Gender-associated gene expression and chromatin accessibility of human urothelium

Benjamin Charles Hopkins

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

The University of Leeds

Faculty of Medicine and Health

September 2019

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

Acknowledgements

Carrying out this PhD has been the most challenging but equally rewarding endeavour. It would not have been possible without all the amazing friends, family, and work colleagues that surround me, and I would like to take this opportunity to acknowledge that support.

First and foremost, I would like to express the sincerest gratitude to my amazing supervisors. Thank you to wonderful and wise Prof. Margaret Knowles giving me the opportunity to work in her lab, and for her guidance, mentorship, and expertise over the last four years. I am equally grateful to my second supervisor Dr. Julie Burns who has always been there to help support and troubleshoot any technical, intellectual, or emotional problems I may have encountered throughout, you truly are the goddess of the lab. Special thanks are due to Dr. Carolyn Hurst who has acted in many ways as an additional supervisor by dedicating her time and parting with wisdom throughout the course of my PhD.

Thank you to the rest of Lab 5 for providing such an excellent environment to carry out my PhD. The expertise, time, and kindness of all current and past members of Lab 5 has been invaluable. I would also like to mention all members of level 8, particularly for their emotional support throughout the writing of this thesis.

Acknowledgements are due to CRUK Cambridge for providing essential bioinformatics training. Thank you also to Dr. Alastair Droop and Dr. Martin Callaghan for their bioinformatics training at Leeds University. I would also like to mention Dr. Amel Saadi, Dr. Matt Care, and Dr. Ron Chen for insightful discussions regarding ChIP-seq, ATAC-seq, and bioinformatics.

I am incredibly lucky to have the most loving family, and I would like to thank all of them for their emotional support. Particularly my incredible mother, Mrs Dawn Hopkins, I couldn't wish for a better mother and you are a huge inspiration to me. I also want to thank my gorgeous boyfriend Dr. Andrew Galloway, meeting you alone has made this PhD the best decision I've ever made.

Thank you to LICAP for generously funding my tuition and research.

I would like to dedicate this thesis to my late father Malcolm Hopkins. I feel his support and encouragement as much now as ever. I miss you dearly.

Abstract

Bladder cancer is 7th most common type of cancer in the UK, and presents up to 4 times more often men than in women, even when adjusting for environmental factors such as smoking and occupation. Bladder cancer also has the highest rates of mutations in chromatin modifier genes compared to any other cancer type. Despite this, studies regarding the epigenome of bladder and bladder cancer are lacking. This study considers the genome-wide transcriptional and chromatin accessibility landscape of healthy urothelium, and aims to identify gender-associated differences promoting the gender biases observed in bladder cancer. The study is the first to establish reliable protocols for chromatin immunoprecipitation (ChIP) of histone marks, and an assay for transposase-accessible chromatin followed by next generation sequencing (ATAC-seq) in normal urothelium, and the first to carry out transcriptional profiling on normal human urothelial cells (NHUC).

Affymetrix HTA2.0 microarrays using three models of healthy urothelium [NHUC, immortalised NHUC (TERT-NHUC), and uncultured healthy urothelial cells (UHUC)], showed that although the majority of differentially expressed (DE) genes between genders are located on the sex chromosomes, five autosomal genes are upregulated in female NHUC that are associated with invasive bladder tumours, inflammation, and hypoxia. Furthermore, each gender showed different transcriptomic perturbations in response to common mutations in a cohort of 102 stage Ta grade 2 tumours, including in tumours with mutations in the X-linked histone demethylase *KDM6A* where females had DE of chromatin regulatory genes but males did not.

ATAC-seq in two male and two female TERT-NHUC lines showed a genome-wide increase in chromatin accessibility in males, which could be seen by increased signal at individual loci and a greater number of overall peaks. Although this difference did not correlate with increased transcriptional activity, cell proliferation, or cell-cycle stage, it did correlate with a global increase of the activating histone marks H3K4me3 and H3K27ac, but not the heterochromatin marker H3K27me3. A reliable ChIP protocol with validated controls was developed for the histone marks H3K4me1, H3K4me3, H3K27ac, and H3K27me3, and provides a foundation for future epigenetic research in bladder cancer.

The results of this study suggest that, although variation between individual donors is greater than between gender groups, future research in bladder should consider genders separately. For instance, therapeutic efforts aimed at targeting chromatin architecture, such as HDAC inhibition, may be more effective in females, particularly when they have acquired mutations in *KDM6A*.

Table of Contents

List of Tables

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Abbreviations

Abbreviation	Name
°C	degrees Celsius
<	less than
=	equal to
>	greater than
~	approximately
3C	chromosome conformation capture
4C	circular chromosome conformation capture
5-aza-dC	5-aza-deoxycytidine
5C	carbon copy chromosome conformation capture
ACTBL2	beta-actin-like protein 2
ANOVA	analysis of variance
AR	androgen receptor
ATAC	assay for transposase accessible chromatin
BBN	butyl-(4-hydroxybutyl) nitrosamine
BCG	Bacillus Calmette-Guerin
bp	base-pair
BS-seq	bisulphite sequencing
BSA	bovine serum albumin
cDNA	complementary DNA
ChIP	chromatin Immunoprecipitation
chr	chromosome
cm	centimetre
CO_2	carbon dioxide
COMPASS	complex of proteins associated with Set1
Ct	cycle threshold
DE	differentially expressed/enriched
dH_2O	deionised water
DHT	5α-dihydrotestosterone
DNA	deoxyribonucleic acid
DNase	deoxyribonuclease
dNTP	deoxynucleotide triphosphates
dsDNA	double stranded DNA
DTT	dithiothreitol
E	efficiency
EDTA	ethylenediaminetetraacetic acid
EGFR	epidermal growth factor receptor
EGTA	egtazic acid,
ENCODE	encyclopaedia of DNA elements

ER	oestrogen receptor
ERE	oestrogen response element
EtOH	ethanol
EZH2	enhancer of zeste homolog 2
FAIRE	formaldehyde-assisted isolation of regulatory elements
FC	fold change
FDR	false discovery rate
FISH	fluorescence *in situ* hybridisation
g	g force / centrifugal force
GAPDH	glyceraldehyde 3-phosphate dehydrogenase
gDNA	genomic DNA
GSEA	gene set enrichment analysis
H_2O	water
H_2SO_4	sulphuric acid
H3	histone 3
H3K27ac	histone 3 lysine 27 acetylation
H3K27me3	histone 3 lysine 27 tri-methylation
H3K4me1	histone 3 lysine 4 mono-methylation
H3K4me3	histone 3 lysine 4 tri-methylation
HAT	histone acetyltransferase
HCl	hydrochloric acid
HDAC	histone deacetylase
HGP	human genome project
Hi-C	carbon copy conformation capture followed by NGS
HMT	histone methyltransferase
HRP	horseradish peroxidase
hrs	hours
IGEPAL	octylphenoxypolyethoxyethanol
IgG	immunoglobulin
IHC	immunohistochemistry
IP	immunoprecipitation
KCl	potassium chloride
KDM6A	lysine demethylase 6A
KMT2A/B/C/D	lysine methyltransferase 2A/B/C/D
LiCl	lithium chloride
LIMMA	linear models for microarray data
M	molar
mAB	monoclonal antibody
MBD	methyl-binding domain
MeDIP	methylated-DNA immunoprecipitation
$MgCl_2$	magnesium chloride

MIBC	muscle-invasive bladder cancer
min	minute(s)
MLL	mixed-lineage leukaemia
mM	millimolar
MNase	micrococcal nuclease
MRE	methylation-sensitive restriction enzyme
MRES	multiple regional epigenetic silencing
mRNA	messenger RNA
MSigDB	molecular signatures database
MTs	metallothioneins
MYT-1	myelin transcription factor 1
N	normality (for acid concentration)
NaCl	sodium chloride
NaDOC	sodium deoxycholate
$NaHCO_3$	sodium bicarbonate
NGS	next-generation sequencing
NHUC	normal human urothelial cells
nm	nanometre
NMIBC	non-muscle-invasive bladder cancer
OD	optical density
pAB	polyclonal antibody
PAGE	polyacrylamide gel electrophoresis
PAH	polycyclic aromatic hydrocarbon
PBS	phosphate buffered saline
PBS-T	PBS-Tween20 0.1%
PCA	principal component analysis
PCR	polymerase chain reaction
PD	population doublings
pH	potential hydrogen
PMSF	phenylmethylsulfonyl fluoride
PRC	polycomb repressor complex
qPCR	quantitative polymerase chain reaction
RIN	RNA integrity number
RIPA	radioimmunopreciptiation buffer
RISC	RNA-induced silencing complex
RMA	robust multi-chip average
Rn	maximum fluorescence
RNA	ribonucleic acid
RNAPII	RNA polymerase II
RNase	ribonuclease
RNasin	ribonuclease Inhibitor

ROC	receiver of operating characteristics
rRNA	ribosomal RNA
RT	room temperature
SDHA	succinate dehydrogenase complex, subunit A
SDS	sodium dodecyl sulphate
shRNA	short hairpin RNA
siRNA	short interfering RNA
snoRNA	small nucleolar RNA
SST	signal space transformation
TAD	topologically associated domain
TCA	Trichloroacetic acid
TD	tagmentation DNA buffer
TE	tris-EDTA
TERT	telomerase reverse transcriptase
TERT-NHUC	normal human urothelial cells immortalised using telomerase
TGS	Tris-glycine-SDS
Tn5	Transposase 5
TNM	tumour node metastasis
Tris	trisaminomethane
TSA	trichostatin A
TURBT	transurethral resection
U	enzyme activity unit(s)
UHUC	uncultured human urothelial cells
VCAN	versican
W	watt(s)
WB	western blot
WHO	World Health Organisation
XCI	X-chromosome inactivation
βME	2-mercaptoethanol
µg	microgram
µl	microlitre
µm	micrometre

Chapter 1
Introduction

1.1 The normal bladder

The urinary bladder is a hollow, smooth-muscular organ with two primary functions; the storage of soluble waste filtered by the kidneys that is transported to the bladder through the bilateral ureters, and the expulsion of this waste through the urethra by the process of micturition (Fry, 2005; Lukacz *et al.*, 2011). The bladder wall is organised into four layers; the mucosa, submucosa, muscularis, and serosa and adventitial layers (Figure 1.1) (Merrill *et al.*, 2016). The inner mucosal layer consists of transitional epithelial cells (termed the urothelium) which function as an impermeable bladder to urine. Beneath is the submucosa (or lamina propria) composed of loose connective tissue containing fibroblasts, adipocytes, interstitial cells of Cajal, vasculature, and nerve terminals, that is proposed to serve as integratory tissue of the urothelium and smooth muscle layers (Merrill *et al.*, 2016). The muscle layer is organised into three smooth-muscle components that are together termed the detrusor, and this is responsible for the contraction of the bladder and subsequent expulsion of urine. The composition of the detrusor differs in the urethra where mechanical contraction for micturition is not required, and is established as an internal urethral sphincter in males but not in females (Fry, 2005; Merrill *et al.*, 2016). The outer serosa layer surrounds the outer bladder wall, and is surrounded by loose connective tissue called adventitia. A healthy bladder stores urine without discomfort with intermittent periods of filling, has a full capacity of 300-400ml in adults, and empties with a strong continuous flow without pain and the absence of blood (Lukacz *et al.*, 2011).

1.2 Bladder cancer

1.2.1 Epidemiology

Bladder cancer is the 7th most common cancer in UK, and is up to 4 times more common in men, therefore making it the 4th and 13th most common cancer type for men and women respectively (NICE, 2017). Although men are more likely to develop bladder cancer, women often present with more advanced disease and have a less favourable outcome. Each year more than 429,000 cases of bladder cancer are diagnosed globally and account for over 165,000 deaths (Ferlay *et al.*, 2015). Due to difficulties of treatment, prevalence, duration and

recurrence of the disease, bladder cancer is one of the most expensive cancers to treat per incidence of cancer, and was estimated to have cost the EU nearly €5 billion in 2012 (Leal *et al.*, 2015). There has been little improvement in the treatment of bladder cancer over the last few decades, with 5-year survival rates of 33% for locally advanced and 3% for metastatic bladder cancers remaining constant for 25 years (Siegel *et al.*, 2017). Incidence of bladder cancer increases with age, with the majority of new cases diagnosed in patients over the age of 60 (Sanli *et al.*, 2017). The predominant symptom of bladder cancer is haematuria, although increased frequency, urgency, and irritation of urination can also occur.

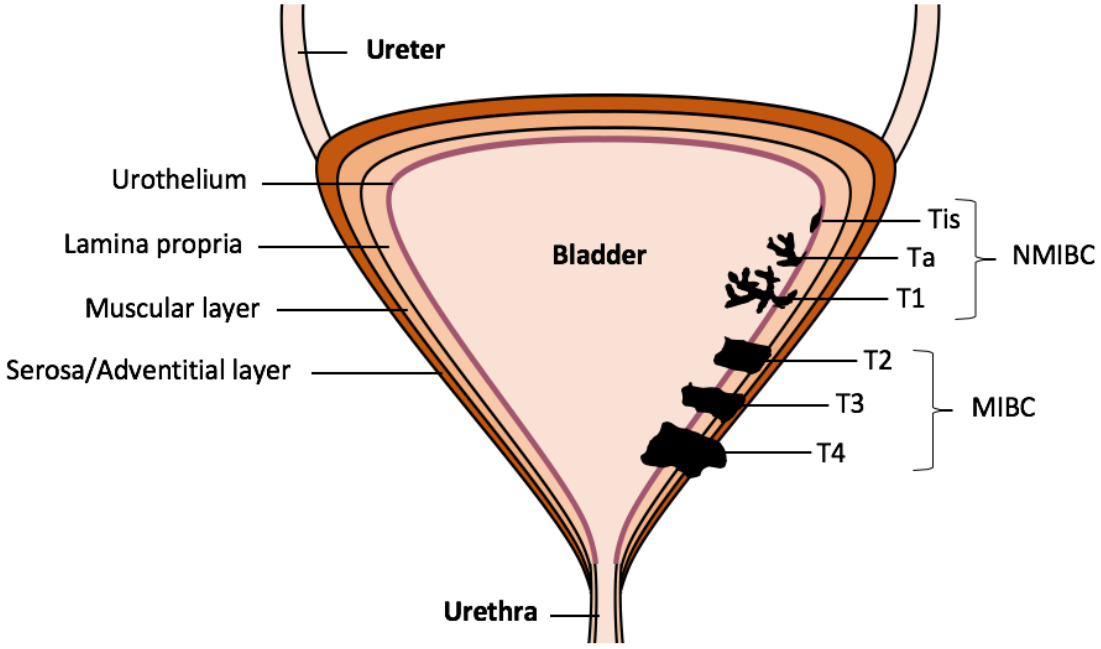


Figure 1.1 Structure of the bladder and TNM classification of bladder cancers

Urine is transported to the bladder from the kidneys through two ureters, and expelled from the bladder through the process of micturition through the urethra. The layers of the bladder from inside to outside include the urothelium, the lamina propria, the muscular layer, and the serosa/adventitial layer. Bladder cancers are categorised by their invasive stage according to the tumour-node-metastasis (TNM) system and are considered to be non muscle-invasive (NMIBC; stages Tis, Ta, and T1), and muscle-invasive (MIBC; T2, T3, and T4).

1.2.2 Aetiology

Smoking is the most notable risk factor for bladder cancer and is attributed to around 50% of cases (Cumberbatch *et al.*, 2016). Both disease incidence and disease-related mortality is greatest for current smokers, is decreased for ex- and second-hand smokers, and is lowest for people who have never smoked. Smoking duration and intensity is also strongly correlated with increased risk of bladder cancer (Cumberbatch *et al.*, 2016; Cumberbatch *et al.*, 2018). Tobacco contains high concentrations of known carcinogens such as polycyclic aromatic hydrocarbons (PAH), aromatic amines, and N-nitroso compounds, whose metabolites are excreted in the urine and cause DNA damage through single- and double-strand breaks, mutations of individual bases, and the formation of DNA adducts (Stern *et al.*, 2009). Furthermore, smoking induces changes to the DNA damage response machinery which propagates the impairment of the host's response to carcinogens. The next largest risk factor for bladder cancer is occupation, where risk is highest for workers exposed to aromatic amines (tobacco, dye, and rubber workers; hairdressers; printers; and leather workers) or PAH (nurses; waiters; aluminium workers, oil/petroleum workers), and lowest for workers in the agricultural sector (Cumberbatch *et al.*, 2015). Due to increased safety legislation and improvements in workplace hygiene, occupational risk of bladder cancer has dropped significantly over time and is believed to contribute to only 5-7% of cases (Cumberbatch *et al.*, 2015). Other risk factors include low-fruit diets, alcohol consumption, consumption of arsenic-polluted water, and air pollution (Sanli *et al.*, 2017). Genetic predisposition to bladder cancer has also been reported, where polymorphisms of genes involved in carcinogen detoxification, such as *N*-acetyl-transferase 2 (NAT2) and glutathione *S*-transferase μ1 (GSTM1), may increase the carcinogenic effects of smoking that promote bladder cancer (Guey *et al.*, 2010; An *et al.*, 2015).

1.2.3 Pathology

Around 90% of cancers arising from the urothelium are transitional cell carcinoma. Less common variants include squamous cell carcinoma, small-cell carcinoma, and adenocarcinoma which are high grade and associated with aggressive and metastatic forms of the disease. Bladder cancer has traditionally been divided into two groups, Non Muscle-Invasive Bladder Cancer (NMIBC) and Muscle-Invasive Bladder Cancer (MIBC), that are categorised by stage and grade according to the Tumour-Node-Metastasis (TNM) system and World Health Organisation (WHO) classification system respectively (Figure 1.1). The TNM system describes the extent of tumour invasion into surrounding tissue and is graded

from Tis (carcinoma *in situ)* to T4 (Brierley *et al.*, 2017). The WHO grading system describes tumour differentiation, where low-grade tumours are highly differentiated (G1 and G2) and bear resemblance to the normal urothelium, and high-grade tumours are poorly differentiated and highly dissimilar to urothelium. Although the WHO grading system has been reclassified in recent years, this project will refer to the older system which classifies tumour grade as G1, G2, and G3 (Moch *et al.*, 2016). NMIBC includes primarily low-grade tumours that occupy only the mucosal layer (urothelium and/or lamina propria; Tis-T1) and these account for 75% of bladder cancer cases. NMIBC are recurrent, infrequently progress to invasive forms, and have a five-year survival rate of over 90%. MIBC includes high-grade tumours that invade into and past the detrusor muscle of the bladder (T1-T4), and are more aggressive with a five-year survival rate of less than 50%.

1.2.4 Treatment

Treatment of bladder cancer depends on tumour stage and grade. A physical examination of the patient followed by cystoscopy/urethroscopy is carried out to assess tumour size, site, number, and severity. For low-grade NMIBC, treatment involves transurethral resection of the bladder tumour (TURBT) followed by chemotherapy by intravesical instillation of mitomycin C. For higher-grade, large or multifocal NMIBC this is further followed by immunotherapy with Bacillus Calmette-Guérin (BCG) to reduce the risk of recurrence and progression (Woldu *et al.*, 2017). These higher grade NMIBC patients (stage T1 and/or grade 3) are particularly difficult to manage due to limitations in predicting whether they may progress to more aggressive disease states, and they have reduced recurrence-free survival and an increased mortality rate (Dalbagni *et al.*, 2009). Following failure of BCG treatment in these patients, further treatment decisions only include additional TURBT and BCG therapy or complete removal of the bladder (cystectomy). Radical cystectomy in many cases is an over-treatment for the disease and comes with the associated risks of major surgery in already vulnerable patients (Denzinger *et al.*, 2008; Woldu *et al.*, 2017). Treatment of MIBC is limited to radical cystectomy combined with cisplatin-based neoadjuvant chemotherapy, and has been shown to provide long-term disease-free survival of 70% for tumours confined to the bladder (Stein *et al.*, 2001). The cisplatin-based therapies used in the treatment of MIBC include MVAC (methotrexate, vinblastine, doxorubicin, and cisplatin) or GC (gemcitabine and cisplatin), which are highly toxic and may promote complications such as reduced immunity and increased infection, ulceration of the digestive tract, and nausea, and can have a toxic death rate of up to 4% (Sr *et al.*, 1992; Study *et al.*, 2000). Treatment of invasive and metastatic bladder cancer had not progressed over the last 20 years. However, recent assessment of immunotherapies has seen FDA

approval of five inhibitors of PD-1 and PD-L1 receptors that are involved in suppression of T-cell activation (Dietrich and Srinivas, 2018). Two of these, atezolizumab and pembrolizumab, are now considered for second-line treatment in patients with metastatic MIBC who have contraindication to cisplatin, and are currently being investigated for first-line treatment (Sanli *et al.*, 2017; Dietrich and Srinivas, 2018).

1.3 Molecular biology of bladder cancer

1.3.1 Chromosomal aberrations

NMIBC and MIBC have distinct molecular profiles. At the level of chromosomal alterations, low-stage and low-grade tumours are generally stable and predominantly have a near-diploid karyotype, whereas high-stage and high-grade tumours are very unstable and are typically aneuploid with many copy number alterations and chromosomal rearrangements (Hurst *et al.*, 2012; Weinstein *et al.*, 2014; Robertson *et al.*, 2017). Stage T1 tumours have varying levels of complexity and can resemble either invasive or non-invasive tumours (Hurst *et al.*, 2012). It is currently unknown what drives the increased chromosome instability in MIBC. However, mutations in the minichromosomal maintenance complex, inactivating mutations of DNA damage and repair genes, and mutations in components of the Cohesin complex have all been described (Knowles and Hurst, 2015).

Despite the aforementioned genomic differences in invasive and non-invasive tumours, commonalities have been found throughout bladder cancer. For instance, loss of chromosome 9 is seen in ~50% of both NMIBC and MIBC (Sanli *et al.*, 2017), and is a predictor of reduced recurrence-free interval. Candidate tumour suppressors implicated on chromosome 9 include cyclin-dependent kinase inhibitor 2A (*CDKN2A*; which encodes $p16^{INK4a}$ and $p14^{ARF}$) and *CDKN2B* (which encodes p15), patched 1 (*PTCH1*), deleted in bladder cancer 1 (*DBC1*), Notch homologue 1 (*NOTCH1*), and tuberous sclerosis 1 (*TSC1*) (Knowles and Hurst, 2015). Loss of *CDKN2A* is particularly critical given that p16 and $p14^{ARF}$ negatively regulate the Rb and p53 tumour suppressor pathways respectively. These are essential in the regulation of cell growth and proliferation and are ubiquitously perturbed in the tumorigenesis of many cancers (Sherr and McCormick, 2002; Sanli *et al.*, 2017). Another critical tumour suppressor encoded on chr9 is TSC1, which forms a complex with TSC2 to negatively regulate the mTOR branch of the PI3K signalling pathway which regulates cell survival and growth and is commonly mis-regulated in cancer, including bladder (Fresno Vara *et al.*, 2004; Knowles and Hurst, 2015)

1.3.2 Common mutations

Invasive and non-invasive bladder tumours are also distinct at the level of somatic mutations (Table 1.1). MIBC has one of the highest mutations rates of any cancer with ~8.2 mutations per megabase (Mb), and is generally characterised by a diverse mutational spectrum but which often includes inactivating mutations of tumour protein p53 (*TP53*), and retinoblastoma 1 (*RB1*) and deletions of *CDKN2A* which result in loss of function of these (LOF) genes. Alterations in at least one of these genes occur in ~89% of MIBC (Weinstein *et al.*, 2014; Robertson *et al.*, 2017). In comparison, NMIBC have only ~1.8 mutations per Mb and are characterised by frequent gain of function (GOF) mutations in fibroblast growth receptor 3 (*FGFR3*), GOF-mutations in phosphatidylinositol-4,5-bisphosphate 3-kinase (*PIK3CA;*), and GOF-mutations in stromal antigen 2 (*STAG2*) and a higher frequency of mutations of lysine demethylase 6A (*KDM6A*), RAS genes, Ras homologue gene family member B (*RHOB*) and *UNC80* (Hurst *et al.*, 2017; Pietzak *et al.*, 2017). Nevertheless, commonalities also exist in the mutational landscape of bladder tumours, including a high rate of mutations of chromatin modifiers and mutations that activate telomerase reverse transcriptase (*TERT).*

The most common mutational events in the progression of bladder cancer are GOF point mutations in the promoter region of telomerase reverse transcriptase (*TERT)*, which are present in ~80% of bladder tumours regardless of stage of grade (Allory *et al.*, 2014; Hurst *et al.*, 2014). Such mutations create consensus binding motifs that allow binding of ETS and TCF transcription factors, which then promote the activation of the *TERT* promoter (Lamb *et al.*, 2013). TERT increases telomere length at the end of chromosomes and is essential for proliferating cells. Expression is therefore downregulated in differentiated cells, and upregulated in undifferentiated cells including cancers. The high mutation rate across all bladder cancer types suggests that *TERT* promoter mutations occur early in the onset of the disease (Knowles and Hurst, 2015). An *in vitro* model commonly used for studying the healthy urothelium includes normal human urothelial cells (NHUC) that are immortalised by retroviral transduction of hTERT (TERT-NHUC) (Chapman *et al.*, 2006; Chapman *et al.*, 2008). This method of immortalisation in NHUC indefinitely increases the number of population doublings (PD) *in vitro* (otherwise limited to ~20PD), without inactivating the p16/Rb pathway or producing chromosomal alterations (Chapman *et al.*, 2006). NHUC are typically immortalised after 3-4 passages in culture, with TERT-NHUC being used up to an additional 10-15 passages following the immortalisation. A comparison of expression profiles between matched pairs of hTERT-immortalised and non-immortalised NHUC found differential expression of 103 genes, 20% of which were known

Polycomb-group targets and involved in differentiation and tumorigenesis (Chapman *et al.*, 2008).

FGFR3 GOF mutations are found in up to 80% of stage Ta tumours, ~40% of T1 and only 15% of MIBC (Di Martino *et al.*, 2012). Mutations of *FGFR3,* most commonly S249C point mutations, constitutively activate the receptor and in MIBC, where mutations are less frequent, upregulation of FGFR3 protein expression is often seen (Di Martino *et al.*, 2012). In TERT-NHUC, expression of mutant FGFR3 leads to activation of the RAS-MAPK (mitogen-activated protein kinase) pathway and phospholipase Cγ (PLCγ), resulting in increased cell survival and proliferation, and suggests *FGFR3* mutations contribute to early urothelial hyperplasia *in vivo* (Di Martino *et al.*, 2009).

Activating mutations of *PIK3CA* are found in ~50% of stage Ta bladder tumours, compared to ~20% in tumour stages ≥T1, and these often co-occur with mutations in *FGFR3* (Knowles *et al.*, 2009). Like the aforementioned loss of *TSC1* on chromosome 9, these mutations promote activation of the PI3K pathway to promote cell growth, proliferation, and survival through receptor tyrosine signalling. Indeed, PI3K pathway activation is not limited to activating mutations of *PIK3CA* and loss of *TSC1*, and mutations and mis-regulation of genes such as AKT serine/threonine kinase, receptor tyrosine kinases ErbB-1 and -2 (*ERBB1/2),* epidermal growth factor receptor (*EGFR*), and phosphatase and tensin homologue (*PTEN*) are seen in bladder cancers, more commonly associated with invasive tumours (Knowles *et al.*, 2009; Knowles and Hurst, 2015; Sanli *et al.*, 2017).

Activating mutations of the RAS family genes *HRAS* and *KRAS* are found in >10% of bladder cancers, although are more commonly associated with low stage tumours (Hurst *et al.*, 2017). However, mutations in RAS and *FGFR3* are mutually exclusive, which implicates them in conferring the same phenotype in bladder cancer, namely, the activation of the RAS-MAPK signalling pathway (Jebar *et al.*, 2005). The RAS-MAPK signalling pathway regulates cell proliferation and differentiation, and although it has long been implicated in the tumorigenesis of many cancers (Wei and Liu, 2002), it is not clear what role MAPK signalling plays in the development of bladder cancer (Knowles and Hurst, 2015).

Table 1.1 Most commonly mutated genes (within coding regions) in bladder cancer

	Hurst *et al.*, 2017	Pietzak *et al.*, 2017	Pietzak *et al.*, 2017	Guo *et al.*, 2013	TCGA 2017
	Ta (%)	Ta (%)	T1 (%)	T1 (%)	MIBC
FGFR3	79	66	30	25	14
PIK3CA	54	36	22	6	22
KDM6A	52	50	43	50	26
STAG2	37	24	22	25	14
KMT2D	30	31	26	10	28
ARID1A	18	25	27	6	25
EP300	18	20	8	16	15
CREBBP	15	23	19	12	12
KMT2C	15	16	5	3	18
RHOB	13	ND	ND	0	11
HRAS	12	2	8	16	9
KMT2A	11	9	11	9	11
TSC1	11	5	22	12	8
BRCA2	10	11	11	0	7
COL11A1	10	ND	ND	0	5
RBM10	10	22	5	0	9
TP53	4	11	35	25	48
FAT1	-2	13	17	0	12
KRAS	2	11	8	6	4
ATM	-1	13	19	3	14
CDKN1A	-1	11	13	0	9
ELF3	-1	ND	ND	12	12
ERCC2	-1	21	13	6	9
ERBB2	0	11	19	3	12
ERBB3	0	9	19	3	10
RB1	0	0	5	9	17

Bladder cancer has one the highest rates of mutations in chromatin-modifier genes compared to any other cancer type, the most common of which include histone demethylase *KDM6A,* histone methyltransferases *KMT2A, KMT2C,* and *KMT2D*, histone acetyltransferases *CREBBP* and *EP300*, components of the SWI/SNF complex (*ARID1A, ARID4A*), Polycomb-group genes *ASXL1* and *AXL2*, the nuclear receptor co-repressor 1 (*NCOR1*), chromodomain helicase DNA-binding proteins *CHD6* and *CHD7*, and a component of the Cohesin complex *STAG2* (Gui *et al.*, 2011a; Guo *et al.*, 2013; Weinstein *et al.*, 2014; Robertson *et al.*, 2017; Hurst *et al.*, 2017). Although these chromatin modifiers are predominantly involved in the activation of genes through epigenetic mechanisms and their inactivation in cancer implicates them as tumour suppressor genes, the mechanism through which they promote tumorigenesis in bladder cancer is poorly studied, and their mutations are not charactised well enough to determine loss or gain of function.

The most commonly mutated chromatin modifier in bladder cancer is KDM6A, an X-linked demethylase of histone H3 lysine 27 di/tri methylation (H3K27me3) and component of the KMT2C/D-COMPASS-like complex, which is mutated in up to 65% of NMIBC and 26% of MIBC. *KDM6A* also escapes X-chromosome inactivation (Greenfield *et al.,* 1998), and exhibits increased expression in female tissues compared to male tissues (Xu *et al.,* 2008; Guo *et al.,* 2013).

KDM6A consists of two domains; a tetratricopeptide repeat (TPR) domain predicted to mediate protein-protein interactions, and a catalytic domain with a zinc-binding domain, linker region, helical region, and a Jumonji C (JmjC) domain (Sengoku and Yokoyama, 2011). However, *KDM6A* mutations in bladder cancer do not show any domain-specific bias and are found throughout the gene. The JmjC domain is characteristic of the α-ketoglutarate-dependent dioxygenase class of HDMs that include KDM6A, KDM6B (also known as JMJD3), and KDM6C (also known as UTY), and is essential for the catalytic activity of these HDMs. Both KDM6A and KDM6B show H3K27 demethylating properties *in vitro* and *in vivo*, and regulate a distinct set of genes (Agger *et al.,* 2007; Jiang *et al.,* 2013). However, KDM6C, which is located on the Y-chromosome and shares 88% sequence homology with KDM6A (98% sequence homology in the JmjC domain), has not been shown to elicit histone demethylase activity *in vivo*. Mass spectrometry studies have shown that KDM6C does have demethylating properties *in vitro* albeit at lower levels than KDM6A, most likely due to a proline residue at position 1214 (JmjC domain) instead of isoleucine as found in KDM6A (Walport *et al.,* 2014).

This intriguing structural homology but apparent divergence in functionality between KDM6A and KDM6C suggests demethylase-independent and gender-specific functions of

KDM6A. Homozygous deletions of *Kdm6a* are lethal during mid-gestation of female mouse embryos (Shpargel *et al.,* 2012; Wernig *et al.,* 2008). However, a subset of their male counterparts did survive to term although with stunted growth and a reduced life span. This rescue is possibly due to KDM6C that is able to compensate for KDM6A demethylase-independent gender-related functions (Shpargel *et al.,* 2012; Wernig *et al.,* 2008). Such studies demonstrate gender-specific developmental roles of KDM6A that are both demethylase-dependent and demethylase-independent. This characteristic may be essential for gender-specific differentiation cues during development, as evidenced by the female bias of *Rhox6* and *Rhox9* regulation by Kdm6a in mouse embryonic stem cells (Berletch *et al.,* 2013). Kdm6a has been further implicated during development including posterior-development, endoderm differentiation, osteogenesis, and adipogenesis through the regulation of genes such as the *Hox*-gene cluster, *Wnt3*, and *Runx2, Osteocalcin*, and *PPARγ2* (Lan *et al.,* 2007; Agger *et al.,* 2007; Jiang *et al.,* 2013; Hemming *et al.,* 2014).

These results suggest a sexual dimorphism for perturbations of KDM6A in bladder cancer, and one may expect that males would see a higher mutation rate in *KDM6A* as females would require mutations in both alleles to have the same tumorigenic effect (van Haaften *et al.*, 2009). However, recent studies have shown that *KDM6A* mutations are more common in low-grade female NMIBC (74% incidence in females compared to 42% in males) and although mutations in *KDM6C* and loss of chrY are also seen in males, it does not compensate for the high incidence rate of *KDM6A* mutations in females. This gender-associated mutational bias of *KDM6A* is only seen in NMIBC and not in MIBC (Hurst *et al.*, 2017; Robertson *et al.*, 2017), and contradicts findings in other cancers such as T-cell acute lymphoblastic leukaemia and medulloblastomas which show a bias for mutations in *KDM6A* in males (Pugh *et al.*, 2012; Robinson *et al.*, 2012; Meulen *et al.*, 2015). Interestingly, high rates of mutation in *KDM6A* are often accompanied by co-occurring mutations in *KMT2C* and *KMT2D* (found in up to 15% and 30% of NMIBC respectively), which implicates perturbations in COMPASS-like complexes and subsequent mis-regulation of enhancers in bladder cancer development (Hurst *et al.*, 2017). Further consideration of epigenetics and the COMPASS-like complexes will be discussed later in this chapter.

1.4 Gender and bladder cancer

1.4.1 Environmental risk

The incidence of bladder cancer is three to four times more common in males than in females (NICE, 2017). Despite this, women predominantly present with more advanced

tumours at the time of diagnosis (Emil *et al.*, 2009; Fajkovic *et al.*, 2011) and although this is attributed to a delayed response to diagnosis of bladder cancer in women, women also have a worse outcome across all stages of disease at presentation (Mungan *et al.*, 2000; Dobruch *et al.*, 2016). The disparity between genders of incidence and risk of mortality still exists when accounting for environmental risk factors (Dobruch *et al.*, 2016). For instance, one study found that a cohort of 2,806 individuals with bladder cancer showed a male to female incidence ratio of 3.9:1, and in the absence of exposure to known risk factors such as smoking, occupation, and urinary tract infections, the male to female incidence ratio dropped to only 2.7:1 (Hartge *et al.*, 1990). An analysis of the prostate, lung, colorectal and ovarian (PLCO) cancer screening trial and the national lung cancer screening trial (NLST) also found that bladder cancer was 3-4.5 times more common in males than females, and this incidence ratio persisted when comparing gender groups with similar exposure to tobacco (Krabbe *et al.*, 2015). Another study found that in 21 distinct geographical locations, smoking rates only partially explained the male incidence bias of bladder cancer, and that risk of bladder cancer was equal in genders exposed to comparable amounts of tobacco (Hemelt *et al.*, 2009). Risk factors regarding occupation and exposure to hair dyes have also been described for bladder cancer, although these are attributed to a minority of cases and gender differences are not implicated (Dobruch *et al.*, 2016).

1.4.2 Microbiome and Metabolism

As the gender biases seen in bladder cancer cannot be explained by environmental risk factors, efforts are being made to decipher biological mechanisms that may be promoting these biases. For instance, studies have shown differences in the composition of bacteria that make up the microbiome of male and female bladders, with preponderance of *Corynebacterium* species in males and Lactobacillales in females, and it is postulated that these differences promote disparate local environments that may promote or inhibit bladder tumorigenesis (Xu *et al.*, 2014). Differences in the urinary microbiome have also been found between patients with and without bladder cancer, and a protective effect in preventing bladder cancer recurrence has been shown for an oral preparation of *Lactobacillus* (Xu *et al.*, 2014; Dobruch *et al.*, 2016). These differences in the bladder microbiome are not attributed to the increased rate of urinary tract infection (UTI) in females, which is instead driven by *E.coli* infections. The increased UTI incidence rate in women is primarily attributed to differences in human anatomy (Foxman, 2010), although a recent study has shown that testosterone dulls the effects of IL-17 to promote longer lasting and chronic UTI in males (Scharff *et al.*, 2019). The relationship between UTI and bladder cancer is not entirely clear, but it is unlikely that the increased rate of UTI in females promotes the intrinsic gender differences of bladder

cancer (Bayne *et al.*, 2018). However, the increased incidence of UTI in women does promote a time-delay for the diagnosis of bladder cancer in women as health professionals often misdiagnose UTI as the cause of haematuria, and may be a significant contributor to the finding that women present with more advanced bladder cancers (Dobruch *et al.*, 2016).

