perform discriminative motif discovery using the upregulated/downregulated loci as the positive sequences and a random set of sequences of the same size sampled from unaffected genes as the negative sequences. The model filters for non-specific repeats that are present in upstream regions of genes. The model along with the positive sequences was used as input to MEME to identify enriched motifs of length between 8-12 bp and 0 or 1 occurrence in the input sequences. Finally, the enriched motifs were searched against a database of known motifs of transcription factors in vertebrates (JASPAR database[59]) using the TOMTOM[60] tool for motif comparison within the MEME package.

4.4.9 ChIP-Seq and ChIRP-Seq

Chromatin Immunoprecipitation was carried according to the Diagenode IDeal ChIP-seq Kit (C01010054). Briefly, 15 million HCT116 cells were directly crosslinked in 1% formaldehyde for 2 minutes and the reaction quenched in 1/10th 1.25M glycine. The cells were washed and lysed and chromatin was sheared using the Diagenode Bioruptor until the chromatin fragments were 200-500bp. Subsequently, IDeal ChIP-seq Kit was applied for ChIP of the sheared chromatin. ChIP samples were performed

using Egr1 monoclonal antibody (Fisher 11594971) and SP1 polyclonal antibody (Abcam ab13370) for scrambled and HINCR1-siRNA treated samples at the 2 hr timepoint in hypoxia. DNA was extracted and purified using the Diagenode I-Pure kit and was subjected to paired-end deep sequencing with read depths in the range of 40-50 M for individual sample. Matched input samples for also generated for each treatment. The reads were aligned using Bowtie 2.0[61] with default parameters. Aligned reads were then used for peak detection using MACS2[62] with default parameters. Peaks were refined to identify sub-peaks using peak splitting algorithms within MACS. Peak annotation and motif search in peak regions was performed using algorithms in Homer. Finally, quantitative comparison of ChIP-Seq datasets was performed using MAnorm.

The standard ChIRP-seq protocol was followed as described by Chu *et al*, 2012[63]. Briefly, 20 million cells per sample HCT116 cells were crosslinked using 1% formaldehyde for 10 min. The reaction was then quenched in glycine (1/10th 1.25M). After washing with PBS and pelleting down, the cells were lysed using Lysis Buffer (50mM Tris-Cl pH7.0, 10 mM EDTA, 1% SDS, supplemented with PMSF, Protease Inhibitor Cocktail and Superase-in). The chromatin was sheared into 100- to 500-bp DNA fragments on a Diagenode Bioruptor and mixed with Hybridisation Buffer (750mM NaCl, 1%SDS, 50mM Tris-Cl pH 7.0, 1mM EDTA, 15% formamide and supplemented as above). The sheared chromatin was incubated with either HINCR1-specific or non-specific DNA probes modified with a TEG linker and Biotin at their 3′ ends. Following incubation for 24 h at RT°C the biotin oligos were pulled down using Streptavidin-C1 magnetic beads and washed in Wash Buffer (2xSSC, 0.5% SDS supplemented with PMSF). The DNA was then treated with RNAases (RNaseH and RNAseA) for 30 mins 37°C in DNA Elution Buffer (50mM NaHC03, 1% SDS) and subsequently with Proteinase K. The DNA was then extracted from the samples using phenol/chloroform and alcohol precipitation, pelleted the next day, air dried and resuspended in Qiagen Elution Buffer. The DNA was then used qPCR analysis and paired-end DNA sequencing was performed on a HiSeq2500. DNA reads were mapped against human genome (hg19) using Bowtie 2, and peak calling was performed using MACS2. Peak annotation was performed using Homer.

HINCR1-specific probes (5′ -> 3′): CCACCTAAGAGCAGAAACTT, GGGGTGATTAGATTACTTGC, TGATAGGTGAGGAGAACTGA, CAAAAAGCTTACTGTTGGCC, CCTTGGAAAAATTGCTGGCT, CTGAGATACACTGAGCCATA, GTCAAAAGAAGAGCAGGGTT,

CAAGTGAGGAAGAATCCACA, ATGGAGAATACTGGGGAAGA, GAGGAGGTTACGTAGATCTT, GTAGCAAACTTGTACACGCA, ACTGTGTTGCTTCAAATGGG, GCAAACAGCAAGTTGGATTC, TTTCATTGGGAGATGAGGAC, CTGAGTGGAGTGTTGATTCT.

Non-specific probes (5′ -> 3′): ATCAACGCCTAACTAGCAGA, GTAGGTTCGTATCGTGGATA, GTGGTGCGATAAGATAAGAG, GCTTATCGCCTAATACAAGG, ATACTCGGCGTGTATATAGC, GCCTACCGATAGACTAATAG, AGAGCGTATAAGAGTGCAAG, AGCAATGAATGACGACGAAC, GAGAGAATACGAGTAAGTGG, GCGAGTACGTTAATTGGTAG, ACTAGCGTAACCATCGGAAT, GTGTAACGTTCGTGCGATAT, AGTCGTATACTCGACAGGAA, GATGTAGTATGCGGATACGT, ACAGTGTATCGTGTTGGTTG

Author Contributions

DM performed the data analysis with contributions from YL, CW, HSL. KD performed the bench experiments. SF, JB provided siRNA expertise. GB, CD, CJM supervised the project. DM and CJM wrote the manuscript with assistance from all the authors.

Acknowledgements

This work was funded by Cancer Research UK (Grant number: C5759/A12328).

4.5 References

1. Kung, J. T. Y., Colognori, D. & Lee, J. T. Long noncoding RNAs: past, present, and future. *Genetics* **193,** 651–669 (2013).
2. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489,** 101–108 (2012).
3. Rinn, J. L. & Chang, H. Y. Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* **81,** 145–166 (2012).
4. Lee, J. T. Epigenetic regulation by long noncoding RNAs. *Science* **338,** 1435–1439 (2012).
5. Zhao, J., Sun, B. K., Erwin, J. A., Song, J.-J. & Lee, J. T. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* **322,** 750–756 (2008).
6. Jeon, Y. & Lee, J. T. YY1 tethers Xist RNA to the inactive X nucleation center. *Cell* **146,** 119–133 (2011).
7. Gong, C. & Maquat, L. E. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. *Nature* **470,** 284–288 (2011).

8. Tripathi, V. *et al.* The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Mol. Cell* **39,** 925–938 (2010).
9. Wang, X. *et al.* Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. *Nature* **454,** 126–130 (2008).
10. Moulder, J. E. & Rockwell, S. Tumor hypoxia: its impact on cancer therapy. **5,** 313–341 (1987).
11. Teicher, B. A. Hypoxia and drug resistance. *Cancer Metastasis Rev.* **13,** 139–168 (1994).
12. Carmeliet, P. *et al.* Role of HIF-1alpha in hypoxia-mediated apoptosis, cell proliferation and tumour angiogenesis. *Nature* **394,** 485–490 (1998).
13. Maxwell, P. H. *et al.* Hypoxia-inducible factor-1 modulates gene expression in solid tumors and influences both angiogenesis and tumor growth. *Proc. Natl. Acad. Sci. U.S.A.* **94,** 8104–8109 (1997).
14. Wilson, W. R., Wilson, W. R., Hay, M. P. & Hay, M. P. Targeting hypoxia in cancer therapy. *Nat. Rev. Cancer* **11,** 393–410 (2011).
15. McKeown, S. R. Defining normoxia, physoxia and hypoxia in tumours-implications for treatment response. *Br J Radiol* **87,** 20130676 (2014).
16. Eustace, A. *et al.* A 26-gene hypoxia signature predicts benefit from hypoxia-modifying therapy in laryngeal cancer but not bladder cancer. *Clin. Cancer Res.* **19,** 4879–4888 (2013).
17. Huang, L. E., Gu, J., Schau, M. & Bunn, H. F. Regulation of hypoxia-inducible factor 1alpha is mediated by an O2-dependent degradation domain via the ubiquitin-proteasome pathway. *Proc. Natl. Acad. Sci. U.S.A.* **95,** 7987–7992 (1998).
18. Gordan, J. D. & Simon, M. C. Hypoxia-inducible factors: central regulators of the tumor phenotype. *Curr. Opin. Genet. Dev.* **17,** 71–77 (2007).
19. Makino, Y. *et al.* Inhibitory PAS domain protein is a negative regulator of hypoxia-inducible gene expression. *Nature* **414,** 550–554 (2001).
20. Tian, H. *et al.* Endothelial PAS domain protein 1 (EPAS1), a transcription factor selectively expressed in endothelial cells. *Genes & Development* **11,** 72–82 (1997).
21. Rong, Y. *et al.* Early growth response gene-1 regulates hypoxia-induced expression of tissue factor in glioblastoma multiforme through hypoxia-inducible factor-1-independent mechanisms. *Cancer Res.* **66,** 7067–7074 (2006).
22. ENCODE Project Consortium *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447,** 799–816 (2007).
23. Hurst, C. D. *et al.* High-resolution analysis of genomic copy number alterations in bladder cancer by microarray-based comparative genomic hybridization. *Oncogene* **23,** 2250–2263 (2004).
24. Matouk, I. J. *et al.* The H19 non-coding RNA is essential for human tumor growth. *PLoS ONE* **2,** e845 (2007).
25. Xue, M., Li, X., Li, Z. & Chen, W. Urothelial carcinoma associated 1 is a hypoxia-inducible factor-1α-targeted long noncoding RNA that enhances hypoxic bladder cancer cell proliferation, migration, and invasion. *Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine* (2014). doi:10.1007/s13277-014-1925-x
26. Lin, E. *et al.* Exon array profiling detects EML4-ALK fusion in breast, colorectal, and non-small cell lung cancers. *Molecular cancer research : MCR* **7,** 1466–1476 (2009).
27. Xi, L. *et al.* Whole genome exon arrays identify differential expression of alternatively spliced, cancer-related genes in lung cancer. *Nucleic Acids Research* **36,** 6535–6547 (2008).

