Interrogating RNA-IP Data to Assess Androgen Receptor Protein Interaction with Long Non-Coding RNA NORAD

Teerapon Sahwangarrom[1], Charlotte Bevan[1*], and Marc Lorentzen[1]

[1]Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, W12 0NN London, United Kingdom
[*]Principle supervisor

ABSTRACT

Background: Prostate cancer is the most common cancer that occurs in men. One of the primary factors of disease progression in the condition is androgens and androgen receptor (AR). Androgens, otherwise referred to as male sex hormones, trigger prostate cancer cells to divide and multiply, whereas ARs are steroid receptors that are essential to gene expression and are activated by binding to androgens. In order to target prostate cancer, pharmacological methods of androgen deprivation therapy have been trialed to date and have aimed to block AR to prevent androgen synthesis. The hormone therapy can suppress prostate cancer, but often cancer can cause mutation of the AR and resultantly adapts to survive without the need to interact with androgens. This phenomenon causes castrate-resistant prostate cancer. Non-coding RNA activated by DNA damage (NORAD), which is highly abundant, conserved long non-coding RNA (lncRNA) has previously been reported to promote the metastasis and cell proliferation of prostate cancer cells. Likewise, the expression of NORAD is altered in the xenograft tissue of castrated mice. Thus, we hypothesized that NORAD expression might be regulated by AR. ***Materials and Methods:*** Three forms of data were used to test the hypothesis: RIP-seq, RNA-seq, and ChIP-seq data. In order to examine sequence alignment, STAR and Salmon aligners were used to analyze RIP-seq and RNA-seq data and the bowtie2 aligner was used to evaluate ChIP-seq data. Differential gene expression analysis using edgeR was performed to investigate the relationship between NORAD and the AR protein. ***Results:*** RIP-seq and RNA-seq analysis demonstrated statistically non-significant FDR results of NORAD. Accordingly, the null hypothesis was accepted; there was no significant difference in NORAD gene expression between androgen treatment and control samples. Additionally, no peak was noted within the NORAD transcript for RIP-seq analysis and at NORAD promoter site for ChIP-seq analysis. This indicated that there were no AR binding sites in RNA and DNA sequences of NORAD. Therefore, there is no supporting evidence to suggest that NORAD is associated with AR protein in our analysis. ***Discussion:*** This study and its findings were limited by small numbers of RIP-seq and RNA-seq data that were deposited into the NCBI database, and these were from only one prostate cancer cell line (LNCaP). Two possible causes for the lack of association between NORAD and the AR protein are gene-gene interaction and epigenetics. The gene expression of NORAD may be affected by significantly higher expressed genes. Likewise, epigenetic change may be influenced by several factors, such as histone modification and DNA methylation, which may ultimately influence NORAD gene expression.

BACKGROUND

Androgens and Androgen Receptors (AR) in Prostate Cancer

Prostate cancer is one of the most frequently occurring cancers in men (1). One of the main causes of prostate cancer progression lies in androgens and the Androgen Receptor (AR). Androgens, otherwise known as male sex hormones, facilitate prostate cancer cell multiplication and tumor growth. ARs are cellular protein receptors that are activated by binding androgen hormones, such as dihydrotestosterone (DHT) and testosterone. Once androgens bind to ARs in the cytoplasm, processes are triggered to facilitate nuclear translation, activated gene expression to regulate apoptosis, cell proliferation and invasion of the prostate cancer (2). In order to hinder prostate cancer progression, an assortment of drugs can be offered, including androgen deprivation therapy (ADT), an approach that aims to inhibit the synthesis of androgen hormones or block the AR (3). The hormone therapy is consequently used to suppress prostate cancer. However, resistance to ADT frequently occurs, and mechanisms include AR amplification, AR mutation to increase sensitivity to low levels of circulating androgens, AR activation by non-androgen ligands (4), expression of constitutely-active AR splice variants, epigenetic modifications and modulation of AR cofactors. Resistance to ADT results in progression to castration-resistant prostate cancer (CRPC) (5), for which

treatment options are severely limited. Thus, identification of novel drug targets is highly desirable.

Long Non-Coding RNAs

One form of ribonucleic acids (RNAs) is long non-coding RNAs (lncRNAs), which usually consist of a minimum of 200 nucleotides that are not translated into protein. Similar to mRNAs, lncRNAs typically have a 3' polyadenylated end and a 5' cap. Likewise, lncRNAs can localize to a cell's cytoplasm or nucleus (6). It has been further indicated that lncRNAs are able to interact with RNA, DNA molecules and transcription factors in the facilitation of numerous biological processes, such as DNA damage regulation (7).

Non-Coding RNA Activated by DNA Damage (NORAD)

NORAD, also known as *LINC00657* in humans, is a ubiquitously expressed and highly conserved lncRNA. NORAD transcript is approximately 5kb in length, without overlapping other genes; the structure starts from a single promoter, which overlaps with a CpG island and terminates with a single canonical poly(A) site (8). Lee *et al.* identified that the function of NORAD is to maintain DNA repair, mitosis, and genomic stability by sequestering and repressing PUMILIO (PUM) proteins. NORAD contains up to 17 PUM binding sites, called Pumilio Response Elements (PREs) (9). PUM1 and PUM2 proteins are well-conserved RNA binding proteins that harbor pivotal regulatory roles in transcription, differentiation, and cell development (10). PUM proteins drive chromosomal instability and can regulate DNA replication by binding to the target transcript, which contains PUMILIO response elements (11). Therefore, through the repression of PUM proteins, NORAD acts as a 'defender' of the genome (12).

