12-2022

Novel Regulators of Human Gene Expression

Alexander Coley

NOVEL REGULATORS OF HUMAN GENE EXPRESSION

A Dissertation

Submitted to the Graduate Faculty of the
University of South Alabama
in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

In

Basic Medical Sciences

by
Alexander Coley
B.S., The University of South Alabama, 2018
December 2022

ACKNOWLEDGEMENTS

I would like to thank my committee members for both their guidance on this dissertation work and their mentorship that has fostered my development as a scientist. I am particularly grateful to Dr. Borchert, who first inspired me to become a scientist during my undergraduate years and continued that mentorship throughout my doctoral education. I appreciate the help and friendship from lab members past and present including Jeffrey DeMeis, Cana Brown, Mika Houserova, Valeria King, Justin Roberts, and Dillon Patterson. Funding was provided in part by NSF RAPID grant NSF2030080 (GMB) and NSF CAREER grant 1350064 (with co-funding provided by the NSF EPSCoR program) (GMB) both awarded by the Division of Molecular and Cellular Biosciences. Also, I am thankful for the mentorship from Dr. Gary Piazza and the Piazza lab members. Most importantly, I thank my wife Lauren who has supported me through the highs and lows of earning a Ph.D. To the children we have been blessed with during graduate school (Rose, Lilah, and Theo), I wouldn’t have been able to accomplish any of this without knowing that you would be there waiting for Dad at the end of the day. There are too many family members to name whose support I have relied on during this time, but my Mom and Dad deserve special mention for their unfailing love for me and support for my education from preschool to Ph.D.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

Appendix Table

LIST OF FIGURES

Figure Page

Appendix Figure

LIST OF ABBREVIATIONS

3C	Chromosome Conformation Capture
Ago	Argonaute Protein
BLAST+	Basic Local Alignment Search Tool
G4	Guanine Quadruplex
IGV	Integrative Genomics Viewer
LG4	Long Guanine Quadruplex Region
miRNA	MicroRNA
mRNA	Messenger RNA
NCBI	National Center for Biotechnology Information
ncRNA	Noncoding RNA
ndRNA	Noncoding-Derived RNA
NGS	Next-Generation Sequencing
nt(s)	Nucleotide(s)
pre-miRNA	Precursor miRNA
pri-miRNA	Primary miRNA
RF	Restriction Fragment
RISC	RNA-Induced Silencing Complex
rRNA	Ribosomal RNA
sdRNA	SnoRNA-Derived RNA
snoRNA	Small Nucleolar RNA
SRA	Sequence Read Archive

TAD	Topologically Associated Domain
TCGA	The Cancer Genome Atlas
TF	Transcription Factor
UTR	Untranslated Region

ABSTRACT

Alexander Coley, Ph.D., University of South Alabama, December 2022. Novel Regulators of Human Gene Expression. Chair of Committee: Glen M. Borchert, Ph.D.

The human genome is rife with regulatory elements that control whether genes are expressed or silenced. While regulatory elements such as epigenomic modifications, transcription factors, promoters, and enhancers are well established, there still remain regulatory elements that are poorly characterized or even undiscovered. In recent years, noncoding RNAs derived from small nucleolar RNAs were identified and linked to gene regulation in a variety of human cancers. Adding to the growing understanding of these sdRNAs, we performed a computational analysis of castration-resistant prostate cancer patient data and identified 38 sdRNAs as significantly overexpressed in CRPC. Two of these, sdRNA-D19B and sdRNA-A24, were found to regulate the expression of tumor suppressor genes CD44 and CDK12 respectively, thus functioning as oncogenic ncRNAs in CRPC. At the level of chromatin conformation, we developed a custom Hi-C analysis pipeline to characterize the interactome of long guanine quadruplex regions. The pipeline successfully identified LG4 contacts with local genes and regulatory elements, indicating that LG4s form topologically associated domains and likely function as a novel mechanism of gene regulation and genomic organization. Taken together, the work presented adds 38 sdRNAs and an entirely new level of gene regulation via LG4 TADs to the growing catalogue of regulators of human gene expression.

CHAPTER I

STATEMENT OF PROBLEM

1.1 Human Gene Expression and Regulation

There are approximately 20,000 protein-coding genes in the human genome, however human cells are wildly differentiated[1]. Containing the same set of genes, epithelial cells of the lung airway form cilia that beat in rhythm to push mucus out of the lungs while natural killer T cells express the T Cell Receptor cell-surface protein to recognize antigens and initiate a cytokine response. The key factor that drives this differentiation is the selective expression of only certain genes within cellular and temporal contexts. It follows that to gain a more complete understanding of the molecular underpinnings of human genetic complexity and pathology, we must gain a more complete understanding of how gene expression is regulated. To address this gap in knowledge, this dissertation's purpose is to discover and characterize novel regulators of human genes.

Before the regulation of gene expression can be discussed, the basics of gene expression must first be established. The basic structure of a protein-coding gene includes the 5' untranslated region (UTR), exons, introns, and the 3' UTR[2] (Fig.1A). Transcription of human messenger RNA (mRNA) is initiated when RNA polymerase II (RNAPII) is recruited to the promoter region that resides directly upstream of a gene's 5'-UTR[3] (Fig.1B). The recruitment of RNAPII is dependent on transcription factors (TF) which bind either immediately upstream of the promoter or at enhancers which can be upwards of 500 kilobases (kb) away from their associated promoter[4]. TFs that bind at enhancer regions engage coactivators and facilitate localization of RNAPII to the associated promoter, where it binds to TFs that are bound near the promoter[3]. These promoter-associated TFs are responsible for allowing RNAPII to recognize the promoter as well as for facilitating chromatin unwinding to enable access to genes for transcription. The resulting chromatin loop forms a transcription "bubble" stabilized by cohesin with RNAPII poised to begin transcribing a gene. Canonically, gene expression flows from DNA to RNA to protein (Fig.1C). Following transcription of the DNA by RNAPII, the resulting pre-mRNA is processed to remove introns and add a 5' cap as well as a polyadenylated tail to the 3' end, producing a mature mRNA. The mature mRNA then exits the nucleus and is translated at ribosomes into protein.

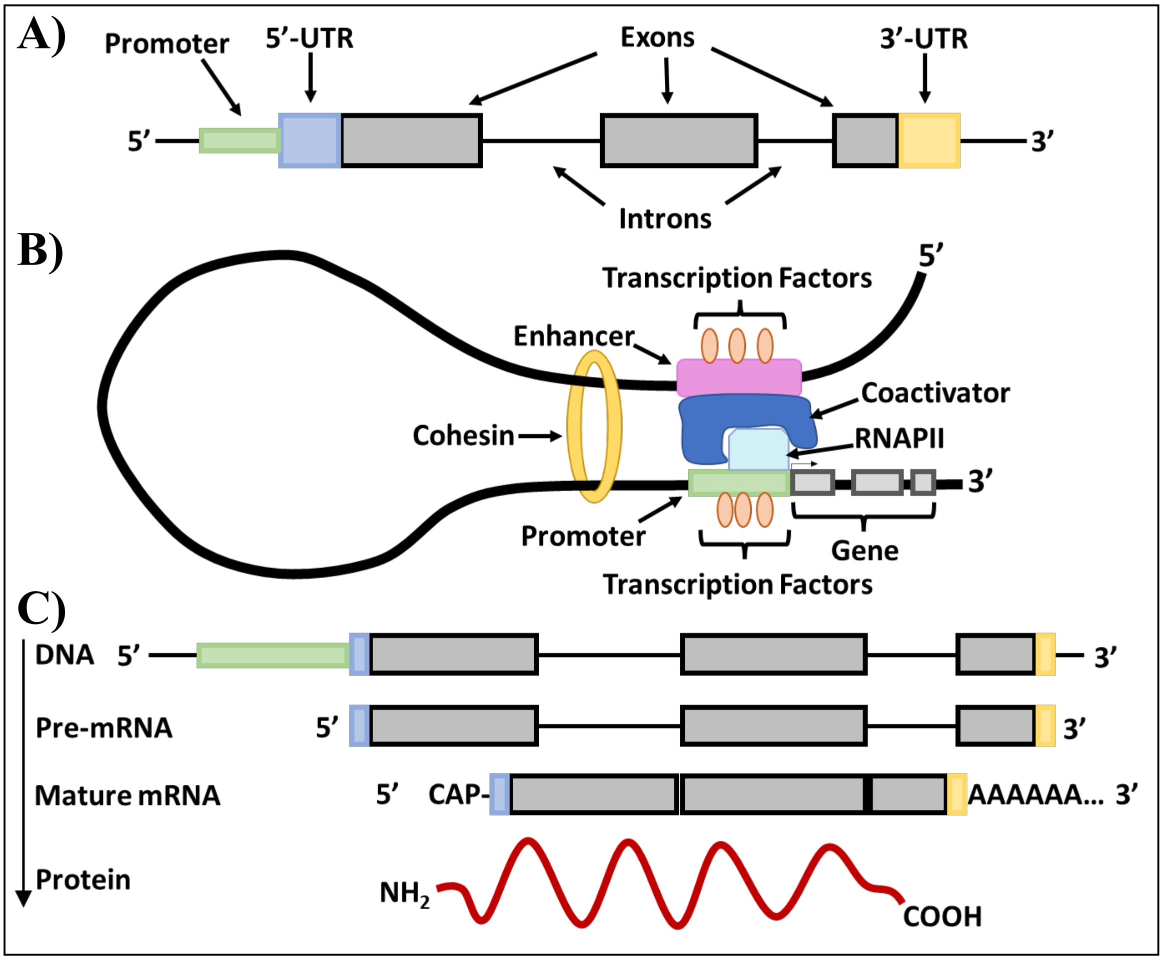


Figure 1. Gene structure and expression. A) The basic structure of a protein-coding gene. Running from 5' to 3', the promoter (green), 5'-UTR (blue), exons (gray boxes), introns (black lines), and 5'-UTR (yellow) compose the gene. **B) The molecular machinery involved in the initiation of transcription.** Transcription Factors (orange) bind to an enhancer (pink), which associates with coactivators (dark blue) to recruit RNAPII (light blue) to an associated promoter region (green) and its bound transcription factors. The transcription machinery is stabilized by cohesion (yellow). Once properly assembled, RNAPII will transcribe the downstream gene (gray boxes). **C) Gene expression from DNA to protein.** Moving stepwise from top to bottom, the DNA sequence of a gene (labeled as in **A**) is transcribed to pre-RNA by the transcription machinery detailed in **B**. Pre-RNA is then modified by processing factors to add a 5' 7-methylguanosine cap, splice out introns, and add a 3' polyadenylated tail resulting in mature mRNA[5]. The mature mRNA is then translated to protein by ribosomal machinery.

Gene regulation occurs at virtually every step in the process of expression[6]. First, the genes, promoters, and enhancer must be accessible by transcription machinery before any gene expression can occur. The 3D organization of the genome, or chromatin confirmation, is tightly controlled and is a key regulator of which genes will be transcribed[7]. Epigenetic control of chromatin conformation is mediated by an arsenal of proteins that can modify histones to expose or condense chromatin, influencing transcriptional activity of the genes located within the affected nucleosome. The importance of chromosome conformation is garnering increased attention, and the result has been an enhanced understanding of how the genome is divided into "neighborhoods" governing regulation[8]. These include topologically-associated domains (TADs), regions of DNA that preferentially self-associate. Typically flanked by CTCF binding regions, TADs constitute functional units of the genome where enhancer-gene interactions are sequestered. This grouping has functional relevance, as TF binding can be confined within TADs leading to the coordinated expression of sets of genes.

Post-transcriptional gene regulation can arise from several mechanisms, but most relevant to this dissertation is regulation mediated by noncoding RNAs (ncRNA). NcRNAs can arise from both exonic and intronic DNA, sometimes as the product of mRNA splicing or processing steps. These RNAs are not destined for translation, but rather fulfill other roles[9]. Some of the larger families of ncRNAs include transfer RNAs (tRNAs) which localize amino acids to the ribosome during translation and ribosomal RNAs (rRNAs) which associated with ribosomal proteins to facilitate translation. Many ncRNAs were consigned to "housekeeping" functions until 1993, when it was first discovered that ncRNAs termed microRNAs (miRNAs) can participate in post-transcriptional gene regulation in C. elegans[10]. The field of miRNA research then greatly increased in popularity following a 2002 publication linking miRNA dysregulation to the chronic lymphoblastic leukemia phenotype[11]. MiRNAs are small ncRNAs of just ~22 nucleotides (nt) in length. Following transcription, pri-miRNAs are processed by a

Microprocessor Complex including the RNase III enzyme DROSHA and DGCR8 into pre-miRNAs, exported from the nucleus into the cytoplasm, and then further processed into mature miRNAs by DICER[12]. Mature miRNAs are loaded on the Argonaute (Ago) protein, which in complex with additional proteins forms the RNA Induced Silencing Complex (RISC)[13]. The RISC, guided by a 2-6 nt "seed-region" of the miRNA, will then typically bind to the 3' UTR of mRNAs. RISC association results either in mRNA degradation if there is a perfect complement or translational repression for a non-perfect complement.

In more recent years, it has become apparent that miRNAs are not the only ncRNAs that exert a regulatory effect. Many of the so-called "housekeeping" ncRNAs, such as tRNAs and rRNAs are actually processed into small ncRNA-derived RNA (ndRNA) fragments that can guide the RISC like miRNAs[14,15]. One novel class of ndRNA arises from the ncRNA family small nucleolar RNAs (snoRNA). SnoRNAs are further categorized as either C/D box snoRNAs or H/ACA box snoRNAs. Canonically, both categories of snoRNA function in the nucleolus to chemically modify pre-rRNAs as a step in rRNA biogenesis, with C/D box snoRNAs coordinating 2'-O-ribose-methylation and H/ACA box snoRNAs coordinating pseudouridylation[16–18] (Fig.2). However, in 2008 it was discovered that snoRNAs are specifically and frequently processed into sno-derived RNAs (sdRNA)[19]. These sdRNAs follow along a similar pathway to the biogenesis and mechanism of action of miRNAs, ultimately forming the RISC and facilitating mRNA post-transcriptional repression[20] (Fig.3).

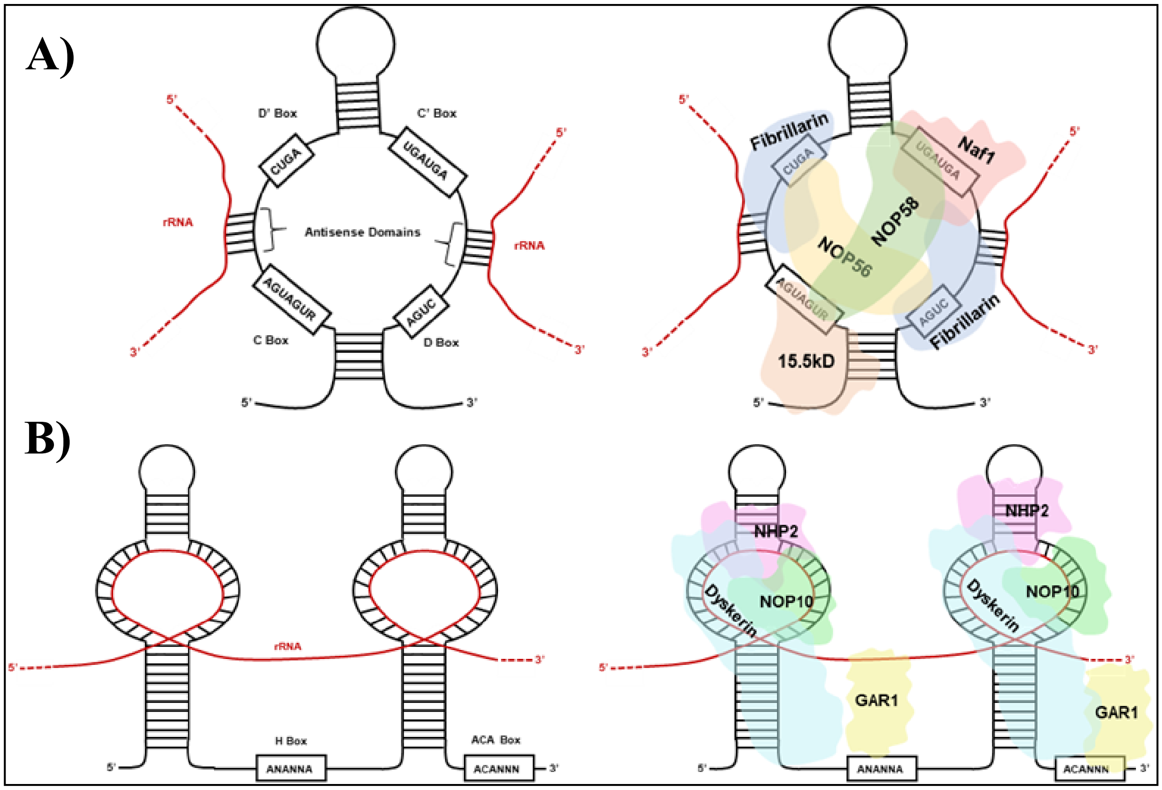


Figure 2. The structure and associated proteins of snoRNAs. A) C/D box snoRNA structure and accessory proteins. (Left) The C/D box snoRNA consists of a 5’ C box (RUGAUGA) motif, a 3’ D box (CUGA) motif, and the C’ and D’ boxes located internally. Antisense domains identify target rRNAs (red) via complementarity. (Right) The C/D box snoRNP complex consists of NOP56, NOP58, 15.5K, and fibrillarin proteins[16]. **B) H/ACA box snoRNA structure and accessory proteins.** (Left) The H/ACA box snoRNA consists of a 3’ ACA box (ACANNN, N= any nt) and two hairpins that target rRNA (red) linked by an H box (ANANNA, N= any nt). (Right) The H/ACA box snoRNP complex consists of NHP2, NOP10, GAR1, and dyskerin proteins[17].

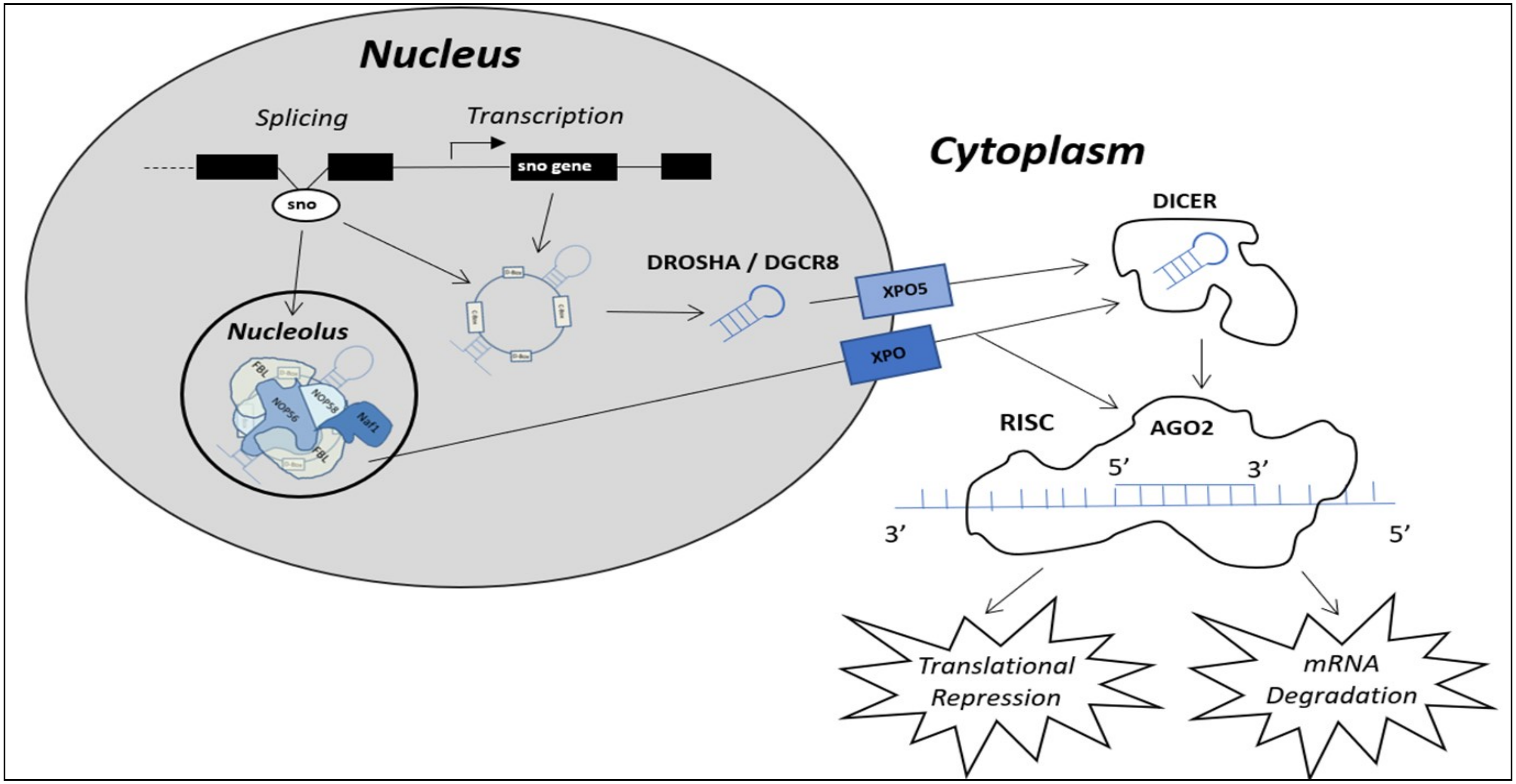


Figure 3. sdRNA Biogenesis and Function. Full length small-nucleolar RNAs (snoRNAs) are generated either as products of transcription or splicing[21-24]. snoRNAs produced by transcription can give rise to microRNA-like sdRNAs which are specifically excised from parent snoRNA transcripts by employment of the classical microRNA (miRNA) processing pathway. This occurs by processing of parent snoRNAs into smaller transcripts by the microprocessor complex which consists of Drosha Ribonuclease III (DROSHA) and DiGeorge syndrome critical region 8 (DGCR8). The intermediate snoRNA then undergoes cytoplasmic exportation via exportin 5 (XPO5). Following this, the smaller cytoplasmic snoRNA is processed by Dicer RNase III endonuclease (DICER) to generate the mature sdRNA which associates with AGO2, leading to the formation of the RNA-induced Silencing complex (RISC). Similar to miRNAs, these sdRNAs function in post-transcriptional gene suppression by antisense binding to target mRNA transcripts within RISC[25]. That said, snoRNAs produced by splicing can also enter the classical miRNA processing pathway. Spliced snoRNAs however can bypass processing from DROSHA/DGCR8 and/or DICER as a result of trafficking to the nucleolus and subsequent processing by the fibrillarin complex followed by cytoplasmic export via a transporter belonging to the Exportin (XPO) family of proteins[26].

Aberrant gene regulation is a critical factor in the onset of a multitude of diseases. With our contemporary understanding of genetics, it is commonly accepted that almost all human diseases are in part influenced by genetic variation[27]. In diseases such as cancers which are more strongly linked to genetic aberrations, the relationship between gene expression and pathology is made clear. Cancer is, by heredity or environmental insult, a genetic disease[28]. Recent studies have shown that abnormal chromosomal conformation can influence the expression of oncogenes and tumor-suppressors leading to phenotypic consequences in cancer[29]. One such example in prostate cancer (PCa) involves aberrant chromatin loop formation driven by Androgen Receptor (AR) that brings distal genes TMPRSS2 and ERG into proximity and facilitates TMPRSS2-ERG fusion, which is a known PCa biomarker and driver of the PCa malignant phenotype[30]. Aberrant post-transcriptional gene regulation by ncRNAs is especially well established as a contributor to cancer[31]. Virtually every cancer type has been linked to at least one misexpressed miRNA that regulates a tumor suppressor or oncogene[32]. Both chromosome conformation and miRNAs have since been used as targets for drug development to alleviate aberrant gene expression driving a given disease[33,34]. Clearly, the identification and characterization of novel gene regulators is of the utmost importance to provide tractable targets for prognosis, diagnosis and even therapeutic interventions.

One such class of novel regulators is the sdRNA. While sdRNAs were first discovered in 2008, they have only become the focus of increasing study over the past decade[19,20]. Studies have been published nearly every year since 2012 that link gene regulation by sdRNAs to cancer phenotypes. Strikingly, many ncRNAs currently annotated as miRNAs actually arise from snoRNAs or snoRNA-like transcripts. In the best case, this is overlooked and the sdRNAs remain misannotated. More commonly, however, once a ncRNA is found to arise from a snoRNA locus it is erroneously disregarded overlooked in subsequent analysis (example text: "Aligned sequences with the following annotations were eliminated (as potential microRNAs): tRNA, snoRNA…")[35]. A review by our lab details the growing catalogue of sdRNAs implicated in human cancer (Table 1)[20]. In one example from our lab, we found that sdRNA-93 is overexpressed in aggressive breast cancer, represses the PIPOX mRNA transcript, and enhances breast cancer cell invasion[36]. Despite the increasing attention given to sdRNAs, there is still a great need for research aimed at characterizing regulatory sdRNAs to provide greater insight into cancer gene regulation and provide tractable targets to improve patient outcomes.

Table 1. Summary of sdRNAs implicated in cancer. SdRNA name includes ensembl gene ID (ENSG) where possible. SdRNA sequence and parental snoRNA name were determined from the provided reference and retrieved via ensembl genome browser. Each sdRNA's previous annotation as a miR, cancer, expression, phenotypic effect, and target were determined from the provided reference. **For sdRNAs arising from two different loci on the same precursor, the 5' sequence precedes "::" and the 3' sequence follows. SdRNAs arising from just one loci are given as a single sequence.

sdRNA	Parental snoRNA	Annotated as miR?	Cancer	Expression	Phenotypic effect	Target
sdRNAs Misannotated as Traditional miRNAs						
sd/miR-664a (ENSG00000281696)	SNORA36B (ENSG00000222370)	Yes	Hepatocellular carcinoma Cervical Cutaneous squamous cell carcinoma	Downregulated in tumor Downregulated in tumor Upregulated in Tumor	Tumor-suppressor Tumor-suppressor Tumor-promoter	AKT2 c-Kit IFR2
sd/miR-1291 (ENSG00000281842)	SNORA2C (ENSG00000221491)	Yes	Pancreatic Pancreatic Renal Cell Carcinoma Prostate Breast	Downregulated in Tumor UNDETERMINED Downregulated in Tumor Downregulated in Tumor Downregulated in metastases	Tumor-suppressor Tumor-suppressor Tumor-suppressor Tumor-suppressor Tumor-suppressor	FOXA2 FOXA2 GLUT1 MED1 UNDETERMINED
sd/miR-1248 (ENSG00000283958)	SNORA81 (ENSG00000221420)	Yes	Prostate	Upregulated aggressive tumor	UNDETERMINED	UNDETERMINED
sd/miR-3651 (ENSG00000281156)	SNORA84 (ENSG00000239183)	Yes	Colorectal Esophageal	Upregulated in tumor Downregulated in tumor	Tumor-promoter UNDETERMINED	TBX1 UNDETERMINED
sd/miR-768 (ENSG00000223224)	SNORD71 (ENSG00000223224)	Yes	Breast Gastric Lung*‡, Breast*, Ovary†, Melanoma†, Liver†, Parotid Gland†, Thyroid Gland†, Large Cell†	UNDETERMINED Downregulated in tumor Downregulated in tumor	UNDETERMINED UNDETERMINED *Tumor-suppressor, †UNDETERMINED	YB-1 UNDETERMINED ‡KRAS
sdRNAs not Previously Annotated as miRNAs						
sd/hsa-sno-HBII-296B	SNORD91B (ENSG00000275084)	No	Pancreatic ductal adenocarcinoma	Downregulated in Tumor	UNDETERMINED	UNDETERMINED
sd/hsa-sno-HBII-85-29	SNORD116-29 (ENSG00000207245)	No	Pancreatic ductal adenocarcinoma	Downregulated in Tumor	UNDETERMINED	UNDETERMINED
sno-miR-28	SNORD28 (ENSG00000274544)	No	Breast	Upregulated in Tumor	Tumor-promoter	TAF9B
sdRNA-93	SNORD93 (ENSG00000221740)	No	Breast	Upregulated in Tumor	Tumor-promoter	PIPOX
sdRNA-D19b	SNORD19b (ENSG00000238862)	No	Prostate	Upregulated in tumor	Tumor-promoter	CD44
sdRNA-A24	SNORA24 (ENSG00000275994)	No	Prostate	Upregulated in tumor	Tumor-promoter	CDK12

Table 1, cont.

miRNAs that Bind Dyskerin						
sd/miR-140 (ENSG00000208017)	Binds Dyskerin	Yes	Prostate	Downregulated in tumor	Tumor-suppressor	BIRC1
sd/miR-151 (ENSG00000254324)	Binds Dyskerin	Yes	Prostate	Downregulated in tumor	Tumor-suppressor	UNDETERMINED
sd/miR-215 (ENSG00000207590)	Binds Dyskerin	Yes	Ovary	Downregulated in Tumor	Tumor-suppressor	XIAP (not confirmed)
			Colorectal	Downregulated in Tumor	Tumor-suppressor	EREG, HOXB9
			Prostate	Downregulated in Tumor	Tumor-suppressor	PGK1 (not confirmed)
			Lung	Downregulated in Tumor	Tumor-suppressor	Leptin, SLC2A5
sd/miR-605 (ENSG00000207813)	Binds Dyskerin	Yes	Melanoma	UNDETERMINED	Tumor-suppressor	INPP4B
			Prostate	UNDETERMINED	Tumor-suppressor	EN2
			Colorectal*, Breast*, Lung †	UNDETERMINED	*Tumor-suppressor, †UNDETERMINED	Mdm2
			Prostate	Downregulated in tumor	UNDETERMINED	UNDETERMINED
			Prostate	UNDETERMINED	UNDETERMINED	UNDETERMINED
miRNAs that Bind Fibrillarin						
sd/miR-16-1 (ENSG00000208006)	Binds Fibrillarin	Yes	Chronic Lymphocytotic Leukemia	Downregulated in tumor	UNDETERMINED	Multiple (not confirmed)
			Gastric	Downregulated in tumor	Tumor-suppressor	TWIST1
			Non-small cell lung cancer	Downregulated in tumor	Tumor-suppressor	TWIST1
			Osteosarcoma	Downregulated in tumor	Tumor-suppressor	FGFR2
			Breast	Downregulated in tumor	Tumor-suppressor	PGK1
sd/miR-27b (ENSG00000207864)	Binds Fibrillarin	Yes	Prostate	Downregulated in tumor	Tumor-suppressor	UNDETERMINED
			Lung	Downregulated in tumor	Tumor-suppressor	LIMK1
			Bladder	Downregulated in tumor	Tumor-suppressor	EN2
sd/miR-31 (ENSG00000199177)	Binds Fibrillarin	Yes	Colorectal	Upregulated in tumor	Tumor-promoter	UNDETERMINED
			Head and neck squamous cell carcinoma	Upregulated in tumor	Tumor-promoter	FIH (not confirmed)
			Lung	Upregulated in tumor	Tumor-promoter	LATS2, PP2R2A
			Glioblastoma	Downregulated in tumor	Tumor-suppressor	RDX
			Melanoma	Downregulated in tumor	Tumor-suppressor	UNDETERMINED
			Prostate	Downregulated in tumor	Tumor-suppressor	UNDETERMINED
sd/let-7g (ENSG00000199150)	Binds Fibrillarin	Yes	Non-small cell lung cancer	Downregulated in tumor	Tumor-suppressor	KRAS (not confirmed)
			Colorectal	Downregulated in tumor	Tumor-suppressor	UNDETERMINED
			Ovary	Downregulated in tumor	Tumor-suppressor	UNDETERMINED
sd/miR-28 (ENSG00000207651)	Binds Fibrillarin	Yes	B-cell Lymphoma	Downregulated in tumor	Tumor-suppressor	MAD2L1, BAG1, RAP1B, RAB23
			Prostate	Downregulated in tumor	Tumor-suppressor	SREBF2
			Breast	Downregulated in tumor	Tumor-suppressor	WSB2

Table 1, cont.

Sno-Derived Piwi-interacting RNAs						
pi-sno75	SNORD75	No	Breast	Downregulated in Tumor	Tumor-suppressor	WDR5
pi-sno74	SNORD74	No	Breast	Downregulated in Tumor	UNDETERMINED	UNDETERMINED
pi-sno44	SNORD44	No	Breast	Downregulated in Tumor	UNDETERMINED	UNDETERMINED
pi-sno78 (sd78-3')	SNORD78	No	Breast	Downregulated in Tumor	UNDETERMINED	UNDETERMINED
			Prostate	Upregulated in metastases	UNDETERMINED	UNDETERMINED
pi-sno81	SNORD81	No	Breast	Downregulated in Tumor	UNDETERMINED	UNDETERMINED
piR-017061 (piR-33686)	SNORD91A (ENSG00000212163)	No	Pancreatic ductal adenocarcinoma	Downregulated in Tumor	UNDETERMINED	UNDETERMINED

A better-known mechanism of gene regulation involves the dispersion of guanine quadruplex (G4) motifs in the genome. G4s are formed when guanine-rich runs of DNA form square planar geometries via hydrogen bonds and stack upon one another to increase stability of the ssDNA structure (Fig.4)[37]. The initial characterization of the G4 helix structure dates back to 1962[38]. However, in more recent years G4 helix formation has been linked to gene expression as well as genomic instability. Interestingly, G4 motifs accumulate in gene promoters and are seen in a higher abundance in cancer genomes at the promoters of oncogenes[39]. Despite detailed study, to date there has been no validated mechanism of action to explain this strong correlation and purported association between G4s and gene expression.

In 2020, our lab took a novel approach to studying G4s by instead considering only long G4 regions (LG4) in the genome[40]. LG4 identification was performed computationally based on the functional LG4 comprising the antibody switch region. 301 total LG4s were identified, and strikingly 217/301 were found to overlap enhancers. It was proposed that LG4s may function as super enhancers, interacting with the promoters of multiple genes to coordinate expression (Fig.5). Potentially, local LG4-DNA interactions could form an LG4 TAD. This putative LG4 TAD would constitute a novel level of gene regulation, where LG4s influence local chromosome conformation to influence gene regulation within their associated TAD.