28. French, P. J. *et al.* Identification of differentially regulated splice variants and novel exons in glial brain tumors using exon expression arrays. *Cancer Res.* **67,** 5635–5642 (2007).
29. Du, Z. *et al.* Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. *Nat. Struct. Mol. Biol.* **20,** 908–913 (2013).
30. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics (Oxford, England)* **8,** 118–127 (2007).
31. Bastian, M., Heymann, S. & Jacomy, M. Gephi : An Open Source Software for Exploring and Manipulating Networks. 1–2 (2009). at <http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154>
32. Winter, S. C. *et al.* Relation of a hypoxia metagene derived from head and neck cancer to prognosis of multiple cancers. *Cancer Res.* **67,** 3441–3449 (2007).
33. Weigand, J. E., Boeckel, J.-N., Gellert, P. & Dimmeler, S. Hypoxia-induced alternative splicing in endothelial cells. *PLoS ONE* **7,** e42697–e42697 (2012).
34. Hang, X. *et al.* Transcription and splicing regulation in human umbilical vein endothelial cells under hypoxic stress conditions by exon array. *BMC Genomics* **10,** 126 (2009).
35. Adélaïde, J. *et al.* Integrated profiling of basal and luminal breast cancers. *Cancer Res.* **67,** 11565–11575 (2007).
36. Oga, A. *et al.* Preferential loss of 5q14-21 in intestinal-type gastric cancer with DNA aneuploidy. *Cytometry* **46,** 57–62 (2001).
37. Voorham, Q. J. M. *et al.* Chromosome 5q loss in colorectal flat adenomas. *Clin. Cancer Res.* **18,** 4560–4569 (2012).
38. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511,** 543–550 (2014).
39. Thai, P. *et al.* Characterization of a novel long noncoding RNA, SCAL1, induced by cigarette smoke and elevated in lung cancer cell lines. *American journal of respiratory cell and molecular biology* **49,** 204–211 (2013).
40. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).
41. Young, M. D., Wakefield, M. J., Smyth, G. K. & Oshlack, A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol* **11,** R14 (2010).
42. Yu, G. *et al.* GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* **26,** 976–978 (2010).
43. Yates, T., Okoniewski, M. J. & Miller, C. J. X:Map: annotation and visualization of genome structure for Affymetrix exon array analysis. *Nucleic Acids Research* **36,** D780–6 (2008).
44. Schäfer, J. & Strimmer, K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat Appl Genet Mol Biol* **4,** Article32 (2005).
45. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **102,** 15545–15550 (2005).
46. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* **507,** 315–322 (2014).
47. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).
48. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487,** 330–337 (2012).
49. Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517,** 576–582 (2015).

50. Davis, C. F. *et al.* The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer Cell* **26,** 319–330 (2014).
51. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499,** 43–49 (2013).
52. Linehan, W. M. *et al.* Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma. *N. Engl. J. Med.* (2015). doi:10.1056/NEJMoa1505917
53. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489,** 519–525 (2012).
54. Cancer Genome Atlas Research Network. Electronic address: schultz@cbio.mskcc.orgCancer Genome Atlas Research Network. The Molecular Taxonomy of Primary Prostate Cancer. *Cell* **163,** 1011–1025 (2015).
55. Cancer Genome Atlas Research Network. Integrated genomic characterization of papillary thyroid carcinoma. *Cell* **159,** 676–690 (2014).
56. Cancer Genome Atlas Research Network *et al.* Integrated genomic characterization of endometrial carcinoma. *Nature* **497,** 67–73 (2013).
57. Therneau, T. M. A package for survival analysis in S. (2014). at <http://cran.r-project.org/web/packages/survival/index.html>
58. Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology* **2,** 28–36 (1994).
59. Portales-Casamar, E. *et al.* JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Research* **38,** D105–10 (2010).
60. Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying similarity between motifs. *Genome Biol* **8,** R24 (2007).
61. Langmead, B. Aligning short sequencing reads with Bowtie. *Curr Protoc Bioinformatics* **Chapter 11,** Unit 11.7 (2010).
62. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9,** R137 (2008).
63. Chu, C., Quinn, J. & Chang, H. Y. Chromatin isolation by RNA purification (ChIRP). *J Vis Exp* (2012). doi:10.3791/3912

Chapter 5. Conclusion

The rather weak correlation between genome size and organism complexity (also known as the C-value paradox) has always intrigued biologists[1]. A possible resolution to this enigma was offered by potential functionality hidden within the non-coding regions of the genome previously considered to be "junk" (or non-functional)[1]. Thus while regions with protein-coding potential occupy only 1.2% of the genome the advent of genome-wide profiling methods such as global RNA-Seq and ChIP-Seq, provided evidence of transcription as well as the widespread binding of proteins (transcription factors and chromatin modifiers) throughout these non-coding regions. It is now estimated that 70–90% of the human genome is transcribed in certain contexts[2].

Another key observation from global transcriptomic studies has been the extent of alternative splicing in coding genes, significantly expanding diversity within the proteome[3]. In general, higher eukaryotes tend to have more alternative splicing with more 95% of human estimated to undergo alternative splicing[3]. Therefore, alternative splicing of coding genes and presence of non-coding genes together are likely to account for the complexity observed in higher organisms.

With the discovery of the highly pervasive nature of eukaryotic transcription, a number of questions have been raised[1]. Firstly, do the non-coding molecules expressed as a result of pervasive transcription have any functional relevance, or are they just transcription noise?[1] Recent studies have revealed a diversity of functional roles for lncRNAs, suggesting that many are likely to be functional[4]. Secondly, how essential are these lncRNAs for organism survival. Since most lncRNAs are expressed in a particular cell type, the focus has been on the role of lncRNA in a tissue-specific context[5,6] and majority of knockout models generated are of tissue-specific lncRNAs[7]. Work in this thesis asked the opposite question: Are there lncRNAs that act as housekeepers and how do they differ from tissue-specific lncRNAs? Since a subset of housekeeping protein-coding genes have been shown to be core essential genes for the survival of the organism[8], it is reasonable to posit that essential lncRNAs would be found among predicted housekeeping lncRNAs.

A small set of 55 HK-lncRNAs were identified with distinct properties. Only a handful of these lincRNAs, including MALAT1 and NEAT1 have been extensively studied[9-11]. It is hoped that this initial catalogue will provide a useful source with which to prioritise future studies. The biological roles predicted from this bioinformatics analysis would need to be verified in the lab, however, work on the lncRNA HINCR1 demonstrates the utility of the bioinformatics approaches and suggests that predictions for these HK-lincRNAs are also likely to be true. Since these HK-lincRNAs are ubiquitously expressed, it is reasonable to suggest that they might be under the control of a common transcription factor. Bioinformatics analysis of potential promoter regions of these lncRNAs did not identify any common motifs (data not shown). However, this question could be studied by investigating publically available ChIP-Seq datasets representing an increasingly large set of ubiquitous transcription factors[12]. The functional activity of lncRNA has been attributed to its secondary structure. However, the structure of lncRNAs are poorly understood, and identification of functional domains in lncRNA structure remains a critical challenge. Therefore, in future it would also be interesting to ask whether these HK-lincRNAs share any common structural motifs in their secondary structures. This may be possible if the HK-lncRNAs have a similar mechanism of action or interact with the same set of proteins. Ultimately, the goal will be to develop classification methods for lncRNAs similar to those established for proteins.