Gender-related metabolic differences have also been attributed to promoting the gender biases in bladder cancer (Dobruch *et al.*, 2016). Glutathione-*S*-transferase M1 (GSTM1) detoxification of tobacco-derived aromatic hydrocarbons, has been shown to influence bladder cancer risk by reducing bladder exposure to carcinogens (Karagas *et al.*, 2005). However, *GTM1*-null genotypes have been found, but these only increase risk of bladder cancer in women that smoke, and not in non-smokers or men (Karagas *et al.*, 2005). Gender-related differences in the metabolic profile of the liver have also been considered as a driver of gender biases in bladder cancer, as carcinogens of the bladder are primarily metabolised in the liver (Dobruch *et al.*, 2016). For example, differential isoform expression of uridine 5'-diphosphoglucuronosyltransferase (UGT) has been described in male and female liver, and is believed to result in differential break-down of aromatic amines derived from tobacco smoke, and therefore promotes differential exposure to DNA damage in the bladder (Zhang, 2013).

1.4.3 Androgen receptor (AR)

A common biological consideration when questioning gender-related differences in biology is the role of sex hormones, in particular androgen and oestrogen receptors (AR and ER respectively).

AR is a ligand-dependent transcription factor responsible for initiating male sexual development and differentiation and maintaining sexual activity and reproductive function. Through binding to androgen ligands such as testosterone and its more potent metabolite 5α-dihydrotestosterone (DHT), AR is translocated to the nucleus and binds to androgen response elements within the promoter regions of target genes to facilitate transcriptional processes often involved in cell growth and survival (Tan *et al.*, 2015a). This is particularly evident in the prostate, where androgens are essential in maintaining the balance of proliferation and apoptosis of prostate cells, perturbation of which leads to the 2nd most common cause of cancer death in men: prostate cancer (Tan *et al.*, 2015a; Siegel *et al.*, 2017). The initiation of over 50% of prostate cancers can be attributed to the AR-dependent upregulation of E-twenty-six (ETS) transcription factors and transmembrane serine protease 2 (TMPRSS2) which leads to cell cycle progression and cell proliferation (Tomlins *et al.*, 2005). As a result, prostate cancer is commonly treated by androgen suppression either by surgical castration (orchiectomy) and/or chemical castration using gonadotropin releasing

hormones such as leuprolide or goserelin (Tan *et al.*, 2015a). However such treatment shows transient success, with patients often relapsing having developed an antiandrogen- or castration-resistant form of the disease.

There have been many studies of the role of AR in bladder cancer, with conflicting results (Li *et al.*, 2017). A plethora of studies using immunohistochemical (IHC) techniques to assess a possible prognostic value of AR in bladder cancer have had conflicting results, with some studies suggesting a correlation between positive AR expression and a lower risk of recurrence (Kil Nam *et al.*, 2014), others suggesting that AR is related to tumour progression (Miyamoto *et al.*, 2012; Mashhadi *et al.*, 2014), and many studies failing to show any prognostic significance of AR expression in bladder cancer (Tan *et al.*, 2015a; Godoy *et al.*, 2016; Li *et al.*, 2017). The largest and most rigorous of these studies carried out IHC on 472 patient samples, using multiple antibodies for AR and both automated and manual scoring systems, and cross-checked results between two institutions in a blinded fashion (Mir *et al.*, 2011). Although the authors acknowledge that the majority of their tumours were invasive, they found very little correlation between AR expression and tumour stage, grade, invasiveness, patient mortality or gender.

Despite a lack of obvious clinical implications for AR in bladder cancer being demonstrated so far, multiple *in vitro* and *in vivo* functional studies for AR in the bladder maintain the attraction of AR as a therapeutic target for bladder cancer. For instance, siRNA knockdown of AR decreased cell proliferation whilst increasing apoptosis and reducing cell migration in T24♀ and 253J♂ bladder cancer cell lines (Wu *et al.*, 2010). The expression of metastatic growth-related gene matrix metallopeptidase 9 (*MMP-9*), B-cell lymphoma-extra large (*Bcl-x*$_L$), and cyclin D1 (*CCND1*) was also assessed by qPCR and showed that these were decreased upon AR knockdown. Further *in vitro* studies have implicated AR in regulating the epidermal growth factor receptor (EGFR)/ERBB2 pathway and β-catenin/TCF/LEF1/Wnt signalling in TCC-SUP♀, J82♂, 5637♂, and UM-UC-3♂ bladder cancer cell lines, and that targeting AR with anti-androgen treatment or knockdown could supress cell proliferation, implicating possible therapeutic targeting of AR (Miyamoto *et al.*, 2007; Zheng *et al.*, 2011; Li *et al.*, 2013). Early studies in mice have shown that treating with Butyl-(4-hydroxybutyl) nitrosamine (BBN) promotes bladder cancer in mice, and male mice develop tumours significantly earlier than their female counterparts (Bertram and Craig, 1972). However, BBN-induced bladder tumour induction time was equal if male mice were castrated, or female mice were treated with testosterone (Bertram and Craig, 1972). Further experiments showed that 92% of male mice and 42% of female mice treated with BBN developed bladder cancer, but AR knockout (AR^{KO}) mice did not develop any tumours

(Miyamoto *et al.*, 2007). The authors then went on to show that 25% of BBN-treated AR^{KO} mice supplemented with DHT developed bladder cancer regardless of gender, and that 50% of castrated wild-type male mice also developed tumours when supplemented with DHT (Miyamoto *et al.*, 2007). Studies on prostate cancer patients who also went on to develop bladder cancer, found that patients given androgen deprivation therapy (ADT) had reduced risk of bladder tumour recurrence (12.5-22% recurrence rate) compared to patients not given ADT (30-50% recurrence rate) (Izumi *et al.*, 2014; Masaki *et al.*, 2017).

These *in vitro* and *in vivo* studies implicate AR in bladder tumour development and suggest potential therapeutic applications. However, the controversial results found in human bladder cancer patients and the lack of correlation with AR function and gender in non-rodent studies mean the hypothesis of AR as the predominant driver of male bladder cancer gender is currently not supported.

1.4.4 Oestrogen receptor (ER)

Oestrogens are predominantly involved in the development and maintenance of key sexual and reproductive characteristics in women, but they are also essential for the biological effects of cardiovascular, musculoskeletal, immune and central nervous systems of both men and women (Heldring *et al.*, 2007). The most potent oestrogen is 17β-estradiol (E2), and its metabolites estrone and estriol are also known to bind to ER. Oestrogens predominantly exert their effects through binding with oestrogen receptors (ERs), which then act as transcription factors by binding to target gene promoters at oestrogen response elements (EREs), or through interacting with other transcription factor complexes such as Fos/Jun and SP-1 to target genes lacking EREs. There are two types of ER, ER-α and ER-β, which are expressed from separate genes on different chromosomes (6q25.1 and 14q23.2, respectively). They consist of three domains, an N-Terminal Domain (NTD, 16% homology), a DNA-binding domain (DBD, 97% homology), and a COOH-terminal ligand-binding domain (LBD, 59% homology) (Heldring *et al.*, 2007; Jia *et al.*, 2015).The two ERs therefore share a high degree of homology and they have similar affinities for E2 and bind the same EREs (although differential binding is also observed). However, the low homology in the NTD domain allows for differential binding to coregulators that results in distinct functionalities of both ERs. Indeed ER-α and ER-β are known to counter each other's effects, suggesting that a biological response to E2 is dependent on the balance of ER-α and ER-β signalling (Liu *et al.*, 2002).

Similarly to AR, ER-α and ER-β have been studied extensively in bladder cancer. Again, there is no general consensus for the role of ER-α in the tumorigenesis of bladder cancer, with many conflicting studies showing that ER-α may or may not be associated with

bladder cancer stage, progression, and recurrence (Shen *et al.*, 2006; Kauffman *et al.*, 2011; Tan *et al.*, 2015b; Godoy *et al.*, 2016). However, unlike ER-α, ER-β has more robustly been associated with higher stages and grades of bladder cancer as well as demonstrating a correlation with disease outcomes (Croft *et al.*, 2005; Han *et al.*, 2012; Miyamoto *et al.*, 2012; Kauffman *et al.*, 2013), leading to the general consensus that ER-β is the primary ER present in bladder cancer (Godoy *et al.*, 2016). A case-study on 224 bladder cancer samples by Shen *et al* showed that only 0.89% of samples were ER-α positive by IHC staining, whereas 63% of samples expressed ER-β (Shen *et al.*, 2006). When grouped into tumour stages ER-β expression was seen in 53%, 55%, 80%, 81%, and 75% of Ta, T1, T2, T3, and T4 tumours respectively, therefore correlated with tumour stage and higher expression in MIBC compared to NMIBC (Shen *et al.*, 2006). A more recent study further confirmed the increased expression of ER-β in bladder cancer as well as a lack of association of both ER-α and AR in 410 patients treated with radical cystectomy for urothelial cell carcinoma, in which over 90% of samples expressed ER-β, but only 2.9% and 0% expressed ER-α and AR respectively (Tan *et al.*, 2015b).

In vitro studies have shown that both ER-α and ER-β are involved in multiple cellular processes in urothelial cells. A study which treated bladder cell lines with E2 and ER-specific agonists (pyrazoletriol for ER-α, and diarylpropionitrile for ER-β), found that E2 upregulated cyclin D1 and cyclin-E to promote cell-cycle progression and cell proliferation, but that this was due to ER-β in primary and immortalised urothelial cells (immortalised using HPV E6, E7, or SV40 large T antigen) whilst ER-α was implicated in bladder cancer cell lines (5637♂ and T24♀) (Teng *et al.*, 2008). Another study showed that three out of four ER-β-positive bladder cancer cell lines (5637♂, RT4♂, and T24♀, but not TCCSUP♀) showed reduced proliferation when treated with the anti-oestrogens raloxifene, 4-hydroxytamoxifen, or the pure anti-estrogen ICI-182,780 (Hoffman *et al.*, 2013). The authors then went on to show that RT4♂ cells treated with raloxifene had activated caspase-4 and poly-ADP ribose polymerase (PARP), which increased apoptosis, and increased expression of B-cell translocation gene (*BTG2*) and decreased cyclin D1 transcription which inhibited cell proliferation (Hoffman *et al.*, 2013).

As with AR, these studies implicate ER in bladder tumorigenesis and potential therapeutic applications. Although there are conflicting results on the potential role of ER-α in bladder cancer, ER-β is commonly upregulated in bladder tumours and has a high correlation with stage and grade. This is particularly interesting given that women more often present with more aggressive bladder tumours, although it is noted that the aforementioned studies do not report a gender disparity in the proportion of tumours that are ER-β+, and

mechanistic insights into ER regulation also do not show sexual dimorphism. Whether ER-β is the primary driver of gender differences in bladder cancer requires further research.

1.4.5 FOXA1

The binding of AR and ER-α to their respective targets has largely been attributed to regulation by FOXA1 (Carroll *et al.*, 2005; Sahu *et al.*, 2011; Hurtado *et al.*, 2011). Forkhead box (FOX) proteins are a family of transcription factors involved in the regulation of genes promoting cell proliferation, growth, differentiation and longevity, especially in development. FOXA proteins can act as pioneering factors that engage condensed chromatin and allow binding of other transcription factors, but are also able to promote chromatin compaction through the recruitment of corepressor complexes in a largely context-dependent manner (Sekiya and Zaret, 2007; Kaestner, 2010).

A key demonstration of the regulation of ER-α binding by FOXA1 came from a study in ER-positive breast cancer cell lines (MCF-7♀, T-47D♀, and ZR75-1♀) (Hurtado *et al.*, 2011). Using ChIP-seq the authors confirmed previous results showing that FOXA1 and ER-α co-occur at sites throughout the genome (Carroll *et al.*, 2005), and showed that this co-occurrence persists at sites unique to each cell line (Hurtado *et al.*, 2011). Although this co-occurrence constituted only 50% of the total of binding sites for FOXA1 and ER-α, KD of FOXA1 in MCF-7♀ cells showed that 90% of all ER-α binding was lost despite no change in the overall expression of ER-α. By using FAIRE-seq, the authors then showed that FOXA1 exerts its role as a pioneering factor by allowing the binding of ER-α to previously inaccessible chromatin. Demonstrating that this relationship is not confined to breast cancer cell lines, ER-α chromatin interactions were shown to be promoted in ovarian and osteosarcoma cell lines with increased FOXA1 expression (Hurtado *et al.*, 2011). A recent meta-analysis of publications regarding FOXA1 in breast cancer showed that FOXA1 was positively associated with ER status and survival outcome, but that its expression significantly predicted poor response to chemotherapy in ER-positive breast cancer patients (Shou *et al.*, 2016)

A similar relationship was shown between FOXA1 and AR in prostate cancer (Sahu *et al.*, 2011). Unlike in breast and bladder cancer, increased expression of FOXA1 is associated with poor prognosis in prostate cancer. Using a combination of ChIP-seq, DNase-seq and gene expression microarrays in LNCaP-1FS cells with FOXA1 KD, it was shown that FOXA1 can act either as a pioneering factor to allow AR binding to inaccessible chromatin, or by masking AR binding sites that require the functional depletion of FOXA1 to then allow AR to bind (Sahu *et al.*, 2011). FOXA1 regulation of AR binding accounted for

a considerable number of binding events, but it was noted that there was still a large subset of AR-chromatin interaction events that did not require FOXA1.

In bladder, expression of FOXA1 and FOXA2 has been shown to play an essential role in urothelial development, with FOXA1 expression perisisting into the adult urothelium (Oottamasathien *et al.*, 2007). However, in bladder cancer FOXA1 expression is inversely correlated with tumour stage and grade (DeGraff *et al.*, 2012; Reddy *et al.*, 2015; Warrick *et al.*, 2016). A study on >600 bladder cancer tissues from 302 patients that underwent cystectomy demonstrated that FOXA1 could act as an independent predictor of patient survival after considering patient age, sex and tumour stage (Reddy *et al.*, 2015). Furthermore, by utilising an inducible Cre-*LoxP* system to knockdown *Foxa1* in the urothelium of adult mice, the authors showed sexually distinct histological characteristics, where adult male mice developed urothelial hyperplasia and female mice developed keratinizing squamous metaplasia of the urothelium (Reddy *et al.*, 2015). The authors also described no apparent differential expression of genes in mice with forced over-expression of AR and ER in the presence or absence of FOXA1, in contrast to the hormone regulatory effects of FOXA1 observed in prostate and breast.

An earlier study showed that reduced FOXA1 was associated with higher stage and grade bladder tumours, where IHC showed positive expression of FOXA1 in 100%, 67%, 59%, 42%, and 34% of Ta, T1, T2, T3, and T4 tumours respectively. The study also used bladder cancer cell lines to show negligible expression of FOXA1 in T24♀, J82♂, 5637♂, 253J♂, UM-UC-3♂ and SCaBER♂, but not RT4♂ cell lines. This coincided with decreased expression of uroplakins (UPK), including the urothelial differentiation marker UPK2 (DeGraff *et al.*, 2012). Conversely, the study then went on to show that KD of FOXA1 in RT4♂ cells increased UPK expression and decreased E-cadherin expression, cell growth and invasion, whereas exogenous expression of FOXA1 in T24♀ decreased UPK expression and increased E-cadherin, cell growth and invasion (DeGraff *et al.*, 2012).

Other cell line studies have implicated FOXA1 in urothelial differentiation alongside GATA3 and PPARγ. When 27 different cell lines were classified into 3 subgroups (basal, luminal, other) according to copy-number alteration, exome, and expression data, FOXA1 and GATA3 were almost exclusively restricted to the luminal subgroup. FOXA1 and GATA3 were also shown to cooperate with PPARγ to reprogramme the basal-like 5637♂ cell line into a more luminal phenotype (Warrick *et al.*, 2016). These results implicate FOXA1 in bladder cancer development and progression, as the luminal subtype comprises the majority of early-stage and non-invasive bladder cancers, whereas the basal subtype is more aggressive and persists almost exclusively in more advanced and invasive stages of the disease

(Weinstein *et al.*, 2014). A recent study utilising FAIRE-seq further demonstrated the role of FOXA1, GATA3 and PPARγ in driving a luminal phenotype in urothelial cells. The study showed that NHUC assumed a non-differentiated basal phenotype maintained by TP63, but upon PPARγ activation by troglitazone, GATA3 and FOXA1 cooperated to drive the expression of luminal marker genes including UPK1A and UPK2 (Fishwick *et al.*, 2017).

1.5 Epigenetics and bladder cancer

1.5.1 Epigenetics

Epigenetics literally means "above genetics", and is the platform by which over 200 specialised cell types, each with stable but transient expression profiles, are able to arise from a single human genome. Epigenetic mechanisms primarily concern post-translational modifications of histones, a core component of chromatin, or modifications to nucleotides in DNA, most notably the methylation of cytosine residues (5mC) (Kouzarides, 2007; Bannister and Kouzarides, 2011; Greer and Shi, 2012). Such modifications to DNA and histones can be highly dynamic and result in changes to both global and local chromatin architecture. This may include promoting "open" euchromatin that is easily accessible to transcriptional machinery and associated with active transcription, or conversely promoting "closed" or heterochromatin that is less accessible and is associated with inactive genes (Kouzarides, 2007; Zentner and Henikoff, 2013). Histone modifications and DNA methylation exert their influence on transcriptional regulation and other processes such as DNA replication, DNA damage response and splicing, by promoting the condensation or relaxation of chromatin either directly or indirectly by recruiting/blocking other factors which themselves alter the state of chromatin (Kouzarides, 2007; Greer and Shi, 2012). These modifications are independent of the DNA sequence itself but can be heritable, thus epigenetics refers to heritable changes to gene expression that are attributed to alterations in chromatin structure but not the underlying genetic code.

1.5.1.1 Chromatin architecture

In eukaryotes, the compaction of genomic DNA (gDNA) into individual cell nuclei is achieved by wrapping DNA around nucleosome structures that accumulate to form chromatin (Kouzarides, 2007). Nucleosomes are 147 base pairs of DNA wound 1.7 times around an octameric histone complex composed of histone proteins H2A, H2B, H3, and H4 arranged into four histone heterodimers (two each of H3-H4 and H2A-H2B) (Luger *et al.*, 1997; Szerlong and Hansen, 2011; Luger *et al.*, 2012). Nucleosomes are connected by a stretch of "linker" DNA which may vary from 20-90bp in length. The orientation of linker

DNA entering or exiting the nucleosome complex can also be influenced by two more linker histones, H1 and H5 (Szerlong and Hansen, 2011). Histone proteins are predominantly globular in structure, with the exception of the N-terminal tails that can be post-translationally modified and alter the structure of the surrounding chromatin.

This chromatin structure of DNA wrapped around octameric nucleosomes has been long known, and on a linear plane has been described as "beads on a string" (Kornberg, 1974). More recent chromosome conformation capture (3C) technologies that are coupled with next-generation sequencing (Hi-C) have shown a sophisticated and highly transient organisation of chromatin within cell nuclei (Rowley and Corces, 2018; Eagen, 2018; Zheng and Xie, 2019). Such studies have shown that, at the level of the nuclei and entire chromosomes (megabase resolution), chromatin segregates into two main compartments that are largely euchromatin and heterochromatin, and at least six sub-compartments that again are euchromatin (2 sub-compartments) and heterochromatin (4 sub-compartments) but with distinct histone modification profiles (Rao *et al.*, 2014). At kilobase resolution, chromatin is organised into topologically associated domains (TADs), which may contain TADs within them, and at the resolution of hundreds of base pairs individual chromatin loops have been visualised that frequently link promoter and enhancer regions to regulate gene transcription (Rao *et al.*, 2014; Greenwald *et al.*, 2019). This intricate structure of chromatin looping is predominately controlled by CTCF and Cohesin complexes (Rao *et al.*, 2014), and although chromatin architecture is known to be transient and change with gene expression (Rowley and Corces, 2018), depletion of CTCF and Cohesin has been shown to eliminate chromatin looping but not dramatically alter global transcription (Rao *et al.*, 2017). Therefore the role of chromatin architecture may not be primarily transcriptional regulation, but other processes such as the maintenance, repair, and replication of DNA (Eagen, 2018; Zheng and Xie, 2019). This description of a transient chromatin structure is most relevant to cell interphase. However, the most dynamic changes are seen as a result of cell-cycle progression, where global chromatin condensation is seen in preparation for mitosis, chromatin relaxation in early G1-phase, and DNA replication is correlated with increased compartmentalisation and TAD insulation, but with chromatin loops remaining intact throughout interphase (Nagano *et al.*, 2017).

1.5.1.2 Histone modifications

Over 100 different types of covalent modifications to residues on N-terminal histone tails have been described, and include well-studied modifications such as acetylation, methylation, and phosphorylation, as well as more unusual modifications such as crotonylation (Bannister and Kouzarides, 2011; Zentner and Henikoff, 2013). The regulation of histone modifications can be highly dynamic due to their regulation by antagonistic histone-modifying proteins. Furthermore, histone modifications may exert their effects by directly influencing the chromatin structure, or by recruiting other chromatin and DNA binding factors (Tessarz and Kouzarides, 2014). As histone marks can perpetuate differential regulation of the chromatin, they can be used to infer types of regulatory regions and their activation state. For instance, enhancer regions are typically marked by H3K4me1, but are considered active when this mark co-occurs with H3K27ac and chromatin accessibility, and inactive if the mark co-occurs with H3K27me3 and inaccessible chromatin. Some of the best characterised post-translational histone modifications are outlined below.

1.5.1.2.1 Acetylation

The first reported histone modification was histone acetylation (Phillips, 1963). However, functional attributions of histone acetylation came later when Allfrey *et al* demonstrated that histone acetylation regulated RNA synthesis *in vitro* by showing that increased amounts of histone arginine-acetylation coincided with an increased uptake of ATP into RNA and decreased inhibition of transcription (Allfrey *et al*., 1964). It is now widely accepted that histone acetylation is associated with active transcription and directly affects chromatin by neutralising the positive charge of lysine residues (Hong *et al*., 1993). This weakens charge-dependent interactions between the DNA and histones and between neighbouring nucleosomes, and promotes a more euchromatic state (Zentner and Henikoff, 2013; Bannister and Kouzarides, 2011). Although individual lysine residues can be acetylated at one location, it is the accumulation of acetylation that promotes euchromatin formation, as an individual acetylation event has minimal impact on DNA-nucleosome interactions or chromatin architecture (Dion *et al*., 2005). Histone acetylation is regulated by histone acetyltransferases (HATs) and histone deacetylases (HDACs) that antagonise one another to allow for highly dynamic control.

1.5.1.2.2 Methylation

Histone methylation is perhaps the best documented of all histone modifications. Methylation occurs on the basic lysine and arginine residues of histone tails. Lysine can undergo mono- (me1), di- (me2) or tri- (me3) methylation with the best-characterised

including those that occur on histone H3 lysine-4 (H3K4), H3K9, H3K27, H3K36, H3K79 and H4K20. Methylation of arginine residues may be mono- (me1), symmetrically dimethylated (me2s) or asymmetrically dimethylated (me2) and includes H4R3, H3R17, H3R26, and H3R2 (Greer and Shi, 2012). Although these methylation events are the most studied, nearly all basic histone-tail residues have been found methylated (Zentner and Henikoff, 2013). Methylation itself does not alter the charge of histone tails and so does not directly alter the chromatin structure. Instead, methylation acts as a platform to recruit or block chromatin binding proteins or transcription factors in order to remodel chromatin structure and alter the transcriptional status of a gene (Kouzarides, 2007). Such proteins often contain methyl-binding domains (MBD) such as PHD fingers, WD40 repeats, CW domains, PWWP domains, and ankyrin repeats (Greer and Shi, 2012). Histone methylation can be associated with both gene activation, such as with H3K4me3 which is found at the promoter of active genes (Schneider *et al.*, 2004) and H3K36me3 which is found within active gene bodies (Bannister *et al.*, 2005), or repression, such as H3K27me3 which is a marker of facultative heterochromatin regions, or H3K9me3 which is a marker of constitutive heterochromatin at the promoters of inactive genes (Trojer and Reinberg, 2007). Common marks for promoter and enhancer regulation can be seen in Figure 1.2. Histone methylation is regulated by histone methyltransferases (HMTs) and histone demethylases (HDMTs), which have much greater specificity than HDACs and HATs.

1.5.1.2.3 Phosphorylation

Histone phosphorylation is similar to histone acetylation in that it is able to directly modify the chromatin architecture by placing negative charges onto positively charged histone tails to promote euchromatin and active transcription, as well as to serve as a recognition platform for phospho-recognition proteins (Bannister and Kouzarides, 2011; Zentner and Henikoff, 2013). For instance, heterochromatin Protein 1 (HP1) is inhibited from binding to H3K9me3 by H3S10p. This results in gene activation, as HP1 no longer recognises the repressive H3K9me3 mark to promote heterochromatin formation (Hirota *et al.*, 2005). Phosphorylation events occur on serine, threonine, and tyrosine residues and are regulated by kinases and phosphatases that phosphorylate and dephosphorylate residues respectively. Unlike acetylation, phosphorylation tends to be site-specific, with individual phosphorylation events capable of influencing chromatin structure. This is likely due to the ability of individual phosphorylation modifications to displace binding of chromatin modifying factors such as HP1 without directly altering nucleosome-DNA charge-dependent

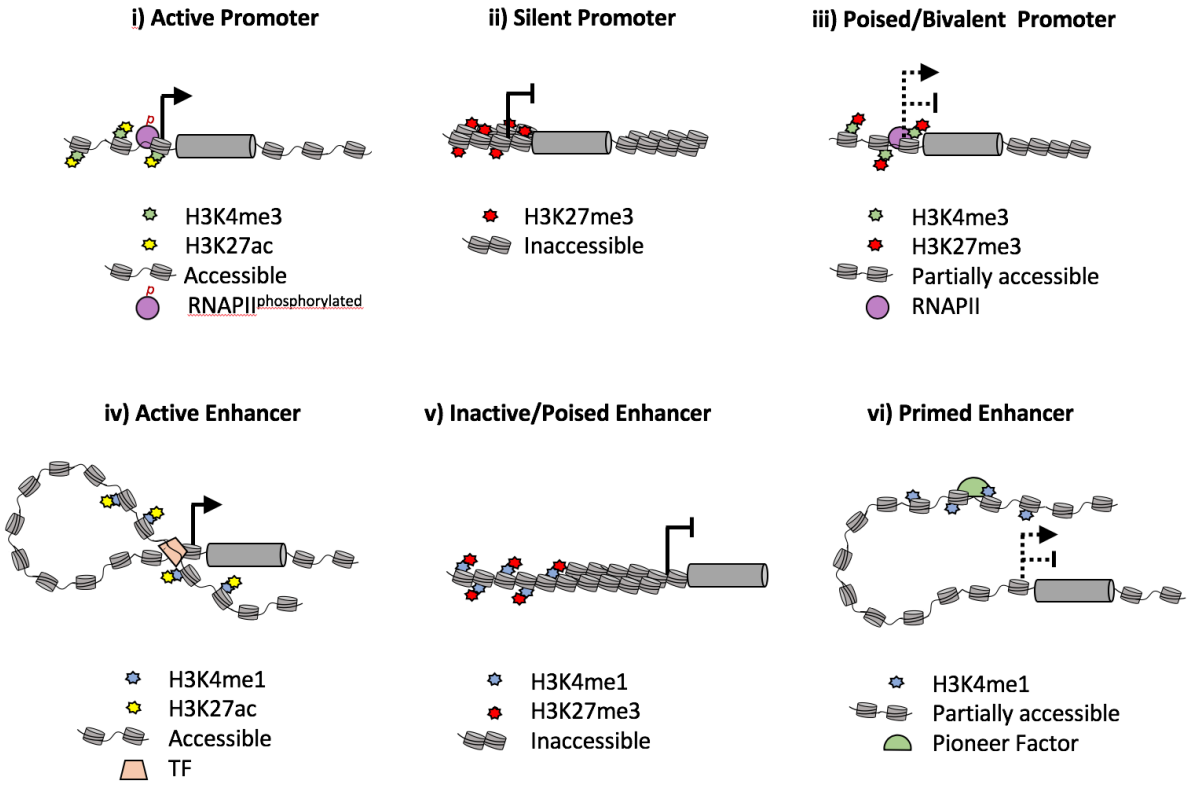


Figure 1.2 Common chromatin markers at promoter and enhancer regions

i) Active promoter regions are marked by H3K4me3 and H3K27ac and have a high degree of chromatin accessibility. Active transcription also requires phosphorylated RNAPII. ii) Silent promoters are inaccessible and are marked by H3K27me3 or H3K9me3. iii) Poised promoters are marked by H3K4me3 and H3K27me3, have non-phosphorylated RNAPII, and can quickly become transcriptionally active, such as in response to stimuli. iv) Active enhancers are marked by H3K4me1 and H3K27ac, and have a high degree of chromatin accessibility. They are also bound by transcription factors and other transcriptional machinery which are brought into close proximity of promoters through chromatin looping. v) Inactive/poised enhancers are marked by H3K4me1 and H3K27me3 and are inaccessible. vi) Primed enhancers are marked by H3K4me, are partially accessible, and often targeted by pioneering factors which often activate the enhancer, but also inactivate the enhancer.

interactions. Like acetylation, the accumulation of phosphorylation is needed to directly influence chromosome architecture (Rossetto *et al.*, 2012). Histone phosphorylation is less well studied compared to acetylation, although it is mostly regarded for its role in DNA damage repair mechanisms (Rossetto *et al.*, 2012). This is similar to more recently discovered histone modifications such as crotonylation, formylation, succinylation and malonylation, which all also neutralise the positive charges of lysine and possibly promote a euchromatin state.

1.5.1.2.4 Ubiquitylation

The previously described histone modifications include very small adjustments to individual residues of histone tails. In contrast to these is ubiquitylation, a modification that entails the binding of a much larger 76 amino acid polypeptide to histone tails. Such a large modification results in architectural changes to the chromatin that may result in transcriptional activation (as with H2BK123ub1 (Lee *et al.*, 2007a) or transcriptional silencing (as with H2AK119ub1 (Hengbin *et al.*, 2004)) by opening up the chromatin or by blocking access to DNA. Ubiquitylation of histones is carried out by E1, E2, and E3 enzymes that provide specificity as well as the degree (mono- or poly-) to which a histone is ubiquitylated (Hershko and Aaron, 1998). The removal of ubiquitin is carried out by isopeptidases, thus antagonising the effects of the previously mentioned enzymes and allowing dynamic regulation of histone ubiquitylation

1.5.1.3 DNA methylation

In mammalian cells, DNA methylation occurs on the 5th carbon of cytosines (5mC) found in CpG dinucleotides (Caiafa and Zampieri, 2005). Although methylation exerts no change in the DNA or chromatin structure itself, it is still considered a key signature for transcriptional repression. A common mode of action for 5mC is to flag areas of the genome to MBD proteins that in turn recruit HDACs to remove histone acetylation and promote a closed chromatin structure (Cedar and Bergman, 2009). 5mC is regulated by three DNA methyltransferases: DNMT1 for maintenance, and DNMT3a and DNMT3b for *de novo* methylation. Although no direct demethylases for 5mC have yet been found in mammals, the ten-eleven translocation (TET) family of proteins TET1, TET2, and TET3 have been found to oxidise 5mC through 5-hydroxymethylation (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC) intermediates before replacement with an unmodified cytosine residue. However, this mode of "active demethylation" is a high-energy conversion that is not thought to be used for the removal of 5mC. Therefore "passive" demethylation is considered to be the primary method for DNA demethylation. Passive demethylation takes

place when DNMT1 is diminished and 5mC is not re-established on the new DNA strand following DNA replication (Klose and Bird, 2006).

1.5.2 Next-generation sequencing techniques and epigenetics

The history of DNA sequencing spans over 40 years, has been extremely rapid in pace, and began with Fredrick Sanger and a technique which won him his second Nobel prize in 1980. Sanger sequencing used a chain terminating procedure to produce radioactively labelled DNA fragments that could be visualised by polyacrylamide gel electrophoresis, and could determine a DNA sequence at single-base resolution (Sanger *et al.*, 1977). The technique was eventually coupled with a "shotgun-sequencing" approach to sequence large fragments of the human genome that were cloned into bacterial artificial chromosomes, which over the course of ten years culminated in the first completed sequence of the human genome (IHGS Consortium, 2004). Following the Human Genome Project (HGP), Sanger sequencing was quickly superseded by "massively parallel" or "next-generation" sequencing (NGS) technologies. Although initially a competitive market, Illumina systems now dominate NGS with platforms that carry out bridge amplification of small-length DNA reads followed by sequencing-by-synthesis through stepwise polymerase-mediated incorporation of fluorescently labelled deoxynucleotides (Shendure *et al.*, 2017). NGS technologies have enabled increasingly higher throughput sequencing at lower costs and increased accuracy, allowing individual laboratories to affordably sequence entire genomes in less than a day. Future single molecule "third-generation" sequencing approaches such as those being developed by PacBio and Nanopore are likely to supersede NGS, as they allow reads megabases in length to be sequenced by smaller and more accessible platforms, and will facilitate *de novo* genome assembly and better transcriptome profiling (Shendure *et al.*, 2017; Marinov, 2018). Nevertheless, NGS technologies have been particularly impactful in the realm of functional genomics where techniques such as ChIP-seq, ATAC-seq, and HiC (amongst many others) have enabled genome-wide studies that characterise varying aspects of molecular biology such as protein occupancy, chromatin accessibility, and chromatin architecture. Indeed, following the HGP, large consortia such as ENCODE, modENCODE, and the Roadmap Epigenetics Consortium have been set up which aim to provide reference epigenomes for all cell and tissue types in human and other species (ENCODE Consortium, 2012; Contrino *et al.*, 2012; Kundaje *et al.*, 2015).

1.5.2.1 DNA-protein interactions by ChIP-seq

Chromatin immunoprecipitation (ChIP) has been the primary method of characterising protein-DNA interactions since its inception, when it was used to show RNA polymerase II (RNAPII) occupation on the *cI* gene and *lac* operon in *E.coli* (Gilmour and Lis,

1984). Although the methodology has changed considerably in that time, the overall principles of the technique have remained unchanged. A typical ChIP protocol involves the chemical cross-linking of proteins to DNA (often using formaldehyde), followed by physical or enzymatic fragmentation of DNA, enrichment of DNA fragments bound to the protein of interest, and subsequent reversal of cross-linking, purification, and analysis of enriched DNA (Figure 1.3) (Marinov and Kundaje, 2018). Analysis of enriched DNA fragments was initially carried out using PCR/qPCR and showed protein enrichment at the level of individual loci (ChIP-qPCR). Genome-scale approaches were then developed which coupled ChIP with microarrays (ChIP-Chip/ChIP-on-Chip), but these offered low resolution and were typically limited to the promoter regions of known genes (Lieb *et al.*, 2001). First efforts to directly sequence ChIP-enriched DNA fragments were made using paired-end tagging (ChIP-PET) (Wei *et al.*, 2006), but were quickly superseded by more efficient library preparation techniques that enabled coupling of ChIP with NGS platforms (ChIP-seq) for truly genome-wide low-resolution mapping of protein occupation (Barski *et al.*, 2007). ChIP-seq has since enabled researchers to characterise the molecular functions of transcription factors, such as their recognition of DNA-binding motifs and how they are associated with histone modifications to regulate gene expression (Jolma *et al.*, 2013; Zhang *et al.*, 2018). Nevertheless, the technique is still limited in its resolution (usually 100-500bp due to the DNA fragmentation procedure) and typically requires a high input of DNA from millions of cells. Various approaches to overcome these limitations have been demonstrated and include: the use of exonuclease to trim ChIP-DNA to precise distances from crosslinking sites (ChIP-exo and ChIP-nexus) (Rhee and Pugh, 2011; He *et al.*, 2015), antibody-targeted controlled cleavage by micrococcal nuclease (MNase) instead of crosslinking and sonication (CUT&RUN) (Skene and Henikoff, 2017), and Tn5 transposase adapter ligation (tagmentation) for library preparation following ChIP (ChIPmentation) (Schmidl *et al.*, 2015). Although these techniques have demonstrated increased resolution using a lower DNA input, they are yet to supersede conventional ChIP-seq protocols and require further optimisation of downstream data analysis pipelines.