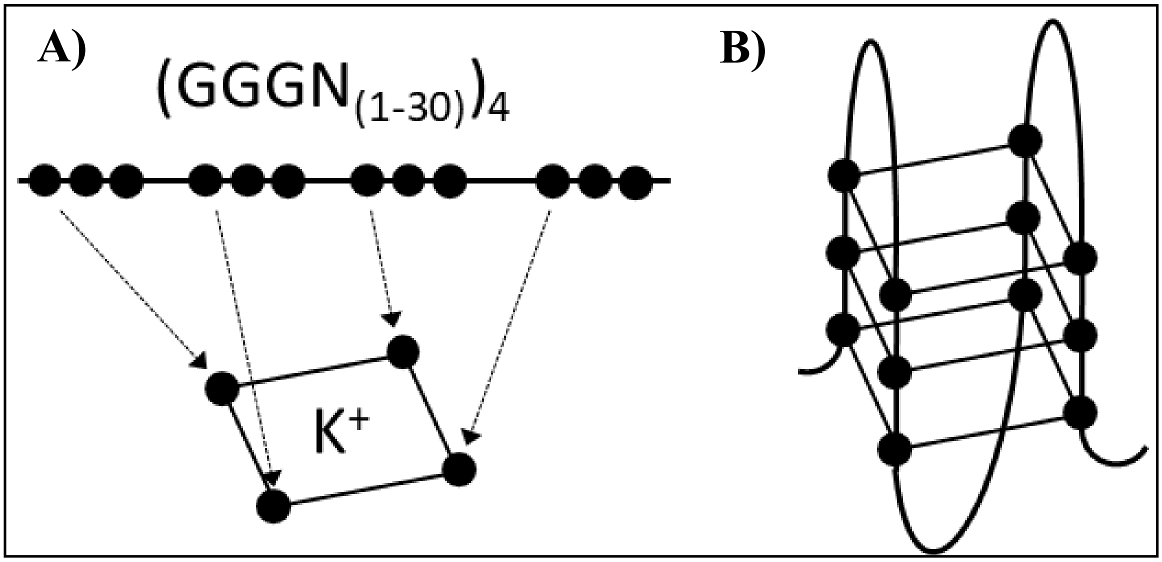
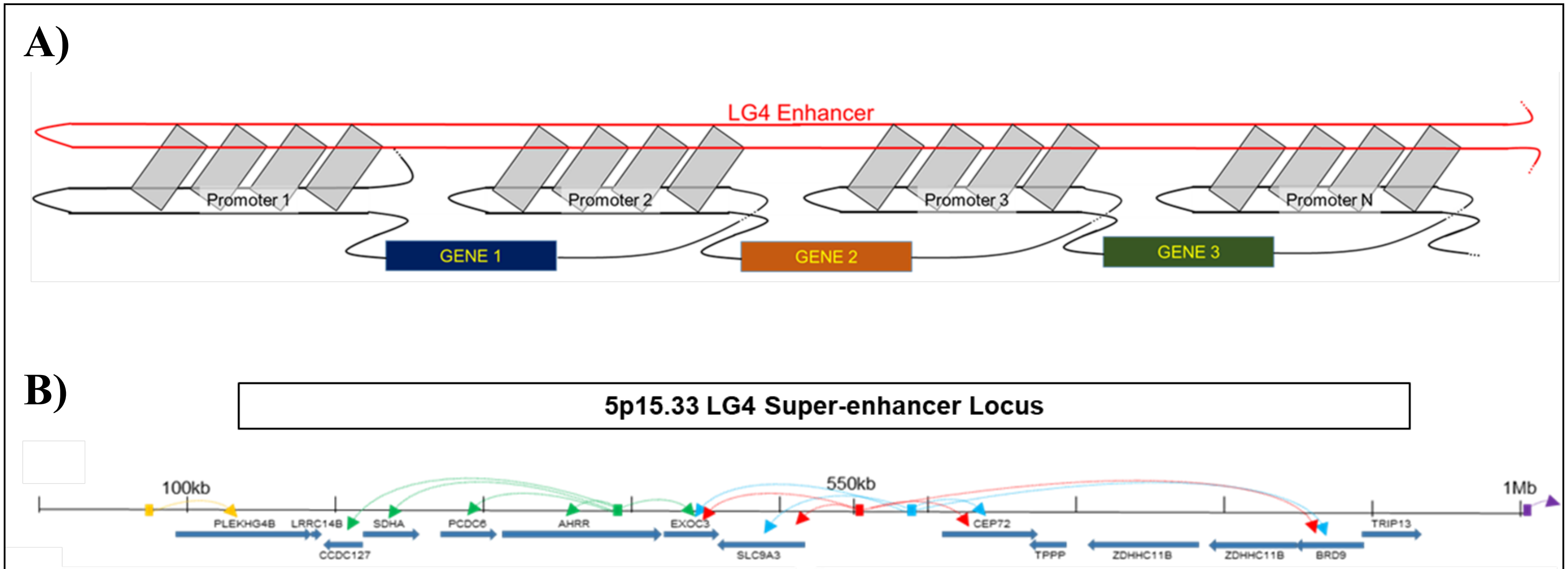


Figure 4. The structure of a guanine quadruplex (G4). A) Linear G-rich regions of DNA form G4s. Stretches of DNA that follow a general formula (top, N = any nt) can associate under physiological conditions to form a tetrad (bottom) stabilized by G-G hydrogen bonding. **B) Single stranded DNA G4 helix.** Consecutive G4s stack to further stabilize the G4 structure and form a ssDNA G4 helix.



LG4	chr5:81537-86535		
Enhancer	chr5:81622-83317	GH05J000081	PLEKHG4B
Enhancer	chr5:389993-391530	GH05J000390	AHRR, CCDC127, EXOC3, PDCD6, SDHA
LG4	chr5:391248-392495		
Enhancer	chr5:392046-397180	GH05J000392	AHRR, CCDC127, EXOC3, PDCD6, SDHA
LG4	chr5:551937-556935		
Enhancer	chr5:553660-554453	GH05J000553	BRD9, CEP72, EXOC3, SLC9A3
Enhancer	chr5:555169-557644	GH05J000555	CEP72, SLC9A3
LG4	chr5:581937-586935		
Enhancer	chr5:585801-587000	GH05J000585	BRD9, CEP72, EXOC3, SLC9A3
Enhancer	chr5:1020384-1022303	GH05J001020	NKD2, SLC12A7, SLC6A18
LG4	chr5:1021899-1026849		
Enhancer	chr5:1025801-1026200	GH05J001025	NKD2, SLC12A7

Figure 5. LG4s Function as Super Enhancers. A) Proposed model for LG4 TAD formation. In our proposed model, LG4 DNA (red) physically interacts with the promoter regions (gray diamonds) of nearby genes (blue, orange, green) to co-regulate gene expression within the LG4 TAD. **B) The 5p15.33 LG4 super enhancer locus.** Supporting the proposed model in **A**, a concentration of enhancer-associated LG4s resides on chromosome 5 (5p15.33). Each LG4 is listed in a colored box alongside its chromosomal position. All enhancers overlapped by each LG4 are listed with their chromosomal position, Genehancer ID (" GH…"), and genes that they regulate. A model is drawn based on enhancer overlap for each LG4, where the LG4 is represented as a box on the genome with arrows pointing towards genes it is predicted to regulate based on enhancer proximity.

The process of identifying novel regulatory elements, whether DNA or RNA, begins with computational biology. For both ncRNAs and regulatory chromatin, a multitude of exploratory strategies exist rooted in next-generation sequencing (NGS) and subsequent computational analysis. Each step from the acquisition of data to its analysis introduces bias, and it is therefore important to understand the choices available to ensure that any discoveries made reflect reality and are not due to artifice.

1.2 Computational Approaches and Challenges of Small ncRNA Discovery

In a given RNA sequencing (RNA-seq) sample, longer RNAs such as rRNA and mRNA will dominate the sequencing depth making it difficult to detect smaller RNAs. This presents an immediate challenge to ncRNA quantification. To overcome this, small RNA-seq workflows include a step where sample RNA is size selected for smaller transcripts and subsequently amplified to provide enough depth for the small ncRNA landscape to be captured[41]. This is a significant point of bias introduction due either to the methods involved in selection or over/under amplification of transcripts. At this point, the sequencing data consists of unidentified reads. The next step in small ncRNA sequence analysis, identical to the analysis of total RNA, is sequence alignment.

A variety of alignment programs have been developed over the years including Basic Local Alignment Search Tool (BLAST), Bowtie2, and Burrows-Wheeler Aligner (BWA) to name a few[42–44]. While differences between these aligners must be considered in experimental design, in general they follow the same stepwise process. Each read from the sample dataset is mapped to a given reference. This presents a challenge to the discovery of any novel RNA transcripts including small ncRNAs as the reference selected will limit what can be discovered. As mentioned previously, many ncRNA databases will erroneously discard transcripts that align to larger ncRNAs making it impossible to identify them using these databases. Care must be taken when building or selecting an ncRNA database so that potentially significant results are not discarded before alignment even takes place. Alignments are commonly generated based on sequence identity between the input and the reference, which is scanned at specified intervals for matches. Often, search parameters are set based on the type of data and the overall experimental design. For example, how an aligner handles percent match, length of alignments, and gapping are all important parameters that need to be considered on a case-by-case basis. Smaller ncRNAs such as miRNAs and sdRNAs can be omitted by the default word-size search parameters of many aligners including the NCBI's BLAST. Adjusting search parameters to include ncRNAs is vital to the selection of the aligner as well as its fine tuning in a small ncRNA discovery experiment.

Following alignment, reads need to be quantified and analyzed to draw conclusions such as differential expression. Many packages exist that will perform statistical analyses on aligned reads automatically such as DESeq2 and baySeq[45,46]. Alternatively, many publications include custom-build analysis pipelines suited to a particular lab's needs in-context. For the identification of sdRNAs, a further step is required to identify whether any fragments are preferentially processed from the full length snoRNA. Our lab recently built a web resource for the identification of ncRNA fragments present in RNA-seq datasets, Short Uncharacterized RNA Fragment Recognition (SURFR)[47,48]. SURFR is an all-in-one tool that aligns given RNA-seq data to a custom ncRNA reference, and then performs a differential expression analysis of the reads aligning to a single full length ncRNA to identify any potential ndRNA. A subsequent wavelet analysis is conducted on the differentially expressed reads within a ncRNA that yields the expression levels, start, and end positions of any putative ndRNAs (Appendix A).

1.3 Computational Approaches and Challenges of DNA Interactome Discovery

Several options exist to assess 3D genomic organization, with the most popular in recent years being Hi-C which integrates previous chromosome conformation capture (3C) approaches with high-throughput NGS data[49]. Hi-C is performed on the entire genome of a sample, allowing for the detection of DNA interactions globally without requiring prior selection of targets like the earlier methods of 3C[50]. One relevant application of Hi-C is the identification of TADs in the genome. The TADs identified using Hi-C often function as regulatory neighborhoods where regulatory interactions (such as promoter-enhancer association) are confined[8]. Thus, Hi-C data is excellently suited to identify novel chromosome conformation motifs and characterize the interactions within.

To generate contact data from a given sample, the Hi-C workflow involves first crosslinking DNA to preserve DNA-DNA interactions[49]. Crosslinked DNA is then biotinylated and digested into restriction fragments (RF) using a restriction enzyme. The RFs are ligated and pulled down for high-throughput sequencing, resulting in Hi-C reads that are composed of pairs of interacting DNA[51]. A hurdle that Hi-C data had to overcome in order to ascend to its current status was the issue of sequencing depth. By design, Hi-C data captures global genomic interactions. To ensure enough interaction data is captured to parse out significant interactions from the ubiquitous self-self and adjacent DNA interactions, typical Hi-C experiments require 200-500 million read pairs to be generated[52]. This is an order of magnitude higher than the average number of reads needed for a typical RNA seq experiment, and consequently Hi-C datasets are typically quite large (~100GB). Hi-C analysis, especially when multiple samples are involved, must be scalable to account for the relatively large size of Hi-C data files compared to other NGS data.

Analysis of Hi-C reads can be performed using publicly available packages such as HiCdat and Juicer[53,54]. However, more so even than with ncRNA NGS analysis, data scientists tend to create custom analysis packages tailored to their specific goals for the Hi-C data. In general, the workflow for Hi-C analysis involves alignment of each fragment to a reference genome to identify each partner constituting an interaction pair[55]. Following fragment identification, reads are typically aggregated based on their chromosomal locus and filtered based on the user's personal criteria. This usually involves the exclusion of highly repetitive DNA or relatively short fragments. The final interactome is often displayed as a contact matrix that graphically represents all interactions across a given genomic interval. Depending on the experimental conditions, interpretation of these interactions can reveal novel regulatory domains that self-associate to control gene expression.

1.4 Hypothesis

We hypothesize that by leveraging publicly available ncRNA-seq and Hi-C NGS data we can characterize novel regulators of the human genome at both the post-transcriptional and chromosome conformation levels.

CHAPTER II

MICRORNA-LIKE SNORNA-DERIVED RNAS (SDRNAS) PROMOTE CASTRATION-RESISTANT PROSTATE CANCER

2.1 Brief Overview

Small nucleolar RNAs (snoRNAs) are specifically and frequently processed into snoRNA-derived RNAs (sdRNAs) that function primarily as miRNA-like regulators of gene expression[19,25]. Over the past decade many sdRNAs have been demonstrated to possess key oncogenic and/or tumor suppressive roles in cancers of diverse tissue types including breast, lung, and the prostate[36,56,57]. The lack of sustainable treatment options for the castration-resistant prostate cancer (CRPC) molecular subtype largely contributes to the fact that prostate cancer (PCa) is the second leading cause of cancer death in American men, behind only lung cancer[58]. In this chapter, we conduct the first ever analysis of the CRPC sdRNAome. Using computational analyses, we identify significantly differentially expressed sdRNAs in CRPC and perform *in vitro* experiments to elucidate their contribution to the CRPC phenotype as well as the target genes that they regulate. The results of this work expand the CRPC regulatory landscape to include sdRNAs as potential new therapeutic targets and/or prognostic indicators.

2.2 Methods

2.2.1 SURFR Alignment and Data Analysis

All samples were acquired from The Cancer Genome Atlas (TCGA) Research network PRAD dataset and are publicly available at https://www.cancer.gov/tcga (accessed on 14 January 2019). The Short Uncharacterized RNA Fragment Recognition (SURFR) tool[47,48] is a publicly available web-based tool that comprehensively profiles ncRNA-derived RNAs from input RNA-seq data. SURFR analysis of TCGA PRAD and normal prostate control returned expression in reads per million (RPM) for each sdRNA detected. Rstudio[59] was used to calculate differential expression and rank each sdRNA by cancer prevalence (% of TCGA samples that expressed the sdRNA) and differential expression. Significant results were constricted to those sdRNAs with ≥2× fold change in prostate cancer and were expressed at ≥30 RPM in a minimum of 50% of TCGA PRAD small RNA-seq files. To confirm SURFR findings, small RNA-seq files were obtained for the TCGA PRAD dataset (https://www.cancer.gov/tcga (accessed on 14 January 2019)). Alignments between snoRNAs and reads were obtained via BLAST+ (available at https://blast.ncbi.nlm.nih.gov/Blast.cgi (accessed on 14 January 2019)) using the following parameters: 100% identity, word_size = 6, ungapped, and e-value = 0.001. The frequency of alignments to putative sdRNA loci across each full length snoRNA was calculated by counting reads rigidly defined as ≥20 nts and perfect matches (100% identity). PC3 cell Ago pulldown data were obtained from the NCBI SRA (www.ncbi.nlm.nih.gov/sra/ (accessed on 14 January 2019)) with the identifier

SRR2966868. Alignments between sdRNAs and Ago pulldown reads were obtained via BLAST+ using the same parameters as listed above.

2.2.2 Validation of sdRNA Expression via Quantitative RT-PCR

Small RNA was isolated using mirVana miRNA Isolation Kit according to the manufacturer's instructions. Real-time, quantitative PCR was performed to validate sdRNA expression using the All-in-One miRNA qRT-PCR Kit (GeneCopoeia). Reactions were performed in triplicate in a 96-well plates using 0.2 µM of each custom forward and universal reverse primers provided in the kit and 1.5 µg of total RNA in nuclease-free water. qRT-PCR was conducted on the iQ-5 Real-Time PCR Detection System (Bio-Rad) with the following settings: initial polymerase activation and DNA denaturation at 95 °C for 10 min, followed by 40 cycles of 95 °C for 10 s, 60 °C for 20 s, 72 °C for 15 s. Specificity of amplifications was verified using melting curves. qRT-PCR primers are listed in the appendix (Appendix B).

2.2.3 Manipulating sdRNA-D19b and -A24 levels

Antisense oligonucleotides were designed to target sdRNAs and ordered as custom IDT® miRNA Inhibitors from IDT (Integrated DNA Technologies, Coralville, IA, USA). Similarly, sdRNA mimics and scrambled controls were ordered as custom miRIDIAN mimics from Dharmacon (GE Healthcare Dharmacon, Inc., Chicago, IL, USA). Mimic and inhibitor sequences are detailed in the appendix (Appendix C). Cell migration, proliferation, and invasion assays were then performed to observe the effects of manipulating sdRNA-D19b and -A24 levels. Human PC3 cells (ATCC, CR L-1435) were cultured at 37 °C in 25 cm2 vented flasks (Corning, Manassas, VA, USA) with DMEM (Corning) supplemented with 10% fetal bovine serum (Corning) and 1%

PenStrep (Corning) in a humidified atmosphere at 5% CO2. For transient transfections, the cells were cultured in 12-well plates and grown to 60% confluency before transfection with mimics or inhibitors using Lipofectamine RNAiMAX (Life Technologies, Carlsbad, CA, USA).

2.2.4 Phenotypic Assays

Proliferation assays. PC3 cells were first transfected with either 100 nmol/L of RNA mimic, antisense RNA (inhibitor), or negative control using Lipofectamine RNAiMAX (Life Technologies, Carlsbad, CA, USA) according to the manufacturer's protocol. The cell number was determined by trypan blue staining and manual counting at 24, 36, and 48 h post-transfection. Proliferation was determined as the relative cell number compared with the vehicle-treated (0.1% DMSO) controls ($n \geq 8$).

Cell migration assays. Scratch assays were used to assess migration. PC3 cells were transfected with inhibitors or mimics in standard Petri dishes (Corning), as described for examining the cell proliferation, then grown to 100% confluence. A 1 cm-wide zone was scratched across the center of each dish, and then images were taken every 3 h using an EVOS XL Core inverted microscope imaging system to assess the rate of migration ($n \geq 3$).

Examining chemoresistance. Following transfection, the cells were incubated for 20 min in 5% CO2 at 37 °C, after which they were treated with paclitaxel (5 nM), dasatinib (50 nM), cisplatin (50 µM), or DMSO control. Cell survival was determined by methylene blue staining and manual counting at 0, 6, 12, 18, and 24 h post-transfection. Viability was determined as the relative live cell number compared with vehicle-treated (0.1% DMSO) controls ($n \geq 3$).

2.2.5 Vector Construction

Unless otherwise indicated, PCR amplifications were performed in 40 µL reactions at standard concentrations (1.5 mM MgCl2, 0.2 mM dNTP, 1x Biolase PCR buffer, 0.5 U Taq (Bioline USA, Inc., Randolph, MA, USA), 0.5 µM each primer) and using standard cycling parameters (94 °C—3 min, (94 °C—30 s 55 °C—30 s, 72 °C—60 s) × 30 cycles, 72 °C—3 min), then, they were cloned into Topo PCR 2.1 (Invitrogen) and sequenced. Antisense reporters were constructed by the standard PCR with primers containing 5′ Xho-I and 3′ Not-I restriction enzyme sites. Following digestion, amplicons were ligated into the Renilla luciferase 3′UTR of psiCheck2 (Promega, Madison, WI, USA) vector linearized with Xho-I and Not-I. Reporter assays were performed as previously[36] described, where the presence of an independently transcribed firefly luciferase in these reporters allowed normalization for transfection efficiency. Primer sequences are detailed in the appendix (Appendix B).

2.2.6 Luciferase Assays

Human embryonic kidney (HEK293) cell line was obtained from GenLantis (San Diego, CA, USA) and cultured in MEM (Mediatech, Herndon, VA, USA) supplemented with 10% fetal bovine serum (Hyclone, Logan, UT, USA), 25 mg/mL streptomycin, and 25 I.U. penicillin (Mediatech). Cells were cultured in a humidified atmosphere with 5% CO2 at 37 °C. For luciferase assays, HEK293 cells were cultured in MEM (10% FBS and 1% PS) in 12-well plates. At 90% confluency, cells were transfected following the Lipofectamine 2000 (Invitrogen, Carlsbad, CA, USA) protocol. At 36 h post transfection, cells were scraped from well bottoms and transferred to 1.5 mL Eppendorf tubes. Eppendorfs were centrifuged at 2000 RCF for 3 min, followed by supernatant aspiration and cell resuspension in 300 μL of PBS. Cells were lysed by freeze thaws and debris removed by centrifuging at 3000 RCF for 3 min. A total of 50 μL of supernatant was transferred to a 96-well MicroLite plate (MTX Lab Systems, Vienna, VA, USA), then, firefly and Renilla luciferase activities were measured using the Dual-glo Luciferase® Reporter System (Promega) and a 96-well plate luminometer (Dynex, Worthing, West Sussex, UK). RLUs were calculated as the quotient of Renilla/firefly RLU and normalized to mock.

2.2.7 Statistical Analyses

Cell proliferation and migration assays. Treatment effects were assessed using a two-tailed Student's t-test at each time point measurement. To assess the longitudinal effects of treatment, a mixed model was utilized to examine the difference across all groups and between each pair of groups for the whole study period. Data were presented as mean ± SD from no less than three independent experiments, and a p value < 0.05 was considered significant. For imaging, five microscopic fields randomly chosen from each assay were counted individually, then, the results were averaged. Luciferase assays. Data are presented as the average intensity ± standard deviation in four independent experiments. Quantitative RT-PCR. Gene expression was calculated via the Delta–Delta cycle threshold method and qRT-PCR data were analyzed by Fisher's exact test.

2.3 Results

2.3.1 *In Silico* Identification of PCa-Overexpressed sdRNAs

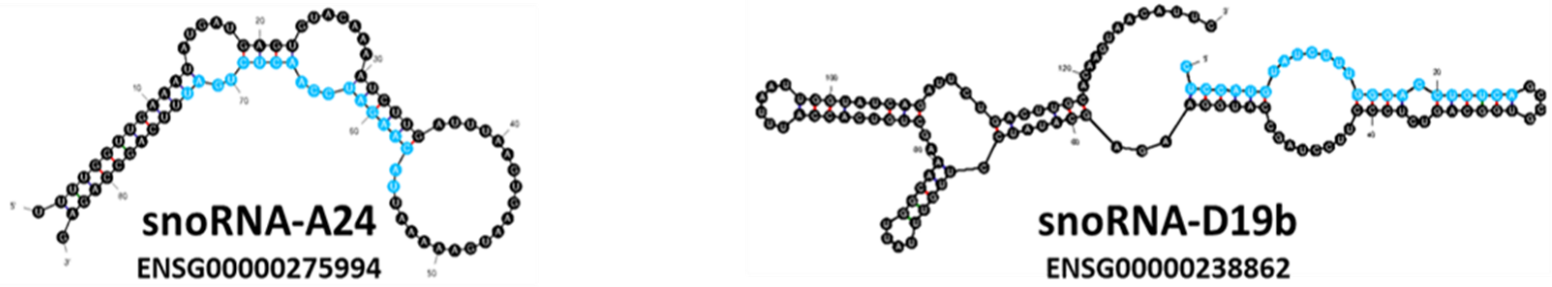
Our lab has recently developed a web resource to identify and quantify noncoding RNA fragments present in small RNA-seq datasets, namely, Short Uncharacterized RNA Fragment Recognition (SURFR). Briefly, SURFR aligns next generation sequencing (NGS) datasets to a frequently updated database of all human ncRNAs, performs a wavelet analysis to specifically determine the location and expression of ncRNA-derived fragments (ndRNAs), and then conducts an expression analysis to identify significantly differentially expressed ndRNAs. We began by utilizing SURFR to determine sdRNA expressions in 489 PCa and 52 normal prostate TCGA patient RNA-seq datasets. This produced a ranked catalogue of significantly differentially expressed sdRNAs in PCa (APPENDIX D). We elected to focus on sdRNA-A24 and sdRNA-D19b for in vitro characterization as: (1) SdRNA-D19b is expressed (avg. 384 RPM) in 91.6% of 489 TCGA PCa samples versus only 42.3% of normal tissue controls (avg. 162 RPM), and sdRNA-A24 is expressed (avg. 711 RPM) in 97.5% of 489 TCGA PCa samples versus only 30.8% of normal tissue controls (avg. 150 RPM) (Figure 6A). (2) Both sdRNA-A24 and sdRNA-D19b are specifically excised from unique, annotated snoRNA parental loci (Figure 6B). (3) RNA-seq analyses indicate they are both expressed in PC3 cells in agreement with our qRT-PCR analyses (data not shown), where they are also found in association with Ago (Figure 6C,D). In summary, sdRNA-A24 and sdRNA-D19b were ultimately selected for experimental interrogation, as they were the only two sdRNAs found in association with Ago in PC3 cells that were expressed in >90% of TCGA PCa samples but <50% of TCGA normal tissue controls (APPENDIX D).

Figure 6. SdRNAs-D19b and -A24. A) SdRNA-A24 and sdRNA-D19b are significantly overexpressed sdRNAs in TCGA prostate cancer patient datasets. The SURFR algorithm[47,48] was used to identify sdRNAs abundantly expressed in prostate cancer patient tumors versus normal prostate. **B) The most thermodynamically stable secondary structures of putative sdRNA producing snoRNAs with sdRNA sequences highlighted in blue as calculated by Mfold[60]**. Common name and Ensembl gene ID for putatively processed snoRNAs are listed below corresponding structures. " Hits" refer to the number of times fragments of putative sdRNA producing snoRNAs perfectly aligned to small RNA-seq reads (PRAD ID: f45a166f-d67b-5de1-8cbd-b5782659457a) from the TCGA prostate cancer dataset. Numbers preceding total numbers of hits correspond to the number of times positions highlighted in blue (putative sdRNAs) perfectly aligned to small RNA-seq reads (e.g., 1380 of 1407 small RNA reads aligning to snoRNA-A24 corresponded to the sequence highlighted in blue). **C) Alignment between the human genome (GRCh38:chr4:118279190-118279320:1) (top), SNORA24 (ENSG00000275994) (upper middle), sdRNA-A24 (SURFR call) (lower middle), and next generation small RNA sequence read (bottom) obtained by Illumina sequencing of PC3 cell Ago immunoprecipitations (SRR2966868) is shown.** The underlined sequence corresponds to the Illumina TruSeq Small RNA adapter RA3. All sequences are in the 5′ to 3′ direction. An asterisk indicates base identity between the snoRNA and genome. Vertical lines indicate identity across all three sequences. **D) Alignment (as in C) between the human genome (GRCh38:chr3:52690744-52690827:1) (top), SNORD19b (ENSG00000238862) (upper middle), sdRNA-D19b (SURFR call) (lower middle), and next generation small RNA sequence read (bottom) obtained by Illumina sequencing of PC3 cell Ago immunoprecipitations.**

A)

	Prevalence (%) in 489 PCa Samples	Average Expression (RPM) in Pca Samples	Prevalence (%) in 52 Tissue Controls	Average Expression (RPM) in Tissue Controls	Differential Expression Fold Change (Cancer/Control)
sdRNA-A24	97.5	711	30.8	150	4.74x
sdRNA-D19b	91.6	384	42.3	162	2.4x
sdRNA-D30	99.6	31067	100.0	19719	1.6x
sdRNA-D61	53.2	215	17.3	119	1.8x

B)



snoRNA-A24
ENSG00000275994
1380 of 1407 hits

snoRNA-D19b
ENSG00000238862
995 of 1007 hits

C)

```
Ch.4                    TGTGAATCTCCATGTATCTTTGGGACCTGTCAGCCGTGGCAGTCTCCCTTCCTAGCCATGGAAGAGCATATCCTTGTTTATTGGCAAAGCTGTCACCATTTAATTGGTATCAGATTCTGACTTGCACAAGTAACATTCACTGTTA
                               |||||||||||||||||||||||||**********************************************************************************************************
snora24                        CTCCATGTATCTTTGGGACCTGTCAGCCGTGGCAGTCTCCCTTCCTAGCCATGGAAGAGCATATCCTTGTTTATTGGCAAAGCTGTCACCATTTAATTGGTATCAGATTCTGACTTGCACAAGTAACATTC
                               |||||||||||||||||||||||||
sdRNA-A24                      CTCCATGTATCTTTGGGACCTGTCA
                               |||||||||||||||||||||||||
SRR2966868.111682636           CTCCATGTATCTTTGGGACCTGTCAGCCTGGAATTCTCGGGTGCCAAGGA
```

D)

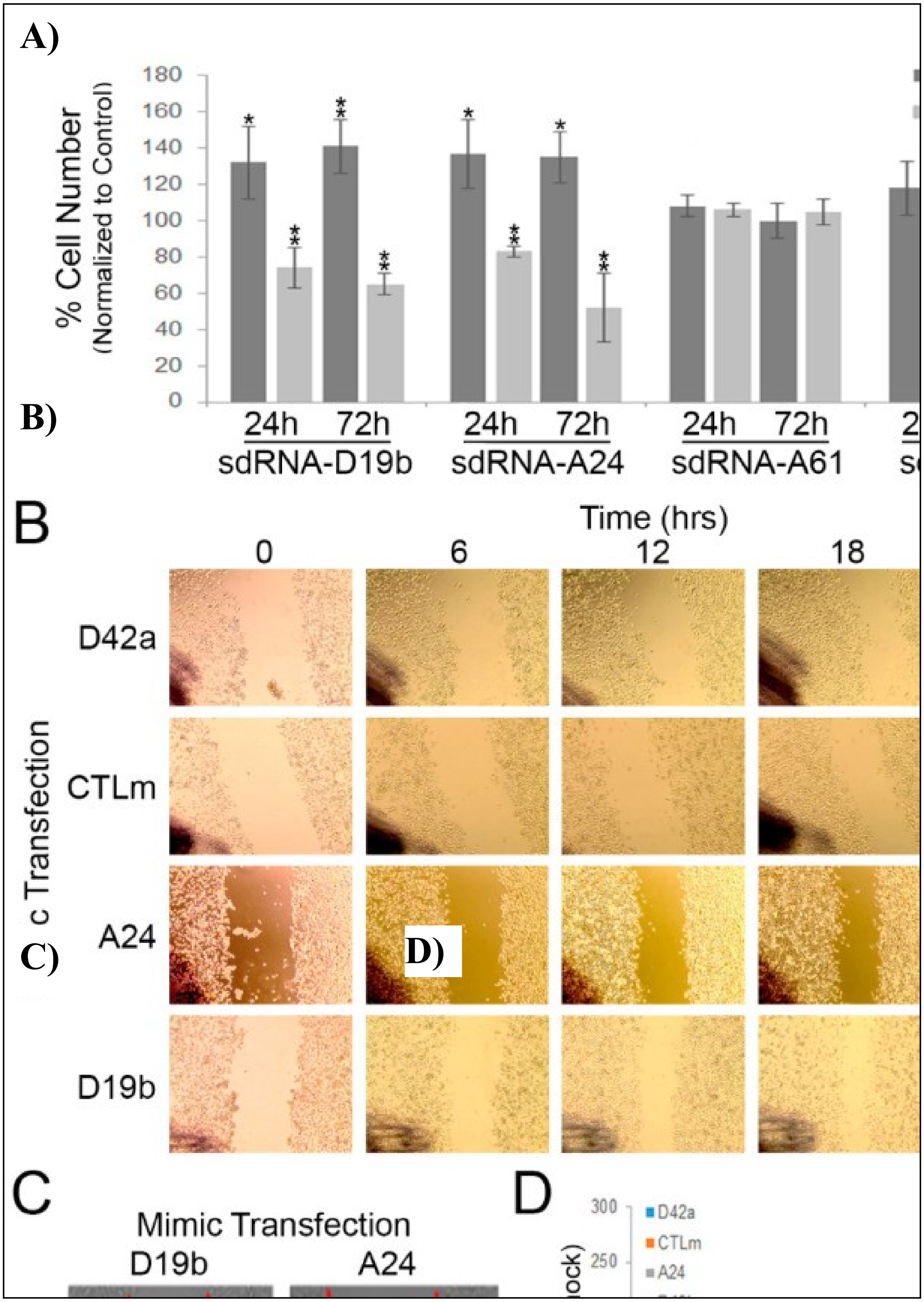
```
Ch.3                    AGAAATGTGATTCTTTCAGATTTTGGTTGAAATATGATGAGTGTACAAAATCTTGATTTAAGTGAATGAAAAATTACAAGATCCAACTCTGATTTCAGCCAGAGATCATCTGAAAGGCAATGT
                                            ****************************************************|||||||||||||||||||||***********
snord19b                                    TTTTGGTTGAAATATGATGAGTGTACAAAATCTTGATTTAAGTGAATGAAAAATTACAAGATCCAACTCTGATTTCAGCCAGAG
                                                                                                |||||||||||||||||||||
sdRNA-D19b                                                                                      ATTACAAGATCCAACTCTGAT
                                                                                                |||||||||||||||||||||
SRR2966868.11768371                                                                           AAATTACAAGATCCAACTCTGATTTGGAATTCTCGGGTGCCAAGGAACTC
```

2.3.2 sdRNA-D19b and sdRNA-A24 Expressions Directly Affect PC3 Cell Proliferation

We selected the PC3 cell line to interrogate the CRPC sdRNAome and determine whether sdRNAs-D19b and -A24 contribute to the CRPC phenotype. PC3 cells are commonly used as a model of aggressive CRPC, as they do not express the androgen receptor, and their growth is independent of androgen signaling[61]. To manipulate sdRNA expression, we used a custom mimic/inhibitor system detailed and validated in a previous publication from our lab[36]. In brief, RNA sequences identical to sdRNA-D19b and sdRNA-A24 were commercially synthesized and used to simulate sdRNA overexpression through transfecting PC3 cells with these specific sdRNA mimics. Conversely, RNAs complementary to sdRNA-D19b or sdRNA-A24 were similarly synthesized and employed as sdRNA inhibitors through transfecting PC3 cells with these specific sdRNA antagomiRs. We first evaluated the effects of manipulating sdRNAs-D19b and -A24 expressions on PC3 proliferation. Excitingly, the misexpression of either sdRNA-D19b or sdRNA-A24 profoundly impacted PC3 proliferation as compared to control sdRNAs (sdRNA-A61 and sdRNA-93), which are not significantly expressed in TCGA PCa samples, but interestingly, were previously shown to positively contribute to breast cancer cell proliferation[36]. The overexpression of sdRNA-D19b increased PC3 cell proliferation by 24% and 32% at 24 and 72 h, respectively (as compared to cells transfected with scrambled controls). Conversely, sdRNA-D19b inhibition reduced PC3 cell proliferation by 22% and 32% at 24 and 72 h, respectively. Similarly, sdRNA-A24 overexpression enhanced PC3 proliferation by ~25% at both 24 and 72 h, and sdRNA-

A24 inhibition decreased proliferation by 14% and 40% at 24 and 72 h, respectively (as compared to cells transfected with scrambled controls). Conversely, PC3 proliferation was not significantly altered following the manipulation of the expressions of two distinct, control sdRNAs expressed in PC3 cells but not differentially expressed in PCa. Collectively, these results indicate functional involvements for both sdRNA-D19b and sdRNA-A24 in PC3 proliferation (Figure 7A).

Figure 7. SdRNA-D19b and -A24 levels significantly impact PC3 cell proliferation and migration. A) PC3 cells were transfected with indicated sdRNA mimic or antagomiR (Anti-sd). Cell counts were performed at 24 and 72 h then normalized to scrambled control transfections (n = 8). * indicates $p \leq 0.05$; ** indicates $p \leq 0.01$; p-values by unpaired two-tailed t-test. **B,C) Representative migration (wound-healing) assays for PC3 cells transfected with the indicated sdRNA mimic.** Wound border closure is indicated by black arrows. **D) PC3 migration assays quantified.** Images were captured at the indicated times (X-axis) and wound healing quantified using ImageJ as % migration normalized to scrambled control ($n \geq 3$). * indicates $p \leq 0.05$; p-values by unpaired two-tailed t-test. D42a, sdRNA-D42a mimic; CTLm, scrambled mimic; A24, sdRNA-A24 mimic; D19b, sdRNA-D19b mimic.