Hypoxia causes dramatic changes to the transcriptomic landscape of cells. These changes are brought about by a number of mechanisms. Among them transcriptional activation by HIFs is well established, as are the post-transcriptional changes brought about by miRNAs such as miR-210. In addition, splicing changes have also been reported in hypoxia. Several genes (e.g. Bnip3[13] and PFKFB3[14], PFKFB4[15], VEGFA[16], Cyr61[17], MAX[18], ADM[19], PDK1[20], PS2[21], TrKA[22], SMN2[23], YT521[24]), have been reported to undergo alternative splicing changes in response to hypoxia (or hypoxia mimic) conditions in human or mouse, but work to date has focused on individual loci, rather than global patterns. Among the few attempts to perform global analyses of the effects of hypoxia on alternative splicing using, predominately, microarray techniques[25-27], the following observations have been reported: Weigand *et al.*, 2012[25] found probesets corresponding to a small subset of genes that were differentially expressed, indicative of alternative splicing when endothelial cells were treated with hypoxia. These included cases of intron retention, exon skipping and alternative promoter usage that in some cases affected the coding sequence. In another study, Hu *et al.*[28] reported hypoxia-

dependent alternative splicing changes in mesenchymal stem cells but found limited overlap with the Weigand *et al.* study, suggesting that cell-specificity may have a significant impact in the cellular response to hypoxia. Finally, reanalysis of publically available microarray data in the context of splicing has shown that cells in stressed environment (such as heat-shock, hypoxia, breast cancer and gliomas) tend to select latent, intronic 5′ splice sites in hundreds of functionally important genes, thus leading to incorporation of premature stop codons with downstream effects on cellular activity[29]. Previous work on Cyr61[17] and ADM[19] has shown splicing out of introns in hypoxic condition for increased protein production. While we observed accurate splicing for many genes with a critical role in hypoxia, the data here suggest a reverse phenomenon that occurs on a global scale, in which cells use intron retention as a means to de-activate protein synthesis. We observed that in certain conditions coding genes could behave as non-coding. Further work would be required to determine whether this coding to non-coding switch is a consequence of mis-splicing, or the result of a novel mechanism. The latter possibility is appealing since cells may use this coding to non-coding switch as one among many mechanisms with which to regulate protein levels. These findings are particularly interesting given that genes belonging to the DNA repair pathway were found to be deactivated. While we show decreases in protein levels for two loci, it will be interesting to see whether global proteomics data in hypoxia indicates widespread de-activation of DNA repair pathway. A number of studies have reported weak or moderate correlation between transcriptomics and proteomics data[30-33]. In addition to post-transcriptional gene expression changes[34,35], an additional biological explanation for this relatively weak correlation is the expression of non-coding transcripts from protein-coding genes. Although a coding to non-coding switch may sound atypical, even in normal tissues a small proportion of protein-coding genes (~800) express non-coding transcript as its major isoform. The proportion increases substantially in in tumours. Overall our data indicate the importance of this phenomenon and suggest that it merits further investigation both in terms of basic cellular biology and in the context of cancer. Genomic annotation databases have sub-divided genes using a hard threshold as either coding or non-coding. These definitions will need to be made more pliable to capture the dynamic nature of the transcriptome. Since all the work was performed on a large cell population, another question that remains unclear is how heterogeneous these changes are across bulk tissue, and whether there is a single population of cells with a constant change, or whether sub-populations exhibit more dramatic switches between coding and non-coding

expression. With the advent of Single cell Sequencing techniques[36] it would now be possible to address this question.

In addition to a switch between coding and non-coding state, many protein-coding genes also switch from one coding isoform to another coding isoform in hypoxia. These genes were significantly enriched for chromatin modifiers (corrected BH *p*-value < 10^{-5}) including the critical epigenetic regulator KDM2A, which was predicted to utilize an alternative promoter under hypoxia, presumably in order to regulate expression. KDM2A is frequently over expressed in Non-Small Cell Lung Cancers (NSCLCs), has been associated with increased proliferation and invasiveness, and activates ERK1/2 via the epigenetic repression of the dual specific phosphatase DUSP3[37]. Depletion of KDM2A in human stem cells has been shown to lead to G1/S cell cycle arrest through de-repression of p15/INK4B and p27/Kip1[38]. We also observed alterations to the exon structure within the CDS of many genes, including multiple kinases. Strikingly, COT kinase/Tpl2 (MAP3K8) was detected in normoxia as a novel short isoform lacking the majority of its C-terminus including the kinase domain and its active site. COT/Tpl2 has a broad range of substrates: its overexpression has been shown to activate ERK1/2, JNK, p38γ and ERK5, and it has been implicated in the activation of NF-κB (recently reviewed in [39]). While the function of the N-terminus of the protein (encoded by the novel short isoform we detected in normoxia) is not known, our data show for the first time that the full length Tpl2 transcript is induced under hypoxia, suggesting a hypoxia dependent modulation of downstream signalling from Tpl2. While KDM2A and MAP3K8 form interesting cases of hypoxia-dependent inclusion of function domains, we also observe cases wherein hypoxia may lead to loss of functional domains as seen in case of IKBIP. Differential splicing in coding region in many cases was inter-linked with differential use of first or last exon illustrating the significance of UTR selection in hypoxia. In summary, our work provides interesting insights into the role of hypoxia in alternative splicing.

As part of another study we also investigated the role of non-coding RNAs in hypoxia. A significant proportion of non-coding genes were found to be dysregulated in hypoxia. Using a combination of bioinformatics approaches, a locus with a critical role in hypoxia was identified and then subjected to experimental characterization in collaboration with Keren Dawson, an SSO in the RNA Biology Group who performed the bench work. There are multiple transcripts expressed from HINCR1 locus in hypoxia. Some of these transcripts are unannotated in existing annotation databases. It is unclear whether

these transcripts have different functional roles. We focused on studying the overall function role of the HINCR1 locus. Therefore, the pool of siRNAs used in our study targeted the common region of all predicted transcripts. The HINCR1 transcripts highly enriched for SINE elements and these SINE elements play functional role in ncRNA. *De novo* motif search in HINCR1 binding peaks did show slight enrichment of SINE elements (data not shown). A particularly exciting result from the study of the HINCR1 locus was the regulatory relationship between HINCR1 and Egr1.

Egr1 has been shown to act as a master regulator of ischemic stress. Ischemia is a restriction in blood supply to tissues leading to reperfusion injury and then triggering a vascular response that results in tissue damage[40]. The immediate effect on tissues affected by ischemia is the deprivation of oxygen. Studies on mice models with ischemic lungs show high levels of Egr1 within 30 minutes of lung ischemia[40]. Egr1 knockout mice that suffer lung ischemia show insufficient induction of inflammatory response genes such as ICAM-1, IL-1-B and MIP-2[40]. Another feature of ischemic lungs is the deposition of fibrin due to activation of procoagulation pathways, which are significantly curtailed in the absence of Egr1, primarily due to reduced induction of TF (tissue factor) and PAI-1 (SERPINE1)[40]. Finally, the Egr1 knockout mice also show higher overall survival[40]. Depending on the cellular context, Egr1 has been shown to have both tumour-suppressive[41] and oncogenic potential[42]. Egr1 is down-regulated in a number of tumour types including lung[43]. In Non-small cell lung cancer (NSCLC), Egr1 expression has been shown to be strongly correlated with PTEN expression and NSCLC patients with higher levels of Egr1 showed better overall survival as compared NSCLC patients with lower levels of Egr1[43]. Over-expression of Egr1 in HCT1299 and A549 NSCLC cell lines reduced cell migration. By contrast, Egr1 expression is induced in prostate[44] and hepatocellular carcinoma[45]. Further, there is growing evidence that Egr1 plays a major role in prostate cancer development making it a suitable drug target[42]. Egr1 has been shown to regulate the levels of IL-8[46], a chemokine that is mainly expressed in metastatic prostate cancer tissues and contributes to tumour growth and angiogenesis.

Given the crucial role of Egr1 in cells, the relationship between HINCR1 and Egr1 is of critical importance. We found a tendency of HINCR1 to bind to Egr1 binding sites. HINCR1 is also able to regulate Egr1 levels, presumably through a HINCR1 binding site. The ability of lncRNAs to regulate transcriptional programs via interaction with transcription factors is not new[4]. The transcription factor SOX2 and the lncRNA RMST

also shows a similar relationship[47]. However, RMST was not found to regulate SOX2[47]. The relationship between Egr1 and HINCR1 is thus harder to unravel. Although we observe strong correlation of gene expression between HINCR1 and Egr1 target genes in a number of expression datasets, only a weak correlation was observed between HINCR1 and Egr1 itself. Due to the complex relationship between HINCR1 and Egr1, it may not be obvious from simple correlation based approaches. Regulatory feedback relationships between transcription factors and ncRNAs have been suggested for other ncRNAs as well[48]. Therefore, in order to accurately predict complex feedback relationships between ncRNAs and transcription factors requires the development of more sophisticated bioinformatics approaches. In addition, the co-expression networks built at gene level and do not take into account potential sources of post-transcriptional regulation. Therefore, there is a need to shift towards transcript (not gene) level co-expression networks and also to including the effect of miRNA-based regulation.

In addition to contributing towards a between understanding of the non-coding transcriptome, this work has also facilitated development of bioinformatics tools and approaches. Bioinformatics methods for identification of alternative splicing changes in RNA-Seq are currently being developed[49], and a consensus on the best approach has yet to be reached[49]. Different tools offer different type of information, and a substantial amount of effort was spent developing a comprehensive pipeline to study alternative splicing. A better visual representation of alternative splicing changes was also developed in collaboration with Chris Smowton, a software engineering postdoc in the Scientific Computing Team. These sorts of representations are suitable for incorporation within DEXSeq package[50] as improved visualizations of DEXSeq output. Similarly, significant effort went into selection of 'candidate' lncRNAs for functional characterization. The bioinformatics approaches used for lncRNA selection proved to be successful in predicting lncRNA-coding gene relationships.