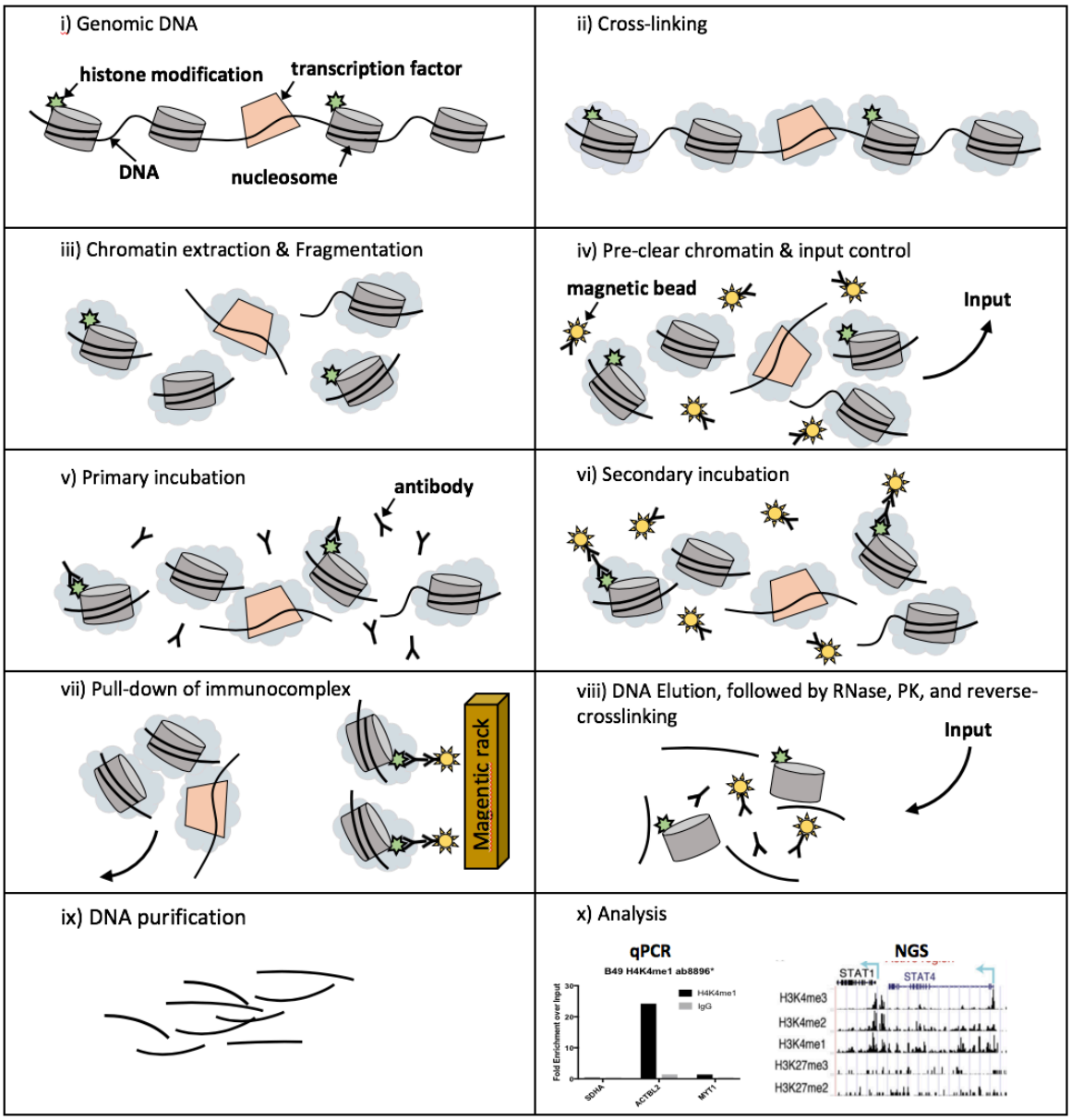


Figure 1.3 Chromatin Immunoprecipitation (ChIP)

ChIP is a commonly used technique to determine DNA-protein interactions. i) A simplified DNA locus with a bound transcription factor and DNA wrapped around histones (nucleosomes), some of which have a post-translational histone modification which will be targeted in this example. ii) ChIP begins with cross-linking, usually with formaldehyde, of live cells to fix protein-DNA interactions. iii) gDNA is extracted from cells and fragmented to produce 100-500bp fragments. iv) Fragmented DNA is incubated with magnetic agarose beads to reduce non-specific binding. Following pre-clearing, a fraction of the sample is kept aside as input control. v) Pre-cleared chromatin is incubated with antibodies targeting the protein of interest, in this case a histone modification. vi) Secondary incubation is carried out using magnetic beads which bind to the primary antibody, resulting in immunocomplexes consisting of magnetic beads, primary antibodies, and histones with modifications of interest bound to DNA fragments. vii) The immunocomplex is pulled down using a magnetic rack, and supernatants are discarded. viii) The immunocomplex is subjected to a series of washes in LiCl, RIPA, and TE buffer. DNA is eluted then treated with RNase, proteinase K (PK), and reverse cross-linked with high salt and temperature. ix) DNA is purified, often using phenol:chloroform protocols. x) ChIP DNA is analysed immediately by qPCR (ChIP-qPCR) to determine locus-specific protein occupation, or used for library preparation and NGS (ChIP-seq) to determine genome-wide protein occupancy.

1.5.2.2 Chromatin accessibility by DNase-seq and ATAC-seq

Accessible chromatin regions constitute only ~3% of the genome but account for over 90% of regions bound by transcription factors, and the majority of transcription factors exclusively bind to open chromatin (Thurman *et al.*, 2012). The depletion of nucleosomes and binding of transcription factors predominantly occurs at regulatory loci such as enhancers, insulators, and the promoters of active genes. Therefore, the regulatory potential of a genetic locus can be inferred by its chromatin accessibility (Liu *et al.*, 2019). Techniques used to determine chromatin accessibility have traditionally been carried out using deoxyribonuclease (DNase) I, an enzyme that degrades DNA by catalysing the hydrolytic cleavage of phosphodiester linkages in the DNA backbone. By limiting digestion of gDNA by DNase I, regions of open chromatin that are displaced by bound transcription factors and mark *cis*-regulatory regions are preferentially digested, and termed DNase I-hypersensitive sites. This method was first applied to identify DNase I-hypersensitive loci at two heat-inducible polypeptide genes in *Drosophila* which were visualised using traditional Southern blot assays (Wu, 1980). This technique was used for over 25 years before genome-scale approaches that used microarray chips (DNase-chip) were developed (Crawford *et al.*, 2006). DNase-Chip was soon superseded by coupling with NGS (DNase-seq) which offered high-throughput genome-wide mapping of DNase I hypersensitive sites (Boyle *et al.*, 2008). DNase-seq was long considered the gold standard for assaying chromatin accessible regions, although other techniques have also been commonly used, such as formaldehyde-assisted isolation of regulatory elements coupled with NGS (FAIRE-seq) which uses a sonication-based approach for determining chromatin accessibility (Giresi *et al.*, 2007), and MNase digestion of gDNA followed by NGS (MNase-seq) to determine nucleosome positioning (Schones *et al.*, 2008). Although long considered the gold standard for assaying chromatin accessibility, a typical DNase-seq protocol requires hundreds of thousands of cells and is a multi-step process that can take days.

A more recent technique, namely the assay for transposase-accessible chromatin using NGS (ATAC-seq) (Figure 1.4), has superseded DNase-seq in becoming the preferred approach to determine genome-wide chromatin accessibility, owing to a quick and easy protocol and low input requirements (Buenrostro *et al.*, 2013). ATAC-seq uses a hyperactive Tn5 transposase that simultaneously cuts DNA and incorporates sequencing adapters. As with DNase-seq, a partial digestion of gDNA is carried out using Tn5 which allows preferential cutting of easily accessible regions whilst steric hinderance by proteins that decorate the DNA prevents cutting and the incorporation of sequencing adapters. ATAC-

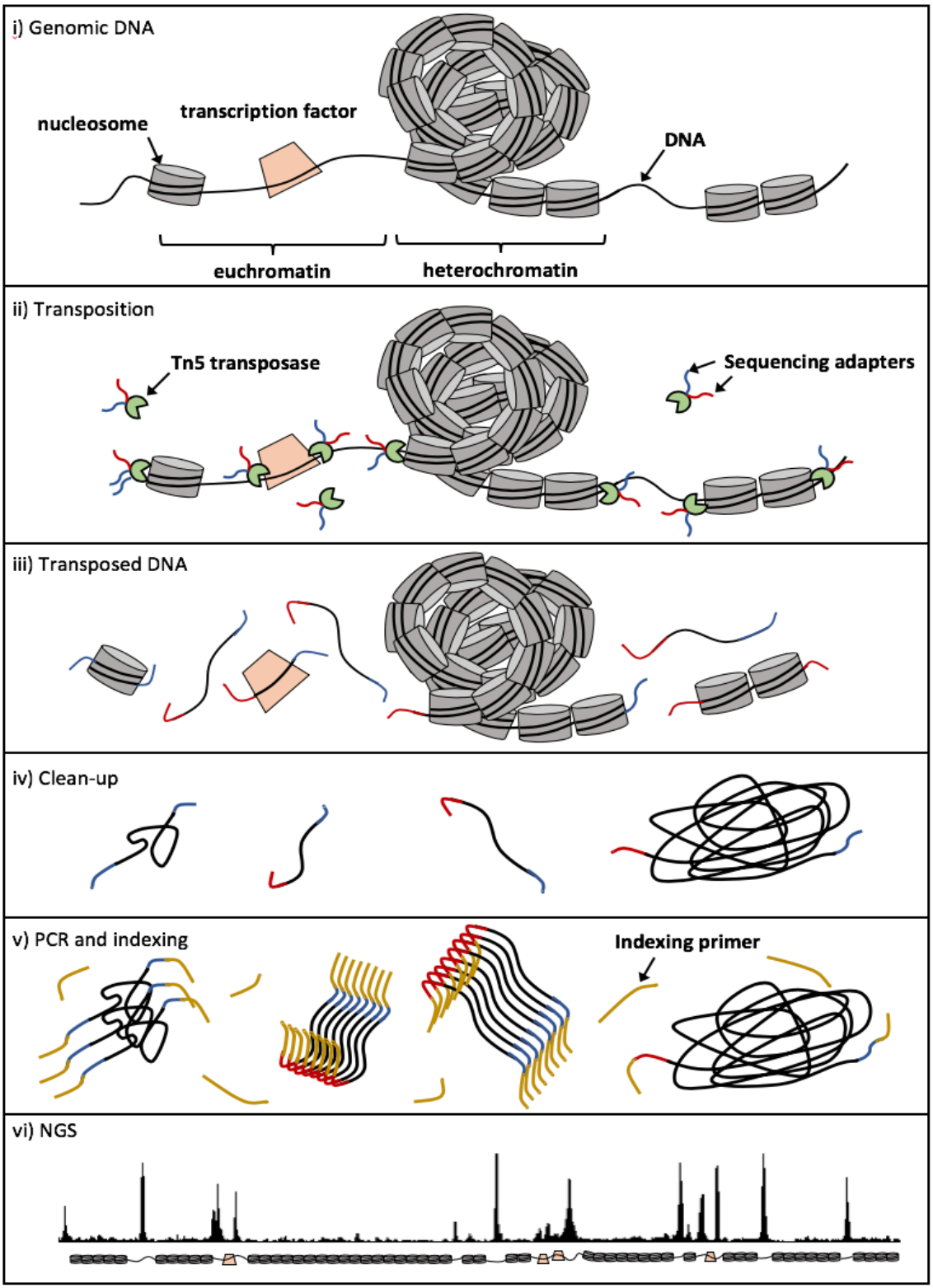


Figure 1.4 Assay for transposase accessible chromatin (ATAC-seq)

Illustration of a standard ATAC-seq protocol. i) Nuclei are isolated from ~50,000 cells. Illustration shows a simplified gDNA locus. ii) Transposition is carried out with a partial digestion of gDNA using Tn5 transposase which preferentially digests accessible chromatin. iii) Tn5 simultaneously incorporates sequencing adapters when cutting DNA; partial digestion limits transposition due to steric hinderance by proteins such as nucleosomes and transcription factors. iv) Transposed DNA is purified using a PCR clean-up kit. v) Transposed DNA is amplified by PCR using indexing primers which bind to library adapter. Following PCR, DNA is purified using a PCR clean-up kit. vi) Amplified libraries are used for NGS.

seq is therefore highly correlated with DNase-seq data and captures similar regulatory information, but requires fewer than 50,000 cells and a protocol that can take as little as 2 hours (Buenrostro *et al.*, 2013; Klemm *et al.*, 2019). As transcription factors recognise and bind to consensus motifs on DNA, chromatin accessible sites identified by ATAC-seq or DNase-seq can also be used to infer the transcription factors bound at *cis*-regulatory sites throughout the genome (Buenrostro *et al.*, 2013; Li *et al.*, 2019)

1.5.2.3 DNA methylation analysis by MRE-seq, MeDIP-seq, BS-seq

There have historically been three main approaches for determining the distribution of 5mC. The first is use of methylation-sensitive restriction enzymes (MRE) that only cut if their DNA recognition sites are unmethylated (such as *Hpa*II or *Not*I) or methylated (such as *Mcr*BC), and was traditionally used to look at individual sites using gel electrophoresis following PCR, or by Southern blotting. MRE has also been coupled with microarray chips (MRE-chip) and NGS (MRE-seq) for genome-wide determination of 5mC, but these techniques offer low resolution as they are limited to CpG sites located within MRE recognition sites (Yong *et al.*, 2016). A more effective technique is methylated-DNA immunoprecipitation (MeDIP), which is similar to ChIP in that it uses antibodies which recognise 5mC to enrich fragmented DNA by immunoprecipitation. MeDIP can be coupled with qPCR to quantify 5mC at individual loci or with NGS (MeDIP-seq) to determine genome-wide 5mC (Down *et al.*, 2008). MeDIP is also limited by a resolution of ~150bp due to DNA fragmentation, and therefore cannot distinguish between single or multiple methylation events, or determine the exact CpG dinucleotide that is methylated (Yong *et al.*, 2016). The gold standard for profiling 5mC is bisulphite sequencing (BS-seq), which treats gDNA with sodium bisulphite to transform unmodified cytosine residues into uracil, which are then converted into thymine by PCR. This treatment is then followed by sequencing, from which 5mC can be inferred from the remaining cytosine residues in the sequence. BS-seq can therefore be coupled to NGS to provide genome-wide mapping of 5mC at single-base resolution. (Cokus *et al.*, 2008). Limitations to BS-seq include the mappability of reads to the genome as a result of high thymine content, and the inability to distinguish between 5mC and 5hmC, which also prevents cytosine conversion to uracil by bisulphite treatment (Yong *et al.*, 2016).

1.5.2.4 Chromosome architecture by 3C technologies

Early studies of chromosome organisation were mainly carried out by fluorescence *in situ* hybridisation (FISH) and were at a microscopic level. Sequencing technologies have since enabled genome-wide determination of DNA-DNA interactions and advanced the understanding of chromatin architecture. These techniques are based on chromosome

conformation capture (3C) technologies, which generally use formaldehyde crosslinking to fix interacting DNA (such as in chromatin loops and TADs), followed by restriction enzyme digestion of crosslinked DNA, and then re-ligation. This results in DNA motifs which were previously just in close proximity to each over now constituting individual DNA fragments (Davies *et al.*, 2017). Early 3C techniques were only coupled with PCR to assess interactions between individual loci (one-to-one) (Dekker, 2002). However, improvements to 3C and sequencing technologies enables the development of circular 3C (4C) to determine all potential interacting partners of any particular loci of interest (one-to-all)(Zhao *et al.*, 2006), and carbon copy 3C (5C) which uses hybridisation and ligation of oligos to look at all interactions within larger loci of interest (many-to-many) (Dostie *et al.*, 2006). The gold standard for the study of chromosome architecture is the combination of 3C and NGS (Hi-C), a typical protocol for which requires biotin fill-in of restriction cut sites, followed by blunt ligation, pulldown of the biotin-marked 3C fragments using streptavidin beads, and the addition of sequencing adapters (Lieberman-Aiden *et al.*, 2009). Hi-C therefore looks at all interacting sites throughout the genome (all-to-all), and has been fundamental in describing the principles of chromatin looping, TADs, and the greater chromosome architecture. Nevertheless, Hi-C is still limited in resolution by the distance between restriction sites, and although resolution down to 1kb has been described by increasing sequencing depth and using 4bp restriction enzymes, such approaches are extremely costly and require billions of reads per sample (Davies *et al.*, 2017).

1.5.3 Epigenetic studies in bladder cancer

1.5.3.1 DNA methylation in bladder cancer

The majority of epigenetic studies in bladder cancer concern DNA methylation (Porten, 2018). Such studies have shown that 50-90% of bladder cancers have hypermethylation at the promoters of genes that regulate tumour suppression, DNA repair, cell cycle, and cell invasion (Yates *et al.*, 2007; Reinert *et al.*, 2011). Hypermethylation of promoters is probably an early event in the development of bladder cancer, with one study that used BS-seq Illumina chips showing that 89% of hypermethylated CpG sites were shared between all stages of bladder cancers, and that the number of methylated loci was correlated with tumour invasiveness (Wolff *et al.*, 2010). DNA hypomethylation in bladder cancer has also been shown at CpG-rich satellite regions such as the long-interspersed retroelement 1 (LINE-1), and is known to promote genomic instability (Flori *et al.*, 1999; Porten, 2018). Most studies of DNA methylation in bladder cancer primarily concern identifying biomarkers as diagnostic and prognostic tools in order to reduce the invasiveness of clinical tests. However, such studies have failed to show consistent results for the prognostic

implications of DNA methylation-based biomarkers, as shown by two large systematic reviews encompassing over 120 studies which found large inconsistency and low reproducibility of DNA methylation biomarkers throughout the literature (Casadevall *et al.*, 2017; Porten, 2018).

1.5.3.2 HDAC and KDM1A in bladder cancer

High levels of HDAC1, HDAC2, and HDAC3 were shown by IHC in 40%, 42% and 59% of samples from 174 bladder cancer patients, with increased HDAC1 and HDAC2 associated with high-grade NMIBCs (Poyet *et al.*, 2014). However, this study did not include normal urothelium as a control. These results correlated with a later study which used IHC on formalin-fixed paraffin-embedded tissues from 271 bladder cancer tumours and 29 normal urothelial samples, to show reduced levels of H3ac, but not H4ac or H3K18ac, in bladder tumours compared to normal urothelium, and that these histone acetylation markers were lower in MIBC compared to NMIBC samples (Ellinger *et al.*, 2016).

The H3K4me1/2 and H3K9me1/2 HMT KDM1A was shown by qPCR and IHC staining to be highly expressed in bladder cancer cells compared to normal bladder tissue (Hayami *et al.*, 2011), and both siRNA KD and chemical inhibition of LSD1 resulted in reduced proliferation of TCCSUP♀ and HT1376♀, but not J82♂ bladder cancer cell lines (Kauffman *et al.*, 2011).

1.5.3.3 NGS-based epigenetic studies in bladder cancer

Although ChIP-seq has been ubiquitously used throughout the literature, including studies ranging from developmental biology to cancer, and in nearly all tissue types, there have been very few examples where ChIP-seq has been used to study epigenetics in the urothelium. For example, of the >15,000 entries spanning 147 cell and tissues types from the ENCODE consortium, only 6 entries include ChIP-seq for histone marks in bladder, all of these have insufficient or extremely low read depth, short read lengths, and low library complexity (ENCODE Consortium, 2012). The Roadmap Epigenomics Project, that aims to characterise the epigenomes of 111 primary cell/tissue types, also does not include bladder samples (Kundaje *et al.*, 2015). In smaller studies, ChIP-seq for H3K27me3 and H3K9me3 was carried out using RT112♀ and T24♀ bladder cancer cell lines, but failed to include essential input controls (Dudziec *et al.*, 2012). Effective normalisation and useful interpretation of this data is therefore impossible, and was demonstrated by the low reproducibility in the study (Dudziec *et al.*, 2012). A recent study which developed a novel approach for carrying out ChIP-seq on formalin-fixed, paraffin-embedded tissue samples (FiT-seq), demonstrated the protocol for enrichment of H3K4me2 in 6 NMIBC tumour

samples (Cejas *et al.*, 2016). However, the anti-H3K4me2 antibody used in this study had previously been shown to preferentially bind to other histone marks in a separate study, including increased affinity for H3K4me1 and H3K4me3 (Rothbart *et al.*, 2015). Other studies have used ChIP-seq for histone marks in bladder cancer cell lines, and will be discussed later (Chen *et al.*, 2015; S. Wu *et al.*, 2016; Ler *et al.*, 2017).

Aside from ChIP-seq for histone modifications, ChIP-seq has been carried out for FOXA1 in RT4♂ cells, again with very low reproducibility between replicates (Warrick *et al.*, 2016). Nevertheless, this study showed that GATA3 and PPARγ cooperate with FOXA1 to promote a luminal phenotype in bladder cancer, a finding that was demonstrated again in a later study which used FAIRE-seq to identify differential FOXA1, GATA3, and p63 binding motifs in chromatin accessible regions in PPARγ-activated, differentiating NHUC (Fishwick *et al.*, 2017). This latter study was also the first of only a few to assess chromatin accessibility in urothelial cells. Another study profiled chromatin accessibility in 410 tumour samples across 23 cancer types (Corces *et al.*, 2018). In this study, ATAC-seq was carried out in 9 male bladder cancer samples and identified ~100,000 chromatin-accessible peaks that were mainly located at distal intergenic and intronic regions, implicating them as enhancers (Corces *et al.*, 2018). A more recent study in mouse also included ATAC-seq on 2 bladder samples, and showed distinct clustering by principal component analysis (PCA) based on chromatin accessibility when compared to 20 other healthy tissue types (Liu *et al.*, 2019).

1.5.3.4 Approaches to infer epigenetic regulation from copy number and gene expression data in bladder cancer

To compensate for the lack of NGS-based epigenetic studies in bladder cancer, methods which use gene expression, DNA copy number, and exome sequencing data have been used to infer epigenetic regulation. For example, as part of a large study that used RNA-seq data from 8928 tumour samples to identify enhancer regions (RNA transcribed at non-coding intronic/intergenic regions inferred from high RNA-seq signal at these regions), 4,102 active enhancers and a ~12.5% increase in enhancer activation was found in 399 bladder tumour samples (Chen *et al.*, 2018). This approach is based on the rationale that active enhancers are transcriptionally "leaky", and therefore fails to identify poised enhancers and active "non-leaky" enhancers that do not transcribe eRNA (de Santa *et al.*, 2010).

A series of studies, all from the Radvanyi laboratory, attempted to compensate for a lack of histone ChIP-seq data in bladder by using copy number data and microarray expression arrays (Stransky *et al.*, 2006; Vallot *et al.*, 2011; Vallot *et al.*, 2015). By identifying differentially expressed regions that are independent of mutation and copy number

alterations in bladder tumours, the authors reasoned that such regions are epigenetically regulated (Stransky *et al.*, 2006). Using 57 bladder tumour samples, an initial study identified 28 copy number-independent differentially expressed regions, then assessed epigenetic regulation in a region of chromosome 3p22.3, where 4 genes (*VILL*, *PLCD1, DLEC1* and *ACAA1*) were downregulated. Using the bladder cancer cell line TCCSUP♀, re-expression of genes in this region could be achieved by treating cells with the HDAC inhibitor trichostatin A (TSA), but not the DNA demethylating agent 5-aza-deoxycytidine (5-aza-dC). ChIP-qPCR showed that H3K9me3 was increased at the promoters of these 4 genes in TCCSUP♀ compared to NHUC, and diminished upon TSA treatment, whereas BS-seq across this region showed no aberrations in DNA methylation (Stransky *et al.*, 2006).

A follow-up study then identified 7 more copy number-independent differentially expressed regions that were downregulated, which constituted a multiple regional epigenetic silencing (MRES) phenotype that was more associated with invasive tumours in their cohort of 57 bladder tumours (Vallot *et al.*, 2011). Again, using BS-seq and ChIP-qPCR, the authors showed that increased H3K9me3 and H3K27me3, but not DNA methylation at the gene promoters in these regions, was associated with transcriptional repression and that H3K9ac was increased at these same gene promoters in NHUC (Vallot *et al.*, 2011). Interestingly, patients with the MRES phenotype identified by Vallot *et al* were also likely to display a previously described carcinoma *in situ* gene expression signature that is associated with progression to MIBC (Dyrskjøt *et al.*, 2004).

Using FISH at 2 of the 7 MRES sites, the Radvanyi laboratory then showed that increased chromatin compaction was correlated with higher H3K27me3 and lower H3ac at these regions in CL1207 bladder cancer cells with the MRES phenotype, compared to RT112♀ cells without the MRES phenotype, or NHUC (Vallot *et al.*, 2015). Increased expression of the polycomb repressor complex 2 (PRC2) H3K27 HMT component EZH2 was also seen in bladder cancer cells and tumours with the MRES phenotype. Although inhibition of EZH2 decreased H3K27me3, it did not relieve chromatin compaction or promote gene expression, and HDACi treatment was able to relieve chromatin compaction and restore gene expression in these regions without decreasing H3K27me3. These results therefore suggest that although H3K27me3 constitutes part of the MRES phenotype, its removal is neither necessary nor sufficient for activating gene transcription or to relieve chromatin compaction (Vallot *et al.*, 2015). By using microarray gene expression analysis and copy number alteration data, this series of studies implicated heterochromatin silencing of multiple large gene regions in the tumorigenesis of MIBC, and used localised ChIP-qPCR of repressive histone marks to support their findings (Stransky *et al.*, 2006; Vallot *et al.*, 2011;

Vallot *et al.*, 2015). Nevertheless, this approach of determining epigenetically regulated regions is intrinsically limited in resolution and is therefore unable to identify epigenetic aberrations that may occur at individual loci such as at *cis*-regulatory regions.

1.5.3.5 KDM6A and the KMT2C/D COMPASS-like complex in bladder cancer

Given the high frequency of *KDM6A* mutations in bladder cancer, recent studies have aimed to demonstrate a role for the loss of KDM6A in bladder tumorigenesis. Results are often discrepant, although a consensus suggests that loss of KDM6A promotes long-term proliferation of urothelial cells (Nickerson *et al.*, 2014; Ahn *et al.*, 2016; Ler *et al.*, 2017; Kaneko and Li, 2018; Lang *et al.*, 2019). Initial results showed that KDM6A mRNA depletion in MGH-U3♂ (*KDM6A*^{WT}) enhanced anchorage independent growth and cell migration whereas the converse was true for overexpression of KDM6A mRNA in T24♀ (*KDM6A*^{mut}), but in both cases monolayer growth was not affected. Loss of KDM6A expression in MGH-U3♂ also increased subcutaneous tumour growth in mice (Nickerson *et al.*, 2014). CRISPR/Cas9 was used to produce *KDM6A* single knock-out (KO), *KDM6C* single KO, and *KDM6A*/*KDM6C* double KO HT-1197♂ (*KDM6A*^{WT}) and UM-UC-3♂ (*KDM6A*^{WT}) cells. Single KO of both *KDM6A* and *KDM6C* resulted in increased long-term proliferation, compared to WT, and a similar increase in long-term proliferation in double KO cells was seen compared to both single KO cells (Ahn *et al.*, 2016). This study suggests that the tumourigenic effect of *KDM6A* mutations are not limited to its demethylase activity, as loss its paralogue *KDM6C* which does not exhibit demethylase activity *in vivo* and only limited demethylase activities in cell-free *in vitro* studies, also promoted cell proliferation (Walport *et al.*, 2014).

The primary role of KDM6A has long been considered that of regulating H3K27me3, although studies that have assessed this in bladder cancer suggest that the primary mode of mechanism of KDM6A may not concern H3K27me3 (Ler *et al.*, 2017; Kaneko and Li, 2018; Lang *et al.*, 2019). IHC has shown a mild but non-significant increase of H3K27me3 in *KMD6A*^{mut} bladder tumours, and GSEA on RNA-seq data showed 41 pathways downregulated and 10 pathways upregulated in KMD6A^{mut} bladder tumours, suggesting increased H3K27me3 may be repressing specific pathways (Ler *et al.*, 2017). However, using knock-down (KD) or ectopic expression of KDM6A in the bladder cancer cell lines RT-4♂ (*KMD6A*^{WT}) and KU-19-19♂ (*KMD6A*^{mut}) showed only mild global changes in H3K27me3 and cell proliferation (Ler *et al.*, 2017). Furthermore, ChIP-seq for H3K27me3 showed that enrichment throughout the genome remained unchanged in KDM6A KD cells, with the exception of mild changes at the promoters of some PRC2 target loci such *PIP5K1B* and *GHR*. *KDM6A*^{mut} cell lines were sensitive to EZH2 inhibition, which only marginally

reduced cell proliferation and global H3K27me3, although a marked effect of EZH2 inhibition was seen in KDM6A-null patient-derived xenograft models (PDX) (Ler *et al.*, 2017).

A more recent study using 6 bladder cancer cell lines with different KDM6A mutation status, showed that loss of KDM6A promotes long-term proliferation but diminished colony formation in low-density seeded colony forming assays, and although differences in KDM6A protein levels were seen between these cells, there were no differences in global H3K27me2/3, H3K27ac, or H3K4me3 (Lang *et al.*, 2019). The study also showed that re-expression of KDM6A in the *KDM6A*^{mut} cells RT112♀ and VM-CUB-1♂ affected a largely distinct set of genes, although gene-set enrichment analysis showed common themes in global gene expression changes, including upregulation of extracellular structure, cell communication, and cell membrane composition and adhesion gene-sets, and downregulation of an RNA biosynthesis gene set (Lang *et al.*, 2019). Finally, the study showed that KDM6A localisation in the nucleus was dependent on KMT2C or KMT2D, as loss of both of these genes resulted in loss of KDM6A localisation in the nucleus and increased localisation in the cytoplasm, therefore implicating a role of the KMT2C/KMT2D COMPASS-like complex (Lang *et al.*, 2019).

One study showed that conditional KO of *Kdm6a* in the urothelium of female mice sharply increased incidence and mortality of BBN-induced bladder cancer, which was not seen for loss of *Kdm6a* in male mice (Kaneko and Li, 2018). Reduced expression of the TP53 target genes *CDKN1A* and *PERP* was identified in Kdm6a-deficient mice. Furthermore, ectopically expressed WT KDM6A, but not catalytically inactive KDM6A, induced expression of *CDKN1A* and *PERP* in the bladder cancer cell line UM-UC-3♂, which also correlated with decreased H3K27me3 at these loci. An analysis of 412 MIBC samples from the cancer genome atlas (TCGA) project showed that both reduced KDM6A expression and mutations in *KDM6A* were associated with reduced disease-free survival in female but not male bladder cancer patients (Kaneko and Li, 2018). This study therefore suggests that KDM6A offers protection to females through epigenetic mechanisms, including p53 target genes.

Collectively, these studies have shown that loss of KDM6A generally promotes the long-term but not short-term cell proliferation, although it is likely that mutations in *KDM6A* do not affect a common set of genes in all bladder cancers, and effects may be dependent upon the larger tumour context. Furthermore, these studies suggest that loss of KDM6A may not primarily be causing perturbations through changes in H3K27me3, although a mild dependence on EZH2 in *KDM6A*^{mut} cells and PDX models does make them more

vulnerable to EZH2 inhibition. Lastly, *KDM6A* likely influences sexual biases seen in bladder cancer as it escapes XCI, offers protection to females against MIBC through epigenetic regulation of p53 targets, and has an increased mutation frequency in female stage TaG2 stage Ta Grade 2, see page 4) NMIBC. The functional mechanism of KDM6A in bladder cancer, and how it may be promoting gender biases still remains to be determined.

These studies concerning KDM6A in bladder have primarily focussed on its role in regulating H3K27me3, but have failed to show this as a major mode of perturbation in bladder cancer. However, KDM6A constitutes part of the *Trithorax*-related (Trr) branch of the complex of proteins associated with Set1 (COMPASS)-like complex, the core catalytic components of which include the H3K4 methyltransferase KMT2C or KMT2D (also known as MLL3 and MLL2/4) (Wang *et al.*, 2017). Mutations in *KMT2C* and *KMT2D* are also commonly seen in bladder cancer, with over 70% of NMIBC having mutations in a least one component of the KMT2C/D COMPASS-like complex (Hurst *et al.*, 2017). There are 6 COMPASS family members in mammals, each with different core components and subunits, which can be divided into three groups; the SET1A/SET1B COMPASS complex which is responsible for bulk H3K4me2 and H3K4me3 throughout the genome, the KMT2A/KMT2B COMPASS-like complex which is necessary for H3K4me3 at gene-specific and bivalent promoters (poised promoters marked by H3K4me3 and H3K27me3), and the KMT2C/D COMPASS-like complexes which are essential for H3K4me1 at enhancers (Sze and Shilatifard, 2016; Meeks and Shilatifard, 2017). COMPASS-like complex mutations in bladder cancer therefore suggest perturbations in the regulation of enhancer regions. Enhancers are *cis*-regulatory regions that promote transcription by bringing transcription factors, RNAPII, and other transcriptional machinery into close proximity to target genes. Enhancers act in a cell-type and/or context-specific manner, can be located within introns or at distal intergenic regions, and act on target genes through chromatin looping (Hu and Tee, 2017). Active enhancers are typically marked by H3K4me1 and H3K27ac, and have a high degree of chromatin accessibility due to the binding of transcription factors, whereas silent enhancers are usually inaccessible and marked by H3K27me3. A third class of poised enhancers are marked by H3K4me1 and H3K27me3 and also have peaks of chromatin accessibility, although at a lower level than their active counterparts (Hu and Tee, 2017).

1.5.3.6 Other COMPASS-complex studies in bladder cancer

Only few studies have looked at other components of COMPASS in bladder, and not exclusively the KMT2C/KMT2D COMPASS-like complex. A core component of all COMPASS complexes, WDR5, is a H3K4 methyltransferase and this was shown by IHC to

be upregulated in ~70% of bladder cancer tissues compared to normal urothelial controls, and was negatively correlated with patient survival (Chen *et al.*, 2015). Gain or loss studies in UM-UC-3♂ and T24♀ bladder cancer cell lines, showed that WDR5 promotes bladder cell proliferation, self-renewal, and resistance to cisplatin. Further microarray and ChIP-qPCR analysis showed that it regulates H3K4me3 at the promoters of target genes (Chen *et al.*, 2015).

In a study comparing mutation profiles of recurrent and primary bladder tumours, the H3K4 methyltransferase *KMT2A* had a significantly increased mutation rate in recurrent tumours and this correlated with increased H3K4me3 (Wu *et al.*, 2016). CRISPR/Cas9 was then used on T24♀ cells to introduce a C4437G mutation into *KMT2A* which resulted in increased transcription and increased H3K4me3 at the promoters of *GATA4* and *ETS1*, and promoted resistance to epirubicin (Wu *et al.*, 2016).

1.6 Epigenetics and gender

Large studies that have carried out genome-wide comparisons of chromatin states between all tissues in males and females have shown that 70% of gender-associated chromatin states pertain to the polycomb-repressed heterochromatin on the X-chromosome (Ernst and Kellis, 2012; Yen and Kellis, 2015). Many gender-associated chromatin states were also identified on chrY, and only very few differences were found on autosomes. The differences seen between genders at the epigenetic level are also reflected at the transcriptional level, where the majority of gender-associated gene expression events occur on X and Y chromosomes (Deluca *et al.*, 2015; Gershoni and Pietrokovski, 2017). These results are unsurprising given XCI for dosage compensation and the lack of chrY in females.

X-chromosome inactivation (XCI) is a mechanism of silencing one of the two X chromosomes in females to enable dosage compensation to equalise X-linked expression between XY males and XX females. Initial XCI happens during development, and once established, the inactive state is inherited by all cell progeny. Therefore, XCI is a key example of mitotic epigenetic inheritance, and demonstrates that changes to the epigenome, even in early development, can be maintained throughout the lifetime of an organism. The initial event in XCI is the expression of the X-linked lncRNA *XIST*, which coats the entire X-chromosome to trigger chromosome-wide silencing and heterochromatin formation (Tukiainen *et al.*, 2017). The accumulation of XIST on X-chromosome genes promotes the expulsion of RNAPII and TFs, and can recruit protein factors such as the SMART/HDAC1-associated repressor protein (SHARP) which enables gene silencing through interactions with transcription corepressors such as NCOR1/2 and the recruitment of HDACs (McHugh

et al., 2015). Recruitment of the polycomb group complexes PRC1 and PRC2 then further the repressive state through H2AK119ub and H3K27me3 respectively, and this has been attributed to maintaining gene silencing during development prior to more fixed epigenetic silencing mechanisms in somatic cells such as DNA methylation (Galupa and Heard, 2018). Interestingly, up to 25% of X-linked genes can escape from XCI, some of which are constitutive escapees, such as *KDM5C* and *KDM6A,* which escape XCI from the outset, whilst others are facultative escapees and are later reactivated in a tissue-specific manner (Tukiainen *et al.*, 2017). However, the mechanisms by which these genes escape XCI remains largely unknown.

