UNIVERSITY OF CALIFORNIA
RIVERSIDE

Investigation of LncRNA Isoform-Specific Alterations and MYC Acetylation in Cancer

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Cell, Molecular and Developmental Biology

by

Michael Hamilton

September 2018

Dissertation Committee:
Dr. Ernest Martinez, Chairperson
Dr. Frances M. Sladek
Dr. Fedor V. Karginov
Dr. Thomas Girke

The Dissertation of Michael Hamilton is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

Special thanks to Dr. Ernest Martinez for his support and advice throughout my years as a graduate student. Without his guidance and friendship, I would not have been as successful moving forward in my career as an academic researcher. Thank you to all the members in the Martinez lab (current and past) that aided me with my experiments and provided me with an excellent work environment. Thank you to the professors on my committee, Drs. Sladek, Girke, Karginov and Martinez. Our yearly meetings were of great value and provided excellent feedback to the advancement of our research.

Dedication

To my wife, Melissa, and our two children Sydney and Dillan that supported me throughout the years.

ABSTRACT OF DISSERTATION

Investigation of LncRNA Isoform-Specific Alterations and MYC Acetylation in Cancer

by

Michael Hamilton

Doctor of Philosophy, Graduate Program in Cell, Molecular and Developmental Biology
University of California, Riverside, September 2018
Dr. Ernest Martinez, Chairperson

The events leading to the development of cancer often involve a series of sequential molecular derailments. The complexities of these derailments are frequently specific not only to the cancer type, but can also be specific to an individual tumor. However, despite these challenges, investigations characterizing the global or universal aberrations seen in human cancers have provided many effective therapeutic strategies to the combat human malignancies. The current dissertation focuses on two neglected areas of study in the cancer biology field, the study of lncRNA isoform-specific alterations and the post-translational modifications of the critically important MYC oncoprotein in cancer.

In the beginning chapters of this dissertation, we explore isoform-specific alterations in a subtype of renal cell carcinomas, known as clear cell renal cell carcinoma (ccRCC). ccRCC is one of the most prevalent cancers within the United States, and can be particularly difficult to treat with conventional therapies. As such, new therapeutics strategies are needed to treat ccRCC in its

later stages of the disease. ccRCC has been shown to have severe aberrant RNA production and processing, lending itself as a prime candidate to explore isoform-specific alterations. Furthermore, using new computational methods, we identified previously uncharacterized events of differential transcript expression and usage in ccRCC implicating several novel genes in the pathology. Discovered within these transcriptomic analyses was a long non-coding RNA, referred to as *HOXA Transcript Antisense RNA, Myeloid-Specific 1* (*HOTAIRM1)*, which was found to be specifically downregulated in ccRCC and regulates key genes involved in the hypoxia pathway. In chapter 3, we investigate *HOTAIRM1* further examining its function in ccRCC and its role in kidney cell differentiation and maintenance.

In the last two chapters, we discuss and evaluate the effects of two disparate forms of MYC regulation. First, we discuss the complex nature of the interplay between MYC expression and numerous lncRNAs, in what is referred to as the lncRNA-MYC network. In this review, we reveal the breadth of the complexities of how MYC is regulated by lncRNAs and how MYC regulates its oncogenic function through the use of lncRNAs. Finally, with the development of innovative transgenic cell lines, expressing different mutant forms of MYC, we demonstrate that different acetylation states of MYC can have gene-selective effects and consequently alter different molecular pathways.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Transcript-specific alterations in cancer

Since the development of next-generation technologies, investigations exploring global gene expression changes in human cancer have been extensive. These analyses have generated significant insights in the universal gene expression changes common among several cancers and also the gene expression changes that are specific to individual cancers. However, despite these advancements little attention has been given to the transcript-specific alterations that occur. While there are likely several reasons for the paucity in transcript-specific investigations, one significant contributing factor was the lack of fast and accurate computational techniques for such analyses.

However, within the last two years new accurate computational methods have been developed to tackle these challenges observed with transcript-level quantifications. One such program, referred to as kallisto, uses a *k*-mer based approach, and it is currently one of the most accurate and expeditious transcript quantification programs available [1]. In short, kallisto works by assigning k-mers to compatibility classes (determines what transcript(s) the *k*-mers could belong to) and then determines the intersection of *k*-mer compatibility classes to assign a read to a specific transcript.

Abnormal RNA splicing and processing is a common occurrence among several human cancers [2-4]. Consequently, such altered RNA production and metabolism can result in significant changes to the function of coding and non-coding RNAs. For example, an alternative spliced isoform of pyruvate kinase,

referred to as *PKM2*, was found to be exclusively expressed in tumor tissues and cancer cell lines and was discovered to be necessary for an enhanced oxidative phosphorylation phenotype [5]. Furthermore, *in vivo* experiments demonstrated that *PKM2* provided a significant tumor growth advantage, and switching to the expression of the other isoform of pyruvate kinase, referred to as *PKM1,* reduced tumorigenicity. This phenomenon of use or exchange of different isoforms in not limited solely to cancer, it can be also highly influential in a normal cellular environment. For example, in a recent study examining the gene regulatory mechanisms of meiosis, they found that "toggling" back-and-forth between translatable and non-translatable transcripts of several genes was shown to be one way in which developing cells modulate protein levels [6].

Many of the past and current studies interested in transcript-specific alterations have focused on identifying novel splicing events and/or exploring the expression changes that occur with individual transcripts. While programs, such as Cufflinks and jSplice, have been developed to address the former, the current dissertation focuses solely on the global expression changes of currently annotated individual transcripts [7-8]. There are two terms frequently used to describe transcript expression changes. The general term referred to as differential transcript expression (DTE) is any change in the expression of a transcript, regardless of its locus of origin, between two conditions. Second, is a narrowly defined type of DTE referred to as differential transcript usage (DTU), which is the proportional change of a transcript relative to the transcripts derived from the same locus [9].

An example of an extreme case of DTU is commonly known as isoform switching. Isoform-switching is defined as a switch in the primary transcript (most abundant) of given a gene, in which a previously minor transcript becomes the most abundant transcript for that gene.

Clear cell renal cell carcinoma

One promising cancer model system to study the events of altered RNA processing and production is in a renal cell carcinoma subtype, known as clear cell renal cell carcinoma (ccRCC). Renal cell carcinomas are one of the top 10 most prevalent cancers within the United States, and ccRCCs comprise the majority of all renal cell carcinomas [10-12]. RCCs are cancers derived predominantly from the proximal tubule cells from the cortex of the kidney and carry a relative good prognosis provided the cancer is discovered in its earlier stages [13, 14]. However, relative to others cancers, RCCs are rather resistant to conventional therapeutic strategies, such as chemotherapy and radiation, and can be particularly difficult treat in their later stages [15].

One characteristic feature of ccRCCs is the frequently mutated von Hippel-Lindau (*VHL*) gene, seen with ~55% of ccRCCs. VHL is an E3 ubiquitin ligase, which tightly regulates the stability of a family of transcription factors, known as hypoxia induced factors (HIFs). HIFs are responsible for the transcription of several genes of which diversely contribute to cancer progression, regulating proliferation, angiogenesis, metabolism and metastasis [16]. Additionally, one

pronounced abnormal RNA processing event seen in ccRCC is widespread transcriptional read-through, in which transcription fails to terminate at the canonical termination site and proceeds to an alternative downstream termination site [3]. These events frequently lead to the formation of chimeric RNAs in ccRCC and are partially attributed to the relatively high frequency mutations observed within the histone methyltransferase, known as SET domain containing 2 (*SETD2*) [3, 17]. Moreover, many of the splicing factors, including many of the SR proteins, have altered expression in ccRCC [18-20]. In a recent example, overexpression of a splicing factor referred to as splicing factor 3b subunit 3 (*SF3B3*) (commonly upregulated in ccRCC tissue samples), increased the inclusion of alternative exon in the EZH2 transcript, and consequently increased the proliferative, migratory and tumorigencity of commensurate renal cell lines [18].

Furthermore, as changes in gene-level expression changes may not fully capture the changes in the transcriptome that result from these aberrant RNA processing events, studies that assess steady-levels of individual transcripts in ccRCC are warranted. Many of the earlier studies exploring transcript-specific changes in ccRCC have relied on gene microarray platforms, using differential exon usage (DEU) as a surrogate to determine changes to transcript abundances [21-23]. However, there are several pitfalls to this approach for identifying DTE and DTU. First, few annotated transcripts of a given gene are typically evaluated for their expression on microarray platforms. Frequently, only the “best-characterized”

transcripts are included in gene microarrays, which could lead to the loss of identifying relevant transcripts pertinent to the tissue or cancer of interest. Second, as a consequence of including few transcripts in microarray platforms, the accuracy of identifying events of DTU is questionable. The percentage change of a transcript, relative to other transcripts for a gene, is highly reliant on a comprehensive list of all transcripts derived from that locus. For example, if only 2 transcripts were evaluated for a gene, the percentage changes of the transcripts would likely be different than if 5 transcripts were considered for that gene, which could lead to dramatically different DTU results. Lastly, and possibly most important, exons are typically shared across more than a single transcript; therefore, many of expression differences discovered with the DEU approach are the result of expression changes of more than one individual transcript. In next chapter, we use a multifaceted approach using new computational techniques, that avoid many of the problems see with previous analyses, to provide a more accurate and comprehensive assessment of transcript-specific alterations observed in ccRCC.

HOXA transcript antisense RNA, myeloid-specific 1

Among the results from the aforementioned computational analyses, examining transcript-specific alterations in ccRCC, our lab discovered an uncharacterized deregulated lncRNA in ccRCC, referred to as HOXA transcript antisense RNA, myeloid-specific 1 (*HOTAIRM1 or HM1*). Fortuitously, concomitant functional

studies exploring the role(s) of *HM1* in developing pluripotent mouse stem cells and human cancer cells was already in progress.

HM1 is a relatively highly conserved lncRNA gene, found among several other lncRNA genes, within the HOXA cluster. *HM1* is located between *HOXA1* and *HOXA2* and produces three major isoforms: two spliced products with either two exons (*HM1-3*) or three exons (*HM1-2-3*) and a full length unspliced transcript. *HM1* was first characterized in 2006 in a study investigating the intergenic transcriptional events within the HOXA cluster [24]. Here, the investigators found a collinear expression of *HM1,* and the other ncRNAs, with the surrounding HOXA genes upon induction with a morphogen, known as retinoic acid. The same initial study also discovered concomitant changes to the epigenetic landscape of the HOXA cluster, with increases in H3K4me2, H3K4me3 and loss of the occupancy of the Polycomb Repressive Complex 2 (PRC2).

In 2009, *HM1* was demonstrated to play a role in myeloid differentiation by attenuating the induction of *HOXA1*, *HOXA4* and other myeloid differentiation genes [25]. At this time, *HM1* was given its name as a myeloid specific lncRNA due to its high expression in differentiating myeloid cells. These initial findings were further supported by later studies establishing that *HM1* regulates cell cycle genes and activates the proximal HOXA genes in differentiating NB4 cells [26, 27]. In an alternative study, also in NB4 cells, *HM1* was demonstrated to modulate myeloid differentiation by controlling the degradation of the PML-RARA oncoprotein. *HM1* was found to act as a miRNA sponge for miR-20a/106b and

miR-125b regulating the expression of key genes (*E2F1* and *DRAM2*) involved in the formation of autophagosomes [28]. However, it is not entirely clear from these studies which of the *HM1* isoforms are responsible for the observed effects, as the loss-of-function assays performed do not exclusively alter only one of the *HM1* isoforms. This remains a vital question to understanding the mechanistic actions of *HM1*, as these studies and our results, support different cellular localizations of the *HM1* isoforms, and most likely different mechanistic roles.

While our mechanistic understanding of *HM1* in differentiating cells has made some steady progress, the function(s) of *HM1* in most human cancers remains largely unknown. *HM1* expression is frequently deregulated in human cancers; however, there is little consistency among the cancers as to the nature of the deregulation. In glioblastomas, pancreatic and leukemic cancers *HM1* expression is frequently upregulated, while colon, breast, head and neck tumors often exhibit a downregulation of *HM1* [29-34]. One study that explored the mechanistic role(s) of *HM1* used FaDu cells (a hypopharyngeal squamous cell carcinoma cell line), and suggested *HM1* was acting as a miR-148a sponge regulating the expression of DLG Associated Protein 1 (*DLGAP1*), a gene known for its role in the assembly and stability of synapses [33]. However, similar to the reasons stated above, many of the *HM1* studies in cancer do not clarify which of the *HM1* transcripts are being evaluated. Moreover, many of the studies also do not consider the absolute expression of *HM1* in these cancers, as differential

expression of *HM1* would be of questionable clinical significance if the expression of *HM1* very low. In chapter 4, we attempt to provide better clarity to the functional role of one of the *HM1* isoforms (*HM1-3)* and investigate its role in ccRCC.

The lncRNA-MYC network in cancer

As our study investigating *HM1* in ccRCC and numerous other studies suggest a functional role of lncRNAs in human cancers, our lab sought review all of the studies linking lncRNAs to the critically important MYC oncoprotein [35]. *MYC*, also commonly known as *c-MYC,* is one of the most studied genes in cancer research. *MYC* encodes a basic helix-loop-helix transcription factor that is critical to numerous cellular activities including: cell growth, proliferation, apoptosis, transformation and metabolism [36-38]. In review of the literature, we discovered a rather complex network of regulation between lncRNAs and MYC.

The *MYC* locus is located on the long arm of human chromosome 8 in a region referred to as a "gene desert," as few protein-coding genes reside in the region. However, there are several lncRNA genes surrounding the *MYC* locus that play a vital role in regulating *MYC* mRNA and protein expression. Discussed in the review, *MYC* expression is regulated at the transcriptional and post-transcriptional levels by lncRNAs. In one example of a neighboring lncRNA gene, the *CCAT1* lncRNA was found to facilitate the transcription of the *MYC* locus by maintaining local chromatin interactions via the recruitment of a

transcription factor, known as CCCTC-Binding factor or CTCF [39]. Additional forms of MYC regulation include post-transcriptional mechanisms such as, lncRNAs acting as miRNA sponges for *MYC* mRNA, and also lncRNAs binding directly to MYC regulate the degradation of the protein. Conversely, we also found studies supporting MYC as a transcription factor for many lncRNA genes.

MYC acetylation

An alternative area of growing interest in MYC research focuses on understanding how post-translational modifications (PTMs) alter MYC stability and activity. PTMs that facilitate the degradation of MYC, such as ubiquitination, sumoylation and phosphorylation, are the subject of great interest, as increased MYC stability is found in many human cancers [40-42]. Additionally, PTMs can also modulate the transcriptional and transformative activities of MYC. For example, studies have shown that MYC phosphorylation at Ser-373 (by protein kinase C ζ) alters prostate tumorigenesis, and also Cdk5-mediated phosphorylation of MYC Ser-62 was essential for transcriptional activation of cyclin B1 [43, 44]. These studies demonstrate phosphorylation of MYC, and likely other post-translational modifications, are key determinants in MYC transcriptional activity and tumorigenesis.

The role of MYC acetylation is less understood, relative to the other PTMs. Histone acetyltransferases (HATs), such as CBP/p300, GCN5 and TIP60, have been shown to physically interact with MYC, and have been thought to facilitate

MYC-mediate transcription via the acetylation of local histones [45-47]. While MYC was also found to be directly acetylated, MYC acetylation did not alter its binding to DNA nor was MYC acetylation observed to affect dimerization with MAX [45, 48]. However, reductions in MYC ubiquitination were observed with concomitant MYC acetylation [45]. As both acetylation and ubiquitination both occur on lysines, it was recently suggested that MYC acetylation is likely interfering with MYC ubiquitination thereby contributing to enhanced MYC stability [42]. There are several lysine residues on the human MYC that are acetylated by p300, including K143, K157, K275, K317, K323 and K371 [49]. Despite these initial findings, it remains unclear whether MYC acetylation alters its transcriptional activity. In the final chapter of the dissertation, we examine the global transcriptomic effects of overexpressing three novel MYC acetylation mutants in Rat1a cells.

Objectives of the dissertation

There were two main objectives to exploring transcript-specific alterations in ccRCC. The first goal was to provide a comprehensive resource of all differentially expressed transcripts, including all putative coding and non-coding transcripts, and identify the key molecular programs altered in ccRCC. The second goal of the study was to highlight the importance of transcript analyses by demonstrating an increased sensitivity to detecting transcriptomic alterations using gene-level and transcript-level analyses in parallel.

In chapter 3, the main goal of investigating *HM1* was to discover its function in human cancer. Discovery of a novel downregulation of *HM1-3* in ccRCC led to our second goal in the study, characterizing the molecular changes that occur with loss and gain-of-function of *HM1* in renal proximal tubule cell lines.

The objectives of reviewing the literature exploring the link between MYC and lncRNAs was to highlight the mechanisms in which lncRNAs regulate *MYC* expression and to explore the MYC-induced lncRNA genes that likely contribute to the transformative abilities of MYC. Finally, the last goal of the dissertation was to explore the suspected gene-selective effects elicited by overexpression of different mutant forms of MYC with altered acetylation states.

References

1. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016; 34: 525-7. doi: 10.1038/nbt.3519.

2. Yanagisawa M, Huveldt D, Kreinest P, Lohse CM, Cheville JC, Parker AS, Copland JA, Anastasiadis PZ. A p120 catenin isoform switch affects Rho activity, induces tumor cell invasion, and predicts metastatic disease. J Biol Chem. 2008; 283: 18344-54. doi: 10.1074/jbc.M801192200.

3. Grosso AR, Leite AP, Carvalho S, Matos MR, Martins FB, Vitor AC, Desterro JM, Carmo-Fonseca M, de Almeida SF. Pervasive transcription read-through promotes aberrant expression of oncogenes and RNA chimeras in renal carcinoma. Elife. 2015; 4. doi: 10.7554/eLife.09214.

4. Han B, Park HK, Ching T, Panneerselvam J, Wang H, Shen Y, Zhang J, Li L, Che R, Garmire L, Fei P. Human DBR1 modulates the recycling of snRNPs to affect alternative RNA splicing and contributes to the suppression of cancer development. Oncogene. 2017; 36: 5382-91. doi: 10.1038/onc.2017.150.

5. Christofk HR, Vander Heiden MG, Harris MH, Ramanathan A, Gerszten RE, Wei R, Fleming MD, Schreiber SL, Cantley LC. The M2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth. Nature. 2008; 452: 230-3. doi: 10.1038/nature06734.

6. Cheng Z, Otto GM, Powers EN, Keskin A, Mertins P, Carr SA, Jovanovic M, Brar GA. Pervasive, Coordinated Protein-Level Changes Driven by Transcript Isoform Switching during Meiosis. Cell. 2018; 172: 910-23 e16. doi: 10.1016/j.cell.2018.01.035.

7. Christinat Y, Pawlowski R, Krek W. jSplice: a high-performance method for accurate prediction of alternative splicing events and its application to large-scale renal cancer transcriptome data. Bioinformatics. 2016; 32: 2111-9. doi: 10.1093/bioinformatics/btw145.

8. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012; 7: 562-78. doi: 10.1038/nprot.2012.016.

9. Froussios K, Mourao K, Schurch NJ, Barton GJ. Identifying differential isoform abundance with RATs: a universal tool and a warning. bioRxiv. 2017. doi: 10.1101/132761.

10. U.S. Cancer Statistics Working Group. (2017). United States Cancer Statistics: 1999-2014 Incidence and Mortality Web-based Report. Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute).

11. American Cancer Society. (2018). Cancer Facts & Figures 2018. Atlanta, Ga: American Cancer Society).

12. Cancer Genome Atlas Research N. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature. 2013; 499: 43-9. doi: 10.1038/nature12222.

13. Bander NH, Finstad CL, Cordon-Cardo C, Ramsawak RD, Vaughan ED, Jr., Whitmore WF, Jr., Oettgen HF, Melamed MR, Old LJ. Analysis of a mouse monoclonal antibody that reacts with a specific region of the human proximal tubule and subsets renal cell carcinomas. Cancer Res. 1989; 49: 6774-80. doi:

14. van den Berg E, van der Hout AH, Oosterhuis JW, Storkel S, Dijkhuizen T, Dam A, Zweers HM, Mensink HJ, Buys CH, de Jong B. Cytogenetic analysis of epithelial renal-cell tumors: relationship with a new histopathological classification. Int J Cancer. 1993; 55: 223-7. doi:

15. Mickisch GH, Roehrich K, Koessig J, Forster S, Tschada RK, Alken PM. Mechanisms and modulation of multidrug resistance in primary human renal cell carcinoma. J Urol. 1990; 144: 755-9. doi:

16. Yang Y, Sun M, Wang L, Jiao B. HIFs, angiogenesis, and cancer. J Cell Biochem. 2013; 114: 967-74. doi: 10.1002/jcb.24438.

17. Pflueger D, Mittmann C, Dehler S, Rubin MA, Moch H, Schraml P. Functional characterization of BC039389-GATM and KLK4-KRSP1 chimeric read-through transcripts which are up-regulated in renal cell cancer. BMC Genomics. 2015; 16: 247. doi: 10.1186/s12864-015-1446-z.

18. Chen K, Xiao H, Zeng J, Yu G, Zhou H, Huang C, Yao W, Xiao W, Hu J, Guan W, Wu L, Huang J, Huang Q, et al. Alternative Splicing of EZH2 pre-mRNA by SF3B3 Contributes to the Tumorigenic Potential of Renal Cancer. Clin Cancer Res. 2017; 23: 3428-41. doi: 10.1158/1078-0432.CCR-16-2020.

19. Kedzierska H, Poplawski P, Hoser G, Rybicka B, Rodzik K, Sokol E, Boguslawska J, Tanski Z, Fogtman A, Koblowska M, Piekielko-Witkowska A. Decreased Expression of SRSF2 Splicing Factor Inhibits Apoptotic Pathways in Renal Cancer. Int J Mol Sci. 2016; 17. doi: 10.3390/ijms17101598.

20. Piekielko-Witkowska A, Wiszomirska H, Wojcicka A, Poplawski P, Boguslawska J, Tanski Z, Nauman A. Disturbed expression of splicing factors in renal cancer affects alternative splicing of apoptosis regulators, oncogenes, and tumor suppressors. PLoS One. 2010; 5: e13690. doi: 10.1371/journal.pone.0013690.

21. Brito GC, Fachel AA, Vettore AL, Vignal GM, Gimba ER, Campos FS, Barcinski MA, Verjovski-Almeida S, Reis EM. Identification of protein-coding and intronic noncoding RNAs down-regulated in clear cell renal carcinoma. Mol Carcinog. 2008; 47: 757-67. doi: 10.1002/mc.20433.

22. Deng M, Blondeau JJ, Schmidt D, Perner S, Muller SC, Ellinger J. Identification of novel differentially expressed lncRNA and mRNA transcripts in clear cell renal cell carcinoma by expression profiling. Genom Data. 2015; 5: 173-5. doi: 10.1016/j.gdata.2015.06.016.

23. Valletti A, Gigante M, Palumbo O, Carella M, Divella C, Sbisa E, Tullo A, Picardi E, D'Erchia AM, Battaglia M, Gesualdo L, Pesole G, Ranieri E. Genome-wide analysis of differentially expressed genes and splicing isoforms in clear cell renal cell carcinoma. PLoS One. 2013; 8: e78452. doi: 10.1371/journal.pone.0078452.

24. Sessa L, Breiling A, Lavorgna G, Silvestri L, Casari G, Orlando V. Noncoding RNA synthesis and loss of Polycomb group repression accompanies the colinear activation of the human HOXA cluster. RNA. 2007; 13: 223-39. doi: 10.1261/rna.266707.

25. Zhang X, Lian Z, Padden C, Gerstein MB, Rozowsky J, Snyder M, Gingeras TR, Kapranov P, Weissman SM, Newburger PE. A myelopoiesis-associated regulatory intergenic noncoding RNA transcript within the human HOXA cluster. Blood. 2009; 113: 2526-34. doi: 10.1182/blood-2008-06-162164.

26. Wang XQ, Dostie J. Reciprocal regulation of chromatin state and architecture by HOTAIRM1 contributes to temporal collinear HOXA gene activation. Nucleic Acids Res. 2017; 45: 1091-104. doi: 10.1093/nar/gkw966.

27. Zhang X, Weissman SM, Newburger PE. Long intergenic non-coding RNA HOTAIRM1 regulates cell cycle progression during myeloid maturation in NB4 human promyelocytic leukemia cells. RNA Biol. 2014; 11: 777-87. doi:

28. Chen ZH, Wang WT, Huang W, Fang K, Sun YM, Liu SR, Luo XQ, Chen YQ. The lncRNA HOTAIRM1 regulates the degradation of PML-RARA oncoprotein and myeloid cell differentiation by enhancing the autophagy pathway. Cell Death Differ. 2017; 24: 212-24. doi: 10.1038/cdd.2016.111.

29. Diaz-Beya M, Brunet S, Nomdedeu J, Pratcorona M, Cordeiro A, Gallardo D, Escoda L, Tormo M, Heras I, Ribera JM, Duarte R, de Llano MP, Bargay J, et al. The lincRNA HOTAIRM1, located in the HOXA genomic region, is expressed in acute myeloid leukemia, impacts prognosis in patients in the intermediate-risk cytogenetic category, and is associated with a distinctive microRNA signature. Oncotarget. 2015; 6: 31613-27. doi: 10.18632/oncotarget.5148.

30. Novak P, Jensen T, Oshiro MM, Wozniak RJ, Nouzova M, Watts GS, Klimecki WT, Kim C, Futscher BW. Epigenetic inactivation of the HOXA gene cluster in breast cancer. Cancer Res. 2006; 66: 10664-70. doi: 10.1158/0008-5472.CAN-06-2761.

31. Wan L, Kong J, Tang J, Wu Y, Xu E, Lai M, Zhang H. HOTAIRM1 as a potential biomarker for diagnosis of colorectal cancer functions the role in the tumour suppressor. J Cell Mol Med. 2016; 20: 2036-44. doi: 10.1111/jcmm.12892.

32. Zhang X, Sun S, Pu JK, Tsang AC, Lee D, Man VO, Lui WM, Wong ST, Leung GK. Long non-coding RNA expression profiles predict clinical phenotypes in glioma. Neurobiol Dis. 2012; 48: 1-8. doi: 10.1016/j.nbd.2012.06.004.

33. Zheng M, Liu X, Zhou Q, Liu G. HOTAIRM1 competed endogenously with miR-148a to regulate DLGAP1 in head and neck tumor cells. Cancer Med. 2018. doi: 10.1002/cam4.1523.

34. Zhou Y, Gong B, Jiang ZL, Zhong S, Liu XC, Dong K, Wu HS, Yang HJ, Zhu SK. Microarray expression profile analysis of long non-coding RNAs in pancreatic ductal adenocarcinoma. Int J Oncol. 2016; 48: 670-80. doi: 10.3892/ijo.2015.3292.

35. Hamilton MJ, Young MD, Sauer S, Martinez E. The interplay of long non-coding RNAs and MYC in cancer. AIMS Biophys. 2015; 2: 794-809. doi: 10.3934/biophy.2015.4.794.

36. Bretones G, Delgado MD, Leon J. Myc and cell cycle control. Biochim Biophys Acta. 2015; 1849: 506-16. doi: 10.1016/j.bbagrm.2014.03.013.

37. Dang CV. MYC, metabolism, cell growth, and tumorigenesis. Cold Spring Harb Perspect Med. 2013; 3. doi: 10.1101/cshperspect.a014217.

38. McMahon SB. MYC and the control of apoptosis. Cold Spring Harb Perspect Med. 2014; 4: a014407. doi: 10.1101/cshperspect.a014407.

39. Xiang JF, Yin QF, Chen T, Zhang Y, Zhang XO, Wu Z, Zhang S, Wang HB, Ge J, Lu X, Yang L, Chen LL. Human colorectal cancer-specific CCAT1-L

lncRNA regulates long-range chromatin interactions at the MYC locus. Cell Res. 2014; 24: 513-31. doi: 10.1038/cr.2014.35.

40. Dohmen RJ. SUMO wrestles down myc. Cell Cycle. 2015; 14: 2551-2. doi: 10.1080/15384101.2015.1062331.

41. Gonzalez-Prieto R, Cuijpers SA, Kumar R, Hendriks IA, Vertegaal AC. c-Myc is targeted to the proteasome for degradation in a SUMOylation-dependent manner, regulated by PIAS1, SENP7 and RNF4. Cell Cycle. 2015; 14: 1859-72. doi: 10.1080/15384101.2015.1040965.

42. Farrell AS, Sears RC. MYC degradation. Cold Spring Harb Perspect Med. 2014; 4. doi: 10.1101/cshperspect.a014365.

43. Kim JY, Valencia T, Abu-Baker S, Linares J, Lee SJ, Yajima T, Chen J, Eroshkin A, Castilla EA, Brill LM, Medvedovic M, Leitges M, Moscat J, et al. c-Myc phosphorylation by PKCzeta represses prostate tumorigenesis. Proc Natl Acad Sci U S A. 2013; 110: 6418-23. doi: 10.1073/pnas.1221799110.

44. Seo HR, Kim J, Bae S, Soh JW, Lee YS. Cdk5-mediated phosphorylation of c-Myc on Ser-62 is essential in transcriptional activation of cyclin B1 by cyclin G1. J Biol Chem. 2008; 283: 15601-10. doi: 10.1074/jbc.M800987200.

45. Vervoorts J, Luscher-Firzlaff JM, Rottmann S, Lilischkis R, Walsemann G, Dohmann K, Austen M, Luscher B. Stimulation of c-MYC transcriptional activity and acetylation by recruitment of the cofactor CBP. EMBO Rep. 2003; 4: 484-90. doi: 10.1038/sj.embor.embor821.

46. Adhikary S, Eilers M. Transcriptional regulation and transformation by Myc proteins. Nat Rev Mol Cell Biol. 2005; 6: 635-45. doi: 10.1038/nrm1703.

47. Faiola F, Liu X, Lo S, Pan S, Zhang K, Lymar E, Farina A, Martinez E. Dual regulation of c-Myc by p300 via acetylation-dependent control of Myc protein turnover and coactivation of Myc-induced transcription. Mol Cell Biol. 2005; 25: 10220-34. doi: 10.1128/MCB.25.23.10220-10234.2005.

48. Patel JH, Du Y, Ard PG, Phillips C, Carella B, Chen CJ, Rakowski C, Chatterjee C, Lieberman PM, Lane WS, Blobel GA, McMahon SB. The c-MYC oncoprotein is a substrate of the acetyltransferases hGCN5/PCAF and TIP60. Mol Cell Biol. 2004; 24: 10826-34. doi: 10.1128/MCB.24.24.10826-10834.2004.

49. Zhang K, Faiola F, Martinez E. Six lysine residues on c-Myc are direct substrates for acetylation by p300. Biochem Biophys Res Commun. 2005; 336: 274-80. doi: 10.1016/j.bbrc.2005.08.0

Chapter 2

Global isoform-specific transcript alterations and deregulated networks in clear cell renal cell carcinoma

Published in Oncotarget. 2018; 9:23670-

23680. https://doi.org/10.18632/oncotarget.25330

Abstract

Extensive genome-wide analyses of deregulated gene expression have now been performed for many types of cancer. However, most studies have focused on deregulation at the gene-level, which may overlook the alterations of specific transcripts for a given gene. Clear cell renal cell carcinoma (ccRCC) is one of the best-characterized and most pervasive renal cancers, and ccRCCs are well-documented to have aberrant RNA processing. In the present study, we examine the extent of this aberrant RNA processing by reporting a comprehensive transcript-level analysis, using the new kallisto-sleuth-RATs pipeline, investigating coding and non-coding differential transcript expression in ccRCC. We analyzed 50 ccRCC tumors and their matched normal samples from The Cancer Genome Altas datasets. We identified 7,339 differentially expressed transcripts and 94 genes exhibiting differential transcript isoform usage in ccRCC. Additionally, transcript-level coexpression network analyses identified vasculature development and the tricarboxylic cycle as the most significantly deregulated networks correlating with ccRCC progression. These analyses uncovered several uncharacterized transcripts, including lncRNAs *FGD5-AS1* and *AL035661.1,* as potential regulators of the tricarboxylic acid cycle associated with ccRCC progression. As ccRCC still presents treatment challenges, our results provide a new resource of potential therapeutics targets and highlight the importance of exploring alternative methodologies in transcriptome-wide studies.

Introduction

Renal cancer is one of the ten most frequently occurring cancers found in both males and females in the United States [1]. In 2018, an estimated 65,340 new cases of renal cancer will be diagnosed within the US with ~96% of them being renal cell carinomas (RCC) [2]. Most RCC tumors originate from the epithelial cells of proximal tubules within the cortex of the kidney, and RCCs carry with them several therapeutics challenges [3, 4]. Specifically, both chemotherapy and radiation treatments are largely ineffective, patients can be frequently asymptotic, and metastatic RCC has a relatively high 5-year mortality rate of >90% [5]. Among the four major histological RCC subtypes, clear cell renal cell carcinoma (ccRCC) is the most common, observed within 75% of cases [6].