In summary this thesis summarizes the use of bioinformatics approaches to propose novel hypotheses on the functional role of non-coding transcriptome and experimental verification of these predictions in collaboration with wet lab scientists. Analysis of deep sequencing datasets have revealed the highly dynamic nature of the transcriptome in hypoxia, the likelihood of presence of essential lncRNAs and helped us decipher the function of a hypoxia-dependent lncRNA. Together this work unravels the multi-faceted functional role of the non-coding transcriptome.

5.1 References

1. Kung, J. T. Y., Colognori, D. & Lee, J. T. Long noncoding RNAs: past, present, and future. *Genetics* **193,** 651–669 (2013).
2. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489,** 101–108 (2012).
3. Keren, H., Lev-Maor, G. & Ast, G. Alternative splicing and evolution: diversification, exon definition and function. *Nat. Rev. Genet.* **11,** 345–355 (2010).
4. Moran, V. A., Perera, R. J. & Khalil, A. M. Emerging functional and mechanistic paradigms of mammalian long non-coding RNAs. *Nucleic Acids Research* **40,** 6391–6400 (2012).
5. Kelley, D. & Rinn, J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol* **13,** R107 (2012).
6. Cabili, M. N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & Development* **25,** 1915–1927 (2011).
7. Sauvageau, M. *et al.* Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *Elife* **2,** e01749 (2013).
8. Hart, T. *et al.* Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. *Mol. Syst. Biol.* **10,** 733–733 (2014).
9. Naganuma, T. & Hirose, T. Paraspeckle formation during the biogenesis of long non-coding RNAs. *RNA Biol* **10,** 456–461 (2013).
10. West, J. A. *et al.* The long noncoding RNAs NEAT1 and MALAT1 bind active chromatin sites. *Mol. Cell* **55,** 791–802 (2014).
11. Ulitsky, I. & Bartel, D. P. lincRNAs: genomics, evolution, and mechanisms. *Cell* **154,** 26–46 (2013).
12. ENCODE Project Consortium *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447,** 799–816 (2007).
13. Gang, H. *et al.* A novel hypoxia-inducible spliced variant of mitochondrial death gene Bnip3 promotes survival of ventricular myocytes. *Circ. Res.* **108,** 1084–1092 (2011).
14. Mykhalchenko, V. G. *et al.* Expression of mouse 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase-3 mRNA alternative splice variants in hypoxia. *Ukr Biokhim Zh (1999)* **80,** 19–25 (2008).
15. Minchenko, O. H., Ogura, T., Opentanova, I. L., Minchenko, D. O. & Esumi, H. Splice isoform of 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase-4: expression and hypoxic regulation. *Mol. Cell. Biochem.* **280,** 227–234 (2005).
16. Salton, M., Voss, T. C. & Misteli, T. Identification by high-throughput imaging of the histone methyltransferase EHMT2 as an epigenetic regulator of VEGFA alternative splicing. *Nucleic Acids Research* **42,** 13662–13673 (2014).
17. Hirschfeld, M., Hausen, zur, A., Bettendorf, H., Jaeger, M. & Stickeler, E. Alternative Splicing of Cyr61 Is Regulated by Hypoxia and Significantly Changed in Breast Cancer. *Cancer Res.* **69,** 2082–2090 (2009).
18. Kemmerer, K. & Weigand, J. E. Hypoxia reduces MAX expression in endothelial cells by unproductive splicing. *FEBS Lett* **588,** 4784–4790 (2014).
19. Sena, J. A. *et al.* HIFs Enhance the Transcriptional Activation and Splicing of Adrenomedullin. *Mol Cancer Res* **12,** 728–741 (2014).
20. Sena, J. A., Wang, L., Heasley, L. E. & Hu, C.-J. Hypoxia regulates alternative splicing of HIF and non-HIF target genes. *Mol Cancer Res* **12,** 1233–1243 (2014).

21. Higashide, S. *et al.* Identification of regulatory cis-acting elements for alternative splicing of presenilin 2 exon 5 under hypoxic stress conditions. *J. Neurochem.* **91,** 1191–1198 (2004).
22. Tacconelli, A. *et al.* TrkA alternative splicing: a regulated tumor-promoting switch in human neuroblastoma. *Cancer Cell* **6,** 347–360 (2004).
23. Bebee, T. W., Dominguez, C. E., Samadzadeh-Tarighat, S., Akehurst, K. L. & Chandler, D. S. Hypoxia is a modifier of SMN2 splicing and disease severity in a severe SMA mouse model. *Hum. Mol. Genet.* **21,** 4301–4313 (2012).
24. Hirschfeld, M. *et al.* Hypoxia-dependent mRNA expression pattern of splicing factor YT521 and its impact on oncological important target gene expression. *Mol. Carcinog.* **53,** 883–892 (2014).
25. Weigand, J. E., Boeckel, J.-N., Gellert, P. & Dimmeler, S. Hypoxia-induced alternative splicing in endothelial cells. *PLoS ONE* **7,** e42697–e42697 (2012).
26. Moller-Levet, C. S. *et al.* Exon array analysis of head and neck cancers identifies a hypoxia related splice variant of LAMA3 associated with a poor prognosis. *PLoS Comp Biol* **5,** e1000571 (2009).
27. Hang, X. *et al.* Transcription and splicing regulation in human umbilical vein endothelial cells under hypoxic stress conditions by exon array. *BMC Genomics* **10,** 126 (2009).
28. Hu, X. *et al.* Severe hypoxia exerts parallel and cell-specific regulation of gene expression and alternative splicing in human mesenchymal stem cells. *BMC Genomics* **15,** 303 (2014).
29. Nevo, Y. *et al.* Genome-wide activation of latent donor splice sites in stress and disease. *Nucleic Acids Research* **40,** 10980–10994 (2012).
30. Tian, Q. *et al.* Integrated genomic and proteomic analyses of gene expression in Mammalian cells. *Mol. Cell Proteomics* **3,** 960–969 (2004).
31. Lu, P., Vogel, C., Wang, R., Yao, X. & Marcotte, E. M. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol.* **25,** 117–124 (2007).
32. Schmidt, M. W., Houseman, A., Ivanov, A. R. & Wolf, D. A. Comparative proteomic and transcriptomic profiling of the fission yeast Schizosaccharomyces pombe. *Mol. Syst. Biol.* **3,** 79 (2007).
33. Jüschke, C. *et al.* Transcriptome and proteome quantification of a tumor model provides novel insights into post-transcriptional gene regulation. *Genome Biol* **14,** r133 (2013).
34. Larsson, O., Tian, B. & Sonenberg, N. Toward a genome-wide landscape of translational control. *Cold Spring Harb Perspect Biol* **5,** a012302 (2013).
35. Vogel, C. *et al.* Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol. Syst. Biol.* **6,** 400 (2010).
36. Saliba, A.-E., Westermann, A. J., Gorski, S. A. & Vogel, J. Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Research* **42,** 8845–8860 (2014).
37. Wagner, K. W. *et al.* KDM2A promotes lung tumorigenesis by epigenetically enhancing ERK1/2 signaling. *J. Clin. Invest.* **123,** 5231–5246 (2013).
38. Gao, R. *et al.* Depletion of histone demethylase KDM2A inhibited cell proliferation of stem cells from apical papilla by de-repression of p15INK4B and p27Kip1. *Mol. Cell. Biochem.* **379,** 115–122 (2013).
39. Vougioukalaki, M. *et al.* Tpl2 kinase signal transduction in inflammation and cancer. *Cancer letters* **304,** 80–89 (2011).
40. Yan, S. F. *et al.* Egr-1, a master switch coordinating upregulation of divergent gene families underlying ischemic stress. *Nat. Med.* **6,** 1355–1361 (2000).
41. Krones-Herzig, A. *et al.* Early growth response 1 acts as a tumor suppressor in vivo and in vitro via regulation of p53. *Cancer Res.* **65,** 5133–5143 (2005).
42. Gitenay, D. & Baron, V. T. Is EGR1 a potential target for prostate cancer

therapy? *Future Oncol* **5,** 993–1003 (2009).
43. Ferraro, B., Bepler, G., Sharma, S., Cantor, A. & Haura, E. B. EGR1 predicts PTEN and survival in patients with non-small-cell lung cancer. *J. Clin. Oncol.* **23,** 1921–1926 (2005).
44. Yang, S.-Z., Eltoum, I. A. & Abdulkadir, S. A. Enhanced EGR1 activity promotes the growth of prostate cancer cells in an androgen-depleted environment. *J. Cell. Biochem.* **97,** 1292–1299 (2006).
45. Peng, W.-X. *et al.* Egr-1 promotes hypoxia-induced autophagy to enhance chemo-resistance of hepatocellular carcinoma cells. *Exp. Cell Res.* (2015). doi:10.1016/j.yexcr.2015.12.006
46. Ma, J. *et al.* Targeted knockdown of EGR-1 inhibits IL-8 production and IL-8-mediated invasion of prostate cancer cells through suppressing EGR-1/NF-kappaB synergy. *J. Biol. Chem.* **284,** 34600–34606 (2009).
47. Ng, S.-Y., Bogu, G. K., Soh, B. S. & Stanton, L. W. The long noncoding RNA RMST interacts with SOX2 to regulate neurogenesis. *Mol. Cell* **51,** 349–359 (2013).
48. Ghosal, S., Das, S. & Chakrabarti, J. Long noncoding RNAs: new players in the molecular mechanism for maintenance and differentiation of pluripotent stem cells. *Stem Cells Dev.* **22,** 2240–2253 (2013).
49. Liu, R., Loraine, A. E. & Dickerson, J. A. Comparisons of computational methods for differential alternative splicing detection using RNA-seq in plant systems. *BMC Bioinformatics* **15,** 364 (2014).
50. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Research* **22,** 2008–2017 (2012).

