Investigations into the roles of R-loops in androgen signalling and prostate cancer

A thesis submitted to the University of Sheffield for the degree of
Doctor of Philosophy

By

Jonathan Luke Griffin MBChB (Hons) MRCS FRCPath

Department of Biosciences

October 2022

Declaration

I, the author, confirm that the Thesis is my own work. I am aware of the University's Guidance on the Use of Unfair Means (www.sheffield.ac.uk/ssid/unfair-means). This work has not previously been presented for an award at this, or any other, university.

Acknowledgments

I must first thank my supervisors Prof Sherif El-Khamisy and Prof Jim Catto. Together you have taught me how to be a researcher and provided me with patience, mentorship and opportunities. I am incredibly grateful for the time and energy you have dedicated to supervising me these last four years. I must also thank the members of the El-Khamisy lab past and present. I couldn’t have hoped for a better group with whom to share this journey. Thank you for keeping things fun despite my moaning about multi-factor designs and linear dynamic ranges.

Thank you also to everyone in the histopathology department at the Royal Hallamshire Hospital in Sheffield. In particular I must thank Dr Jon Bury, Dr Yota Kitsanta, Dr Katie MacDonald and Dr Malee Fernando for supporting and encouraging me to pursue a clinical academic career.

I was fortunate enough to supervise two master’s students, Lorna Gilroy-Turner and Abbie Fisher, and a Lister summer student Clara Chen during my work on this thesis. Data generated by Lorna and Clara under my supervision appears in this thesis in figures 4.17 and 6.4 respectively. I was also fortunate to collaborate with Professor Ester Hammond from the University of Oxford. Prof Hammond provided the RNA-seq data that I analysed with RloopTools in figures 3.18 and 3.19. Dr Hannah Crane from the El-Khamisy lab provided the RNA-seq data used in Figure 3.21.

Finally, and most importantly, to my wife Eleanore and our three amazing children Rosie, Bash and Ophelia. Without you none of this means anything. Thank you for your constant support, encouragement and love.

Publications arising from this thesis

Griffin J, Chen Y, Catto JWF, El-Khamisy S. Gene of the month: NKX3.1. J Clin Pathol. 2022 Jun;75(6):361-364. doi: 10.1136/jclinpath-2021-208073. Epub 2022 Jan 7. PMID: 34996754.

Ramachandran S, Ma TS, **Griffin J**, Ng N, Foskolou IP, Hwang MS, Victori P, Cheng WC, Buffa FM, Leszczynska KB, El-Khamisy SF, Gromak N, Hammond EM. Hypoxia-induced SETX links replication stress with the unfolded protein response. Nat Commun. 2021 Jun 17;12(1):3686. doi: 10.1038/s41467-021-24066-z. PMID: 34140498.

Funding that supported this thesis

Jean Shanks Foundation/ The Pathological Society of Great Britain and Ireland Clinical PhD Research Training Fellowship (CPHD-2018-01)

The Pathological Society of Great Britain and Ireland Equipment Grant for purchase of a RotorGene qPCR machine (EG-1019-02)

The Pathological Society of Great Britain and Ireland Equipment Grant for purchase of an Eppendorf refrigerated centrifuge (EGS-0921-01)

The more I see, the more I know. The more I know, the less I understand

-The Changingman by Paul Weller

Abstract

Background

Prostate cancer is the 2nd commonest cause of male cancer death in the UK. Growth of both normal and neoplastic prostate requires androgen signalling. However, androgen signalling mechanisms have not been fully elucidated. Recently, R-loops (three stranded, co-transcriptional DNA:RNA hybrids) have been implicated in DNA damage and regulation of gene expression. The aim of this project was to evaluate what role R-loops play in androgen regulated gene expression.

Methods

Prostate cancer cells were stimulated with androgens and the resulting changes in R-loop dynamics were assessed by 1) Predictive modelling based on nucleotide sequence contexts favouring R-loop formation and 2) DNA:RNA immunoprecipitation with next generation sequencing (DRIP-seq). The functional roles of R-loops were probed by overexpression of RNAse H1, a well characterised R-loop resolvase. R-loop interacting proteins were identified by DNA:RNA immunoprecipitation and Western blotting. The functions of these proteins were examined by siRNA knockdown experiments, DNA damage response assays, analysis of gene expression and chromatin immunoprecipitation.

Results

Predicted R-loop forming sequences were higher in genes downregulated in metastatic castration resistant prostate cancer. In LNCaP cells stimulated with androgens there was no correlation between gene expression and R-loop formation. R-loop dynamics did not change at androgen receptor bound enhancers. RNA-seq in cells overexpressing RNase H1 and treated with DHT showed unchanged canonical androgen signalling but did show potential off-target effects of RNase H1 overexpression. Two androgen responsive genes – NKX3.1 and KLK3 – accumulated R-loops. Only NKX3.1 accumulated R-loops with androgen treatment. Knock down of the DEAD box helicases DDX5 and DDX17 affected NKX3.1 and KLK3 gene expression and dysregulated global R-loop levels. DDX17 was recruited to the NKX3.1 R-loop but depletion of DDX5/DDX17 did not have locus-specific R-loop effects.

Conclusions

R-loops do not have general roles in androgen regulated gene expression but they may play a role in regulating specific loci.

List of figures

Chapter 4

Chapter 6

List of tables

List of appendices

Abbreviations

ANOVA: Analysis of Variance

ADT: Androgen Deprivation Therapy

AR: Androgen Receptor

BAM: Binary Alignment Map

BED: Browser Extensible Data

CRPC: Castration resistant prostate cancer

ChIP: Chromatin immunoprecipitation

DAPI: 4,6-diamidino-2-phenylindole

DNA: Deoxyribonucleic Acid

DHT: Dihydrotestosterone

DRIP: DNA:RNA hybrid Immunoprecipitation

DSB: Double strand DNA break

EGTA: Ethylene glycol-bis(β-aminoethyl ether)-N,N,N′,N′-tetraacetic acid

EDTA: Ethylenediaminetetraacetic acid

FDR: False discovery rate

GFP: Green Fluorescent Protein

PFA: Paraformaldehyde

RNA pol 2: RNA polymerase 2

PAGE: Polyacrylamide gel electrophoresis

PCR: Polymerase Chain Reaction

RNA: Ribonucleic Acid

RNH1: RNase H1

SAM: Sequence Alignment Map

SSB: Single strand DNA break

SDS: Sodium dodecyl sulphate

SPA: Supra-physiological Androgen

TEMED: Tetramethylethylenediamine

TE: Tris-EDTA pH8

Table of contents

Chapter 1: Introduction

1.1 Prostate cancer

1.1.1 Diagnosis and treatment of prostate cancer

Prostate cancer causes approximately 11,000 deaths annually in the United Kingdom making it the 2nd commonest cause of cancer death in men. Despite this, three quarters of men diagnosed with prostate cancer will survive at least 10 years after diagnosis (CRUK, 2021). In addition, 25-60% of men over 60 have clinically undetected or latent prostate cancer in autopsy studies (Kimura *et al.*, 2021). Central to understanding this wide spectrum of disease, from undetectable to lethal, is the clinical, pathological and molecular classification of prostate cancer. Upon presenting with suspected prostate cancer a man will undergo clinical, radiological and biopsy/histopathological assessment (Mottet *et al.*, 2021). This approach is directed at determining how far the tumour has spread (tumour stage; summarised in **table 1.1**) and how aggressive the tumour is (tumour grade). Like other cancers, tumour grading of prostate cancer is performed by histological assessment of tissue sampled by biopsy and/or surgical resection. Based on the degree of differentiation seen at histopathological examination, a Gleason score is assigned to tumour. The Gleason system has been in use since first developed by Donald Gleason in 1966 (Gleason, 1966). Contemporary use of the Gleason score involves a pathologist assigning two grades to prostate cancer material using architectural features seen at low to medium power microscopic examination. For historical reasons the individual grade range spans 3 – 5 and the two scores are expressed as a sum (e.g. 3+4=7) with higher scores indicating more aggressive cancer. The two grades assigned by the pathologist represent the two most abundant grades present in a tumour. In this way 3+4=7 and 4+3=7 are not equivalent - the second score indicates a higher proportion of grade 4 cancer and this correlated with higher cancer stage and earlier recurrence after radical treatment in a large retrospective series (Burdick *et al.*, 2009). Gleason grading can be confusing for clinicians and patients and was therefore recently updated to classify patients into one of five grade groups (Egevad *et al.*, 2016). In this approach Gleason grades are grouped together to reflect their ability to predict recurrence and overall survival. This approach is now used as part of the UK NICE recommendations to guide management decisions (National Institute for Health and Care Excellence [NICE], 2019).

Table 1.1 Staging of prostate cancer. DRE: Digital rectal examination

Tumour stage	**Description**
TUMOUR	
T1 – Clinically undetectable tumour (e.g. not palpable by DRE)	
T1a	Incidental tumour in < 5% of submitted tissue
T1b	Incidental tumour in > 5% of submitted tissue
T1c	Needle biopsy detected tumour
T2 – Palpable tumour confined to prostate	
T2a	Tumour involves up to 50% of one prostate lobe
T2b	Tumour involves >50% of one prostate lobe
T2c	Tumour involves both prostate lobes
T3 – Tumour invades through the prostatic capsule	
T3a	Tumour extends through prostatic capsule
T3b	Tumour invades seminal vesicles
T4 – Tumour invades adjacent structures not including seminal vesicles	
T4	e.g. invades bladder, pelvic side wall, external sphincter
LYMPH NODES	
N0	No involvement of regional lymph nodes
N1	Tumour spread to regional lymph nodes
METASTASIS	
M0	No metastasis (excluding regional lymph nodes)
M1a	Metastasis to non-regional lymph nodes
M1b	Metastasis to bone
M1c	Metastasis to other sites (e.g. visceral)

The decision making around treatment for prostate cancer is complex. Treatment decisions are informed by a multitude of factors including: The stage and grade of the cancer, a patient's physical fitness, presence of comorbidities, life expectancy, patient wishes and consideration of the potential harms of over-treatment (National Institute for Health and Care Excellence [NICE], 2019; Mottet *et al.*, 2021). For example, an incidentally detected low stage low grade cancer in a patient with multiple comorbidities is unlikely to cause the patient significant harm and one of their comorbidities may be more significant in terms of their health span and eventual cause of death. Conversely, a higher stage, higher grade tumour in a medically fit man would be more likely to cause long term morbidity and mortality, and could be managed more aggressively by either surgical removal of the prostate or radiotherapy. In between these two extremes is the approach of active surveillance where a low-risk cancer is serially monitored radiologically to detect and treat progression early in the disease course.

Approximately 20% of patients who present with localised prostate cancer progress to metastatic prostate cancer with metastases commonly involving local lymph nodes, bone, distant lymph nodes and the liver (Gandaglia *et al.*, 2014; Xie *et al.*, 2017). In addition, 10% of advanced prostate cancer cases are metastatic at index presentation. Minimising and controlling prostate cancer spread is achieved through androgen deprivation therapy (ADT). Pioneering work by Charles Huggins established that castration resulted in a marked decrease in prostate size, radiological resolution of metastases and symptomatic improvement (Huggins, Stevens and Hodges, 1941). This was the first study to link androgen signalling and prostate cancer growth. Since then, chemical castration using luteinizing hormone-releasing hormone agonists has replaced physical castration. Chemical castration causes a reduction in serum testosterone by blocking signalling from the hypothalamus to the pituitary to inhibit the release of luteinising hormone. ADT is effective at reducing tumour burden and secondary complications of metastatic cancer (Crawford *et al.*, 2019; Wang, Lee and Armstrong, 2022). However, almost all advanced prostate cancer treated with ADT eventually becomes castration resistant (castration resistant prostate cancer; CRPC). Treatment options for this stage of prostate cancer include androgen receptor inhibitors such as Enzalutamide, which was shown to improve survival in men with CRPC after completing conventional chemotherapy in the AFFIRM trial (Scher *et al.*, 2012).

Despite this, CRPC is incurable. In the PREVAIL trial patients treated with enzalutamide had a higher median survival (36 months vs. 31 months) but 80% of patients had died by five year's follow up (Armstrong *et al.*, 2020). Treatment resistance is therefore a major unsolved problem in prostate cancer treatment. Prostate cancer growth, treatment and treatment resistance all depend on the androgen receptor and androgen signalling (Fujita and Nonomura, 2019) so understanding androgen signalling pathways is central to improving outcomes in prostate cancer.

1.1.2 Androgen signalling in prostate development and prostate cancer

Development of the prostate gland begins in the embryo at 9-10 weeks post gestation. Primitive prostatic epithelium arises from the embryonic urogenital sinus and eventually forms acini and ducts lined by a characteristic double layer of inner luminal cells surrounded by basal cells (Cunha *et al.*, 2018). Prostate development is driven by androgens, principally testosterone, secreted from the Leydig cells of the testis. Testosterone is converted to its active metabolite dihydrotestosterone in androgen responsive tissues including the prostate. As a cholesterol derived molecule, testosterone diffuses freely across the cell membrane where it binds to the androgen receptor (AR). Testosterone binding promotes dissociation of AR from the chaperone heat shock protein 90 (HSP90) and translocation of AR to the nucleus. Here, the androgen receptor dimerises and binds specific DNA sequences termed androgen response elements (**figure 1.1**). In addition to DNA binding, a topoisomerase 2 mediated physiological transient double stranded DNA break is required for efficient androgen receptor activity (Ju *et al.*, 2006). Chromatin immunoprecipitation coupled with high throughput sequencing (ChIP-chip and ChIP-seq) studies revealed that most androgen receptor binding occurs in distal gene regions. These regions are often marked by bidirectional short RNA production and the H3K27 acetylation histone modification indicating that AR binds enhancers (Toropainen *et al.*, 2016). The current model of androgen receptor transcriptional activation suggests chromosomal looping of AR bound enhancers to promoters of androgen responsive genes to initiation transcription (Wu *et al.*, 2011). Despite these characteristics, there are approximately 10 times more AR binding sites than genes regulated by androgen signalling. Using conservative criteria of what constitutes an AR binding site and a massively parallel enhancer characterisation technique called STARR-seq, Huang *et al* (2021) classified AR bound enhancers into inactive, constitutive and androgen induced based on their ability to induce transcription. Inducible AR bound enhancers were closer to their target genes, formed more chromatin loops and had a greater effect on transcriptional output as measured by target gene expression down regulation after CRISPR interference of the enhancer region.

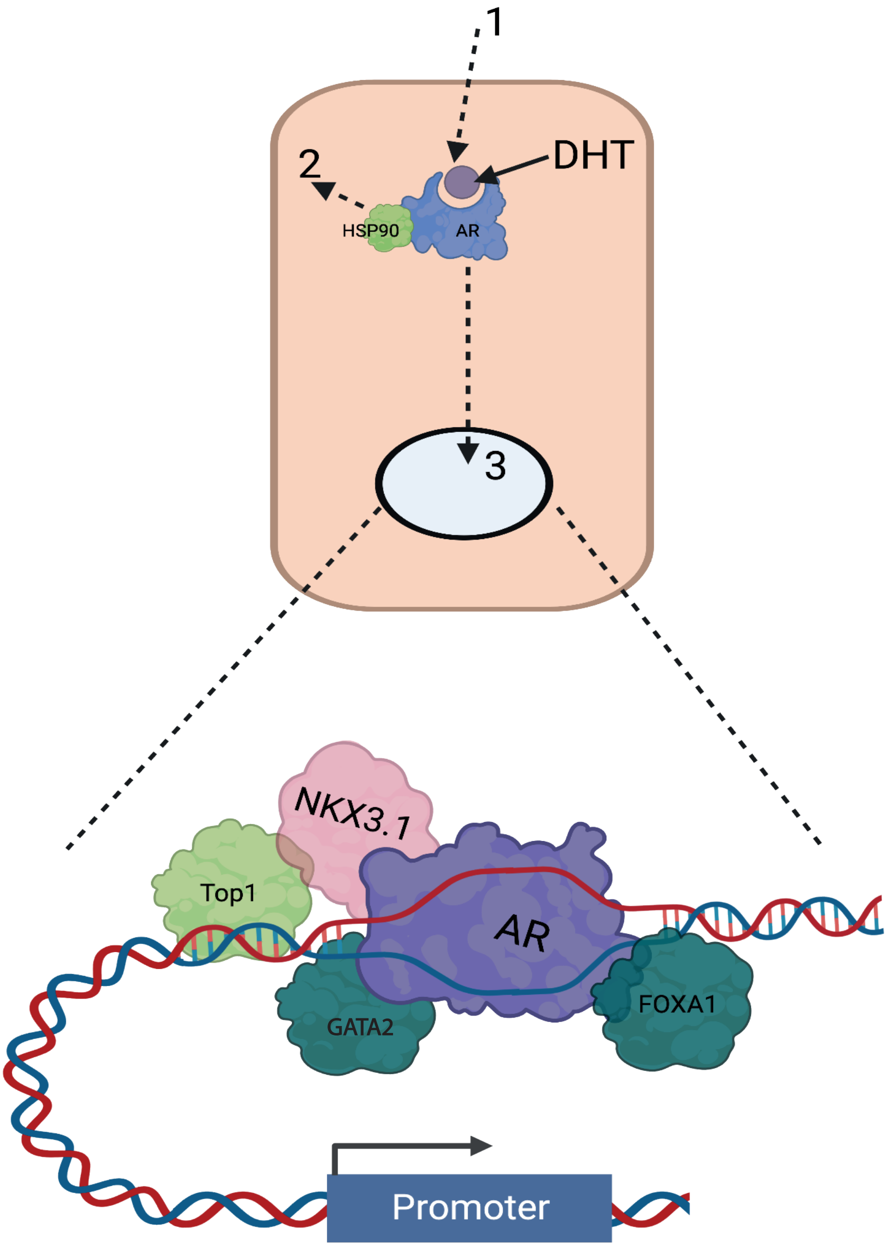


Figure 1.1 Androgen receptor signalling
1: Dihydrotestosterone (DHT) diffuses freely across the cell membrane and binds to the androgen receptor (AR). 2: HSP90 dissociates from the AR. 3: The AR translocates to the nucleus to regulate transcription of androgen responsive genes through chromatin looping beween enhancers and promoters. AR binding sites are pre-bound by pioneer factors including FOXA1, GATA2, Topoisomerase 1 (Top1) and NKX3.1 which mediate chromatin accessibility and AR activity.

To exert its function, the androgen receptor requires additional pioneer factors. These proteins associate with and decondense chromatin to create a more open chromatin environment and improve the chance of transcription factor binding. Four important AR-associated pioneer factors have emerged: GATA2, HOXB13, FOXA1 and NKX3-1 (Tan *et al.*, 2012; Hankey, Chen and Wang, 2020). GATA2, HOXB13 and FOXA1 all associate with and modify chromatin to create a favourable state for AR binding and activity (Hankey, Chen and Wang, 2020). Furthermore, GATA2 maintains chromatin looping between some enhancers and their target gene prior to androgen receptor binding (Wu *et al.*, 2014). These three factors have increased expression with progression to CRPC and this most likely represents an adaptation to androgen deprivation therapy whereby androgen regulated transcription can continue despite low androgen levels. FOXA1 is also commonly mutated in prostate cancer with mutations clustering around the forkhead domain observed in 4% of the prostate cancer TCGA cohort which examined 333 primary prostate cancer cases (Veluvolu *et al.*, 2015). A similar FOXA1 mutation rate was seen in castration resistant prostate cancer (Grasso *et al.*, 2012). In the TCGA cohort a higher level of androgen receptor signalling was observed in the FOXA1 mutated cases.

NKX3-1 has also been characterised as a pioneer factor however its role at enhancers differs from GATA2, HOXB13 and FOXA1. Like these proteins, NKX3-1 is resident at promoters before androgen stimulation of LNCaP cells. Upon androgen receptor binding, NKX3-1 interacts with topoisomerase 1, an enzyme responsible for making physiological single stranded DNA breaks to relieve the torsional stress of transcription (Puc *et al.*, 2015). At AR bound enhancers this single strand break is necessary for production of eRNA and activation of target androgen regulated genes. Whether this is a genome wide phenomenon and whether the SSB is necessary for all eRNA production was not explored in the study as specific loci were used to model androgen signaling.

1.1.3 Mechanisms of resistance to androgen deprivation therapy

Prostate cancer adapts to maintain growth and cell survival under the low androgen conditions of androgen deprivation therapy. Importantly, under castrate condition, androgen receptor signalling is still the primary mechanism by which prostate cancer continues to survive. The mechanisms underlying this adaptation include upregulation of pioneer factors and changes to the androgen receptor at genomic and transcriptomic levels. At the genomic level, the androgen receptor can undergo point mutations. Nearly half of these increase androgen receptor activity either through allowing promiscuous binding of other steroids such as glucocorticoids or by rendering the receptor constitutively active (Shi *et al.*, 2002). Other mutations increase the interaction with pioneer factors or decrease the binding of inhibitory factors. Alternatively, the androgen receptor locus can undergo copy number amplification, thereby increasing the use of scarce androgen through increased AR transcript and protein levels. In a large cohort of CRPC patients AR amplification was seen in 70% of cases (Quigley *et al.*, 2018). Interestingly the most commonly amplified region was an androgen receptor enhancer and patients with this amplification had higher levels of androgen receptor mRNA. This enhancer amplification is therefore the commonest reason for transcriptomic upregulation of the androgen receptor. The androgen receptor can also undergo alternative splicing. The most common splice variant is AR-V7, formed by transcription of a cryptic exon 3. This variant has transcriptional activity independent of ligand binding and activates a programme of upregulated DNA damage repair associated with resistance to ionising radiation (Kounatidou *et al.*, 2019).

1.1.4 Androgen signalling and DNA damage

Damage from endogenous and exogenous sources is a constant threat to the integrity of DNA. It is estimated that a cell undergoes 10,000 single strand breaks per day due to endogenous processes such as generation of reaction oxygen species, topoisomerase activity during transcription and chemical modification of DNA bases. Left unrepaired, these breaks can be converted to double strand breaks leading to mutational events including point mutations, insertion and deletions of genomic material and large chromosomal rearrangements/ translocations. In addition to endogenous sources of DNA damage, there are multiple exogenous threats to DNA including pyrimidine dimers formed by UV light exposure, aromatic DNA adducts from cigarette smoke and double strand breaks from ionising radiation (Ciccia and Elledge, 2010). Recently DNA alkylation by certain species of *E.Coli* in the colonic microbiome has also been described (Pleguezuelos-Manzano *et al.*, 2020).

As described earlier, physiological androgen signalling is dependent on programmed DNA double strand breaks. This has been proposed as the reason for an upregulation of DNA damage repair pathways upon androgen stimulation (Polkinghorn *et al.*, 2013). In this study, primary prostate cancer cases with high androgen signalling also had increased expression of DNA damage repair genes and prostate cancer cells treated with androgens exhibited resistance to ionising radiation. Specifically, the classical non-homologous end joining (NHEJ) pathway was upregulated by androgens and downregulated by androgen deprivation. It is factors from the NHEJ pathway that are recruited to androgen receptor induced double strand breaks (Ju *et al.*, 2006). This pathway allows the repair of DNA damage throughout the cell cycle (*c.f.* homologous recombination which requires DNA replication to have taken place). However, NHEJ is error prone and can result in loss of genetic material. A further study demonstrated a reciprocal relationship between androgen receptor signalling and DNA damage repair (Goodwin *et al.*, 2013). In this work the authors observed an increase in transcription of the canonical androgen regulated genes TMPRSS2 and FKBP5 in response to ionising radiation in the absence of androgen stimulation. Furthermore, androgen deprivation downregulated the activity of DNAPKcs, a key component of the NHEJ pathway. Androgen stimulation rescued this downregulation. The downregulation of DNA damage repair by ADT is synergistic with ionising radiation and this is the proposed mechanism of action of this common combined treatment strategy in prostate cancer.

Paradoxically, increased androgen signalling has also been associated with increased DNA damage. An early study described how the common TMPRSS2-ERG fusion is an androgen induced genomic rearrangement (Haffner *et al.*, 2010). In a study of enzalutamide-resistant patient derived xenografts, treatment with supra physiological testosterone reduced tumour volume and improved survival compared to placebo. DNA damage repair pathways were downregulated in this context and both homologous recombination (HR) and NHEJ pathways showed downregulation at the transcriptional level (Lam *et al.*, 2019). A second study showed that androgen receptor amplification increased the amount of DNA damage induced by supraphysiological testosterone. Again, both HR and NHEJ pathways were downregulated. Translating these findings to data from an ongoing clinical trial the authors found that patients with HR deficient tumours had a better response to supraphysiological androgen therapy. In cell culture models deficient in BRCA2, supraphysiological androgen was synergistic with PARP inhibition (Chatterjee *et al.*, 2019). A further study showed that androgen re-supplementation after maximal treatment for CRPC appears to have clinical benefit. In a phase 2 study of 30 patients who had already received enzalutamide for CRPC, treatment with androgens reduced blood plasma levels of prostate specific antigen, a marker of disease burden in prostate cancer (Teply *et al.*, 2018). To date, no study has reported progression-free or overall survival rates for supraphysiological androgen therapy in the setting of CRPC.

Together, these studies describe a complex relationship between androgen signalling and DNA damage. The increased DNA damage seen with supra physiological testosterone is dependent on androgen receptor amplification, a common occurrence in CRPC. It could be that enhanced androgen receptor binding causes an overwhelming number of 'programmed' DNA double strand breaks. However, this does not explain the downregulation of DNA damage repair pathways by supraphysiological androgen signalling and this mechanism remains to be explored. Conversely, ADT is a well-established treatment for advanced prostate cancer and has the effect of down regulating DNA damage repair. The paradoxically similar responses of androgen deprivation and androgen supplementation are most likely due to differences in the underlying biological context at the point treatment is initiated.

1.2 R-loops

R-loops are three-stranded nucleic acid structures composed of a DNA:RNA hybrid and a looped out DNA strand (Thomas, White and Davis, 1976). R-loops form co-transcriptionally when nascently transcribed RNA exits the RNA polymerase complex then re-anneals to its complementary DNA template. These nucleic acid hybrids are found throughout all kingdoms of life and play physiological roles in plants (Xu *et al.*, 2021), yeast (Cornelio *et al.*, 2017), bacteria (Lin *et al.*, 2010), viruses (Wongsurawat *et al.*, 2020) and mammals (reviewed in Crossley, Bocek and Cimprich, 2019). Their accumulation, resolution and regulation are important in physiological processes and their perturbation has important pathological consequences.

1.2.1 Factors favouring the formation of R-loops

R-loops are separate from the short (10-15 bp) DNA:RNA hybrid formed in the transcription bubble and may extend for thousands of kilobases behind the elongating RNA polymerase (Malig *et al.*, 2020). The distribution R-loops across the genome is non-random. These structures tend to accumulate in gene promoter (Ginno *et al.*, 2013a) and transcription termination regions (Skourti-Stathaki, Kamieniarz-Gdula and Proudfoot, 2014) as well as across the rRNA loci (Abraham *et al.*, 2020). These regions share common DNA sequence characteristics, most notably GC-skew. Regions of high GC skew contain a higher proportion of guanine nucleotides compared to cytosine in the coding/non-template DNA strand at promoters and termination regions (Ginno *et al.*, 2013b). *In vitro* transcription experiments showed that, beyond GC-skew, guanine clustering was also required for efficient R-loop formation (Roy and Lieber, 2009). Once formed, R-loops are stable structures owing to the higher energy required to dissociate DNA:RNA hybrids compared to DNA duplexes (Lesnik and Freier, 1995). In addition, guanine and cytosine associate via three hydrogen bonds compared to two hydrogen bonds in AT base pairs.

In addition to sequence context, two further genomic features contribute to R-loop formation. The first is negative super helicity. This unwinding of DNA supercoils behind the RNA polymerase increases DNA accessibility and thus increases the likelihood of complementary RNA finding its DNA target (Stolz *et al.*, 2019). The final feature associated with R-loop formation is RNA dwell time. This is defined as the time that an RNA sequence is in proximity to its complementary template. Like negative super helicity, increased dwell time increases the likelihood of the RNA re-annealing and forming an R-loop. Dwell time can be increased by RNA polymerase 2 stalling. In neuroblastoma MYCN was shown to interact with BRCA1 and regulate RNA polymerase 2 elongation. In the absence of BRCA1 RNA polymerase 2 accumulated at the 5' end of genes and this was accompanied by an increase in DRIP-seq signal indicative of R-loop accumulation (Herold *et al.*, 2019). A similar mechanism was seen in BRCA1 deficient breast cancer where BRCA1 promoted productive transcription and its absence was associated with RNA polymerase 2 stalling, R-loop formation and DNA damage (Zhang *et al.*, 2017). In addition to stalling of RNA polymerase 2, defects in RNA splicing have also been linked to R-loop formation. Bonnet *et al.* (2017) proposed that introns provide a layer of protection against R-loop formation. By inserting an intron into an otherwise intronless gene in yeast they decreased R-loop formation as also demonstrated that human intronless genes were more R-loop prone than genes containing R-loops. The removal of introns by co-transcriptional splicing renders the processed RNA non-complementary to its DNA template which decreases the likelihood of re-annealing. In support of this, knock down of splicing factors has been used to generate R-loops experimentally and cells with spliceosome mutations also accumulate more R-loops (Nguyen *et al.*, 2018; Goulielmaki *et al.*, 2021; Saha *et al.*, 2022).

1.2.2 Physiological roles of R-loops

R-loop preferentially form in important gene regulatory regions such as gene promoter and transcription termination regions. On this basis the roles of R-loops in the physiological regulation of epigenetic processes has been extensively investigated. R-loops have been implicated in CpG methylation, histone modifications and transcriptional dynamics.

1.2.2.1 R-loops and CpG methylation

The formation and resolution of R-loops must have an associated energetic cost to the cell as specific proteins are deployed for R-loop removal. Therefore, this cost must be beneficial or at least not detrimental in evolutionary terms for genes to have evolved characteristics that promote R-loop formation. In mapping R-loop forming sequences across the genome, Kuznetsov *et al* (2018) observed that only 28% of pseudogenes contain at least one DNA sequence highly likely to form R-loops compared to 79% of protein coding genes. This suggests that R-loops may play a role in the transcription of functional, protein coding genes. In line with this R-loops were found to map to the promoter regions of genes in genome-wide R-loop mapping studies. One feature of promoter regions is the presence of CpG islands where stretches of CpG dinucleotides are maintained in a demethylated state. Methylation of these regions is associated with chromatin condensation and gene silencing therefore demethylated CpG islands favour transcription of the associated genes. The presence of GC skew and R-loop formation was correlated with unmethylated CpG islands and mechanistic studies suggested that R-loops ‘protect’ CpG islands from the methyltransferase DNMT3B1 thereby maintaining these regions in a demethylated state (Ginno *et al.*, 2012, 2013b). Further work demonstrated that R-loops may act as a signal and docking site for the demethylase ten-eleven translocation1 (TET1). GADD45A, identified as an R-loop binding protein in S9.6 immunoprecipitation experiments (Cristini *et al.*, 2018) was recruited to R-loops and recruited TET1 in turn (Arab *et al.*, 2019). Whilst this study demonstrated the genome-wide reduction of TET1 recruitment at promoters upon RNase H1 over expression, the corresponding effect on gene expression was only characterised for a single locus. However, a further study uncovered the upstream recruitment of DHX33, a DEAH-box helicase, as necessary for GADD45A recruitment and subsequent TET1 recruitment to enable demethylation. Importantly DHX33 knock down reduced RNA polymerase 2 occupancy at promoters and reduced the transcription of these genes (Feng *et al.*, 2020). Sabino *et al* (2022) used CRISPR to target TET to specific loci thereby locally increasing 5-hydroxymethylcytosine (the first intermediate in demethylation of cytosine). This had no effect on transcription but did increase R-loops at the locus. Together these studies suggest that there is a bidirectional relationship between R-loop formation and cytosine methylation and demethylation.

1.2.2.2 R-loops and histone modifications

In the nucleus DNA is packaged by wrapping around histones. A 147 bp length of DNA wraps 1.65 times around a histone octamer and is joined to other histone-DNA complexes by a 10-60 bp length of linker DNA. The N-terminal tails of histones can undergo modification such as methylation and acetylation and many of these processes are associated with characteristic chromatin states that suppress or permit transcription. By mapping R-loops genome-wide, Sanz *et al* (2016) correlated R-loop occupancy with histone modifications associated with active promoters and an accessible, transcribed chromatin state including H3K4me2, H3K4me3 and H3K27 acetylation. Interestingly, overall gene expression levels as measured by RNA-seq were not a determinant of R-loop occupancy in promoters containing at least one R-loop. The association of accessible chromatin and R-loop formation could be cause or effect: Accessible chromatin could permit transcription therefore increasing changes in DNA:RNA hybrid formation. Alternatively, R-loops could act as an epigenetic signal to recruit histone modifying enzymes. A study in mouse embryonic stem cells showed colocalization of R-loops and the histone acetyltransferase complex Tip60-p400 which promotes transcription at a subset of genes. The reduction of R-loops by RNase H1 over expression reduced the recruitment of Tip60-p400 and favoured silencing of these genes by association with the polycomb repressive complex (PRC). This implies that for histone acetylation, R-loops act to recruit effector enzymes rather than the R-loops occurring secondary to the modification. There are two caveats to this conclusion. Firstly, the study was performed in stem cells which have a different transcriptional profile to mature differentiated somatic cells and so the findings may not be generalisable. Indeed Tip60-p400 can silence genes that induce differentiation in stem cells (Chen *et al.*, 2013). To date, the R-loop occupancy of genes repressed by Tip60-p400 has not been studied. Secondly, the authors did not measure the change in histone acetylation with the RNAse H1 induced reduction in Tip60-p400 recruitment so an off-target effect of the absence of Tip60-p400 cannot be ruled out. By contrast R-loops have also been associated with establishing a repressive chromatin environment at the 3' end of genes to facilitate transcription termination in a subset of genes (Skourti-Stathaki, Kamieniarz-Gdula and Proudfoot, 2014). In this study R-loops formed over RNA polymerase 2 pausing regions 3' of the poly A signal. This promoted antisense transcription and formation of dsRNA between the antisense RNA and R-loop RNA. This enabled cross talk between dsRNA processing factors such as DICER and the histone demethylase G9a/GLP which established the repressive H3K9me2 histone modification. These studies demonstrate how the genomic context can influence whether R-loop formation is repressive or permissive to transcription.

1.2.2.3 R-loops, non-coding RNA, and chromatin accessibility

Non-coding RNA are transcripts that do not encode a functional protein product but can regulate gene expression through interactions with chromatin. Over the last ten years multiple examples of a link between non-coding RNA and R-loops have been described. Two studies have examined anti-sense transcripts and R-loop formation. Work by Boque-Sastre *et al.* (2015) showed bidirectional transcription at the vimentin transcription start site which gave rise to an anti-sense transcript VIM-AS1. This transcript participated in R-loop formation which maintained accessible chromatin and allowed binding of NF-KB, stimulating transcription of the protein coding vimentin transcript. This study used the sensitive R-loop foot printing method where bisulfite treatment enables strand specific and nucleotide resolution mapping of R-loops thus demonstrating with certainty that it is the antisense transcript that forms the R-loop. Inspired by this work, Gibbons *et al.* (2018) demonstrated an R-loop in the GC-skewed promoter region of GATA3, a key pioneer factor for the differentiation of T-helper 2 cells. In contrast to the vimentin locus, the GATA3 antisense transcript does not overlap with its partner gene transcription start site. Overexpression of GATA3-AS1 potentiated GATA3 expression and histone markers of accessible chromatin were detected in the shared promoter region. However, the authors did not resolve the R-loop formed by GATA3-AS1 by over expression of RNase H1 or an R-loop helicase and as such the functional relevance of R-loop formation at this locus needs further validation. In a genome-wide approach, the Proudfoot lab mapped strand specific R-loops and observed a significant overlap with antisense transcripts (Tan-Wong, Dhir and Proudfoot, 2019). Furthermore, antisense transcription and antisense R-loops were particularly sensitive to RNase H1 overexpression and their reduction resulted in a decreased recruitment of general transcription factors as measured by ChIP. Together these observations implied that R-loops can act as promoters for antisense transcripts and highlight another mechanism by which R-loops regulate transcription.

1.2.2.4 R-loops as a torsional stress absorber

As mentioned earlier, transcription can drive negative supercoiling of DNA upstream of RNA polymerase. The chromatin fibre can be viewed as fixed between two points owing to the mass of chromatin up- and down-stream of a gene that is being transcribed. This limits the amount of torsional stress that can absorbed by the DNA fibre. R-loops have been proposed as a torsional stress absorber (Stolz, Sulthana, Stella R. Hartono, *et al.*, 2019; Chedin and Benham, 2020). By using thermodynamic modelling and *in vitro* transcription Stolz *et al.* demonstrated how R-loops could relax twenty times the torsional stress as the effect of DNA wrapping around histones. In this model, the looped out DNA strand was wound around the outside of the DNA:RNA hybrid creating a three-stranded helix. As well as absorbing stress, this model suggests that R-loops could store and release helical stress and the release of this stress upon R-loop resolution could create a more favourable context for the re-establishment of nucleosomes after transcription of a locus has finished. This mechanism has not been experimentally tested however.

1.2.3 Mechanisms of R-loop resolution

Given the thermodynamic stability of R-loops, cells have evolved mechanisms to regulate the formation and resolution of these structures. Proteins such as the DEAD or DEAH box family of helicases and senataxin can specifically unwind DNA:RNA hybrids. Furthermore, RNase H1 and H2 function specifically to hydrolyse RNA hybridised to DNA. Topoisomerase 1 can also function to reduce R-loop accumulation by resolving the negative superhelicity that forms in DNA behind processive RNA polymerase.

1.2.3.1 The DEAD and DEAH box helicases as R-loop processing factors

The DEAD-box family of RNA helicases are characterised by a shared domain containing the Asp-Glu-Ala-Asp motif. Originally characterised as ATP-dependent RNA helicases, the roles for this group of helicases have include RNA processing, co-transcriptional splicing and transcriptional co-activation (Linder and Jankowsky, 2011). Over the last five years numerous studies have shown that DEAD-box helicases can interact with R-loops and may be important for regulating R-loop formation and resolution. DDX5 was identified as an R-loop interacting protein in a study mapping the R-loop interactome (Cristini *et al.*, 2018). Using laser micro irradiation induced DNA damage Yu *et al* (2020) showed that effective DSB repair required DDX5 to resolve R-loops formed near the site of damage. A second study identified a DDX5-BRCA2 interaction at DSBs with BRCA2 enhancing the recruitment of DDX5 to peri-DSB R-loops, their unwinding and subsequent recruitment of Rad51 to facilitate homologous recombination. DDX5 has also been identified as a player in the regulation of gene expression when reprogramming mouse embryonic fibroblasts (Li *et al.*, 2020). In this study DDX5 resolved R-loops at pluripotency genes, reducing their expression and driving differentiation. Contrary to the R-loop resolving role described in these studies a recent pre-print which has not yet been peer reviewed demonstrated how DDX5 expression is repressed in hypoxia yet replacement of DDX5 in this context led to higher levels of R-loops (Leszczynska *et al.*, 2022). This points to context-specific activity of DDX5. Using DRIP-seq in U2OS cells, Villarreal and colleagues (2020) showed that DDX5 knockdown increased the number of R-loops without upregulating transcription of genes that overlapped with R-loop peaks. The paradoxical similar effect on R-loops by DDX5 reduction or increase may be related to the underlying transcriptional context observed in hypoxia (Ramachandran *et al.*, 2021).

The R-loop resolving capability of DDX5 is tightly regulated and dependent upon upstream factors. Mersaoui *et al* (2019) identified PRMT5 as the factor responsible for methylating arginine residues on DDX5. In the absence of PRMT5 (and therefore unmethylated arginines), R-loops accumulated globally and at specific loci. This was accompanied by an increase in RNA polymerase 2 occupancy at the transcription start site and an increase in double strand breaks. Importantly the authors demonstrated that DDX5 has specific DNA:RNA hybrid unwinding activity suggesting that its dysregulation is primarily via R-loops rather than a secondary effect of dysfunctional RNA processing. Numerous other members of the DEAD-box helicase family have been implicated in R-loop homeostasis with roles in DNA damage and gene expression regulation. These are summarised in **table 1.2.**

Table 1.2 DEAD box helicases in R-loop homeostasis.

AID: Activation induced cytidine deaminase, CSR: Class switch recombination, DSB: Double strand break, KD: Knock-down LPS: Lipopolysaccharide, PAR: Poly-ADP ribose, PDAC: Pancreatic ductal adenocarcinoma, RESTAT: All trans retinol 13,14 reductase SMN: Survival motor neuron, TRC: transcription replication conflict.

Reference	**Helicase**	**Disease process/model**	**Mechanism summary**
Song *et al.*, 2017	DDX21	Oestrogen stimulation of MCF7 cells	DDX21 KD increased R-loops at oestrogen responsive genes DDX21 deacetylation by SIRT7 increases R-loop helicase efficiency
Hodroj *et al.*, 2017	DDX19	HeLa cells	DDX19 recruited to DSB ~30 min after DNA damage to resolve R-loops at transcription-replication conflicts
Sridhara *et al.*, 2017	DDX23	U2OS cells; adenoid cystic carcinoma	DDX23 recruited to R-loop forming promoters to resolve stalled RNA polymerase 2
Ribeiro de Almeida *et al.*, 2018	DDX1	Mouse CH12 cells modelling CSR	DDX1 unwinds RNA-G4 structures at IgH locus to facilitate R-loop formation and AID recruitment
Argaud *et al.*, 2019	DDX21	LPS treatment of HEK293 cells	DDX21 KD reduced ENPP2 response to LPS through R-loop accumulation. DDX21 localisation to ENPP2 TSS required functional ENPP2 enhancer
Pérez-Calero *et al.*, 2020	DDX39B	HeLa cells	DDX39B KD increased promoter R-loops and DNA damage secondary to TRC
Pinter *et al.*, 2021	DDX19A	3T3 mouse fibroblast cells	DDX19A recruitment to and resolution of R-loops supports LSD1 recruitment to histones to repress gene expression

Lin *et al.*, 2022	DDX18	U2OS cells	DDX18 is recruited to DNA damage sites in a PAR dependent manner. DDX18 KD increased R-loop induced DNA damage
Karyka *et al.*, 2022	DDX21	SMA patient cell lines; SMN deficient mouse neurons	DDX21 reduced in motors neurons lacking SMN. DDX21 KD increased R-loops and nucleolar DNA damage
Tu *et al.*, 2022	DDX39B	PDAC cell culture and patient derived organoids	High RESTAT expression in hypoxic PDAC recruits DDX39B to resolve R-loops, reduce replication stress and drive resistance to Gemcitabine

The DEAH box family of RNA processing enzymes have also been associated with R-loop homeostasis. Two studies of DHX9 presented contradictory functions for this factor. Using a topoisomerase 1 inhibitor – camptothecin (CPT) – to induce R-loops, Cristini *et al* (2018) found that DHX9 knock down increased the formation of R-loops. This was associated with an increase in γH2AX accumulation, in keeping with an increase in DNA damage. However, work from the Hiom lab (Chakraborty, Huang and Hiom, 2018) showed that DHX9 knockdown *reduced* R-loop accumulation when R-loops were induced by concurrent knock down of RNA splicing factor SFPQ. The authors proposed a model where DHX9 unwinds RNA secondary structure to allow its processing by splicing factors. In the absence of DHX9 secondary structure persists which inhibits re-annealing of the nascent RNA to its DNA template. Whilst these two studies appear contradictory, the key difference is the method of R-loop generation. CPT treatment and splicing factor deficiency make nascent RNA available for R-loop formation by different mechanisms (intron removal vs. helical stress and RNA polymerase 2 stalling respectively). This highlights how the underlying source of R-loops may influence the activity of R-loop resolving factors.

1.2.3.2 Senataxin

Senataxin (SETX) is a DNA:RNA helicase with significant homology to its yeast counterpart Sen1. Studies in yeast showed that Sen1 can unwind and resolve R-loops and similar effects were reported in mammalian cells. Initially characterised as resolving R-loops at transcription termination sites to facilitate transcription termination (Skourti-Stathaki, Proudfoot and Gromak, 2011) SETX was later shown to localise to R-loops across the genome including in promoter regions and at common fragile sites (Jurga *et al.*, 2021). In addition SETX has been detected at double strand breaks with associated R-loop formation. Its role here is the coordination of Rad51 to allow efficient homologous recombination. Importantly SETX mutations are seen in the neurodegenerative disorder Amyotrophic Lateral Sclerosis type 4 (ALS4) linking R-loops with this condition. Interestingly, in mutant SETX in ALS4 is associated with a decrease in R-loop burden, a deprotection of genes from methylation and subsequent altered transcriptional programmes that favour TGF-beta transcription(Grunseich *et al.*, 2018). Recently, regulation of SETX has been studied by the El-Khamisy lab. This work showed that a USP11 – KEAP1 circuit regulated ubiquitination of

SETX and that loss of USP11 was associated with R-loop induced DNA damage (Jurga *et al.*, 2021). USP11 was also implicated in a study of MYCN amplified neuroblastoma (Herold *et al.*, 2019). Here MYCN over expression reduced R-loop formation and USP11 depletion was associated with the retention of MYCN on chromatin. The overall effect in the two studies was similar: that USP11 loss was associated with an increase in R-loop burden but via different mechanisms.

1.2.3.3 Topoisomerase 1

The function of topoisomerase 1 (top1) in R-loop homeostasis has been studied by siRNA knock down. Manzo *et al.* (2018) performed stranded genome-wide characterisation of R-loops in the setting of top1 knock down. Unexpectedly this showed that there were more R-loops lost than gained upon top1 knockdown. However, the R-loop gain that was observed was cotranscriptional and centred on gene bodies, commensurate with disruption of RNA polymerase 2 elongation as the R-loop initiating event. Another study used a similar methodology and found that DRIP-seq signal increased at transcription termination sites with top1 depletion (Promonet *et al.*, 2020). This increase in signal coincided with an increase in DNA damage as measured by γH2AX ChIP-seq and iBLESS, a method of mapping double strand breaks. These findings supported data from an earlier study where Top1 knock down led to an increase in γH2AX foci and a higher incidence of transcription-replication collisions, a key source of R-loop mediated double strand breaks (Tuduri *et al.*, 2009). DNA damage was reduced by overexpressing RNase H1 implicating R-loops in the process however the authors did not measure R-loop levels directly so indirect effects of RNase H1 overexpression cannot be ruled out.

1.2.4 Endogenous and exogenous initiators of R-loop formation

Many of the mechanisms of R-loop formation and homeostasis described above relied upon the induction of R-loops through stalling of RNA polymerase via CPT induced Top1 cleavage complexes or by inhibition of RNA splicing. Other cellular insults have also been associated with R-loop accumulation. Hypo-osmotic stress was shown to induce R-loop formation and a DNA damage response in nucleoli. Interestingly, no increase in DNA strand breaks was detected using comet assay. The comet assay is a sensitive method for detecting DNA damage including single strand breaks, double strand breaks an abasic sites. The nucleolus occupies a small proportion of the total nuclear volume therefore DNA damage limited only to the nucleolus as reported in this study may have been at a level below the limit of detection for this assay. This provides an explanation for the lack of DNA damage detected by the comet assay. ATR was however recruited to the osmotically induced R-loops and this activated a signalling cascade to reduce transcription of rRNA, presumably to limit cell damage under stressed conditions (Velichko *et al.*, 2019). Hypoxia is another external condition that can induce R-loop formation. Under hypoxic conditions, colorectal cancer cells were more prone to R-loop formation. Apparently as an adaptive response, SETX was upregulated and depletion of SETX led to a greater accumulation of R-loops under hypoxic conditions and an increase in DNA damage (Ramachandran *et al.*, 2021). These findings correlated with clinical data from The Cancer Genome Atlas where SETX was found to be over expressed in tumours demonstrating a hypoxic gene expression signature. Lastly, reactive oxygen species (ROS) have been shown to induce R-loop formation. Using a system that induces locus specific ROS production, Teng *et al.* (2018) demonstrated induction of R-loops at the same locus. The presence of a DNA:RNA hybrid recruited Cockayne syndrome B (CSB) protein to regulate DNA damage repair. The authors did not explore how ROS trigger R-loop formation initially. Previous work has shown that a single strand break can initiate R-loop formation in an *in vitro* transcription system (Roy *et al.*, 2010) so if a similar lesion was caused by ROS, this could provide a mechanistic explanation for ROS induced R-loops. Alternatively, as ROS can stall RNA polymerase 2 (Kolbanovskiy *et al.*, 2017), this could create an environment favouring R-loop formation. These hypotheses remain untested currently.

1.2.5 Pathological consequences of R-loops (figure 1.2)

1.2.5.1 R-loops and DNA damage

DNA damage is one of the main pathological consequences of dysregulated R-loop homeostasis. The underlying mechanisms can be broadly grouped as transcription-replication collisions, activation of transcription associated nucleotide excision repair or cytidine deaminase activity, typified by the AID/APOBEC family of enzymes. Transcription-replication collision/conflict (TRC) is thought to arise as an R-loop presents an obstacle to an oncoming DNA replication fork. Importantly head-on collisions specifically cause double strand breaks (Hamperl *et al.*, 2017). Chappidi *et al.* (2020) proposed that DNA supercoiling occurs between transcription and replication machinery and if this remains unresolved then replication fork collapse occurs with subsequent DNA damage.

Transcription coupled nucleotide excision repair (TC-NER) is another common cause of R-loop mediated DNA damage and can occur throughout the cell cycle, in contrast to TRC. The looped-out DNA strand of an R-loop presents a substrate to bulky DNA lesions that are canonically repaired by TC-NER. Factors such as XPF and XPG can process R-loop and cells lacking these proteins are more susceptible to R-loop associated DNA damage induced by CPT (Sollier *et al.*, 2014). Defects in the NER pathway can lead to an accumulation of cytoplasmic R-loop fragments, implicating this R-loop DNA repair pathway in chronic inflammation and fibrosis (Chatzidoukaki *et al.*, 2021). The activity of XPF and XPG can also combine with topoisomerase 1 activity, converting single strand breaks into double strand breaks (Cristini *et al.*, 2019).

In an example of physiological R-loop regulated DNA damage, R-loops formed in the class-switch region of the IgH locus recruit AID. Then, via deamination of cytosine to uracil, AID creates double strand breaks which facilitate class switch recombination (Ribeiro de Almeida *et al.*, 2018). Although the related APOBEC family of enzymes has been proposed as a pathological effector of R-loop mediated DNA damage, no peer-reviewed evidence of this has been published. In a recent pre-print study which has not been peer reviewed (McCann *et al.*, 2021) APOBEC3B was shown to interact with R-loops *in vitro and in vivo.* Furthermore, depletion of APOBEC3B was associated with increased R-loops. Whilst a direct effect of APOBEC3B on the looped-out DNA strand could not be assessed, there was an association between splicing factor mutations (associated with R-loop formation) and APOBEC mutational signatures in TCGA sequencing data. Whilst these data don't definitively implicate APOBEC in the damage of ssDNA associate with R-loops they do provide an additional dimension in the complex interplay between transcription, R-loop formation and DNA damage.

1.2.5.2 R-loops in cancer

Given the association of R-loops with fundamental cellular processes such as transcription regulation and with the pathological process of DNA damage it is unsurprising that R-loops have been implicated in the pathogenesis of cancer. R-loops have been studied in breast cancer where oestrogen stimulation of breast cancer cells was shown to cause an R-loop mediated increase in DNA damage (Stork *et al.*, 2016). R-loop mapping by DRIP-seq in patient tissues demonstrated an increase in R-loop accumulation in cancers with mutant BRCA1. R-loops were increased at a specific genomic locus overlapping an oestrogen induced enhancer which decreased enhancer interactions with its target genes. This interfered with normal differentiation of breast epithelial cells and directed cells instead towards the more aggressive basal subtype (Chiang *et al.*, 2019). In Ewing sarcoma, an aggressive bone and soft tissue cancer, the characteristic EWS-FLI1 mutation was responsible for driving high rates of transcription. This increase R-loop levels which then acted as a sink for BRCA1, sequestering this factor (Gorthi *et al.*, 2018). The resulting functional BRCA1 deficiency sensitised the cells to PARP inhibition highlighting how R-loop mediated pathology could also present therapeutic opportunities. A similar mechanism was observed in another soft tissue cancer - synovial sarcoma - typified by an S18-SSX1 gene fusion. Here the gene fusion increased R-loops and replication stress and mediated a sensitivity to treatment with ATR inhibitor (Jones *et al.*, 2017).

Bauer *et al.* (2020) linked gastric inflammation initiated by *H. Pylori* infection to increased R-loop accumulation and subsequent replication stress induced DNA damage. In this setting, the inflammatory response associated with *H Pylori* infection upregulated NF-KB signalling. The associated upregulation in transcription increased global R-loops levels and induced replication stress. The study also showed that RNase H1 overexpression could rescue this phenotype implicating R-loop in this specific DNA damage mechanism. As *H. Pylori* is a major cause of gastric cancer this offers the possibility of R-loop mediated DNA damage acting as an early or initiating event in this malignancy. Lastly, a study using the ATAC-seq data from TCGA showed that the presence of R-loops can act as a mechanism to maintain open chromatin in cancer (Guo *et al.*, 2021). The chromatin binding of NR4A1 stalled RNA polymerase 2 and R-loops were maintained across the gene bodies of immediate early response genes (IEGs), a group of genes rapidly transcribed in the event of cell stress. The stalled RNA polymerase 2 keeps these genes in a state of partial transcription which can then be rapidly completed by dissociation of NR4A1 from chromatin. Interestingly overexpression of RNase H1 increased the expression of the same early response genes implying interplay between R-loops formation and transcriptional repression. Alternatively, RNase H1 over expression could have generated cell stress and led to the activation of IEGs.

In summary, multiple lines of evidence implicate R-loop formation in physiological and pathological processes. R-loops have been associated with DNA damage in diverse experimental systems and appear to be a response to a variety of cellular stressors. Moreover, R-loops are pivotal in regulating gene expression, and their dysregulation can lead to altered transcriptional programmes and cellular dedifferentiation.

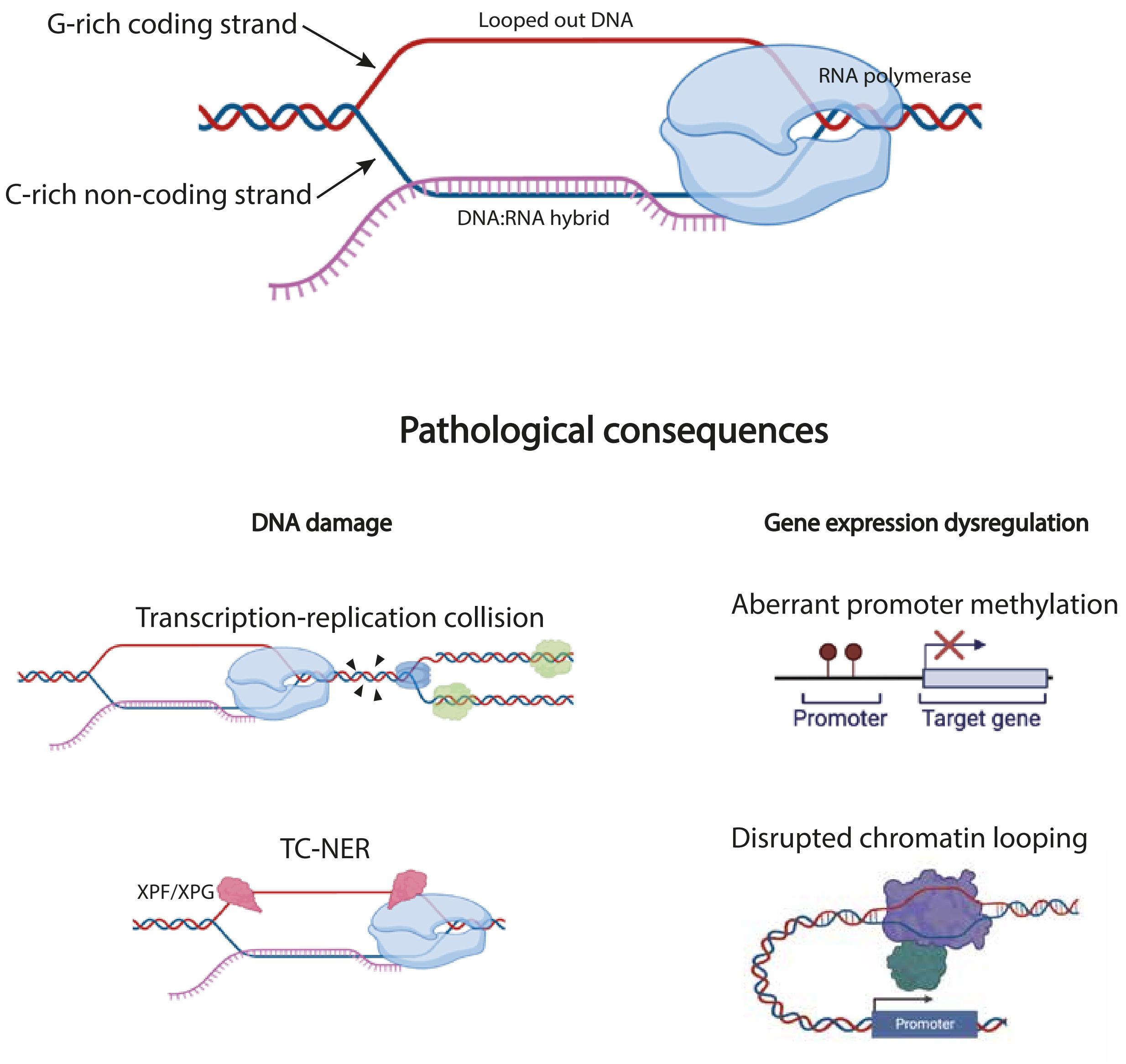

G-rich coding strand
Looped out DNA
RNA polymerase
C-rich non-coding strand
DNA:RNA hybrid
Pathological consequences
DNA damage
Gene expression dysregulation
Transcription-replication collision
Aberrant promoter methylation
Promoter
Target gene
TC-NER
XPF/XPG
Disrupted chromatin looping
Promoter

Figure 1.2 Pathological consequences of R-loops
TC-NER: Transcription coupled nucleotide excision repair.

1.3 R-loops, androgen signalling and prostate cancer – a possible connection?

Few studies have addressed the relationship between androgen signalling and R-loops. Kumar Gupta et al (2018) demonstrated that DHT treatment could bring the TMPRSS2 and ERG loci in proximity to one another and then antisense R-loop formation by a non-coding RNA transcribed in *cis* could facilitate the fusion between these two genes. Importantly, this fusion is clinically relevant being present in ~50% of prostate cancer cases. However, this study did not address how androgens might affect gene expression through R-loop homeostasis. In a separate study of gene expression regulation, oestrogen or retinoic acid were used to induce liganded nuclear receptor mediated transcription. The authors saw a reduction in chromatin loops between enhancers and promoters after treating extracted DNA with recombinant RNase H1 and also demonstrated R-loop formation upon transcriptional induction (Pezone *et al.*, 2019). However, no *in vivo* resolution of R-loops was performed (i.e. by RNase H1 over expression) and so the functional relevance of R-loop homeostasis and nuclear receptor mediated chromatin looping could not be determined.

1.4 Synopsis

The aims of this thesis are to investigate links between androgen signalling and R-loop formation in prostate cancer cells. R-loops are formed co-transcriptionally and can affect gene expression and initiate DNA damage. Androgen signalling works through modulating transcriptional programmes and its modulation can cause DNA damage in prostate cancer cells. A link between androgen signalling and R-loops could uncover novel mechanisms governing prostate cancer pathophysiology. In chapter three I developed a computational tool to predict the prevalence of R-loop forming sequences in different androgen signalling contexts. In chapters four and five I investigate the genome-wide distribution of androgen signalling induced R-loops and test their functional significance. In chapter six I examine the mechanisms modulating androgen induced R-loops. Finally, I discuss these findings in the wider context of the R-loop and androgen signalling literature in chapter seven.

Chapter 2: Materials and Methods

2.1 Solutions, reagents and equipment

2.1.1 Solutions

Phosphate buffered saline (PBS)

One PBS tablet was dissolved in 1 L ddH$_2$O then autoclaved. PBS with Tween (PBST) was made by adding Tween-20 to a final concentration. of 0.01%.

Tris buffered saline (TBS)

Tris base (24.2 g) and NaCl (81.8 g) were dissolved in one litre of ddH$_2$O. HCl was added to pH 7.9.

TBS with Tween (TBST) was made by adding Tween-20 to a final concentration of 0.1%.

Sodium Acetate; 3 M pH 5.2

24.6 g of sodium acetate powder was dissolved in 70 ml ddH$_2$O. The pH was adjusted to 5.2 using glacial acetic acid and the volume adjusted to 100 ml with ddH$_2$O. The solution was then filter sterilised.

SDS-PAGE running buffer (10x)

Tris base (30.3 g), glycine (187.7 g) and SDS (10 g) were dissolved in one litre of ddH$_2$O.

Semi Dry Western blot Transfer Buffer (10x)

5x TransBlot Turbo (BioRad) stock buffer was diluted as per the manufacturer's instructions by adding 200 ml of stock buffer to 600 ml ddH$_2$O and 200 ml 100% ethanol.

Tris borate EDTA pH 8.0 (TBE; 10x)

Tris base (108 g) and boric acid (54 g) were dissolved in one litre of ddH$_2$O. The pH was adjusted to 8 using 0.5 M EDTA.

Sodium dodecyl Sulphate (SDS) 10%

10 g SDS powder was dissolved in 100 ml ddH$_2$O.

Tris 1.5 M pH 8.8

Tris base (90.8g) was dissolved in 400 ml ddH_2O. The pH was adjusted to 8.8 using HCl and the solution made up to a final volume of 500 ml.

Tris 1 M pH 6.8

Tris base (60.6 g) was dissolved in 400 ml ddH_2O. The pH was adjusted to 6.8 using HCl and the solution made up to a final volume of 500 ml.

Ammonium persulphate (APS) 10%

1 g of APS powder was dissolved in 10 ml of ddH_2O and aliquoted. Aliquots were stored at -20 °C.

Protein loading buffer 5x

SDS powder (1 g), bromophenol blue (500 mg) and DTT powder (771.25 mg) were dissolved in 2.5 ml ddH_2O, 2.5 ml 250 mM Tris-HCl (pH 6.8) and 5 ml 50% glycerol.

Paraformaldehyde 37% (PFA)

1.85 g of paraformaldehyde powder (Sigma) was suspended in 3.5 ml ddH_2O. 10 µL of 10 M KOH was added and the solution was heated by microwave until the PFA had dissolved. Solution volume was made up to 5 ml. PFA was then diluted in PBS according to experimental conditions. All steps were performed in a fume hood. PFA was made fresh for each use.

Luria broth

LB powder (14 g) was dissolved in 400 ml ddH_2O. The solution was autoclaved.

Luria broth agar

LB agar powder (8 g) was dissolved in 400 ml ddH_2O. The solution was autoclaved.

All solutions were stored at room temperature unless otherwise indicated.

2.1.2 Reagents, chemicals and kits

Reagent	Manufacturer
1 KB Plus DNA ladder	New England Biolabs
Acetone (99.9%, molecular biology grade)	Fisher Scientific
AMPure XP beads	Beckman Coulter
Charcoal-stripped foetal bovine serum	Gibco
Dihydrotestosterone	Sigma-Aldrich/Merck
DMSO	Sigma
Ethanol (99.9%, molecular biology grade)	Fisher Scientific
Foetal bovine serum	Gibco
Glycogen	Invitrogen
High-capacity cDNA kit	Invitrogen
ImmunMount mounting solution	Invitrogen
L-glutamine	Sigma
Lipofectamine 2000	ThermoFisher
Lithium chloride	Invitrogen
Luria broth (LB) and LB agar	ThermoFisher
Methanol (99.9%, molecular biology grade)	Fisher Scientific
Midiprep plasmid isolation kit	Qiagen
Miniprep plasmid isolation kit	Qiagen
Molecular biology grade water	Promega
NP-40	Sigma
Penicillin-Streptomycin	Sigma
Phenol:Chloroform:Isoamyl alcohol (25:24:1)	Invitrogen
Precision plus protein ladder	Bio-Rad
Protein A/G beads	ThermoFisher
Quantinova qPCR mastermix	Qiagen
RNeasy RNA extraction kit	Qiagen
RPMI-1640	Sigma

Sodium deoxycholate	Sigma
Sodium lauryl sarcosinate	Sigma
TE buffer	Invitrogen
Triton-X	Sigma
Tween-20	Sigma
xGen stubby adapters and primers	Integrated DNA Technology (IDT)

2.1.3 Equipment

Equipment	Manufacturer
Heraeus Pico 17 tabletop centrifuge	ThermoFisher
Refrigerated centrifuge	Centurion
TE300 inverted fluorescence microscope	Nikon
Phase contrast microscope	Nikon
Rotor gene qPCR machine	Qiagen
ChemiDoc MP imaging system	BioRad
TransBlot Turbo transfer system	BioRad
SSL3 Gyro Rocker	Stuart
SRT6 Roller Mixer	Stuart
Tube revolver	ThermoFisher
UVP crosslinker	AnalytikJenka
Thermal cycler	3Prime
pH meter	Jenway
Thermomixer	Eppendorf
Bioruptor Pico sonicator	Diagenode
Nanodrop	ThermoFisher
Bioanalyser	Agilent
Qubit fluorometer	ThermoFisher

2.2 Molecular biology techniques

2.2.1 Cell culture

LNCaP cells were acquired from the American Type Culture Collection and routinely tested for mycoplasma. LNCaP cells were maintained in RPMI-1640 media supplemented with 10% foetal calf serum, 2 mM L-glutamine and 100 units/ml of penicillin and 0.1 mg/ml streptomycin in T-75 or T-175 flasks. Cells were passaged at 80% confluency by discarding the media, washing once with warm sterile PBS and adding trypsin (2 ml for T-75, 4 ml for T-175) for 2 minutes at 37 °C. Trypsin activity was stopped by adding four times the trypsin volume of fresh media and the cell suspension aspirated then pelleted by centrifuge (100 *g*, 5 minutes, room temperature). The media was discarded and the cell pellet resuspended by gentle pipetting in 1 ml media. This was then further diluted to achieve cell concentrations appropriate for individual experiments. A Neubauer chamber was used to count cells. Low passage number (< 20) cells were used for all experiments. Biological replicates were grown in separate flasks and separated by at least two passages from their parental cell culture flask. Media and trypsin volumes were scaled up or down according to culture vessel surface area. For immunofluorescence experiments cells were grown on glass coverslips.

For androgen stimulation experiments cells were seeded on day one in full media. After 24 hours the media was removed and the cells were gently washed with warm PBS to remove traces of hormone-containing media. Fresh media containing charcoal stripped serum was then added and the plates placed back in the incubator for 48 hours prior to androgen stimulation. Dihydrotestosterone (DHT) was added to the media at the concentrations and timepoints indicated in the text and figure legends.

2.2.2 DNA:RNA hybrid immunoprecipitation (DRIP)

Extraction of genomic DNA

DRIP was performed as per the protocol by Sanz and Chedin (2019) with minor modifications. Two million LNCaP cells were seeded onto 15 cm culture plates, and grown as described above. Media was removed from culture dishes and cells washed with 10 ml of ice-cold PBS. Cells were then scraped in 5 ml of ice-cold PBS and the cell suspension aliquoted for DRIP and parallel RNA and/or protein extraction. The cell suspension was centrifuged at 200 *g* for 5 mins at 4 °C to pellet the cells then resuspended in 1.6 ml of TE buffer. Fifty microlitres of 20% of SDS and 10 uL of 10 mg/ml proteinase K were added, the suspension was mixed gently by inversion then incubated at 37 °C for 16 hours. The DNA lysate was poured into a 15 ml MaxExtract high-density tube (Qiagen) and mixed with 1.6 ml phenol/chloroform/isoamyl alcohol by inversion. This was centrifuged at 1500 g until the supernatant was colourless. The supernatant was poured into a 15 ml tube containing 4 ml 100% ethanol and 160 µL sodium acetate (3 M, pH 5.2). Genomic DNA was precipitated by end-over-end mixing for at least 10 minutes. The DNA was then spooled out of the solution using a P1000 pipette tip and moved to 1.5 ml microcentrifuge tube. Three 10-minute washes with 70 % ethanol were done to remove residual salts and contaminants. The DNA was then dried until translucent and 125 µL TE buffer added. The DNA was allowed to rehydrate on ice for 1 hour then resuspended using a cut 200 µL tip. Following a further hour on ice, 80 – 100 µL DNA was digested for 16 hours at 37 °C using a cocktail of restriction enzymes (SSP, BSRGI, ECoRI, HindIII, XbaI; 30 units each) in the presence of 200 µg/ml BSA, 1 x NEB buffer 2.1 and 1 mM spermidine in a total volume of 150 µL.