One of the characteristic features of ccRCC is the frequently mutated von Hippel-Lindau (VHL) gene, found within ~50% of ccRCC tumors, or loss of the short arm of chromosome 3 [7-10]. Loss of a functional VHL protein, a E3 ubiquitin ligase, results in enhanced stability of a family of transcription factors, known as hypoxia inducible transcription factors (HIFs) [11]. As a result of elevated HIF proteins, changes to expression levels of several HIF responsive genes can occur, such as vascular endothelial growth factor (*VEGF*), MET proto-oncogene (*c-MET*), and transforming growth factor (*TGF*), altering the pro-angiogenic, invasive and proliferative characteristics of cancer cells. With the advent of large-platform and high-throughout techniques, we have greatly improved our understanding of the VHL/HIF pathway, and we have expanded beyond this classical model to reveal

other key molecular events that occur in ccRCC. In a recent comprehensive study examining ccRCC, an integrative pathway analysis showed one the most frequently mutated subnetworks were genes that influence the epigenetic landscape, such as *PBRM1* and genes in the PBAF SWI/SNF chromatin remodeling complex [7].

However, despite the shift to global gene expression profiling, little attention has been given to examining transcript-specific changes in ccRCC and other cancers, possibly due to the additional computational constraints compared to conventional gene-level analyses. Aberrant transcript isoforms from altered transcription initiation, termination and RNA processing (including altered alternative splicing) are well-documented phenomena found within many cancers [8, 12-15]. Furthermore, abnormal RNA processing events can have profound effects on coding and non-coding RNA species [16, 17]. In a recent example, inactivation of a histone methyltransferase, known as SET domain containing 2 (*SETD2*), was discovered to be one of the inciting causes of widespread transcriptional read-through and abnormal RNA chimera production found in ccRCC [16].

With the advent of alignment-free RNA-Seq quantification algorithms, larger scale and more comprehensive transcript-level analyses can now be performed with a smaller computational footprint. An example is kallisto, one of the fastest and most accurate transcript-level quantification programs. Instead of more time consuming read alignments, it uses a *k*-mer approach for quantifying the

abundance of transcripts in RNA-seq experiments [18]. More recently, two R packages, sleuth and RATs (Relative Abundance of Transcripts), were developed that exploit the bootstrap estimates from kallisto to identify events of differential transcript expression and differential transcript usage, respectively [19, 20]. Differential transcript expression (DTE) is any change in the relative abundance of a transcript between two conditions. Alternatively, differential transcript usage (DTU) is the proportional change of the transcripts that a gene encodes. For example, DTU can frequently result in isoform-switching, in which the major isoform (most abundant) "switches" with an alternative transcript, and thereby that isoform is longer the major isoform of that particular gene. To our knowledge, there are relatively few transcriptome-level studies examining differential transcript expression in ccRCC, and these studies have either relied on microarray platforms or focused largely on one aspect of differential transcript usage (e.g. differential splicing) [21-26]. Importantly, transcript-level analyses can add greater resolution to a transcriptome-wide study, as significant DTE can evade traditional gene-level analysis techniques.

The current study uses a multifaceted approach with new highly accurate computational methods, not employed by previous studies, quantifying all transcript-level alterations in ccRCC, and places these alterations in context key biological pathways involved in ccRCC progression (Figure 2.1A). In doing so, we identified several previously uncharacterized deregulated genes implicated in ccRCC. We analyzed 100 RNA-seq datasets (50 matched pair samples) from

The Cancer Genome Altas (TCGA) with kallisto to quantify all putative coding and non-coding transcripts, sleuth to identify significant differentially expressed transcripts (DETs) and RATs to discover events of differential transcript usage (DTU). We identified 7,339 DETs and 94 DTU genes of which 68 genes are uncharacterized. Furthermore, we performed a comparative differential expression analysis, using both gene-level and transcript-level analyses, and identified novel deregulated genes in ccRCC. Additionally, we performed one of the first weighted transcript-level coexpression network analyses in ccRCC. Using WGCNA, we found that transcript networks controlling vascular development and TCA cycle were most significantly deregulated and correlated with ccRCC tumor stage. These analyses identified several uncharacterized genes as potential modulators of pathways deregulated in ccRCC.

Materials and Methods

Transcript quantification and differential expression analysis

A total of 100 fastq RNA-seq files (50 primary ccRCC and 50 normal adjacent renal samples were downloaded from The Cancer Genome Atlas (TCGA) legacy archive (https://portal.gdc.cancer.gov/legacy-archive/search/f). Human cDNA and ncRNA FASTA formatted transcript files (Ensembl v89 annotation) were acquired form the Ensembl ftp site (https://www.ensembl.org/info/data/ftp/index.html), and merged to create a master file of all putative coding and non-coding transcripts. All quantification

and differential expression analyses were performed using the kallisto-sleuth pipeline. Using the default settings, kallisto was used to create an index for quantification using the aforementioned FASTA master file. Subsequently, kallisto was used to quantify all putative transcripts using 50 bootstrap samples. Differential expression analysis was performed with sleuth using the Wald test with a cutoff of q-value of 0.005. RATs was performed using the read counts and bootstrap values calculated from kallisto. As ccRCC is a highly heterogeneous cancer, and there are 4 major subtypes of ccRCC, a replicate reproducibility of 0.25 was used in the analysis. All other parameters remained on default settings.

For the edgeR analysis, alignment of the fastq files was performed first with HISAT2 using the hg38 human assembly [27-29]. Read counting was performed using the summarizeOverlaps package, with union mode [30]. Using the read counts, an edgeR analysis was performed using the default settings. The entire pipeline was performed within the systemPipeR package [31].

Weighted coexpression network analysis

All 217,082 TPM transcripts quantifications were initially filtered for an average absolute expression of >1 TPM. Subsequently, 10,000 of the most variable transcripts, using the mean absolute deviation, were used for the proceeding WGCNA pipeline [32]. A soft thresholding power of 6 was used in a signed transcript coexpression network framework. All other parameters remained on the default recommended settings. ccRCC correlated coexpression networks

were exported to VisANT with an adjacency threshold 0.08 for visualization purposes [33]. For the gene-level Metascape analysis (http://metascape.org) of each of the network modules, genes were considered only once in the analysis, regardless of the numbers of transcripts derived from the gene.

Primer design and quantitiative PCR

Primers sequences were designed using Primer3 plus (http://primer3plus.com/cgi-bin/dev/primer3plus.cgi) using the default qPCR settings. When possible, primers were designed over exon junctions to avoid capturing unannotated alternative transcripts. All primers were synthesized by Integrated DNA Technologies. Twelve matched paired ccRCC RNA samples were acquired form Origene. Origene RNA samples were verified for quality and quantity using gel electrophoresis and the Thermoscientific Nanodrop2000 spectrophotometer. cDNA was synthesized using 1ug of total RNA using the iScript reverse transcription supermix (Biorad, Irvine, CA) according to the manufacturer's instructions. Quantitative PCR was performed using the Biorad iQ SYBR green supermix and a Biorad CFX Connect thermocylcer (Biorad, Irvine, CA) and analyzed using the CFX manager software. Using a single threshold Cq determination, the Livak method was employed for all gene expression analyses. All qPCR analyses were normalized to *PPIA*, as *PPIA* was shown to be a suitable reference gene when comparing normal adjacent tissue to ccRCC tumor tissue [34, 35].

Results

Global identification and validation of DTE in ccRCC

From the kallisto analysis, a total of 217,082 transcripts quantifications (160,717 protein-coding and 56,365 non-coding) for each of the 100 samples were used in the differential expression analyses, comparing 50 normal adjacent renal samples against 50 ccRCC samples. Using the Wald test, with a $\log_2$ transformation, 90,002 transcripts passed the initial filtering process used by the sleuth R package. With a q-value of <0.005, we identified 32,642 DETs, encoded by 14,767 genes (Supplemental Figure 2.1). With additional filtering, using the bias estimator, referred to as the beta value of >1 or <-1 and an average absolute transcript expression of >1 TPM, 7,339 high confidence DETs were identified (Figure 2.1B).

Gene ontology analyses using the express analysis in Metascape of the unique genes encoding the DETs are consistent with previous reports [25]. There is significant enrichment of gene sets and GO terms related to the immune response for the 3,366 upregulated DETs (encoded by 2,023 genes). Conversely, there is enrichment in GO terms related to metabolic processes and transport of small molecules and ions for the 3,973 downregulatd DETs (encoded by 2,518 genes). Previously reported and contained within the 7,339 DETs, is ras-related C3 botulinum toxin substrate 1 (*RAC1*), that shows a statistically significant downregulation of one of its transcripts, ENST00000356142.4 (Supplemental Figure 2.2) [13]. ENST00000356142.4 contains an additional

exon, referred to as exon 3b that is frequently spliced out in ccRCC. The most abundant *RAC1* transcript, ENST00000348035.8, is unaffected in ccRCC.

As mutations in key epigenetic modifiers, such as *SETD2*, *PBRM1* and *BAP1*, among ccRCCs have demonstrated to have significant effects on the epigenetic landscape and consequently splicing events, we compared the DETs observed in the current study against 6,207 RefSeq transcripts previously found to have defects in splicing and intron retention [14]. Among the 6,207 transcripts, 6,070 transcripts were readily converted to an ensembl annotation, and 1,857 transcripts were identified as differentially expressed. In a similar study, among 30 genes found to have a deficiency in H3K36me3 and *SETD2*-mediated alternative splicing [15], we found 27 of these genes to have at least one DET in the current study (using an FDR <0.005).

Among the 7,339 DETs discovered (4,470 individual loci), ~89% were protein-coding (6,546 transcripts) and ~11% were non-coding (793 transcripts) (Figure 2.1C, left). These DETs represented only ~4% and ~1% of the total putative protein-coding and non-coding transcripts, respectively (Figure 2.1C, right). Further characterization of the DETs showed that the number of transcripts affected remained relatively static, regardless of the number of putative transcripts derived from a given gene (Figure 2.1D). With genes encoding ≥2 transcripts, >80% of the genes had ≤3 detectable DETs.

Lastly, as previous gene-level expression analyses may not have detected some cases of DTE, we performed a comparative differential expression analysis of the

matched pair samples evaluating the results of edgeR and sleuth (Figure 2.2A) [24]. For the gene-level edgeR analysis, read counts were generated within the systemPipeR package, using HISAT2 for the alignment of the sequence reads and summarizeOverlaps for the generation of the gene counts. With thresholds of >2 fold change and FDR<0.005, edgeR identified 5,665 differentially expressed genes (DEGs). In an alternative gene-level analysis, using kallisto generated gene counts, the sleuth gene-level analysis discovered 6,441 DEGs, with a beta value of >1 or <-1 and a FDR<0.005. Among the 4470 genes, encoding the 7,339 DETs (described above), a total of 1,159 genes were found exclusively within the sleuth transcript-level analysis. Interestingly, only ~4% (51 genes) of the 1,159 genes harbored both upregulated and downregulated DETs. A moderate degree of overlap was observed between the four differential expression analyses, sharing 1,581 genes in common. Similarly, all gene-level analyses shared 1,932 genes in common, while the kallisto gene-level and our edgeR analyses had the most in common, sharing 3,632 DEGs.

One example of significant differentially expressed transcripts, not detected by gene-level analyses and not identified by previous ccRCC studies, are derived from Pleckstrin homology like domain family B member 2 (*PHLDB2*) known commonly for its association with vascular dementia (Figure 2.2B) [35]. *PHLDB2* encodes for 18 putative transcripts, and two transcripts ENST00000393923.7 and ENST00000393925.7 are downregulated in ccRCC. ENST00000393923.7 is the most abundant protein-coding *PHLDB2* transcript, and it is the most

significantly downregulated in ccRCC (Figure 2.2C). ENST00000393925.7 is a slightly less abundant *PHLDB2* transcript, and it is unaffected in ccRCC. Evaluation of the tumor/normal TPM ratios of the 50 matched pair samples showed that patients with a high degree of ENST00000393923.7 downregulation exhibited lower survival rates over ~12 years (p=0.0015, Figure 2.2D). Two additional examples of genes harboring DETs, solute carrier family 37 member 3 (*SLC37A3*) and high-density lipoprotein binding protein (*HDLBP*) were also found to correlate with patient survival (Supplemental Figure 2.3). ENST00000393923.7 downregulation was validated using transcript-specific qPCR with 12 independent matched pair ccRCC samples (Figure 2.2E). Using a Wilcoxon signed-rank test, ENST00000393923.7 was found to be significantly downregulated in ccRCC with a median downregulation of ~6.3 fold change. No statistically significant difference was observed with ENST00000393925.7.

Weighted transcript-level coexpression network analysis

As our previous analyses suggest some transcripts derived from the same gene exhibiting different expression profiles, we sought to better understand the isoform-specific changes occurring within ccRCCs. Therefore, we pursued a weighted coexpression network analysis using the calculated transcript quantifications as a framework. Using WGCNA and the calculated TPM values from 10,000 of the most variable transcripts, a coexpression network was performed across five stages of ccRCC progression (normal, stage I, stage II, stage III, stage IV). A total of 26 coexpression modules were identified (Figure

2.3A), with 7 coexpression modules highly correlated with ccRCC progression (pearson coefficient >0.5 or <-0.5 and p<0.05). Using the Reactome, KEGG pathway, CORUM gene sets and the conventional GO terms, a Metascape analysis was performed separately with each of the 7 correlated coexpression modules. Among the 4 positively correlated coexpression modules, vascular development, ribosome, cytokine signaling and collagen formation were the most enriched terms found within each of the modules. Conversely, the 3 negatively correlated coexpression modules revealed TCA cycle, extracellular matrix organization and organic acid catabolic processes as the most significant terms. Identified within each of the modules were transcripts with the highest module membership, as these transcripts are likely extensively connected intramodular hubs (Figure 2.3A). These transcripts included: ENST00000381125.8 encoded by Phosphofructokinase, Platelet (*PFKP*), ENST00000356892.3 encoded by SAM And SH3 domain containing 3 (*SASH3*), ENST00000225430.8 encoded by Ribosomal Protein L19 (*RPL19*), ENST00000296388.9 encoded by Prolyl 3-Hydroxylase 1 (*P3H1*), ENST00000295887.5 encoded by CDP-Diacylglycerol Synthase 1 (*CDS1*), ENST00000257290.9 encoded by Platelet Derived Growth Factor Receptor Alpha (*PDGFRA*), and ENST00000354775.4 encoded by Aldehyde Dehydrodenase 9 Family Member 1 (*ALDH9A1*).

Further characterization of the coexpression networks showed that the majority of the transcripts comprising the networks, and all the transcripts used in the network construction, were encoded from separate individuals genes

(Supplemental figure 2.4). Additionally, validation of the network and gene set analyses showed 24 out of the top 30 coexpressed transcripts (transcripts with high adjacency scores) contained within the vascular development coexpression module, are derived from genes comprising the core signature angiogenesis genes described previously (Figure 2.3B, left) [36]. Moreover, among the top 30 correlated transcripts contained within the TCA coexpression module, 28 transcripts are produced by genes previously discovered as being downregulated in ccRCC (Figure 2.3B, right) [37]. The remaining transcripts, ENST00000424349.1 encoded by FGD5 antisense RNA 1 (*FGD5-AS1*) and ENST00000620459.1 encoded by *AL035661.1* are uncharacterized lncRNAs highly downregulated in ccRCC.

Differential transcript usage in ccRCC

Using the kallisto transcript abundances, the RATs R package identified 97 events of differential transcript usage (Figure 2.4A, left, Supplemental Figure 2.5). These 97 transcripts were identified using the RATs transcript-level test, which examines each transcript individually and then merges the transcript information to form a gene-level finding. Alternatively, the gene-level DTU test, which collectively evaluates the transcripts of a gene, identified only 26 DTU genes (Figure 2.4A, right, Supplemental Figure 2.5). Among both transcript-level and gene-level DTU tests, 7 DTU genes (*AP1M2*, *CAB39L*, *CCDC146*, *C16orf89*, *DAB2*, *MAPK8IP1*, *FGFR2*) have been identified previously [25, 26]. Collectively, 94 DTU genes (68 uncharacterized DTU genes) in total were

discovered (using both DTU tests) when comparing normal adjacent and ccRCC tissues. No statistically significant GO terms were enriched within the 94 DTU genes, using a corrected p-value. However, the Metascape analysis showed the top GO term (p=0.0007) was carboxylic acid transport, supporting previous results demonstrating metabolic derangements as a cornerstone of ccRCC [7, 38]. Seven DTU genes were found to have a carboxylic acid transport GO classification, which included: *AGXT*, *SLC38A5*, *SLC9A4*, *SLC3A2*, *UNC13B*, *FABP6* and *FOLR1*.

Examination of the DTU events showed that non-primary (i.e. non-major) isoform switches are more frequent than primary isoform switches in ccRCC (Figure 2.4B). On average, we identified approximately twice as many non-primary isoform switches relative to primary isoform switches. Among the 8 primary isoform switches (in common between the DTU tests), all of them also had non-primary isoform switches. The DTU genes (described previously) *AP1M2*, *DAB2* and *FGFR2* exhibited both primary and non-primary isoform switching events (Supplementary Figure 2.6-7). Constituting the majority of DTU genes, a total of 76 DTU protein-coding genes were observed. The remaining DTU genes encompassed 11 ncRNA and 7 unclassified genes. Two examples of mostly uncharacterized DTU genes, with high isoform-switch frequencies, were *FOLR1* and *BABAM2* (Figure 2.4C, Supplemental Figure 2.6). *FOLR1,* known as folate receptor 1, produces 4 putative transcripts, and was found to be one of the most significant primary isoform switches. ENST00000393676.4 has an alternative 5'

end and is the most abundant *FOLR1* transcript in normal renal tissue (Figure 2.4D); however, ENST00000393681.6 switches with ENST00000393676.4 becoming the most abundant or primary *FOLR1* transcript in ccRCC. *FOLR1* had the highest isoform-switch frequency with 61% of ccRCC samples exhibiting the primary isoform-switch (Figure 2.4E). *BABAM2* encodes for a component of the BRCA1-A complex, and it produces 11 putative transcripts, 4 of which were eligible for DTU analysis. ENST00000436924.5 was the only *BABAM2* transcript to show a significant proportional increase in its abundance in ccRCC, becoming the second most abundant *BABAM2* transcript in ccRCC (Supplemental Figure 2.6).

Discussion

In the current study, we identified the global isoform-specific alterations in ccRCC and explored the deregulated networks implicated in ccRCC progression. Using the kallisto-sleuth pipeline, we discovered 7,339 DETs of which ~90% of the transcripts were derived from protein-coding genes. Additionally, comparative differential expression and coexpression network analyses aided in the discovery of several potentially clinically relevant genes and the major deregulated networks in ccRCC progression. Lastly, we discovered 68 uncharacterized high-frequency DTU genes in ccRCC with a suggested enrichment of genes involved in metabolic function.

Differential exon usage (DEU) has frequently been used as an inference for DTE

in ccRCC; however, this approach could present challenges in identifying DETs among transcripts sharing exons [21, 23, 25]. Additionally, gene-level expression analyses could potentially overlook deregulated transcripts from clinically relevant genes that give rise to multiple transcripts. Therefore, we sought to identify deregulated transcripts and cognate genes that were not discovered readily by gene-level analyses by using novel methods that are not subject to the disadvantages of the DEU approach. In a typical gene-level analysis, all exonic reads from a gene are consolidated and used to determine if the expression of a gene is altered between two conditions. However, this approach could be disadvantageous in specific circumstances. One potential pitfall to a gene-level analysis is that if the other transcripts from the same gene are of similar abundance to the DET, then a conventional gene-level analysis may not detect a gene-level difference between the two conditions. Additionally, while isoform switching was found to be a relatively rare occurrence in ccRCC, isoform switching could also account for a "masking" of a relevant gene. *PHLD2, HDLBP* and *SLC37A3* are examples of this "masking" effect, in which DTE was not detected using conventional gene-level analyses. While we acknowledge that the degree of overlap between gene-level and transcript-level analyses could vary greatly depending on methodology and experimental thresholds, the current study highlights the importance of considering transcript-level analyses in comprehensive transcriptome-wide studies. Lastly, comparisons with previous studies, focused on *SETD2* mutational status/H3K36me3 prevalence of ccRCC

tumors and the resulting effects on splicing [14, 15], suggest that genes subject to splicing defects can also harbor DETs. However, additional studies with large cohorts of mutation-specific ccRCCs are needed to determine isoform-specific expression changes that may be dependent on mutational status. As only 12 ccRCC tumors had a mutated *SETD2*, in the current study, our findings largely reflect *SETD2*-independent isoform-specific changes.

The discovery of two uncharacterized transcripts encoded by lncRNAs genes *FGD5-AS1* and *AL035661.1* identified in the network analysis suggest these lncRNAs transcripts could be potential regulators of TCA cycle genes or alternatively regulated by a common factor. These lncRNAs could be of particular importance to understanding ccRCC because of their implications in metabolic function. However, further investigation is needed, as the function of these lncRNAs is unknown. Another interesting transcript found within the TCA cycle coexpression module, identified with the highest module membership, is ENST00000295887.5 encoded by *CDS1*. *CDS1* encodes an integral membrane enzyme, located on the membranes of the mitochondrion and endoplasmic reticulum, that catalyzes the conversion of phosphatidic acid into CDP-diacylaglycerol [39, 40]. *CDS1* is uncharacterized in ccRCC and there is limited information on its role in cancer; however, in a recent study, *CDS1* was suggested to potentiate limitless growth and genomic instability in breast cancer [41]. We identified a total of 94 genes exhibiting differential transcript usage in ccRCC of which 7 DTU genes were reported previously [25, 26]. However, when

considering the findings of an alternative study [24], which also evaluated lower frequency isoform-switches, the current study identified 26 DTU genes in common. Therefore, the differences observed in the DTU genes are likely attributed to different computational techniques/thresholds and/or the use of different transcript annotations [19]. While our findings show that the majority of isoform switching events involves non-primary isoforms, which is consistent with a previous result [24], alterations in the expression of non-primary isoforms could still be clinically relevant, as supported by the survival analyses seen with the non-primary *SLC37A3* and *HDLBP* deregulated transcripts. However, the mechanisms involved require further investigation. Recent studies have illustrated how isoform-specific alterations could be highly influential in ccRCC and other cancers. For instance, alternatively spliced isoforms of VHL were shown to alter VHL binding affinity to components of the p53 pathway [42]. Additionally, isoform-switching events have been demonstrated to alter the invasive properties of cancer cells [14, 43]. From our analyses and previous similar studies, mentioned above, it is highly suggestive that isoform-specific deregulations are a critical part to characterizing and understanding the molecular underpinnings of ccRCC, and suggest that isoform-level transcriptomic analyses should more generally be considered to obtain a more comprehensive view of the genetic deregulations in cancer.

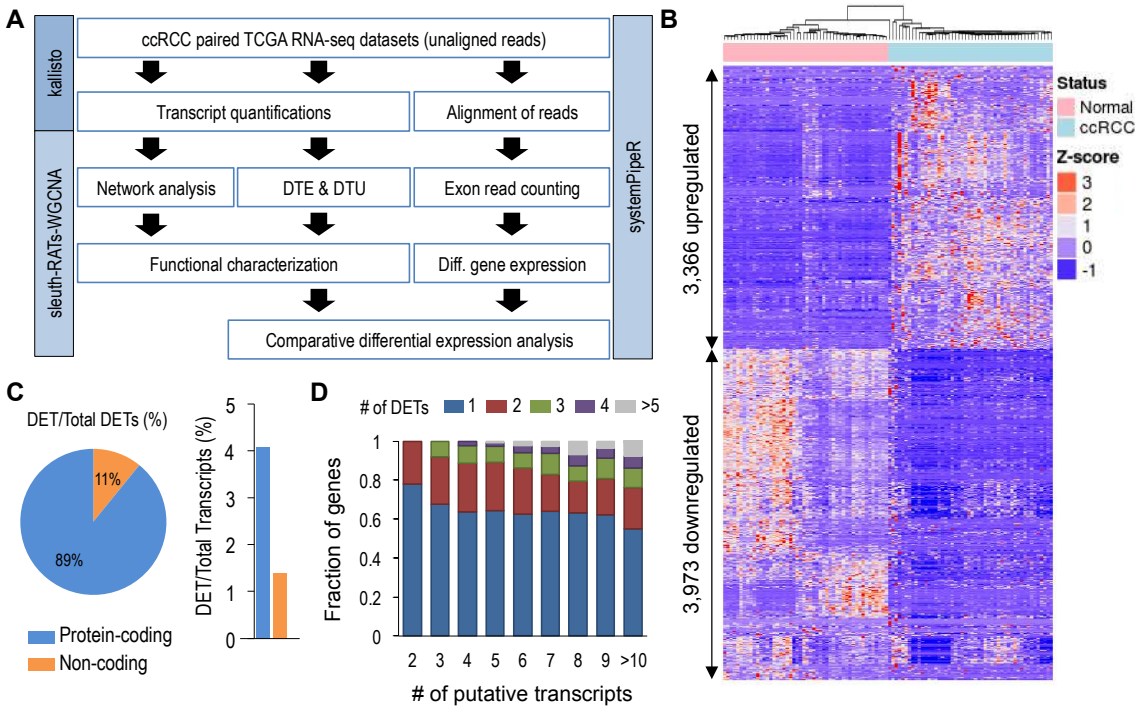


Figure 2.1. Global identification of differential transcript expression in ccRCC. **A**- Overview of pipeline used in identification and characterization of DTE and DTU in ccRCC. **B-** Unsupervised hierarchical clustering of 7,339 DETs identified using sleuth (FDR <0.005 and beta value of <-1 or >1). **C-** Percentage of protein-coding and non-coding genes encoding the 7,339 DETs identified using sleuth. **D-** Proportion of genes with *n* identified DETs relative to total number of encoded transcripts

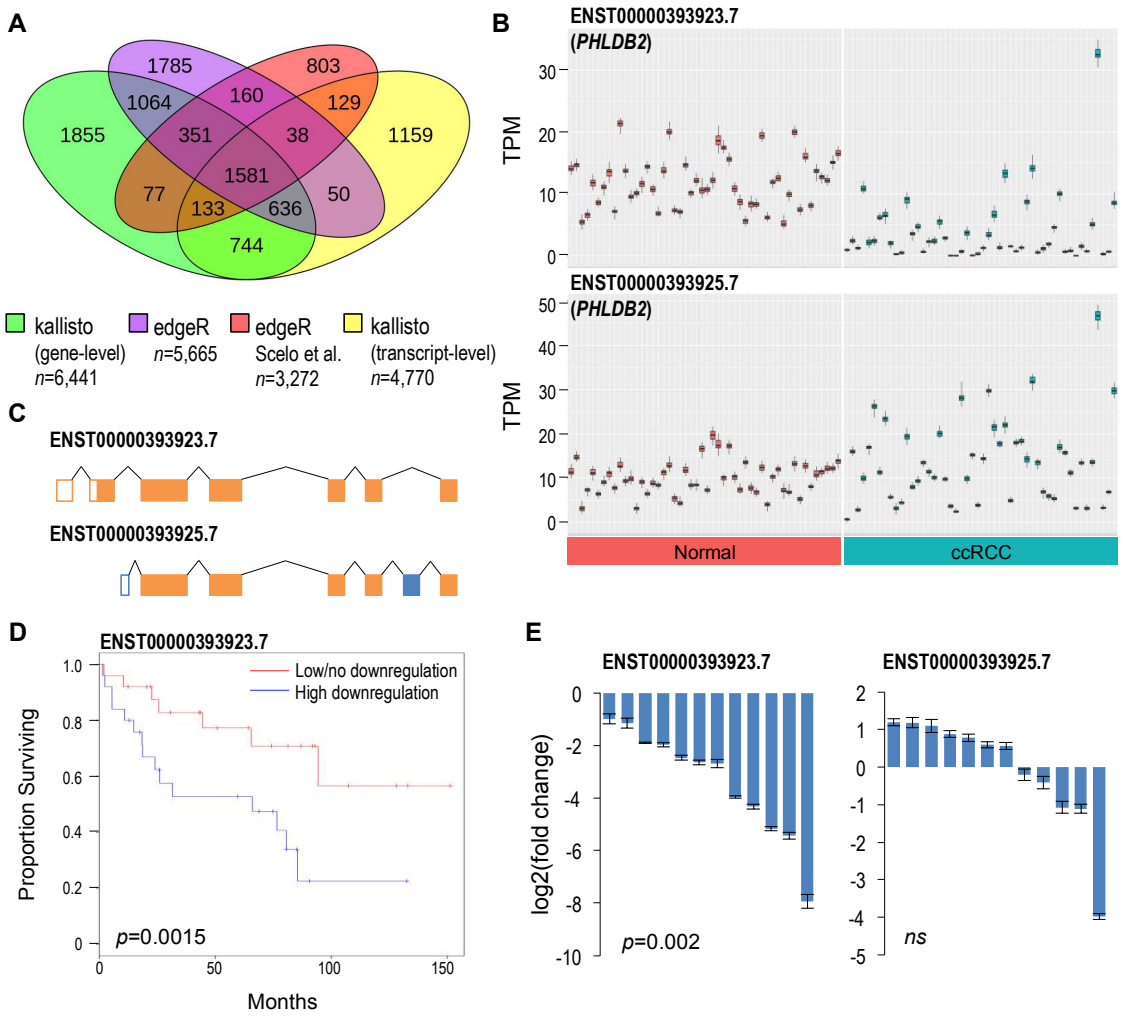


Figure 2.2. Comparative differential expression analysis identifies novel genes implicated in ccRCC. **A**- Comparison of DEGs/DTE genes discovered with sleuth, edgeR, and a previous study by Scelo et al. **B-** Transcript abundances in normal renal and ccRCC tissues for the two most abundant *PHLDB2* transcripts. Each box plot represents 50 calculated bootstrap values of an individual sample (red = normal, blue = ccRCC). **C-** ENST00000393923.7 harbors an alternative exon 1 and 2 and excludes exon 6 of

ENST00000393925.7. Differences colored in blue. **D-** Kaplan-Meier plot assessing survival of patients with high vs low/no ENST00000393923.7 downregulation. Median T/N ratio was used to partition samples into low/no and high downregulation groups. Log rank test was used to calculate statistical significance. **E-** qPCR validation of *PHLDB2* DTE showing log2 fold change of 12 ccRCC tissues relative to their normal adjacent tissues. Results normalized to *PPIA* reference gene. Two-tailed Wilcoxon signed-rank test was used to determine statistical significance. Error bars = average standard deviation of technical replicates of pair samples. ns = non-significant (>0.05).

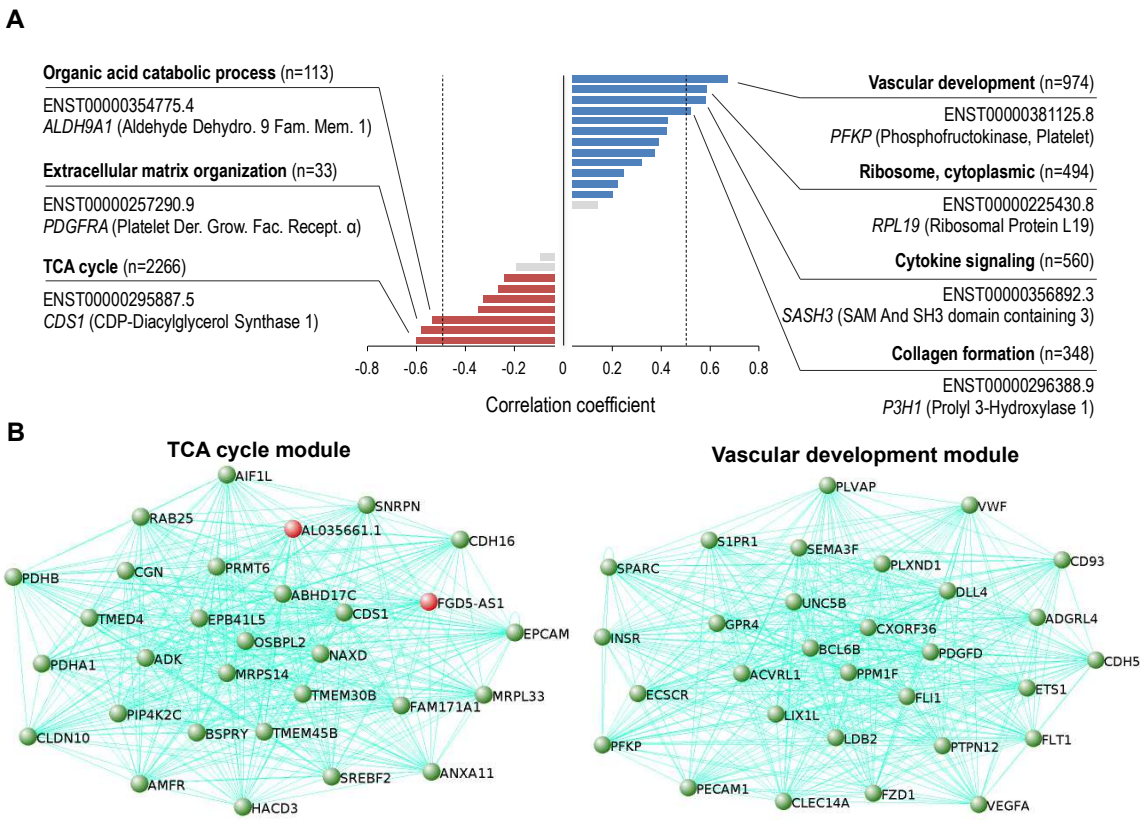


Figure 2.3. Vascular development and TCA cycle coexpression modules are the highest correlated networks in ccRCC progression. **A**- ccRCC correlated coexpression network modules identified with WGCNA. Using a correlation coefficient of >0.5 or <-0.5 and p<0.05, 4 positively correlated networks (blue bars, right of dotted line), and 3 negatively networks were identified to be in ccRCC (red bars, left of dotted line). Networks with no significant correlation with ccRCC (grey, p>0.05). Most significant GO term for each module shown in bold, and the transcript with the highest module membership shown below. **B-** Top 30 highest coexpressed transcripts (gene names shown) within the TCA cycle (left) and vascular development modules

(right). Novel genes highlighted in red.