XCI is essential for sexual differentiation during female developmental biology. Another key epigenetic regulatory event for sexual differentiation during development is that of the Y-linked gene *SRY*, which is important in inducing testis differentiation and propagating the male-phenotype (Kuroki and Tachibana, 2018). In mice, the expression of *Syr* is finely tuned throughout development, by regulation of repressive H3K9 methylation across this locus, where H3K9 methylation is maintained by the EHT2-complex and is essential in preventing lethality in the early embryo, but demethylated by KMT3C during the sex-determination process to promote testes development (Kuroki and Tachibana, 2018)

The majority of studies that compare somatic sex differences in epigenetics relate to neurobiology and are predominantly focused on DNA modifications (McCarthy *et al.*, 2017). However, one study has reported histone modification biases in the bed nucleus of the terminal stria and preoptic area of mice (Shen *et al.*, 2015). This sexually dimorphic region of the brain was shown by ChIP-seq to have differential enrichment of H3K4me3 at 248 loci that were associated with synaptic function, mainly at TSS in females. The differential enrichment was not correlated with increased expression of these genes when tested by RT-qPCR which corroborated with an earlier finding of minimal gene expression differences in the hypothalamus of male and female mice (Rinn *et al.*, 2004). Given these results, it was hypothesised that these genes exist in a primed state of activation to allow for quick changes in gene expression (Shen *et al.*, 2015).

Two studies by Waxman *et al* reported differences in the chromatin state of male and female mouse livers, particularly at distal intergenic chromatin-accessible regions (Ling *et al.*, 2010; Sugathan and Waxman, 2013). These regions primarily had a bias for H3K4me1 and H3K27ac in male mouse liver, and were associated with FOXA pioneer factors that were proposed to facilitate male-enriched STAT5 binding. Furthermore, a previously identified set of 1000 gender-related differentially expressed genes did not display differences in chromatin state at their TSS or within their gene bodies, which indicated that their differential

regulation was driven by gender-related activation of distal enhancer regions (Ling *et al.*, 2010; Sugathan and Waxman, 2013). However, genes that were only expressed in males were repressed by H3K27me3 across the gene body in females, although this was not reciprocated at the loci of female-expressed genes in males (Sugathan and Waxman, 2013). Interestingly, a study in humans which used microarray analysis to compare transcriptional profiles between 112 male and 112 female liver samples, identified 1249 gender-associated differentially expressed genes, which are likely associated with differential chromatin states as was found in mouse (Zhang *et al.*, 2011). Of these gender-associated differentially expressed genes, 70% were upregulated in females, and although X and Y chromosomes showed typical female and male expression biases, males also showed a bias for expression of the zinc finger protein cluster on chr19. Functional enrichment analysis showed that female-associated genes were involved in chromatin and epigenetic processes, lipid metabolic pathways, and cell junctions and projections, whereas male-associated genes were related to sexual reproduction (Zhang *et al.*, 2011).

ATAC-seq was used on CD4+ T-cells from healthy donors and identified 66,344 chromatin-accessible sites (Qu *et al.*, 2015). Of these sites, 92.8% were shared between individuals and over time, therefore showing high fidelity in the regulatory landscape. Of the accessible sites that differed, ~25% were identified in the same donors but at different time points, therefore reflecting dynamic epigenetic regulation (Qu *et al.*, 2015). The remaining differential chromatin-accessible sites showed stable inter-individual differences, with 4.8% of differential peaks (0.3% of all peaks) attributed to gender. The majority of gender-associated chromatin-accessible regions were located on the X and Y chromosomes, indicative of dosage compensation, XCI, and chrY expression. Only a few autosomal gender-associated chromatin-accessible regions were located on autosomes, which also showed differential transcription factor occupancy, such as IRF family members in males and CST6 in females (Qu *et al.*, 2015).

Large-scale gene expression studies have also shown that the majority of gender-associated gene expression differences are X-linked and Y-linked (Deluca *et al.*, 2015; Gershoni and Pietrokovski, 2017). For instance, a comparison of transcriptional regulation between tissues and individuals across 29 tissue types from 175 post-mortem donors from the Genotype-Tissue Expression (GTEx) project found that the majority of gender-associated gene expression differences between tissues were located on X and Y chromosomes (Deluca *et al.*, 2015). Breast tissue had the most gender-associated gene expression differences with 715 differentially expressed autosomal genes, whereas adipose and muscle tissues, which had the second greatest number, only had 12 differentially

expressed autosomal genes (Deluca *et al.*, 2015). A later study which also used GTEx data characterised the sex-differential transcriptome across 53 tissues from 544 adult post-mortem donors (Gershoni and Pietrokovski, 2017). This study also found that mammary glands had the greatest number of gender-associated differentially expressed genes (6123 genes), followed by skeletal muscle, skin, adipose, and heart which all had over 100 gender-associated differentially expressed genes. Again, much of the sexual dimorphism was associated with the sex chromosomes. Pathway enrichment analysis showed that female-biased expression was generally associated with obesity, muscular diseases, cardiomyopathy, metabolism, and adipogenesis, whereas male-biased expression was generally associated with glucose metabolism and muscle contraction (Gershoni and Pietrokovski, 2017). Although this study did include samples obtained from bladder, these are not likely to be of the urothelium as it is known that the urothelium is lost as early as two hours after death (M.A Knowles, personal communication).

Smaller gene expression studies of male and female hypothalamus, kidney, liver, and heart from adult mice have also shown that gender-associated differentially expressed genes are mainly located on the sex chromosomes (Rinn *et al.*, 2004). Few gender-associated autosomal gene expression differences were found in kidney, liver, and heart, and these were mainly upregulation of cytochrome P450 family members in female mice, implicating differential regulation in metabolism (Rinn *et al.*, 2004; Isensee *et al.*, 2008).

1.7 Project aims and objectives

Genome-wide studies assessing histone modifications and chromatin accessibility in bladder cancer are lagging behind other fields, with very few such studies having been carried out on normal urothelium. The need to address this lack of epigenetic data is urgent, given the high mutation rate of chromatin modifying genes in bladder cancer. Furthermore, the majority of functional studies in bladder have been carried out in male cells and mice. Indeed, research in other disease types also has a bias towards the use of male samples. This issue is particularly important in bladder cancer where gender biases in risk remain largely unexplained. Interestingly, the third most commonly mutated gene in bladder cancer, *KDM6A*, is likely to intrinsically promote sexual dimorphism in normal urothelium given its ability to escape XCI, and biases in the mutation rate of *KDM6A* have also been found in female NMIBC. A mutational bias in an XCI-escapee chromatin modifier indicates possible intrinsic differences in the epigenomes of male and female urothelium. This exploratory study seeks to provide essential transcriptional and epigenomic data from normal urothelial cells to the wider research community. This will require optimising techniques such as ChIP

and ATAC-seq that have not previously been used in the TERT-NHUC cell line model. Furthermore, the project will compare male and female urothelium to identify gender-associated transcriptional and chromatin accessibility differences intrinsic to urothelium that may promote the gender biases seen in bladder cancer.

The specific aims of this project are as follows:

- To carry out gene-expression analysis using microarray chips in normal human urothelial cells NHUC, TERT-NHUC, and uncultured human urothelial cells (UHUC).
- To identify gender-associated gene expression differences between male and female normal urothelial cells, and a cohort of TaG2 NMIBC samples.
- To optimise a protocol for ATAC-seq in male and female TERT-NHUC.
- To establish a bioinformatic pipeline for the analysis of ATAC-seq data.
- To carry out ATAC-seq in TERT-NHUC and identify gender-associated chromatin accessible loci.
- To optimise a ChIP protocol in TERT-NHUC for the histone marks H3K4me1, H3K4me3, H3K27ac, and H3K27me3.

Chapter 2
Materials and Methods

2.1 Cell lines used in this study

Table 2.1 Cell lines used in this study with information including their origin, gender, and age.

Cell Line	Cell Type	Gender	Age
TERT-NHUC B (B-TERT)	Normal Ureter	Male	66
TERT-NHUC C (C-TERT)	Normal Ureter	Male	80
TERT-NHUC H (H-TERT)	Normal Ureter & Renal Pelvis	Female	62
TERT-NHUC K (K-TERT)	Normal Ureter & Renal Pelvis	Female	68
NHUC 206	Normal Ureter	Male	unknown
NHUC 258	Normal Ureter	Male	unknown
NHUC 262	Normal Ureter	Female	51
NHUC 269	Normal Ureter	Female	unknown
NHUC R630	Normal Ureter	Male	60
UHUC R657	Normal Ureter	Female	81
UHUC R658	Normal Ureter	Female	68
UHUC R661	Normal Ureter	Female	73
UHUC R660	Normal Ureter	Male	57
UHUC R664	Normal Ureter	Male	43
UHUC R684	Normal Ureter	Male	58

2.2 Preparation of uncultured human urothelial cells (UHUC) and ethics

Urological samples (renal pelvis, ureter, bladder) were transported from surgical theatre (St James's University Hospital, Leeds) in sterile universals containing 15 ml of sterile Transport Medium (Hanks Balanced Salt Solution (Sigma-Aldrich #H9269), 10mM HEPES, 1% Penicillin-Streptomycin (Sigma-Aldrich #P0781), and 20KIU/ml Aprotinin (Sigma-Aldrich #A3428). The sample was poured into a 10cm petri dish, and medium changed to remove excess blood. Adipose and connective tissue was dissected and removed using sterile

scissors and forceps. Samples were cut to a final size of ~0.5cm x ~0.5cm, then transferred to a universal with 15ml of sterile Stripper Medium (Transport Medium with 0.1% EDTA) and incubated overnight at 4°C. The sample was poured into a 10cm petri dish, and urothelium was scraped/pulled off gently using fine point forceps. Urothelial cell clumps were transferred back into the same tube used for stripping, and then pipetted up and down to disaggregate partially. Cells were pelleted by centrifugation at 500 x g for 4min and supernatant removed. Cells were then resuspended in freezing mix (Growth medium, 10% DMSO, and 10% FBS) and stored in liquid nitrogen. Clinical samples and associated clinical data were sourced from the Leeds Multidisciplinary Research Tissue Bank (Leeds East Research Ethics Committee reference: 10/H1306/7). All patients gave written informed consent.

2.3 Cell Culture for NHUC and TERT NHUC

Cell culturing protocols were carried out using aseptic techniques and performed in a Biomat class II laminar flow hood (MAT). Cells stored in cryo-vials in liquid nitrogen were rapidly thawed at 37°C and recovered into 10ml growth medium (Keratinocyte Growth Medium Kit 2 with supplements (Bovine Pituitary Extract 0.004ml/ml, Epidermal Growth Factor (recombinant human) 0.125ng/ml, Insulin (recombinant human) 5µg/ml, Hydrocortisone 0.33µg/ml, Epinephrine 0.39µg/ml, Transferrin (recombinant human) 10µg/ml, CaCl2 0.09mM)) then pelleted by centrifugation at 1000 x g. Cells were then re-suspended in appropriate volume of growth medium and plated into either 25cm^2 or 75cm^2 vented Primaria^{TM} flasks (Corning). Cells were incubated at 37°C with 5% CO_2 in a humidified incubator (Sanyo), and medium was changed every 2/3 days.

2.4 Cell passaging

Cells in 75cm^2 flasks were split at 70% to 80% confluence. Medium was aspirated and cells were washed in calcium and magnesium-free phosphate buffered saline (PBS) followed by a primary incubation in 10ml of PBS-EDTA 0.1% for 2min at 37°C, followed by 1ml of trypsin-0.02% EDTA (TV; Sigma-Aldrich #T3924) until cells detached. For NHU-TERT cells 100µl of 1x Trypsin Inhibitor (Sigma-Aldrich, #T6522) was used following trypsin treatment, and cells were then suspended in 10ml media and centrifuged at 1000 x g for 4min. For other cell lines, trypsin was inhibited by FCS in cell medium. Cells were suspended in 10ml of medium and a fraction further diluted in 15ml fresh medium in a sterile 75cm^2 flask.

2.5 Cell growth curve assay

Single-cell suspensions were prepared as described in section 2.4, and 5 x 10^4 cells were seeded into single wells of a Primaria 6 well plate (Corning). Cells were seeded to allow for triplicate reads over 12 days, and cultured as outlined in section 2.3. For counting, single-cell suspensions were prepared in 1ml of PBS, and then diluted 1:100 in triplicate into 10ml of Isoton® II Diluent (Beckman Coulter, #8448011). Cells were counted using a Beckman Coulter Z2 Particle Counter and Size Analyser, where particles between 10µm - 34µm were counted as single cells.

2.6 Protein extraction

Protein was extracted from cells in 75cm^2 flasks at 70% confluence using an extraction solution consisting of 250µl RIPA buffer (PBS, 1% Triton X100, 1mM EDTA, 0.5% sodium deoxycholate, 0.1% SDS), 6.25µl protease inhibitor cocktail (Sigma-Aldrich, P8340) and 2.5µl phosphatase inhibitor cocktail (Sigma-Aldrich, P5726,) and incubated on ice for 5min. Total cell extracts were centrifuged at 12,000 x g for 10min at 4°C, and supernatants containing protein lysates were collected and quantified by Bradford assay and stored at -80°C.

2.7 Protein quantification (Bradford assay)

Protein concentrations were measured using the Bio-Rad protein assay (Bio-Rad, #500-0006) following the manufacturer's instructions. The assay utilises the Beer-Lambert Law and therefore requires producing a standard curve of known protein concentrations, in this case Bovine Serum Albumin (BSA), to calculate unknown concentrations of protein produced from extraction. Absorbance was measured on a Bio-Rad SmartSpec Plus spectrophotometer at a wavelength of 595nm.

2.8 Acid histone extraction and purification

When analysing histones by immunoblotting, histones were extracted and purified as described by Shechter *et al,* 2007. Roughly 5 x 10^6 cells were trypsinised to produce single-cell suspensions in PBS, and pelleted in 1.5ml tubes by centrifugation at 300 x g at 4°C for 5min. Cell pellets were resuspended in 1ml of ice-cold hypotonic lysis buffer (1mM KCl, 1.5mM $MgCl_2$, 1mM DTT, 10mM Tris-HCl pH 8.0, with the addition of protease and phosphatase inhibitors (Sigma-Aldrich P8340, and Sigma-Aldrich P5726 respectively) just

before use) and incubated on a rotator at 4°C for 30min. Intact nuclei were pelleted by centrifugation at 10,000 x g for 10min at 4°C and supernatant discarded. Nuclei pellets were re-suspended in 400µl 0.2M H_2SO_4 and rotated for 30min at 4°C. Nuclear debris was then pelleted by centrifugation at 16,000 x g for 10min at 4°C and the supernatant transferred to a fresh 1.5ml tube. 132µl of TCA was added drop-wise to supernatant, followed by overnight incubation on ice. Histones were then pelleted by centrifugation at 16,000 x g for 10min at 4°C and supernatant discarded. Histone pellets were gently washed with 400µl ice-cold acetone without disturbing the pellet and then centrifuged again at 16,000 x g for 10min at 4°C and supernatant discarded (this step was then repeated). Pellets were air-dried at room temperature (RT) for 20min followed by re-suspension in 50µl dH_2O. Histones were quantified by measuring absorbance on a NanoDrop™ 8000 UV-Vis Spectrophotometer (Labtech) at a wavelength of 230nm, where a 1:10 dilution of 1µg/µl histones gives an OD of 0.42.

2.9 Western blot analysis

20µl of 1.5µg/µl total protein in PBS and 5µl 5X WB-loading buffer (0.02% bromophenol blue, 1% SDS, 30% glycerol, 250mM Tris-HCl pH 6.8, and freshly added 0.7M 2-mercaptoethanol (βME)) was denatured by heating at 100°C for 5min. Protein was then separated by SDS-PAGE using 10% or 12% polyacrylamide mini-PROTEAN® gels (Bio-Rad, #4568033/#4568044) in TGS buffer (Bio-Rad) at 3W per gel, and transferred to 0.2mm nitrocellulose membrane (Bio-Rad #170-4159) using Trans-Blot® Turbo™ Transfer System (Bio-Rad). Membranes were blocked with 5% BSA or 4% milk-powder in PBS-Tween20 0.1% (PBS-T) for 1hr at RT, then incubated overnight at 4°C with diluted primary antibody (see Table 2.4) in 1% BSA or 2% milk in PBS-T. Following primary antibody incubation, membranes were washed 4 x 10min in PBS-T. Membranes were then incubated with 1:5000 HRP-conjugated anti-rabbit (Southern Biotech #4010-05) or anti-mouse (Bio-Rad #170-6516) secondary antibodies for 1hr at RT, then washed again with PBS-T. All incubation and wash steps were carried out on a shaking platform. Proteins were visualised using Luminata Forte HRP Substrate (Millipore #WBLUF0500) and ChemiDoc MP System and Image Lab software (Bio-Rad). Membranes were stripped at 55°C for 45min in stripping buffer (10M urea, 0.1M Tris-HCl pH 6.8) then washed 4 x 10min PBS-T, blocked and then re-hybridised to examine expression of other proteins. See Table 2.4 for list of antibodies and concentrations.

2.10 Chromatin immunoprecipitation (ChIP)

Approximately 5 x 10^6 cells were trypsinised and re-suspended in medium and then cross-linked in 1% formaldehyde and incubated on a rotator at RT for 10min. Glycine was added to a final concentration of 0.125M and incubated for 10min at RT to stop fixation and quench unreacted formaldehyde. Cross-linked cells were washed twice in ice-cold PBS and then suspended in ChIP-Lysis Buffer (1% SDS, 10mM EDTA, 50mM Tris-HCl pH 8.0, protease inhibitors). 5 x 10^6 lysed cells were sonicated using a Bioruptor® (Diagenode) in 300ul ChIP-Lysis Buffer with cycles of 30s on/off on HIGH to produce fragments between 100-400bp. Chromatin samples were then clarified by centrifugation at 16,000 x g and supernatants were transferred to a clean 1.5ml tube and diluted with 300µl ChIP-IP Buffer (2mM EDTA, 150mM NaCl, 1% Triton X-100, Tris-HCl pH 8.0). Chromatin was quantified using Nanodrop™ and for each immunoprecipitation (IP), 100µg of chromatin was aliquoted and adjusted to a volume of 500µl with ChIP-IP buffer. Chromatin was pre-cleared by incubating with 50µl of washed protein A magnetic beads (Bio-Rad #161-4023) for 2hrs at 4°C on a rotator. Prior to IP a fraction of the pre-cleared chromatin was kept aside to use as an input control. Chromatin samples were then incubated overnight at 4°C on a rotator with 4µl/10µl of respective antibody (anti-H3 (Abcam #ab1791), anti-H3K4me1 (Abcam #ab8895), anti-H3K4me3 (Cell Signaling #9751), anti-H3K27me3 (Cell Signaling #9733), anti-H3K27ac (Active Motif #39685), or anti-IgG (Santa-Cruz Biotechnology #sc-2027), followed by another 4°C incubation for 2hrs with 100µl Protein-A magnetic beads (Bio-Rad #161-4023). The immunocomplexes then underwent a series of ice-cold washes including 4 x 1ml washes with RIPA buffer (1mM EDTA, 0.5mM EGTA, 140mM NaCl, 1% Triton X-100, 0.1% NaDOC, 0.1% SDS, 10mM Tris-HCl pH 8.0), 2 x 1ml washes with LiCl buffer (1mM EDTA, 0.17M LiCl, 0.5% NP40, 0.5% NaDOC, 10mM Tris-HCl pH 8.0), and 2 x 1ml washes with TE buffer (1mM EDTA, 10mM Tris-HCl pH 8.0). The immunocomplexes were then eluted using 500µl of Elution buffer (0.1M $NaHCO_3$, 1% SDS) for 30min at RT on a rotator. Both the immunocomplexes and input were then incubated with 10µg RNase (Invitrogen, #12091-021) for 1hr at 37°C. NaCl was added to a final concentration of 0.1M along with 10µg Proteinase K (Sigma-Aldrich, #P2308), and a final overnight incubation at 65°C was carried out to reverse crosslinks. DNA was purified using phenol-chloroform extraction.

2.11 Phenol-Chloroform purification of DNA

An equal volume of phenol:chloroform (Sigma-Aldrich, #P2069) was added to DNA-containing solution and then vigorously shaken for 5min followed by centrifugation at 16,000 x g for 5min at RT. The aqueous phase was then transferred to a fresh 1.5ml tube and an equal volume of chloroform added. Tubes were shaken vigorously for 5min and then centrifuged at 16,000 x g for 5min at RT. The aqueous phase was transferred to a fresh 1.5ml tube and 20µg of glycogen (Sigma-Aldrich, #G1508) and 50µl of 3M sodium acetate pH 5.2 added. 1ml of ice-cold 100% Ethanol (EtOH) was added and then incubated at -20°C for 45min. DNA was pelleted by centrifugation at 16,000 x g for 20min at 4°C. The EtOH was then aspirated and the pellet washed with 1ml ice-cold 70% EtOH, then centrifuged at 16,000 x g for 15min at 4°C. All EtOH was aspirated and DNA pellet air-dried for 10min at RT. DNA was re-suspended in 50ul dH_2O, or for ChIP, in 50µl 0.1x TE pH 8.0.

2.12 RNA extraction and cDNA synthesis

RNA was extracted from cell lines as in section 2.13, purified using a RNeasy Mini Kit (Qiagen, #74106) according to the manufacturer's instructions and quantified by measuring absorbance at 260nm using NanoDrop™. Reverse transcription was carried out by incubating 1-5µg RNA with 0.5mM random hexamer primers (Life Technologies) and 0.625mM dNTP at 65°C for 5min, then cooled to 4°C. The RNA was then treated with 10mM DTT and 10U/µl SuperScript® II Reverse Transcriptase with respective buffer (Invitrogen, #100004925) and then incubated at 42°C for 50min, followed by 70°C for 15min and then 4°C for 5min. Samples were then diluted 1:10 with dH_2O and stored at -80°C.

2.13 RNA preparation for microarrays

RNA was collected from 1 x T75cm^3 flask by washing cells with ice-cold PBS and then lysing with 350µl of RLT buffer (Qiagen, #74106) with βME. RNA was purified using Qiagen RNeasy® Mini Kit (Qiagen, #74106) spin columns according to the manufacturer's instructions then quantified using Nanodrop™. RNA samples then underwent DNase treatment (Invitrogen, #18068015) by incubating up to 10µg of RNA with 40U RNasin, 1x digest buffer and 3U DNase I, adjusted to a total volume of 40µl with RNase-free dH_2O, then incubated at RT for 15min. DNase reactions were then stopped with the addition of 4µl of 25mM EDTA. 64µl of RNase-free dH_2O was added to each RNA sample followed by 350µl of RLT with βME and 250µl of absolute EtOH, then once again purified using the

Qiagen spin column. RNA was eluted in 30µl of RNase-free dH_2O. Samples then underwent quantification and quality checks by Nanodrop™ and TapeStation (Agilent) analysis, where RNA was considered for microarrays once a 260/280nm absorbance ratio >2 was measured by Nanodrop™, and a RIN value >8 was measured by TapeStation. An example RNA profile following Nanodrop™ and TapeStation analysis can be found in Appendix A. Microarrays for NHU-TERT cells were carried out in biological triplicate (from three independent flasks of cells).)

2.14 Microarray procedure – conducted by Affymetrix (Thermo Fisher, UK)

Total RNA was amplified and subject to reverse transcription using the Affymetrix GenChip® WT PLUS Reagent Kit according to the manufacturer's instructions. The resulting cDNA was quantified using Nanodrop™, then normalised and hybridised onto an Affymetrix Human Transcriptome 2.0 microarray chip for 16 hours at 45°C. Microarrays were then washed and stained using the Affymetrix GeneChip® Hybridisation Wash and Stain Kit according to the manufacturer's instructions and using the Affymetrix GeneChip® Fluidics station 450. Microarrays were scanned using the Affymetrix GeneChip® 7G microarray scanner. Quality assessment was carried out using the Affymetrix Expression Console Software.

2.15 Analysis of microarray data

Quality assessment, normalisation, and probe summarisation of microarray data was carried out using the Affymetrix Expression Console. Normalisation and summarisation were carried out using the Signal Space Transformation – Robust Multi-chip Average methods (SST-RMA) and used to convert raw CEL files into CHP files. SST normalises data by adjusting probe intensity levels and using GC4 to reduce background noise through GC count levelling, and the RMA algorithm is used to carry out summarisation and fit a robust linear model at the probe level to minimise the effect of probe-specific affinity differences (Irizarry *et al.*, 2003). The series of quality control tests include: histograms and boxplots on signal intensity of all probes for each sample, line graphs for hybridisation controls (spike-controls, positive housekeeping exon vs negative housekeeping intron controls), principal component analysis (PCA) and Pearson's Correlation coefficient between samples. The Affymetrix Transcriptome Analysis Console was then used to carry out an unpaired Analysis of Variance (ANOVA) to determine which genes had undergone changes in expression

between groups. Differential expression analysis was also conducted using Linear Models for Microarray Data (LIMMA) analysis in R using the "limma", "oligo", and "affy" packages downloaded from Bioconductor (Ritchie *et al.*, 2015). For both ANOVA and LIMMA analysis, genes were considered differentially expressed if they exhibited a fold change ≥1.5, and a p-value of ≤0.05. Gene Set Enrichment Analysis (GSEA) was carried out using the GSEA v3.0 software downloaded from the Broad Institute and using gene sets obtained from the molecular signatures database (MSigDB v6.1). Metacore™ was used to also examine differentially altered pathways between experimental groups.

2.16 PCR and agarose gel electrophoresis

Gel electrophoresis was carried out using 0.8-1.2% agarose gels (depending on the size of DNA fragments) cast using 1x TBE (Alfa Aesar) with 0.6μg/ml ethidium bromide. 500ng 100bp DNA ladder was used in the first lane to visualise size of DNA fragments. Samples were prepared with 1μg of DNA loaded with 2ul gel loading dye (New England Biolabs, #B70215) and adjusted to 12μl with H_2O. 10μl of each sample was loaded per well. For samples with low concentrations of DNA, 1x SYBR Green (Invitrogen, #S7563) was also added and ethidium bromide was not used in the gel. Gels were electrophoresed in 1x TBE buffer at 80-120V (depending on the size of the gel) and imaged using Bio-Rad ChemiDoc MP System and Image Lab software (Bio-Rad)

2.17 Primer design and testing

Primers were designed using NCBI Primer-BLAST and using Thermodynamic Template Alignment and purchased from Sigma-Aldrich. An appropriate FASTA sequence was submitted that encompassed the DNA/RNA region of interest and primers were selected to ensure a product size of 100-250bp, a primer melting temperature of 58-62°C, a primer GC content of 40-60%, a minimum primer length of 18bp, a self-complementarity score of 0, and no possible off-target amplicons. Primers designed to test gene expression were targeted at exon-exon junctions, designed to span more than 1 exon, and to only encompass exons that are expressed in all mRNA isoforms of that gene in order to avoid biases resulting from alternative splicing and DNA contamination.

Primer-pair efficiency and specificity were tested using qPCR (section 2.19) to produce an efficiency plot of Ct (cycle threshold) vs DNA concentration. A serial dilution of gDNA/cDNA (500ng, 250ng, 125ng, 62.5ng, and 31.25ng) was used to produce the efficiency data, and H_2O was used as a negative control to show no primer-dimer

amplification and contamination. 500ng gDNA was used as a negative control for primers assessing mRNA expression. Primer efficiency (E) was calculated using the gradient of the standard curve produced from the efficiency plot, and the equation $E = 10^{(-1/Gradient)} - 1$, where an E value of between 0.85-1.15 was considered as a pass for that primer pair. Example efficiency plots and melt curves for a successful and unsuccessful primer pair can be seen in Figure 2.1. Efficiency plots for primer pairs used in this study can be found in the Appendix.

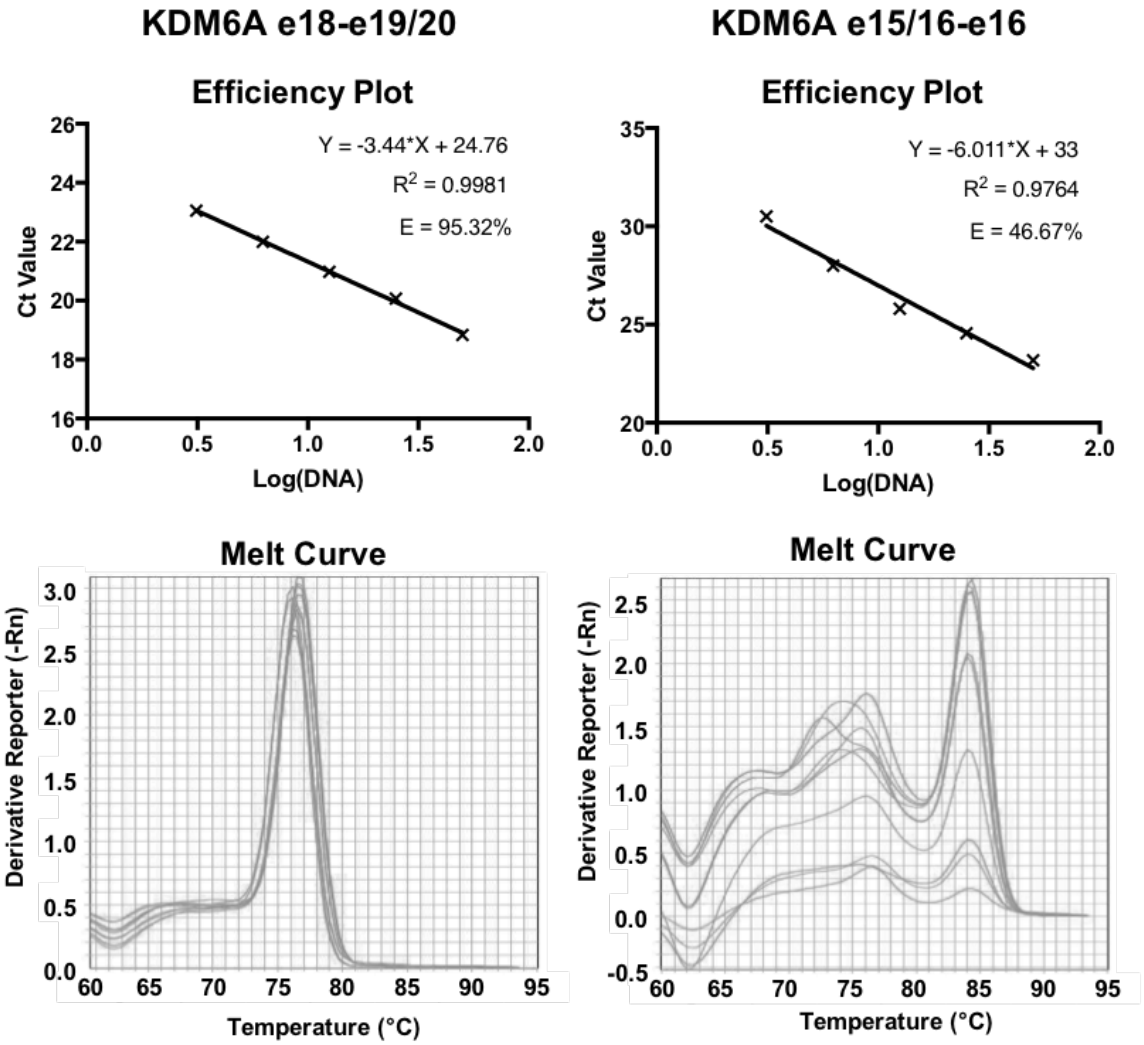


Figure 2.1: Example efficiency plots and melt curves.

The KDM6A e18-e19/20 primer pair shows a good E value of 95.32% (top left) and a single peak on the melt curve (bottom left) and is therefore considered appropriate for further use in qPCR analysis. The KDM6A e15/16-e16 primer pair shows a low E value of 46.76% (top right) and multiple peaks on the melt curve therefore suggesting primer dimerization or off target amplification (bottom right), and is therefore not appropriate for use in qPCR analysis.

2.18 Primers used in this study

All primers used in this study were purchased from Sigma-Aldrich as dry desalted oligos. Upon delivery oligos were resuspended in PCR-grade H_2O in a PCR cabinet to produce an oligo concentration of 100μM and left to stand for 30min. A fraction was then diluted to 25μM to be used for PCR experiments. Primers were stored at -80°C.

Table 2.2 Primer oligos used in this study and their targets

Gene	Application	Target	Direction	Oligo Sequence (5'→3')
SDHA	ChIP & ATAC qPCR	promoter	Forward	CTTCGGTCTGGGCGATCC
		promoter	Reverse	GACGGTGGCGTTAAGGGAA
	ATAC qPCR	exon9	Forward	CCTCCCCACCGTGCATTATA
		intron9	Reverse	TCTAAAGAGACAACTGCGAGGT
GAPDH	ATAC qPCR	promoter	Forward	TCTGCTGAGTCACCTTCGAAC
		promoter	Reverse	CATTACTGTCTTCTCCCCGCA
	ATAC qPCR	exon5/intron5 junction	Forward	GAGTCCACTGGCGTCTTCAC
		intron5/exon6 junction	Reverse	CCCTGCAAATGAGCCTACAG
MYT-1	ChIP qPCR	intron1	Forward	GGAGAGTGGATCCCGGTTTT
		intron1	Reverse	TGCAGACGACAATTAGGGCC
ACTBL2	ChIP qPCR	enhancer	Forward	CACACAAAAGTAAGGCCATGT
		enhancer	Reverse	CAGCCTGCCTCAATAGTACAA
KDM6A	mRNA qPCR	exon18	Forward	TTCACCATACCCTCCCTTGC
		exon19/20 junction	Reverse	AGAAAAGTCCCAGGTCTAACTTAA
EZH2	mRNA qPCR	exon8	Forward	CCTCCTGAATGTACCCCCAAC
		exon8/9 junction	Reverse	TGAAAAGGATGTAGGAAGCAGTC

2.19 Quantitative polymerase chain reaction (qPCR)

Relative mRNA expression and ChIP enrichment was determined using SYBR™ Green Master Mix (Applied Biosystems, #4309155) with primers from Table 2.2. Reactions were carried out with 0.3μM of each primer, 2μl of cDNA/ChIP DNA, and 2X SYBR™ Green buffer to a total volume of 20μl. The PCR running protocol was as follows: 50°C for 2min, 95°C for 10min, and 40 cycles of 95°C for 15s and 60°C for 1min. Following the PCR reaction a melt curve was generated by measuring absorbance between 60°C to 95°C at a 1% temperature increase. Reactions were performed in duplicate/triplicate using a QuantStudio 5 Real-Time PCR System (ThermoFisher). mRNA expression was normalised to *SDHA* (succinate dehydrogenase complex flavoprotein subunit A) using ΔCt method and quantified as fold-enrichment over a normal human urothelial cell line.

2.20 Micrococcal Nuclease (MNase) Digestion Assay

MNase Digestion Assays were performed on NHU-TERT cells over 11 different time points. 10 x 75cm^2 flasks of 80% confluence were required for each assay. Cells were trypsinised and pelleted as described in section 2.4 then re-suspended in 5ml of ice-cold Nuclear Buffer A (85mM KCl, 0.5mM spermidine, 0.2mM EDTA, 250μM PMSF, 5.5% sucrose, 10mM Tris-HCl pH 7.6) and mixed well. 5ml of ice-cold Nuclear Buffer B (85mM KCl, 0.5mM spermidine, 0.2mM EDTA, 250μM PMSF, 5.5% sucrose, 0.1% NP-40, 10mM Tris-HCl pH 7.6) was added and mixed thoroughly then incubated on ice for 5min. Nuclei were then pelleted by centrifugation at 1000 x g for 5min at 4°C. Nuclear pellets were resuspended in 10ml ice-cold Nuclear Release Buffer (85mM KCl, 1.5mM $CaCl_2$, 3mM $MgCl_2$ 250μM PMSF, 5.5% sucrose, and 10mM Tris-HCl pH 7.6) and immediately pelleted by centrifugation at 1000 x g for 5min at 4°C. Pellets were then re-suspended in 500μl Nuclear Release Buffer and kept on ice. Nuclei concentration was measured by Nanodrop™ at A_{260}, and 10ug of nuclei distributed into fresh 1.5ml tubes (one for each time point), and adjusted to a total volume of 500ul with Nuclear Release Buffer and addition of 20ug RNase. 100U of MNase (ThermoFisher, #88216) was then added to each sample and incubated at RT for the required amount of time for the assay. Digestion was stopped with the addition of 300ul of Genomic Lysis Buffer (300mM NaCl, 20mM EDTA, 1% SDS). Samples were then incubated overnight at 55°C with 10μg of Proteinase K. DNA was then purified using phenol:chloroform and suspended in 20ul dH_2O as described in section 2.11. DNA concentration was determined by Nanodrop™ and visualised by gel electrophoresis.