The digested DNA was purified by adding 100 µL water and 250 µL phenol/chloroform/isoamyl alcohol then centrifuged in a light phase lock gel tube for 10 minutes at 16000 g. The supernatant was added to 2.5 volumes (625 µL) 100% ethanol, 1/10 volume (25 µL) sodium acetate and 1.5 µL glycogen then mixed by inversion. This was incubated at -20 °C for overnight to precipitate the DNA. The tube was then centrifuged for 35 minutes at 16000 g at 4 °C. The supernatant was removed and 200 µL of ice-cold 70% ethanol added. A second centrifuge step was then done for 10 minutes at 16000 g at 4 °C. The supernatant was removed and the DNA pellet allowed to dry. The DNA pellet was resuspended in 50 µL TE buffer and stored at -80 °C until immunoprecipitation.

Immunoprecipitation

8 µg of DNA was diluted in 500 µL TE buffer and a 50 µL aliquot saved as input for qPCR. The sample was incubated with 52 µL 10x DRIP binding buffer (1.4M sodium chloride, 100 mM sodium phosphate, 0.5% Triton X-100 in nuclease free water) and 20 µg S9.6 antibody overnight at 4 °C with rotation. The next day, 90 µL of protein G beads were washed twice with 700 µL 1x DRIP binding buffer then the beads mixed with the sample and incubated for 2 hours at 4 °C with rotation. Following incubation, the beads were washed twice with 700

µL 1x DRIP binding buffer. Bound R-loops were eluted by resuspending the beads in 300 µL DRIP elution buffer (50 mM Tris pH8, 10 mM EDTA pH 8, 0.5% SDS in nuclease free water) with 14 µL of 10mg/ml proteinase K and incubating at 55 °C for 45 minutes with 10 RPM rotation. Following magnetic separation of the beads, the supernatant was transferred to phase lock gel tubes and purified following the same procedure for earlier purification of genomic DNA. The resulting DNA pellet was resuspended in 50 µL TE buffer and stored at -80 °C until qPCR and/or sequencing.

2.2.3 Library preparation and high throughput sequencing for DRIP-seq

Following quality control by qPCR, immunoprecipitated genomic DNA was sonicated for 12 cycles (15 seconds on/60 seconds off) on a Bioruptor Pico (Diagenode). Sonicated DNA was end repaired by incubation with NEBnext end repair enzyme (New England Biolabs) for 30 mins at room temperature. The reaction was cleaned up using AMPure XP beads (Beckman Coulter) and one on-bead wash with 80% ethanol. Next, A-tailing was done using NEB Klenow fragment exo- and incubation at 37 C for 30 mins. After a second Ampure XP bead clean up, IDT xGen stubby adapters (Integrated DNA Technologies) were ligated using NEB quick ligase. After a further Ampure XP bead clean up the library was eluted in 20 µL of 10 mM Tris-HCl (pH 8). Half of this volume was used for library amplification using the IDT xGen index primer mix and Phusion Flash PCR master mix (Thermo Fisher Scientific). The remaining volume was saved in case of failed library amplification. The following programme was used on a thermal cycler: 1 cycle for 30 seconds at 98 C; 10 cycles of 10 seconds at 98 C, 30 seconds at 60 C, 30 seconds at 72 °C; 1 cycle for 5 minutes at 72 C. A two stage clean up of the amplified libraries was then performed to select fragments between 200 and 500 bp. Library concentration and fragment size distribution were then checked using a Qubit and Bioanalyser respectively. Sequencing (150 bp, paired end) was performed by Novogene UK on a Novaseq 6000 system (Illumina).

2.2.4 Chromatin Immunoprecipitation

Cell harvesting and paraformaldehyde fixation

LNCaP cells (2 x 10^6) were seeded in two 15 cm culture dishes per biological replicate. Treatment conditions are described in figure legends. Cells were fixed by discarding media then adding 12 ml 1% paraformaldehyde and incubating at room temperature for 10 minutes. The formaldehyde was quenched by adding glycine to a final concentration of 125 mM. This was mixed for 5 minutes then the fixed cells were washed twice in ice cold PBS and scraped in 15 ml PBS. Cells were pelleted by centrifuge (200 *g*; 5 minutes; 4 °C), the supernatant discarded and the remaining cell pellet was snap frozen in liquid nitrogen and stored at -80 °C prior to lysis.

Lysis

Cells were lysed by resuspending the pellet in five pellet volumes of ChIP lysis buffer 1 (50 mM HEPES-KOH pH 7.5, 140 mM NaCl, 1mM EDTA, 10% glycerol, 0.5% NP-40, 0.25% Triton X-100) and incubating on ice for 5 minutes with intermittent mixing by inversion. Nuclei were pelleted by centrifugation (3000 *g* for 5 minutes at 4 °C) and the pellet resuspended in five pellet volumes of ChIP lysis buffer 2 (10 mM Tris-HCl pH 7.4, 200 mM NaCl, 1mM EDTA pH 8, 0.5 mM EGTA pH 8). This suspension was incubated at room temperature for 10 minutes with 20 RPM rotation then pelleted by centrifugation (1500 *g* for 5 minutes at 4 °C). The pellet was resuspended in 200 µL ChIP lysis buffer 3 (10 mM Tris-HCl pH 7.4, 200 mM NaCl, 1mM EDTA pH 8, 0.5 mM EGTA pH 8, 0.1% sodium deoxycholate, 0.5% N-laurylsarcosine) and incubated on ice for 5 minutes. All lysis buffers were supplemented with 1x protease inhibitor and 1x phosphatase inhibitor.

The lysate was sonicated (four cycles of 20 seconds on/ 40 seconds off) using a Diagenode Bioruptor Pico and centrifuged (20,000 *g* for 15 minutes at 4 °C). The supernatant was transferred to a new tube and 10 µL removed to check sonication efficiency and DNA concentration. Protein-DNA crosslinks were reversed by adding 190 µL ChIP elution buffer and 8 µL 5M NaCl then incubating at 65 °C for 16 hours. RNA contaminants were removed by adding 200 µL TE buffer/4 mM $CaCl_2$, 8 µL RNAse A (10 mg/ml) and incubating at 37 °C for 30 minutes with 800 RPM shaking. Eight microlitres of 10 mg/ml proteinase K was added and the sample incubated for a further 2 hours at 55 °C with 800 RPM shaking. Phenol-chloroform DNA extraction and ethanol/sodium acetate precipitation was then done as described in section 2.2.2 (DNA:RNA immunoprecipitation). A 5 µL aliquot was used to check sonication efficiency on a 1.2% agarose/ethidium bromide gel (see section 2.2.11). DNA concentration and purity were analysed using a Qubit and NanoDrop Spectrophotomer respectively.

Immunoprecipitation

Thirty micrograms of sonicated chromatin was diluted with 4x the lysate volume of ChIP dilution buffer (16.7 mM Tris-HCl pH 7.4, 167 mM NaCl, 1.2 mM EDTA pH 8, 1.1% Triton X-100, 1x protease and phosphatase inhibitors). This was incubated with either 2 µg antibody or IgG at 4 °C for 16 hours with 20 RPM rotation. A 1% volume of lysate was saved as input for qPCR. Thirty microlitres of Protein A or G Dynabeads were washed with 200 µL ChIP dilution buffer, resuspended in 30 µL ChIP dilution buffer and added to the lysate/antibody mixture. This was incubated at 4 °C for 2 hours with 20 RPM rotation. The beads were then incubated in a series of wash buffers (low-salt, high-salt and lithium chloride buffers; 5 minutes each with 20 RPM rotation at room temperature). Between washes beads were immobilised on a magnetic rack, the supernatant discarded and beads resuspended in the appropriate buffer. Immunoprecipitated DNA was eluted by resuspending the beads in 100 µL ChIP elution buffer and incubated at 65 °C for 30 minutes with 1000 RPM shaking. This was repeated once. The 200 µL eluate was then subjected to reverse crosslinking, RNAse A and proteinase K treatment, and phenol-chloroform DNA purification as described above. The 10 µL purified eluate was diluted to a total volume of 50 µL with TE buffer prior to qPCR.

2.2.5 RNA extraction and reverse transcription

RNA was extracted using the RNeasy kit (Qiagen) as per the manufacturer's instructions. RNA concentration and purity were checked using a NanoDrop spectrophotometer. Reverse transcription to cDNA was done using the High Capacity cDNA kit (Thermo Fisher) following the instructions for use. Briefly, 500 – 1000 ng of RNA in a 10 µL volume was added to 10 µL of 2x reverse transcription mastermix containing reverse transcription buffer, dNTP mix, random primers and reverse transcriptase. The final concentration of dNTP was 4 mM. The 20 µL containing RNA and mastermix was incubated in a thermal cycler with settings 25 °C for 10 minutes, 37 °C for 120 minutes, 85 °C for 5 minutes and a final hold at 4 °C. The resulting cDNA was stored at -20 °C.

2.2.6 Library preparation for RNA-seq

Stranded RNA-seq libraries (paired end, 150 bp reads) were prepared by Novagene UK from extracted RNA. Sequencing was performed on a Novaseq 6000 platform (Illumina).

2.2.7 Quantitative qPCR

GAPDH was used used as a housekeeping gene as described previously (H. Zhao *et al.*, 2018). For quantification of gene expression, primers were designed to span exon-exon boundaries to reduce spurious results by contamination with residual genomic DNA. Primers were designed using PrimerQuest (https://eu.idtdna.com/PrimerQuest/Home/) and checked for off target hybridisation against HG38 genomic sequence using Primer Blast (https://www.ncbi.nlm.nih.gov/tools/primer-blast/). Primers were ordered from Integrated DNA Technologies (UK). The preferred Refseq transcript was used for RT-qPCR primers. DRIP-qPCR primers were designed from the FASTA sequence of R-loop forming regions from either DRIP-seq or qmRLFS data. Detailed primer information is given in appendix 1.

For immunoprecipitation followed by qPCR, standards were created by combining equal volumes of input samples. This combined sample was denoted standard 1 and was then serially diluted 1:10 with nuclease free water to make 4 standards in total. A mastermix was made by combining n x 10 µL Quantinova mix, n x 2.8 µL combined forward and reverse primers and n x 5.2 µL nuclease free water where n is the total number of samples in a single qPCR run (standards + inputs + samples). Eighteen microlitres of mastermix were added to each tube. Two microlitres of standard, input or sample were added to each qPCR tube. The final primer concentration was 0.7 µM.

A similar procedure was followed for RT-qPCR except standards were made by combining aliquots of the condition with the highest anticipated expression followed by serial dilution. The mastermix was made by combining n x 10 µL Quantinova mix, n x 2.8 µL combined forward and reverse primers and n x 2.2 µL nuclease free water where n is the total number of samples in the run and 15 µL of mastermix added to each tube. A 5 µL volume of standard or sample was added to each tube. Reactions were run as technical duplicates. Filter pipette tips were used throughout.

Samples were run analysed using a Rotargene 6000 qPCR machine (Qiagen) with the following settings:

- Initial denaturation: 95 °C for 10 minutes.
- 50 cycles of:
  - Denaturation: 95 °C for 10 secs.
  - Annealing: 5 °C below primer melting temperature for 15 secs. Primer melting temperatures were calculated using OligoAnalyzer from Integrated DNA Technologies (www.eu.idtdna.com).
  - Extension: 72 °C for 30 secs.
- Melt curve: 72 °C - 95 °C in 1 °C increments at 5 sec intervals with continuous fluorescence acquisition.

The standards were analysed using the 'auto find threshold' command in the Rotor Gene Q software package. The resulting R^2 and efficiency values were used to assess primer specificity and technical reproducibility. Melt curves were also examined for every qPCR run. The thresholding procedure was repeated for each individual primer producing calibrated cycle threshold (Ct) values. For DRIP- or ChIP-qPCR the Pfaffl method (Pfaffl, 2001) was used to calculate % enrichment over input samples using the formula 100 x $2^{(\text{corrected Ct input value} - \text{sample Ct value})}$. The corrected Ct input value was calculated by subtracting $\log_2 10$ (3.32) from the input Ct value. For RT-qPCR the delta-delta Ct method was used to calculate fold change in expression after normalisation against housekeeping genes (Livak and Schmittgen, 2001). In the case of two-factor experimental designs, the standard curve method was used to allow comparison across multiple combinations of conditions to one baseline condition.

2.2.8 Immunofluorescence

Cells were grown on glass coverslips in 6 well plates then transferred to 24 well plates prior to androgen treatment. Coverslips were washed twice with ice-cold PBS then fixed with 500 µL 4% formaldehyde for 10 minutes at room temperature. Following three 500 µL PBS washes, cells were permeabilised with 0.1% Triton X-100 for 10 minutes at room temperature. For experiments using the S9.6 antibody cells were fixed and permeabilised in 500 µL ice-cold methanol/acetone (2:1) for 10 minutes at -20 °C. Cells were washed with 500 µL PBS then blocked using 500 µL of 3% BSA in PBST (PBS + 0.01% Tween-20) for 1 hour at room temperature. The primary antibody was added and the coverslips incubated overnight at 4 °C. Details of antibodies are provided in appendix 7. The next morning the antibody was removed and coverslips were washed three times with 500 µL 3% BSA in PBST. The secondary antibody was added and coverslips incubated at room temperature for one hour in the dark. Three 500 µL PBST washes were done then the coverslips incubated with 1:1000 DAPI for 15 minutes in the dark. After two PBST washes and a final PBS wash, coverslips were dried on paper towel for 5 minutes then mounted onto glass slides using Immun-Mount (Thermo Fisher). Slides were stored in the dark at 4 °C until imaging. A Nikon Eclipse Ti-2 fluorescence microscope with a 63 X oil immersion lens was used for imaging. NIS-elements was used for image acquisition employing a region of interest 1952 x 1952 µM (0.11 µM/pixel). Details of microscope settings for each antibody and experiment are provided in appendix 6. Segmentation and quantification were performed in FIJI using custom scripts for generation of nuclear masks and measurement of fluorescence intensity. Data were normalised to the median of a baseline control condition specific to each experiment (details given in figure legends) as described by Caicedo *et al.* (2017).

2.2.9 Slot blot

Genomic DNA was extracted as per the DRIP procedure described above. For each experiment three samples were prepared each comprising one microgram of DNA in 100 µL final volume of TE buffer. One sample was treated with 2 µL recombinant RNAse H1 for 24 hours at 37 °C. Nylon membrane was cut to size and soaked in TBS for 10 minutes. The membrane was placed in the slot blot apparatus and, after securing the screws, slots that were not required were covered with tape. A 300 mbar vacuum was applied to the membrane for one minute then switched off. One hundred microlitres of sample was loaded into each slot and the vacuum applied at 300 mbar for 2 minutes then 600 mbar for 2 minutes. Whilst the vacuum was still on, the screws were untightened, and the membrane removed. All samples were then dried for 2 minutes at room temperature then nucleic acids cross-linked to the membrane by exposure to 120000 uJ/cm2 in a UV cross-linker. The membrane was cut to separate the samples for S9.6 and dsDNA staining as necessary. Blocking, primary and secondary antibody incubation and visualisation then followed the same procedure as Western blotting described in section 2.2.10 . BioRad ImageLab was to quantify intensity of staining. Each band was normalised to its corresponding dsDNA control. Fold change relative to vehicle treated cells was calculated by dividing this normalised value by the mean value of vehicle treated cells.

2.2.10 SDS-PAGE and Western blot

Cells were washed with ice-cold PBS, harvested by scraping and pelleted by centrifugation (200 *g* at 4 °C for 5 minutes). The supernatant was removed and the cell pellet either snap frozen and stored at -80 °C or taken directly for lysis. The cell pellet was lysed in 100 µL lysis buffer (20mM HEPES pH 7.4, 80mM NaCl, 2mM $MgCl_2$, triton X-100 1%) on ice for 20 minutes with mixing by vortex at 5 minute intervals. The lysate was pelleted by centrifuge (20000 *g* at 4 °C for 15 minutes) and the supernatant transferred to a new tube. The protein content was determined by Bradford assay.

A 4-20% gradient gel was poured into a BioRad gel cassette following the protocol of Miller et al (2016). Equal volumes of low and high percentage gel mix were made (**table 2.1** details the reagents used for each percentage gel). Three millilitres of low percentage gel were aspirated into a 5 ml stripette followed by 3 ml of high percentage gel. A 0.5 ml air bubble was then aspirated to mix the interface and the gel mix pipetted into the gel cassette. The gel was levelled by adding 500 µl isopropanol and allowed to polymerise for 20 minutes. After removing isopropanol, a 4% stacking gel (680 µl 30% acrylamide, 500 µl 1M tris pH 6.8, 3.736 ml ddH_2O, 40 µl 10% SDS, 40 µl 10% APS and 10 µl TEMED) was poured on top of the gradient gel and a gel comb added. The BioRad cassette was placed into the gel tank and 1x running buffer was added. Protein samples were diluted to 20 µg in 24 µl and 6 µl of 5x protein loading buffer added. The diluted samples were heated to 95 °C for 5 minutes. The 30 µl volume was pipetted into the well. Three microlitres of protein ladder (BioRad) was loaded into one well and remaining empty wells were loaded with 30 µl of 1x protein loading buffer. The gel was run at 100 V until samples were aligned at the interface of the stacking and resolving gels (10-15 minutes). The voltage was increased to 170 V and the gel run for a further 60-90 minutes.

Table 2.1 Reagent volumes for 4/20% gradient gel

	4% gel	**20% gel**
ddH_2O	1.79 ml	192 µL
30% Acrylamide	400 µL	2 ml
1.5 M Tris pH 8.8	750 µL	750 µL
10% SDS	30 µL	30 µL
10% APS	30 µL	30 µL
TEMED	2.4 µL	1.2 µL

Following PAGE the gel was removed and a semi-dry transfer performed using the BioRad Trans-Blot® Turbo™ Transfer System. The resulting nitrocellulose membrane was blocked in 5% milk (0.5g milk powder in 10 ml 1x TBST) at room temperature for 1 hour in a 50 ml tube with gentle rotation. The membrane was then incubated with primary antibody at 4 °C overnight with gentle rotation. The list of antibodies and their concentrations is given in appendix 7. The membrane was washed three times with 1x TBST for 5 minutes per wash. The membrane was incubated with a secondary antibody (IgG-HRP conjugate; 1:4000) corresponding to the species of the primary antibody for 1 hour at room temperature with gentle rotation. Finally, the membrane was washed three times with 1 x TBST and incubated with electrochemiluminescence (ECL) solution for 1 minute in the dark. The membrane was imaged on a BioRad ChemiDoc MP. The protein ladder was imaged using colorimetric mode. Protein bands were imaged using high sensitivity mode. Exposure times were set automatically by the ImageLab software. Files were exported in .tiff format for analysis.

2.2.11 Agarose gel electrophoresis

A 1% agarose gel was used unless otherwise stated. One gram of agarose was dissolved in 100 ml of TBE by heating. When cooled, 2 µL of ethidium bromide was added for a final concentration of 0.2 ug/ml. The gel was poured and allowed to set. Samples were prepared for loading by mixing 5.5 µL of sample with 1.1 µL of 6x gel loading dye. Six microlitres of sample were loaded onto the gel and electrophoresis performed at 80V for 90 minutes. The gel was imaged on a BioRad ChemiDoc MP using the ‘Ethidium Bromide’ setting.

2.2.12 Preparation of chemically competent DH5α *E. Coli*

DH5α *E. Coli* were streaked onto a plain LB plate and incubated at 37 °C overnight. The next morning one colony was inoculated into 5 ml LB and incubated overnight at 37 °C with 225 RPM shaking. One ml of this starter culture was added to 200 ml LB and incubated at 37 °C with 225 RPM shaking until the OD_{600} was 0.4. The culture was chilled on ice for 10 minutes then centrifuged (15 minutes, 200 RPM, 4 °C). The supernatant was discarded and the pellet resuspended in 66.6 ml of buffer 1 (10 mM RbCl, 50 mM $MnCl_2$.4H2O, 30 mM KOAc, 10 mM $CaCl_2$, 15% v/v glycerol, pH 5.8). This was incubated on ice for 45 minutes then centrifuged (15 minutes, 200 RPM, 4 °C). The supernatant was discarded and the pellet resuspended in 8 ml buffer 2 (10 mM RbCl, 10 mM MOPS, 75 mM $CaCl_2$, 15% v/v glycerol, pH 6.8) then incubated on ice for 15 minutes. The cells were then aliquoted, snap frozen in liquid nitrogen and stored at -80 °C.

2.2.13 Bacterial transformation and culture

One microlitre of plasmid DNA was added to 50 µL of chemically competent DH5α *E. Coli* and incubated for 30 minutes on ice. Heat shock was done at 42 °C for 45 seconds then the cells returned to ice for 2 minutes. The cells were resuspended in 500 µL SOC media and incubated at 37 °C for 45 minutes with 225 RPM in a shaking incubator. The cell suspension was centrifuged at 4000 RPM for 2 minutes, supernatant aspirated and the pellet resuspended in a total volume of 100 µL SOC media. This was divided equally between two LB agar plates with appropriate antibiotic and incubated at 37 °C overnight. Plasmids from AddGene (MA, USA) arrived as glycerol stabs. These were streaked onto LB agar plates and incubated at 37 °C overnight. The next morning individual colonies were picked for culture using a sterile pipette tip. Single colonies were inoculated into appropriate media and incubated in 5 ml of Luria broth for 8 hours at 37 °C with 225 RPM shaking. Five hundred microlitres of this starter culture were inoculated into 50 ml of Luria broth and incubated overnight at 37 °C with 225 RPM shaking. The following morning bacteria were pelleted by centrifugation (5000 *g*; 15 minutes; 4 °C) then either snap frozen in liquid nitrogen or processed directly by midi prep. Aseptic technique was used throughout.

2.2.14 Plasmid midi-prep

Plasmids were extracted and purified from bacterial culture using a midi-prep kit (Qiagen) following the manufacturer's instructions with one minor modification: The silica spin column was incubated with elution buffer for 10 minutes to improve plasmid yield.

2.2.15 Plasmid and siRNA Transfections

A ratio of 1:2 for the mass of DNA in ug: lipofectamine 2000 in µL was used for plasmid transfections. For 6 cm plates, 6 ug of plasmid to 12 µL lipofectamine 2000 were diluted in separate tubes containing 250 µL Optimem serum free media. These were incubated at room temperature for 5 minutes then combined and mixed well. The plasmid/lipofectamine mix was incubated at room temperature for a further 30 minutes to allow transfection complexes to form. These were then added drop wise to cells at 70-80% confluency. Reactions were scaled up or down according to the surface area of culture vessel.

For siRNA transfections, onTarget siRNA pools (Horizon, UK) were used. These comprise 4 siRNA directed against the target of interest. Sequences are provided in appendix 3. A corresponding non-targeting siRNA (siNT) was used as a control. All siRNA were used at a final concentration of 30 nM. For a single well of a six well plate 3 µL of siRNA and 5 µL of lipofectamine 2000 were each separately diluted in 125 µL Optimem serum-free media. After 5 minutes at room temperature these two solutions were combined and incubated at room temperature for 30 minutes to allow transfection complexes to form. This was added drop wise to the cells to a final volume of 1.5 ml. Reactions were scaled up or down according to the size of the culture vessel. Plasmid maps are provided in appendix 2.

2.2.16 S9.6 immunoprecipitation

These experiments were based on the protocols of Cristini *et al* (2018) and Pinter *et al* (2021). Cells from two 15 cm plates were washed with 15 ml ice cold PBS, scraped in 10 ml ice cold PBS and pelleted by centrifuge (200 *g*; 5 minutes; 4 °C). Each pellet was resuspended in 1 ml of lysis buffer (85 mM KCl, 5 mM HEPES pH 8, 0.5% NP-40 with 1x protease inhibitor) and split into three tubes of equal volume. Cells were lysed on ice for 15 minutes with intermittent inversion to mix. Nuclei were pelleted by centrifuge (15,000 *g;* 2 minutes; 4 °C) and the supernatant saved as the cytoplasmic fraction. Each pellet was resuspended in 250 µL resuspension buffer (RSB: 10 mM Tris-Hcl pH 7.5, 200 mM NaCl, 2.5 mM MgCl2) with 0.2 % sodium deoxycholate, 0.1% SDS, 0.05% sodium lauroyl-sarcosinate, 0.5% Triton X-100 and 1x protease inhibitor. Each lysate was transferred to a separate sonication tube and sonicated for 3 cycles (30 seconds on; 30 seconds off) on a Bioruptor Pico sonicator (Diagenode). The sonicated lysates were recombined, and the protein concentration measured by Bradford assay.

Lysate corresponding to 750 µg of protein was diluted in a total volume of 3 ml RSB with 0.5% Triton X-100 (RSBT). An aliquot corresponding to 1 % input was taken from the lysate and stored at -20 °C. The remaining lysate was aliquoted for 1) agarose gel electrophoresis to check DNA fragment size and sonication efficiency and 2) SDS-PAGE and Western blot to confirm successful fractionation of nucleus and cytoplasm.

The lysate was pre-cleared by incubation with 1 µg mouse IgG and 35 µL protein G Dynabeads. The sample was then split into three equal volumes and, after magnetic separation, the supernatant was either treated with RNase H1 overnight at 37 °C (RNase H treated control to test antibody specificity) or simply incubated at 37 °C overnight. The samples were then incubated with 10 ng RNase A for 1 hour at 37 °C. Ten micrograms S9.6 or 10 µg IgG were added and incubated at 4 °C with gentle rotation. Simultaneously, 35 µL protein G Dynabeads per sample were pre-conditioned by incubation with RSBT for 1 hour at 4 °C with gentle rotation. These were added to the three samples and incubated for a further 2 hours at 4 °C.

Following immunoprecipitation, the beads were washed four times with RSBT and two times with RSB. Proteins were eluted by resuspending the beads in 35 µL protein loading buffer and heating to 70 °C for 10 minutes with intermittent mixing by vortex. The eluate was snap frozen and stored at -80 °C. SDS-PAGE and Western blot were done as per section 2.2.10.

2.3 Bioinformatics

All software, packages and libraries are listed in appendix 5.

2.3.1 Access to Ensembl data

The biomaRt R package (Durinck *et al.*, 2005) was used to access genomic coordinates and gene metadata/ annotations using gene-level Ensembl identifiers.

2.3.2 Calculation of GC skew

Nucleotide sequences for each gene of interest were retrieved using the BSGenome package in R (Pagès, 2022). A 200 bp sliding window was moved along each nucleotide sequence and the frequency of each nucleotide recorded using functions from the Biostrings package in R (Pagès *et al.*, 2020). GC skew for each window was calculated as described previously (Ginno *et al.*, 2013b; Hartono, Korf and Chédin, 2015):

$$\frac{G-C}{G+C}$$

The GC skew was plotted at two genomics features: promoters (defined as 2000 bp upstream to 1000 bp downstream of the TSS) and terminators (2000 bp either side of the TTS).

2.3.3 Prediction and analysis of R-loop forming sequences

The Quantitative Model of R-loop Forming Sequences (qmRLFS) tool was developed to predict R-loop forming sequences based on sequence context. Based on the biochemical property of G-clustering described in chapter 1, qmRLFS annotates a nucleotide sequence as an RLFS if an R-loop initiation zone (RIZ) and an R-loop extension zone (REZ) are present (**figure 2.1**). This tool has been used to generate an *in-silico* genome-wide dataset of R-loop forming sequences (Wongsurawat *et al.*, 2012; Kuznetsov *et al.*, 2018) and has been validated against in-vivo R-loop mapping techniques such as DRIP and R-ChIP. Subsequent studies used this map of RLFS to design and validate in-vivo experiments where qPCR primers were designed against predicted RLFS for use in DRIP-qPCR experiments (Li *et al.*, 2015; Lambo *et al.*, 2019; Zhang *et al.*, 2020). The authors of qmRLFS have made a partial version of the *in-silico* dataset available on their website (http://rloop.bii.astar.edu.sg/?pg2=stats). This version suffers from incomplete coverage of the genome and is based on hg19 reference genome coordinates. To overcome these limitations, the qmRLFS tool was downloaded and used to compile a genome wide BED file of RLFS for the hg38 reference genome. FASTA files of chromosomes 1-22, X, Y and the mitochondrial chromosome from the hg38 reference genome were retrieved from ftp://hgdownload.soe.ucsc.edu/goldenPath/hg38/chromosomes/ on 23rd March 2020. The FASTA files were processed by the qmRLFS tool in a Linux/Bash environment using m1 and m2 model parameters. Unique predicted RLFS for both DNA strands were stored as a BED file. All subsequent steps were performed in R.

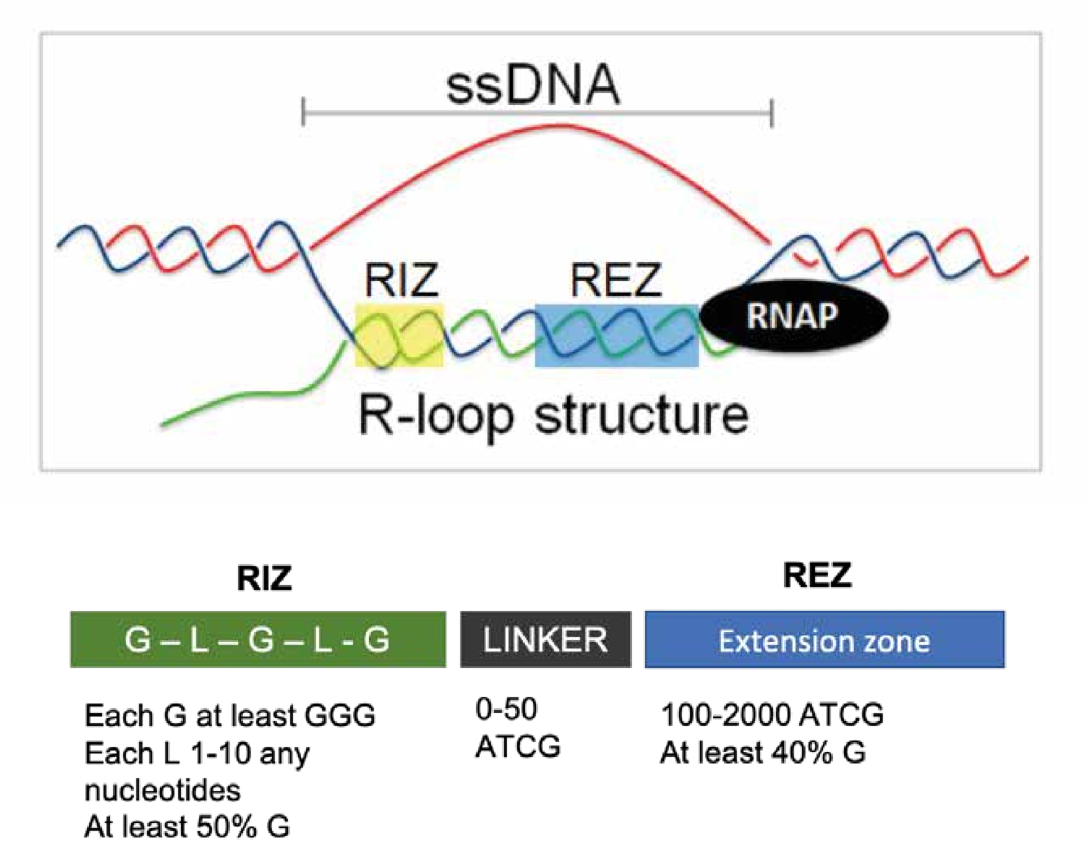


Figure 2.1 Quantitative model of R-loop forming sequences model
RIZ: R-loop initiation zone, REZ: R-loop extension zone, RNAP: RNA polymerase
Adapted from Wongsuwarat et al (2012). Used with permission.

The BED file of all RLFS was imported into R and protein coding genes +/- a 2 kb 5' and 3' flank that overlapped an RLFS by at least 1 bp were counted using functions from the GenomicRanges Bioconductor package. Overlaps had to be in a co-transcriptional orientation to the gene to be counted. All gene information (e.g. genomic co-ordinates, % GC content, HGNC gene symbol) were acquired from Ensembl via the biomaRt package using Ensembl version 100). The fitdistr package was used to assess Poisson and negative binomial distributions of the number of RLFS per gene. The MASS package was used for negative binomial regression when comparing two groups of differentially expressed genes.

2.3.4 Acquisition of publicly available sequencing data

Publicly available sequencing data was downloaded from the gene expression omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/). Where available, pre-processed gene-level count data was used. For LNCaP and VCaP GRO-seq datasets the FASTQ files were downloaded from the sequence read archive (SRA; https://www.ncbi.nlm.nih.gov/sra) using fastq-dump from the SRA toolkit. Publicly available datasets used in this work are listed in appendix 4.

2.3.5 GRO-seq data processing and analysis

Raw read quality was assessed using FASTQC and BWA was used for alignment to GRCh38.p13 reference genome. FeatureCounts was used to count reads within a custom gene model comprising the genomic coordinates of all annotated genes with the 5' 1kb removed to avoid spuriously counting paused RNA polymerase 2.

2.3.6 TCGA data processing

Data from TCGA was obtained using the TCGA Biolinks package in R (Colaprico *et al.*, 2016).

2.3.7 DRIP-seq data processing and analysis

FASTQ files of paired end reads were retrieved from Novogene (Cambridge, UK) following sequencing. Quality control of raw sequencing data was done with FastQC (Andrews, 2010). Adapter sequence removal and quality trimming were done with cutadapt (Martin, 2011). The reads were then aligned to the GRCh38.p13 using BWA mem and duplicates removed using samtools (Li *et al.*, 2009). MACS2 was used to call peaks for all samples with input DNA used as background. Default arguments were used in BAM paired end mode with broad peak calling.

Reproducibility of DRIP-seq signal was compared between and within groups of replicates as described previously (Stork *et al.*, 2016). Differential peak analysis was performed using DiffBind (Ross-Innes *et al.*, 2012) and a consensus peakset of peaks present in at least 2 of 3 replicates from each condition. An FDR threshold of < 0.05 and fold change of +/- 2 were used as cut offs to call significant enrichment between conditions. The default count normalisation method was used, duplicate reads were ignored and paired end mode was used.

ChIPpeakAnno was used to annotate significantly enriched peaks using Ensembl hg38 gene models (version 100). Metagene plots were created using DeepTools.

2.3.8 RNA-seq data processing and analysis

RNA-seq data from large clinical cohorts (e.g. TCGA) were analysed using the approach described by Li *et al.* (2022) owing to the high false positive rate of DESeq2 in this context. Counts were normalised using edgeR to give each gene from each sample a counts per million (CPM) value. CPM were compared between conditions using the Wilcoxon rank sum test then the false discovery rate controlled by the Benjamini-Hochberg method. A list of 1187 stromal genes (Marzec *et al.*, 2021) were excluded from the analysis to avoid confounding the differential expression analysis between prostate-confined and metastatic tumours.

For the RNA-seq data in chapter 5, FASTQ files from strand specific RNA-seq were retrieved from Novogene (Cambridge, UK). After quality control with fastQC, reads were pseudoaligned to the transcriptome using Kallisto. These pseudoalignments were imported into R using tximport. For differential expression analysis in cell culture models, DESeq2 was used (Love, Huber and Anders, 2014).

Chapter 3: *In*-silico Analysis of R-loop Forming Sequences Across the Androgen Signalling Landscape of Prostate Cancer

3.1 Introduction

In chapter 1 I described the genomic and transcriptional conditions that favour R-loop formation. Nucleotide sequence composition, in particular the guanine content of the coding DNA strand plays a crucial role in the establishment of R-loops (Ginno *et al.*, 2012, 2013b; Hartono, Korf and Chédin, 2015). These nucleotide characteristics were incorporated into a predictive model, Quantitative Model of R-loop Forming Sequences (QmRLFS) by Wongsurawat *et al.* (2012). The RLFS model uses characteristics of the non-template/coding DNA sequence to predict genomic regions with a tendency to form R-loops, principally loci with G clustering and a G-rich sequence (Roy and Lieber, 2009). The model comprises an R-loop initiation zone (RIZ) and R-loop elongation zone (REZ), both of which must be present to ensure stable R-loop formation. QmRLFS is implemented in Python and takes FASTA nucleotide sequence files as input then scans the nucleotide sequence for features that favour R-loop formation. Sequences that fulfil criteria for R-loop formation are recorded as an R-loop forming sequence. In later work using the same model this group showed that protein coding genes had a greater number of RLFS than pseudogenes (Kuznetsov *et al.*, 2018). This implies that genes with RLFS are under selective pressure and adds weight to the idea that R-loops are part of the expression regulation mechanism for many protein coding genes.

The QmRLFS tool has been validated against DRIP-seq datasets and showed a sensitivity and specificity of 90% and 72% respectively when compared with genome-wide mapping techniques such as DRIP-seq (Kuznetsov *et al.*, 2018). In this study the authors also predicted R-loops using the qmRLFS model and validated these predictions for specific loci using DRIP-qPCR. Furthermore qmRLFS has been implemented as a web server which can be queried through a web browser (Jenjaroenpun *et al.*, 2015). Subsequent studies have used predicted RLFS to successfully design qPCR primers for individual R-loop loci (Li *et al.*, 2015;

Lambo *et al.*, 2019; Zhang *et al.*, 2020). However, no studies have asked whether differentially expressed genes might have different levels of RLFS.

3.1.1 Hypothesis

In this chapter I tested the hypothesis that the nucleotide sequence context of androgen regulated genes might be associated with their expression. I reasoned that if R-loops were involved in regulation of gene expression then the presence of R-loop forming sequences would be different between up and down regulated genes and/or between differentially expressed genes and a random background null distribution drawn from all expressed genes. As discussed in chapter one androgen signalling can up and down regulate gene expression. In this chapter I asked whether there were differences in the R-loop forming potential of genes between genes up and down regulated by androgens.

Objectives

- Develop a per-gene measures of RLFS to allow comparison of RLFS density between groups of differentially expressed genes
- Use prostate cancer datasets from cell culture and clinical studies to ask if RLFS differ in the context of androgen activity

3.2 Results

3.2.1 Comparison of nucleotide sequence characteristics of genes differentially regulated by androgens in cell culture models reveals differences in GC content and GC skew

The nucleotide sequence of a gene plays a central role in regulating that gene's expression. The proportion of bases in a gene that are guanine or cytosine (percentage GC content) is particularly important. Analysis of ~20,000 RNA-seq datasets by Zrimec et al (2020) showed that the GC content was higher in the promoter and termination regions of expressed genes compared to 5' and 3' UTR implying roles in transcriptional regulation. This corresponds to well-described CpG islands around the TSS of mammalian genes which facilitate transcription by favouring transcription factor binding and chromatin accessibility (Deaton and Bird, 2011). To date, to my knowledge, a possible relationship between GC content and the androgenic regulation of gene expression in prostate cancer has not been investigated. I selected publicly available GRO-seq datasets from LNCaP (Wang *et al.*, 2011) and VCaP (Toropainen *et al.*, 2016) cells exposed to 100 nM DHT for 2 hours. GRO-seq measures nascently transcribed RNA. I chose this timeframe and the measurement of nascent rather than processed mature RNA so the datasets would reflect genes that were truly androgen regulated and limit contamination by genes whose expression is a result of downstream signalling after the initial response to androgens.

The LNCaP cell line is androgen responsive but has a normal androgen receptor copy number (Horoszewicz *et al.*, 1983). The VCaP cell line is a model of advanced prostate cancer derived from a vertebral metastasis (Korenchuk *et al.*, 2002). The cell line harbours an amplification of the AR locus and is responsive to treatment with androgens. In the VCaP dataset, 1575 genes were upregulated by DHT and 1519 genes were downregulated. These genes had a higher percentage GC content compared to upregulated genes (48.6% vs. 42.6% in upregulated genes; [$p < 2.2 \times 10^{-16}$; Wilcoxon rank-sum test]; **Figure 3.1A and 3.1B**). To investigate this in more detail I examined GC skew in the coding/non-template strand of each of the groups of genes. GC skew is linked with R-loop formation and consequently gene regulation by modulating chromatin structure, CpG methylation and transcriptional dynamics at the promoter and terminator regions. These characteristics were discussed fully in the introduction and methods chapters.

The GC skew plots presented in **figure 3.1C and D** allow a visual comparison of GC skew at promoter and terminator regions. Both up and down regulated genes showed an increase in GC skew starting approximately 250 bp upstream of and peaking at the transcription start site as described previously (Ginno *et al.*, 2013a). Within terminator regions a peak of GC skew was observed around the transcription termination site as described previously (Ginno *et al.*, 2013a; Skourti-Stathaki, Kamieniarz-Gdula and Proudfoot, 2014). No qualitative differences in the GC skew profile were seen between up and down regulated genes.

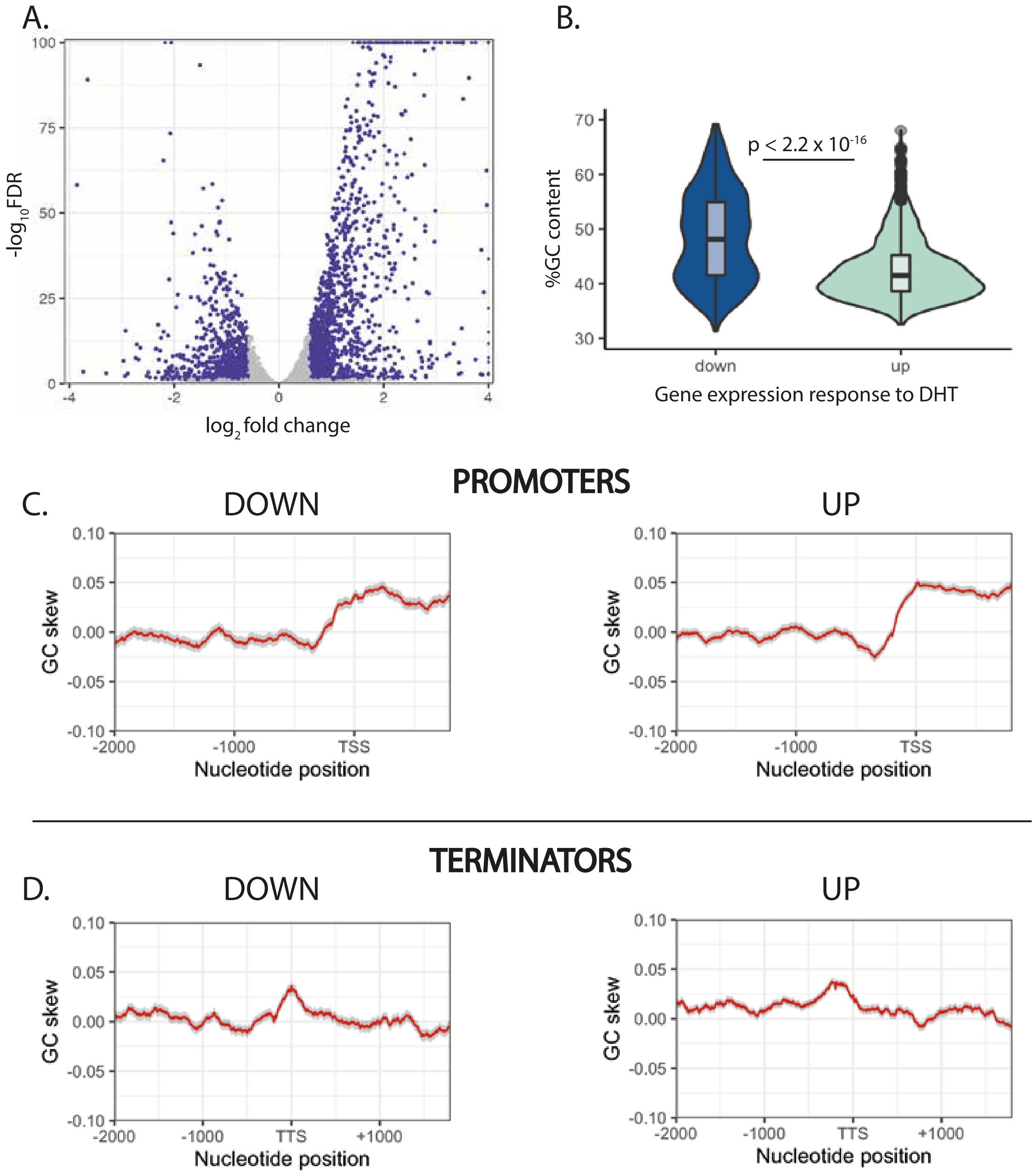


Figure 3.1 GC content and GC skew of differentially expressed genes in VCaP cells treated with 100 nM DHT for 2 hours

A. Volcano plot of differentially nascently transcribed RNA from GRO-seq. Purple dots represent genes with FDR < 0.05 and absolute $\log_2$ fold change > 0.585. B. Violin plot of %GC content of genes up and downregulated by DHT. P value from Wilcoxon rank-sum test. C. GC skew in promoter regions of up and down regulated genes. Red line indicates mean skew at each position; grey ribbon is standard error of the mean. D. GC skew in transcription termination regions of up and down regulated genes. TSS: Transcription start site, TTS: Transcription termination site.

The same analysis was performed with GRO-seq data from the LNCaP model. After two hours' DHT exposure there were fewer differentially expressed genes compared to VCaP cells (421 vs. 1713, **figure 3.2A**). A comparison of differentially expressed genes in VCaP and LNCaP cells is given in **table 3.1**. In LNCaP cells upregulated genes had a significantly higher percentage GC content compared to down regulated genes (42.5% vs. 40.7%, upregulated vs. down regulated genes; [p = 0.003; Wilcoxon rank-sum test], **figure 3.2B**). Like VCaP cells, the GC skew profiles of differentially expressed genes did not differ between up and down regulated genes at the promoters of terminators (**figure 3.2C and D)**

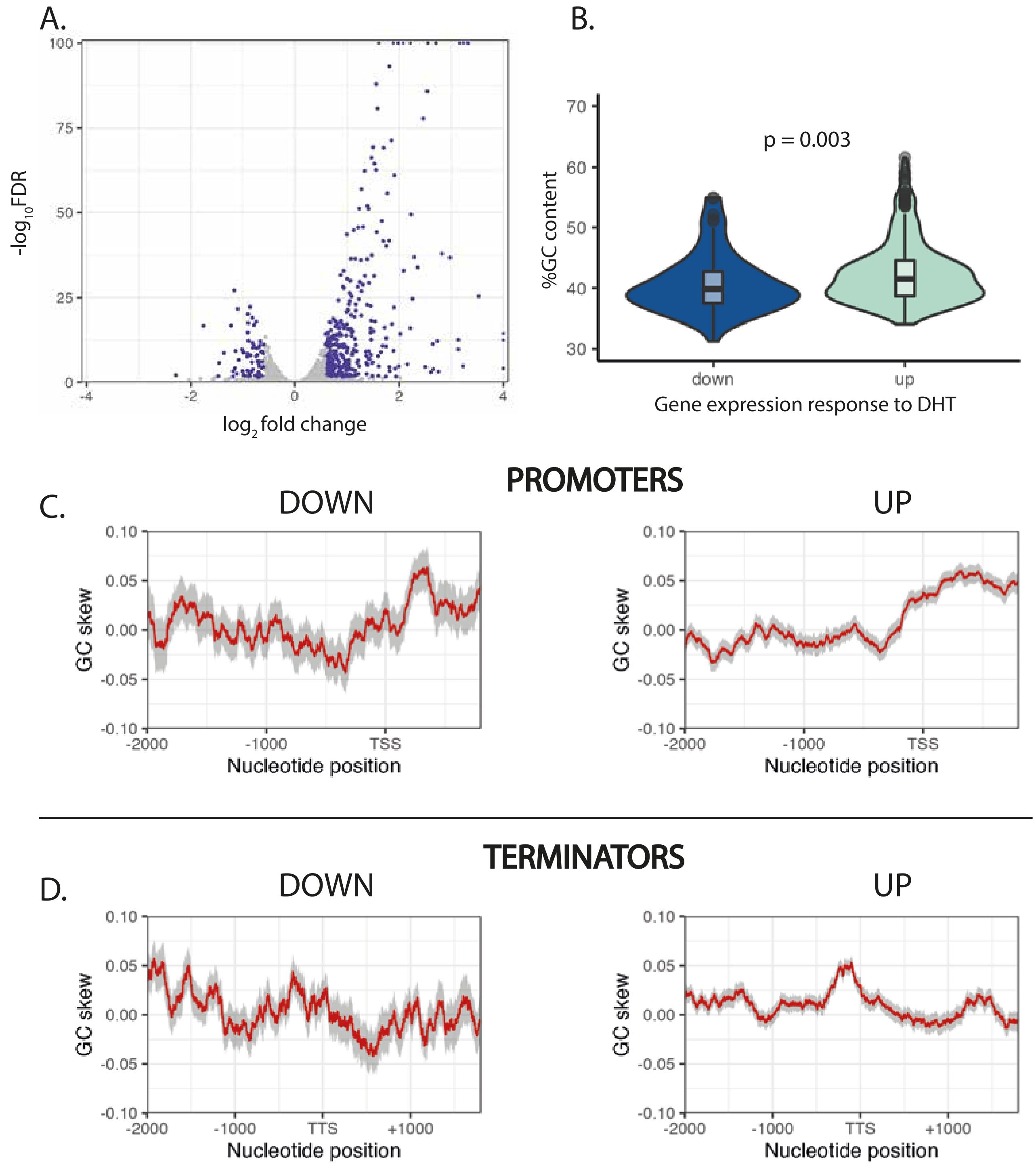


Figure 3.2 GC content and GC skew of differentially expressed genes in LNCaP cells treated with 100 nM DHT for 2 hours

A. Volcano plot of differentially nascently transcribed RNA from GRO-seq. Purple dots represent genes with FDR < 0.05 and absolute $\log_2$ fold change > 0.585. B. Violin plot of %GC content of genes up and downregulated by DHT. P value from Wilcoxon rank-sum test. C. GC skew in promoter regions of up and down regulated genes. Red line indicates mean skew at each position; grey ribbon is standard error of the mean. D. GC skew in transcription termination regions of up and down regulated genes. TSS: Transcription start site, TTS: Transcription termination site.

Table 3.1 Top ten up and down regulated genes in LNCaP and VCaP cells in response to DHT. Genes differentially regulated in both cell lines are highlighted in bold.

LNCaP – Upregulated genes		**VCaP – Upregulated genes**	
Gene name	Fold change	Gene name	Fold change
CHRNA2	5.2	ORM2	6.1
TTN	4.3	**NPPC**	5.9
NPPC	4.2	RDH10	5.9
SLC26A3	4.2	AGR3	5.6
PLA2G5	3.5	STEAP4	5.6
TMPRSS2	3.3	**FKBP5**	4.9
KLK2	3.3	ADAMTS8	4.9
FKBP5	3.2	KRT73	4.8
STEAP4	3.2	MME	4.7
NKX3-1	3.1	SLC2A5	4.6
LNCaP – Downregulated genes		**VCaP – Downregulated genes**	
GLP1R	-2.2	CXCR6	-4.6
FGD5	-2.0	TRIB1	-4.1
TFCP2	-1.8	IHH	-3.8
PLD1	-1.7	SLC3A1	-3.7
SLC6A12	-1.5	FOLH1	-3.6
LRRC31	-1.5	IL1F10	-3.3
PGR	-1.5	ASZ1	-2.9
COL3A1	-1.4	FAM240A	-2.9
DAPK1	-1.4	GCOM1	-2.9
MAN1A1	-1.3	FZD10	-2.8

Together these results suggest that genes differentially expressed in response to androgen have different % GC content but the pattern of %GC content is different depending on the signalling context. The androgen receptor amplification of VCaP cells manifested as higher fold changes in the most significantly differentially expressed genes (both up and down regulated) compared to LNCaP cells indicating increased AR activity. This raised an interesting question: Is the GC content and R-loop forming potential of genes regulated by high AR activity different from those regulated by normal AR activity and if so, could R-loops play a mechanistic role?

3.2.2 Development of gene-level RLFS metrics

I next asked if the differences in GC content between up- and down- regulated genes in LNCaP and VCaP cells corresponded with a difference of R-loop forming potential in those cell culture models. To facilitate this I developed three metrics based on the quantitative model of R-loop Forming Sequences (qmRLFS) tool:

1. Proportion of genes containing at least one RLFS
2. Number of RLFS per gene
3. Percent coverage by RLFS

3.2.2.1 Proportion of genes containing at least one RLFS

The qmRLFS was used to assign RLFS to genes in a strand specific manner (Wongsurawat *et al.*, 2012). The FASTA sequence of the human genome build hg38 was processed by the qmRLFS Python script. This generated 671357 individual RLFS across the whole genome. Merging RLFS that overlapped by at least 1 bp created a list of 231502 intervals. Of 19928 protein coding genes, 15734 (78.9%) contained at least one merged RLFS (mRLFS) that overlapped with its genomic co-ordinates and was co-directional with transcription. These genes were denoted as RLFS positive (81.5% of all protein coding genes); genes with no overlapping mRLFS were denoted mRLFS negative (18.5% of all protein coding genes; **Figure 3.3A**). Both RLFS positive and RLFS negative genes showed an increase in GC skew approximately 250 bp 5’ to the TTS as described previously (Ginno *et al.*, 2013b). However, RLFS positive genes had a higher maximum skew value (~0.05 vs. ~0.025) in keeping with the relationship between GC skew and R-loop formation (**Figure 3.3B and C**).

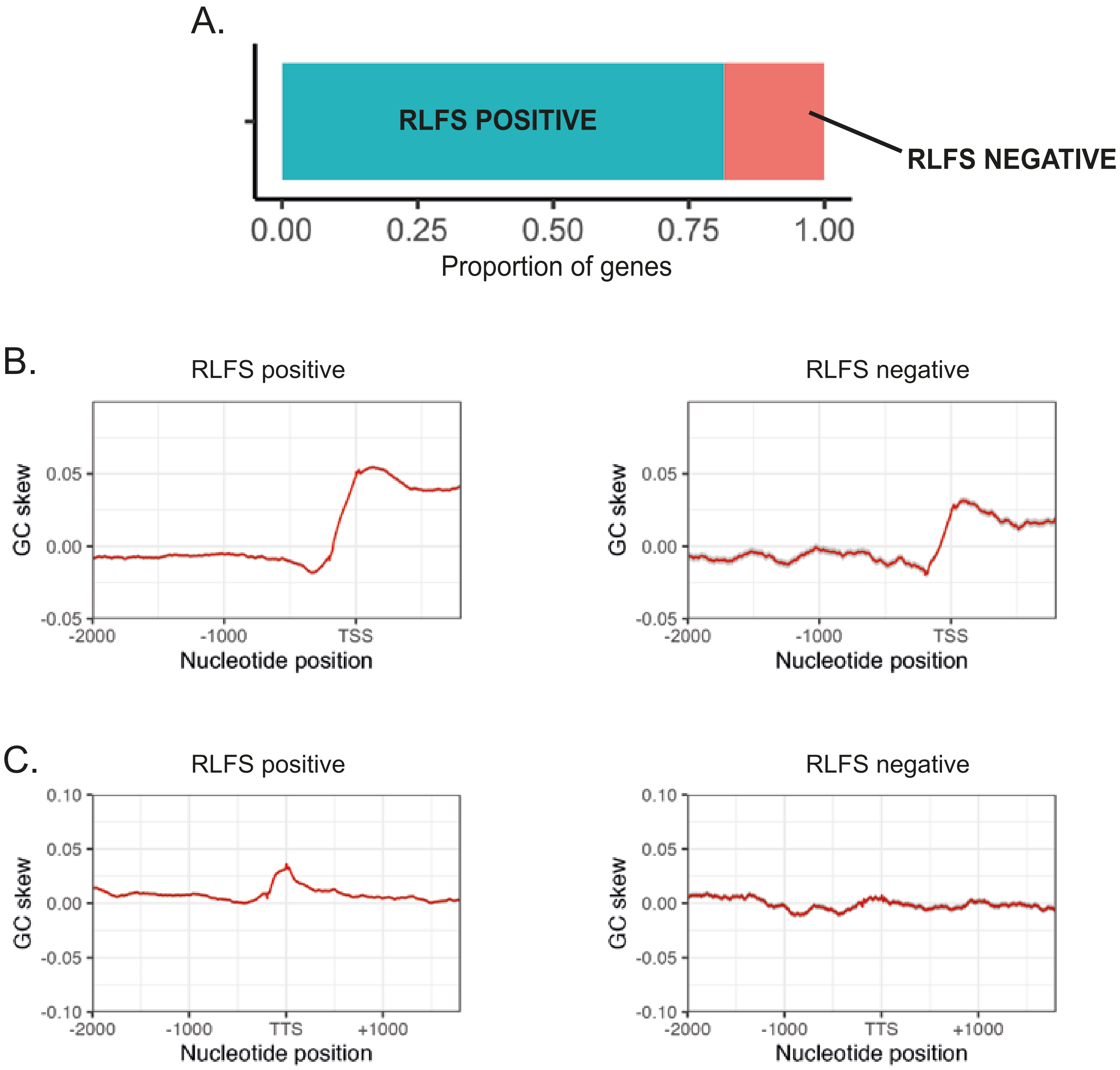


Figure 3.3 RLFS positive and negative protein coding genes have different GC skew profiles
A. Proportion of genes with at least one RLFS (green bar) and no RLFS (red bar) in all protein coding genes (n = 19680). B. and C. GC skew profiles for RLFS positive and RLFS negative protein coding genes in promoter and termination regions respectively.

3.2.2.2 The negative binomial distribution accurately models the number of RLFS per gene

The proportion of mRLFS positive genes in a dataset gives a general indication of whether a set of genes is more prone to form R-loops. However, this metric doesn't contain information about the propensity of the mRLFS positive genes to form R-loops. Two datasets could have identical proportions of mRLFS positive genes but one of these datasets could contain genes with low numbers of RLFS per gene whilst the second dataset could contain genes with high numbers of RLFS.

Examination of the distribution of mRLFS in protein coding genes showed that 95% had 14 mRLFS or fewer. The range was wide however: 62 genes had at least 50 mRLFS and 10 genes had at least 100 The number of mRLFS per gene is classified as count data – it can only take values that are non-negative integers. As such this metric should not be analysed by methods typical for continuous data such as the t test or Wilcoxon rank-sum test (Kirkwood and Sterne, 2011). I sought to find a regression model that I could use to assess the differences between two datasets (e.g. between up and down regulated genes). I initially evaluated the Poisson distribution as a model for mRLFS per gene across all genes. Visual inspection of the predicted Poisson distribution vs. the observed distribution suggested that this model over-estimated mRLFS per gene (**Figure 3.4A**). Furthermore, the variance was approximately 6 times the mean, implying overdispersion of this distribution. RNA-seq data is typically over dispersed and this has been successfully modelled using the negative binomial distribution in the DESeq2 differential expression analysis package (Love, Huber and Anders, 2014). The negative binomial distribution had a better fit for mRLFS per gene as demonstrated visually (**Figure 3.4B**) and through a lower Akaike Information Criterion, a comparative measure of how well a model fits a dataset (Dziak *et al.*, 2020). Negative binomial regression with the group membership (e.g. up or down regulated gene expression) as a model coefficient could be used to ask if the number of RLFS differ between two groups of genes. If the distribution of mRLFS in the two groups is significantly different then the group coefficient will have a low p-value. $p < 0.05$ was set as the significance level.

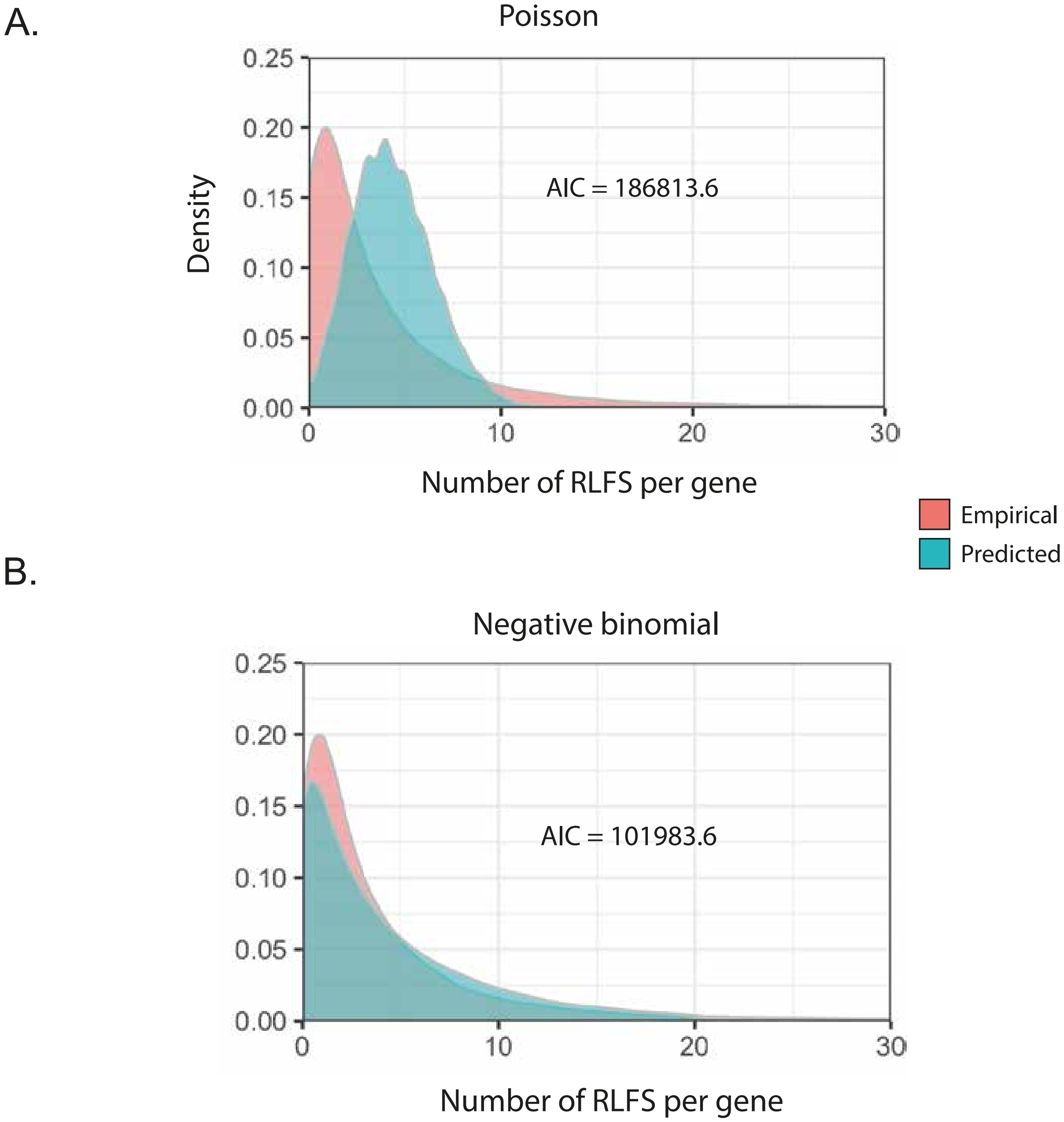


Figure 3.4 The negative binomial distribution effectively models the distribution of RLFS per gene
A. Poisson distribution fit to RLFS per gene for all protein coding genes. B. Negative binomial distribution fit to the same data. AIC: Akaike Information Criterion. Red curve ('empirical') is the known number of RLFS per gene for all protein coding genes. The green curve is the predicted distribution for each model.

3.2.2.3 Percent of gene coverage by RLFS

Gene length has previously been shown to influence R-loop accumulation and longer genes accumulated more R-loops in response to topoisomerase 1 knock down by siRNA (Manzo *et al.*, 2018) and in aged Drosophila melanogaster photoreceptor cells (Jauregui-Lozano *et al.*, 2022). In the mRLFS dataset, the number of mRLFS per gene and gene length were positively correlated (Spearman correlation coefficient = 0.16, $p < 2.2 \times 10^{-16}$; **Figure 3.5A**). To avoid the number of RLFS acting simply as a surrogate for gene length I defined a further RLFS density metric that accounted for gene length: percentage of gene length covered by mRLFS (fig 3.5B). This metric also allows the ranking of individual genes by their R-loop forming potential which could be useful for selecting candidate genes for further study including measuring locus specific R-loop abundance between conditions.

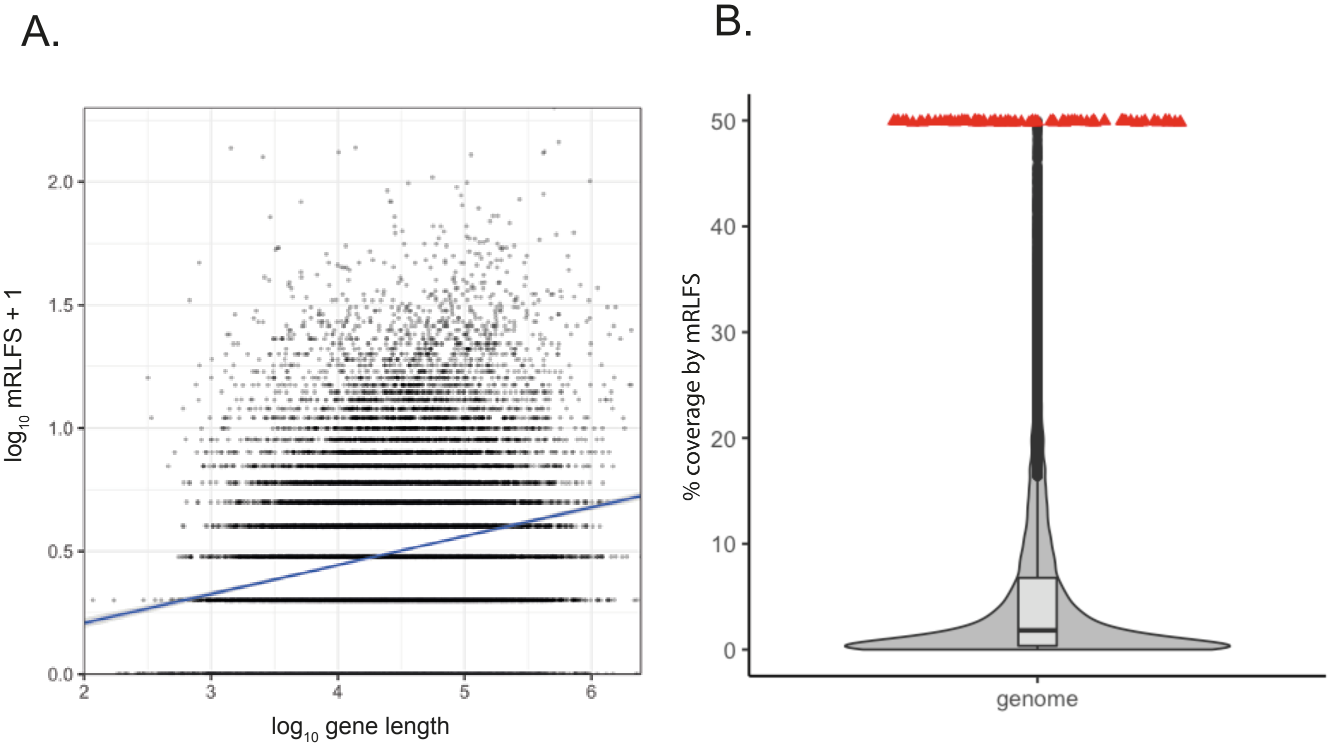


Figure 3.5 Percentage gene coverage by RLFS
A. Correlation between number of RLFS per gene and gene length. B. Violin plot of percentage gene coverage by RLFS for all protein coding genes. The y-axis has been censored at 50% to avoid a small number of extreme values obscuring the majority of the distribution. Genes with percentage RLFS coverage greater than 50% are indicated by red triangles.

3.2.3 Genes downregulated by DHT in VCaP cells have more RLFS than upregulated genes

I applied the RLFS metrics to the VCaP GRO-seq dataset. After 2 hours treatment with 100 nM DHT up and down regulated genes contained a similar proportion RLFS positive genes (**Figure 3.6A**). However down regulated genes had significantly more RLFS per gene and significantly higher percent coverage by RLFS (**Figure 3.6B and C**). To set this in the context of genes not differentially regulated by androgens I used a random sampling approach to generate a background null gene set. For each differentially expressed gene set an identical number of genes were selected at random from non-differentially expressed genes and labelled the background set. This set was combined with the differentially expressed gene set and the group labels (e.g. ‘up regulated’ and ‘background’) randomly shuffled. From this shuffled set three metrics were generated to represent population average measurements of each of the three RLFS metrics: Proportion of RLFS positive, the negative binomial coefficient from regression with the shuffled groups and median RLFS percent coverage. This procedure was repeated 1000 times to generate 1000 shuffled datasets. A histogram of each population average metric was plotted together with the original value from the differentially expressed gene set in question (visualised as a red vertical line in **figure 3.7**). This approach showed that in VCaP cells both up and down regulated genes had a greater proportion of RLFS genes compared to a null background, as indicated by the vertical red line towards the extreme right of the histogram (**Figure 3.7A**). By contrast, only down regulated genes differed significantly in their negative binomial coefficient distribution (**Figure 3.7B**). Both gene sets had markedly different RLFS percent coverage than the background null but with opposite direction: down regulated genes had a greater median RLFS percent coverage whereas this metric was much lower for up regulated genes (**Figure 3.7C**). Together this implies that in VCaP cells androgen regulated genes are generally more likely to have at least one RLFS. However, the number of RLFS and how much of the gene the RLFS cover varies between up and downregulated genes.