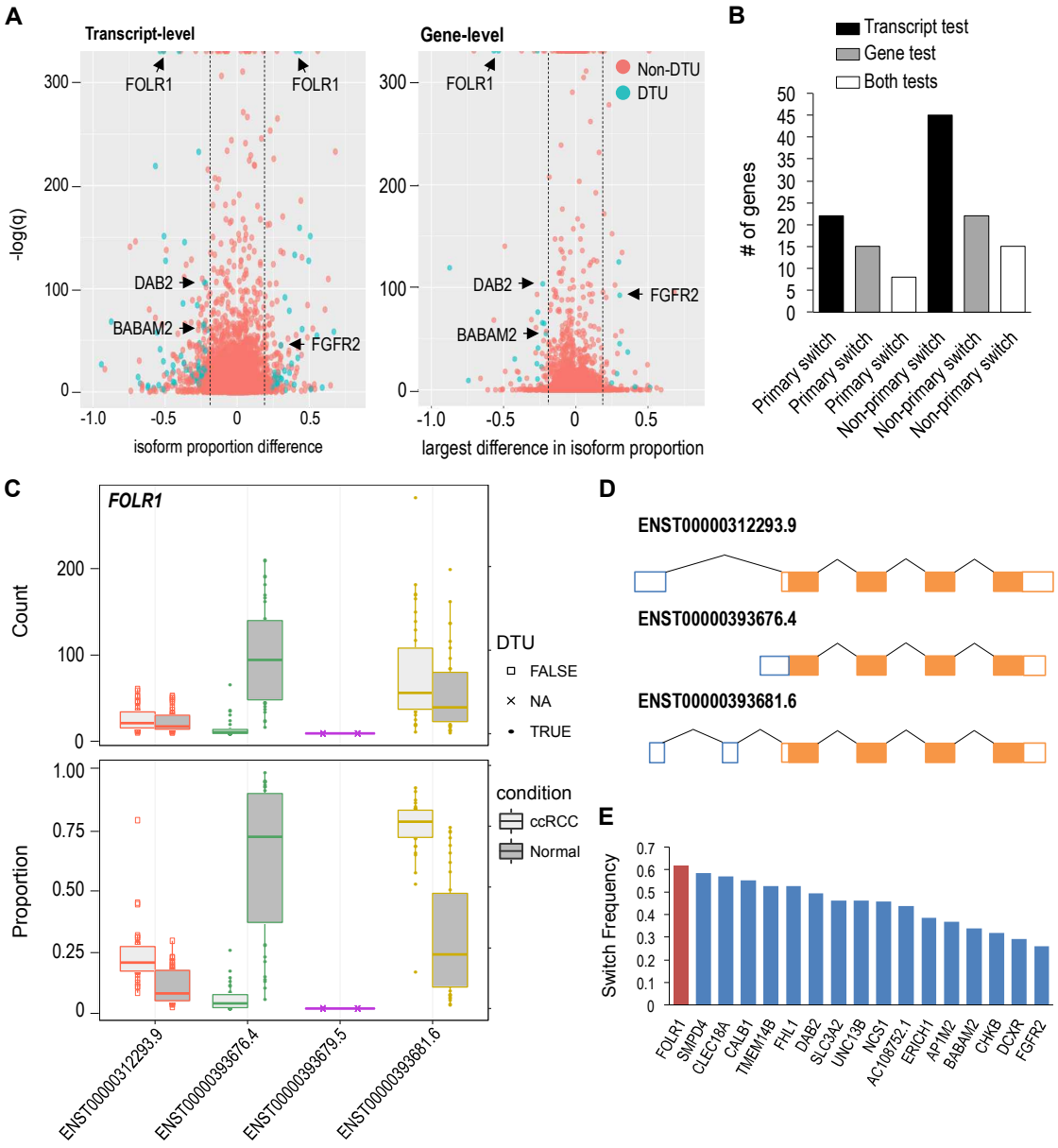


Figure 2.4. Few high frequency DTU genes observed in ccRCC. A- Transcript and gene-level tests using RATs to identify DTU events in ccRCC (red dot = non-DTU, blue dot = DTU). **B-** Number of primary and non-primary isoform switches discovered in ccRCC. "Both" represents the number of shared DTU genes identified in both the transcript and gene-level tests. **C-** *FOLR1* exhibiting

significant proportional isoforms changes in ccRCC. Circle = significant DTU. Square = tested in DTU analysis, but not significant. X = did not meet abundance threshold for DTU anlaysis. **D-** Schematic of *FOLR1* transcripts analyzed in DTU analysis. **E-** Frequency of *FOLR1* and 17 other isoform switches shared between both DTU tests.

References

1. U.S. Cancer Statistics Working Group. (2017). United States Cancer Statistics: 1999-2014 Incidence and Mortality Web-based Report. (Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute).

2. American Cancer Society. (2018). Cancer Facts & Figures 2018. (Atlanta, Ga: American Cancer Society).

3. Bander NH, Finstad CL, Cordon-Cardo C, Ramsawak RD, Vaughan ED, Jr., Whitmore WF, Jr., Oettgen HF, Melamed MR, Old LJ. Analysis of a mouse monoclonal antibody that reacts with a specific region of the human proximal tubule and subsets renal cell carcinomas. Cancer Res. 1989; 49: 6774-80. doi:

4. van den Berg E, van der Hout AH, Oosterhuis JW, Storkel S, Dijkhuizen T, Dam A, Zweers HM, Mensink HJ, Buys CH, de Jong B. Cytogenetic analysis of epithelial renal-cell tumors: relationship with a new histopathological classification. Int J Cancer. 1993; 55: 223-7. doi:

5. Kosary C, McLaughlin J. Kidney and renal pelvis. In: Miller BA, Ries LAG, Hankey BF, et al., eds. SEER cancer statistics review, 1973-1990. NIH publication. 1993; 93- 2789: XI.1-XI.22. doi:

6. Linehan WM, Walther MM, Zbar B. The genetic basis of cancer of the kidney. J Urol. 2003; 170: 2163-72. doi: 10.1097/01.ju.0000096060.92397.ed.

7. Cancer Genome Atlas Research N. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature. 2013; 499: 43-9. doi: 10.1038/nature12222.

8. Gnarra JR, Tory K, Weng Y, Schmidt L, Wei MH, Li H, Latif F, Liu S, Chen F, Duh FM, et al. Mutations of the VHL tumour suppressor gene in renal carcinoma. Nat Genet. 1994; 7: 85-90. doi: 10.1038/ng0594-85.

9. Linehan WM, Zbar B. Loss of DNA sequences on chromosome 3 in renal cell carcinoma. Urology. 1987; 30: 404. doi:

10. Long JP, Anglard P, Gnarra JR, Walther MM, Merino MJ, Liu S, Lerman MI, Zbar B, Linehan WM. The use of molecular genetic analysis in the diagnosis of renal cell carcinoma. World J Urol. 1994; 12: 69-73. doi:

11. Ohh M, Park CW, Ivan M, Hoffman MA, Kim TY, Huang LE, Pavletich N, Chau V, Kaelin WG. Ubiquitination of hypoxia-inducible factor requires direct

binding to the beta-domain of the von Hippel-Lindau protein. Nat Cell Biol. 2000; 2: 423-7. doi: 10.1038/35017054.

12. Pflueger D, Mittmann C, Dehler S, Rubin MA, Moch H, Schraml P. Functional characterization of BC039389-GATM and KLK4-KRSP1 chimeric read-through transcripts which are up-regulated in renal cell cancer. BMC Genomics. 2015; 16: 247. doi: 10.1186/s12864-015-1446-z.

13. Piekielko-Witkowska A, Wiszomirska H, Wojcicka A, Poplawski P, Boguslawska J, Tanski Z, Nauman A. Disturbed expression of splicing factors in renal cancer affects alternative splicing of apoptosis regulators, oncogenes, and tumor suppressors. PLoS One. 2010; 5: e13690. doi: 10.1371/journal.pone.0013690.

14. Simon JM, Hacker KE, Singh D, Brannon AR, Parker JS, Weiser M, Ho TH, Kuan PF, Jonasch E, Furey TS, Prins JF, Lieb JD, Rathmell WK, et al. Variation in chromatin accessibility in human kidney cancer links H3K36 methyltransferase loss with widespread RNA processing defects. Genome Res. 2014; 24: 241-50. doi: 10.1101/gr.158253.113.

15. Ho TH, Park IY, Zhao H, Tong P, Champion MD, Yan H, Monzon FA, Hoang A, Tamboli P, Parker AS, Joseph RW, Qiao W, Dykema K, et al. High-resolution profiling of histone h3 lysine 36 trimethylation in metastatic renal cell carcinoma. Oncogene. 2016; 35: 1565-74. doi: 10.1038/onc.2015.221.

16. Grosso AR, Leite AP, Carvalho S, Matos MR, Martins FB, Vitor AC, Desterro JM, Carmo-Fonseca M, de Almeida SF. Pervasive transcription read-through promotes aberrant expression of oncogenes and RNA chimeras in renal carcinoma. Elife. 2015; 4. doi: 10.7554/eLife.09214.

17. Yanagisawa M, Huveldt D, Kreinest P, Lohse CM, Cheville JC, Parker AS, Copland JA, Anastasiadis PZ. A p120 catenin isoform switch affects Rho activity, induces tumor cell invasion, and predicts metastatic disease. J Biol Chem. 2008; 283: 18344-54. doi: 10.1074/jbc.M801192200.

18. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016; 34: 525-7. doi: 10.1038/nbt.3519.

19. Froussios K, Mourão K, Simpson GG, Barton GJ, Schurch NJ. Identifying differential isoform abundance with RATs: a universal tool and a warning. bioRxiv. 2017. doi: 10.1101/132761.

21. Brito GC, Fachel AA, Vettore AL, Vignal GM, Gimba ER, Campos FS, Barcinski MA, Verjovski-Almeida S, Reis EM. Identification of protein-coding and

intronic noncoding RNAs down-regulated in clear cell renal carcinoma. Mol Carcinog. 2008; 47: 757-67. doi: 10.1002/mc.20433.

22. Christinat Y, Pawlowski R, Krek W. jSplice: a high-performance method for accurate prediction of alternative splicing events and its application to large-scale renal cancer transcriptome data. Bioinformatics. 2016; 32: 2111-9. doi: 10.1093/bioinformatics/btw145.

23. Deng M, Blondeau JJ, Schmidt D, Perner S, Muller SC, Ellinger J. Identification of novel differentially expressed lncRNA and mRNA transcripts in clear cell renal cell carcinoma by expression profiling. Genom Data. 2015; 5: 173-5. doi: 10.1016/j.gdata.2015.06.016.

24. Scelo G, Riazalhosseini Y, Greger L, Letourneau L, Gonzalez-Porta M, Wozniak MB, Bourgey M, Harnden P, Egevad L, Jackson SM, Karimzadeh M, Arseneault M, Lepage P, et al. Variation in genomic landscape of clear cell renal cell carcinoma across Europe. Nat Commun. 2014; 5: 5135. doi: 10.1038/ncomms6135.

25. Valletti A, Gigante M, Palumbo O, Carella M, Divella C, Sbisa E, Tullo A, Picardi E, D'Erchia AM, Battaglia M, Gesualdo L, Pesole G, Ranieri E. Genome-wide analysis of differentially expressed genes and splicing isoforms in clear cell renal cell carcinoma. PLoS One. 2013; 8: e78452. doi: 10.1371/journal.pone.0078452.

26. Zhao Q, Caballero OL, Davis ID, Jonasch E, Tamboli P, Yung WK, Weinstein JN, Kenna Shaw for Trn, Strausberg RL, Yao J. Tumor-specific isoform switch of the fibroblast growth factor receptor 2 underlies the mesenchymal and malignant phenotypes of clear cell renal cell carcinomas. Clin Cancer Res. 2013; 19: 2460-72. doi: 10.1158/1078-0432.CCR-12-3708.

27. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015; 12: 357-60. doi: 10.1038/nmeth.3317.
38. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 2012; 40: 4288-97. doi: 10.1093/nar/gks042.

28. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26: 139-40. doi: 10.1093/bioinformatics/btp616.
29. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013; 9: e1003118. doi: 10.1371/journal.pcbi.1003118.

30. TW HB, Girke T. systemPipeR: NGS workflow and report generation environment. BMC Bioinformatics. 2016; 17: 388. doi: 10.1186/s12859-016-1241-0.

31. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008; 9: 559. doi: 10.1186/1471-2105-9-559.

32. Hu Z, Mellor J, Wu J, DeLisi C. VisANT: an online visualization and analysis tool for biological interaction data. BMC Bioinformatics. 2004; 5: 17. doi: 10.1186/1471-2105-5-17.

33. Dupasquier S, Delmarcelle AS, Marbaix E, Cosyns JP, Courtoy PJ, Pierreux CE. Validation of housekeeping gene and impact on normalized gene expression in clear cell renal cell carcinoma: critical reassessment of YBX3/ZONAB/CSDA expression. BMC Mol Biol. 2014; 15: 9. doi: 10.1186/1471-2199-15-9.

34. Jung M, Ramankulov A, Roigas J, Johannsen M, Ringsdorf M, Kristiansen G, Jung K. In search of suitable reference genes for gene expression studies of human renal cell carcinoma by real-time PCR. BMC Mol Biol. 2007; 8: 47. doi: 10.1186/1471-2199-8-47.

35. Eom S, Lee C. Functions of intronic nucleotide variants in the gene encoding pleckstrin homology like domain beta 2 (PHLDB2) on susceptibility to vascular dementia. World J Biol Psychiatry. 2013; 14: 227-32. doi: 10.3109/15622975.2011.630407.

36. Masiero M, Simoes FC, Han HD, Snell C, Peterkin T, Bridges E, Mangala LS, Wu SY, Pradeep S, Li D, Han C, Dalton H, Lopez-Berestein G, et al. A core human primary tumor angiogenesis signature identifies the endothelial orphan receptor ELTD1 as a key regulator of angiogenesis. Cancer Cell. 2013; 24: 229-41. doi: 10.1016/j.ccr.2013.06.004.

37. Gatto F, Nookaew I, Nielsen J. Chromosome 3p loss of heterozygosity is associated with a unique metabolic network in clear cell renal carcinoma. Proc Natl Acad Sci U S A. 2014; 111: E866-75. doi: 10.1073/pnas.1319196111.

38. Hakimi AA, Reznik E, Lee CH, Creighton CJ, Brannon AR, Luna A, Aksoy BA, Liu EM, Shen R, Lee W, Chen Y, Stirdivant SM, Russo P, et al. An Integrated Metabolic Atlas of Clear Cell Renal Cell Carcinoma. Cancer Cell. 2016; 29: 104-16. doi: 10.1016/j.ccell.2015.12.004.

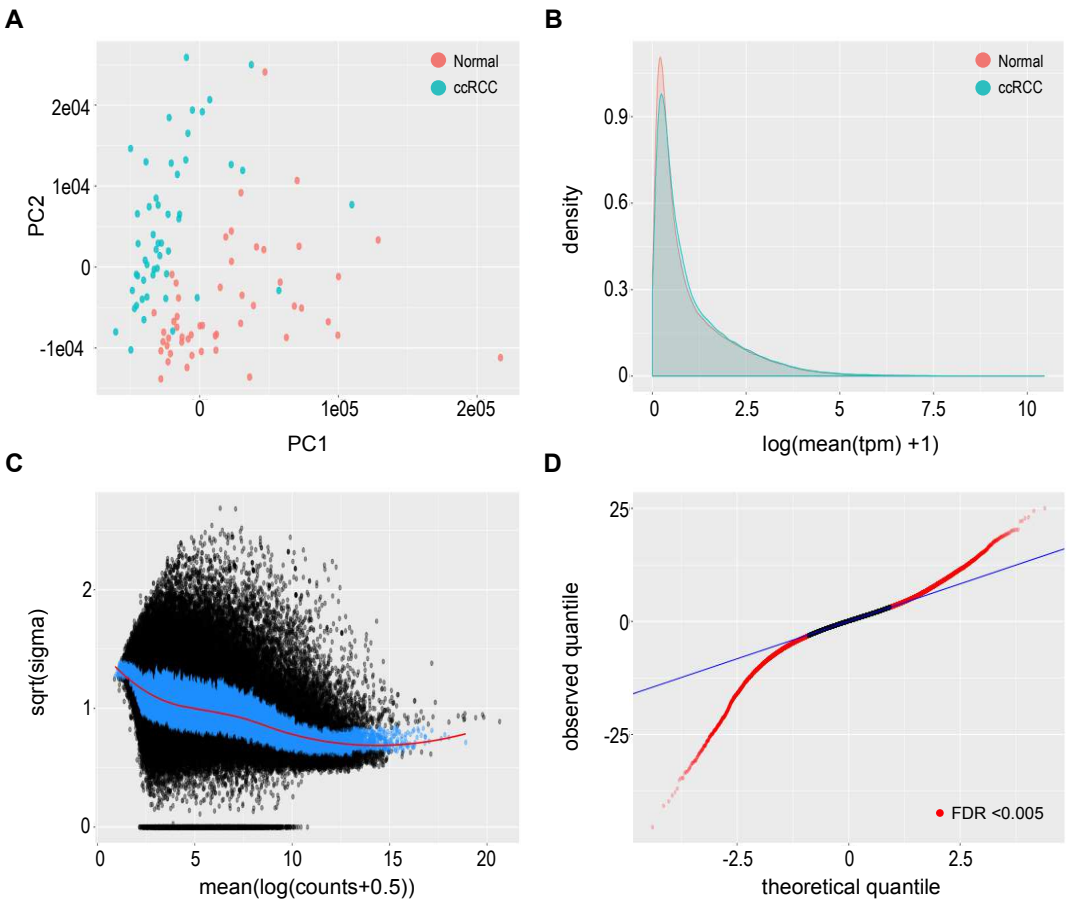
39. Heacock AM, Agranoff BW. CDP-diacylglycerol synthase from mammalian tissues. Biochim Biophys Acta. 1997; 1348: 166-72. doi:

40. Letts VA, Klig LS, Bae-Lee M, Carman GM, Henry SA. Isolation of the yeast structural gene for the membrane-associated enzyme phosphatidylserine synthase. Proc Natl Acad Sci U S A. 1983; 80: 7279-83. doi:

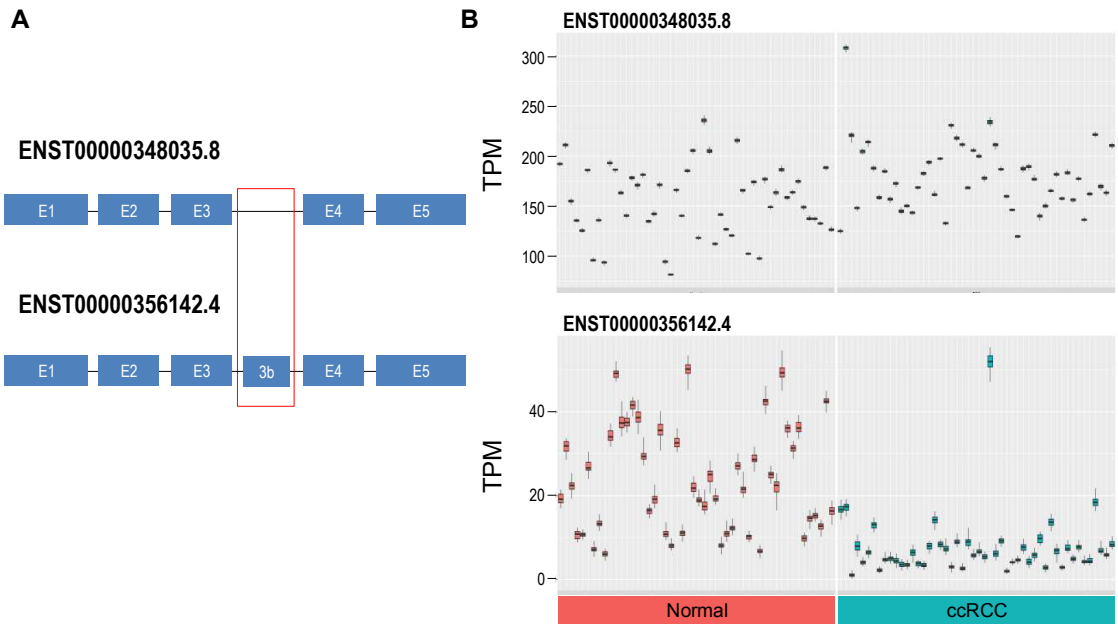
41. Cook AC, Tuck AB, McCarthy S, Turner JG, Irby RB, Bloom GC, Yeatman TJ, Chambers AF. Osteopontin induces multiple changes in gene expression that reflect the six "hallmarks of cancer" in a model of breast cancer progression. Mol Carcinog. 2005; 43: 225-36. doi: 10.1002/mc.20105.

42. Minervini G, Mazzotta GM, Masiero A, Sartori E, Corra S, Potenza E, Costa R, Tosatto SC. Isoform-specific interactions of the von Hippel-Lindau tumor suppressor protein. Sci Rep. 2015; 5: 12605. doi: 10.1038/srep12605.

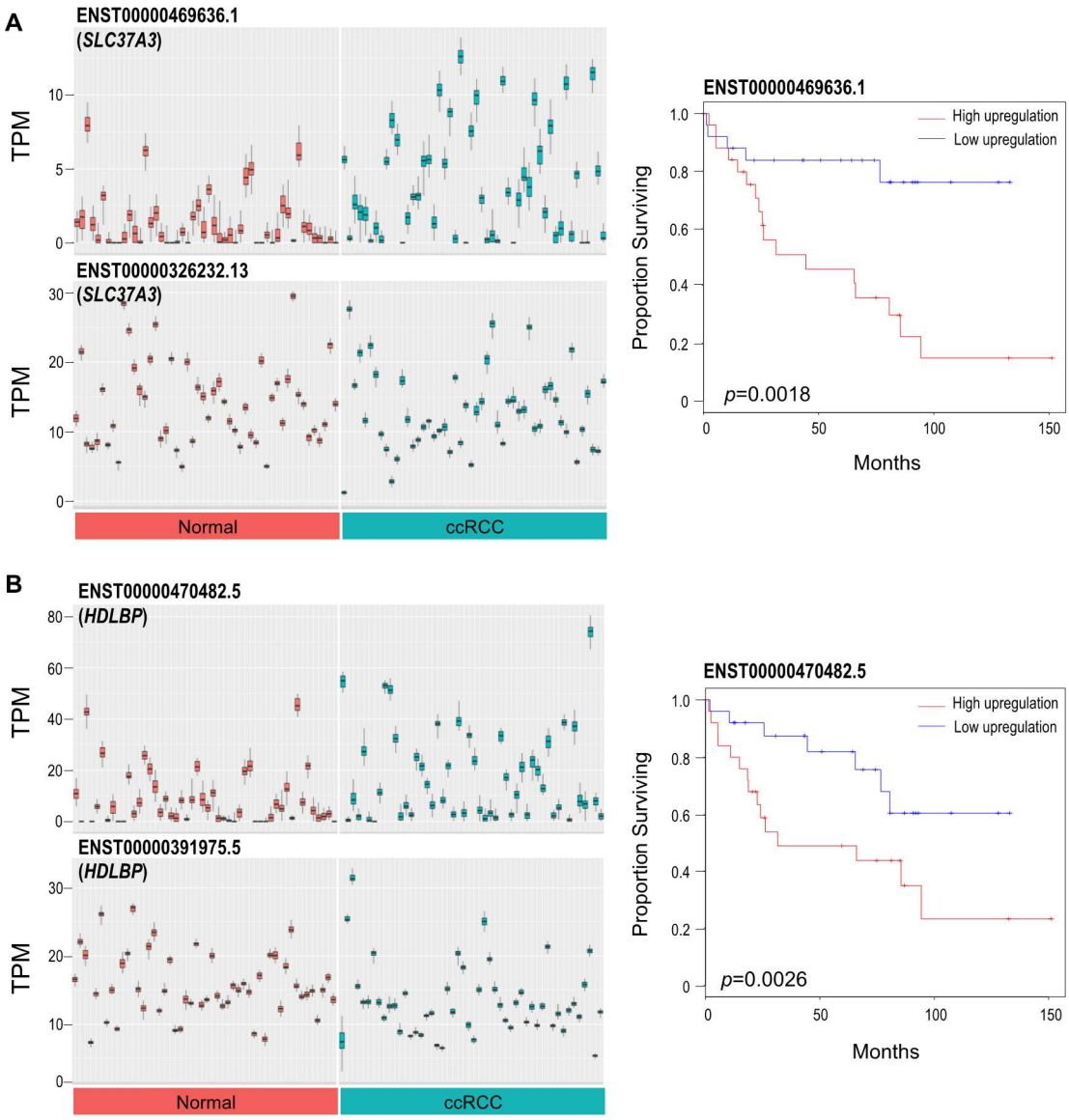
43. Lu H, Liu J, Liu S, Zeng J, Ding D, Carstens RP, Cong Y, Xu X, Guo W. Exo70 isoform switching upon epithelial-mesenchymal transition mediates cancer cell invasion. Dev Cell. 2013; 27: 560-73. doi: 10.1016/j.devcel.2013.10.020.



Supplemental figure 2.1. Assessment of calculated transcript abundances and sleuth differential expression analysis. **A**- Principal component analysis of TPM abundances to assess for outliers. **B-** Distributions of transcript abundances by tissue status (normal vs ccRCC). **C-** Mean-variance of transcripts modeled by sleuth (blue dots represent transcripts used in shrinkage estimation). **D-** Q-Q plot assessing abundance distributions between normal and ccRCC samples (red dot = FDR <0.005).

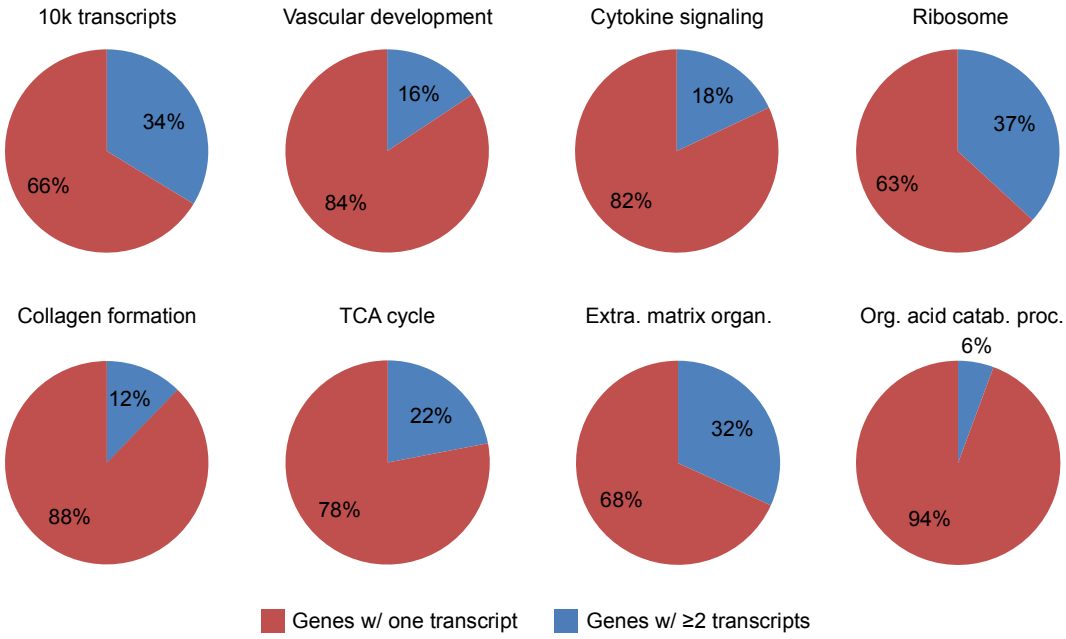


Supplemental figure 2.2. ENST00000356142.4 (*RAC1)* transcript downregulated in ccRCC. **A**- Schematic of most abundant protein-coding *RAC1* transcripts in normal renal tissue. ENST00000356142.4 contains an additional exon, referred as exon 3b (enclosed in red box). **B-** *RAC1* transcript abundances in normal renal and ccRCC tissues. ENST00000356142.4 is downregulated in ccRCC. Each box plot represents 50 calculated bootstrap values of an individual sample (red = normal, blue = ccRCC).

