Coordinated post-transcriptional regulation by microRNAs and RNA-binding proteins

by

Akiko Sekikawa

A thesis submitted in conformity with the requirements for the degree of Master of Science

Department of Molecular Genetics

University of Toronto

© Copyright by Akiko Sekikawa (2013)
Coordinated post-transcriptional regulation by microRNAs and RNA-binding proteins

Akiko Sekikawa

Master of Science

Department of Molecular Genetics

University of Toronto

2013

Abstract

Both microRNAs (miRNAs) and RNA-binding proteins (RBPs) regulate post-transcriptional events, but the post-transcriptional contribution to the global mammalian transcriptomes is still not well understood. In this study we study the synergistic interaction between microRNAs that inhibit gene production, and a special RBP, HuR, that positively regulates mRNA stability. We examined their relationship in terms of spatial, conservational and expressional perspective. We show comprehensive mapping of HuR binding sites by combination of its structural and sequential preferences; and cross-platform normalization method within a process of refining miRNA and HuR binding site mapping. Finally, we observed co-evolution of miRNA and HuR binding sites by looking at their proximity and conservation levels. Collectively, our data suggest that mammalian microRNAs and HuR, with seemingly opposing regulatory effects, cooperatively regulate their mutual targets.
Acknowledgements

I would like to thank Dr. Zhaolei Zhang for mentoring me. His thoughtful advices and supports always motivated and encouraged me throughout my studies. I am also grateful to Dr. Jeffery Wrana for providing me with chance to study in the University of Toronto and for broadening my perspective on my project by advising from views of different research area. I also appreciate the guidance and comments from my Supervisory Committee, Dr. Benjamin Blencowe and Dr. Jason Moffat.

I would like to especially grateful to TaeHyung Kim and Yue Li from the Zhang laboratory for their advices and helps on data analysis and learning computational skills. I am also grateful to Ka-Chun Wong and a former Zhang laboratory member Dr. Jingjin Li for advises and helping me revising this thesis. Among the members of the Wrana laboratory, I am particularly thankful to Dr. Ian Taylor and Yu Liu on giving me advises on my project directions. I would like to show regards and appreciation to my friends from the Zhang laboratory and Wrana laboratory for working with and supporting me.

I would also like to thank following people on checking grammars and comments to complete this thesis; Brian Law, Kota Hatta, Nhat Tran and Emilie Nham.
# Table of contents

Abstract .......................................................................................................................... ii
Acknowledgements ........................................................................................................ iii
Table of contents ........................................................................................................... iv
List of tables .................................................................................................................. v
List of figures .................................................................................................................. vi
List of abbreviations ....................................................................................................... vii

1. Introduction .................................................................................................................. 1
   1.1. mRNA regulation and stability ........................................................................... 1
   1.2. HuR mediated post-transcriptional regulation ................................................... 4
   1.3. MicroRNA mediated post-transcriptional regulation ......................................... 7
   1.4. Coordinated post-transcriptional regulation by microRNAs and RBPs ............. 9

2. Rationale ..................................................................................................................... 11

3. Materials and Methods ............................................................................................. 12
   3.1. Initial Dataset Processing .................................................................................. 12
   3.2. Accessibility ....................................................................................................... 13
   3.3. Binding motif prediction .................................................................................... 14
   3.4. Combining results of accessibility and motif search ......................................... 15
   3.5. Expression Data ............................................................................................... 15
   3.6. Conservation of RBP/miRNA targets ................................................................ 16
   3.7. Protein-protein interaction data ....................................................................... 19

4. Results ....................................................................................................................... 20
   4.1. RBP and MicroRNA binding sites mapping ...................................................... 21
   4.2. Cross-platform normalization and expression data integration ......................... 30
   4.3. Integration of expression data and common targets of HuR and microRNA ..... 35
   4.4. Conservation of target sites .............................................................................. 38
   4.5. Integration of PPI network with binding sites prediction/expression data ...... 48

5. Discussion and Conclusion ......................................................................................... 52
   5.1. Comprehensive RBP binding site prediction ...................................................... 52
   5.2. Expression data integration refines binding site prediction ............................. 54
   5.3. HuR and AGO/miRNA mediated targets post-transcriptional regulation ........ 56
   5.4. Conclusion .......................................................................................................... 58

6. References .................................................................................................................. 61
List of tables

Table 1 Result of curations for small RNA-seq and mRNA-seq data ........................................... 32
Table 2 Pearson correlation of same/similar-cell line expression data............................................. 34
Table 3 Results of DAVID functional analysis ................................................................................. 37
Table 4 Portion of significantly expressed genes is similar in all targeting states ....................... 51
List of figures

Figure 1 Study pipeline ........................................................................................................................................................................... 20
Figure 2 Accessibility and 7mer score peaks around cross-linking site on HuR binding reads ...................................................................................................................................................................................................... 23
Figure 3 3-D plots of 7mer score, accessibility and read position for HuR ........................................................................................................ 24
Figure 4 Accessibility and 7mer scores are significantly higher at HuR crosslinking sites ........................................................................ 25
Figure 5 Determination of cut-offs of HuR target reads from comparison of boxplots ........................................................................ 26
Figure 6 Accessibility peaks around cross-linking site on AGO binding reads .................................................................................... 28
Figure 7 Accessibility is significantly higher at AGO crosslinking sites ............................................................................................... 29
Figure 8 Determination of cut-offs of AGO target reads from comparison of boxplots ....................................................................... 29
Figure 9 Number of miRNA binding sites have similar distribution as accessibility ........................................................................... 30
Figure 10 Percentile rank shows high significance on cross-platform normalization (small RNAseq) ...................................................................................................................................................................................................... 33
Figure 11 Percentile rank shows high significance on cross-platform normalization (RNAseq) ...................................................................................................................................................................................................... 34
Figure 12 Results Number of predicted HuR and AGO targets after cut-offs ........................................................................................ 36
Figure 13 Boxplot HuR and AGO target expressions ............................................................................................................................. 37
Figure 14 HuR has lower conservation scores around binding regions, being opposed to AGO ...................................................................................................................................................................................................... 39
Figure 15 Average conservation score plot after expression data cutoff ............................................................................................ 40
Figure 16 HuR RNA recognition motifs are highly conserved in primate species ................................................................................ 42
Figure 17 Orthologous sequences share similar accessibility distributions as human HuR targets ...................................................................................................................................................................................................... 43
Figure 18 Orthologous sequences share similar 7mer scores as human HuR targets ........................................................................ 44
Figure 19 No. of insertions associated with lower conservation of HuR binding sites ........................................................................... 45
Figure 20 HuR binding motif lengths are species specific ......................................................................................................................... 46
Figure 21 HuR and miRNA binding sites in close proximity are more conserved .................................................................................. 47
Figure 22 Selected Cytoscape visualized tissue specific PPIs ................................................................................................................ 49
Figure 23 Human PPI network overlaid with HEK293 HuR knockdown data .......................................................................................... 50
## List of abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGO</td>
<td>Argonaute</td>
</tr>
<tr>
<td>APA</td>
<td>Alternative polyadenylation</td>
</tr>
<tr>
<td>ARE</td>
<td>AU-rich element</td>
</tr>
<tr>
<td>AU-rich</td>
<td>Adenylate-uridylate-rich</td>
</tr>
<tr>
<td>cDNA</td>
<td>Complementary DNA</td>
</tr>
<tr>
<td>CDS</td>
<td>Coding regions</td>
</tr>
<tr>
<td>CLIP-seq</td>
<td>Cross-linking immunoprecipitation-high-throughput</td>
</tr>
<tr>
<td>CR-APA</td>
<td>Coding region-alternative polyadenylation</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>EST</td>
<td>Expressed sequence tag</td>
</tr>
<tr>
<td>FPKM</td>
<td>Fragments per kilobase of exon per million fragments mapped</td>
</tr>
<tr>
<td>GO</td>
<td>Gene ontology</td>
</tr>
<tr>
<td>HITS-clip</td>
<td>High-throughput sequencing and crosslinking immunoprecipitation</td>
</tr>
<tr>
<td>HNS</td>
<td>HuR Nucelocytoplasmic Shuttling sequence</td>
</tr>
<tr>
<td>HOMER</td>
<td>Hypergeometric Optimization of Motif EnRichment</td>
</tr>
<tr>
<td>HuR</td>
<td>Human antigen R</td>
</tr>
<tr>
<td>IRES</td>
<td>Ribosome entry site</td>
</tr>
<tr>
<td>MEME</td>
<td>Multiple EM for Motif Elicitation</td>
</tr>
<tr>
<td>miRNA</td>
<td>microRNA</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger RNA</td>
</tr>
<tr>
<td>PAR-CLIP</td>
<td>Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation</td>
</tr>
<tr>
<td>PPI</td>
<td>Protein-protein interaction</td>
</tr>
<tr>
<td>RBP</td>
<td>RNA-binding protein</td>
</tr>
<tr>
<td>RIP-chip</td>
<td>RNA-Binding Protein Immunoprecipitation-Microarray</td>
</tr>
<tr>
<td>RISC</td>
<td>RNA-induced silencing complex</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic acid</td>
</tr>
<tr>
<td>RNAi</td>
<td>RNA interference</td>
</tr>
<tr>
<td>RNAseq</td>
<td>RNA Sequencing</td>
</tr>
<tr>
<td>RRM</td>
<td>RNA recognition motif</td>
</tr>
<tr>
<td>siRNA</td>
<td>Small interfering RNA</td>
</tr>
<tr>
<td>UCSC</td>
<td>University of California, Santa Cruz</td>
</tr>
<tr>
<td>UTR</td>
<td>Untranslated region</td>
</tr>
<tr>
<td>UTR-APA</td>
<td>3'UTR-alternative polyadenylation</td>
</tr>
<tr>
<td>UV</td>
<td>Ultraviolet</td>
</tr>
</tbody>
</table>
1. Introduction