Moreover, recent studies have documented that NORAD is overexpressed in multiple cancers, such as bladder cancer (13), colorectal cancer (14), pancreatic cancer (15), and prostate cancer (16). NORAD is associated with reduced survival of prostate cancer patients across multiple patient data sets, and it is thought that it may reduce response to DNA damaging therapeutics, such as chemotherapy and PARP inhibitors (C. Fletcher, personal communication). Its potential role in prostate cancer progression and AR signaling is supported by the observation that NORAD expression is significantly increased in patient-derived xenografts following castration (Wei Yuan, Institute of Cancer Research, personal communication).

Accordingly, it was hypothesized that NORAD expression may be regulated by the AR protein. The purpose of this research was to determine whether NORAD interacts with the AR protein and clarify the relationship between NORAD and ARs. In order to progress this aim, three forms of data were retrieved: RNA immunoprecipitation (RIP-seq), RNA-seq, and Chromatin immunoprecipitation (ChIP-seq) data.

Materials and Methods

RNA-binding Protein Immunoprecipitation (RIP) Data Collection, Sequence Alignment, and IGV

RNA immunoprecipitation (RIP-seq) is a sequencing method that is used to identify sites where proteins are bound to the RNA sequence (17). The raw sequence of RIP-seq data was accessed from the National Center for Biotechnology Information (NCBI) database in the GEO repository under accession number GSE100710. Two types of samples were present, anti-androgen receptor (antiAR) and anti-IgG (control); each sample had two replicates. The raw sequence data, which is from LNCaP cells, was assessed for quality using the FASTQC format (18). Subsequently, the data set was trimmed of low-quality reads and clipped of sequencing adapters using trimmomatic (19). Next, the sequence reads were aligned to the hg38 version (GRCh38.p13) of the human reference genome with STAR (20) and Salmon (21) aligners.

When applying the STAR aligner, the sequence reads were mapped to the human genome. A count matrix was associated with the genes and constructed using the featureCounts function, a tool derived from the Subread package (22). Following this, the count matrix was used to perform differential expression gene analysis (Figure 1A).

Once the sequence reads were aligned to the human genome, the Salmon aligner was applied. Transcript-level estimates for RIP-seq data were imported to R programming using the tximport package (23). Following this, the gene count table was used to analyze differential expression gene analysis (Figure 1A).

In order to investigate whether AR protein binding sites are present in the RNA sequence of the NORAD gene, peak calling analysis was conducted using macs2 (24). In this process, the aligned reads between treatment (anti-AR) and control (anti-IgG) samples were compared and regions in the human genome that were enriched with the aligned sequence were identified. The peaks were considered significant when the false discovery rate (FDR) or adjusted P value was equal to or less than 0.05.

RNA-seq Data Collection, Sequence Alignment, and Differential Expression Gene Analysis

RNA sequencing (RNA-seq) is a method used to examine the quantity of RNA in a biological sample using next-generation sequencing (NGS). The raw sequence data for RNA-seq was sourced from NCBI in the GEO repository under accession number GSE102306. Two types of biological samples were used, treated and untreated androgen hormone samples,

with each sample encompassing two replicates. The treated replicate samples that have been treated with 1 nM androgen hormone (R1881) for 24 hours. The raw sequence data, which is from LNCaP cell line, was examined for quality using the FASTQC tool (18). Following this, the data was trimmed of low-quality reads and clipped of sequencing adapters using trimmomatic (19). Subsequently, the sequence reads were aligned to hg38 version (GRCh38.p13) of the human reference genome with STAR and Salmon aligners.

Once the STAR aligner was used, the sequence reads were mapped to the human genome, a count matrix was associated with the genes and produced using the featureCounts tool. Next, the count matrix, with genes, was applied to DEG analysis (Figure 1A).

In case of the Salmon aligner, the newly sequenced reads were applied to transcript-level quantification of the RNA-seq data and imported to R programming using the tximport package. Following this, the count data, with genes, was used to perform DEG analysis (Figure 1A).

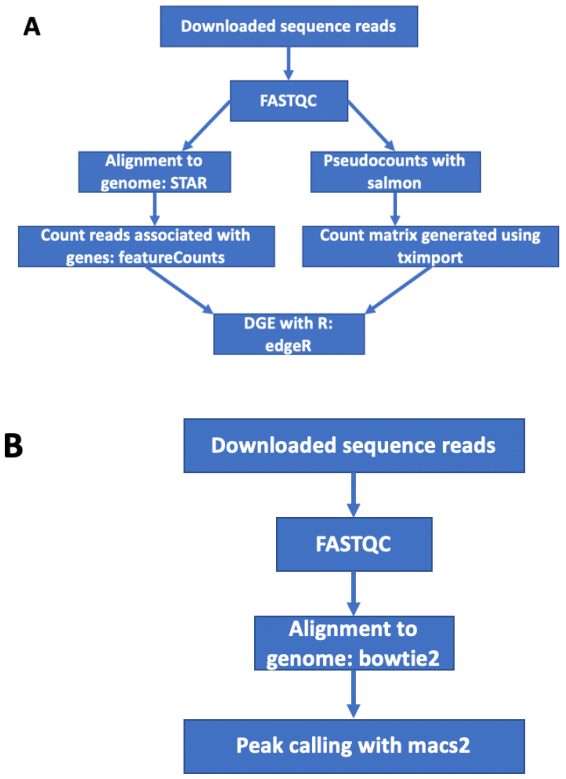

Figure 1. Data analysis pipelines. Three forms of data were used for the analysis: RIP-seq, RNA-seq, and ChIP-seq data. (A) Analysis pipeline for RIP-seq and RNA-seq data. (B) Analysis workflow for ChIP-seq data.