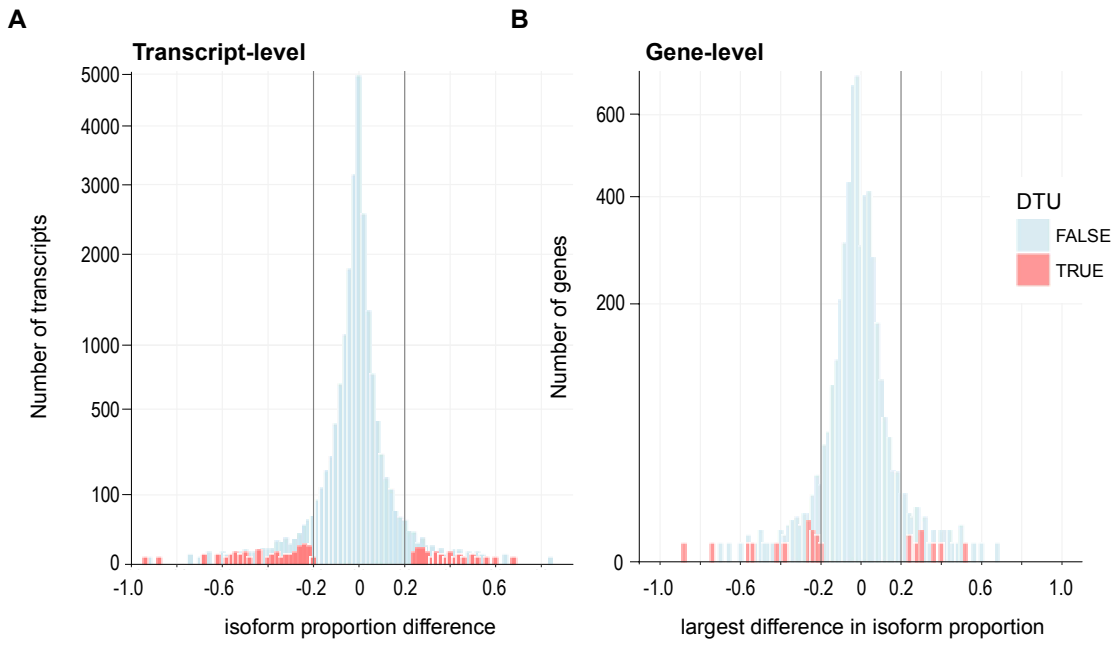


Supplemental figure 2.3. *SLC37A3* and *HDLBP* upregulated transcripts in ccRCC. **A**- *SLC37A3* transcript abundances in normal renal and ccRCC tissues (left). ENST00000469636.1 upregulated in ccRCC and correlated with patient survival (right). **B-** *HDLBP* transcript abundances in normal renal and ccRCC tissues (left). ENST00000470482.5 upregulated in ccRCC and correlated with patient survival (right). Each box plot represents 50 calculated bootstrap values

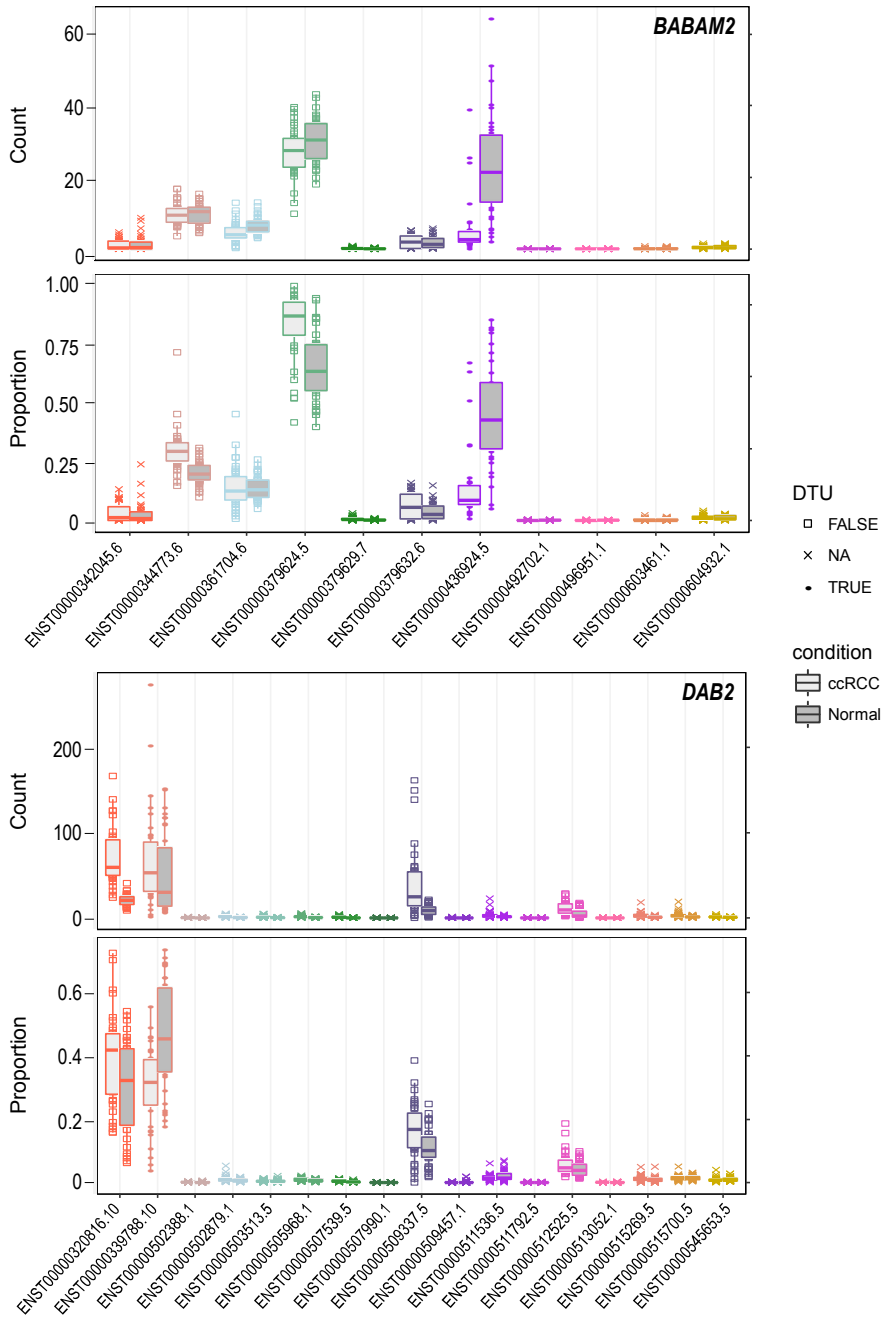
of an individual sample (red = normal, blue = ccRCC). Median T/N ratio was used to partition samples into low and high upregulation groups. Log rank test was used to calculate statistical significance.



Supplemental figure 2.4. Coexpression modules comprised mostly of transcripts encoded by unique genes. Assessment of transcripts used in the construction of the network analysis (top left) and the transcripts comprising ccRCC correlated modules. Red = genes with one transcript. Blue = gene with ≥2 transcripts.

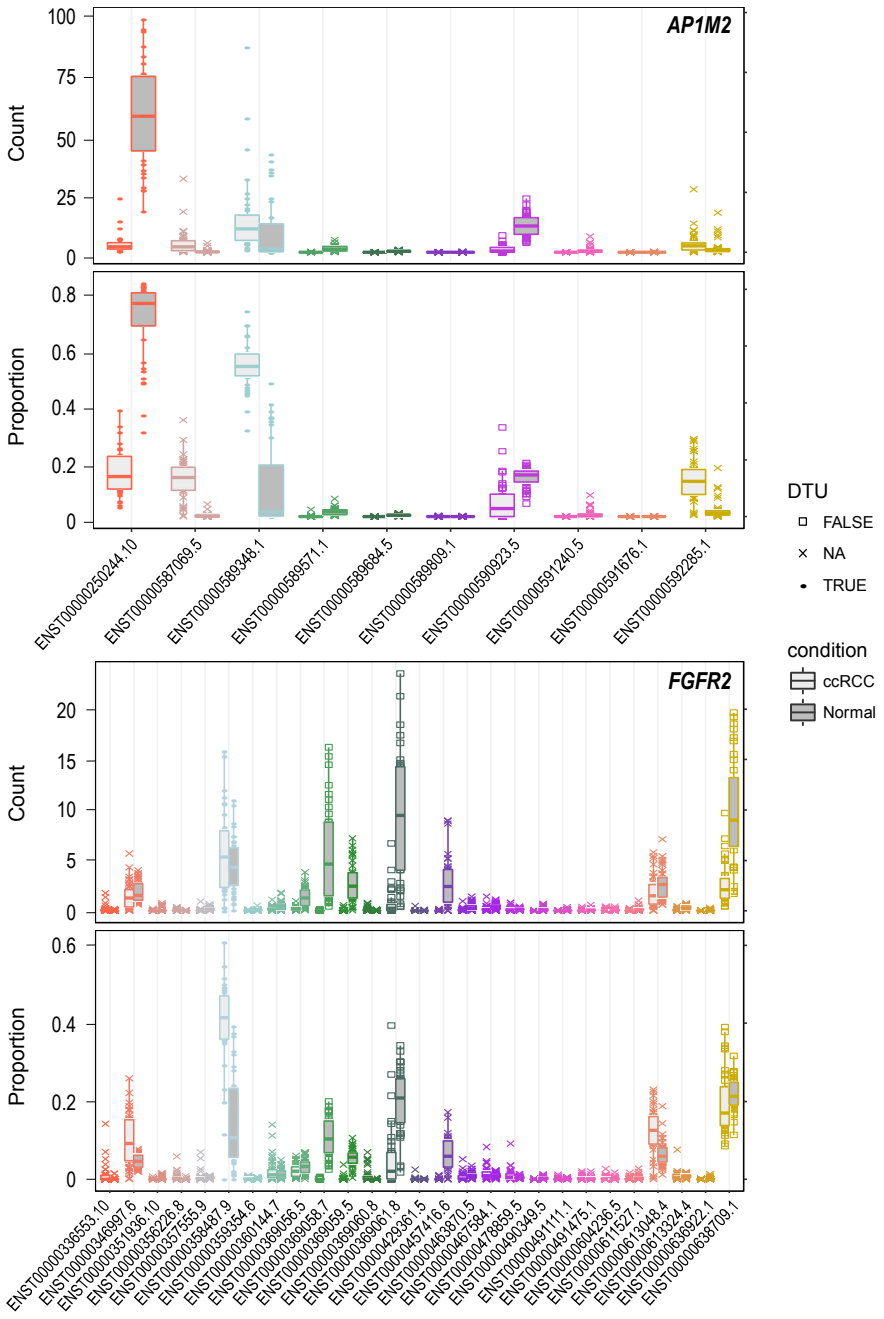


Supplemental figure 2.5. Density distributions of DTU genes relative to isoform proportion differences. **A**- Number of DTU transcripts and their isoform proportion differences discovered using RATs. **B-** Number of DTU genes and their isoform proportion differences discovered using RATs. Red bar = DTU transcript/gene. Blue bar = non-DTU transcript/gene. Y-axis is square root compressed.



Supplemental figure 2.6. *BABAM2* and *DAB2* DTU in ccRCC. Significant proportional increase observed in *BABAM2* isoform ENST00000436924.5 in ccRCC (top). Significant proportional increase observed in DAB2 isoform

ENST00000339788.10 in ccRCC. Circle = significant DTU. Square = tested in DTU analysis, but not significant. X = did not meet abundance threshold for DTU analysis.



Supplemental figure 2.7. *AP1M2* and *FGFR2* DTU in ccRCC. Primary isoform switch observed in *AP1M2* in ccRCC (top). A significant proportional decrease observed in *FGFR2* isoform ENST00000358487.9 in ccRCC. Circle = significant

DTU. Square = tested in DTU analysis, but not significant. X = did not meet abundance threshold for

Chapter 3

HOTAIRM1 lncRNA alters the hypoxia pathway in clear cell renal cell carcinoma and regulates kidney cell differentiation

Abstract

Investigations into long non-coding RNAs (lncRNAs) in cancer biology are leading to intriguing insights into the molecular and phenotypic aberrations that occur when lncRNAs are deregulated. *HOXA Transcript Antisense RNA, Myeloid-Specific 1*, or *HOTAIRM1*, is a highly conserved lncRNA that has only recently been implicated in cancer. In our present study, we report a novel finding that shows an alternative transcript of *HOTAIRM1*, referred to as *HM1-3*, is specifically and pervasively downregulated in >90% of clear-cell renal cell carcinomas (ccRCCs). *HM1-3* was found to localize predominantly to the cytoplasm and was commonly suppressed in human cell lines. Targeted knockdown of *HM1* in CAKI-1 cells, has limited effects on RNA steady-state levels, but suggests *HM1,* specifically *HM1-3*, regulates key genes involved in the hypoxia pathway. Increases in *ANGTPL4* and *DDAH1* expression were observed with concomitant increases in HIF1α protein expression with *HM1* knockdown. Lastly, assessment of *HM1* in differentiating mES cells into kidney progenitor cells showed significant increases in *HM1* expression, while knockdown of *HM1* in kidney progenitor cells resulted in suppression of kidney differentiation genes. Collectively, these data suggest *HM1* as an important regulator of ccRCC pathogenesis and also normal kidney differentiation.

Introduction

Kidney and renal pelvis cancers are among the most pervasive cancers found within the United States [1]. Renal cell carcinomas (RCCs) comprise >90% of kidney cancers, which have been shown to be particularly difficult to treat with conventional therapies [2-4]. Investigations exploring the most frequently occurring RCC, known as clear cell renal carinaoma (ccRCC), suggest long non-coding RNAs (lncRNAs) could serve as new highly specific therapeutic targets. LncRNAs have a greater tissue-specific expression relative to protein-coding genes, and recent transcriptomic analyses examining lncRNAs have discovered several potentially important deregulated lncRNAs in ccRCC [5, 6]. One prominent example is the lncRNA *PVT1,* which regulates the stability of the MYC oncoprotein [7]. *PVT1* is part of a network of lncRNAs that modulates MYC activity and consequently the VHL-HIF axis by affecting the binding partners of HIF1α and HIF2α in ccRCC [8-11].

However, few lncRNAs have been extensively explored in ccRCC. In a recent study examining the isoform-specific transcript alterations in ccRCC, *HOTAIRM1* or *HM1,* was suggested to be deregulated in ccRCC [12], specifically the shorter spliced *HM1* isoform, referred to as *HM1-3* (Figure 3.1A). The *HM1* locus is located within the HOXA cluster between *HOXA1* and *HOXA2*, and *HM1* is best known for its role in the differentiation of promyelocytic leukemia cells [13, 14]. However, contrary to its original designated name, *HM1* expression is observed in numerous developing and fully differentiated tissue and cell types, and *HM1*

expression is altered in several human cancers [15-18]. Consequently, *HM1* has caught the attention of the cancer biology field. While the mechanistic role(s) of *HM1* in cancer are largely unknown, recent evidence suggests *HM1* regulates the autophagy pathway acting as a miRNA sponge [19].

In the current study, we have expanded on our previous work by examining *HM1-3* expression in ccRCC, and attempted to dissect the functional importance of *HM1-3* in ccRCC. Using a well-established ccRCC cell line (CAKI-1), we demonstrate that *HM1-3* is regulating the hypoxia-responsive genes possibly via the modulation of the HIF proteins. Furthermore, we show *HM1* is an important player in kidney cell differentiation and maintenance of the differentiated kidney state.

Materials and Methods

Cell culture

The HK-2, ACHN and CAKI-1 cell lines were acquired from ATCC and were cultured as recommended. The HK-2 cell line was cultured in keratinocyte serum-free medium supplied with 0.05 mg/ml bovine pituitary extract and 5 ng/ml human recombinant epidermal growth factor (Invitogen, Carlsbad, CA). The ACHN cell line was cultured in Dulbecco's modified Eagle's medium supplemented with 10% FBS (Gibco, Grand Island, USA). The CAKI-1 cell line was cultured in McCoy's 5a Modified Medium supplemented with 10% FBS

(Gibco, Grand Island, USA). All cultures were maintained in a humidified incubator with 5% CO2 at 37°C.

RNA extraction and quantitative PCR

Cells were collected using 0.25% trypsin and RNA was extracted using the GeneJet RNA purification kit (Thermoscientific, Carlsbad, CA) per manufacturer's recommendations. DNA was digested using the Rnase-Free DNase set (Qiagen, Valencia, CA) for 1 hour on the column according to the manufacturer's instructions. Extracted RNA was verified for quality and quantity using gel electrophoresis and the Thermoscientific Nanodrop2000 spectrophotometer. cDNA was synthesized using 1ug of total RNA using the iScript reverse transcription supermix (Biorad, Irvine, CA) according to the manufacturer's instructions. Quantitative PCR was performed using the Biorad iQ SYBR green supermix and a Biorad CFX Connect thermocylcer (Biorad, Irvine, CA) and analyzed using the CFX manager software. Using a single threshold Cq determination, the Livak method was employed for all gene expression analyses. Furthermore, all expression analyses were normalized to *PPIA*, as *PPIA* was to found to be the most suitable reference gene when comparing normal adjacent tissue to ccRCC tumor tissue [20, 21], and little change was observed with *HM1* knockdown. Three technical replicates of each biological replicate were performed for every qPCR reaction. All Origene cDNA panels (CSRT30, HKRT102) and match paired RNA (Supplemental Table 3.1) samples were ran using the aforementioned protocols, reagents and instrumentation.

Primer design

Primers sequences were obtained either from qPrimerDepot (https://primerdepot.nci.nih.gov/) or designed using Primer3 Plus (www.primer3plus.com/) using the qPCR settings (Supplemental Table 3.2). All primers were synthesized by Integrated DNA Technologies.

Compartment lysis and expression analysis

Approximately 1 million cells were collected using 0.25% trypsin and spun at 400g for 3 minutes to pellet the cells. A total volume 175 ul of cytoplasmic lysis (50mM TrisCl pH 8.0, 140 mM NacCL, 1.5 mM MgCl2, 0.5% P-40, 1mM DTT) was used to resuspend the cells, and cells were then incubated on ice for 5 minutes for the HK-2 and CAKI-1 cells and 35 minutes for the ACHN cells. Following incubation, lysate was spun at 300g for 2 minutes at 4C. Supernatant and pellet were separated, and RNA was extracted as previous described. RNA was eluted with equal volumes from RNA extraction columns, and equal volumes of cytoplasmic and nuclear RNA were used for the reverse transcription reaction.

siRNA construction and transfection

All custom siRNAs sequences (Supplemental Table 3.3) were constructed using the MIT Whitehead software (http://sirna.wi.mit.edu/), and then used to make Silencer Select siRNAs synthesized by Ambion (Carlsbad, CA, USA). The validated Silencer Select negative siRNA #2 was used as negative control in transient knockdown assays. siRNAs were transfected using Lipofectamine3000

per manufacturer's recommendations to a final concentration of 100 nM. Cells were collected 60-72 hours following transfection.

RNA-seq analysis

Transfected CAKI-1 cells were trypsinized and RNA was extracted, as stated above. RNA quality and quantity were evaluated with a bioanalyzer and Thermoscientific Nanodrop2000 spectrophotometer. Single-end read RNA-seq libraries were constructed using the NEBNext Ultra Directional RNA library prep kit for Illumina per manufacturer's protocol. Samples were multiplexed and sequenced with the NEX-seq Illumina sequencing platform.

Western blot

Transfected CAKI-1 cells were scraped and lysed using a standard RIPA buffer. Lysate was spun at max speed and the supernatant was collected. Protein concentration was determined using Bradford reagent. A total of 10-20 ug of total protein was loaded and subjected to SDS-page. Protein was transferred to a nitrocellulose membrane using the Biorad Trans-Blot Turbo for 45 minutes at 25V. Membranes were blocked with milk for 1 hour and then probed with primary antibodies, β-actin at 1:7500, DDAH1 at 1:1000, VHL at 1:3000, and HIF1α at 1:3000, overnight. The next day, three washes with 1X TBST at 10 minutes were conducted. Subsequently, anti-mouse and anti-rabbit secondary HRP antibodies incubated with membrane for 1 hour at room temperature. Then, three washes with 1X TBST at 10 minutes were conducted and the membrane was exposed to autography and developed.

Bioinformatic analyses

A total of 614 fastq RNA-seq files were downloaded from TCGA legacy archive website (https://portal.gdc.cancer.gov/legacy-archive/search/f). Human cDNA and ncRNA FASTA formatted transcript files (Ensembl v89 annotation) were acquired form the Ensembl ftp site (https://www.ensembl.org/info/data/ftp/index.html), and merged to create a master file of all putative coding and non-coding transcripts.

Transcript quantifications and differential expression analyses were performed using the cufflink suite (TCGA data) or the kallisto-sleuth pipeline (RNA-seq analysis) [22-24]. Cufflinks was used to obtain transcript quantifications [23]. Calculated transcript quantifications were then used to generated tumor/normal ratios. A two-tailed Wilcoxon signed rank test was performed to determine statistical significance. Cuffdiff was used to confirm differential expression. Using the default settings, kallisto was used to create an index for quantification using the aforementioned FASTA master file. Subsequently, kallisto was used to quantify all putative transcripts using 50 bootstrap samples. Differential expression analysis was performed with sleuth using the Wald test with a cutoff of q-value <0.05 and beta >0.5.

For the gene-level analyses, alignment of the fastq files was performed first with HISAT2 using the hg38 human assembly [25]. Read counting was performed using the summarizeOverlaps package, with union mode [26]. Using the read counts, an edgeR analysis was performed using the default settings [27, 28].

The entire pipeline was performed within the systemPipeR package [29]. Normalization of the gene counts was performed using DESeq2 and then subsequently used in consensus clustering to determine the number molecular subtypes in ccRCC [30]. Consensus clustering was performed using the ConsensusClusterPlus R package [31]. A total 1,000 of the most variable genes, based on mean absolute deviation were used in the clustering generating consensus matrices for *k*=2-7. Number of molecular subtypes was determined based on the consensus matrices and the cumulative distribution functions for each *k*.

Results

HM1-3 downregulated in ccRCC

Using a commercially available multiple tissue cDNA array, qPCR was used to examine the expression levels of three experimentally validated *HM1* transcripts between 8 cancer tissues relative to their respective normal anatomical tissues (Figure 3.1A-B, Supplemental Figure 3.1A-B). Discovered in this initial survey was a novel finding showing a significant downregulation of *HM1-3* expression in RCCs. *HM1-3* downregulation was also observed in breast and colorectal cancers, which is consistent with previous reports [15, 32]. Further examination of *HM1-3* expression in RCC, in a second independent cDNA array, demonstrated *HM1-3* downregulation was shown to be restricted largely to ccRCC (Figure 3.1C). An average ~5.5 fold downregulation in *HM1-3* expression

was seen when comparing 9 normal renal tissue samples to 21 ccRCC samples. No statistically significant *HM1-3* downregulation was observed in papillary renal cell carcinomas (pRCC). These results were supported further using 12 ccRCC matched pair samples, which showed 11 ccRCC samples with a *HM1-3* downregulation relative to their normal adjacent tissue (Figure 3.1D). No statistically significant differentially expression was found for *HM1-2-3* and *Unspliced HM1.*

To further substantiate *HM1-3* downregulation in ccRCC, 614 RNA-seq datasets (72 normal and 542 ccRCC samples) from TCGA were bioinformatically examined. Evaluation of *HM1* FPKM tumor/normal ratios, using 50 matched paired samples contained within these data, showed *HM1-3* to be the only significantly downregulated *HM1* transcript, as determined with a Wilcoxon signed ranked test (Figure 3.1E). This result was later validated in a Cuffdiff analysis. Among ~250,000 transcripts identified with cufflinks, using the forementioned matched pair samples, *HM1-3* was found within 1,710 differentially expressed transcripts identified in the analysis, using a threshold of >2 fold change and <0.2% FDR. Subsequently, absolute levels of *HM1* transcripts in normal renal tissue were examined. A composite trace of 72 merged alignments files from normal renal tissue and the transcript quantifications using cufflinks shows *HM1-3* is the most abundant *HM1* transcript in normal renal tissue (Figure 3.1F, top). These results were validated with PCR using 12 normal renal tissue samples (Figure 3.1F, bottom).

As ccRCC is a highly herterogeneous cancer, *HM1-3* expression was explored within the different subtypes of ccRCC. Consensus clustering was performed using gene-level read counts from the 542 ccRCC samples, which showed four distinct subtypes of ccRCC (Figure 3.1G, left, Supplemental Figure 3.2) consistent with previous findings [6, 33]. Among the four subtypes identified are the two major molecularly distinct ccA and ccB subtypes, in addition to the mixed ccA/ccB and distal tubule subtypes. Using a two-tailed Student's t-test, a significant *HM1-3* downregulation would found within all subtypes of ccRCC (Figure 3.1G, right).

Characterization of the HM1 transcripts in proximal tubule renal cells

In effort to identify a suitable *in vitro* model, four proximal tubule renal cell lines were investigated for their *HM1* expression (Figure 3.2). Using qPCR, absolute levels of *HM1-3* and *HM2-3* expression were observed to be low in HRPTEpC, HK-2 and ACHN cells, with *HM1-3* copies per cell (Figure 3.2A). CAKI-1 had the most copies of *HM1-3*, with approximately 10 copies per cell with each of the spliced *HM1* transcripts. Alternatively, significantly higher levels of *Unspliced HM1* were observed within HK-2, ACHN and CAKI-1 cell lines. Other cell lines were also evaluated for *HM1* absolute expression, such as MCF10A, MCF7 and DAOY, all of which showed similar or lower levels of expression (data not shown). Subsequently, subcellular fractionation followed by qPCR, showed both spliced transcripts, *HM1-3* and *HM2-3,* were found predominantly within the cytoplasm, while *Unspliced HM1* was observed mostly within the nucleus in all

three renal cell lines (Figure 3.2B). *HOXA1* mRNA and *MALAT1* lncRNA were used as cytoplasmic and nuclear controls, respectively.

HM1 knockdown alters hypoxia pathway in CAKI-1 cells

As CAKI-1 cells had the most *HM1-3* expression among the cell lines, a stranded RNA-seq analysis was performed evaluating the transcriptomic effects of a targeted *HM1* knockdown in CAKI-1 cells (Figure 3.3). Collectively, 40 genes were differentially expressed with *HM1* knockdown between edgeR and sleuth analyses (Figure 3.3A-B). Using edgeR, 28 genes were differentially expressed (16 upregulated and 12 downregulated) with *HM1* knockdown, with a 1.25 fold change and 5% FDR threshold. Alternatively, using kallisto gene counts and sleuth for differential expression analysis, 14 differentially expressed genes (8 upregulated and 6 downregulated) were observed with a 0.5 bias estimator value and 5% FDR threshold. Only *DDAH1* and *MELTF* were found in common between both edgeR and sleuth analyses. As the upregulated DEGs identified in the RNA-seq showed fewer type I errors (seen below), a comprehensive Enrichr analysis was performing using the 22 identified upregulated genes. No statistically significant GO terms were identified.

A total of 10 upregulated DEGs identified between the two RNA-seq analyses were partially randomly selected for qPCR validation. Among the 10 genes, 8 genes were confirmed with qPCR: *ADAM19*, *H3F3C*, *CUTA*, *DDAH1, GXYLT1, ZC3H18, ANGPTL4* and *CDKN1C.* Conversely, only 2 of the 7 downregulated DEGs were validated using qPCR (Figure 3.3C). In a second validation, using

one inefficient *HM1* siRNA (HM1 siRNA #1) and a more efficient *HM1* siRNA (HM1 siRNA #3), *DDAH1* and *ANGPTL4* were the only genes consistently deregulated with efficient *HM1* knockdown conditions (Figure 3.3D). In effort to discern which *HM1* transcript(s) were contributing to the *DDAH1* and *ANGPTL4* upregulation, a *HM1* knockdown was performed in CAKI-1 cells constitutively overexpressing the *HM1-3* transcript (Figure 3.3D, left). With *HM1* knockdown, an attenuation of the *ANGPTL4* and *DDAH1* upregulation was observed in the *HM1-3* overexpressing cells relative to the parental CAKI-1 cells (Figure 3.3D, right). The pre-mRNA *ANGPTL4* and *DDAH1* expression was also evaluated to explore a possible mechanism for the observed upregulation of *ANGPTL4* and *DDAH1* steady-state expression. A concomitant upregulation in *ANGPTL4* precursor was observed with *HM1* knockdown, while the *DDAH1* precursor was unaffected (Figure 3.3E).

As the Enrichr analysis of the identified DEGs suggested a connection to HIF1α, and *ANGPTL4* is known target of HIF1α, protein expression of ANGPTL4, DDAH1 and the HIFs were evaluated (Supplemental Table 3.4). With *HM1* knockdown, significant increases in HIF1α were observed (Figure 3.3E). However, no statistically significant changes were observed with DDAH1 and VHL expression, and ANGPTL4 protein expression is still under investigation. Lastly, evaluation of the tumor/normal ratios of TPM values (generated using kallisto), show a highly significant upregulation in *ANGPTL4* expression in ccRCC relative to normal adjacent tissue (Figure 3.3F, left). Alternatively,

DDAH1 expression exhibited a significant downregulation in ccRCC relative to normal adjacent tissue (Figure 3.3F, right). Statistical significance was determined using the Wilcoxon signed-rank test for the match paired samples.

To explore *HM1*, *DDAH1* and *ANGPTL4* further, we exposed CAKI-1 cells to 100 uM cobalt chloride, a mimic for hypoxia-induced stress. As anticapted, we observed a significant induction of HIF1α protein (not shown) and the HIF1α responsive gene, *ANGPTL4,* at approximately 4 hours of exposure to cobalt chloride (Figure 3.4). Little expression change was observed with *DDAH1.* Alternatively, similar levels of reduction in all three isoforms of *HM1* and *HIF1α* were observed starting at 4 hours of exposure.

HM1 knockdown dedifferentiates mouse kidney progenitor cells (results produced by M. Young)

To examine a suspected role of *HM1* in kidney differentiation, mES cells were differentiated into kidney progenitor cells, as previously described (Figure 3.5A) [34]. Gene expression levels of mesoderm-commitment, intermediate mesoderm and metanephric mesenchyme markers demonstrated successful timed induction of the corresponding markers, signifying differentiation of mES into kidney progenitor cells (Figure 3.5B). Evaluation of all three *HM1* isoforms showed peak expression at the end of the induction process at day 8 (Figure 3.5C). Knockdown *HM1* in kidney progenitor cells showed slight reductions in *OSR1* and *GDNF* expression and a large reduction in *PAX2* expression (Figure 3.5D).

No other significant change in kidney differentiation markers were observed with *HM1* knockdown.

Discussion

In the current study, we provide new insights in the functional role(s) of the *HM1* lncRNA in kidney biology. In our results, we identify a novel and highly pervasive downregulation of *HM1-3* in ccRCCs, not previously reported. We demonstrate that *HM1-3* is the only *HM1* isoform downregulated in ccRCC, and it is the most abundant *HM1* transcript found in normal kidney tissue. Furthermore, we show that *HM1-3* is largely localized to the cytoplasm in renal proximal tubule cells, and knockdown of *HM1* alters the hypoxia pathway in ccRCC cells. Lastly, we provide evidence suggesting a functional role of *HM1* in kidney differentiation and maintenance.

As both HIF1α protein expression and *ANGPTL4* expression, a HIF1α target gene, are increased with *HM1* knockdown, these findings strongly suggest *HM1* regulates the hypoxia pathway in ccRCC cells. However, it remains unclears whether the induction of HIF1α is responsible for the increase in *ANGPTL4* expression. Preliminary evidence suggests that HIF1α is not responsible for increased *ANGPTL4* expression, as *HIF1α* knockdown also elicits an induction of *ANGPTL4* expression (data not shown). Currently, HIF1β, HIF2α and aryl hyrdrocarbon receptor (AHR) are being explored to explain the *ANGPTL4* induction seen with *HM1* knockdown, as all of these proteins are likely transcription factors for *ANGPTL4.* Furthermore, *HM1* expression is reduced

with cobalt chloride exposure, mimicking the expression profile seen with *HIF1α* expression, suggesting *HM1* is part of the hypoxia-induced response of ccRCC cells. Additionally, as the HM1 also increases *ANGPTL4* pre-mRNA, *HM1* is possibly changing the transcriptional output of *ANGPTL4* and not affecting the degradation of the *ANGPTL4* mRNA. *DDAH1* was similarly upregulated with *HM1* knockdown; however, *DDAH1* pre-mRNA and protein were not affected and *DDAH1* is highly downregulated in ccRCC. These data suggest that *DDAH1* is not contributing to the ccRCC pathology and could be a compensatory response of the cells to *HM1* knockdown.

The findings observed from differentiating mES cells suggest the role of *HM1* extends more broadly to an essential kidney biological function. As the expression profiles of *HM1* with differentiating mES mirror the expression profiles of *WNT11*, *GDNF* and *CDH11*, HM1 is possibly involved in the later metanephric mesenchyme commitment stage of kidney cell differentiation. Additionally, *HM1* expression appears to be necessary to maintain the kidney progenitor state, as loss of *HM1* expression with knockdown shows a suppression of kidney differentiation markers, *OSR1*, *GDNF* and *PAX2*. As the knockdown of *HM1* does not preferentially or exclusively affect *HM1-3*, it is unclear which of the *HM1* transcript is responsible for the maintenance of the differentiated state. However, it is suspected that *HM1-3* is likely the transcript responsible for loss of the differentiated state, as it is significantly more abundant than the unspliced *HM1* transcript.

Collectively, the study provides new evidence of a cell-type specific regulatory function of *HM1* influencing the hypoxia pathway and differentiation state in kidney cells. Immediate future studies exploring the mechanisms of *HM1* would be of greatest interest, as it remains unknown how *HM1* functions in kidney cells.