Stability is an important keyword to describe both transcriptional and post-transcriptional regulation, and is one of the main focuses in gene regulation. A simple PubMed search shows that a keyword “mRNA stability” gives 2115 hits for 2012, which is double the amount from ten years ago, which tells us growing popularity in researches in mRNA regulations. Messenger RNA regulation is roughly divided into splicing, export from the nucleus to the cytoplasm, and translation, where stability and expression levels are closely related. Especially, microRNAs (miRNAs) and RNA-binding proteins (RBPs) are critical factors to understand post-transcriptional regulation and stability controls particularly at the 3' untranslated region (UTR). This introduction will summarise general understanding about how mRNA stability and gene expression are controlled by different regulators at pre- and post-transcriptional levels, followed by a review on the main topic of this thesis on miRNA and RBPs mediated post-transcriptional events.

1.1. mRNA regulation and stability

Stability of mRNAs is determined in various steps including 1) length of introns/3'UTRs, 2) alternative splicing and polyadenylation (APA), and 3) trans-factors that bind to mRNAs. Firstly, the control of stability and expression start at the gene constitution level. An inverse proportional relationship between intron length and expression levels was observed in Caenorhabditis elegans and Homo sapiens. In 2003,
housekeeping genes were reported their compactness in regard to not only introns but also untranslated regions and total gene, arguing their extension and confirmation of previous results on shorter introns of highly expressed genes\(^4\). Both studies reasoned that short length of introns/genes would be due to selection against transposable element insertion into introns and towards elimination of introns for highly expressed genes, respectively.

Secondly, post-transcriptional modifications including splicing and polyadenylation are other sources of stability and expression level controls\(^9\). Up to 95% of human multi-exon genes are known to be alternatively spliced from mRNA-seq data, and computational analysis of cDNA/ESTs showed that \(\sim 54\%\) of human transcripts have multiple polyadenylation sites\(^10,11\). In mRNA splicing, alternative exons are shorter than constitutive exons and have shorter flanking introns, and exon inclusion favoured by short introns was seen in CD44 transcript isoforms. This is consistent with higher expression of genes with short introns\(^4,5,12\). The impacts of APA events are divided in two types depending on where it happens, internal introns/exons (CR-APA) or 3’UTR (UTR-APA). The CR-APA can change the gene expression qualitatively as it leads to protein isoforms with distinctly different constitutions, while the UTR-APA potentially change expression quantitatively because number of \(cis\)-regulatory elements or microRNA (miRNA) binding site changes by the difference in the length of 3’ UTR, which may affect the expression level\(^6,7\).

Importantly, RNA stability control is also facilitated by \textit{trans}-acting elements in the 3’UTR including RNA-binding proteins (RBPs) and microRNAs (miRNAs). As discussed above, 3’ UTRs are important in terms of mRNA stability. Adenylate-uridylate-rich
elements (AREs) are a type of cis-regulatory element in the 3’UTR that regulate the stability and are known to be targeted by RBPs. It is known to have canonical sequences, AUUUA or UUAUUUAUUU that lies in length of 50-150 nucleotide sequences, and 9% of cellular mRNAs contains AREs\textsuperscript{13,14}. AU-rich sequences were first discovered in the human and mouse TNF gene\textsuperscript{14}. ARE binding protein-mediated stability controls have been reported in some RBPs. While many of them including TTP, BRF1, BRF2, AURF1, KSRP have a destabilizing effect to their targets, Hu protein families stabilize their targets\textsuperscript{8,15}. \textit{C-fos} is one of well-characterized genes in regards to the ARE in its 3’UTR and has been used in different studies to reveal characteristics of ARE including its canonical sequence identification\textsuperscript{16}. In 2001 Chen \textit{et al.} examined these RBPs that promote destabilization of their targets such as AUF1, AUBP, KSRB and TTP binding on the \textit{c-fos} mRNA by recruiting the exosome for target degradation\textsuperscript{17}. The \textit{c-fos} mRNA was previously revealed to be a target by multiple RBPs suggesting that ARE can have multiple binders. RBPs are also known to have multiple targets\textsuperscript{8}. Argonaute (AGO) is another type of RBP that regulates mRNAs at 3’UTRs by recruiting small non-coding RNAs such as miRNAs and small interfering RNAs (siRNAs) in RNA-mediated gene silencing mechanisms in eukaryotes\textsuperscript{18,19}. The AGO protein has three domains, PAZ, MID and PIWI, where PAZ binds to single-stranded RNA\textsuperscript{20}. It is a key component of RNA-induced silencing complex (RISC), in which it binds to these small RNAs using the PAZ and MID domains. Then the RISC binds to its targets in a sequence-dependent manner, and the endonucleolytic activity of the PIWI domain cause target RNA cleavage\textsuperscript{20,21}. In summary, gene expression and stability are controlled in different stages of gene regulation. However, it is undeniable that post-transcriptional events play important roles in these stages; especially 3’end regulation
by RNA-binding proteins and miRNAs is crucial to determine the fate of transcripts, which will be further discussed in the following sections.

1.2. HuR mediated post-transcriptional regulation

HuR is one of the RBPs mentioned above that binds to AREs to regulate stability of its targets. Recent studies on HuR have revealed its importance by showing roles in all stages of the post-transcriptional regulation from splicing to translation\textsuperscript{22,23}. HuR is a part of Hu protein family, and expressed broadly across tissues, while other members of the family are known to be neuron or nervous tissue specific\textsuperscript{24}. HuR is also known as a type of ARE binding protein, and Benoit et al. discovered in 2010 that two of its three RNA recognition motifs (RRMs), RRM1 and RRM2 are known to bind to AU-rich sequences on 3’ UTR of mRNAs from x-ray crystal structure\textsuperscript{25}. In 2011, it was reported that HuR binding to its targets is initiated by its homodimerization. HuR binds both to AU-rich and U-rich sequences with higher affinity to U-rich RNAs\textsuperscript{26}. The RRM3 is connected to RRM2 by a basic hinge region. The role of RRM3 is maintenance of the stability of the RNA-protein complex and/or polyA tail binding\textsuperscript{22}.

HuR is predominantly located in the nucleus and facilitates shuttling itself between the nucleus and cytoplasm through HNS (HuR Nucleocytoplasmic Shuttling sequence) in the hinge region\textsuperscript{27,28}. Both import and export of HuR are mediated by transportins, Trn1 and Trn2. Two protein ligands of HuR, pp32 and APRIL, signals of which are recognized by
the export receptor, CRM1, are also involved in HuR import\textsuperscript{29,30}. In the nucleus, HuR is involved in post-transcriptional modification including splicing and polyadenylation. Neuron-specific alternative splicing activity of the Hu family was observed in neurofibromatosis type 1 (NF1)\textsuperscript{31}. Izquierdo also reported that HuR prevents a splicing factor U2AF65 from alternative splicing of apoptosis-promoting factor Fas, which results in absence of exon 6\textsuperscript{32}. Lebedeva \textit{et al.} compared RNA-seq data before and after siRNA knockdown of HuR to find HuR dependent alternative splicing, and found 30 reduced and 21 increased exons\textsuperscript{33}. HuR activity in APA is also indicated. Zhu \textit{et al.} showed evidence of Hu family proteins blocking cleavage and polyadenylation using \textit{in vitro} polyadenylation assay with HeLa cell nuclear extract and recombinant Hu proteins\textsuperscript{34}. HuR also auto-regulates its alternative polyadenylated isoform through competition with TTP\textsuperscript{35}.