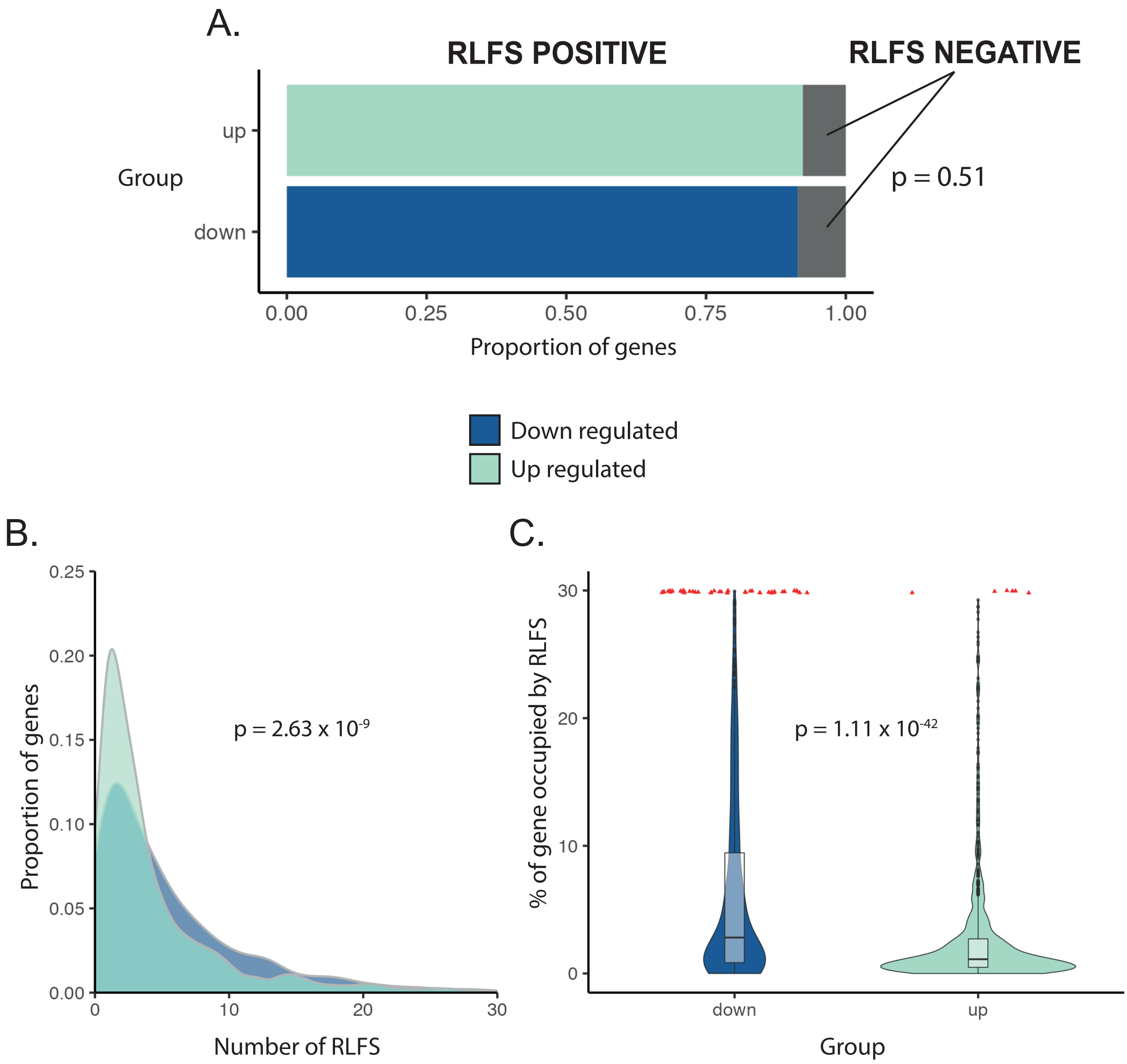


Figure 3.6 Genes down regulated by DHT in VCaP cells have more R-loop forming sequences
A. Proportion of up and down regulated genes that are RLFS positive. P value is from Chi squared test. B. Density plot of RLFS per gene in up and down regulated genes. P value is derived from the coefficient of the negative binomial regression model. C. Violin plot of RLFS percent coverage in up and down regulated genes. P value is from Wilcoxon rank-sum test. Red triangles indicate genes with greater than 30% RLFS coverage. RLFS: R-loop forming sequences.

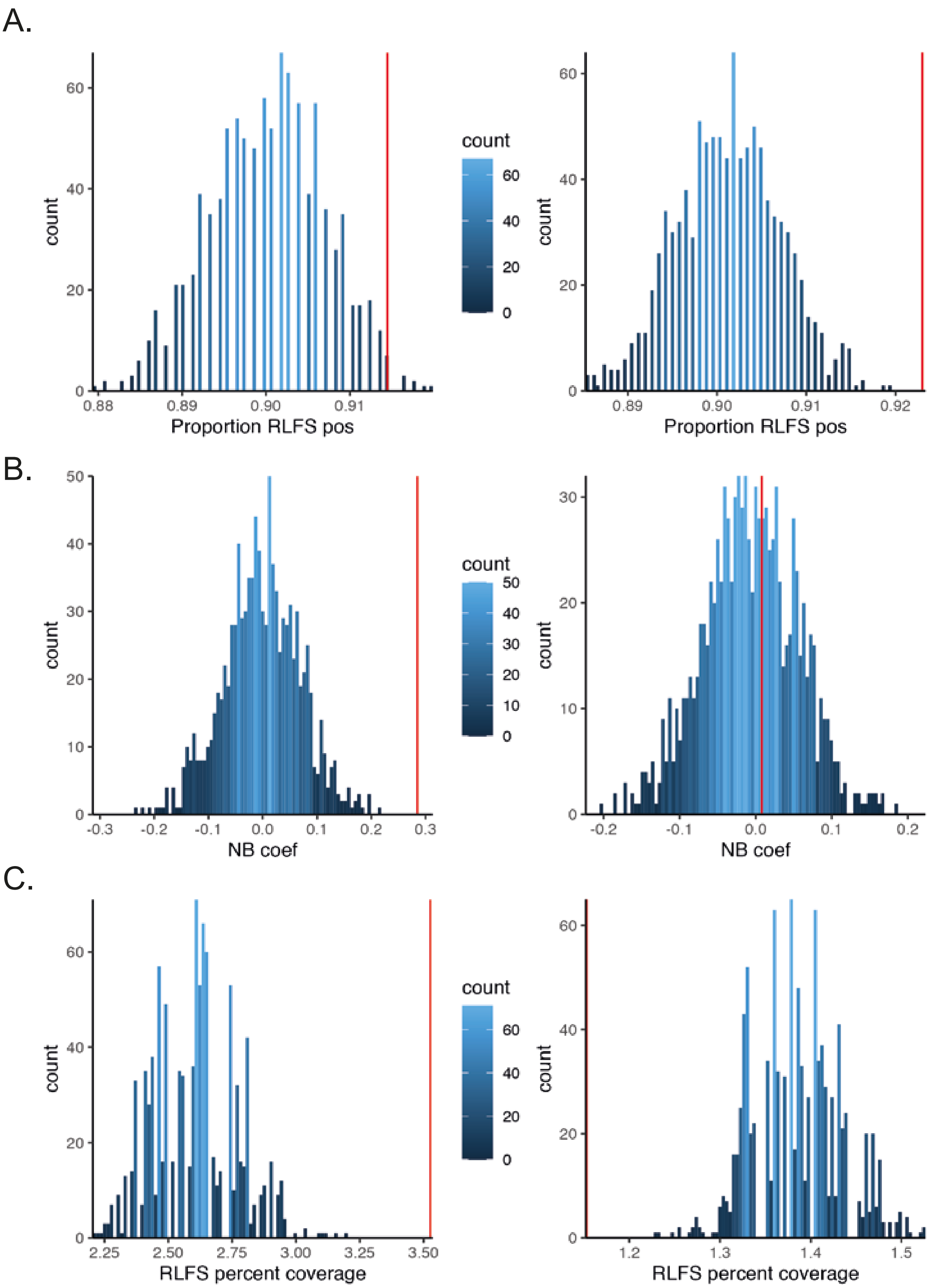


Figure 3.7 Comparison of RLFS metrics in differentially expressed genes and a background null gene set in VCaP cells

All plots are histograms of random sampling for each metric. A. Proportion of RLFS positive genes. B. Coefficient of negative binomial regression model with group membership shuffled for each sampling iteration. C. RLFS percent coverage. Down regulated genes are on the left, up regulated genes on the right. The red line in each plot indicates the value of that metric in the corresponding gene set.

3.2.4 Genes downregulated by DHT in LNCaP cells have fewer RLFS than upregulated genes

Applying the same metrics to the LNCaP differentially expressed gene set I found that the downregulated gene set had fewer RLFS positive genes but this difference was not significant (**Figure 3.8A**). Upregulated genes had more RLFS per gene and a higher percent coverage by RLFS but the magnitude of change was different compared to the pattern seen in VCaP cells (**Figure 3.8B and C**). Comparison with the background null distribution showed that for the proportion of RLFS positive and negative binomial coefficient metrics, down and up regulated genes were either side of the null values with downregulated genes typically to the left extreme on these plots (indicating lower values) and upregulated genes having higher values than the background null. Interestingly the median percent coverage by RLFS was lower than the null in both gene sets (**Figure 3.9A-C**). The data imply that whilst upregulated genes have more RLFS per gene this is driven partially by downregulated genes having fewer RLFS positive genes.

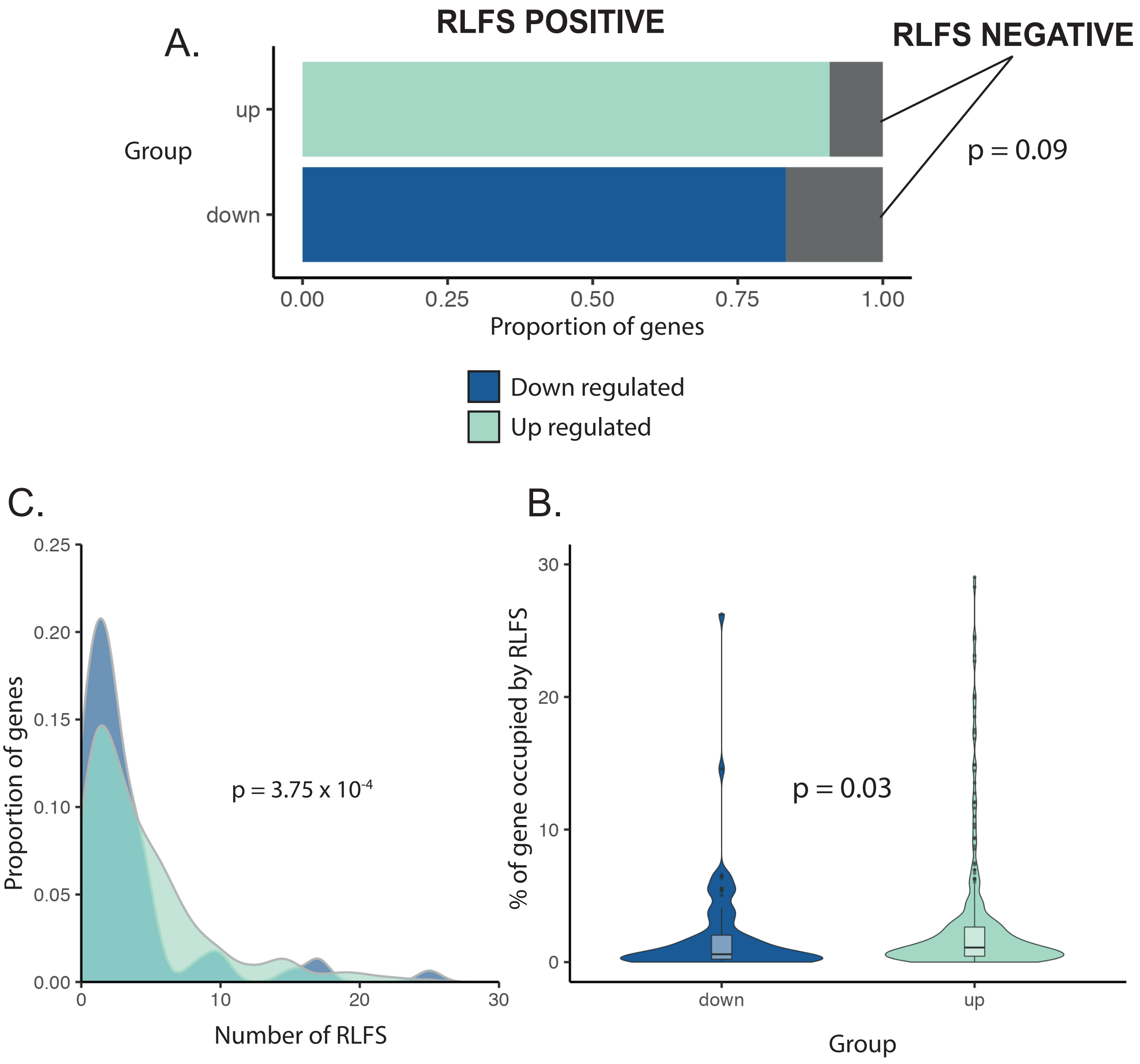


Figure 3.8 Genes up regulated by DHT in LNCaP cells have more R-loop forming sequences
A. Proportion of up and down regulated genes that are RLFS positive. P value is from Chi squared test. B. Density plot of RLFS per gene in up and down regulated genes. P value is derived from the coefficient of the negative binomial regression model. C. Violin plot of RLFS percent coverage in up and down regulated genes. P value is from Wilcoxon rank-sum test. RLFS: R-loop forming sequences.

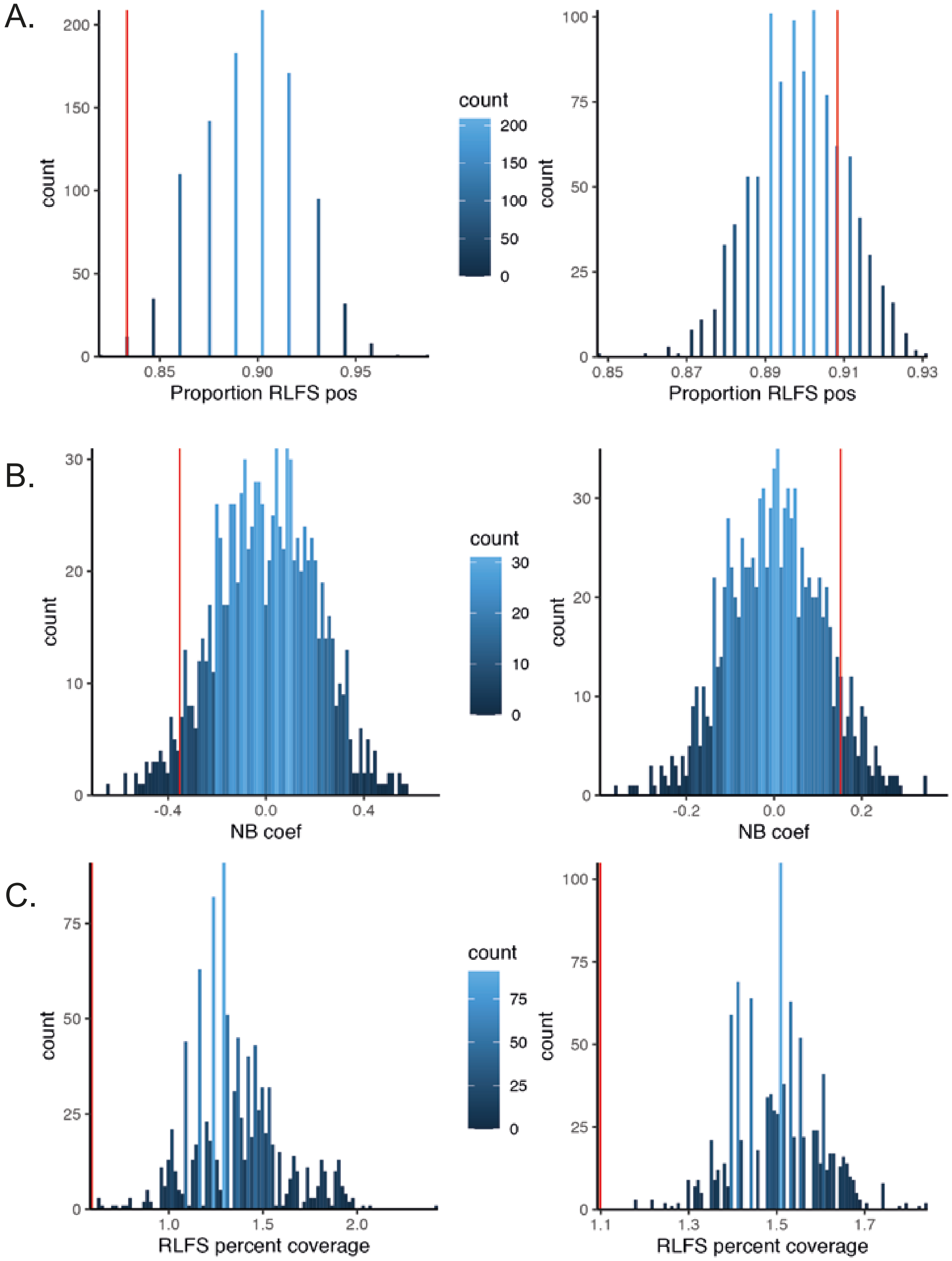


Figure 3.9 Comparison of RLFS metrics in differentially expressed genes and a background null gene set in LNCaP cells

All plots are histograms of random sampling for each metric. A. Proportion of RLFS positive genes. B. Coefficient of negative binomial regression model with group membership shuffled for each sampling iteration. C. RLFS percent coverage. Down regulated genes are on the left, up regulated genes on the right. The red line in each plot indicates the value of that metric in the corresponding gene set.

3.2.5 Genes downregulated by DHT in VCaP cells drive the differences in RLFS between LNCaP and VCaP cells

To resolve the differences in RLFS between VCaP and LNCaP datasets I asked which upregulated genes were shared between the two cell lines. Two hundred and forty-six genes were significantly upregulated in both cell lines; this group contained well-described androgen up regulated genes such as TMPRSS2, KLK2, KLK3, NKX3-1, MBOAT2 and FKBP5. I have denoted this group 'shared upregulated genes'. Upregulated genes unique to each cell line comprised 30% of the LNCaP upregulated gene set and 81% of the VCaP upregulated gene set reflecting the larger number of differentially expressed genes in VCaP cells (**Figure 3.10A**). In addition, there was a moderate and statistically significant correlation between fold change for the shared upregulated genes in each cell line suggesting that in this common set of genes the magnitude of androgen activity effect was similar (figure 3.10B). In both cell lines, shared upregulated genes had a higher average fold change than non-shared genes indicating that this core set of genes was generally more responsive to androgen signalling after a short exposure to DHT (**Figure 3.10C and D**).

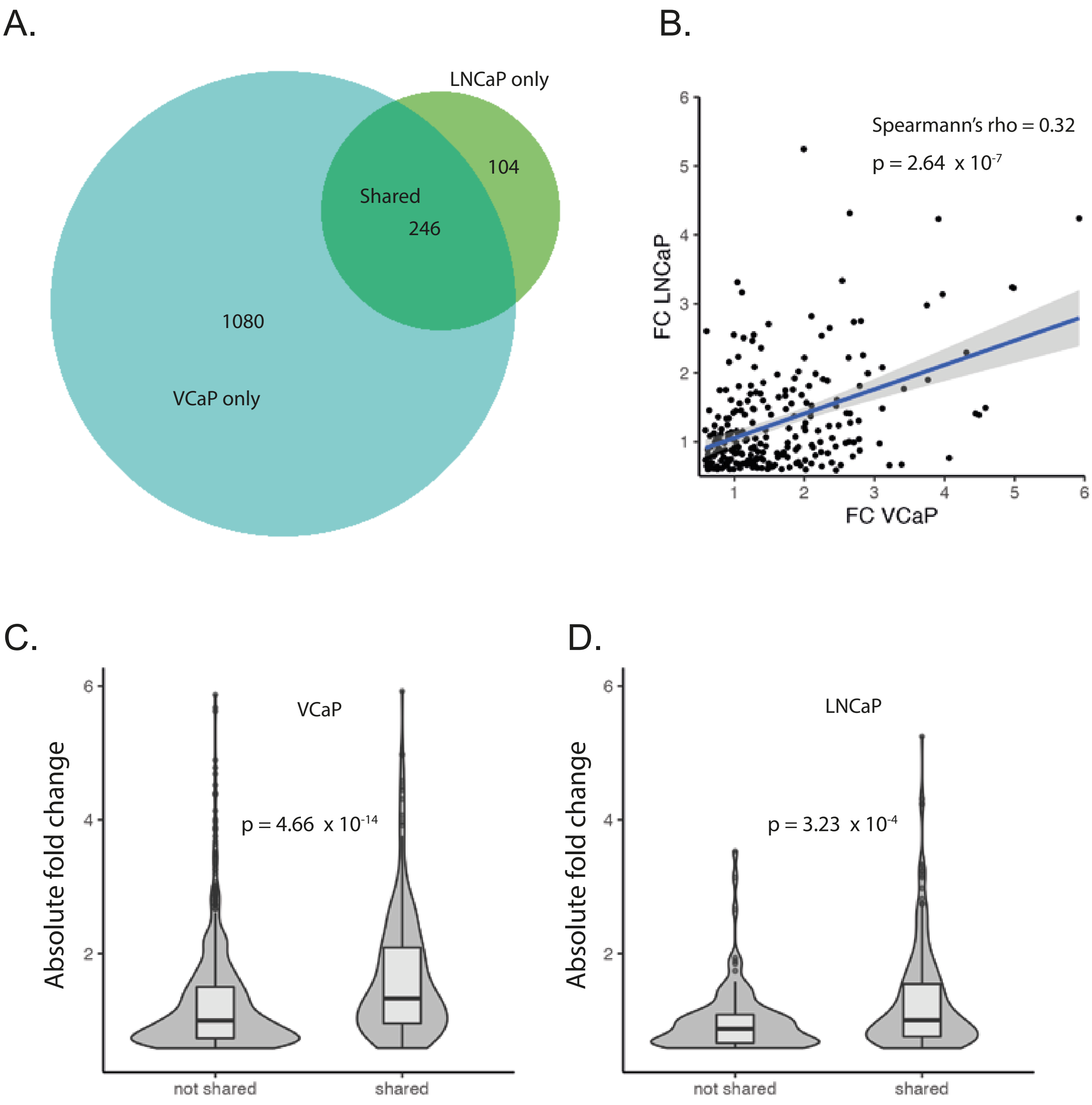


Figure 3.10 Shared upregulated genes in VCaP and LNCaP cells have a more marked response to DHT A. Venn diagram showing the intersection of upregulated genes between VCaP and LNCaP cells after 100 nM DHT. B. Correlation of gene expression fold change of shared upregulated genes in VCaP and LNCaP cells. C. Comparison of fold change of upregulated genes unique to VCaP cells and those shared with LNCaP cells. D. Comparison of fold change of upregulated genes unique to LNCaP cells and those shared with VCaP cells. P values in C and D are from WIlcoxon rank-sum test

The proportion of downregulated gene sets unique to each cell line (‘non-shared genes’) was higher in both VCaP (97% of genes) and LNCaP cells (41% of genes, **figure 3.11A**). In contrast to upregulated genes, no correlation was observed between absolute fold change of the shared genes in this group and there was no significant difference in the average fold change between shared and non-shared genes in either cell line (**figure 3.11B-D**). Next, I compared the RLFS density metric between shared and non-shared genes for each cell line. Only downregulated genes in the VCaP dataset had a significantly different number of RLFS per gene. In this comparison the downregulated genes unique to VCaP cells had more RLFS per gene than downregulated genes shared with LNCaP cells (**Figure 3.12A-D**). In summary, the larger number of downregulated genes in VCaP cells coupled with a greater proportion of unique downregulated genes compared to upregulated genes seems to account for the differences in RLFS seen between the two cell lines modelling response to DHT.

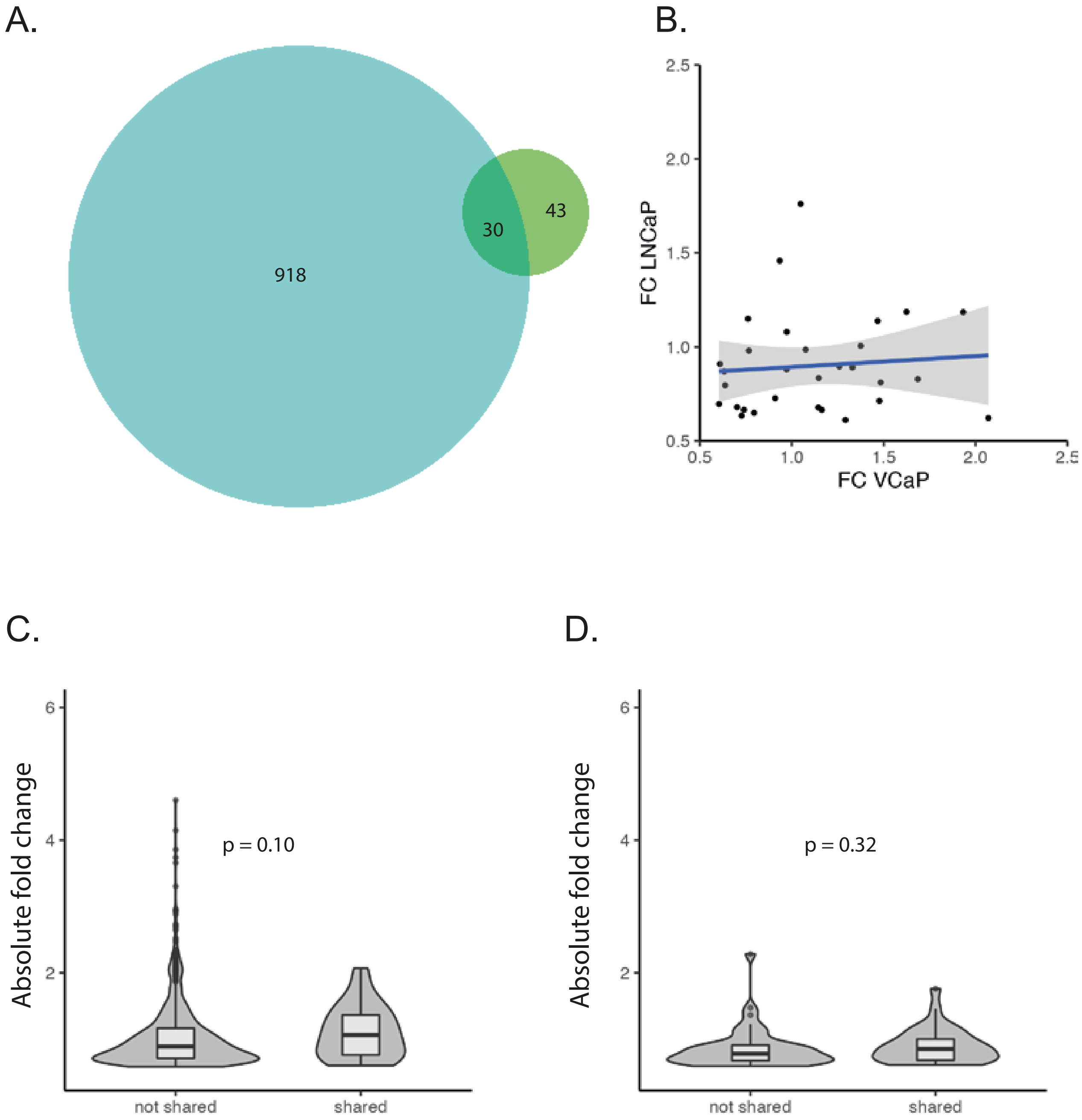


Figure 3.11 Shared downregulated genes in VCaP and LNCaP cells have similar responses to DHT
A. Venn diagram showing the intersection of downregulated genes between VCaP and LNCaP cells after 100 nM DHT. B. Correlation of gene expression fold change in VCaP and LNCaP cells. C. Comparison of fold change of upregulated genes unique to VCaP cells and those shared with LNCaP cells. D. Comparison of fold change of upregulated genes unique to LNCaP cells and those shared with VCaP cells. P values in C and D are from WIlcoxon rank-sum test. Absolute fold change is used for consistency with figure 3.9.

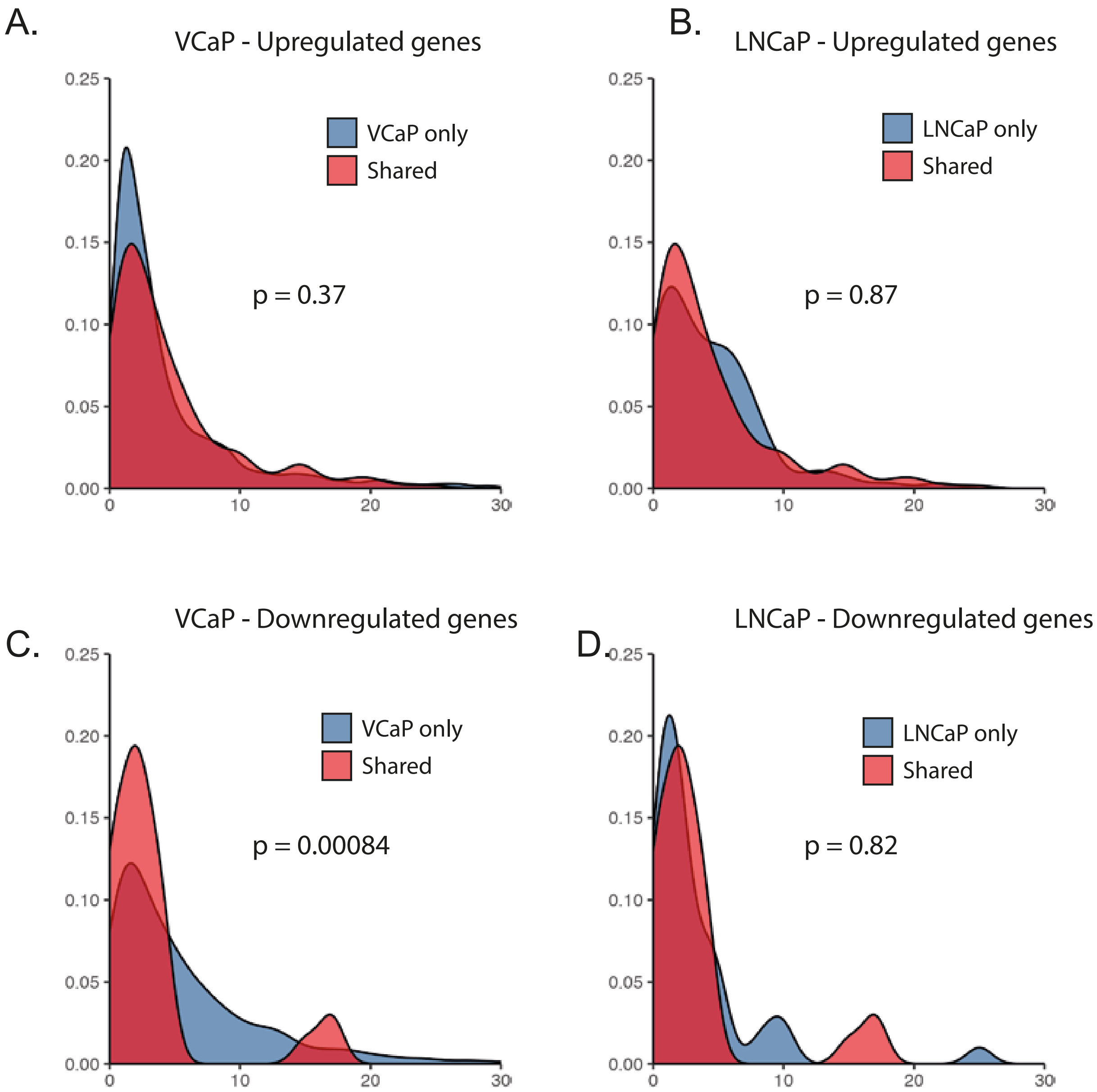


Figure 3.12 Differences in RLFS metrics are driven by the downregulated genes in VCaP cells
A - D. Density plots of shared or unqiue differentially expressed genes. P values are derived from the negative binomial regression models for each comparison.

3.2.6 Metastatic castration resistant prostate cancer cases with androgen receptor amplification downregulate genes with high RLFS burden

I next asked if the association of increased androgen activity and the resultant downregulated genes having higher RLFS metrics compared to upregulated genes was also present in clinical samples. The West Coast Dream Team collaboration characterised 99 metastatic castration resistant prostate cancer (mCRPC) cases with RNA-seq, whole genome sequencing and bisulphite sequencing (Quigley *et al.*, 2018). A key finding from this study was amplification of the androgen receptor locus and its associated enhancer region. This is thought to arise from widespread structural variation as an adaptation to the selective pressure of androgen deprivation which is a common first line treatment of advanced/ metastatic prostate cancer (Viswanathan *et al.*, 2018). I used the TCGA cohort of 333 primary localised prostate cancers as the comparison group for calling differentially expressed genes to represent changes in gene expression between primary and metastatic cancer. At an absolute fold change threshold of two there were 7986 differentially expressed genes (5526 upregulated, 2460 downregulated). Grouping the TCGA and WCDT datasets by the 100 most differentially expressed genes showed that primary and metastatic cases clustered separately (**Figure 3.13A**). In addition, the mRNA expression of the androgen receptor was significantly higher in the metastatic cohort (**Figure 3.13B**) confirming that there was likely to be increased androgen signalling in this group. I examined the biological significance of these gene expression changes with gene ontology analysis (**Figure 3.13C**). The 20 most significant biological processes in upregulated genes included epigenetic mechanisms such as chromatin organisation, nucleosome assembly and gene silencing. Epigenetic adaptation has been implicated in castration resistance. One study showed how ASCL1 was upregulated by androgen deprivation and this favoured a switch to a neuronal-like stem cell phenotype (Nouruzi *et al.*, 2022). This has clinical relevance as CRPC often progresses to an untreatable neuro endocrine phenotype. Importantly in my comparison ASCL1 had a five-fold higher mean expression in the mCRPC cases compared to primary prostate cancer. Consistent with this, some neurological processes were included in the top biological process terms including 'detection of chemical stimulus', implying that some of the mCRPC samples had progressed to a neuronal-like phenotype. The final significant gene ontology finding of note was 'negative regulation of apoptosis' implying a potential mechanism for evading cell death and increasing cell survival.

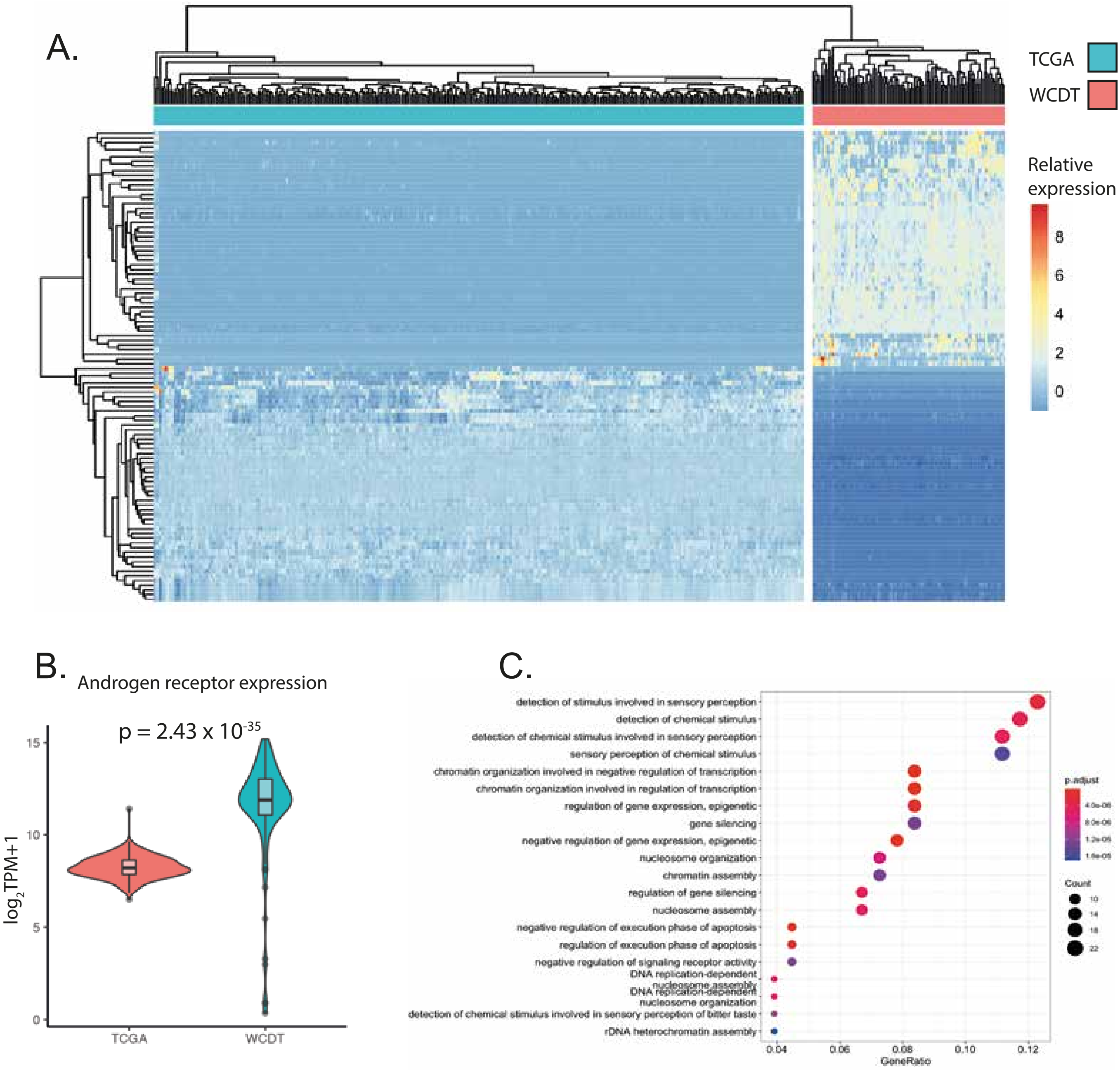


Figure 3.13 Androgen signalling in mCRPC samples drives gene expression programmes
A. Heatmap of top and bottom 50 differentially expressed genes in primary PCa (TCGA) vs. mCRPC (WCDT) samples. B. Violin plot comparing androgen receptor mRNA expression in TCGA and WCDT datasets. C. Top 20 biological process gene ontology terms from genes upregulated in the WCDT dataset.

Analysis of RLFS metrics showed a significant reduction of RLFS positive genes amongst upregulated genes. Approximately 70% of upregulated genes contained at least one RLFS whereas 93% of down regulated genes were RLFS positive (**Figure 3.14A**). This was reflected in a higher number of RLFS per gene and a significantly higher percent coverage by RLFS in down regulated genes (**Figure 3.14B and C**). For all three metrics the up and down regulated gene sets were at the opposite extremes of the background null distribution histograms (vertical red lines in **Figure 3.15A-C**). Furthermore, the RLFS data observed in this dataset was mirrored by GC skew profiles of up and down regulated genes. In promoter regions the peak GC skew value was lower in upregulated genes and in transcription termination regions the characteristic peak of positive GC skew seen around the termination site was lower in upregulated genes (**Figure 3.16**). This suggests that R-loop formation could be reduced at both 5' and 3' ends of genes upregulated with the emergence of castration resistance compared to genes which are down regulated. The clinical sample data also suggests that increased androgen activity favours the downregulation of genes with high levels of RLFS. This corresponds with the data presented in **figures 3.6 -3.9** where downregulated genes in VCaP cells tended to have higher RLFS metrics than up regulated genes and a background null set of genes.

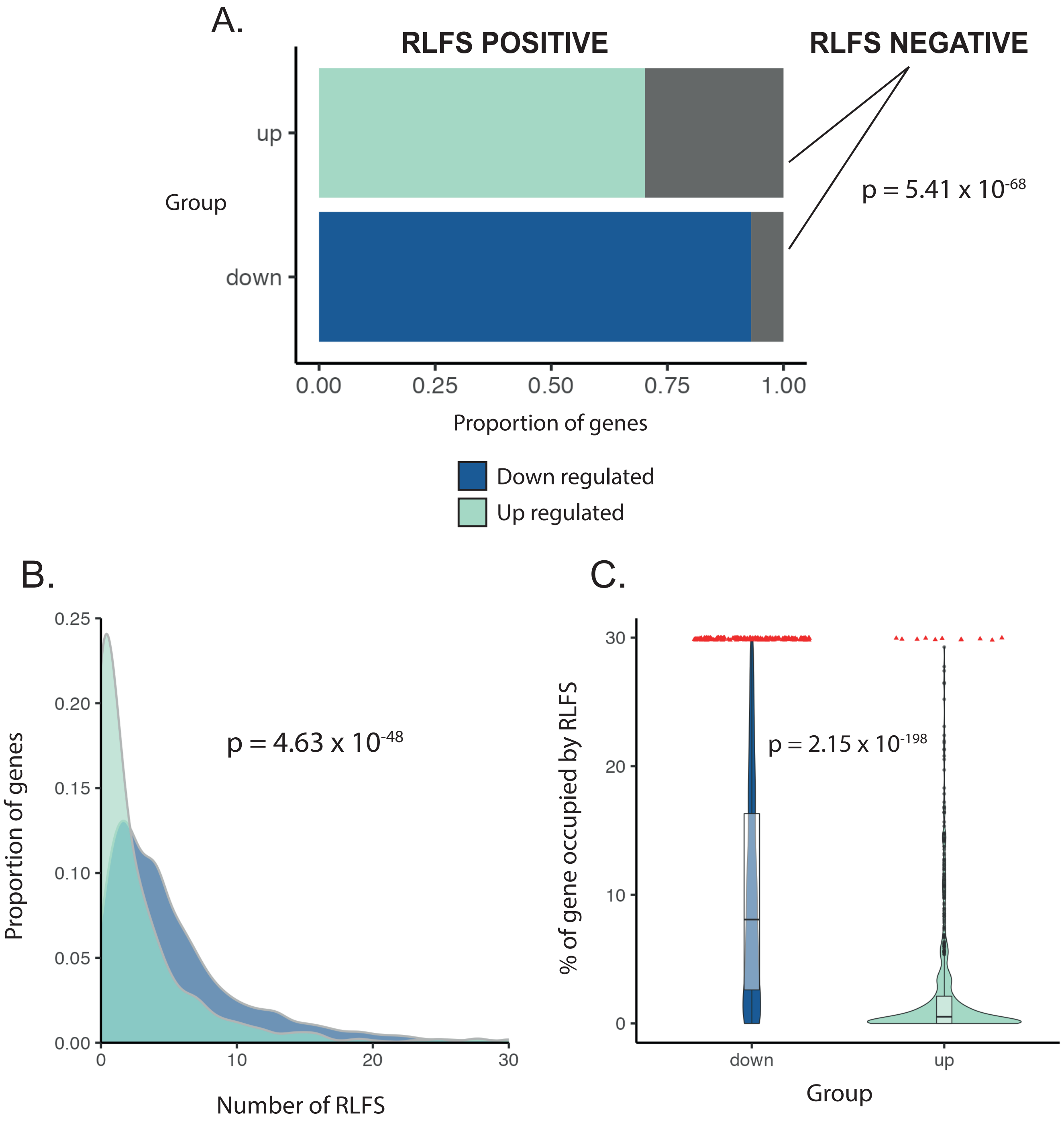


Figure 3.14 Genes upregulated in mCRPC samples have lower levels of R-loop forming sequences. A. Proportion of RLFS positive genes in up and down regulated genes. B. Distributiopn of number of RLFS per gene in up and down regulated genes. C. Percent coverage by RLFS of up and down regulated genes. p-values are derived from the same statistical tests in figures 3.6 and 3.8.

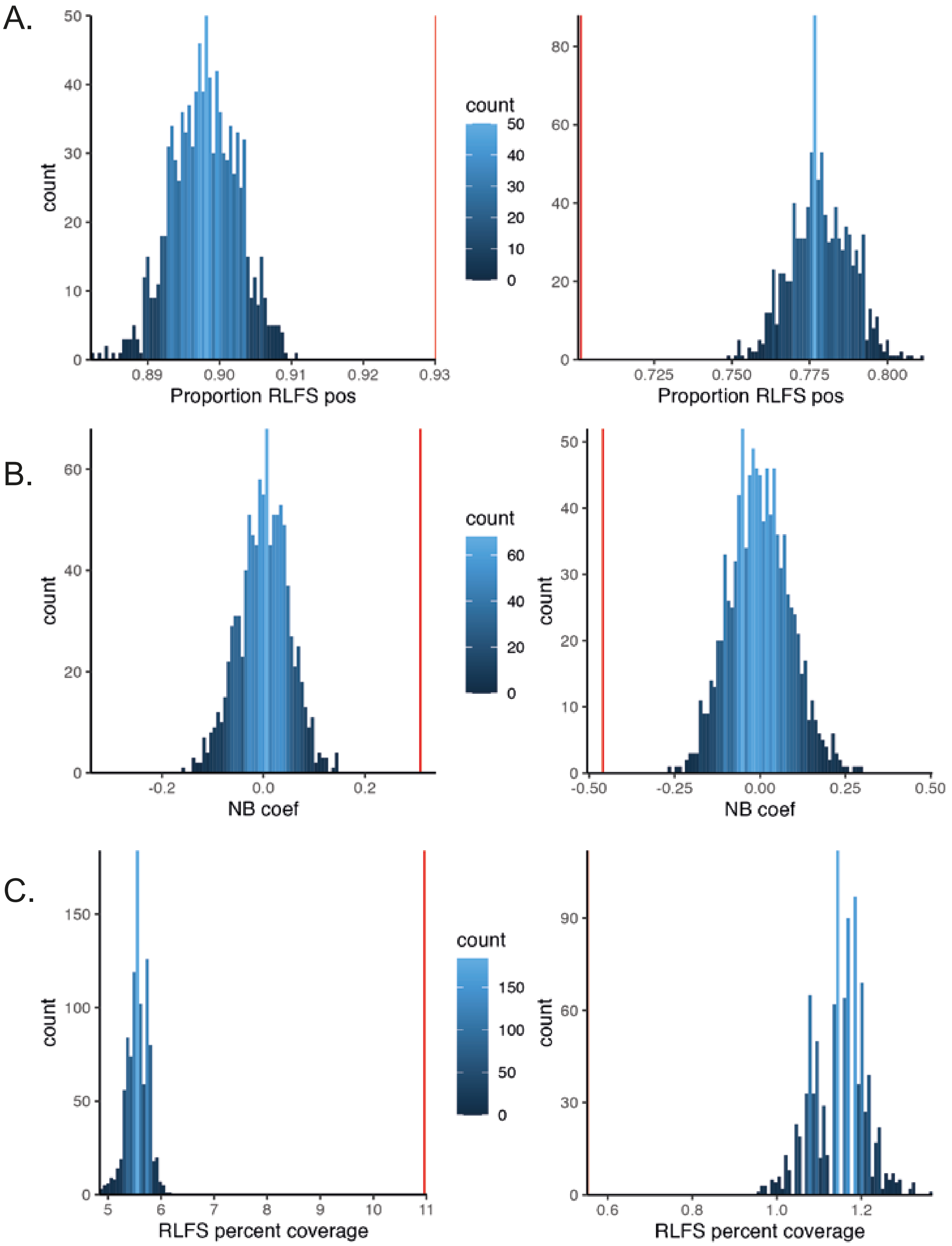


Figure 3.15 Comparison of RLFS in up and down regulated genes with background null distribution
All plots are histograms of random sampling for each metric. A. Proportion of RLFS positive genes. B. Coefficient of negative binomial regression model with group membership shuffled for each sampling iteration. C. RLFS percent coverage. Down regulated genes are on the left, up regulated genes on the right. The red line in each plot indicates the value of that metric in the corresponding gene set.

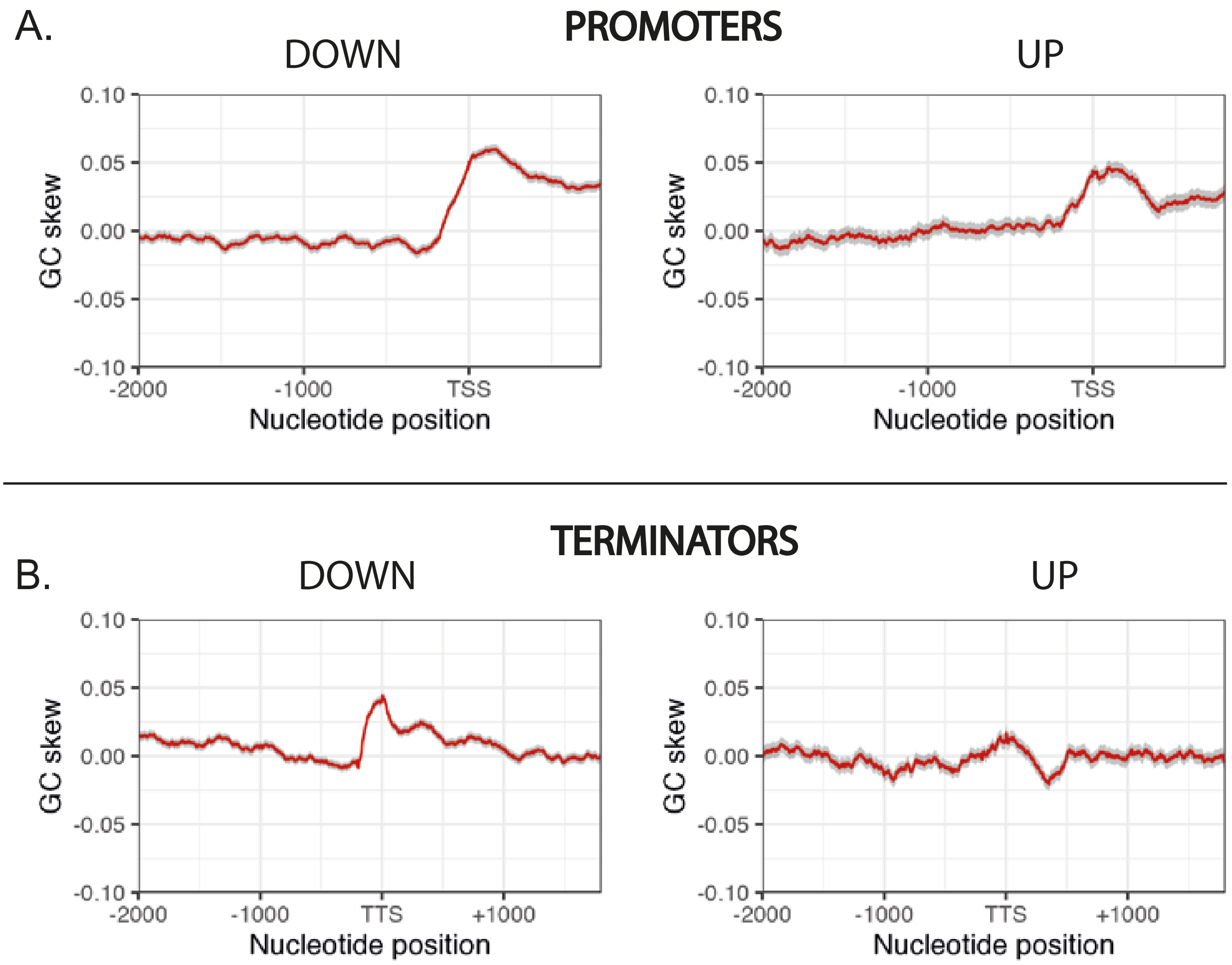
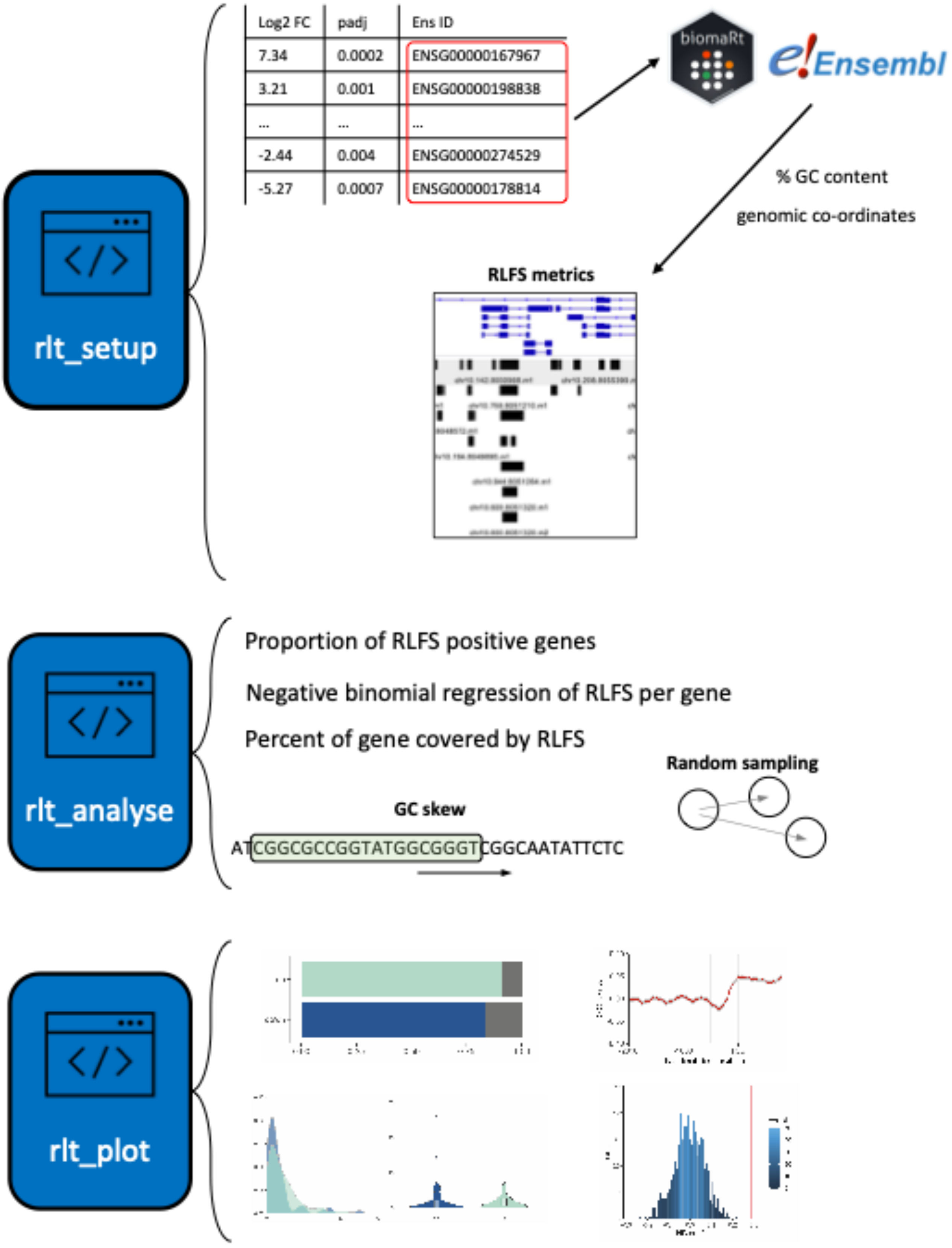


Figure 3.16 GC skew is reduced at TSS and TTS of genes upregulated in mCRPC
A. GC skew plot of the promtoer region in down (left) and up (right) regulated genes in mCRPC compared to primary PCa. B. GC skew plot pf transcrpition termination region of down (left) and up (right) regulated genes in mCRPC compared to primary PCa. TSS: Transcription start site; TTS: Transcription termination site

3.2.7 Development of RloopTools: An R package to apply RLFS metrics to differential gene expression datasets

In the preceding sections I have demonstrated bioinformatics tools that help characterise the R-loop forming potential of differentially expressed genes in different androgen signalling contexts. The use cases for these tools include characterisation of differentially expressed genes from a diverse range of conditions that perturb or manipulate transcription. Examples include gene knock down by siRNA or shRNA, CRISPR knock-out, drug treatment, and window of opportunity clinical trials where tissue samples are taken before and after treatment. To explore different use cases of RLFS densities I developed an R package named RloopTools that encapsulates the functions and parameters required for this analysis. RloopTools is designed to take a set of genes from a gene expression experiment and analyse the RLFS densities of different groups. The output comprises useful figures and comparisons together with a ranked list of genes and their RLFS metrics. The tool is implemented in three simple commands: Firstly, rlt_setup retrieves genomic co-ordinates and annotations for the relevant Ensembl build. Next rlt_analyse combines information from Ensembl, the set of merged RLFS and the list of differentially expressed genes provided by the user. At this stage the FDR and fold change values to call genes differentially expressed can be submitted by the user. Lastly, the output of rlt_analyse is inputted to rlt_plot to produce RLFS metric plots, GC skew plots and histograms comparing the differentially expressed genes' RLFS metrics with a background null set (summarised in **figure 3.17**). Below I demonstrate three examples of how the package can be applied to differential expression datasets.



Log2 FC	padj	Ens ID
7.34	0.0002	ENSG00000167967
3.21	0.001	ENSG00000198838
...	...	...
-2.44	0.004	ENSG00000274529
-5.27	0.0007	ENSG00000178814

rlt_setup
biomaRt
e! Ensembl
% GC content
genomic co-ordinates
RLFS metrics
rlt_analyse
Proportion of RLFS positive genes
Negative binomial regression of RLFS per gene
Percent of gene covered by RLFS
Random sampling
GC skew
ATCGGCGCCGGTATGGCGGGTCGGCAATATTCTC
rlt_plot

Figure 3.17 Structure and functionality of the RloopTools R package

3.2.7.1 Genes regulated by hypoxia-induced senataxin in lung cancer

Senataxin (SETX) is a 302 kDa protein with RNA and DNA helicase functions (Groh *et al.*, 2017). SETX has been implicated in R-loop homeostasis genome-wide. At the transcription termination sites SETX activity allows nascent RNA to be degraded by XRN2 thereby 'torpedo-ing' RNA polymerase 2 from chromatin and ending the transcription cycle (Skourti-Stathaki, Proudfoot and Gromak, 2011). In support of this model, SETX depletion in a mouse model of circadian rhythm regulation led to increased transcriptional readthrough at a subset of genes (Padmanabhan *et al.*, 2012). More recently the Hammond group showed that SETX expression was induced by hypoxia both in cell culture and using a hypoxia gene expression signature in colorectal and lung cancer TCGA datasets (Ramachandran *et al.*, 2021). I applied R-loopTools to an RNA-seq dataset from this study where RKO colorectal cancer cells were cultured in hypoxic conditions (0.1% O_2) and SETX depleted by siRNA. The gene list of differentially expressed genes between an siRNA targeting SETX and a control siRNA were kindly provided by Prof Hammond (Oxford, UK). There was no significant difference of any of the three RLFS metrics between up and downregulated genes (**Figure 3.18A-C**). However, comparing up and downregulated genes with an equal background set of genes showed that differentially expressed genes – those up or down regulated by SETX + hypoxia – had a higher negative binomial regression coefficient suggesting that these genes had more RLFS. Concordantly, differentially expressed genes had a higher proportion of RLFS positive genes and a higher % coverage by R-loop forming sequences (**Figure 3.19A-C**). Together these data imply that the genes regulated by SETX in the context of hypoxia could be regulated via R-loops and that the propensity of a gene to form R-loops could in part determine their transcription under certain types of cellular stress. Importantly the Hammond group validated the predicted accumulation of R-loops by S9.6 and RNase H immunofluorescence experiments under the same conditions as the RNA-seq experiment.

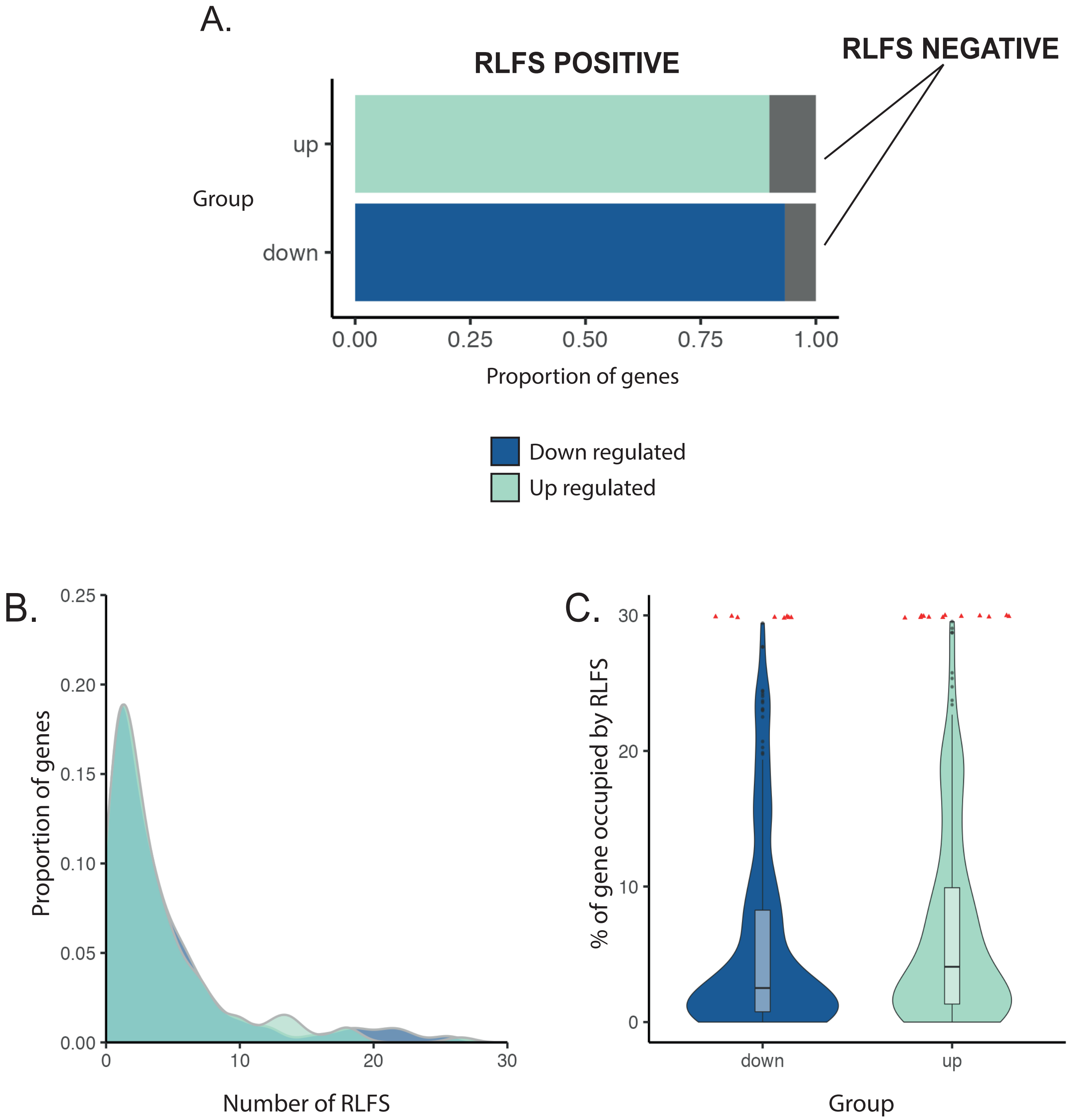


Figure 3.18 Genes differentially expressed by siSETX in the context of hypoxia show no difference in RLFS metrics

A. Proportion of RLFS positive genes in up and down regulated genes. B. Distributiopn of number of RLFS per gene in up and down regulated genes. C. Percent coverage by RLFS of up and down regulated genes. p-values are derived from the same statistical tests in figures 3.6 and 3.8.

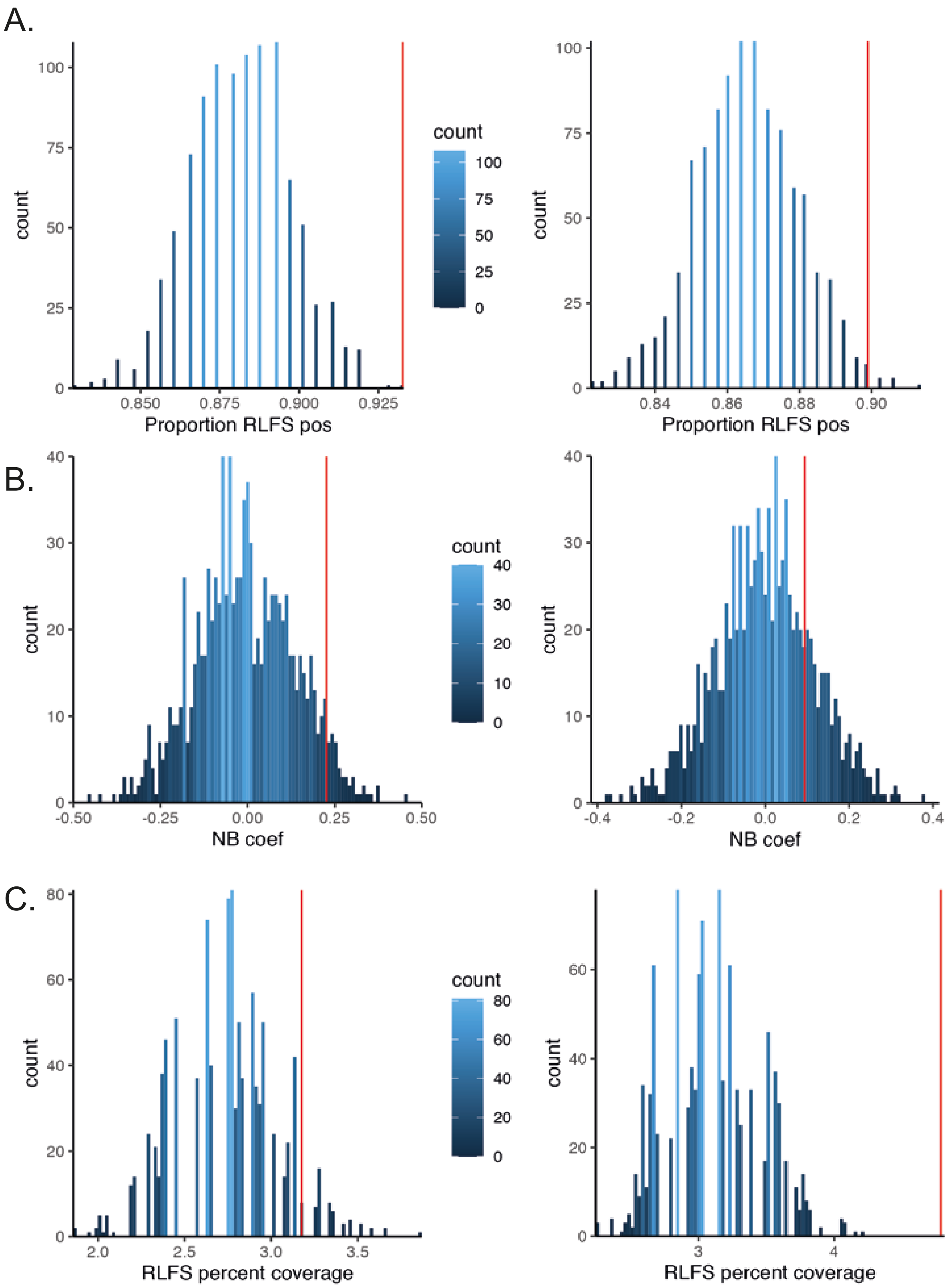


Figure 3.19 Gene differentially by siSETX in the context of hypoxia have higher RLFS metrics than a background null set of genes

All plots are histograms of random sampling for each metric. A. Proportion of RLFS positive genes. B. Coefficient of negative binomial regression model with group membership shuffled for each sampling iteration. C. RLFS percent coverage. Down regulated genes are on the left, up regulated genes on the right. The red line in each plot indicates the value of that metric in the corresponding gene set.