ChIP-seq Data Collection, Sequence Alignment, Peak Calling, and IGV

ChIP-seq is a method that is used to identify protein interactions in a reflection of DNA sequences. The ChIP-seq dataset was accessed from the NCBI database in the GEO repository under accession numbers GSM1069682 and GSM2219854. The ChIP-seq data of GSM1069682 consists of two biological replicate samples that have been treated with 10 nM androgen hormone (DHT) for four hours and two control samples (input). Likewise, the accession number GSM2219854 includes two replicate samples that have been treated with 100 nM androgen hormone (DHT) for two hours and two control samples (input). All of the raw data, which are from LNCaP cell lines, were electronically downloaded using the fastq-dump system. Subsequently, the raw sequence data was analyzed using the FASTQC tool as part of a quality control assessment. The sequence reads were then trimmed and clipped of low-quality reads and adapters using the trimmomatic application. Trimmed reads were mapped to the human reference genome (hg38) using Bowtie2 (25). The peak calling process was performed using macs2 by comparing the aligned reads between treated and untreated androgen hormone samples. This was conducted to identify regions in the human genome that have been enriched with the aligned reads. The peaks were considered significant when FDR < 0.05 (Figure 1B).

Differential Expression Gene Analysis of RIP-seq and RNA-seq Data

When using the RIP-seq data, the two types of condition were classified into anti-AR and anti-IgG. Each condition contained two biological replicates. Anti-AR replicates, the sample has treated with 0.1 nM androgen hormone (R1881) for six hours and AR antibody (specific antibody), were classed as the treatment group. Comparatively, anti-IgG replicates, the sample has been treated with 0.1 nM androgen hormone (R1881) for six hours and rabbit IgG antibody (non-specific antibody), were classed as the control group.

RNA-seq data were collected from both of the two biological replicates in each condition (treated and untreated androgen hormone). The treated androgen hormone replicates (LNCaP cell lines treated with androgen hormone) were categorized as the treatment group. In comparison, the untreated androgen hormone replicates (LNCaP cell lines that were not treated with androgen hormone) were classified as the control group.

In order to determine whether the NORAD gene interacts with AR proteins, differential expression gene analysis was conducted using the edgeR application. EdgeR is underpinned by the concept of a generalized linear model (GLM), which compares how dichotomous groups (treatment and control) affect continuous variables (the count data) and make predictions for determining differential expression. The GLM assumes that the count data is not normally distributed. As the GLM approach offers generous flexibility, it is more popular than the classic

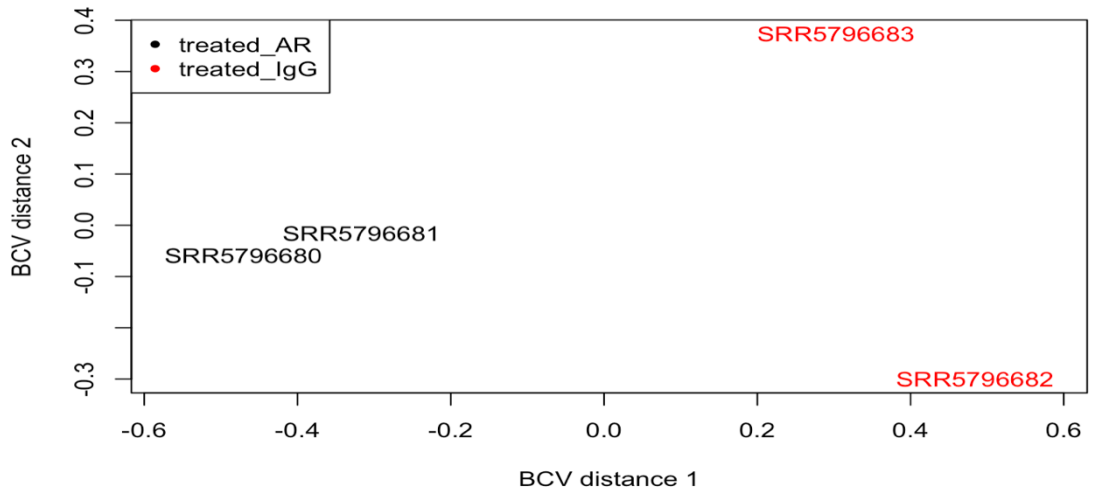

Figure 2. Multidimensional scaling plot of RIP-seq data. BCV Distance 1 (x-axis) separates the treated_AR (black) and treated_IgG (red) samples. BCV Distance 2 (y-axis) corresponds to sample replicates. Treated_AR samples (black) refer to samples that were treated with androgen hormone and AR antibody. However, treated_IgG samples (red) were samples that were treated with androgen hormone and IgG antibody.

method. The likelihood ratio test and quasi-likelihood F-test are two forms of analysis employed in the framework of GLM (26). The quasi-likelihood method was utilized in this research as it provides a more robust and reliable error rate control when the number of sample replicates is small. Prior to edgeR analysis, genes with very low counts were removed; these were classed as genes with less than 100 counts per million (cpm) in at least two samples. Next, the count data was normalized using the trimmed mean of M-values (TMM) method in order to draw fair gene comparisons between samples.

Statistical analysis was conducted using a model matrix that was split into the experimental and control groups. Subsequently, edgeR was used to compare gene expression values between the two conditions using a t-test. A criterion was determined to differentiate significantly expressed genes between two conditions using the ratio of log2 fold change. The latter was classed as the log ratio of expression values between treatment and control samples when it was greater than one whilst considering the FDR values, which were calculated using the Benjamini-Hochberg method to trim false positives that were lower than 0.05.

To provide a visual representation of the differential expression gene analysis for RIP-seq and RNA-seq data, volcano plots were constructed using the Enhanced Volcano package (27) in R programming (28).

Results

RIP-seq Analysis Reveals that the NORAD is not Associated with AR Protein

In order to identify whether any AR binding sites exist in the RNA sequence of the NORAD gene, RIP-seq analysis was conducted. Following the filtering and normalization steps of the count data, we constructed a multidimensional scaling plot to visualize the relationship between samples and replicates (Figure 2). The biological coefficient of variation (BCV) defines the gene variation between RIP samples and their replicates. BCV Distance 1, the treatment (treated_AR) and control (treated_IgG) samples were separated, while BCV Distance 2 separated the biological replicates. The Distance 2 between the two replicates of the control sample (treated_IgG) was greater than between the two replicates of the treatement sample (treated_AR). This indicates that the gene abundance of the control sample (anti-IgG) was more varied than the treatment sample (anti-AR) as a nonspecific antibody (IgG) bound to the cell randomly when compared to a specific antibody (anti-AR).