As mentioned above, the HuR protein has nuclear export sequence (NES)-like sequences that are used to shuttle its mRNA between the nucleus and cytoplasm in a CRM1-dependent manner\textsuperscript{36}. This activity is not limited to its own mRNA but also export of other ARE-containing mRNA can be regulated by HuR in this CRM1-dependent manner. This HuR activity of translocation of mRNAs to the cytoplasm has been shown in \textit{c-fos} and CD83\textsuperscript{29,37-39}. Recently, unspliced foamy virus mRNAs were found to be bound by HuR to regulate nuclear export in a similar CRM1-dependent manner\textsuperscript{40}. Another activity of HuR is its involvement in translational regulation in the cytoplasm. Different studies suggested that HuR is involved in translation in multiple ways. Translation of p35 mRNA is enhanced by HuR in response to UV light\textsuperscript{41}. Durie \textit{et al.} reported that HuR interacts with an internal ribosome entry site (IRES) in the 5’UTR of XIAP mRNA to regulate translation\textsuperscript{42}. ARE-
dependent translation of tumour necrosis factor (TNF)-α is mediated by competition between HuR and TTP. When TTP is phosphorylated, its affinity to ARE is decreased and hence HuR-mediated initiation of TNF mRNA translation is enhanced\textsuperscript{43}.

Another cytoplasmic HuR activity is its regulation of mRNA stability through AREs in the 3’UTR. This was first discovered in an \textit{in vivo} study on \textit{c-fos} and \textit{c-jun} ARE in mouse cells. Peng \textit{et al.} confirmed this role of HuR from three pieces of evidences; (i) inhibition of \textit{c-fos} ARE-mediated mRNA decay paralleled by elevation of HuR concentration; (ii) increased formation of RNP complex of HuR and \textit{c-fos} resulting from redistribution of HuR from the nucleus to cytoplasm by transcription blockage; and (iii) differential regulation of mRNA decay mediated by \textit{c-fos} and non-ARE-containing \textit{c-jun}\textsuperscript{44}. Chen \textit{et al.} showed the selective targeting of HuR by comparing \textit{c-fos} and other mRNAs that contain ARE sequence of AUUUA, and any two of three HuR RRM domains being sufficient to stabilize its targets\textsuperscript{45}. The binding motif of HuR was later refined to NNUUNNUUU, and its binding property to single-stranded RNA was discovered\textsuperscript{46}. Among different functions of HuR, this mRNA stability determination are especially important due to its binding to AU-rich sequences in 3’UTR and its positive effect on target mRNA stability particularly because miRNAs also have preference for AU-rich nucleotide composition near their binding sites for efficient negative regulation of their targets\textsuperscript{47}, which will be further discussed in the following section.
1.3. MicroRNA mediated post-transcriptional regulation

In contrast to HuR’s positive regulation on its targets, AGO is well known for its involvement in negative regulation, namely RNA interference (RNAi)\(^{48}\). AGO has been found in various organisms including plants, flies, nematodes and human. 8 Ago genes have been identified in human. The AGO protein has three domains, PAZ, PIWI and MID, as explained briefly above. PAZ and PIWI domains have RNase III-type and RNase-H-like enzyme properties\(^{48,49}\). AGO forms the RISC with single-stranded non-coding RNAs, a siRNA or miRNA in the RNAi pathway\(^{18,19}\).

A class of small non-coding RNA involved in RNAi pathway, miRNAs will be the focus in this section. AGO in miRNA-induced silencing complex (miRISC) interacts with TNRC6 protein to promote deadenylation of miRNA target mRNAs to trigger mRNA decay\(^{50}\). The first miRNA discovered was lin-4 in C. elegans\(^{51}\). lin-4 functions as a negative regulator of itself and determines the timing of developmental events of C. elegans. Although LIN-14 protein levels normally transition from high to low towards the later stage of C. elegans development, lin-4 mutants show an abnormally high level of LIN-14 protein in late in development\(^{52}\). Discovery of miRNA was achieved through the investigation of this mechanism of negative regulation of level of LIN-14 by lin-4\(^{51}\). The nomenclature of miRNA was introduced when two small temporal RNAs, lin-4 and let-7 were determined to be members of small noncoding RNAs 21-24 nucleotides in length\(^{53}\). At the beginning, it was believed that miRNAs were not involved in the RNAi pathway due to extensive complementarity between siRNAs and their targets. However, precursor
miRNAs were found to be cleaved by Dicer and enter RNAi pathway by forming the miRNA/RISC complex to bind to the targets\textsuperscript{54,55}.

To date, 1540 human miRNAs have been reported in miRBase\textsuperscript{56}. RNAi event of miRNAs generally causes mRNA degradation or repression of translation although there are some cases that result in opposite effects\textsuperscript{57-59}. Ribosome profiling shows that mRNA degradation is dominant over translational repression\textsuperscript{60}. In mammals, approximately 60\% of mRNAs are presumably targeted by miRNAs\textsuperscript{61}. Many miRNA loci are polycistronic with multiple miRNAs found in close proximity. RNA polymerase II transcribes these miRNA together, and Drosha and DGCR8 cleaves the transcript at the stem of hairpin structures, resulting in small hairpins, pri-miRNAs. These pri-miRNAs are exported from the nucleus to cytoplasm. The export is mediated by expotin 5. These pri-miRNAs are cleaved into short dsRNAs by Dicer to be loaded into the RISC complex for targeting mRNAs, as mentioned above\textsuperscript{62,63}.

Base pairing of miRNA with mRNA requires its 5’ seed region centred on 2-7\textsuperscript{th} nucleotides\textsuperscript{62}. This seed region is highly conserved in metazoan miRNAs\textsuperscript{64}, and 7-8 nt matches are conserved in the most common type of miRNA targeting\textsuperscript{62}. More than half of the human protein-coding genes that are targeted by miRNAs are under selective pressure to maintain 3’ UTR pairing\textsuperscript{65}. Bartel reviewed additional features that would influence miRNA-targeting efficacy including: binding sites near the stop codon, binding sites away from the centre of long UTRs, existence of AU-rich nucleotide composition near binding sites, and co-expression of miRNAs that have their binding sites close to each other\textsuperscript{62}. 
However, it remains unknown how these features influence miRNA targeting, and hence gene regulation.

1.4. Coordinated post-transcriptional regulation by microRNAs and RBPs

A few studies reported a relationship between miRNAs and RBPs/AREs. Jacobsen et al. reported a possible synergistic relationship between miRNA and ARE in positive regulation of gene expression\textsuperscript{66}. Possible inhibition of some miRNAs by binding of miRNA binding sites by dead end 1 (Dnd1) in human cells was also reported\textsuperscript{67}. Recently, two papers described a relationship between miRNAs and HuR. Glorian reported UV light mediated loss of cooperative effects between miR-19 and HuR on translation rate of Ras homolog B to down-regulate their target\textsuperscript{68}. On the other hand, Tominaga showed competitive regulation of nucleolin by HuR and miR-494\textsuperscript{69}. Nucleolin harbours both HuR and miR-494 binding sites, and its translation is promoted by HuR. However, when miR-494 was overexpressed, negative regulation of nucleolin was observed. Furthermore, Epis et al. showed that HuR antagonises the microRNA down-regulation effect in prostate cancer cells\textsuperscript{70}.

Different types of miRNA and HuR mediated events have been reported, and the relationship between these two factors would provide a more comprehensive understanding on how gene expression is tuned and controlled. Recently, Mukerjee et al. suggested that overall distance between HuR and miRNA binding sites influences its target expression.
levels\textsuperscript{71}, which provides evidence that miRNAs and HuR possibly have global relationship on gene expression. Among various processes of gene expression that has been discussed so far, this global relationship between \textit{cis}-regulatory elements and/or their \textit{trans}-regulators are still not well understood. Detailed investigations would provide more understanding about how post-transcriptional events are regulated. This thesis will discuss about insights into extensive miRNAs and HuR binding sites mapping and endeavours to dissect their relationship by integrating expression data and conservation data.
2. Rationale

In response to recent studies on various relationship between HuR and miRNA mediated post-transcriptional regulations where they work independently or cooperatively to influence their mutual targets in various studies\textsuperscript{68-70,72}, it would be possible to hypothesize that HuR and miRNA can be globally related to each other in post-transcriptional regulations. Investigations of these two factors would be very intriguing to understand how they are related. This analysis can be done effectively using appropriately predicted HuR and miRNA binding sites. To obtain these sites, there has been various \textit{in vivo} assays have been introduced using RBP-RNA co-IP techniques. RIP-chip assay, which is combination of this co-IP technique and microarray, would be the attempt of identification of RBP binding sites at earlier time\textsuperscript{73}. With the introduction of next-generation sequencing\textsuperscript{74}, this co-IP technique was combined with this sequencing technique called CLIP-seq\textsuperscript{75}. This study aimed to map HuR and miRNA binding sites using this \textit{in vivo} co-IP assay high-throughput sequencing technique as well as focusing their sequence preferences, followed by further investigations in terms of conservation and gene expression of their binding sites to characterize the relationship between HuR and miRNA regulation of common targets.
3. Materials and Methods

3.1. Initial Dataset Processing

AGO2 binding data from the HEK293 PAR-CLIP assay was collected from a study by Hafner et al.\textsuperscript{76}. They clustered reads that overlapped by at least one nucleotide, and ranked the clusters by the number of thymine to cytosine mutations to exclude the background clusters with low mutation rates, resulting in 17,319 high-confidence reads. Similarly, HuR binding data in HEK293 cells from the PAR-CLIP assay was obtained from online resources. Kishore et al. used a combination of a Gaussian mixture model approach and HuR knockdown assays to obtain the 5000 most confident reads\textsuperscript{77}. These author-filtered AGO and HuR binding reads were used for further analysis.