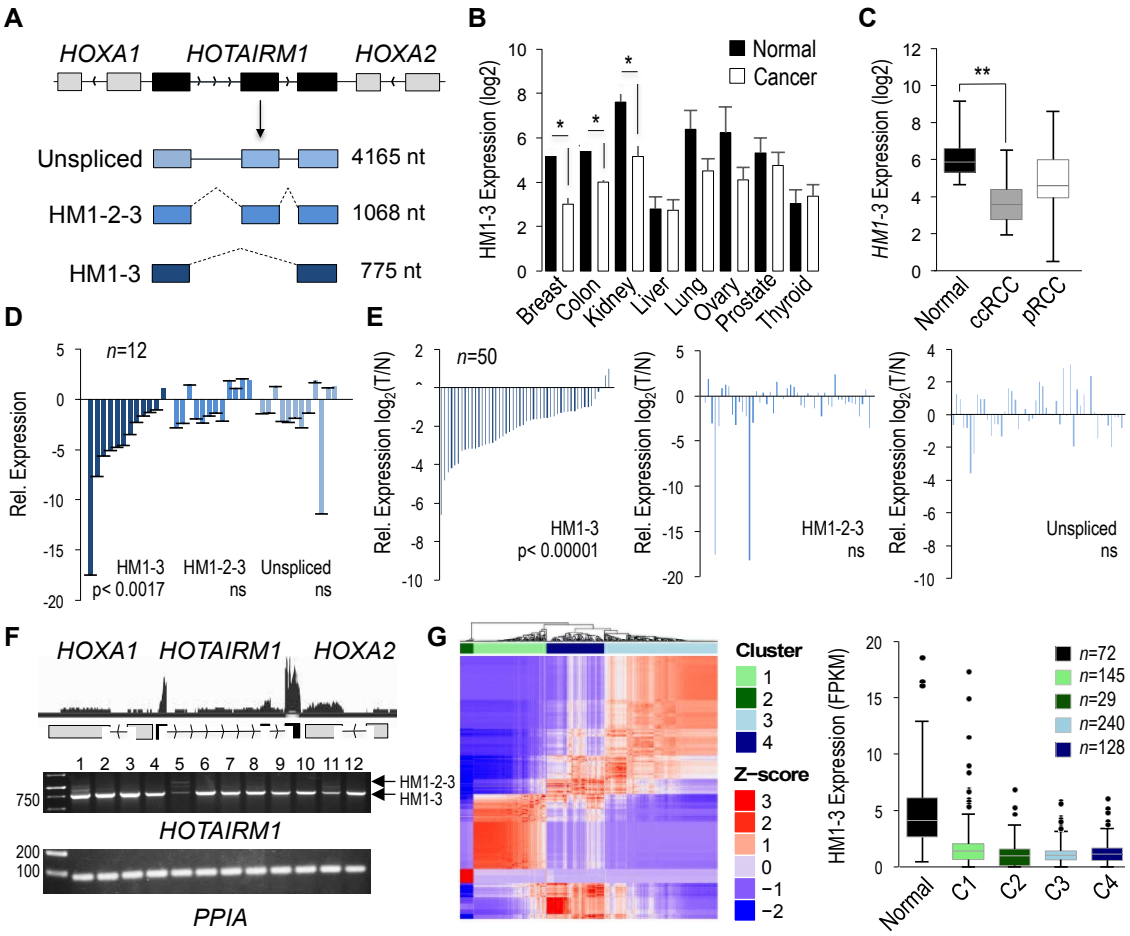


Figure 3.1. Reduced *HM1-3* expression in ccRCC. **A-** The *HOTAIRM1* locus is located in the *HOXA* gene cluster between *HOXA1* and *HOXA2* and produces two spliced transcripts (*HM1-3, HM1-2-3)* and one unspliced transcript. **B-** Relative expression of *HM1-3* in eight human cancers relative to their respective normal tissue. *HM1-3* levels were analyzed by qPCR and normalized commercially to β-actin. **C-** Analysis of *HM1-3* expression by qPCR in normal tissues (n=9) versus ccRCC (n=21) and pRCC (n=10, papillary renal cell carcinoma) tumors was performed as in panel B**.** **D-** Analysis of *HM1-3* expression by qPCR in 12 ccRCC matched pair samples. Fold changes in

expression in tumor vs normal (ΔΔCt) are indicated**. E-** FKPM tumor/normal ratios of 50 ccRCC match paired TCGA samples for all *HM1* transcripts**. F-** Quantification of *HM1* transcripts within normal adjacent samples. Shown is the merged trace of 72 normal renal RNA-seq datasets and validation using 12 independent normal renal samples using PCR, showing relative abundance of the *HM1* transcripts**. G-** Unsupervised consensus clustering identified 4 distinct ccRCC subtypes (left). *HM1-3* expression with respect to their ccRCC subtype classification (right). Statistical significance was determined by using two-tailed Student's t-test for panels B and C and the Wilcoxon signed-rank test for the match paired samples in panel D and E (* $p<0.01$, ** $p<0.005$, ns >0.01).

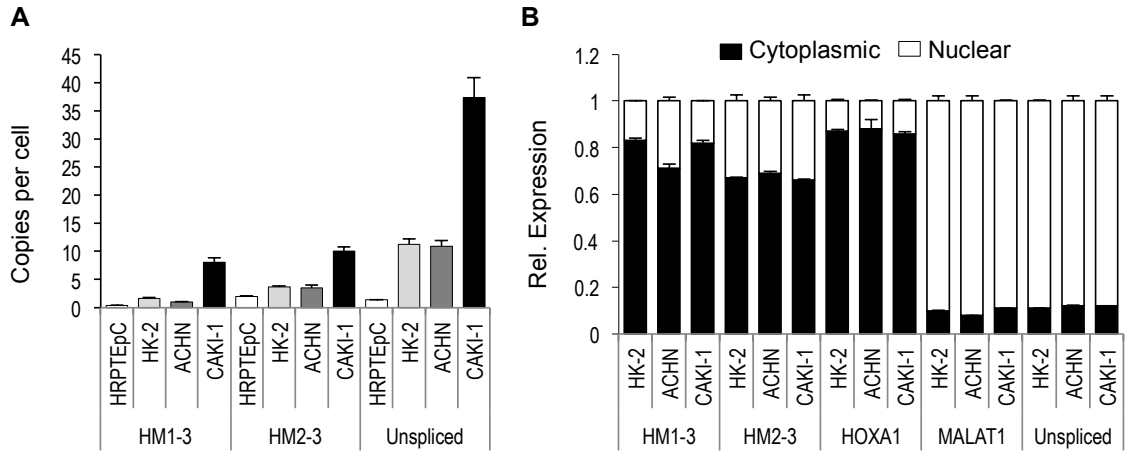


Figure 3.2. Characterization of *HM1* transcripts in proximal renal tubule cell lines. A - Copies per cell of *HM1* transcripts using qPCR. Error bars represent SEM across three biological replicates. HRPTEpC copies per cell inferred based on 7 pg total RNA per cell. **B -** Subcellular localization of *HM1* transcripts. *MALAT1* serves as a nuclear transcript control and *HOXA1* serves as a cytoplasm transcript control. Error bars represent SEM across three technical replicates.

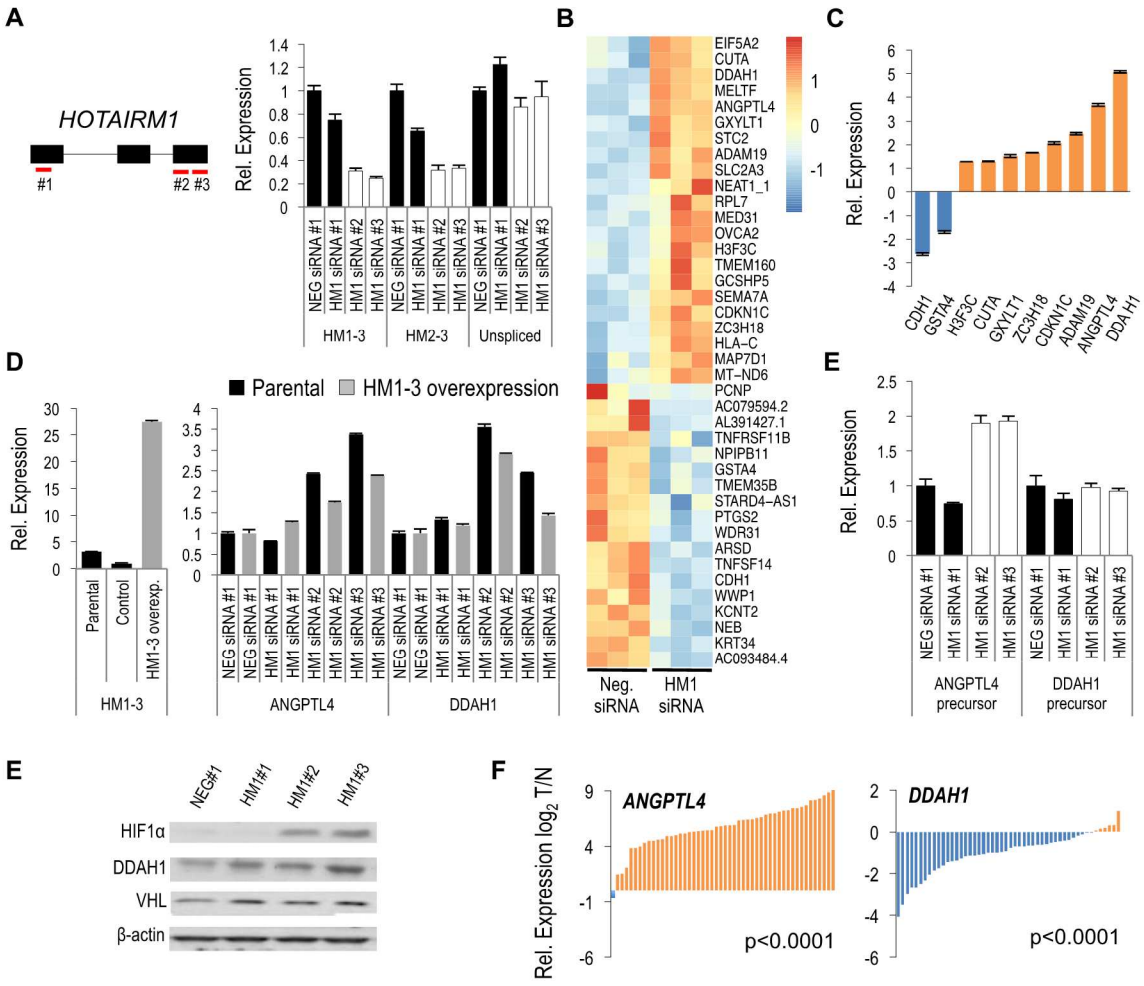


Figure 3.3. Hypoxia pathway altered with *HM1* knockdown in CAKI-1 cells. **A**– *HM1* knockdown with three *HM1* siRNAs. **B**– A total of 40 DEGs collectively identified between edgeR (fold change >1.25 and FDR <0.05) and sleuth analyses (β >0.5 and FDR <0.05) in RNA-seq experiment using *HM1* siRNA #2. **C**– qPCR validation of 10 identified DEGs in RNA-seq using *HM1* siRNA #2. **D** – Validation of *ANGPTL4* and *DDAH1* upregulation with *HM1* knockdown in parental and *HM1-3* overexpression CAKI-1 cells. **E**– *ANGPTL4* and *DDAH1* precursor expression with *HM1* knockdown. **F**– Western blot evaluating HIF1α,

DDAH1 and VHL protein expression with *HM1* knockdown. **G**- TPM tumor/normal mRNA ratios of 50 ccRCC match paired TCGA samples for *ANGPTL4* and *DDAH1.*

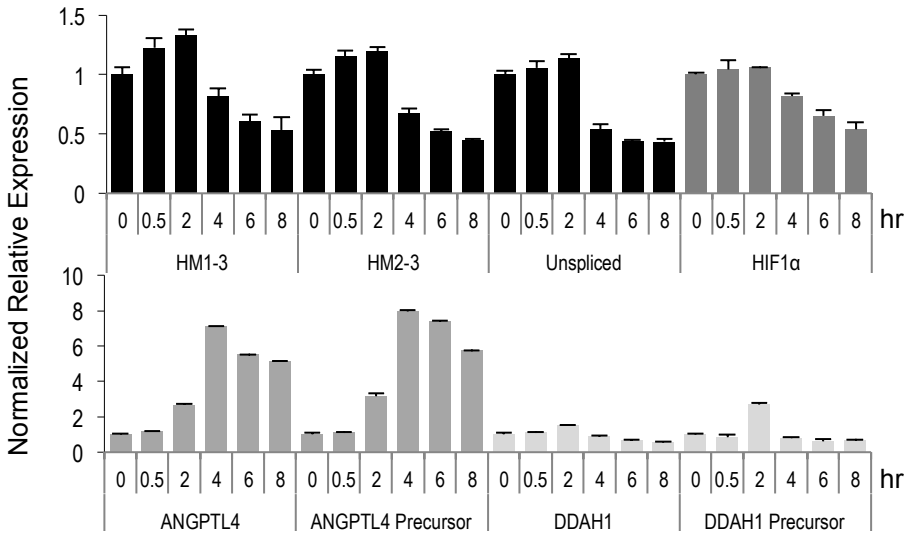


Figure 3.4. *HM1*, *ANGPTL4* and *HIF1α* RNA expression altered with cobalt chloride exposure. CAKI-1 cells exposed to 100uM cobalt chloride at increasing time intervals. Relative expression normalized to *PPIA* expression.

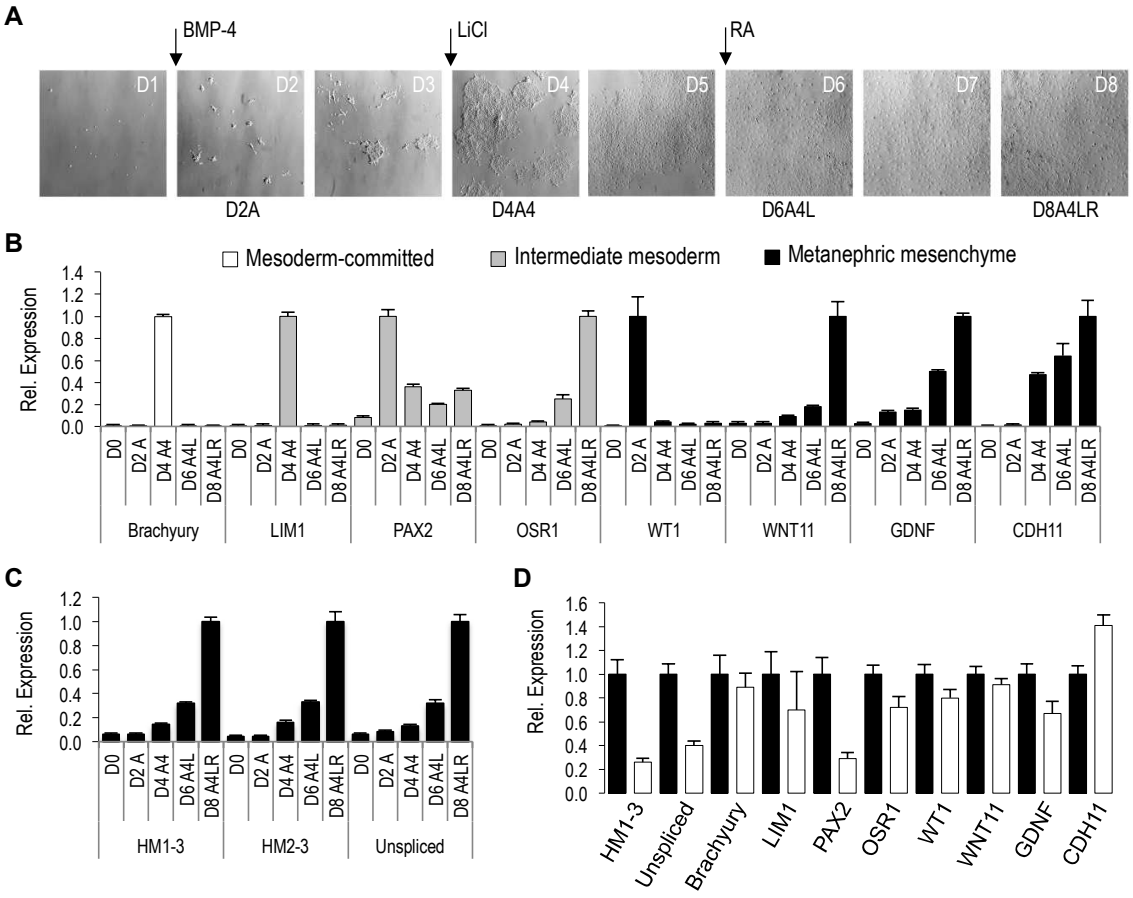


Figure 3.5. ***HM1*** **knockdown dedifferentiates mouse kidney progenitor cells.** **A –** mES cells sequentially exposed to activin A (Day 0, D0), BMP-4 (Day 2, D2), LiCL (Day 4, D4) and RA (Day 6, D6) to differentiate into kidney progenitor cells. A=activin A, 4=BMP-4, L=LiCL, R=RA. **B –** Gene expression of mesoderm-commitment, intermediate mesoderm and metanephric mesenchyme markers during differentiating time course. **C**- *HOTAIRM1* expression during differentiating time course. **D**- *HM1* knockdown in kidney progenitor cells and its effects on kidney differentiation markers. Gene expression normalized to

POLR2A, GAPDH and *ACTB.* Error bars represent the SEM across three biological replicates.

References

1. U.S. Cancer Statistics Working Group. (2017). United States Cancer Statistics: 1999-2014 Incidence and Mortality Web-based Report. Atlanta: US Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute.

2. Seitz W, Karcher KH, Binder W. Radiotherapy of metastatic renal cell carcinoma. Semin Surg Oncol. 1988; 4: 100-2. doi:

3. Ferguson RE, Jackson SM, Stanley AJ, Joyce AD, Harnden P, Morrison EE, Patel PM, Phillips RM, Selby PJ, Banks RE. Intrinsic chemotherapy resistance to the tubulin-binding antimitotic agents in renal cell carcinoma. Int J Cancer. 2005; 115: 155-63. doi: 10.1002/ijc.20816.

4. American Cancer Society. (2018). Cancer Facts & Figures 2018. Atlanta, Ga: American Cancer Society.

5. Blondeau JJ, Deng M, Syring I, Schrodter S, Schmidt D, Perner S, Muller SC, Ellinger J. Identification of novel long non-coding RNAs in clear cell renal cell carcinoma. Clin Epigenetics. 2015; 7: 10. doi: 10.1186/s13148-015-0047-7.

6. Malouf GG, Zhang J, Yuan Y, Comperat E, Roupret M, Cussenot O, Chen Y, Thompson EJ, Tannir NM, Weinstein JN, Valero V, Khayat D, Spano JP, et al. Characterization of long non-coding RNA transcriptome in clear-cell renal cell carcinoma by next-generation deep sequencing. Mol Oncol. 2015; 9: 32-43. doi: 10.1016/j.molonc.2014.07.007.

7. Tseng YY, Moriarity BS, Gong W, Akiyama R, Tiwari A, Kawakami H, Ronning P, Reuland B, Guenther K, Beadnell TC, Essig J, Otto GM, O'Sullivan MG, et al. PVT1 dependence in cancer with MYC copy-number increase. Nature. 2014; 512: 82-6. doi: 10.1038/nature13311.

8. Hamilton MJ, Young MD, Sauer S, Martinez E. The interplay of long non-coding RNAs and MYC in cancer. AIMS Biophys. 2015; 2: 794-809. doi: 10.3934/biophy.2015.4.794.

9. Gordan JD, Bertout JA, Hu CJ, Diehl JA, Simon MC. HIF-2alpha promotes hypoxic cell proliferation by enhancing c-myc transcriptional activity. Cancer Cell. 2007; 11: 335-47. doi: 10.1016/j.ccr.2007.02.006.

10. Gordan JD, Lal P, Dondeti VR, Letrero R, Parekh KN, Oquendo CE, Greenberg RA, Flaherty KT, Rathmell WK, Keith B, Simon MC, Nathanson KL. HIF-alpha effects on c-Myc distinguish two subtypes of sporadic VHL-deficient

clear cell renal carcinoma. Cancer Cell. 2008; 14: 435-46. doi: 10.1016/j.ccr.2008.10.016.

11. Grampp S, Platt JL, Lauer V, Salama R, Kranz F, Neumann VK, Wach S, Stohr C, Hartmann A, Eckardt KU, Ratcliffe PJ, Mole DR, Schodel J. Genetic variation at the 8q24.21 renal cancer susceptibility locus affects HIF binding to a MYC enhancer. Nat Commun. 2016; 7: 13183. doi: 10.1038/ncomms13183.

12. Hamilton MJ, Girke T, Martinez E. Global isoform-specific transcript alterations and deregulated networks in clear cell renal cell carcinoma. Oncotarget. 2018; 9: 23670-80. doi: 10.18632/oncotarget.25330.

13. Zhang X, Lian Z, Padden C, Gerstein MB, Rozowsky J, Snyder M, Gingeras TR, Kapranov P, Weissman SM, Newburger PE. A myelopoiesis-associated regulatory intergenic noncoding RNA transcript within the human HOXA cluster. Blood. 2009; 113: 2526-34. doi: 10.1182/blood-2008-06-162164.

14. Zhang X, Weissman SM, Newburger PE. Long intergenic non-coding RNA HOTAIRM1 regulates cell cycle progression during myeloid maturation in NB4 human promyelocytic leukemia cells. RNA Biol. 2014; 11: 777-87. doi:

15. Wan L, Kong J, Tang J, Wu Y, Xu E, Lai M, Zhang H. HOTAIRM1 as a potential biomarker for diagnosis of colorectal cancer functions the role in the tumour suppressor. J Cell Mol Med. 2016; 20: 2036-44. doi: 10.1111/jcmm.12892.

16. Zhang X, Sun S, Pu JK, Tsang AC, Lee D, Man VO, Lui WM, Wong ST, Leung GK. Long non-coding RNA expression profiles predict clinical phenotypes in glioma. Neurobiol Dis. 2012; 48: 1-8. doi: 10.1016/j.nbd.2012.06.004.

17. Zhou Y, Gong B, Jiang ZL, Zhong S, Liu XC, Dong K, Wu HS, Yang HJ, Zhu SK. Microarray expression profile analysis of long non-coding RNAs in pancreatic ductal adenocarcinoma. Int J Oncol. 2016; 48: 670-80. doi: 10.3892/ijo.2015.3292.

18. Diaz-Beya M, Brunet S, Nomdedeu J, Pratcorona M, Cordeiro A, Gallardo D, Escoda L, Tormo M, Heras I, Ribera JM, Duarte R, de Llano MP, Bargay J, et al. The lincRNA HOTAIRM1, located in the HOXA genomic region, is expressed in acute myeloid leukemia, impacts prognosis in patients in the intermediate-risk cytogenetic category, and is associated with a distinctive microRNA signature. Oncotarget. 2015; 6: 31613-27. doi: 10.18632/oncotarget.5148.

19. Chen ZH, Wang WT, Huang W, Fang K, Sun YM, Liu SR, Luo XQ, Chen YQ. The lncRNA HOTAIRM1 regulates the degradation of PML-RARA

oncoprotein and myeloid cell differentiation by enhancing the autophagy pathway. Cell Death Differ. 2017; 24: 212-24. doi: 10.1038/cdd.2016.111.

20. Dupasquier S, Delmarcelle AS, Marbaix E, Cosyns JP, Courtoy PJ, Pierreux CE. Validation of housekeeping gene and impact on normalized gene expression in clear cell renal cell carcinoma: critical reassessment of YBX3/ZONAB/CSDA expression. BMC Mol Biol. 2014; 15: 9. doi: 10.1186/1471-2199-15-9.

21. Jung M, Ramankulov A, Roigas J, Johannsen M, Ringsdorf M, Kristiansen G, Jung K. In search of suitable reference genes for gene expression studies of human renal cell carcinoma by real-time PCR. BMC Mol Biol. 2007; 8: 47. doi: 10.1186/1471-2199-8-47.

22. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016; 34: 525-7. doi: 10.1038/nbt.3519.

23. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012; 7: 562-78. doi: 10.1038/nprot.2012.016.

24. Pimentel H, Bray NL, Puente S, Melsted P, Pachter L. Differential analysis of RNA-seq incorporating quantification uncertainty. Nat Methods. 2017; 14: 687-90. doi: 10.1038/nmeth.4324.

25. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015; 12: 357-60. doi: 10.1038/nmeth.3317.

26. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013; 9: e1003118. doi: 10.1371/journal.pcbi.1003118.

27. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26: 139-40. doi: 10.1093/bioinformatics/btp616.

28. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 2012; 40: 4288-97. doi: 10.1093/nar/gks042.

29. TW HB, Girke T. systemPipeR: NGS workflow and report generation environment. BMC Bioinformatics. 2016; 17: 388. doi: 10.1186/s12859-016-1241-0.

30. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014; 15: 550. doi: 10.1186/s13059-014-0550-8.

31. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics. 2010; 26: 1572-3. doi: 10.1093/bioinformatics/btq170.

32. Novak P, Jensen T, Oshiro MM, Wozniak RJ, Nouzova M, Watts GS, Klimecki WT, Kim C, Futscher BW. Epigenetic inactivation of the HOXA gene cluster in breast cancer. Cancer Res. 2006; 66: 10664-70. doi: 10.1158/0008-5472.CAN-06-2761.

33. Cancer Genome Atlas Research N. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature. 2013; 499: 43-9. doi: 10.1038/nature12222.

34. Nishikawa M, Yanagawa N, Kojima N, Yuri S, Hauser PV, Jo OD, Yanagawa N. Stepwise renal lineage differentiation of mouse embryonic stem cells tracing in vivo development. Biochem Biophys Res Commun. 2012; 417: 897-902. doi: 10.1016/j.bbrc.2011.12.071.

Supplemental table 3.1. Origene ccRCC match paired samples.

Match Pair #	Catalog Number	Case ID	Sample Classification
1	CR563036	CU0000000807	Tumor
	CR562944	CU0000000807	Normal
2	CR559247	CU0000006303	Tumor
	CR561460	CU0000006303	Normal
3	CR560088	CI0000010082	Tumor
	CR560086	CI0000010082	Normal
4	CR561100	CI0000000216	Tumor
	CR559748	CI0000000216	Normal
5	CR560856	CI0000005561	Tumor
	CR560857	CI0000005561	Normal
6	CR560960	CI0000005877	Tumor
	CR560957	CI0000005877	Normal
7	CR560907	CI0000006155	Tumor
	CR560906	CI0000006155	Normal
8	CR559302	CI0000006640	Tumor
	CR560658	CI0000006640	Normal
9	CR559596	CI0000009997	Tumor
	CR560141	CI0000009997	Normal
10	CR561841	CU0000011475	Tumor
	CR559696	CU0000011475	Normal
11	CR559768	CU0000012615	Tumor
	CR561775	CU0000012615	Normal
12	CR559682	CU0000012830	Tumor
	CR561813	CU0000012830	Normal

Supplemental table 3.2. qPCR primer sequences

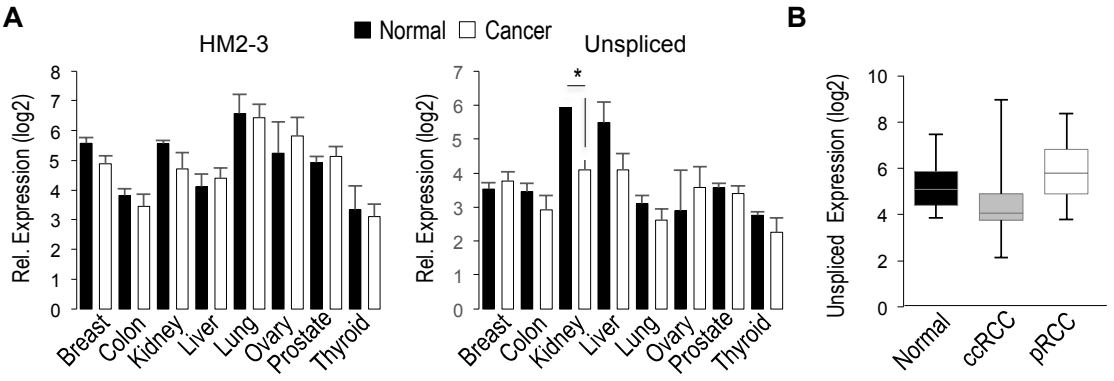
Gene/Transcript	Primer Sequence
HM1-3	AAGATGAACTGGCGAGAGGTC
	TTTCAAACACCCACATTTCAACC
HM2-3	CATCGCGTTGTCATTGGAAC
	TTCAGGCAAAACAGACCGTGA
Unspliced	GCAACAACCCAGTGACACAC
	TGCTTCGAAGTCAGGTTAGC
HOXA1	TTGACCCAGGTAGCCGTACT
	TCTTCTCCAGCGCAGACTTT
MALAT1	AGGGACTGGAGCTGCTTTTATC
	TGAACCAAAGCTGCACTGTG
RPL7	GCCATATATTGCATGGGGGTAC
	TGCCATAACCACGCTTGTAG
ADAM19	AGCACTTGCCCCAAAGTTTC
	AGCTCAAGGAAAGGGAGAAGC
H3F3C	TGGTGGGTCTGTTGGAAGATAC
	TGTCTTTGGGCATGATGGTG
MAP7D1	AAGGAGGCTGTGCAGAAAGAG
	AGAAGCCATTCTCCTGGTGTG
CUTA	TGCAGCCTTTGTTACTTGCC
	AATCTGAGGGATGAGGTTGACG
DDAH1	AGTGAATCTGCACAGAAGGC
	ACAGTGAGTTTGTCGTAGCG
GXYLT1	TGGCTCATGCATGTAATCCC
	ACTTTGTCACCTAGGCTGGTC
ZC3H18	GAGGACGATGATGGAGAAATCG
	ACTGGGGTCCTTCACTTCAC
ANGPTL4	CGCCATTTTTGGTGAACTGC
	TTGAAGTCCACTGAGCCATCG
CDKN1C	ACGCACTAGCTCGGTTATTG
	GCTACAGCTTGTGAGTGACC
PTGS2	ATGATTGCCCGACTCCCTTG
	TGGGGATCAGGGATGAACTTTC
ARSD	CAGCATCTTCACGCAGCAC
	TGCGGGGTCGTGTAATGAAC
GSTA4	CAGTTGTACAAGTTGCAGGATGG
	TCCCGTCAATTTCAACCATGGG
CDH1	AATGGGGCAATCGCTTCAAG
	ACCACCAGCAACGTGATTTC
WDR31	TGGCTGCTTTGAACTCAGAC
	TGGTGATCTCATGTTCATGTCC
TNFRSF11B	ATTTGGAGTGGTGCAAGCTG
	AGGGTGCTTTAGATGACGTCTC

Supplemental table 3.2 continued. qPCR primer sequences

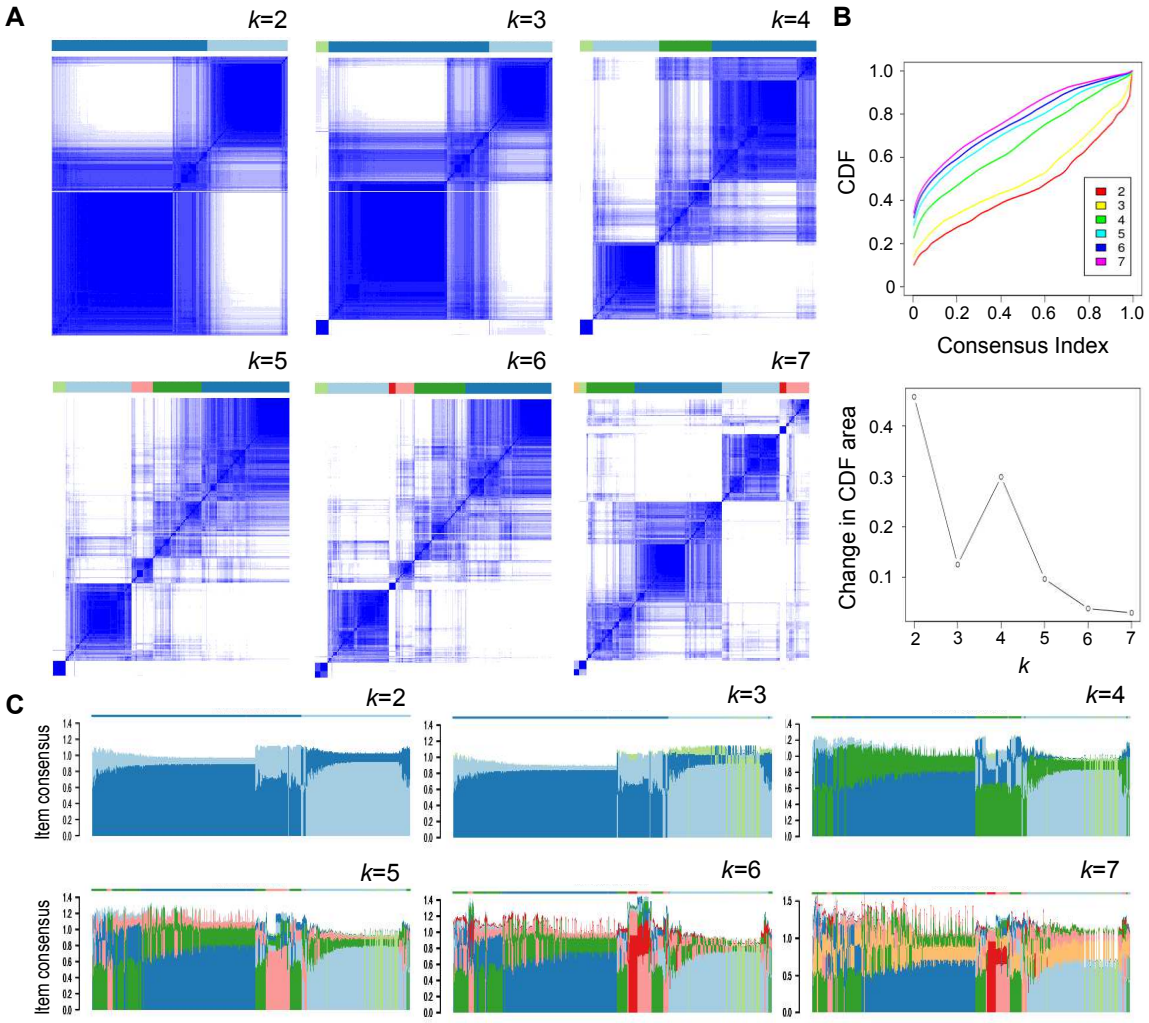
Gene/Transcript	Primer Sequence
WWP1	CTGCCGATGACACTGTTAATGG
	TACTGGAGTACCCGTGACAG
ANGPTL4 precursor	ACAGCTGGCATTCATGGAAG
	AGTGACCAGGAAGACGCTTTC
DDAH1 precursor	AGGCCCTAACTGCTCTTCAAAG
	TCTACCCTGTCAAATGCCATCC
HIF1α	ACAGTAACCAACCTCAGTGTGG
	ATGGGTGAGGAATGGGTTCAC

Supplemental table 3.3. siRNA sequences

siRNA name	siRNA sequence (5'->3')
HM1 siRNA #1	UCAAUGAAAGAUGAACUGGtt
HM1 siRNA #2	CUGGGAGAUUAAUCAACCAtt
HM1 siRNA #3	GGAGACUGGUAGCUUAUUAaa



Supplemental Figure 3.1. *HM2-3 and Unspliced* transcripts unaffected in ccRCC. **(A)** Relative expression of *HM1-2-3* and *Unspliced* transcripts in eight human cancers relative to their respective normal tissue. *HM1-2-3* and *Unspliced* transcripts levels were analyzed by qPCR and normalized commercially to β-actin. **(B)** Analysis of *Unspliced HM1* expression by qPCR in normal tissues (n=9) versus ccRCC (n=21) and pRCC (n=10, papillary renal cell carcinoma) tumors was performed as in panel B. No significant *Unspliced HM1* differential expression observed in ccRCC or pRCC. Statistical significance was determined by using two-tailed Student's t-test for all panels (* $p<0.05$).