A)
% Cell Number
(Normalized to Control)
180
160
140
120
100
80
60
40
20
0
24h 72h
sdRNA-D19b
24h 72h
sdRNA-A24
24h 72h
sdRNA-A61
2
s
B)
B
Time (hrs)
0
6
12
18
c Transfection
D42a
CTLm
A24
D19b
C)
D)
C
Mimic Transfection
D19b
A24
D
300
250
ock)
D42a
CTLm
A24

2.3.3 sdRNA-D19b Overexpression Enhances PC3 Cell Migration

Uncontrolled cell proliferation is a key cellular process during oncogenesis and is recognized as a hallmark of cancer[62]. Another vital hallmark is the acquisition of migratory capabilities, enabling primary tumors to exit their local environment and give rise to metastases. These metastases are primarily responsible for patient mortality[63]. CRPC is notoriously metastatic, a characteristic largely responsible for its associated high morbidity. As such, we next assessed whether sdRNAs-D19b and -A24 similarly contribute to PC3 cell migration via the wound-healing assay. In this method, a "scratch" was introduced to bisect confluent cells in a culture dish following sdRNA mimic, inhibitor, or scrambled control transfection (Figure 7B,C). We found neither sdRNA-D19b, sdRNA-A24 inhibition, nor sdRNA-A24 overexpression significantly altered the PC3 migration as compared to the controls. Notably, we similarly found neither inhibition nor overexpression of a sdRNA significantly overexpressed in TCGA PCa samples (APPENDIX D) but not expressed in PC3 cells (sdRNA-D42a) significantly altered PC3 migration. In striking contrast, however, we found sdRNA-D19b overexpression markedly increased migration (avg 86.8%) between 6 h and 24 h (Figure 7D).

2.3.4 sdRNA-D19b and sdRNA-A24 Manipulations Alter Drug Sensitivities In Vitro

To assess the potential role of sdRNAs-D19b and -A24 in modulating PCa drug resistance, we examined treatment with three cytotoxic agents, paclitaxel, cisplatin, and dasatinib, to encompass a range of mechanisms of action of drugs typically leveraged to treat CRPC. PC3 cells were treated with one of the chemotherapeutic drugs and either sdRNA mimic, inhibitor, or scrambled control, and then the cells were enumerated every 6 h to assess the impact of sdRNA expression on chemoresistance. Neither overexpression nor inhibition of sdRNA-D19b significantly altered PC3 sensitivity to paclitaxel. In contrast, sdRNA-A24 overexpression improved PC3 resistance to paclitaxel, increasing cell viability between 28.9% and 70.3% at all time points as compared to controls and, although not statistically significant, the sdRNA-A24 inhibition reciprocally sensitized PC3 cells to paclitaxel by 43.2% and 23.9% at 18 and 24 h, respectively (Figure 8A). Conversely, sdRNA-D19b overexpression markedly desensitized PC3 cells to dasatinib treatment, increasing cell viability by over 3-fold at 24 h as compared to controls, whereas neither sdRNA-D19b inhibition nor sdRNA-A24 overexpression nor inhibition produced any discernable effect (Figure 8B). Together, these results clearly support a significant, albeit complex, role for sdRNAs in PC3 drug resistance and strongly imply that sdRNA-D19b and sdRNA-A24 occupy different mechanistic roles in greater drug resistance.

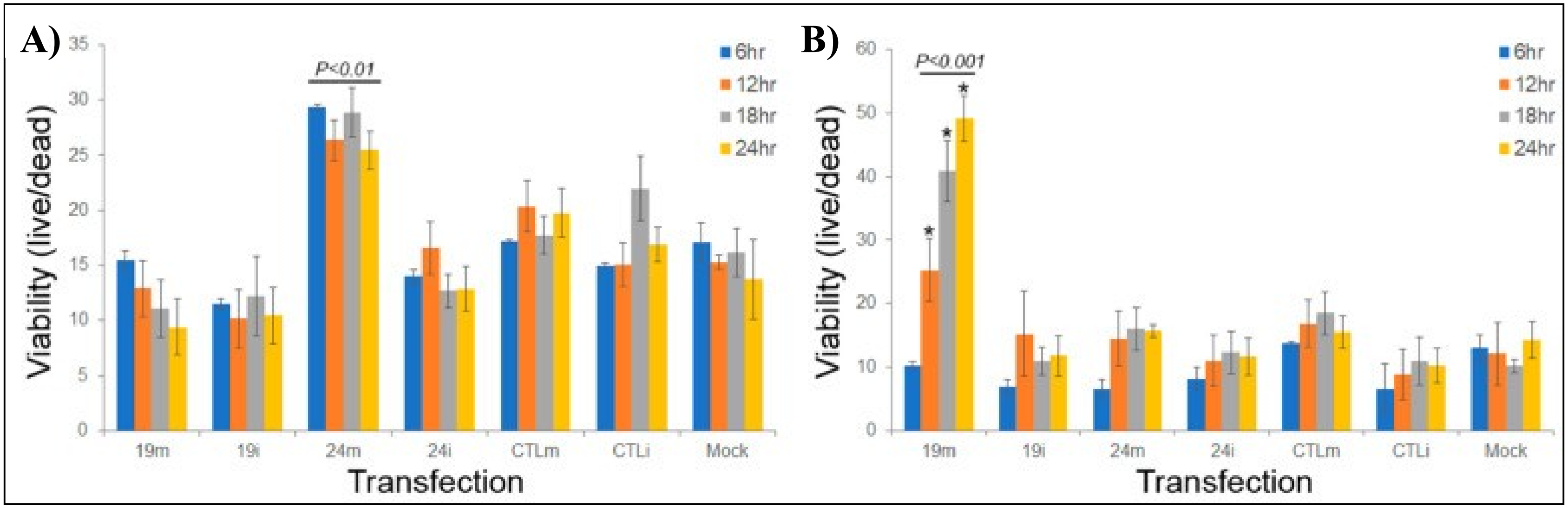
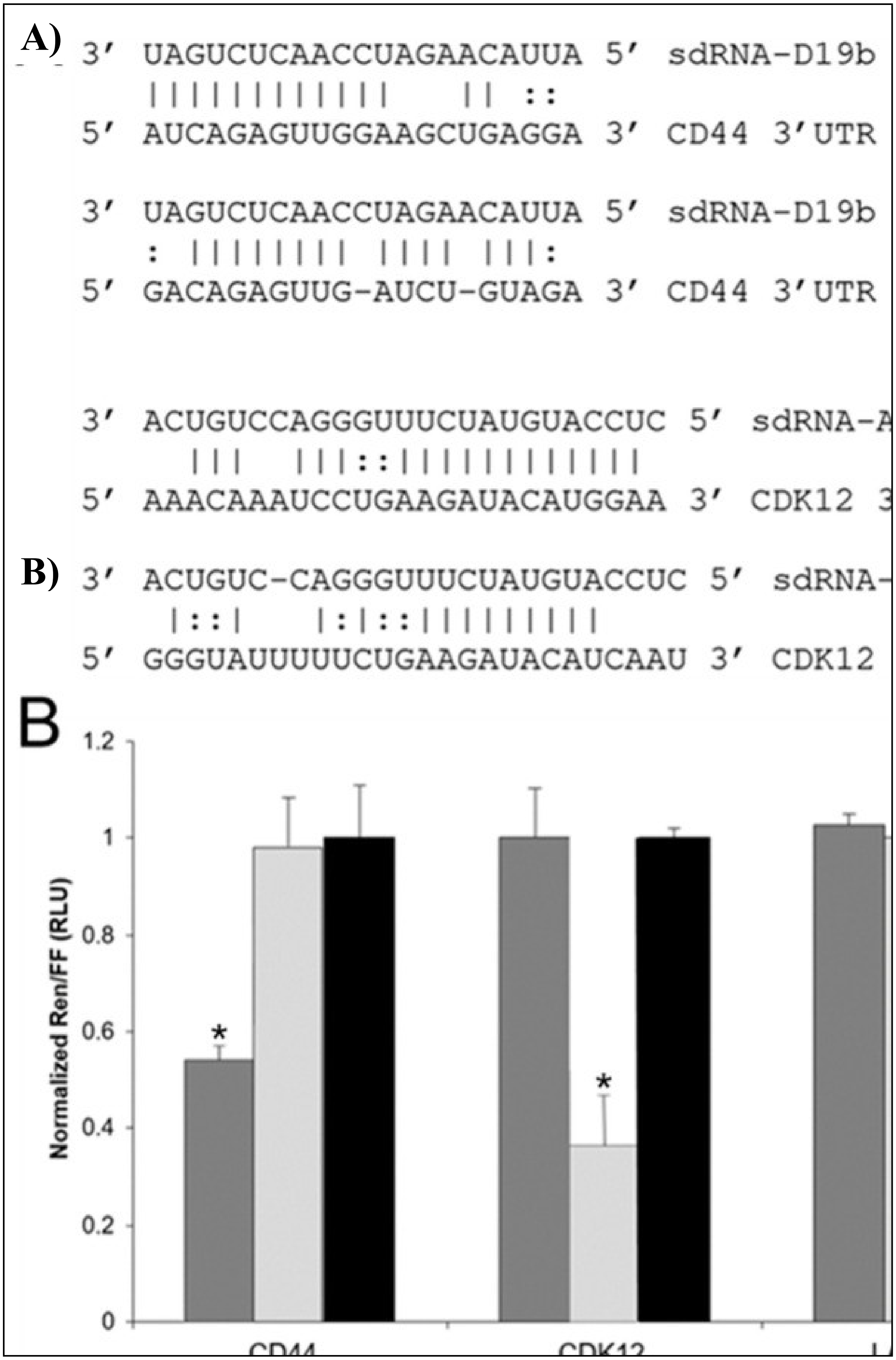


Figure 8. SdRNA overexpression protects PC3 cells from chemotherapeutic agents. Cells were cultured in 24-well plates and transfected at 70% confluency with mimics or inhibitors. Following transfection, cells were treated with **A)** paclitaxel (5 nM) or **B)** dasatinib (50 nM). Cell death was quantified every 6 h for 24 h total using ImageJ and methylene blue dead cell staining. 19 m, sdRNA-D19b mimic; 19i, sdRNA-D19b inhibitor; 24 m, sdRNA-A24 mimic; 24i, sdRNA-A24 inhibitor; CTLm, scrambled mimic; CTLi, scrambled inhibitor; Mock, vehicle-treated control. ($n \geq 3$). * indicates $p < 0.001$; p-values by unpaired two-tailed t-test as compared to Mock.

2.3.5 sdRNA-D19b and sdRNA-A24 Target the 3′UTRs of CD44 and CDK12, Respectively

Putative mRNA targets were identified using a strategy previously developed by our group[36] that (1) limits potential targets to those predicted by multiple algorithms and (2) confirms target mRNAs are expressed in PC3 cell RNA-seq datasets. Employing this streamlined methodology readily yielded marked candidates for both sdRNA-D19b and -A24 regulation (APPENDIX E,F), and we selected the most notable of these for further validation *in vitro*. The highest scoring target mRNA identified for sdRNA-D19b (containing two notable 3'UTR complementarities) is a known regulator of PCa proliferation and migration and the cell adhesion glycoprotein CD44[64] (Figure 9A, top). Similarly, the highest scoring target mRNA identified for sdRNA-A24 (also containing two notable 3'UTR complementarities, one bearing 100% complementarity to sdRNA-A24 nucleotides 2 through 18) is a known tumor suppressor mutated in ~6% of patients with metastatic CRPC, CDK12[65,66] (Figure 9A, bottom). Importantly, sdRNA-D19b mimic transfection of PC3 cells' silenced expression from a standard Renilla luciferase reporter containing the principle putative CD44 3′UTR target sites by more than 40%, as compared to the control and sdRNA-A24 mimic transfections. Conversely, the sdRNA-A24 mimic transfection of PC3 cells' silenced expression from a standard Renilla luciferase reporter containing the principle CDK12 3′UTR target sites by ~70%, as compared to control and sdRNA-D19b mimic transfections (Figure 9B).

Figure 9. SdRNA-D19b and sdRNA-A24 mRNA targets. A) Alignments between putative 3′UTR target sites with sdRNAs-D19b (top) and -A24 (bottom). Vertical lines indicate Watson-Crick basepair. Dotted lines indicate G:U basepair. TS1, target site 1. TS2, target site 2. **B) SdRNAs-D19b and -A24 specifically repress luciferase expression from mRNAs containing CD44 and CDK12 target sites in their 3′UTRs.** SdRNA mimics and luciferase reporters with target sequences (bottom) and/or controls (LACTA refers to beta galactosidase control sequence) were constructed and cotransfected, as previously described[36]. * indicates $p < 0.01$; p-values by unpaired two-tailed t-test as compared to LACTA excepting LACTA compared to CD44.



```
A) 3’ UAGUCUCAACCUAGAACAUUA 5’ sdRNA-D19b
      ||||||||||||   ||  ::
   5’ AUCAGAGUUGGAGCUGAGGA 3’ CD44 3’UTR

   3’ UAGUCUCAACCUAGAACAUUA 5’ sdRNA-D19b
      : |||||||| |||| |||:
   5’ GACAGAGUUG-AUCU-GUAGA 3’ CD44 3’UTR


   3’ ACUGUCCAGGGUUUCUAUGUACCUC 5’ sdRNA-A
        |||  |||::||||||||||||
   5’ AAACAAAUCCUGAAGAUACAUGGAA 3’ CDK12 3

B) 3’ ACUGUC-CAGGGUUUCUAUGUACCUC 5’ sdRNA-
       |::|    |:|::|||||||||
   5’ GGGUAUUUUUCUGAAGAUACAUCAAU 3’ CDK12
```

CHAPTER III

LONG G-QUADRUPLEX (LG4) DNA REGIONS FORM TOPOLOGICALLY ASSOCIATED DOMAINS AND POTENTIALLY FUNCTION AS SUPER ENHANCERS

3.1 Brief Overview

G4 motifs are located throughout the genome, concentrating primarily in the promoter regions of genes[67]. While minimal G4s have been studied extensively, our lab published the first genome-wide analysis focused on identifying and characterizing long G4 (LG4) regions[40]. We identified 301 such LG4s and provided evidence that they are concentrated in gene-rich regions near enhancers and that they can form intramolecular contacts termed "G4 kissing" (G4K) loops. It has been well established that minimal G4s play a role in gene regulation, though the precise mechanism has yet to be fully elucidated[37,39]. We proposed that LG4s interact with local genetic elements to form TADs that likely function as super enhancers.

To validate this claim and make precise predictions of the LG4 interactome, this chapter outlines the development and implementation of a custom Hi-C analysis pipeline whose purpose is to mine chromatin conformation data to identify LG4 interactions. Hi-C data captures all DNA interactions globally within a cell in the form of read-pairs: two interacting genomic loci captured in a read together[49]. By constructing a pipeline tailored to the analysis LG4 interactions within a putative TAD region, the work presented in this chapter has defined the Chromosome 5 LG4 TAD and established a reliable pipeline for the Borchert Lab to leverage for Hi-C analysis.

3.2 Methods

Unless stated otherwise, all data processing and analysis was performed on the high-performance computing (HPC) platform provided by the Alabama Supercomputer Authority (https://www.asc.edu). Additionally, development and quality testing was conducted using Rstudio[59]. All sequence alignments were manually validated, and peak loci were confirmed via Ensembl genome browser during development.

3.2.1 Development of Novel Approach

We have developed a custom analysis pipeline to enable efficient and high-confidence predictions of LG4 interactions using Hi-C NGS data. The pipeline can be divided into three primary steps: alignment of HiC reads to LG4 sequences, alignment of unknown read pairs to the putative LG4 TAD, and peak calling (Figure 10).

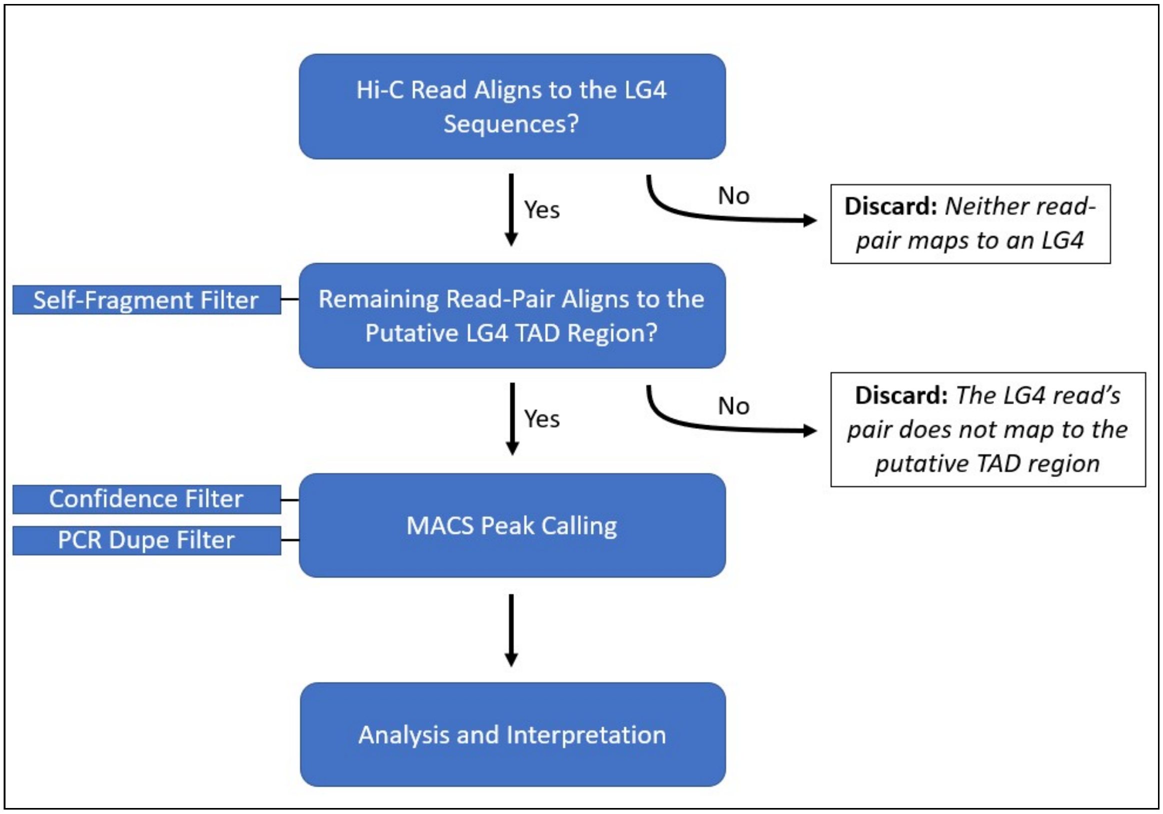

Hi-C Read Aligns to the LG4 Sequences?
Yes
No
Discard: *Neither read-pair maps to an LG4*
Self-Fragment Filter
Remaining Read-Pair Aligns to the Putative LG4 TAD Region?
Yes
No
Discard: *The LG4 read's pair does not map to the putative TAD region*
Confidence Filter
PCR Dupe Filter
MACS Peak Calling
Analysis and Interpretation

Figure 10. Hi-C Analysis Workflow. An overview of the custom Hi-C analysis pipeline is provided. Input Hi-C data is first aligned to the LG4 sequence to select reads where at least one read pair maps to the LG4. The remaining reads are masked to exclude the LG4-aligning read fragment and then aligned to the LG4 TAD region to select reads for which the non-LG4 read pair corresponds to the LG4 TAD. All resulting alignments are subjected to peak calling which serves the dual purposes of removing duplicate reads to reduce PCR bias and resolving the raw interaction data into distinct peaks prioritized by confidence (fold enrichment score). The final peaks are then analyzed to determine whether an LG4 TAD is formed, and the nature of the contacts made within the TAD.

Hi-C Sample Selection and Pre-Processing: Raw HiC data obtained either by the user or from a publicly available database (e.g. NCBI SRA) is stored in a shared Borchert Lab directory on the Alabama Supercomputer Cluster. All files are converted to FASTA format and zipped to prepare for later analysis steps.

Hi-C Read Alignment to LG4 Sequences: The Basic Local Alignment Search Tool (BLAST+) is used to align all Hi-C files to a given LG4 sequence. BLAST performs sequence similarity alignment of query sequences to a given database, resulting in alignments or "hits" that pass a predetermined confidence threshold[42]. BLAST+

applies a “word” extension heuristic where small sequences from a query sequence (referred to as a ”word”) are aligned to the reference database. Perfect word alignments are extended and scored based on the identity between query and database as well as the length of the alignment. This method of local alignment affords significantly less computational demand than global alignment, an advantage required for large-scale analyses such as this. BLAST parameters are configured to allow only perfect alignments, an expect value (e-values) threshold of 0.01, one target sequence per hit, and a word-size window of 6 nucleotides (nt). Additionally, hits are filtered to discard all alignments <30 nt to filter out alignments to short repeat regions. The resulting reads are comprised of at least one read pair that aligns to the LG4 of interest.

Read-Pair Alignment to the Putative LG4 TAD Region: The majority of enhancer-promoter interactions occur over a maximum distance of 500 kb[4]. As such, putative LG4 TAD regions are designed by retrieving the 500 kb up and downstream flank sequences for each LG4. To preclude self-alignments, the LG4 sequence as well as the 5 kb up and downstream flank are masked within the putative LG4 TAD (APPENDIX G). The analysis pipeline is then executed (APPENDIX H). First, all alignments to the LG4 in question are passed through a ≥30 nt cutoff filter and compiled (APPENDIX I,J). Before aligning the remaining Hi-C reads to this putative LG4 TAD database, the full length Hi-C read is retrieved (APPENDIX K,L) and the read-pair that corresponds to the LG4 is masked (APPENDIX M) to limit the search to only the unidentified read-pair. BLAST+ alignment of the masked Hi-C query to the putative LG4 TAD database is conducted using the same parameters as before (APPENDIX N,O). Hits are subsequently filtered to discard all that fail a <50 nt threshold. All remaining hits are compiled into an output table for the user that provides the sample of origin, read ID, read sequence, LG4 sequence, target sequence, and target chromosomal position for each Hi-C read that passed analysis. Additionally, a file in browser extensible data (bed) format is produced that contains the chromosomal position and read ID for each target sequence to be used in the next analysis step (APPENDIX P).

Peak Calling: Bed files are then passed to the Model-Based Analysis of ChIP-Seq (MACS) program for peak calling[68] (APPENDIX Q). Peak-calling resolves given reads into distinct regions of genomic enrichment. Designed for ChIP-Seq, MACS is well-suited to resolve our pipeline's Hi-C alignments into peaks within the putative LG4 TAD genomic region[69]. MACS first pre-processes the input data by removing duplicate reads that share the same start and stop position to account for PCR amplification bias as well as overrepresentation by a sample if multiple datasets are used. The optional step of model-building is disabled, as this feature is designed for ChIP-seq analysis and is not relevant to our Hi-C data. MACS searches for peaks by detecting hits that pile-up within a given genomic region. Peaks are called if a region has a p-value < $10e^{-5}$ compared to local hit bias (λ). A fold-enrichment score (FE) is calculated for each potential peak based on the raw number of hits in a peak (pile-up) divided by the variance λ. The resulting peaks are filtered to reject those with a FE < 6.0. The final peaks are exported as a table that reports information about each significant peak and a bed file containing each significant peak and its chromosomal position are exported for the next step in analysis.

Analysis and Interpretation: The presence or absence of significant peaks reveals whether the LG4 in question makes physical contact within its putative TAD. The output data can be further interpreted to determine which genomic features are in contact with the LG4. The table can be used to manually search genome browsers such as Ensembl[70] and identify the LG4 interaction partner. Alternatively (and preferentially), the exported bed file can be loaded directly into the Integrative Genomics Viewer (IGV)[71] where results can be visualized against the genome. This second method allows for more rapid analysis of the results as a whole and can be easily modified to compare peaks against the user's genomic features of interest by creating a custom bed file and uploading it to IGV alongside the HiC peaks.

3.2.2 Utilizing the Borchert Lab Hi-C Pipeline to Define the Chromosome 5 LG4 TAD

The Hi-C pipeline is used here to identify significant interactions within a putative LG4 TAD located on Chromosome 5.

Datasets: All Hi-C datasets were acquired from NCBI Sequence Read Archive (SRA) (https://www.ncbi.nlm.nih.gov/sra). All data was converted to FASTA format using the fastx toolkit (https://github.com/agordon/fastx_toolkit).

The LG4 sequence was obtained using Ensembl biomart[70] reference GRCh38, and formated into a blast database using BLAST+. The putative LG4 TAD sequence was obtained from Ensembl biomart[70] reference GRCh38 and formatted into a blast database using BLAST+.

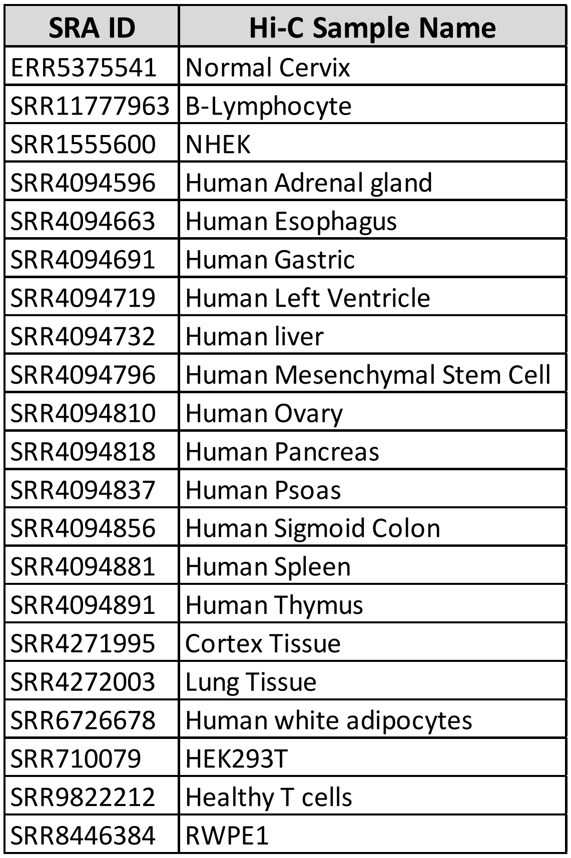
HiC Analysis: The pipeline was executed in a linux environment hosted on the Alabama Supercomputer Authority's HPC platform (https://www.asc.edu).

3.3 Results

3.3.1 Hi-C Analysis

To assess whether LG4s are capable of forming TADs and potentially function as super enhancers, we decided to focus on an LG4 located on Chromosome 5 that is embedded in a locus of genes whose dysregulation has been implicated in disease states including cystic fibrosis[72,73]. Additionally, genes located within a reasonable enhancer interaction range of 500 kb centered on the LG4 frequently fuse with one another. Using TumorFusions[74] to mine TCGA data for cancer-associated gene fusions, four unique Tier 1 (highest confidence) gene fusions were detected within the region (Appendix R). This indicates that these genes are brought into close proximity, a process which could potentially be mediated by an LG4 TAD according to our proposed LG4 super enhancer model (Figure 5A). In order to assess the status of LG4 TAD interactions under normal, non-pathological conditions, 21 Hi-C data files collected from diverse normal human tissue samples were downloaded from the NCBI SRA for analysis (Table 2).

Table 2. Hi-C Samples Acquired from NCBI SRA



SRA ID	Hi-C Sample Name
ERR5375541	Normal Cervix
SRR11777963	B-Lymphocyte
SRR1555600	NHEK
SRR4094596	Human Adrenal gland
SRR4094663	Human Esophagus
SRR4094691	Human Gastric
SRR4094719	Human Left Ventricle
SRR4094732	Human liver
SRR4094796	Human Mesenchymal Stem Cell
SRR4094810	Human Ovary
SRR4094818	Human Pancreas
SRR4094837	Human Psoas
SRR4094856	Human Sigmoid Colon
SRR4094881	Human Spleen
SRR4094891	Human Thymus
SRR4271995	Cortex Tissue
SRR4272003	Lung Tissue
SRR6726678	Human white adipocytes
SRR710079	HEK293T
SRR9822212	Healthy T cells
SRR8446384	RWPE1

The Hi-C Analysis Pipeline searched each of the 21 samples for reads that constitute interactions between the Chromosome 5 LG4 and genomic loci within 500 kb of the LG4. MACS peak calling yielded 76 unique interaction loci, 17 of which passed the fold enrichment score (FE) cutoff of 6.0 and were counted as significant (Figure 11). These LG4-interacting regions included 8 unique protein-coding genes: PLEKHG4, SDHA, PDCD6, AHRR, EXOC3, SLC9A3, CEP72, and TRIP13. The Ensembl Regulatory Build[70], which contains positional information on experimentally-defined regulatory elements in the human genome, was downloaded to assess whether the LG4-interacting regions represent regulatory elements. In total, the Chromosome 5 LG4 TAD includes interactions between the LG4 and 6 unique CTCF binding sites, 2 unique enhancers, and 10 unique promoters/promoter flanking regions (Table 3).

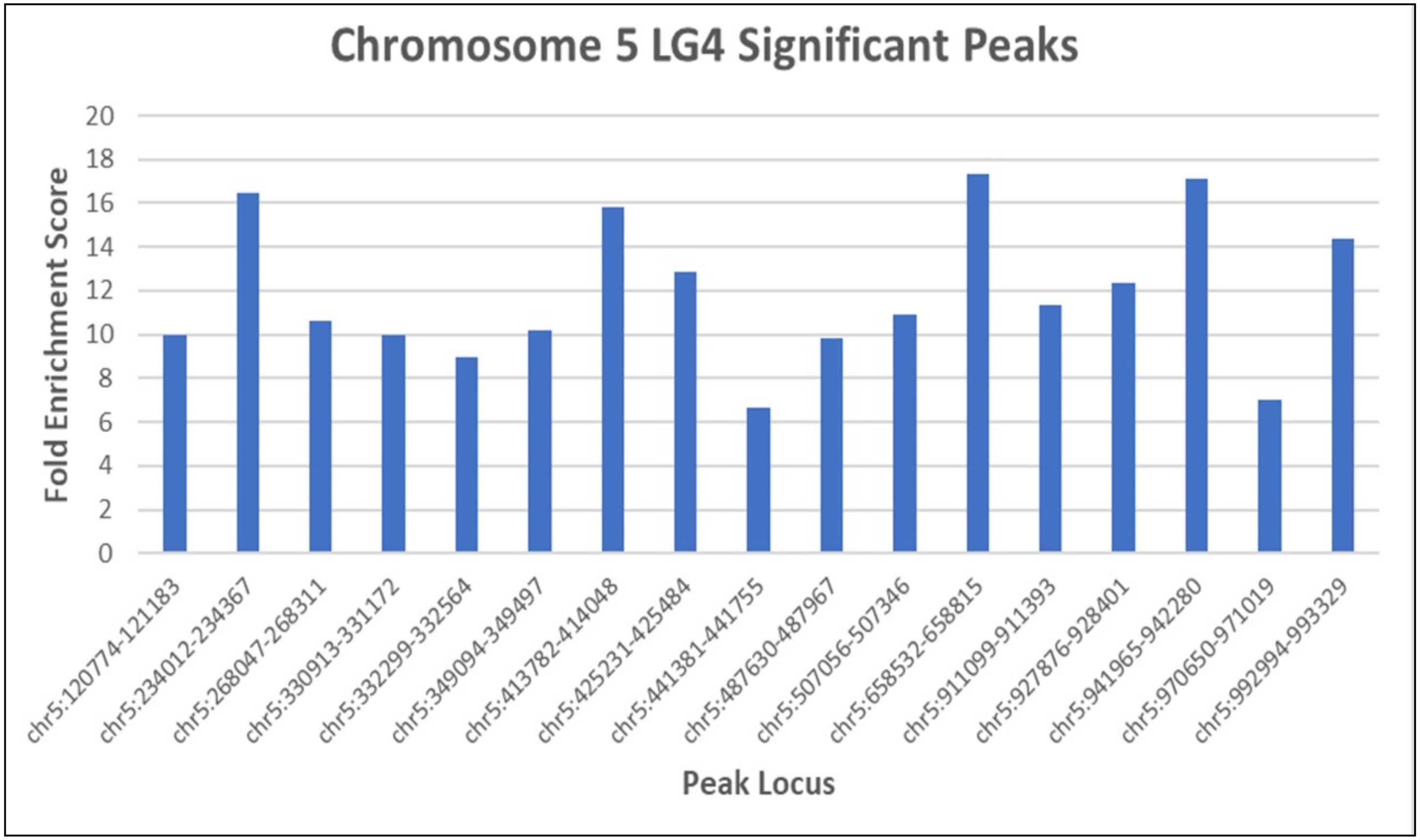
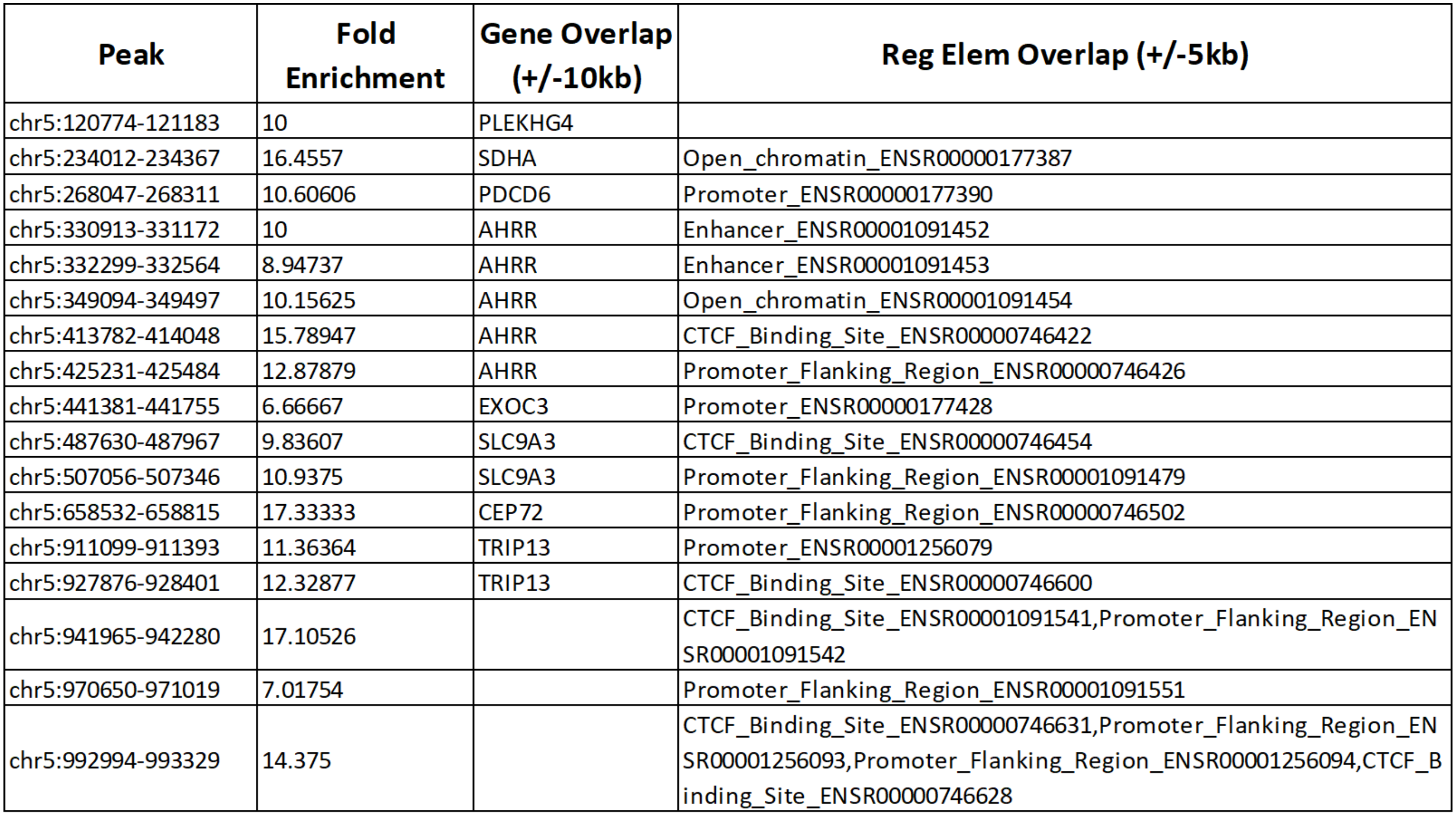


Figure 11. Chr 5 LG4 Significant Interaction Loci. Following analysis using the Borchert Lab Hi-C pipeline and MACS peak calling, read fragments that align to the Chromosome 5 LG4 TAD were resolved into distinct peaks with a confidence cutoff FE ≥ 6. Significantly enriched peak loci are reported as " chromosome : start position – stop position" .