3.2.7.2 Gene expression subtypes in muscle invasive bladder cancer

Muscle invasive bladder cancer (MIBC) is an aggressive malignancy with a five year survival rate of 40-50% that has not improved over 20 years despite a fall in incidence (Eylert *et al.*, 2014). Over the last ten years multiple research groups have worked to define a molecular classification of MIBC through gene expression signatures (Choi *et al.*, 2014; Robertson *et al.*, 2017; Kamoun *et al.*, 2020). Whilst terminology varies these signatures group MIBC into cancers that recapitulate the gene expression profiles of the superficial umbrella cell lining of the bladder ('luminal' and 'luminal papillary'), those with luminal features and a prominent immune cell infiltrate ('luminal infilitrated') and cancers more like the cell layer in contact with underlying lamina propria ('basal'). These subtypes are associated with differing responses to neoadjuvant cisplatin-based chemotherapy. Basal subtype tumours tend to have a favourable response whereas luminal and luminal papillary tumours derive no benefit (Choi *et al.*, 2014; Seiler *et al.*, 2017). It is therefore important to understand the biological differences between tumour subtypes and the associated difference in chemosensitivity. To make a two-group comparison I used RNA-seq data from the luminal/luminal papillary (n = 168) and basal subtype (n = 144) tumours from The Cancer Genome Atlas MIBC dataset. In luminal tumours 373 genes were upregulated and 857 genes downregulated compared to basal tumours. Upregulated genes were more likely to be RLFS positive and had a statistically significantly higher number of RLFS per gene and higher percent coverage by RLFS (**Figure 3.20**) although the absolute difference between subtypes was small. As luminal/luminal papillary tumour are less responsive to cisplatin chemotherapy I next tested if the observed RLFS characteristics were associated with cisplatin resistance in other settings.

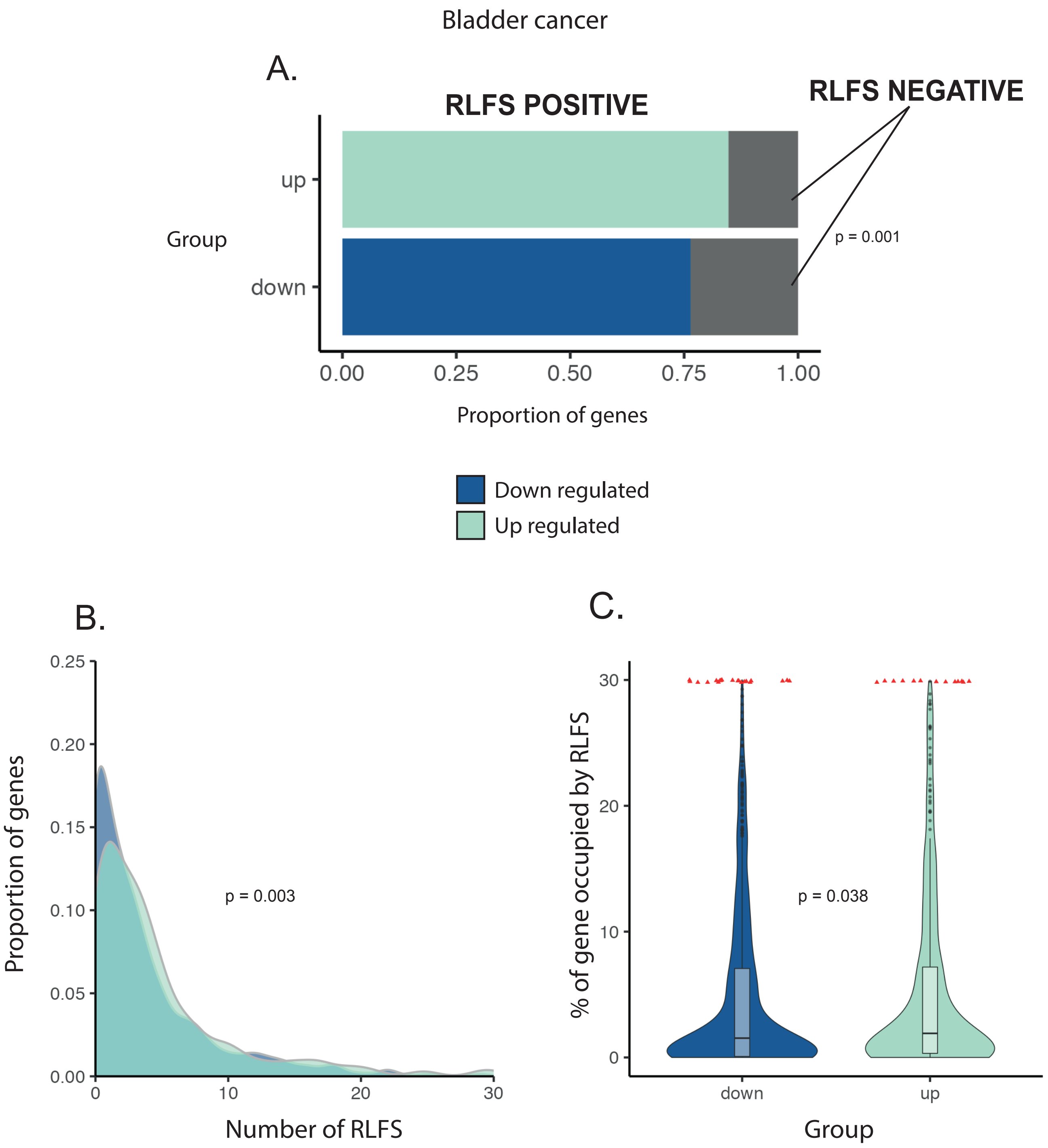


Figure 3.20 Genes upregulated in luminal subtype muscle invasive bladder cancer have more RLFS than downregulated genes
A. Proportion of RLFS positive genes in up and down regulated genes. B. Distributiopn of number of RLFS per gene in up and down regulated genes. C. Percent coverage by RLFS of up and down regulated genes. p-values are derived from the same statistical tests in figures 3.6 and 3.8.

3.2.7.3 Differences in gene expression upon acquisition of cisplatin resistance in head and neck cancer

Head and neck squamous cell carcinoma (HNSCC) is a malignancy of increasing incidence with a five-year survival rate of 50-60%. Human papilloma virus (HPV) infection defines two subtypes of HNSCC that differ at a molecular and clinical level (Lawrence *et al.*, 2015; Johnson *et al.*, 2020). HPV negative HNSCC has a worse prognosis compared to HPV positive disease. Loco regional spread and recurrence is treated with cisplatin but resistance often occurs, after which few treatment options are available (Griso *et al.*, 2022). To model cisplatin resistance Dr Hannah Crane in the El-Khamisy lab cultured HPV negative HNSCC cells (SCC89) in cisplatin containing media for three months. To characterise this model Dr Crane performed RNA-seq of parental and cisplatin resistance clones and kindly provided the list of differentially expressed genes. I applied RloopTools to this differential expression dataset. In total, 427 genes were differentially expressed (226 upregulated; 201 downregulated). Compared to upregulated genes, down regulated genes were less likely to be RLFS positive, had a lower RLFS density and a lower percentage coverage by RLFS (**Figure 3.21A-C**). This pattern was similar to that of the cisplatin-resistant luminal subtype of muscle invasive bladder cancer. This raised the possibility that cisplatin resistance was associated with downregulation of genes containing fewer RLFS. Comparing the up and down regulated genes in the MIBC and HNSCC datasets revealed no significant overlap of the gene sets (**Figure 3.22**). This implies that there is not simply a shared set of dysregulated genes observed with the emergence of cisplatin resistance accounting for the similarity in RLFS metrics between the two datasets.

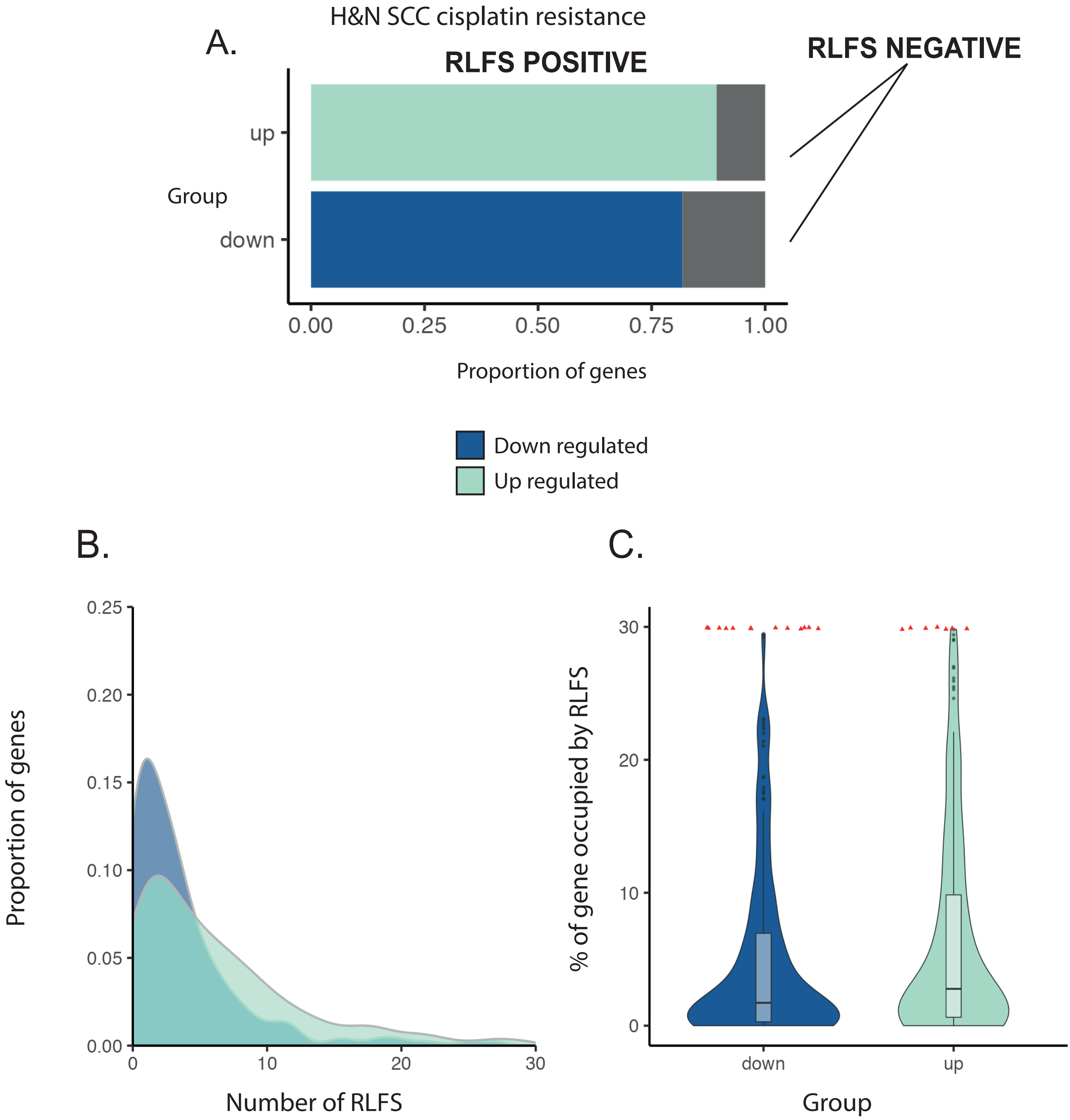


Figure 3.21 Genes upregulated in a cisplatin resistance HPV- HNSCC cell line have more RLFS than downregulated genes
A. Proportion of RLFS positive genes in up and down regulated genes. B. Distributiopn of number of RLFS per gene in up and down regulated genes. C. Percent coverage by RLFS of up and down regulated genes. p-values are derived from the same statistical tests in figures 3.6 and 3.8.

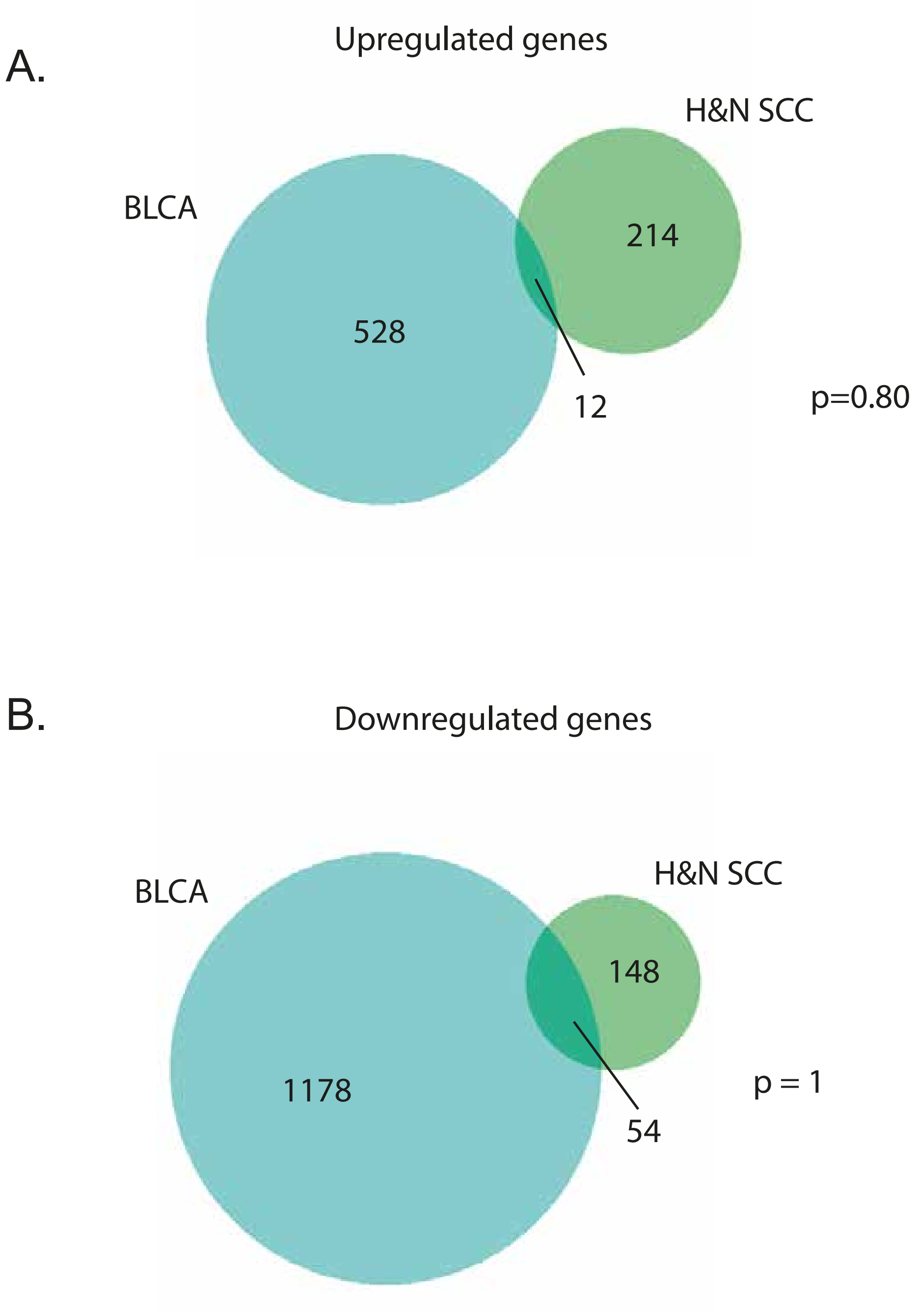


Figure 3.22 Cisplatin resistant HNSCC cells and luminal subtype bladder cancer share few differentially expressed genes.
A. Venn diagram of upregulated genes in bladder cancer and H&N SCC. B. Venn diagram of downregulated genes in bladder cancer and H&N SCC. P values are from hypergeometric test.

3.4 Discussion

In this chapter I have:

1. Generated gene level RLFS metrics that can be used to characterise groups of genes
2. Tested whether RLFS metrics differ between up and down regulated genes in different androgen signalling contexts
3. Combined these metrics and methods into an R package: RloopTools
4. Tested RloopTools using differential expression datasets from a diverse group of biological contexts

3.4.1 Genes downregulated by high androgen activity have higher levels of RLFS

The VCaP cell culture model and West Coast Dream Team CRPC samples both had downregulated gene sets with high RLFS levels compared to upregulated genes and a background null set of genes. As discussed in chapter 1 R-loops have been implicated in gene expression regulation by multiple mechanisms. That downregulated genes have a higher propensity to form R-loops, at least by sequence context, suggests that these genes might rely more on R-loop homeostasis for their regulation compared to upregulated genes. The androgen receptor can repress gene expression by many different mechanisms including association and displacement of transcription factors at the protein level, direct interaction with the promoters of repressed genes, reducing the expression of other transcription factors and also non-genomic activation of signalling pathways with downstream indirect effects on gene expression (Grosse, Bartsch and Baniahmad, 2012). Of these mechanisms R-loop homeostasis is most relevant to repression involving the interaction of the androgen receptor at the promoters of repressed genes owing to the enrichment of R-loops at promoters. A key example of AR-promoter mediated repression is the gene CDH1 which encodes E-cadherin, whose loss is associated with epithelial to mesenchymal transition and metastatic behaviour in many tumour types (Liu *et al.*, 2008). In this study the authors showed that AR recruitment to the CDH1 promoter together with histone deacetylation activity were required for repression of E-cadherin expression. The CDH1 gene was significantly repressed by androgen activity in the VCaP dataset but not the LNCaP dataset. The promoter region contains multiple RLFS sequences making this an interesting target for further characterisation of androgen receptor signalling in different androgen receptor activity contexts.

The differences in RLFS between up and down regulated genes may also represent a selective adaptation instead of or in addition to gene regulatory roles. Recently, supra physiological androgen (SPA) treatment has been proposed as an alternative treatment for prostate cancer. Chatterjee et al (2019) showed that the increased androgen signalling activity induced by SPA reduced the expression of non-homologous end joining DNA damage repair genes previously associated with androgen signalling at physiological androgen levels. A concomitant increase in DNA damage was seen and this was potentiated by PARP inhibition. Interestingly, the largest responses were seen in LNCaP^{AR} cells which have been transduced with additional copies of the androgen receptor to model increased androgen signalling activity. DNA damage is one of the main pathological mechanisms of R-loop accumulation. This can occur through transcription replication collisions resulting in double strand breaks or through the looped out single stranded DNA component of an R-loop acting as a single stranded substrate for nucleotide excision repair factors such as XPF. This latter mechanism has been associated with oestrogen induced R-loop mediated DNA damage (Stork *et al.*, 2016). From my data, I could speculate that the downregulation of genes with more RLFS in the context of increased androgen signalling was a protective mechanism to avoid androgen induced DNA damage. This could be adaptive whereby androgen induced DNA damage upregulated an R-loop resolving factor which repressed the R-loop prone genes. Alternatively, the mechanism could be selective where cells accumulate lethal levels of DNA damage in response to increased androgen activity and the remaining cells survived through gene expression programmes that are less R-loop prone. These survival mechanisms would have to be a balance between reducing the expression of genes that might lead to excessive DNA damage vs. maintaining gene expression programmes that allowed continued survival and metabolic adaptation.

Another interesting observation was the upregulation of genes with more RLFS in LNCaP cells. This is contrary to the observation in cells and clinical samples with a higher baseline AR activity through AR amplification. The SPA data from the Nelson group (Chatterjee *et al.*, 2019) shows that LNCaP^{AR} cells have a greater magnitude of response to androgens across a range of concentrations and exposure duration. This was reflected in my comparison of LNCaP and VCaP GRO-seq data there were more genes differentially expressed in the VCaP dataset and the fold changes of the most up and down regulated genes had a higher magnitude in VCaP cells. From this I suggest that there is a threshold of R-loop formation in response to androgens which is crossed with a higher availability of the AR through increases in AR copy number. Beyond this threshold the balance of up and downregulated genes changes with respect to R-loop forming potential possibly by the mechanisms described above. This would imply that in the context of a non-amplified AR locus, there is a tendency to expression of genes with higher R-loop forming potential. This correlates with the observations of Stork et al (2016) who showed a significant association of gene expression and R-loop formation in breast cancer cells stimulated with oestrogen.

3.4.2 Limitations of inferring biological behaviour from nucleotide sequence analysis

The observations in this chapter are all based on the qmRLFS model which infers R-loop formation from nucleotide sequence features. Multiple groups have established that GC skew favours R-loop formation (Ginno *et al.*, 2013a; Hartono, Korf and Chédin, 2015; Crossley *et al.*, 2020) but moreover that G-clustering is an essential nucleotide characteristic (Roy and Lieber, 2009) in R-loop initiation and it is this observation that is encapsulated in the qmRLFS tool (Wongsurawat *et al.*, 2012). By extending the use of this tool to provide per-gene read outs of R-loop forming potential I have enabled the analysis of multiple differential gene expression datasets. However, this model assumes that increased transcription of these genes (e.g if specific genes are upregulated in response to a stimulus) will always lead to an increase in R-loops if those genes are particularly R-loop prone. Whilst predicted RLFS have been validated against DRIP-seq and with DRIP-qPCR for a selection of loci, many of these loci have stable R-loops and the correlation between predicted RLFS and dynamic R-loops is less certain. A specific criticism of qmRLFS is its performance at recognising transcription termination R-loops. The Chedin group developed Rlooper, a command line tool that uses a free energy equilibrium model to assign a probability of R-loop stability to a nucleotide sequence (Stolz, Sulthana, Stella R. Hartono, *et al.*, 2019). Validation of this model using long read bisulfite sequencing to identify the asymmetric looped out DNA strand characteristic of R-loops showed that Rlooper and qmRLFS correctly predicted R-loop formation at promoter region but that Rlooper was more accurate at transcription termination sites (Malig *et al.*, 2020). However, this comparison was made for selected loci cloned into vectors and PCR amplified so cannot be generalised genome-wide. Furthermore, the Rlooper tool requires a superhelicity parameter for R-loop prediction. This parameter is presumably dynamic based on the chromatin architecture and transcriptional state of each specific locus. Therefore choosing an optimal value for this parameter across thousands of genes would be challenging and could influence the likelihood of a sequence being predicted to form R-loops.

3.4.3 Possible links between senataxin, cisplatin resistance and R-loops identified by the RloopTools package

Despite the limitations of the RloopTools package, I have demonstrated how three RLFS metrics can be used to characterise differentially expressed genes in diverse contexts. The data from SETX knockdown in the setting of tumour hypoxia validates the model as SETX is a known R-loop helicase and the genes differentially expressed by its loss have a greater propensity to from R-loops. The finding of up regulated genes containing more RLFS in cisplatin-resistant cells across two unrelated cancer types is interesting as this may indicate an additional mechanism of cisplatin resistance. In a genome wide CRISPR screen, loss of SETX was associated with an increased sensitivity to cisplatin (Olivieri *et al.*, 2020). As SETX loss is associated with R-loop accumulation and DNA damage (Groh *et al.*, 2017; Jurga *et al.*, 2021) it may be that R-loop accumulation and cisplatin treatment are synergistic in producing lethal DNA damage. Supporting this hypothesis Andrews et al (2018) showed that loss of SAN1, a regulator of SETX activity increased the sensitivity of HeLa cells to cisplatin and other intra-strand cross linking agents. An earlier study demonstrated RNA polymerase 2 stalling at cisplatin induced DNA damage lesions (Jung and Lippard, 2006) with the polymerase stably retained on DNA. This would conceivably increase the dwell time of the nascent RNA at cisplatin adducts and lead to an increase propensity of R-loop formation similar to that seen in transcription blockage caused by Top1 inhibition (El Hage *et al.*, 2010). However these lines of evidence are contrary to my observation that upregulated genes are *more* R-loop prone than down regulated genes. If cisplatin related DNA lesions increased R-loop formation it might be expected that genes with lower R-loop forming potential would be transcribed to circumvent excessive R-loop mediated DNA damage. One explanation is that in cisplatin resistance there is more efficient DNA intra-strand crosslink repair (Wynne *et al.*, 2007) and this could allow cells to maintain transcription of essential genes that are also prone to R-loop formation.

3.4.4 Future development of RloopTools

The RloopTools package has some limitations. Currently only simple dichotomous comparisons are available. This caters for many differential expression experiments however it excludes experiments where multiple conditions are compared with a baseline or time course experiments. In addition, use of the package currently requires knowledge of R programming to run and whilst the commands and output are designed to be simple to use, this could represent a barrier to entry to produce a simple analysis and list of RLFS prone genes to test experimentally. A future development could include deployment of the package as a web based Shiny package, enabling access to the package without R programming knowledge. More than 50 shiny apps with molecular biology applications are published annually (Kasprzak *et al.*, 2020) illustrating the popularity of making sophisticated analyses accessible through a web browser interface. Certain modifications would be needed to achieve this for RloopTools most notably the inclusion of server-side static data containing gene and RLFS annotations that users could query by inputting their own differentially expressed gene lists.

Chapter 4. The androgen regulated gene NKX3.1 accumulates R-loops in response to androgen stimulation

4.1 Introduction

R-loop formation is a co-transcriptional event requiring nascently transcribed RNA to re-anneal to its complementary DNA template. Androgen signalling is the central driver of prostate cancer growth and development, and binding of ligand bound androgen receptor to DNA is accompanied by activation of specific transcriptional programmes. At the transcriptional level this manifests as changes in levels of nascent and mature RNA. In MCF-7 breast cancer cells, oestrogen receptor signalling was shown to cause an increase in R-loop accumulation and subsequent DNA damage (Stork *et al.*, 2016), illustrating how hormone bound nuclear receptors can contribute to mutagenic processes. This finding of oestrogen induced R-loop accumulation was contradicted by a second study that used T47D cells. R-loops did not increase globally with oestrogen treatment however a reduction in R-loop accumulation was observed at loci that required RING1B for oestrogen induced transcription. A third study showed R-loop formation at oestrogen responsive genes that blocked transcription by acting as a stalling force on RNA polymerase 2 (Song *et al.*, 2017). Given these varied and contradictory effects of oestrogen receptor signalling on the causes and effects of R-loops I hypothesised that androgen signalling would affect R-loop accumulation at androgen responsive loci. In addition I hypothesised that androgen stimulation might change R-loop accumulation at androgen receptor binding sites which are known to transcribe enhancer RNA, a potential source of R-loop formation (Hsieh *et al.*, 2014; Wang *et al.*, 2021).

4.2 Results

4.2.1 Androgen stimulation does not change global R-loop levels in LNCaP cells

I began by asking if androgenic stimulation of transcription in prostate cancer cells would change global R-loop abundance. LNCaP cells were used as they represent a metastatic prostate cancer model that has intact androgen signalling and can model the transcriptional response to androgens (Abate-Shen and Nunes de Almeida, 2022). Cells were androgen starved by culturing in hormone free media for 48 hours then exposed to 100 nM DHT for 2 hours or 24 hours, or treated with an equivalent concentration of vehicle (methanol) for 24 hours. These timepoints were chosen to represent early and sustained transcriptional activation respectively. Genomic DNA was extracted and assayed by slot blot. In this technique DNA is applied to a positively charged Nylon membrane and then probed with the S9.6 antibody to detect R-loops (Stork *et al.*, 2016; Ramirez *et al.*, 2021). Quantification of band intensities revealed no change in global R-loop levels at either timepoint compared to vehicle-treated cells. Importantly, the S9.6 signal was significantly reduced by *in vitro* RNase H1 pre-treatment confirming specificity of the assay for R-loops (**Figure 4.1**). Slot blot cannot give an indication of locus-specific R-loop changes and a large proportion of R-loop signal comes from abundantly transcribed ribosomal DNA which could obscure more subtle changes in R-loop occupancy of other genomic loci (Wahba *et al.*, 2011; Nadel *et al.*, 2015; Abraham *et al.*, 2020). I therefore decided to perform DNA:RNA Immunoprecipitation coupled to high throughput sequencing (DRIP-seq) for a higher resolution analysis of androgen responsive R-loops.

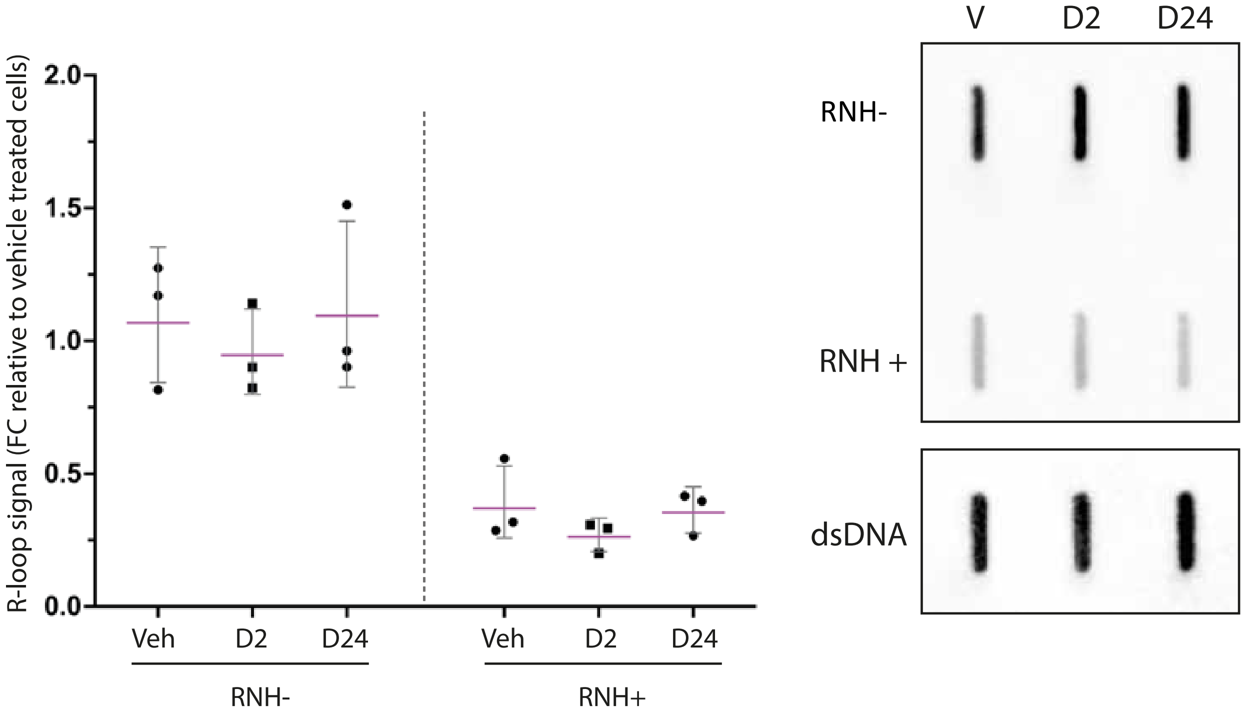


Figure 4.1 Global R-loop levels in LNCaP cells are unchanged by androgen treatment
LNCaP cells were treated with 100 nM DHT or vehicle for the indicated timepoints. Left: Quantification of S9.6 signal. All values were normalised to the mean of vehicle treated cells and to the double stranded DNA (dsDNA) loading control. Right: Representative image of slot blot. Student t test was used to dervive p values for individual comparisons between RNH- and RNH+ conditions. RNH: RNase H1 treatment prior to slot blot. V: Vehicle only, D2: 2hr DHT, D24: 24hr DHT.

4.2.2 Optimisation and quality control of DNA:RNA immunoprecipitation (DRIP) and library preparation

DNA:RNA hybrid immunoprecipitation (DRIP) uses the S9.6 antibody (Boguslawski *et al.*, 1986) to purify regions of the genome containing R-loops. The S9.6 antibody has high affinity for DNA:RNA hybrids (Phillips *et al.*, 2013) and recognises a six base pair hybrid, making contact with three consecutive RNA bases and six DNA bases in the minor groove. Importantly this binding mechanism is similar to that of RNase H1, which resolves and is specific for R-loops *in vivo* (Bou-Nader *et al.*, 2022). Following purification, R-loop levels can be measured at specific loci by qPCR or genome-wide by next generation sequencing.

Following immunoprecipitation, qPCR was performed with primers designed to amplify two regions prone to form R-loops (RPL13A and TFPT; positive control loci) and one region characteristically low in R-loops (SNRPN; negative control locus). The percentage of input (described in chapter 2) for each positive locus was divided by that of the negative control locus to give a fold change of immunoprecipitation efficiency. Whilst this was variable, there was enrichment of known R-loop forming regions in all samples indicating successful immunoprecipitation (**figure 4.2**). Furthermore, treatment of the extracted DNA with RNAse H1 prior to immunoprecipitation resulted in a marked reduction of the DRIP signal. RNase H1 recognises and cleaves the RNA strand in R-loops (Nowotny *et al.*, 2005) and provides a sensitivity control in DRIP experiments.

After the initial library preparation, the bioanalyser trace for DRIP-seq libraries showed evidence of PCR 'bubbles' in all samples. A PCR bubble occurs when a sequencing library is over-amplified and library products anneal to each other rather than primers in later amplification cycles (Illumina, 2021). I repeated the PCR amplification with the remaining unamplified library using 10 cycles. The bioanalyser trace now showed that libraries were of an appropriate fragment size distribution with a peak fragment size of ~300bp. Furthermore, the single peak implied an absence of over-amplification, PCR bubbles and primer dimers (**figure 4.3**).

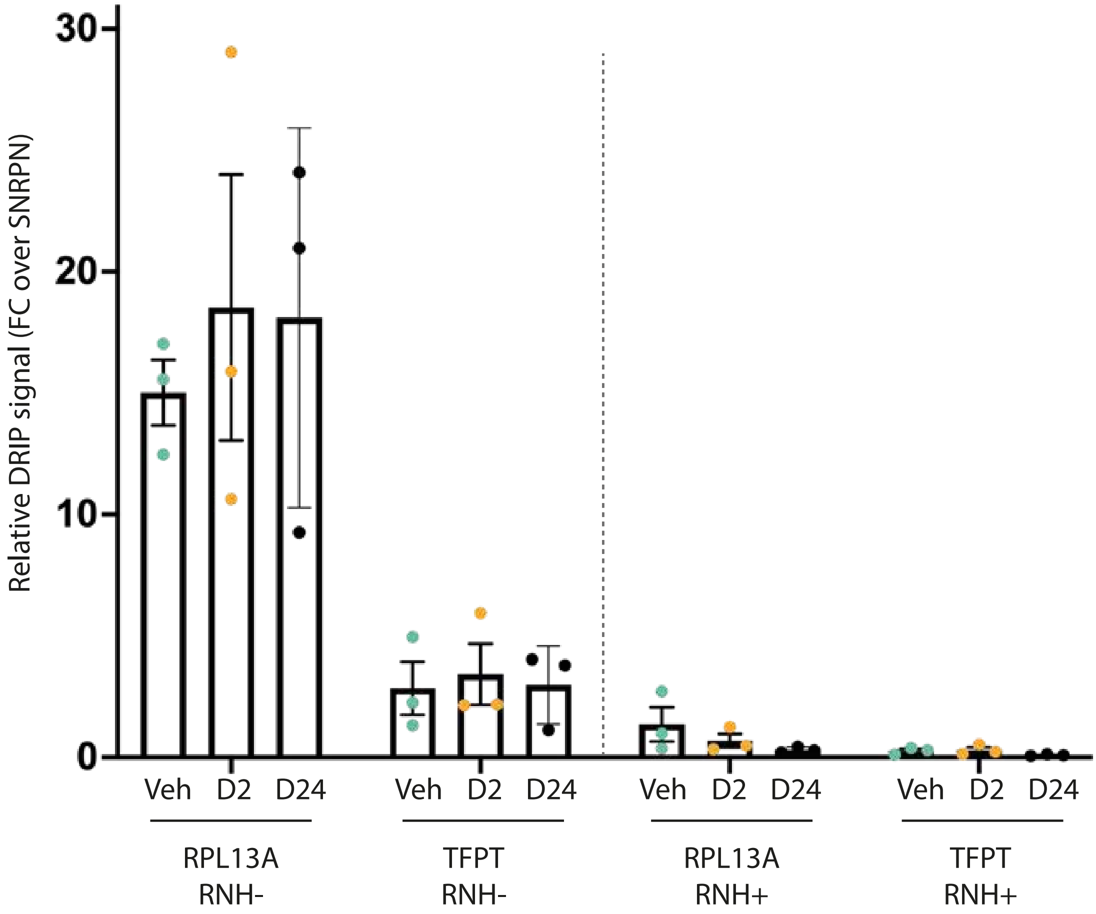


Figure 4.2 Pre-sequencing quality control: Assessment of DRIP efficiency
DRIP-qCR using primers that amplify RPL13A, TFPT R-loop loci. DRIP efficiency is presented as fold change (FC) over the SNRPN locus. Treatment with recombinant RNase H1 (RNH+) was used as a specificity control.

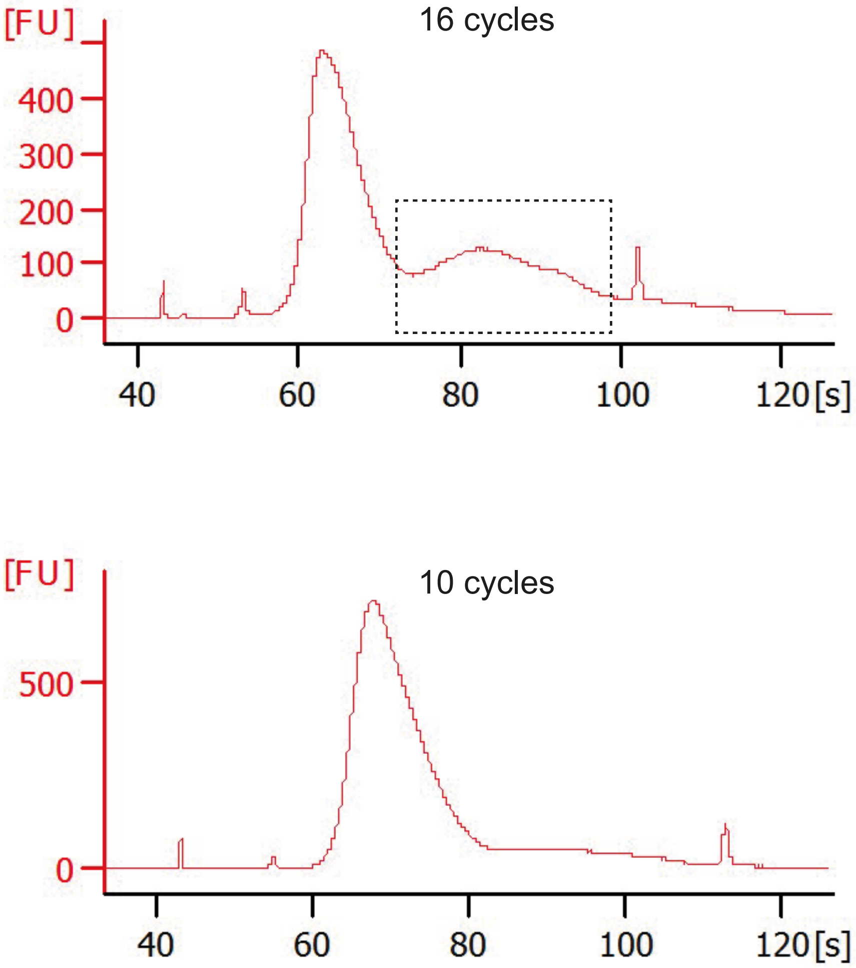


Figure 4.3 Pre-sequencing quality control: Recognition and resolution of library PCR bubbles
Representative bioanalyser traces of DRIP-seq library after 16 and 10 cycles respectively. The dotted line box indicates the PCR bubble in the 16 cycle library. FU: Fluorescence units

4.2.3 Quality control following sequencing

Following adapter removal and read alignment the consistency between biological replicates was assessed by counting reads over 10kb bins and normalising to the total number of reads (counts per million). Replicates for each condition were similar and the correlation was reduced by RNase H treatment indicating true R-loop signal (**figure 4.4A**). Read counts within significant peaks called by MACS2 were concordant between conditions (**figure 4.4B**) and principal component analysis of read counts in within called peaks showed that biological replicates from the same condition clustered together (**figure 4.4C**). As a final quality control, the DRIP-seq signal was examined in IGV. This showed RNase H sensitive enrichment of R-loop signal at regions of previously R-loop occupancy including beta-actin, RPL13A, TFPT and the chromosome 21 ribosomal RNA locus (**figure 4.5**). Importantly the magnitude of change was greater at the RNA45S locus (~150 CPM) than the other loci (~1 CPM). This in agreement with published data describing the abundance of R-loops at ribosomal RNA/ nucleolar loci (Velichko *et al.*, 2019; Zhou *et al.*, 2020) and demonstrated the dynamic range of this DRIP-seq experiment. Interestingly the reduction in R-loop signal with RNAse H1 treatment was less marked at the ribosomal RNA locus which may reflect the increased occupancy of R-loops at this position. Taken together these data indicate that immunoprecipitation and sequencing detected *bona fide* R-loops in a reproducible manner allowing for comparisons between conditions.

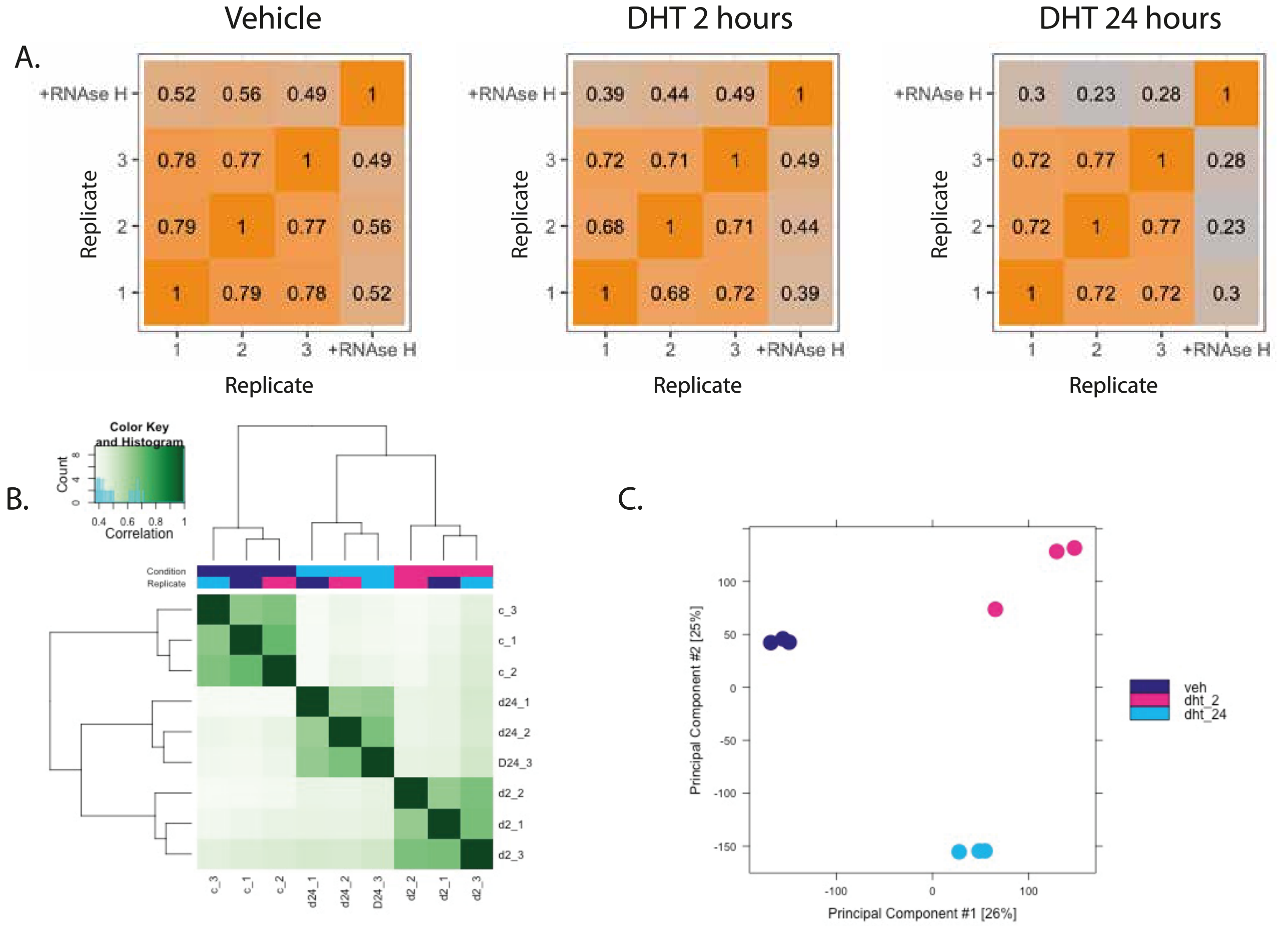


Figure 4.4 Post-sequencing quality control
A. Heatmap of Pearson correlation coefficients of read counts over 10kb bins comparing each biological replcaite for each treatment condition. B. Heatmap of Pearson correlation coefficient of read counts in each R-loop peak called by MACS2. C. Principal component analysis plot demonstrating clustering of biological repicates from each treatment condition

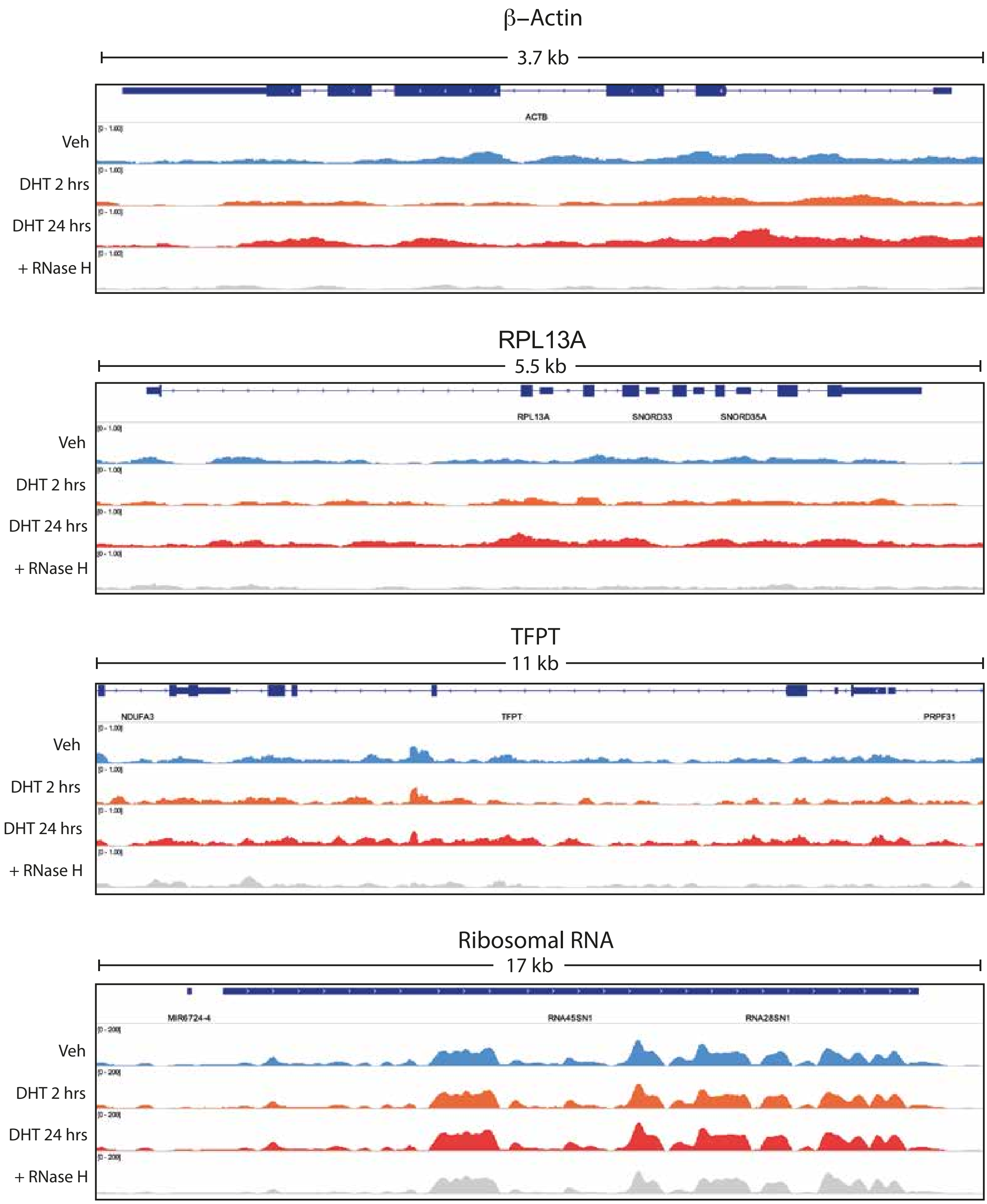


Figure 4.5 Representative IGV tracks of R-loop hotspots
DRIP-seq profiles of four R-loop hotspots (beta actin, TFPT, RPL13A and the chromsome 21 ribosomal RNA array). Grey tracks labelled +RNAse H indicate samples treated with recombinant RNAse H1 prior to DRIP which act as a sensitivity control.

4.2.4 R-loop losses outnumber R-loop gains after androgen treatment.

R-loops peaks were called using MACS2. The genomic coordinates of peaks present in at least two of three biological replicates were merged to create a per-condition peak set. These peak sets were combined yielding a consensus peak set of 47,516 peaks with a median peak width of 646 bp (**figure 4.6**) in keeping with previous DRIP-seq datasets (Halász *et al.*, 2017; Villarreal *et al.*, 2020). Differential enrichment analysis of consensus R-loop peaks was done using DESeq2. This revealed that loss of R-loops at both time points after DHT outnumbered gains (**figure 4.7A and B**). At two- and 24-hours after DHT there were 2.3 and 2.7 times fewer R-loop compared to vehicle treated cells. Despite there being a similar number of differentially enriched R-loops at each time points, only 302 (~20% of 2 hour differential R-loops) differential R-loops were shared between the two timepoints (**figure 4.7C**). R-loops with reduced enrichment after DHT had a greater median width at both time points, and occupied a greater total genomic space (2 hours: 3.15 vs. 1.21 Mb; 24 hours: 3.69 vs. 1.19 Mb). After 2 hours DHT treatment ~80% of R-loop losses and ~70% R-loop gains were in genic regions. With 24 hours DHT treatment ~75% and ~70% of R-loop losses and gains respectively were in genic regions (**figure 4.8A**). Within genic regions there was a significantly higher proportion of R-loop loss in promoters at both timepoints. This pattern was reversed in introns at 2 hours after DHT where a greater proportion of gains were found (**figure 4.8B**). In addition, there was a significantly higher proportion of R-loops gained by 2 hr DHT treatment in intergenic regions. I explore the genic and intergenic R-loop data separately below.

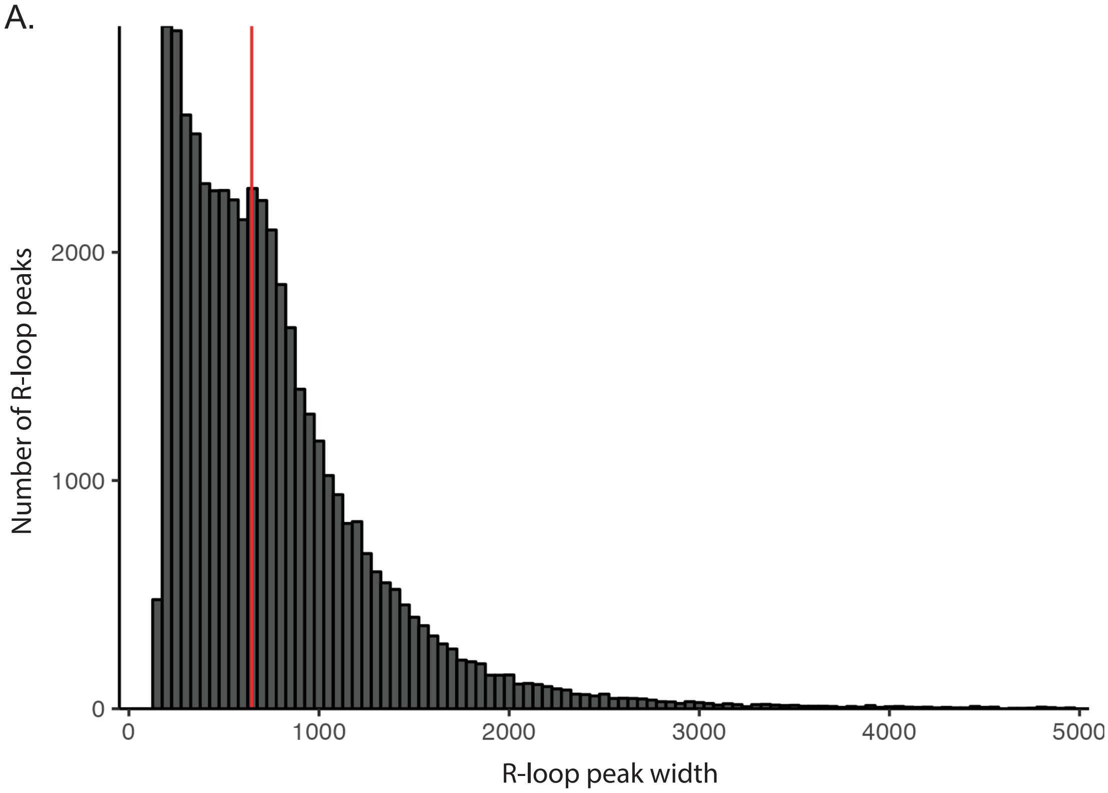


Figure 4.6 Median R-loop peak width of consensus peak set
A. Histogram of the distribution of R-loop peak widths. Median (646 bp) is indicated by the vertical red line. Peaks present in at least 2 of 3 biological replicates in each condition were merged if they overlapped by at least 1 bp to construct the consensus peak set.

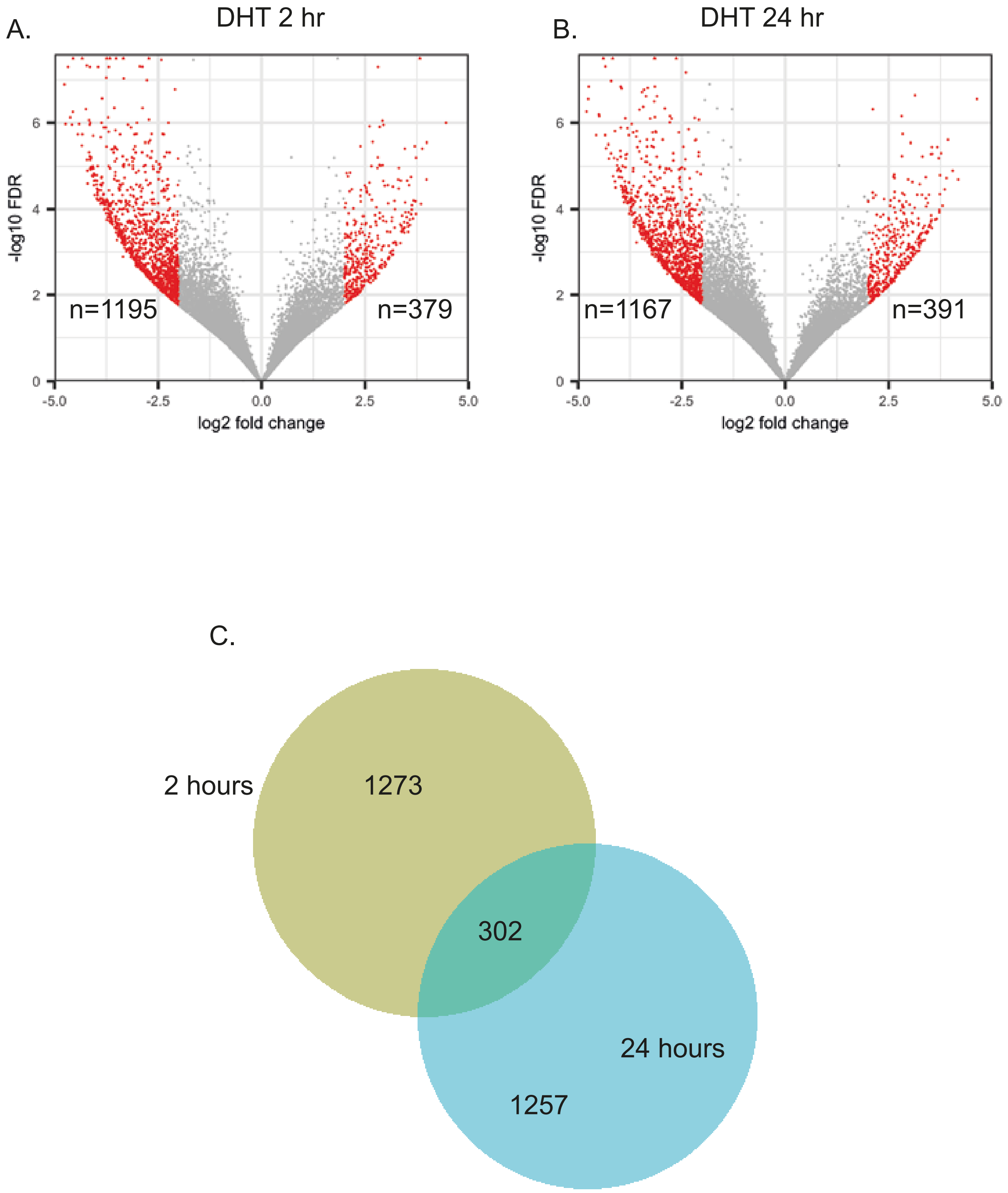


Figure 4.7 R-loop losses outnumber R-loop gains after androgen treatment
A. and B. Volcano plot of R-loop peaks significantly differentially enriched after 2 and 24 hr DHT treatment respectively. C. Venn diagram of the intersection of differential R-loops after 2 and 24 hr DHT.

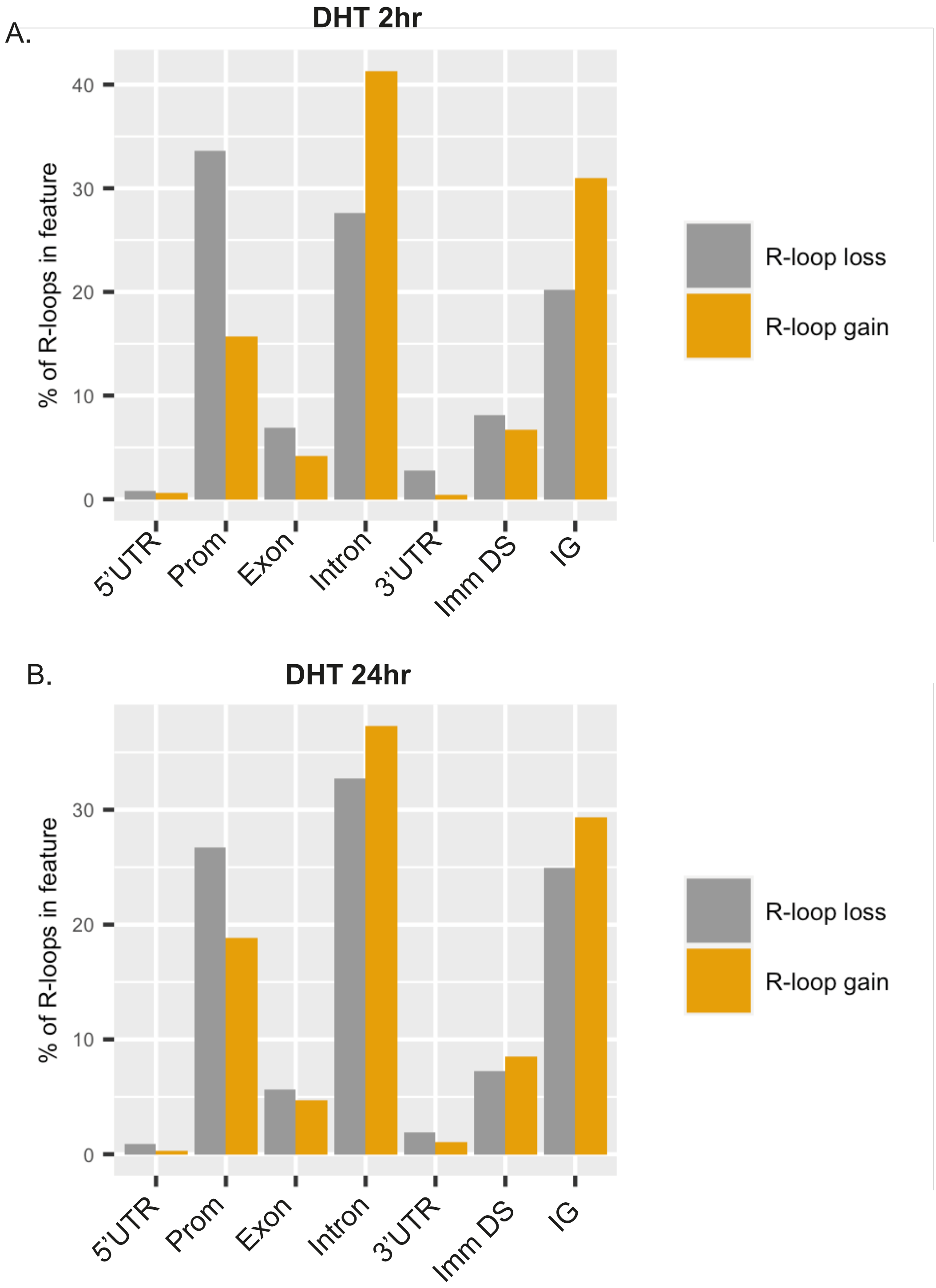


Figure 4.8 Distirbution of R-loop loss and R-loop gain across genomic features
A and B. R-loop loss and gain after 2 and 24 hr DHT respectively. R-loop loss was defned as R-loop peaks present in vehicle treated cells but no longer present after DHT treatment. R-loop gain was defined as R-loops present in DHT treated cells not present in vehicle treated cells.

4.2.5 Integration of DRIP-seq and gene expression data shows that most dynamic R-loop changes are not in androgen responsive genes

To determine if the observed pattern of genic R-loop gains and losses correlated with gene expression levels I used two publicly available gene expression datasets. For the two-hour timepoint I used the same GRO-seq dataset as described in chapter 3 (Wang *et al.*, 2011) and for the 24 hour timepoint I selected an RNA-seq dataset where LNCaP cells had been treated with DHT for 24 hours (Yuan *et al.*, 2019). The majority of differentially expressed genes ($\log_2$ fold change +/- 0.585 and FDR < 0.05) in both GRO-seq (2 hr DHT) and RNA-seq (24 hr DHT) did not overlap with differentially enriched R-loops and there was no correlation between gene expression fold change and R-loop enrichment (**figure 4.9 A and B**). Of 421 differentially expressed genes at 2 hr after DHT 45 (10.7%) also had significant differential R-loop enrichment. After 24 hr DHR 126/1476 (8.5%) of differentially expressed genes had differential R-loop enrichment. Within these groups there was also no correlation between gene expression and R-loop enrichment.

I next tested whether differentially expressed genes had different R-loop profiles in general. After 2 hr DHT treatment ~70% of up and down regulated genes contained at least one R-loop peak (**figure 4.10A**), similar to the proportions of R-loop positive genes described previously in neural stem cells (Thongthip *et al.*, 2022). Metagene plots showed that the DRIP-seq profiles were the same for all three treatment conditions in up and down regulated differentially expressed genes (**figure 4.10B**). Importantly these profiles followed a similar pattern to previously published DRIP-seq experiments with promoter enrichment of DRIP-seq signal and a reduction in signal at the TSS followed by a peak (Stork *et al.*, 2016; Promonet *et al.*, 2020; Wang *et al.*, 2021). After 24 hr DHT treatment there were significantly fewer R-loop positive genes in the upregulated gene set compared to downregulated genes (**figure 4.11A**). This was reflected in the metagene plot around the transcription start site where the vehicle treated DRIP-seq signal was higher than in either DHT treatments (**figure 4.11B**). This was also commensurate with the observed loss of promoter R-loops in **figure 4.8**. I attempted to functionally characterise R-loop positive and R-loop negative genes in GRO-seq and RNA-seq datasets using gene ontology analysis but none of these groups contained significantly enriched biological pathways. Taken together, these results imply that androgen treatment has a complex effect on R-loop dynamics. Most

R-loops are not androgen responsive and those that are do not necessarily correlate with androgen induced transcription.

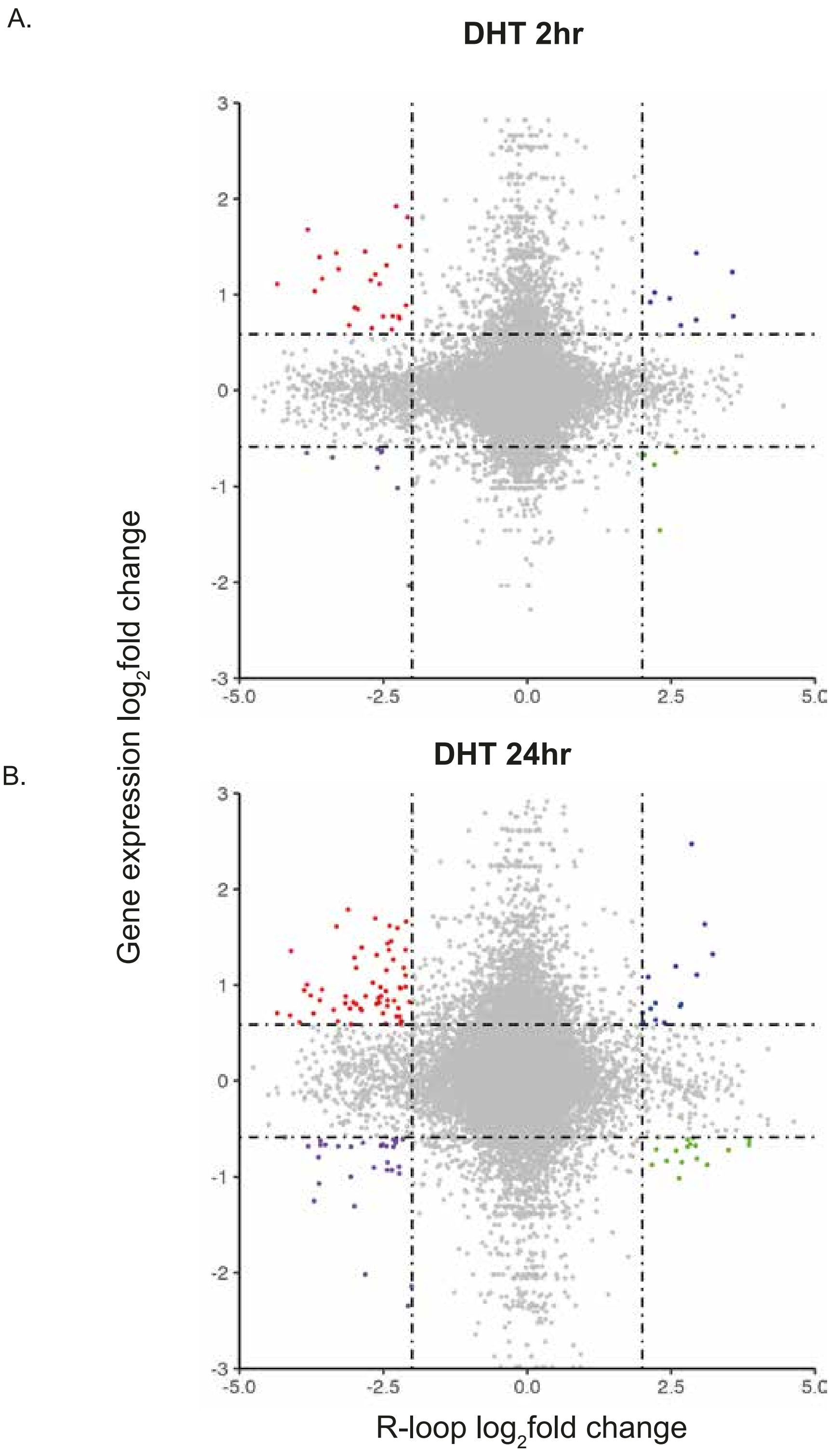


Figure 4.9 The relationship between change in gene epression and change in R_loop enrichment after DHT treatment

R-loops were assigned to a gene if they overlapped by at least 1 base pair. Each dot represents a gene - R-loop pair. A and B. DHT treatment for 2 and 24 hr respectively.