Supplemental Figure 3.2. Four major subtypes identified in ccRCC. A- Consensus matrices generated for *k*=2-7, using 1000 of the most variable genes. **B-** Cumulative distribution functions (CDF) for each *k* of the consensus matrix (top). Greatest relative change in the area under the CDF curves observed from *k=3* to *k=4* (bottom). **C-** Mean item consensus value for each cluster at a given *k.*

Supplemental table 4.3. Enrichr analysis results of 22 upregulated DEGs

Name	p-value	q-value	Z-score
SMAD_19615063_ChIP-ChIP_OVARY_Human	0.008122	0.9594	-4.09
HIF1A_21447827_ChIP-Seq_MCF-7_Human	0.005031	0.9594	-3.36
TP63_17297297_ChIP-ChIP_HaCaT_Human	0.04311	0.9594	-4.05
NANOG_18700969_ChIP-ChIP_MESCs_Mouse	0.05435	0.9594	-4.13
SOX9_22984422_ChIP-ChIP_TESTIS_Rat	0.06812	0.9594	-4.35
NANOG_21062744_ChIP-ChIP_HESCs_Human	0.01235	0.9594	-2.46
ERG_21242973_ChIP-ChIP_JURKAT_Human	0.04803	0.9594	-2.9

Chapter 4

The interplay of long non-coding RNAs and MYC in cancer

Published in AIMS Biophysics. 2015; 2(4): 794-809

doi: 10.3934/biophy.2015.4.794

Abstract

Long non-coding RNAs (lncRNAs) are a class of RNA molecules that are changing how researchers view eukaryotic gene regulation. Once considered to be non-functional products of low-level aberrant transcription from non-coding regions of the genome, lncRNAs are now viewed as important epigenetic regulators and several lncRNAs have now been demonstrated to be critical players in the development and/or maintenance of cancer. Similarly, the emerging variety of interactions between lncRNAs and MYC, a well-known oncogenic transcription factor linked to most types of cancer, have caught the attention of many biomedical researchers. Investigations exploring the dynamic interactions between lncRNAs and MYC, referred to as the lncRNA-MYC network, have proven to be especially complex. Genome-wide studies have shown that MYC transcriptionally regulates many lncRNA genes. Conversely, recent reports identified lncRNAs that regulate MYC expression both at the transcriptional and post-transcriptional levels. These findings are of particular interest because they suggest roles of lncRNAs as regulators of MYC oncogenic functions and the possibility that targeting lncRNAs could represent a novel avenue to cancer treatment. Here, we briefly review the current understanding of how lncRNAs regulate chromatin structure and gene transcription, and then focus on the new developments in the emerging field exploring the lncRNA-MYC network in cancer.

Introduction

In recent years, the investigations exploring the importance of how long non-coding RNAs (lncRNAs) influence epigenetic modifications and chromatin structure has truly been paradigm shifting in our fundamental understanding of how transcription is regulated in higher eukaryotes. Once considered to be "transcriptional noise" inherent to the large genomes of higher eukaryotic organisms, lncRNAs are now viewed as critical regulators of complex genomes and have added another layer of complexity to the molecular mechanisms that govern gene regulation. In humans there is ~2 times more genes that produce lncRNAs, an estimated ~48,000 lncRNA genes [1], than protein-coding genes, and only a very small fraction of these lncRNA genes have been characterized [1].

While lncRNAs are only a subset of the non-coding transcriptome, over the last several years these mysterious RNAs have stepped into the limelight. In particular, a topic of great interest has been how dysregulation of lncRNAs leads to the inappropriate epigenetic regulation of critical genes that are involved in the development and/or maintenance of cancer. Recent evidence suggests that MYC, a well-studied oncogenic transcription factor that is deregulated in most types of cancer and controls many cellular processes, including cell growth, metabolism, proliferation, differentiation and apoptosis [2-6], is an important mediator in the transcription of lncRNAs [7, 8]. In turn, new evidence suggests that lncRNAs can also control the expression of MYC [9]. In this review, we

briefly discuss our current understanding of the basic features of lncRNAs and how they regulate the epigenetic landscape and then focus on the emerging dynamic relationships between MYC and several lncRNAs as they pertain to cancer.

Characteristics of lncRNAs: structure and function

The generally accepted definition of a lncRNA is an RNA molecule longer than 200 nucleotides that does not code for a protein [10]. With the arrival of genome-wide platforms, such as microarrays and next-generation sequencing, and more sophisticated computational analyses of genome-wide data, the exploration of the non-coding transcriptome has developed a strong foothold in molecular research laboratories [11-14]. In a recent publication, it was estimated that there are ~110,000 different lncRNA transcripts within the human genome, with ~80,000 of these considered to be high confidence lncRNA transcripts, representing ~48,000 genes [1]. Through the use of sophisticated computational analyses, these high confidence lncRNAs transcripts were shown to have very limited or no appreciable coding potential [1]. LncRNAs share many similarities to protein-coding transcripts. LncRNAs undergo similar co-transcriptional and post-transcriptional processing. Many lncRNA transcripts are transcribed by RNA polymerase II (although some are transcribed by RNA polymerase III) [15], they share the same canonical splice sites and polyadenylation terminal signals and frequently contain a 5' cap and a polyadenylated 3'-end [16,17].

Importantly, there are also some notable dissimilarities between lncRNAs and mRNAs. LncRNAs can undergo some unconventional processing [18-20] and tend to have a higher degree of tissue-specific expression relative to protein-coding genes [14, 21, 22]. Additionally, the primary sequences of lncRNAs tend to be less conserved across species [23]. These data suggest that the structure (rather than sequence) of lncRNAs may be of greater importance when exploring the function of lncRNAs, but this topic of RNA biology remains challenging. LncRNAs that have been extensively studied, such as HOTAIR and MALAT1, have provided some of the initial insights into the importance of the structural features of lncRNAs [23-27]. Recent evidence has shown that secondary structural elements of lncRNAs that are evolutionarily conserved contain important protein-binding domains [27]. Several methodologies have been developed in an effort to determine the secondary and tertiary structures of RNA molecules [28]. Some of the most noteworthy techniques have used either specific nucleolytic enzymes or chemical modifications of the RNA molecules followed by sequencing [29-33]. Once sequenced, the secondary structure of the RNA molecules can be determined from the ends of the reads using advanced computational tools (Figure 4.1). For example, the technique frequently referred to as Structure-seq employs the use of dimethyl sulfate (DMS) which penetrates the cells and methylates N1 of adenines and N3 of cytosines when not involved in Watson-Crick base pairing [29, 35]. Reverse transcription of the DMS-methylated RNA, results in reverse transcriptase stopping at the methylated

bases generating cDNAs of different lengths that are then sequenced. From these data, the secondary structure of RNA molecules can be determined on a genome-wide level using computational models. However, most techniques for the structural analysis of RNA molecules involve *in vitro* conditions that may not retain the structural characteristics of lncRNAs *in vivo* [36]. In spite of the challenges in RNA structural biology, the elucidation of lncRNA structure appears to be of vital importance, if we are to fully understand the functional capabilities of lncRNAs.

LncRNAs have a myriad of functions within eukaryotic cells. One of the best-understood and studied functions of lncRNAs is how they modulate gene expression. LncRNAs have been described as “fine-tuners” of gene regulatory networks regulating gene expression both at the transcriptional and post-transcriptional levels via a variety of distinct mechanisms [37, 38]. LncRNAs can be cis-acting, regulating chromatin structure and transcription of neighboring genes, and/or trans-acting, regulating the transcription of genes at distant locations within the genome [10, 38]. There are three broad functional classifications for lncRNAs; they can serve as decoys, scaffolds and/or guides (Figure 4.2). In simple terms, lncRNAs that serve as decoys can associate with both regulatory RNAs and proteins, such as miRNAs, DNA-binding proteins and histone-modifying enzymes, and prevent their binding to specific target mRNAs or chromatin loci and/or inhibit their enzymatic activity [10, 37, 39, 40]. In a recent example, MALAT1 was shown to mediate the mRNA levels of serum

release factor (SRF), an important transcription factor in myogenesis, by acting as a sponge or competing endogenous RNA (ceRNA) for miR-133 [40]. Scaffold lnRNAs provide a platform onto which different molecular interactions can occur, such as protein-protein interactions, including interactions of distant chromatin loci [41]. For example, HOTAIR has been demonstrated to aid in protein ubiquitination by acting as a scaffold for E3 ubiquitin ligases containing Dzip3 and Mex3b RNA-binding domains and also their substrates, Ataxin-1 and Snurportin-1, respectively [41]. Lastly, guide lncRNAs aid in the recruitment of protein complexes to specific locations within the cell, such as recruitment of epigenetic modifying enzymes to specific genes [10, 37, 38, 42]. The Xist lncRNA and its role in dosage compensation is a prime example of a guide lncRNA. Xist has been shown to interact with the SHARP-SMRT complex and recruit it to the X chromosome, thereby activating HDAC3 leading to histone deacetylation and exclusion of RNA polymerase from the X chromosome [42]. As more lncRNAs are characterized, we suspect these functional classifications will change with the discovery of novel lncRNA cellular functions. All three of these broad functional classifications of lncRNAs are found within the lncRNA-MYC network, described below.

The lncRNA-MYC network

Over the last decade as lncRNAs have been drawing the attention of more researchers, as too has the lncRNA-MYC network gained the attention of many

investigators. In two recent reviews, the lncRNA-MYC network has been described: either by examining how lncRNAs influence MYC expression [9], or selectivity summarizing some of the interactions seen within the human lncRNA-MYC regulatory network [43]. Here, we will further expand on what is known about the lncRNA-MYC network by providing a comprehensive summary of the molecular interactions within this regulatory network (Table 4.1), with an emphasis on recent developments in the field demonstrating functional relationships between cancer-associated lncRNAs and MYC.

PVT1

We will begin by describing recent findings suggesting a reciprocal relationship between MYC (formerly c-MYC) and a well-known lncRNA, known as PVT1. Both the *MYC* and *PVT1* genes are located in the 8q24 chromosomal region, which is frequently referred to as a "gene desert" because it contains few protein-coding genes (Figure 4.3). However, several lncRNA genes have been discovered within this region. The 8q24 region has also been of particular interest because it is a frequent region of genomic alterations, including amplifications and translocation breakpoints, in several different types of cancer [44]. Moreover, aberrant overexpression of PVT1 has been discovered in many different human cancers [45]. As previous mentioned, MYC is an oncogenic transcription factor and can either activate or repress transcription [46]. In a recent study, *PVT1* was shown to contain two non-canonical MYC-binding sites, which were found to be important for the binding of both MYC and its paralog

MYCN (formerly N-MYC) to the promoter region of *PVT1*, with changes in H4 acetylation and *PVT1* mRNA production correlating with changes in MYCN occupancy at the PVT1 promoter [47]. As suggested by the authors, this study demonstrates that *PVT1* is a likely downstream target of MYCN. Conversely, PVT1 lncRNA has been shown to be important in the regulation of MYC expression. Through the use of chromosome engineering in mice and both loss and gain-of-function analyses in different human cancer cell lines, it was demonstrated that PVT1 was required for high MYC protein expression, via its capacity to protect MYC from phosphorylation and subsequent degradation [48]. PVT1 is an exceptionally interesting lncRNA, both in its ability to physically interact with and regulate MYC and its pivotal roles in many cancers, making it an attractive therapeutic target to combat different cancers. For a more extensive review of PVT1 and its oncogenic features, we will refer to a recent review by Colombo et al. [45].

The CCAT family

The colon cancer associated transcripts (CCATs) are a collection lncRNAs located on different chromosomes that have been both associated with and functionally demonstrated to be involved in the development of human colorectal cancers (CRC). Specifically, three of the best-characterized CCAT lncRNAs are CCAT1 (also known as CARLo5), CCAT2 and CCAT6. CCAT6, also known as MYCLo2, will be discussed below in the MYCLos section. While the *CCAT6* gene is located on chromosome 7, *CCAT1* and *CCAT2* are located in the gene

desert region of 8q24, near *MYC* and *PVT1*. With the use of genome-wide association (GWA) studies, the 8q24 region has been implicated in CRC [49-51]. From these GWA studies, CCAT1 was later identified and characterized as being a highly specific marker for CRC [52].

The interplay between MYC and CCAT1 involves many complex molecular interactions. Contained within the 8q24 region are several chromatin-looping interactions that have been shown to be tissue-specific [53] and have been suggested to regulate MYC expression [53-58]. One of the most studied structural elements found in the 8q24 region is an enhancer region located ~335 kb upstream of *MYC*, frequently referred as MYC-335 [55-57]. Located ~180 kb upstream of MYC-335 is *CCAT1*, and this region is considered to be a super-enhancer (Figure 4.3). A recent study showed a long-range physical interaction between MYC-335 and the promoter of *CCAT1*, suggesting that MYC-335 is important for CCAT1 expression [54]. Moreover, it was later demonstrated that a long isoform of CCAT1, referred to as CCAT1-L, was important in the maintenance of the chromatin interaction via its role in the recruitment of a transcription factor, called CCCTC-Binding factor or CTCF [58]. Moreover, CCAT1 has been suggested to also regulate MYC post-transcriptionally. Deng et al., found that CCAT1 was deregulated in hepatocellular carcinoma, and CCAT1 expression correlated with the progression of the malignancy and poor prognosis [59]. With the use of RNA immunoprecipitation, CCAT1 was discovered to function as a let-7 miRNA sponge, thereby disinhibiting MYC [59].

Adding to the complexity, MYC has been shown to bind to the promoter of *CCAT1* and upregulate its expression and promote proliferation and invasion of colon and gastric cancer cells [60, 61].

Also important to the regulation of MYC expression is the CCAT2 lncRNA. CCAT2 is transcribed from MYC-335, described above, and CCAT2 is overexpressed in CRC and has been shown to promote tumor growth and metastasis [62]. Moreover, in the same study, CCAT2 was also shown to upregulate transcription of *MYC* through the recruitment of TCF7L2 [62]. Recently, additional CCAT lncRNAs have been discovered; however, it remains unclear whether these novel CCAT lncRNAs are part of the lncRNA-MYC regulatory network [63]. Altogether, the CCAT lncRNA family is proving to be complex and important in the involvement of colorectal cancer and in the regulation of MYC expression.

MYCLos

MYCLos is as collective term for several lncRNAs included within the CCAT family, coined by a research group examining the importance of these lncRNAs in human CRC. In the original study, conducted by Kim et al., a microarray analysis was used to profile ~33,000 lncRNAs in both normal and CRC samples [63]. Their results revealed thousands of lncRNAs to be differentially expressed, including the CCAT1 and CCAT2 lncRNAs that had previously been suggested to be important in several stages of CRC [52, 54, 58, 62, 64, 65]. To further narrow their search and to isolate the lncRNAs that were both differentially

expressed in CRC and regulated by MYC, they examined the effects of MYC knockdown in different CRC cell lines. From these experiments, they identified three lncRNAs, referred to as MYCLo-1, MYCLo2 (also known as CCAT6), and MYCLo-3, that were transcriptionally upregulated by MYC. They later confirmed that MYCLos had influential roles in cell proliferation and cell cycle progression by regulating the expression of *CDKN1A* and *CDKN2B*, known gene targets of MYC. In a follow-up study by the same research group, three additional lncRNAs were identified, named MYCLo-4, MYCLo5, and MYCLo6, and were repressed by MYC. Similar to MYCLos1-3, MYCLos4-6 were also found to influence cell proliferation and cell cycle progression, by regulating the expression of MYC target genes [66]. Collectively, MYCLos are a newly identified class of MYC-regulated lncRNAs, with some of them having an oncogenic role (MYCLos 1-3) and others having a tumor suppressor role (MYCLos 4-6). In the future, it will be important to determine how universal these lncRNAs are to the functions of MYC and whether a similar regulation of MYCLos by MYC is observed in other cancers.

The PCAT family

The prostate cancer associated transcripts (PCATs) are another class of lncRNAs within the lncRNA-MYC network. Three of the better-characterized PCAT lnRNAs are PCAT1, PCAT8 (also known as PRNCR1 and CARLo-3) and PCAT9 (also known as PCGEM1). While *PCAT9* is located on chromosome 2, *PCAT1* and *PCAT8* are located ~715 kb and ~645 kb upstream of *MYC*,

respectively. As mentioned above, lncRNAs have been shown to regulate the transcription of the *MYC* gene and the stability of the MYC protein. More recent evidence suggests that lncRNAs may also influence MYC protein expression at the mRNA level. In a recent study in prostate cancer cells, it was shown that PCAT1 attenuates the downregulation of MYC protein expression (but not mRNA amount or stability) by interfering with miR-34a [67], a known miRNA that regulates MYC expression by targeting the MYC mRNA 3'UTR [68-70]. Although many lncRNAs act as sponges to sequester miRNAs away from their mRNA targets [71, 72], the investigators were unable to identify any putative miR-34a binding site in PCAT1. Therefore, it was suggested that PCAT1 indirectly affects miR-34a post-transcriptional regulation of MYC [67]. While PCAT1 does appear to be directly involved in the lncRNA-MYC network, it is unclear if PCAT8 is also part of this network; however, PCAT8 has been associated with both prostate and colorectal cancers [73, 74].

In a study by Hung etl al., PCAT9, also known as prostate cancer gene expression marker 1 (PCGEM1), was found to be an important transcriptional mediator of many metabolic pathways in prostate cancer cells [75]. With chromatin isolation by RNA purification (ChIRP), a technique developed to examine specific RNA-DNA interactions [76], it was demonstrated that PCAT9 physically interacts with the promoters of metabolic genes, and that PCAT9 expression affected cell-cycle progression and proliferation [75]. Also discovered, PCAT9 was found to bind to MYC and that upon knockdown of

PCAT9 recruitment of MYC to metabolic genes was diminished [75]. To date, PCAT9 is the only lncRNA that been shown to bind to MYC and promote its transactivation activity thereby affecting the metabolism of cancer cells.

GAS5

The growth arrest-specific 5 (GAS5) lncRNA is a functionally diverse lncRNA [77], that is transcribed from chromosome 1. GAS5 has been suggested to be a tumor suppressor, implicated in several human cancers [78-82]. Similar to PCAT1, GAS5 has been suggested to affect MYC expression at the mRNA level. In a recent study, GAS5 was shown to bind to both the eIF4E translation initiation factor and the MYC mRNA thereby inhibiting translation of MYC [82]. However, further investigation is needed to determine mechanistically how GAS5 is suppressing the translation of MYC mRNA. This study provides another example of the diversity of mechanisms by which lncRNAs regulate the expression of MYC.

GHET1

Gastric carcinoma proliferation enhancing transcript 1 (GHET1) is an unspliced lncRNA transcribed from chromosome 7 that has been implicated in gastric and bladder cancers [83, 84]. First discovered by Yang et al., GHET1 was found to be upregulated in gastric carcinoma clinical samples and higher levels of GHET1 expression correlated with a poor survival rate [84]. Knockdown of GHET1 was shown to inhibit proliferation rates of gastric carcinoma cells. Conversely, overexpression of GHET1 promoted cell proliferation rates *in vitro* and tumor

growth *in vivo*. With the use of different immunoprecipitation techniques, GHET1 was shown to physically interact with insulin-like growth factor 2 mRNA binding protein 1 (IGF2BP1) and also promote the binding of IGF2BP1 to MYC mRNA aiding in its stabilization [84]. The MYC mRNA is a very unstable mRNA that is rapidly degraded, and IGF2BP1 is part of a protein complex that has been shown to promote its stability [85, 86]. In the future, it would be interesting to see if GHET1 maintains this same mechanistic relationship with MYC in other malignancies.

H19

Imprinted maternally expressed transcript, known as H19, is a lncRNA expressed only from the maternal allele on chromosome 11 that has been to shown to be essential for human tumor growth and metastasis [87, 88]. Moreover, H19 has been demonstrated to be functionally important in several human cancers [89-94]. While it has been known for many years that H19 is a key player in many human malignancies, it was only recent that a functional link between H19 and MYC had been discovered. MYC was found to bind to E-boxes located in the *H19* promoter and assist in histone acetylation, thereby promoting H19 expression [94]. Many of these findings were recapitulated in a later study [93]. Interestingly, H19 is predominantly a cytoplasmic lncRNA, and recently has been demonstrated to be important in muscle differentiation by acting as a molecular sponge for the let-7 miRNA [96]. Furthermore, the role of H19 in metastasis was elucidated later in ovarian cancer cells were H19 was discovered to interfere with

let-7 mediated downregulation of MYC mRNA and protein levels [97]. Collectively, H19 is one of the most pervasive dysregulated lncRNAs seen in human cancer, and to date it is one of only a few lncRNAs that feeds into a positive feedback loop with MYC, by being transcriptionally upregulated by MYC and post-transcriptionally disinhibiting MYC mRNA degradation.

TUSC8

A relatively uncharacterized lncRNA, referred to as tumor suppressor candidate 8 (TUSC8) located on chromosome 13 has also be suggested to modulate the expression of MYC. In a study by Liao et al., TUSC8 was found to be downregulated in cervical cancer, and TUSC8 expression was found to correlate with the progression of the cervical cancer and patient survival rate. In HeLa, SiHA and HCC94 cells, overexpression of TUSC8 was discovered to diminish both MYC mRNA and protein levels and decrease proliferation rates, while knockdown of TUSC8 had an opposite effect in both MYC expression and proliferation rates [98]. However, the mechanisms of how TUSC8 regulates the expression of MYC is unclear and these observed effects on MYC expression could potentially be indirect.

Conclusion

Our understanding of the dynamic regulatory relationship between lncRNAs and MYC remains in its infancy. However, just within the past year there have been several studies exploring this potentially invaluable relationship found within

many human cancers. It is not surprising that MYC would transcriptionally regulate many lncRNAs, and it is especially interesting that MYC oncogenic functions could be mediated through the regulation of specific lncRNAs. Given the crucial role of MYC in many cancers, these findings suggest that MYC-regulated lncRNAs and also lncRNAs that regulate MYC could be potential valuable targets in the treatment of many human cancers. MYCN is another interesting protein of the lncRNA-MYC network that is garnering attention. New studies have been conducted exploring a functional connection between MYCN and lncRNAs implicated in cancer [99-103]. Given its importance in the nervous system and mesenchymal tissues [104, 105], like MYC, MYCN could also mediate some of its oncogenic functions through the regulation of lncRNAs. Currently, we are still left with many unanswered questions concerning the importance of the lncRNA-MYC regulatory network in the development and/or maintenance of cancer. Specifically, it would be interesting to know how pervasive these regulatory networks are and whether the same or distinct molecular interactions exist in different malignancies. Given the sheer number of different lncRNA genes/loci, which give rise to an even larger number of lncRNA transcripts, and the fact that many of these lncRNAs are expressed both in a temporal and tissue-specific manner [14, 21, 22], one could postulate the existence of many more lncRNAs that could be regulated by MYC in a context-dependent manner. Altogether, future investigations in understanding this

complex regulatory network could serve to provide critical insights in the biology underlying the many different types of cancers.

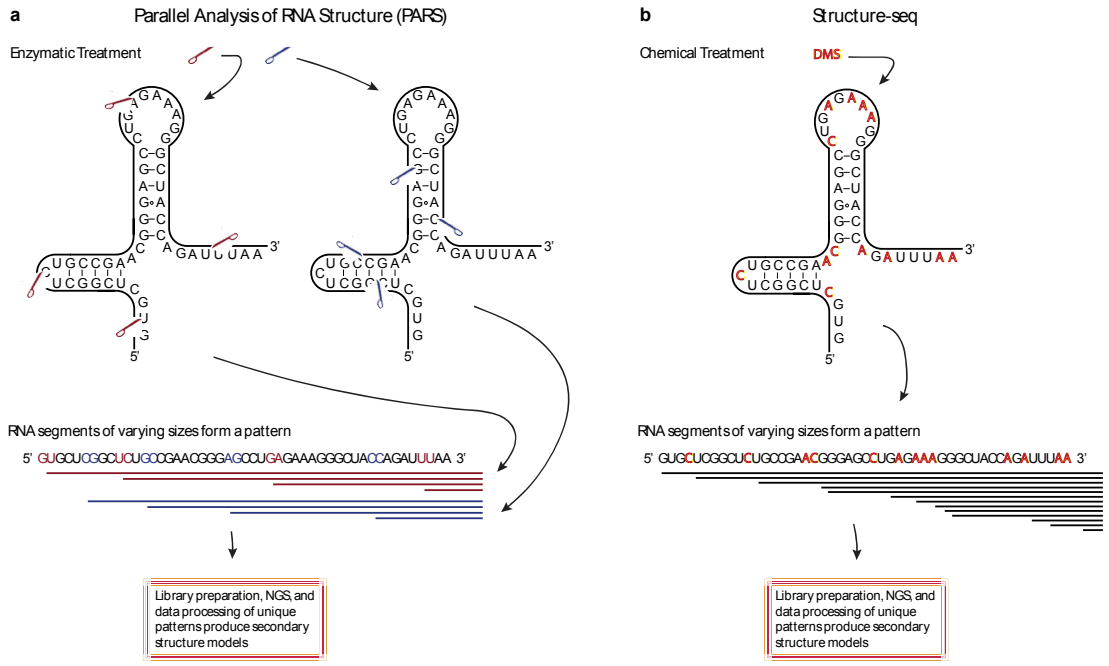


Figure 4.1. Overview of two methodologies used to determine RNA secondary structure. A- Parallel analysis of RNA structure (PARS) uses an *in vitro* enzymatic treatment with single strand (S1 nuclease, red scissor) and double strand (RNase V1, blue scissor) cutters to generate two pools of digested RNA. Once digested, adaptor sequences are ligated to the cleavage sites, converted into a cDNA library and subject to next-generation sequencing (NGS). Cleavages sites, identified from the sequencing data, will provide the locations of double stranded RNA regions (seen from the RNase V1 cleavage sites) or single stranded regions (seen from the S1 nuclease cleavage sites). Collectively, from these data secondary structure of RNA molecules can be determined. **B-** An *in vivo* chemical treatment, named Structure-seq, uses DMS to selectively methylate available adenines and cytosines (denoted by red letters). Reverse

transcriptase activity stops one nucleotide before reaching the methylated adenine or cytosine. A cDNA library is constructed and subject to NGS. As a result, the signature of discernable stop sites can be used to infer secondary structure from NGS data.

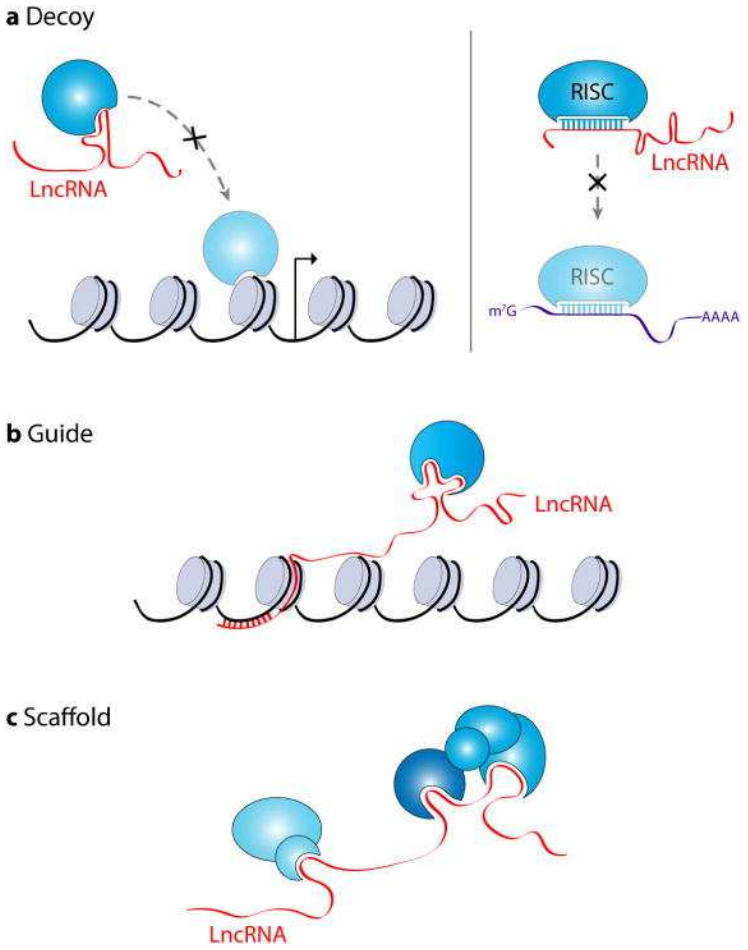


Figure 4.2. Functional categories of lncRNAs. There are three broad functional classifications for lncRNAs. **A-** LncRNAs can act as decoys sequestering either proteins and/or regulatory RNAs, such as miRNAs, away from their targets or cellular locations. **B-** LncRNAs can also be key players in the recruitment of proteins, such chromatin-modifying enzymes, to specific genomic locations thereby influencing transcriptional events. **C-** LncRNAs can provide a platform or scaffold to facilitate different molecular interactions, such as protein-protein interactions.