Table 3. Chromosome 5 LG4 TAD Interactions with Genes and Regulatory Elements



Peak	Fold Enrichment	Gene Overlap (+/-10kb)	Reg Elem Overlap (+/-5kb)
chr5:120774-121183	10	PLEKHG4	
chr5:234012-234367	16.4557	SDHA	Open_chromatin_ENSR00000177387
chr5:268047-268311	10.60606	PDCD6	Promoter_ENSR00000177390
chr5:330913-331172	10	AHRR	Enhancer_ENSR00001091452
chr5:332299-332564	8.94737	AHRR	Enhancer_ENSR00001091453
chr5:349094-349497	10.15625	AHRR	Open_chromatin_ENSR00001091454
chr5:413782-414048	15.78947	AHRR	CTCF_Binding_Site_ENSR00000746422
chr5:425231-425484	12.87879	AHRR	Promoter_Flanking_Region_ENSR00000746426
chr5:441381-441755	6.66667	EXOC3	Promoter_ENSR00000177428
chr5:487630-487967	9.83607	SLC9A3	CTCF_Binding_Site_ENSR00000746454
chr5:507056-507346	10.9375	SLC9A3	Promoter_Flanking_Region_ENSR00001091479
chr5:658532-658815	17.33333	CEP72	Promoter_Flanking_Region_ENSR00000746502
chr5:911099-911393	11.36364	TRIP13	Promoter_ENSR00001256079
chr5:927876-928401	12.32877	TRIP13	CTCF_Binding_Site_ENSR00000746600
chr5:941965-942280	17.10526		CTCF_Binding_Site_ENSR00001091541,Promoter_Flanking_Region_ENSR00001091542
chr5:970650-971019	7.01754		Promoter_Flanking_Region_ENSR00001091551
chr5:992994-993329	14.375		CTCF_Binding_Site_ENSR00000746631,Promoter_Flanking_Region_ENSR00001256093,Promoter_Flanking_Region_ENSR00001256094,CTCF_Binding_Site_ENSR00000746628

Strikingly, of the numerous contacts between the LG4 and its TAD, every target locus overlaps at least one gene or regulatory element. This supports our proposed model suggesting that the LG4 functions as a super enhancer, regulating genes within its TAD by regulating access to promoters and enhancers. Informed by the results of our Hi-C analysis, we have constructed a TAD map that documents the physical interactions established by this LG4 within its TAD (Figure 12).

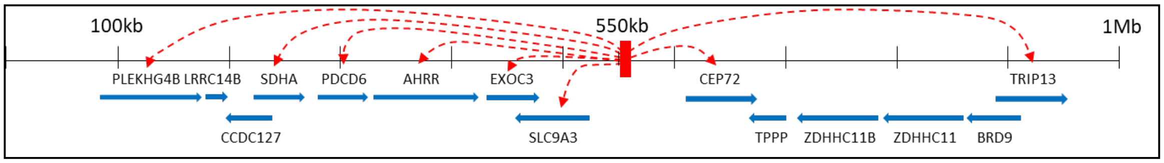


Figure 12. The Chromosome 5 TAD. Graphical representation of the Chromosome 5 TAD gene interactions based on Hi-C data analyzed by the Borchert lab Hi-C pipeline.

CHAPTER IV

DISCUSSION AND CONCLUSIONS

Epigenetic organization of chromatin, promoter and enhancer accessibility, transcription factor expression and activation, and microRNA-loaded RISC post-transcriptional silencing constitute some of the best known and studied regulatory phenomena. Disease states such as cancer can result from aberrant regulation by just a few of these networks, underpinning their importance in the maintenance of cellular homeostasis[75,76]. While there is much that we do understand, still there are novel regulatory elements and pathways uncovered every year. In recent years, phase-separated transcription bubbles have emerged as likely drivers of gene expression and transcriptional co-regulation[77]. Novel links between miRNAs and malignant phenotypes are published multiple times a year as reviewed by Peng *et al.*[31] in 2016 and Si *et al.*[78] in 2019. Clearly there is a need for further study to characterize regulatory elements. The work presented here has expanded our knowledge of regulatory sdRNAs and established the LG4 TAD as a novel regulatory element. The former constitutes a relatively novel phenomenon where ncRNAs previously considered to only carry out "housekeeper" functions are now understood to be processed into regulatory ndRNAs[20,79]. The latter is a completely novel phenomenon first proposed by our lab where long G4-rich regions of the genome interact with nearby genes and regulatory elements to form TADs potentially functioning as super enhancers[40].

Numerous miRNAs have now been characterized as master regulators of oncogenes and tumor suppressors[80,81]. With the preponderance of studies implicating miRNAs in virtually all cancer types, aberrant miRNA expression has been rightfully proposed to constitute a hallmark of cancer[75]. Similarly, over the past decade, a growing number of studies have suggested sdRNAs could likewise play significant roles in malignancy[82]. In light of this, we explored the potential for sdRNAs to function similarly in other cancer types, leading to the identification of 38 significantly differentially expressed sdRNAs in CRPC and the characterization of direct roles for sdRNAs-D19b and -A24 in modulating CRPC. A core characteristic of CRPC is enhanced metastasis, a factor largely responsible for the marked morbidity and high death rate among men in the US[83]. As such, the striking phenotypic consequences associated with manipulating sdRNA-D19b and sdRNA-A24 expressions described in this work (e.g., sdRNA-D19b overexpression results in an ~100% increase in PC3 migration) strongly indicate an important role occupied by sdRNAs in promoting CRPC malignant traits.

Excitingly, the highest scoring target mRNA identified for sdRNA-D19b is a known regulator of PCa proliferation and migration, namely, the cell adhesion glycoprotein CD44[64] (Figure 9A, top). Similarly, the highest scoring target mRNA identified for sdRNA-A24 is CDK12, a known tumor suppressor mutated in ~6% of patients with metastatic castration-resistant PCa[65,66] (Figure 9A, bottom). Of note, a loss of CD44 expression is frequently associated with enhanced PCa progression and markedly promotes PCa metastasis[84]. In agreement with this, our work strongly suggests that sdRNA-D19b can directly suppress CD44 expression, and importantly demonstrates that sdRNA-D19b overexpression markedly increases PC3 cell migration in vitro. Also of note, the loss of the sdRNA-A24 target gene CDK12 in CRPC defines a clinically relevant subclass of CRPC that is characteristically hyper-aggressive[65]. CDK12 is a cyclin-dependent kinase that promotes genomic stability through various DNA repair pathways, and a loss of CDK12 expression in PCa enhances genomic mutagenicity, resulting in an aggressive and treatment-resistant phenotype[85]. In this study, we demonstrated that CDK12 is directly regulated by sdRNA-A24, and that sdRNA-A24 overexpression significantly desensitizes PC3 cells to treatment with the microtubule-stabilizing agent, paclitaxel. Interestingly, miR-613 was recently reported to similarly directly modulate paclitaxel resistance via targeting CDK12 in human breast cancer[86].

In summary, with tools such as SURFR[47,48] having only recently made the intensive interrogation of sdRNAomes widely available, we suggest that the identification of relevant sdRNA contributions to malignancy will accelerate in the near future and lead to the development of novel therapies and diagnostics based on sdRNAs. It is important to note, however, that direct validation and characterization of sdRNA-D19b and/or -A24 misexpressions (in addition to larger patient sample cohorts) will clearly be required to establish the utility of one or both of these sdRNAs as viable biomarkers. In short, considerably more extensive groundwork must be laid before these (or any) sdRNAs can be fashioned as tractable drug targets for cancer therapy or as diagnostic/prognostic markers similar to cutting-edge miRNA translational applications[87,88]. That said, we do suggest the work presented here does begin to expand the CRPC regulatory landscape to include sdRNAs as potential new therapeutic targets and/or prognostic indicators through identifying sdRNA-D19b and sdRNA-A24 as likely contributors to CRPC, an aggressive molecular subtype of PCa for which there are currently only limited options for therapy.

In addition to the discovery of novel sdRNAs involved in post-transcriptional regulation of tumor suppressors in CRPC, we also provide evidence for a novel regulatory element: the LG4 TAD super enhancer. In a previous publication, we demonstrated that long stretches of G-quadruplexes can be found in gene-rich enhancer-dense regions of the human genome[40]. Additionally, we showed that these LG4s can form interacting structures that would have implications on chromatin accessibility and therefore potentially impact gene expression. We proposed a model by which LG4s form TADs, isolating regions of DNA to co-regulate expression (Figure 5A). Here, we have verified this claim by analyzing publicly available Hi-C data to confirm LG4-TAD contacts. Hi-C reads consist of DNA interaction pairs, or read pairs, that capture DNA-DNA interactions genome-wide[55]. We developed a custom Hi-C analysis pipeline that allowed for the analysis of multiple Hi-C datasets in a single run to produce high-confidence interaction predictions for a given LG4.

Following Hi-C analysis, an LG4 on Chromosome 5 was found to interact with 8 unique genes, 6 unique CTCF binding sites, 2 unique enhancers, and 10 unique promoters/promoter flanking regions (Table 3). These results confirm the ability of LG4s to associate with nearby genes and regulatory elements, forming LG4 TADs. Further, all significant contacts were associated with at least one gene or regulatory element, strongly suggesting that the LG4 TAD represents a previously undescribed level of gene regulation that is driven by the formation of LG4-DNA interactions.

The work presented here defines 38 sdRNAs that contribute to the CRPC regulatory sdRNAome, two of which (-A24 and -d19B) were fully characterized and implicated as tumor promoting sdRNAs via a biologically coherent mechanism. Also, a completely novel level of gene regulation, the LG4 TAD, was validated using a high-throughput custom Hi-C analysis pipeline. Taken together, these novel genomic regulators support the notion that there are mechanisms of gene regulation that remain uncharacterized. This has consequences both for our basic understanding of gene expression and in pathology, as currently undiscovered regulatory elements may be responsible for the more complex and treatment-recalcitrant diseases. In the case of sdRNAs such as sdRNA-A24 and sdRNA-D19B, it is urgent that regulatory ndRNAs such as these be assembled into panels for diagnostic and prognostic studies. It has been demonstrated that miRNAs are readily available in patient serum, and this is true as well for ndRNAs such as sdRNAs[89,90]. With a wealth of potential prognostic/diagnostic ndRNAs being uncovered yearly, the next logical step is to assess their ability to benefit patients as a panel. Additionally, as with miRNAs, sdRNA-based therapeutics deserve further study. While regulatory ncRNA based therapeutics may still be out of reach in the clinic, the advent of mRNA vaccines during Covid may springboard miRNA and sdRNA-based therapies for cancer and other diseases.

The implications of LG4 TADs are numerous. In the context of disease, it could be that these regulatory elements break down in the disease state leading to a loss of normal gene regulation. This could potentially result from mutations that alter the LG4-TAD interaction thereby changing the regulatory landscape within the TAD. Further research is needed to understand the impact of LG4 TAD loss, but much can be gleaned from our understanding of TADs in general[29]. In our lab, the Hi-C analysis pipeline's LG4 interaction predictions are being used to inform primer design for our proprietary assay EQUIP-seq. This method is designed to pull down a promoter region and identify

DNA interactions. By performing EQUIP-seq on the predicted LG4 interaction partners we can confirm the LG4 TAD *in vitro*, thus providing an assay perfectly suited to assess LG4 function and dysfunction in the context of disease.

In conclusion, through the work presented we have identified and characterized novel regulators of human genes. We hope that this work bolsters our understanding of gene regulation and highlights the fact that there is still much that we do not understand, necessitating more exploratory studies such as this one. Additionally, the specific regulatory elements identified here provide novel targets for diagnostic, prognostic, and therapeutic developments aimed at diseases driven by aberrant gene regulation.

REFERENCES

1. Salzberg SL. Open questions: how many genes do we have? *BMC Biol*. 2018;16(1):94. doi:10.1186/S12915-018-0564-X

2. Polyak K, Meyerson M. *Overview: Gene Structure*. 6th ed. (Kufe D, Pollock R, Weichselbaum R, eds.). Holland-Frei; 2003. Accessed June 1, 2022. https://www-ncbi-nlm-nih-gov.libproxy.usouthal.edu/books/NBK12983/

3. Cramer P. Organization and regulation of gene transcription. *Nature*. 2019;573(7772):45-54. doi:10.1038/s41586-019-1517-4

4. van Arensbergen J, van Steensel B, Bussemaker HJ. In search of the determinants of enhancer–promoter interaction specificity. *Trends Cell Biol*. 2014;24(11):695-702. doi:10.1016/J.TCB.2014.07.004

5. Bentley DL. Coupling mRNA processing with transcription in time and space. *Nat Rev Genet.* 2014;15(3):163-175. doi:10.1038/nrg3662

6. Lee TI, Young RA. Transcriptional regulation and its misregulation in disease. *Cell*. 2013;152(6):1237-1251. doi:10.1016/J.CELL.2013.02.014

7. Kouzarides T. Chromatin modifications and their function. *Cell*. 2007;128(4):693-705. doi:10.1016/J.CELL.2007.02.005

8. Hnisz D, Day DS, Young RA. Insulated neighborhoods: structural and functional units of mammalian gene control. *Cell*. 2016;167(5):1188-1200. doi:10.1016/J.CELL.2016.10.024

9. Cech TR, Steitz JA. The noncoding RNA revolution—trashing old rules to forge new ones. *Cell*. 2014;157(1):77-94. doi:10.1016/J.CELL.2014.03.008

10. Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*. 1993;75(5):843-854. doi:10.1016/0092-8674(93)90529-Y

11. Calin GA, Dumitru CD, Shimizu M, Bichi R, Zupo S, Noch E, Aldler H, Rattan S, Keating M, Rai K, Rassenti L, Kipps T, Negrini M, Bullrich F, Croce CM. Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proc Natl Acad Sci*. 2002;99(24):15524-15529. doi:10.1073/PNAS.242606799

12. Annese T, Tamma R, De Giorgis M, Ribatti D. microRNAs biogenesis, functions and role in tumor angiogenesis. *Front Oncol*. 2020;10:581007. doi:10.3389/FONC.2020.581007

13. Pratt AJ, MacRae IJ. The RNA-induced silencing complex: a versatile gene-silencing machine. *J Biol Chem*. 2009;284(27):17897-17901. doi:10.1074/JBC.R900012200

14. Kuscu C, Kumar P, Kiran M, Su Z, Malik A, Dutta A. tRNA fragments (tRFs) guide Ago to regulate gene expression post-transcriptionally in a Dicer-independent manner. *RNA*. 2018;24(8):1093-1105. doi:10.1261/RNA.066126.118

15. Guan L, Grigoriev A. Computational meta-analysis of ribosomal RNA fragments: potential targets and interaction mechanisms. *Nucleic Acids Res*. 2021;49(7):4085-4103. doi:10.1093/NAR/GKAB190

16. Kiss-László Z, Henry Y, Bachellerie JP, Caizergues-Ferrer M, Kiss T. Site-specific ribose methylation of preribosomal RNA: a novel function for small nucleolar RNAs. *Cell*. 1996;85(7):1077-1088. doi:10.1016/S0092-8674(00)81308-

2

17. Ganot P, Bortolin ML, Kiss T. Site-specific pseudouridine formation in preribosomal RNA is guided by small nucleolar RNAs. *Cell*. 1997;89(5):799-809. doi:10.1016/S0092-8674(00)80263-9

18. Wajahat M, Bracken CP, Orang A. Emerging functions for snoRNAs and snoRNA-derived fragments. *Int J Mol Sci*. 2021;22(19):10193. doi:10.3390/IJMS221910193

19. Kawaji H, Nakamura M, Takahashi Y, Sandelin A, Katayama S, Fukuda S, Daub CO, Kai C, Kawai J, Yasuda J, Carninci P, Hayashizaki Y. Hidden layers of human small RNAs. *BMC Genomics.* 2008;9:157. doi:10.1186/1471-2164-9-157

20. Coley AB, DeMeis JD, Chaudhary NY, Borchert GM. Small nucleolar derived RNAs as regulators of human cancer. *Preprints*. 2022;2022060005. doi:10.20944/PREPRINTS202206.0005.V1

21. Richard P, Kiss AM, Darzacq X, Kiss T. Cotranscriptional recognition of human intronic box H/ACA snoRNAs occurs in a splicing-independent manner. *Mol Cell Biol*. 2006;26(7):2540-2549. doi:10.1128/MCB.26.7.2540-2549.2006

22. Leverette RD, Andrews MT, Maxwell ES. Mouse U14 snRNA is a processed intron of the cognate hsc70 heat shock pre-messenger RNA. *Cell*. 1992;71(7):1215-1221. doi:10.1016/S0092-8674(05)80069-8

23. Fragapane P, Prislei S, Michienzi A, Caffarelli E, Bozzoni I. A novel small nucleolar RNA (U16) is encoded inside a ribosomal protein intron and originates by processing of the pre-mRNA. *EMBO J*. 1993;12(7):2921-2928. doi:10.1002/J.1460-2075.1993.TB05954.X

24. Tycowski KT, Shu M Di, Steitz JA. A mammalian gene with introns instead of exons generating stable RNA products. *Nature*. 1996;379(6564):464-466. doi:10.1038/379464a0

25. Ender C, Krek A, Friedländer MR, Beitzinger M, Weinmann L, Chen W, Pfeffer S, Rajewsky N, Meister G. A human snoRNA with microRNA-like functions. *Mol Cell*. 2008;32(4):519-528. doi:10.1016/J.MOLCEL.2008.10.017

26. Bai B, Yegnasubramanian S, Wheelan SJ, Laiho M. RNA-Seq of the nucleolus reveals abundant SNORD44-derived small RNAs. *PLoS One*. 2014;9(9):e107519. doi:10.1371/JOURNAL.PONE.0107519

27. Claussnitzer M, Cho JH, Collins R, Cox NJ, Dermitzakis ET, Hurles ME, Kathiresan S, Kenny EE, Lindgren CM, MacArthur DG, North KN, Plon SE, Rehm HL, Risch N, Rotimi CN, Shendure J, Soranzo N, McCarthy MI. A brief history of human disease genetics. *Nature*. 2020;577(7789):179-189. doi:10.1038/s41586-019-1879-7

28. The Genetics of Cancer. National Cancer Institute. Updated October 12, 2017. Accessed June 24, 2022. https://www.cancer.gov/about-cancer/causes-prevention/genetics

29. Jia R, Chai P, Zhang H, Fan X. Novel insights into chromosomal conformations in cancer. *Mol Cancer.* 2017;16(1):173. doi:10.1186/S12943-017-0741-5

30. Wu D, Zhang C, Shen Y, Nephew KP, Wang Q. Androgen receptor-driven chromatin looping in prostate cancer. *Trends Endocrinol Metab*. 2011;22(12):474-480. doi:10.1016/J.TEM.2011.07.006

31. Peng Y, Croce CM. The role of MicroRNAs in human cancer. *Signal Transduct Target Ther*. 2016;1:15004. doi:10.1038/SIGTRANS.2015.4

32. Hayes J, Peruzzi PP, Lawler S. MicroRNAs in cancer: biomarkers, functions and therapy. *Trends Mol Med*. 2014;20(8):460-469. doi:10.1016/J.MOLMED.2014.06.005

33. Yao Z, Chen Y, Cao W, Shyh-Chang N. Chromatin-modifying drugs and metabolites in cell fate control. *Cell Prolif*. 2020;53(11):e12898. doi:10.1111/CPR.12898

34. Hanna J, Hossain GS, Kocerha J. The potential for microRNA therapeutics and clinical research. *Front Genet*. 2019;10:478. doi:10.3389/FGENE.2019.00478/BIBTEX

35. Lai EC, Tomancak P, Williams RW, Rubin GM. Computational identification of Drosophila microRNA genes. *Genome Biol*. 2003;4(7):R42. doi:10.1186/GB-2003-4-7-R42

36. Patterson DG, Roberts JT, King VM, Houserova D, Barnhill EC, Crucello A, Polska CJ, Brantley LW, Kaufman GC, Nguyen M, Santana MW, Schiller IA, Spicciani JS, Zapata AK, Miller MM, Sherman TD, Ma R, Zhao H, Arora R, Coley AB, Zeidan MM, Tan M, Xi Y, Borchert GM. Human snoRNA-93 is processed into a microRNA-like RNA that promotes breast cancer cell invasion. *NPJ Breast Cancer*. 2017;3(1). doi:10.1038/S41523-017-0032-8

37. Spiegel J, Adhikari S, Balasubramanian S. The structure and function of DNA G-quadruplexes. *Trends Chem*. 2020;2(2):123-136. doi:10.1016/j.trechm.2019.07.002

38. Gellert M, Lipsett MN, Davies DR. Helix formation by guanylic acid. *Proc Natl Acad Sci U S A*. 1962;48(12):2013-2018. doi:10.1073/PNAS.48.12.2013

39. Hänsel-Hertsch R, Beraldi D, Lensing S V, Marsico G, Zyner K, Parry A, Di Antonio M, Pike J, Kimura H, Narita M, Tannahill D, Balasubramanian S. G-quadruplex structures mark human regulatory chromatin. *Nat Genet*. 2016;48(10):1267-1272. doi:10.1038/ng.3662

40. Williams JD, Houserova D, Johnson BR, Dyniewski B, Berroyer A, French H, Barchie AA, Bilbrey DD, Demeis JD, Ghee KR, Hughes AG, Kreitz NW, McInnis CH, Pudner SC, Reeves MN, Stahly AN, Turcu A, Watters BC, Daly GT, Langley RJ, Gillespie MN, Prakash A, Larson ED, Kasukurthi MV, Huang J, Jinks-Robertson S, Borchert GM. Characterization of long G4-rich enhancer-associated genomic regions engaging in a novel loop:loop "G4 Kissing" interaction. *Nucleic Acids Res*. 2020;48(11):5907-5925. doi:10.1093/NAR/GKAA357

41. Lopez JP, Diallo A, Cruceanu C, Fiori LM, Laboissiere S, Guillet I, Fontaine J, Ragoussis J, Benes V, Turecki G, Ernst C. Biomarker discovery: quantification of microRNAs and other small non-coding RNAs using next generation sequencing. *BMC Med Genomics*. 2015;8:35. doi:10.1186/S12920-015-0109-X

42. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009;10:421. doi:10.1186/1471-2105-10-421

43. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9(4):357-359. doi:10.1038/NMETH.1923

44. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler

transform. *Bioinformatics*. 2009;25(14):1754-1760. doi:10.1093/BIOINFORMATICS/BTP324

45. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550. doi:10.1186/S13059-014-0550-8

46. Hardcastle TJ, Kelly KA. BaySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*. 2010;11:422. doi:10.1186/1471-2105-11-422

47. Kasukurthi MV, Houserova D, Huang Y, Barchie AA, Roberts JT, Li D, Wu B, Huang J, Borchert GM. SALTS – SURFR (sncRNA) and LAGOOn (lncRNA) transcriptomics suite. Published online February 10, 2021. doi:10.1101/2021.02.08.430280

48. Kasukurthi MV, Zhang D, Housevera M, Huang Y, Tan S, Ma B, Li D, Benton R, Lin J, Li S, Borchert GM, Huang J. SURFr: algorithm for identification and analysis of ncRNA-derived RNAs. *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Published online November 1, 2019:1504-1507. doi:10.1109/BIBM47256.2019.8983074

49. Pal K, Forcato M, Ferrari F. Hi-C analysis: from data generation to integration. *Biophys Rev*. 2018;11(1):67-78. doi:10.1007/S12551-018-0489-1

50. Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J. Comprehensive mapping of long-range interactions reveals folding

principles of the human genome. *Science*. 2009;326(5950):289-293. doi:10.1126/SCIENCE.1181369

51. van Berkum NL, Lieberman-Aiden E, Williams L, Imakaev M, Gnirke A, Mirny LA, Dekker J, Lander ES. Hi-C: a method to study the three-dimensional architecture of genomes. *J Vis Exp*. 2010;(39):1869. doi:10.3791/1869

52. Cameron CJF, Dostie J, Blanchette M. HIFI: estimating DNA-DNA interaction frequency from Hi-C data at restriction-fragment resolution. *Genome Biol*. 2020;21(1):11. doi:10.1186/S13059-019-1913-Y

53. Schmid MW, Grob S, Grossniklaus U. HiCdat: a fast and easy-to-use Hi-C data analysis tool. *BMC Bioinformatics*. 2015;16(1):277. doi:10.1186/S12859-015-0678-X

54. Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, Aiden EL. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst*. 2016;3(1):95-98. doi:10.1016/J.CELS.2016.07.002

55. Lajoie BR, Dekker J, Kaplan N. The hitchhiker's guide to Hi-C analysis: practical guidelines. *Methods*. 2015;72:65-75. doi:10.1016/J.YMETH.2014.10.031

56. Subramani A, Alsidawi S, Jagannathan S, Sumita K, Sasaki AT, Aronow B, Warnick RE, Lawler S, Driscoll JJ. The brain microenvironment negatively regulates miRNA-768-3p to promote K-ras expression and lung cancer metastasis. *Sci Rep*. 2013;3:2392. doi:10.1038/SREP02392

57. Cai Q, Zhao A, Ren L, Chen J, Liao K, Wang Z, Zhang W. MicroRNA-1291 mediates cell proliferation and tumorigenesis by downregulating MED1 in prostate cancer. *Oncol Lett*. 2019;17(3):3253-3260. doi:10.3892/OL.2019.9980

58. Fujita K, Nonomura N. Role of androgen receptor in prostate cancer: a review. *World J Mens Health*. 2019;37(3):288-295. doi:10.5534/WJMH.180040

59. Rstudio. Rstudio: integrated development environment for R. Published online 2021. http://www.rstudio.com/

60. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*. 2003;31(13):3406-3415. doi:10.1093/NAR/GKG595

61. Tai S, Sun Y, Squires JM, Zhang H, Oh WK, Liang CZ, Huang J. PC3 is a cell line characteristic of prostatic small cell carcinoma. *Prostate*. 2011;71(15):1668-1679. doi:10.1002/PROS.21383

62. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell*. 2011;144(5):646-674. doi:10.1016/J.CELL.2011.02.013

63. Fares J, Fares MY, Khachfe HH, Salhab HA, Fares Y. Molecular principles of metastasis: a hallmark of cancer revisited. *Signal Transduct Target Ther*. 2020;5(1):28. doi:10.1038/s41392-020-0134-x

64. Li W, Qian L, Lin J, Huang G, Hao N, Wei X, Wang W, Liang J. CD44 regulates prostate cancer proliferation, invasion and migration via PDK1 and PFKFB4. *Oncotarget*. 2017;8(39):65143-65151. doi:10.18632/ONCOTARGET.17821

65. Schweizer MT, Ha G, Gulati R, Brown LC, McKay RR, Dorff T, Hoge ACH, Reichel J, Vats P, Kilari D, Patel V, Oh WK, Chinnaiyan A, Pritchard CC, Armstrong AJ, Montgomery RB, Alva A. CDK12-mutated prostate cancer: clinical outcomes with standard therapies and immune checkpoint blockade. *JCO Precis Oncol*. 2020;4:382-392. doi:10.1200/PO.19.00383

66. Giacinti S, Poti G, Roberto M, Macrini S, Bassanelli M, Di Pietro F, Aschelter

AM, Ceribelli A, Ruggeri EM, Marchetti P. Molecular basis of drug resistance and insights for new treatment approaches in mCRPC. *Anticancer Res*. 2018;38(11):6029-6039. doi:10.21873/ANTICANRES.12953

67. Lago S, Nadai M, Cernilogar FM, Kazerani M, Domíniguez Moreno H, Schotta G, Richter SN. Promoter G-quadruplexes and transcription factors cooperate to shape the cell type-specific transcriptome. *Nat Commun*. 2021;12(1):3885. doi:10.1038/s41467-021-24198-2

68. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W, Liu XS. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*. 2008;9(9):R137. doi:10.1186/GB-2008-9-9-R137

69. Feng J, Liu T, Zhang Y. Using MACS to identify peaks from ChIP-Seq data. *Curr Protoc Bioinformatics*. 2011;Chapter 2:Unit2.14-2.14. doi:10.1002/0471250953.BI0214S34

70. Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Austine-Orimoloye O, Azov AG, Barnes I, Bennett R, Berry A, Bhai J, Bignell A, Billis K, Boddu S, Brooks L, Charkhchi M, Cummins C, Da Rin Fioretto L, Davidson C, Dodiya K, Donaldson S, El Houdaigui B, El Naboulsi T, Fatima R, Giron CG, Genez T, Martinez JG, Guijarro-Clarke C, Gymer A, Hardy M, Hollis Z, Hourlier T, Hunt T, Juettemann T, Kaikala V, Kay M, Lavidas I, Le T, Lemos D, Marugán JC, Mohanan S, Mushtaq A, Naven M, Ogeh DN, Parker A, Parton A, Perry M, Piližota I, Prosovetskaia I, Sakthivel MP, Salam AIA, Schmitt BM, Schuilenburg H, Sheppard D, Pérez-Silva JG, Stark W, Steed E, Sutinen K, Sukumaran R, Sumathipala D, Suner MM, Szpak M, Thormann A, Tricomi FF,

Urbina-Gómez D, Veidenberg A, Walsh TA, Walts B, Willhoft N, Winterbottom A, Wass E, Chakiachvili M, Flint B, Frankish A, Giorgetti S, Haggerty L, Hunt SE, IIsley GR, Loveland JE, Martin FJ, Moore B, Mudge JM, Muffato M, Perry E, Ruffier M, Tate J, Thybert D, Trevanion SJ, Dyer S, Harrison PW, Howe KL, Yates AD, Zerbino DR, Flicek P. Ensembl 2022. *Nucleic Acids Res*. 2022;50(D1):D988-D995. doi:10.1093/NAR/GKAB1049

71. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. *Nat Biotechnol*. 2011;29(1):24-26. doi:10.1038/NBT.1754

72. Corvol H, Blackman SM, Boëlle PY, Gallins PJ, Pace RG, Stonebraker JR, Accurso FJ, Clement A, Collaco JM, Dang H, Dang AT, Franca A, Gong J, Guillot L, Keenan K, Li W, Lin F, Patrone M V, Raraigh KS, Sun L, Zhou YH, O'Neal WK, Sontag MK, Levy H, Durie PR, Rommens JM, Drumm ML, Wright FA, Strug LJ, Cutting GR, Knowles MR. Genome-wide association meta-analysis identifies five modifier loci of lung disease severity in cystic fibrosis. *Nat Commun*. 2015;6:382. doi:10.1038/ncomms9382

73. Dang H, Polineni D, Pace RG, Stonebraker JR, Corvol H, Cutting GR, Drumm ML, Strug LJ, O'Neal WK, Knowles MR. Mining GWAS and eQTL data for CF lung disease modifiers by gene expression imputation. *PLoS One*. 2020;15(11):e0239189. doi:10.1371/JOURNAL.PONE.0239189

74. Hu X, Wang Q, Tang M, Barthel F, Amin S, Yoshihara K, Lang FM, Martinez-Ledesma E, Lee SH, Zheng S, Verhaak RGW. TumorFusions: an integrative resource for cancer-associated transcript fusions. *Nucleic Acids Res*.