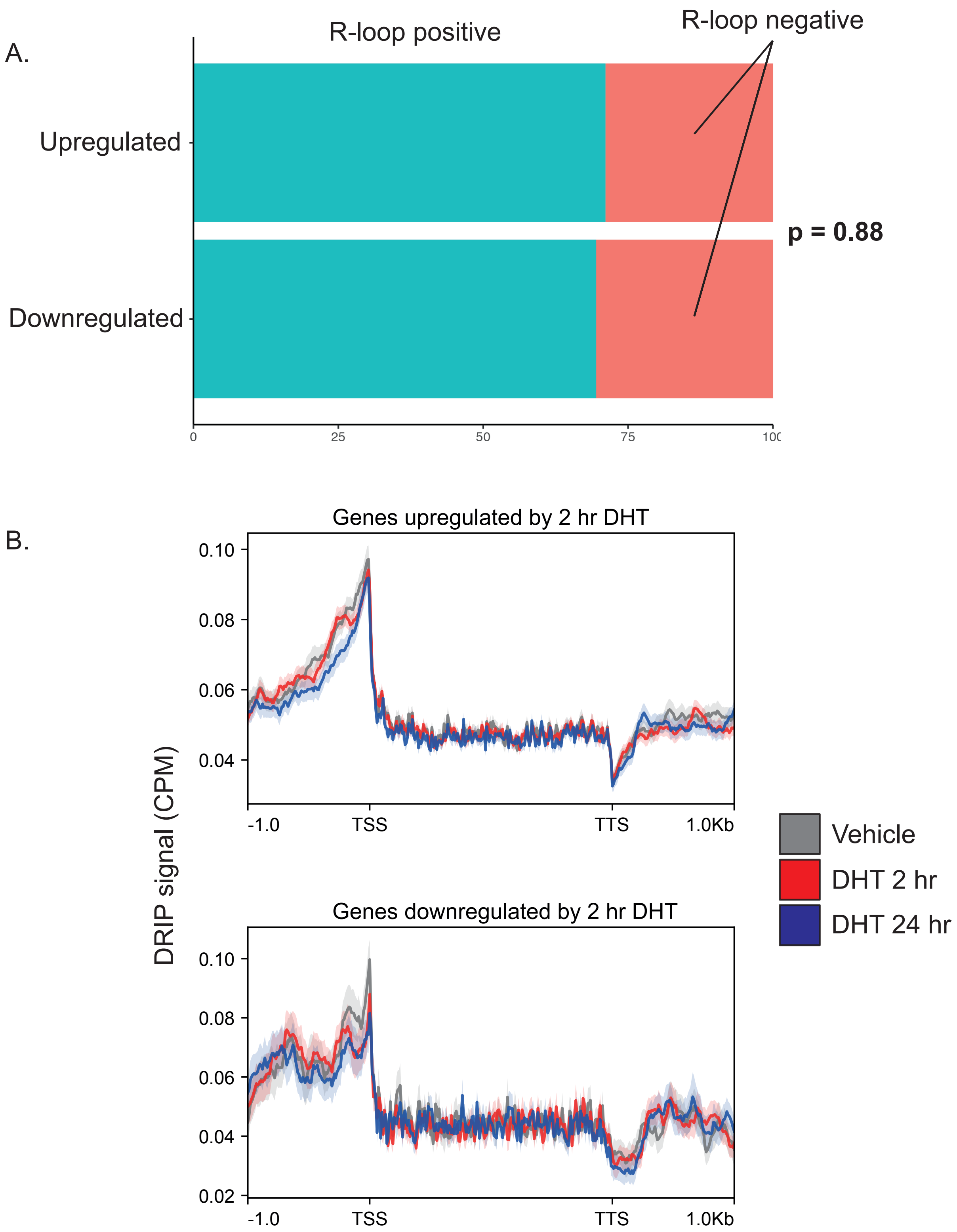


Figure 4.10 DRIP-seq profiles of genes differentially epressed by 2hr DHT
A. The proportion of up and down regulated genes that are R-loop positive or negative. R-loop positive genes were defined as genes with at least one R-loop peak B. DRIP-seq metagene profiles in genes up and down regulated by 2 hr DHT. TSS: Transcription start site, TTS: Transcription termination site. Ensembl 100 gene models were used. TSS and TTS were defined as the 5' and 3' coordinate of the primary Ensembl transcript.

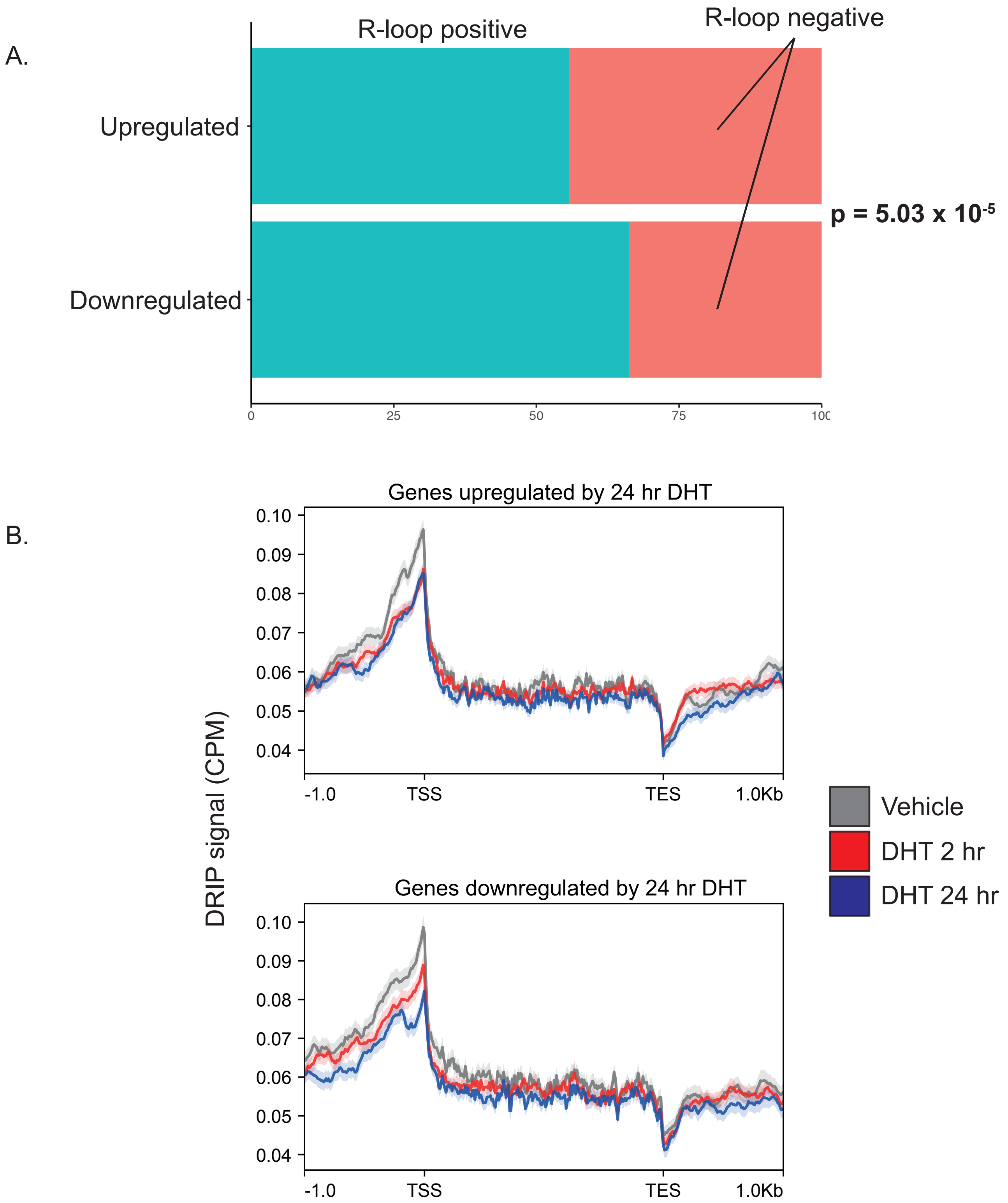


Figure 4.11 DRIP-seq profiles of genes differentially epressed by 24 hr DHT
A. The proportion of up and down regulated genes that are R-loop positiveor negative. B. DRIP-seq metagene profiles in genes up and down regulated by 24 hr DHT. TSS: Transcription start site, TTS: Transcription termination site. Ensembl 100 gene models were used. TSS and TTS were defined as the 5’ and 3’ coordinate of the primary Ensembl transcript.

4.2.6 DRIP-seq signal is reduced at androgen receptor binding sites

I next investigated intergenic R-loops. As R-loops have been demonstrated at enhancers (Wang *et al.*, 2021; Wulfridge and Sarma, 2021) and the androgen receptor is known to bind enhancers. I therefore hypothesised that intergenic R-loops might reflect transcription of enhancer RNA (eRNA) at androgen receptor binding sites. To test this I used publicly available androgen receptor ChIP-seq data from LNCaP cells treated with DHT for 2 hours (Malinen *et al.*, 2017). 19,672 AR peaks were present in this dataset, in keeping with the number of AR peaks observed in other studies. I observed a reduction in DRIP-seq signal around the centre of these peaks (**figure 4.12A**). Notably, the reduction was more marked in cells treated for 24 hr with DHT. I next tested if this reduction in DRIP-seq signal was a result of the underlying DNA sequence or of modulation of R-loops at that site. GC skew plots showed a negative GC skew 5′ of the androgen binding site followed by a rapid increase in GC skew to a positive value at the binding site itself. This pattern was present on plus and minus strands (**figure 4.12B**). The change in GC skew started approximately 250 bp away from the AR binding site implying that the 15 nucleotide dihexameric repeat plus 3 nucleotide spacer of the classical androgen response element is flanked by asymmetrical GC skew characteristics with low skew 5′ of the binding site and high skew on the 3′ side.

A study in LNCaP cells demonstrated that less than 10% of AR binding sites display enhancer activity upon androgen stimulation and a similar proportion display activity in the absence of androgens (Huang *et al.*, 2021). This study used self-transcribing active regulatory region sequencing (STARR-seq). This technique leverages massively parallel cloning of putative enhancer sequences into a plasmid which allows enhancers to transcribe themselves. This library is transfected into cells and the mRNA readout is the enhancer sequence itself which facilitates the quantitative measurement of enhancer activity in response to a stimulus. I downloaded bed files from this STARR-seq experiment of androgen receptor binding sites classified as inactive, constitutively active and inducible (by DHT). Plotting my DRIP-seq profiles separately for these regions showed a reduction in DRIP-seq signal around the centre of STARR-seq defined AR binding sites for vehicle and DHT treated cells (**figure 4.13A-C)**. Interestingly, the profile was the same for all three classes of enhancer. Overall, these data show that R-loops are reduced at androgen receptor binding sites but this effect is unchanged by androgen stimulation. This implies that binding of androgen receptor at these sites doesn't influence R-loop levels.

Another study by the same group showed that androgen receptor binding sites are commonly mutated in prostate cancer. In this study the authors used AR ChIP-seq profiles from LNCaP cells and clinical cancer samples to define androgen receptor binding sites and demonstrated a higher density of SNVs at these sites compared to other transcription factors and chromatin binding proteins (Morova *et al.*, 2020). The proposed mechanism was a steric hindrance effect of the androgen receptor blocking the DNA damage repair machinery from repairing transcription associated DNA damage. A second study showed a significant peak of single nucleotide variants at AR binding sites in primary and metastatic prostate cancer (Huang *et al.*, 2021). Using the commonly mutated AR sites (n=938) from the first of these studies as the centre point for metagene analysis I observed a variable DRIP-seq signal profile over the 4 kb region surrounding the site of the mutation. There was no obvious association of DRIP-seq signal and the mutation location (**figure 4.14**). These data suggest that changes in R-loop profile around ARBs does not change at sites of androgen receptor associated DNA damage and therefore R-loops are unlikely to play a role in this phenomenon.

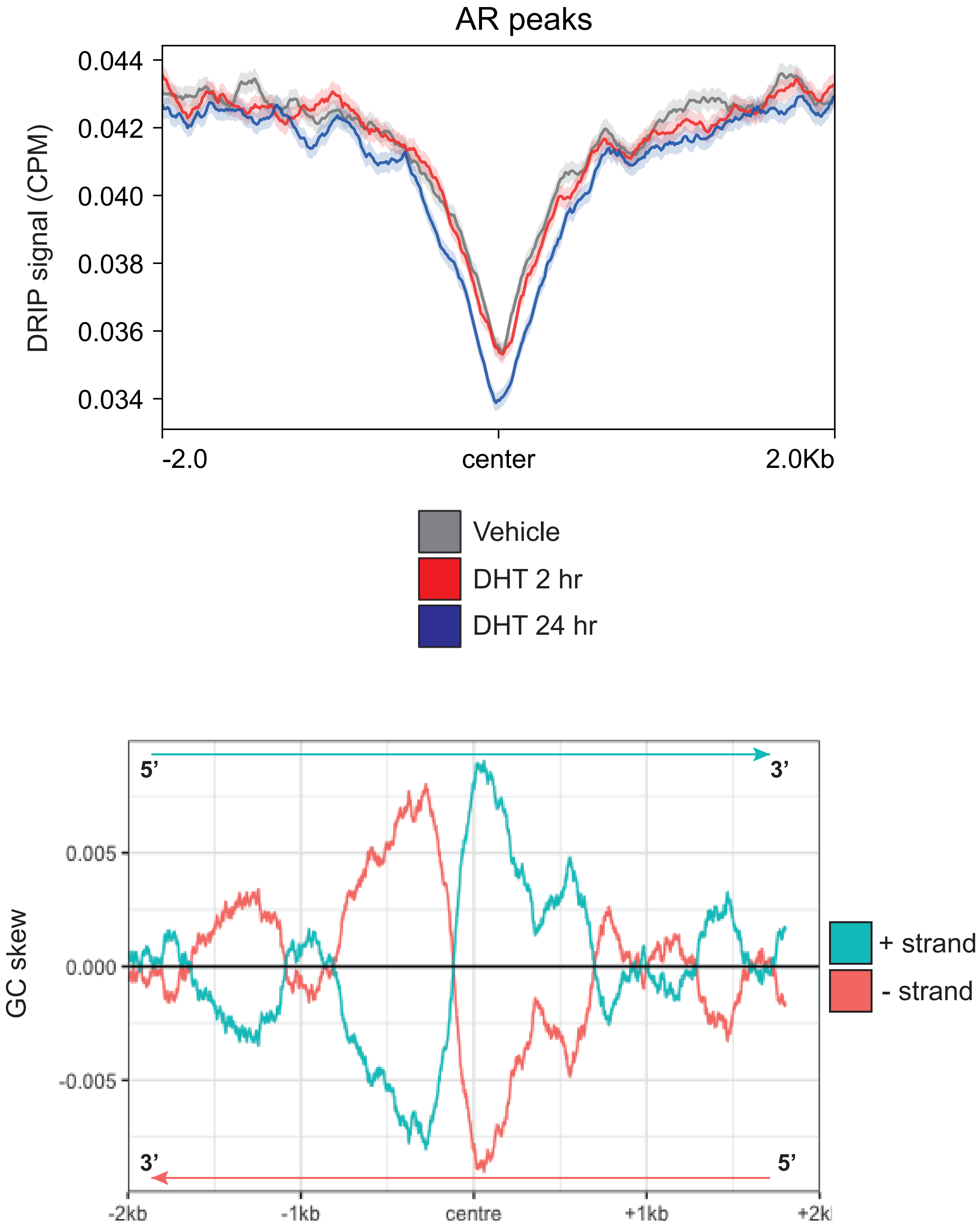


Figure 4.12 DRIP-seq profiles around androgen receptor binding sites
A. Metagene of DRIP-seq signal centered around AR binding sites (n=19672) in LNCaP cells. B. GC skew profile over the same regions as in A.

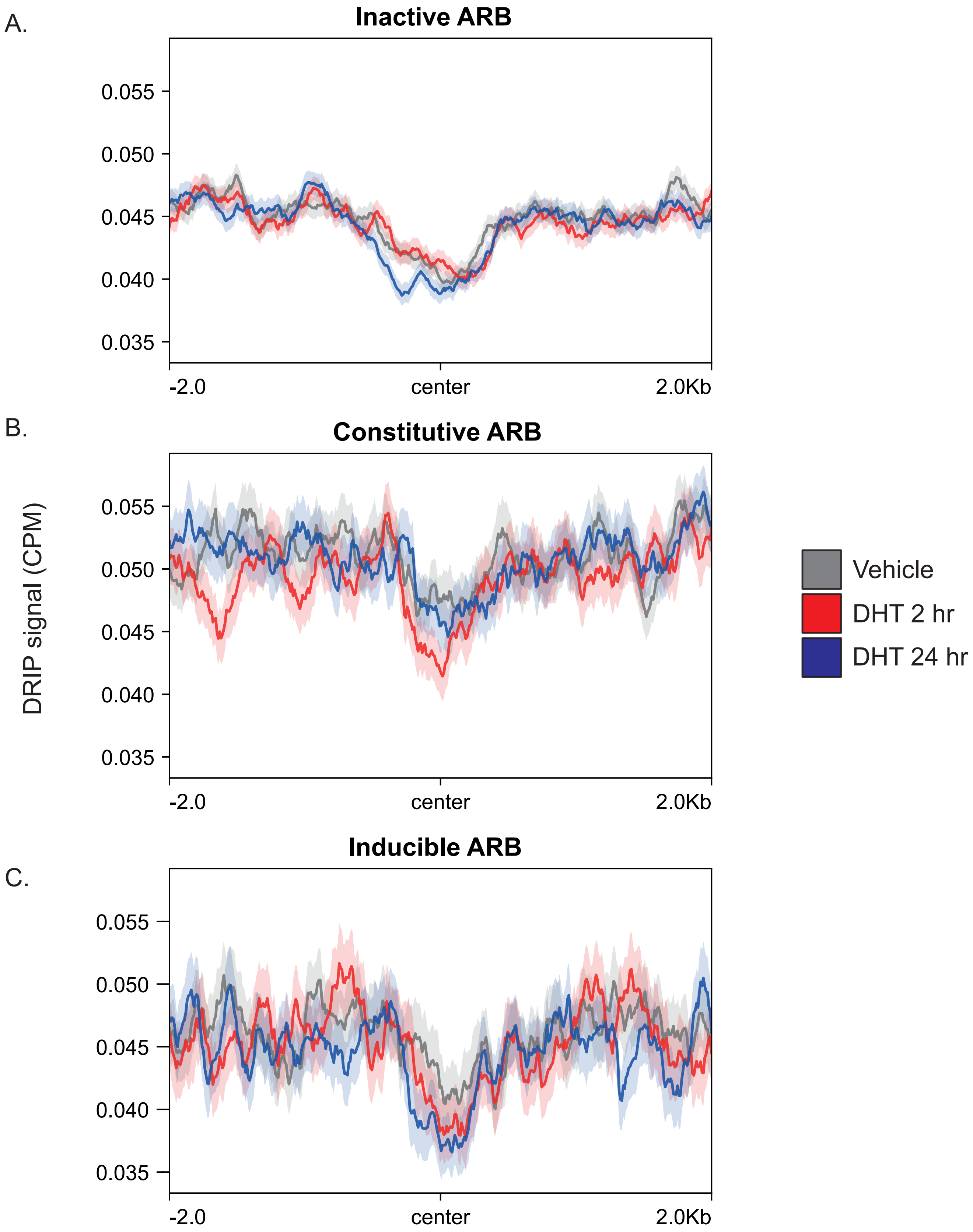


Figure 4.13 DRIP-seq profile around AR bound enhancers from STARR-seq in LNCaP cells
A. Inactive AR bound enhancers (n = 2479). B. Constitutively active AR bound enhancers (n = 465). C. AR bound enhancers induced by DHT (n = 286).

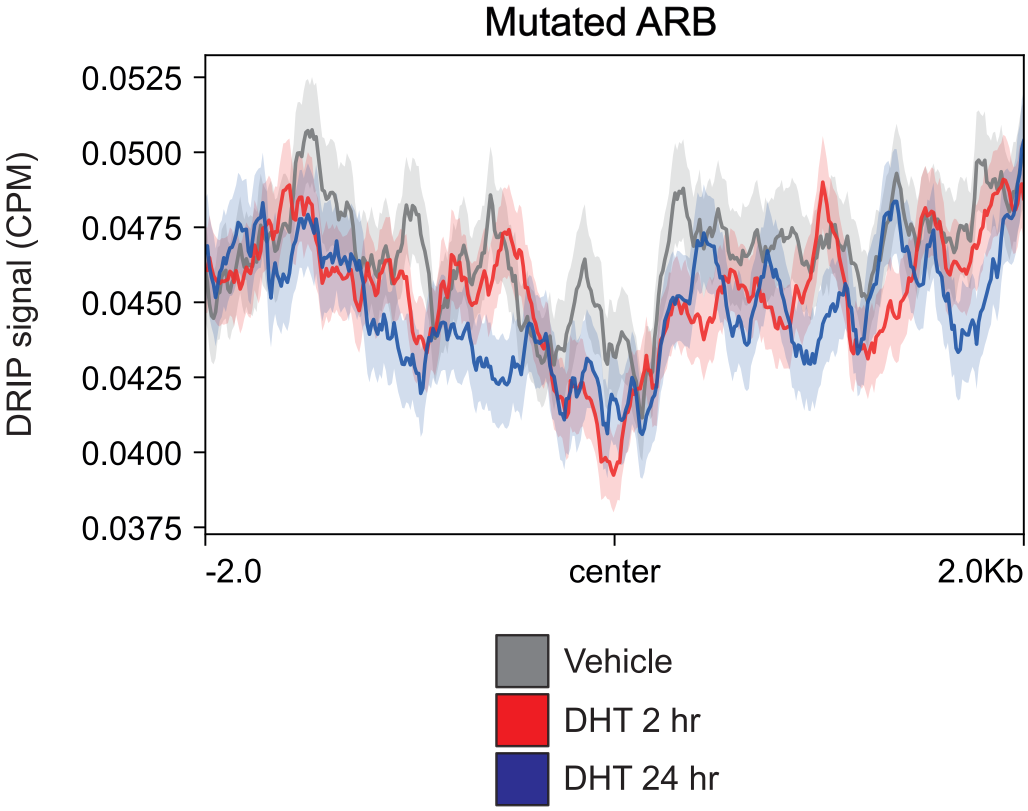


Figure 4.14 DRIP-seq profiles around commonly mutated AR binding sites
Genomic coordinates of commonly mutated androgen receptor binding sites (ARB) were taken from Morova *et al* (2020). DRIP-seq signal was plotted using DeepTools using the mutated ARBs as the centre point and a flank of 2 kb up- and downstream.

4.2.7 Validation of DRIP-seq results

I next attempted to validate the DRIP-seq findings with DRIP-qPCR for specific loci. Eight loci were selected to reflect the combinations of R-loop dynamics, transcriptional activity and genomic context (**table 4.1**). Primers were designed to target R-loops peaks differentially enriched by androgen treatment in these genes.

Table 4.1 Loci selected for DRIP-qPCR validation

Locus	Feature	Genomic coordinates	Response to DHT in DRIP-seq
Intergenic			
IG1	Intergenic	chr6:11964835-11965537	Upregulated
IG2	Intergenic	chr7:128,095,001-128,096,896	Downregulated
Genic from DRIP-seq			
CBX1	Promoter	chr17:48,100,544-48,101,663	Upregulated
NATL8	last exon/3’ UTR	chr4:2,062,417-2,064,976	Upregulated
KLF14	Promoter	chr7:130,733,343-130,734,414	Downregulated
TUSC1	Promoter	chr9:25,678,069-25,678,533	Downregulated
Genic androgen responsive genes			
NKX3-1	Promoter	chr8:23,681,260-23,683,212	Upregulated
KLK3	Gene body	chr19:50,854,676-50,860,987	No change

I first tested the R-loop accumulation in four genes (CBX1, NATL8, KLF14 and TUSC1) and two intergenic regions. CBX1 is involved in heterochromatin formation and its promoter is differentially enriched for the repressive chromatin mark H3K27me3 in prostate cancer tissues compared to normal samples (Ngollo *et al.*, 2017). NATL8 transcript levels were increased in the progression to castration resistant prostate cancer. The protein product N-acteyl aspartate synthetase catalyses the production of N-acetyl aspartate which has recently been shown to accumulate in cell culture models of CRPC compared to their isogenic parental counterparts (Salji *et al.*, 2022). KLF14 was upregulated in a mouse model of CRPC which in turn led to the upregulation of pathways protective of oxidative stress (Luo *et al.*, 2019).

In the DRIP-seq dataset the CBX1 and NATL8 R-loops were enriched after 2hr and 24 hr DHT treatment however no statistically significant increase in DRIP-qPCR signal was observed in the validation experiment. Conversely, KLF14 and TUSC1 both displayed a loss of R-loop enrichment upon DHT treatment in the DRIP-seq dataset but the reverse pattern was observed in DRIP-qPCR validation. Neither of the selected intergenic R-loop regions showed differential R-loop enrichment upon treatment with DHT in the DRIP-qPCR validation experiment (**figure 4.15**). One of the unexpected findings in the DRIP-seq data was that transcription either at the nascent or mature transcript level did not correlation with R-loop accumulation. However, one androgen regulated gene – NKX3.1 – did show correlation of transcriptional and R-loop upregulation in response to androgens. A second well-characterised androgen responsive gene – KLK3 – did not exhibit any change in R-loop dynamics in response to DHT despite an increase in transcript levels. I successfully validated both findings using DRIP-qPCR (**figure 4.16 A and B**).

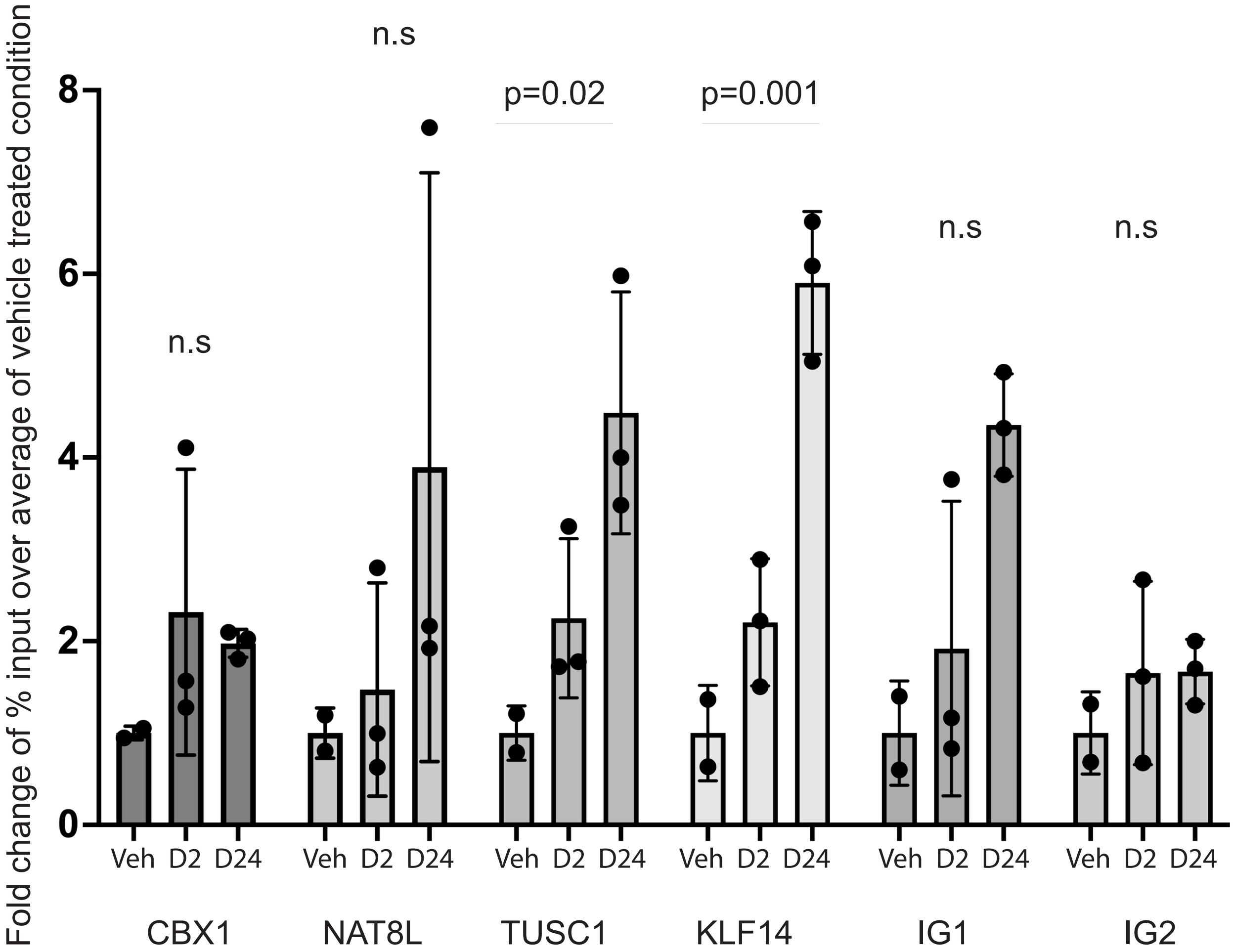


Figure 4.15 Validation of androgen responsive R-loop loci by DRIP-qPCR
LNCaP cells were stimulated with 100 nM DHT for 2 or 24 hr and R-loop occupancy at the indicated loci determined by DRIP-qPCR. All values are normalised to the vehicle treated condition for each locus. p-values derived from one-way ANOVA with Dunnet correction for multiple testing. n.s: not significant. Veh: Vehicle treated, D2: 100 nM DHT for 2 hr, D24: 100 nM DHT for 24 hr. IG1: Intergenic 1 locus, IG2: Intergenic 2 locus. See table 4.1 for further details

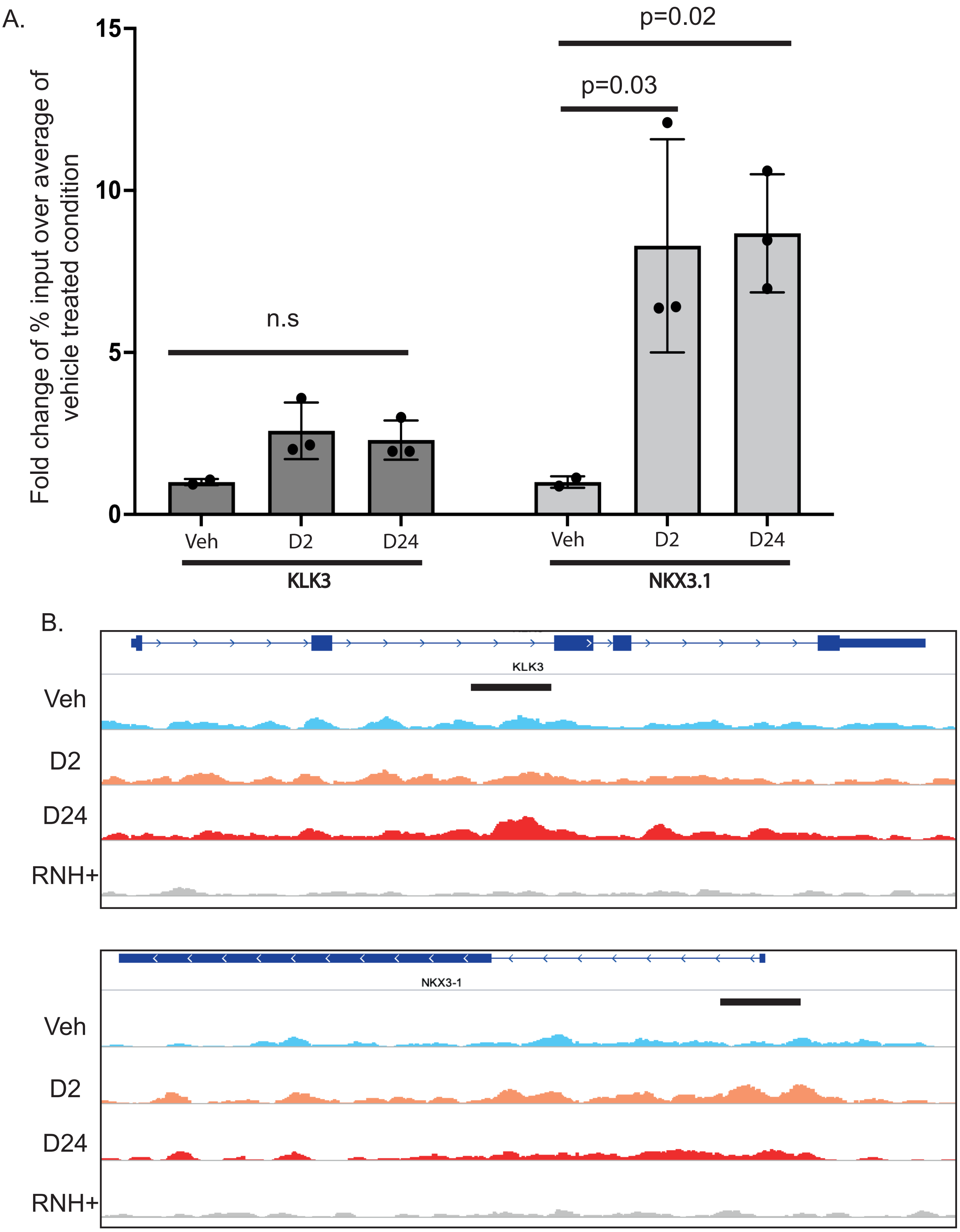


Figure 4.16 Validation of NKX3.1 and KLK3 R-loop dynamics by DRIP-qPCR
A. DRIP-qPCR of the KLK3 and NKX3.1 R-loop loci. P values are derived from one-way ANOVA with Dunnet test for multiple correction. n.s: not significant B. IGV tracks of DRIP-seq signal for KLK3 and NKX3.1. The black bar indicates the region amplified by qPCR primers. Veh: Vehicle treated cells, D2: DHT 2hrs, D24: DHT 24hrs, RNH+: Sample pre-treated with RNase H1 as a specificity control.

4.2.8 Attempts at alternative methods of R-loop profiling

Given the unexpected lack of association between transcriptional output and R-loop dynamics together with the difficulties in validating my results by qPCR I decided to attempt orthogonal methods of R-loop profiling. DRIP-seq is the most widely-used method of genome-wide R-loop profiling. A recent review showed that 58% of all R-loop profiling experiments published to date used DRIP-seq and the remaining studies used one of ten other techniques (R. Lin *et al.*, 2022). However, despite its popularity DRIP-seq has some limitations. Firstly, DRIP-seq is not a stranded technique and therefore R-loops that map to genomic annotations on both strands such as a gene within a gene, bidirectional promoters or closely arranged 5’ and 3’ gene ends create difficulties for deciding which gene that R-loop originated from. Secondly, genomic DNA is fragmented using five restriction enzymes in DRIP-seq. This creates fragments ranging 0.2 – 5 kb in length. The cut sites of these enzymes is biased towards intergenic regions creating smaller intergenic fragments and larger genic fragments (Halász *et al.*, 2017). Furthermore, the cutting efficiency of each enzyme is difficult to determine and given that most restriction enzymes don’t efficiently cleave R-loops (Kisiala *et al.*, 2020), chromatin fragments pulled down by the S9.6 antibody may contain an R-loop but also contain long B form DNA tails that will then proceed to library preparation and be spuriously classed as part of an R-loop. Indeed, DRIP protocols that use sonication rather than enzyme digest report narrower median R-loop peak lengths (Sanz, Castillo-Guzman and Chédin, 2021). At the beginning of this project, DRIP-seq using enzyme digest was the most well described procedure however over the last four years multiple other approaches have been described and the DRIP technique has been refined. I attempted to validate my DRIP-seq findings using a modification of DRIP-seq called sonication-DRIP or sDRIP (Sanz, Castillo-Guzman and Chédin, 2021). This method is described fully in chapter 2. Briefly, cells are harvested, lysed and treated with proteinase K prior to genomic DNA extraction by phenol-chloroform and ethanol precipitation. The lysate is then sonicated to produce 300-500 bp DNA fragments and this lysate proceeds directly to immunoprecipitation with the S9.6 antibody. Second strand synthesis using dUTP, RNase H treatment and library preparation with uracil N- glycosylase is employed to give the library strand specificity. The underlying principle is that sonication removes the looped-out strand of DNA present in an R-loop and RNase H removed the RNA moiety, leaving a single strand of DNA for amplification.

These experiments were performed in conjunction with Lorna Gilroy-Turner, a master's student who I supervised. We first attempted to optimise sonication conditions. We found that the genomic DNA was very sensitive to small changes in sonication conditions. Initial tests of 15 cycles 30s on/ 30s off using a DIagenode Bioruptor caused over fragmentation of DNA with most fragments less than 100 bp and not suitable for immunoprecipitation. We tried five, six and seven cycles of sonication (**figure 4.17A**). Six cycles gave the best distribution of DNA fragments and this DNA was taken forward for immunoprecipitation. Using the same primer pairs for the positive control loci (RPL13A and TFPT) and the negative control locus (SNRPN) we found a generally low immunoprecipitation efficiency with RPL13A showing an identical enrichment to the negative control locus SNRPN (**figure 4.17B**; *c.f.* **figure 4.2** where RPL13A was enriched 15-20 fold more than SNRPN). Given the difficulties of sonication I also attempted a strand specific version of DRIP (DRIPc) where DNA fragmentation is done with restriction enzymes then immunoprecipitated with S9.6 as per the well-described DRIP method. The sample is then treated with DNAse leaving just the RNA moiety from R-loops which is taken forward for strand-specific 2nd strand synthesis and library preparation yielding stranded and higher resolution DRIP-seq libraries (Sanz and Chédin, 2019; Stolz, Sulthana, Stella R. Hartono, *et al.*, 2019). Using one sample from each experimental condition (vehicle treated, DHT for 2 hours and DHT for 24 hours) I observed a robust enrichment of DRIP signal at the RPL13A locus for the 24hr sample but very little enrichment for vehicle treated or 2 hour DHT samples. The TFPT locus showed a fold change of less than one at all timepoints (**figure 4.17C**) indicating that RNA recovered from these R-loop loci had been lost in reverse transcription. Both sonication DRIP and DRIPc required large numbers of cells per condition (3 and 5 15 cm plates respectively) and correspondingly large volumes of reagents for DNA extraction, purification and immunoprecipitation. In addition, each experiment takes 10 days from seeding cells to immunoprecipitated DNA that can be analysed by qPCR or sequencing. Given these time and resource parameters, I made the judgement that there was insufficient benefit in optimising an alternative R-loop profiling strategy and these experiments were abandoned.

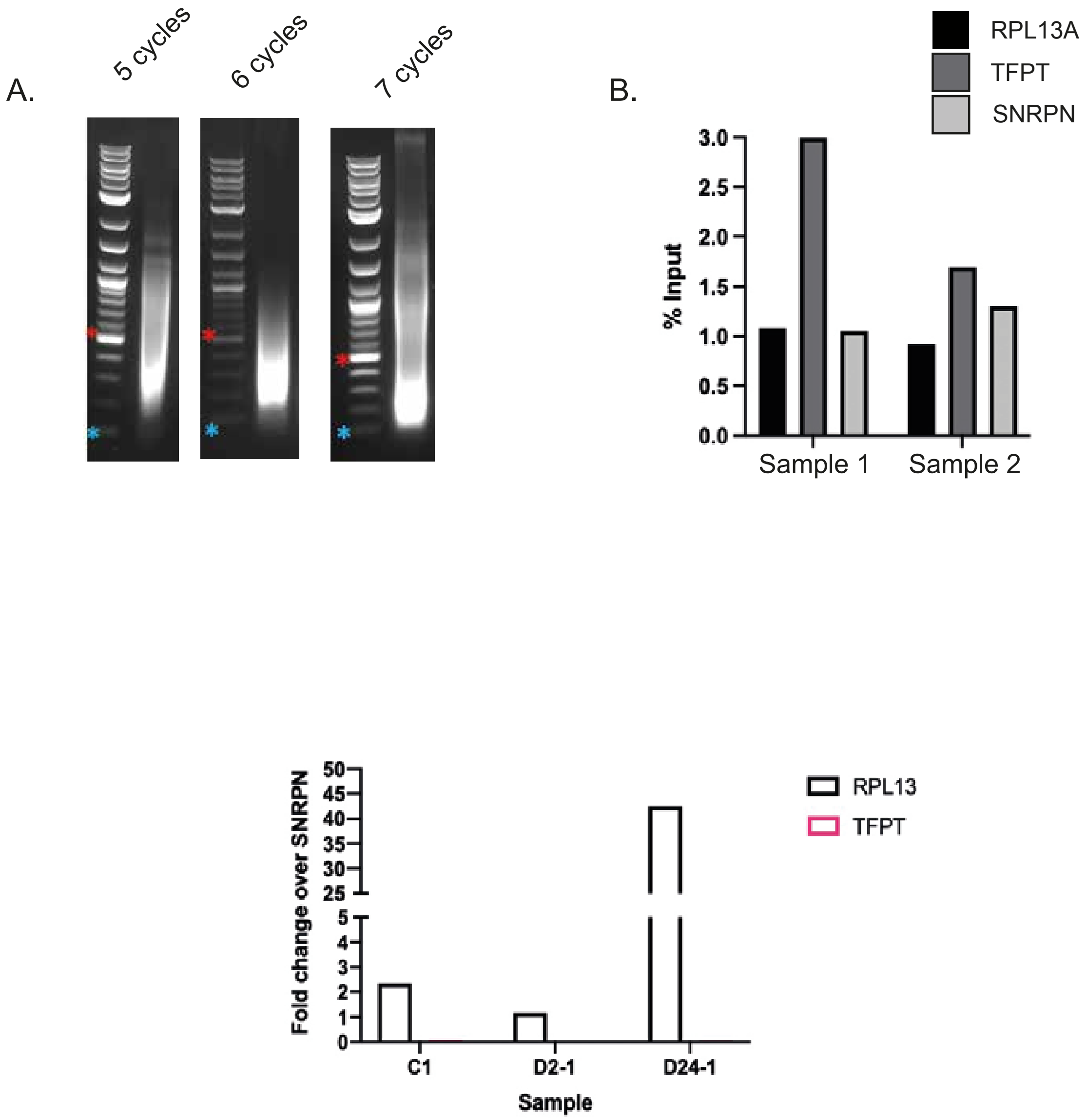


Figure 4.17 Attempted optimisation of strand specific R-loop mapping techniques
A. Sonicated genomic DNA analysed on a 0.8% agarose/ ethidium bromide gel. Blue and red asterisks denote 100 and 500 bp markers respectively. Sonication cycle numbers are indicated above the corresponding gel image. B. DRIP-qPCR for R-loop loci in RPL13A, TFPT and SNRPN. C. Fold change in DRIP signal over the negative control locus (SNRPN) following immunoprecipitation and reverse transcription.

4.3 Discussion

4.3.1 R-loop formation decreases with DHT treatment and does not correlate with transcription

In this chapter I have used DRIP-seq to delineate the first genome-wide map of R-loop dynamics in response to androgen stimulation. Unexpectedly, DHT treatment induced a net loss of R-loops and that loss appeared to be more prevalent in promoter regions. Moreover, no significant correlation with transcription was seen. These findings suggest that the relationship between androgens and R-loops is not as simple as androgens increasing transcription and therefore increasing R-loop levels. Instead, the loss of R-loops in the setting of upregulated transcription and *vice versa* (displayed in **figure 4.9**) implies a degree of R-loop regulation. This regulation could occur by modulation of the factors that resolve R-loops. Alternatively R-loop regulation could be influenced by the up- or down-regulation of RNA processing factors that make R-loop formation by nascent RNA less likely by changing RNA secondary structure thereby altering nascent RNA dwell time and the likelihood of R-loop formation. My findings are in contrast to the result of oestrogen stimulation of breast cancer MCF-7 cells (Sanz *et al.*, 2016). In this study oestrogen exposure for 2 and 24 hours increased R-loop levels globally as measured by S9.6 slot and also increased the number of R-loops peaks detected by DRIP-seq. Furthermore, R-loops were upregulated in canonical oestrogen responsive genes such as SLC7A5 and GREB. The oestrogen and androgen receptors are both type 1 nuclear hormone transcription factors that exert their function through ligand binding, translocation to the nucleus and activation of transcriptional programmes. Interestingly, there was a correlation between transcription and R-loop occupancy only at the 2 hour timepoint and the correlation coefficient was 0.16 but statistically significant. By 24 hours of oestrogen stimulation there was no correlation between mature transcript levels and R-loop accumulation. My data suggest that an increase in R-loops is not a general phenomenon of nuclear hormone signalling. One limitation of both my DRIP-seq data and that from the stimulation of MCF-7 cells is the use of one cell line. Whilst the LNCaP cell line is commonly used for studies of androgen signalling (Abate-Shen and Nunes de Almeida, 2022), it might be that different androgen signalling contexts such as androgen receptor amplification present in VCaP or LNCaP-AR cells (Chatterjee *et al.*, 2019) could promote transcriptional programmes with different R-loop dynamics.

Few R-loop mapping studies have investigated the relationship between differential gene expression and differential R-loop formation in response to a stimulus. Recently Li et al (2022) showed that approximately 50% of differentially expressed genes had a positively correlated differential enrichment in R-loops across a time course of reprogramming mouse embryonic fibroblasts to induced pluripotent stem cells. Reprogramming is accompanied by large scale epigenetic and gene expression changes within a cell (Buganim *et al.*, 2012) and so represents a large scale shift in the composition of differentially expressed gene sets. In the Li dataset approximately 5000 genes were differentially expressed between day 3 and day 0 of reprogramming. By contrast relatively few genes are differentially expressed by androgen treatment of LNCaP cells: 461 genes and 747 genes at 2 and 24 hours after DHT respectively in the datasets used in this chapter. These data point to most dynamic R-loop changes being present in non-differentially expressed genes. R-loops have been implicated in the regulation of epigenetic mechanisms such as CpG island methylation (Arab *et al.*, 2019) and histone modifications (Argaud *et al.*, 2019). One explanation for my data could be that R-loops are differentially regulated at genes whose expression is not regulated by androgens to ensure these genes remain in a steady state of transcription in the face of attempted direct or indirect transcriptional up regulation. As R-loop accumulation can act as a negative or positive regulator of transcription dynamics (Chédin, 2016) their formation could act to suppress androgen signalling activity at certain genes. Whilst this is speculative it could be tested by selecting genes with differential enrichment of R-loops by DHT without a change in nascent or mature transcript levels. By manipulating R-loops through over-expression or local targeting of an R-loop resolving factor such as RNase H1 (Abraham *et al.*, 2020) I could then determine if the formation and resolution of these specific R-loops was important for maintaining a steady state of transcription. Unfortunately, as I was unable to validate such loci by DRIP-qPCR, I could not take this hypothesis forward for testing.

4.3.2 R-loops are reduced at androgen receptor binding sites

As there was no correlation between gene transcript levels and R-loop accumulation in response to DHT I next asked if R-loops might associate with enhancer RNA transcription at androgen receptor binding sites. By integrating DRIP-seq profiles and androgen receptor bound enhancers identified from a functional screen of ARB enhancers I showed a decrease in R-loop signal at the point of AR binding regardless of the class of enhancer or treatment with DHT. Furthermore, I showed that GC skew is not symmetrical around ARBS. Unfortunately, I was unable to investigate this further as attempts at optimising strand specific R-loop profiling were unsuccessful. A study of stranded R-loop formation recently revealed that active and poised enhancers in mouse embryonic stem cells form unidirectional R-loops despite active bidirectional nascent transcription. This correlates with the unidirectional positive GC skew that I observed at AR bound enhancers and this warrants further investigation to determine if there is corresponding unidirectional R-loop formation and if this influences AR bound enhancer activity.

4.3.3. Difficulties in validation of DRIP-seq by DRIP-qPCR

The quality control data I have presented for DRIP-seq (**figures 4.2 – 4.5**) strongly suggests that this experiment detected *bone fide* R-loops. However, I was unable to validate many of the selected loci by DRIP-qPCR. The most significant global result of R-loop reduction upon treatment with DHT could not be validated by qPCR and this hindered my ability to take these loci forward for further characterisation e.g. by ChIP-qPCR for dynamic histone modifications, R-loop binding proteins and the effect of R-loop resolution by the overexpression of R-loop helicases. The reasons for this failed validation are not clear. The main difference between DRIP-seq and DRIP-qPCR is the DNA fragment size that is analysed. In DRIP-seq DNA is fragmented by sonication after immunoprecipitation. In DRIP-qPCR no further DNA fragmentation is performed after the initial restriction enzyme digest. Therefore, different loci may have a different sized fragments from which their much smaller amplicon is amplified. However, this is similar to RT-qPCR where long mRNA transcripts are reverse transcribed then probed with primers amplifying relatively short amplicons. The extra fragmentation step of DRIP-seq may have introduced bias into which loci were represented in the final analysed data however this does not correlate with the robust quality control data I presented for this experiment. As I couldn't find a reasonable explanation for the differences in DRIP-seq I attempted two alternative methods of DRIP that removed enzyme digestion in the case of sDRIP or generated higher-resolution strand-specific R-loop maps in the case of DRIPc. Neither of these approaches were successful. The sonication parameters were difficult to optimise in sDRIP and genomic DNA extracted from LNCaP cells was sensitive to small changes in the number of sonication cycles. This was reflected in the lack of DRIP signal for RPL13A; shown to be an R-loop ‘hotspot’ in my own DRIP-seq data and numerous published studies (e.g. Kotsantis *et al.*, 2016; Stork *et al.*, 2016; Sanz and Chédin, 2019). In the case of DRIPc I could not reliably reverse transcribe RNA isolated from R-loops immunoprecipitated by the S9.6 antibody. This may have been due to RNase activity either endogenously within the cell lysate or introduced as a contaminant during the experimental procedure. A second possibility is that the DNase digestion step had unintended activity on the RNA moiety of extracted R-loops and reduced the RNA yield.

4.3.3 NKX3.1 is an androgen regulated gene with androgen regulated R-loops

DRIP-qPCR did successfully validate two R-loop loci: NKX3.1 and KLK3. NKX3.1 is a prostate specific tumour suppressor with roles in coordinating the DNA damage repair response, modulating the mitochondrial response to reactive oxygen species and also an androgen receptor pioneer factor (reviewed in Griffin *et al.*, 2022). NKX3.1 expression is increased by androgen signalling and is well characterised as a direct androgen receptor target gene (He *et al.*, 1997). It is thought to act as a prostate specification and differentiation factor, an idea supported by experiments in mice where ectopic NKX3.1 expression in seminal vesical respecified this tissue to prostate (Dutta *et al.*, 2016). Furthermore, genomic loss followed by inactivation/silencing of the remaining allele or post transcriptional silencing is a common event in progression to castration resistant prostate cancer (Sooreshjani *et al.*, 2021). KLK3 is one the most well characterised androgen responsive genes and frequently used as a model locus for studying androgen receptor activity (Clark *et al.*, 2008, 2013; Hsieh *et al.*, 2014). KLK3 encodes prostate specific antigen (PSA), secreted from prostate tissue after injury, inflammation and in prostate cancer. PSA can be detected in the blood and has been posited as a screening test for prostate cancer however a large meta-analysis of PSA-based screening demonstrated no overall 10-year survival benefit in men screened using PSA vs. standard care (Ilic *et al.*, 2018). PSA is recommended for use by NICE in follow-up after radical treatment to detect disease recurrence and/or progression (National Institute for Health and Care Excellence [NICE], 2019).

My DRIP-seq and DRIP-qPCR data showed that R-loops were upregulated by DHT treatment at 2 and 24 hours, particularly in the promoter region of NKX3.1. However, for KLK3 no change in R-loop occupancy was observed with DHT treatment, despite gene expression induction by DHT. Both loci showed a defined R-loop peak implying that R-loops form by transcription through the NKX3.1 locus but R-loops are present at a steady state in KLK3. This could mean that R-loop homeostasis plays different roles at these two loci even though they are both transcribed in response to androgens. Interestingly, other androgen responsive loci such as TMPRSS2, FKBP5 and ELOVL5 also formed R-loops in their promoters but these did not change dynamically with DHT treatment. Overall, this suggests that an increase in R-loop formation is restricted at androgen responsive loci. This could relate to efficient RNA processing or the activities of R-loop resolving enzymes at these loci possibly to avoid R-loop related DNA damage or the stalling effect of R-loops on RNA polymerase 2. NKX3.1 and KLK3 provide a useful pair of loci to explore this disconnect between androgen induced transcription and R-loop formation and I explore this in chapter 6.

Chapter 5. RNAse H1 overexpression alters the transcriptional programme of prostate cancer cells

5.1 Introduction

In chapters 3 and 4 I identified potential changes in R-loop occupancy across the genome under different androgen signalling conditions. As I could only validate two of the DRIP-seq loci by DRIP-qPCR (KLK3 an NKX3.1) I sought an orthogonal approach for asking if androgen induced R-loop formation was involved in androgen regulated gene expression. RNAse H1 over-expression has previously been used to probe the functional roles of R-loops in gene expression (Chen *et al.*, 2015; Argaud *et al.*, 2019; Sabino *et al.*, 2022). RNAse H1 is a ribonuclease that specifically cleaves RNA-DNA hybrids containing at least four consecutive ribonucleotides differentiating it from RNAse H2 that removes single ribonucleotides embedded in DNA (Nowotny *et al.*, 2007). I hypothesised that RNase H1 overexpression would reduce R-loop formation. If androgen regulated genes required R-loop formation as part of their transcription cycle their transcription would be dysregulated. This would manifest as altered mRNA levels when treated with DHT in the context of RNAse H1 overexpression.

Objectives

- Demonstrate successful RNase H1 overexpression in LNCaP cells
- Analyse differential gene expression in the setting of RNase H1 overexpression using RNA-seq
- Characterise genes with dysregulated expression

5.2 Results

5.2.1 RNase H1 can be over expressed in LNCaP cells to reduce global R-loop abundance

LNCaP cells were treated with 100 nM DHT or vehicle for 24 hours on a background of transient overexpression of a GFP-only (empty vector GFP; EV-GFP) or a RNase H1-GFP construct (RNH-GFP) (experimental outline is given in **figure 5.1A**). Transfection conditions were first optimised then successful overexpression of RNase H1 was confirmed by Western blot and live-cell immunofluorescence (**figure 5.1B**).

The functional effect of RNase H1 over expression was assessed by immunofluorescence using the S9.6 antibody (**figure 5.1C**). Cells transfected with the EV-GFP construct retained a strong nuclear S9.6 signal. This signal was reduced by transfection with the RNH1-GFP plasmid. Interestingly, *in vitro* treatment of the fixed and permeabilised cells with recombinant RNase H isolated from *E. Coli* further reduced the S9.6 signal in both the EV-GFP and RNH-GFP conditions. This implies that whilst over expression of RNH1 does reduce R-loop levels, there are some R-loops that may be resistant to RNH1 over expression in the native intra-cellular environment.

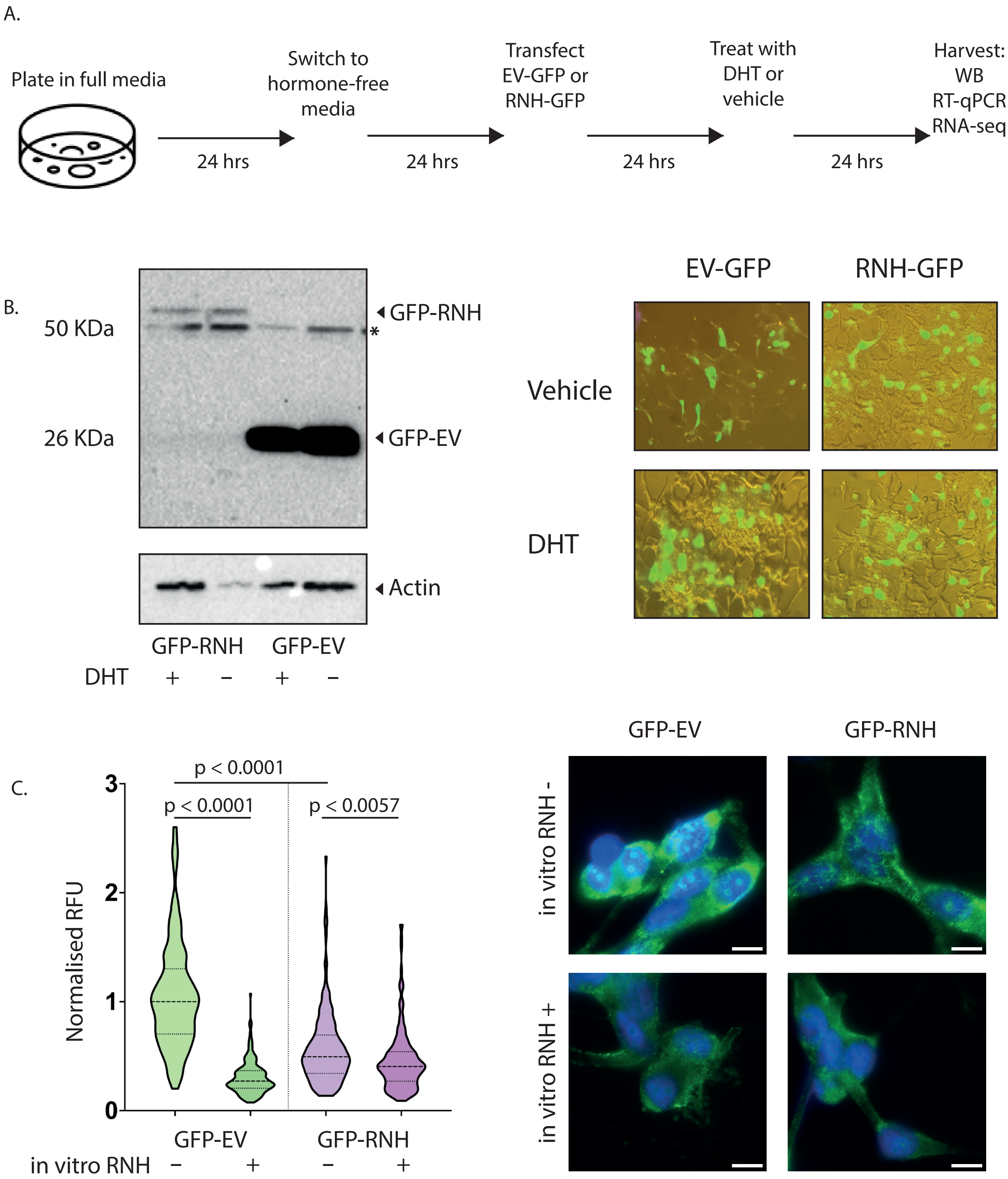


Figure 5.1 RNase H1 can be over expressed in LNCaP cells and reduces global R-loop abundance A. Experimental strategy. WB: Western blot. B. Left: Western blot confirming over expression of the RNH or EV constructs. The asterisk denotes a non-specific band which may represent a GFP dimer. Right: Live cell immunofluoresence was used as an initial qaulity control to ensure GFP expression prior to harvesting. C. Left: Violin plot of S9.6 immunofluorescence signal in cells with over expression of GFP-EV (green) or GFP-RNH (purple) subsequently treated *in vitro* with recombinant RNase H1. Pooled data from three biological repeats. P values are derived from one-way ANOVA and post-hoc Tukey test for all pairwise comparisons with correction for multiple testing. Right: Representative immunofluorescence images. Scale bar = 10 μm.

5.2.2 Quality control of stranded RNA-seq

Having established successful over-expression of RNase H1 in LNCaP cells I next tested if this influenced transcriptome-wide androgen regulated gene expression using stranded RNA-seq. Four combinations of conditions were represented: EV-GFP + vehicle for 24 hr, EV-GFP + 100 nM DHT for 24 hr, RNH-GFP + vehicle for 24 hr and RNH-GFP + 100 nM DHT for 24 hr. An initial analysis of RNase H gene expression demonstrated a robust increase in RNase H mRNA levels compared to empty vector treated cells. Importantly, the addition of DHT did not affect RNase H expression in cells transfected with the EV-GFP construct (**fig 5.2A**). Principal component analysis demonstrated clustering of biological replicates, implying good experimental reproducibility (**fig 5.2B**). Furthermore, the first two principal components (PC1 and PC2 in **fig 5.2B** respectively) accounted for 94% of the total variance. Interestingly the four combinations of conditions occupied different quadrants of the PCA plot and there was no clustering by RNAse H1 overexpression status. This suggests that RNase H1 over expression and DHT treatment may have separate effects. As further quality control biological replicates of each condition showed significant pairwise correlation of expression across all transcribed genes (**fig 5.2C**). Taken together these data show that the experimental setup was reproducible.

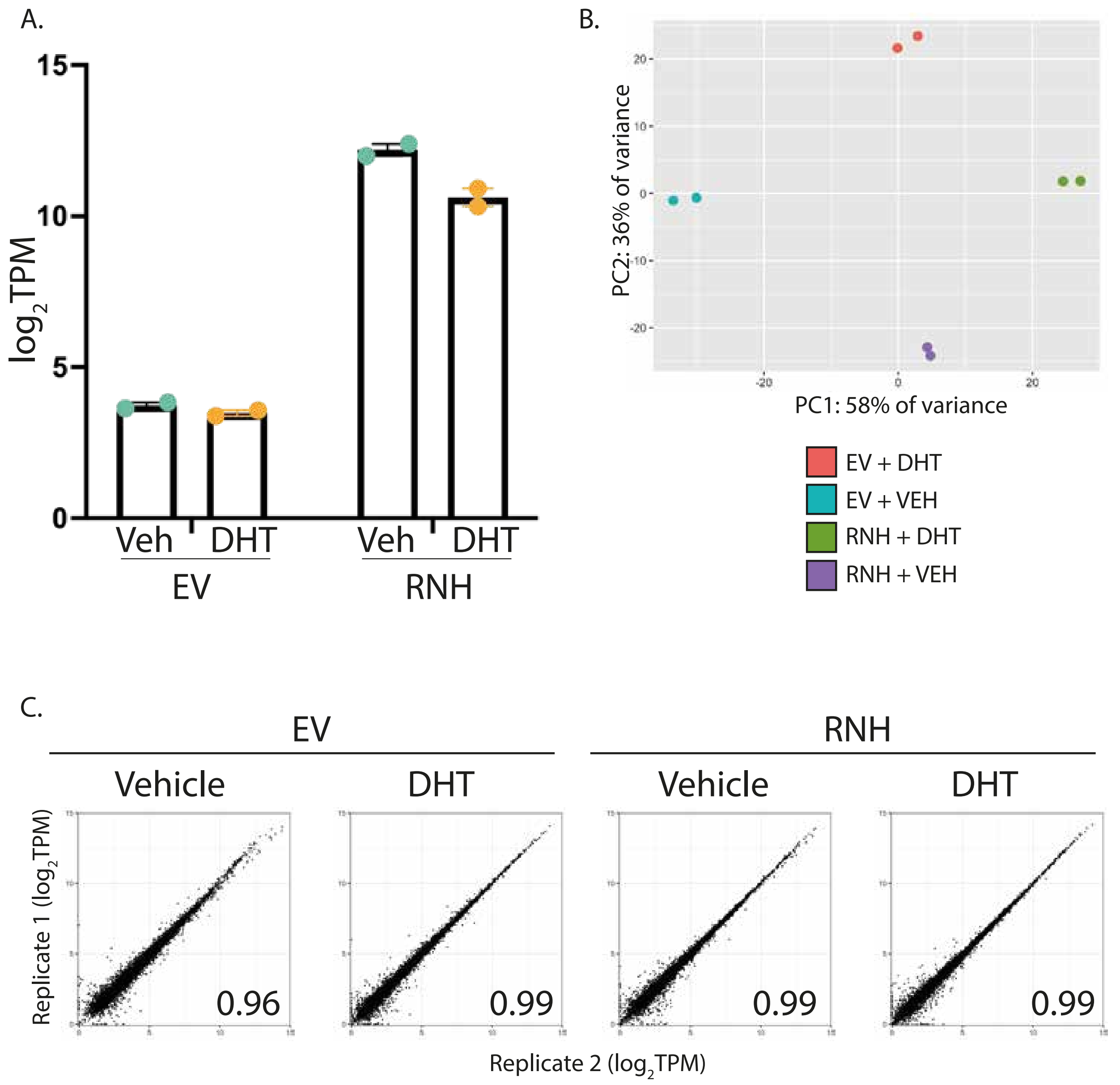


Figure 5.2 Quality control of stranded RNA-seq data
A.Normalised mRNA expression levels of RNase H1 from RNA-seq data. B. Principal component analysis plot showing clustering of biological replicates from the same condition. C. Correlation of expression levels (log2 TPM) between replicate 1 (x axis) and replicate 2 (y axis) for each condition. Pearson correlation coefficient is given for each comparison.

5.2.3 Overexpression of RNase H has varied effects on androgen regulated gene expression

To investigate the effect of RNase H1 overexpression on androgen regulated gene expression I used the R package DEseq2 (Love, Huber and Anders, 2014) to quantify fold changes in gene expression. I specified the design formula as

$$\sim condition + treatment + condition * treatment$$

where condition is over expression of empty vector or RNase H1 and treatment is vehicle or DHT. The final term (condition*treatment) is included to test the combined effect of RNase H1 overexpression and DHT treatment within the linear model used to generate fold change values between comparisons. When LNCaP cells were stimulated with DHT for 24 hours following transfection with the empty vector GFP construct, 881 genes were upregulated and 1739 genes were down regulated (**figure 5.3A**). Gene ontology analysis using the EnrichGO R package showed that upregulated genes were enriched for biological process terms including 'steroid biosynthetic process', 'cholesterol biosynthetic process' and terms relating to mitochondrial gene transcription and translation as described previously (Massie *et al.*, 2011; **figure 5.3B**). This shows that transfection with the empty vector did not significantly alter the baseline response to androgen. I next used interaction plots to test if RNase H1 over expression changed the androgen response of four well-characterised androgen responsive genes: NKX3.1, KLK3, TMPRSS2 and FKBP5. Expression of all four genes was upregulated in response to androgen and RNase H1 overexpression did not change the magnitude or direction of response **(figure 5.4)**

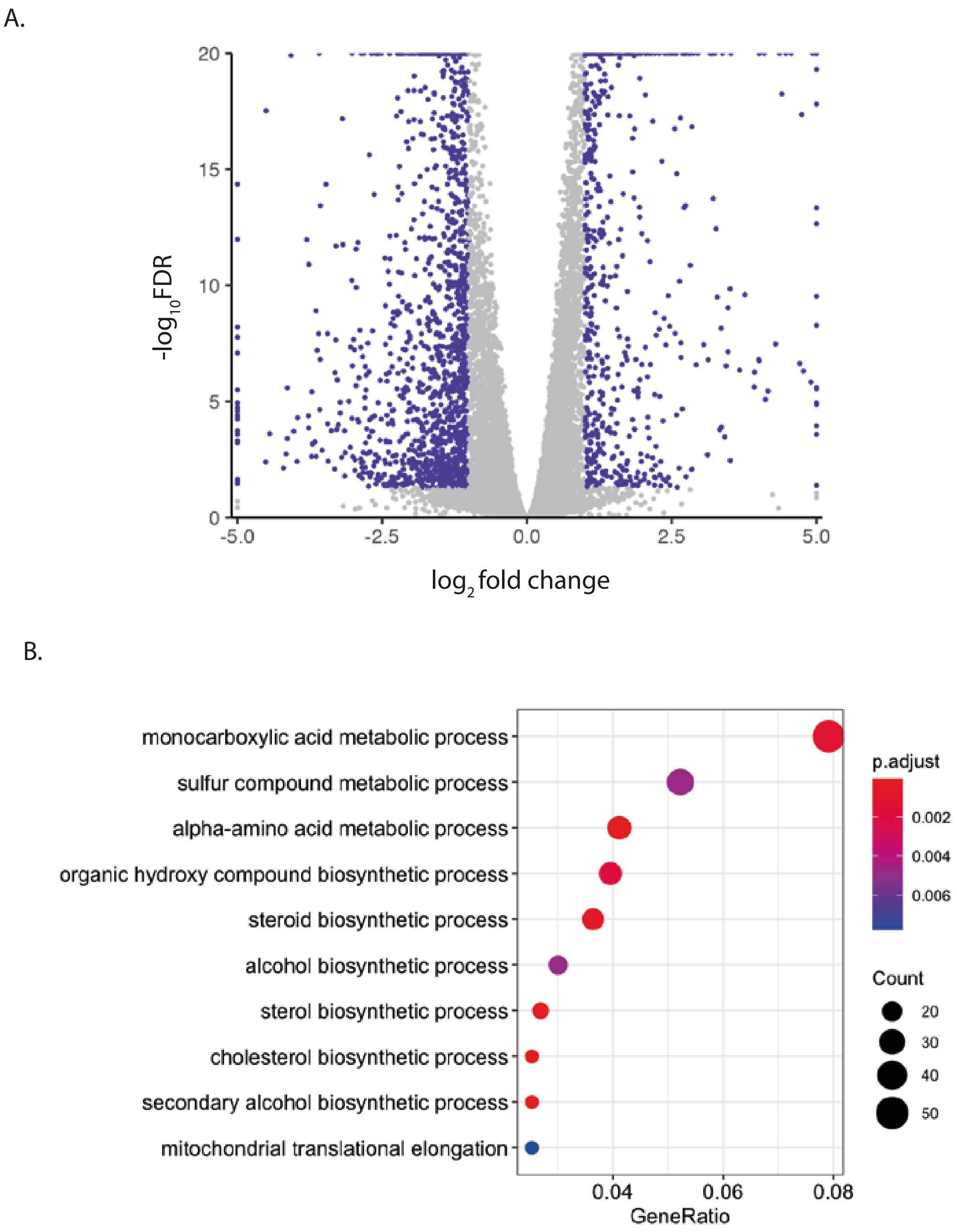


Figure 5.3 Transfection with empty vector does not alter the androgen repsonse of LNCaP cells A.Volcano plot of genes significantly differentially expressed in response to 100 nM DHT (absolute log2 fold change > 1, adjusted p value < 0.05). B. Dotplot of biological processes enriched in upregulated genes from EnrichGO gene ontology analysis.