Although all the reads obtained from online resources were already annotated, each read from the AGO and HuR binding data was double-checked for its annotation and sequences. First, whole human genome sequences with hg18 refseq annotation were downloaded from the UCSC Genome Browser\textsuperscript{78}. Each sequence has exons and introns noted in upper and lower cases, respectively. Taking advantage of this notation, genomic and transcript coordinates of exons were calculated. Using this exon map, the genomic and transcript coordinates of the reads were also calculated to classify them as 5’ UTR, CDS, or 3’ UTR. Those read coordinates or sequences that did not match between the original dataset and the exon map were discarded. Then 80 nucleotides were added both upstream and downstream of each read to compare with RBP binding and background regions.
As negative datasets, two types of random sequences were prepared. One was prepared by shuffling the bases in the positive datasets. For the other, sequences with the same length as the positive dataset were randomly retrieved from the genome, keeping the same GC content and gene region as the original read, i.e. if the read was from the 3’ UTR, this random sequence was also chosen from 3’ UTR sequences.

3.2. Accessibility

RBP binding site accessibility was calculated for PAR-CLIP data using a programme called RNAplfold\textsuperscript{79}. RNAplfold calculates the probabilities of base pairing in RNA sequences by determining the local structure within the given window size to computationally determine the stability of the RNA secondary structure. For a given length $L$ of a secondary structure, the average probability of pairing of $u$ consecutive bases is calculated within a fixed-size window $W$. This programme was originally implemented to search for non-coding RNA binding sites, and its effectiveness on siRNAs and RBPs binding to their targets was shown by both Tafer \textit{et al.} and Li \textit{et al.}\textsuperscript{80,81}. Parameter values of $L=40$ and $W=80$ were used based on the previous optimization in these studies for both RIP-chip and PAR-CLIP datasets, while the value of $u$ was determined based on the motif lengths for both types of data. The lengths from motif prediction by MEME were used as $u$ for RIP-chip datasets, while $u=7$ was used for PAR-CLIP datasets.
3.3. Binding motif prediction

*HuR binding sites prediction*

HuR binding reads from PAR-CLIP data were evaluated using a 7-mer score that provides sequence preference scores generated from an *in vivo* RNAcompete assay\(^82\). This assay is characterised by the design of the microarray chip, which has RNA oligonucleotides prepared using two different deBruijn sequences, which is cyclic sequences of given alphabet with a specific size and length\(^83\), to specifically identify unknown RBP binding preferences by scoring combinations of 7 nucleotide sequences according to the microarray intensities. For each HuR binding read and their random sequences, these 7-mer scores are assigned to each 7 nucleotides of the sequences in sliding window manner.

*MicroRNA binding sites prediction*

To determine AGO binding sites, rather than finding AGO’s consensus binding motifs like other RBPs, the fact that it binds to its targets together with miRNAs in RISC was exploited. miRNA binding sites were predicted using TargetScan\(^65\), according to stringent seed pairing, site number, and site type. TargetScan also considers the site context, including features such as local AU contribution and distance from end of the UTR.
3.4. Combining results of accessibility and motif search

Cut-offs for accessibility and 7-mer scores were decided based on comparison between boxplots of positive and negative sets for both RIP-chip and PAR-CLIP data. A similar method is described in Tafer et al.\textsuperscript{80}. To achieve more precise binding site predictions, the threshold was set relatively strictly by choosing the upper whisker of the boxplots of the negative datasets. The calculation was done using the following equations in MATLAB;

\[
\text{Upper whisker} = q2 + 1.5 \times (q2-q1)
\]

Where \(q1\) and \(q2\) are the 25th and 75th percentiles of given data,

To further specify the RBP binding regions on PAR-CLIP reads, Wilcoxon’s ranksum test was applied between positive and negative datasets for both accessibility and 7-mer scores, and the significant peaks were chosen as the RBP binding regions. For RIP-chip data, the probe sequences with higher accessibility scores than the cut-off and MEME motifs matches were sought.

3.5. Expression Data

Raw miRNA and mRNA expression data was curated from GEO and ArrayExpress for microarray and RNA-seq assays. The data was limited to those cell lines/tissue types
that have both miRNA and mRNA expression data for both assay types. Illumina platform data were gathered for mRNA and small RNA-seq.

The small RNA and mRNA-seq data was also processed primarily for more accurate comparison between expression data from different studies. FPKM (fragment per kilobase of exon per million fragments mapped) of mRNA-seq data was obtained by Yue Li. Small RNA-seq adapters were removed from sequencing reads the ccatadapt software\textsuperscript{84}. For both the small RNA and mRNA-seq data, reads were mapped to the genome using Bowtie, and then annotation and expression levels in FPKM were obtained using cufflinks and tophat\textsuperscript{85-87}. After obtaining FPKM expression for both miRNAs and mRNAs, two types of normalizations (percentile rank and quantile) on all the data set were performed, followed by calculating the Pearson correlation between same samples from different platforms to check for data quality after normalization. Then, the HuR and miRNA target predictions were refined by checking if HuR, miRNAs, and their targets are expressed.

3.6. Conservation of RBP/miRNA targets

Conservation scores

Conservation scores calculated by the software phastCon were downloaded from the UCSC Genome Browser\textsuperscript{88}. phastCon provides scores based on hidden Markov models to identify conserved elements in a multiple sequence alignment from primates and
mammalian species\textsuperscript{89}. Scores were assigned to each base position in all the reads, including their flanking sequences, and their randomized versions.

\textit{Sequence alignment}

To further analyze HuR targets conservation score result, orthologous sequences were searched for HuR binding reads and their random sets with the same GC content. Sequence alignment data between human and a primate were downloaded from the UCSC genome browser for four primate species: rhesus macaque, orangutan, chimpanzee, and marmoset\textsuperscript{78,90-92}. These four human-primate alignments were combined into one multiple alignment dataset, adding "-" for missing bases.

\textit{Binding motifs comparison}

HOMER (Hypergeometric Optimization of Motif EnRichment) was executed to visually determine similarity between the motifs of the human HuR binding reads and the orthologous sequences in the four primates\textsuperscript{93}. This tool was implemented specifically for next-generation sequencing analysis to discover RNA or DNA motifs by using hypergeometric enrichment, based on the assumption that at most one motif occurs in each sequence. It uses two groups of data to determine the difference in terms of higher enrichment of motifs in one relative to another.
Accessibility calculations for orthologous sequences of human HuR targets

Orthologous sequences to human HuR binding reads and their GC random sets, accessibility was calculated. The parameter settings for RNAplfold were the same as human HuR binding data mentioned in section 2.2.

Insertion and deletion rates calculations

Using the sequence alignments for human and other primates created using the method shown above, number of insertion and deletions of bases were counted and average indels were calculated for each of following regions: upstream sequence flanking sequence, upstream read sequence, binding region, downstream read sequence, and downstream flanking sequence. A base position was determined to have a deletion when other primate(s) have a base where human does not, whereas an insertion occurred when one or more primates are missing a base found in human.

HuR amino acid sequence alignment and protein structure prediction

HuR amino acid sequences for human and the four primates were obtained from the uniprot database\textsuperscript{94}, and were aligned using the multiple sequence alignment program ClustalX\textsuperscript{95}. Using the amino acid sequences, orthologous binding interface structures were predicted using SWISS-MODEL workspace The SWISS-MODEL workspace performs
homology modelling of proteins using known related protein structures as a template\textsuperscript{96}. Human HuR RRM crystal structure from PDB (PDB ID: 4EGL) was used as a template\textsuperscript{97}.

3.7. Protein-protein interaction data

Datasets curation and data combining

HuR/miRNA-mRNA (or its protein product)-protein interaction study has been ongoing. 6 protein-protein interaction (PPI) databases (Biogrid\textsuperscript{98}, DIP\textsuperscript{99}, HPRD\textsuperscript{100}, IntAct\textsuperscript{101}, iRefweb\textsuperscript{102}, and MINT\textsuperscript{103}) were merged together according to the following criteria: 1) The interaction has at least one Pubmed ID of its original data, i.e. it does not much with Pubmed IDs of these 6 PPI databases. 2) The interaction has to be physical or direct.
4. Results

To investigate the relationship between RBP and miRNA binding sites, their binding sites were predicted by applying the following criteria to in vivo mRNA targets of RBPs; accessibility and motif matches. This prediction was refined by confirming expression of miRNAs and RBPs and their predicted target. HuR and miRNA targets were further analysed for their conservation levels and their interplay with interactomes in PPI network (See Materials and Method section for details). The study pipeline is shown as Figure 1 below.