2018;46:D1144-D1149. doi:10.1093/NAR/GKX1018

75. Zaheer U, Faheem M, Qadri I, Begum N, Yassine HM, Al Thani AA, Mathew S. Expression profile of microRNA: an emerging hallmark of cancer. *Curr Pharm Des*. 2019;25(6):642-653. doi:10.2174/1386207322666190325122821

76. Boltsis I, Grosveld F, Giraud G, Kolovos P. Chromatin conformation in development and disease. *Front Cell Dev Biol*. 2021;9:723859. doi:10.3389/FCELL.2021.723859

77. Hnisz D, Shrinivas K, Young RA, Chakraborty AK, Sharp PA. A phase separation model for transcriptional control. *Cell*. 2017;169(1):13-23. doi:10.1016/J.CELL.2017.02.007

78. Si W, Shen J, Zheng H, Fan W. The role and mechanisms of action of microRNAs in cancer drug resistance. *Clin Epigenetics*. 2019;11(1):25. doi:10.1186/S13148-018-0587-8

79. Liang J, Wen J, Huang Z, Chen XP, Zhang BX, Chu L. Small nucleolar RNAs: insight into their function in cancer. *Front Oncol*. 2019;9:587. doi:10.3389/FONC.2019.00587

80. Calin GA, Dumitru CD, Shimizu M, Bichi R, Zupo S, Noch E, Aldler H, Rattan S, Keating M, Rai K, Rassenti L, Kipps T, Negrini M, Bullrich F, Croce CM. Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proc Natl Acad Sci*. 2002;99(24):15524-15529. doi:10.1073/PNAS.242606799

81. Acunzo M, Romano G, Wernicke D, Croce CM. MicroRNA and cancer – a brief overview. *Adv Biol Regul*. 2015;57:1-9. doi:10.1016/J.JBIOR.2014.09.013

82. Martens-Uzunova ES, Hoogstrate Y, Kalsbeek A, Pigmans B, Vredenbregt-van den Berg M, Dits N, Nielsen SJ, Baker A, Visakorpi T, Bangma C, Jenster G. C/D-box snoRNA-derived RNA production is associated with malignant transformation and metastatic progression in prostate cancer. *Oncotarget*. 2015;6(19):17430-17444. doi:10.18632/ONCOTARGET.4172

83. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer Statistics, 2021. *CA Cancer J Clin*. 2021;71(1):7-33. doi:10.3322/CAAC.21654

84. Iczkowski KA. Cell adhesion molecule CD44: its functional roles in prostate cancer. *Am J Transl Res*. 2010;3(1):1-7. Accessed December 12, 2021. /pmc/articles/PMC2981422/

85. Wu YM, Cieślik M, Lonigro RJ, Vats P, Reimers MA, Cao X, Ning Y, Wang L, Kunju LP, de Sarkar N, Heath EI, Chou J, Feng FY, Nelson PS, de Bono JS, Zou W, Montgomery B, Alva A, PCF/SU2C International Prostate Cancer Dream Team, Robinson DR, Chinnaiyan AM. Inactivation of CDK12 delineates a distinct immunogenic class of advanced prostate cancer. *Cell*. 2018;173(7):1770-1782.e14. doi:10.1016/J.CELL.2018.04.034

86. Mei J, Liu Y, Yu X, Hao L, Ma T, Zhan Q, Zhang Y, Zhu Y. YWHAZ interacts with DAAM1 to promote cell migration in breast cancer. *Cell Death Discov*. 2021;7(1):221. doi:10.1038/S41420-021-00609-7

87. Rupaimoole R, Slack FJ. MicroRNA therapeutics: towards a new era for the management of cancer and other diseases. *Nat Rev Drug Discov*. 2017;16(3):203-222. doi:10.1038/NRD.2016.246

88. Chakraborty C, Sharma AR, Sharma G, Lee SS. Therapeutic advances of

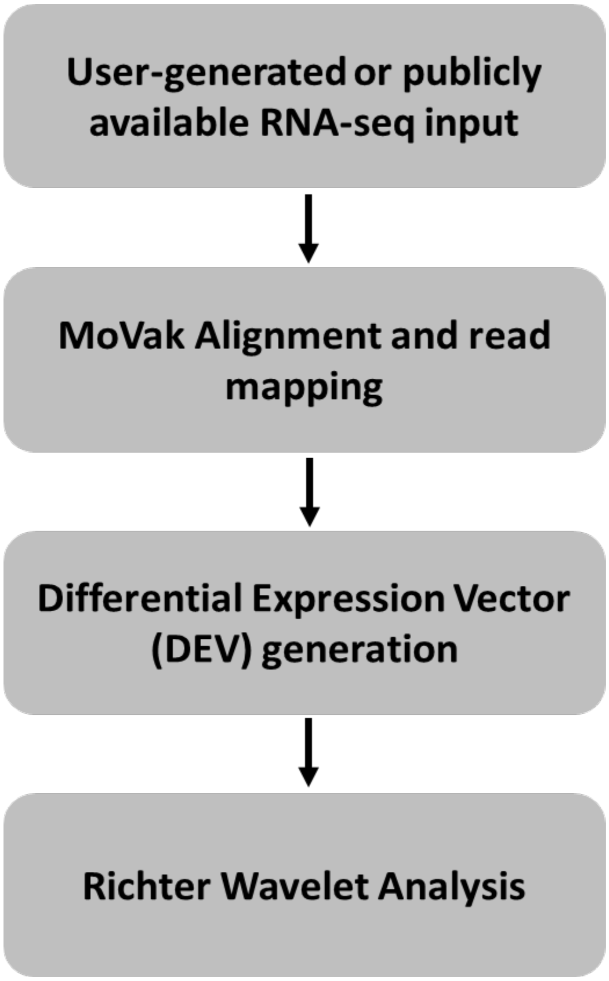
miRNAs: a preclinical and clinical update. *J Adv Res*. 2021;28:127-138. doi:10.1016/J.JARE.2020.08.012

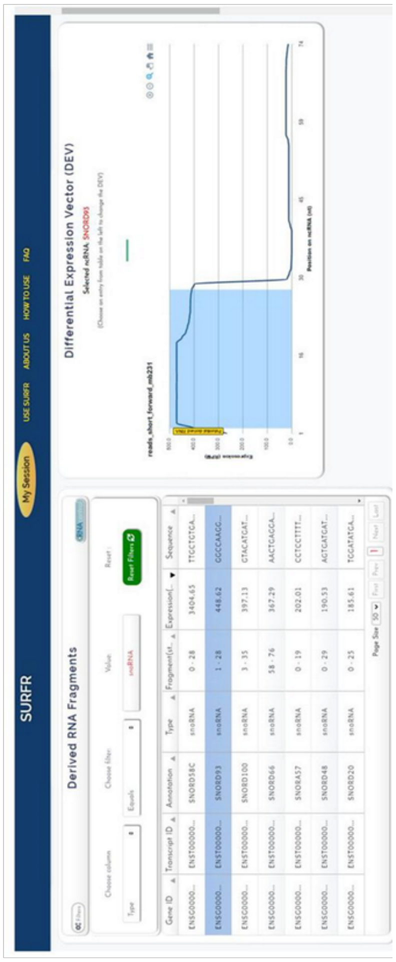
89. Weber JA, Baxter DH, Zhang S, Huang DY, Huang KH, Lee MJ, Galas DJ, Wang K. The microRNA spectrum in 12 body fluids. *Clin Chem*. 2010;56(11):1733-1741. doi:10.1373/CLINCHEM.2010.147405

90. Dhahbi JM, Spindler SR, Atamna H, Boffelli D, Mote P, Martin DIK. 5-YRNA fragments derived by processing of transcripts from specific YRNA genes and pseudogenes are abundant in human serum and plasma. *Physiol Genomics*. 2013;45(21):990-998. doi:10.1152/PHYSIOLGENOMICS.00129.2013

APPENDICES

Appendix A: Short Uncharacterized RNA Fragment Recognition (SURFR) Workflow (left) and Output (right)


User-generated or publicly available RNA-seq input
MoVak Alignment and read mapping
Differential Expression Vector (DEV) generation
Richter Wavelet Analysis



Derived RNA Fragments

Gene ID	Transcript ID	Annotation	Type	Fragment(st...	Expression(...	Sequence
ENSG0000...	ENST00000...	SNORD58C	snoRNA	0 - 28	3404.65	TTGCTGTGCA...
ENSG0000...	ENST00000...	SNORD93	snoRNA	1 - 28	448.62	GCCCAAGG...
ENSG0000...	ENST00000...	SNORD100	snoRNA	3 - 35	397.13	GTACATGAT...
ENSG0000...	ENST00000...	SNORD66	snoRNA	58 - 76	367.29	AACTGACGA...
ENSG0000...	ENST00000...	SNORA57	snoRNA	0 - 19	202.01	CCTCCTTTT...
ENSG0000...	ENST00000...	SNORD48	snoRNA	0 - 29	190.53	AGTGATGAT...
ENSG0000...	ENST00000...	SNORD20	snoRNA	0 - 25	185.61	TCGATATGA...

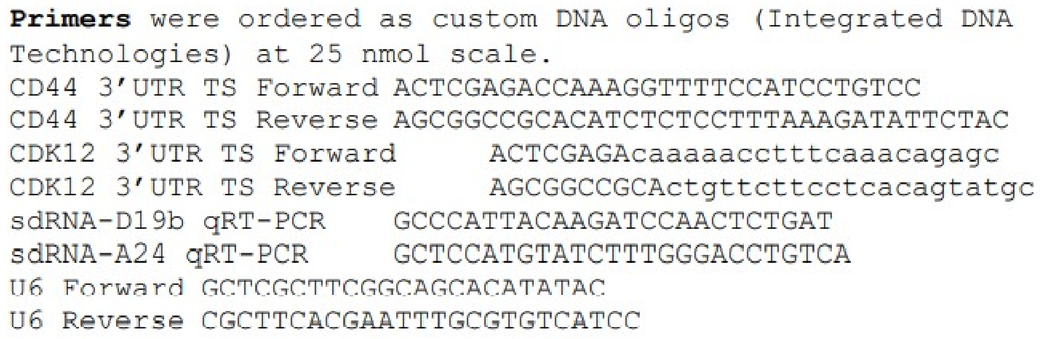
Differential Expression Vector (DEV)

Selected ncRNA: SNORD93

(Choose an entry from table on the left to change the DEV)

Figure A1: Short Uncharacterized RNA Fragment Recognition (SURFR) Workflow (left) and Output (right). A flowchart depicting the major steps of SURFR is displayed (left). An example of a typical SURFR output is displayed (right).

Appendix B: qRT-PCR primers for sdRNAs

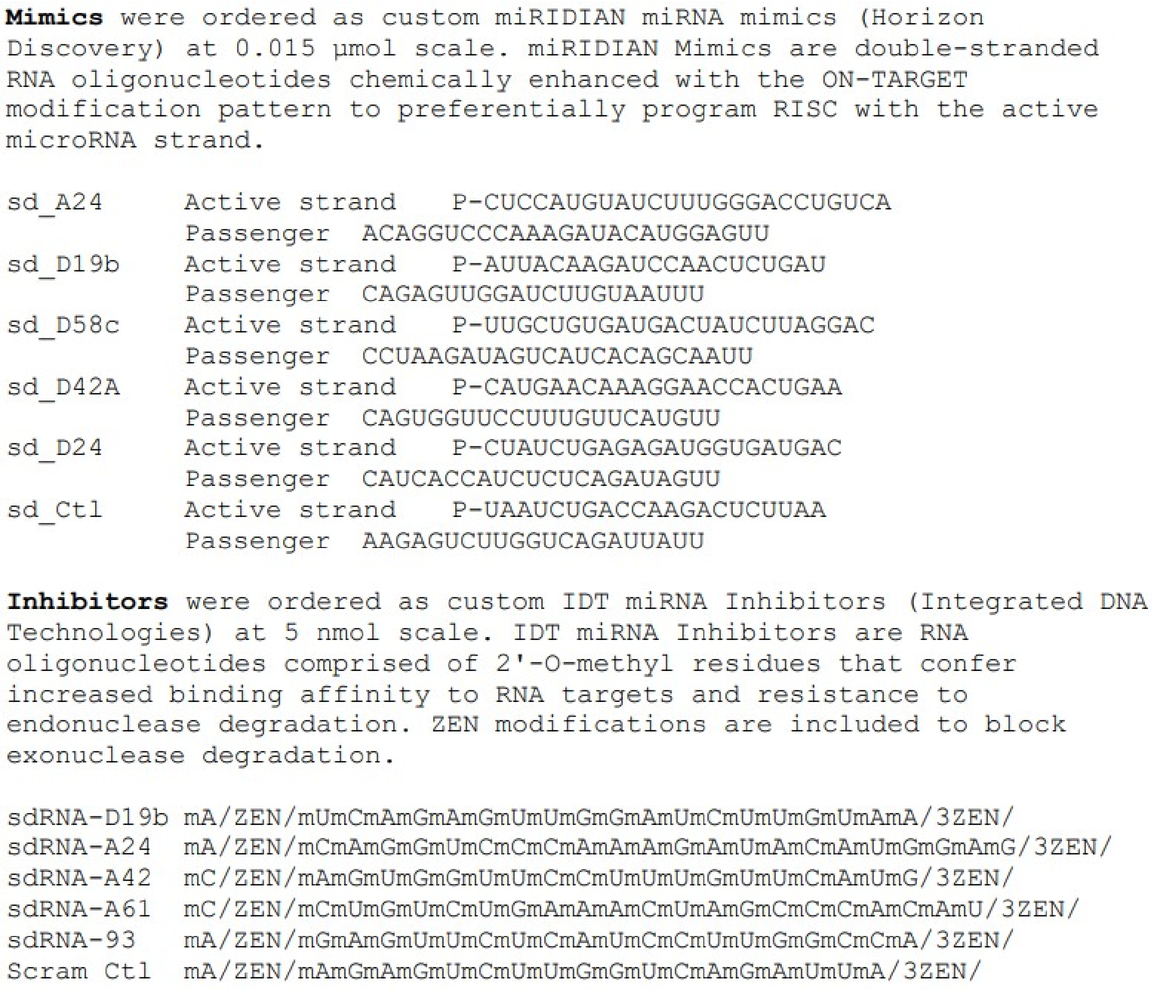


Primers were ordered as custom DNA oligos (Integrated DNA Technologies) at 25 nmol scale.

CD44 3’UTR TS Forward	ACTCGAGACCAAAGGTTTTCCATCCTGTCC
CD44 3’UTR TS Reverse	AGCGGCCGCACATCTCTCCTTTAAAGATATTCTAC
CDK12 3’UTR TS Forward	ACTCGAGAcaaaaacctttcaaacagagc
CDK12 3’UTR TS Reverse	AGCGGCCGCActgttcttcctcacagtatgc
sdRNA-D19b qRT-PCR	GCCCATTACAAGATCCAACTCTGAT
sdRNA-A24 qRT-PCR	GCTCCATGTATCTTTGGGACCTGTCA
U6 Forward	GCTCGCTTCGGCAGCACATATAC
U6 Reverse	CGCTTCACGAATTTGCGTGTCATCC

Figure A2: qRT-PCR primers for sdRNAs. Sequences corresponding to each primer are provided on the right-hand side corresponding to the primer ID.

Appendix C: Mimic and Inhibitor Sequences for sdRNAs



Mimics were ordered as custom miRIDIAN miRNA mimics (Horizon Discovery) at 0.015 µmol scale. miRIDIAN Mimics are double-stranded RNA oligonucleotides chemically enhanced with the ON-TARGET modification pattern to preferentially program RISC with the active microRNA strand.

sd_A24	Active strand	P-CUCCAUGUAUCUUUGGGACCUGUCA
	Passenger	ACAGGUCCCAAAGAUACAUGGAGUU
sd_D19b	Active strand	P-AUUACAAGAUCCAACUCUGAU
	Passenger	CAGAGUUGGAUCUUGUAAUUU
sd_D58c	Active strand	P-UUGCUGUGAUGACUAUCUUAGGAC
	Passenger	CCUAAGAUAGUCAUCACAGCAAUU
sd_D42A	Active strand	P-CAUGAACAAAGGAACCACUGAA
	Passenger	CAGUGGUUCCUUUGUUCAUGUU
sd_D24	Active strand	P-CUAUCUGAGAGAUGGUGAUGAC
	Passenger	CAUCACCAUCUCUCAGAUAGUU
sd_Ctl	Active strand	P-UAAUCUGACCAAGACUCUUAA
	Passenger	AAGAGUCUUGGUCAGAUUAUU

Inhibitors were ordered as custom IDT miRNA Inhibitors (Integrated DNA Technologies) at 5 nmol scale. IDT miRNA Inhibitors are RNA oligonucleotides comprised of 2'-O-methyl residues that confer increased binding affinity to RNA targets and resistance to endonuclease degradation. ZEN modifications are included to block exonuclease degradation.

sdRNA-D19b	mA/ZEN/mUmCmAmGmAmGmUmUmGmGmAmUmCmUmUmGmUmAmA/3ZEN/
sdRNA-A24	mA/ZEN/mCmAmGmGmUmCmCmCmAmAmAmGmAmUmAmCmAmUmGmGmAmG/3ZEN/
sdRNA-A42	mC/ZEN/mAmGmUmGmGmUmUmCmCmUmUmUmGmUmUmCmAmUmG/3ZEN/
sdRNA-A61	mC/ZEN/mCmUmGmUmCmUmGmAmAmAmCmUmAmGmCmCmCmAmCmAmU/3ZEN/
sdRNA-93	mA/ZEN/mGmAmGmUmUmCmUmCmAmUmCmCmUmUmGmGmCmCmA/3ZEN/
Scram Ctl	mA/ZEN/mAmGmAmGmUmCmUmUmGmGmUmCmAmGmAmUmUmA/3ZEN/

Figure A3: Mimic and Inhibitor Sequences for sdRNAs. Sequences are provided for each sdRNA mimic (top) and sdRNA inhibitor (bottom).

Appendix D: TCGA PRAD tumor sample SURFR snoRNA analysis.

Table A1: TCGA PRAD tumor sample SURFR snoRNA analysis.

Ensembl Gene ID	SnoRNA	Start Position	Stop Position	Length	Cancer Average Reads per Million (RPM)	Cancer File Count (/489)	% Cancer File Expression	Control Average Reads per Million (RPM)	Control File Count (/52)	% Control File Expression	Cancer/Control Fold Change	% Cancer File - % Control File Expressions	Sequence (5' to 3')
ENSG00000206754	SNORD101	3	26	24	185	444	90.8	< 30*	0	0.0	6.17	90.8	UUGAAUGAUGACUUUAAUUGUCGG
ENSG0000275043	SNORD25	43	65	23	1312	470	96.1	235	42	80.8	5.59	15.3	CGUGAGGAUAAAUAACUCUGAGG
ENSG0000199753	SNORD104	45	70	26	10102	489	100.0	1896	52	100.0	5.33	0.0	CGGGUGAUGCGAACUGGAGUCUGAGC
ENSG00000201823	SNORD48	38	61	24	774	282	57.7	150	14	26.9	5.18	30.7	UGAUGCCAUCACCGCAGCGCUCUG
ENSG00000264549	SNORD95	39	62	24	1623	448	91.6	333	52	100.0	4.88	-8.4	UGCUGAAAUCCAGAGGCUGUUUCU
ENSG00000275994	SNORA24	1	25	25	711	477	97.5	150	16	30.8	4.75	66.8	CUCCAUGUAUCUUUGGGACCUGUCA
ENSG00000207280	SNORD20	2	24	23	628	486	99.4	136	26	50.0	4.61	49.4	GGAUAUGAUGACUGAUUACCUGA
ENSG00000283551	SNORD98	40	64	25	1874	488	99.8	428	52	100.0	4.37	-0.2	GCAGUGUGGAACACAAUGAACUGAA
ENSG00000275996	SNORD27	2	22	21	598	267	54.6	144	23	44.2	4.16	10.4	CUCCAUGAUGAACACAAAAUG
ENSG00000275996	SNORD27	47	68	22	1233	483	98.8	305	52	100.0	4.05	-1.2	GUGAUGUCAUCUUACUACUGAG
ENSG00000276314	SNORD107	2	29	28	129	255	52.1	34	3	5.8	3.79	46.4	GUUCAUGAUGACACAGGACCUUGUCUGA
ENSG00000199477	SNORA31	2	24	23	1470	393	80.4	395	16	30.8	3.72	49.6	UGCAUCCACUGAUAGACCUUGAA
ENSG00000212452	SNORD69	26	48	23	479	362	74.0	129	10	19.2	3.72	54.8	GGAUCUGACUGACUGUGCUGAGU
ENSG00000226572	SNORD57	2	25	24	420	469	95.9	113	23	44.2	3.71	51.7	GGAGGUGAUGAACUGUCUGAGCCU
ENSG00000206630	SNORD60	2	26	25	1350	489	100.0	376	49	94.2	3.59	5.8	GUCUGUGAUGAAUUGCUUUGACUUC
ENSG00000206622	SNORA69	112	131	20	336	346	70.8	98	5	9.6	3.42	61.1	GACAGAUUGACAUGGACAAU
ENSG00000281859	SNORD38B	46	65	20	102	414	84.7	< 30*	0	0.0	3.40	84.7	GAAGAUAAAGUGUGUCUGAG
ENSG00000221539	SNORD99	49	68	20	415	371	75.9	131	3	5.8	3.16	70.1	AUGCGGAUAUGGGACUGAGA
ENSG00000238917	SNORD10	121	140	20	364	416	85.1	119	13	25.0	3.06	60.1	UCAGUCUUUGUACUCUGAGA
ENSG00000221116	SNORD110	53	73	21	299	449	91.8	100	15	28.8	3.00	63.0	GAUGUCUCCAUGUCUCUGAGC
ENSG00000235408	SNORA71B	141	160	20	277	280	57.3	93	5	9.6	2.99	47.6	UCUGGAGCUUUCGUACAUGC
ENSG00000263934	SNORD3A	196	215	20	3494	484	99.0	1205	43	82.7	2.90	16.3	GAGAGAACGCGGUCUGAGUG
ENSG00000206979	SNORD61	52	71	20	215	260	53.2	77	5	9.6	2.80	43.6	UCCUCUAAGAAGUUCUGAGC
ENSG00000202031	SNORD38A	1	19	19	499	489	100.0	185	46	88.5	2.70	11.5	UCUCGUGAUGAAAACUCUG
ENSG00000239039	SNORD13	83	104	22	252	325	66.5	94	7	13.5	2.67	53.0	UGGGCACAUUACCCGUCUGACC
ENSG00000206989	SNORD63	40	61	22	293	480	98.2	110	45	86.5	2.67	11.6	AACGUGUGGAAAACUAAUGACU
ENSG00000239043	SNORD127	5	27	23	217	264	54.0	84	2	3.8	2.58	50.1	AACUGUGAUGAAAGAUUUGGUCU
ENSG00000266300	SNORD52	38	64	27	276	362	74.0	109	4	7.7	2.53	66.3	GGUCAUGAUGUCAAAACUAAGUUCUGA
ENSG00000263764	SNORD43	2	22	21	395	484	99.0	157	40	76.9	2.52	22.1	ACAGAUGAUGAACUUAUUGAC
ENSG00000238942	SNORD2	31	52	22	285	374	76.5	115	48	92.3	2.48	-15.8	CUGACCUGAAAUGAAGAGAAUA
ENSG00000264591	SNORD84	39	66	28	638	417	85.3	260	35	67.3	2.45	18.0	CGCAGUGAUGACCCUCAUCUAUCACCCU
ENSG00000249020	SNORA58	114	134	21	254	378	77.3	108	11	21.2	2.37	56.1	AUUGCAGGACUCUAAACAUUU
ENSG00000238862	SNORD19B	55	73	19	384	448	91.6	163	22	42.3	2.36	49.3	UACAAGAUCCAACUCUGAU
ENSG00000238649	SNORD42A	41	60	20	281	425	86.9	124	13	25.0	2.27	61.9	UGAACAAAGGAACCACUGAA
ENSG00000207031	SNORD59A	50	71	22	164	414	84.7	73	24	46.2	2.25	38.5	GAAGCCACAUUUAGGUACUGAG
ENSG00000206656	SNORD116-17	72	91	20	639	324	66.3	286	6	11.5	2.24	54.7	AUCCUCGUCGAACUGAGGUC
ENSG00000206680	SNORD21	71	90	20	374	466	95.3	175	44	84.6	2.14	10.7	GUUUCAAGACGGGACUGAUG
ENSG00000212447	SNORD90	79	97	19	872	390	79.8	416	25	48.1	2.10	31.7	CCUACUGUGGAAUCUGAAG

Appendix E: mRNA Target Prediction Results for sdRNA-D19B

Table A2: mRNA Target Prediction Results for sdRNA-D19B.

sdRNA-D19b								
Gene stable ID	HGNC sym	Chr	Gene start	Gene end	Gene description	Best Alignment length	Best Alignment %ID	Multiple Target Sites (Y/N)
ENSG00000163435	ELF3	1	2.02E+08	2.02E+08	E74 like ETS transcription factor 3 [Source:HGNC Symbol;Acc:HGNC:3318]	21	95.238	Y
ENSG00000164588	HCN1	5	45254948	45696498	hyperpolarization activated cyclic nucleotide gated potassium channel 1 [Source:HGNC Symbol;Acc:HGNC:4845]	21	85.714	Y
ENSG00000005700	IBTK	6	82169986	82247754	inhibitor of Bruton tyrosine kinase [Source:HGNC Symbol;Acc:HGNC:17853]	20	90	Y
ENSG00000168961	LGALS9	17	27629798	27649560	galectin 9 [Source:HGNC Symbol;Acc:HGNC:6570]	19	89.474	Y
ENSG00000026508	**CD44**	**11**	**35138882**	**35232402**	**CD44 molecule (Indian blood group) [Source:HGNC Symbol;Acc:HGNC:1681]**	17	88.235	Y
ENSG00000108671	PSMD11	17	32444379	32483319	proteasome 26S subunit, non-ATPase 11 [Source:HGNC Symbol;Acc:HGNC:9556]	17	94.118	Y
ENSG00000205189	ZBTB10	8	80485619	80526265	zinc finger and BTB domain containing 10 [Source:HGNC Symbol;Acc:HGNC:30953]	17	94.118	Y
ENSG00000172869	DMXL1	5	1.19E+08	1.19E+08	Dmx like 1 [Source:HGNC Symbol;Acc:HGNC:2937]	16	93.75	Y
ENSG00000099954	CECR2	22	17359949	17558151	CECR2 histone acetyl-lysine reader [Source:HGNC Symbol;Acc:HGNC:1840]	15	93.333	Y
ENSG00000112297	CRYBG1	6	1.06E+08	1.07E+08	crystallin beta-gamma domain containing 1 [Source:HGNC Symbol;Acc:HGNC:356]	15	93.333	Y
ENSG00000112159	MDN1	6	89642498	89819794	midasin AAA ATPase 1 [Source:HGNC Symbol;Acc:HGNC:18302]	15	93.333	Y
ENSG00000108510	MED13	17	61942605	62065278	mediator complex subunit 13 [Source:HGNC Symbol;Acc:HGNC:22474]	15	93.333	Y
ENSG00000118217	ATF6	1	1.62E+08	1.62E+08	activating transcription factor 6 [Source:HGNC Symbol;Acc:HGNC:791]	14	92.857	Y
ENSG00000088682	COQ9	16	57447425	57461270	coenzyme Q9 [Source:HGNC Symbol;Acc:HGNC:25302]	14	92.857	Y
ENSG00000083097	DOP1A	6	83067666	83171350	DOP1 leucine zipper like protein A [Source:HGNC Symbol;Acc:HGNC:21194]	14	100	Y
ENSG00000197563	PIGN	18	61905255	62187118	phosphatidylinositol glycan anchor biosynthesis class N [Source:HGNC Symbol;Acc:HGNC:8967]	14	92.857	Y
ENSG00000145734	BDP1	5	71455651	71567820	B double prime 1, subunit of RNA polymerase III transcription initiation factor IIIB [HGNC Symbol;Acc:HGNC:13652]	13	92.308	Y
ENSG00000122591	FAM126A	7	22889371	23014130	family with sequence similarity 126 member A [Source:HGNC Symbol;Acc:HGNC:24587]	13	100	Y
ENSG00000113163	CERT1	5	75356345	75512138	ceramide transporter 1 [Source:HGNC Symbol;Acc:HGNC:2205]	12	91.667	Y
ENSG00000086065	CHMP5	9	33264879	33282070	charged multivesicular body protein 5 [Source:HGNC Symbol;Acc:HGNC:26942]	12	100	Y
ENSG00000204209	**DAXX**	**6**	**33318558**	**33323016**	**death domain associated protein [Source:HGNC Symbol;Acc:HGNC:2681]**	12	91.667	Y
ENSG00000137821	LRRC49	15	70853239	71053658	leucine rich repeat containing 49 [Source:HGNC Symbol;Acc:HGNC:25965]	12	100	Y
ENSG00000186868	**MAPT**	**17**	**45894527**	**46028334**	**microtubule associated protein tau [Source:HGNC Symbol;Acc:HGNC:6893]**	12	91.667	Y
ENSG00000113569	NUP155	5	37288137	37371106	nucleoporin 155 [Source:HGNC Symbol;Acc:HGNC:8063]	12	100	Y
ENSG00000225190	PLEKHM1	17	45435900	45490749	pleckstrin homology and RUN domain containing M1 [Source:HGNC Symbol;Acc:HGNC:29017]	12	91.667	Y
ENSG00000159788	**RGS12**	**4**	**3293021**	**3439913**	**regulator of G protein signaling 12 [Source:HGNC Symbol;Acc:HGNC:9994]**	12	100	Y
ENSG00000198315	ZKSCAN8	6	28141883	28159460	zinc finger with KRAB and SCAN domains 8 [Source:HGNC Symbol;Acc:HGNC:12983]	12	100	Y
ENSG00000164323	CFAP97	4	1.85E+08	1.85E+08	cilia and flagella associated protein 97 [Source:HGNC Symbol;Acc:HGNC:29276]	11	100	Y
ENSG00000150760	DOCK1	10	1.27E+08	1.27E+08	dedicator of cytokinesis 1 [Source:HGNC Symbol;Acc:HGNC:2987]	11	100	Y
ENSG00000151491	EPS8	12	15620134	15882329	epidermal growth factor receptor pathway substrate 8 [Source:HGNC Symbol;Acc:HGNC:3420]	11	100	Y
ENSG00000145907	G3BP1	5	1.52E+08	1.52E+08	G3BP stress granule assembly factor 1 [Source:HGNC Symbol;Acc:HGNC:30292]	11	100	Y
ENSG00000261609	GAN	16	81314944	81390809	gigaxonin [Source:HGNC Symbol;Acc:HGNC:4137]	11	100	Y
ENSG00000160410	SHKBP1	19	40576853	40591399	SH3KBP1 binding protein 1 [Source:HGNC Symbol;Acc:HGNC:19214]	11	100	Y
ENSG00000065923	SLC9A7	X	46599251	46759118	solute carrier family 9 member A7 [Source:HGNC Symbol;Acc:HGNC:17123]	11	100	Y
ENSG00000073584	SMARCE1	17	40624962	40648654	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily e, member 1 [HGNC:11109]	11	100	Y
ENSG00000176438	SYNE3	14	95407266	95516650	spectrin repeat containing nuclear envelope family member 3 [Source:HGNC Symbol;Acc:HGNC:19861]	11	100	Y
ENSG00000137501	SYTL2	11	85694224	85811159	synaptotagmin like 2 [Source:HGNC Symbol;Acc:HGNC:15585]	11	100	Y
ENSG00000111602	TIMELESS	12	56416363	56449426	timeless circadian regulator [Source:HGNC Symbol;Acc:HGNC:11813]	11	100	Y
ENSG00000198551	ZNF627	19	11559374	11619161	zinc finger protein 627 [Source:HGNC Symbol;Acc:HGNC:30570]	11	100	Y
ENSG00000109323	MANBA	4	1.03E+08	1.03E+08	mannosidase beta [Source:HGNC Symbol;Acc:HGNC:6831]	23	86.957	N
ENSG00000104967	NOVA2	19	45933734	45974044	NOVA alternative splicing regulator 2 [Source:HGNC Symbol;Acc:HGNC:7887]	21	85.714	N
ENSG00000230124	ACBD6	1	1.8E+08	1.81E+08	acyl-CoA binding domain containing 6 [Source:HGNC Symbol;Acc:HGNC:23339]	20	85	N
ENSG00000178163	ZNF518B	4	10439880	10457426	zinc finger protein 518B [Source:HGNC Symbol;Acc:HGNC:29365]	20	90	N
ENSG00000147180	ZNF711	X	85243991	85273362	zinc finger protein 711 [Source:HGNC Symbol;Acc:HGNC:13128]	20	85	N
ENSG00000197959	DNM3	1	1.72E+08	1.72E+08	dynamin 3 [Source:HGNC Symbol;Acc:HGNC:29125]	19	89.474	N
ENSG00000187715	KBTBD12	3	1.28E+08	1.28E+08	kelch repeat and BTB domain containing 12 [Source:HGNC Symbol;Acc:HGNC:25731]	19	89.474	N
ENSG00000204852	TCTN1	12	1.11E+08	1.11E+08	tectonic family member 1 [Source:HGNC Symbol;Acc:HGNC:26113]	19	89.474	N
ENSG00000185100	ADSS1	14	1.05E+08	1.05E+08	adenylosuccinate synthase 1 [Source:HGNC Symbol;Acc:HGNC:20093]	18	88.889	N

Table A2, cont.

ENSG00000153774	CFDP1	16	75293698	75433503	craniofacial development protein 1 [Source:HGNC Symbol;Acc:HGNC:1873]	18	88.889	N
ENSG00000162601	**MYSM1**	**1**	**58643440**	**58700077**	**Myb like, SWIRM and MPN domains 1 [Source:HGNC Symbol;Acc:HGNC:29401]**	18	88.889	N
ENSG00000091844	**RGS17**	**6**	**1.53E+08**	**1.53E+08**	**regulator of G protein signaling 17 [Source:HGNC Symbol;Acc:HGNC:14088]**	18	88.889	N
ENSG00000154240	CEP112	17	65635537	66192133	centrosomal protein 112 [Source:HGNC Symbol;Acc:HGNC:28514]	17	88.235	N
ENSG00000170296	GABARAP	17	7240008	7242449	GABA type A receptor-associated protein [Source:HGNC Symbol;Acc:HGNC:4067]	17	94.118	N
ENSG00000100784	RPS6KA5	14	90847861	91060641	ribosomal protein S6 kinase A5 [Source:HGNC Symbol;Acc:HGNC:10434]	17	88.235	N
ENSG00000130234	ACE2	X	15494566	15607236	angiotensin converting enzyme 2 [Source:HGNC Symbol;Acc:HGNC:13557]	16	93.75	N
ENSG00000112685	EXOC2	6	485154	693139	exocyst complex component 2 [Source:HGNC Symbol;Acc:HGNC:24968]	16	93.75	N
ENSG00000248383	PCDHAC1	5	1.41E+08	1.41E+08	protocadherin alpha subfamily C, 1 [Source:HGNC Symbol;Acc:HGNC:8676]	16	93.75	N
ENSG00000171865	RNASEH1	2	3541430	3558333	ribonuclease H1 [Source:HGNC Symbol;Acc:HGNC:18466]	16	93.75	N
ENSG00000066923	STAG3	7	1E+08	1E+08	stromal antigen 3 [Source:HGNC Symbol;Acc:HGNC:11356]	16	93.75	N
ENSG00000156709	AIFM1	X	1.3E+08	1.3E+08	apoptosis inducing factor mitochondria associated 1 [Source:HGNC Symbol;Acc:HGNC:8768]	15	93.333	N
ENSG00000172331	BPGM	7	1.35E+08	1.35E+08	bisphosphoglycerate mutase [Source:HGNC Symbol;Acc:HGNC:1093]	15	93.333	N
ENSG00000088881	EBF4	20	2692874	2760108	EBF family member 4 [Source:HGNC Symbol;Acc:HGNC:29278]	15	93.333	N
ENSG00000090863	GLG1	16	74447427	74607144	golgi glycoprotein 1 [Source:HGNC Symbol;Acc:HGNC:4316]	15	93.333	N
ENSG00000213625	LEPROT	1	65420587	65436007	leptin receptor overlapping transcript [Source:HGNC Symbol;Acc:HGNC:29477]	15	93.333	N
ENSG00000155530	LRGUK	7	1.34E+08	1.34E+08	leucine rich repeats and guanylate kinase domain containing [Source:HGNC Symbol;Acc:HGNC:21964]	15	93.333	N
ENSG00000123213	NLN	5	65722205	65871725	neurolysin [Source:HGNC Symbol;Acc:HGNC:16058]	15	93.333	N
ENSG00000134853	PDGFRA	4	54229280	54298245	platelet derived growth factor receptor alpha [Source:HGNC Symbol;Acc:HGNC:8803]	15	93.333	N
ENSG00000103479	RBL2	16	53433977	53491648	RB transcriptional corepressor like 2 [Source:HGNC Symbol;Acc:HGNC:9894]	15	93.333	N
ENSG00000101695	RNF125	18	32018825	32073219	ring finger protein 125 [Source:HGNC Symbol;Acc:HGNC:21150]	15	93.333	N
ENSG00000180739	S1PR5	19	10512742	10517931	sphingosine-1-phosphate receptor 5 [Source:HGNC Symbol;Acc:HGNC:14299]	15	93.333	N
ENSG00000105866	SP4	7	21428043	21514822	Sp4 transcription factor [Source:HGNC Symbol;Acc:HGNC:11209]	15	93.333	N
ENSG00000185518	SV2B	15	91099950	91302565	synaptic vesicle glycoprotein 2B [Source:HGNC Symbol;Acc:HGNC:16874]	15	93.333	N
ENSG00000275895	U2AF1L5	21	6484623	6499261	U2 small nuclear RNA auxiliary factor 1 like 5 [Source:HGNC Symbol;Acc:HGNC:51830]	15	93.333	N
ENSG00000267041	ZNF850	19	36714383	36772825	zinc finger protein 850 [Source:HGNC Symbol;Acc:HGNC:27994]	15	93.333	N
ENSG00000206560	ANKRD28	3	15667236	15859771	ankyrin repeat domain 28 [Source:HGNC Symbol;Acc:HGNC:29024]	14	92.857	N
ENSG00000115355	CCDC88A	2	55287842	55419895	coiled-coil domain containing 88A [Source:HGNC Symbol;Acc:HGNC:25523]	14	92.857	N
ENSG00000261210	CLEC19A	16	19285731	19322145	C-type lectin domain containing 19A [Source:HGNC Symbol;Acc:HGNC:34522]	14	92.857	N
ENSG00000165732	DDX21	10	68956135	68985068	DExD-box helicase 21 [Source:HGNC Symbol;Acc:HGNC:2744]	14	92.857	N
ENSG00000127955	GNAI1	7	79768028	80226181	G protein subunit alpha i1 [Source:HGNC Symbol;Acc:HGNC:4384]	14	92.857	N
ENSG00000128944	KNSTRN	15	40382721	40394288	kinetochore localized astrin (SPAG5) binding protein [Source:HGNC Symbol;Acc:HGNC:30767]	14	92.857	N
ENSG00000243709	LEFTY1	1	2.26E+08	2.26E+08	left-right determination factor 1 [Source:HGNC Symbol;Acc:HGNC:6552]	14	100	N
ENSG00000196712	**NF1**	**17**	**31094927**	**31382116**	**neurofibromin 1 [Source:HGNC Symbol;Acc:HGNC:7765]**	14	92.857	N
ENSG00000167081	**PBX3**	**9**	**1.26E+08**	**1.26E+08**	**PBX homeobox 3 [Source:HGNC Symbol;Acc:HGNC:8634]**	14	92.857	N
ENSG00000198933	TBKBP1	17	47694161	47712052	TBK1 binding protein 1 [Source:HGNC Symbol;Acc:HGNC:30140]	14	92.857	N
ENSG00000184056	VPS33B	15	90998673	91022603	VPS33B late endosome and lysosome associated [Source:HGNC Symbol;Acc:HGNC:12712]	14	92.857	N
ENSG00000114439	BBX	3	1.08E+08	1.08E+08	BBX high mobility group box domain containing [Source:HGNC Symbol;Acc:HGNC:14422]	13	100	N
ENSG00000164120	**HPGD**	**4**	**1.74E+08**	**1.75E+08**	**15-hydroxyprostaglandin dehydrogenase [Source:HGNC Symbol;Acc:HGNC:5154]**	13	100	N
ENSG00000032742	IFT88	13	20567138	20691444	intraflagellar transport 88 [Source:HGNC Symbol;Acc:HGNC:20606]	13	100	N
ENSG00000162981	LRATD1	2	14632700	14650814	LRAT domain containing 1 [Source:HGNC Symbol;Acc:HGNC:20743]	13	100	N
ENSG00000114125	RNF7	3	1.42E+08	1.42E+08	ring finger protein 7 [Source:HGNC Symbol;Acc:HGNC:10070]	13	100	N
ENSG00000170381	SEMA3E	7	83363238	83649139	semaphorin 3E [Source:HGNC Symbol;Acc:HGNC:10727]	13	100	N
ENSG00000213588	ZBTB9	6	33453970	33457544	zinc finger and BTB domain containing 9 [Source:HGNC Symbol;Acc:HGNC:28323]	13	100	N
ENSG00000120697	ALG5	13	36949738	37000763	ALG5 dolichyl-phosphate beta-glucosyltransferase [Source:HGNC Symbol;Acc:HGNC:20266]	12	100	N
ENSG00000128335	APOL2	22	36226209	36239954	apolipoprotein L2 [Source:HGNC Symbol;Acc:HGNC:619]	12	100	N
ENSG00000126453	BCL2L12	19	49665142	49673916	BCL2 like 12 [Source:HGNC Symbol;Acc:HGNC:13787]	12	100	N
ENSG00000187068	C3orf70	3	1.85E+08	1.85E+08	chromosome 3 open reading frame 70 [Source:HGNC Symbol;Acc:HGNC:33731]	12	100	N
ENSG00000079691	CARMIL1	6	25279078	25620530	capping protein regulator and myosin 1 linker 1 [Source:HGNC Symbol;Acc:HGNC:21581]	12	100	N
ENSG00000145241	CENPC	4	67468762	67545503	centromere protein C [Source:HGNC Symbol;Acc:HGNC:1854]	12	100	N
ENSG00000128512	DOCK4	7	1.12E+08	1.12E+08	dedicator of cytokinesis 4 [Source:HGNC Symbol;Acc:HGNC:19192]	12	100	N

Table A2, cont.