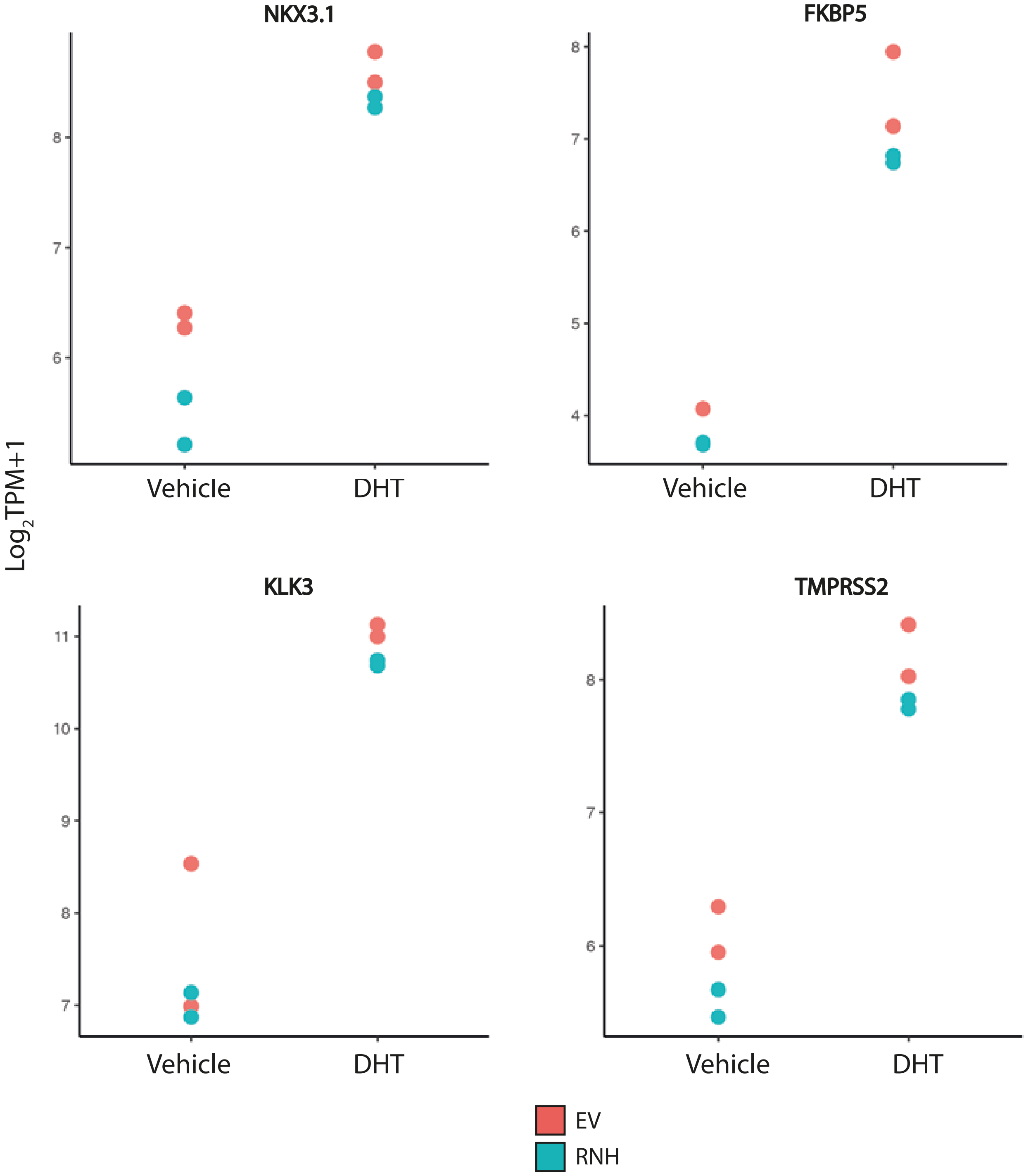


Figure 5.4 RNAse H overexpression does not alter expression of canonical androgen induced genes Interaction plots of expression of NKX3.1, FKBP5, KLK3 and TMPRSS2 showing an increase in mRNA abundance after the addition of 100 nM DHT for 24 hours not significantly modified by RNH1 over expression.

As differential expression analysis can only assess one dichotomous comparison at a time, I selected a series of different comparisons or 'contrasts' for analysis to probe whether RNAse H1 over expression affected the expression of any androgen regulated genes.

5.2.3.1 RNase H1 over expression does not significantly alter expression of genes also differentially expressed when transfected with empty vector

To test the overall similarity of androgen regulated genes between empty vector (EV) and RNAse H1 over expression (RNH1-OE) gene sets I asked what proportion of genes were significantly differentially expressed in both conditions after treatment with DHT. Of differentially expressed genes in the EV condition, 976/2179 (45%) were shared with genes differentially expressed in the RNH1-OE condition. These shared genes had significantly correlated fold changes (**figure 5.5**) implying that RNase H1 overexpression doesn't change the magnitude of response for this gene set. However, there were four genes where the direction of expression changed between EV and RNH1-OE conditions (**table 5.1** and annotated in **figure 5.5**). Of these four genes, two (ALX4 and RUNX2) showed a decrease in expression with DHT treatment in the EV condition. With RNH1-OE these genes had a reduced expression level when treated with vehicle compared to the same treatment with EV. Their expression rose when treated with DHT under RNH-OE1 conditions but this rise was smaller than the corresponding fall under EV conditions when assessed using the interaction plots in **figure 5.6**. This is not reflected in the fold change values in table 5.1. Here the rise in expression in RNH1-OE appears greater than the fall in expression under EV. This is due to very low count values in RNH-OE1 vehicle treated cells making the divisor in the fold change calculation less than 1 and therefore spuriously increasing the fold change value.

The remaining two genes – ITIH6 and LRRC37A16P – showed large variance in the EV vehicle and EV DHT samples respectively (**figure 5.6)**. The effect of RNase H1 overexpression is therefore difficult to discern. Notably all expression levels for these two genes were less than 1 $\log_2$TPM+1 indicating generally low expression of these genes in all conditions. Genes with low expression are known to have more variable expression in RNA-seq experiments (Love, Huber and Anders, 2014) and so these results are likely to represent experimental/ technical variation rather than a true biological effect.

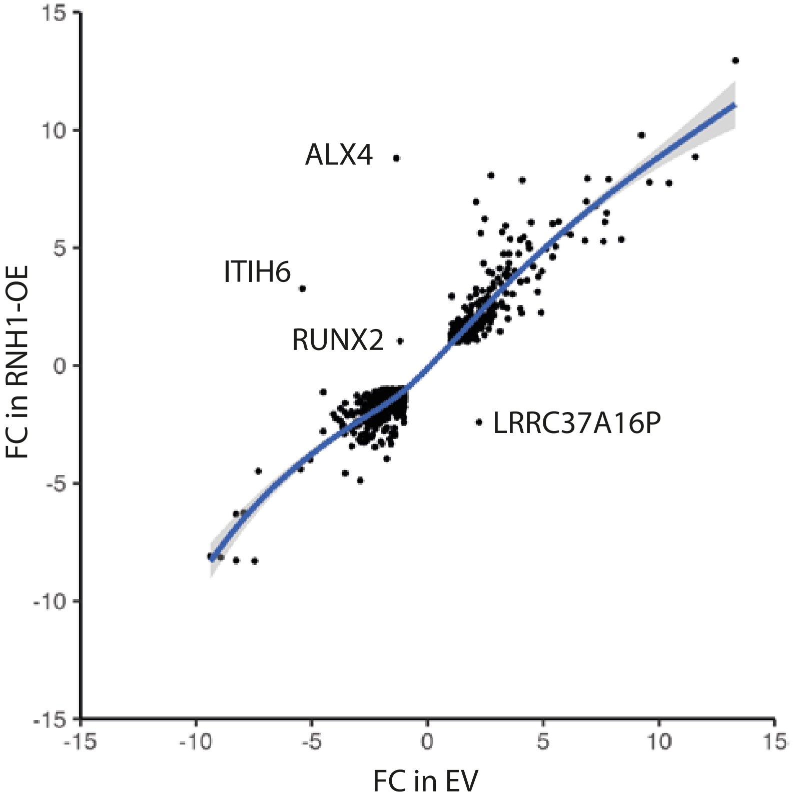


Figure 5.5 Rnase H1 over expression does not alter the magnitude or direction of relative mRNA abundance in genes differentially expressed in empty vector and RNH1 over expression conditions

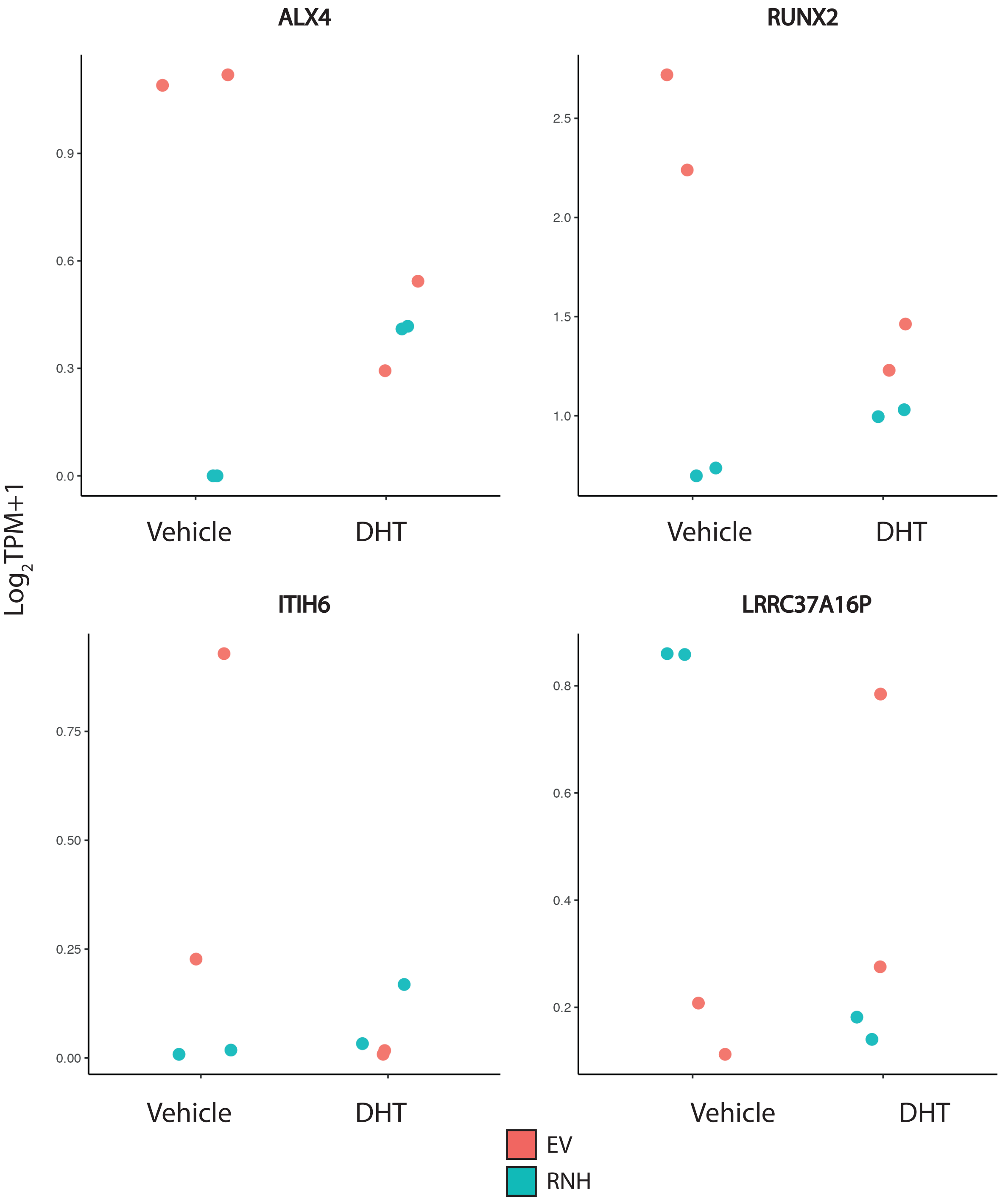


Figure 5.6 Interaction plots of four genes where the direction of expression changed with RNH1 over expression

Table 5.1 Genes differentially expressed in both EV and RNH1-OE conditions whose fold change direction changes between the two datasets. * indicates genes where change in fold change is likely to be due to experimental rather than biological variation. See also figure 5.6. FC = $\log_2$ fold change.

Gene	Biotype	FC in EV	FC in RNH1-OE	Change with RNH OE
RUNX2	Protein coding	-1.18	+1.04	Increase
ALX4	Protein coding	-1.34	+8.81	Increase
ITIH6	Protein coding	-5.42	+3.26	Increase*
LRRC37A16P	Transcribed unprocessed pseudogene	+2.22	-2.40	Decrease*

5.2.3.2 RNase H1 overexpression reduces expression of genes involved in mitochondrial processes

I next characterised the differentially expressed genes unique to the EV and RNH1-OE conditions. Fifty-five percent (n = 1203) of the EV gene set showed no differential expression in the RNH1-OE gene set. Of these genes there was a similar proportion of up and down regulated genes compared to the complete EV gene set (69% vs. 73% of each set downregulated) indicating that the effect of RNAse H1 overexpression was not biased towards up or down regulated genes. These genes also had a significantly lower fold change magnitude (e.g. down regulated genes had a less negative fold change and upregulated genes had a less positive fold change) compared to genes that were differentially expressed under EV and RNH1-OE conditions (**figure 5.7**). This suggests that the largest effects of androgen stimulation are not changed by RNase H1 overexpression. Further characterisation by gene ontology analysis showed that shared genes were enriched for previously described androgen regulated processes including steroid, sterol and cholesterol biosynthetic processes (**figure 5.8A**). Interestingly, genes that were uniquely differentially expressed in the EV gene set were enriched for mitochondrial biological processes, protein translation elongation and termination and MHC class 2 protein complex assembly (**figure 5.8B**). This implies that these processes were down regulated by the combination of RNase H1 overexpression and DHT treatment. RNase H1 is active in the nucleus and the mitochondria, and recently has been shown to regulate transcription of mitochondrial genes (Reyes *et al.*, 2020).

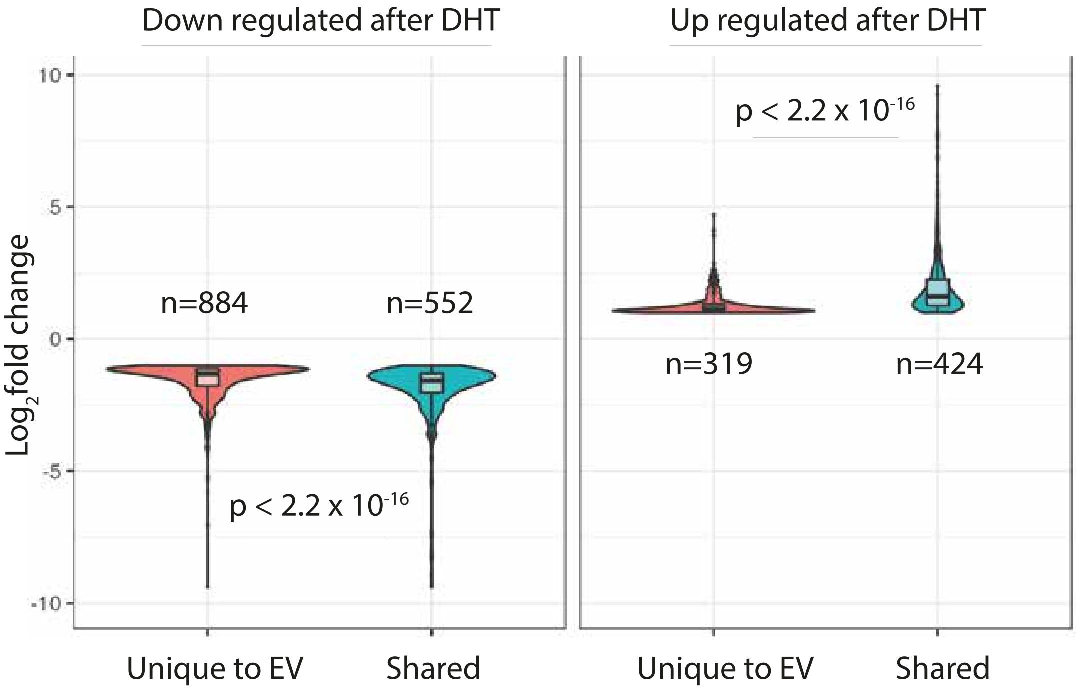


Figure 5.7 Genes uniquely differentially expressed by DHT under EV conditions have a lower magnitude of expression compared to those differentially expressed under EV and RNH1-OE conditions

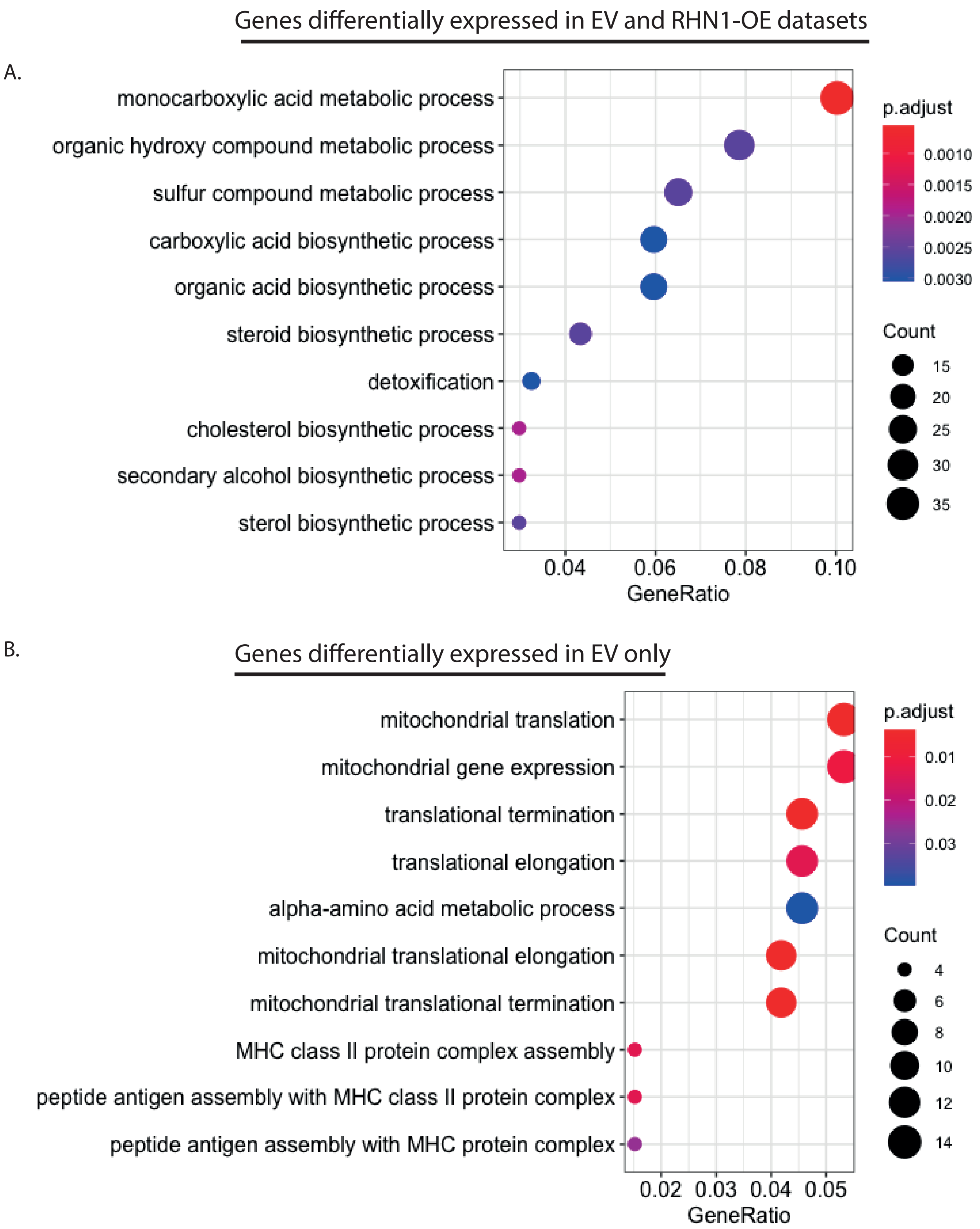


Figure 5.8 Gene ontology analysis
A. Biological processes enriched in genes differentially expressed after DHT in EV and RNH1-OE datasets
B. Biological processes enriched in genes differentially expressed after DHT in EV dataset only

I applied the same analysis to genes that were uniquely differentially expressed in RNH1-OE cells after DHT treatment. There were fewer genes differentially expressed in this set compared to the EV gene set (1319 vs. 2179). Of these 1319 genes, 343 genes (26%) were unique to RNH1-OE. These genes had a lower magnitude fold change compared to the shared gene set. Whilst this was statistically significant (**figure 5.9**), the difference in median log2 fold changes was less than that seen when comparing shared and unique genes in EV cells. Gene set enrichment analysis showed that only one biological process – 'neurotransmitter reuptake' was enriched in the unique RNH1-OE gene set. Only four of the 343 unique genes contributed to this enrichment, implying that the remaining genes were not part of specific pathways.

To determine if the effects of RNase H1 overexpression were related to R-loops or to downstream changes in signalling secondary to the combination of RNH1-OE and DHT I used RloopTools to characterise the unique gene sets described above (**figure 5.10 A-C**). Compared to EV-transfected cells, RNH1-OE cells had a similar proportion of RLFS positive differentially expressed genes. The EV differential expression gene set had a significantly higher number of RLFS per gene but did not have a significantly different percent coverage by RLFS. These data suggest that the genes uniquely expressed by EV and DHT have a greater propensity to form R-loops. This could mean that the effects observed with RNase H1 overexpression do not represent an interaction between R-loop resolution and DHT signalling but are the result of cell stress from RNase H1 over expression.

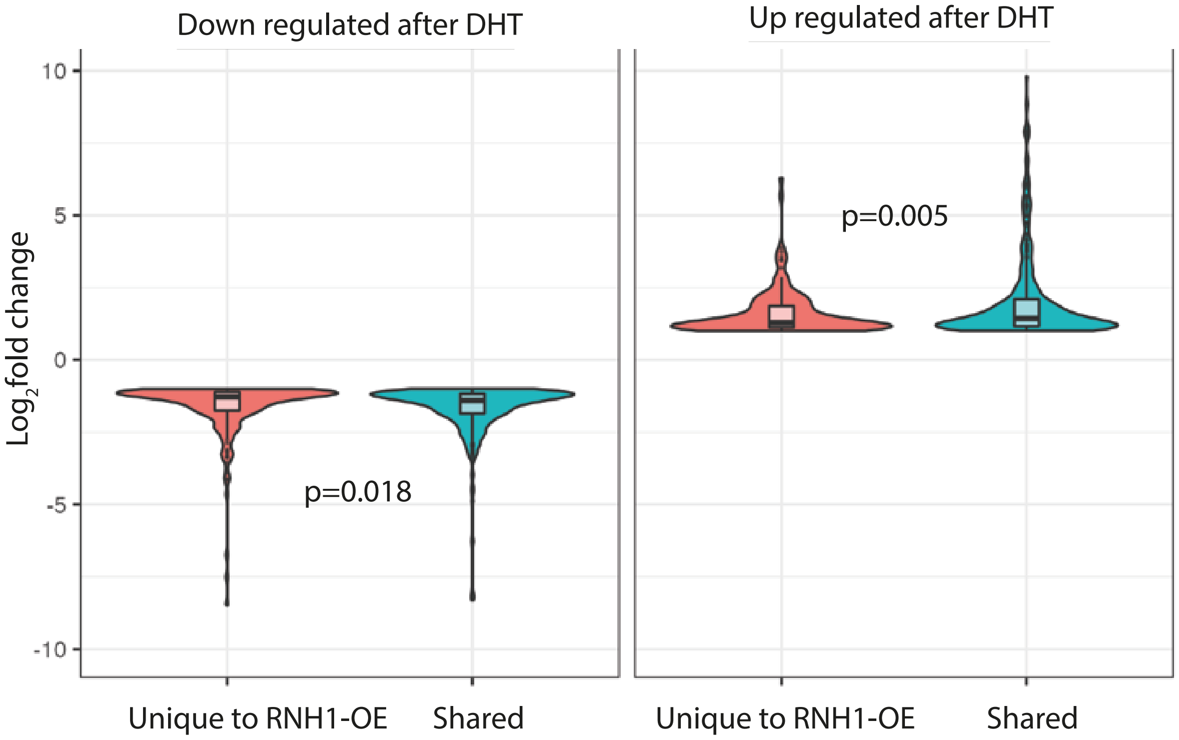


Figure 5.9 Genes uniquely differentially expressed by DHT under RNH1-OE conditions have a lower magnitude of expression compared to those differentially expressed under EV and RNH1-OE conditions

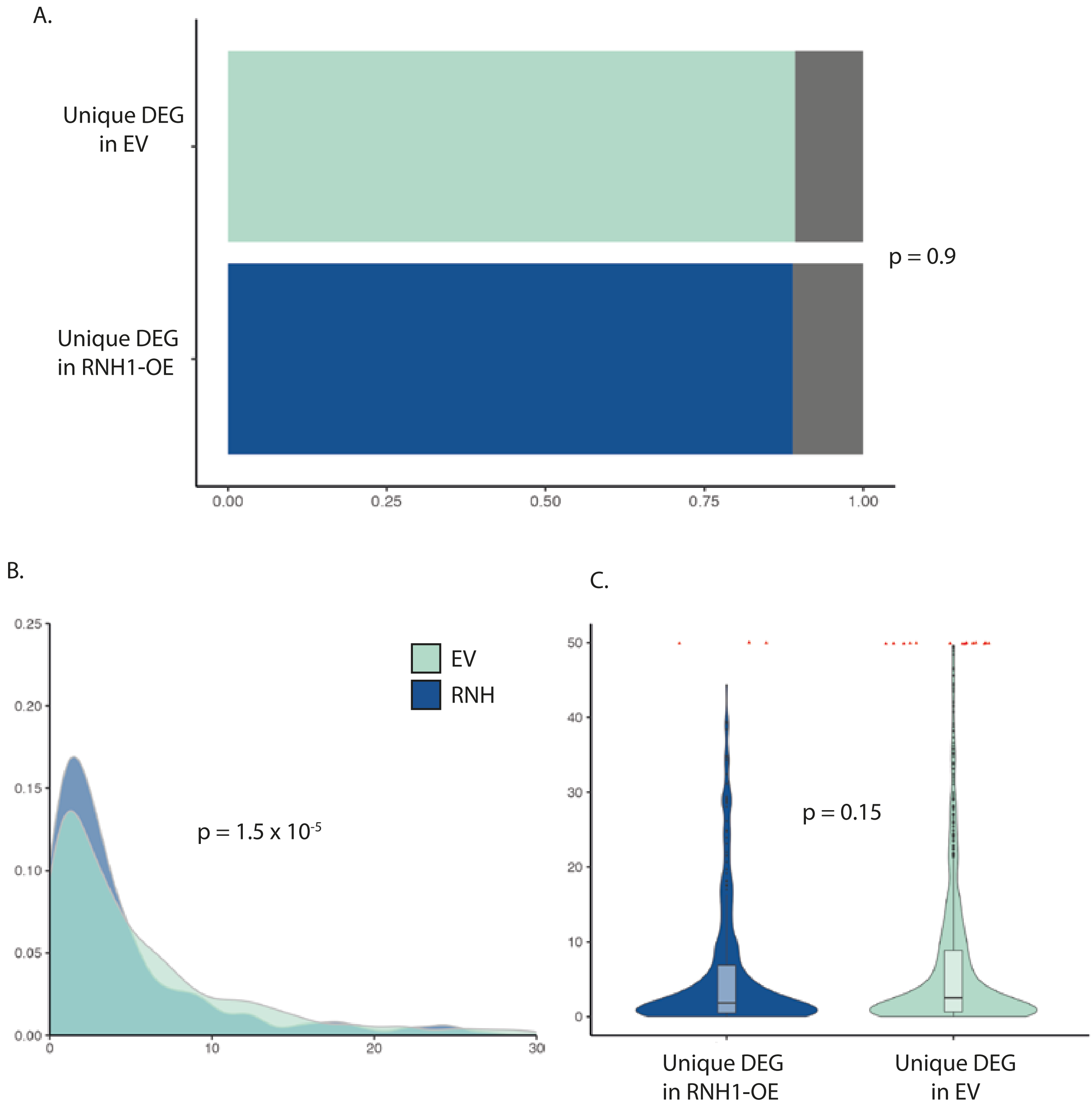


Figure 5.10 RloopTools analysis of genes uniquely differentially expressed by DHT in empty vector conditions vs. genes uniquely expressed in RNH1-OE conditions
A. Proportion of each gene set with at least one RLFS ('RLFS positive') B. Density plot of number of RLFS per gene. P-value is derived from negative binomial regression. C. Violin plot of percentage coverage by RLFS.

5.2.3.3 RNase H1 overexpression is associated with downregulation of genes involved in viral response immune pathways

To further examine the effect of RNase H1 overexpression on gene expression in LNCaP cells I performed two additional differential expression analyses: 1) Empty vector vs. RNH1-OE in the setting of vehicle treatment and 2) empty vector vs. RNH1-OE with DHT treatment. Both analyses showed that RNH1-OE was associated with a markedly higher number of downregulated genes compared to upregulated genes. Approximately 94% and 92% of differentially expressed genes were downregulated by RNH1-OE in vehicle and DHT treated conditions respectively (**figure 5.11 A and B**). Of these genes there was a marked overlap of genes downregulated in both vehicle and DHT treated groups implying that RNase H1 overexpression was the major effect in determining these genes' repression (**figure 5.11C**). Furthermore, classifying this shared set of RNH downregulated genes (n=262) by their response to DHT under EV conditions showed that most genes (63%) were downregulated by DHT and only three genes were originally upregulated by DHT treatment. This adds further evidence that RNase H1 overexpression did not affect the transcription of DHT-upregulated genes. In addition, RNH1-OE reduced the expression of 94 genes not significantly differentially expressed by DHT indicating additional effects of RNase H1 overexpression on housekeeping genes in LNCaP cells (**figure 5.11D**). Representative examples of the expression of individual genes are given in **figure 5.12**. These show the reduction of gene expression by RNase H1 over expression in genes normally repressed by DHT (ZBTB7C and TRANK1) and all in genes not normally regulated by DHT (HLA-B and IFI6). Characterisation of this shared gene set with gene ontology analysis showed that immune signalling biological processes such as 'response to virus', 'response to interferon' and 'regulation of cytokine production' were enriched (**figure 5.13**). Given that most of this gene set was down regulated by RNH1-OE, I can infer that these immune signalling processes were also downregulated under these conditions. Possible explanations for this are explored in the discussion (section 5.3)

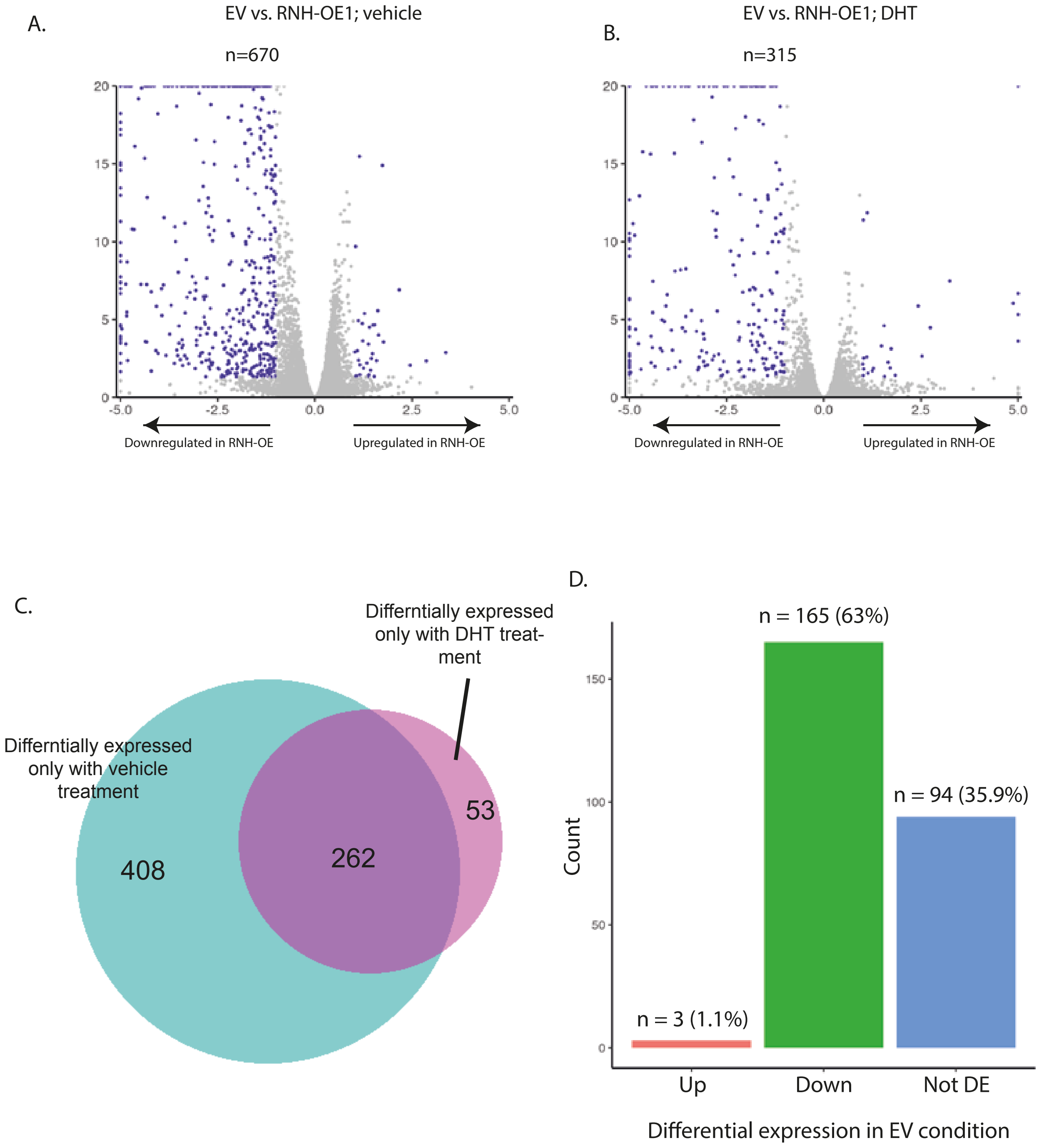


Figure 5.11 Gene downregulation is the major effect of RNAse H1 overexpression in vehicle treated conditions
A. and B. Volcano plots of genes differentially expressed by RNH1-OE with vehicle or DHT treatment respectively. C Venn diagram showing the 262 genes downregulated by RNH1-OE in both treatments. D. Relationship between shared gene set from C and baseline expression in response to DHT.

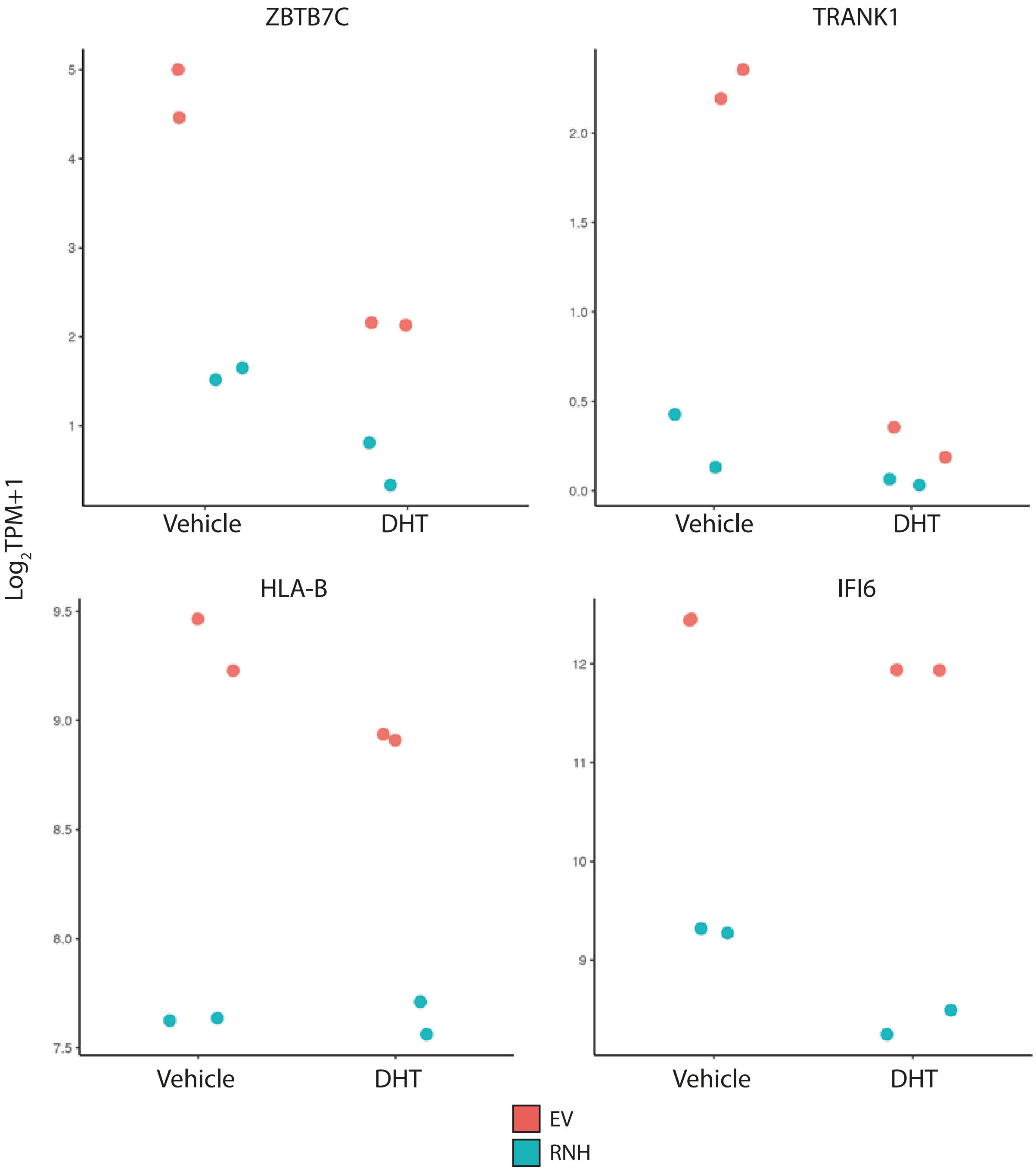


Figure 5.12 Representative examples of gene expression downregulated by RNH1-OE
Upper row is two genes normally repressed by DHT. Lower row is genes whose expression is not changed by DHT but is reduced by RNH1-OE.

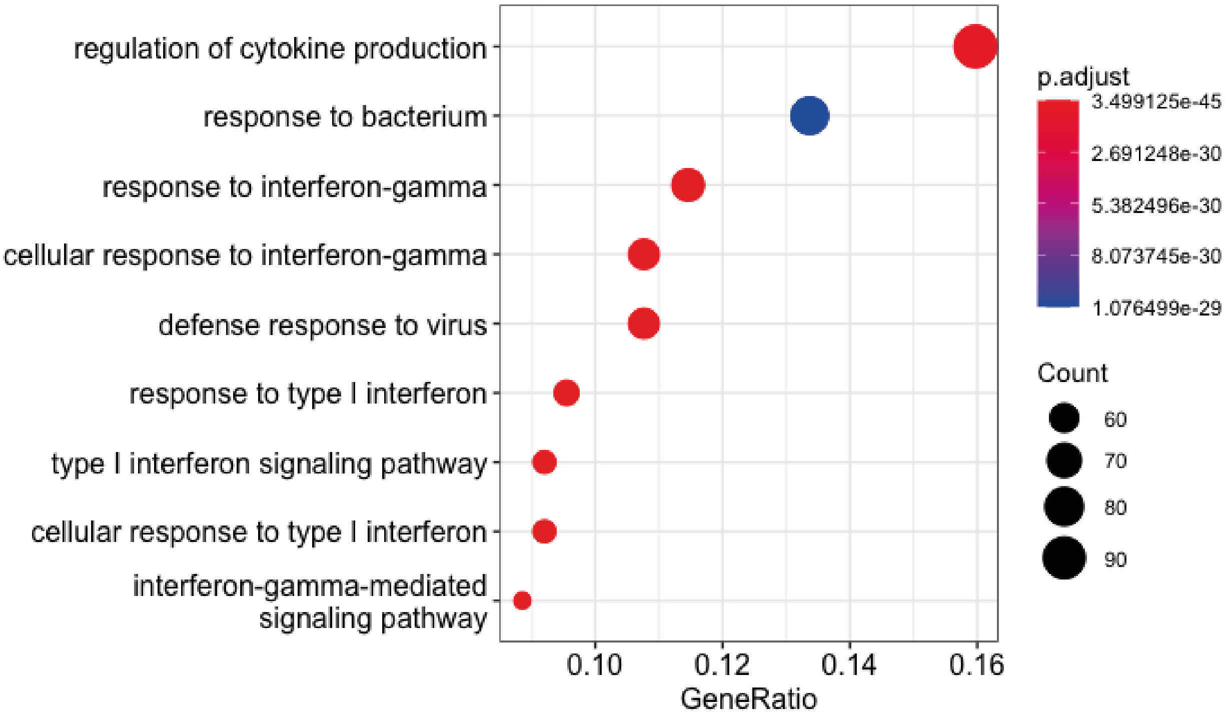


Figure 5.13 Genes downregulated by RNH1-OE are enriched for immune signalling biological processes

Lastly, I used RloopTools to characterise this gene set. I compared the shared gene set of genes differentially expressed by RNH1-OE in vehicle or DHT treated cells with a random equally sized sample (n=262) drawn from the expressed genes in this experiment. Unexpectedly the RNH1-OE differentially expressed gene set had fewer RLFS positive genes and a lower number of RLFS per gene than the randomly selected sample. The percent coverage of RLFS was not significantly different between the two groups (**figure 5.14 A-C**).

Taken together these data indicate five findings:

1. RNase H1 over expression does not significantly affect the induction of androgen induced genes.
2. RNase H1 overexpression is associated with an overall reduction in gene expression.
3. The combination of DHT and RNase H1 overexpression may downregulate gens involved in mitochondrial processes.
4. Genes dysregulated by RNase H1 overexpression regardless of androgen signalling are involved in immunity related processes.
5. The propensity to form R-loops (as measured by RloopTools) does not correlate with RNase H1 induced gene expression dysregulation.

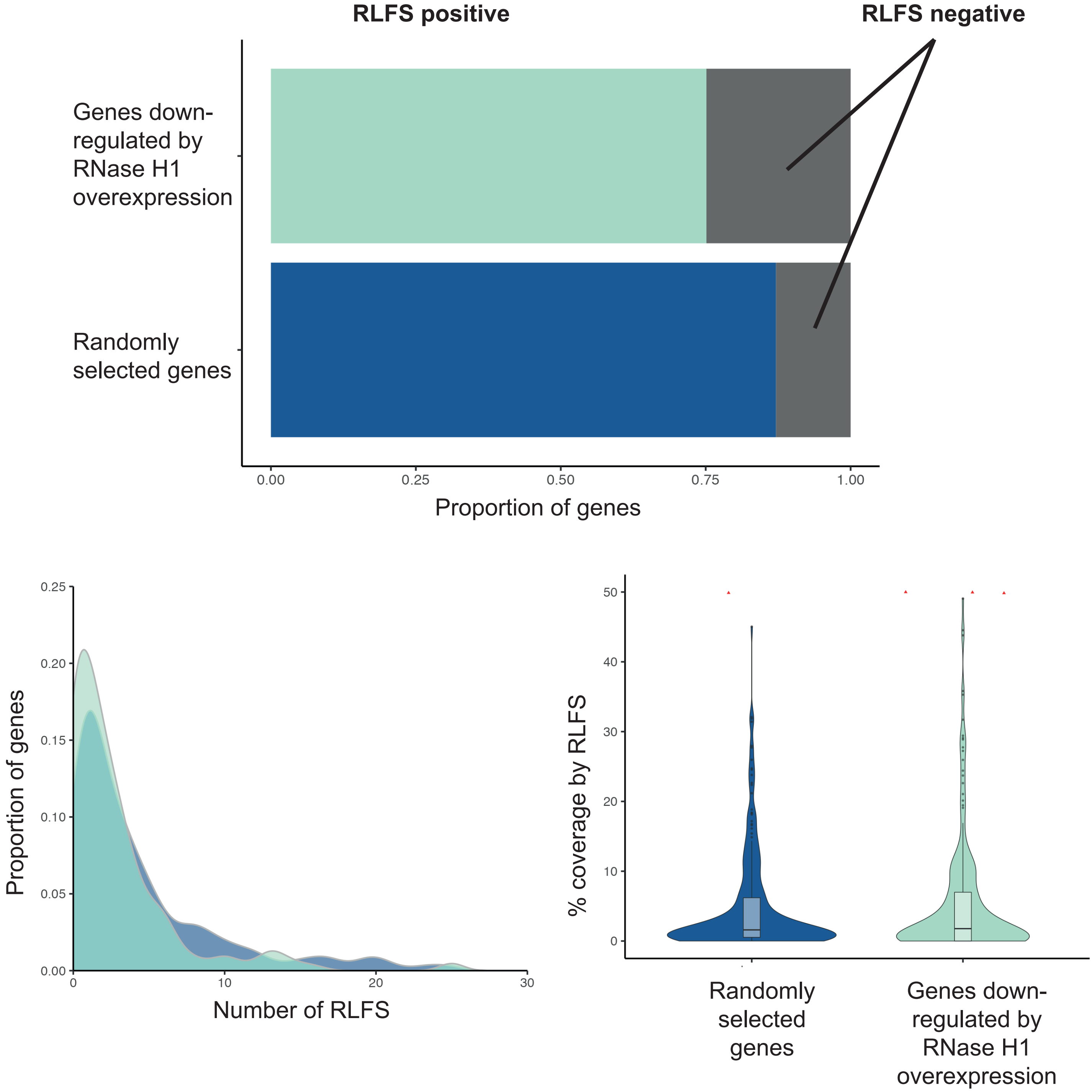


Figure 5.14 RloopTools analysis of genes downregulated by RNH1-OE compared to an equally sized random sample

A. Proportion of each gene set with at least one RLFS ('RLFS positive') B. Density plot of number of RLFS per gene. P-value is derived from negative binomial regression. C. Violin plot of percentage coverage by RLFS.

5.2.4 RNase H1 overexpression does not resolve R-loops at androgen induced genes

I next asked if overexpressed RNase H1 acted at loci known to form R-loops in LNCaP cells. I performed DRIP under the same conditions as the RNA-seq experiment described above and qPCR to interrogate any changes in R-loop occupancy at specific loci. For androgen responsive loci I used the same primers described in chapter four for promoter R-loops of NKX3.1 and KLK3. I also assessed R-loops at four R-loop prone genes that are not androgen responsive: RPL13A, TFPT, Actin and the ribosomal locus 28S. I first confirmed that RNase H1 was still over expressed in the 15cm plate format required for DRIP and that scaling up of the transfection had not affected the experimental set up (**figure 5.15A)**. DRIP-qPCR showed no change in R-loop occupancy at androgen responsive genes or at the RPL13A, TFPT and actin loci (**figure 5.15 B and C**). At 28S however there was a significant increase in DRIP signal in cells overexpressing RNase H1 and treated with vehicle compared to the corresponding empty vector cells. Interestingly, this increase was not seen in RNase H1 over expressing cells treated with DHT. I could not correlate this with 28S rRNA expression from the RNA-seq experiment as ribosomal transcripts had been excluded in the library preparation stage of this experiment by standard polyA capture methods (S. Zhao *et al.*, 2018).

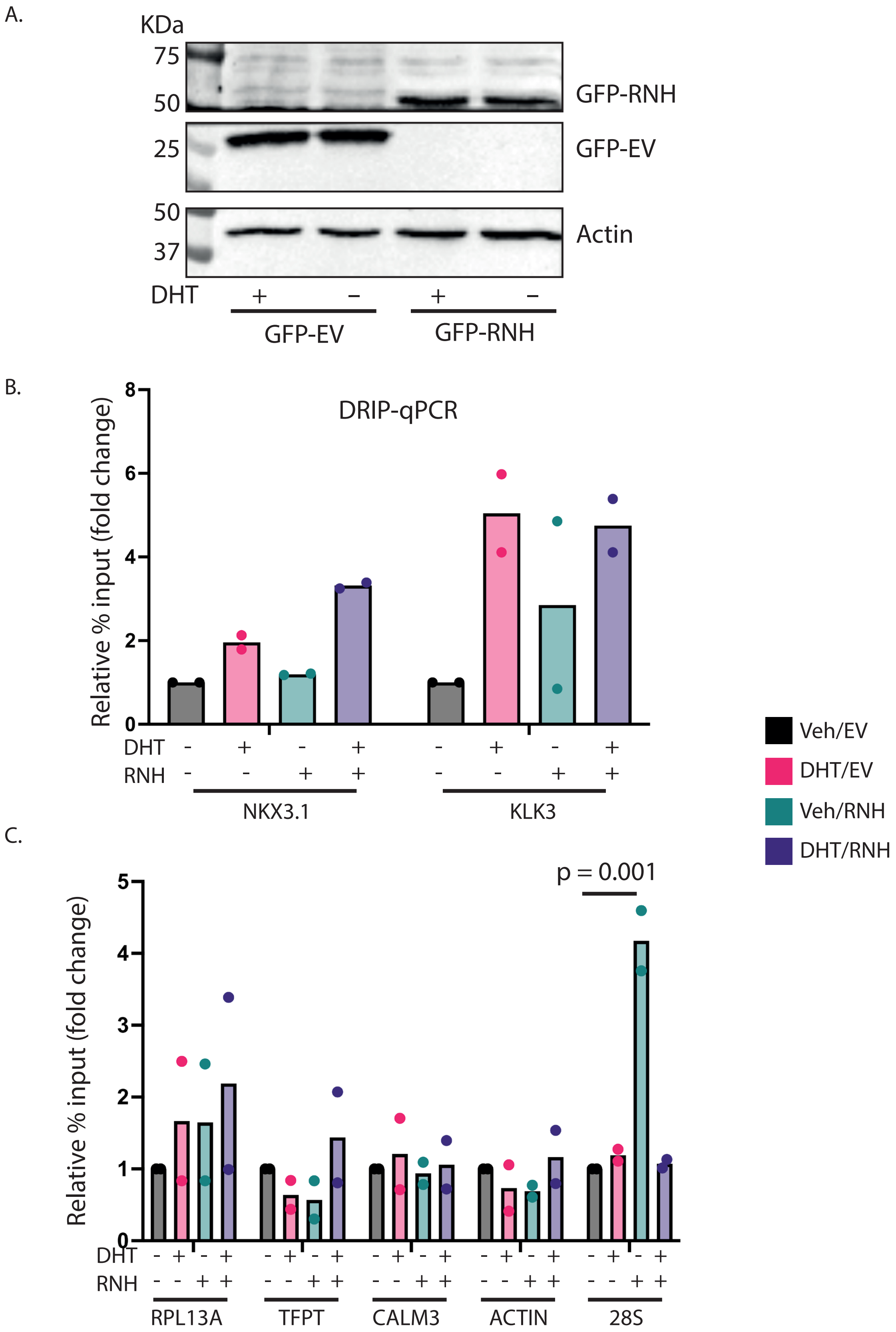


Figure 5.15 RNase H1 overexpression does not resolve R-loops at androgen induced genes
A. Western blot of whole cell lysates probed with anti-GFP to detect EV-GFP (~25 KDa) or RNH1-GFP (~50 KDa). Actin was used as a loading control. B. DRIP-qPCR of NKX3.1 and KLK3 loci. C. DRIP-qPCR of RPL13A, TFPT, CALM3, ACTIN and 28S. All DRIP-qPCR values are fold change relative to empty vector vehicle treated cells for each primer pair. DHT+: 100 nM DHT for 24 hrs, DHT- or Veh: Vehicle treatment.

5.3 Discussion

In this chapter I have shown:

- RNase H1 over expression interferes with the normal androgen-induced expression of mitochondrial genes
- RNase H1 over expression does not alter the expression or R-loop occupancy of canonical androgen induced genes such as NKX3.1 and KLK3
- RNase H1 over expression reduces overall transcriptional activity and may have effects not directly related to resolution of R-loops

5.3.1 Mitochondrial gene down regulation by RNase H1 over expression

To my knowledge, this is one of the first transcriptome-wide analyses of gene expression changes in the context of RNase H1 over expression. The major effect I observed was a reduction in genes associated with mitochondrial function but without a change in the expression of androgen induced genes such as NKX3.1 or KLK3. RNase H1 exhibits DNA:RNA hybrid resolving activity in the mitochondria and nucleus. Normal mitochondrial DNA replication and mitochondrial gene expression both require R-loop formation and resolution (Holt, 2019; Reyes *et al.*, 2020). In mitochondrial DNA replication the DNA:RNA hybrid acts as a primer. In Progressive External Ophthalmolplegia with Mitochondrial DNA Deletions, Autosomal Recessive 2 (PEOB2) syndrome patients present with adult onset progressive spino-cerebellar ataxia, muscle weakness and exercise intolerance (OMIM, 2022). PEOB2 is characterised by RNAse H1 mutations and the accumulation of mitochondrial DNA aggregates indicative of defective replication. Unexpectedly, an RNase H1 inactivating mutation was associated with a *reduction* in mitochondrial R-loops in PEOB2 patient derived fibroblasts (Akman *et al.*, 2016). The reasons for this apparently counter-intuitive finding are not clear however. RNAse H1 mutations have also recently been implicated in reducing transcript levels of mitochondrial ribosomal RNA and mitochondrially transcribed elements of the electron transport chain (Reyes *et al.*, 2020). Within the nucleus RNase H1 is primarily an R-loop resolvase; its putative role in Okazaki fragment removal has recently been challenged by data suggesting that other nucleases such as Dna2, Exo1 and Rad27are more important in RNA primer removal and that RNase H1 is active throughout the cell cycle to resolve cell stress induced R-loops (Lockhart *et al.*, 2019).

As described in the introduction, R-loops form in genic regions, ribosomal DNA arrays, centromeres and telomeres as well as mitochondria. Given the widespread distribution of R-loops in the nucleus and mitochondria, it follows that R-loop resolution by RNAse H1 over expression could have wide ranging effects on the cell. In my data, there appears to be a bias towards RNase H1 overexpression preferentially interfering with genes important for mitochondrial function. As RNase H1 localises to the mitochondria and nucleus (Shen *et al.*, 2017) this raises an important question as to why this discrepancy towards mitochondrial genes was observed. In validating my experimental approach I showed that GFP signal was seen in nuclei and that S9.6 signal was reduced within nuclei segmented by DAPI staining (**figure 5.1).** This implies that RNase H1 was present within the nucleus. In addition, many of the dysregulated genes (e.g. mitochondrial ribosomal RNA genes) are transcribed from the nucleus but are active in the mitochondria. If their resolution was due to the direct effect of RNase H1 then this supports nuclear RNase H1 activity. Two factors argue against this however. First, the genes preferentially dysregulated by RNase H1 overexpression had lower levels of R-loop forming sequences, making them less likely to form R-loops thereby suggesting that their downregulation was not a direct effect of RNase H1 on R-loops in regulatory regions such as promoters but potentially owing to a knock-on effect of dysregulated transcription of other genes. Second, DRIP-qPCR showed no resolution of R-loops at specific loci after RNase H1 over expression. This could be due to the continuous transcription of housekeeping genes such as actin or androgen regulated gene transcription such as NKX3.1 overcoming the R-loop resolving effect of RNase H1. Formation of DNA:RNA hybrids will rely on a balance between transcription rate, RNA processing and R-loop processing/ resolution. A second possibly is that RNase H1 is excluded from chromatin at these loci to avoid deleterious R-loop resolution. The signals and mechanisms that determine which R-loops RNase H1 is dispatched to remain to be elucidated. One study suggested that topoisomerase 1 and RNase H1 are partially functionally redundant (Shen *et al.*, 2017) at least in resolving nucleolar R-loops formed after CPT treatment. Topoisomerase 1 has previously been implicated in androgen activated transcription being a binding partner of NKX3.1 (Bowen *et al.*, 2007) and having activity at the AR-bound enhancers where DNA nicking is required for enhancer RNA transcription and androgen regulated gene transcription (Puc *et al.*, 2015). I therefore speculate that RNase H1 activity at androgen responsive genes might be downregulated or inhibited owing to the activity of topoisomerase 1. If no such inhibition is present at mitochondrial genes then the resulting aggregate gene expression profile would manifest as a majority mitochondrial response as seen in my data.

5.3.2 Downregulation of viral response biological processes by RNase H1 over expression

A further unexpected finding was the apparent downregulation of genes with viral response biological processes by RNase H1 over expression. This implies that in the absence of RNase H1 over expression there is some baseline viral response activity within the cell. DNA:RNA hybrids have been described in the cytoplasm of cells infected with bacteria or viruses (Rigby *et al.*, 2014; Koo *et al.*, 2015) and these out of place nucleic acids can trigger an inflammatory response. Recently two papers have described cytoplasmic DNA:RNA hybrids arising from endogenous processes. Chatzidoukaki et al (2021) demonstrated an accumulation of cytoplasmic R-loops and single stranded DNA fragments in ERCC1 deficient mouse pancreas and correlated this with an anti-viral immune response and subsequent chronic pancreatitis. A separate study implicated RNA polymerase 3 (responsible for transcription of tRNA, 5S rRNA, spliceosome RNA and some microRNA) in the production of cytoplasmic R-loops which again gave rise to an anti-viral like response (Koo *et al.*, 2015). Importantly the authors localised R-loops to the cytoplasm by co-staining with mitochondrial markers to ensure they were not spuriously recognising mitochondrial R-loops as cytoplasmic. In relation to my data, RNase H1 over expression could have reduced nuclear and mitochondrial R-loop production thereby reducing the basal levels of DNA:RNA hybrids released into the cytoplasm. This in turn may have reduced the requirement for a viral-like response as observed in my gene ontology characterisation.

5.3.3 Unintended consequences of RNase H1 overexpression

The third major finding in this chapter is that transcription was generally downregulated by RNAse H1 overexpression as evidenced by the ~95% of genes with reduced expression upon RNAse H1 overexpression compared to cells transfected with empty vector alone. Two thirds of these genes had their activity reduced by DHT under baseline conditions where they were transfected with the empty vector. These genes therefore had a relatively high expression level without androgen stimulation but the response to DHT was abolished by

RNase H1 overexpression and their level of transcription was reduced in both vehicle and DHT treated cells. A second group of genes were not regulated by DHT but their steady state mRNA levels were reduced by RNase H1. There are many examples in the R-loop literature of RNase H1 overexpression being used to demonstrate resolution of a phenotype and therefore attribute R-loops as the causal factor of this phenotype. Multiple studies have investigated the relationship between RNA helicases and DNA damage. Hodroj et al (2017) showed that DDX19 knock down was associated with R-loop accumulation and DNA damage. RNase H1 overexpression rescued DDX19 mediated impaired DNA synthesis as measured by CldU incorporation. However, the authors did not ask if RNase H1 overexpression rescued the DNA damage observed with DDX19 knockdown. A separate study suggested that DHX9 protected cells from CPT-induced R-loop mediated damage (Cristini *et al.*, 2018). Whilst RNase H1 overexpression reduced the recruitment of DHX9 to the R-loop prone gene actin, the authors did not demonstrate resolution of R-loops or γH2AX foci in this gene. Immunofluorescence for S9.6 and γH2AX was reduced by RNAse H1 overexpression but this represented the general level of these markers. In a third example researchers used RNase H1 overexpression again rescued a DNA synthesis defect, this time induced by loss of Fanconi anameia pathway genes in primordial germ cells (Yang *et al.*, 2022). Interestingly, GFP-RNase H1 overexpression did not reduce basal R-loop levels and immunofluorescence images showed GFP signal in a network around the nucleus as well as in the nuclear region itself suggestive of RNase H1 recruitment to mitochondria. None of these studies measured the effect of RNase H1 overexpression on transcription activity within the cells. If the same reduction in transcription was observed as was present in my data then this could account for a resolution of phenotype through reduction of available RNA transcripts globally or in a large number of genes which would have the effect of reducing R-loop formation. This is similar to the use of 5,6-dichloro-1-beta-ribofuranosylbenzimidazole (DRB) to inhibit transcription and show that R-loops are transcription dependent (Jurga *et al.*, 2021). The use of RNAse H1 overexpression is subtly different however. Rather than reducing transcription, RNase H1 overexpression is intended to remove R-loops to demonstrate it was R-loop formation specifically that was responsible for the phenotype being examined. If the effect is actually transcriptional inhibition then this represents a non-specific action of RNase H1 where transcriptional reduction would reduce R-loops in general rather than necessarily at biologically relevant loci. This transcriptional reduction could arise from cell stress owing to ribosomal disruption (Abraham *et al.*, 2020), mitochondrial dysfunction or dysregulation of transcription requiring R-loop formation and resolution (Pérez-Calero *et al.*, 2020).

In summary I have identified possible off target effects of RNAse H1, a commonly used control in R-loop biology. Other studies have identified cell stress effects of RNase H1 over expression. In two studies there was an *increase* in γH2AX foci with RNase H1 over expression without any other exogenous DNA damaging agents (Shen *et al.*, 2017; Landsverk *et al.*, 2019) and separately in mouse haematopoietic stem cells 'non-specific cell toxicity' was observed which precluded the use of transfected RNase H1 (Shi *et al.*, 2017). Furthermore, RNase H1 over expression in cells transfected with a non-targeting siRNA showed a decrease in DNA synthesis (Prendergast *et al.*, 2020) which could represent an additional off target effect. Together, the literature and my data suggest that RNase H1 overexpression may not be a reliable indicator of R-loop specific phenotypes owing off downstream effects on transcription and mitochondrial function. Its use requires careful optimisation and extensive orthogonal controls to ensure any R-loop resolution is specific and not simply a change in transcriptional output resulting from non-specific cell stress.

Chapter 6. The DEAD-box Helicases DDX5 and DDX17 have varied effects on androgen regulated transcription and R-loop accumulation

6.1 Introduction

As described in chapter one, the DEAD-box helicases have recently been implicated in R-loop homeostasis. DDX5 and DDX17 are two such DEAD box helicases that also have roles in androgen signalling. These closely related helicases are paralogues and have 90% sequence homology of the central protein core. The N- and C- terminal domains are less conserved. In LNCaP cells DDX5 was shown to interact with the androgen receptor and functional studies suggested that DDX5 functioned as an androgen receptor co-activator at least for the androgen responsive model locus KLK3 (Clark *et al.*, 2008). A separate study used an siRNA that targeted DDX5 and DDX17 in combination and investigated the effects of DDX5/17 knockdown on oestrogen and androgen regulated signalling (Samaan *et al.*, 2014). This work implicated these helicases in alternative splicing of GSK3β. DDX5/17 knockdown led to the expression of an alternative GSK3β isoform which reduced androgen receptor protein levels and reduced androgen signalling. Only one study has specifically identified DDX17 as an androgen receptor co-factor. This study used androgen response element DNA oligonucleotides to pull down the androgen receptor and associated proteins. Whilst this detected an interaction between DDX17 and the androgen receptor, no further mechanistic investigation was done (Wong *et al.*, 2009).

Given the overlapping roles of DDX5 and DDX17 in androgen receptor signalling and in R-loop homeostasis I decided to test if DDX5 and DDX17 exert their roles in androgen regulated gene expression through R-loops.

6.2 Aims and objectives

The aims of this chapter were to:

- Determine if DDX5 and DDX17 interact with R-loops in prostate cancer cells and if this interaction is dependent on androgens
- Examine the effect of loss of DDX5 and DDX17 on the transcriptional response to androgens
- Examine the effect of DDX5 and DDX17 loss on androgen mediated DNA damage
- Probe the interaction between dead box helicases and chromatin and whether R-loops play a role in modulating this

6.3 Results

6.3.1 DDX17 and DDX5 interact with R-loops in LNCaP cells

The GEPIA web portal (Tang *et al.*, 2017) was used to analyse data from The Cancer Genome Atlas prostate cancer cohort (Veluvolu *et al.*, 2015) and the GTeX project (GTeX consortium, 2020). The GTeX project has comprehensively measured gene expression in normal tissues donated after death and provides a useful reference level for comparisons with cohorts of cancer cases. DDX17 had significantly lower expression in the prostate cancer tissues from TCGA. DDX5 was also expressed at a lower level in cancer cases but this difference was not significant (**figure 6.1B**)

Two studies have used high-throughput methods to identify proteins that interact with R-loops and therefore may be involved in R-loop regulation. Using S9.6 immunoprecipitation coupled to mass spectrometry, the Gromak lab identified 846 R-loop interacting proteins including splicing factors, DNA binding proteins, chromatin remodelling factors and RNA helicases (Cristini *et al.*, 2018). The Cheung lab used an orthogonal approach of pulling down proteins that associated with biotinylated DNA:RNA hybrids composed of sequences from two well-described R-loops (Wang *et al.*, 2018). A 25% overlap was observed between the two datasets with the differences attributed to the different cell lines and approaches to protein enrichment. In particular, the use of two *in vitro* transcribed R-loop sequences by the Cheung lab may have biased the capture of R-loop binding proteins to those with specific sequence preferences.

The publicly available mass spectrometry data from Cristini *et al.* (2018) was re-analysed (**figure 6.1A**). This revealed that DDX5 and DDX17 were significantly enriched in the S9.6 immunoprecipitation fraction with fold change over isotype matched control pulldown of ~10.5 and ~9.5 respectively. S9.6 immunoprecipitation coupled to SDS-PAGE and immunoblotting was used to ask if DDX5 and DDX17 are R-loop associated proteins in LNCaP cells. Successful cell lysis, sonication and fractionation was first confirmed using antibodies against tubulin and histone H3 to detect cytoplasmic and nuclear fractions respectively (**figure 6.1C and D**). RNA polymerase 2 was successfully immunoprecipitated and a reduction in signal was seen in the IgG and RNase H controls indicating that S9.6 was binding to nascently transcribed RNA that was still in association with RNA pol 2 rather than mature mRNA that had been release from chromatin (**figure 6.1E**). Bands corresponding to DDX5 and DDX17 were seen in the nuclear lysates immunoprecipitated with the S9.6 antibody. Importantly no signal was seen using a non-specific IgG antibody and a marked reduction in signal was observed after treatment with RNase H (**figure 6.2**). Qualitatively, there was no change in DDX5 or DDX17 signal with two or 24 hour androgen treatment. Together these results indicate that S9.6 immunoprecipitation successfully isolated *bona fide* R-loops and that DDX5 and DDX17 are indeed R-loop interacting proteins.