When using the STAR aligner, a volcano plot was constructed to provide a visual representation of the differential expression analysis (Figure 3A). As shown, each red-colored dot represents significantly differentially-expressed genes between the two conditions. Red dots on the left side of the graph represent RNAs with lower binding to AR than to IgG, whereas red dots on the right side of the plot represent RNAs with higher binding to AR than to IgG. Comparatively, black, green, and blue dots represent genes with no significant expressed based on the specified criteria of FDR < 0.05 and |log2FC| >= 1.0. A total of 493 significantly-differentially AR-bound versus IgG-bound RNAs were identified: 189 showing increased binding to AR versus IgG and 304 showing reduced AR binding versus IgG. Among

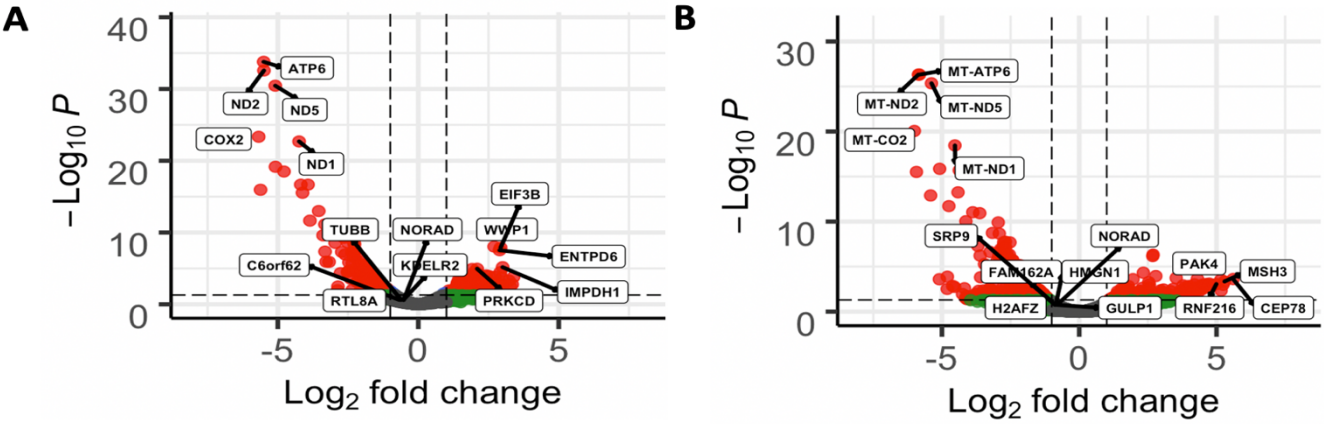

Figure 3. RIP-seq analysis and demonstration of differential gene expression gene (edgeR) using STAR and Salmon aligners. (A) A volcano plot depicting DEG analysis using STAR. The x-axis represents log 2 fold change, while the y-axis portrays negative adjust p-value (FDR). Log2 fold change shows log ratio of gene expression values between treatment and control samples. (B) Volcano plot representation of DEG analysis using Salmon.

the upregulated genes the most significantly expressed gene was WWP1 (FDR = 8.98E-09, logFC = 2.70), while ATP 6 (FDR = 1.63E-34, logFC = 5.50) was the most significant downregulated gene. However, NORAD was not identified as having increased association with AR as compared to IgG control: FDR = 0.25 and logFC = -0.52.

In the Salmon aligner analysis, 686 significantly expressed genes were identified, including 243 upregulated and 443 downregulated genes, (Figure 3B). The most significantly expressed upregulated gene was EIF3B (FDR = 4.71E-07, logFC = 2.69) while MT-ATP6 (FDR = 8.59E-31, logFC = -5.82) was the most significant downregulated gene. Again, NORAD was not identified as having increased association with AR as compared to IgG control: FDR =0.18 and logFC = -0.81.

To demonstrate and identify whether there are AR protein binding sites in the RNA sequence of NORAD, peak calling was performed using macs2 (24). Integrative Genomics Viewer (IGV) (29) was used to plot the peak calling result (Figure 6A). AS shown, the x-axis represents genes of the human genome, whereas the y-axis shows the number of aligned reads that were mapped to the genome. The NORAD gene was located between position 36,045,299 and 36,051,018 bases in chromosome 20 (Figure 6A, red box). If an interation between AR protein and the transcript exists in this data set, a greater number of reads should be identified in the AR antibody condition as compared to the IgG control condition for the given gene (Treated_rep2). Decreased reads in the AR antibody condition compared to the IgG condition may indicate non-specific binding of an RNA to IgG (Treated_rep1).