2.21 Guava Cell Cycle Analysis

Cells were plated in 75cm^2 flasks such that upon collection they were either at 70% confluence or full confluence, harvested under normal conditions and counted. 5 x 10^5 cells were pipetted onto a chilled round-bottomed 96-well plate and pelleted by centrifugation at 450 x g for 10min at 4°C with brake on low. Cells were then washed twice by resuspending in 200µl ice-cold PBS and pelleted by centrifugation at 450 x g for 10min at 4°C with brake on low. Supernatants were gently removed and cells were then fixed by resuspending in 200µl 70% EtOH and incubated whilst shaking at low speed at 4°C for >1hr. During this time Guava Cell Cycle Reagent (Millipore #4500-0220) was warmed to RT. Cells were pelleted by centrifugation at 450 x g for 10min at 4°C and supernatant was discarded, and then washed in 200µl ice-cold-PBS and again pelleted with supernatant discarded. Cell pellets were resuspended in 200µl of Guava Cell Cycle Reagent and incubated in the dark for 30min at RT. Samples were then acquired on a Guava EasyCyte System (Millipore) and analysed using FCSalyzer.

2.22 Assay for Transposase Accessible Chromatin (ATAC-seq)

Library preparation was essentially carried out as described by Buenrostro *et al* with minor modifications (Buenrostro *et al.*, 2013). 2x10^5 NHU-TERT cells were pelleted by centrifugation at 500 x g for 5min at 4°C. The cell pellet was re-suspended in ice-cold lysis buffer (10mM NaCl, 3mM $MgCl_2$, 0.1% IGEPAL, 10mM Tris-HCl pH 7.4) by gently pipetting up and down 5 times followed by a light vortex and repeating. Lysed cells were pelleted by centrifuging at 500 x g for 10min at 4°C. The supernatant was removed and nuclei were re-suspended in 25µl H_2O and counted using a haemocytometer. 1x10^5 nuclei were used for the digestion reaction by incubating at 37°C for 30min with transposase reaction mix (2.5µl Tn5 transposase and 25µl TD buffer, Illumina #15028212) and H_2O to a final volume of 50µl. Following transposition, the sample was purified using Qiagen MinElute kit (#28004) according to the manufacturer's instructions and eluted in 10µl H_2O. Following purification, ATAC libraries were amplified by PCR. Reactions were carried out with 10µl transposed DNA, 1.25µM ATAC primer 1, 1.25µM ATAC primer 2 (barcoded, Table 2.3), and 25µl NEBNext High-Fidelity PCR mix (NEB, #M0541), and made to a final volume of 50µl. with H_2O. Cycling conditions were as follows: 98°C for 30sec, n x (98°C for 10sec, 63°C for 30sec, 72°C for 1min), hold 4°C, where 'n' is the number of PCR cycles, unique for each sample. To determine number of additional PCR samples required to amplify each ATAC library, an initial amplification of 5 cycles was followed by qPCR with the addition of

SYBR-green (Invitrogen, #S7563), and the number of additional cycles was calculated by taking the cycle number at which 1/3 of maximum fluorescence (Rn) occurred. The total number of cycles varied with each sample and ranged between 4-9. Libraries were then size-selected and purified to remove PCR by-products and select fragments below 1000bp. ATAC libraries were gently mixed with 0.5x volume of AMPure beads and incubated for 10min at RT. Using a magnetic rack, the supernatant was removed, mixed with 1.8X post-PCR volume of AMPure beads and incubated for 10min at RT. Using a magnetic rack, the supernatant was discarded and beads were washed with 80% ethanol. After removal of all ethanol, beads were air-dried for 10min at RT before finally being re-suspended in 20µl H_2O. Purified libraries were stored at -80°C until further use. Quantification of ATAC libraries was carried out using KAPA Biosystems library quantification kit for Illumina (#KK4873) according to the manufacturer's instructions. Libraries were quality checked using TapeStation and then pooled for sequencing. Sequencing was carried out using the Illumina NextSeq500 with 75bp paired-end reads at the Leeds university sequencing facilities at St James's University Hospital.

Table 2.3 Indexing sequences used for each ATAC-seq library sample

Library	Index Sequence
ATAC primer Ad1	AATGATACGGCGACCACCGAGATCTACACTCGTCGGCAGCGTC AGATGTG
B-TERT Ad2.1	CAAGCAGAAGACGGCATACGAGATTCGCCTTAGTCTCGTGGGC TCGGAGATGT
B-TERT Ad2.5	CAAGCAGAAGACGGCATACGAGATAGGAGTCCGTCTCGTGGG CTCGGAGATGT
C-TERT Ad2.2	CAAGCAGAAGACGGCATACGAGATCTAGTACGGTCTCGTGGG CTCGGAGATGT
C-TERT Ad2.6	CAAGCAGAAGACGGCATACGAGATCATGCCTAGTCTCGTGGGC TCGGAGATGT
H-TERT Ad2.3	CAAGCAGAAGACGGCATACGAGATTTCTGCCTGTCTCGTGGGC TCGGAGATGT
H-TERT Ad2.7	CAAGCAGAAGACGGCATACGAGATGTAGAGAGGTCTCGTGGG CTCGGAGATGT
K-TERT Ad2.4	CAAGCAGAAGACGGCATACGAGATGCTCAGGAGTCTCGTGGG CTCGGAGATGT
K-TERT Ad2.8	CAAGCAGAAGACGGCATACGAGATCCTCTCTGGTCTCGTGGGC TCGGAGATGT

2.23 Analysis of ATAC-seq Data

ATAC-seq data was analysed using an in-house pipeline that includes the following tools: FastQC v0.11.5, TrimGalore v0.4.4, Bowtie2 v2.3, Samtools v2.26.0, Bedtools v2.26.0, Picard 2.9.2, Deeptools v2.4.0, Macs2 v2.1.1.20160309, and R v3.4. Briefly, the pipeline includes an initial quality assessment using FastQC followed by adapter trimming using TrimGalore with parameters -a CTGTCTCTTATACACATCT -A CTGTCTCTTATACACATCT -q 30 --minimum-length 20. Reads are aligned to the hg19 reference genome using Bowtie2 with "-very-sensitive" parameters, a maximum fragment length of 2000bp and ensuring a high mapping quality (MAPQ score >30), and then filtered to remove mitochondrial DNA and black-listed regions. Duplicate reads are removed using Picard tools. Peaks are called using Macs with parameters –nomodel –shift -100 --extsize 200 for ATAC-peaks, and annotated in R using ChIPpeakAnno where intergenic peaks are assigned to genes that fall within 100kb from the TSS. Differential binding analysis is carried out using DiffBind, and Bedtools is used to find overlapping peaks. Samtools is used for file format conversions.

2.24 TapeStation Analysis

Following ATAC-seq library preparation (section 2.22) and RNA preparation for microarray analysis (section 2.13), DNA and RNA was subjected to quality assessment using TapeStation analysis. Samples were run on the Agilent Technologies 2200 TapeStation and analysed using the 2200 TapeStation Software Version A.01.04.

2.24.1 TapeStation analysis of DNA

For DNA TapeStation analysis, 2µl of library sample, H_2O, or High Sensitivity D1000 DNA Ladder (Agilent Technologies, #5067-5587) was added to 2µl of High Sensitivity D1000 Sample Buffer (Agilent Technologies, #5067-5585), and run on the TapeStation using a High Sensitivity D1000 ScreenTape (Agilent Technologies, #5067-5584).

2.24.2 TapeStation analysis of RNA

For RNA TapeStation analysis, 1µl of RNA sample or RNA ScreenTape Ladder (Agilent Technologies, #5067-5578) was incubated with 5µl of RNA ScreenTape Sample Buffer (Agilent Technologies, #5067-5577) for 3min at 72°C and then immediately cooled on ice. Samples were run on the TapeStation using the RNA ScreenTape (Agilent Technologies, #5067-5576).

2.25 Antibodies

Table 2.4 List of antibodies used in this study

Target	Company	Catalogue No.	Host	Clonality	Technique	Dilution
DICER	Cell Signaling	#5362	Rabbit	mAb	WB	1:1000
EGFR	Bethyl Labs	A300-388A	Rabbit	pAb	WB	1:3000
H3K27ac	Active Motif	#39685	Mouse	mAb	WB/ChIP	1:1000/ 1:55
H3K27me3	Cell Signaling	#9733	Rabbit	mAb	WB/ChIP	1:1000/ 1:55
H3K4me1	Abcam	ab8895	Rabbit	pAb	ChIP	1:140
H3K4me3	Cell Signaling	#9751	Rabbit	mAb	WB/ChIP	1:1000/ 1:140
Histone H3	Abcam	ab1791	Rabbit	pAb	WB/ChIP	1:10000 /1:140
IgG	Santa Cruz	sc-2027	Rabbit		ChIP	1:140
α-Tubulin	Bio-Rad	MCA77G	Rat	mAb	WB	1:2000
β-Actin	Santa Cruz	sc-81178	Mouse	mAb	WB	1:5000
Anti-Mouse IgG - HRP Conjugate	Bio-Rad	#1706516	Goat		WB	1:5000
Anti-Rabbit IgG - HRP Conjugate	Southern Biotech	#4010-05	Goat		WB	1:5000

2.26 Suppliers

Company details of suppliers used in this study are outlines in Table 2.5. Unless otherwise stated in the methodology, reagents were purchased from Thermo Fisher Scientific.

Table 2.5 Company information of suppliers used in this study

Company	Address
Abcam	Cambridge, UK
Active Motif	Carlsbad, California, USA
Affymetrix	Santa Clara, California, USA
Agilent Technologies	Santa Clara, California, USA
Alfa Aesar	Burlington, Massachusetts, USA
Applied Biosytems	Foster City, California, USA
Beckman Coulter	Brea, California, USA
Bio-Rad	Hercules, California, USA
Cell Signalling Technologies	Danvers, Massachusetts, USA
Corning	Corning, New York, USA
Diagenode	Rue Bois Saint-Jean, Seraing, Belgium
Illumina	San Diego, California, USA
Invitrogen	Carlsbad, California, USA
KAPA biosystems	St. Louis, Missouri, USA
Life Technologies	Carlsbad, California, USA
Merck	Darmstadtm Germany
Metacore	London, UK
Millipore	Burlington, Massachusetts, USA
New England Biolabs	Ipswich, Massachusetts, USA
Qiagen	Venlo, Netherlands
Santa-Cruz Biotechnology	Dallas, Texas, USA
Sigma-Aldrich	St. Louis, Missouri, USA
Southern Biotech	Birmingham, Alabama, USA
Thermo Fisher Scientific	Waltham, Massachusetts, USA

Chapter 3
Gender-related differences in the transcriptome of Normal Human Urothelial Cells (NHUC) and TaG2 non muscle-invasive bladder cancer (NMIBC) tumours

3.1 Introduction

Despite the well described gender bias observed in the incidence of bladder cancer, few studies regarding gender differences at the transcriptomic and epigenetic level of healthy and diseased bladder have been carried out. Large consortium efforts such as FANTOM, ENCODE and the GTEx projects which aim to profile the transcriptomes and epigenomes of all major tissue and cell types, contain little data for bladder (ENCODE Consortium, 2012; Forrest *et al.*, 2014; Deluca *et al.*, 2015). This study aimed to assess gender-related transcriptomic and epigenetic differences in the bladder. In this chapter, microarray analysis was used to assess three types of normal human urothelial cells (NHUC) (including primary NHUC, telomerase-immortalised NHUC (TERT-NHUC) and uncultured normal urothelial cells (UHUC)), and a cohort of 102 stage Ta grade 2 (TaG2) non-invasive bladder tumours.

A previous study that assessed the sex-differential transcriptome of 53 human tissues from post-mortem donors using data acquired from the GTEx project included 6 male and 5 female bladder samples (Deluca *et al.*, 2015). However, separate studies on recently deceased donors have shown that the bladder urothelium is lost as early as 2 hours after death (M. Knowles, personal communication), and it is therefore unlikely that the bladder samples collected in the GTEx project are from the urothelium but instead from the underlying detrusor muscle. Nevertheless, Deluca *et al* identified only 10 differentially expressed (DE) genes (where DE was defined as $P \leq 0.05$ and $FC \geq 1.1$), all of which were located on chrY and were DE in at least 44 other tissue types (Gershoni and Pietrokovski, 2017). Interestingly, no genes that commonly escape X-chromosome inactivation (XCI) were identified in the analysis for bladder. These results therefore suggested that there is no gender bias at the transcriptomic level that is unique in healthy bladder tissue. Another study in mice assessed differential gene expression in the bladder detrusor between mature and aged mice and compared how this differed between genders (Kamei *et al.*, 2018). The authors identified a total of 1480 DE genes (where DE was defined as $P \leq 0.05$ and $FC \geq 2$) between the bladders of mature and aged mice. However, when separated into gender groups, only 45

DE genes were identified in both male and female aged comparisons (Kamei *et al.*, 2018). These results indicate a differential response to ageing in the bladders of male and female mice, although this may be attributed to alterations in sex hormone status as a result of ageing. Both of the aforementioned studies were limited to the bladder detrusor muscle, and therefore the transcriptional comparison between genders in this study is likely the first to be done using urothelial samples. One may hypothesise minimal differences in the expression profiles of healthy male and female urothelium, with any DE genes being located on the sex chromosomes. However, given the differential response to ageing in the detrusor of male and female mice, it may be expected that each gender responds differently upon acquisition of mutations in the urothelium in bladder cancer.

Genome-wide transcriptional analysis is commonly carried out using microarray chips or RNA-seq, and is used to assess differential gene expression between biological groups. This study used the Affymetrix GeneChip™ Human Transcriptome Array 2.0 (HTA 2.0), a high-resolution microarray composed of over 6 million distinct probes that target exons and splice junctions to enable the transcriptional profiling of all known gene transcripts (Xu *et al.*, 2011). Each exon is targeted by ~10 pbobes, and each splice-site by ~4 probs, and are then arranged into probe-sets to summarize the data into gene-level, exon-level, and splice-junction probe-sets. This study concerns the gene-level probe set for gene transcription, with each probe-set being referred to as individual probes hereafter. Genome-wide gene expression was initially compared between two male and two female TERT-NHUC lines. This would provide differential gene expression and functional enrichment data to complement later ATAC-seq and ChIP-seq analyses in these cell lines. Although an initial assessment of gender-associated differential gene expression was made at this stage, further analyses to complement these results included using non-immortalised NHUC and a separate analysis using uncultured healthy urothelial cells (UHUC). This Chapter also takes advantage of microarray data generated in our laboratory from 102 TaG2 tumour samples to identify gender-associated DE genes that may persist or arise during the development of bladder cancer, and assess whether gender groups show different differentially expressed genes in samples with mutations in *KDM6A* and other commonly mutated genes in bladder cancer.

3.2 Results

3.2.1 Microarray analysis of TERT-NHUC

The majority of this study concerns mapping the epigenetic landscape of TERT-NHUC by ATAC-seq and ChIP-seq. To complement the next generation-sequencing (NGS) data, microarray analysis was carried out to ascertain the transcriptional profiles of these cells as well as determine gender-associated differences at the transcriptomic level. RNA was collected in biological triplicate from two male (B & C) and two female (H & K) TERT-NHUC lines then sent to Tepnel Pharma Services (TPS) for hybridisation onto HTA 2.0 and generation of data as CEL files. Differential expression (DE) analysis was carried out using the Linear Models for Microarray Data Approach (LIMMA), and differentially enriched pathways were determined using Gene-Set Enrichment Analysis (GSEA) using the gene-sets curated from the GO, KEGG, Reactome, and Hallmarks databases.

3.2.1.1 Quality assessment of TERT-NHUC RNA samples prior to microarray

Following RNA purification samples underwent quality assessment (QA) and quantification using TapeStation and Nanodrop™. The primary QA concerned determining RNA integrity values (RIN) by calculating the ratio of 28S and 18S rRNA, where a RIN value >8 indicates a good quality of RNA extraction and purification (Imbeaud *et al.*, 2005). All RNA samples prepared in this study displayed a RIN value >8 by TapeStation analysis as shown by the representative samples in Figure 3.1A. RNA has a maximum absorbance at 260nm (A_{260}), therefore quantification of RNA can be determined by measuring at this absorbance using Nanodrop™. Absorbance readings can also be taken at 280nm to determine protein contamination and 230nm for contaminants such as salts and phenols. Therefore, additional QA using Nanodrop™ can be also carried out by calculating absorbance ratios, where an A_{260}/A_{280} value >2.1 indicates a highly pure RNA sample and a A_{260}/A_{230} ratio >1.5 indicates no contamination from salts and phenols. All RNA samples prepared in this study were highly pure and showed sufficient quantity for microarray analysis, as shown by the representative examples in Figure 3.1B.

3.2.1.2 Quality assessment of TERT-NHUC microarray data

Raw data was provided as CEL files from TPS, and processed using the Affymetrix Transcription Analysis Console software, which enables normalisation by SST-RMA (see Methods) and QA. The QA involves a series of tests which include generating histograms and box plots for the signal intensity of all probes for each sample, a principal component analysis (PCA) plot, and line graphs for labelling and hybridisation controls (Figure 3.2).

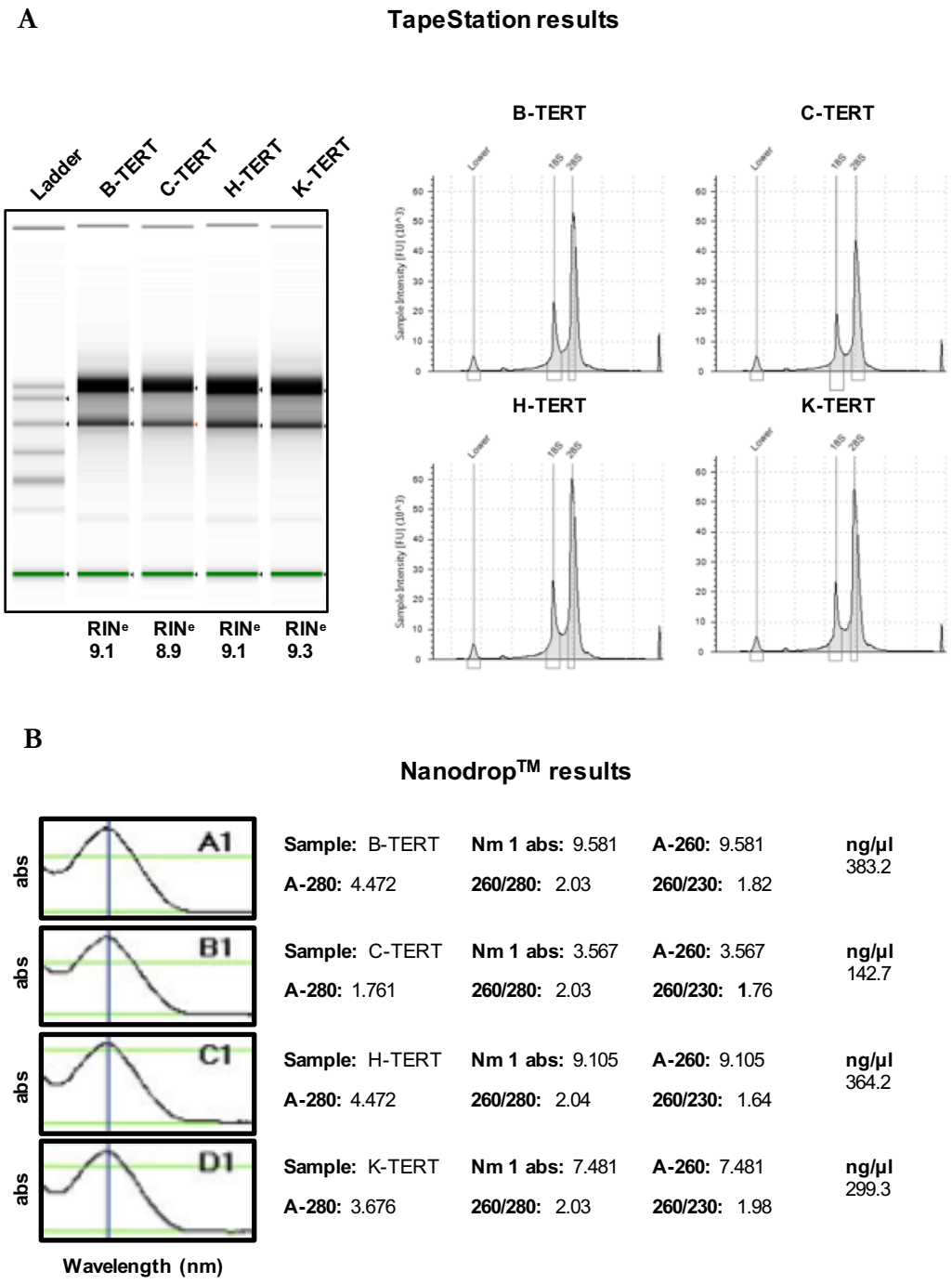


Figure 3.1 Representative RNA profiles for the TERT-NHUC samples used for microarray analysis

(A) TapeStation analysis was carried out on purified RNA to obtain RIN values and visualise potential degradation. Samples were considered suitable for microarray analysis when RIN values were >8.0. **(B)** Nanodrop was used to quantify RNA samples and assess purity. Samples were considered suitable for microarray analysis when an A_{260}/A_{280} ratio >2.1 and an A_{260}/A_{230} ratio >1.5 was attained.

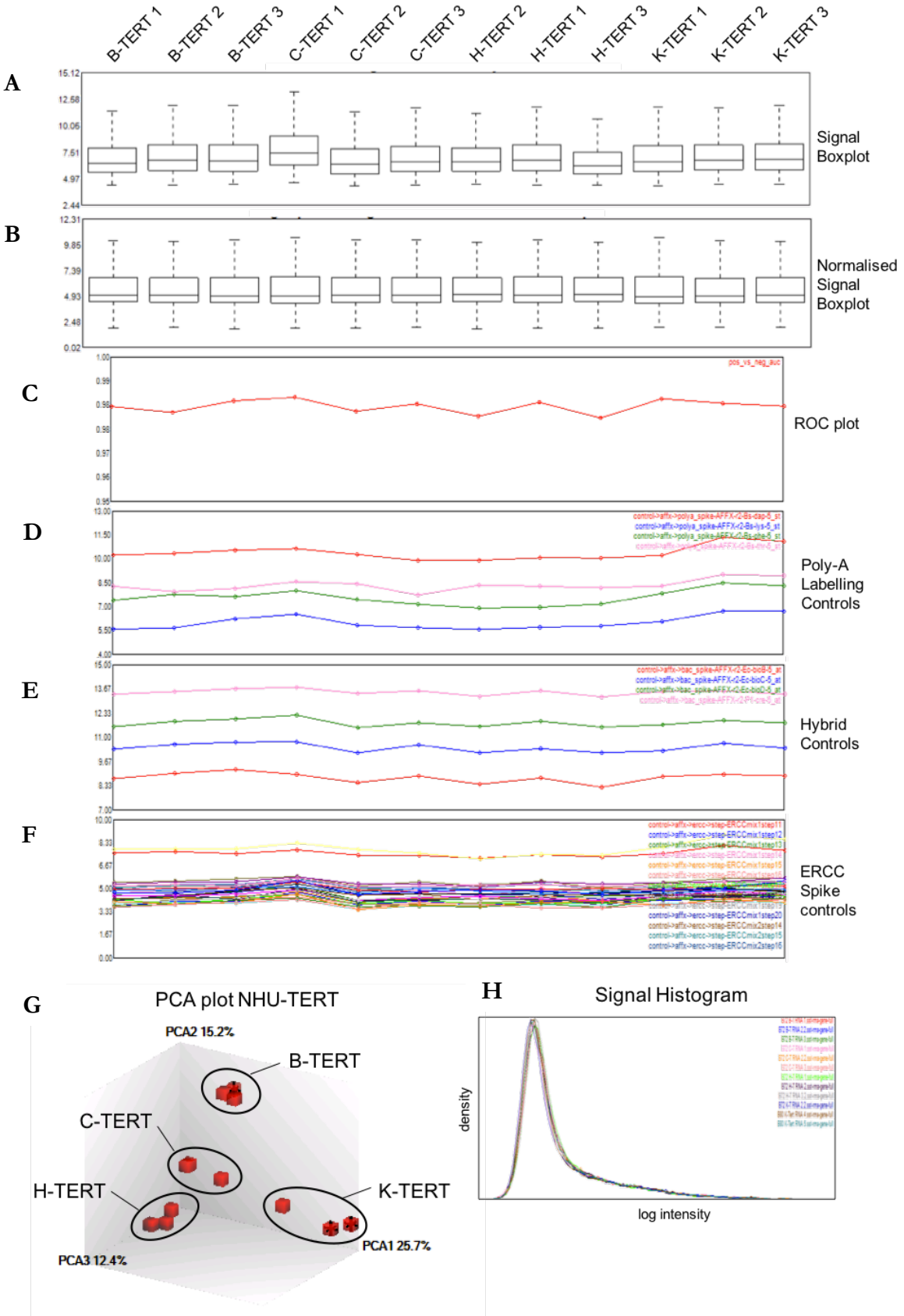


Figure 3.2 QA analysis of TERT-NHUC microarray data

QA on microarray data was carried out using the ThermoFisher Transcription Analysis Console. QA plots include; **(A)** a box plot of signal intensity, **(B)** a box plot of normalised signal intensity, **(C)** a receiver for operating characteristics (ROC) plot, **(D)** a line graph of poly-A labelling controls, **(E)** a line graph of hybridisation controls, **(F)** a line graph of ERCC spike controls, **(G)** a principal component analysis (PCA) plot, and **(H)** a histogram of log signal intensities.

Box plots of log signal-intensity for each sample were generated prior to, and after normalisation. The box plots show minimal differences in signal distribution between samples before normalisation and nearly identical signal distribution between samples following SST-RMA normalisation (Figure 3.2A & B). Equal distribution of signal intensity between samples is also seen in the signal intensity histogram (Figure 3.2H).

Using positive and negative control probes, a receiver for operating characteristics (ROC) plot can be generated. A ROC plot compares signal values for positive (constitutively expressed exons of GAPDH and β-actin) and negative (putative introns) controls, where the assumption is that signal generated from negative controls is a measure of false positives and signal generated from positive controls is a measure of true positives. When plotted, a value of 1 signifies perfect separation of positive and negative controls, with values ~0.85 typical for this array type. The RNA samples prepared here all showed ROC values greater than 0.98 indicating exceptional separation of controls (Figure 3.2C).

Poly-A labelling controls are used to monitor the target labelling process. These include lys, phe, thr and dap gene sets (absent from eukaryotic genomes) at increasing concentrations, and are spiked in prior to RNA hybridisation to HTA 2.0 to assess the overall success of the target preparation steps (Figure 3.2D). Samples were comparable and showed the expected pattern of signal intensity for each of the Poly-A labelling controls.

Hybridisation controls are used to test the sample hybridisation efficiency onto the gene expression array, and include ec-BioB, ec-BioC, ec-BioD and P1-Cre (absent from eukaryotic genomes) that are spiked into the hybridisation cocktail independent of the RNA sample preparation. Efficient hybridisation onto the arrays is attained if these hybridisation controls show increasing signal values for all samples to reflect increasing concentrations, as is observed in Figure 3.2E.

ERCC RNA spike-in controls include 92 polyadenylated transcripts that are used to assess the dynamic range, lower limit of detection, and fold-change response of the platform used, and ensure these limits are similar between all samples. ERCC spike-in controls showed comparable signal between the TERT-NHUC samples (Figure 3.2F).

Finally, a PCA plot was generated to display correlations between samples by considering where variability in the data is derived (Figure 3.2G). Variability is plotted on three axes, where the majority of the variability in the data is captured by PCA1 (25.7%), followed by PCA2 (15.2%) to capture the variability that was not considered in PCA1, and again by PCA3 (12.4%) for as much of the variability as was not accounted for by PCA1 and PCA2. As expected, the PCA plot showed that replicates from the same cell line group

together (Figure 3.2G). However, the PCA plot also shows that the majority of variance is the result of differences that are donor-related and not gender-related.

Overall the QA analyses for the TERT-NHUC samples used in this study were of a high quality, indicating comparable signal distribution, effective hybridisation and labelling of samples during array preparation, and a high correlation between replicate samples. QA data for microarray analyses on NHUC and UHUC, and K-TERT-NHUC samples that failed QA, are shown in Appendix A-1.

3.2.1.3 Differential expression analysis of male and female TERT-NHUC

Differential expression analysis between gender groups from the microarray data was carried out using LIMMA, which identifies DE genes by fitting probes to a linear model (Ritchie et al., 2015). The LIMMA analysis identified 507 probes with a fold change (FC) ≥1.5 and a P-value ≤0.05, of which 279 probes were upregulated in male TERT-NHUC (255 autosomal; 18 chrY and 6 chrX), and 228 probes were upregulated in female TERT-NHUC (205 autosomal; 23 chrX) (Figure 3.3; Table 3.1, Table 3.2; and Appendix A-2)

A previous study found only 10 genes that were differentially expressed in bladder, all of which were enriched in males and located on chrY. However, these samples were collected from post-mortem donors and are therefore likely to be relevant to the bladder detrusor muscle and not the urothelium (Gershoni and Pietrokovski, 2017). Nevertheless, the DE gene lists obtained from the TERT-NHUC microarray data also show these same 10 genes enriched in male TERT-NHUC, and include *DDX3Y, EIF1AY, KDM5D, NLGN4Y, RPS4Y1, TTTY15, TXLNGY, USP9Y, UTY,* and *ZFY* (Gershoni and Pietrokovski, 2017)(Figure 3.3; Appendix A-2). The most DE autosomal genes with >10 FC expression in male TERT-NHUC include *ANXA6, MMP1,* and *IL33* (Figure 3.3; Table 3.1). Interestingly, a high number of small nucleolar RNAs (snoRNAs) and ribosomal RNAs (rRNAs) showed increased expression in male TERT-NHUC (Appendix A-4), indicating altered regulation and processing of ribosomal components (Dupuis-Sandoval et al., 2015).

The most DE genes in females include genes that commonly escape X-chromosome inactivation (XCI) such as*, DDX3X, ELF1AX, ELF2S3, KDM5C, KDM6A, STS*, *SMC1A, USP9X, VGLL1 XG,* and *XIST* (Tukiainen et al., 2017) (Figure 3.3; Appendix A-3). However, other chrX genes that have not previously been characterised as escaping XCI were also upregulated and include *TMEM47, RPS6KA6, SLC38A5,* and *KLF8*. The most DE autosomal genes with >10 FC expression in females include *MIR4273, FRG2C,* and *ALOX15B* (Figure 3.3; Table 3.2). Of particular interest was *UCA1*, which showed a 9.99

FC in female TERT-NHUC, and is a commonly upregulated ncRNA in bladder cancer that promotes cell proliferation and migration (Luo et al., 2017; Lebrun et al., 2018)

To identify whether differentially expressed genes pertain to common pathways and molecular functions, the Database for Annotation, Visualisation, and Integrated Discovery (DAVID) was used (Huang et al., 2009). DAVID enabled a functional annotation the DE genes obtained by LIMMA by assigning them to gene-sets curated from the KEGG (Kyoto Encyclopedia of Genes and Genomes), Reactome, and GO (Gene Ontology) databases (Table 3.3 and Table 3.4). DAVID identified 39 terms associated with genes upregulated in male TERT-NHUC, and 27 terms associated with genes upregulated in female TERT-NHUC, where enriched terms were considered significant when $P \leq 0.05$. For male TERT-NHUC, DE genes were predominantly immune-related gene-sets including response to interferon that was mainly attributed to increased *IFITM1, IFITM2,* and *IFITM3,* and Rhematoid arthritis and immune response gene-sets that were mainly attributed to increased *CXCL5, CXCL6,* and *CXCL8* (Table 3.3). Male TERT-NHUC upregulated genes were also associated with blood vessel morphogenesis, cardiac muscle differentiation, and angiogenesis, which were mainly attributed to *CCBE1, EFNB2,* and *GREM1*, and the aforementioned snoRNAs were associated with ribosomal biogenesis. DE genes in female TERT-NHUC were associated with many different gene-sets including; cell-to-cell communication, gap junction, and connexon assembly (attributed to *GJA5, GJB2,* and *GJB6*); extracellular matrix assembly and cell adhesion (attributed to *LAMA1, VCAN,* and *PXDN*); and response to estradiol and progesterone (attributed to *GJB2, TXNIP,* and *NCOA1*) (Table3.4).

The DE gene lists generated by LIMMA were also analysed using Metacore (by Clarivate Analytics) to determine transcription factors (TFs) and their respective targets that may be upregulated (Appendix A-15). Only GATA6 was identified in male TERT-NHUC, and was associated with the upregulation of *CLDN11, DPP1, DKK-1, SARG, PDEF, PIB4,* and *TN-C*. There were no TFs found to be upregulated in female TERT-NHUC.

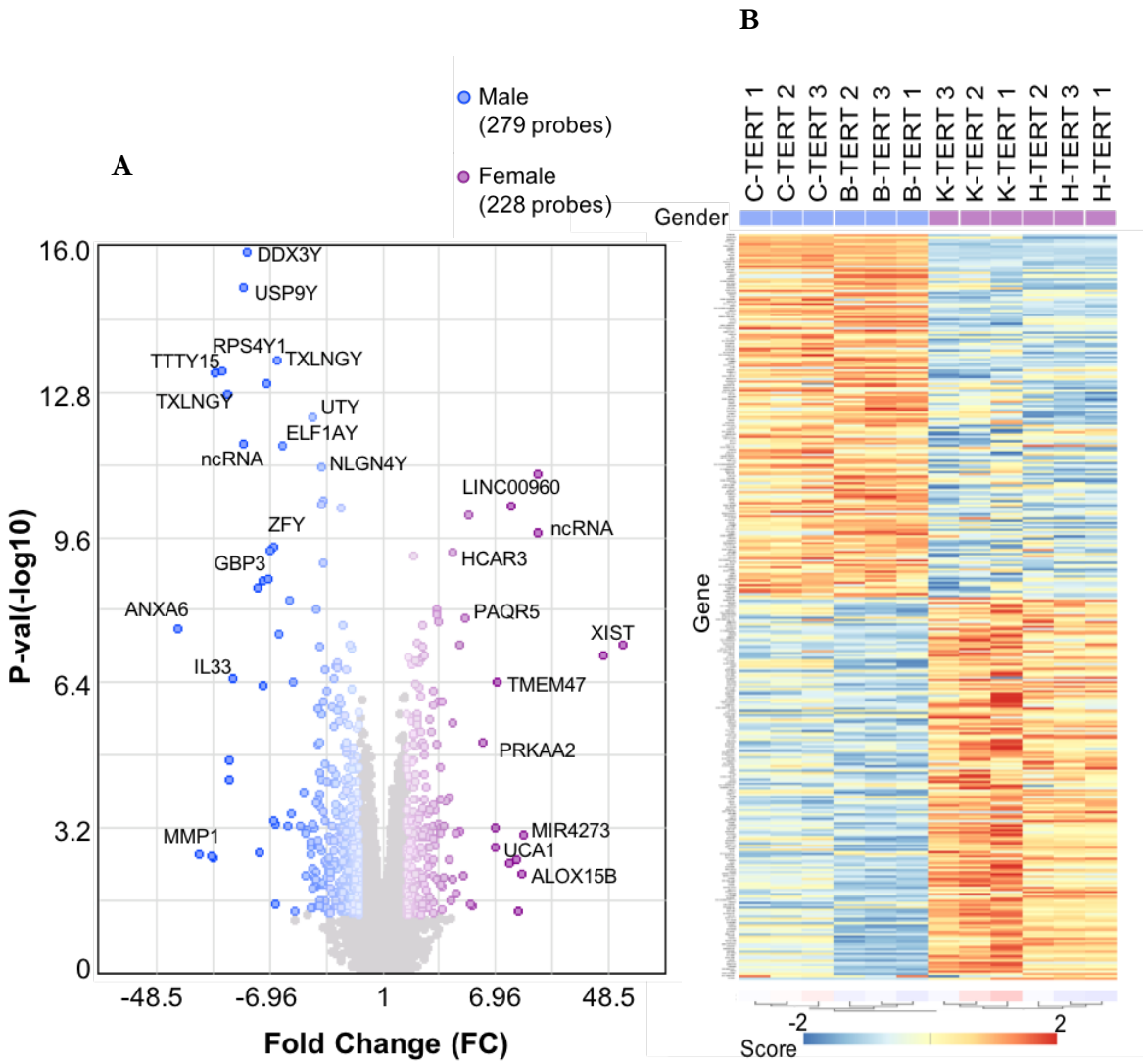


Figure 3.3 Volcano plot and heatmap of differentially expressed probes/genes between male and female TERT-NHUC

Gender-associated differential gene expression was determined by carrying out LIMMA between male and female TERT-NHUC HTA2.0 microarray data. Probes were considered differentially expressed (DE) when attaining ≥1.5-fold change (FC) and a P-value ≤ 0.05. **A)** Data is presented as a volcano plot of P-value vs fold change, with differentially expressed probes coloured blue for male DE probes and purple for female DE probes **B)** Heatmap of Z-scores for DE genes with hierarchical clustering of samples.