Chapter 6. Appendix

6.1 Supplementary Figures for Chapter 2

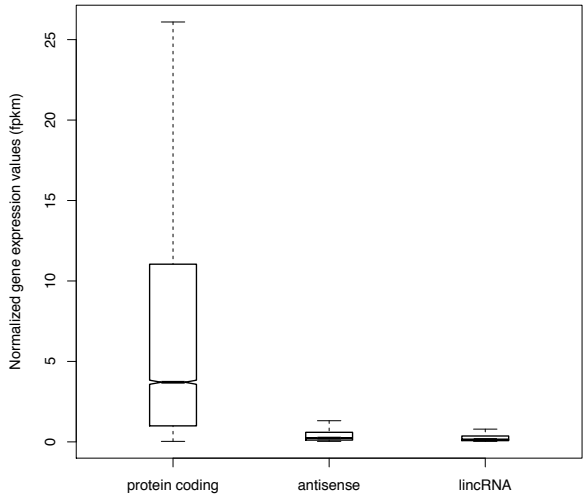

Figure S2.1 Comparison of abundance of protein-coding transcripts, antisense transcripts and lincRNAs in the BodyMap RNA-Seq data.

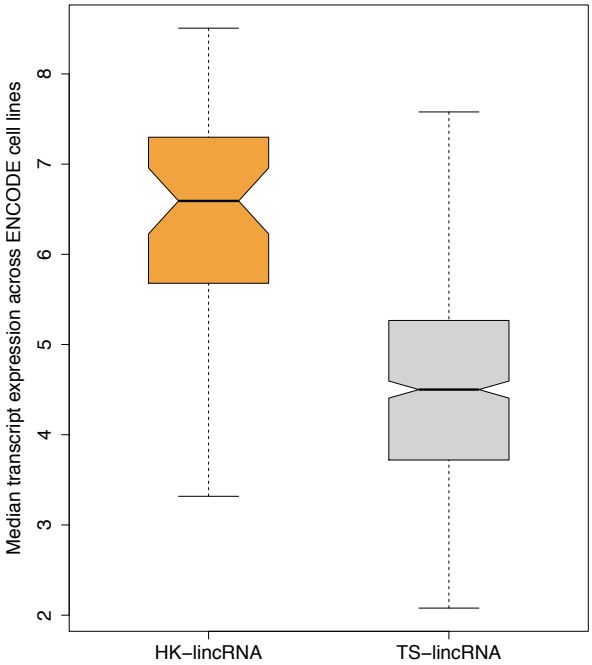

Figure S2.2 Comparison of median transcript expression levels, across ENCODE cell lines, of HK-lincRNAs and TS-lincRNAs.

6.2 Supplementary Figures for Chapter 3

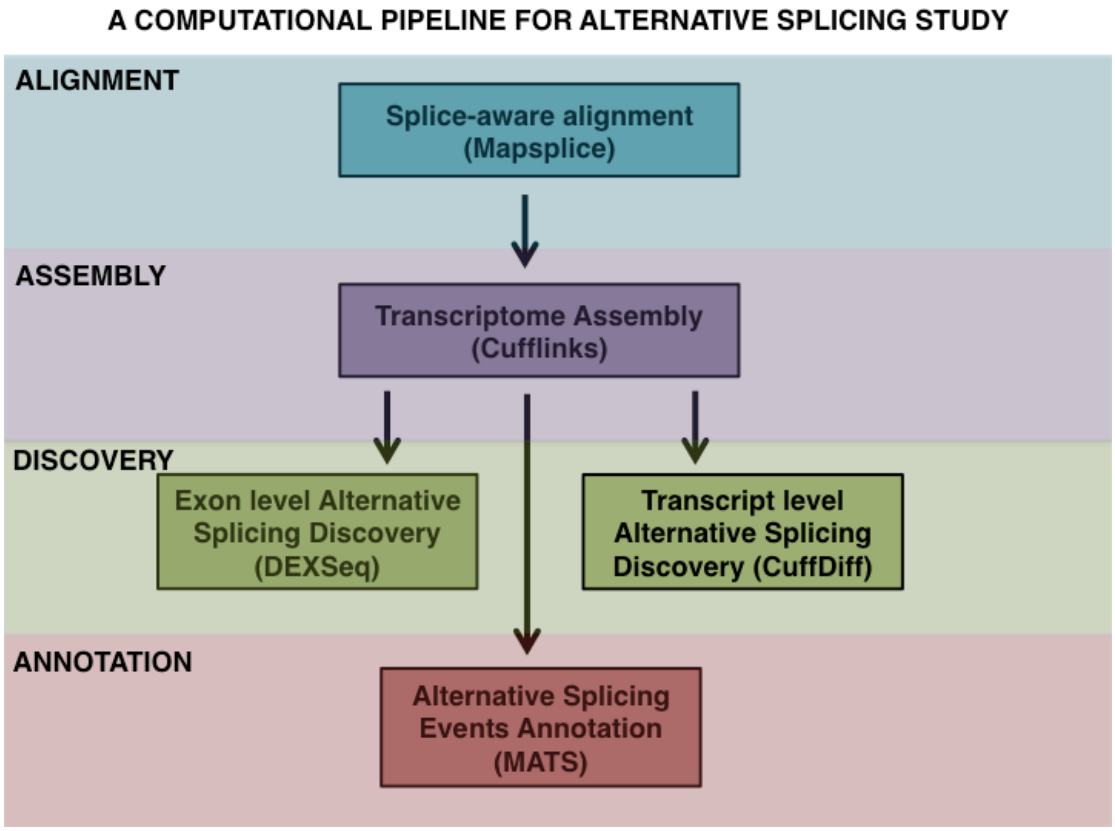

Figure S3.1 The pipeline used for discovery and annotation of alternative splicing events in hypoxia.