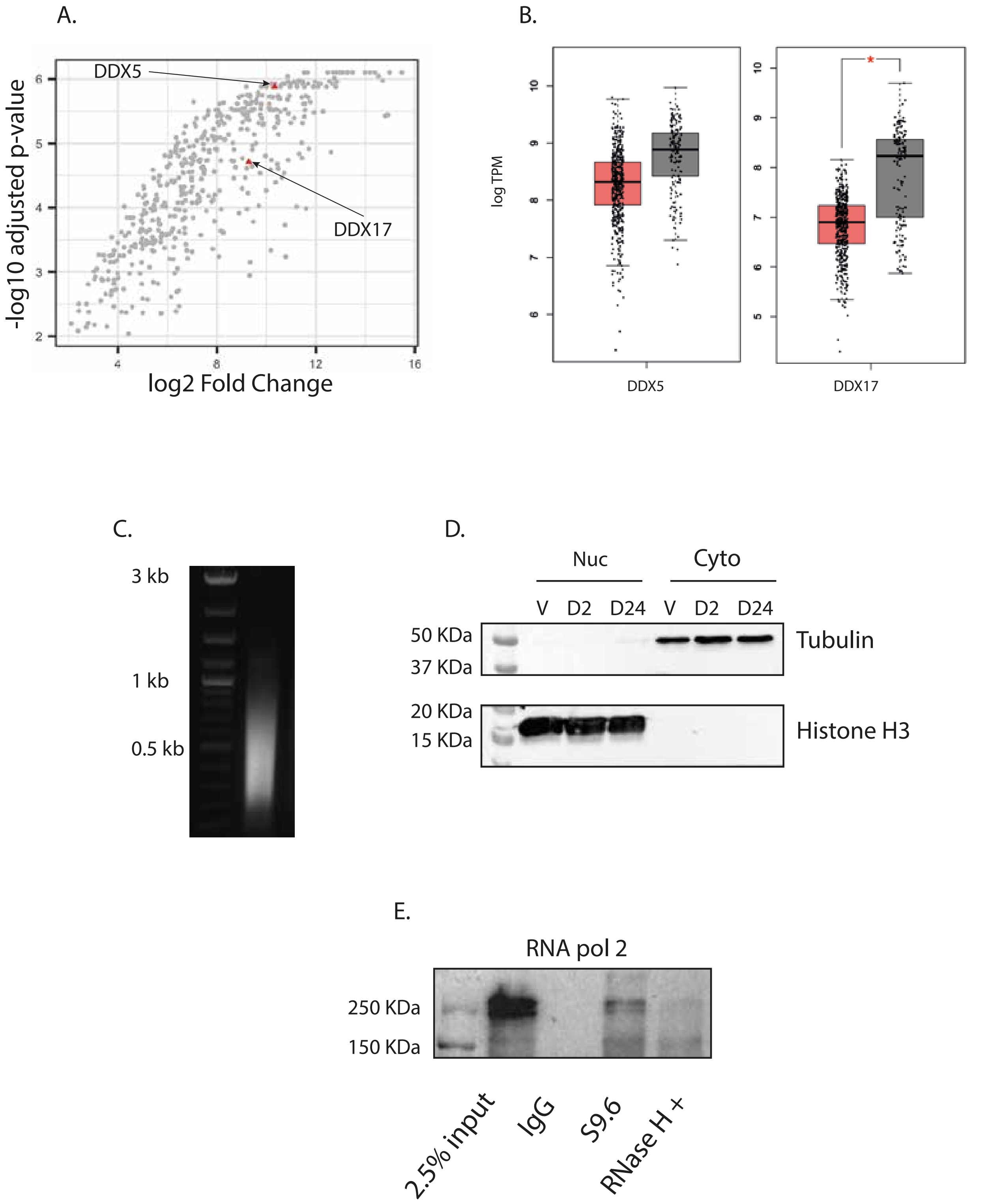


Figure 6.1 Quality control of S9.6 immunoprecipitation
A. S9.6 IP-MS data from Cristini et al (2018) with DDX17 and DDX5 highlighted (red triangles).
B. mRNA expression levels of DDX17 and DDX5 in normal (grey boxplot) and TCGA prostate cancer cases (red boxplot). C. Quality control of S9.6 IP. Agarose gel electrophoresis demonstrating DNA fragment size after 4 cycles of sonication. D. western blot of cytoplasmic and nuclear fractions probed with tubulin and histone H3. E. proteins immunoprecipitated by S9.6 probed with RNA polymerase 2 antibody. Cyto and Nuc denote cytoplasmic and nuclear fractions.

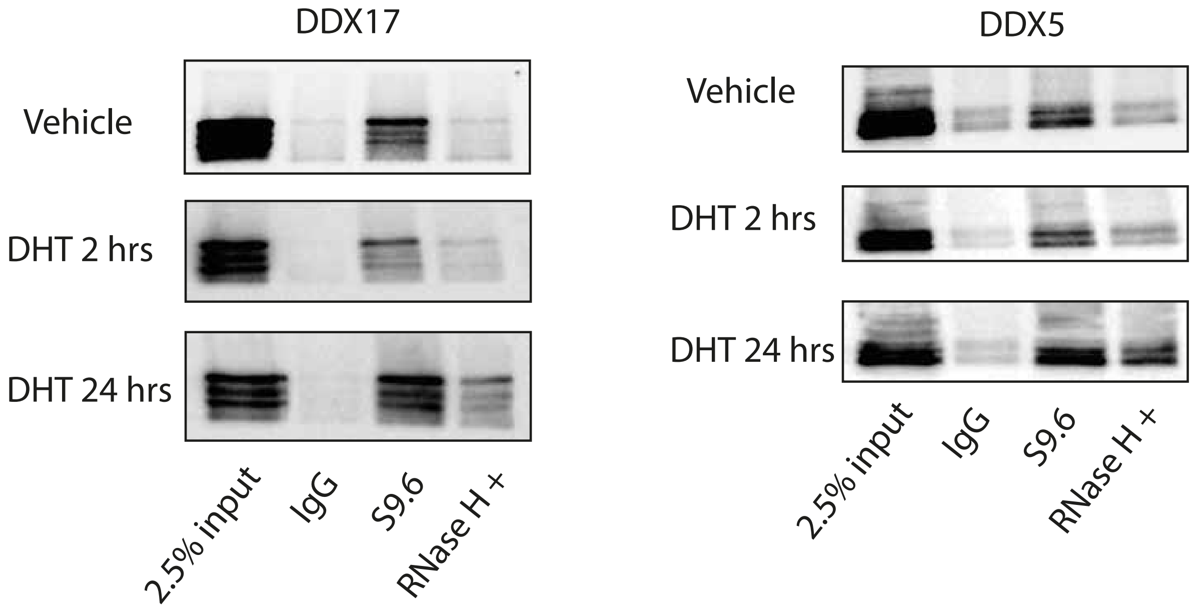


Figure 6.2 DDX17 and DDX5 interact with R-loops in LNCaP cells
Western blot of proteins immunoprecipitated by S9.6 antibody probed with anti-DDX17 (left) and anti-DDX5 (right) antibodies.

6.3.2 The androgen regulated genes KLK3 and NKX3.1 exhibit distinct changes in the androgen response in the context of DDX5 or DDX17 knockdown

To assess the role of DDX5 and DDX17 in androgen signalling, the expression of each helicase was reduced using siRNA. Successful knock down at the mRNA level was confirmed by RT-qPCR, which also showed that androgen stimulation does not affect mRNA levels of DDX5 or DDX17 (**figure 6.3A**). SDS-PAGE and Western blot demonstrated a reduction in the protein level of DDX5 and DDX17 (**figure 6.3B**). Interestingly there was upregulation of DDX17 in response to DDX5 knock down and *vice versa* but this was seen solely at the mRNA level and did not translate to a protein-level alteration.

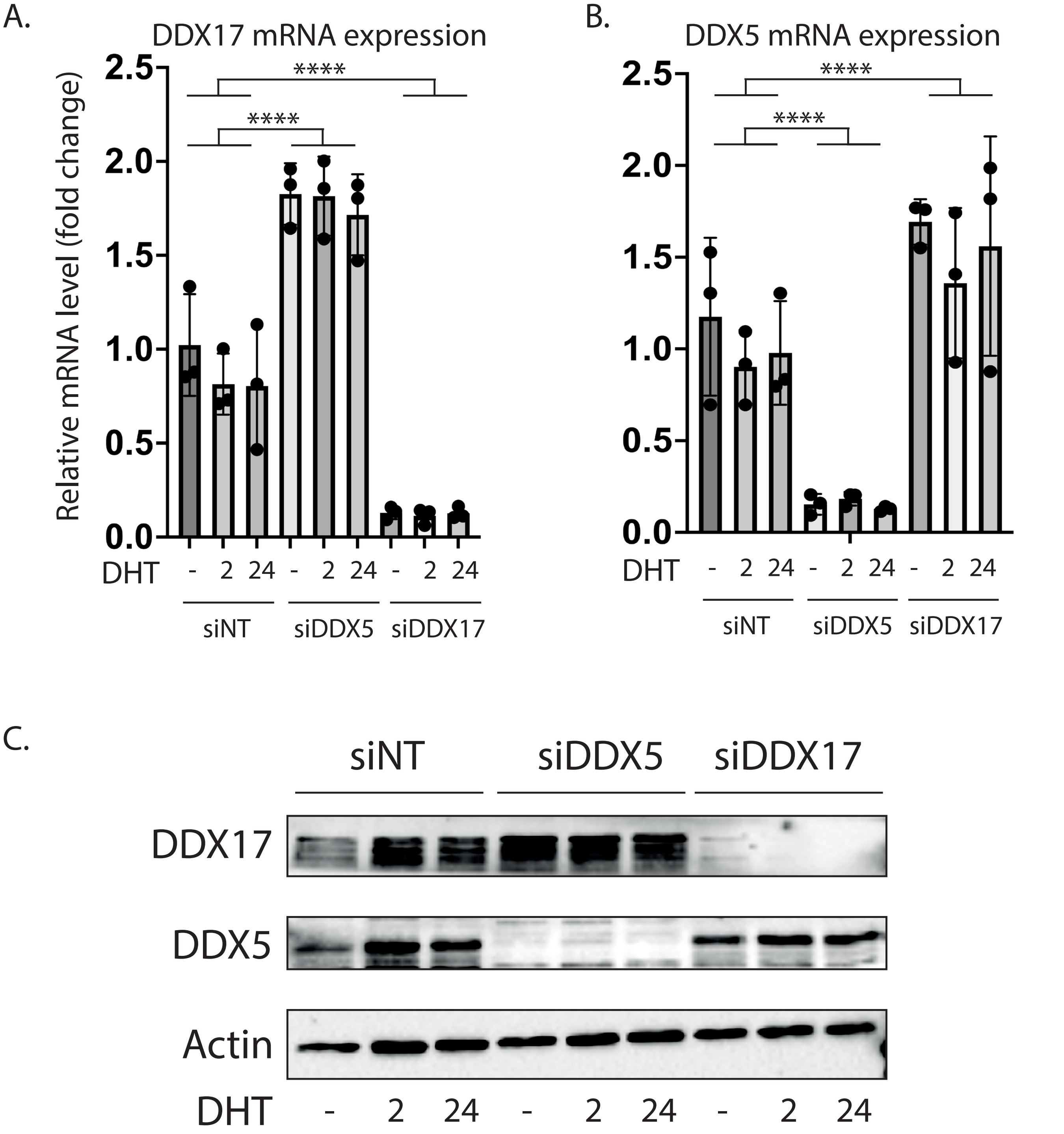


Figure 6.3 Successful knock down of DDX17 and DDX5.
A. and B. RT-qPCR for gene expression of DDX17 and DDX5 respectively. No difference in expression was seen between treatment conditions for each siRNA. Therefore comparisons were made between pooled observations for each siRNA using the t test. C. Western blot using antibodies against DDX5 and DDX17. Actin was used as a loading control. Lysates were run on a 4-20% gradient polyacrylamide gel. DHT: 100 nM DHT for 2 or 24 hours; - indicates treatment with vehicle only. ****: $p < 0.001$.

This experiment had a two-factor design where the observation of interest was the interaction between androgen signalling and loss of DDX5 or DDX17. Specifically, the question was: Does loss of DDX5 or DDX17 modify the effect of DHT on expression of NKX3.1 and/or KLK3? To properly answer this, gene expression values from RT-qPCR were determined using the standard curve method. The values were normalised against a house keeping gene (GAPDH). The fold change of these normalised values was then determined relative to the gene expression in cells transfected with non-targeting siRNA and treated with vehicle only. This provided one baseline measurement as a comparator and allowed me to ask if knock down of either helicase had a general effect on gene expression or if the effect was a specific interaction with androgen stimulation.

No significant change in NKX3.1 expression was seen in cells treated with vehicle or DHT for 2 hr. However, knock down of DDX17 led to a higher expression of NKX3.1 after 24 hr DHT when compared to non-targeting siRNA at the same timepoint as suggested by a significant interaction effect (p = 0.017) using two-way ANOVA (difference in absolute fold change at 24 hr siNT vs. siDDX17: 1.93 [95% CI: 0.53 – 3.33, adjusted p = 0.0072; Šídák's multiple comparisons test, **Figure 6.4A**). By contrast, DDX5 knock down increased the expression of NKX3.1 at both treatment times and in vehicle-treated cells when compared with baseline conditions but without evidence of interaction between knock down and DHT treatment (**Figure 6.4B**).

For KLK3, no change in the response to DHT was observed when DDX17 was depleted (**Figure 6.4C**). Upon DDX5 knock down there was a decrease in the KLK3 mRNA level after 24 hours DHT treatment (interaction effect p = 0.0004 by two-way ANOVA; difference in absolute fold change at 24 hr siNT vs. siDDX17: 5.74 [95% CI: 3.69 – 7.82], p < 0.0001; Šídák's multiple comparisons test, **Figure 6.4D**).

Together, these data suggest that loss of DDX5 is associated with generally higher expression of NKX3.1 but a reduction in the response to androgen by KLK3. Conversely DDX17 loss increases the androgen response of NKX3.1 but has no effect on the androgen response of KLK3.

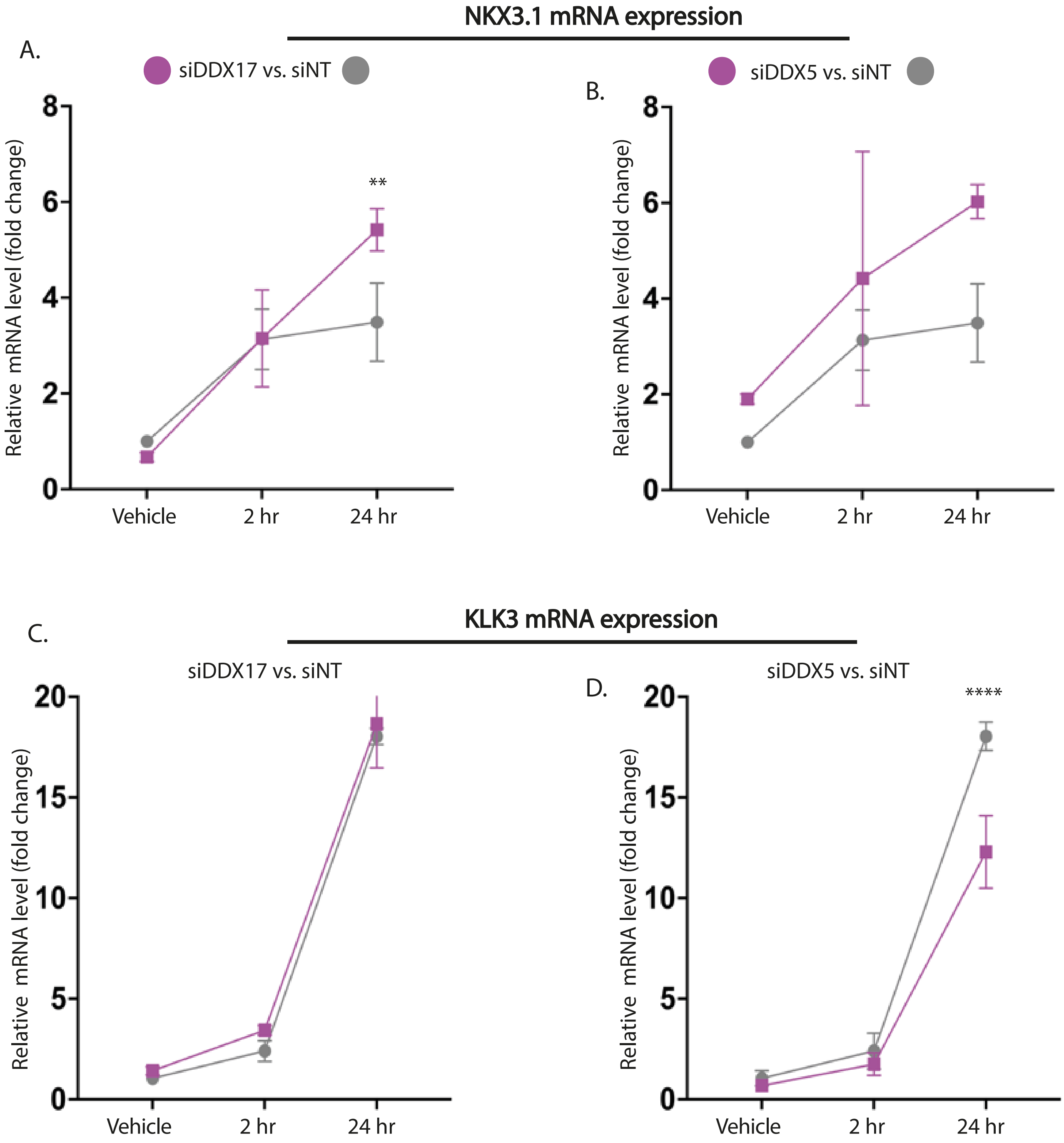


Figure 6.4 Knockdown of DDX5 and DDX17 have gene-specific effects on androgen regulated gene expression.

A. - D. Interaction plots of mRNA fold change in response to 100 nM DHT for 2 or 24 hours for NKX3.1 (upper) and KLK3 (lower). DDX5 or DDX17 were knocked down using siRNA for 48 hours prior to DHT treatment. Fold change values are relative to LNCaP cells transfected with non-targeting siRNA (siNT) and treated with vehicle (methanol) for 24 hours. In all plots siNT is denoted by grey points/lines, and siDDX5 or siDX17 by purple points/lines. Error bars are standard error of the mean. The interaction of knockdown and DHT was tested by two-way ANOVA. Statistical significance was determined by Šídák's multiple comaprisons test. **: $p < 0.01$; ***: $p < 0.001$

6.3.3 Chromatin occupancy of DDX5 and DDX17

I next used chromatin immunoprecipitation (ChIP) coupled to qPCR to ask if DDX5 and/or DDX17 are recruited to the androgen inducible R-loops in KLK3 and NKX3.1. LNCaP cells were treated with vehicle or DHT and proteins then cross linked to chromatin with paraformaldehyde (PFA). I first optimised PFA fixation parameters and sonication settings to yield chromatin fragments 150 – 400 bp in length (**figure 6.5A**). Immunoprecipitation was done using the same DDX5 and DDX17 antibodies described above. DDX17 showed a four-fold enrichment at the NKX3.1 R-loop region after 2 hours of DHT treatment but returned to baseline after 24 hours DHT treatment. No dynamic changes in DDX17 recruitment were observed at the KLK3 R-loop region. Furthermore, no dynamic DDX17 recruitment was demonstrated at the RPL13A and TFPT R-loop regions which are not androgen induced. Importantly the ChIP-qPCR signal was higher for DDX17 ChIP compared to an isotype matched IgG control implying that the anti-DDX17 antibody was detecting genuine chromatin binding events and not background noise (**figure 6.5B)**. By contrast there was no dynamic recruitment of DDX5 to any of the same R-loop forming loci. With this antibody there was more variability in the isotype control IgG signal (**Figure 5.6**). This implies either that there was less DDX5 binding to chromatin in general or that technical experimental issues such as poor antibody specificity or loss of antibody binding during processing made the result unreliable. Given the pre-existing literature showing DDX5 recruitment to chromatin (Yu *et al.*, 2020; Suthapot *et al.*, 2022) it is likely that technical issues prevented meaningful analysis of this experiment. In summary, DDX17 is dynamically recruited to the NKX3.1 R-loop forming region with short term androgen exposure.

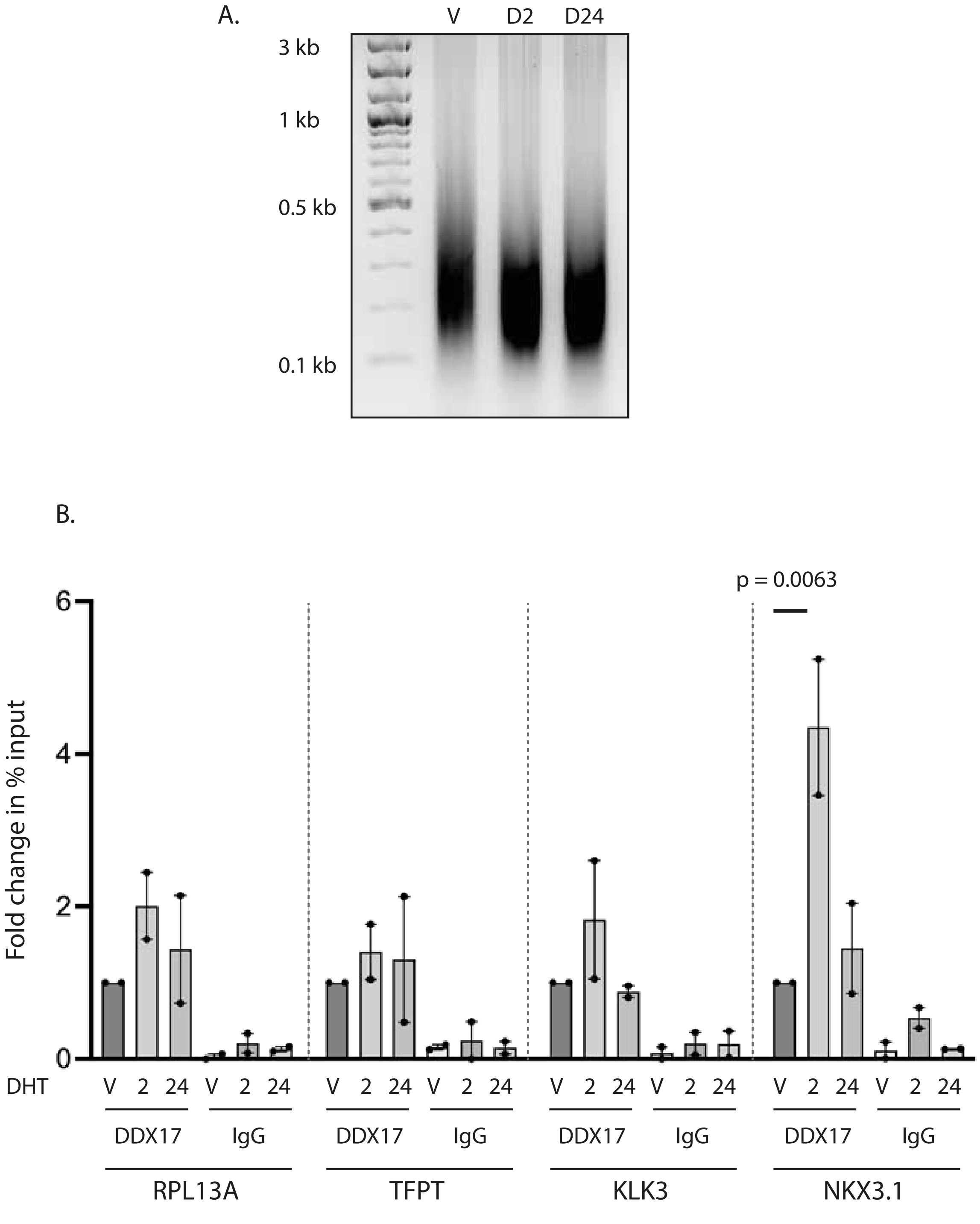


Figure 6.5 Chromatin occupancy of DDX17 at androgen responsive loci
A. Representative image of sonicated cross-linked chromatin run on a 1.5% agarose gel containing 0.5 ug/ml Ethidium Bromide. V: Vehicle (ethanol) for 24 hours; D2: 100 nM DHT for 2 hours; D24: 100 nM DHT for 24 hours. B. ChIP-qPCR of DDX17 chromatin occupancy after 2 and 24 hours DHT treatment at the loci indicated. C. Similar to B but for DDX5.

C.

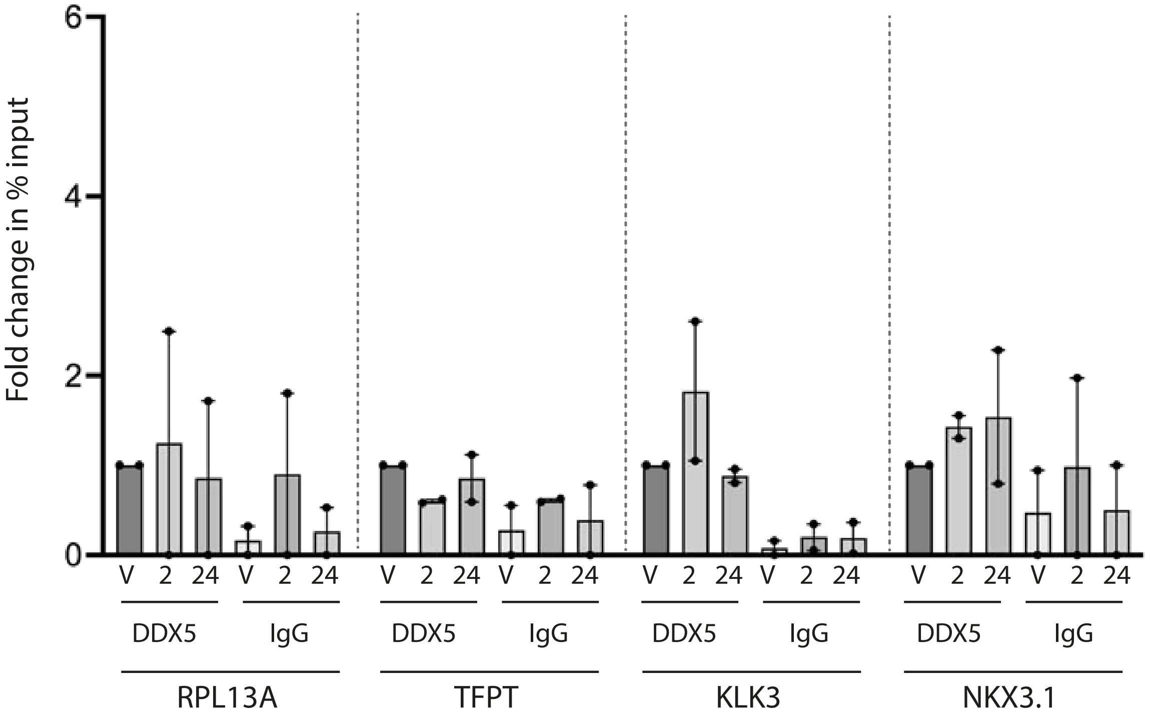


Figure 6.6 Chromatin occupancy of DDX5 at androgen responsive loci
A. Representative image of sonicated cross-linked chromatin run on a 1.5% agarose gel containing 0.5 ug/ml Ethidium Bromide. V: Vehicle (ethanol) for 24 hours; D2: 100 nM DHT for 2 hours; D24: 100 nM DHT for 24 hours. B. ChIP-qPCR of DDX5 chromatin occupancy after 2 and 24 hours DHT treatment at the loci indicated.

6.3.4 DDX5 and DDX17 knockdown have differing effects on global R-loop accumulation

Having established that DDX5 and DDX17 interact with R-loops, that DDX17 was dynamically recruited to NKX3.1 and that these helicases had specific effects on androgen regulated gene expression I next tested whether knockdown of either helicase with siRNA would affect R-loop accumulation at a whole-cell level and whether there was any interaction between knockdown and androgen stimulation. LNCaP cells were depleted of DDX5 or DDX17 by siRNA or transfected with non-targeting siRNA (siNT) then cultured in hormone-free media. Forty-eight hours later cells were treated with DHT for two or 24 hours. Global R-loop levels were assayed by immunofluorescence using the S9.6 antibody. Nuclear S9.6 immunofluorescence signal did not change with androgen treatment in siNT transfected cells, confirming the slot blot results presented in **figure 3.1**. By contrast, the baseline S9.6 signal increased in DDX5 knockdown cells treated with vehicle only. Moreover, there was an interaction between DDX5 knockdown and androgen exposure at both the two and 24 hour timepoints with S9.6 signal increasing at both timepoints compared to vehicle treatment (**figure 6.7)**. DDX17 knockdown was associated with an increase in S9.6 signal at all treatment timepoints but there was no interaction between loss of DDX17 and androgen treatment. Together these data add further evidence to the roles of DDX5 and DDX17 as R-loop interacting proteins. Furthermore, the interaction between DDX5 and androgen signalling suggested that DDX5 might act to resolve androgen induced R-loops. DDX17 may play a more general R-loop resolving role in the cell. Importantly treatment of cells with recombinant RNAse H1 after fixation markedly reduced the S9.6 signal in all conditions, confirming the specificity of the S9.6 antibody.

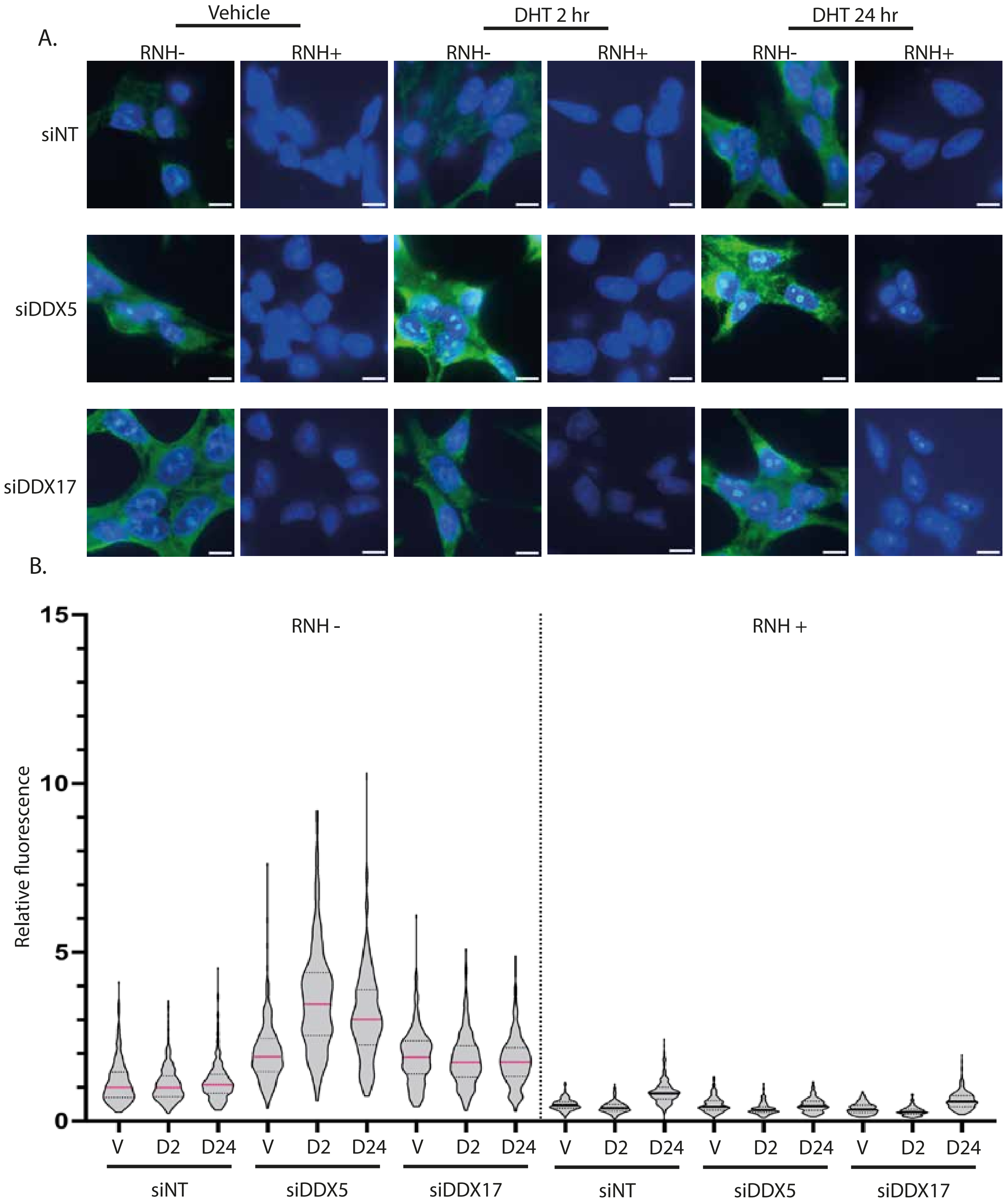


Figure 6.7 Knockdown of DDX5 causes androgen dependent R-loop accumulation
DDX5 and DDX17 were depleted in LNCaP cells using siRNA and cells treated with vehicle (V), 100 nM DHT for 2 (D2) or 24 (D24) hours. R-loops were detected by S9.6 immunofluorescence. Treatment of fixed cells with recombinant RNase H1 for 24 hr (RNH+) was used as a specifity control. A. Representative immunofluorescence microscopy images. B. Quantification of immunofluorescence signal from three biological repeats per condition. All data was normalised by dividing the raw integrated density by the median raw integrated density of the baseline condition: vehicle/RNH-/siNT. Scale bar = 10 µm.

6.3.5 Loss of DDX5 or DDX17 reduce R-loop occupancy at the KLK3 locus whereas only DDX5 affects R-loops at NKX3.1

To probe the interaction of androgen signalling and DDX5 and DDX17 further I performed DRIP-qPCR for the KLK3 and NKX3.1 R-loop loci in the context of DDX5 or DDX17 knockdown and androgen stimulation. In cells transfected with siNT, I could reproduce the DRIP signal observed in chapter three for the NKX3.1 and KLK3 loci: The NKX3.1 promoter R-loop was induced by DHT at 2 hours and remaining at 24 hours albeit at a lower level to the earlier timepoint. The KLK3 R-loop was slightly increased at 2 hours and returned to baseline by 24 hours.

At the NKX3.1 R-loop locus DDX5 knockdown was associated with a reduction in R-loop occupancy at all treatment timepoints and the characteristic induction of this locus was lost. By contrast siRNA targeting DDX17 had no effect on the R-loop profile of NKX3.1. At the KLK3 locus knockdown of either helicase reduced the R-loop signal at all timepoints when compared to cells transfected with non-targeting siRNA and treated with vehicle. There was a two- to five-fold reduction in DRIP signal with siDDX17 and a five- to ten-fold reduction with siDDX5. These data imply that DDX5 and DDX17 have locus specific effects on R-loop homeostasis (**figure 6.8**).

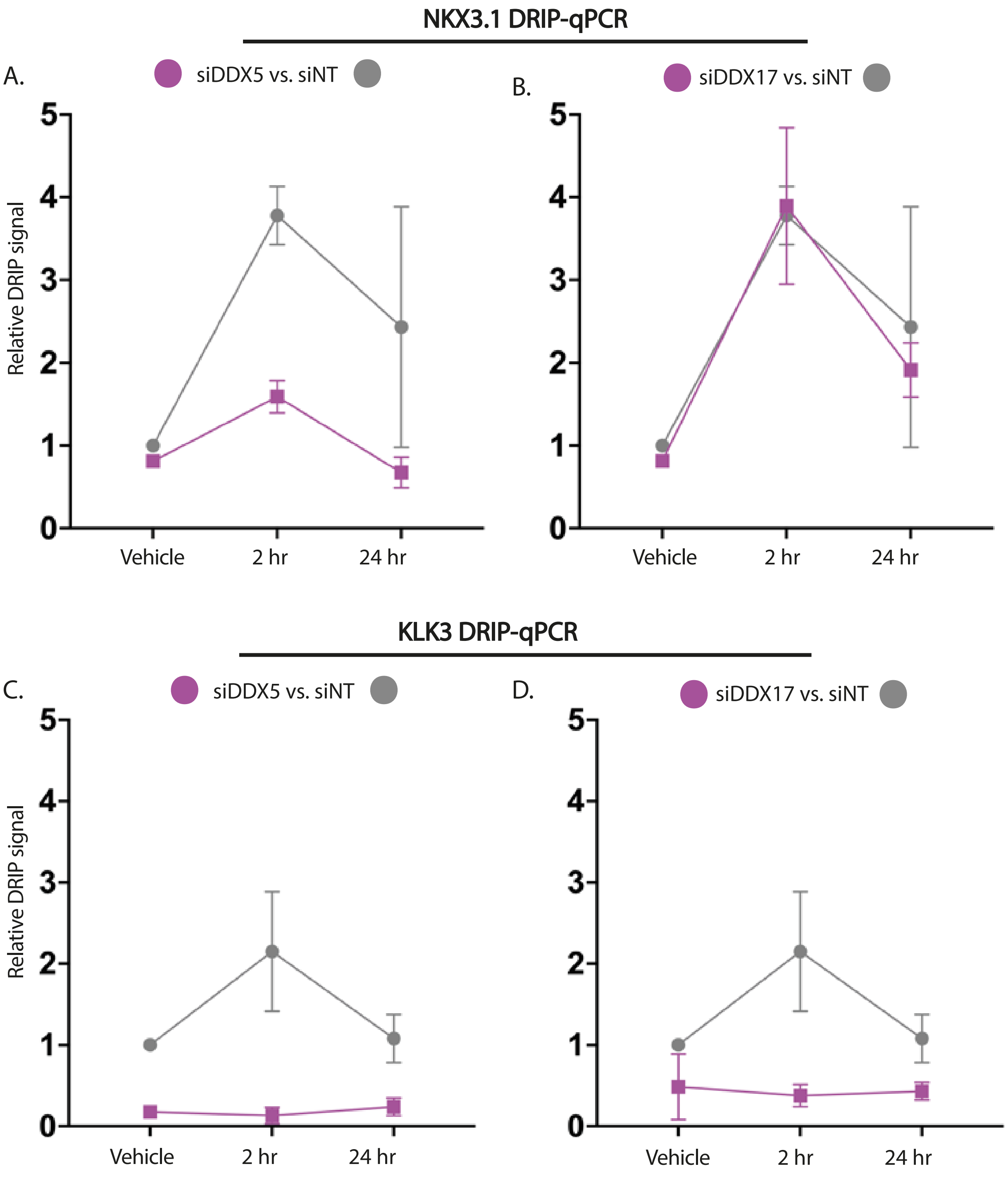


Figure 6.8 Locus specific changes in R-loop accumulation with DDX5 and DDX17 knockdown LNCaP cells had DDX5 or DDX17 depleted by siRNA then 2hr or 24 hr with DHT was done. DRIP using the S9.6 antibody coupled to qPCR for the indicated loci was used to assess dynamic R-loop changes.

6.3.6 DDX5 and DDX17 KD effect on androgen induced DNA damage

Finally, I tested whether the combination of helicase loss and androgen stimulation was associated with an increase in DNA damage. As described in the introduction androgen deprivation and androgen stimulation have both been implicated in causing DNA damage and androgen signalling has also been linked with the upregulation of DNA damage repair pathways. Excessive R-loop formation has also been suggested as a cause of DNA damage either through replication transcription collisions or through recruitment of nucleotide excision repair pathway factors such as XPF or XPG. After transfection with non-targeting siRNA, stimulation with DHT led to a transient rise in γH2AX signal followed by resolution to lower than vehicle treatment only by 24 hours. This is in keeping with previously described androgen mediated DNA repair dynamics (Goodwin *et al.*, 2013; Puc, Aggarwal and Rosenfeld, 2017). With DDX5 knockdown there was a non-significant rise in γH2AX signal under vehicle treatment conditions but prolonged (24 hours) treatment with DHT failed to resolve the signal resulting in a small but significant increase in DNA damage at this timepoint compared to cells treated with non-targeting siRNA. By contrast, DDX17 knock down did not cause an increase in γH2AX signal (**figure 6.9)**.

I then used 53BP1 immunofluorescence to validate these findings. 53BP1 is another marker of double strand breaks. Mechanistically it recruits the non-homologous end joining (NHEJ) repair machinery to ensure double strand break repair in the G1 phase of the cell cycle. 53BP1 signal has been previously shown to increase with androgen mediated DNA damage (Chatterjee *et al.*, 2019). Similar to γH2AX, the 53BP1 signal showed a decrease after 24 hours DHT treatment with non-targeting siRNA but a failure to resolve DNA damage after 24 hours DHT when DDX5 had been knocked down (**figure 6.10**). In addition, DDX17 knock down was associated with a significant reduction in 53BP1 signal after 24 hours DHT treatment. This reduction was present in comparisons with both DDX17 knockdown/vehicle treatment and with non-targeting siRNA/vehicle treatment. In addition, 53BP1 signal under vehicle treated conditions was lower with siDDX17 compared to non-targeting siRNA. These data suggest that DDX17 may have certain roles in NHEJ which are only apparent when 53BP1 – the signalling factor for this method of repair – is probed. The discrepancy in γH2AX and 53BP1 signal in the context of DDX17 loss could be accounted for by G2/S phase DNA damage repair signalling which would only be detected (in aggregate with G1 signal) with γH2AX.

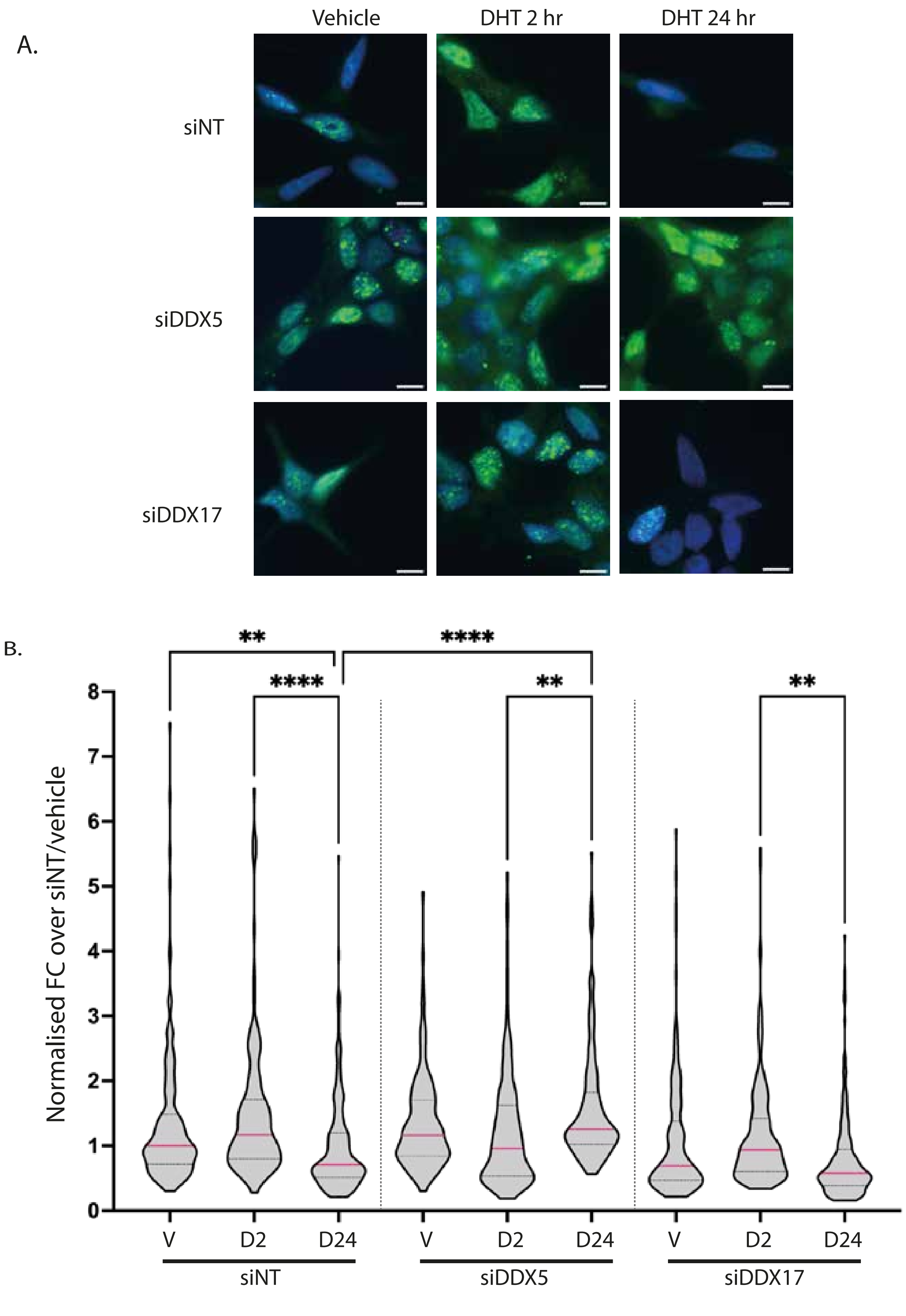


Figure 6.9 Knockdown of DDX17 and DDX5 cause increased γH2AX levels
DDX5 and DDX17 were depleted in LNCaP cells using siRNA and cells treated with 100 nM DHT for 2 or 24 hours. Immunofluorescence using an antibody against γH2AX phosphorylated at Serine 139 was used to detect double strand breaks. A. Representative immunofluorescence microscopy images. B. Quantification of immunofluorescence signal from four biological repeats per condition. All data was normalised by dividing the raw integrated density by the median raw integrated density of the baseline condition: vehicle/siNT.
Scale bar = 10 μm.

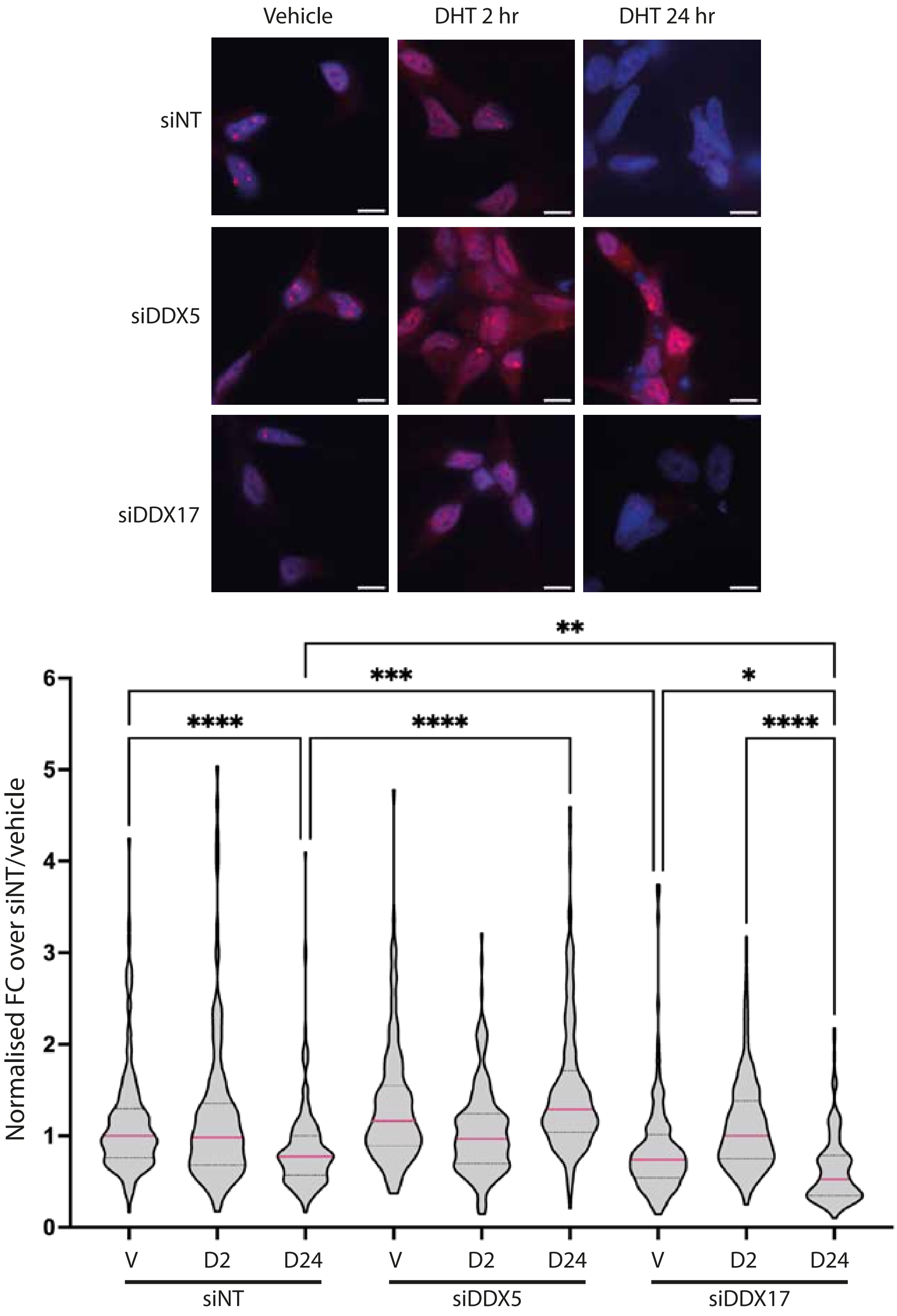


Figure 6.10 Knockdown of DDX17 and DDX5 cause increased 53BP1 levels
DDX5 and DDX17 were depleted in LNCaP cells using siRNA and cells treated with 100 nM DHT for 2 or 24 hours. Immunofluorescence using an antibody against 53BP1 was used to detect double strand breaks. A. Representative immunofluorescence microscopy images. B. Quantification of immunofluorescence signal from four biological repeats per condition. All data was normalised by dividing the raw integrated density by the median raw integrated density of the baseline condition: vehicle/siNT. Scale bar = 10 μm.

6.4 Discussion

In this chapter I have shown that two closely related DEAD box helicases – DDX5 and DDX17 – have locus specific effects on androgen regulated gene expression. Both helicases are R-loop interacting proteins and their depletion increases R-loop levels, in an androgen dependent manner in the case of DDX5 together with an associated accumulation of DNA damage. Whilst DDX17 is recruited to an androgen inducible R-loop region in the NKX3.1 promoter, depletion of DDX17 did not change R-loop levels here.

6.4.1 DDX5 has locus specific effects on androgen induced gene expression

DDX5 was previously shown to be an androgen receptor co-activator and knock down of DDX5 reduced the response of KLK3 transcription to DHT stimulation (Clark *et al.*, 2008). My data confirm this finding (**figure 6.4D**) however I found that DDX5 knockdown had a general inductive effect on NKX3.1 transcription but did not exhibit an interaction with androgen signalling. This means that even under androgen starved condition, the basal expression of NKX3.1 was increased with DDX5 depletion. This suggests that DDX5-androgen receptor co-operation does not produce the same outcome at every locus. DDX5 depletion has been implicated in R-loop mediated DNA damage in other model systems (Kang *et al.*, 2021; Sessa *et al.*, 2021; Saha *et al.*, 2022) and I observed a global increase in androgen induced DNA damage in the setting of DDX5 knock down as measured by γH2AX and 53BP1 accumulation. Whether this is due to an increase in unresolved transcription induced R-loops or to reduced resolution of DNA:RNA hybrids at double strand breaks is a question that requires further investigation. NKX3.1 has previously been shown to participate in the DNA damage response. In a mouse model of NKX3.1 depletion, γH2AX signal was higher after irradiation (Zhang *et al.*, 2016) and ATM and ATR phosphorylation were reduced by NKX3.1 knockdown in LNCaP cells (Bowen and Gelmann, 2010). Given these DNA damage repair roles of NKX3.1, I speculate that the increase of NKX3.1 expression after DDX5 depletion may be a response to androgen induced DDX5 mediated DNA damage. I attempted to charactertise the levels of DNA damage at the NKX3.1 promoter in the region of androgen induced R-loop formation by γH2AX ChIP-qPCR in the context of DDX5 or DDX17 depletion. This experiment was technically unsuccessful as the non-specific isotype matched control immunoprecipitation produced similar qPCR results to the DDX5 and DDX17 immunoprecipitations. This may have been related to technical issues with the experiment or could represent a lack of DNA damage at the loci tested.

6.4.2 DDX17 knockdown potentiates the androgen induced transcription of NKX3.1

As paralogues, the functions of DDX5 and DDX17 have been considered redundant in the context of ribosome biogenesis (Jalal, Uhlmann-Schiffler and Stahl, 2007). My data suggests that in the setting of androgen regulated transcription these two helicases have distinct functions. Using a PSA promoter luciferase reporter Clark et al (2008) showed that DDX17 knockdown failed to recapitulate the reduction in promoter activity seen with DDX5 knockdown. To my knowledge my data is the first evidence of DDX17 having an androgen dependent gene regulatory role and contrary to DDX5, depletion of DDX17 *increased* the androgen response of NKX3.1 Furthermore, DDX17 was recruited to the NKX3.1 promoter by short term androgen stimulation. DDX5 ChIP-qPCR yielded unreliable results owing to a possibly lack of specifity of the DDX5 antibody in this context so I could not discern if DDX5 was also differentially recruited to chromatin at the KLK3 and NKX3.1 loci. Suthapot *et al* (2022) used ChIP-seq to characterise DDX5 and DDX17 occupancy genome-wide in NTERA cells before and after neural differentiation. Interestingly, there was minimal overlap between DDX5 and DDX17 binding sites and differences in gene expression between DDX5 and DDX17 knockdown implying that these two helicases regulate different genes unless in the setting of neural differentiation. This study contains the only DDX17 ChIP-seq dataset on the gene expression omnibus. A similar experiment in LNCaP cells under androgen stimulation combined with gene expression profiling and DRIP-seq would help delineate the different contributions of DDX5, DDX17 and R-loops to androgen regulated gene expression.

6.4.3 DDX5 and DDX17 knock down do not increase R-loop occupancy at KLK3 or NKX3.1 loci

I hypothesised that DDX17 might resolve androgen induced R-loops at this locus and this might enhance transcription through resolution of RNA polymerase 2 pausing. However, there was no change in the R-loop profile at NKX3.1 with DDX17 knockdown. This implies either that DDX17 is not modulating R-loops at this locus or that another factor is counter-balancing the loss of DDX17 to maintain androgen induced R-loops at NKX3.1. Interestingly DDX17 depletion reduced the R-loop occupancy at the KLK3 R-loop locus and DDX5 depletion reduced R-loop levels at both NKX3.1 and KLK3. This is contrary to most of the previously described roles of DDX5 in R-loop homeostasis as, typically, an increase in R-loop levels is seen with DDX 5 loss. It is tempting to assume that DNA:RNA helicases would unwind DNA:RNA hybrids and by reducing the level of a DEAD box helicase, R-loop abundance would increase. In a recent preprint report, Leszczynska *et al* (2022) showed that hypoxic conditions in a murine model of glioblastoma reduced DDX5 expression but this was associated with *reduced* global R-loop levels. These findings mirror my data showing reduced R-loops at specific loci with DDX5 reduction by siRNA. Authors of another study suggested that a DEAH-box helicase – DHX9 – promoted the accumulation of R-loops and knockdown of DHX9 reduced the accumulation of R-loops from concomitant knockdown of splicing factors (Chakraborty, Huang and Hiom, 2018). The proposed model suggested that DHX9 unwinds RNA secondary structure to facilitate the binding of RNA-associated proteins. With the combined absence of DHX9 and RNA binding proteins such as splicing factors, the nascent RNA had a prolonged dwell time in proximity to complementary DNA and could re-anneal as an R-loop. However with depletion of only DHX9 RNA secondary structure was not resolved making it harder for the RNA to re anneal to complementary DNA. I can speculate that a similar mechanism might explain my data: DDX5 and DDX17 were originally characterised as RNA processing factors (Wong *et al.*, 2009; Samaan *et al.*, 2014) and therefore their reduction by knockdown may lead to a higher incidence of RNA secondary structure thus reducing the likelihood of R-loop formation. It is important to note that both DDX5 and DDX17 have RNA:RNA re-annealing activity in addition to their unwinding activity (Rössler, Straka and Stahl, 2001). Their actions are therefore likely to be context specific: unwinding and annealing of RNA duplexes and propagation or resolution of R-loops will depend on the interaction between nucleotide sequence, local chromatin environment and downstream signalling.

In summary I have shown that DDX5 and DDX17 have locus specific roles in androgen regulated gene expression but that these are not secondary to R-loop homeostasis at least in the genes' promoter regions. Dead box helicases have also been shown to have G-quadruplex unwinding capabilities. G-quadruplexes form in G rich DNA and often occur in the looped-out coding strand of R-loops. G-quadruplexes can also form in RNA, regulating

RNA secondary structure and metabolism. DDX5 and DDX17 were both identified as binding RNA G-quadruplexes present in the 5’ UTR of NRAS (Herdy *et al.*, 2018). Little research has been undertaken into androgen signalling and RNA G-quadruplexes however stabilisation of DNA quadruplexes in the androgen receptor promoter region led to down regulation of androgen signalling and has been suggested as an adjunct to androgen blocking therapies (Tassinari *et al.*, 2018). G quadruplex formation and resolution may provide an explanation for my DDX17 data and this could be investigated through the use of G-quadruplex stabilisers and G4 ChIP-seq.

Chapter 7. General Discussion and Future Work

7.1 Summary of results

The aim of this project was to investigate the roles that R-loops might play in androgen signalling in prostate cancer. To date, no published studies have investigated this. Starting with a computational analysis of R-loop forming sequences I found that genes up and down regulated by DHT had different RLFS characteristics. Applying this method to clinical prostate cancer cohorts revealed that genes whose expression is repressed in CRPC tended to have a higher burden of RLFS and I speculate that this is an adaptive mechanism to limit the deleterious consequences of R-loop accumulation. In addition, in analysing these data I developed an R package that characterises gene lists by their R-loop forming potential.

To quantify the R-loop burden associated with androgen regulated signalling I measured R-loops in LNCaP stimulated with DHT. Bulk analysis of genomic DNA by slot demonstrated that there were no large scale changes in R-loop levels contrary to what was found in similar experiments using oestrogen in MCF-7 cells (Stork *et al.*, 2016). To obtain locus specific data I used DRIP-seq under the same conditions. This demonstrated no correlation between androgen regulated transcription and R-loop dynamics and no pathway specific patterns of R-loop dynamics. Furthermore, I observed a marked reduction in R-loop signal at AR bound enhancers but this was not influenced by DHT treatment. Despite attempts at validating these findings I could only identify two loci that replicated the DRIP-seq findings by DRIP-qPCR. Interestingly only one of these - NKX3.1 - showed a correlation between transcription and R-loop accumulation. KLK3 R-loops remained constant implying a basal level of R-loop occupancy that is unchanged by increased transcription through this locus. These basal R-loop could act to maintain an open chromatin state for rapid transcription of this locus. However, NKX3.1 is equally rapidly transcribed by androgen signalling. This lead me to ask why did the R-loop dynamics differ between these two transcriptionally similar loci?

In an attempt to answer this question I over expressed RNase H1 and performed RNA-seq on the basis that resolution of these R-loops by this DNA:RNA hybrid specific nuclease would alter the DHT transcriptional response. Whilst this overexpression decreased R-loop according to S9.6 immunofluorescence data, the transcriptional response of androgen regulated genes remained the same. The major effect of RNase H1 was instead to interfere with mitochondrial processes and this may have dwarfed any other effects discernible through RNA-seq. In addition, there was a net downregulation of genes by RNase H1 over expression. These genes were enriched for biological processes associated with immune signalling. Although not the primary focus of this thesis, undesirable side effects of RNase H1 over expression deserve further investigation to determine if this represents a systematic bias in the current R-loop evidence base.

RNase H1 over expression did not affect the R-loop levels at the model loci NKX3.1 and KLK3. I therefore assessed if the R-loop resolving roles of DEAD box helicases DDX5 and DDX17 might intersect with androgen signalling. Depletion of each helicase increased S9.6 immunofluorescence signal indicating an increase in R-loop and for DDX5 this effect was synergistic with androgen treatment. However, knock down of neither helicase affected R-loops at my two model loci. This was surprising as DDX17 was specifically enriched at the NKX3.1 R-loop in response to DHT and its depletion increased R-loop levels at this locus.

In summary, I conclude that whilst androgen signalling does alter R-loop levels in prostate cancer cells, I have not established a specific mechanism by which this occurs and R-loop levels do not correlate with altered transcriptional output. Overall, it is unlikely that R-loops play a significant role in androgen regulated gene expression at least under the conditions tested.

7.2 Limitations of this work and future directions

7.2.1 The transcriptional context of prostate cancer in different disease stages

I selected LNCaP cells as my experimental system as they have been widely used to dissect androgen signalling (Abate-Shen and Nunes de Almeida, 2022) and also because of the availability of publicly available sequencing generated with this cell line. LNCaP cells represent metastatic prostate cancer that is still sensitive to androgen deprivation therapy. Whilst administration of DHT induces a robust androgen signalling response in this cell line, other models are associated with greater changes in transcription. VCaP and LNCaP-AR cells have a greater response to DHT both in terms of the number of differentially expressed genes and the magnitude of gene expression fold change. These models could be interrogated by the same method described here to determine if a certain level of transcription is required to reveal an interaction between androgens and R-loops. Likewise in CRPC, the transcriptional programme is different from ADT-responsive prostate cancer. The West Coast Dream Team CRPC sequencing initiative identified a subset of treated CRPC with therapeutic resistance and a neuroendocrine transcriptomic profile (Aggarwal *et al.*, 2019). This subtype represents the end stage of CRPC where cell survival has finally become uncoupled from androgen signalling. As a novel entity few models exist for probing the genomic and transcriptomic regulation of this subtype. However, it would be interesting to profile the R-loop landscape in this subtype and test whether R-loop played a role in the distinct transcriptional profiles present in these tumours.

7.2.2 DDX5 and DDX17 are not redundant in androgen signalling

In the context of ribosome biogenesis, Jalal, Uhlmann-Schiffler and Stahl (2007) concluded that DDX5 and DDX17 had redundant roles as knockdown of both helicases was required to see an effect on ribosomal RNA production. My data show that DDX5 and DDX17 have locus specific effects in the context of androgen regulated gene expression. Further work could characterise these effects genome-wide. RNA-seq under similar conditions to the data shown in figure 6.7 (DDX5 or DDX17 knock with vehicle or DHT treatment) would delineate which genes were dependent on each helicase for transcription. In addition, genome-wide R-loop profiling would answer the question of whether DDX17 was regulating R-loops elsewhere in the genome.

Whilst DDX17 didn’t modulate R-loop levels at the NKX3-1 locus it may have exerted its transcriptional effects through RNA G-quadruplex (rG4) homeostasis instead. DDX17 has been shown to interact with RNA G quadruplexes (Herdy *et al.*, 2018) and these structures could be immunoprecipitated using an anti G quadruplex antibody. The relative levels of these structures, comparing DDX17 KD to control cells, could then be assessed by sequencing or qPCR. Similarly, G quadruplex ligands specific for rG4 such as carboxypyridostatin (Ribeiro de Almeida *et al.*, 2018) could be used to test the effect of stabilising these structures on androgen regulated gene expression. Alternatively other factors may have been transcriptionally upregulated or recruited to chromatin in response to DDX17 knockdown.

7.2.3 Further development of the RloopTools package

I have developed a bioinformatics tool that can be used as an initial screen of two groups of genes divided by some characteristic such as up or down regulation, bound or not by a transcription factor or different DNA accessibility profiles. RloopTools uses sequence context to infer the likelihood of R-loop formation. Other bioinformatics tools have used sequence context to predict RNA polymerase 2 pausing (Zrimec *et al.*, 2020), gene expression profiles (Vlaming *et al.*, 2022) and G-quadruplex formation (Lee *et al.*, 2020). I could expand the findings from RloopTools to include some of these additional metrics generated from sequence context alone. RloopTools relies on the assumption that transcription and R-loop accumulation are correlated. Whilst I did not observe this correlation in my DRIP-seq data I did show that genes dysregulated by senataxin knock down have higher levels of RLFS. In addition, many studies have demonstrated a link between transcription levels and R-loop accumulation (Ginno *et al.*, 2013a; Skourti-Stathaki *et al.*, 2019; Crossley *et al.*, 2020). I speculate that the magnitude of transcriptional change induced by DHT in LNCaP cells may not have been high enough for R-loop formation and transcription to be coupled. Despite these limitations, RloopTools can provide a useful screen allowing researchers to decide if R-loops are worth pursuing as a mechanistic explanation for phenomena observed in sequencing datasets. Furthermore, the returned genomic coordinates of RLFS regions allows design of primers for exploratory DRIP-qPCR experiments instead of or prior to more resource intensive R-loop sequencing techniques.

7.2.4 The S9.6 antibody

The final limitation of this work is the use of the S9.6 antibody to detect R-loops. S9.6 was first described by Boguslawski *et al.* (1986). This antibody has been used extensively in R-loop research and has been described as the ‘workhouse’ antibody in this field (Tan-Wong, Dhir and Proudfoot, 2019). The antibody has previously been shown to have high specificity for R-loops (Phillips *et al.*, 2013) and a recent crystal structure of S9.6 demonstrated this specific interaction with a DNA:RNA hybrid (Bou-Nader *et al.*, 2022). However, research groups have raised concerns about S9.6 recognising double stranded RNA in addition to

DNA:RNA hybrids (Smolka *et al.*, 2021). Throughout this thesis I have used the control of either recombinant RNase H1 treatment or RNase H1 over expression to show that any R-loop signal observed is genuine. In addition, over the course of this thesis, numerous novel methods of mapping R-loops have been described either based around the S9.6 antibody or based around recombinant RNase H1 with various fusion proteins for purification and isolation of bound R-loops. This latter approach also has problems as RNase H doesn't necessarily recognise all R-loops and so will bias the recovery of hybrids (Malig *et al.*, 2020). However, further validation of my findings could be undertaken using an alternative method to S9.6

7.3 Summary

R-loops are dynamic non-B DNA structures with roles in DNA damage, its repair, and regulating gene expression. In this thesis I have not found a global effect linking R-loop formation and androgen signalling in prostate cancer. I have identified NKX3.1, an important prostate specific gene as a locus with dynamic R-loop accumulation in response to androgens and androgen regulated recruitment of DDX17, a helicase with roles in R-loop and RNA processing. These factors can be investigated further in the different transcriptional contexts of prostate cancer progression.

References

Abate-Shen, C. and Nunes de Almeida, F. (2022) ‘Establishment of the LNCaP Cell Line – The Dawn of an Era for Prostate Cancer Research’, *Cancer Research*, 82(9), pp. 1689–1691. doi: 10.1158/0008-5472.CAN-22-1065.

Abraham, K. J. *et al.* (2020) ‘Nucleolar RNA polymerase II drives ribosome biogenesis.’, *Nature*. Springer US, 585(7824), pp. 298–302. doi: 10.1038/s41586-020-2497-0.

Aggarwal, R. R. *et al.* (2019) ‘Whole-Genome and Transcriptional Analysis of Treatment-Emergent Small-Cell Neuroendocrine Prostate Cancer Demonstrates Intraclass Heterogeneity’, *Molecular Cancer Research*, 17(6), pp. 1235–1240. doi: 10.1158/1541-7786.MCR-18-1101.

Akman, G. *et al.* (2016) ‘Pathological ribonuclease H1 causes R-loop depletion and aberrant DNA segregation in mitochondria’, *Proceedings of the National Academy of Sciences of the United States of America*, 113(30), pp. E4276–E4285. doi: 10.1073/pnas.1600537113.

Andrews, A. M. *et al.* (2018) ‘A senataxin-associated exonuclease SAN1 is required for resistance to DNA interstrand cross-links’, *Nature Communications*. Springer US, 9(1). doi: 10.1038/s41467-018-05008-8.

Andrews, S. (2010) *FastQC: a quality control tool for high throughput sequence data.*

Arab, K. *et al.* (2019) ‘GADD45A binds R-loops and recruits TET1 to CpG island promoters’, *Nature Genetics*. Springer US, 51(2), pp. 217–223. doi: 10.1038/s41588-018-0306-6.

Argaud, D. *et al.* (2019) ‘Enhancer-mediated enrichment of interacting JMJD3–DDX21 to ENPP2 locus prevents R-loop formation and promotes transcription’, *Nucleic Acids Research*, 47(16), pp. 8424–8438. doi: 10.1093/nar/gkz560.

Armstrong, A. J. *et al.* (2020) ‘Five-year Survival Prediction and Safety Outcomes with Enzalutamide in Men with Chemotherapy-naïve Metastatic Castration-resistant Prostate Cancer from the PREVAIL Trial’, *European Urology*. European Association of Urology, 78(3), pp. 347–357. doi: 10.1016/j.eururo.2020.04.061.

Bauer, M. *et al.* (2020) ‘The ALPK1/TIFA/NF-κB axis links a bacterial carcinogen to R-loop-induced replication stress’, *Nature Communications*. Springer US, 11(1), p. 5117. doi: 10.1038/s41467-020-18857-z.

Boguslawski, S. J. *et al.* (1986) ‘Characterization of monoclonal antibody to DNA · RNA and its application to immunodetection of hybrids', *Journal of Immunological Methods*, 89(1), pp. 123–130. doi: 10.1016/0022-1759(86)90040-2.

Bonnet, A. *et al.* (2017) 'Introns Protect Eukaryotic Genomes from Transcription-Associated Genetic Instability', *Molecular Cell*, 67(4), pp. 608-621.e6. doi: 10.1016/j.molcel.2017.07.002.

Boque-Sastre, R. *et al.* (2015) 'Head-to-head antisense transcription and R-loop formation promotes transcriptional activation', *Proceedings of the National Academy of Sciences of the United States of America*, 112(18), pp. 5785–5790. doi: 10.1073/pnas.1421197112.

Bou-Nader, C. *et al.* (2022) 'Structural basis of R-loop recognition by the S9.6 monoclonal antibody', *Nature Communications*. Springer US, 13(1), pp. 1–14. doi: 10.1038/s41467-022-29187-7.

Bowen, C. *et al.* (2007) 'NKX3.1 homeodomain protein binds to topoisomerase I and enhances its activity', *Cancer Research*, 67(2), pp. 455–464. doi: 10.1158/0008-5472.CAN-06-1591.

Bowen, C. and Gelmann, E. P. (2010) 'NKX3.1 activates cellular response to DNA damage', *Cancer Research*, 70(8), pp. 3089–3097. doi: 10.1158/0008-5472.CAN-09-3138.

Buganim, Y. *et al.* (2012) 'Single-cell gene expression analyses of cellular reprogramming reveal a stochastic early and hierarchic late phase', *Cell*, 150(6), pp. 1209–1222. doi: 10.1016/j.cell.2012.08.023.Single-cell.

Burdick, M. J. *et al.* (2009) 'Comparison of Biochemical Relapse-Free Survival Between Primary Gleason Score 3 and Primary Gleason Score 4 for Biopsy Gleason Score 7 Prostate Cancer', *International Journal of Radiation Oncology Biology Physics*, 73(5), pp. 1439–1445. doi: 10.1016/j.ijrobp.2008.07.033.