ENSG00000274290	H2BC6	6	26172059	26184655	H2B clustered histone 6 [Source:HGNC Symbol;Acc:HGNC:4753]	12	100	N
ENSG00000100221	JOSD1	22	38685543	38701556	Josephin domain containing 1 [Source:HGNC Symbol;Acc:HGNC:28953]	12	100	N
ENSG00000272333	KMT2B	19	35717973	35738880	lysine methyltransferase 2B [Source:HGNC Symbol;Acc:HGNC:15840]	12	100	N
ENSG00000196549	MME	3	1.55E+08	1.55E+08	membrane metalloendopeptidase [Source:HGNC Symbol;Acc:HGNC:7154]	12	100	N
ENSG00000129422	**MTUS1**	**8**	**17643795**	**17801094**	**microtubule associated scaffold protein 1 [Source:HGNC Symbol;Acc:HGNC:29789]**	12	100	N
ENSG00000213619	NDUFS3	11	47565336	47584562	NADH:ubiquinone oxidoreductase core subunit S3 [Source:HGNC Symbol;Acc:HGNC:7710]	12	100	N
ENSG00000091622	PITPNM3	17	6451263	6556555	PITPNM family member 3 [Source:HGNC Symbol;Acc:HGNC:21043]	12	100	N
ENSG00000138032	PPM1B	2	44167969	44244384	protein phosphatase, Mg2+/Mn2+ dependent 1B [Source:HGNC Symbol;Acc:HGNC:9276]	12	100	N
ENSG00000027075	PRKCH	14	61187559	61550976	protein kinase C eta [Source:HGNC Symbol;Acc:HGNC:9403]	12	100	N
ENSG00000175467	SART1	11	65961728	65980137	spliceosome associated factor 1, recruiter of U4/U6.U5 tri-snRNP [Source:HGNC Symbol;Acc:HGNC:10538]	12	100	N
ENSG00000108061	SHOC2	10	1.11E+08	1.11E+08	SHOC2 leucine rich repeat scaffold protein [Source:HGNC Symbol;Acc:HGNC:15454]	12	100	N
ENSG00000122912	SLC25A16	10	68477998	68527523	solute carrier family 25 member 16 [Source:HGNC Symbol;Acc:HGNC:10986]	12	100	N
ENSG00000138050	THUMPD2	2	39736060	39779267	THUMP domain containing 2 [Source:HGNC Symbol;Acc:HGNC:14890]	12	100	N
ENSG00000114742	WDR48	3	39052013	39096671	WD repeat domain 48 [Source:HGNC Symbol;Acc:HGNC:30914]	12	100	N
ENSG00000177054	ZDHHC13	11	19117099	19176422	zinc finger DHHC-type palmitoyltransferase 13 [Source:HGNC Symbol;Acc:HGNC:18413]	12	100	N
ENSG00000164199	ADGRV1	5	90529344	91164437	adhesion G protein-coupled receptor V1 [Source:HGNC Symbol;Acc:HGNC:17416]	11	100	N
ENSG00000159461	AMFR	16	56361452	56425545	autocrine motility factor receptor [Source:HGNC Symbol;Acc:HGNC:463]	11	100	N
ENSG00000169604	ANTXR1	2	69013176	69249327	ANTXR cell adhesion molecule 1 [Source:HGNC Symbol;Acc:HGNC:21014]	11	100	N
ENSG00000149182	ARFGAP2	11	47164299	47177125	ADP ribosylation factor GTPase activating protein 2 [Source:HGNC Symbol;Acc:HGNC:13504]	11	100	N
ENSG00000183337	**BCOR**	**X**	**40049815**	**40177329**	**BCL6 corepressor [Source:HGNC Symbol;Acc:HGNC:20893]**	11	100	N
ENSG00000055130	CUL1	7	1.49E+08	1.49E+08	cullin 1 [Source:HGNC Symbol;Acc:HGNC:2551]	11	100	N
ENSG00000164934	DCAF13	8	1.03E+08	1.03E+08	DDB1 and CUL4 associated factor 13 [Source:HGNC Symbol;Acc:HGNC:24535]	11	100	N
ENSG00000167986	DDB1	11	61299451	61342596	damage specific DNA binding protein 1 [Source:HGNC Symbol;Acc:HGNC:2717]	11	100	N
ENSG00000086189	DIMT1	5	62347284	62403943	DIM1 rRNA methyltransferase and ribosome maturation factor [Source:HGNC Symbol;Acc:HGNC:30217]	11	100	N
ENSG00000138246	DNAJC13	3	1.32E+08	1.33E+08	DnaJ heat shock protein family (Hsp40) member C13 [Source:HGNC Symbol;Acc:HGNC:30343]	11	100	N
ENSG00000161960	EIF4A1	17	7572824	7579006	eukaryotic translation initiation factor 4A1 [Source:HGNC Symbol;Acc:HGNC:3282]	11	100	N
ENSG00000171723	GPHN	14	66507407	67181803	gephyrin [Source:HGNC Symbol;Acc:HGNC:15465]	11	100	N
ENSG00000166503	HDGFL3	15	83112738	83207823	HDGF like 3 [Source:HGNC Symbol;Acc:HGNC:24937]	11	100	N
ENSG00000086696	**HSD17B2**	**16**	**82035004**	**82098534**	**hydroxysteroid 17-beta dehydrogenase 2 [Source:HGNC Symbol;Acc:HGNC:5211]**	11	100	N
ENSG00000169592	INO80E	16	29995715	30005793	INO80 complex subunit E [Source:HGNC Symbol;Acc:HGNC:26905]	11	100	N
ENSG00000143493	INTS7	1	2.12E+08	2.12E+08	integrator complex subunit 7 [Source:HGNC Symbol;Acc:HGNC:24484]	11	100	N
ENSG00000135709	KIAA0513	16	85027782	85094230	KIAA0513 [Source:HGNC Symbol;Acc:HGNC:29058]	11	100	N
ENSG00000105835	**NAMPT**	**7**	**1.06E+08**	**1.06E+08**	**nicotinamide phosphoribosyltransferase [Source:HGNC Symbol;Acc:HGNC:30092]**	11	100	N
ENSG00000136448	NMT1	17	44957992	45109016	N-myristoyltransferase 1 [Source:HGNC Symbol;Acc:HGNC:7857]	11	100	N
ENSG00000161542	PRPSAP1	17	76309478	76384521	phosphoribosyl pyrophosphate synthetase associated protein 1 [Source:HGNC Symbol;Acc:HGNC:9466]	11	100	N
ENSG00000166224	SGPL1	10	70815948	70881184	sphingosine-1-phosphate lyase 1 [Source:HGNC Symbol;Acc:HGNC:10817]	11	100	N
ENSG00000170921	TANC2	17	62966235	63427703	tetratricopeptide repeat, ankyrin repeat and coiled-coil containing 2 [Source:HGNC Symbol;Acc:HGNC:30212]	11	100	N
ENSG00000185361	TNFAIP8L1	19	4639516	4655568	TNF alpha induced protein 8 like 1 [Source:HGNC Symbol;Acc:HGNC:28279]	11	100	N
ENSG00000198900	TOP1	20	41028822	41124487	DNA topoisomerase I [Source:HGNC Symbol;Acc:HGNC:11986]	11	100	N
ENSG00000143337	TOR1AIP1	1	1.8E+08	1.8E+08	torsin 1A interacting protein 1 [Source:HGNC Symbol;Acc:HGNC:29456]	11	100	N
ENSG00000140416	TPM1	15	63042632	63071915	tropomyosin 1 [Source:HGNC Symbol;Acc:HGNC:12010]	11	100	N
ENSG00000221926	**TRIM16**	**17**	**15627960**	**15684311**	**tripartite motif containing 16 [Source:HGNC Symbol;Acc:HGNC:17241]**	11	100	N
ENSG00000159459	UBR1	15	42942897	43106113	ubiquitin protein ligase E3 component n-recognin 1 [Source:HGNC Symbol;Acc:HGNC:16808]	11	100	N
ENSG00000134987	WDR36	5	1.11E+08	1.11E+08	WD repeat domain 36 [Source:HGNC Symbol;Acc:HGNC:30696]	11	100	N
ENSG00000138658	ZGRF1	4	1.13E+08	1.13E+08	zinc finger GRF-type containing 1 [Source:HGNC Symbol;Acc:HGNC:25654]	11	100	N
ENSG00000159905	ZNF221	19	43951223	43967709	zinc finger protein 221 [Source:HGNC Symbol;Acc:HGNC:13014]	11	100	N
ENSG00000267680	ZNF224	19	44094361	44109886	zinc finger protein 224 [Source:HGNC Symbol;Acc:HGNC:13017]	11	100	N
ENSG00000256294	ZNF225	19	44112181	44134822	zinc finger protein 225 [Source:HGNC Symbol;Acc:HGNC:13018]	11	100	N
ENSG00000197951	ZNF71	19	56595300	56626481	zinc finger protein 71 [Source:HGNC Symbol;Acc:HGNC:13141]	11	100	N

Appendix F: mRNA Target Prediction Results for sdRNA-A24B

Table A3: mRNA Target Prediction Results for sdRNA-A24B.

sdRNA-A24								
Gene stable ID	**HGNC sym**	**Chr**	**Gene start**	**Gene end**	**Gene description**	**Best Alignment length**	**Best Alignment %ID**	**Multiple Target Sites**
ENSG00000124942	**AHNAK**	**11**	**62433542**	**62556235**	**AHNAK nucleoprotein [Source:HGNC Symbol;Acc:HGNC:347]**	25	80	Y
ENSG00000167258	**CDK12**	**17**	**39461486**	**39564907**	**cyclin dependent kinase 12 [Source:HGNC Symbol;Acc:HGNC:24224]**	22	81.818	Y
ENSG00000185513	L3MBTL1	20	43489442	43550954	L3MBTL histone methyl-lysine binding protein 1 [Source:HGNC Symbol;Acc:HGNC:15905]	22	81.818	Y
ENSG00000164631	ZNF12	7	6688433	6706947	zinc finger protein 12 [Source:HGNC Symbol;Acc:HGNC:12902]	22	86.364	Y
ENSG00000143341	HMCN1	1	1.86E+08	1.86E+08	hemicentin 1 [Source:HGNC Symbol;Acc:HGNC:19194]	21	85.714	Y
ENSG00000185658	BRWD1	21	39184176	39321559	bromodomain and WD repeat domain containing 1 [Source:HGNC Symbol;Acc:HGNC:12760]	20	90	Y
ENSG00000140836	**ZFHX3**	**16**	**72782885**	**73891871**	**zinc finger homeobox 3 [Source:HGNC Symbol;Acc:HGNC:777]**	19	89.474	Y
ENSG00000109771	LRP2BP	4	1.85E+08	1.85E+08	LRP2 binding protein [Source:HGNC Symbol;Acc:HGNC:25434]	18	88.889	Y
ENSG00000158411	MITD1	2	99161427	99181058	microtubule interacting and trafficking domain containing 1 [Source:HGNC Symbol;Acc:HGNC:25207]	18	88.889	Y
ENSG00000005810	MYCBP2	13	77042474	77327094	MYC binding protein 2 [Source:HGNC Symbol;Acc:HGNC:23386]	18	88.889	Y
ENSG00000173821	RNF213	17	80260852	80398794	ring finger protein 213 [Source:HGNC Symbol;Acc:HGNC:14539]	18	88.889	Y
ENSG00000056277	ZNF280C	X	1.3E+08	1.3E+08	zinc finger protein 280C [Source:HGNC Symbol;Acc:HGNC:25955]	18	88.889	Y
ENSG00000108651	UTP6	17	31860904	31901708	UTP6 small subunit processome component [Source:HGNC Symbol;Acc:HGNC:18279]	17	94.118	Y
ENSG00000164684	ZNF704	8	80628451	80874781	zinc finger protein 704 [Source:HGNC Symbol;Acc:HGNC:32291]	17	94.118	Y
ENSG00000165119	HNRNPK	9	83968083	83980616	heterogeneous nuclear ribonucleoprotein K [Source:HGNC Symbol;Acc:HGNC:5044]	16	93.75	Y
ENSG00000153885	KCTD15	19	33795933	33815763	potassium channel tetramerization domain containing 15 [Source:HGNC Symbol;Acc:HGNC:23297]	16	93.75	Y
ENSG00000196262	PPIA	7	44796680	44824564	peptidylprolyl isomerase A [Source:HGNC Symbol;Acc:HGNC:9253]	16	93.75	Y
ENSG00000107863	**ARHGAP2:**	**10**	**24583609**	**24723887**	**Rho GTPase activating protein 21 [Source:HGNC Symbol;Acc:HGNC:23725]**	15	86.667	Y
ENSG00000112297	CRYBG1	6	1.06E+08	1.07E+08	crystallin beta-gamma domain containing 1 [Source:HGNC Symbol;Acc:HGNC:356]	15	93.333	Y
ENSG00000158290	**CUL4B**	**X**	**1.21E+08**	**1.21E+08**	**cullin 4B [Source:HGNC Symbol;Acc:HGNC:2555]**	15	93.333	Y
ENSG00000046604	DSG2	18	31498177	31549008	desmoglein 2 [Source:HGNC Symbol;Acc:HGNC:3049]	15	93.333	Y
ENSG00000121957	GPSM2	1	1.09E+08	1.09E+08	G protein signaling modulator 2 [Source:HGNC Symbol;Acc:HGNC:29501]	15	93.333	Y
ENSG00000116678	LEPR	1	65420652	65641559	leptin receptor [Source:HGNC Symbol;Acc:HGNC:6554]	15	93.333	Y
ENSG00000070018	LRP6	12	12116025	12267044	LDL receptor related protein 6 [Source:HGNC Symbol;Acc:HGNC:6698]	15	93.333	Y
ENSG00000140396	NCOA2	8	70109782	70403808	nuclear receptor coactivator 2 [Source:HGNC Symbol;Acc:HGNC:7669]	15	93.333	Y
ENSG00000172320	OR5A1	11	59436469	59451380	olfactory receptor family 5 subfamily A member 1 [Source:HGNC Symbol;Acc:HGNC:8319]	15	93.333	Y
ENSG00000109475	RPL34	4	1.09E+08	1.09E+08	ribosomal protein L34 [Source:HGNC Symbol;Acc:HGNC:10340]	15	93.333	Y
ENSG00000113300	CNOT6	5	1.8E+08	1.81E+08	CCR4-NOT transcription complex subunit 6 [Source:HGNC Symbol;Acc:HGNC:14099]	14	92.857	Y
ENSG00000198408	OGA	10	1.02E+08	1.02E+08	O-GlcNAcase [Source:HGNC Symbol;Acc:HGNC:7056]	14	100	Y
ENSG00000079387	SENP1	12	48042897	48106079	SUMO specific peptidase 1 [Source:HGNC Symbol;Acc:HGNC:17927]	14	100	Y
ENSG00000183049	CAMK1D	10	12349547	12835545	calcium/calmodulin dependent protein kinase ID [Source:HGNC Symbol;Acc:HGNC:19341]	13	100	Y
ENSG00000156345	CDK20	9	87966441	87974753	cyclin dependent kinase 20 [Source:HGNC Symbol;Acc:HGNC:21420]	13	100	Y
ENSG00000139219	COL2A1	12	47972967	48004554	collagen type II alpha 1 chain [Source:HGNC Symbol;Acc:HGNC:2200]	13	92.308	Y
ENSG00000188153	COL4A5	X	1.08E+08	1.09E+08	collagen type IV alpha 5 chain [Source:HGNC Symbol;Acc:HGNC:2207]	13	92.308	Y
ENSG00000137770	CTDSPL2	15	44427622	44529038	CTD small phosphatase like 2 [Source:HGNC Symbol;Acc:HGNC:26936]	13	100	Y
ENSG00000129422	MTUS1	8	17643795	17801094	microtubule associated scaffold protein 1 [Source:HGNC Symbol;Acc:HGNC:29789]	13	100	Y
ENSG00000116833	NR5A2	1	2E+08	2E+08	nuclear receptor subfamily 5 group A member 2 [Source:HGNC Symbol;Acc:HGNC:7984]	13	100	Y
ENSG00000107771	CCSER2	10	84328586	84518521	coiled-coil serine rich protein 2 [Source:HGNC Symbol;Acc:HGNC:29197]	12	100	Y
ENSG00000114270	COL7A1	3	48564073	48595329	collagen type VII alpha 1 chain [Source:HGNC Symbol;Acc:HGNC:2214]	12	91.667	Y
ENSG00000109861	CTSC	11	88265069	88359684	cathepsin C [Source:HGNC Symbol;Acc:HGNC:2528]	12	100	Y
ENSG00000198947	DMD	X	31097677	33339609	dystrophin [Source:HGNC Symbol;Acc:HGNC:2928]	12	91.667	Y
ENSG00000134109	EDEM1	3	5187646	5219958	ER degradation enhancing alpha-mannosidase like protein 1 [Source:HGNC Symbol;Acc:HGNC:18967]	12	100	Y
ENSG00000132205	EMILIN2	18	2847006	2916003	elastin microfibril interfacer 2 [Source:HGNC Symbol;Acc:HGNC:19881]	12	100	Y
ENSG00000130244	FAM98C	19	38403093	38409088	family with sequence similarity 98 member C [Source:HGNC Symbol;Acc:HGNC:27119]	12	100	Y
ENSG00000115159	GPD2	2	1.56E+08	1.57E+08	glycerol-3-phosphate dehydrogenase 2 [Source:HGNC Symbol;Acc:HGNC:4456]	12	100	Y
ENSG00000166503	HDGFL3	15	83112738	83207823	HDGF like 3 [Source:HGNC Symbol;Acc:HGNC:24937]	12	100	Y
ENSG00000091136	LAMB1	7	1.08E+08	1.08E+08	laminin subunit beta 1 [Source:HGNC Symbol;Acc:HGNC:6486]	12	100	Y

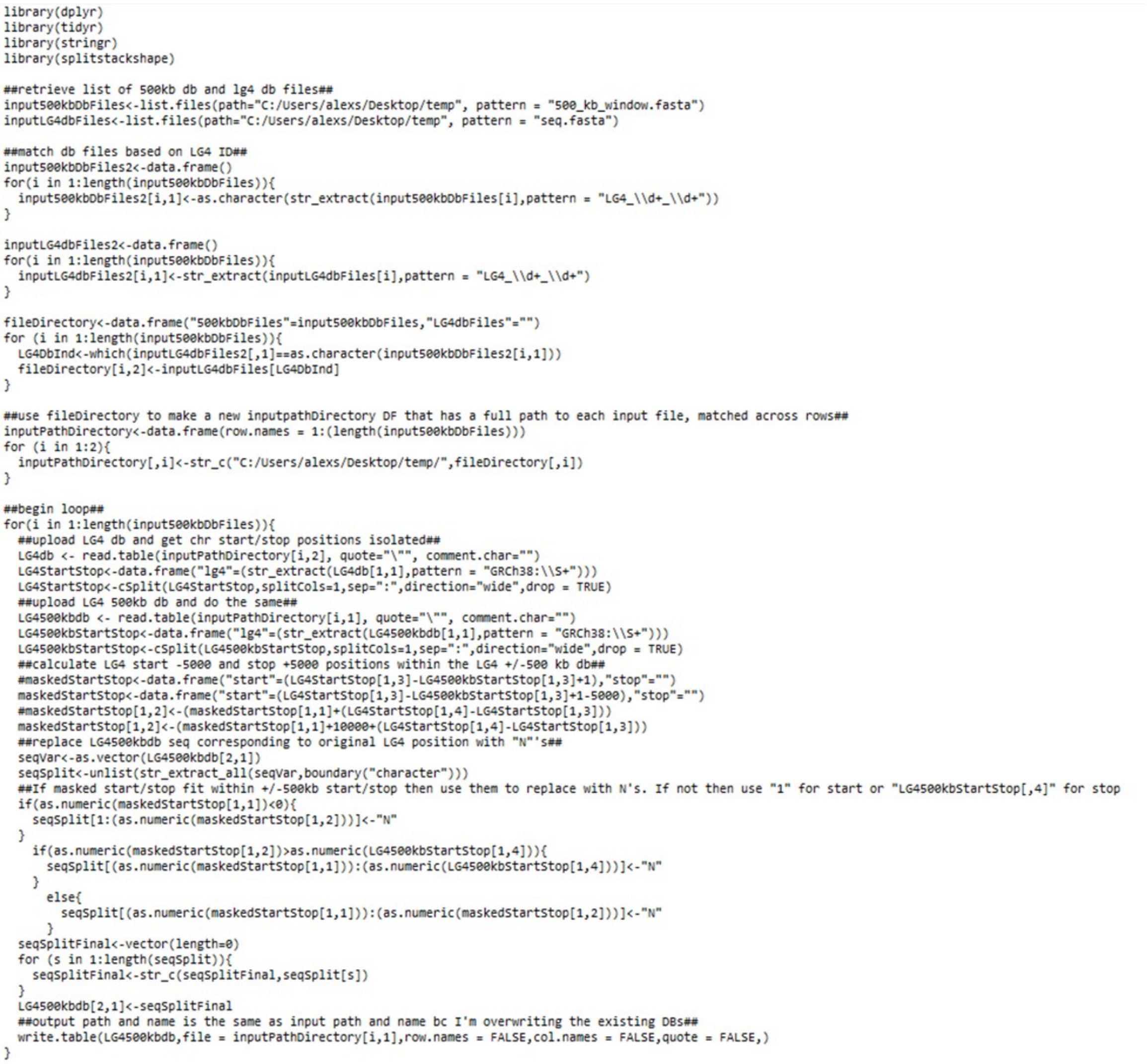
Table A3, cont.

ENSG00000154237	LRRK1	15	1.01E+08	1.01E+08	leucine rich repeat kinase 1 [Source:HGNC Symbol;Acc:HGNC:18608]	12	100	Y
ENSG00000105926	PALS2	7	24573268	24694193	protein associated with LIN7 2, MAGUK family member [Source:HGNC Symbol;Acc:HGNC:18167]	12	100	Y
ENSG00000112096	SOD2	6	1.6E+08	1.6E+08	superoxide dismutase 2 [Source:HGNC Symbol;Acc:HGNC:11180]	12	100	Y
ENSG00000159433	STARD9	15	42575606	42720998	StAR related lipid transfer domain containing 9 [Source:HGNC Symbol;Acc:HGNC:19162]	12	100	Y
ENSG00000196233	LCOR	10	96832254	96995956	ligand dependent nuclear receptor corepressor [Source:HGNC Symbol;Acc:HGNC:29503]	11	100	Y
ENSG00000143127	ITGA10	1	1.46E+08	1.46E+08	integrin subunit alpha 10 [Source:HGNC Symbol;Acc:HGNC:6135]	25	84	N
ENSG00000135454	B4GALNT1	12	57623409	57633239	beta-1,4-N-acetyl-galactosaminyltransferase 1 [Source:HGNC Symbol;Acc:HGNC:4117]	22	86.364	N
ENSG00000083807	SLC27A5	19	58479512	58512413	solute carrier family 27 member 5 [Source:HGNC Symbol;Acc:HGNC:10999]	22	86.364	N
ENSG00000095383	TBC1D2	9	98199011	98255649	TBC1 domain family member 2 [Source:HGNC Symbol;Acc:HGNC:18026]	22	86.364	N
ENSG00000105711	SCN1B	19	35030470	35040449	sodium voltage-gated channel beta subunit 1 [Source:HGNC Symbol;Acc:HGNC:10586]	21	85.714	N
ENSG00000198570	RD3	1	2.11E+08	2.11E+08	RD3 regulator of GUCY2D [Source:HGNC Symbol;Acc:HGNC:19689]	20	85	N
ENSG00000129675	ARHGEF6	X	1.37E+08	1.37E+08	Rac/Cdc42 guanine nucleotide exchange factor 6 [Source:HGNC Symbol;Acc:HGNC:685]	19	89.474	N
ENSG00000058668	ATP2B4	1	2.04E+08	2.04E+08	ATPase plasma membrane Ca2+ transporting 4 [Source:HGNC Symbol;Acc:HGNC:817]	19	89.474	N
ENSG00000188878	FBF1	17	75909574	75941140	Fas binding factor 1 [Source:HGNC Symbol;Acc:HGNC:24674]	19	89.474	N
ENSG00000152061	RABGAP1L	1	1.74E+08	1.75E+08	RAB GTPase activating protein 1 like [Source:HGNC Symbol;Acc:HGNC:24663]	19	89.474	N
ENSG00000152076	CCDC74B	2	1.3E+08	1.3E+08	coiled-coil domain containing 74B [Source:HGNC Symbol;Acc:HGNC:25267]	18	88.889	N
ENSG00000111817	DSE	6	1.16E+08	1.16E+08	dermatan sulfate epimerase [Source:HGNC Symbol;Acc:HGNC:21144]	17	94.118	N
ENSG00000144648	ACKR2	3	42804752	42887974	atypical chemokine receptor 2 [Source:HGNC Symbol;Acc:HGNC:1565]	16	93.75	N
ENSG00000104899	AMH	19	2249309	2252073	anti-Mullerian hormone [Source:HGNC Symbol;Acc:HGNC:464]	16	93.75	N
ENSG00000162493	PDPN	1	13583465	13617957	podoplanin [Source:HGNC Symbol;Acc:HGNC:29602]	16	93.75	N
ENSG00000115762	PLEKHB2	2	1.31E+08	1.31E+08	pleckstrin homology domain containing B2 [Source:HGNC Symbol;Acc:HGNC:19236]	16	93.75	N
ENSG00000064933	PMS1	2	1.9E+08	1.9E+08	PMS1 homolog 1, mismatch repair system component [Source:HGNC Symbol;Acc:HGNC:9121]	16	93.75	N
ENSG00000125835	SNRPB	20	2461634	2470853	small nuclear ribonucleoprotein polypeptides B and B1 [Source:HGNC Symbol;Acc:HGNC:11153]	16	93.75	N
ENSG00000131748	STARD3	17	39637090	39664201	StAR related lipid transfer domain containing 3 [Source:HGNC Symbol;Acc:HGNC:17579]	16	93.75	N
ENSG00000118046	STK11	19	1177558	1228431	serine/threonine kinase 11 [Source:HGNC Symbol;Acc:HGNC:11389]	16	93.75	N
ENSG00000182521	TBPL2	14	55413541	55456726	TATA-box binding protein like 2 [Source:HGNC Symbol;Acc:HGNC:19841]	16	93.75	N
ENSG00000105197	TIMM50	19	39480412	39493785	translocase of inner mitochondrial membrane 50 [Source:HGNC Symbol;Acc:HGNC:23656]	16	93.75	N
ENSG00000136205	TNS3	7	47275154	47582558	tensin 3 [Source:HGNC Symbol;Acc:HGNC:21616]	16	93.75	N
ENSG00000099904	ZDHHC8	22	20129456	20148007	zinc finger DHHC-type palmitoyltransferase 8 [Source:HGNC Symbol;Acc:HGNC:18474]	16	93.75	N
ENSG00000151150	ANK3	10	60026298	60733490	ankyrin 3 [Source:HGNC Symbol;Acc:HGNC:494]	15	93.333	N
ENSG00000169814	BTD	3	15601341	15722311	biotinidase [Source:HGNC Symbol;Acc:HGNC:1122]	15	93.333	N
ENSG00000137200	CMTR1	6	37433219	37482827	cap methyltransferase 1 [Source:HGNC Symbol;Acc:HGNC:21077]	15	93.333	N
ENSG00000169372	CRADD	12	93677375	93894840	CASP2 and RIPK1 domain containing adaptor with death domain [Source:HGNC Symbol;Acc:HGNC:2340]	15	93.333	N
ENSG00000154511	DIPK1A	1	92832737	92961522	divergent protein kinase domain 1A [Source:HGNC Symbol;Acc:HGNC:32213]	15	93.333	N
ENSG00000159131	GART	21	33503931	33543491	phosphoribosylglycinamide formyltransferase [HGNC Symbol;Acc:HGNC:4163]	15	93.333	N
ENSG00000176454	LPCAT4	15	34358633	34367196	lysophosphatidylcholine acyltransferase 4 [Source:HGNC Symbol;Acc:HGNC:30059]	15	93.333	N
ENSG00000133030	MPRIP	17	17042457	17217679	myosin phosphatase Rho interacting protein [Source:HGNC Symbol;Acc:HGNC:30321]	15	93.333	N
ENSG00000204316	MRPL38	17	75898644	75905093	mitochondrial ribosomal protein L38 [Source:HGNC Symbol;Acc:HGNC:14033]	15	93.333	N
ENSG00000106443	PHF14	7	10973336	11169623	PHD finger protein 14 [Source:HGNC Symbol;Acc:HGNC:22203]	15	93.333	N
ENSG00000123739	PLA2G12A	4	1.1E+08	1.1E+08	phospholipase A2 group XIIA [Source:HGNC Symbol;Acc:HGNC:18554]	15	100	N
ENSG00000150687	PRSS23	11	86791059	86952910	serine protease 23 [Source:HGNC Symbol;Acc:HGNC:14370]	15	93.333	N
ENSG00000131242	RAB11FIP4	17	31391675	31538211	RAB11 family interacting protein 4 [Source:HGNC Symbol;Acc:HGNC:30267]	15	93.333	N
ENSG00000011454	RABGAP1	9	1.23E+08	1.23E+08	RAB GTPase activating protein 1 [Source:HGNC Symbol;Acc:HGNC:17155]	15	93.333	N
ENSG00000100106	TRIOBP	22	37697048	37776556	TRIO and F-actin binding protein [Source:HGNC Symbol;Acc:HGNC:17009]	15	93.333	N
ENSG00000151718	WWC2	4	1.83E+08	1.83E+08	WW and C2 domain containing 2 [Source:HGNC Symbol;Acc:HGNC:24148]	15	93.333	N
ENSG00000180626	ZNF594	17	5179535	5191868	zinc finger protein 594 [Source:HGNC Symbol;Acc:HGNC:29392]	15	93.333	N
ENSG00000117020	AKT3	1	2.43E+08	2.44E+08	AKT serine/threonine kinase 3 [Source:HGNC Symbol;Acc:HGNC:393]	14	92.857	N
ENSG00000055609	KMT2C	7	1.52E+08	1.52E+08	lysine methyltransferase 2C [Source:HGNC Symbol;Acc:HGNC:13726]	14	92.857	N
ENSG00000172339	ALG14	1	94974405	95072951	ALG14 UDP-N-acetylglucosaminyltransferase subunit [Source:HGNC Symbol;Acc:HGNC:28287]	13	100	N

Table A3, cont.