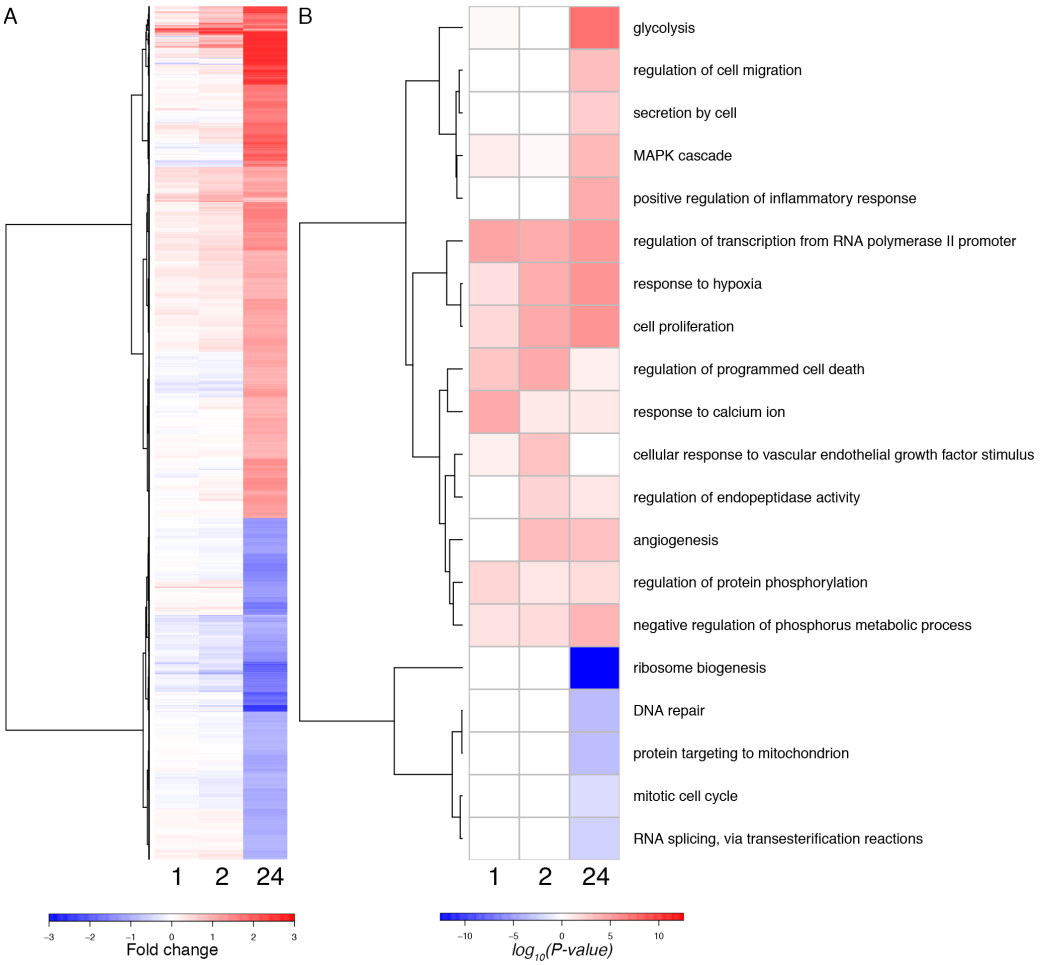

Figure S3.2 Global transcriptional changes for protein-coding loci in response to hypoxia. (A) Hypoxia-dependent differentially expressed protein-coding genes at 1, 2, or 24 hours. Rows represent the log2 fold changes (mean of three biological replicates) plotted relative to $t = 0$. Columns are ordered according to time point. (B) Key biological processes that are up- or down-regulated in response to hypoxia at 1, 2, or 24 hours. Rows represent the $-log_{10}(P\text{-value})$ of up-regulated GO terms and $log_{10}(P\text{-value})$ of down-regulated GO terms. Columns are ordered according to time point. Only non-redundant biological processes below a BH corrected *p*-value cutoff of 0.01 are represented. See also Supplementary Table 3.1.

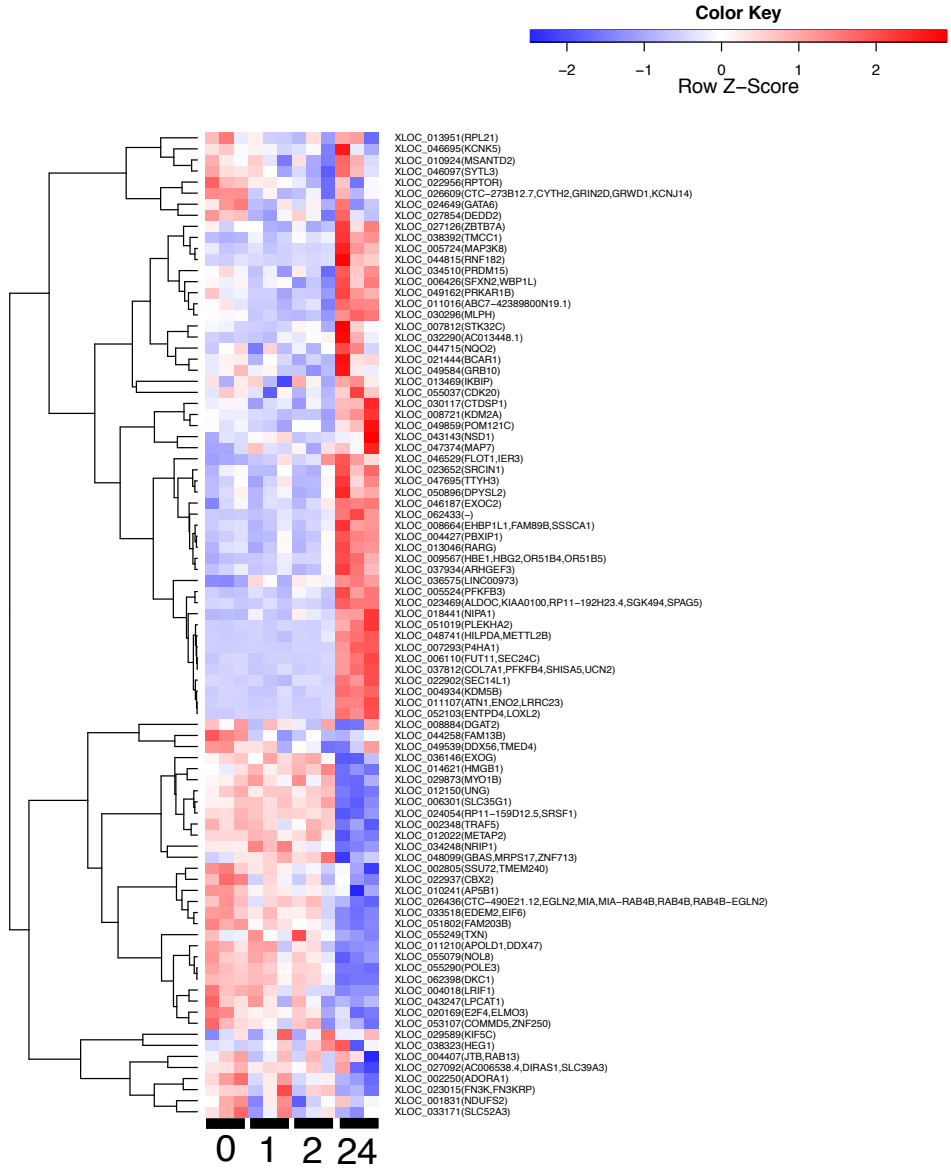

Figure S3.3 Expression profile of genes undergoing differential promoter usage in hypoxia. Rows represent genes, columns ordered by timepoint. Cells are coloured by the RPKM z-score.

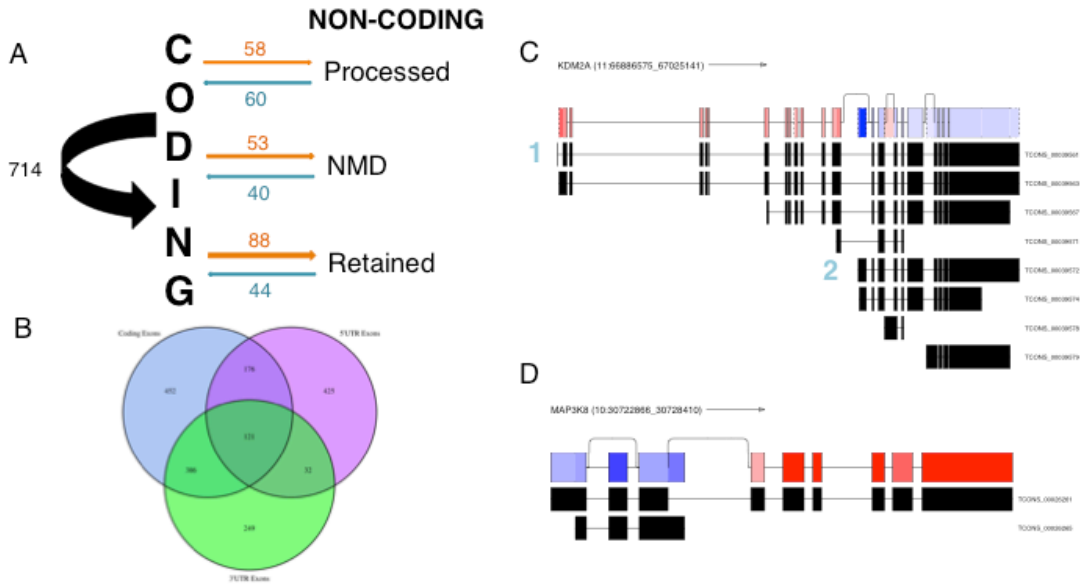

Figure S3.4 Alternative splicing changes in response to hypoxia. (A) Change in major isoform class between hypoxia and and normoxia samples; Orange: enriched in hypoxia samples. (B) No. of protein-coding genes which have differentially used exons at 24 hours in hypoxia relative to 0 hour time-point as identified by DEXSeq. The genes have been grouped according to exon annotations derived from Ensembl (v74) as either Coding Exons, 5′ UTR and 3′ UTR. (C) Predicted transcript structures are represented in black. Normalized fold changes between normoxia (blue) and hypoxia (red) are shown in colour in the top row of each plot, with exon connectivity as determined from the RNA sequencing data used to generate exon links. (C) KDM2A. A novel alternate isoform of the lysine specific demethylase KDM2A lacking the Jumanji domain was expressed in normoxia (1). Elevated levels of the canonical full-length transcripts were observed in hypoxia (2). (D) A novel short isoform of MAP3K8/Tpl2/Cot Kinase missing the kinase domain was detected in normoxia, with the canonical full-length isoform being detected only in hypoxia.