Table 3.1 The top 25 upregulated autosomal genes in male TERT-NHUC (determined by male vs female TERT-NHUC LIMMA analysis)

Symbol	Gene name	Chr	FC	P-value
ANXA6	annexin A6	chr5	29.92	<0.01
MMP1	matrix metallopeptidase 1	chr11	20.18	<0.01
IL33	interleukin 33	chr9	12.87	<0.01
GBP3	guanylate binding protein 3	chr1	7.30	<0.01
SLC16A4	solute carrier family 16, member 4	chr1	7.28	<0.01
EYA4	EYA transcriptional coactivator and phosphatase 4	chr6	6.81	<0.01
FKBP10	FK506 binding protein 10	chr17	6.56	<0.01
DSG3	desmoglein 3	chr18	6.33	0.03
KRTAP2-3	keratin associated protein 2-3	chr17	4.71	<0.01
PLCB4	phospholipase C, beta 4	chr20	4.47	0.04
KRT6A	keratin 6A, type II	chr12	4.27	<0.01
SAA1	serum amyloid A1	chr11	4.16	0.05
KRT6C	keratin 6C, type II	chr12	3.84	<0.01
CSF3	colony stimulating factor 3	chr17	3.72	0.04
MRGPRX3	MAS-related GPR, member X3	chr11	3.59	<0.01
KCCAT198	renal clear cell carcinoma-associated transcript 198	chr12	3.53	<0.01
SORL1	sortilin-related receptor, L(DLR class) A repeats containing	chr11	3.38	<0.01
KRT6B	keratin 6B, type II	chr12	3.31	<0.01
SLFN11	schlafen family member 11	chr17	3.28	0.01
TAGLN	transgelin	chr11	3.26	0.01
CCDC144B	coiled-coil domain containing 144B (pseudogene)	chr17	3.24	0.01
SHISA2	shisa family member 2	chr13	3.20	<0.01
CDH11	cadherin 11, type 2, OB-cadherin (osteoblast)	chr16	3.18	0.04
TNC	tenascin C	chr9	3.15	<0.01

Table 3.2 The top 25 upregulated autosomal genes in female TERT-NHUC (determined by male vs female TERT-NHUC LIMMA analysis)

Symbol	Gene name	Chr	FC	P-value
MIR4273	microRNA 4273	chr3	14.06	<0.01
FRG2C	FSHD region gene 2 family, member C	chr3	13.12	<0.01
ALOX15B	arachidonate 15-lipoxygenase, type B	chr17	10.83	0.01
UCA1	urothelial cancer associated 1 (non-protein coding)	chr19	9.99	<0.01
ADGRL4	adhesion G protein-coupled receptor L4	chr1	7.00	<0.01
NLRP2	NLR family, pyrin domain containing 2	chr19	6.75	<0.01
LINC00960	long intergenic non-protein coding RNA 960	chr3	6.67	<0.01
PRKAA2	protein kinase, AMP-activated, alpha 2 catalytic subunit	chr1	5.82	<0.01
AKT3	v-akt murine thymoma viral oncogene homolog 3	chr1	4.08	0.01
PAQR5	progestin and adipoQ receptor family member V	chr15	4.05	<0.01
GJB6	gap junction protein beta 6	chr13	3.71	<0.01
SORT1	sortilin 1	chr1	3.39	0.01
HCAR3	hydroxycarboxylic acid receptor 3	chr12	3.27	<0.01
TXNIP	thioredoxin interacting protein	chr1	3.16	<0.01
DPY19L2P1	DPY19L2 pseudogene 1	chr7	3.11	<0.01
LINC01296	long intergenic non-protein coding RNA 1296	chr22	2.98	<0.01
TMOD2	tropomodulin 2 (neuronal)	chr15	2.98	<0.01
VCAN	versican	chr5	2.97	<0.01
PEG10	paternally expressed 10	chr7	2.90	<0.01
LAMA1	laminin, alpha 1	chr18	2.87	<0.01
LOC102723854	uncharacterized LOC102723854	chr2	2.75	0.04
GPX3	glutathione peroxidase 3	chr5	2.75	<0.01
PLXDC2	plexin domain containing 2	chr10	2.66	<0.01
ACAA2	acetyl-CoA acyltransferase 2	chr18	2.64	<0.01

Table 3.3 Top 25 (of 39) enriched terms identified by DAVID for male TERT-NHUC DE expressed genes identified by LIMMA

ID	Genset name	DE genes in gene list	P-value
GO:0035455	response to interferon-alpha	LAMP3, IFITM1, IFITM2, IFITM3	0.000
GO:0051607	defense response to virus	IFITM1, IFITM2, IFITM3, IL33, OAS2, SLFN11, IFNK, DNAJC3, GBP3	0.000
GO:0034341	response to interferon-gamma	KYNU, IFITM1, IFITM2, IFITM3	0.001
GO:0048845	venous blood vessel morphogenesis	EFNB2, CCBE1, HEG1	0.001
GO:0060337	type I interferon signaling pathway	IFITM1, IFITM2, IFITM3, OAS2, IFI6	0.002
R-HSA-909733	Interferon alpha/beta signaling	IFITM1, IFITM2, IFITM3, OAS2, IFI6	0.002
GO:0035456	response to interferon-beta	IFITM1, IFITM2, IFITM3	0.002
GO:2000727	positive regulation of cardiac muscle cell differentiation	MYOCD, EFNB2, GREM1	0.003
GO:0009607	response to biotic stimulus	IFITM1, IFITM2, IFITM3	0.006
hsa05323	Rheumatoid arthritis	CXCL5, CXCL8, CXCL6, LTB, MMP1	0.007
GO:0009653	anatomical structure morphogenesis	EYA4, HOXA1, FAT1, EFNB2, MAB21L1	0.007
GO:0006955	immune response	CSF3, CXCL5, IFITM2, IFITM3, CXCL8, CXCL6, OAS2, IL7R, LTB, IFI6	0.008
GO:0031012	extracellular matrix	SERPINF1, F3, TNC, TGFBI, TGM2, HIST1H4F, ABI3BP, MMP1	0.009
GO:0046597	negative regulation of viral entry into host cell	IFITM1, IFITM2, IFITM3	0.009
GO:0005518	collagen binding	TGFBI, CCBE1, ABI3BP, SRGN	0.012
GO:0005925	focal adhesion	ANXA6, CYBA, LPXN, FAT1, CD46, TNC, EFNB2, TGM2, DPP4	0.012
GO:0009615	response to virus	IFITM1, IFITM2, IFITM3, OAS2, IFNK	0.013
GO:0008284	positive regulation of cell proliferation	CSF3, KRT6A, CXCL5, MYOCD, TNC, EFNB2, GREM1, MAB21L1, DPP4, IL31RA	0.014
GO:0045766	positive regulation of angiogenesis	GATA6, F3, CCBE1, CXCL8, GREM1	0.015
GO:0044267	cellular protein metabolic process	SAA1, GATA6, TGFBI, HIST1H4F, MMP1	0.016
hsa03008	Ribosome biogenesis in eukaryotes	SNORD3A, SNORD3C, SNORD3B-1, SNORD3B-2	0.018

Table 3.4 Top 25 (of 27) enriched terms identified by DAVID for female TERT-NHUC DE expressed genes identified by LIMMA

ID	Genset name	DE genes in gene list	P-value
R-HSA-373760	L1CAM interactions	LAMA1, CNTN1, EPHB2	0.004
GO:0008152	metabolic process	ENPP5, ACAA2, GSTM3, UGT8, LPCAT2, UGT1A1, DXO	0.005
GO:0007154	cell communication	ENPP5, GJB6, GJA5, GJB2	0.005
hsa04722	:Neurotrophin signaling pathway	IRAK2, RPS6KA6, MAPK13, SORT1, SHC3, AKT3	0.006
GO:0006954	inflammatory response	PRKD1, IRAK2, LXN, ITGB6, F2RL1, SCN9A, PARP4, NLRP3, NLRP2, LY75-CD302	0.008
GO:0016491	oxidoreductase activity	DHRS2, CYP24A1, FAR2, PDPR, SCCPDH, EGLN3, SCD5	0.009
GO:0006897	endocytosis	SNX9, MRC2, SORT1, DPYSL2, SNX33, LY75-CD302	0.009
GO:0007160	cell-matrix adhesion	ITGB8, NPNT, ITGB6, BCAM, SGCE	0.009
GO:0032355	response to estradiol	TXNIP, NCOA1, ASS1, CAT, GJB2	0.010
GO:0070542	response to fatty acid	ASS1, UCP2, CAT	0.010
GO:0042744	hydrogen peroxide catabolic process	PXDN, GPX3, CAT	0.014
GO:0005922	connexon complex	GJB6, GJA5, GJB2	0.015
GO:0007155	cell adhesion	LAMA1, ITGB8, ITGB6, DSC3, RHOB, CNTN1, BCAM, VCAN, SPOCK1, NEO1	0.026
GO:0098869	cellular oxidant detoxification	PXDN, FAM213A, GPX3, CAT	0.026
R-HSA-3296197	Hydroxycarboxylic acid-binding receptors	HCAR3, HCAR2	0.027
GO:0016050	vesicle organization	SNX9, SORT1, SNX33	0.027
GO:0051092	positive regulation of NF-kappaB transcription factor activity	PRKD1, IRAK2, CLU, CAT, NLRP3	0.034
GO:0030198	extracellular matrix organization	LAMA1, PXDN, ITGB8, NPNT, ITGB6, VCAN	0.034
GO:0005578	proteinaceous extracellular matrix	LAMA1, PXDN, NPNT, PI3, VCAN, SPOCK1, MMP2	0.036
GO:0006663	platelet activating factor biosynthetic process	LPCAT2, CHPT1	0.036

3.2.1.4 Gene-set enrichment analysis for gender-enriched gene sets

To determine if any commonly annotated pathways, molecular functions, or cellular components were differentially enriched between male and female TERT-NHUC, gene-set enrichment analysis (GSEA) was carried out on the entire HTA 2.0 gene list and using annotated gene-sets available from the molecular signature database (MSigDB) (Subramanian *et al.*, 2005). In particular, MsigDB gene-sets curated from KEGG, Reactome, and GO databases, as well as the Hallmark gene-set of the most well-defined biological states and processes, were used for analysis (Figure 3.4 & Figure 3.5). GSEA analysis therefore identifies enriched gene-sets by assessing differences across all genes throughout the genome, and as not restricted to DE gene lists as in the DAVID analysis. Gene-sets were considered enriched when a P-value ≤ 0.05 was attained. However, it is noted that no gene-sets were identified with a false discovery rate (FDR; Q-value) ≤ 0.1. Following GSEA, heatmaps with hierarchical clustering of samples were also generated using the enriched gene-set lists, and are shown next to their respective gene-sets throughout the following figures.

The GSEA identified 58 gene-sets enriched in male TERT-NHUC, 56 of which were from the GO gene-sets, and only one from each of the KEGG ("vascular smooth muscle contraction") and Reactome ("cell-to-cell communication") gene-sets (Figure 3.4, Appendix A-5). No gene-sets from the Hallmarks curated list were enriched in male TERT-NHUC. Despite the high number of enriched gene-sets, there was no general theme of biological processes, functions, or cellular components. The GO analysis does include a number of development and differentiation related gene-sets, although these involve tissues and cells encompassing the respiratory system, kidneys, digestive system, mesenchyme and stem cells and were attributed to a similar set of genes including *MYOCD, HOXA5, WNT5A,* and *LRP6* (Appendix A-5). A number of gene-sets linked to both positive and negative regulation immune-related pathways were also enriched, as well as gene-sets involved in muscle regulation. However, the top two most enriched GO gene-sets include the regulation of protein modifications, mainly pertaining to ubiquitination (Figure 3.4). Inspection of these gene-sets showed that the gene-set enrichment was due to a small number of ubiquitin regulating genes including the ubiquitin peptidases *USP9Y, USP2, USP45, USP17L6P, USP40,* and the deubiquitinases *OTUB2* and *TNFAIP3,* which were upregulated in male TERT-NHUC.

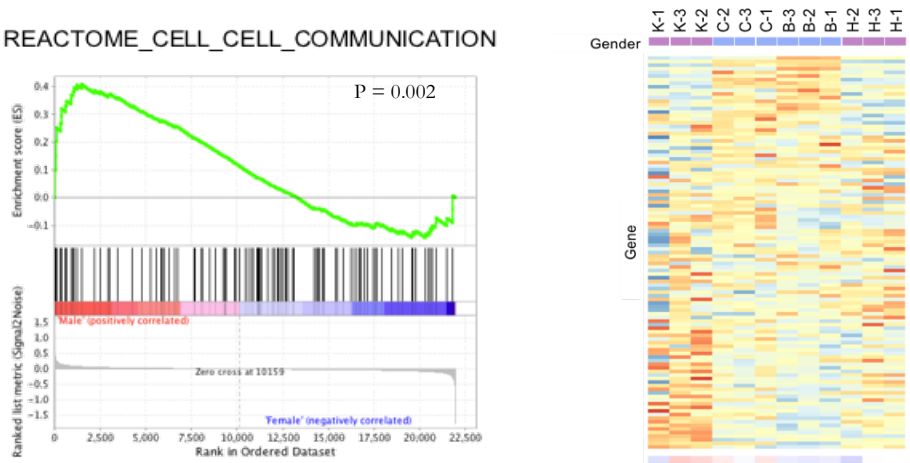
For female TERT-NHUC 76 gene-sets were enriched, with only one gene-set enriched from each of the KEGG ("T-Cell Receptor Signalling"), Reactome ("signalling by FGFR in disease") and Hallmark ("UV response up") curated gene-set lists (Figure 3.5; Appendix A-6). Signalling by FGFR was particularly interesting given that mutations of

FGFR3 are the most common type of mutation in bladder cancer. Inspection of this gene-set showed that 36 of 116 genes in this list contributed to enrichment, and included genes such as CDKN1A, AKT2/3, FGF2/3, MAPK1, MAPK2K1/2, GAB1/2 and KRAS, but did not include the majority of FGFs or their receptors. As with the male enriched gene-sets, females showed a high number of GO enriched gene-sets but lacked overall biological themes. Again, the GO analysis identified a number of development related gene-sets enriched in female TERT-NHUC, this time pertaining to spinal cord, appendage, ear, reproductive system, and germ cells. Enrichment of many gene-sets related to metabolic and biosynthetic processes of fatty acids, steroids, monocarboxylic acids and hydroxy compounds was also observed, and with genes attributed to promoting these pathways including ALOX15B, ALOX5, PRKAA2, PTGS1, and SCD5.

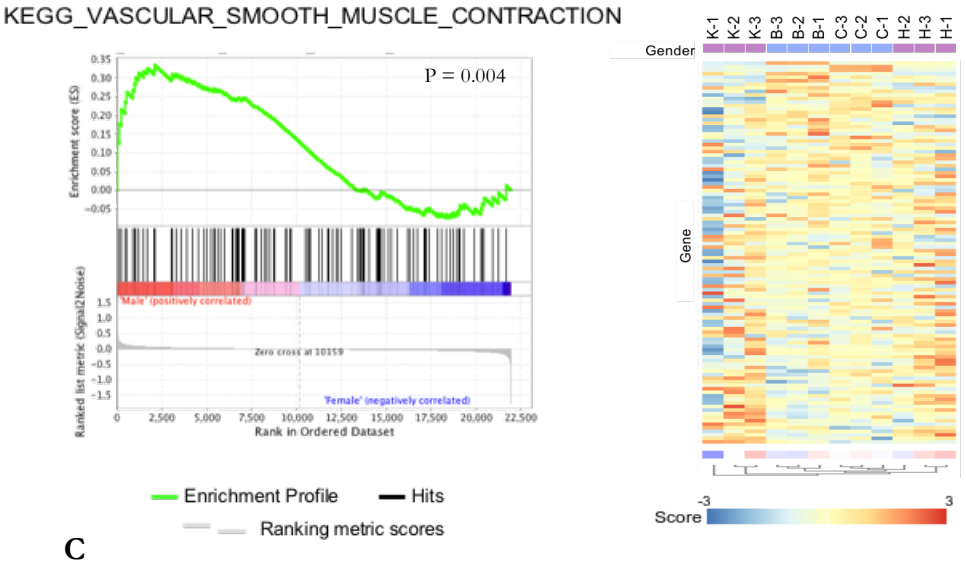
Despite the aforementioned results of enriched gene-sets, the accompanying heatmaps show that enrichment of each gene-set may not primarily have been driven by gender differences, and instead by differential expression of genes in K-TERT (Figure 3.4 and Figure 3.5). Clustering of samples using the gene list from each gene-set showed that male (B and C) TERT-NHUC cluster together, but that the female (H and K) TERT-NHUC separate into distinct groups, with H-TERT clustering more with the male TERT-NHUC. This is particularly apparent for the female enriched gene-sets, where a strong enrichment of genes in K-TERT can be seen but not in male TERT-NHUC or H-TERT. Therefore K-TERT alone is likely driving the enriched gene-set phenotype in in females (Figure 3.5).

The GSEA between male and female TERT-NHUC identified a wide and varied range of enriched gene-sets for each gender. Although enrichment within each gender lacked general biological themes, males showed enrichment for a number of immunogenic and muscle regulating gene-sets, whereas females showed enrichment of metabolic and biosynthetic processes. Both genders had enrichment of gene-sets pertaining to development, although varying in cell and tissue type. It is noted that, although enriched gene-sets attained low P-values, Q-values remained high and there was little concordance of enriched gene-sets between the different databases. Furthermore, it is likely that enriched gene-sets may be attributed to differences in a single cell line (K-TERT), and not between genders, as seen by the accompanying heatmaps for all gene-sets.

A



B

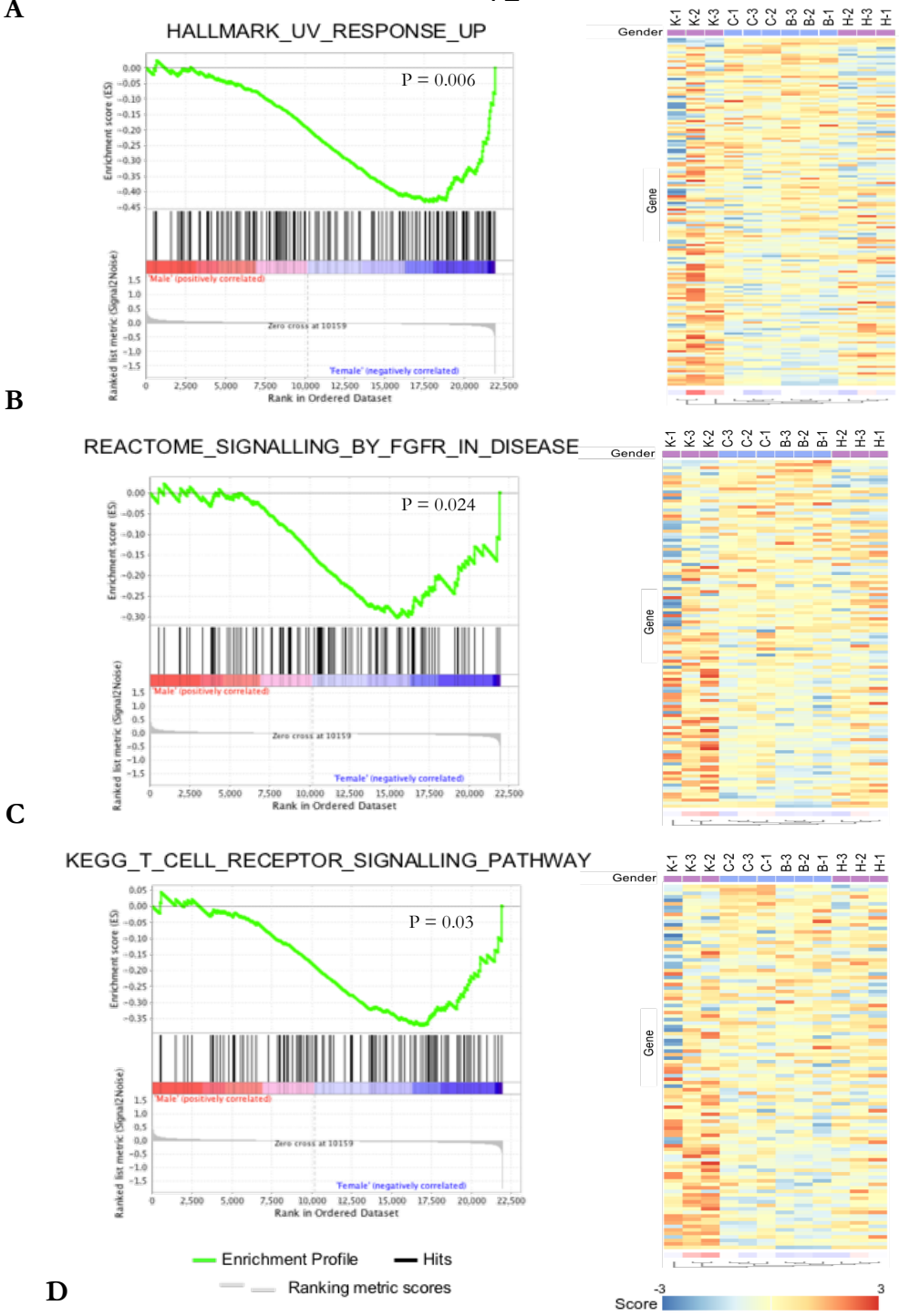


C

Enriched GO Terms in Male TERT-NHUC (top 5 of 56)	P-Value
GO_PROTEIN_MODIFICATION_BY_SMALL_PROTEIN_REMOVAL	0.000
GO_UBIQUITIN_LIKE_PROTEIN_SPECIFIC_PROTEASE_ACTIVITY	0.000
GO_PROTEIN_FOLDING	0.027
GO_COLUMNAR_CUBOIDAL_EPITHELIAL_CELL_DIFFERENTIATION	0.002
GO_REGULATION_OF_G_PROTEIN_COUPLED_RECEPTOR_PROTEIN_SIGNALING_PATHWAY	0.002
GO_TRANSLATIONAL_INITIATION	0.010
GO_RESPONSE_TO_RETINOIC_ACID	0.004
GO_ENDOPLASMIC_RETICULUM_LUMEN	0.000
GO_SYNAPSE_ORGANIZATION	0.024
GO_POSITIVE_REGULATION_OF_BINDING	0.038

Figure 3.4 Male TERT-NHUC enriched pathways identified using GSEA

All probes from the HTA 2.0 microarray were used to carry out Gene-set enrichment analysis (GSEA) for male vs female TERT-NHUC. GSEA was used to identify enriched gene-sets obtained from MSigDB that were curated from **A)** Reactome, **B)** KEGG, and **C)** GO databases. No gene sets from the Hallmarks list were enriched in male TERT-NHUC. Enriched gene sets were used to generate heatmaps of Z-scores with hierarchical clustering of samples, and are shown to the right of their respective GSEA plot. Gene sets were considered enriched when a P-value ≤ 0.05 was attained.



Enriched GO Terms in Female TERT-NHUC (top 10 of 73)	P-Value
GO_NEGATIVE_REGULATION_OF_INTRACELLULAR_TRANSPORT	0.000
GO_ACTIN_FILAMENT_BINDING	0.012
GO_INTRACELLULAR_RECEPTOR_SIGNALING_PATHWAY	0.000
GO_REGULATION_OF_CIRCADIAN_RHYTHM	0.000
GO_STEROID_HORMONE_MEDIATED_SIGNALING_PATHWAY	0.000
GO_VACUOLAR_LUMEN	0.006
GO_NEGATIVE_REGULATION_OF_CYTOPLASMIC_TRANSPORT	0.008
GO_GERM_CELL_DEVELOPMENT	0.000
GO_REGULATION_OF_PHOSPHATASE_ACTIVITY	0.022
GO_FATTY_ACID_METABOLIC_PROCESS	0.000

Figure 3.5 Female TERT-NHUC enriched pathways identified using GSEA

All probes from the HTA 2.0 microarray were used to carry out Gene-set enrichment analysis (GSEA) for male vs female TERT-NHUC. GSEA was used to identify enriched gene-sets obtained from MSigDB, including **A)** the Hallmarks gene set, and gene sets curated from **B)** Reactome, **C)** KEGG, and **D)** GO databases. Enriched gene sets were used to generate heatmaps of Z-scores with hierarchical clustering of samples, and are shown to the right of their respective GSEA plot. Gene sets were considered enriched when a P-value ≤ 0.05 was attained.

3.2.1.5 LIMMA and GSEA analysis on TERT-NHUC and NHUC

So far, a comparison of expression profiles in healthy urothelium had been limited to two male and two female TERT-NHUC. This comparison is useful in itself and complements data generated by ATAC-seq, but it is limited by the number of samples used for each gender. To overcome this, expression profiles were also obtained for three male and two female non-immortalised NHUC (direct from patients, not cultured) and three male and three female uncultured normal urothelial cells (UHUC). Combining these expression profiles with those from the TERT-NHUC would improve confidence in the findings. However, as QA of samples identified increased variance that resulted in a disparate grouping of UHUC from NHUC and TERT-NHUC, it was not appropriate to include the UHUC into this gender comparison analysis (Appendix A-1). Unlike UHUC, TERT-NHUC and NHUC do cluster together by PCA. However, it is noted that previous studies have shown that although immortalisation of NHUC by hTERT does not produce chromosomal alterations, differences in the expression profiles of matched pairs of immortalised and non-immobilised NHUC included genes involved in differentiation and tumorigenesis, associated with regulation by PCG (Chapman *et al.*, 2006; Chapman *et al.*, 2008). Nevertheless, an additional gender-based comparison of healthy urothelial cells was carried out using NHUC/TERT-NHUC, giving a total of 5 males and 4 females that distinctly clustered together by PCA (Appendix A-1). This would better identify gender-associated gene expression differences, instead of differences deriving from individual donors (Figure 3.6).

Differential expression analysis using LIMMA between male and female NHUC/TERT-NHUC identified 210 probes with a FC $\geq$ 1.5 and a P-value $\leq$ 0.05 (Figure 3.6A, Appendix A-7, Appendix A-8). Of these, 103 probes were upregulated in male NHUC/TERT-NHUC (83 autosomal, 19 chrY and 1 chrX) and 107 probes were upregulated in female NHUC/TERT-NHUC (92 autosomal, 15 chrX). Therefore, the number of DE probes decreased by over half when including NHUC compared to the previous analysis. The heatmap of DE genes shows that although male and female NHUC/TERT-NHUC cluster into distinct groups, the gender difference is strongest between TERT-NHUC and weaker between NHUC (Figure 3.6B).

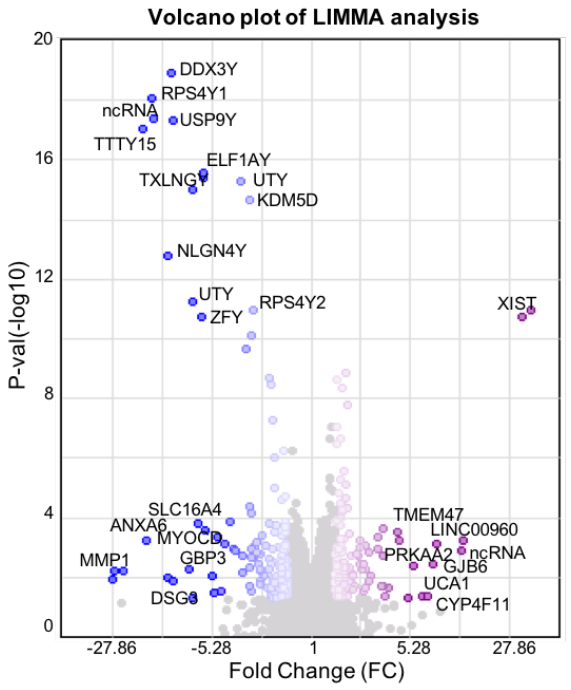
As before, the 10 gender-associated DE chrY genes identified in the literature (Gershoni and Pietrokovski, 2017) were DE following LIMMA analysis when including the NHUC. *ANXA6* and *MMP1* were again the most differentially expressed autosomal genes in males, but not *IL33* (Figure 3.6C) . Only 2 snoRNAs (*SNORD111* & *SNORD14D*) and 2 rRNAs (*RPS4Y1* and *RPS4Y2*) were DE in males when including NHUC. For females, many of the XCI genes were again upregulated, with notable exceptions including *DDX3X,*

VGLL1, XG, and *KDM6A*. Furthermore, *UCA1* was still included in the most differentially expressed autosomal genes (Figure 3.6D). Overall there was good consistency for the top DE genes following LIMMA analysis, where 8 of the top 10 DE autosomal genes identified for each gender when including NHUC were found among the top 25 most DE autosomal genes in the previous analysis.

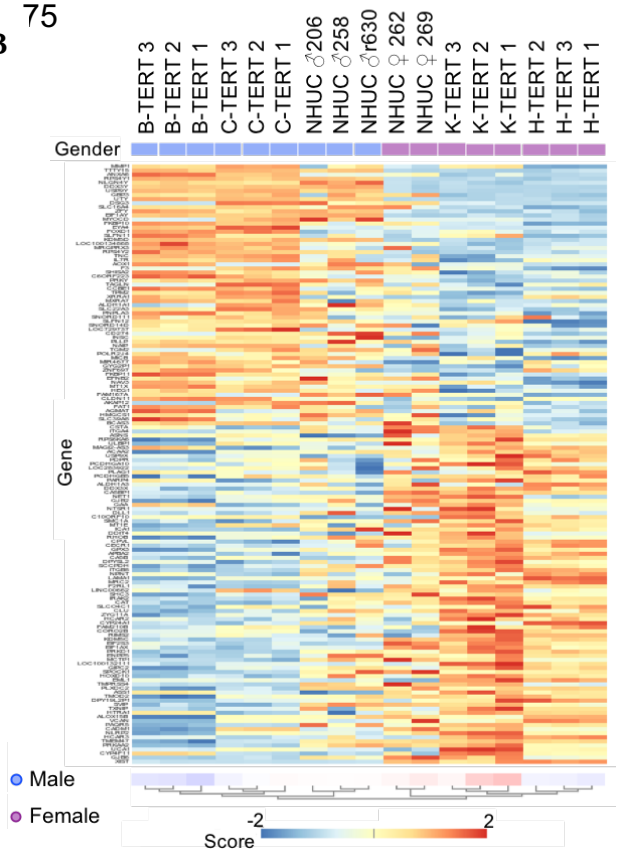
DAVID was again used to identify common pathways and molecular functions associated with gender-related DE genes obtained by LIMMA in NHUC/TERT-NHUC (Figure 3.6E & F). DAVID identified 5 terms associated with male NHUC/TERT-NHUC DE genes, and 17 terms associated with female NHUC/TERT-NHUC, where enriched terms were considered significant when $P \leq 0.05$. Male terms included venous blood and lymph vessel formation, cardiac muscle differentiation, and focal adhesion, and all terms included EFNB2 in their gene lists (Figure 3.6E). Interestingly, Metacore analysis of active transcription factors for male NHUC/TERT-NHUC DE genes identified only MYOCD (Appendix A-15), a TF commonly expressed in cardiac, aorta, and vasculature tissues, and which was also identified by DAVID as constituting part of the cardiac muscle differentiation gene list. Metacore did not identify any active TFs for female NHUC/TERT-NHUC. Female terms identified by DAVID included cell-to-cell communication and cell junctions (attributed to GJB2 and GJB6), extracellular matrix organisation (attributed to LAMA and VCAN), and response to estradiol (Figure 3.6F). These terms identified by DAVID in male and female NHUC/TERT-NHUC were also seen in the previous analysis with TERT-NHUC, however terms pertaining to immune-related pathways and ribosome biogenesis were no longer seen in males, and terms related to metabolic processes were no longer seen in females.

GSEA on the full HTA 2.0 gene list produced vastly different results when including NHUC (Figure 3.6C & D). Males showed only two enriched gene-sets including "Diabetic Pathways" from Reactome, and "Leukocyte Transendothelial Migration" from KEGG, neither of which were identified in the previous analysis (Figure 3.6C). Females showed enrichment of 17 gene-sets, with only 1 identified in the previous GSEA, namely the "UV response up" from the Hallmarks gene-set list (Figure 3.7; Appendix A-8). A closer inspection of this curated gene-set showed that only a few of the gene-lists that constituted this set were related to UV response, with the majority of the constituting gene-sets related to a single study regarding fibroblast response to human cytomegalovirus infection. An additional Hallmark gene-set pertaining to fatty acid metabolism was also enriched in females when including NHUC, which does coincide with GO gene-sets identified from the previous GSEA on TERT-NHUC.

A



B



C

Top 10 (of 83) male DE autosomal

Symbol	Gene name	Chr	FC	p-value
MMP1	matrix metallopeptidase 1	chr11	27.30	0.01
ANXA6	annexin A6	chr5	15.67	0.00
GBP3	guanylate binding protein 3	chr1	7.69	0.01
DSG3	desmoglein 3	chr18	7.25	0.05
SLC16A4	solute carrier family 16, member 4	chr1	6.68	0.00
KRTAP2-3	keratin associated protein 2-3	chr17	5.20	0.01
FKBP10	FK506 binding protein 10	chr17	4.81	0.00
MYOCD	myocardin	chr17	4.31	0.00
EYA4	EYA transcriptional coactivator and phosphatase 4	chr6	3.97	0.00
FOXD1	forkhead box D1	chr5	3.20	0.01

D

Top 10 (of 92) female DE autosomal genes

Symbol	Gene name	Chr	FC	p-value
LINC00960	long intergenic non-protein coding RNA 960	chr3	8.48	0.00
CYP4F11	cytochrome P450, family 4, subfamily F, polypeptide 11	chr19	6.97	0.04
UCA1	urothelial cancer associated 1 (non-protein coding)	chr19	5.74	0.05
PRKAA2	protein kinase, AMP-activated, alpha 2 catalytic subunit	chr1	5.54	0.00
GJB6	gap junction protein beta 6	chr13	5.38	0.00
LINC01296	long intergenic non-protein coding RNA 1296	chr22	3.45	0.04
HCAR3	hydroxycarboxylic acid receptor 3	chr12	3.39	0.00
NLRP2	NLR family, pyrin domain containing 2	chr19	3.31	0.00
CADM1	cell adhesion molecule 1	chr11	3.30	0.02
PAQR5	progestin and adipoQ receptor family member V	chr15	3.22	0.02

E

Top 5 (of 5) male enriched terms by DAVID

ID	Genset name	DE genes in gene list	p-value
GO:0048845	venous blood vessel morphogenesis	EFNB2, CCBE1, HEG1	0.000
GO:0009653	anatomical structure morphogenesis	EYA4, FAT1, EFNB2, MAB21L1	0.003
GO:0005925	focal adhesion	ANXA6, FAT1, TNC, EFNB2, AKAP12, TGM2	0.007
GO:0001945	lymph vessel development	EFNB2, HEG1	0.025
GO:2000727	positive regulation of cardiac muscle cell differentiation	MYOCD, EFNB2	0.031

F

Top 5 (of 17) female enriched terms by DAVID

ID	Genset name	DE genes in gene list	p-value
GO:0043065	positive regulation of apoptotic process	TXNIP, DDX3X, ALDH1A3, CLU, RHOB, NTSR1, NET1	0.002
GO:0032355	response to estradiol	TXNIP, ASS1, CAT, GJB2	0.008
GO:0001649	osteoblast differentiation	CYP24A1, MRC2, VCAN, CAT	0.011
GO:0030198	extracellular matrix organization	LAMA1, ITGB8, NPNT, VCAN, ITGA4	0.011
GO:0007154	cell communication	ENPP5, GJB6, GJB2	0.012

Figure 3.6 Differential gene expression analysis between male and female NHUC/TERT-NHUC

Gender-associated differential gene expression was determined by carrying out LIMMA analysis between male and female NHUC/TERT-NHUC HTA2.0 microarray data. Probes were considered differentially expressed (DE) when attaining ≥1.5-fold change (FC) and a P-value ≤ 0.05. **A)** Data is presented as a volcano plot of P-value vs fold change, with differentially expressed probes coloured blue for male DE probes and purple for female DE probes **B)** Heatmap of Z-scores for DE genes with hierarchical clustering of samples. **C)** Top 10 autosomal genes upregulated in males. **D)** Top 10 autosomal genes upregulated in females. **E)** Top 5 Terms identified by DAVID analysis on male DE gene list. **F)** Top 5 Terms identified by DAVID analysis on female DE gene list.