ENSG00000175224	ATG13	11	46617527	46674818	autophagy related 13 [Source:HGNC Symbol;Acc:HGNC:29091]	13	100	N
ENSG00000204217	BMPR2	2	2.02E+08	2.03E+08	bone morphogenetic protein receptor type 2 [Source:HGNC Symbol;Acc:HGNC:1078]	13	100	N
ENSG00000163840	DTX3L	3	1.23E+08	1.23E+08	deltex E3 ubiquitin ligase 3L [Source:HGNC Symbol;Acc:HGNC:30323]	13	100	N
ENSG00000188906	LRRK2	12	40196744	40369285	leucine rich repeat kinase 2 [Source:HGNC Symbol;Acc:HGNC:18618]	13	100	N
ENSG00000152256	PDK1	2	1.73E+08	1.73E+08	pyruvate dehydrogenase kinase 1 [Source:HGNC Symbol;Acc:HGNC:8809]	13	100	N
ENSG00000121716	PILRB	7	1E+08	1E+08	paired immunoglobin like type 2 receptor beta [Source:HGNC Symbol;Acc:HGNC:18297]	13	100	N
ENSG00000132275	RRP8	11	6595072	6603616	ribosomal RNA processing 8 [Source:HGNC Symbol;Acc:HGNC:29030]	13	100	N
ENSG00000004939	SLC4A1	17	44248390	44268141	solute carrier family 4 member 1 (Diego blood group) [Source:HGNC Symbol;Acc:HGNC:11027]	13	100	N
ENSG00000102710	SUPT20H	13	37009312	37059713	SPT20 homolog, SAGA complex component [Source:HGNC Symbol;Acc:HGNC:20596]	13	100	N
ENSG00000147687	TATDN1	8	1.24E+08	1.25E+08	TatD DNase domain containing 1 [Source:HGNC Symbol;Acc:HGNC:24220]	13	100	N
ENSG00000198270	TMEM116	12	1.12E+08	1.12E+08	transmembrane protein 116 [Source:HGNC Symbol;Acc:HGNC:25084]	13	100	N
ENSG00000100284	TOM1	22	35299275	35347992	target of myb1 membrane trafficking protein [Source:HGNC Symbol;Acc:HGNC:11982]	13	100	N
ENSG00000108599	AKAP10	17	19904302	19978343	A-kinase anchoring protein 10 [Source:HGNC Symbol;Acc:HGNC:368]	12	100	N
ENSG00000131620	ANO1	11	69985907	70189530	anoctamin 1 [Source:HGNC Symbol;Acc:HGNC:21625]	12	100	N
ENSG00000101639	CEP192	18	12991362	13125052	centrosomal protein 192 [Source:HGNC Symbol;Acc:HGNC:25515]	12	100	N
ENSG00000113163	CERT1	5	75356345	75512138	ceramide transporter 1 [Source:HGNC Symbol;Acc:HGNC:2205]	12	100	N
ENSG00000213923	CSNK1E	22	38290691	38318084	casein kinase 1 epsilon [Source:HGNC Symbol;Acc:HGNC:2453]	12	100	N
ENSG00000187954	CYHR1	8	1.44E+08	1.44E+08	cysteine and histidine rich 1 [Source:HGNC Symbol;Acc:HGNC:17806]	12	100	N
ENSG00000136848	DAB2IP	9	1.22E+08	1.22E+08	DAB2 interacting protein [Source:HGNC Symbol;Acc:HGNC:17294]	12	100	N
ENSG00000197635	**DPP4**	**2**	**1.62E+08**	**1.62E+08**	**dipeptidyl peptidase 4 [Source:HGNC Symbol;Acc:HGNC:3009]**	12	100	N
ENSG00000087502	ERGIC2	12	29337352	29381189	ERGIC and golgi 2 [Source:HGNC Symbol;Acc:HGNC:30208]	12	100	N
ENSG00000177150	FAM210A	18	13663347	13726663	family with sequence similarity 210 member A [Source:HGNC Symbol;Acc:HGNC:28346]	12	100	N
ENSG00000187239	FNBP1	9	1.3E+08	1.3E+08	formin binding protein 1 [Source:HGNC Symbol;Acc:HGNC:17069]	12	100	N
ENSG00000156650	KAT6B	10	74824927	75032624	lysine acetyltransferase 6B [Source:HGNC Symbol;Acc:HGNC:17582]	12	100	N
ENSG00000124159	MATN4	20	45293445	45308529	matrilin 4 [Source:HGNC Symbol;Acc:HGNC:6910]	12	100	N
ENSG00000119684	MLH3	14	75013769	75051532	mutL homolog 3 [Source:HGNC Symbol;Acc:HGNC:7128]	12	100	N
ENSG00000159256	MORC3	21	36320189	36386148	MORC family CW-type zinc finger 3 [Source:HGNC Symbol;Acc:HGNC:23572]	12	100	N
ENSG00000147684	NDUFB9	8	1.25E+08	1.25E+08	NADH:ubiquinone oxidoreductase subunit B9 [Source:HGNC Symbol;Acc:HGNC:7704]	12	100	N
ENSG00000171631	P2RY6	11	73264498	73305103	pyrimidinergic receptor P2Y6 [Source:HGNC Symbol;Acc:HGNC:8543]	12	100	N
ENSG00000066379	POLR1H	6	30058899	30064909	RNA polymerase I subunit H [Source:HGNC Symbol;Acc:HGNC:13182]	12	100	N
ENSG00000059915	PSD	10	1.02E+08	1.02E+08	pleckstrin and Sec7 domain containing [Source:HGNC Symbol;Acc:HGNC:9507]	12	100	N
ENSG00000132334	**PTPRE**	**10**	**1.28E+08**	**1.28E+08**	**protein tyrosine phosphatase receptor type E [Source:HGNC Symbol;Acc:HGNC:9669]**	12	100	N
ENSG00000144724	PTPRG	3	61561569	62297609	protein tyrosine phosphatase receptor type G [Source:HGNC Symbol;Acc:HGNC:9671]	12	100	N
ENSG00000187024	PTRH1	9	1.28E+08	1.28E+08	peptidyl-tRNA hydrolase 1 homolog [Source:HGNC Symbol;Acc:HGNC:27039]	12	100	N
ENSG00000136104	RNASEH2E	13	50909747	51024120	ribonuclease H2 subunit B [Source:HGNC Symbol;Acc:HGNC:25671]	12	100	N
ENSG00000197713	RPE	2	2.1E+08	2.1E+08	ribulose-5-phosphate-3-epimerase [Source:HGNC Symbol;Acc:HGNC:10293]	12	100	N
ENSG00000172809	RPL38	17	74203582	74210655	ribosomal protein L38 [Source:HGNC Symbol;Acc:HGNC:10349]	12	100	N
ENSG00000052749	RRP12	10	97356358	97426076	ribosomal RNA processing 12 homolog [Source:HGNC Symbol;Acc:HGNC:29100]	12	100	N
ENSG00000123453	SARDH	9	1.34E+08	1.34E+08	sarcosine dehydrogenase [Source:HGNC Symbol;Acc:HGNC:10536]	12	100	N
ENSG00000138674	SEC31A	4	82818509	82901166	SEC31 homolog A, COPII coat complex component [Source:HGNC Symbol;Acc:HGNC:17052]	12	100	N
ENSG00000086300	SNX10	7	26291862	26374383	sorting nexin 10 [Source:HGNC Symbol;Acc:HGNC:14974]	12	100	N
ENSG00000137642	SORL1	11	1.21E+08	1.22E+08	sortilin related receptor 1 [Source:HGNC Symbol;Acc:HGNC:11185]	12	100	N
ENSG00000176148	TCP11L1	11	33039417	33105943	t-complex 11 like 1 [Source:HGNC Symbol;Acc:HGNC:25655]	12	100	N
ENSG00000164168	TMEM184	4	1.48E+08	1.48E+08	transmembrane protein 184C [Source:HGNC Symbol;Acc:HGNC:25587]	12	100	N
ENSG00000139668	WDFY2	13	51584455	51767709	WD repeat and FYVE domain containing 2 [Source:HGNC Symbol;Acc:HGNC:20482]	12	100	N
ENSG00000139985	ADAM21	14	70417107	70460427	ADAM metallopeptidase domain 21 [Source:HGNC Symbol;Acc:HGNC:200]	11	100	N

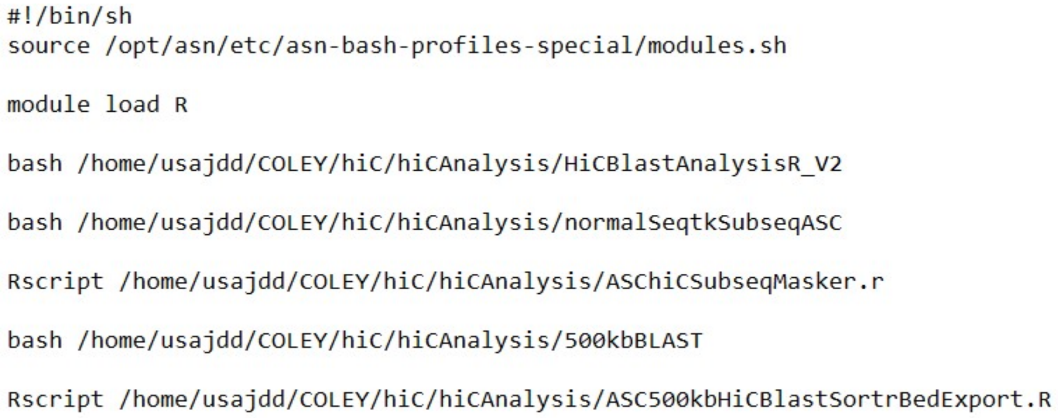
Appendix G: LG4 TAD Masker R Script



```r
library(dplyr)
library(tidyr)
library(stringr)
library(splitstackshape)

##retrieve list of 500kb db and lg4 db files##
input500kbDbFiles<-list.files(path="C:/Users/alexs/Desktop/temp", pattern = "500_kb_window.fasta")
inputLG4dbFiles<-list.files(path="C:/Users/alexs/Desktop/temp", pattern = "seq.fasta")

##match db files based on LG4 ID##
input500kbDbFiles2<-data.frame()
for(i in 1:length(input500kbDbFiles)){
  input500kbDbFiles2[i,1]<-as.character(str_extract(input500kbDbFiles[i],pattern = "LG4_\\d+_\\d+"))
}

inputLG4dbFiles2<-data.frame()
for(i in 1:length(input500kbDbFiles)){
  inputLG4dbFiles2[i,1]<-str_extract(inputLG4dbFiles[i],pattern = "LG4_\\d+_\\d+")
}

fileDirectory<-data.frame("500kbDbFiles"=input500kbDbFiles,"LG4dbFiles"="")
for (i in 1:length(input500kbDbFiles)){
  LG4DbInd<-which(inputLG4dbFiles2[,1]==as.character(input500kbDbFiles2[i,1]))
  fileDirectory[i,2]<-inputLG4dbFiles[LG4DbInd]
}

##use fileDirectory to make a new inputpathDirectory DF that has a full path to each input file, matched across rows##
inputPathDirectory<-data.frame(row.names = 1:(length(input500kbDbFiles)))
for (i in 1:2){
  inputPathDirectory[,i]<-str_c("C:/Users/alexs/Desktop/temp/",fileDirectory[,i])
}

##begin loop##
for(i in 1:length(input500kbDbFiles)){
  ##upload LG4 db and get chr start/stop positions isolated##
  LG4db <- read.table(inputPathDirectory[i,2], quote="\"", comment.char="")
  LG4StartStop<-data.frame("lg4"=(str_extract(LG4db[1,1],pattern = "GRCh38:\\S+")))
  LG4StartStop<-cSplit(LG4StartStop,splitCols=1,sep=":",direction="wide",drop = TRUE)
  ##upload LG4 500kb db and do the same##
  LG4500kbdb <- read.table(inputPathDirectory[i,1], quote="\"", comment.char="")
  LG4500kbStartStop<-data.frame("lg4"=(str_extract(LG4500kbdb[1,1],pattern = "GRCh38:\\S+")))
  LG4500kbStartStop<-cSplit(LG4500kbStartStop,splitCols=1,sep=":",direction="wide",drop = TRUE)
  ##calculate LG4 start -5000 and stop +5000 positions within the LG4 +/-500 kb db##
  #maskedStartStop<-data.frame("start"=(LG4StartStop[1,3]-LG4500kbStartStop[1,3]+1),"stop"="")
  maskedStartStop<-data.frame("start"=(LG4StartStop[1,3]-LG4500kbStartStop[1,3]+1-5000),"stop"="")
  #maskedStartStop[1,2]<-(maskedStartStop[1,1]+(LG4StartStop[1,4]-LG4StartStop[1,3]))
  maskedStartStop[1,2]<-(maskedStartStop[1,1]+10000+(LG4StartStop[1,4]-LG4StartStop[1,3]))
  ##replace LG4500kbdb seq corresponding to original LG4 position with "N"'s##
  seqVar<-as.vector(LG4500kbdb[2,1])
  seqSplit<-unlist(str_extract_all(seqVar,boundary("character")))
  ##If masked start/stop fit within +/-500kb start/stop then use them to replace with N's. If not then use "1" for start or "LG4500kbStartStop[,4]" for stop
  if(as.numeric(maskedStartStop[1,1])<0){
    seqSplit[1:(as.numeric(maskedStartStop[1,2]))]<-"N"
  }
    if(as.numeric(maskedStartStop[1,2])>as.numeric(LG4500kbStartStop[1,4])){
      seqSplit[(as.numeric(maskedStartStop[1,1])):(as.numeric(LG4500kbStartStop[1,4]))]<-"N"
    }
      else{
        seqSplit[(as.numeric(maskedStartStop[1,1])):(as.numeric(maskedStartStop[1,2]))]<-"N"
      }
  seqSplitFinal<-vector(length=0)
  for (s in 1:length(seqSplit)){
    seqSplitFinal<-str_c(seqSplitFinal,seqSplit[s])
  }
  LG4500kbdb[2,1]<-seqSplitFinal
  ##output path and name is the same as input path and name bc I'm overwriting the existing DBs##
  write.table(LG4500kbdb,file = inputPathDirectory[i,1],row.names = FALSE,col.names = FALSE,quote = FALSE,)
}
```

Figure A4: LG4 TAD Masker R Script. R script encoding the LG4 TAD Masker step of the HiC pipeline.

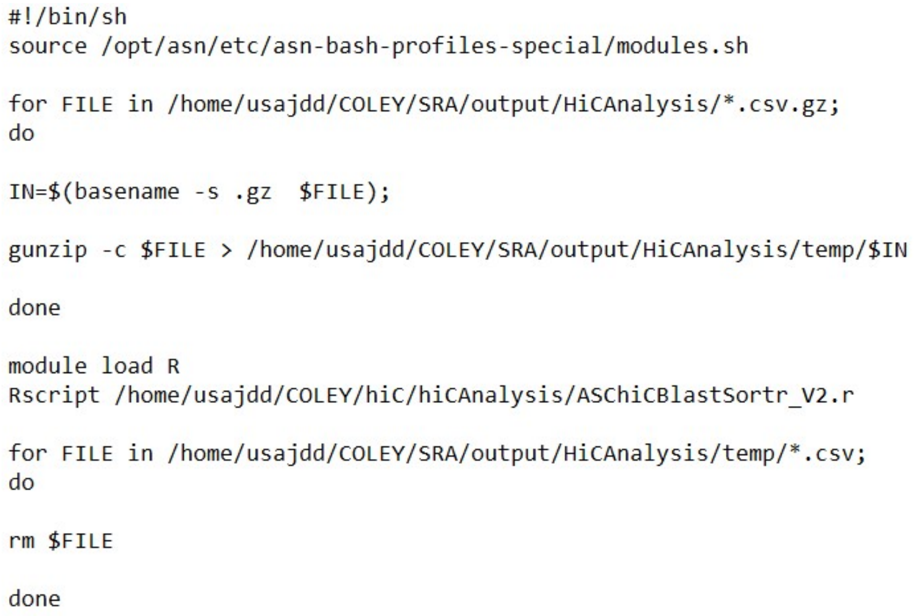
Appendix H: Execute HiC Analysis Bash Script



```
#!/bin/sh
source /opt/asn/etc/asn-bash-profiles-special/modules.sh

module load R

bash /home/usajdd/COLEY/hiC/hiCAnalysis/HiCBlastAnalysisR_V2

bash /home/usajdd/COLEY/hiC/hiCAnalysis/normalSeqtkSubseqASC

Rscript /home/usajdd/COLEY/hiC/hiCAnalysis/ASChiCSubseqMasker.r

bash /home/usajdd/COLEY/hiC/hiCAnalysis/500kbBLAST

Rscript /home/usajdd/COLEY/hiC/hiCAnalysis/ASC500kbHiCBlastSortrBedExport.R
```

Figure A5: Execute HiC Analysis Bash Script. Bash script encoding the overall execution of the HiC pipeline components.

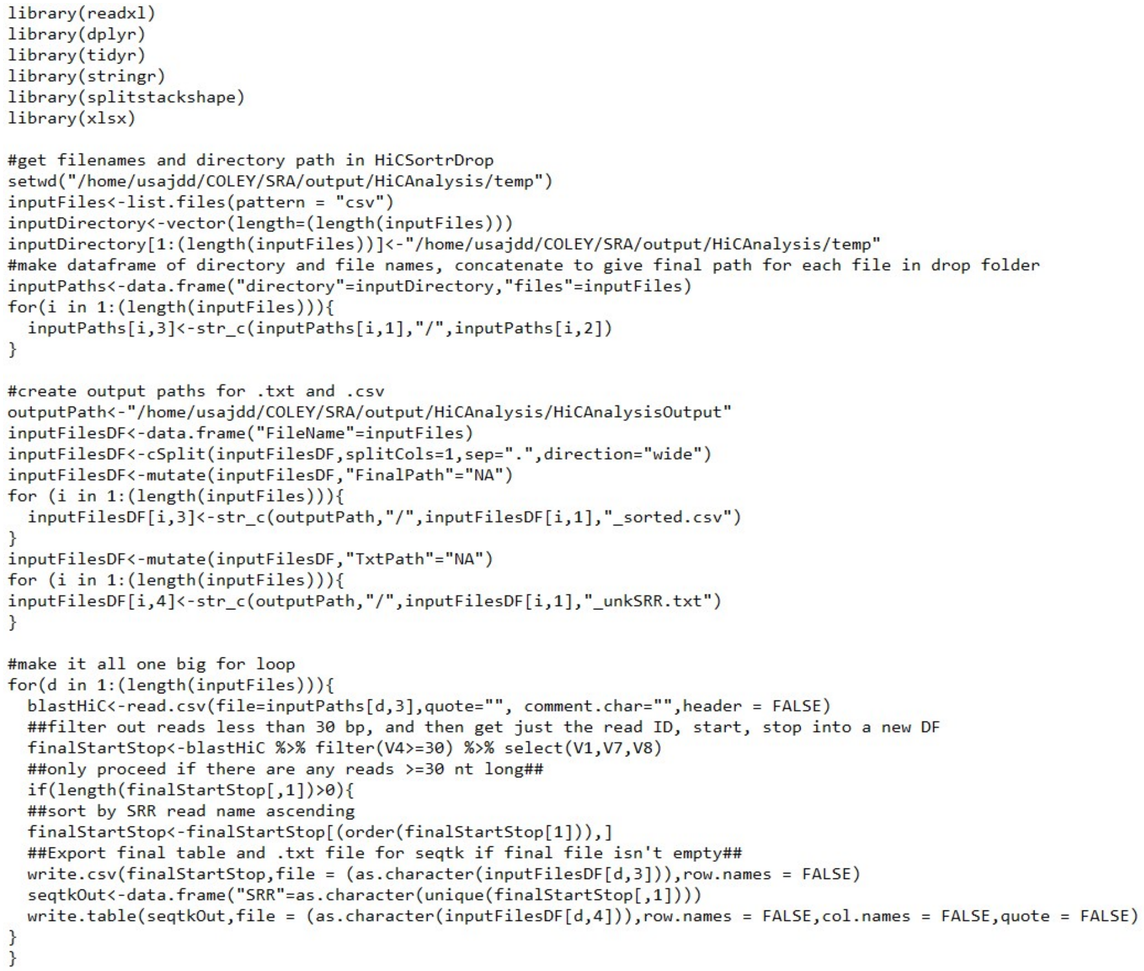
Appendix I: HiC Blast Analysis Bash Script



```
#!/bin/sh
source /opt/asn/etc/asn-bash-profiles-special/modules.sh

for FILE in /home/usajdd/COLEY/SRA/output/HiCAnalysis/*.csv.gz;
do

IN=$(basename -s .gz  $FILE);

gunzip -c $FILE > /home/usajdd/COLEY/SRA/output/HiCAnalysis/temp/$IN

done

module load R
Rscript /home/usajdd/COLEY/hiC/hiCAnalysis/ASChiCBlastSortr_V2.r

for FILE in /home/usajdd/COLEY/SRA/output/HiCAnalysis/temp/*.csv;
do

rm $FILE

done
```

Figure A6: HiC Blast Analysis Bash Script. Bash script encoding the BLAST analysis step of the HiC pipeline.

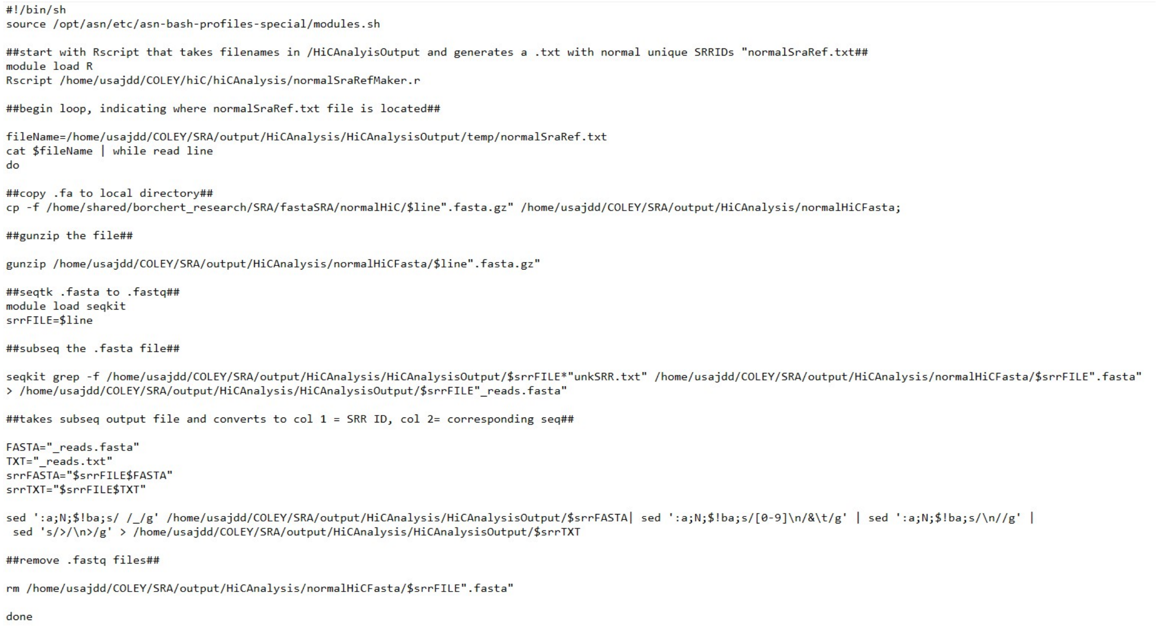
Appendix J: HiC Blast SortR R Script



```
library(readxl)
library(dplyr)
library(tidyr)
library(stringr)
library(splitstackshape)
library(xlsx)

#get filenames and directory path in HiCSortrDrop
setwd("/home/usajdd/COLEY/SRA/output/HiCAnalysis/temp")
inputFiles<-list.files(pattern = "csv")
inputDirectory<-vector(length=(length(inputFiles)))
inputDirectory[1:(length(inputFiles))]<-"/home/usajdd/COLEY/SRA/output/HiCAnalysis/temp"
#make dataframe of directory and file names, concatenate to give final path for each file in drop folder
inputPaths<-data.frame("directory"=inputDirectory,"files"=inputFiles)
for(i in 1:(length(inputFiles))){
  inputPaths[i,3]<-str_c(inputPaths[i,1],"/",inputPaths[i,2])
}

#create output paths for .txt and .csv
outputPath<-"/home/usajdd/COLEY/SRA/output/HiCAnalysis/HiCAnalysisOutput"
inputFilesDF<-data.frame("FileName"=inputFiles)
inputFilesDF<-cSplit(inputFilesDF,splitCols=1,sep=".",direction="wide")
inputFilesDF<-mutate(inputFilesDF,"FinalPath"="NA")
for (i in 1:(length(inputFiles))){
  inputFilesDF[i,3]<-str_c(outputPath,"/",inputFilesDF[i,1],"_sorted.csv")
}
inputFilesDF<-mutate(inputFilesDF,"TxtPath"="NA")
for (i in 1:(length(inputFiles))){
inputFilesDF[i,4]<-str_c(outputPath,"/",inputFilesDF[i,1],"_unkSRR.txt")
}

#make it all one big for loop
for(d in 1:(length(inputFiles))){
  blastHiC<-read.csv(file=inputPaths[d,3],quote="", comment.char="",header = FALSE)
  ##filter out reads less than 30 bp, and then get just the read ID, start, stop into a new DF
  finalStartStop<-blastHiC %>% filter(V4>=30) %>% select(V1,V7,V8)
  ##only proceed if there are any reads >=30 nt long##
  if(length(finalStartStop[,1])>0){
  ##sort by SRR read name ascending
  finalStartStop<-finalStartStop[(order(finalStartStop[1])),]
  ##Export final table and .txt file for seqtk if final file isn't empty##
  write.csv(finalStartStop,file = (as.character(inputFilesDF[d,3])),row.names = FALSE)
  seqtkOut<-data.frame("SRR"=as.character(unique(finalStartStop[,1])))
  write.table(seqtkOut,file = (as.character(inputFilesDF[d,4])),row.names = FALSE,col.names = FALSE,quote = FALSE)
}
}
```

Figure A7: HiC Blast SortR R Script. R script encoding the BLAST sortR analysis step of the HiC pipeline.

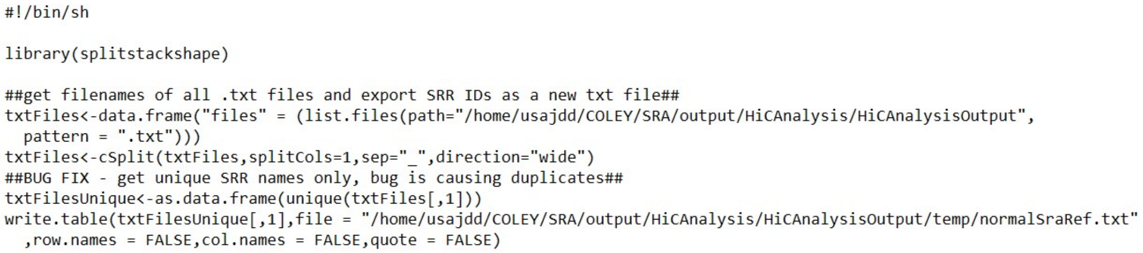
Appendix K: HiC Seqtk Bash Script



```
#!/bin/sh
source /opt/asn/etc/asn-bash-profiles-special/modules.sh

##start with Rscript that takes filenames in /HiCAnalyisOutput and generates a .txt with normal unique SRRIDs "normalSraRef.txt##
module load R
Rscript /home/usajdd/COLEY/hiC/hiCAnalysis/normalSraRefMaker.r

##begin loop, indicating where normalSraRef.txt file is located##

fileName=/home/usajdd/COLEY/SRA/output/HiCAnalysis/HiCAnalysisOutput/temp/normalSraRef.txt
cat $fileName | while read line
do

##copy .fa to local directory##
cp -f /home/shared/borchert_research/SRA/fastaSRA/normalHiC/$line".fasta.gz" /home/usajdd/COLEY/SRA/output/HiCAnalysis/normalHiCFasta;

##gunzip the file##

gunzip /home/usajdd/COLEY/SRA/output/HiCAnalysis/normalHiCFasta/$line".fasta.gz"

##seqtk .fasta to .fastq##
module load seqkit
srrFILE=$line

##subseq the .fasta file##

seqkit grep -f /home/usajdd/COLEY/SRA/output/HiCAnalysis/HiCAnalysisOutput/$srrFILE*"unkSRR.txt" /home/usajdd/COLEY/SRA/output/HiCAnalysis/normalHiCFasta/$srrFILE".fasta"
> /home/usajdd/COLEY/SRA/output/HiCAnalysis/HiCAnalysisOutput/$srrFILE"_reads.fasta"

##takes subseq output file and converts to col 1 = SRR ID, col 2= corresponding seq##

FASTA="_reads.fasta"
TXT="_reads.txt"
srrFASTA="$srrFILE$FASTA"
srrTXT="$srrFILE$TXT"

sed ':a;N;$!ba;s/ /_/g' /home/usajdd/COLEY/SRA/output/HiCAnalysis/HiCAnalysisOutput/$srrFASTA| sed ':a;N;$!ba;s/[0-9]\n/&\t/g' | sed ':a;N;$!ba;s/\n//g' |
 sed 's/>/\n>/g' > /home/usajdd/COLEY/SRA/output/HiCAnalysis/HiCAnalysisOutput/$srrTXT

##remove .fastq files##

rm /home/usajdd/COLEY/SRA/output/HiCAnalysis/normalHiCFasta/$srrFILE".fasta"

done
```

Figure A8: HiC Seqtk Bash Script. Bash script encoding the seqtk sequence retrieval step of the HiC pipeline.

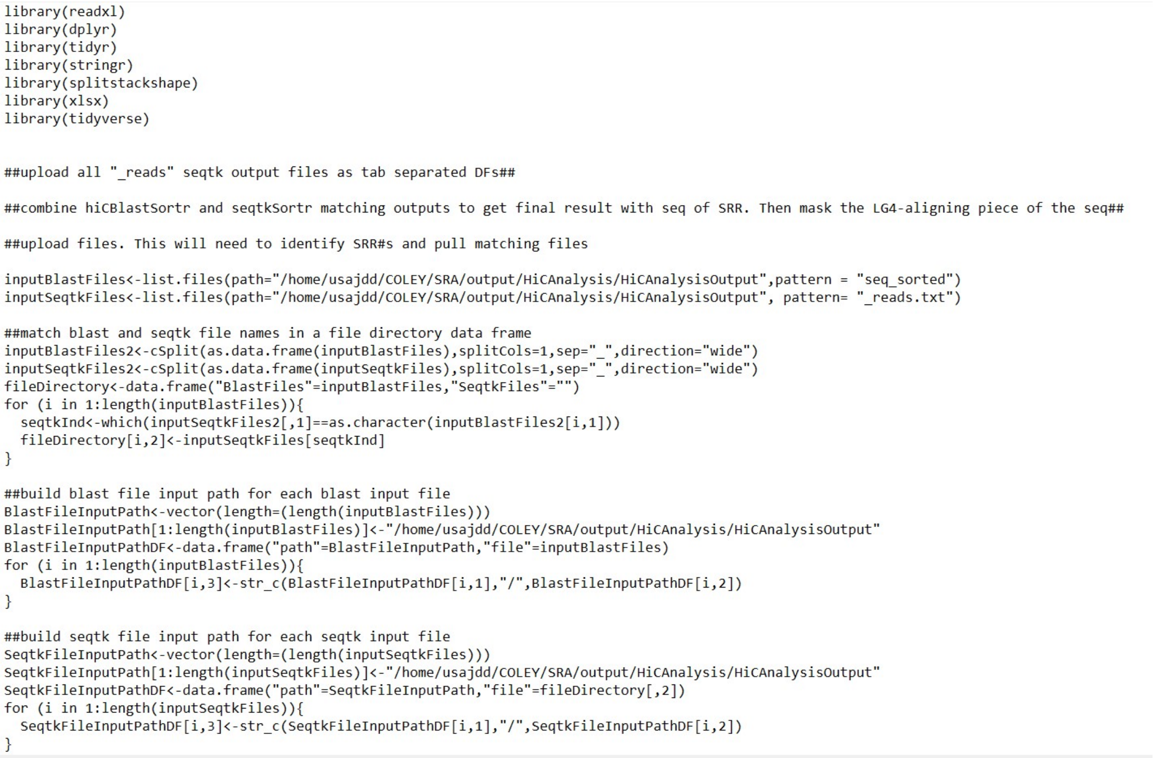
Appendix L: HiC SRA Reference Maker R Script



```
#!/bin/sh

library(splitstackshape)

##get filenames of all .txt files and export SRR IDs as a new txt file##
txtFiles<-data.frame("files" = (list.files(path="/home/usajdd/COLEY/SRA/output/HiCAnalysis/HiCAnalysisOutput",
  pattern = ".txt")))
txtFiles<-cSplit(txtFiles,splitCols=1,sep="_",direction="wide")
##BUG FIX - get unique SRR names only, bug is causing duplicates##
txtFilesUnique<-as.data.frame(unique(txtFiles[,1]))
write.table(txtFilesUnique[,1],file = "/home/usajdd/COLEY/SRA/output/HiCAnalysis/HiCAnalysisOutput/temp/normalSraRef.txt"
  ,row.names = FALSE,col.names = FALSE,quote = FALSE)
```

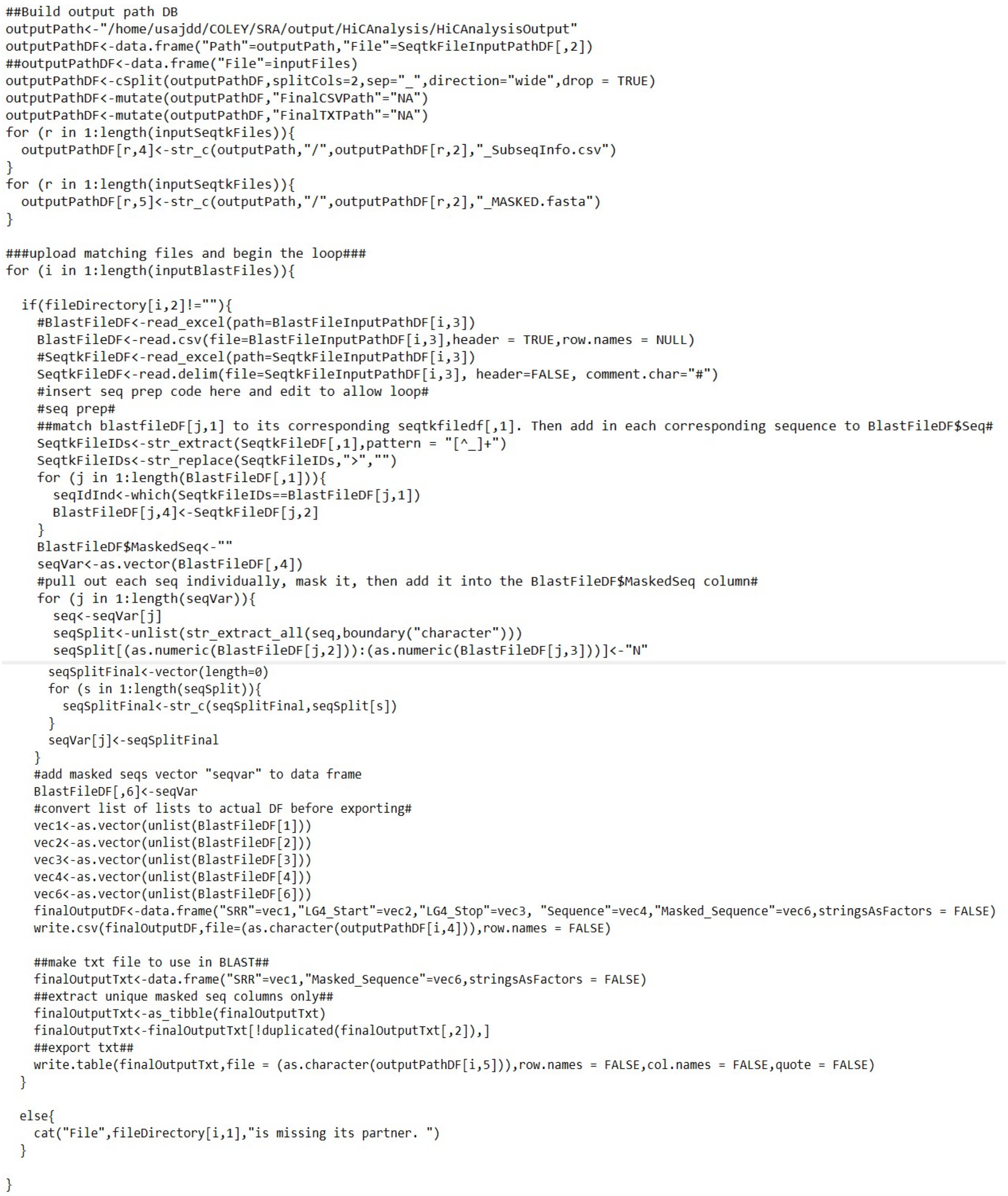
Figure A9: HiC SRA Reference Maker R Script. R script encoding the SRA reference step of the HiC pipeline.