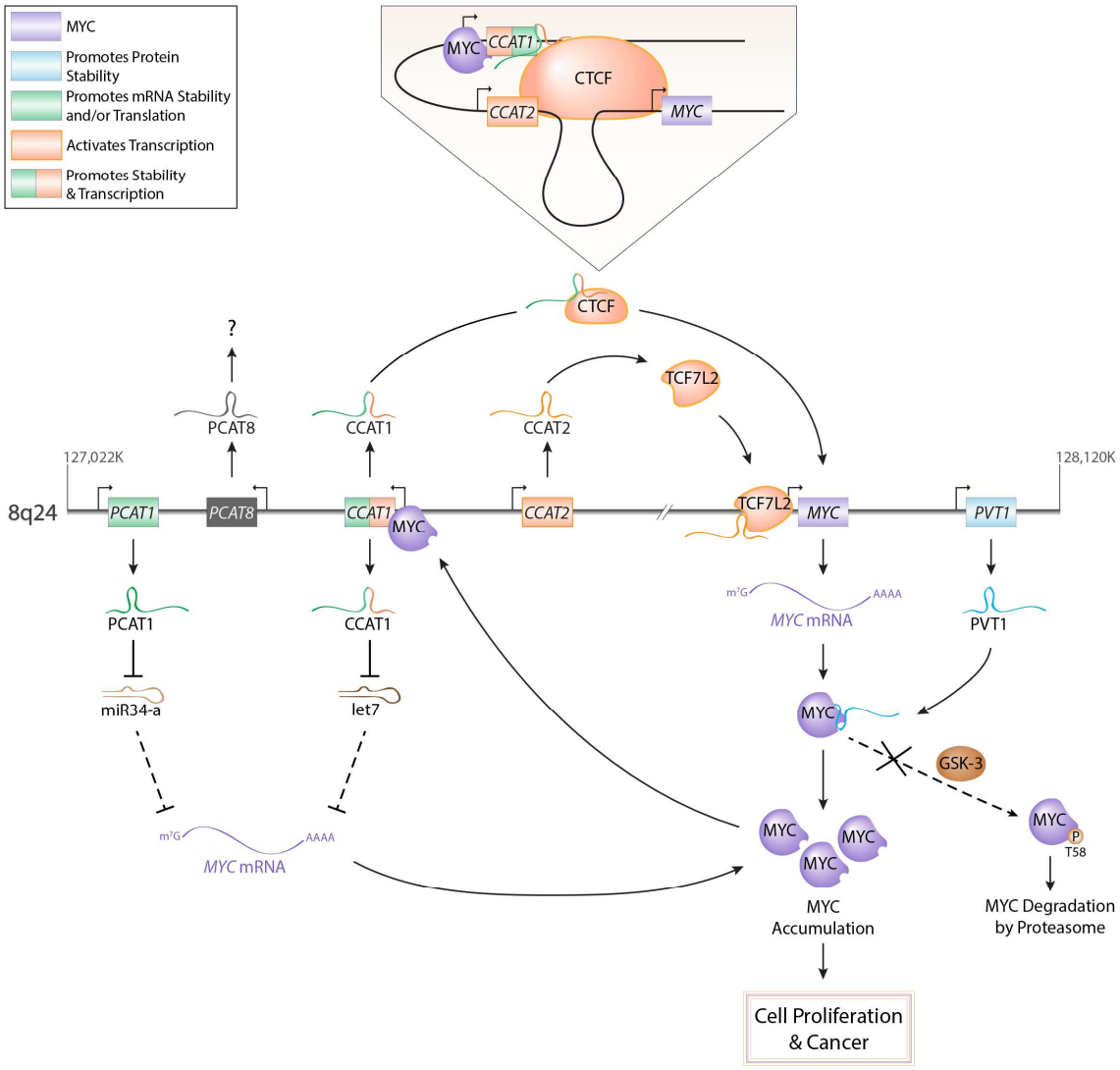


Figure 4.3. Molecular interactions of the lncRNA-MYC network at the 8q24 genomic region. CCAT1 both transcriptionally (through chromatin interactions) and post-transcriptionally (through titration of let-7) regulates *MYC* expression. CCAT2 recruits TCFL2 to the *MYC* promoter aiding in transcriptional activation. PCAT1 prevents miR34-a mediated translational repression. PVT1 binds to MYC preventing threonine-58 phosphorylation by glycogen synthase kinase 3 (GSK3) and subsequent MYC degradation. Collectively, all of the flanking

lncRNAs promote the accumulation of MYC; therefore, when these lncRNA are inappropriately upregulated, MYC-dependent malignancies can develop.

Table 4.1. Summary of the participants of the lncRNA-MYC network.

LncRNA	Aliases	Functional category	Chromosome	Regulated by MYC	Regulates *MYC*	Classification	Function	References
PVT1	LINC00079, onco-lncRNA-100	Decoy	8	Upregulated	Yes	Oncogene	Binds MYC preventing phosphorylation and degradation	[45],[47],[48]
CCAT Family								
CCAT1	CARLo-5,onco-lncRNA-40	Guide, scaffold, decoy	8	Upregulated	Yes	Oncogene	Facilitates transcription of *MYC*, titrates let-7 away preventing *MYC* degradation	[52], [54],[58-61], [63-65]
CCAT2	NCCP1, LINC00873	Guide	8	Unknown	Yes	Oncogene	Facilitates transcription of *MYC* via the recruitment of TCFL2	[62]
MYCLo-1	AK021907	Unknown	20	Upregulated	Unknown	Oncogene	-	[63]
MYCLo-2	CCAT6, AC074389.9	Unknown	7	Upregulated	Unknown	Oncogene	-	[63]
MYCLo-3	KTN1-AS1 , C14orf33	Unknown	14	Upregulated	Unknown	Oncogene	-	[63]
MYCLo-4	JX046912	Unknown	9	Downregulated	Unknown	Tumor suppressor	-	[66]
MYCLo-5	JX046913, JX046914	Unknown	3	Downregulated	Unknown	Tumor suppressor	-	[66]
MYCLo-6	JX046915	Unknown	3	Downregulated	Unknown	Tumor suppressor	-	[66]
PCAT Family								
PCAT1	PCA1	Unknown	8	Unknown	Yes*	Oncogene	Interferes with miR34-a preventing MYC translational repression	[67]
PCAT9	PCGEM1, LINC00071	Guide	2	Unknown	Yes	Oncogene	Increases MYC transactivation activity	[75]
GAS5	SNHG2	Unknown	1	Unknown	Yes	Tumor suppressor	Prevents translation of *MYC* mRNA	[82]
GHET1	-	Guide, scaffold	7	Unknown	Yes	Oncogene	Promotes stability of *MYC* mRNA via the recruitment of IGF2BP1	[84]
H19	ASM, BWS, WT2, ASM1, D11S813E, LINC00008	Decoy	11	Upregulated	Yes	Tumor suppressor/ oncogene	Titrates let-7 away preventing MYC mRNA degradation	[93], [94], [97]
TUSC8	XLOC_010588, LINC01071	Unknown	13	Unknown	Yes*	Tumor suppressor	-	[98]

*Unclear whether effect is direct or indirect

References

1. Volders PJ, Verheggen K, Menschaert G, et al. (2015) An update on LNCipedia: a database for annotated human lncRNA sequences. *Nucleic Acids Res* 43: D174-180.

2. Amati B, Frank SR, Donjerkovic D, et al. (2001) Function of the c-Myc oncoprotein in chromatin remodeling and transcription. *Biochim Biophys Acta* 1471: M135-145.

3. Bretones G, Delgado MD, Leon J (2015) Myc and cell cycle control. *Biochim Biophys Acta* 1849: 506-516.

4. Dang CV (2013) MYC, metabolism, cell growth, and tumorigenesis. *Cold Spring Harb Perspect Med* 3.

5. McMahon SB (2014) MYC and the control of apoptosis. *Cold Spring Harb Perspect Med* 4: a014407.

6. Vennstrom B, Sheiness D, Zabielski J, et al. (1982) Isolation and characterization of c-myc, a cellular homolog of the oncogene (v-myc) of avian myelocytomatosis virus strain 29. *J Virol* 42: 773-779.

7. Zheng GX, Do BT, Webster DE, et al. (2014) Dicer-microRNA-Myc circuit promotes transcription of hundreds of long noncoding RNAs. *Nat Struct Mol Biol* 21: 585-590.

8. Winkle M, van den Berg A, Tayari M, et al. (2015) Long noncoding RNAs as a novel component of the Myc transcriptional network. *FASEB J* 29: 2338-2346.

9. Xiang JF, Yang L, Chen LL (2015) The long noncoding RNA regulation at the MYC locus. *Curr Opin Genet Dev* 33: 41-48.

10. Rinn JL, Chang HY (2012) Genome regulation by long noncoding RNAs. *Annu Rev Biochem* 81: 145-166.

11. Kapranov P, Cawley SE, Drenkow J, et al. (2002) Large-scale transcriptional activity in chromosomes 21 and 22. *Science* 296: 916-919.

12. Rinn JL, Euskirchen G, Bertone P, et al. (2003) The transcriptional activity of human Chromosome 22. *Genes Dev* 17: 529-540.

13. Guttman M, Garber M, Levin JZ, et al. (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* 28: 503-510.

14. Cabili MN, Trapnell C, Goff L, et al. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* 25: 1915-1927.

15. Wu J, Okada T, Fukushima T, et al. (2012) A novel hypoxic stress-responsive long non-coding RNA transcribed by RNA polymerase III in Arabidopsis. *RNA Biol* 9: 302-313.

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

17. Gardini A, Shiekhattar R (2015) The many faces of long noncoding RNAs. *FEBS J* 282: 1647-1657.

18. Wilusz JE, Spector DL (2010) An unexpected ending: noncanonical 3' end processing mechanisms. *RNA* 16: 259-266.

19. Zhang Y, Yang L, Chen LL (2014) Life without A tail: new formats of long noncoding RNAs. *Int J Biochem Cell Biol* 54: 338-349.

20. Peart N, Sataluri A, Baillat D, et al. (2013) Non-mRNA 3' end formation: how the other half lives. *Wiley Interdiscip Rev RNA* 4: 491-506.

21. Ravasi T, Suzuki H, Pang KC, et al. (2006) Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. *Genome Res* 16: 11-19.

22. Djebali S, Davis CA, Merkel A, et al. (2012) Landscape of transcription in human cells. *Nature* 489: 101-108.

23. He S, Liu S, Zhu H (2011) The sequence, structure and evolutionary features of HOTAIR in mammals. *BMC Evol Biol* 11: 102.

24. Brown JA, Bulkley D, Wang J, et al. (2014) Structural insights into the stabilization of MALAT1 noncoding RNA by a bipartite triple helix. *Nat Struct Mol Biol* 21: 633-640.

25. Brown JA, Valenstein ML, Yario TA, et al. (2012) Formation of triple-helical structures by the 3'-end sequences of MALAT1 and MENbeta noncoding RNAs. *Proc Natl Acad Sci U S A* 109: 19202-19207.

26. Smith MA, Gesell T, Stadler PF, et al. (2013) Widespread purifying selection on RNA structure in mammals. *Nucleic Acids Res* 41: 8220-8236.

27. Somarowthu S, Legiewicz M, Chillon I, et al. (2015) HOTAIR forms an intricate and modular secondary structure. *Mol Cell* 58: 353-361.

28. Mortimer SA, Kidwell MA, Doudna JA (2014) Insights into RNA structure and function from genome-wide studies. *Nat Rev Genet* 15: 469-479.

29. Ding Y, Tang Y, Kwok CK, et al. (2014) In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* 505: 696-700.

30. Kertesz M, Wan Y, Mazor E, et al. (2010) Genome-wide measurement of RNA secondary structure in yeast. *Nature* 467: 103-107.

31. Lucks JB, Mortimer SA, Trapnell C, et al. (2011) Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proc Natl Acad Sci U S A* 108: 11063-11068.

32. Seetin MG, Kladwang W, Bida JP, et al. (2014) Massively parallel RNA chemical mapping with a reduced bias MAP-seq protocol. *Methods Mol Biol* 1086: 95-117.

33. Underwood JG, Uzilov AV, Katzman S, et al. (2010) FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat Methods* 7: 995-1001.

34. Aviran S, Trapnell C, Lucks JB, et al. (2011) Modeling and automation of sequencing-based characterization of RNA structure. *Proc Natl Acad Sci U S A* 108: 11069-11074.

35. Rouskin S, Zubradt M, Washietl S, et al. (2014) Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* 505: 701-705.

36. Novikova IV, Hennelly SP, Sanbonmatsu KY (2013) Tackling structures of long noncoding RNAs. *Int J Mol Sci* 14: 23672-23684.

37. Fatica A, Bozzoni I (2014) Long non-coding RNAs: new players in cell differentiation and development. *Nat Rev Genet* 15: 7-21.

38. Batista PJ, Chang HY (2013) Long noncoding RNAs: cellular address codes in development and disease. *Cell* 152: 1298-1307.

39. Batista PJ, Chang HY (2013) Cytotopic localization by long noncoding RNAs. *Curr Opin Cell Biol* 25: 195-199.

40. Han X, Yang F, Cao H, et al. (2015) Malat1 regulates serum response factor through miR-133 as a competing endogenous RNA in myogenesis. *FASEB J* 29: 3054-3064.

41. Yoon JH, Abdelmohsen K, Kim J, et al. (2013) Scaffold function of long non-coding RNA HOTAIR in protein ubiquitination. *Nat Commun* 4: 2939.

42. McHugh CA, Chen CK, Chow A, et al. (2015) The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature* 521: 232-236.

43. Deng K, Guo X, Wang H, et al. (2014) The lncRNA-MYC regulatory network in cancer. *Tumour Biol* 35: 9497-9503.

44. Beroukhim R, Mermel CH, Porter D, et al. (2010) The landscape of somatic copy-number alteration across human cancers. *Nature* 463: 899-905.

45. Colombo T, Farina L, Macino G, et al. (2015) PVT1: a rising star among oncogenic long noncoding RNAs. *Biomed Res Int* 2015: 304208.

46. Walz S, Lorenzin F, Morton J, et al. (2014) Activation and repression by oncogenic MYC shape tumour-specific gene expression profiles. *Nature* 511: 483-487.

47. Carramusa L, Contino F, Ferro A, et al. (2007) The PVT-1 oncogene is a Myc protein target that is overexpressed in transformed cells. *J Cell Physiol* 213: 511-518.

48. Tseng YY, Moriarity BS, Gong W, et al. (2014) PVT1 dependence in cancer with MYC copy-number increase. *Nature* 512: 82-86.

49. Zanke BW, Greenwood CM, Rangrej J, et al. (2007) Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat Genet* 39: 989-994.

50. Tenesa A, Farrington SM, Prendergast JG, et al. (2008) Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nat Genet* 40: 631-637.

51. Tomlinson I, Webb E, Carvajal-Carmona L, et al. (2007) A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat Genet* 39: 984-988.

52. Nissan A, Stojadinovic A, Mitrani-Rosenbaum S, et al. (2012) Colon cancer associated transcript-1: a novel RNA expressed in malignant and pre-malignant human tissues. *Int J Cancer* 130: 1598-1606.

53. Ahmadiyeh N, Pomerantz MM, Grisanzio C, et al. (2010) 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC. *Proc Natl Acad Sci U S A* 107: 9742-9746.

54. Kim T, Cui R, Jeon YJ, et al. (2014) Long-range interaction and correlation between MYC enhancer and oncogenic long noncoding RNA CARLo-5. *Proc Natl Acad Sci U S A* 111: 4173-4178.

55. Pomerantz MM, Ahmadiyeh N, Jia L, et al. (2009) The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat Genet* 41: 882-884.

56. Sur IK, Hallikas O, Vaharautio A, et al. (2012) Mice lacking a Myc enhancer that includes human SNP rs6983267 are resistant to intestinal tumors. *Science* 338: 1360-1363.

57. Tuupanen S, Turunen M, Lehtonen R, et al. (2009) The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nat Genet* 41: 885-890.

58. Xiang JF, Yin QF, Chen T, et al. (2014) Human colorectal cancer-specific CCAT1-L lncRNA regulates long-range chromatin interactions at the MYC locus. *Cell Res* 24: 513-531.

59. Deng L, Yang SB, Xu FF, et al. (2015) Long noncoding RNA CCAT1 promotes hepatocellular carcinoma progression by functioning as let-7 sponge. *J Exp Clin Cancer Res* 34: 18.

60. He X, Tan X, Wang X, et al. (2014) C-Myc-activated long noncoding RNA CCAT1 promotes colon cancer cell proliferation and invasion. *Tumour Biol* 35: 12181-12188.

61. Yang F, Xue X, Bi J, et al. (2013) Long noncoding RNA CCAT1, which could be activated by c-Myc, promotes the progression of gastric carcinoma. *J Cancer Res Clin Oncol* 139: 437-445.

62. Ling H, Spizzo R, Atlasi Y, et al. (2013) CCAT2, a novel noncoding RNA mapping to 8q24, underlies metastatic progression and chromosomal instability in colon cancer. *Genome Res* 23: 1446-1461.

63. Kim T, Jeon YJ, Cui R, et al. (2015) Role of MYC-regulated long noncoding RNAs in cell cycle regulation and tumorigenesis. *J Natl Cancer Inst* 107.

64. Ye Z, Zhou M, Tian B, et al. (2015) Expression of lncRNA-CCAT1, E-cadherin and N-cadherin in colorectal cancer and its clinical significance. *Int J Clin Exp Med* 8: 3707-3715.

65. Alaiyan B, Ilyayev N, Stojadinovic A, et al. (2013) Differential expression of colon cancer associated transcript1 (CCAT1) along the colonic adenoma-carcinoma sequence. *BMC Cancer* 13: 196.

66. Kim T, Cui R, Jeon YJ, et al. (2015) MYC-repressed long noncoding RNAs antagonize MYC-induced cell proliferation and cell cycle progression. *Oncotarget*.

67. Prensner JR, Chen W, Han S, et al. (2014) The long non-coding RNA PCAT-1 promotes prostate cancer cell proliferation through cMyc. *Neoplasia* 16: 900-908.

68. Yamamura S, Saini S, Majid S, et al. (2012) MicroRNA-34a modulates c-Myc transcriptional complexes to suppress malignancy in human prostate cancer cells. *PLoS One* 7: e29722.

69. Siemens H, Jackstadt R, Hunten S, et al. (2011) miR-34 and SNAIL form a double-negative feedback loop to regulate epithelial-mesenchymal transitions. *Cell Cycle* 10: 4256-4271.

70. Benassi B, Flavin R, Marchionni L, et al. (2012) MYC is activated by USP2a-mediated modulation of microRNAs in prostate cancer. *Cancer Discov* 2: 236-247.

71. Poliseno L, Salmena L, Zhang J, et al. (2010) A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* 465: 1033-1038.

72. Salmena L, Poliseno L, Tay Y, et al. (2011) A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* 146: 353-358.

73. Ge X, Chen Y, Liao X, et al. (2013) Overexpression of long noncoding RNA PCAT-1 is a novel biomarker of poor prognosis in patients with colorectal cancer. *Med Oncol* 30: 588.

74. Chung S, Nakagawa H, Uemura M, et al. (2011) Association of a novel long non-coding RNA in 8q24 with prostate cancer susceptibility. *Cancer Sci* 102: 245-252.

75. Hung CL, Wang LY, Yu YL, et al. (2014) A long noncoding RNA connects c-Myc to tumor metabolism. *Proc Natl Acad Sci U S A* 111: 18697-18702.

76. Chu C, Qu K, Zhong FL, et al. (2011) Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. *Mol Cell* 44: 667-678.

77. Pickard MR, Williams GT (2015) Molecular and Cellular Mechanisms of Action of Tumour Suppressor GAS5 LncRNA. *Genes (Basel)* 6: 484-499.

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

79. Mourtada-Maarabouni M, Pickard MR, Hedge VL, et al. (2009) GAS5, a non-protein-coding RNA, controls apoptosis and is downregulated in breast cancer. *Oncogene* 28: 195-208.

80. Pickard MR, Mourtada-Maarabouni M, Williams GT (2013) Long non-coding RNA GAS5 regulates apoptosis in prostate cancer cell lines. *Biochim Biophys Acta* 1832: 1613-1623.

81. Sun M, Jin FY, Xia R, et al. (2014) Decreased expression of long noncoding RNA GAS5 indicates a poor prognosis and promotes cell proliferation in gastric cancer. *BMC Cancer* 14: 319.

82. Tu ZQ, Li RJ, Mei JZ, et al. (2014) Down-regulation of long non-coding RNA GAS5 is associated with the prognosis of hepatocellular carcinoma. *Int J Clin Exp Pathol* 7: 4303-4309.

83. Li LJ, Zhu JL, Bao WS, et al. (2014) Long noncoding RNA GHET1 promotes the development of bladder cancer. *Int J Clin Exp Pathol* 7: 7196-7205.

84. Yang F, Xue X, Zheng L, et al. (2014) Long non-coding RNA GHET1 promotes gastric carcinoma cell proliferation by increasing c-Myc mRNA stability. *FEBS J* 281: 802-813.

85. Lemm I, Ross J (2002) Regulation of c-myc mRNA decay by translational pausing in a coding region instability determinant. *Mol Cell Biol* 22: 3959-3969.

86. Weidensdorfer D, Stohr N, Baude A, et al. (2009) Control of c-myc mRNA stability by IGF2BP1-associated cytoplasmic RNPs. *RNA* 15: 104-115.

87. Matouk IJ, DeGroot N, Mezan S, et al. (2007) The H19 non-coding RNA is essential for human tumor growth. *PLoS One* 2: e845.

88. Matouk IJ, Raveh E, Abu-lail R, et al. (2014) Oncofetal H19 RNA promotes tumor metastasis. *Biochim Biophys Acta* 1843: 1414-1426.

89. Jiang X, Yan Y, Hu M, et al. (2015) Increased level of H19 long noncoding RNA promotes invasion, angiogenesis, and stemness of glioblastoma cells. *J Neurosurg*: 1-8.

90. Lottin S, Adriaenssens E, Dupressoir T, et al. (2002) Overexpression of an ectopic H19 gene enhances the tumorigenic properties of breast cancer cells. *Carcinogenesis* 23: 1885-1895.

91. Luo M, Li Z, Wang W, et al. (2013) Long non-coding RNA H19 increases bladder cancer metastasis by associating with EZH2 and inhibiting E-cadherin expression. *Cancer Lett* 333: 213-221.

92. Ma C, Nong K, Zhu H, et al. (2014) H19 promotes pancreatic cancer metastasis by derepressing let-7's suppression on its target HMGA2-mediated EMT. *Tumour Biol* 35: 9163-9169.

93. Zhang EB, Han L, Yin DD, et al. (2014) c-Myc-induced, long, noncoding H19 affects cell proliferation and predicts a poor prognosis in patients with gastric cancer. *Med Oncol* 31: 914.

94. Barsyte-Lovejoy D, Lau SK, Boutros PC, et al. (2006) The c-Myc oncogene directly induces the H19 noncoding RNA by allele-specific binding to potentiate tumorigenesis. *Cancer Res* 66: 5330-5337.

95. Shi Y, Wang Y, Luan W, et al. (2014) Long non-coding RNA H19 promotes glioma cell invasion by deriving miR-675. *PLoS One* 9: e86295.

96. Kallen AN, Zhou XB, Xu J, et al. (2013) The imprinted H19 lncRNA antagonizes let-7 microRNAs. *Mol Cell* 52: 101-112.

97. Yan L, Zhou J, Gao Y, et al. (2015) Regulation of tumor cell migration and invasion by the H19/let-7 axis is antagonized by metformin-induced DNA methylation. *Oncogene* 34: 3076-3084.

98. Liao LM, Sun XY, Liu AW, et al. (2014) Low expression of long noncoding XLOC_010588 indicates a poor prognosis and promotes proliferation through upregulation of c-Myc in cervical cancer. *Gynecol Oncol* 133: 616-623.

99. Mestdagh P, Fredlund E, Pattyn F, et al. (2010) An integrative genomics screen uncovers ncRNA T-UCR functions in neuroblastoma tumours. *Oncogene* 29: 3583-3592.

100. Atmadibrata B, Liu PY, Sokolowski N, et al. (2014) The novel long noncoding RNA linc00467 promotes cell survival but is down-regulated by N-Myc. *PLoS One* 9: e88112.

101. Tee AE, Ling D, Nelson C, et al. (2014) The histone demethylase JMJD1A induces cell migration and invasion by up-regulating the expression of the long noncoding RNA MALAT1. *Oncotarget* 5: 1793-1804.

102. Liu PY, Erriquez D, Marshall GM, et al. (2014) Effects of a novel long noncoding RNA, lncUSMycN, on N-Myc expression and neuroblastoma progression. *J Natl Cancer Inst* 106.

103. Vadie N, Saayman S, Lenox A, et al. (2015) MYCNOS functions as an antisense RNA regulating MYCN. *RNA Biol* 12: 893-899.

104. Stanton BR, Perkins AS, Tessarollo L, et al. (1992) Loss of N-myc function results in embryonic lethality and failure of the epithelial component of the embryo to develop. *Genes Dev* 6: 2235-2247.

105. Stanton BR, Parada LF (1992) The N-myc proto-oncogene: developmental expression and in vivo site-directed mutagenesis. *Brain Pathol* 2: 71-83.

Chapter 5

The transcriptomic effects of MYC acetylation in Rat1a cells

Abstract

The MYC oncoprotein has a long-standing and extensively explored history in the cancer biology field. It is one of few proteins that are almost universally deregulated in human cancers. One area of gaining interest in cancer studies is examining the functional role(s) of post-translational modifications of highly influential proteins, such as MYC and p53. Acetylation is one such example of a post-translational modification largely overlooked in MYC studies. While widely known for its role in recruiting histone acetyltransferases to stimulate transcription, recent studies have also demonstrated MYC to be a substrate for histone acetyltransferases. These preliminary investigations have lead to interest in the role MYC acetylation has on its transcriptional activity. In current study, we sought characterize how different MYC acetylation states alter the transcriptome. Using RNA-seq analysis, we discovered MYC acetylation state has gene and pathway-specific effects, providing the first evidence showing that MYC acetylation has the capacity to change the transcriptome of cells.

Introduction

The MYC oncoprotein is a central player in most human cancers, and the mechanisms how its transcriptional activity is regulated is a subject of intense investigation and debate. MYC is a basic helix-loop-helix transcription factor, and it is largely seen as an activator of transcription through its recruitment of histone acetyltransferases or HATs [1-3]. However, MYC-mediated

transcriptional regulation is far more complex, as MYC has multiple binding partners to modulate its transcriptional activity and MYC has also been shown to be a suppressor of transcriptional activity [4]. Furthermore, MYC has also been shown to be a direct substrate for HATs [1, 5, 6]. However, the functional role of MYC acetylation on its stability and transcriptional activity is largely unknown.

MYC binds and is acetylated by HATs, such as: lysine acteyltransferase 2A (GCN5) and 5 (TIP60), and also E1A binding protein p300/CREB binding protien (p300/CBP) [1, 5, 6]. In the preliminary studies exploring MYC acetylation, MYC acetylation was found to stabilize MYC [1, 3, 7]. However, it was discovered that p300 and CBP was found to stabilize MYC independent of its HAT activity [3]. Furthermore, MYC acetylation did not affect its binding to DNA nor its dimerization with MAX [1, 5].

In the current study, we explore the global gene expression changes that occur with altered MYC acetylation states. To achieve our goal, an RNA-seq analysis was performed comparing the transcriptomic profiles of Rat1a cell lines with different expression levels and/or acetylation states of MYC, comprising a wild-type MYC overexpression cell line and three other cell lines overexpressing mutant forms of MYC, with missense mutations converting lysines residues (K149, K158 and K323) into arginine residues. Conversion into arginine of these residues results in a loss of acetylation at these residues positions.

Materials and Methods

RNA-seq analysis (conducted by Matthew Hurd, unpublished)

Rat1a cells were trypsinized and the RNA was extracted using the RNeasy Qiagen RNA extraction kit per manufacturer's protocol. RNA quality and quantity were evaluated with a bioanalyzer and Thermoscientific Nanodrop2000 spectrophotometer. Single-end read RNA-seq libraries were constructed using the NEBNext Ultra Directional RNA library prep kit for Illumina. Samples were multiplexed and sequenced with the HiSeq2500 HT sequencing platform.

Read alignment and differential expression analysis

Alignment of the sequencing reads was performed with HISAT2 using the mouse m38 assembly acquired from the Ensembl ftp site (https://www.ensembl.org/info/data/ftp/index.html) [8]. Read counting was performed using the summarizeOverlaps package, with union mode [9]. Using the read counts, an edgeR analysis was performed using the default settings [10, 11]. The entire pipeline was performed within the systemPipeR package [12]. Gene ontology analysis was conducted with the DEGs in Metascape using default settings (http://metascape.org/gp/index.html#/main/step1) [13].

Gene set enrichment analysis

Read counts generated from the aforementioned summarizeOverlaps analysis, were normalized using DESeq2 [14]. Subsequently, all normalized counts were used in the GSEA software using default settings [15, 16]. Significant gene set were define with an FWER <0.05.

Results

Transcriptomic effects of MYC overexpression in Rat1a cells

Comparison of the Rat1a cell lines overexpressing wild-type (WT) MYC relative to the Rat1a cell with empty control vector (endogenous MYC levels) revealed a significant number of differentially expressed genes with several molecular pathways affected. A total of 916 DEGs were identified, using a >2 fold change and 7% FDR. Surprisingly, more downregulated genes than upregulated genes were observed; a total of 235 upregulated genes and 681 downregulated genes were observed with MYC overexpression. Gene ontology analysis, using Metascape, with all of the identified DEGs showed terms related to wounding, extracellular matrix organization, regulation of cellular component movement and neuron projection development to be the most significantly enriched terms (Table 5.1). Segregated GO analyses rendered similar results to the collective GO analysis with all the DEGs. GO analysis with only upregulated DEGs showed no significantly enriched terms using a q-value of <0.05 (Table 5.2). Alternatively, GO analysis with only the downregulated DEGs rendered enriched terms related to extracellular matrix organization and cellular component movement (Table 5.3).

Transcriptomic effects of overexpression of MYC mutants in Rat1a cells

Differential expression analysis comparing wild-type MYC overexpressing cells relative each of the mutant cell lines revealed both similarities and differences in the identified DEGs (Figure 5.1A, B). The K149 mutant showed the greatest

number of DEGs, with 22 upregulated and 186 downregulated genes. Among these DEGs 11 upregulated and 92 downregulated genes were unique to the K149 mutant. The K149 mutant showed the greatest similarity with the K323 mutant, sharing 93 differential expressed genes in common. Moreover, the K323 mutant had 18 upregulated and 18 downregulated not found in the other mutants. Lastly, the K158 had the highest number of upregulated genes, with 47 genes. The K158 mutant had 53 unique DEGs, and had least number of DEGs in common among the mutants.

GO analysis of the all of the discovered K149 mutant DEGs showed chemotaxis, regulation of cytosolic calcium concentration and prostate gland epithelium morphogenesis among the top enriched ontology terms (Figure 5.2A). GSEA using the Hallmarks gene sets showed a significant enrichment in 6 gene sets (Figure 5.2B, FWER <0.05). The MYC target gene gene sets were among the most enriched, exhibiting significant increases in expression in the K149 mutant. Gene sets related to the interferon response and oxidative phosphorylation were also found to be enriched in the K149 mutant. Alternatively, 2 gene set showed enrichment in the WT relative to the K149 mutant; these gene sets include hedgehog signaling and UV response.

For the K158 mutant, actin filament and filopodium assembly and arginine transport were among the most enriched ontology terms. Similarly, GSEA implicated genes related to the cytoskeleton as being enriched in the K158. Only 3 gene sets were identified as enriched. Two gene sets, mitotic spindle

assembly and WNT β-catenin signaling, were enriched in the K158 mutant. Interesting the EMT gene set was enriched in the WT relative to the K158 mutant, indicating a decreased expression of several EMT genes in the K158 mutant.

As anticipated, the GO analysis and GSEA for the K323 mutant was similar to the K149 mutant. Like the K149 mutant, GO analysis of the K323 mutant showed leukocyte migration, response the virus and wounding as the top enriched ontology terms. Similarly, GSEA showed enrichment in the interferon and TGF-β signaling gene sets. However, the K323 mutant diverged from the K149 mutant showing an enrichment of the oxidative phosphorylation gene set with the WT, while the oxidative phosphorylation gene set was enriched in the K149 mutant.

Discussion

These data provide two essential pieces of information debated among the MYC research field. First, these data highlight that enhanced MYC expression does not solely activate transcription or increase RNA steady-state levels. On the contrary, the data demonstrates that with overexpression of MYC, decreases in RNA steady-state levels are more likely. Second, these data provide the first evidence suggesting MYC acetylation is influential in its activity. The data supports shared changes in gene expression between the three MYC mutants, and it also demonstrates distinct gene and pathway-specific effects.

One of the most noteworthy findings was seen with the gene set enrichment analyses. While sharing several altered genes in common, the K149 and K323 MYC mutants showed opposing changes to genes involved in oxidative phosphorylation. These molecular changes discovered are reflected in the unique phenotypic characteristics seen in the mutant cell lines (data not shown).

Another interesting finding from data was from the K149 mutant. In the GSEA, the K149 MYC mutant was the only mutant that increased MYC target genes. While the K158 and K323 MYC mutants largely inhibited MYC regulatory function or “broke” normal MYC function, the K149 MYC mutant was the only mutant that enhanced MYC function.

It is known that changes in MYC expression will elicit changes to many biological pathways [17-19]; however, these data suggest that MYC acetylation could be serving to “tip the balance” towards one biological function to another. Altogether, it will be interesting in the future to observe how this information will serve the medical field, as new therapeutics strategies could be developed targeting MYC acetylation in effort to modulate only parts of MYC function.