**Figure 1 Study pipeline**
This project was started from HuR and miRNA binding sites mapping followed by identification of mutual targets, conservation analysis and integration of expression/PPI data.
4.1. RBP and MicroRNA binding sites mapping

**RBP binding site mapping from PAR-CLIP data**

HuR and miRNA binding sites prediction was carried out focusing on their preferences on single-stranded RNA and sequence motifs. These preferences were analyzed on *in vivo* HuR and AGO binding data\(^{76,77}\). PAR-CLIP was chosen as reliable assay. This assay stands out from other RNP-coIP assay combined with high-throuput sequencing (RIP-seq) methods by providing marker to identify the T to C mutated cross-linked site by incorporating 4-thiouridine to RNA sequences\(^76\). Prior to starting the analysis, as negative controls, two types of random sequences were prepared to compare with RBP binding sets by randomizing bases of reads (base random) and by randomly chosen genomic sequences keeping the same GC content as reads (GC random). The purpose of having similar/same composition in random sets as bound sets was to avoid sequence bias that can be caused when sequence preferences are examined.

Single-strandedness of RBP binding reads and the random sets denoted as accessibility was predicted using RNAplfold that examine local RNA structure within a certain frame size\(^79\). Sequence preferences of HuR and miRNAs on HuR and AGO binding reads were examined using 7-mer scores and Targetscan, respectively. 7-mer scores show levels of affinity of RBPs including HuR to oligonucleotides of 7 nt in length derived from microarray assay designed for RBP binding specificity called RNAcompete\(^82\). Because they directly examined RBP binding to these sequences experimentally, 7-mer scores would be
more reliable for HuR binding sites prediction. AGO is known for its activity with miRNAs to bind to their targets. The fact that miRNAs canonically bind to their targets by base-pairing of their seed sequence enables binding sequences to be predicted more efficiently than predicting the binding motif of AGO itself.

HuR accessibility scores were presented in boxplots for each base position, and showed a peak around the cross-linking site in the middle of PAR-CLIP reads around the cross-linking sites (Figure 2 a-c). No peaks in base and GC random sets supports the correctness of the result as well as a previous study showing higher accessibility at HuR binding sites. Similarly, 7-mer scores were assigned to each 7 nt of reads and random sets, and is shown in boxplots (Figure 2 d-f). 7-mer scores results on binding and random sets showed the similar trend as accessibility, and this consistency between two scores shows the existence of binding sites around the cross-linking sites. Figure 3 shows 3-D plots of accessibility, 7mer score and corresponding base positions for positive and negative sets. As the plots show, high accessibility is located predominantly around the middle of reads with high accessibility in the positive set, whereas no peak of 7mer scores observed in the negative set. Comparing the datasets by Wilcoxon rank sum test, significant sites around the cross-linking site at 12th-31st bases were chosen as possible binding sites of HuR (Figure 4). To further improve HuR target prediction, thresholds were set for both accessibility and 7mer score that were derived by box-plotting all the values separately for each dataset and a higher whisker of random sets as the cut-off point (Figure 5). As a result, 1492 predicted HuR binding sites were identified.
Figure 2 Accessibility and 7mer score peaks around cross-linking site on HuR binding reads
All the measured accessibility and 7-mer scores from all the HuR binding reads were plotted into boxplot at each corresponding base position. Left is from positive sets having crosslinking site at 0th base position with read sequences located from -19th to 20th (Red bars). (a)(d) HuR binding sets. Both accessibility and 7-mer scores show same distributions with a peak around cross-linking sites. (b)(c) Random sets with the same GC contents as binding sets. (e)(f) Random sets prepared by shuffling the binding set sequences.
Figure 3 3-D plots of 7mer score, accessibility and read position for HuR
HuR binding set data on the top and random set with the same GC content on the bottom. Colour bar shows the degree of 7mer score. (a) High 7mer score with warm colours are located in the middle of the reads, especially at the higher accessibility region, in positive sets. (b) 7mer score is almost entirely low regardless position and accessibility in negative sets.

3D plot of accessibility and 7mer score at each base position for reads and surrounding sequences

(a)

(b)
Figure 4 Accessibility and 7mer scores are significantly higher at HuR crosslinking sites

Wilcoxon rank sum tests were examined between HuR binding set and two types of its random sets. P-values were log transformed and plotted for each base position. The gaps in the middle of plots are due to very low p-values.

Log transformation of Wilcoxon rank sum test between original and random sets

(a) HuR Accessibility (gc)
(b) HuR Accessibility (base)
(c) HuR 7mer score (gc)
(d) HuR 7mer score (base)
Figure 5 Determination of cut-offs of HuR target reads from comparison of boxplots
The cut-offs were determined by choosing higher score at whiskers of boxplots of negative sets, shown in a red broken lines.
AGO-miRNA binding site mapping was done in a similar manner as HuR binding site mapping by calculating accessibility of AGO binding reads from PAR-CLIP assay\textsuperscript{104}. Figure 6 shows boxplots of accessibility scores at each base position of AGO binding reads and the flanking sequences. Although AGO binding sets have higher accessibility around cross-linking sites like it was seen in HuR, two peaks were detected at the 8\textsuperscript{th}-27\textsuperscript{th} and 26\textsuperscript{th}-30\textsuperscript{th} positions that also showed significance by Wilcoxon rank sum test (Figure 7). As AGO binding to its target initially occur at PAZ and MID with small RNA as a guide sequence\textsuperscript{20}, these two peaks are possibly binding sites for these domains. Threshold of accessibility scores for AGO was derived in the same way as that of HuR by boxplotting all scores together (Figure 8). To refine this AGO binding site prediction, miRNA target site prediction was done on its binding reads by TargetScan\textsuperscript{65}. A large amount of predicted targets were located at the position starting at 22\textsuperscript{nd} bases (Figure 9), which matches highly accessible sites. Therefore, those AGO binding reads with both high accessibility and miRNA binding sites predicted at the 8\textsuperscript{th}-30\textsuperscript{th} position of reads were chosen as predicted miRNA/AGO targets, the resultant number being 3312.
Figure 6 Accessibility peaks around cross-linking site on AGO binding reads
All the measured accessibility scores of all the AGO binding reads were plotted into boxplot at each corresponding base position. Crosslinking site is located at 0th base position with read sequences located from -20th to 20th (Red bars). (a) AGO binding sets. (b) Random set with the same GC contents as binding sets. (c) Random set prepared by shuffling the binding set sequences. In positive datasets, there are two peaks which presumably binding sites for PAZ and MID domains of AGO.

Boxplots of accessibility on AGO targets
**Figure 7 Accessibility is significantly higher at AGO crosslinking sites**
Wilcoxon rank sum tests were examined between AGO binding set and two types of its random sets. P-values were log transformed and plotted for each base position.

**Figure 8 Determination of cut-offs of AGO target reads from comparison of boxplots**
The cut-offs were determined by choosing higher score at whiskers of boxplots of negative sets, shown in a red broken lines.
Figure 9 Number of miRNA binding sites have similar distribution as accessibility
Sum of numbers of predicted miRNA binding sites at each base position on all the reads were calculated and plotted. Predicted binding sites are concentrated around the cross-linking site in the middle of reads, which is in between two accessibility peaks in binding set. The red bar denotes location of read sequences.

4.2. Cross-platform normalization and expression data integration

RNA-seq data cross-platform normalization

For the purpose of refining HuR and miRNA binding sites predictions by confirming target expression in HEK293, data curation on mRNAseq and small RNAseq studies was carried out Another purpose of this expression data analysis is to see RNAseq data from
different platforms can be normalized to be compared each other. Among 29 and 103 available samples for small RNAseq and mRNAseq data of Illumina’s platforms, the samples chosen were mainly cell line expression data, which would be expected to be more comparable between miRNA and mRNA expression data over tissue or disease data. 6 cell lines (HEK293, HEK293T, HepG2, K-562, MCF-7 and MDA-MB-231) are available for both small RNAseq and RNAseq data (Table 1). Two types of normalizations (percentile rank and quantile) on all the data set were examined after obtaining FPKM expression for both miRNAs and mRNAs. The Pearson correlation between same samples from different platforms was then checked to see how normalization works. The samples chosen to check for correlation were MCF-7 and HEK293(T) for miRNA and mRNA expression data, respectively. The reason for comparing HEK293 and HEK293T for mRNA data was due to the absence of same cell line data for any cell line type and these embryonic kidney cell lines were expected to have the most similar expression profiles compared to other cell lines. Scatter plots for these miRNA and mRNA expression data are in Figure 10 and 11, and the results of Pearson correlation are shown in Table 2. As shown in the table, percentile rank normalization improves the correlations very well whereas quantile normalization degrades. The lower correlation for miRNA data relative to mRNA data can be reasoned by different methods of small RNA extraction from cells used. However, it should be noted that Pearson correlation of 0.7 for percentile ranked data would be considerably good. Because of property of percentile rank normalization being just a ranking of expressed genes, the relationships between genes in terms of relative expression level to each other would not be modified. Therefore, this normalization method can be considered very useful for RNA-seq data, and it would be interesting to examine this with platforms from different companies.
RNA-seq data were curated for both small RNAs and mRNAs. Data were searched among different platforms, however it was limited to Illumina platforms for more consistency in data.