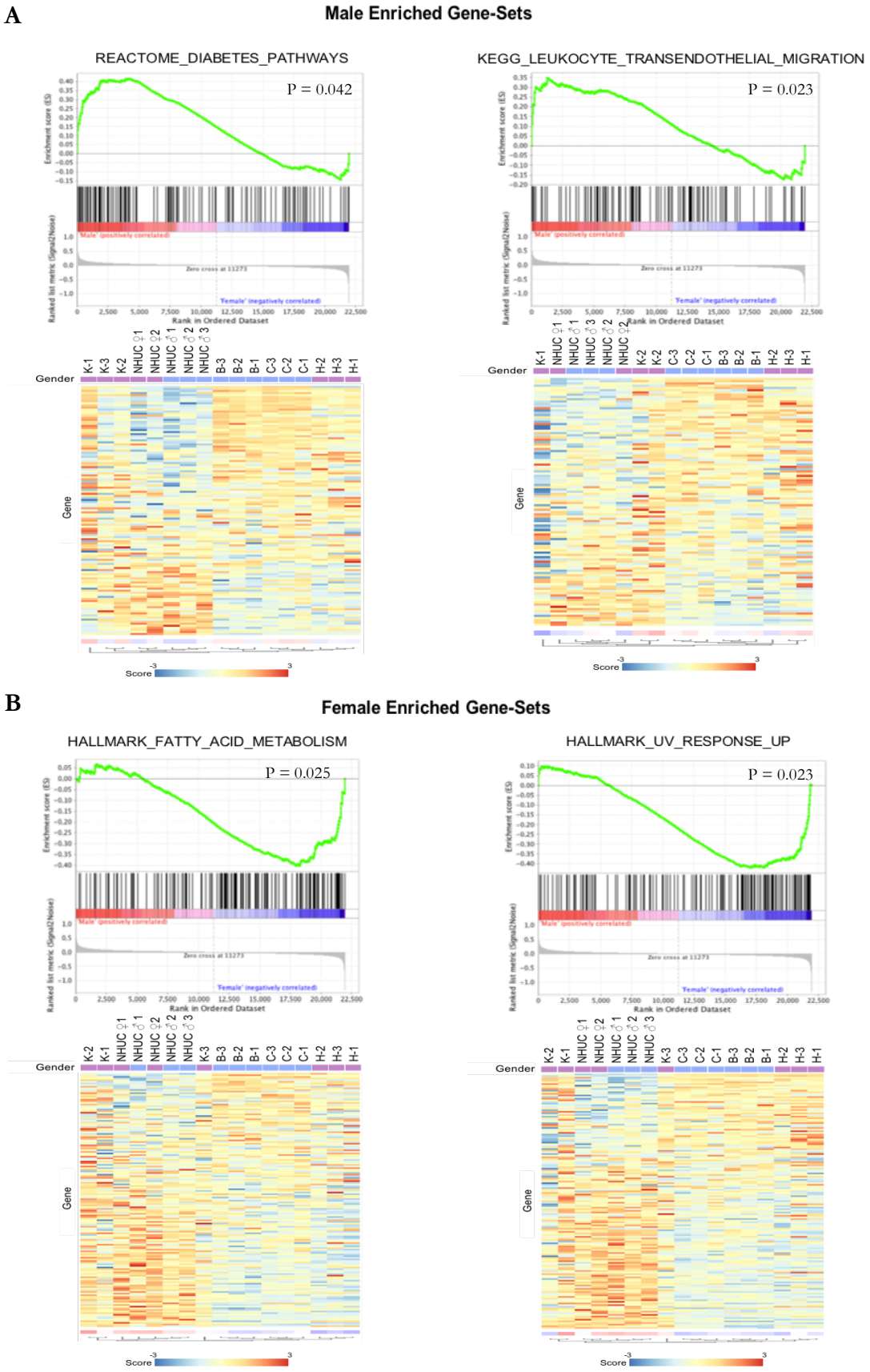


Figure 3.7 GSEA between male and female NHUC/TERT-NHUC

All probes from the HTA 2.0 microarray were used to carry out GSEA for male vs female NHUC/TERT-NHUC for gene-sets obtained from MSigDB. **A)** Male enriched gene-sets. **B)** Female enriched gene sets. Enriched gene-sets were used to generate heatmaps of Z-scores with hierarchical clustering of samples, and are shown to the right of their respective GSEA plot. Gene sets were considered enriched when a P-value ≤ 0.05 was attained.

As with the previous analysis, the heatmaps with hierarchical clustering of samples in NHUC/TERT-NHUC showed little correlation with their respective GSEA plots. Samples predominantly clustered into NHUC and TERT-NHUC groups rather than by gender. Furthermore, K-TERT clustered with the NHUC groups instead of the expected TERT-NHUC group. This therefore indicates that the enriched gene-sets may be attributed more to the immortalised state of cells, or differences in individual donor phenotypes, rather than between genders.

After adding NHUC into the gender comparison of expression profiles, a smaller number of overall DE genes was observed. The majority of DE genes located on sex-chromosomes were seen in both analyses, as well as the top DE autosomal genes, and similar terms were also identified by DAVID. However, there was a dramatically reduced number of gene-sets enriched following GSEA when including NHUC, with little concordance with the previous analysis. Furthermore, differences may be attributed to the immortalised state of samples or to variations in individual donors, rather than to gender.

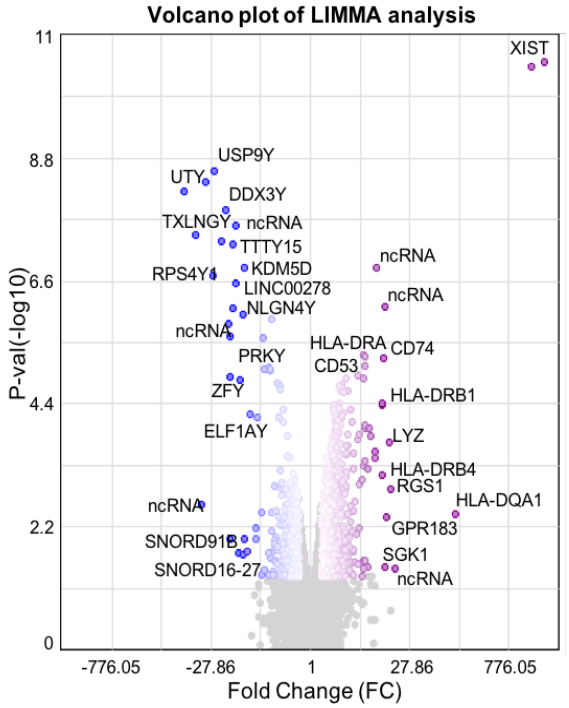
3.2.2 Gender-related transcriptome analysis in uncultured human urothelial cells (UHUC)

Microarray analysis was also used to generate expression profiles of three male and three female UHUC. As these are non-proliferative cells and PCA analysis (Appendix A-1) showed clear separation of their overall expression profiles from the cultured cells, these were analysed separately (Figure 3.8).

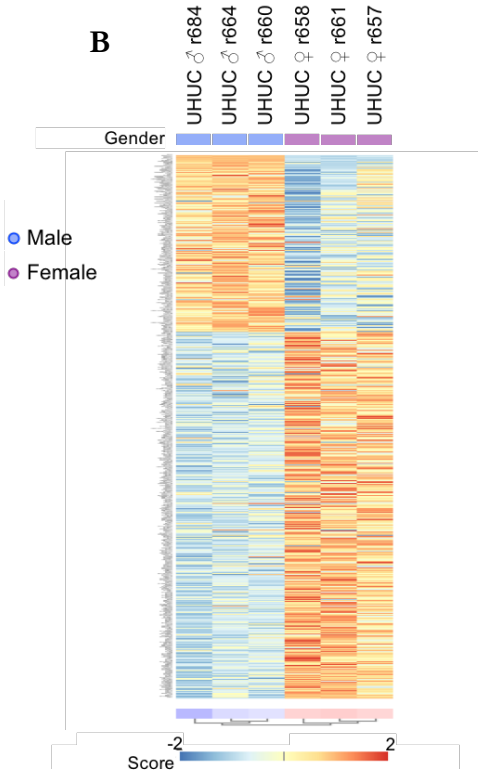
Differential expression analysis between male and female UHUC using LIMMA identified 848 DE probes with ≥1.5-fold change (FC) and a P-value ≤0.05. Of these, 253 probes were upregulated in male UHUC (217 autosomal; 26 chrY and 10 chrX), and 585 probes were upregulated in female UHUC (569 autosomal; 26 chrX) (Figure 3.8; Appendix A-10 & E-11.

The previously identified 10 chrY genes were again among the top 20 most DE genes in males following LIMMA analysis on UHUC. Of the 253 DE probes in males, 50 were related to ribosomal regulating genes, 43 of which were snoRNAs (Figure 3.8A; Appendix A-10). Indeed, 5 of the top 10 DE autosomal genes were snoRNAs and include *SNORD91B, SNORD116-27, SNORA22, SNORD93* and *SNORD115-15* (Figure 3.8C). Other protein-coding DE autosomal genes in males included *GPNMB, CAPNS2,* and *TBX2*. However,

A



B



C

Top 10 (of 217) male DE autosomal genes

Symbol	Gene name	Chr	FC	p-value
SNORD91B	small nucleolar RNA, C/D box 91B	chr17	9.43	0.02
GPNMB	glycoprotein (transmembrane) nmb	chr7	9.01	0.01
SNORD116-27	small nucleolar RNA, C/D box 116-27	chr15	5.10	0.05
CAPNS2	calpain, small subunit 2	chr16	4.73	0.04
TBX2	T-box 2	chr17	4.56	0.01
SNORA22	small nucleolar RNA, H/ACA box 22	chr7	4.09	0.05
SNORD93	small nucleolar RNA, C/D box 93	chr7	4.04	0.02
UGT2B7	UDP glucuronosyltransferase 2 family, polypeptide B7	chr4	3.74	0.00
SNORD115-15	small nucleolar RNA, C/D box 115-15	chr15	3.66	0.01
TAC3	tachykinin 3	chr12	3.54	0.04

D

Top 10 (of 569) female DE autosomal genes

Symbol	Gene name	Chr	FC	p-value
HLA-DQA1	major histocompatibility complex, class II, DQ alpha 1	chr6	128.12	0.00
RGS1	regulator of G-protein signaling 1	chr1	14.82	0.00
LYZ	lysozyme	chr12	14.06	0.00
GPR183	G protein-coupled receptor 183	chr13	12.76	0.00
SGK1	serum/glucocorticoid regulated kinase 1	chr6	12.30	0.03
CD74	CD74 molecule, major histocompatibility complex, class II	chr5	11.43	0.00
FGL2	fibrinogen-like 2	chr7	9.85	0.00
HLA-DRB1	major histocompatibility complex, class II, DR beta 1	chr6	9.23	0.00
SRGN	serglycin	chr10	7.85	0.00
C8orf4	chromosome 8 open reading frame 4	chr8	6.93	0.03

E

Top 5 (of 24) male enriched terms by DAVID

ID	Genset name	DE genes in gene list	p-value
GO:0009058	biosynthetic process	GOT2, FASN, PCYT2, PBLD	0.001
R-HSA-2426168	Activation of gene expression by SREBF (SREBP)	SCD, DHCR7, FASN, SREBF2	0.004
hsa01212	Fatty acid metabolism	SCD, FASN, ACAT2, ACAA1	0.004
hsa01210	2-Oxocarboxylic acid metabolism	GOT2, BCAT2, IDH2	0.006
GO:0006629	lipid metabolic process	CEL, CPNE1, ACAT2, UGT2B7, PLCXD1,	0.006

F

Top 5 (of 124) female enriched terms by DAVID

ID	Genset name	DE genes in gene list	p-value
GO:0016045	detection of bacterium	HLA-DRB1, HLA-DRB4, HLA-A, HLA-B	0.002
GO:0032715	negative regulation of interleukin-6 production	HAVCR2, IRAK3, NCKAP1L, ZC3H12A, TNFAIP3	0.002
GO:0045028	G-protein coupled purinergic nucleotide receptor activity	P2RY12, GPR34, P2RY14, GPR171	0.002
R-HSA-2172127	Innate Immune System	KLRC4-KLRK1, HLA-C, HLA-B, TYROBP, B2M	0.002
GO:0070098	chemokine-mediated signaling pathway	CXCR4, CCR1, CXCR6, CXCL8, CCL4L2, CCL4,	0.002

Figure 3.8 Differential gene expression analysis between male and female uncultured healthy-urothelial cells (UHUC)

Gender-associated differential gene expression was determined by carrying out LIMMA analysis between male and female UHUC HTA2.0 microarray data. Probes were considered differentially expressed (DE) when attaining ≥1.5-fold change (FC) and a P-value ≤ 0.05. **A)** Data is presented as a volcano plot of P-value vs fold change, with differentially expressed probes coloured blue for male DE probes and purple for female DE probes **B)** Heatmap of Z-scores for DE genes with hierarchical clustering of samples. **C)** Top 10 autosomal genes upregulated in males. **D)** Top 10 autosomal genes upregulated in females. **E)** Top 5 Terms identified by DAVID analysis on male DE gene list. **F)** Top 5 Terms identified by DAVID analysis on female DE gene list.

unlike in cultured NHUC, *ANXA6* or *MMP1* were not DE (Figure 3.8B; Appendix A-10). DAVID analysis showed that upregulated genes in male UHUC were associated with metabolic processes such as fatty acid metabolism, 2-oxocarboxilic acid metabolism, and the biosynthetic process, as well as activation of gene expression through upregulation of Zinc-finger proteins and the SREBP2 TF (Figure 3.8E). Metacore also showed that SREBP2 was the only TF upregulated in the male UHUC gene list, and was associated with upregulation of *FASN, SCD, DHCR7* and *SREBP2*-precursor (Appendix A-15). Interestingly, a previous study showed that increased FGFR3 in bladder cancer cell lines regulated sterol and lipid biosynthesis through increased expression of *SREBP2* and its downstream targets *SCD* and *FASN*, which promoted tumour growth and survival (Du *et al.*, 2012).

The most striking result from the LIMMA analysis in UHUC was the high number of immune-related genes that were DE in females. Of the 595 DE probes in females, 132 pertained to immune markers including components of the major histocompatibility complex (HLAs), interferons, chemokines, CD receptors, and more (Figure 3.8A & D; Appendix A-11). This was further reflected in the DAVID analysis which showed that genes upregulated in female UHUC were mainly associated with immune response pathways (Figure 3.8F). Many XCI genes that were upregulated in cultured female NHUC were not upregulated in UHUC, with only *DDX3X, STS,* and *XIST* found to be DE. Furthermore, the top autosomal genes DE in cultured NHUC were also not observed as DE in UHUC, including *UCA1*. Metacore also showed upregulation of BLIMP1 and RUNX3 transcription factors which have both been implicated in immune-related processes (Tellier *et al.*, 2016; Shan *et al.*, 2017) (Appendix A-15).

GSEA on the full HTA 2.0 gene list in UHUC revealed 32 gene-sets enriched in males compared to 179 gene-sets in females for gene-sets curated from GO, KEGG, Reactome and Hallmarks (Figure 3.9). For males, enriched pathways pertained to ribosome biogenesis, translational activity, and regulation of RNA, which coincides with the high number of DE snoRNAs and rRNAs following LIMMA, although these did not constitute part of the identified gene-sets (Figure 3.8C). Indeed, DAVID analysis of a gene-list consisting only of the snoRNAs upregulated in male UHUC showed that these do not constitute any of the gene-lists provided by the KEGG, Reactome, or GO databases, and therefore enrichment of ribosome-related pathways that include snoRNA is neglected by GSEA. Female UHUC showed enrichment of numerous immune-related gene-sets by GSEA, which complemented the high number of immune-related DE genes identified by LIMMA (Figure 3.8D).

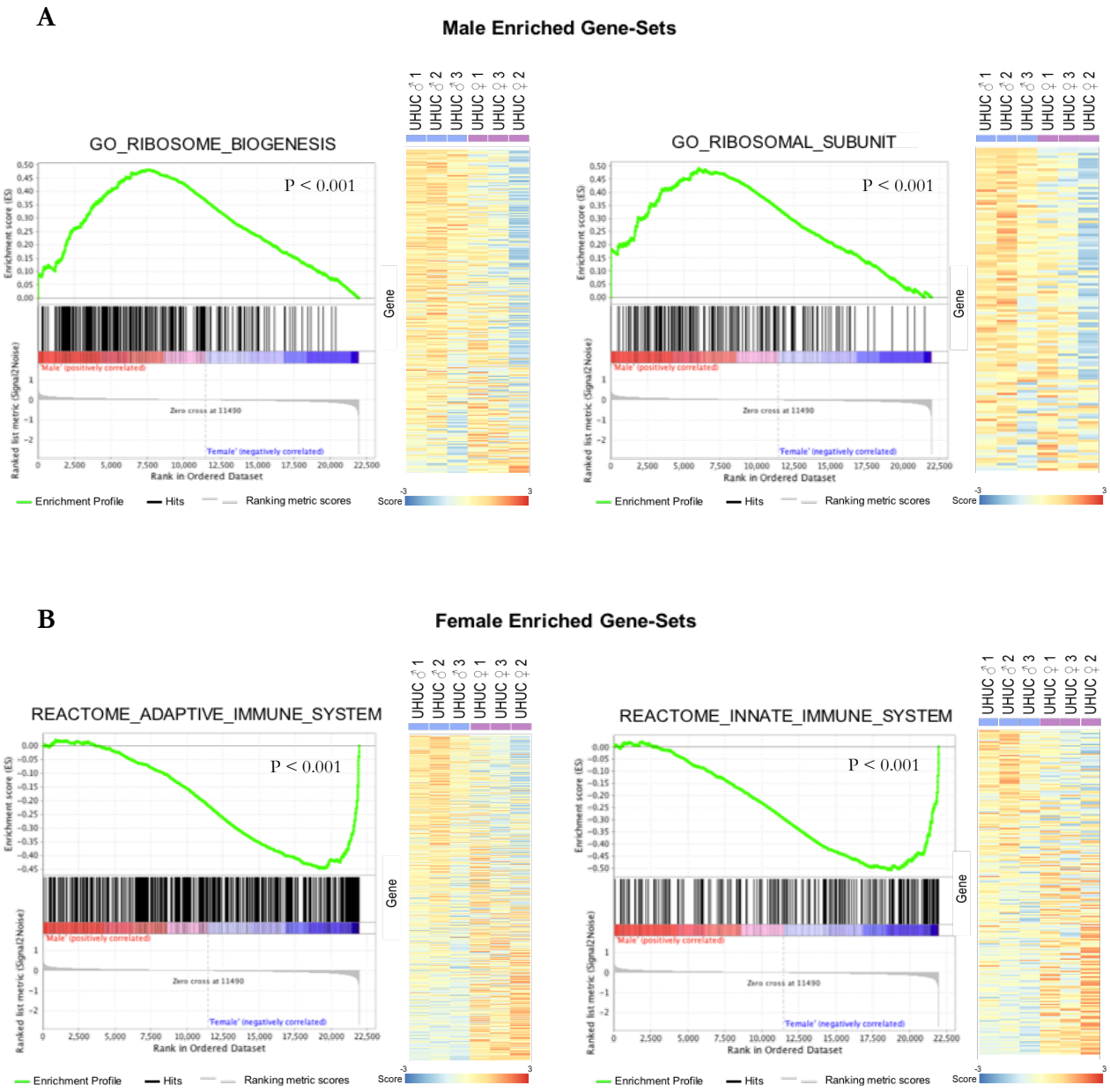


Figure 3.9 GSEA between male and female UHUC

All probes from the HTA 2.0 microarray were used to carry out GSEA for male vs female UHUC for gene-sets obtained from MSigDB. **A)** Male enriched gene-sets. **B)** Female enriched gene sets. Enriched genes-sets were used to generate heatmaps of Z-scores with hierarchical clustering of samples, and are shown to the right of their respective GSEA plot. Gene sets were considered enriched when a P-value ≤ 0.05 was attained.

Heatmaps with hierarchical clustering of samples showed grouping of samples of each gender. However, a single female UHUC (UHUC ♀ 2) appeared as an outlier in male enriched groups and may have been promoting the gender-associated enrichments of these gene-sets, although this was not the case for female enriched gene-sets where a clearer divide between males and females was shown.

Overall the expression profiles obtained from UHUC demonstrated increased regulation of ribosome biogenesis and RNA regulation in males, and a high immune-like state in female UHUC. For females, these results may indicate possible infiltration of immune cells into the urothelium, possibly originating from sample preparation. The results obtained from UHUC largely did not coincide with those from cultured NHUC. Indeed, DAVID and GSEA showed immune-related processes were associated more with male in NHUC, and females in UHUC. However, consistency between NHUC and UHUC was shown for DE genes located on sex chromosomes and the upregulation of snoRNAs in males.

3.2.3 Gender-related transcriptome analysis in TaG2 bladder tumours

A separate project undertaken in the lab concerned identifying genomic subtypes of non-muscle-invasive bladder cancer (NMIBC) and this utilised microarray analysis to generate expression profiles for 102 stage Ta grade 2 (TaG2) tumours (Hurst *et al.*, 2017 and unpublished data). This therefore provided an additional opportunity to assess differential gene expression between genders in bladder cancer, and to compare differential expression related to common chromatin-modifier mutations, in particular those pertaining to *KDM6A* and COMPASS-like complexes.

Of the 102 TaG2 samples analysed, 68 were from males and 34 from females. Differential expression analysis using LIMMA identified 72 DE probes with ≥1.5-fold change (FC) and a P-value ≤0.05. Of these, 28 probes were higher in male tumours (5 autosomal, 22 chrY and 2 chrX) and 44 probes were higher in female tumours (32 autosomal, 12 chrX) (Figure 3.10A). The previously mentioned 10 chrY genes were again among the most DE genes in male TaG2 following LIMMA analysis. Only 6 DE genes were not located on chrY: *STAG2, DAB1, DHCR24, TMEM97, LOC283788,* and *FRG1BP* (Figure 3.10C). These results are therefore in contrast to previous analyses in NHUC, most noticeably the lack of DE snoRNAs, rRNAs, *ANXA6* or *MMP1*. Many of the XCI genes that were DE in NHUC were no longer DE in TaG2, with only *XIST* and *STS* showing consistency. However, other XCI that have been documented in the literature (Tukiainen *et al.*, 2017) were DE in TaG2, including *ARSD, MXRA5, PNPLA4,* and *PUDP*. As with male tumours,

the autosomal genes commonly DE in NHUC were not DE in NMIBC. However, three immunogenic-related genes (*JCHAIN, IGHV3,* and *A2M)* and a number of metabolic-related genes, including cytochrome P450s, heparan sulphates and G-protein coupled receptors, were upregulated in female tumours.

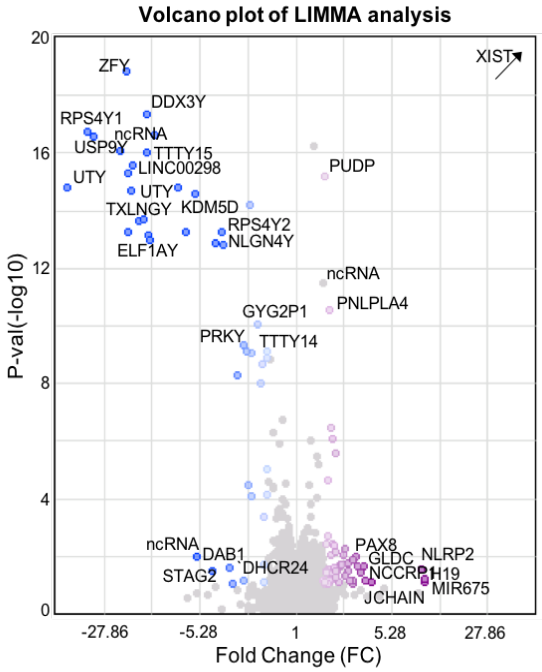
DAVID analysis on gender-associated DE genes in male TaG2 identified 11 pathways that were mainly related to ribosome and translation activity, and histone demethylase activity, and were attributed to upregulation of *RPS4Y2* and *RPS4Y1*, and *UTY* and *KDM5D* respectively, all of which are located on chrY. For gender-associated DE genes in female TaG2, DAVID identified 22 terms that were mainly related to IgA and immune response (attributed mainly to *JCHAIN, IGHA1,* and *IGHV-23),* metabolic pathways involving *CYP24A1* (response to vitamin D, and oxidative stress), and cell adhesion. Metacore analysis did not identify any upregulated transcription factors in either male or female TaG2 tumours.

GSEA only identified 12 gene-sets that were enriched in males compared to 175 gene-sets in females from gene-sets curated from GO, KEGG, Reactome, and Hallmarks (Figure 3.11). For males, enriched pathways predominately pertained to ribosome biogenesis and translational activity, although gene-sets for post-translational modification of proteins (most likely involved in ubiquitination) were also observed. The more numerous gene-sets observed in female tumours were predominantly immune (pertaining to leukocyte and lymphocyte regulation, immune response, B- and T-cell activation, cytokine production and phagocytosis) and development (pertaining to placental, mammary gland, embryonic, retina, skeletal system, sensory organs and more) related gene-sets. However, hierarchical clustering of gene-lists obtained from these gene sets failed to separate male and female samples, and enrichment of genes in these gene-sets could not be seen in heatmaps ordered by gender (Appendix A-14).

3.2.4 Differential gender-associated response to mutations in TaG2 tumors

Another attempt to uncover biological differences associated with gender was to compare differential gene-expression profiles that occur when acquiring mutations in genes of interest (GOI) in bladder cancer. The hypothesis was that each gender should show similar differential gene expression profiles upon acquiring the same gene mutation, similar to ageing in the bladder detrusor of mice (Kamei *et al.*, 2018). The aforementioned NMIBC samples also included a subset of 49 samples that were also used for exome sequencing, and this study took advantage of this cohort by carrying out differential gene expression analysis between mutant and wild type (WT) GOI for each gender separately.

A



C

The 6 male DE genes not on ChrY

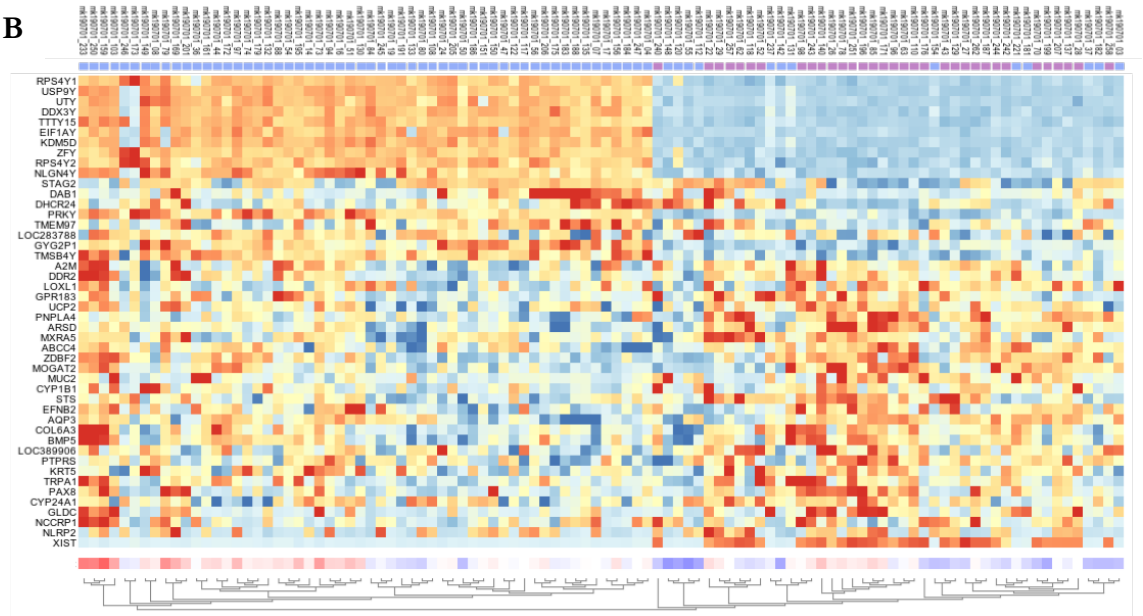
Symbol	Gene name	Chr	FC	p-value
STAG2	stromal antigen 2	chrX	2.99	0.01
DAB1	Dab, reelin signal transducer, homolog 1 (Drosophila)	chr1	2.48	0.05
DHCR24	24-dehydrocholesterol reductase	chr1	2.14	0.04
TMEM97	transmembrane protein 97	chr17	1.57	0.04
LOC283788	FSHD region gene 1 pseudogene	chrUn	1.53	0.00
FRG1BP	FSHD region gene 1 family member B, pseudogene	chr20	1.50	0.00

D

Top 10 (of 32) female DE autosomal genes

Symbol	Gene name	Chr	FC	p-value
H19	H19, imprinted maternally expressed transcript (non-protein coding)	chr11	6.42	0.04
NLRP2	NLR family, pyrin domain containing 2	chr19	6.14	0.01
JCHAIN	joining chain of multimeric IgA and IgM	chr4	3.01	0.04
NCCRP1	non-specific cytotoxic cell receptor protein 1 homolog (zebrafish)	chr19	2.71	0.03
GLDC	glycine dehydrogenase (decarboxylating)	chr9	2.55	0.02
CYP24A1	cytochrome P450, family 24, subfamily A, polypeptide 1	chr20	2.44	0.01
PAX8	paired box 8	chr2	2.40	0.00
TRPA1	transient receptor potential cation channel, subfamily A, member 1	chr8	2.30	0.00
KRT5	keratin 5, type II	chr12	2.30	0.04
HS6ST3	heparan sulfate 6-O-sulfotransferase 3	chr13	2.18	0.04

B



E

Top 5 (of 10) male enriched terms by DAVID

ID	Genset name	DE genes in gene list	p-value
GO:0006413	translational initiation	EIF1AY, DDX3Y, RPS4Y1	0.006
GO:0032452	histone demethylase activity	UTY, KDM5D	0.020
R-HSA-3214842	histone demethylase activity	UTY, KDM5D	0.027
R-HSA-72695	Formation of the ternary complex, and subsequently, the 43S complex	RPS4Y2, RPS4Y1	0.029
GO:0022627	cytosolic small ribosomal subunit	RPS4Y2, RPS4Y1	0.031

F

Top 5 (of 22) female enriched terms by DAVID

ID	Genset name	DE genes in gene list	p-value
GO:0034987	immunoglobulin receptor binding	IGHV3-23, JCHAIN, IGHA1	0.001
GO:0072562	blood microparticle	A2M, IGHV3-23, JCHAIN, IGHA1	0.002
GO:0071751	secretory IgA immunoglobulin complex	JCHAIN, IGHA1	0.005
GO:0071752	secretory dimeric IgA immunoglobulin complex	JCHAIN, IGHA1	0.005
GO:0071748	monomeric IgA immunoglobulin complex	JCHAIN, IGHA1	0.005

Figure 3.10 Differential gene expression analysis in male and female TaG2 tumours

Gender-associated differential gene expression was determined by carrying out LIMMA analysis between male and female TaG2 HTA2.0 microarray data. Probes were considered differentially expressed (DE) when attaining ≥1.5-fold change (FC) and a P-value ≤ 0.05. **A)** Volcano plot of P-value vs fold change, with differentially expressed probes coloured blue for male DE probes and purple for female DE probes **B)** Heatmap of Z-scores for DE genes with hierarchical clustering of samples. **C)** Top 10 autosomal genes upregulated in males. **D)** Top 10 autosomal genes upregulated in females. **E)** Top 5 terms identified by DAVID analysis on male DE gene list. **F)** Top 5 terms identified by DAVID analysis on female DE gene list.

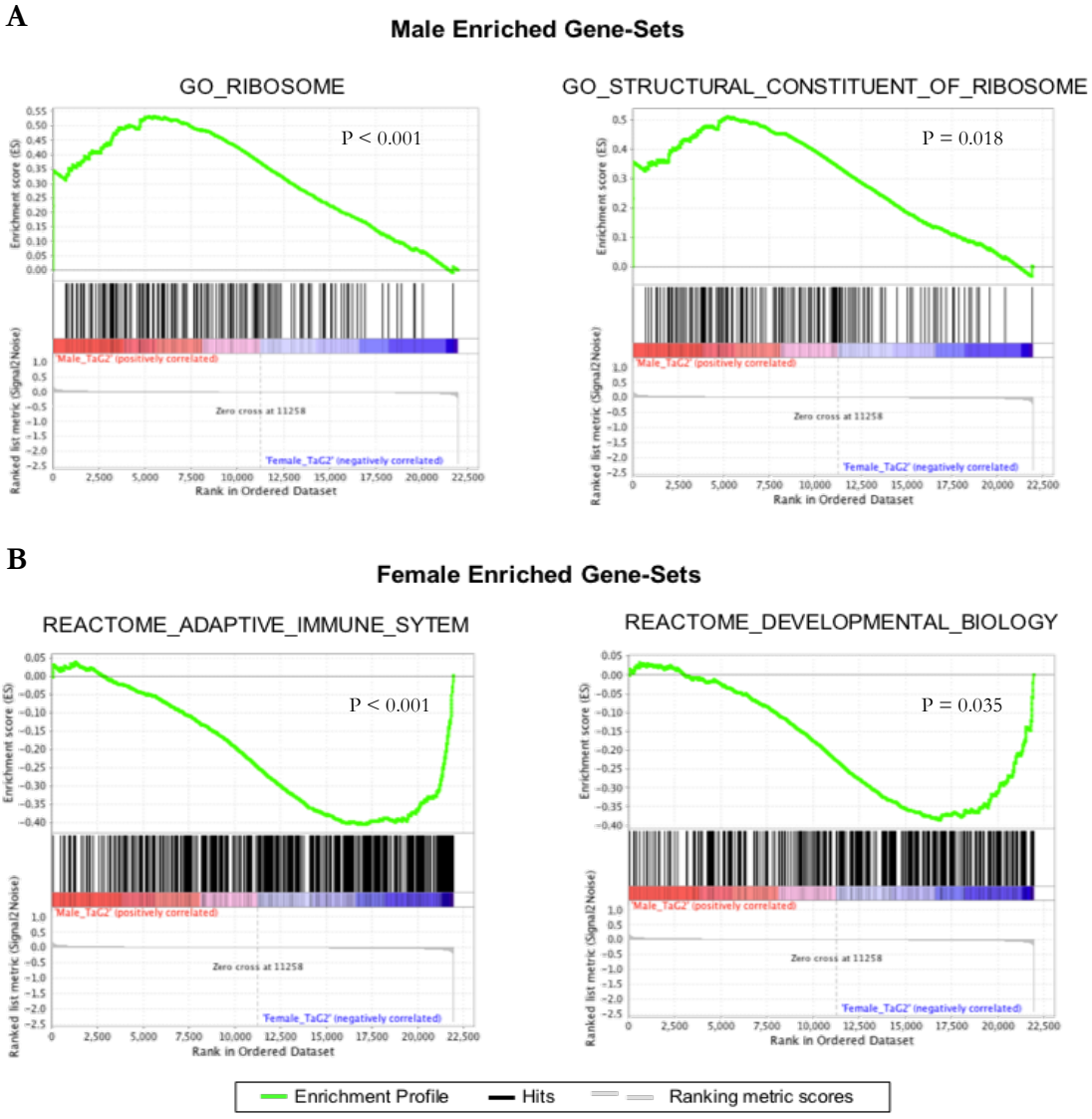


Figure 3.11 GSEA between male and female TaG2 tumours

All probes from the HTA 2.0 microarray were used to carry out GSEA for male vs female UHUC for gene-sets obtained from MSigDB. **A)** Male enriched gene-sets. **B)** Female enriched gene sets. Gene sets were considered enriched when a P-value ≤0.05 was attained. Enriched genes-sets were also used to generate heatmaps of Z-scores with hierarchical clustering of samples, and can be found in Appendix A -14.

The subset of 49 NMIBC samples with accompanying microarray and exome sequencing data included 31 male and 18 female TaG2 tumour samples. An oncoplot for the most commonly mutated genes in this cohort was generated and showed that *FGFR3, PIK3CA,* and *KDM6A* were amongst the most commonly mutated genes followed by multiple chromatin-modifying proteins including *KMT2C, KMT2D, STAG2, EP300, ARID1A,* and *CREBBP* (Figure 3.12).

KDM6A was the third most commonly mutated gene in this cohort, with 13 males and 12 females showing mutation in this gene (KDM6Amut); the remaining 18 male and 6 female TaG2 samples were therefore WT for KDM6A (KDM6AWT). Differential gene expression analysis of KDM6Amut and KDM6AWT samples within gender groups identified 288 DE genes in male-KDM6Amut, and 369 DE genes in female-KDM6Amut (Figure 3.13A & B). However, only 31 genes were identified as differentially expressed in both of the analyses, showing that samples with mutations in KDM6A predominately display differential expression of different genes in each gender (Figure 3.13B). The top DE genes in male-KDM6Amut included SERPING1, KRT13, CRISP3, CSTA, and UTY, and DAVID analysis showed that DE genes in male KDM6Amut were associated with 50 gene-sets that were mainly related to membrane assembly (attributed to DE of cytochrome P450), histone demethylase activity (attributed to DE of KDM6A, UTY, ARID5B, and KDM5D), and synapse regulation (attributed to DE of protocadherins) (Figure 3.13C). The top female-KDM6Amut DE genes include UPK1A, GDA, NTSE, MIR31HG, and NLRP2, and DAVID analysis showed that DE genes in female KDM6Amut were associated with 141 gene-sets that were predominately associated with chromatin and nucleosome regulation (driven by DE of the HIST1 gene cluster), but also cell-to-cell communication and adhesion, and cell cycle regulation (Figure 3.13D). DE genes identified in both male and female Ta-KDM6Amut included KDM6A, ANXA1, CD44, KRT8, and DUXAP10.

These aforementioned gene expression differences in KDM6Amut were assigned as gender-specific if they were identified as DE by LIMMA in only one gender (Figure 3.13B). However, a closer look at the heatmap shows that there are subgroups of DE genes found in both males and females that were considered as gender-specific in this analysis Figure 3.13A. For instance, a subgroup of KDM6Amut and KDM6AWT males had increased expression of a set of genes that were highly upregulated in female KDM6Amut, but were considered as DE only in female KDM6Amut (Figure 3.13A, box v). Nevertheless, these results show that gene expression changes upon mutation of KDM6A in TaG2 are generally different between genders.