Table 5.1. Gene Ontology analysis of all DEGs identified with MYC overexpression

	Gene Count	%	Log10(P)	Log10(q)
response to wounding	62	7.74	-11.67	-7.47
extracellular matrix organization	36	4.49	-10.96	-7.06
positive regulation of cellular component movement	57	7.12	-10.73	-7.01
negative regulation of cellular component movement	40	4.99	-10.19	-6.69
regulation of neuron projection development	59	7.37	-10.19	-6.69
muscle structure development	61	7.62	-8.72	-5.71
blood vessel morphogenesis	55	6.87	-8.69	-5.71
ossification	42	5.24	-8.64	-5.70
actin filament-based process	62	7.74	-8.18	-5.34
collagen biosynthetic process	14	1.75	-7.99	-5.18
negative regulation of cell differentiation	64	7.99	-7.97	-5.18
mesenchyme development	33	4.12	-7.96	-5.18
metal ion homeostasis	57	7.12	-7.37	-4.70
response to mechanical stimulus	34	4.24	-7.01	-4.40
response to growth factor	60	7.49	-6.93	-4.34
positive regulation of fibroblast proliferation	15	1.87	-6.81	-4.24
tissue morphogenesis	56	6.99	-6.78	-4.22
cell-cell adhesion	57	7.12	-6.74	-4.20
positive regulation of reactive oxygen species metabolic process	18	2.25	-6.64	-4.14
positive regulation of lipid kinase activity	10	1.25	-6.55	-4.10

Table 5.2. Gene Ontology analysis of upregulated DEGs identified with MYC overexpression

	Gene Count	%	Log10(P)	Log10(q)
negative regulation of cell cycle arrest	4	2.19	-4.36	-0.17
glutathione metabolic process	5	2.73	-3.92	-0.10
cofactor metabolic process	14	7.65	-3.77	-0.10
regeneration	10	5.46	-3.46	0.00
response to copper ion	4	2.19	-3.05	0.00
negative regulation of growth	9	4.92	-2.93	0.00
regulation of telomerase activity	4	2.19	-2.89	0.00
response to vitamin D	4	2.19	-2.80	0.00
positive regulation of bone resorption	3	1.64	-2.78	0.00
cellular response to lipid	15	8.20	-2.68	0.00
response to zinc ion	4	2.19	-2.60	0.00
cranial nerve morphogenesis	3	1.64	-2.60	0.00
leukocyte migration	9	4.92	-2.57	0.00
monocarboxylic acid metabolic process	12	6.56	-2.47	0.00
monocarboxylic acid transport	6	3.28	-2.46	0.00
acylglycerol homeostasis	3	1.64	-2.44	0.00
organic cyclic compound catabolic process	11	6.01	-2.26	0.00
regulation of release of sequestered calcium ion into cytosol	4	2.19	-2.15	0.00
response to insulin	8	4.37	-2.08	0.00

Table 5.3. Gene Ontology analysis of downregulated DEGs identified with MYC overexpression

	Gene Count	%	Log10(P)	Log10(q)
extracellular matrix organization	36	5.83	-14.29	-10.10
positive regulation of cellular component movement	50	8.09	-11.44	-7.72
response to wounding	52	8.41	-11.17	-7.57
regulation of neuron projection development	49	7.93	-9.60	-6.35
collagen biosynthetic process	14	2.27	-9.44	-6.24
supramolecular fiber organization	52	8.41	-8.87	-5.78
negative regulation of cellular component movement	32	5.18	-8.60	-5.65
cell-cell adhesion	52	8.41	-8.54	-5.65
muscle structure development	51	8.25	-8.51	-5.65
response to growth factor	54	8.74	-8.50	-5.65
positive regulation of reactive oxygen species metabolic process	18	2.91	-8.35	-5.55
positive regulation of MAPK cascade	43	6.96	-8.00	-5.26
ossification	34	5.50	-7.57	-4.91
heart development	47	7.61	-7.40	-4.79
negative regulation of cell differentiation	52	8.41	-7.28	-4.70
tissue morphogenesis	47	7.61	-6.81	-4.26
response to tumor necrosis factor	25	4.05	-6.58	-4.07
response to reactive oxygen species	27	4.37	-6.49	-4.00
positive regulation of lipid kinase activity	9	1.46	-6.45	-3.98
smooth muscle cell proliferation	20	3.24	-6.34	-3.91

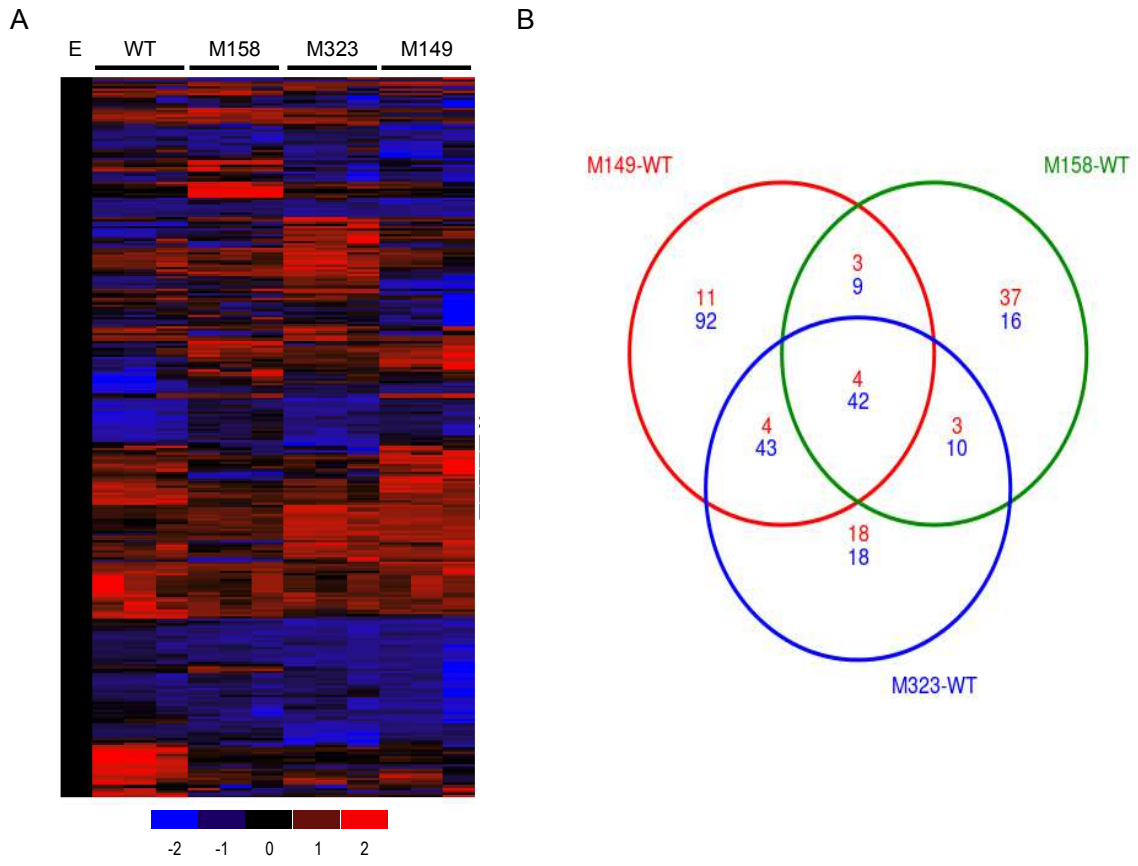


Figure 5.1. MYC acetylated mutants exhibit gene-selective effects. A– Unsupervised hierarchical clustering of differentially expressed genes identified comparing overexpression of wild-type MYC (WT) relative to MYC overexpression mutants (M149, M158, or M323), using edgeR (fold change >2 and FDR <7%). Relative expression normalized to empty control (E). **B–** Venn diagram of DEGs identifying genes in common and unique between the MYC mutants. Red – number of upregulated genes. Blue – number of downregulated genes.

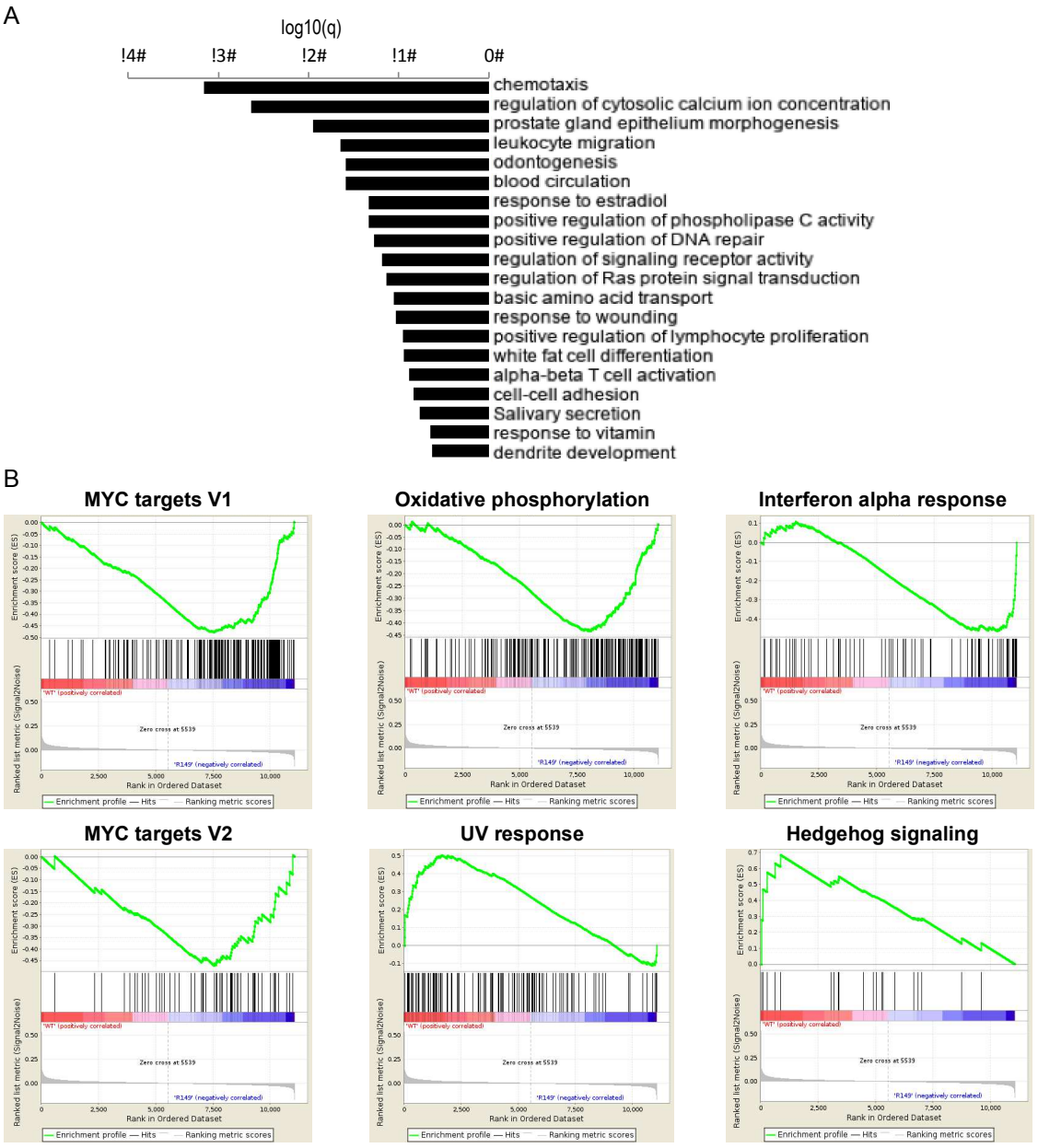


Figure 5.2. MYC-149 mutant alters chemotaxis pathway, upregulates oxidative phosphorylation and MYC-induced target genes. A – Metascape express gene ontology analysis using all 208 DEGs identified in edgeR analysis (fold >2 and FDR <7%) comparing MYC wild-type relative to the MYC-149 mutant. **B –** GSEA with Hallmarks gene sets comparing MYC wild-type relative

to the MYC-149 mutant (FWER p-value <0.05).

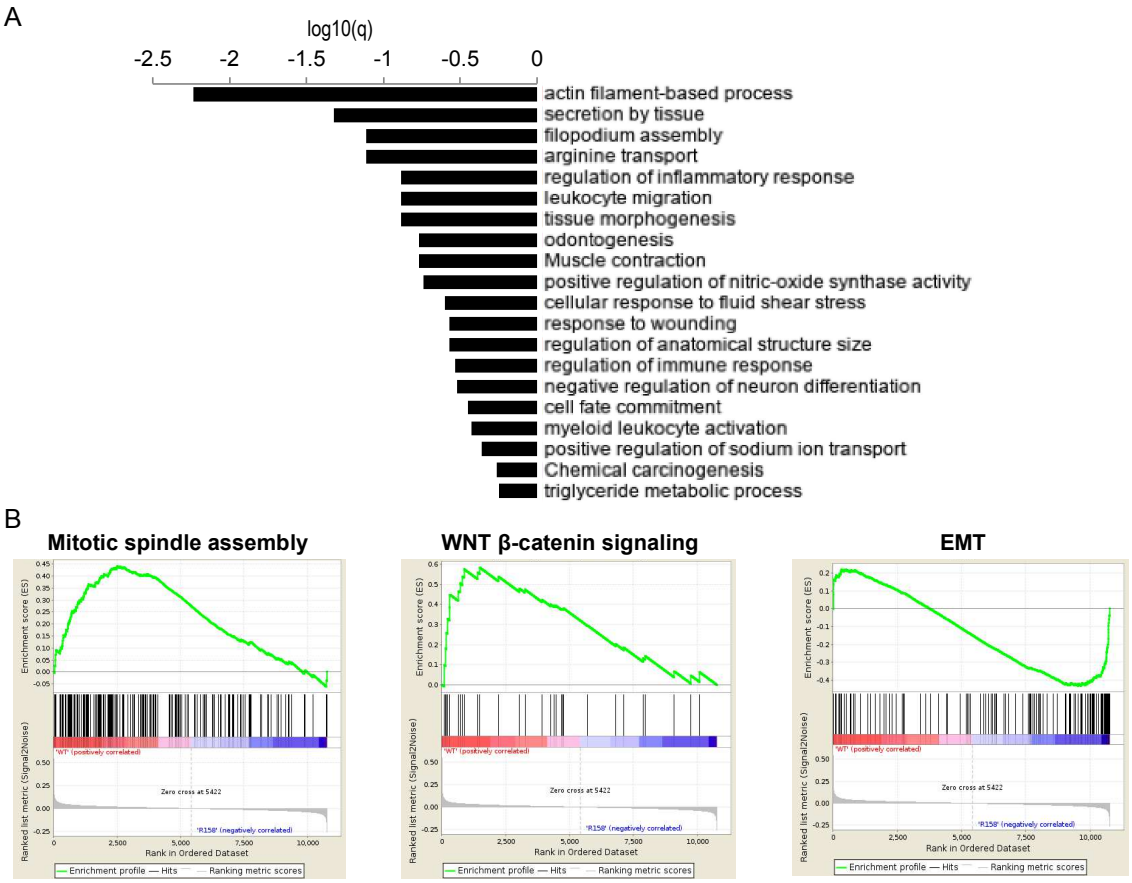


Figure 5.3. MYC-158 mutant downregulates actin/mitotic spindle assembly genes and upregulates EMT genes. **A –** Metascape express gene ontology analysis using all 124 DEGs identified in edgeR analysis (fold >2 and FDR <7%) comparing MYC wild-type relative to the MYC-149 mutant. **B –** GSEA with Hallmarks gene sets comparing MYC wild-type relative to the MYC-158 mutant (FWER p-value <0.05).

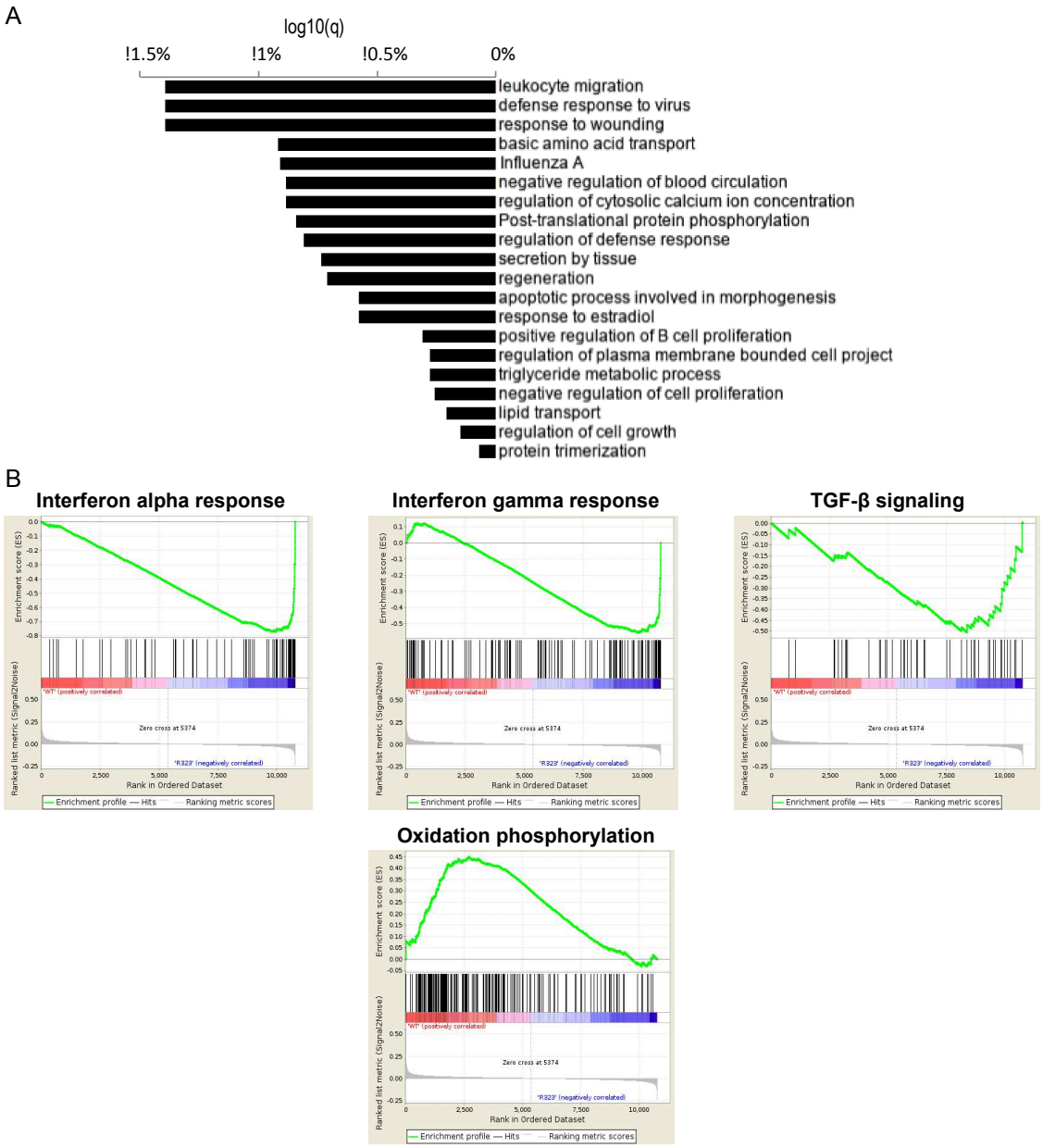


Figure 5.4. MYC-323 mutant upregulates interferon response pathways and downregulates oxidative phosphorylation genes. **A –** Metascape express gene ontology analysis using all 142 DEGs identified in edgeR analysis (fold >2 and FDR <7%) comparing MYC wild-type relative to the MYC-323 mutant. **B –** GSEA with Hallmarks gene sets comparing MYC wild-type relative to the MYC-

323 mutant (FWER p-value <0.05).

References

1. Vervoorts J, Luscher-Firzlaff JM, Rottmann S, Lilischkis R, Walsemann G, Dohmann K, Austen M, Luscher B. Stimulation of c-MYC transcriptional activity and acetylation by recruitment of the cofactor CBP. EMBO Rep. 2003; 4: 484-90. doi: 10.1038/sj.embor.embor821.

2. Adhikary S, Eilers M. Transcriptional regulation and transformation by Myc proteins. Nat Rev Mol Cell Biol. 2005; 6: 635-45. doi: 10.1038/nrm1703.

3. Faiola F, Liu X, Lo S, Pan S, Zhang K, Lymar E, Farina A, Martinez E. Dual regulation of c-Myc by p300 via acetylation-dependent control of Myc protein turnover and coactivation of Myc-induced transcription. Mol Cell Biol. 2005; 25: 10220-34. doi: 10.1128/MCB.25.23.10220-10234.2005.

4. Yang W, Shen J, Wu M, Arsura M, FitzGerald M, Suldan Z, Kim DW, Hofmann CS, Pianetti S, Romieu-Mourez R, Freedman LP, Sonenshein GE. Repression of transcription of the p27(Kip1) cyclin-dependent kinase inhibitor gene by c-Myc. Oncogene. 2001; 20: 1688-702. doi: 10.1038/sj.onc.1204245.

5. Patel JH, Du Y, Ard PG, Phillips C, Carella B, Chen CJ, Rakowski C, Chatterjee C, Lieberman PM, Lane WS, Blobel GA, McMahon SB. The c-MYC oncoprotein is a substrate of the acetyltransferases hGCN5/PCAF and TIP60. Mol Cell Biol. 2004; 24: 10826-34. doi: 10.1128/MCB.24.24.10826-10834.2004.

6. Zhang K, Faiola F, Martinez E. Six lysine residues on c-Myc are direct substrates for acetylation by p300. Biochem Biophys Res Commun. 2005; 336: 274-80. doi: 10.1016/j.bbrc.2005.08.075.

7. Farrell AS, Sears RC. MYC degradation. Cold Spring Harb Perspect Med. 2014; 4. doi: 10.1101/cshperspect.a014365.

8. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015; 12: 357-60. doi: 10.1038/nmeth.3317.

9. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013; 9: e1003118. doi: 10.1371/journal.pcbi.1003118.

10. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26: 139-40. doi: 10.1093/bioinformatics/btp616.

11. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 2012; 40: 4288-97. doi: 10.1093/nar/gks042.

12. TW HB, Girke T. systemPipeR: NGS workflow and report generation environment. BMC Bioinformatics. 2016; 17: 388. doi: 10.1186/s12859-016-1241-0.

13. Tripathi S, Pohl MO, Zhou Y, Rodriguez-Frandsen A, Wang G, Stein DA, Moulton HM, DeJesus P, Che J, Mulder LC, Yanguez E, Andenmatten D, Pache L, et al. Meta- and Orthogonal Integration of Influenza "OMICs" Data Defines a Role for UBR4 in Virus Budding. Cell Host Microbe. 2015; 18: 723-35. doi: 10.1016/j.chom.2015.11.002.

14. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014; 15: 550. doi: 10.1186/s13059-014-0550-8.

15. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005; 102: 15545-50. doi: 10.1073/pnas.0506580102.

16. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E, Houstis N, Daly MJ, Patterson N, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet. 2003; 34: 267-73. doi: 10.1038/ng1180.

17. Bretones G, Delgado MD, Leon J. Myc and cell cycle control. Biochim Biophys Acta. 2015; 1849: 506-16. doi: 10.1016/j.bbagrm.2014.03.013.

18. Dang CV. MYC, metabolism, cell growth, and tumorigenesis. Cold Spring Harb Perspect Med. 2013; 3. doi: 10.1101/cshperspect.a014217.

19. McMahon SB. MYC and the control of apoptosis. Cold Spring Harb Perspect Med. 2014; 4: a014407. doi: 10.1101/cshperspect.a014407.

Chapter 6

Conclusion

The current dissertation focuses on two areas of research largely overlooked in the cancer biology field: the isoform-specific alterations that in occur in human cancer and the regulatory mechanisms controlling MYC expression and activity. While mostly descriptive, the studies presented provide several novel findings and a foundation for future investigations interested in examining the molecular derailments of ccRCC and the mechanisms regulating MYC activity.

In chapter 2, we take an unbiased computational and experiment approach exploring the transcript-specific changes observed in ccRCC. While previously undertaken by other labs, our lab implemented a series of new computational methods to reliably identify differential expressed transcripts and the molecular pathways altered in ccRCC. One of the most interesting findings from our study was the discovery of lncRNAs, *FGD5-AS1* and *AL035661.1,* of which have no known function and were among the top downregulated genes within ccRCC. *FGD5-AS1* and *AL035661.1* were among the top 30 coexpressed genes implicated in TCA cycle, and would be excellent candidates for future functional studies in ccRCC.

Additionally, the study also highlights the unspoken issues seen with different differential expression analyses. In the study, we used gene-level analyses, such as edgeR and sleuth, and observed rather large differences in the number of discovered DEGs between ccRCC tumors and normal adjacent renal tissue. Furthermore, we also discovered transcript-level analyses would also render unique deregulated genes, not detected by any of the gene-level approaches.

As discussed in chapter 3, we suspect that the lack of sensitivity seen with gene-level analyses is in part attributed to the different alterations, and sometimes opposing changes, observed with transcripts derived from the same gene. In other words, in many of the cases we observed differential transcript expression with only one of the transcripts from a gene. This circumstance presents a situational, but fundamental, issue with gene-level analyses because if the unaffected transcripts are of suitable abundance, gene-level analyses could "overlook" the deregulation of the gene.

In the future, it would be interesting to examine larger cohorts of match paired samples to evaluate how mutational status may influence the differential transcript expression analysis. The current study used 50 match paired samples, and looks at the transcript alterations independent of mutational status. As mutations in *SETD2,* and deficiencies in H3K36me3 are commonplace and known to influence splicing events in ccRCC, a transcript-level analysis segregating the paired samples according to these alterations could render different results than the findings presented in this dissertation [1, 2].

Findings from aforementioned computational analyses identified and confirmed the downregulation of *HM1* in ccRCC. In chapter 4, we provide the first evidence of a highly specific and pervasive downregulation of one transcript of *HM1*, referred to as *HM1-3*. The study also provides preliminary evidence to suggest that *HM1-3* regulates the hypoxic response in ccRCC through the suppression of HIF protein levels. However, the mechanisms of how *HM1-3* regulates HIF

protein levels and consequently hypoxia-responsive genes are unknown. As *HM1-3* is almost exclusively found within the cytoplasm, *HM1-3* likely regulates the HIF proteins via a post-transcriptional mechanism. *HM1* has been suggested to be a miRNA sponge; however, as the mRNA levels of the *HIF* genes are unaffected with *HM1* knockdown, this mechanism seems less likely [3, 4]. Moreover, the absolute abundance of *HM1* is rather low in most cell types; therefore, it appears more likely that *HM1-3* is affecting either the production or the stability of the HIF proteins.

We found that steady-state RNA levels are largely unaffected with *HM1* knockdown in CAKI-1 cells. However, hypoxia-responsive genes, such as *ANGPTL4*, were among the most differentially expressed with *HM1* knockdown. Evaluation of the *ANGPTL4* pre-mRNA also showed an upregulation with *HM1* knockdown suggesting that the increase in *ANGPTL4* steady-state levels is likely an increase in transcription of *ANGPTL4.* Furthermore, it is highly suspected that many of the steady-level RNA changes observed with *HM1* knockdown are secondary effects and consequences of the changes seen with the HIF protein levels. Alternatively, genes, such as *DDAH1*, are likely upregulated through a different mechanism. As *DDAH1* pre-mRNA is unaffected with *HM1* knockdown and it is the largest upregulation observed with *HM1* knockdown, *HM1* may be a direct target of *HM1.* However, no appreciable change was observed with DDAH1 protein level with *HM1* knockdown. The time of collection of the cells after the *HM1* knockdown could account for the absence of change in *DDAH1*

protein levels. However, additional studies are needed to determine whether DDHAH1 protein levels are upregulated. Additionally, as *DDAH1* is heavily downregulated in ccRCC tissues, the *DDAH1* upregulation seen could be a compensatory response of the cells to *HM1* knockdown, and may not, in the end, contribute to the ccRCC pathology.

Immediate studies that elucidate the mechanisms of how *HM1-3* regulates HIF proteins levels are warranted. Additionally, a more global proteomics assessment in the presence of *HM1* knockdown may reveal a broader influence on the proteomic landscape. Furthermore, studies using reagents, such as cycloheximide, in the present of the *HM1* knockdown would likely resolve whether *HM1* regulates protein production, stability or both. Lastly, more detailed investigations exploring *HM1* in developing kidney cells are needed. The experiments conducted (by Matt Young) show an increase in *HM1* expression with kidney cell differentiation and a loss of kidney differentiation markers with *HM1* knockdown in kidney progenitor cells. Loss of *HM1* during the course of kidney cell differentiation would solidify the necessity of *HM1* in kidney cell development. Additionally, studies carried out to more terminally differentiated kidney cell states would also been beneficial in further implicating *HM1* function in normal kidney biology.

In chapters 4 and 5, we explore the regulatory mechanisms controlling MYC expression and activity. In chapter 4, we review the growing literature supporting a complex regulatory network between MYC and several lncRNAs. In our review

of the literature, we highlight several experimental studies that show lncRNAs, many of them in close proximity to the MYC locus, regulating the transcription of MYC and also serving as miRNA sponges to increase MYC mRNA levels [5]. As lncRNAs exhibit a higher degree of tissue-specific expression, it would be interesting to learn if the same lncRNA regulatory functions of MYC extent to multiple tissues or cell types [6, 7]. Additionally, as *PVT1* was discovered to bind to MYC, thereby controlling its degradation in cytoplasm [8], future studies examining other MYC binding lncRNAs are of great interest. LncRNAs are best known for their function in transcriptional regulation, and to our knowledge there are no investigations exploring lncRNAs bound to MYC in the nucleus [9].

Finally, in chapter 5, we investigated the transcriptomic changes that occur when MYC acetylation states are altered in Rat1a. Missense mutations in ectopically expressed MYC were created to convert lysines at positions 149, 158 and 323 to arginines. These mutations in turn inhibit acetylation of MYC at these residues positions. However, it is possible that the missense mutations themselves could be responsible for the transcriptomic changes seen and not necessarily the changes in MYC acetylation. Additionally, another disadvantage to this experimental approach is that the mutations are not specific to inhibiting only MYC acetylation. Conversion of the lysines into arginines also affects ubiquitination at these residues; therefore, our study cannot rule out that the transcriptomic changes discovered are solely attributed to MYC acetylation. However, there are a couple of experiments that would help support MYC

acetylation as a mediator of its transcriptional activity and thereby in part responsible for the observed gene-selective effects. First, ChIP experiments, using the antibodies specific to the MYC mutants, should be consistent with the gene-selective expression changes found. In other words, expression increases of a gene should be accompanied by changes in MYC occupancy on the promoter of the same gene. These findings would support that the gene-selective effects are direct effects of MYC and not secondary effects through an intermediate factor. Second, electromobility shift assays could be performed using promoter sequences, identified above, with *in vitro* generated acetylated MYC to see if MYC binding affinities are altered for the respective promoter sequences.

Altogether, the findings presented in the dissertation provide new insights into complexities of the derailments seen in cancer and highlight a potentially new mechanism regulating MYC activity. It is our hope that the information gathered from our studies will provide a backdrop for investigators in the future to develop experimental strategies to help elucidate our unanswered and also novel therapeutic strategies to combat human cancer.

References

1. Simon JM, Hacker KE, Singh D, Brannon AR, Parker JS, Weiser M, Ho TH, Kuan PF, Jonasch E, Furey TS, Prins JF, Lieb JD, Rathmell WK, et al. Variation in chromatin accessibility in human kidney cancer links H3K36 methyltransferase loss with widespread RNA processing defects. Genome Res. 2014; 24: 241-50. doi: 10.1101/gr.158253.113.

2. Ho TH, Park IY, Zhao H, Tong P, Champion MD, Yan H, Monzon FA, Hoang A, Tamboli P, Parker AS, Joseph RW, Qiao W, Dykema K, et al. High-resolution profiling of histone h3 lysine 36 trimethylation in metastatic renal cell carcinoma. Oncogene. 2016; 35: 1565-74. doi: 10.1038/onc.2015.221.

3. Chen ZH, Wang WT, Huang W, Fang K, Sun YM, Liu SR, Luo XQ, Chen YQ. The lncRNA HOTAIRM1 regulates the degradation of PML-RARA oncoprotein and myeloid cell differentiation by enhancing the autophagy pathway. Cell Death Differ. 2017; 24: 212-24. doi: 10.1038/cdd.2016.111.

4. Zheng M, Liu X, Zhou Q, Liu G. HOTAIRM1 competed endogenously with miR-148a to regulate DLGAP1 in head and neck tumor cells. Cancer Med. 2018. doi: 10.1002/cam4.1523.

5. Hamilton MJ, Young MD, Sauer S, Martinez E. The interplay of long non-coding RNAs and MYC in cancer. AIMS Biophys. 2015; 2: 794-809. doi: 10.3934/biophy.2015.4.794.

6. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, et al. Landscape of transcription in human cells. Nature. 2012; 489: 101-8. doi: 10.1038/nature11233.

7. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 2011; 25: 1915-27. doi: 10.1101/gad.17446611.

8. Tseng YY, Moriarity BS, Gong W, Akiyama R, Tiwari A, Kawakami H, Ronning P, Reuland B, Guenther K, Beadnell TC, Essig J, Otto GM, O'Sullivan MG, et al. PVT1 dependence in cancer with MYC copy-number increase. Nature. 2014; 512: 82-6. doi: 10.1038/nature13311.

9. Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. Annu Rev Biochem. 2012; 81: 145-66. doi: 10.1146/annurev-biochem-051410-092902.