RNA-seq Analysis Confirms That There is no Relationship Between the NORAD and AR Proteins

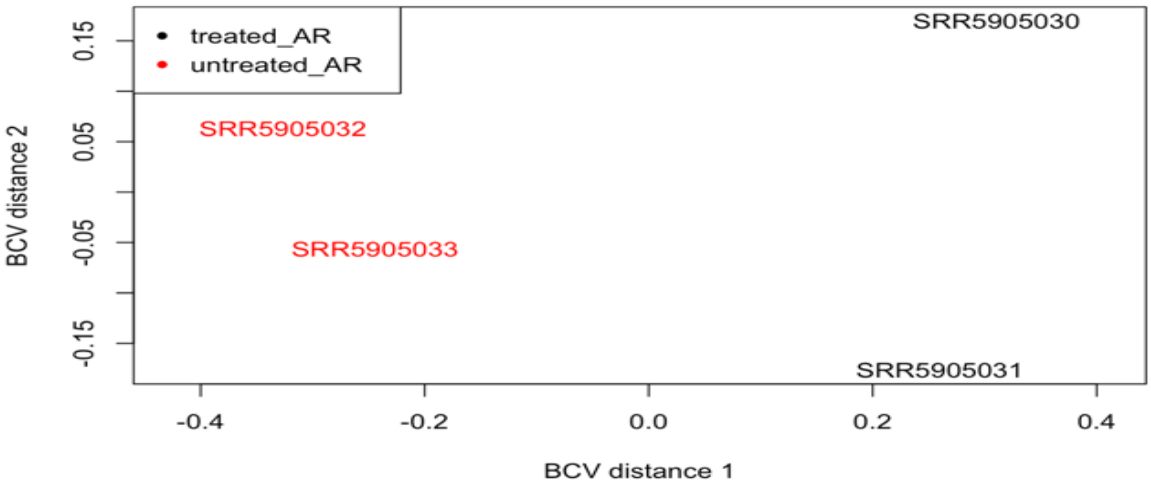

Figure 4. Multidimensional scaling plot of RNA-seq data. BCV Distance 1 (x-axis) separates treated_AR (black) and untreated_AR (red) samples. BCV Distance 2 (y-axis) corresponds to sample replicates. Treated_AR samples (black) correspond to LNCaP cell lines that were treated with androgen hormone, whereas untreated_AR samples (red) were LNCaP cell lines that were not treated with androgen hormone.

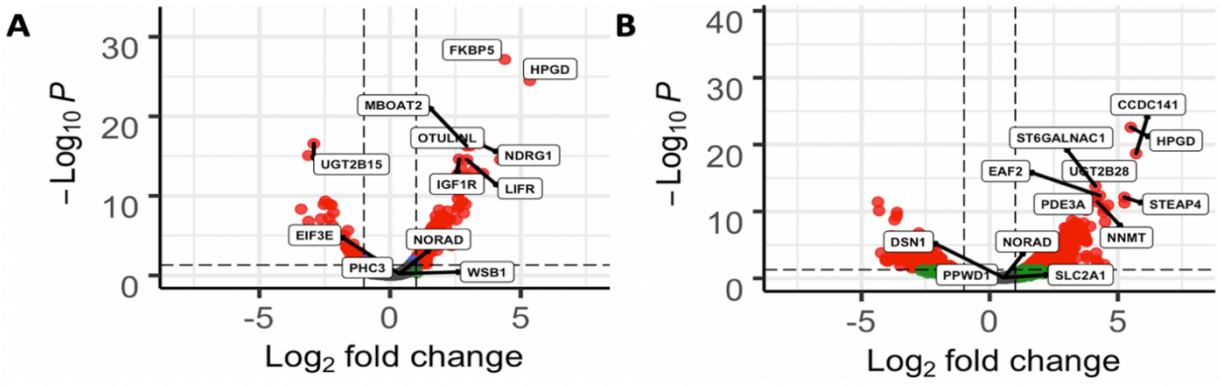

Figure 5. RNA-seq analysis result of differential expression gene analysis (edgeR) using STAR and Salmon aligners. (A) A volcano plot demonstrating the results of the DEG analysis. The x-axis represents log2 fold change, while the y-axis depicts negative log adjusted p-value (FDR). Log2 fold change shows log ratio of gene expression values between treatment and control samples. (B) Volcano plot depeicting DEG analysis using Salmon.

To develop the evidence-base necessary to verify a lack of association between NORAD and AR protein, RNA-seq data was analyzed. The RNA-seq data included two replicates of the treated sample (LNCaP cell lines treated with synthetic androgen hormone, DHT), and two replicates of the control (samples that were not treated with DHT). After trimmed reads were aligned to the human genome (hg38) using STAR and Salmon aligners, edgeR was used to perform in a differential expression gene analysis. Steps of filtering and normalization of the count data were conducted and multidimentional scaling (MDS) plot was developed to represent sample and replicate relationships (Figure 4). BCV Distance 1 in the treatement sample was greater than the control sample.The difference of the Distance 2 between two replicates of the treated sample was higher than the control group.

Following the application of the STAR aligner, a volcano plot was constructed to illustrate the differential expression analysis (Figure 5A). A considerable 217 significantly differentially-expressed genes were identified, including 166 androgen-upregulated and 51 androgen-downregulated genes. The most significantly expressed upregulated gene was FKBP5 (FDR = 7.12E-28, logFC = 4.39), while UGT2B15 (FDR= 2.75E-17, logFC = -2.92) was the most significant among the downregulated genes. However, NORAD was not significantly differentially-expressed between DHT-treated and untreated LNCaP cells: FDR = 0.59 and log FC = 0.36.

When the Salmon aligner was utilized, a total of 598 significantly expressed genes were identified, including 358 upregulated and 240 downregulated genes (Figure 5B). TTN (FDR = 1.28E-34, logFC = 8.83) was the most significantly expressed gene of the upregulated genes, while ACKR3 (FDR = 3.95E-12, logFC = -5.82) was the most significant among the downregulated genes. However, NORAD was not considered significant as FDR = 0.74 and logFC = 0.56.

ChIP-seq Analysis Clarifies the Relationship Between NORAD and the AR Protein

Integrative Genomics Viewer (IGV) was used to visualize the peak calling (Figure 6B). The NORAD gene is located in chromosome 20, between the bases of 36,045,299 to 36,051,018 (Figure 6B, orange box). The first sample replicate of LNCaP cell was treated with 10nM DHT for four hours (LNCaP_DHT) and showed a peak at the promoter region of NORAD (Figure 6B, yellow box). However, other samples depicted no peaks at the promoter area of NORAD.