Caicedo, J. C. *et al.* (2017) 'Data-analysis strategies for image-based cell profiling', *Nature Methods*, 14(9), pp. 849–863. doi: 10.1038/nmeth.4397.

Chakraborty, P., Huang, J. T. J. and Hiom, K. (2018) 'DHX9 helicase promotes R-loop formation in cells with impaired RNA splicing.', *Nature communications*. Springer US, 9(1), p. 4346. doi: 10.1038/s41467-018-06677-1.

Chappidi, N. *et al.* (2020) 'Fork Cleavage-Religation Cycle and Active Transcription Mediate Replication Restart after Fork Stalling at Co-transcriptional R-Loops', *Molecular Cell*. Elsevier Inc., 77(3), pp. 528-541.e8. doi: 10.1016/j.molcel.2019.10.026.

Chatterjee, P. *et al.* (2019) 'Supraphysiological androgens suppress prostate cancer growth through androgen receptor–mediated DNA damage', *Journal of Clinical Investigation*, 129(10), pp. 4245–4260. doi: 10.1172/jci127613.

Chatzidoukaki, O. *et al.* (2021) 'R-loops trigger the release of cytoplasmic ssDNAs leading to chronic inflammation upon DNA damage', *Science Advances*, 7(47). doi: 10.1126/sciadv.abj5769.

Chédin, F. (2016) 'Nascent Connections: R-Loops and Chromatin Patterning.', *Trends in genetics : TIG*, 32(12), pp. 828–838. doi: 10.1016/j.tig.2016.10.002.

Chedin, F. and Benham, C. J. (2020) 'Emerging roles for R-loop structures in the management of topological stress.', *The Journal of biological chemistry*. doi: 10.1074/jbc.REV119.006364.

Chen, P. B. *et al.* (2013) 'Hdac6 regulates Tip60-p400 function in stem cells', *eLife*, 2013(2), pp. 1–25. doi: 10.7554/eLife.01557.

Chen, P. B. *et al.* (2015) 'R loops regulate promoter-proximal chromatin architecture and cellular differentiation', *Nature Structural and Molecular Biology*, 22(12), pp. 999–1007. doi: 10.1038/nsmb.3122.

Chiang, H.-C. *et al.* (2019) 'BRCA1-associated R-loop affects transcription and differentiation in breast luminal epithelial cells', *Nucleic Acids Research*. Oxford University Press, 47(10), pp. 5086–5099. doi: 10.1093/nar/gkz262.

Choi, W. *et al.* (2014) 'Identification of Distinct Basal and Luminal Subtypes of Muscle-Invasive Bladder Cancer with Different Sensitivities to Frontline Chemotherapy', *Cancer Cell*. Elsevier Inc., 25(2), pp. 152–165. doi: 10.1016/j.ccr.2014.01.009.

Ciccia, A. and Elledge, S. J. (2010) 'The DNA Damage Response: Making It Safe to Play with Knives', *Molecular Cell*. Elsevier Inc., 40(2), pp. 179–204. doi: 10.1016/j.molcel.2010.09.019.

Clark, E. L. *et al.* (2008) 'The RNA helicase p68 is a novel androgen receptor coactivator involved in splicing and is overexpressed in prostate cancer', *Cancer Research*, 68(19), pp. 7938–7946. doi: 10.1158/0008-5472.CAN-08-0932.

Clark, E. L. *et al.* (2013) 'p68/DdX5 Supports β-Catenin & RNAP II during Androgen Receptor Mediated Transcription in Prostate Cancer', *PLoS ONE*, 8(1), pp. 1–10. doi: 10.1371/journal.pone.0054150.

Colaprico, A. *et al.* (2016) 'TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data', *Nucleic Acids Research*, 44(8), p. e71. doi: 10.1093/nar/gkv1507.

Cornelio, D. A. *et al.* (2017) 'Both R-loop removal and ribonucleotide excision repair activities of RNase H2 contribute substantially to chromosome stability', *DNA Repair*. Elsevier B.V., 52, pp. 110–114. doi: 10.1016/j.dnarep.2017.02.012.

Crawford, E. D. *et al.* (2019) 'Androgen-targeted therapy in men with prostate cancer: evolving practice and future considerations', *Prostate Cancer and Prostatic Diseases*. Springer US, 22(1), pp. 24–38. doi: 10.1038/s41391-018-0079-0.

Cristini, A. *et al.* (2018) 'RNA/DNA Hybrid Interactome Identifies DXH9 as a Molecular Player in Transcriptional Termination and R-Loop-Associated DNA Damage', *Cell Reports*. ElsevierCompany., 23(6), pp. 1891–1905. doi: 10.1016/j.celrep.2018.04.025.

Cristini, A. *et al.* (2019) 'Dual Processing of R-Loops and Topoisomerase I Induces Transcription-Dependent DNA Double-Strand Breaks.', *Cell reports*, 28(12), pp. 3167-3181.e6. doi: 10.1016/j.celrep.2019.08.041.

Crossley, M. P. *et al.* (2020) 'qDRIP: a method to quantitatively assess RNA-DNA hybrid formation genome-wide.', *Nucleic acids research*. Oxford University Press, pp. 1–19. doi: 10.1093/nar/gkaa500.

Crossley, M. P., Bocek, M. and Cimprich, K. A. (2019) 'R-Loops as Cellular Regulators and Genomic Threats', *Molecular Cell*. Elsevier Inc., 73(3), pp. 398–411. doi: 10.1016/j.molcel.2019.01.024.

CRUK (2021) *Prostate cancer statistics*. Available at: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/prostate-cancer (Accessed: 28 June 2021).

Cunha, G. R. *et al.* (2018) 'Development of the human prostate', *Differentiation*, 103(1), pp. 24–45. doi: 10.1016/j.diff.2018.08.005.

Deaton, A. M. and Bird, A. (2011) 'CpG islands and the regulation of transcription', *Genes and Development*, 25(10), pp. 1010–1022. doi: 10.1101/gad.2037511.

Durinck, S. *et al.* (2005) 'BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis', *Bioinformatics*, 21(16), pp. 3439–3440. doi: 10.1093/bioinformatics/bti525.

Dutta, A. *et al.* (2016) 'Identification of an NKX3.1-G9a-UTY transcriptional regulatory network that controls prostate differentiation', *Science*, 352(6293), pp. 1576–1580. doi: 10.1126/science.aad9512.

Dziak, J. J. *et al.* (2020) 'Sensitivity and specificity of information criteria', *Briefings in Bioinformatics*, 21(2), pp. 553–565. doi: 10.1093/bib/bbz016.

Egevad, L. *et al.* (2016) ‘International Society of Urological Pathology (ISUP) grading of prostate cancer – An ISUP consensus on contemporary grading’, *Apmis*, 124(6), pp. 433–435. doi: 10.1111/apm.12533.

Eylert, M. F. *et al.* (2014) ‘Falling bladder cancer incidence from 1990 to 2009 is not producing universal mortality improvements’, *Journal of Clinical Urology*, 7(2), pp. 90–98. doi: 10.1177/2051415813492724.

Feng, W. *et al.* (2020) ‘DHX33 recruits Gadd45a to cause DNA demethylation and regulate a subset of gene transcription’, *Molecular and Cellular Biology*, (April). doi: 10.1128/mcb.00460-19.

Fujita, K. and Nonomura, N. (2019) ‘Role of androgen receptor in prostate cancer: A review’, *World Journal of Men?s Health*, 37(3), pp. 288–295. doi: 10.5534/wjmh.180040.

Gandaglia, G. *et al.* (2014) ‘Distribution of metastatic sites in patients with prostate cancer: A population-based analysis’, *Prostate*, 74(2), pp. 210–216. doi: 10.1002/pros.22742.

Gibbons, H. R. *et al.* (2018) ‘Divergent lncRNA GATA3-AS1 Regulates GATA3 Transcription in T-Helper 2 Cells’, *Frontiers in immunology*, 9(October), p. 2512. doi: 10.3389/fimmu.2018.02512.

Ginno, P. A. *et al.* (2012) ‘R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters.’, *Molecular cell*, 45(6), pp. 814–25.

Ginno, P. A. *et al.* (2013a) ‘GC skew at the 5’ and 3’ ends of human genes links R-loop formation to epigenetic regulation and transcription termination’, *Genome Research*, 23(10), pp. 1590–1600. doi: 10.1101/gr.158436.113.

Ginno, P. A. *et al.* (2013b) ‘GC skew at the 5’ and 3’ ends of human genes links R-loop formation to epigenetic regulation and transcription termination’, *Genome Research*, 23(10), pp. 1590–1600. doi: 10.1101/gr.158436.113.

Gleason, D. F. (1966) ‘Classification of prostatic carcinomas.’, *Cancer chemotherapy reports*, 50(3), pp. 125–8.

Goodwin, J. F. *et al.* (2013) ‘A Hormone–DNA Repair Circuit Governs the Response to Genotoxic Insult’, *Cancer Discovery*, 3(11), pp. 1254–1271. doi: 10.1158/2159-8290.CD-13-0108.

Gorthi, A. *et al.* (2018) ‘EWS-FLI1 increases transcription to cause R-Loops and block BRCA1 repair in Ewing sarcoma’, *Nature*. Nature Publishing Group, 555(7696), pp. 387–391. doi: 10.1038/nature25748.

Goulielmaki, E. *et al.* (2021) ‘The splicing factor XAB2 interacts with ERCC1-XPF and XPG for R-loop processing’, *Nature Communications*. Springer US, 12(1). doi: 10.1038/s41467-021-23505-1.

Grasso, C. S. *et al.* (2012) ‘The mutational landscape of lethal castration-resistant prostate cancer’, *Nature*, 487(7406), pp. 239–243. doi: 10.1038/nature11125.

Griffin, J. *et al.* (2022) ‘Gene of the month: NKX3.1’, *Journal of clinical pathology*, 75(6), pp. 361–364. doi: 10.1136/jclinpath-2021-208073.

Griso, A. B. *et al.* (2022) ‘Mechanisms of Cisplatin Resistance in HPV Negative Head and Neck Squamous Cell Carcinomas’, *Cells*, 11(3). doi: 10.3390/cells11030561.

Groh, M. *et al.* (2017) ‘Senataxin: Genome Guardian at the Interface of Transcription and Neurodegeneration’, *Journal of Molecular Biology*. Elsevier Ltd, 429(21), pp. 3181–3195. doi: 10.1016/j.jmb.2016.10.021.

Grosse, A., Bartsch, S. and Baniahmad, A. (2012) ‘Androgen receptor-mediated gene repression’, *Molecular and Cellular Endocrinology*. Elsevier Ireland Ltd, 352(1–2), pp. 46–56. doi: 10.1016/j.mce.2011.06.032.

Grunseich, C. *et al.* (2018) ‘Senataxin Mutation Reveals How R-Loops Promote Transcription by Blocking DNA Methylation at Gene Promoters’, *Molecular Cell*. Elsevier Inc., 69(3), pp. 426-437.e7. doi: 10.1016/j.molcel.2017.12.030.

GTeX consortium (2020) ‘The GTEx Consortium atlas of genetic regulatory effects across human tissues’, *Science*, 369(6509), pp. 1318–1330. doi: 10.1126/science.aaz1776.

Guo, H. *et al.* (2021) ‘NR4A1 regulates expression of immediate early genes, suppressing replication stress in cancer’, *Molecular Cell*. Elsevier Inc., 81(19), pp. 4041-4058.e15. doi: 10.1016/j.molcel.2021.09.016.

Gupta, S. K., Luo, L. and Yen, L. (2018) ‘RNA-mediated gene fusion in mammalian cells.’, *Proceedings of the National Academy of Sciences of the United States of America*, 115(52), pp. E12295–E12304. doi: 10.1073/pnas.1814704115.

Haffner, M. C. *et al.* (2010) ‘Androgen-induced TOP2B-mediated double-strand breaks and prostate cancer gene rearrangements.’, *Nature genetics*, 42(8), pp. 668–75. doi: 10.1038/ng.613.

El Hage, A. *et al.* (2010) ‘Loss of Topoisomerase I leads to R-loop-mediated transcriptional blocks during ribosomal RNA synthesis’, *Genes and Development*, 24(14), pp. 1546–1558. doi: 10.1101/gad.573310.

Halász, L. *et al.* (2017) 'RNA-DNA hybrid (R-loop) immunoprecipitation mapping: an analytical workflow to evaluate inherent biases.', *Genome research*, 27(6), pp. 1063–1073. doi: 10.1101/gr.219394.116.

Hamperl, S. *et al.* (2017) 'Transcription-Replication Conflict Orientation Modulates R-Loop Levels and Activates Distinct DNA Damage Responses', *Cell*. Elsevier Inc., 170(4), pp. 774-786.e19. doi: 10.1016/j.cell.2017.07.043.

Hankey, W., Chen, Z. and Wang, Q. (2020) 'Shaping Chromatin States in Prostate Cancer by Pioneer Transcription Factors', *Cancer Research*, 80(12), pp. 2427–2436. doi: 10.1158/0008-5472.CAN-19-3447.

Hartono, S. R., Korf, I. F. and Chédin, F. (2015) 'GC skew is a conserved property of unmethylated CpG island promoters across vertebrates', *Nucleic Acids Research*, 43(20), pp. 9729–9741. doi: 10.1093/nar/gkv811.

He, W. W. *et al.* (1997) 'A novel human prostate-specific, androgen-regulated homeobox gene (NKX3.1) that maps to 8p21, a region frequently deleted in prostate cancer', *Genomics*, 43(1), pp. 69–77. doi: 10.1006/geno.1997.4715.

Herdy, B. *et al.* (2018) 'Analysis of NRAS RNA G-quadruplex binding proteins reveals DDX3X as a novel interactor of cellular G-quadruplex containing transcripts', *Nucleic Acids Research*. Oxford University Press, 46(21), pp. 11592–11604. doi: 10.1093/nar/gky861.

Herold, S. *et al.* (2019) 'Recruitment of BRCA1 limits MYCN-driven accumulation of stalled RNA polymerase', *Nature*. Springer US, 567(7749), pp. 545–549. doi: 10.1038/s41586-019-1030-9.

Hodroj, D. *et al.* (2017) ' An ATR -dependent function for the Ddx19 RNA helicase in nuclear R-loop metabolism ', *The EMBO Journal*, 36(9), pp. 1182–1198. doi: 10.15252/embj.201695131.

Holt, I. J. (2019) 'Survey and summary: The mitochondrial R-loop', *Nucleic Acids Research*. Oxford University Press, 47(11), pp. 5480–5489. doi: 10.1093/nar/gkz277.

Horoszewicz, J. S. *et al.* (1983) 'LNCaP model of human prostatic carcinoma.', *Cancer research*, 43(4), pp. 1809–18.

Hsieh, C. *et al.* (2014) 'Enhancer RNAs participate in androgen receptor-driven looping that selectively enhances gene activation.', *Proceedings of the National Academy of Sciences of the United States of America*, 111(20), pp. 7319–24. doi: 10.1073/pnas.1324151111.

Huang, C. C. F. *et al.* (2021) 'Functional mapping of androgen receptor enhancer activity', *Genome Biology*. Genome Biology, 22(1), pp. 1–26. doi: 10.1186/s13059-021-02339-6.

Huggins, C., Stevens, R. and Hodges, C. (1941) 'Studies on prostate cancer 2: The effect of castration on advanced carcinoma of the prostate gland', *Archives of Surgery*, 43(2), p. 209. doi: 10.1001/archsurg.1941.01210140043004.

Ilic, D. *et al.* (2018) 'Prostate cancer screening with prostate-specific antigen (PSA) test: A systematic review and meta-analysis', *BMJ (Online)*, 362, pp. 1–12. doi: 10.1136/bmj.k3519.

Illumina (2021) *Bubble products in sequencing libraries: causes, identification, and workflow recommendations*. Available at: https://emea.support.illumina.com/bulletins/2019/10/bubble-products-in-sequencing-libraries--causes--identification-.html (Accessed: 1 September 2022).

Jalal, C., Uhlmann-Schiffler, H. and Stahl, H. (2007) 'Redundant role of DEAD box proteins p68 (Ddx5) and p72/p82 (Ddx17) in ribosome biogenesis and cell proliferation', *Nucleic Acids Research*, 35(11), pp. 3590–3601. doi: 10.1093/nar/gkm058.

Jauregui-Lozano, J. *et al.* (2022) 'Proper control of R-loop homeostasis is required for maintenance of gene expression and neuronal function during aging', *Aging Cell*, 21(2), pp. 1–13. doi: 10.1111/acel.13554.

Jenjaroenpun, P. *et al.* (2015) 'QmRLFS-finder: A model, web server and stand-alone tool for prediction and analysis of R-loop forming sequences', *Nucleic Acids Research*, 43(W1), pp. W527–W534. doi: 10.1093/nar/gkv344.

Johnson, D. E. *et al.* (2020) 'Head and neck squamous cell carcinoma', *Nature Reviews Disease Primers*, 6(1). doi: 10.1038/s41572-020-00224-3.

Jones, S. E. *et al.* (2017) 'ATR Is a therapeutic target in synovial sarcoma', *Cancer Research*, 77(24), pp. 7014–7026. doi: 10.1158/0008-5472.CAN-17-2056.

Ju, B.-G. *et al.* (2006) 'A topoisomerase IIbeta-mediated dsDNA break required for regulated transcription.', *Science (New York, N.Y.)*, 312(5781), pp. 1798–802. doi: 10.1126/science.1127196.

Jung, Y. and Lippard, S. J. (2006) 'RNA polymerase II blockage by cisplatin-damaged DNA: Stability and polyubiquitylation of stalled polymerase', *Journal of Biological Chemistry*. Â© 2006 ASBMB. Currently published by Elsevier Inc; originally published by American Society for Biochemistry and Molecular Biology., 281(3), pp. 1361–1370. doi: 10.1074/jbc.M509688200.

Jurga, M. *et al.* (2021) 'USP11 controls R-loops by regulating senataxin proteostasis', *Nature*

Communications. Springer US, 12(1). doi: 10.1038/s41467-021-25459-w.

Kamoun, A. *et al.* (2020) 'A Consensus Molecular Classification of Muscle-invasive Bladder Cancer.', *European urology*, 77(4), pp. 420–433. doi: 10.1016/j.eururo.2019.09.006.

Kang, H. J. *et al.* (2021) 'Thrap3 promotes R-loop resolution via interaction with methylated DDX5', *Experimental and Molecular Medicine*. Springer US, 53(10), pp. 1602–1611. doi: 10.1038/s12276-021-00689-6.

Karyka, E. *et al.* (2022) 'SMN-deficient cells exhibit increased ribosomal DNA damage', *Life Science Alliance*, 5(8), pp. 1–19. doi: 10.26508/lsa.202101145.

Kasprzak, P. *et al.* (2020) 'Six Years of Shiny in Research - Collaborative Development of Web Tools in R', *R Journal*, 12, pp. 1–23. doi: 10.32614/rj-2021-004.

Kimura, T. *et al.* (2021) 'Global trends of latent prostate cancer in autopsy studies', *Cancers*, 13(2), pp. 1–11. doi: 10.3390/cancers13020359.

Kirkwood, B. and Sterne, J. (2011) *Essential Medical Statistics*. 2nd edn. Blackwell Publishing.

Kisiala, M. *et al.* (2020) 'Restriction endonucleases that cleave RNA/DNA heteroduplexes bind dsDNA in A-like conformation', *Nucleic Acids Research*. Oxford University Press, 48(12), pp. 6954–6969. doi: 10.1093/nar/gkaa403.

Kolbanovskiy, M. *et al.* (2017) 'The Nonbulky DNA Lesions Spiroiminodihydantoin and 5-Guanidinohydantoin Significantly Block Human RNA Polymerase II Elongation in Vitro', *Biochemistry*, 56(24), pp. 3008–3018. doi: 10.1021/acs.biochem.7b00295.

Koo, C. X. G. E. *et al.* (2015) 'RNA polymerase III regulates cytosolic RNA:DNA hybrids and intracellular microRNA expression', *Journal of Biological Chemistry*, 290(12), pp. 7463–7473. doi: 10.1074/jbc.M115.636365.

Korenchuk, S. *et al.* (no date) 'VCaP, a cell-based model system of human prostate cancer.', *In vivo (Athens, Greece)*, 15(2), pp. 163–8.

Kotsantis, P. *et al.* (2016) 'Increased global transcription activity as a mechanism of replication stress in cancer', *Nature Communications*, 7. doi: 10.1038/ncomms13087.

Kounatidou, E. *et al.* (2019) 'A novel CRISPR-engineered prostate cancer cell line defines the AR-V transcriptome and identifies PARP inhibitor sensitivities', *Nucleic Acids Research*, (1). doi: 10.1093/nar/gkz286.

Kuznetsov, V. A. *et al.* (2018) 'Toward predictive R-loop computational biology: Genome-scale prediction of R-loops reveals their association with complex promoter structures, G-quadruplexes and transcriptionally active enhancers', *Nucleic Acids Research*. Oxford

University Press, 46(15), pp. 7566–7585. doi: 10.1093/nar/gky554.

Lam, H. M. *et al.* (2019) ‘Durable Response of Enzalutamide-resistant Prostate Cancer to Supraphysiological Testosterone Is Associated with a Multifaceted Growth Suppression and Impaired DNA Damage Response Transcriptomic Program in Patient-derived Xenografts’, *European Urology*. European Association of Urology, 77(2), pp. 144–155. doi: 10.1016/j.eururo.2019.05.042.

Lambo, S. *et al.* (2019) ‘The molecular landscape of ETMR at diagnosis and relapse’, *Nature*. Springer US, 576(December 2018). doi: 10.1038/s41586-019-1815-x.

Landsverk, H. B. *et al.* (2019) ‘Regulation of ATR activity via the RNA polymerase II associated factors CDC73 and PNUTS-PP1’, *Nucleic Acids Research*, 47(4), pp. 1797–1813. doi: 10.1093/nar/gky1233.

Lawrence, M. S. *et al.* (2015) ‘Comprehensive genomic characterization of head and neck squamous cell carcinomas’, *Nature*, 517(7536), pp. 576–582. doi: 10.1038/nature14129.

Lee, C. Y. *et al.* (2020) ‘R-loop induced G-quadruplex in non-template promotes transcription by successive R-loop formation’, *Nature Communications*. Springer US, 11(1), pp. 1–15. doi: 10.1038/s41467-020-17176-7.

Lesnik, E. A. and Freier, S. M. (1995) ‘Relative Thermodynamic Stability of DNA, RNA, and DNA:RNA Hybrid Duplexes: Relationship with Base Composition and Structure’, *Biochemistry*, 34(34), pp. 10807–10815. doi: 10.1021/bi00034a013.

Leszczynska, K. B. *et al.* (2022) ‘Hypoxia-mediated regulation of DDX5 through decreased chromatin accessibility and post-translational targeting restricts R-loop accumulation’, *bioRxiv*, p. 2022.04.30.490097. doi: 10.1101/2022.04.30.490097.

Li, H. *et al.* (2009) ‘The Sequence Alignment/Map format and SAMtools.’, *Bioinformatics (Oxford, England)*, 25(16), pp. 2078–9. doi: 10.1093/bioinformatics/btp352.

Li, H. *et al.* (2022) ‘The Cumulative Formation of R-loop Interacts with Histone Modifications to Shape Cell Reprogramming’, *International Journal of Molecular Sciences*, 23(3). doi: 10.3390/ijms23031567.

Li, M. *et al.* (2015) ‘RECQ5-dependent SUMOylation of DNA topoisomerase I prevents transcription-associated genome instability’, *Nature Communications*. Nature Publishing Group, 6, pp. 1–13. doi: 10.1038/ncomms7720.

Li, Y. *et al.* (2020) ‘R-loops coordinate with SOX2 in regulating reprogramming to pluripotency’, *Science Advances*, 6(24). doi: 10.1126/sciadv.aba0777.

Li, Y. *et al.* (2022) ‘Exaggerated false positives by popular differential expression methods when analyzing human population samples’, *Genome Biology*. BioMed Central, 23(1), pp. 1–13. doi: 10.1186/s13059-022-02648-4.

Lin, R. *et al.* (2022) ‘R-loopBase: A knowledgebase for genome-wide R-loop formation and regulation’, *Nucleic Acids Research*. Oxford University Press, 50(D1), pp. D303–D315. doi: 10.1093/nar/gkab1103.

Lin, W. L. *et al.* (2022) ‘DDX18 prevents R-loop-induced DNA damage and genome instability via PARP-1’, *Cell Reports*. The Authors, 40(3), p. 111089. doi: 10.1016/j.celrep.2022.111089.

Lin, Y. *et al.* (2010) ‘R loops stimulate genetic instability of CTG·CAG repeats’, *Proceedings of the National Academy of Sciences of the United States of America*, 107(2), pp. 692–697. doi: 10.1073/pnas.0909740107.

Linder, P. and Jankowsky, E. (2011) ‘From unwinding to clamping ĝ€" the DEAD box RNA helicase family’, *Nature Reviews Molecular Cell Biology*. Nature Publishing Group, 12(8), pp. 505–516. doi: 10.1038/nrm3154.

Liu, Y.-N. *et al.* (2008) ‘Activated Androgen Receptor Downregulates E-Cadherin Gene Expression and Promotes Tumor Metastasis’, *Molecular and Cellular Biology*, 28(23), pp. 7096–7108. doi: 10.1128/mcb.00449-08.

Livak, K. J. and Schmittgen, T. D. (2001) ‘Analysis of relative gene expression data using real-time quantitative PCR and the 2-ΔΔCT method’, *Methods*, 25(4), pp. 402–408. doi: 10.1006/meth.2001.1262.

Lockhart, A. *et al.* (2019) ‘RNase H1 and H2 Are Differentially Regulated to Process RNA-DNA Hybrids.’, *Cell reports*. ElsevierCompany., 29(9), pp. 2890-2900.e5. doi: 10.1016/j.celrep.2019.10.108.

Love, M. I., Huber, W. and Anders, S. (2014) ‘Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2’, *Genome Biology*, 15(12), pp. 1–21. doi: 10.1186/s13059-014-0550-8.

Luo, X. hui *et al.* (2019) ‘KLF14 potentiates oxidative adaptation via modulating HO-1 signaling in castrate-resistant prostate cancer’, *Endocrine-Related Cancer*, 26(1), pp. 181–195. doi: 10.1530/ERC-18-0383.

Malig, M. *et al.* (2020) ‘Ultra-Deep Coverage Single-Molecule R-loop Footprinting Reveals Principles of R-loop Formation’, *Journal of Molecular Biology*. Elsevier Ltd. doi: 10.1016/j.jmb.2020.02.014.

Malinen, M. *et al.* (2017) ‘Crosstalk between androgen and pro-inflammatory signaling remodels androgen receptor and NF-κB cistrome to reprogram the prostate cancer cell transcriptome’, *Nucleic Acids Research*, 45(2), pp. 619–630. doi: 10.1093/nar/gkw855.

Manzo, S. G. *et al.* (2018) ‘DNA Topoisomerase I differentially modulates R-loops across the human genome’, *Genome Biology*. Genome Biology, 19(1), pp. 1–18. doi: 10.1186/s13059-018-1478-1.

Martin, M. (2011) ‘Cutadapt removes adapter sequences from high-throughput sequencing reads’, *EMBnet.journal*, 17(1), p. 10. doi: 10.14806/ej.17.1.200.

Marzec, J. *et al.* (2021) ‘The transcriptomic landscape of prostate cancer development and progression: An integrative analysis’, *Cancers*, 13(2), pp. 1–24. doi: 10.3390/cancers13020345.

Massie, C. E. *et al.* (2011) ‘The androgen receptor fuels prostate cancer by regulating central metabolism and biosynthesis’, *EMBO Journal*. Nature Publishing Group, 30(13), pp. 2719–2733. doi: 10.1038/emboj.2011.158.

McCann, J. L. *et al.* (2021) ‘R-loop homeostasis and cancer mutagenesis promoted by the DNA cytosine deaminase APOBEC3B’, *bioRxiv*, p. 2021.08.30.458235.

Mersaoui, S. Y. *et al.* (2019) ‘ Arginine methylation of the DDX 5 helicase RGG / RG motif by PRMT 5 regulates resolution of RNA:DNA hybrids ’, *The EMBO Journal*, 38(15), pp. 1–20. doi: 10.15252/embj.2018100986.

Miller, A. J., Roman, B. and Norstrom, E. (2016) ‘A method for easily customizable gradient gel electrophoresis’, *Analytical Biochemistry*. Elsevier Inc, 509, pp. 12–14. doi: 10.1016/j.ab.2016.07.003.

Morova, T. *et al.* (2020) ‘Androgen receptor-binding sites are highly mutated in prostate cancer’, *Nature Communications*. Springer US, 11(1). doi: 10.1038/s41467-020-14644-y.

Mottet, N. *et al.* (2021) ‘EAU-EANM-ESTRO-ESUR-SIOG Guidelines on Prostate Cancer—2020 Update. Part 1: Screening, Diagnosis, and Local Treatment with Curative Intent’, *European Urology*, 79(2), pp. 243–262. doi: 10.1016/j.eururo.2020.09.042.

Nadel, J. *et al.* (2015) ‘RNA:DNA hybrids in the human genome have distinctive nucleotide characteristics, chromatin composition, and transcriptional relationships’, *Epigenetics and Chromatin*. BioMed Central, 8(1), pp. 1–19. doi: 10.1186/s13072-015-0040-6.

National Institute for Health and Care Excellence [NICE] (2019) ‘Prostate cancer: diagnosis and management’.

Ngollo, M. *et al.* (2017) 'Global analysis of H3K27me3 as an epigenetic marker in prostate cancer progression', *BMC Cancer*. BMC Cancer, 17(1), pp. 1–8. doi: 10.1186/s12885-017-3256-y.

Nguyen, H. D. *et al.* (2018) 'Spliceosome Mutations Induce R Loop-Associated Sensitivity to ATR Inhibition in Myelodysplastic Syndromes.', *Cancer research*, 78(18), pp. 5363–5374. doi: 10.1158/0008-5472.CAN-17-3970.

Nouruzi, S. *et al.* (2022) 'ASCL1 activates neuronal stem cell-like lineage programming through remodeling of the chromatin landscape in prostate cancer', *Nature Communications*. Springer US, 13(1), p. 2282. doi: 10.1038/s41467-022-29963-5.

Nowotny, M. *et al.* (2005) 'Crystal structures of RNase H bound to an RNA/DNA hybrid: Substrate specificity and metal-dependent catalysis', *Cell*, 121(7), pp. 1005–1016. doi: 10.1016/j.cell.2005.04.024.

Nowotny, M. *et al.* (2007) 'Structure of Human RNase H1 Complexed with an RNA/DNA Hybrid: Insight into HIV Reverse Transcription', *Molecular Cell*, 28(2), pp. 264–276. doi: 10.1016/j.molcel.2007.08.015.

Olivieri, M. *et al.* (2020) 'A Genetic Map of the Response to DNA Damage in Human Cells', *Cell*, 182(2), pp. 481-496.e21. doi: 10.1016/j.cell.2020.05.040.

OMIM (no date) *PROGRESSIVE EXTERNAL OPHTHALMOPLEGIA WITH MITOCHONDRIAL DNA DELETIONS, AUTOSOMAL RECESSIVE 2; PEOB2*. Available at: https://www.omim.org/entry/616479 (Accessed: 14 September 2022).

Padmanabhan, K. *et al.* (2012) 'Feedback Regulation of Transcriptional Termination by the Mammalian Circadian Clock PERIOD Complex', *Science*, 337(6094), pp. 599–602. doi: 10.1126/science.1221592.

Pagès, H. *et al.* (2020) 'Biostrings: Efficient manipulation of biological strings.' R Package.

Pagès, H. (2022) 'BSgenome: Software infrastructure for efficient representation of full genomes and their SNPs.' R Package.

Pérez-Calero, C. *et al.* (2020) 'UAP56/DDX39B is a major cotranscriptional RNA–DNA helicase that unwinds harmful R loops genome-wide', *Genes and Development*, 34(13–14), pp. 1–15. doi: 10.1101/GAD.336024.119.

Pezone, A. *et al.* (2019) 'RNA Stabilizes Transcription-Dependent Chromatin Loops Induced By Nuclear Hormones', *Scientific Reports*. Springer US, 9(1), pp. 1–12. doi: 10.1038/s41598-019-40123-6.

Pfaffl, M. W. (2001) ‘A new mathematical model for relative quantification in real-time RT-PCR.’, *Nucleic acids research*, 29(9), p. e45. doi: 10.1093/nar/29.9.e45.

Phillips, D. D. *et al.* (2013) ‘The sub-nanomolar binding of DNA-RNA hybrids by the single-chain Fv fragment of antibody S9.6.’, *Journal of molecular recognition : JMR*, 26(8), pp. 376–81. doi: 10.1002/jmr.2284.

Pinter, S. *et al.* (2021) ‘A functional LSD1 coregulator screen reveals a novel transcriptional regulatory cascade connecting R-loop homeostasis with epigenetic regulation’, *Nucleic Acids Research*. Oxford University Press, pp. 1–21. doi: 10.1093/nar/gkab180.

Pleguezuelos-Manzano, C. *et al.* (2020) ‘Mutational signature in colorectal cancer caused by genotoxic pks + E. coli’, *Nature*, 580(7802), pp. 269–273. doi: 10.1038/s41586-020-2080-8.

Polkinghorn, W. R. *et al.* (2013) ‘Androgen receptor signaling regulates DNA repair in prostate cancers’, *Cancer Discovery*, 3(11), pp. 1245–1253. doi: 10.1158/2159-8290.CD-13-0172.

Prendergast, L. *et al.* (2020) ‘Resolution of R-loops by INO80 promotes DNA replication and maintains cancer cell proliferation and viability’, *Nature Communications*. Springer US, 11(1), pp. 1–18. doi: 10.1038/s41467-020-18306-x.

Promonet, A. *et al.* (2020) ‘Topoisomerase 1 prevents replication stress at R-loop-enriched transcription termination sites’, *Nature Communications*. Springer US, 11(1). doi: 10.1038/s41467-020-17858-2.

Puc, J. *et al.* (2015) ‘Ligand-dependent enhancer activation regulated by topoisomerase-I activity’, *Cell*. Elsevier, 160(3), pp. 367–380. doi: 10.1016/j.cell.2014.12.023.

Puc, J., Aggarwal, A. K. and Rosenfeld, M. G. (2017) ‘Physiological functions of programmed DNA breaks in signal-induced transcription.’, *Nature reviews. Molecular cell biology*, 18(8), pp. 471–476. doi: 10.1038/nrm.2017.43.

Quigley, D. A. *et al.* (2018) ‘Genomic Hallmarks and Structural Variation in Metastatic Prostate Cancer’, *Cell*, 174(3), pp. 758-769.e9. doi: 10.1016/j.cell.2018.06.039.

Ramachandran, S. *et al.* (2021) ‘Hypoxia-induced SETX links replication stress with the unfolded protein response’, *Nature Communications*, 12(1), pp. 1–14. doi: 10.1038/s41467-021-24066-z.

Ramirez, P. *et al.* (2021) ‘R-loop analysis by dot-blot’, *Journal of Visualized Experiments*, 2021(167), pp. 1–14. doi: 10.3791/62069.

Reyes, A. *et al.* (2020) ‘RNase H1 Regulates Mitochondrial Transcription and Translation via the Degradation of 7S RNA’, *Frontiers in Genetics*, 10(January), pp. 1–11. doi: 10.3389/fgene.2019.01393.

Ribeiro de Almeida, C. *et al.* (2018) ‘RNA Helicase DDX1 Converts RNA G-Quadruplex Structures into R-Loops to Promote IgH Class Switch Recombination’, *Molecular Cell*. Elsevier Inc., 70(4), pp. 650-662.e8. doi: 10.1016/j.molcel.2018.04.001.

Rigby, R. E. *et al.* (2014) ‘RNA:DNA hybrids are a novel molecular pattern sensed by TLR9’, *EMBO Journal*, 33(6), pp. 542–558. doi: 10.1002/embj.201386117.

Robertson, A. G. *et al.* (2017) ‘Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer’, *Cell*, 171(3), pp. 540-556.e25. doi: 10.1016/j.cell.2017.09.007.

Ross-Innes, C. S. *et al.* (2012) ‘Differential oestrogen receptor binding is associated with clinical outcome in breast cancer’, *Nature*. Nature Publishing Group, 481(7381), pp. 389–393. doi: 10.1038/nature10730.

Rössler, O. G., Straka, A. and Stahl, H. (2001) ‘Rearrangement of structured RNA via branch migration structures catalysed by the highly related DEAD-box proteins p68 and p72’, *Nucleic Acids Research*, 29(10), pp. 2088–2096. doi: 10.1093/nar/29.10.2088.

Roy, D. *et al.* (2010) ‘Competition between the RNA Transcript and the Nontemplate DNA Strand during R-Loop Formation In Vitro: a Nick Can Serve as a Strong R-Loop Initiation Site’, *Molecular and Cellular Biology*, 30(1), pp. 146–159. doi: 10.1128/mcb.00897-09.

Roy, D. and Lieber, M. R. (2009) ‘G Clustering Is Important for the Initiation of Transcription-Induced R-Loops In Vitro, whereas High G Density without Clustering Is Sufficient Thereafter’, *Molecular and Cellular Biology*, 29(11), pp. 3124–3133. doi: 10.1128/mcb.00139-09.

Sabino, J. C. *et al.* (2022) ‘Epigenetic reprogramming by TET enzymes impacts co-transcriptional R-loops’, *eLife*, 11, pp. 1–22. doi: 10.7554/ELIFE.69476.

Saha, S. *et al.* (2022) ‘Resolution of R-loops by topoisomerase III-β (TOP3B) in coordination with the DEAD-box helicase DDX5.’, *Cell reports*. ElsevierCompany., 40(2), p. 111067. doi: 10.1016/j.celrep.2022.111067.

Salji, M. J. *et al.* (2022) ‘Multi-omics & pathway analysis identify potential roles for tumor N-acetyl aspartate accumulation in murine models of castration-resistant prostate cancer’, *iScience*, 25(4). doi: 10.1016/j.isci.2022.104056.

Samaan, S. *et al.* (2014) ‘The Ddx5 and Ddx17 RNA helicases are cornerstones in the complex regulatory array of steroid hormone-signaling pathways’, *Nucleic Acids Research*, 42(4), pp. 2197–2207. doi: 10.1093/nar/gkt1216.

Sanz, L. A. *et al.* (2016) 'Prevalent, Dynamic, and Conserved R-Loop Structures Associate with Specific Epigenomic Signatures in Mammals.', *Molecular cell*, 63(1), pp. 167–78. doi: 10.1016/j.molcel.2016.05.032.

Sanz, L. A., Castillo-Guzman, D. and Chédin, F. (2021) 'Mapping r-loops and rna:Dna hybrids with s9.6-based immunoprecipitation methods', *Journal of Visualized Experiments*, 2021(174), pp. 1–15. doi: 10.3791/62455.

Sanz, L. A. and Chédin, F. (2019) 'High-resolution, strand-specific R-loop mapping via S9.6-based DNA-RNA immunoprecipitation and high-throughput sequencing.', *Nature protocols*. Nature Publishing Group, 14(6), pp. 1734–1755. doi: 10.1038/s41596-019-0159-1.

Scher, H. I. *et al.* (2012) 'Increased Survival with Enzalutamide in Prostate Cancer after Chemotherapy', *New England Journal of Medicine*, 367(13), pp. 1187–1197. doi: 10.1056/nejmoa1207506.

Seiler, R. *et al.* (2017) 'Impact of Molecular Subtypes in Muscle-invasive Bladder Cancer on Predicting Response and Survival after Neoadjuvant Chemotherapy', *European Urology*, 72(4), pp. 544–554. doi: 10.1016/j.eururo.2017.03.030.

Sessa, G. *et al.* (2021) 'BRCA2 promotes DNA-RNA hybrid resolution by DDX5 helicase at DNA breaks to facilitate their repair‡', *The EMBO Journal*, 40(7), pp. 1–25. doi: 10.15252/embj.2020106018.

Shen, W. *et al.* (2017) 'Dynamic nucleoplasmic and nucleolar localization of mammalian RNase H1 in response to RNAP I transcriptional R-loops.', *Nucleic acids research*. Oxford University Press, 45(18), pp. 10672–10692. doi: 10.1093/nar/gkx710.

Shi, W. *et al.* (2017) 'Ssb1 and Ssb2 cooperate to regulate mouse hematopoietic stem and progenitor cells by resolving replicative stress', *Blood*, 129(18), pp. 2471–2478. doi: 10.1182/blood-2016-06-725093.

Shi, X.-B. *et al.* (2002) 'Functional analysis of 44 mutant androgen receptors from human prostate cancer.', *Cancer research*, 62(5), pp. 1496–502.

Skourti-Stathaki, K. *et al.* (2019) 'R-Loops Enhance Polycomb Repression at a Subset of Developmental Regulator Genes', *Molecular Cell*. Elsevier Inc., 73(5), pp. 930-945.e4. doi: 10.1016/j.molcel.2018.12.016.

Skourti-Stathaki, K., Kamieniarz-Gdula, K. and Proudfoot, N. J. (2014) 'R-loops induce repressive chromatin marks over mammalian gene terminators', *Nature*. Nature Publishing

Group, 516(7531), pp. 436–439. doi: 10.1038/nature13787.

Skourti-Stathaki, K., Proudfoot, N. J. and Gromak, N. (2011) ‘Human senataxin resolves RNA/DNA hybrids formed at transcriptional pause sites to promote Xrn2-dependent termination.’, *Molecular cell*. Elsevier Inc., 42(6), pp. 794–805. doi: 10.1016/j.molcel.2011.04.026.

Smolka, J. A. *et al.* (2021) ‘Recognition of rna by the s9.6 antibody creates pervasive artifacts when imaging rna:Dna hybrids’, *Journal of Cell Biology*, 220(6). doi: 10.1083/jcb.202004079.

Sollier, J. *et al.* (2014) ‘Transcription-coupled nucleotide excision repair factors promote R-loop-induced genome instability.’, *Molecular cell*, 56(6), pp. 777–85. doi: 10.1016/j.molcel.2014.10.020.

Song, C. *et al.* (2017) ‘SIRT7 and the DEAD-box helicase DDX21 cooperate to resolve genomic R loops and safeguard genome stability’, *Genes and Development*, 31(13), pp. 1370–1381. doi: 10.1101/gad.300624.117.

Sooreshjani, M. A. *et al.* (2021) ‘Limk2-nkx3.1 engagement promotes castration-resistant prostate cancer’, *Cancers*, 13(10), pp. 1–24. doi: 10.3390/cancers13102324.

Sridhara, S. C. *et al.* (2017) ‘Transcription Dynamics Prevent RNA-Mediated Genomic Instability through SRPK2-Dependent DDX23 Phosphorylation’, *Cell Reports*, 18(2), pp. 334–343. doi: 10.1016/j.celrep.2016.12.050.

Stolz, R., Sulthana, S., Hartono, Stella R, *et al.* (2019) ‘Interplay between DNA sequence and negative superhelicity drives R-loop structures.’, *Proceedings of the National Academy of Sciences of the United States of America*, pp. 1–10. doi: 10.1073/pnas.1819476116.

Stolz, R., Sulthana, S., Hartono, Stella R., *et al.* (2019) ‘Interplay between DNA sequence and negative superhelicity drives R-loop structures’, *Proceedings of the National Academy of Sciences of the United States of America*, 116(13), pp. 6260–6269. doi: 10.1073/pnas.1819476116.

Stork, C. T. *et al.* (2016) ‘Co-transcriptional R-loops are the main cause of estrogen-induced DNA damage.’, *eLife*, 5, pp. 1–21. doi: 10.7554/eLife.17548.

Suthapot, P. *et al.* (2022) ‘The RNA helicases DDX5 and DDX17 facilitate neural differentiation of human pluripotent stem cells NTERA2’, *Life Sciences*. Elsevier Inc., 291(January), p. 120298. doi: 10.1016/j.lfs.2021.120298.

Tan-Wong, S. M., Dhir, S. and Proudfoot, N. J. (2019) ‘R-Loops Promote Antisense Transcription across the Mammalian Genome.’, *Molecular cell*. Elsevier Inc., pp. 1–17. doi:

10.1016/j.molcel.2019.10.002.

Tan, P. Y. *et al.* (2012) ‘Integration of Regulatory Networks by NKX3-1 Promotes Androgen-Dependent Prostate Cancer Survival’, *Molecular and Cellular Biology*, 32(2), pp. 399–414. doi: 10.1128/mcb.05958-11.

Tang, Z. *et al.* (2017) ‘GEPIA: A web server for cancer and normal gene expression profiling and interactive analyses’, *Nucleic Acids Research*, 45(W1), pp. W98–W102. doi: 10.1093/nar/gkx247.

Tassinari, M. *et al.* (2018) ‘Down-Regulation of the Androgen Receptor by G-Quadruplex Ligands Sensitizes Castration-Resistant Prostate Cancer Cells to Enzalutamide’, *Journal of Medicinal Chemistry*, 61(19), pp. 8625–8638. doi: 10.1021/acs.jmedchem.8b00502.

Teng, Y. *et al.* (2018) ‘ROS-induced R loops trigger a transcription-coupled but BRCA1/2-independent homologous recombination pathway through CSB’, *Nature Communications*. Springer US, 9(1). doi: 10.1038/s41467-018-06586-3.

Teply, B. A. *et al.* (2018) ‘Bipolar androgen therapy in men with metastatic castration-resistant prostate cancer after progression on enzalutamide: an open-label, phase 2, multicohort study’, *The Lancet Oncology*. Elsevier Ltd, 19(1), pp. 76–86. doi: 10.1016/S1470-2045(17)30906-3.

Thomas, M., White, R. L. and Davis, R. W. (1976) ‘Hybridization of RNA to double stranded DNA: Formation of R loops’, *Proceedings of the National Academy of Sciences of the United States of America*, 73(7), pp. 2294–2298. doi: 10.1073/pnas.73.7.2294.

Thongthip, S. *et al.* (2022) ‘Relationships between genome-wide R-loop distribution and classes of recurrent DNA breaks in neural stem/progenitor cells’, *Scientific Reports*. Nature Publishing Group UK, 12(1), pp. 1–12. doi: 10.1038/s41598-022-17452-0.

Toropainen, S. *et al.* (2016) ‘Global analysis of transcription in castration-resistant prostate cancer cells uncovers active enhancers and direct androgen receptor targets’, *Scientific Reports*. Nature Publishing Group, 6(September), pp. 15–18. doi: 10.1038/srep33510.

Tu, Q. *et al.* (2022) ‘RETSAT associates with DDX39B to promote fork restarting and resistance to gemcitabine based chemotherapy in pancreatic ductal adenocarcinoma’, *Journal of Experimental & Clinical Cancer Research*. BioMed Central, pp. 1–22. doi: 10.1186/s13046-022-02490-3.

Tuduri, S. *et al.* (2009) ‘Topoisomerase I suppresses genomic instability by preventing interference between replication and transcription’, *Nature Cell Biology*. Nature Publishing

Group, 11(11), pp. 1315–1324. doi: 10.1038/ncb1984.

Velichko, A. K. *et al.* (2019) ‘Hypoosmotic stress induces R loop formation in nucleoli and ATR/ATM-dependent silencing of nucleolar transcription’, *Nucleic Acids Research*. Oxford University Press, 47(13), pp. 6811–6825. doi: 10.1093/nar/gkz436.

Veluvolu, U. *et al.* (2015) ‘The Molecular Taxonomy of Primary Prostate Cancer’, *Cell*, 163(4), pp. 1011–1025. doi: 10.1016/j.cell.2015.10.025.

Villarreal, O. D. *et al.* (2020) ‘Genome-wide R-loop analysis defines unique roles for DDX5, XRN2, and PRMT5 in DNA/RNA hybrid resolution’, *Life Science Alliance*, 3(10), pp. 1–14. doi: 10.26508/LSA.202000762.

Viswanathan, S. R. *et al.* (2018) ‘Structural Alterations Driving Castration-Resistant Prostate Cancer Revealed by Linked-Read Genome Sequencing’, *Cell*, 174(2), pp. 433-447.e19. doi: 10.1016/j.cell.2018.05.036.

Vlaming, H. *et al.* (2022) ‘Screening thousands of transcribed coding and non-coding regions reveals sequence determinants of RNA polymerase II elongation potential’, *Nature Structural & Molecular Biology*. Springer US, (1), p. 2021.06.01.446655. doi: 10.1038/s41594-022-00785-9.

Wahba, L. *et al.* (2011) ‘RNase H and Multiple RNA Biogenesis Factors Cooperate to Prevent RNA:DNA Hybrids from Generating Genome Instability’, *Molecular Cell*. Elsevier, 44(6), pp. 978–988. doi: 10.1016/j.molcel.2011.10.017.

Wang, D. *et al.* (2011) ‘Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA’, *Nature*. Nature Publishing Group, 474(7351), pp. 390–397. doi: 10.1038/nature10006.

Wang, E. C., Lee, W. R. and Armstrong, A. J. (2022) ‘Second generation anti-androgens and androgen deprivation therapy with radiation therapy in the definitive management of high-risk prostate cancer’, *Prostate Cancer and Prostatic Diseases*. Springer US, (April). doi: 10.1038/s41391-022-00598-3.

Wang, I. X. *et al.* (2018) ‘Human proteins that interact with RNA/DNA hybrids’, *Genome Research*, 28(9), pp. 1405–1414. doi: 10.1101/gr.237362.118.

Wang, K. *et al.* (2021) ‘Genomic profiling of native R loops with a DNA-RNA hybrid recognition sensor’, *Science Advances*, 7(8), pp. 1–18. doi: 10.1126/sciadv.abe3516.

Wong, H. Y. *et al.* (2009) ‘DNA dependent recruitment of DDX17 and other interacting proteins by the human androgen receptor’, *Biochimica et Biophysica Acta - Proteins and*

Proteomics. Elsevier B.V., 1794(2), pp. 193–198. doi: 10.1016/j.bbapap.2008.11.001.

Wongsurawat, T. *et al.* (2012) ‘Quantitative model of R-loop forming structures reveals a novel level of RNA-DNA interactome complexity’, *Nucleic Acids Research*, 40(2). doi: 10.1093/nar/gkr1075.

Wongsurawat, T. *et al.* (2020) ‘R-loop-forming Sequences Analysis in Thousands of Viral Genomes Identify A New Common Element in Herpesviruses’, *Scientific Reports*. Springer US, 10(1), pp. 1–9. doi: 10.1038/s41598-020-63101-9.

Wu, D. *et al.* (2011) ‘Androgen receptor-driven chromatin looping in prostate cancer’, *Trends in Endocrinology and Metabolism*. Elsevier Ltd, 22(12), pp. 474–480. doi: 10.1016/j.tem.2011.07.006.

Wu, D. *et al.* (2014) ‘Three-tiered role of the pioneer factor GATA2 in promoting androgen-dependent gene expression in prostate cancer’, *Nucleic Acids Research*, 42(6), pp. 3607–3622. doi: 10.1093/nar/gkt1382.

Wulfridge, P. and Sarma, K. (2021) ‘A nuclease-and bisulfite-based strategy captures strand-specific r-loops genomewide’, *eLife*, 10, pp. 1–15. doi: 10.7554/eLife.65146.

Wynne, P. *et al.* (2007) ‘Enhanced repair of DNA interstrand crosslinking in ovarian cancer cells from patients following treatment with platinum-based chemotherapy’, *British Journal of Cancer*, 97(7), pp. 927–933. doi: 10.1038/sj.bjc.6603973.

Xie, W. *et al.* (2017) ‘Metastasis-free survival is a strong Surrogate of overall survival in localized prostate cancer’, *Journal of Clinical Oncology*, 35(27), pp. 3097–3104. doi: 10.1200/JCO.2017.73.9987.

Xu, C. *et al.* (2021) ‘R-loop resolution promotes co-transcriptional chromatin silencing’, *Nature Communications*. Springer US, 12(1). doi: 10.1038/s41467-021-22083-6.

Yang, Y. *et al.* (2022) ‘Transcription-replication conflicts in primordial germ cells necessitate the Fanconi anemia pathway to safeguard genome stability’, *Proceedings of the National Academy of Sciences of the United States of America*, 119(34), pp. 1–11. doi: 10.1073/pnas.2203208119.

Yu, Z. *et al.* (2020) ‘DDX5 resolves R-loops at DNA double-strand breaks to promote DNA repair and avoid chromosomal deletions’, *NAR Cancer*, 2(3), pp. 1–19. doi: 10.1093/narcan/zcaa028.

Yuan, F. *et al.* (2019) ‘Molecular determinants for enzalutamide-induced transcription in prostate cancer’, *Nucleic Acids Research*. Oxford University Press, 47(19), pp. 10104–10114.

doi: 10.1093/nar/gkz790.

Zhang, H. *et al.* (2016) 'Nkx3.1 controls the DNA repair response in the mouse prostate', *Prostate*, 76(4), pp. 402–408. doi: 10.1002/pros.23131.

Zhang, L. H. *et al.* (2020) 'The SUMOylated METTL8 Induces R-loop and Tumorigenesis via m3C', *iScience*. Elsevier Inc., 23(3), p. 100968. doi: 10.1016/j.isci.2020.100968.

Zhang, X. *et al.* (2017) 'Attenuation of RNA polymerase II pausing mitigates BRCA1-associated R-loop accumulation and tumorigenesis', *Nature Communications*. Nature Publishing Group, 8(May 2017), pp. 1–11. doi: 10.1038/ncomms15908.

Zhao, H. *et al.* (2018) 'Identification of valid reference genes for mRNA and microRNA normalisation in prostate cancer cell lines', *Scientific Reports*. Springer US, 8(1), pp. 1–13. doi: 10.1038/s41598-018-19458-z.

Zhao, S. *et al.* (2018) 'Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: PolyA+ selection versus rRNA depletion', *Scientific Reports*. Springer US, 8(1), pp. 1–12. doi: 10.1038/s41598-018-23226-4.

Zhou, H. *et al.* (2020) 'H3K9 Demethylation-Induced R-Loop Accumulation Is Linked to Disorganized Nucleoli', *Frontiers in Genetics*, 11(February), pp. 1–15. doi: 10.3389/fgene.2020.00043.

Zrimec, J. *et al.* (2020) 'Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure', *Nature Communications*. Springer US, 11(1), p. 6141. doi: 10.1038/s41467-020-19921-4.

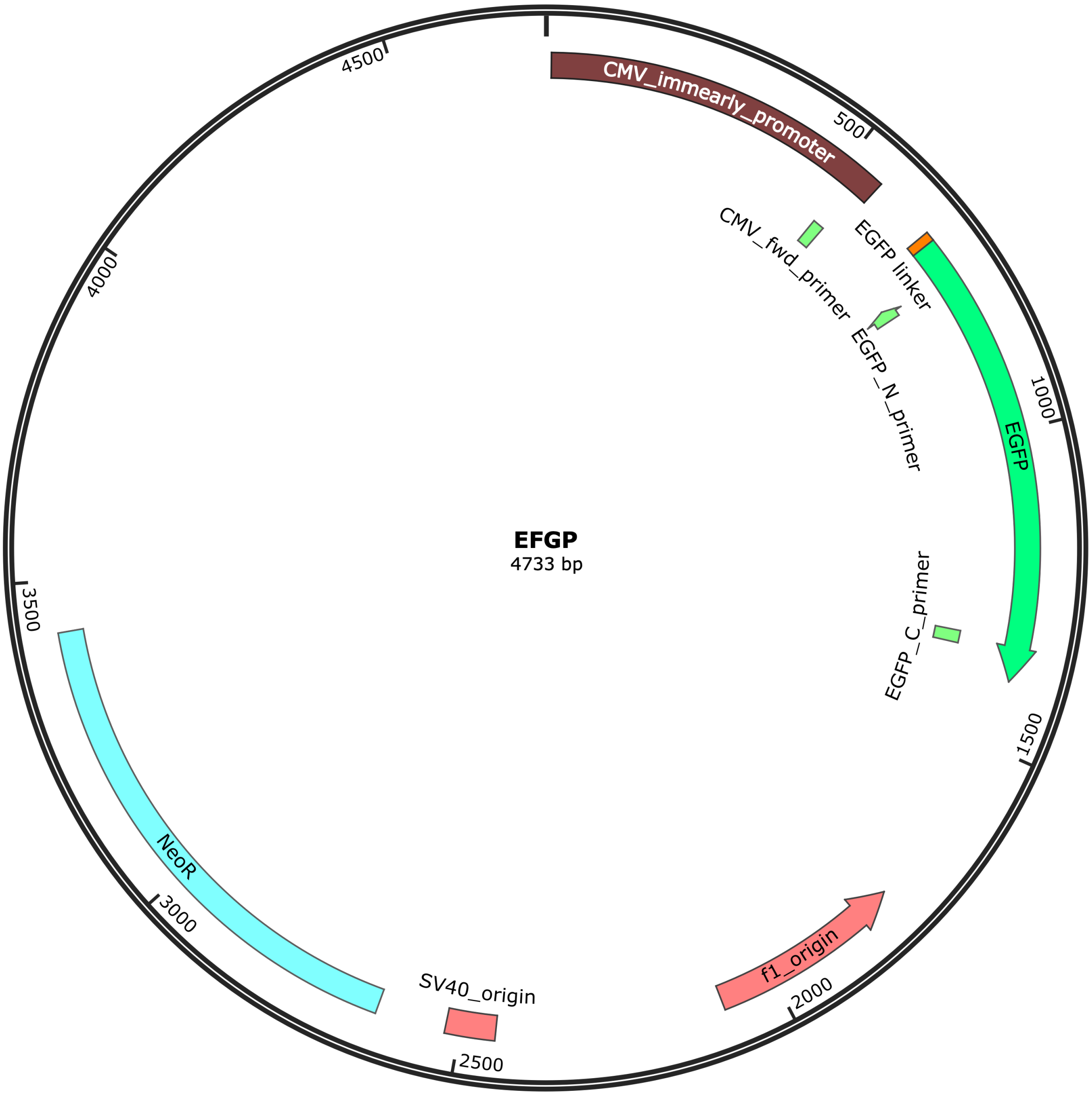
Appendix 1: Primers

All primers were ordered from Integrated DNA Technologies.

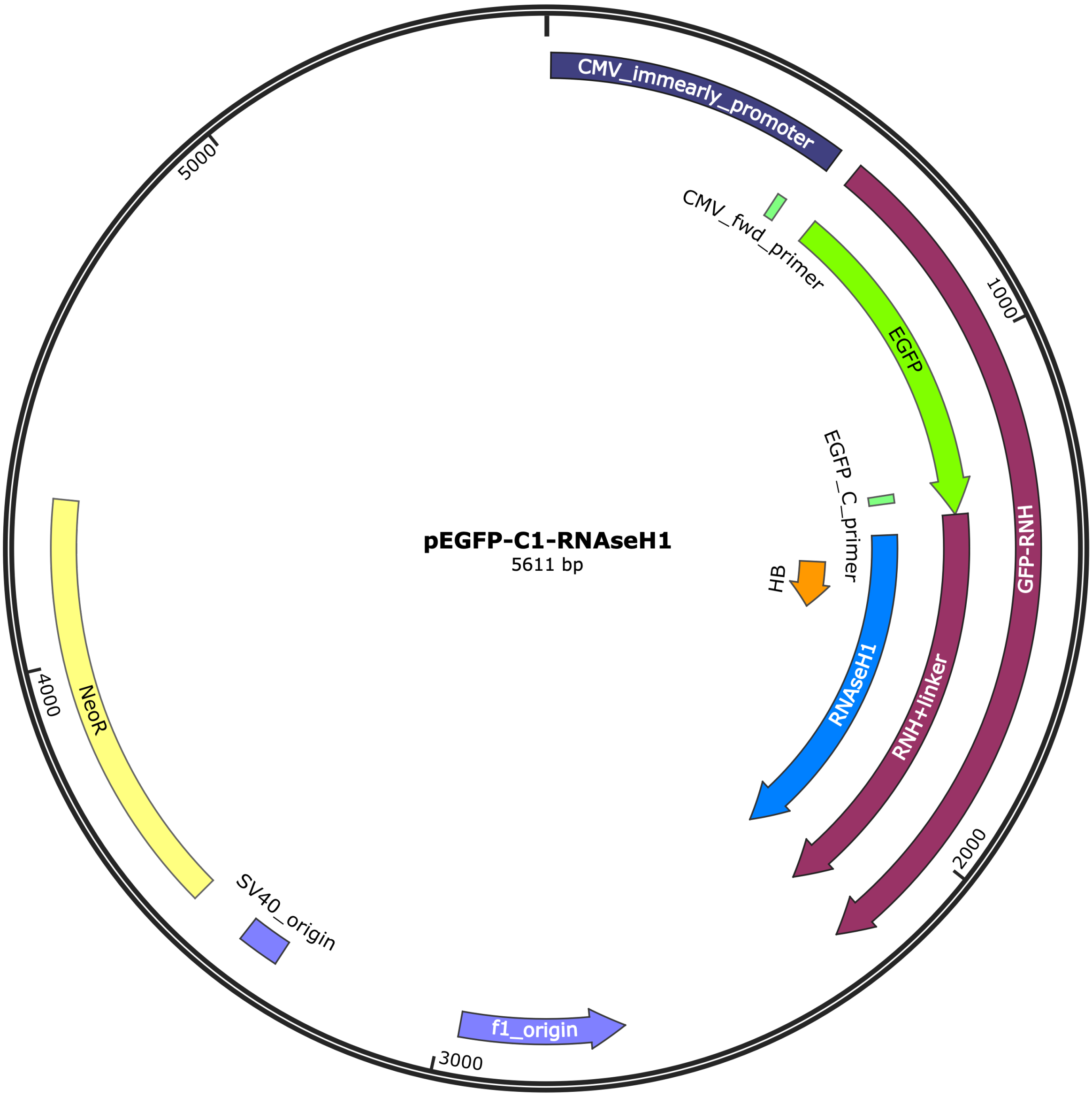
Sequences are given as 5’ to 3’.

Target	Forward Sequence	Reverse Sequence	Application
GAPDH	AGGTCGGAGTCAACGGATTT	ATGAAGGGGTCATTGATGGCA	RT-qPCR
NKX3.1	AGAGCCAGAGGGAGGACG	GACCCCAAGTGCCTTTCTGG	RT-qPCR
KLK3	ACTGCATCAGGAACAAAAGCG	AGCTGTGGCTGACCTGAAAT	RT-qPCR
NKX3.1	CAGGCTCATTCTGGGCTTTA	CCAGGAGAGAGAGTCCACTTAT	DRIP/ChIP-qPCR
KLK3	GCCCATGTCTGTTTCTCTATGT	CAGAGAGAATGAAAGGGCAGAG	DRIP/ChIP-qPCR
RPL13A	AGGTGCCTTGCTCACAGAGT	GGTTGCATTGCCCTCATTAC	DRIP/ChIP-qPCR
TFPT	TCTGGGAGTCCAAGCAGACT	AAGGAGCCACTGAAGGGTTT	DRIP/ChIP-qPCR
SNRPN	GCCAAATGAGTGAGGATGGT	TCCTCTCTGCCTGACTCCAT	DRIP/ChIP-qPCR
CBX1	ACCCGAGGTTTGTAACTGTATT	CAATTCACTCGACGTTACTCCT	DRIP-qPCR
NATL8	GGCATTTGGTCTGGGAGTAG	CTGCTCAGAGTAGTGGTCAAAG	DRIP-qPCR
KLF14	TCTGGTGGGTTCTCTGGA	GAGGTGCGACGACTTGTAATA	DRIP-qPCR
TUSC1	GCTGAACAGCAAAGCACTC	AAAGGAGGCCGGGAATTT	DRIP-qPCR
IG1	GAGCAGGTGAGAGAGAGAGATT	GGCCTAAAGGGAAGTGTGTAAG	DRIP-qPCR
IG2	GTCTACCACACAGGGATGTTTC	CTCTCCTTCCTCCCTCCTAAA	DRIP-qPCR
IG3	GGGAATGAGGCTATTGGTAAGG	GACATTCTCTGGAAGGACTTGG	DRIP-qPCR

Appendix 2: Plasmid maps



EGFP-EV plasmid



EGFP-RNH plasmid

Appendix 3: siRNA sequences

siRNA Name	Accession number	Sequence
siDDX17-1	J-013450-09	CGAUAGAGCUGGUUAUGCU
siDDX17-2	J-013450-10	CGAUAGAGCUGGUUAUGCU
siDDX17-3	J-013450-11	CAAAUGCAGUGUAGAGCUA
siDDX17-4	J-013450-12	GGAGUGCAUUUGAUAGUUA
siDDX5-1	J-003774-05	GCAAAUGUCAUGGAUGUUA
siDDX5-2	J-003774-06	CAACCUACCUUGUCCUUGA
siDDX5-3	J-003774-07	GCAUGUCGCUUGAAGUCUA
siDDX5-4	J-003774-08	CCAAAUAUGCACAAUGGUA
siNT (Non-targeting control)	D-001810-04-05	UGGUUUACAUGUUUUCCUA

All siRNA were modified with 3’ UU overhangs.

Appendix 4: Publicly available data used in this work

Author/Group and Year	PubMed ID	Data Type	Experiment Description
Toropainen et al, 2016	27641228	GRO-seq; FASTQ	VCaP cells treated with 100 nM DHT or vehicle for 2 hours
Niskanen et al, 2017	27672034	GRO-seq; FASTQ	LNCaP cells treated with 100 nM DHT or vehicle for 2 hours
The Cancer Genome Atlas, Prostate Cancer 2015	26544944	RNA-seq; counts	333 patients with primary, localised prostate cancer
West Coast Dream Team, 2018	30033370	RNA-seq; counts	99 patients with metastatic, castration resistant prostate cancer
The Cancer Genome Atlas, Bladder Cancer, 2018	28988769	RNA-seq; counts	412 patients with muscle invasive bladder cancer
Ramachandran et al, 2021	34140498	RNA-seq; DE genes list	RKO cells cultured in 0.1 % O_2 +/- Senataxin depletion by siRNA
Hannah Crane, 2021 (unpublished)	NA	RNA-seq; DE genes list	HPV- H&N SCC cisplatin resistant cell line and parental counterpart
Yuan *et al.*, 2019	31501863	RNA-seq; FASTQ	LNCaP cells treated with 100 nM DHT or vehicle for 24 hr
Malinen *et al.*, 2017	27672034	BED file of genomic intervals	LNCaP cells treated with 100 nM DHT or vehicle for 2 hr
Huang *et al.*, 2021	33975627	BED file of genomic intervals	STARR-seq AR binding sites from LNCaP cells treated with 100 nM DHT or vehicle for 2 hr
Morova *et al.*, 2020	32047165	BED file of genomic intervals	Mutated AR binding sites from whole genome sequencing of primary prostate cancer

DE: Differentially expressed, H&N SCC: Head and neck squamous cell carcinoma

Appendix 5: Software, packages and libraries used in this work

Library/Package Name	**Version**
Bash/ Unix shell	
Burrow-Wheels Aligner (BWA)	0.7.17
Samtools	1.14
Bedtools	2.29.2
Cutadapt	3.4
Kallisto	0.46.1
Deeptools	3.5.0
FastQC	0.11.9
R	
biomaRt	2.46.3
BSGenome	1.64.0
clusterProfiler	3.18.1
DESeq2	1.30.1
edgeR	3.32.1
fitdistrplus	1.1
GenomicFeatures	1.42.3
GenomicRanges	1.42.0
ggplot2	3.3.6
MASS	7.3.56
pheatmap	1.0.12
Rmarkdown	2.14
Rstudio	1.2.5033 (“Orange blossom”), running R 4.0.3
SummarizedExperiment	1.20.0
TCGA biolinks	2.25.2
Tidyverse	1.3.2

Appendix 6: Fluorescence microscopy settings

Fluorophore	**Exposure time (ms)**	**Intensity (%)**
	S9.6 immunofluorescence	
FITC	150	25
Texas Red	20	25
DAPI	15	15
	γH2AX and 53BP1 immunofluorescence	
FITC	10	50
Texas Red	20	25
DAPI	15	15

Appendix 7: Antibodies used in immunofluorescence and Western blotting

Antibody	Concentration	Manufacturer	Species
Immunofluorescence			
S9.6	1:500	Isolated from hybridoma by BioServ (Sheffield, UK)	Mouse
γH2AX	1:1000	Abcam	Mouse
53BP1	1:1000	Abcam	Mouse
GFP	1:1000	Abcam	Rabbit
488 conjugated secondary antibody	1:1000	Abcam	Mouse/Rabbit
596 conjugated secondary antibody	1:1000	Abcam	Mouse/Rabbit
Western blot			
DDX5	1:1000	ProteinTech	Rabbit
DDX17	1:1000	ProteinTech	Rabbit
dsDNA	1:500	ProteinTech	Mouse
GFP	1:2000	Abcam	Rabbit
HRP-IgG conjugated secondary antibody	1:4000	BioRad	Mouse/Rabbit