<table>
<thead>
<tr>
<th>miRNA expression profiles from small RNA-seq</th>
<th>mRNA expression profiles from RNA-seq</th>
</tr>
</thead>
<tbody>
<tr>
<td>HEK293T</td>
<td>Illumina Genome Analyzer</td>
</tr>
<tr>
<td>MCF-7</td>
<td>Illumina Genome Analyzer II</td>
</tr>
<tr>
<td>HEK293</td>
<td>Illumina Genome Analyzer</td>
</tr>
<tr>
<td>HeLa</td>
<td>Illumina Genome Analyzer</td>
</tr>
<tr>
<td>MDA-MB-231</td>
<td>Illumina Genome Analyzer</td>
</tr>
<tr>
<td>HepG2</td>
<td>Illumina Genome Analyzer II</td>
</tr>
<tr>
<td>K-562</td>
<td>Illumina Genome Analyzer</td>
</tr>
</tbody>
</table>
Figure 10 Percentile rank shows high significance on cross-platform normalization (small RNAseq)
MCF-7 expression data from small RNAseq from two Illumina platforms are shown in scatter plots with/without normalization. RHO and P-value are from Pearson correlation. Percentile rank normalization improves both distribution of plot and correlation for small RNAseq data.

Plot between two MCF-7 miRNA expressions from different studies
Figure 11 Percentile rank shows high significance on cross-platform normalization (RNAseq)
HEK293 and HEK293T expression data from small RNAseq from two Illumina platforms are shown
in scatter plots with/without normalization. RHO and P-value are from Pearson correlation. Percentile
rank normalization improves both distribution of plot and correlation for RNAseq data.

Plot between HEK293 and HEK293T mRNA expressions from different studies

Table 2 Pearson correlation of same/similar-cell line expression data
Pearson correlation coefficients were calculated between two similar cell line expression data from
different RNAseq platforms to examine comparability of data. Percentile rank normalization seems to
improve correlation between two datasets.

<table>
<thead>
<tr>
<th>Samples</th>
<th>Platforms</th>
<th>Normalization</th>
<th>PHO</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCF-7 miRNA expression</td>
<td>Illumina Genome Analyzer II vs</td>
<td>FPKM</td>
<td>0.5061</td>
</tr>
<tr>
<td></td>
<td>Illumina Genome Analyzer</td>
<td>Quantile</td>
<td>0.4084</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Percentile</td>
<td>0.7222</td>
</tr>
<tr>
<td>HEK293/HEK293T mRNA expression</td>
<td>Illumina Genome Analyzer vs</td>
<td>FPKM</td>
<td>0.9179</td>
</tr>
<tr>
<td></td>
<td>Illumina Genome Analyzer</td>
<td>Quantile</td>
<td>0.7688</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Percentile</td>
<td>0.9301</td>
</tr>
</tbody>
</table>
4.3. Integration of expression data and common targets of HuR and microRNA

Using percentile rank normalized miRNA-seq and mRNA-seq expression data of HEK293 cells, expression of mRNAs and miRNAs that have HuR and/or miRNAs binding sites predicted in section 4.1 was confirmed to have further precise identification of the predictions. As a result, a total of 1105 and 1861 reads were left as high confidence HuR and miRNA targets, respectively (Figure 12). Among them, 227 transcripts are predictedly bound by both of them. To determine, if these 227 transcripts are not mutually targets by random chances, hypergeometric test was carried out, in which the resultant p-value 7.271932e-17 shows not. The boxplots in Figure 13 show the expression levels of targets by HuR only, AGO only and both. As expected, the median expression of HuR only targets was higher than that of AGO only. However, mutual targets had a median slightly higher than HuR only targets. For mutual targets, gene symbols and description of each gene were checked on a webpage called GeneCards, and many of them were RAS-related or zinc finger-related genes. DAVID functional analysis showed the results of involvement of HuR and miRNA targets in protein complexes and transcription (Table 3).
After determination of accessibility and sequence preferences of HuR and miRNAs, number of predicted targets resulted in 1105 and 1861. 227 transcripts were found to be their mutual targets.
Figure 13 Boxplot HuR and AGO target expressions

HuR/AGO target expressions were plotted into boxplot for HuR only, AGO only and both targeted. Median (in red line) of HuR is higher than that of AGO as expected, however, targets of both shows higher median comparing to HuR.

Table 3 Results of DAVID functional analysis

Mutual targets of HuR and miRNAs were examined on the DAVID functional analysis platform and top two most enriched GO clusters are protein complex and transcription.

Results of DAVID functional analysis

Top 2 clustered Gene Ontology (GO)
4.4. Conservation of target sites

Conservation score distribution of miRNA and HuR target sites

To examine the importance of HuR and miRNA binding sites by checking conservation, conservation scores were sought for each base of reads. These scores were downloaded from the UCSC genome browser and were calculated using phastCons. The software calculates the score based on the hidden Markov model to identify evolutionary conserved elements in a multiple alignment, and the scores calculated by phastCons are noted in the range of 0-1. Average conservation scores from primate genome alignment at each base position of AGO and HuR binding reads after accessibility and 7mer/targetscan cutoff were plotted in Figure 14. From its broad expression across tissues, HuR can be expected to play important roles in post-transcriptional regulation on its targets although conservation of HuR binding sites has not been investigated previously. Unexpectedly, HuR binding sites show lower conservation levels when comparing to its flanking sequences. The scales of conservation scores on HuR binding reads were lower in vertebrates and mammals compared to primates. Having a peak around predicted AGO binding regions in the distribution of conservation among its binding reads was as expected from previous reports of conserved miRNA and AGO binding sites. Also, both GC random sets of HuR and AGO showed no distinct peaks/trend in sequences. Therefore, possibilities of any errors in the scores or processes of assigning these scores to reads would be excluded. These conservation scores were plotted again after narrowing the number of HuR and AGO binding reads after another cutoff based on whether miRNA/HuR and their targets are expressed or
not (Figure 15). However, there was no change in distributions. Therefore, it is highly possible that HuR binding sites are not well conserved.

**Figure 14 HuR has lower conservation scores around binding regions, being opposed to AGO**

Average conservation score at each base position was plotted for HuR and AGO binding reads and their flanking sequences. Red bars represent the read and triangle shows the location of crosslinking site.

Plot of average conservation score at each base position before cutoff with expression data
Figure 15 Average conservation score plot after expression data cutoff

Average conservation score at each base position was plotted for HuR and AGO binding reads and their flanking sequences after removing reads that are not expressed based on HEK293 RNAseq data. Red bars represent the read and triangle shows the location of crosslinking site. Distribution does not change too much from figure 6, which implies that HuR binding sites are not well conserved, whereas AGO binding site shows higher conservation around the crosslinking sites.

Insights to HuR binding sites conservation in primate species

To dissect the low conservation of HuR binding sites, further study on orthologous sequences of HuR and its target sites was done in four primate species (Pan troglodytes, Macaca mulatta, Pongo pygmaeus abelii, Callithrix jacchus). Amino acid sequences of HuR ortholog ELAVL1 in these four species were obtained from PROSITE to confirm the similarities in sequence and predicted binding interface structure using clustalx and
SWISS-MODEL\textsuperscript{110,111}, respectively. By confirming the similarity in domain and amino acid sequences, it was expected to examine if HuR in other primate species has ability to bind its targets with similar preferences as human HuR. Except for the marmoset – the most distant organism from human among those four species – all showed very high similarities (84.24% for marmoset, 98.78% for rest of species) with human in terms of their domain structures of RRM 1 and 2 that are predicted from the amino acid sequences using known human domain structures as template (Figure 16), implying the similarity in targeting profiles and motifs. Having said that, accessibility and 7-mer scores were checked at orthologous regions of predicted human HuR target sequences to examine whether primates HuR has similar binding feature as human HuR. Figures 17 and 18 shows very similar distributions of both accessibility and 7-mer scores, which indicate that these sequences have the potential to be targeted by HuR as well. To further investigate these orthologous sequences to take a closer look at differences between HuR and miRNA target sequences of human and other primate species, all the sequences were aligned along human HuR and miRNA binding read sequences to calculate insertion and deletion rates. As it can be seen in Figures 19, miRNA binding regions had low indel rates, while HuR had higher indel rates around its binding regions, thus explaining the lower conservation of HuR target sites. Furthermore, binding motifs were sought in these orthologous sequences using the motif discovery tool HOMER\textsuperscript{93}. Like those seen in predicted human motifs (Figure 20), the AU-rich motif of HuR was also seen in orthologous sequences with a shorter length that can be explained by higher indel rates in the HuR binding sequences.
Figure 16 HuR RNA recognition motifs are highly conserved in primate species

Structures of RRM1 and 2 were predicted using SWISS-MODEL for four primate species based on homology with known human HuR structures.
Figure 17 Orthologous sequences share similar accessibility distributions as human HuR targets
All the measured marmoset accessibility scores were dumped into boxplot at each corresponding human target base position. Crosslinking site is at 0th base position with read sequences located from -19th to 20th. Orthologous sequences all four primate species showed very similar results as human HuR binding reads.
Figure 18 Orthologous sequences share similar 7mer scores as human HuR targets
The boxplots show high 7mer score at the orthologous sequences of predicted human HuR binding sites, indicating that enrichment of A and U at binding sites is conserved in primates species.
Figure 19 No. of insertions associated with lower conservation of HuR binding sites
While AGO binding keeps similar indel levels as other regions of read sequences, HuR has higher
indel levels at their binding regions. Especially, number of insertions on HuR binding reads is higher
comparing to deletions.