Discussion

The aim of this research was to investigate the relationship between NORAD and the AR protein. An analysis was performed using RIP-seq, RNA-seq, and ChIP-seq data. As raw sequence data were used, three types of sequence aligner were utilized, including STAR, Salmon, and bowtie2; each were used to align sequence reads to the human reference genome. Severals previous studies had indicated that STAR provides a high accuracy, speed, and efficiency in mapping the sequence reads to the reference genome when compared to other aligners such as Tophat and HISAT, The STAR, a general aligner, uses maximal mappable prefixes (MMPs) in the first stages of seed searching prior to stages of seed clustering and stitching to create complete RIP-seq and RNA-seq reads. Moreover, the STAR detects chimeric transcripts, canonical junctions, and noncanonical splices. This indicates a strong rationale for its use in this study. However, Patro et al. developed the Salmon aligner, which is a method for quantifying RNA expression transcripts accurately and at a faster rate. The Salmon method employs quasi-mapping; this includes mapping the sequence reads to transcriptome in the first step and applying statistical models to quantify the transcript expression levels in the second step. Compared to the STAR method, the Salmon aligner is quicker and is less memory intensive as it uses

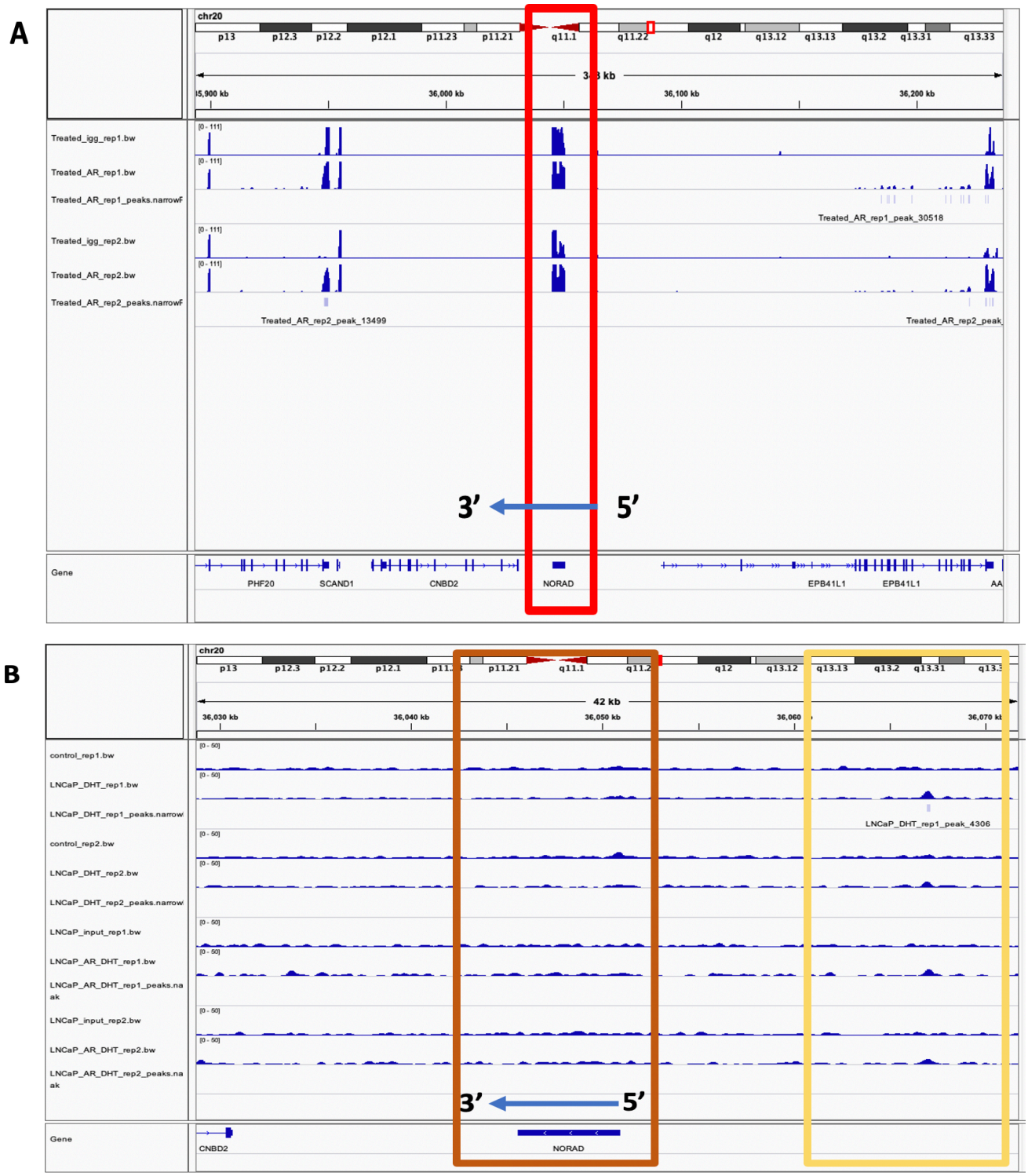

Figure 6. IGV of RIP-seq and RNA-seq data. (A) IGV of RIP-seq data. The x-axis represents the human reference genome (hg38 version), whereas the y-axis represents the quantity of sequence reads from the test samples that match the human genome using STAR alignment. The red box demonstrates the NORAD gene's position on chromosome 20, where it is located from bases 36,045,299 to 36,051,018. Peak calling was performed using macs2, the aligned reads in the treatment (Treated_AR) and control (Treated_igg) samples were compared to identify regions that align with the human genome. Treated_AR_rep1_peaks.narrowPeak and Treated_AR_rep2_peaks.narrowPeak represent peaks used to identify AR binding sites that held an RNA sequence with FDR < 0.05. (B) IGV of ChIP-seq data. Peak calling analysis was conducted using macs2. Mapped reads of treatment (LNCaP_DHT and LNCaP_AR_DHT) and control (control and LNCaP_input) samples were compared in order to identify peak regions that aligned with the human genome. Unlike LNCaP_AR_DHT, LNCaP_DHT samples were treated with 10nM DHT for four hours, whereas LNCaP_AR_DHT samples were treated with 100nM DHT for two hours. LNCaP_DHT_rep1_peaks.narrowPeak, LNCaP_DHT_rep2_peaks.narrowPeak, LNCaP_AR_DHT_rep1_peaks.narrowPeak, and LNCaP_AR_DHT_rep2_peaks.narrowPeak represent peaks that indicate AR binding sities in the DNA sequence when FDR < 0.05. One peak was identified in the LNCaP_DHT_rep1 sample; no other peaks were identified (yellow box)