6.3 Supplementary Figures for Chapter 4

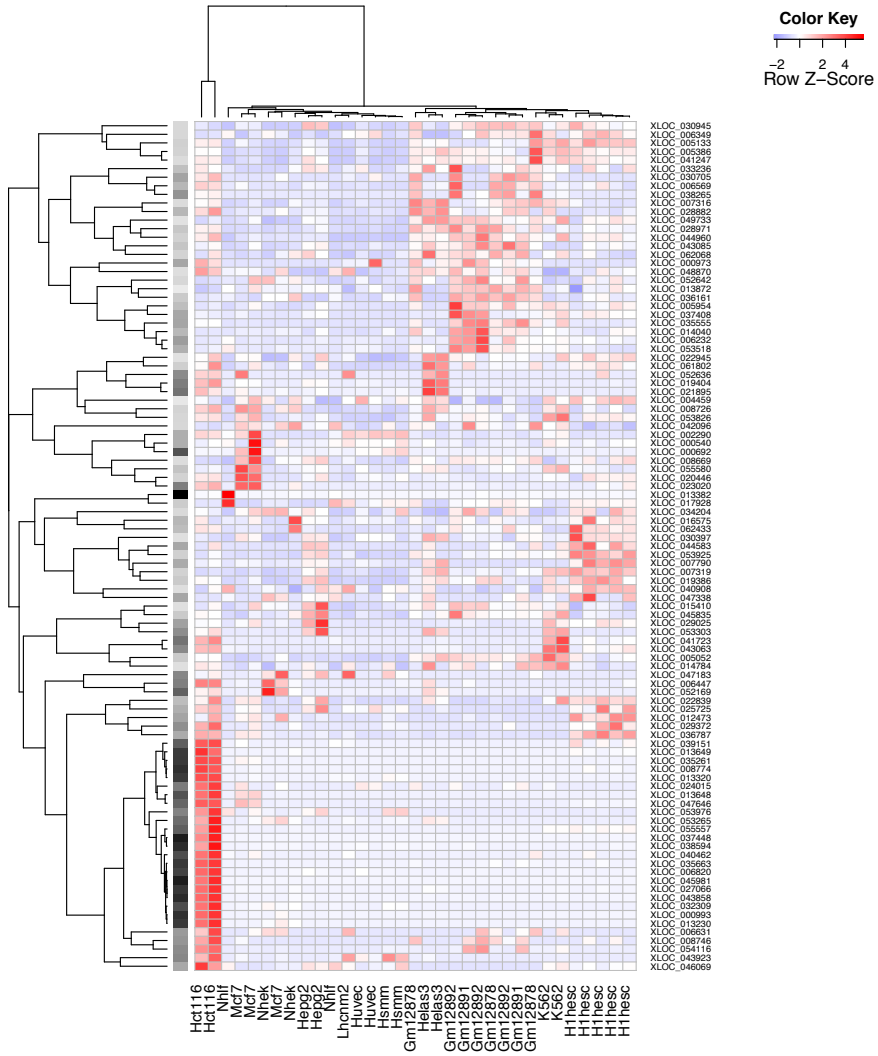

Figure S4.1 Expression profile of novel genes in ENCODE Caltech dataset. Novel genes with detectable expression across 14 cell lines including HCT116 in paired-end RNA-Seq data obtained from ENCODE Caltech dataset. Rows represent novel genes and columns represent untreated cell lines. All cell lines other than Lhcnm2 have more than one biological replicates. Cells in the heatmap are coloured by the RPKM *z*-score. The side bar represents the coefficient of variation of gene expression calculated from the RPKM values. Darker cells correspond to higher values and are indicative of greater cell line-specific expression of the gene.

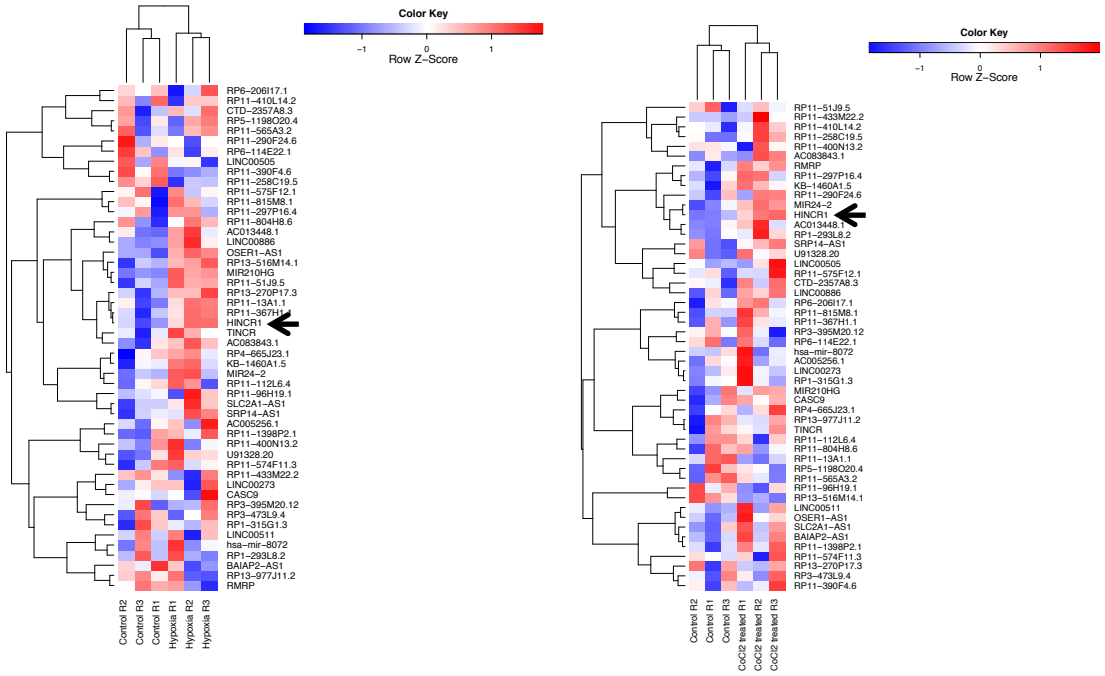

Figure S4.2 Expression profile of Hypoxia Induced Non-coding RNAs (HINCRs) from our study in publically available exon array data of HUVEC cells on treatment of hypoxia. Only 50 out of 59 HINCRs with distinct ENSEMBL ID and reliable probesets were used for this analysis. HINCR1 was found to be induced in HUVEC cells treated with a) hypoxia (t-test p-value < 0.02) and b) 1% CoCl2 (t-test p-value < 0.005).

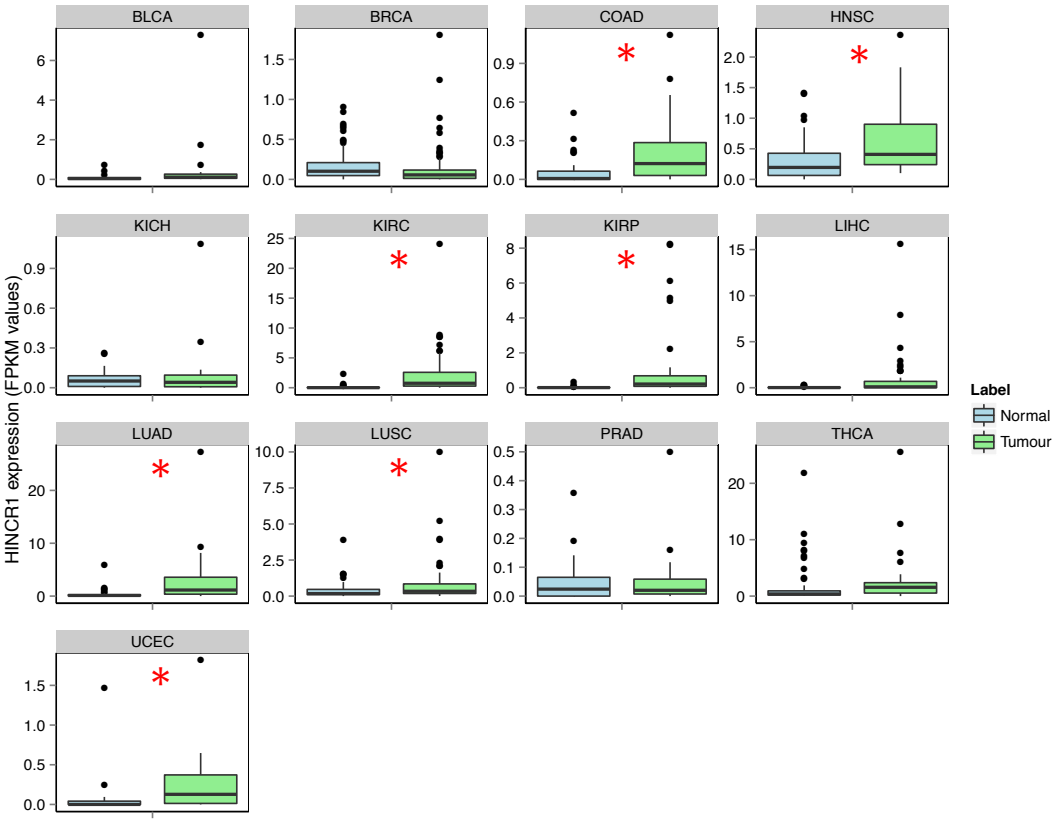

Figure S4.3 Comparison of HINCR1 expression level between matched normal and tumour samples across 13 different tumour types obtained from TCGA. * indicates tumour type wherein HINCR1 levels were found to be significantly different between the matched samples using paired t-test (p-value < 0.01).