Boxplots of number of deletions at different position in target reads
and their flanking sequences

Boxplots of number of insertions at different position in target reads
and their flanking sequences
**Figure 20 HuR binding motif lengths are species specific**
Motifs were predicted for human HuR targets and their orthologs using Homer. As it can be expected from higher insertion rates in binding regions (Figure 23 and 24), human HuR targets and their closest chimpanzee orthologs have longer motifs than the others.

Motifs from predicted human HuR targets and their orthologous sequences in primates

![Motifs](image)

- **Human**
- **Chimpanzee**
- **Orangutan**
- **Macaque**
- **Marmoset**

*Conservation of miRNA and HuR binding sites in close proximity*

To determine the relationship between HuR and miRNA binding sites in terms of distances between them and evolution, using phastCons conservation scores assigned for HuR and miRNA binding reads, score distributions of mutual targets of these RBPs were checked. HuR and miRNA binding reads were divided into three groups according to the following types: (1) a HuR binding read that overlaps within 30 nucleotides up- and down-
stream of AGO binding reads; (2) HuR and miRNA binding reads are located in the same gene, excluding case 1; (3) everything else. As Figure 21 shows, although the conservation score distributions stay the same as overall datasets, scales are higher in case 1 than cases 2 and 3 for both HuR and miRNA target sites. In fact, case 2 also has a higher scale than case 3. This possibly indicates relationship between HuR and miRNAs during evolution of their target sites.

Figure 21 HuR and miRNA binding sites in close proximity are more conserved
Predicted HuR and miRNA binding reads were divided in three groups according to their different proximity; 1) miRNA and HuR binding reads overlaps/nearby(within 30nt); 2) miRNA and HuR binding reads are on the same genes; 3) miRNA and HuR binding reads are on different genes.
4.5. Integration of PPI network with binding sites prediction/expression data

To study the impact of HuR and AGO on overall PPI network, especially on their targets and their interactoms, HuR/miRNA-mRNA (or its protein products)-protein interaction study was attempted. Because HuR and miRNA target mRNAs were identified using HEK293, cell line specific targets were determined having this identification as bottleneck. To begin with, 6 human PPI databases (Biogrid, DIP, HPRD, IntAct, iRefweb, MINT) were merged together to create as precise and large database as possible to prevent any loss of HuR/miRNA targets. As a result, there are 82,748 interactions in the merged PPI data.

Using this merged PPI data, tissue-specific PPI data were obtained based on existence of expressions from RNAseq data that were processed in section 3.2 (HEK293, HEK293T, HepG2, K-562, MCF-7 and MDA-MB-231). These tissue specific PPI network were visualized using Cytoscape. Nodes were coloured according to their expression levels and grouped according to how they are targeted by HuR/AGO (AGO only, HuR only, both, or none, Figure 22). However, it was difficult to see the difference in colours of nodes for each group.
Figure 22 Selected Cytoscape visualized tissue specific PPIs
Each tissue specific PPIs according to mRNAseq data were visualized using Cytoscape. Green and Orange colours show lower and higher expression levels, respectively. PPIs were grouped into 4 groups depending on if they are targeted by HuR, AGO, both, or none of them.
To investigate more about this slight effect on PPI network, siRNA HuR knockdown exon array data and its mock sample data from HEK293 were obtained from an online resource deposited by Mukherjee et al. The different colours in Figure 23 represent p-values resulted from t-test between log2 transformed expression levels of siRNA HuR and control mock data. The colour gradient towards green is set to start p-value <0.05 so that it is easier to visually recognize which proteins are significantly differentially expressed. However, the portion of significant proteins looks scattered equally in all the group (AGO targets, HuR targets, mutual targets, none), and about 20% of each cluster has P<0.05 (Table 4). Overall, this indistinguishable difference on HuR and miRNA targets might explain their roll as a fine tuner of gene expression in post-transcriptional regulation.

Figure 23 Human PPI network overlaid with HEK293 HuR knockdown data
Combined human PPI data was overlaid with result of t-test from comparison between siRNA HuR and its mock expression data, and visualised by Cytoscape. The colour gradient towards green starts from p<0.05, and that towards red is P-value>0.05.
Table 4 Portion of significantly expressed genes is similar in all targeting states
Log2 transformed expression levels of siRNA HuR and its mock data were compared by t-test. Overall, there is small difference in portion of significant and non-significant genes in any group.

<table>
<thead>
<tr>
<th></th>
<th>P-value &lt;0.05 (t-test)</th>
<th>P-value &gt;0.05 (t-test)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGO targets</td>
<td>248 (21.14%)</td>
<td>925 (78.86%)</td>
<td>1173</td>
</tr>
<tr>
<td>HuR targets</td>
<td>125 (18.88%)</td>
<td>537 (81.12%)</td>
<td>662</td>
</tr>
<tr>
<td>Mutual targets</td>
<td>41 (18.06%)</td>
<td>186 (81.94%)</td>
<td>227</td>
</tr>
<tr>
<td>Not targeted</td>
<td>2720 (19.55%)</td>
<td>11191 (80.45%)</td>
<td>13911</td>
</tr>
</tbody>
</table>
5. Discussion and Conclusion

5.1. Comprehensive RBP binding site prediction

The method used in this study for RBP binding site prediction was inspired by the idea to focus on multiple binding properties. Two focuses prior to RBP binding site mapping were sequence and structure preferences of RNA. In the case of HuR, it is known to bind to AU or U-rich sequences on the single-stranded 3’UTR of mRNAs\textsuperscript{25,26,46}. In fact, although there have been many studies that reported HuR motifs, most of them only focus on a single binding property that drives them. Gao \textit{et al.} studied binding motifs of the human ELAVL protein family, based on the known U-rich binding sequences on \textit{c-fos} and \textit{c-myc}\textsuperscript{113}. Their attempt at HuR motif discovery is one of the earliest studies, and hence was biased towards known binding sequences seen in a number of mRNAs. In 2004, two studies derived an HuR motif from the secondary structure of HuR binding sites through an HuR-mRNAs immunoprecipitation assay\textsuperscript{46,114}. Although these two studies showed one of the most comprehensively examined HuR binding features based on genome-wide microarray data, it may be slightly inaccurate to determine motif based solely on structural preferences, as this RNP-coIP assay would still contain a significant amount of noise compared to the assay that uses cross-linking technology (CLIP assay). Therefore, in addition to the two focuses mentioned above, the data from a type of CLIP-seq—PAR-CLIP, which involves incorporation of 4-thiouridine into RNA sequences and makes coIP more accurate by causing T to C mutations at cross-linking sites\textsuperscript{77,104}—was employed in this study.
With the invention of RNAcompete assay technology, which focuses on finding and scoring sequence preference of RBPs using specific microarrays designed to examine all possible 7-mer sequences, RBP binding activity can be more accurately determined compared to conventional motif discovery tools. The 7-mer scores would be particularly useful for HuR binding site discovery because HuR binding motif length can vary, considering variations in length were seen in motifs listed in RBPDB. As in Figure 3, having a peak of these 7-mer scores around the middle of HuR binding reads was also seen in the study of Kishore et al. In addition to sequence preferences, single-stranded regions of mRNA sequences were determined using RNAplfold, which had previously shown its value in RBP binding site prediction. Although accessibility by RNAplfold has not been scored on high-throughput sequencing data before, the fact that it shows very similar distribution to 7-mer scores supports its power in HuR binding site prediction, and also implies the synergistic effect between the two scores.

In the same way as HuR, AGO/miRNA binding sites were predicted by focusing on sequence and structure preferences. The importance of site accessibility had been discussed previously by Kertsz et al., and RNAplfold was also used to find single stranded regions on AGO binding PAR-CLIP reads. Due to its role with miRNAs on target binding in RISC, AGO was assumed to have the same binding sequence preference as miRNA, which was predicted by TargetScan. Among several other tools, TargetScan would be expected to predict miRNA targets more comprehensively, having functions to examine site context. As was similarly seen in HuR targets sites, AGO/miRNA binding reads also showed peaks of both accessibility and number of predicted miRNA binding sites.
Having confirmed the consistency between distributions of different scores calculated for PAR-CLIP reads, the efficacy of accessibility, 7-mer scores and TargetScan on RBP and miRNA binding sites has been indicated. At the same time, the approach to RBP binding site mapping by looking into sequence and structure preferences is possibly considered to predict RBP binding sites effectively. However, because all the thresholds were set as strictly as possible by setting three types of threshold on single-strandedness, sequence preferences and expression, there could be some false negatives. Tuning of these thresholds would make it possible to predict the targets more accurately and hence provide larger number of predicted targets.