pseudoalignment. Due to the swift performance and precision of the sequence alignment, the approaches were used and the results compared to differential expression analysis. In the case of ChIP-seq, we used bowtie2, which is an alignment tool that is popular and suitable for ChIP-seq data (30) as the ChIP-seq has normally very short reads and bowtie2 supports sequence reads from a minimum of 50 base pairs.

The RIP-seq analysis demonstrated that the NORAD gene was not associated with the AR protein. Both the STAR and salmon aligners conveyed this trend and NORAD had a negative logFC (anti-AR/anti-IgG). This indicates that NORAD gene expression in the control sample (anti-IgG), was higher than the tratement sample (anti-AR). Furthermore, the FDR value was higher than 0.05, indicating that the null hypothesis should be accepted. By way of explanation, there is no significant difference in NORAD gene expression between the two conditions. In addition, in term of AR protein interaction with NORAD from the RIP-seq data, an interaction within the NORAD transcript was considered. IGV (Figure 6A) showed that there were increased reads in the AR antibody condition (Treated_AR_rep2) compared to the IgG condition (Treated_igg_rep2), whereas increased reads in the IgG condition (Treated_igg_rep1) compared to the AR condition (Treated_AR_rep1) also were identified. These were possibly indicative of nonspecific binding.

In order to verify that no significant association existed between the NORAD gene and AR protein, RNA-seq data was analyzed. The analysis confirmed that there was no association between the two factors. Both the STAR and Salmon aligners depicted a similar trend. NORAD provided a positive logFC (treatment/control), which indicated that gene expression in the treatment sample was greater than in the control sample. The FDR value was more than 0.05, which showed that there was no difference in NORAD gene expression between the treatment and control samples. Moreover, FKBP5 (FDR = 7.12E-28, logFC = 4.39), HPGD (FDR = 3.25E-25, logFC = 5.35), KCNMA1 (FDR = 1.20E-4, logFC = 1.41), SAT (FDR = 7.96E-4, logFC = 1.53), HEBP2 (FDR = 2.63E-3, logFC = 1.14), TPM1 (FDR = 7.03E-3, logFC = 1.02), which were identified in this study, corresponded with previous publicatioin (31). These genes were strongly (FKBP5,HPGD) and moderately (KCNMA1, SAT, HEBP2, TPM1) upregulated by R1881.

ChIP-seq data was analyzed to investigate whether there were AR binding sites in the DNA sequence of NORAD. The result of ChIP-seq analysis showed that there were no AR binding sites in the NORAD's DNA sequence; this was determined as several samples showed no peaks when sample values were compared to the promoter region of the NORAD gene. Therefore, it has been concluded that the NORAD gene was not related to the AR protein.

There are possible causes for the findings of this research and a lack of association between the NORAD and AR proteins; these are gene-gene interaction (epistasis) and epigenetics. Firstly, the NORAD gene expression may be suppressed by some specific highly expressed genes. The relationship between NORAD and significantly expressed genes should hold an suppressive interaction; this would mean that the representation of highly expressed genes is increased by androgen hormone and NORAD gene expression is inhibited. Secondly, the NORAD gene may have been affected by epigenetics. Epigenetic change can be influenced by several factors, including environmental elements that may influence the culture LNCaP cell lines and treat them with synthetic androgen hormone (DHT) at different times and concentrations. In addition, histone modification, DNA methylation, and non-coding RNA are associated with gene silencing and are able to induce epigenetic change. Moreover, the NORAD gene may have interaction with an adaptor protein. The adaptor protein contain protein-binding motifs, which influence signal transduction pathways. These could be why only one sample showed a peak at the NORAD promoter during the ChIP-seq analysis (Figure 6B, LNCaP_DHT rep1). However, other samples did not identify any peaks. An amalgamation of these causes cold underly the lack of association found in this study.

The reliability of the findings in this study could have been improved if the public database had provided a greater number of samples for RIP-seq and RNA-seq data analysis. However, accuracy was encouraged by using different types of aligner, including STAR and Salmon for the analysis of RIP-seq and RNA-seq data. The findings of these two aligners showed a similar trend of differential expression gene analysis, which indicates that the NORAD gene is not associated with the AR protein. Therefore, there is no supporting RIP-seq, RNA-seq, or ChIP-seq analysis evidence that NORAD is related to AR protein in this study. As androgens and ARs are factors associated with prostate cancer progression, NORAD may not be directly involved in the progression. This contradicts the findings of Zhang et al., which indicated that NORAD activates cell proliferation and migration in prostate cancer.

Because this study was limited by a few samples of data set available and these were from one type of prostate cancer cell line (LNCaP), further avenues of research could clarify AR binds to the NORAD promoter, including ChIP-PCR for AR association with NORAD promoter and enhancer regions across a timecourse of androgen treatments and in different prostate cancer cell lines. In addition, ChIP-PCR could be performed in patient tumor tissue from prostatectomy (32).