Appendix M: HiC Subseq Masker R Script



```
library(readxl)
library(dplyr)
library(tidyr)
library(stringr)
library(splitstackshape)
library(xlsx)
library(tidyverse)


##upload all "_reads" seqtk output files as tab separated DFs##

##combine hiCBlastSortr and seqtkSortr matching outputs to get final result with seq of SRR. Then mask the LG4-aligning piece of the seq##

##upload files. This will need to identify SRR#s and pull matching files

inputBlastFiles<-list.files(path="/home/usajdd/COLEY/SRA/output/HiCAnalysis/HiCAnalysisOutput",pattern = "seq_sorted")
inputSeqtkFiles<-list.files(path="/home/usajdd/COLEY/SRA/output/HiCAnalysis/HiCAnalysisOutput", pattern= "_reads.txt")

##match blast and seqtk file names in a file directory data frame
inputBlastFiles2<-cSplit(as.data.frame(inputBlastFiles),splitCols=1,sep="_",direction="wide")
inputSeqtkFiles2<-cSplit(as.data.frame(inputSeqtkFiles),splitCols=1,sep="_",direction="wide")
fileDirectory<-data.frame("BlastFiles"=inputBlastFiles,"SeqtkFiles"="")
for (i in 1:length(inputBlastFiles)){
  seqtkInd<-which(inputSeqtkFiles2[,1]==as.character(inputBlastFiles2[i,1]))
  fileDirectory[i,2]<-inputSeqtkFiles[seqtkInd]
}

##build blast file input path for each blast input file
BlastFileInputPath<-vector(length=(length(inputBlastFiles)))
BlastFileInputPath[1:length(inputBlastFiles)]<-"/home/usajdd/COLEY/SRA/output/HiCAnalysis/HiCAnalysisOutput"
BlastFileInputPathDF<-data.frame("path"=BlastFileInputPath,"file"=inputBlastFiles)
for (i in 1:length(inputBlastFiles)){
  BlastFileInputPathDF[i,3]<-str_c(BlastFileInputPathDF[i,1],"/",BlastFileInputPathDF[i,2])
}

##build seqtk file input path for each seqtk input file
SeqtkFileInputPath<-vector(length=(length(inputSeqtkFiles)))
SeqtkFileInputPath[1:length(inputSeqtkFiles)]<-"/home/usajdd/COLEY/SRA/output/HiCAnalysis/HiCAnalysisOutput"
SeqtkFileInputPathDF<-data.frame("path"=SeqtkFileInputPath,"file"=fileDirectory[,2])
for (i in 1:length(inputSeqtkFiles)){
  SeqtkFileInputPathDF[i,3]<-str_c(SeqtkFileInputPathDF[i,1],"/",SeqtkFileInputPathDF[i,2])
}
```

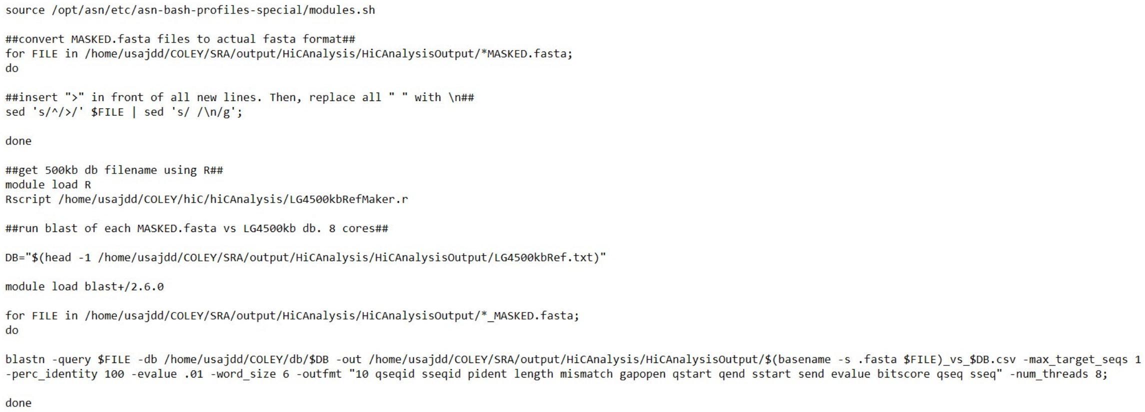
Figure A10: HiC Subseq Masker R Script. R script encoding the masking step of the HiC pipeline.



```
##Build output path DB
outputPath<-"/home/usajdd/COLEY/SRA/output/HiCAnalysis/HiCAnalysisOutput"
outputPathDF<-data.frame("Path"=outputPath,"File"=SeqtkFileInputPathDF[,2])
##outputPathDF<-data.frame("File"=inputFiles)
outputPathDF<-cSplit(outputPathDF,splitCols=2,sep="_",direction="wide",drop = TRUE)
outputPathDF<-mutate(outputPathDF,"FinalCSVPath"="NA")
outputPathDF<-mutate(outputPathDF,"FinalTXTPath"="NA")
for (r in 1:length(inputSeqtkFiles)){
  outputPathDF[r,4]<-str_c(outputPath,"/",outputPathDF[r,2],"_SubseqInfo.csv")
}
for (r in 1:length(inputSeqtkFiles)){
  outputPathDF[r,5]<-str_c(outputPath,"/",outputPathDF[r,2],"_MASKED.fasta")
}

###upload matching files and begin the loop###
for (i in 1:length(inputBlastFiles)){

  if(fileDirectory[i,2]!=""){
    #BlastFileDF<-read_excel(path=BlastFileInputPathDF[i,3])
    BlastFileDF<-read.csv(file=BlastFileInputPathDF[i,3],header = TRUE,row.names = NULL)
    #SeqtkFileDF<-read_excel(path=SeqtkFileInputPathDF[i,3])
    SeqtkFileDF<-read.delim(file=SeqtkFileInputPathDF[i,3], header=FALSE, comment.char="#")
    #insert seq prep code here and edit to allow loop#
    #seq prep#
    ##match blastfileDF[j,1] to its corresponding seqtkfiledf[,1]. Then add in each corresponding sequence to BlastFileDF$Seq#
    SeqtkFileIDs<-str_extract(SeqtkFileDF[,1],pattern = "[^_]+")
    SeqtkFileIDs<-str_replace(SeqtkFileIDs,">","")
    for (j in 1:length(BlastFileDF[,1])){
      seqIdInd<-which(SeqtkFileIDs==BlastFileDF[j,1])
      BlastFileDF[j,4]<-SeqtkFileDF[j,2]
    }
    BlastFileDF$MaskedSeq<-""
    seqVar<-as.vector(BlastFileDF[,4])
    #pull out each seq individually, mask it, then add it into the BlastFileDF$MaskedSeq column#
    for (j in 1:length(seqVar)){
      seq<-seqVar[j]
      seqSplit<-unlist(str_extract_all(seq,boundary("character")))
      seqSplit[(as.numeric(BlastFileDF[j,2])):(as.numeric(BlastFileDF[j,3]))]<-"N"
      seqSplitFinal<-vector(length=0)
      for (s in 1:length(seqSplit)){
        seqSplitFinal<-str_c(seqSplitFinal,seqSplit[s])
      }
      seqVar[j]<-seqSplitFinal
    }
    #add masked seqs vector "seqvar" to data frame
    BlastFileDF[,6]<-seqVar
    #convert list of lists to actual DF before exporting#
    vec1<-as.vector(unlist(BlastFileDF[1]))
    vec2<-as.vector(unlist(BlastFileDF[2]))
    vec3<-as.vector(unlist(BlastFileDF[3]))
    vec4<-as.vector(unlist(BlastFileDF[4]))
    vec6<-as.vector(unlist(BlastFileDF[6]))
    finalOutputDF<-data.frame("SRR"=vec1,"LG4_Start"=vec2,"LG4_Stop"=vec3, "Sequence"=vec4,"Masked_Sequence"=vec6,stringsAsFactors = FALSE)
    write.csv(finalOutputDF,file=(as.character(outputPathDF[i,4])),row.names = FALSE)

    ##make txt file to use in BLAST##
    finalOutputTxt<-data.frame("SRR"=vec1,"Masked_Sequence"=vec6,stringsAsFactors = FALSE)
    ##extract unique masked seq columns only##
    finalOutputTxt<-as_tibble(finalOutputTxt)
    finalOutputTxt<-finalOutputTxt[!duplicated(finalOutputTxt[,2]),]
    ##export txt##
    write.table(finalOutputTxt,file = (as.character(outputPathDF[i,5])),row.names = FALSE,col.names = FALSE,quote = FALSE)
  }

  else{
    cat("File",fileDirectory[i,1],"is missing its partner. ")
  }

}
```

Figure A10, cont.

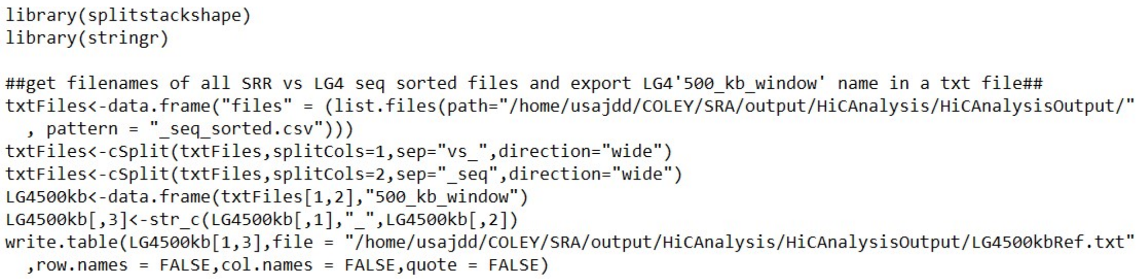
Appendix N: HiC 500kb Blast Bash Script



```
source /opt/asn/etc/asn-bash-profiles-special/modules.sh

##convert MASKED.fasta files to actual fasta format##
for FILE in /home/usajdd/COLEY/SRA/output/HiCAnalysis/HiCAnalysisOutput/*MASKED.fasta;
do

##insert ">" in front of all new lines. Then, replace all " " with \n##
sed 's/^/>/' $FILE | sed 's/ /\n/g';

done

##get 500kb db filename using R##
module load R
Rscript /home/usajdd/COLEY/hiC/hiCAnalysis/LG4500kbRefMaker.r

##run blast of each MASKED.fasta vs LG4500kb db. 8 cores##

DB="$(head -1 /home/usajdd/COLEY/SRA/output/HiCAnalysis/HiCAnalysisOutput/LG4500kbRef.txt)"

module load blast+/2.6.0

for FILE in /home/usajdd/COLEY/SRA/output/HiCAnalysis/HiCAnalysisOutput/*_MASKED.fasta;
do

blastn -query $FILE -db /home/usajdd/COLEY/db/$DB -out /home/usajdd/COLEY/SRA/output/HiCAnalysis/HiCAnalysisOutput/$(basename -s .fasta $FILE)_vs_$DB.csv -max_target_seqs 1
-perc_identity 100 -evalue .01 -word_size 6 -outfmt "10 qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore qseq sseq" -num_threads 8;

done
```

Figure A11: HiC 500kb Blast Bash Script. Bash script encoding the 500 kb reference BLAST step of the HiC pipeline.

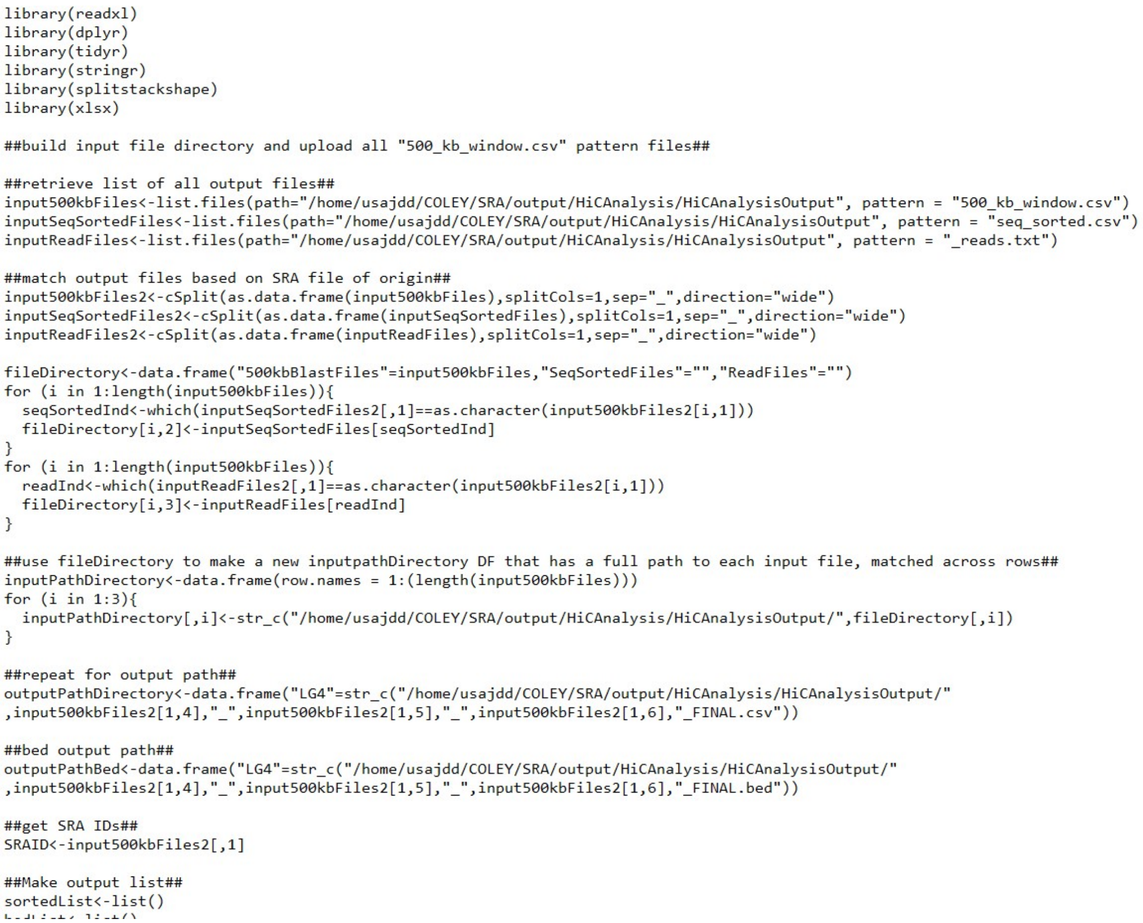
Appendix O: HiC 500kb Blast Reference Maker R Script



```
library(splitstackshape)
library(stringr)

##get filenames of all SRR vs LG4 seq sorted files and export LG4'500_kb_window' name in a txt file##
txtFiles<-data.frame("files" = (list.files(path="/home/usajdd/COLEY/SRA/output/HiCAnalysis/HiCAnalysisOutput/"
  , pattern = "_seq_sorted.csv")))
txtFiles<-cSplit(txtFiles,splitCols=1,sep="vs_",direction="wide")
txtFiles<-cSplit(txtFiles,splitCols=2,sep="_seq",direction="wide")
LG4500kb<-data.frame(txtFiles[1,2],"500_kb_window")
LG4500kb[,3]<-str_c(LG4500kb[,1],"_",LG4500kb[,2])
write.table(LG4500kb[1,3],file = "/home/usajdd/COLEY/SRA/output/HiCAnalysis/HiCAnalysisOutput/LG4500kbRef.txt"
  ,row.names = FALSE,col.names = FALSE,quote = FALSE)
```

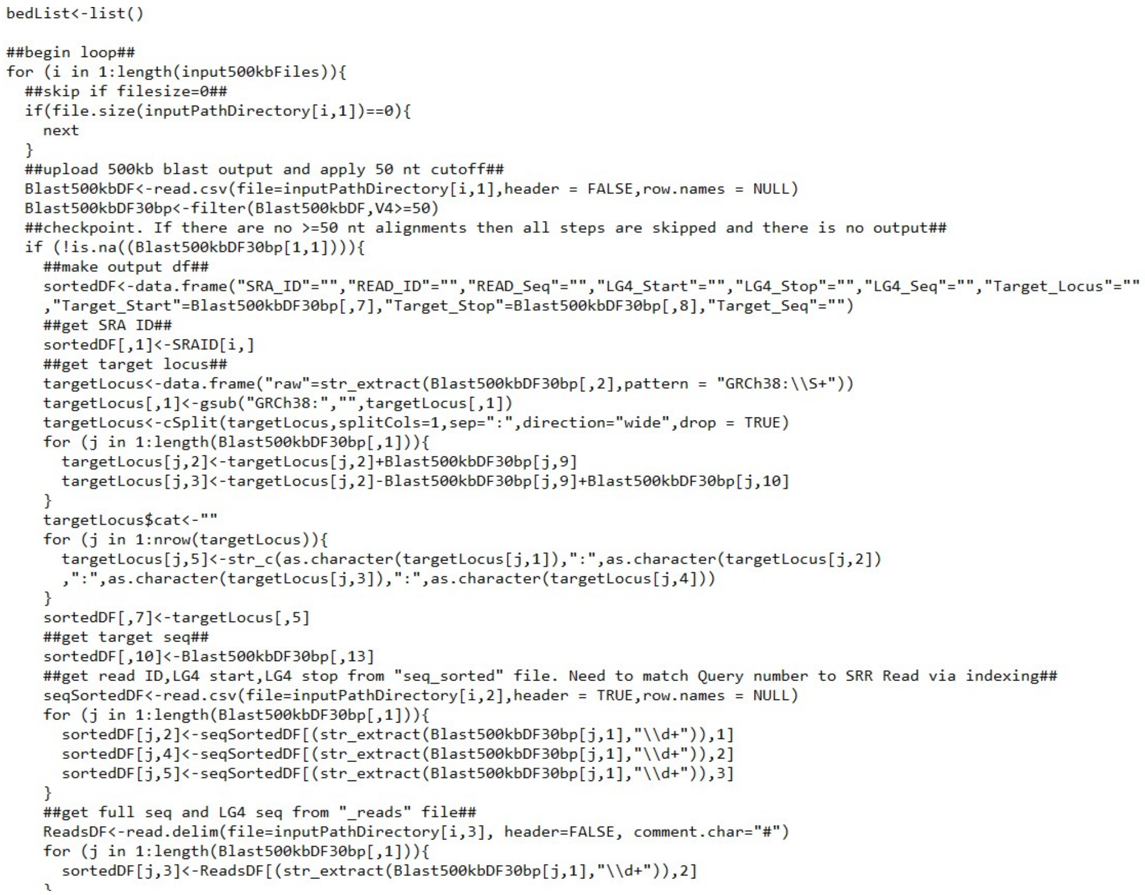
Figure A12: HiC 500kb Blast Reference Maker R Script. R script encoding the 500kb reference maker step of the HiC pipeline.

Appendix P: HiC 500kb Blast Sorter Bash Script



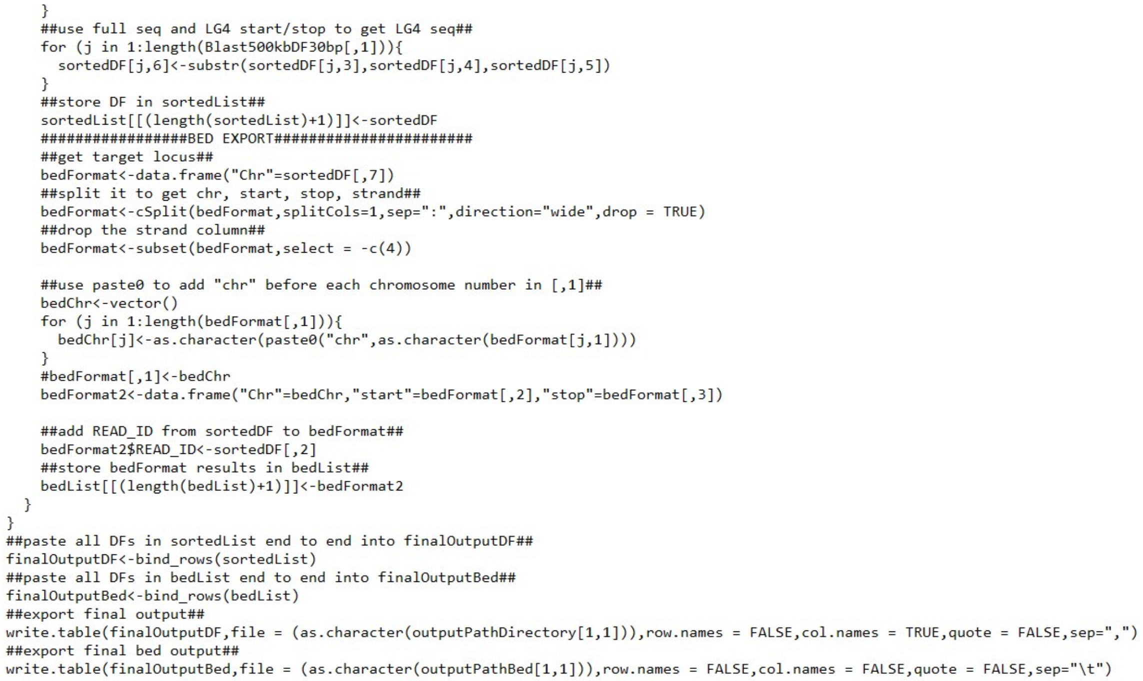
```
library(readxl)
library(dplyr)
library(tidyr)
library(stringr)
library(splitstackshape)
library(xlsx)

##build input file directory and upload all "500_kb_window.csv" pattern files##

##retrieve list of all output files##
input500kbFiles<-list.files(path="/home/usajdd/COLEY/SRA/output/HiCAnalysis/HiCAnalysisOutput", pattern = "500_kb_window.csv")
inputSeqSortedFiles<-list.files(path="/home/usajdd/COLEY/SRA/output/HiCAnalysis/HiCAnalysisOutput", pattern = "seq_sorted.csv")
inputReadFiles<-list.files(path="/home/usajdd/COLEY/SRA/output/HiCAnalysis/HiCAnalysisOutput", pattern = "_reads.txt")

##match output files based on SRA file of origin##
input500kbFiles2<-cSplit(as.data.frame(input500kbFiles),splitCols=1,sep="_",direction="wide")
inputSeqSortedFiles2<-cSplit(as.data.frame(inputSeqSortedFiles),splitCols=1,sep="_",direction="wide")
inputReadFiles2<-cSplit(as.data.frame(inputReadFiles),splitCols=1,sep="_",direction="wide")

fileDirectory<-data.frame("500kbBlastFiles"=input500kbFiles,"SeqSortedFiles"="","ReadFiles"="")
for (i in 1:length(input500kbFiles)){
  seqSortedInd<-which(inputSeqSortedFiles2[,1]==as.character(input500kbFiles2[i,1]))
  fileDirectory[i,2]<-inputSeqSortedFiles[seqSortedInd]
}
for (i in 1:length(input500kbFiles)){
  readInd<-which(inputReadFiles2[,1]==as.character(input500kbFiles2[i,1]))
  fileDirectory[i,3]<-inputReadFiles[readInd]
}

##use fileDirectory to make a new inputpathDirectory DF that has a full path to each input file, matched across rows##
inputPathDirectory<-data.frame(row.names = 1:(length(input500kbFiles)))
for (i in 1:3){
  inputPathDirectory[,i]<-str_c("/home/usajdd/COLEY/SRA/output/HiCAnalysis/HiCAnalysisOutput/",fileDirectory[,i])
}

##repeat for output path##
outputPathDirectory<-data.frame("LG4"=str_c("/home/usajdd/COLEY/SRA/output/HiCAnalysis/HiCAnalysisOutput/"
,input500kbFiles2[1,4],"_",input500kbFiles2[1,5],"_",input500kbFiles2[1,6],"_FINAL.csv"))

##bed output path##
outputPathBed<-data.frame("LG4"=str_c("/home/usajdd/COLEY/SRA/output/HiCAnalysis/HiCAnalysisOutput/"
,input500kbFiles2[1,4],"_",input500kbFiles2[1,5],"_",input500kbFiles2[1,6],"_FINAL.bed"))

##get SRA IDs##
SRAID<-input500kbFiles2[,1]

##Make output list##
sortedList<-list()
bedList<-list()
```

Figure A13: HiC 500kb Blast Sorter Bash Script. Bash script encoding the 500kb BLAST sorter step of the HiC pipeline.



```
bedList<-list()

##begin loop##
for (i in 1:length(input500kbFiles)){
  ##skip if filesize=0##
  if(file.size(inputPathDirectory[i,1])==0){
    next
  }
  ##upload 500kb blast output and apply 50 nt cutoff##
  Blast500kbDF<-read.csv(file=inputPathDirectory[i,1],header = FALSE,row.names = NULL)
  Blast500kbDF30bp<-filter(Blast500kbDF,V4>=50)
  ##checkpoint. If there are no >=50 nt alignments then all steps are skipped and there is no output##
  if (!is.na((Blast500kbDF30bp[1,1]))){
    ##make output df##
    sortedDF<-data.frame("SRA_ID"="","READ_ID"="","READ_Seq"="","LG4_Start"="","LG4_Stop"="","LG4_Seq"="","Target_Locus"=""
    ,"Target_Start"=Blast500kbDF30bp[,7],"Target_Stop"=Blast500kbDF30bp[,8],"Target_Seq"="")
    ##get SRA ID##
    sortedDF[,1]<-SRAID[i,]
    ##get target locus##
    targetLocus<-data.frame("raw"=str_extract(Blast500kbDF30bp[,2],pattern = "GRCh38:\\S+"))
    targetLocus[,1]<-gsub("GRCh38:","",targetLocus[,1])
    targetLocus<-cSplit(targetLocus,splitCols=1,sep=":",direction="wide",drop = TRUE)
    for (j in 1:length(Blast500kbDF30bp[,1])){
      targetLocus[j,2]<-targetLocus[j,2]+Blast500kbDF30bp[j,9]
      targetLocus[j,3]<-targetLocus[j,2]-Blast500kbDF30bp[j,9]+Blast500kbDF30bp[j,10]
    }
    targetLocus$cat<-""
    for (j in 1:nrow(targetLocus)){
      targetLocus[j,5]<-str_c(as.character(targetLocus[j,1]),":",as.character(targetLocus[j,2])
      ,":",as.character(targetLocus[j,3]),":",as.character(targetLocus[j,4]))
    }
    sortedDF[,7]<-targetLocus[,5]
    ##get target seq##
    sortedDF[,10]<-Blast500kbDF30bp[,13]
    ##get read ID,LG4 start,LG4 stop from "seq_sorted" file. Need to match Query number to SRR Read via indexing##
    seqSortedDF<-read.csv(file=inputPathDirectory[i,2],header = TRUE,row.names = NULL)
    for (j in 1:length(Blast500kbDF30bp[,1])){
      sortedDF[j,2]<-seqSortedDF[(str_extract(Blast500kbDF30bp[j,1],"\\d+")),1]
      sortedDF[j,4]<-seqSortedDF[(str_extract(Blast500kbDF30bp[j,1],"\\d+")),2]
      sortedDF[j,5]<-seqSortedDF[(str_extract(Blast500kbDF30bp[j,1],"\\d+")),3]
    }
    ##get full seq and LG4 seq from "_reads" file##
    ReadsDF<-read.delim(file=inputPathDirectory[i,3], header=FALSE, comment.char="#")
    for (j in 1:length(Blast500kbDF30bp[,1])){
      sortedDF[j,3]<-ReadsDF[(str_extract(Blast500kbDF30bp[j,1],"\\d+")),2]
    }
```

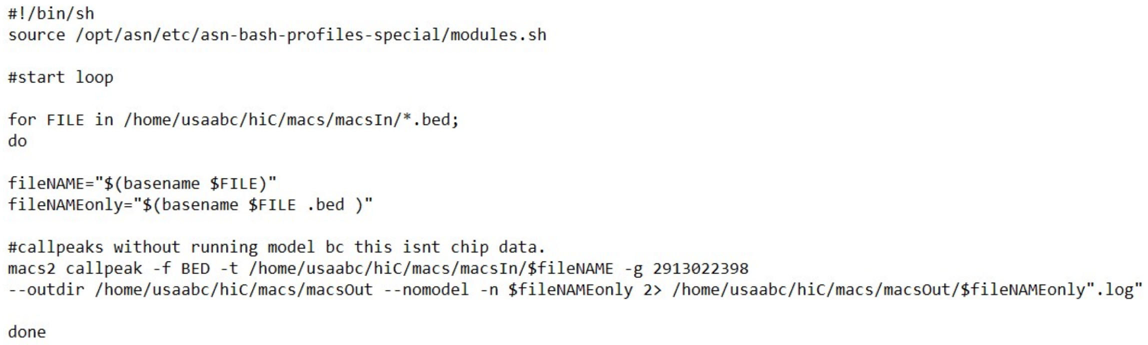
Figure A13, cont.



```
    }
    ##use full seq and LG4 start/stop to get LG4 seq##
    for (j in 1:length(Blast500kbDF30bp[,1])){
      sortedDF[j,6]<-substr(sortedDF[j,3],sortedDF[j,4],sortedDF[j,5])
    }
    ##store DF in sortedList##
    sortedList[[(length(sortedList)+1)]]<-sortedDF
    #################BED EXPORT#######################
    ##get target locus##
    bedFormat<-data.frame("Chr"=sortedDF[,7])
    ##split it to get chr, start, stop, strand##
    bedFormat<-cSplit(bedFormat,splitCols=1,sep=":",direction="wide",drop = TRUE)
    ##drop the strand column##
    bedFormat<-subset(bedFormat,select = -c(4))

    ##use paste0 to add "chr" before each chromosome number in [,1]##
    bedChr<-vector()
    for (j in 1:length(bedFormat[,1])){
      bedChr[j]<-as.character(paste0("chr",as.character(bedFormat[j,1])))
    }
    #bedFormat[,1]<-bedChr
    bedFormat2<-data.frame("Chr"=bedChr,"start"=bedFormat[,2],"stop"=bedFormat[,3])

    ##add READ_ID from sortedDF to bedFormat##
    bedFormat2$READ_ID<-sortedDF[,2]
    ##store bedFormat results in bedList##
    bedList[[(length(bedList)+1)]]<-bedFormat2
  }
}
##paste all DFs in sortedList end to end into finalOutputDF##
finalOutputDF<-bind_rows(sortedList)
##paste all DFs in bedList end to end into finalOutputBed##
finalOutputBed<-bind_rows(bedList)
##export final output##
write.table(finalOutputDF,file = (as.character(outputPathDirectory[1,1])),row.names = FALSE,col.names = TRUE,quote = FALSE,sep=",")
##export final bed output##
write.table(finalOutputBed,file = (as.character(outputPathBed[1,1])),row.names = FALSE,col.names = FALSE,quote = FALSE,sep="\t")
```

Figure A13, cont.

Appendix Q: MACS Peak Caller Bash Script



```
#!/bin/sh
source /opt/asn/etc/asn-bash-profiles-special/modules.sh

#start loop

for FILE in /home/usaabc/hiC/macs/macsIn/*.bed;
do

fileNAME="$(basename $FILE)"
fileNAMEonly="$(basename $FILE .bed )"

#callpeaks without running model bc this isnt chip data.
macs2 callpeak -f BED -t /home/usaabc/hiC/macs/macsIn/$fileNAME -g 2913022398
--outdir /home/usaabc/hiC/macs/macsOut --nomodel -n $fileNAMEonly 2> /home/usaabc/hiC/macs/macsOut/$fileNAMEonly".log"

done
```

Figure A14: MACS Peak Caller Bash Script. Bash script encoding the MACS peak calling step of the HiC pipeline.

Appendix R: TumorFusions Database Tier 1 Fusions Within the Chr 5 LG4 TAD

Table A4: TumorFusions Database Tier 1 Fusions Within the Chr 5 LG4 TAD.

Gene	Unique TCGA Tier 1 Fusions
PLEKHG4B	
LRRC14B	
CCDC127	SDHA (TCGA.S9.A6UA.01A), SDHA (TCGA.96.7544.01A)
SDHA	CCDC127 (TCGA.S9.A6UA.01A), CCDC127 (TCGA.96.7544.01A)
PCDC6	
AHRR	
EXOC3	CEP72 (TCGA.V5.AASW.01A)
SLC9A3	
CEP72	EXOC3 (TCGA.V5.AASW.01A), BRD9 (TCGA.90.6837.01A)
TPPP	
ZDHHC11B	
ZDHHC11	
BRD9	CEP72 (TCGA.90.6837.01A)
TRIP13	
NKD2	
SLC12A7	

BIOGRAPHICAL SKETCH

Name of Author: Alexander Coley

Graduate and Undergraduate School Attended:

University of South Alabama, Mobile, Alabama

Degrees Awarded:

Bachelor of Science in Biomedical Sciences, 2018, Mobile, Alabama

Doctor of Philosophy in Basic Medical Sciences, 2020, Mobile, Alabama

Publications:

Coley, A.B.; Stahly, A.N.; Kasukurthi, M.V.; Barchie, A.A.; Hutcheson, S.B.; Houserova, D.; Huang, Y.; Watters, B.C.; King, V.M.; Dean, M.A.; Roberts, J.T.; DeMeis, J.D.; Amin, K.V.; McInnis, C.H.; Godang, N.L.; Wright, R.M.; Haider, D.F.; Piracha, N.B.; Brown, C.L.; Ijaz, Z.M.; Li, S.; Xi, Y.; McDonald, O.G.; Huang, J.; Borchert, G.M. MicroRNA-like snoRNA-Derived RNAs (sdRNAs) Promote Castration-Resistant Prostate Cancer. *Cells* 2022, 11, 1302. https://doi.org/10.3390/cells11081302

Coley, A.B.; Ward, A.; Keeton, A.B; Chen, X.; Maxuitenko, Y.; Prakash, A.; Li, F.; Foote, J.B.; Buchsbaum, D.J.; Piazza, G.A. Pan-RAS inhibitors: Hitting multiple RAS isozymes with one stone. *Advances in Cancer Research* 2022, 153, 131-168. https://doi.org/10.1016/bs.acr.2021.07.009

Ward, A.B; Keeton, A.B; Chen, X.; Mattox, T.E.; **Coley, A.B.**; Maxuitenko, Y.Y.; Buchsbaum, D.J.; Randall, T.D.; Zhou, G.; Piazza, G.A. Enhancing anticancer activity of checkpoint immunotherapy by targeting RAS. *MedComm* 2020, Sep;1(2), 121-128. https://doi.org 10.1002/mco2.10.

Piazza, G.A.; Ward, A.; Chen, X.; Maxuitenko, Y.; **Coley, A.B.**; Aboelella, N.S.; Buchsbaum, D.J.; Boyd, M.R.; Keeton, A.B; Zhou, G. PDE5 and PDE10 inhibition activates cGMP/PKG signaling to block Wnt/β-catenin transcription, cancer cell growth, and tumor immunity. *Drug Discovery Today* 2020, 25(8), 1521-1527. https://doi.org/10.1016/j.drudis.2020.06.008

Patterson, D.G.; Roberts, J.T.; King, V.M.; Houserova, D.; Barnhill, E.C.; Crucello, A.; Polska, C.J.; **Coley, A.B.**; Zeidan, M.; Brantley, L.W.; Kaufman, G.C.; Nguyen, M.; Santana, M.W.; Schiller, I.A.; Spicciani, J.S.; Zapata, A.K.; Miller, M.M.; Sherman, T.D.; Ma, R.; Zhao, H.; Arora, R.; Tan, M.; Xi, Y.;

Borchert, G.M. Human snoRNA-93 is processed into a microRNA-like RNA that promotes breast cancer cell invasion. *npj Breast Cancer* 2017, 3, 25. https://doi.org/10.1038/s41523-017-0032-8

Amin, S.V.; Roberts, J.T.; Patterson, D.G.; **Coley, A.B.**; Allred, J.A.; Denner, J.M.; Johnson, J.P.; Mullen, G.E.; O'Neal, T.K.; Smith, J.T. Cardin, S.E.; Carr, H.T.; Carr, S.L.; Cowart, H.E.; DaCosta, D.H.; Herring, B.R.; King, V.M.; Polska, C.J.; Ward, E.E.; Wise, A.A.; McAllister, K.N.; Chevalier, D.; Spector, M.P,; Borchert, G.M. Novel small RNA (sRNA) landscape of the starvation-stress response transcriptome of Salmonella enterica serovar typhimurium. RNA Biology 2016, 13:3, 331-342. https://doi.org/10.1080/15476286.2016.1144010