5.2. Expression data integration refines binding site prediction

As one aspect of this study, cross-platform normalization of expression data from different studies and platforms was another challenge. With a large amount of expression data available from online resources, one can expect there to be high demand of integrated cross-platform normalization techniques dealing with noise from differences in experimental environments or platforms, and efforts have already been made in previous studies to address this\textsuperscript{117-119}. Prior to further study into predicted HuR and AGO/miRNA binding sites and how these proteins or miRNAs expression level influences their target, large amounts of microarray and RNA-seq data on mRNAs and miRNAs were curated from online databases. The first attempt at cross-platform normalization was done with microarray data. I decided to make this process as simple as possible to keep better track of and avoid possible
modifications resulting from normalization, which would also help the analysis stay realistic. Z-score normalization was chosen to test whole microarray datasets altogether in a second round of normalization after the conventional normalization of each dataset. Reasons for this choice was because it allows for some missing values coming from a different list of annotations depending on platforms, and also because it preserves the range of data. However, this simple normalization can only overcome differences from experimental environment and platforms from the same company, not platforms from different companies. The cause of this possibly is due to more noise from microarray data than the range z-score normalization can handle.

Reflecting on the result of microarray data normalization, cross-platform normalization was tested on RNA-seq data of various Illumina platforms. After normal FPKM normalization of each dataset, both quantile and percentile normalizations were examined. The results show clear improvement in correlation between the same cell line expression data from different studies and platforms after percentile normalization, whereas lowering of correlation was seen after quantile normalization. Furthermore, because percentile normalization is calculated per sample, it keeps internal relationship between gene expressions in each sample and still provides the same scale between different samples. I therefore propose percentile normalization as a novel and yet very simple technique to normalize cross-platform RNA-seq data with high efficacy. Although Illumina platforms were used for the purpose of this study, this cross-platform normalization could be further improved and useful if it is optimised for platforms from different companies.
5.3. HuR and AGO/miRNA mediated targets post-transcriptional regulation

This study explored the post-transcriptional regulation mediated by RBPs and miRNAs at 3’UTR of mRNAs from various aspects including conservation, PPI and gene ontology (GO) analysis based on the novel prediction methods of binding sites as shown above. In previous studies, HuR and miRNAs have been shown to have their activities broadly across tissues\textsuperscript{107,120}. The results of PPI study and GO analyses on their overall targets suggested their activity as global regulators rather than their role in specific gene sets or pathways. However, conservation and comparative genomics studies on HuR targets had a rather conflicting result with respect to its cross tissue activities, which is its low conservation at its binding sites relative to the flanking sequences. Although miRNAs have also been shown to have broad activities in different tissues, their binding sites are generally known to be conserved, which was also suggested from the result in section 3.4. All the results from accessibility, 7-mer scores and domain structure similarity showed similarity in binding interfaces and sequence preferences of HuR in primates species. However, alignment of orthologous sequences of primates to human HuR binding reads showed high insertion and deletion rates around binding regions and this affects the change in motif length although it stays U-rich in all the species considered. This phenomenon suggests two possible reasons for it. The first possibility is that HuR does not have a strict preference on its motif length, which can be supported by various lengths reported from previous studies listed in RBPDB\textsuperscript{115}. However, because conservation scores calculated between mammals and vertebrates showed lower scales compared to primates, it is also possible that constant changes in HuR binding sites occured in different species. Nonetheless, HuR’s broad activity
in various tissue types implies importance of its targets sites. The relationship between its binding preference and evolution of HuR target sites requires further investigation.

Furthermore, conservation scores also suggested the possibility of co-evolution of HuR and miRNA binding sites located in close proximity. Compared to those binding sites that are distant from each other or are not located in the same gene, binding regions of both HuR and AGO/miRNA showed higher conservation. Grimson et al. previously showed that high local AU-rich context influences positively on miRNA targeting specificity\(^4^7\). This effect of AU-rich context on miRNA targeting and high conservation of HuR and AGO binding sites closely located to each other would suggest HuR’s involvement on miRNA regulation, which was also previously indicated in a study on HuR-dependent loading of miRNA RISC upon UV exposure\(^6^8\). The potential cooperative regulation of HuR and miRNA might also explain the relatively small effect on their targets on PPI network in terms of gene expression level difference before and after HuR knockdown because of the opposite effects of HuR and miRNAs on their targets. However, if post-transcriptional regulation by HuR and miRNAs is controlled upon fine tuning and balance between their opposite effects on their targets, further investigations on target expression levels in patients of diseases including various types of cancers to identify any mis-regulations would be very interesting to be carried out.
5.4. Conclusion

Post-transcriptional regulation is a crucial process to determine gene expression and RNA stability. Especially, 3’UTRs contain various cis-regulatory elements to control the process including RBP and miRNA binding sites. HuR is one of RBPs that binds to AREs to control post-transcriptional events at different stages\textsuperscript{22,23}. Among these functions, a role of HuR as positive regulator of mRNA stability is very interesting by contrast with AGO-miRNA as negative regulator. Recent studies suggested the coordinated mRNA regulations by these two factors can happen in cooperative or competitive manners\textsuperscript{68,69}. The observation of influence on gene expression levels by distance between miRNA and HuR binding sites\textsuperscript{71} implies that their binding sites features are also important factors to determine the relationship between them. Combination of multiple known binding features including accessibility and sequence preferences allowed establishing novel comprehensive mapping method of miRNA and HuR binding sites. Further exploration of these binding sites from evolutionary, expressive and functional aspects suggested possible relationship between HuR and miRNA binding sites that are in close proximity, as well as unusual conservation distribution of HuR binding regions.

Conservation study showed that HuR binding sites have lower conservation level comparing to flanking sequences. Calculations of insertion and deletion rates among primates suggested that higher indel rates caused lower conservation scores at HuR binding sites. Together with various lengths of predicted binding motifs, HuR is considered to have size unspecific binding sites, which is also suggested by Uren et al.\textsuperscript{121}. However, considering
the fact that its broad functions and expression\textsuperscript{107,120}, as well as similarity binding interface between human and other primates, this relatively low specificity of HuR targeting remains to be questioned.

Differences in conservation level depending on the distance between miRNA and HuR binding sites suggest co-evolution of these target sites. This implies global relationship and possible coordinated activities between these two factors. As mentioned in the previous section, this relationship between miRNA and HuR binding sites is supported by the fact that miRNAs functions more efficiently with AU-rich context within 30 nt of its binding sites\textsuperscript{47}. Furthermore, discussing about previous paragraph again, HuR binding sites have high indel rates and low conservation even though the orthologous sequences also have higher accessibility and prefer AU-rich context, which implies these regions are still \textit{cis}-regulatory elements. If HuR has size unspecific binding preference as suggested by Uren \textit{et al.}\textsuperscript{121}, it might be possible to hypothesize that the length of HuR binding sites might affect binding efficiency or outcome of HuR target transcript, which reasons why previous studies reported various coordination between HuR and miRNAs\textsuperscript{68,69}. Finally, together with this possible relationship between HuR and miRNAs, the cooperativity between miRNAs and a \textit{pumilio} (PUM) which also targets AREs in 3’UTR\textsuperscript{122} suggests that further investigation on whether other ARE binding proteins also cooperate with miRNAs would be interesting to be examined.

To conclude, this study described comprehensive HuR and miRNA binding site mapping by focusing their binding preferences on single-strandedness of mRNAs and
sequences on *in vivo* binding data. This mapping data also showed its uniquely low conservation of HuR binding sites and possible co-evolutions between HuR and miRNA binding sites located in close proximity. However, large amount of loss from original RBP binding data after cut-offs from accessibility, motif finding and checking expression of predicted target mRNAs might have caused more false negatives. Tuning of thresholds would allow binding sites to be examined more precisely. After investigation on the low conservation of HuR binding sites, although the results suggested that non-specificity on binding motif length, because of constant changes at the sites implied from the lower conservation on species in phylogenetically more distant relationship with human, inconsistency between HuR as a important regulator and conservation was observed. This possibly is the result of positive selections and more investigations would be required to determine this. Finally, an investigation on relationship between conservation and distances between HuR and miRNA binding sites suggested the possible co-evolution of these sites located nearby. As this suggestion was only based on one observation, more investigations would be required to conclude this result.
6. References


57. Vasudevan, S., Tong, Y. & Steitz, J. A. Switching from repression to activation:


76. Hafner, M., Lianoglou, S., Tuschl, T. & Betel, D. Genome-wide identification of
miRNA targets by PAR-CLIP. *Methods* 1–12 (2012).

doi:10.1016/j.ymeth.2012.08.006