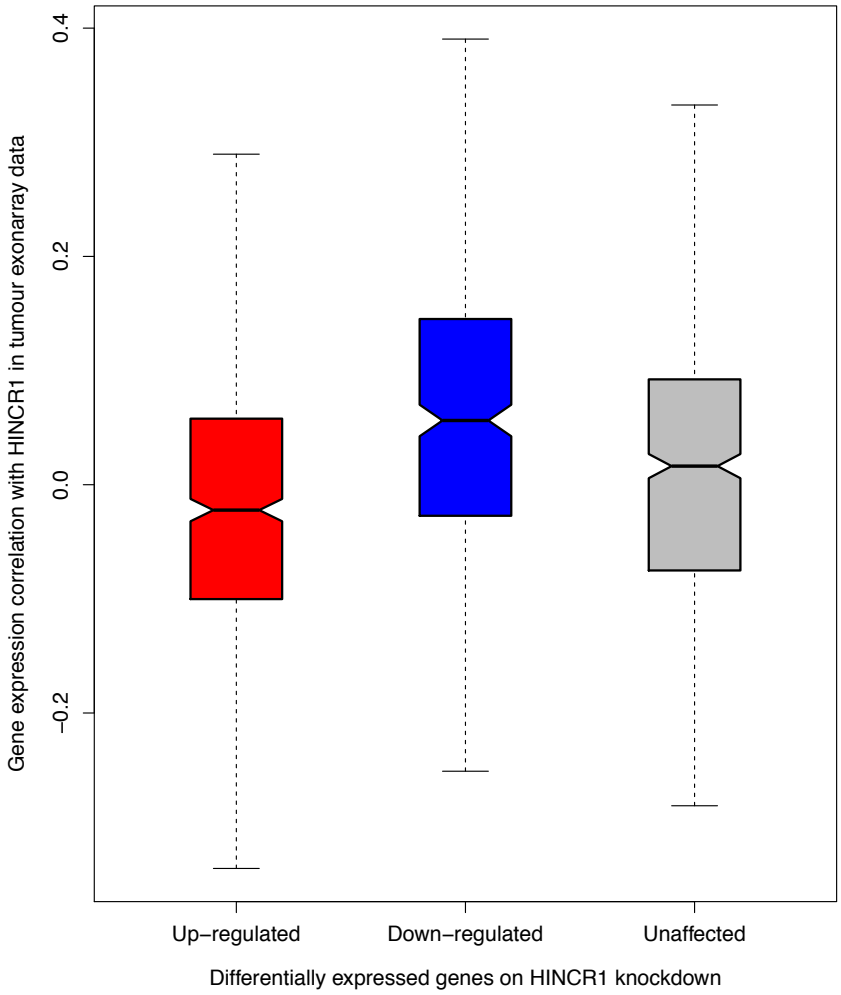

Figure S4.4 Gene expression correlation of up-regulated genes, down-regulated genes and unaffected genes (on knockdown of HINCR1 at 24 hr time-point) with HINCR1 in independent tumour exon array datasets. The notch in each boxplot represents the confidence interval of the median. The down-regulated genes had a more significant positive correlation with HINCR1 as compared to the unaffected genes/up-regulated genes (Wilcox test; p-value < 10^{-6}).

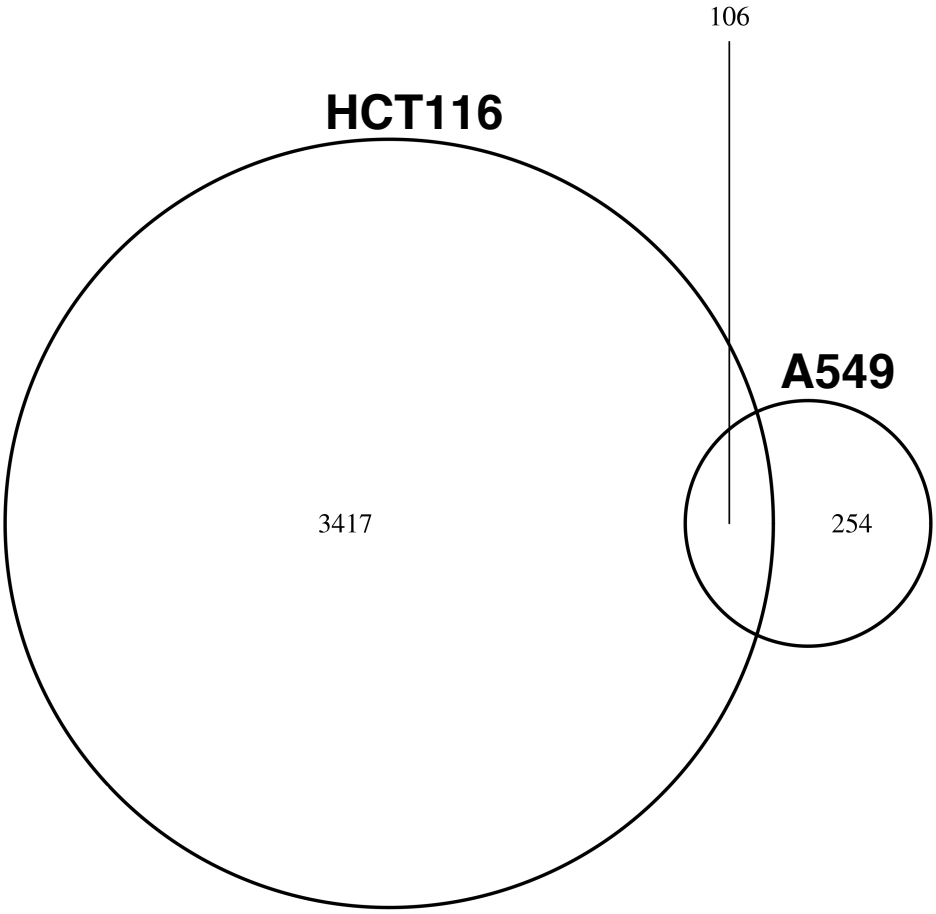

Figure S4.5 Overlap between differentially expressed genes detected on knockdown of HINCR1 in HCT116 and A549 cells. For both the cell lines RNA was sequenced at 0, 2 and 24 hrs in hypoxia after treating cells with HINCR1-siRNAs and scrambled-siRNAs separately. A linear model was developed in edgeR to identify gene expression changes across the time-course in response to HINCR1 knockdown after negating the effect of hypoxia.

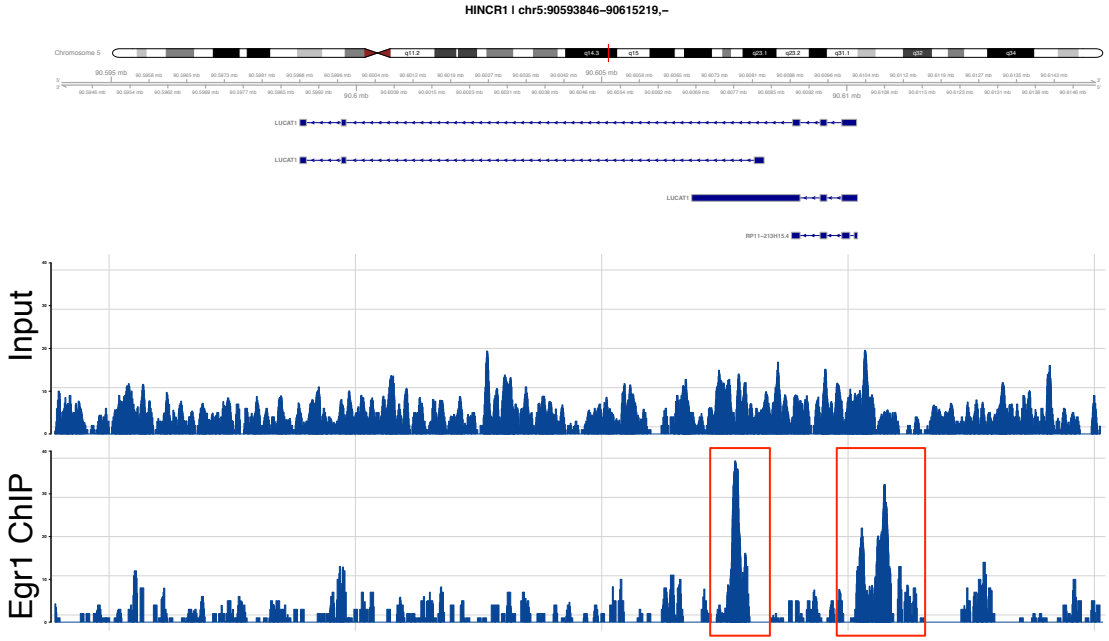

Figure S4.6 Egr1 binding peaks at the HINCR1 locus at the 2 hr timepoint. The plot indicates the HINCR1 transcripts annotated in Ensembl(v74). The Egr1 peaks predicted by MACS2 are indicated in red. The plot also displays the matched input DNA sequenced along the Egr1 ChIP.