Acknowledgements

I would like to thank the head of Androgen Signaling and Prostate Cancer laboratory, Professor Charlotte Bevan, for the precious opportunity to conduct research under her guidance. We also wish to express our thanks to all of the laboratory members for their helpful feedback, which has been invaluable to developing and expanding our knowledge. A special thanks is provided to Marc Lorentzen as a day-to-day supervisor for all of his guidance and support during the project.

Reference:

1. C. Hoey *et al.*, Circulating miRNAs as non-invasive biomarkers to predict aggressive prostate cancer after radical prostatectomy. *J Transl Med* **17**, 173 (2019).

2. Z. Culig, F. R. Santer, Androgen receptor signaling in prostate cancer. *Cancer Metastasis Rev* **33**, 413-427 (2014).

3. B. C. Thomas, D. E. Neal, Androgen deprivation treatment in prostate cancer. *BMJ* **346**, e8555 (2013).

4. T. Chandrasekar, J. C. Yang, A. C. Gao, C. P. Evans, Mechanisms of resistance in castration-resistant prostate cancer (CRPC). *Transl Androl Urol* **4**, 365-380 (2015).

5. T. M. Amaral, D. Macedo, I. Fernandes, L. Costa, Castration-resistant prostate cancer: mechanisms, targets, and treatment. *Prostate Cancer* **2012**, 327253 (2012).

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

7. R. Li, H. Zhu, Y. Luo, Understanding the Functions of Long Non-Coding RNAs through Their Higher-Order Structures. *Int J Mol Sci* **17**, (2016).

8. A. Tichon *et al.*, A conserved abundant cytoplasmic long noncoding RNA modulates repression by Pumilio proteins in human cells. *Nat Commun* **7**, 12209 (2016).

9. S. Lee *et al.*, Noncoding RNA NORAD Regulates Genomic Stability by Sequestering PUMILIO Proteins. *Cell* **164**, 69-80 (2016).

10. D. S. Spassov, R. Jurecic, Cloning and comparative sequence analysis of PUM1 and PUM2 genes, human members of the Pumilio family of RNA-binding proteins. *Gene* **299**, 195-204 (2002).

11. E. Zlotorynski, Non-coding RNA: Decoy pumilio for genomic stability. *Nat Rev Mol Cell Biol* **17**, 68 (2016).

12. M. M. Elguindy *et al.*, PUMILIO, but not RBMX, binding is required for regulation of genomic stability by noncoding RNA NORAD. *Elife* **8**, (2019).

13. Q. Li *et al.*, High expression of long noncoding RNA NORAD indicates a poor prognosis and promotes clinical progression and metastasis in bladder cancer. *Urol Oncol* **36**, 310 e315-310 e322 (2018).

14. L. Wang *et al.*, Overexpression of long noncoding RNA NORAD in colorectal cancer associates with tumor progression. *Onco Targets Ther* **11**, 6757-6766 (2018).

15. H. Li *et al.*, Long noncoding RNA NORAD, a novel competing endogenous RNA, enhances the hypoxia-induced epithelial-mesenchymal transition to promote metastasis in pancreatic cancer. *Mol Cancer* **16**, 169 (2017).

16. H. Zhang, H. Guo, Long non-coding RNA NORAD induces cell proliferation and migration in prostate cancer. *J Int Med Res* **47**, 3898-3904 (2019).

17. F. Zambelli, G. Pavesi, RIP-Seq data analysis to determine RNA-protein associations. *Methods Mol Biol* **1269**, 293-303 (2015).

18. S. Andrews, FastQC: a quality control for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc, (2010).

19. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).

20. A. Dobin *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).

21. R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, C. Kingsford, Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**, 417-419 (2017).

22. Y. Liao, G. K. Smyth, W. Shi, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930 (2014).

23. C. Soneson, M. I. Love, M. D. Robinson, Differential analyses for RNA-seq:transcript-level estimates improve gene-level inferences [version 1; peer review:2 approved]. *F1000Research* **4,** 1521 (2015).

24. Y. Zhang *et al.*, Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).

25. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359 (2012).

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

27. K. S. R. Blighe, M. Lewis, EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labelling. Available online at: https://github.com/kevinblighe, (2018).

28. R Core Team, R: A language and environment for statistical computing. Available online at: http://www.R-project.org/, (2013).

29. H. Thorvaldsdottir, J. T. Robinson, J. P. Mesirov, Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**, 178-192 (2013).

30. Q. Zhang *et al.*, Systematic evaluation of the impact of ChIP-seq read designs on genome coverage, peak identification, and allele-specific binding detection. *BMC Bioinformatics* **17**, 96 (2016).

31. S. Ngan *et al.*, Microarray coupled to quantitative RT-PCR analysis of androgen-regulated genes in human LNCaP prostate cancer cells. *Oncogene* **28**, 2051-2063 (2009).

32. A. A. Singh *et al.*, Optimized ChIP-seq method facilitates transcription factor profiling in human tumors. *Life Sci Alliance* **2**, e201800115 (2019).

ABBREVIATIONS

AR, Androgen receptor; NORAD, Non-coding RNA activated by DNA damage; lncRNA, Long non-coding RNA; RIP-seq, RNA immunoprecipitation sequencing; ChIP-seq, Chromatin immunoprecipitation sequencing; STAR, Spliced Transcripts Alignment to a Reference; FDR, False Discovery Rate; logFC, log Fold Change; logCPM, log counts per million; LR, Likelihood ratio; DEG, Differential expression gene analysis; ADT, Androgen deprivation therapy; DHT, Dihydrotestosterone; MMPs, Maximal mappable prefixes; IGV, Integrative Genomics Viewer; DHT, Dihydrotestosterone; MDS, Multidimentional scaling; TMM, Trimmed mean of M-value; GLM, Generalized linear model;, BCV, Biological coefficient of variation; IgG, Immunoglobulin G;, NGS, Next generation sequencing; R1881, Methyltrienolone.