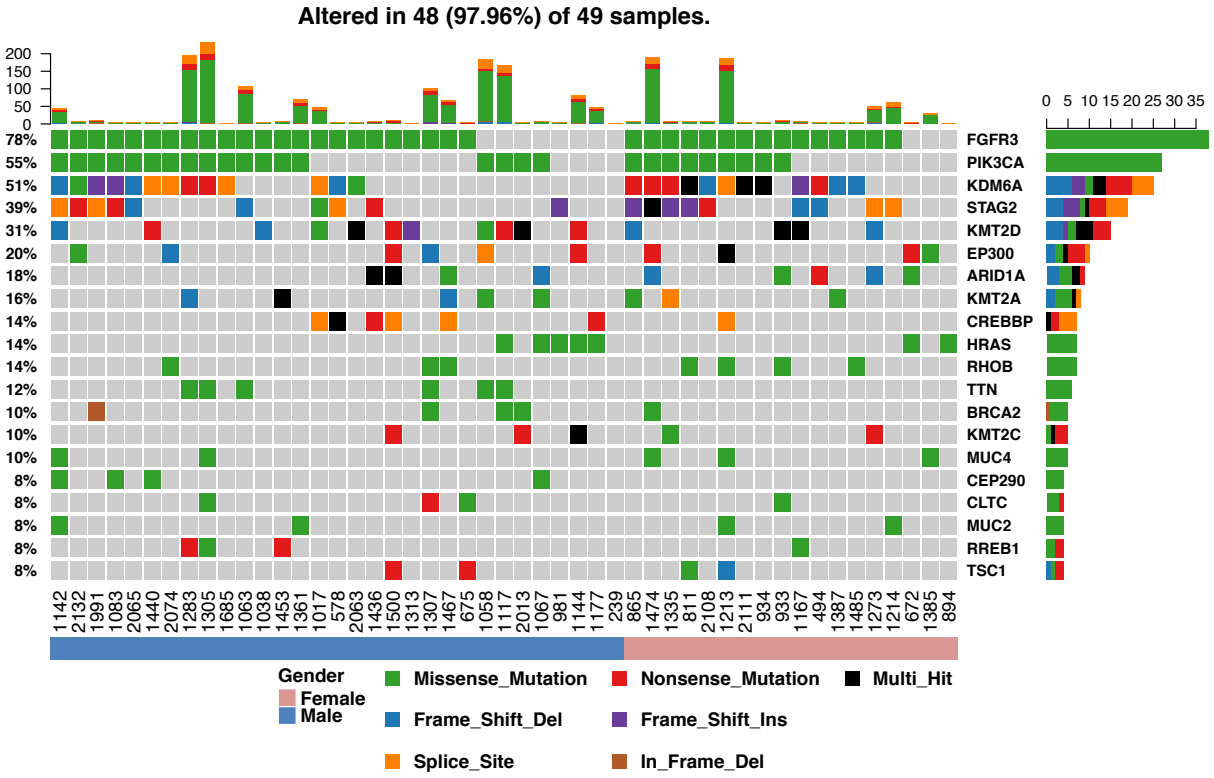
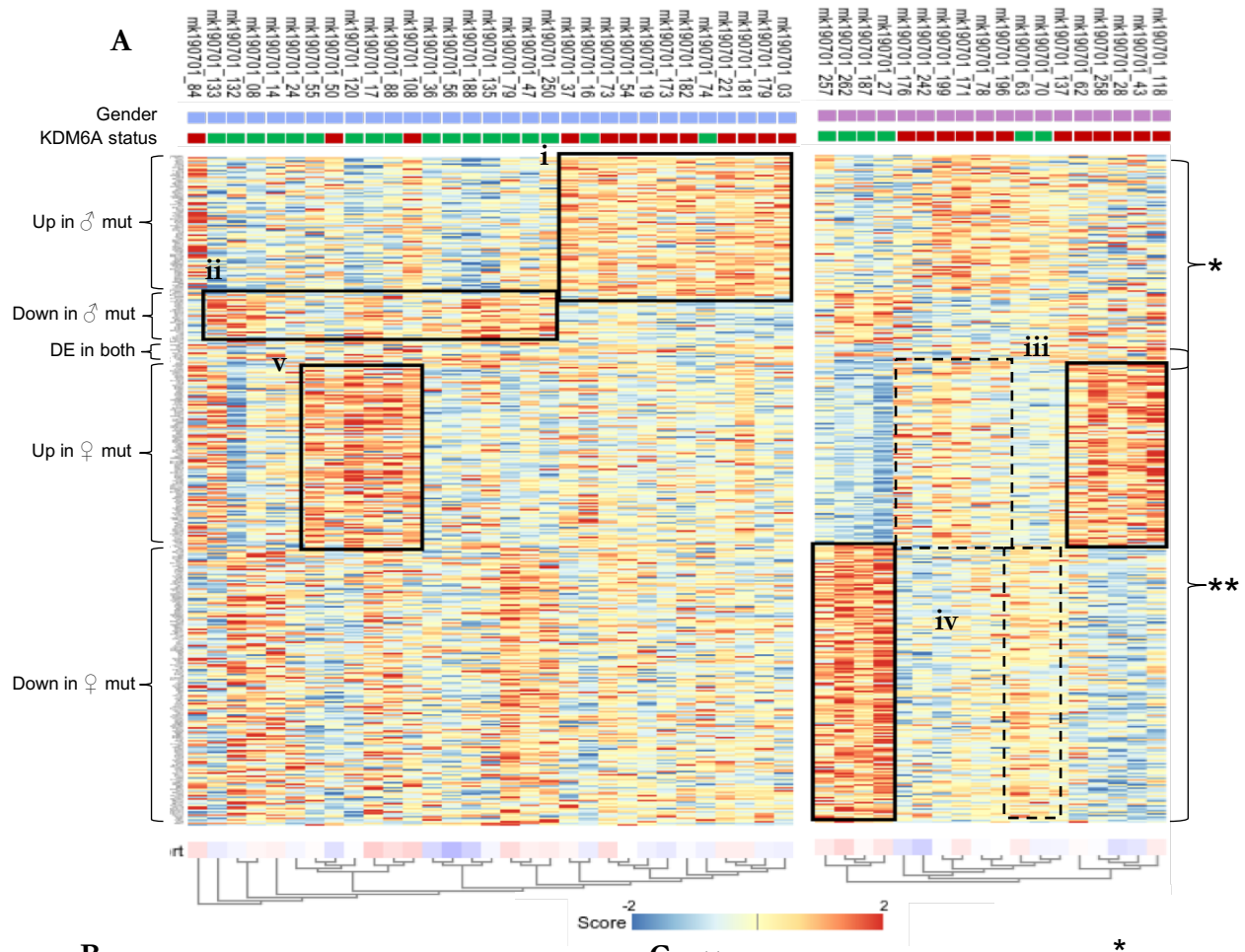


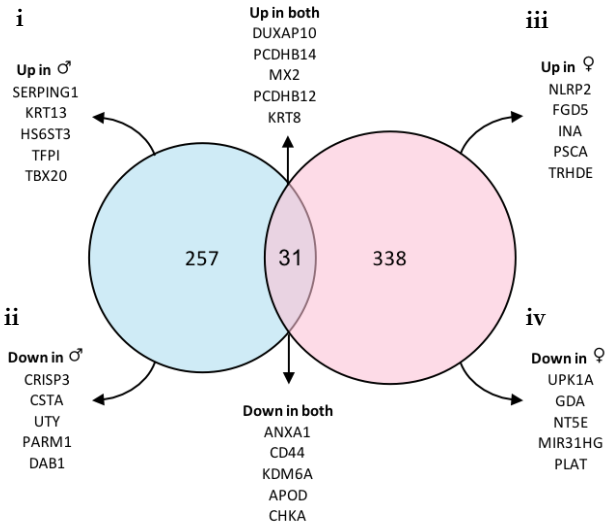
Figure 3.12 Oncoplot of common mutations in TaG2 tumour samples.

Oncoplot showing distribution of mutations in 49 TaG2 bladder tumours. Data are derived from exome sequencing of 31 male (blue) and 18 female (pink) TaG2 samples which were also used for microarray analysis. Left panel shows distribution of gene alterations by sample; missense (green), nonsense (red), frame-shift deletion (blue), frame-shift insertion (purple), splice site (orange), in-frame deletion (brown), and multi-hit (black) mutations are shown. Grey indicates no identified mutations. Right panel histogram, shows percentage of samples with confirmed hits for each gene. Top panel bar chart, shows the number and types of total mutations for each TaG2 sample. Colour coding is consistent throughout.

GSEA was carried out between *KDM6A*^{mut} and *KDM6A*^{wt} samples within gender groups, and showed only 3 gene-sets enriched in male-*KDM6A*^{mut} and 8 gene-sets enriched in female-*KDM6A*^{mut} with a P-value ≤ 0.05, all of which were gene-sets curated from the GO database (Figure 3.14). For male-*KDM6A*^{mut} this included gene-sets pertaining to hexosyl-transferases, retinoic acid response, and postsynaptic membrane. However, when allowing for enriched gene-sets with a P-value ≤ 0.1, gene-sets related to RNA polymerase II and the transcription factor complex were also identified for male-*KDM6A*^{mut}. Enriched female-*KDM6A*^{mut} gene-sets pertained to chromatin architecture and regulation, and the transcription factor complex. Therefore, if allowing for enriched gene-sets with a P-value ≤ 0.1, commonalities exist between genders upon mutation of *KDM6A* with regard to regulation of the transcription factor complex. However, differences still persisted, predominately with respect to differential regulation of chromatin in female-*KDM6A*^{mut} which was not observed in male-*KDM6A*^{mut}.

These results show that changes in gene expression profiles for TaG2 tumours with mutations in *KDM6A*, are different in males and females. This is particularly notable in the differences in genes that regulate chromatin architecture and transcription in female *KDM6A*^{mut}, whereas male *KDM6A*^{mut} have a more varied DE gene set. Similar differential responses to mutations between genders were also observed for other commonly mutated genes in TaG2, including *STAG2, PIK3CA,* and *FGFR3*, where only 80 out of 729, 28 out of 521, and 66 out of 1089 DE genes were shared between gender groups respectively (Appendix A-16).





C Top 5 (of 50) male enriched terms by DAVID *

ID	Geneset name	DE genes in KDM6A mut Males	p-value
GO:0031090	organelle membrane	CYP3A5, CYP4X1, CYP3A7, FA2H, TFPI, EPHX1, CYP4F3, CYP4B1	0.000
R-HSA-3296197	Hydroxycarboxylic acid-binding receptors	HCAR3, HCAR2, HCAR1	0.000
GO:0007416	synapse assembly	CEL, PCDHB5, NLGN4Y, PCDHB14, PCDHB13, SLITRK6	0.001
GO:0016324	apical plasma membrane	OCLN, DSG2, SLC12A2, BST2, MUC20, ANXA1, DUOX1, IGFBP2, AMOTL1, CEACAM1, SLC46A1	0.002
GO:0032452	histone demethylase activity	KDM6A, UTY, ARID5B, KDM5D	0.002
GO:0010951	negative regulation of endopeptidase activity	BST2, SERPINB8, TFPI, SERPING1, TIMP4, CSTA, A2ML1	0.003

D Top 5 (of 141) female enriched terms by DAVID **

ID	Geneset name	DE genes in KDM6A mut Females	p-value
GO:0001666	response to hypoxia	NOX4, PLAT, LDHA, ITGA2, EGLN1, CXCL12, DDIT4, CITED2, ADM, PLOD2, PTK2B, ABAT, ANGPT2, ANGPTL4	0.000
GO:0007155	cell adhesion	PCDHB3, EFNB2, PCDHB4, FERMT1, ITGB5, ITGA2, CNTNAP3B, PCDHB12, AJAP1, PCDHB11, CXCL12, LAMA1, AZGP1, TNFAIP6, CD44, SORBS2, WISP3, CX3CR1, TGFBI, CNTN1, CHL1	0.000
R-HSA-2299718	Condensation of Prophase Chromosomes	CCNB1, HIST1H3J, CDK1, HIST1H2BB, HIST1H2BM, HIST1H4L, HIST1H3D, HIST1H4D, HIST1H3I	0.000
GO:0000786	nucleosome	HIST1H3J, HIST2H2AB, HIST1H2BB, HIST1H2BM, HIST1H4L, HIST1H3D, HIST1H4D, HIST1H3I	0.001
GO:0098641	cadherin binding involved in cell-cell adhesion	HIST1H3J, LIMA1, LDHA, SHTN1, DIAPH3, ANXA1, PFKP, NOTCH3, CCNB2, PKP2, CAPG, HIST1H3D, TJP2, HIST1H3I	0.001

Figure 3.13 DE analysis for KDM6A^{mut} vs WT male and female TaG2 tumours.

LIMMA analysis was carried out within gender groups between *KDM6A*^{mut} and WT TaG2, and used to create a *KDM6*A^{mut} DE gene-list which consisted of genes up in male *KDM6A*^{mut}, genes down in male *KDM6*A^{mut}, genes up in female *KDM6A*^{mut}, genes down in female *KDM6A*^{mut}, and genes DE in both male and female *KDM6A*^{mut}. **A)** This *KDM6A*^{mut} DE gene-list was then used to generate a heatmap with hierarchical clustering of samples. Boxed regions indicate genes up in i) male *KDM6A*^{mut}, ii) male *KDM6A*^{WT}, iii) female *KDM6A*^{mut}, iv) female *KDM6A*^{WT}, and v) genes up in a subset of male TaG2 that were also up in female KDM6A^{mut}. Samples are labelled by gender (males, blue; females, pink), and *KDM6A* status (*KDM6*A^{WT}, green; *KDM6*A^{mut}, red). **B)** Venn diagram of the *KDM6A*^{mut} DE gene-list shows the number of unique DE genes in each gender, and the number that were shared between genders. The top 5 upregulated and downregulated genes for each comparison and those shared are also displayed. **C-D)** DAVID analysis showing top 5 terms associated with the DE genes in *KDM6A*^{mut} male (**C)** and female (**D)** TaG2.

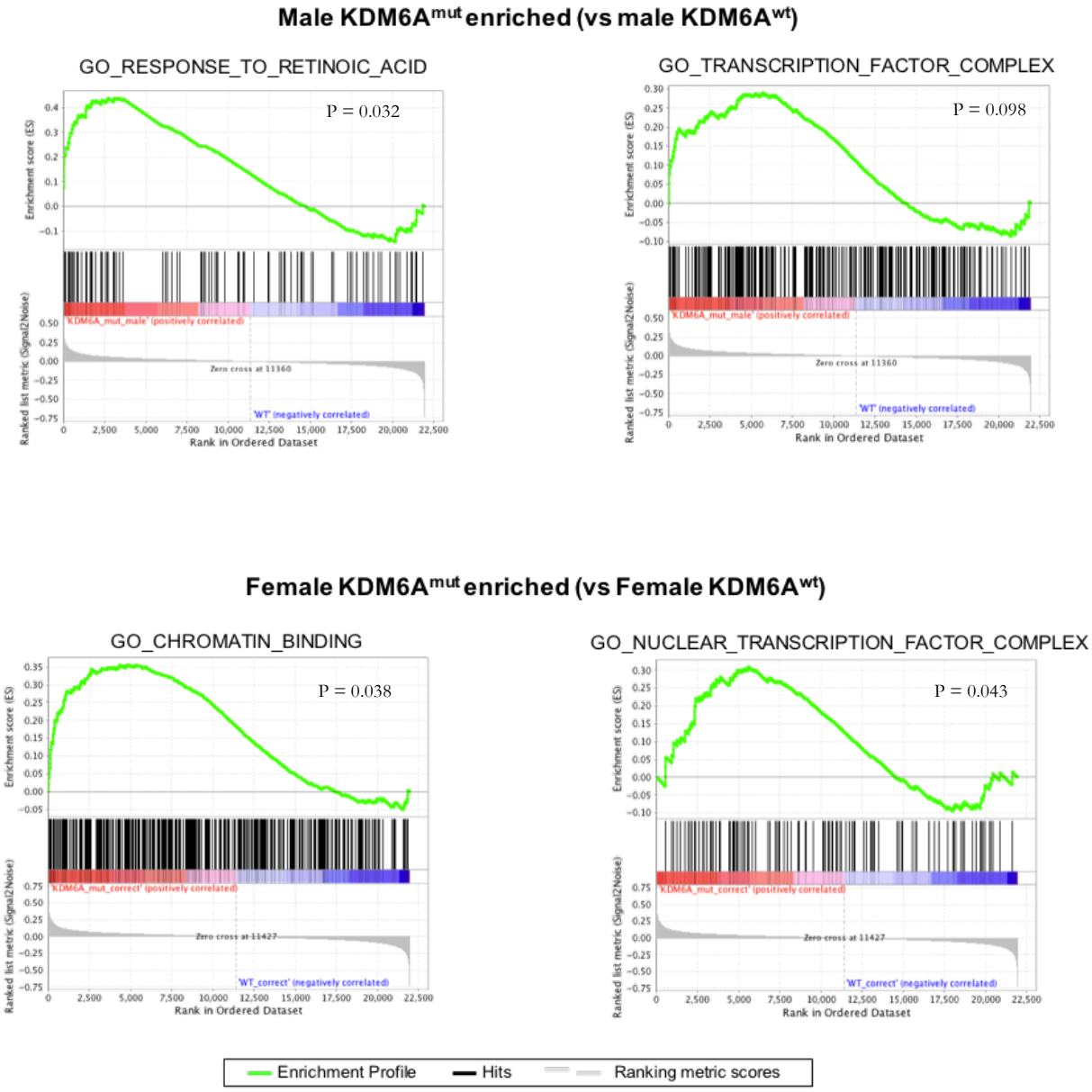


Figure 3.14 GSEA between KDM6A mutant and WT TaG2 tumours

All probes from the HTA 2.0 microarray were used to carry out KDM6A^{mut} vs WT TaG2 tumours, in males and female separately and using gene-sets obtained from MSigDB. **A)** Two gene-sets enriched in male KDM6A^{mut} vs male KDM6A^{WT} TaG2. **B)** Two gene-sets enriched in male KDM6A^{mut} vs male KDM6A^{WT} TaG2. Gene sets were considered enriched when a P-value ≤0.05 was attained, although top right GSEA has P-value >0.05.

Discussion

This Chapter set out to determine whether there were any intrinsic differences in gene expression between healthy male and healthy female urothelial cells that might promote the biases in incidence observed in bladder cancer. As sex differences primarily stem from inequality in expression of genes located on the sex chromosomes, the study predominately focused on identifying the few autosomal genes that may be differentially expressed (Arnold, 2017), although upregulated chrY and XCI genes were documented. Due to limited availability of any one sample type, different NHUC types were used for analysis, and included immortalised, non-immortalised, and uncultured cells. The different models in themselves resulted in a degree of diversity and generated distinct expression profiles by microarray analysis, indicating that this approach may be limited.

Nevertheless, a common set of genes was identified as differentially expressed between genders in all comparisons. For males, this amounted to 12 genes that were upregulated in all NHUC comparisons, and constituted those previously identified in the literature (Gershoni and Pietrokovski, 2017), as well as two additional chrY genes, namely *PRKY* and *RP54Y2*. This observation also persisted into the DE analysis in stage Ta tumour samples where all 12 chrY genes remained upregulated in males. However, given that all these genes were identified as differentially expressed in no fewer than 40 distinct tissue types, they are likely to be of little interest to the transcriptomic profile of bladder, and should instead be considered as intrinsic to the transcriptome of the male phenotype (Gershoni and Pietrokovski, 2017). Nevertheless, as this paper considered only post-mortem donor samples, it likely that the findings in the bladder were only relevant to the detrusor muscle, making the results from this Chapter likely the first to identify these chrY genes as upregulated in male urothelium. Furthermore, loss of chrY is commonly observed in bladder cancers of all stages and grades, and is likely an early event in the evolution of the disease (Sauter *et al.*, 1995; Minner *et al.*, 2010), therefore implicating a loss of these 12 chrY genes in the early onset of bladder cancer.

Although no single species was identified as DE in all three NHUC comparisons, snoRNAs and rRNAs were ubiquitously upregulated in male NHUC and this was particularly apparent in TERT-NHUC and UHUC. snoRNAs can be classified into two families: box C/D snoRNAs and box H/ACA snoRNAs. The canonical functions of snoRNAs include the 2-O'-methylation (box C/D) and pseudouridylation (box H/ACA) of ribosomal and small nuclear RNAs respectively, although both types are also known to regulate alternative pre-mRNA splicing, recognise polyadenylation sites, and regulate chromatin remodelling

(McMahon *et al.*, 2015; Kufel and Grzechnik, 2019). Therefore, snoRNAs are able to indirectly affect countless cellular processes through transcription via alternative splicing, or through translation via ribosome biogenesis. Unsurprisingly, snoRNAs have been implicated in the tumorigenesis of multiple cancers. For instance SNORD50A and SNORD50B are commonly lost in cancers of the prostate, lung, liver, and skin (Siprashvili *et al.*, 2015). In lung cancer, snoRNAs were found to be upregulated compared to healthy lung, and 6 snoRNAs were linked to overall survival (SNORA47, SNORA68, SNORA78, SNORA21, SNORD28 and SNORD66) (Gao *et al.*, 2015). However, in each of these studies there was no relationship between snoRNA expression with age, race or gender (Siprashvili *et al.*, 2015; Gao *et al.*, 2015). One study in multiple sclerosis observed differential expression of small-non-coding RNAs (sncRNA) in patients that underwent acute cycles of relapse (neurological disability) and remission (recovery phase) (Muñoz-Culla *et al.*, 2016). The authors found that females showed more DE sncRNAs in both relapse and remission states compared to males, many of which included snoRNAs (Muñoz-Culla *et al.*, 2016).

The NHUC results of this study predominately showed DE of box C/D snoRNAs which also coincided with the DE of many rRNAs. This, combined with a lack of differential alternative splicing events and enrichment of gene sets pertaining to ribosomal biogenesis in UHUC, indicates that increased snoRNA expression is modulating the ribosome of male NHUC. As one may expect, hyperactivity of ribosomal biogenesis can result in increased protein biosynthesis (Pelava *et al.*, 2016). However, changes in the nucleolar structure of the ribosome can also result in differential translation of mRNAs under normal conditions, without increasing protein biosynthesis (Barna *et al.*, 2008). Other components that interact with snoRNAs to modulate ribosomal biogenesis, such as NOP58, NOP56, 15.5K, FBL and DDX21, were not shown to be DE in this study (Bustelo and Dosil, 2018). It is currently unknown whether upregulation of snoRNAs and rRNAs in this study may be promoting increased protein biosynthesis or differential translational regulation of mRNA. However, as the gender comparison in TaG2 tumours also showed enrichment of gene-sets pertaining to ribosome biogenesis and translational regulation, it may be hypothesised that the differential modulation of ribosomes in healthy males is promoting a tumorigenic phenotype that persists in NMIBC. It is noted that differential expression of snoRNAs was not seen in male TaG2 tumours, and although ribosomal-related gene-sets were enriched, snoRNAs did not constitute part of these lists. However, this may simply reflect an absence of snoRNAs in the MSigDB.

Female NHUC showed 7 DE genes in all three comparisons. Interestingly, only two of these (*XIST* and *DDX3X*) are located on chrX and are known to escape XCI, although

more XCI genes were identified within each separate NHUC analysis. The remaining 5 autosomal genes that are upregulated in female NHUC include *CYP24A1*, *DPYSL2*, *NLRP2*, *MT1E,* and *VCAN*. These seemingly unrelated genes have different roles and functions in cell biology and do not together constitute parts of any known gene-sets curated in MsigDB or Metacore. However, each has been implicated in tumorigenesis to varying degrees, including in bladder, and will be discussed below.

CYP24A1 belongs to the cytochrome P450 family of enzymes and is a hydroxylase of 25-hydroxyvitamin D_3, and therefore regulates the amount of active vitamin D in cells. Genetic variation in *CYP24A1* was linked to risk and aggressiveness in a Korean cohort of prostate cancers, and decreased expression inversely correlates with melanoma (Oh *et al.*, 2014; Brożyna *et al.*, 2014). Low levels of CYP24A1 expression have been reported in normal bladder cells and TERT-NHUC, where it was also shown to regulate levels of vitamin D (Hertting *et al.*, 2010). Further studies have shown that vitamin D receptor, which is positively regulated by vitamin D, is negatively correlated with overall survival, metastasis, and tumour severity in bladder cancer (Jóźwicki *et al.*, 2015). Therefore, a tumour-suppressive role of CYP24A1 can be inferred for bladder cancer. However, it is noted that DAVID identified genes associated with response to vitamin D in female TaG2 DE genes, which does not support a hypothesis of CYP24A1 preventing bladder cancer in females, but does support a hypothesis of preventing them from progressing to more aggressive forms.

DPYSL2 (also known as CRMP2) belongs to the CRMP family of proteins that were thought to be exclusively expressed in the nervous system, and plays an essential role in axonogenesis through modulating microtubule dynamics (Inatome *et al.*, 2000). Major roles for DPYSL2 outside of the nervous system are largely undocumented. However, DPYSL2 has been implicated as a possible biomarker due to increased expression in colorectal carcinoma (C.C. Wu *et al.*, 2008), and a recent study also documented decreased expression of DPYSL2 in a cohort of breast cancer tissues (Shimada *et al.*, 2014). *DPYSL2* also constituted part of hypoxic gene-sets in two separate studies (Chi *et al.*, 2006; Yang *et al.*, 2017), the latter of which included a set of 24 genes that compose a signature for hypoxia in muscle-invasive bladder cancer (MIBC).

Metallothioneins (MTs) are involved in metalloregulatory processes, and therefore protect cells against metal toxicity and oxidative stress, and participate in the regulation of cell growth, proliferation and differentiation (Si and Lang, 2018). In bladder, increased expression of MTs has been implicated in cisplatin resistance. Early results demonstrated that a cisplatin-resistant subline of RT112 cells had increased expression of the MT2 family of metallothioneins compared to their non-resistant parental counterpart, although there was

no observable change in the expression of the MT1 family (Siegsmund *et al.*, 1999). A later study corroborated with these findings by showing that patients with increased expression of MTs had a poorer survival rate, and demonstrated a significant disadvantage in response to cisplatin-based chemotherapy (Wülfing *et al.*, 2007). In a later paper, metallothionein 1E (*MT1E*) and versican (*VCAN*) were found to be positively associated with migration in a radial migration assay of 40 bladder cancer cell lines, and were then shown to be correlated with tumour stage and severity of disease in 62 bladder tumours (Wu *et al.*, 2008). The authors also demonstrated that knock-down of MT1E in the bladder cancer cell lines 253J♂ and T24♀ decreased wound healing capabilities and reduced cell proliferation (Wu *et al.*, 2008). A more recent study has shown that the MT-1 family of metallothioneins, including MTE1, are upregulated in non-differentiated NHUC and downregulated in differentiated NHUC (McNeill *et al.*, 2019). Furthermore, the authors also showed the induction of all MT-1 metallothioneins by cadmium, a carcinogen associated with bladder cancer. This increased expression of MT-1 due to cadmium ion exposure receded over time, but persisted for longer in differentiated NHUC (McNeill *et al.*, 2019). The increased expression of *MT1E* in female NHUC in indicates that these cells may be primed for de-differentiation and proliferation that contributes to more aggressive tumorigenesis compared to males.

VCAN is a structural proteoglycan component of the extracellular matrix and has been implicated in cell adhesion, migration and proliferation, and the invasive and metastatic signatures of many cancers. (Sotoodehnejadnematalahi and Burke, 2013). Moreover, VCAN is considered a central component of cancer-related inflammation due to its ability to bind to a plethora of chemokines, cell adhesion receptors and growth factor receptors. Further to the aforementioned results of *VCAN* involvement in bladder cancer migration and disease severity (Wu *et al.*, 2008), VCAN was shown to promote the metastasis of bladder tumours to the lung in a manner that is dependent on macrophage recruitment by the cytokine CCL2 to the tumour site (Said *et al.*, 2012). The authors ultimately showed that RhoGDI2 acts as a metastatic suppressor in bladder cancer by inhibiting VCAN expression, which in turn reduces inflammation of the tumour microenvironment (Said *et al.*, 2012). Although increased expression of VCAN was observed in NHUC in this study, no correlation was observed with the other components of this metastatic pathway. However, upregulation of VCAN does coincide with increased expression of other immunogenic markers in females, especially in the UHUC.

NLRP1 is a member of the NOD-like receptor family of proteins, which form multi-protein complexes that constitute an essential part of inflammasomes (Chavarría-Smith and Vance, 2015). Processing of IL-1β and IL-18 precursors depends on cytosolic caspase-1

activation, which is tightly regulated by NLRP-inflammasomes. These inflammasomes are themselves activated through the NF-ϰB pathway, TLR4/MuD88 signalling or through the assembly of NLRP3 multi-protein complexes that are facilitated by PAMPs (pathogen-associated molecular patterns), DAMPs (damage-associated molecular patterns), ATP and other toxins (Karan, 2018). The majority of the literature regarding NLRP-inflammasomes concerns NLRP3, which has been shown to promote tumorigenesis in head and neck cancers and oral squamous carcinoma, and supress metastasis in colorectal cancer. However, NLRP1 has been implicated in tumorigenesis of melanoma, by enhancing caspase-1 mediated inflammasome activation that supresses apoptosis and promotes metastasis (Zhai *et al.*, 2017). Perturbations of NLRP1 have also been observed in other diseases such as vitiligo, rheumatoid arthritis, systemic sclerosis and Crohn disease (Finger *et al.*, 2012).

Although these 5 autosomal genes that were upregulated in female NHUC are seemingly unrelated, a connection between them can be postulated. For instance, hypoxia commonly induces oxidative stress and inflammation, and is linked to a plethora of diseases including arthritis, sleep apnoea, neurodegeneration and cancer (Tafani *et al.*, 2016; Snyder *et al.*, 2017; McGarry *et al.*, 2018; Li *et al.*, 2018). Female NHUC showed increased expression of *MT1E* and *DPYSL2*, which are associated with hypoxia and oxidative stress, and increased expression of *VCAN* and *NLRP1,* which can be associated with inflammation. Interestingly, three of these genes (*VCAN, MT1E* and *DPYSL2*) have also been linked to muscle-invasive bladder cancer. (Wu *et al.*, 2008; Wülfing *et al.*, 2007; Said *et al.*, 2012; Yang *et al.*, 2017).

Under normoxic conditions, the hypoxia-inducible transcription factor (HIFα) is hydroxylated by the 2-OGDD dioxygenase EgLN isoenzymes, which marks it for degradation by the von Hippel-Lindau ubiquitin ligase complex. Hypoxia inactivates EgLN, resulting in HIFα stabilisation and its association with HIFβ, which ultimately allows for the transcriptional activation of genes that promote resistance to low oxygen levels (Kaelin, 2008). A recent study found that hypoxia induces H3K27me3 in embryonic kidney, breast cancer and neuroblastoma cell lines, and that H3K27me3 is increased in tissues known to be hypoxic such as the kidney, splenic germinal centres and thymus, but not in oxygen-rich tissues such as the heart (Chakraborty *et al.*, 2019). The study also found that the catalytic domain of KDM6A (also a 2-OGDD dioxygenase) has a high affinity for oxygen. This made KDM6A sensitive to hypoxia and resulted in increased H3K27me3 that prevented transcriptional reprogramming in differentiating cells (Chakraborty *et al.*, 2019). Interestingly, DAVID showed that DE genes in female $KDM6A^{mut}$ TaG2 were most associated with response to hypoxia, implicating a role for KDM6A in hypoxia during the tumorigenesis of female bladder cancer.

KDM6A is commonly mutated in bladder cancer, with a higher mutation rate in female Ta tumours compared to males (Hurst *et al.*, 2017). This was also observed in this study. Furthermore, loss of Kdm6a in female mice increased BBN-induced bladder cancer risk and mortality compared to normal female control mice. However, no significant differences in incidence and mortality were observed with loss of Kdm6a in male mice (Kaneko and Li, 2018). Transcriptomic analysis of 412 muscle-invasive bladder cancers showed that expression of *KDM6A* was higher in females compared to males (likely owing to *KDM6A* escaping XCI and the high number of $KDM6A^{WT}$ females in the study), but also showed that decreased *KDM6A* expression in females was associated with more advanced bladder cancer and predicted poor disease-free survival of BC patients (Kaneko and Li, 2018).

One may therefore speculate that female NHUC cells are primed for MIBC, as they have a mildly hypoxic phenotype even when cultured under normoxic conditions. Indeed this mildly hypoxic phenotype may be promoting an inflammatory response in the urethra, as observed by the increased expression of immunogenic markers in female UHUC. Furthermore, this more easily aquired hypoxic state would promote H3K27me3 transcriptional repression of KDM6A target genes due to oxygen deficiency inhbiting KDM6A demethylase activity. It can be further speculated that female NHUC predominantly require repression of KDM6A target genes to become cancerous, and this is achieved either through mutational perturbation of KDM6A or hypoxia-driven inhibition of KDM6A demethylase acitivity.

The gender comparisons made in the cohort of TaG2 tumour samples correlated with some of the findings in NHUC. Males showed enrichment of gene-sets relating to ribosome biogenesis and translational regulation in all NHUC. Furthermore, the enrichment of gene-sets for post-translational modifications regarding ubiquitination was also observed in male tumours, which coincided with results in TERT-NHUC. However, this may pertain to an early-stage tumorigenic phenotype as a result of the GOF-mutation of the h*TERT* promoter acquired early in the development of bladder cancer (Allory *et al.*, 2014), and possibly indicates a limitation of the TERT-NHUC model. Female tumours showed enrichment of immunogenic gene-sets and gene-sets related to development, again coinciding with findings in NHUC. Of the 5 DE autosomal genes, only *NLRP2* and *CYP24A1* were upregulated in female NMIBC. This is perhaps interesting given that these were the only two genes out of the five not documented to be associated with MIBC. The overlap of gene sets between NHUC and NMIBC indicated that biases present in healthy bladder cells persist into at least early-stage non-aggressive cancer.

The final comparison was between normal and mutated genes of interest within gender groups in NMIBC. These results showed very little overlap of DE genes upon acquiring mutations in *KDM6A, FGFR3, PIK3CA* and *STAG2* between genders. This indicates a differential phenotypic response between genders upon the acquisition of a given mutation in NMIBC. This may have been expected for mutations in *KDM6A*, as it has been speculated that this gene is likely to regulate distinct sets of genes in males and females (Hurst *et al.*, 2017; Kaneko and Li, 2018). However, as this differential response was observed with other genes, it cannot be inferred that this is specific to *KDM6A*. This is true especially when mutations of this gene co-occur with many others in this cohort, particularly *FGFR3*.

The greatest limitation of this gender comparison in NHUC is that of sample size. To compensate for the limited number of available samples, gender comparisons were carried out between several models, which in itself has limitations given intrinsic differences in proliferation, differentiation, and the immortalised state of each model. Furthermore, it was shown that variation in expression profiles of individual donors was greater than the variation attributed to gender differences, and throughout the analyses it could be seen that individual donors were biasing gene expression differences attributed to gender. This was particularly the case for K-TERT in the TERT-NHUC gender comparison, which showed distinct expression profiles throughout analyses and was the predominant driver of female-associated enriched gene sets. Indeed, when including TERT-NHUC into the analysis, K-TERT showed expression profiles more similar to NHUC rather than to their immortalised counterparts. Another example of an individual donor skewing results was also seen in UHUC, where one individual female UHUC showed a greater decrease in the expression of genes for male-associated enriched gene-sets than the other female UHUC. A solution to this issue may be to include a greater number of samples, which would reduce the biasing effects of individual donors. However, when carrying out a gender comparison on over 100 TaG2 samples, hierarchical clustering of samples using gene lists from supposed gender-associated enriched gene-sets failed to distinguish males from females. This suggests that simply increasing sample size may not drastically improve confidence in findings, as gender differences in the expression profile of the urothelium are only subtle when compared to the more pronounced differences between individuals. Nevertheless, consistent differences between genders were shown throughout the several models of healthy urothelium, and a large cohort of TaG2 bladder tumours showed DE genes in mutants were different for each gender. Together, these results suggest that although subtle, gender differences in the urothelium exist, and further research into bladder and bladder cancer should consider genders separately. For instance, female KDM6A^{mut} TaG2 tumours showed differential

expression of genes that regulate chromatin architecture whereas their male counterparts did not, and therefore it may be hypothesised that treatments targeting chromatin architecture (such as HDAC inhibition), would be more effective in female KDM6A^{mut} than in male KDM6A^{mut}.

To further improve confidence in the findings of this study, validation of the microarray data should be carried out by reverse-transcription coupled to qPCR (RT-qPCR). This would be done by correlating the expression of genes shown by microarray and RT-qPCR to improve confidence in genes called as differentially expressed validate relative fold change values. Throughout the analysis, a high number of genes were found as DE under the parameters used in this study. To reduce the high number of genes, a more stringent measure of differential expression should be used, such as increasing the relative fold-change threshold. By increasing the stringency for differential expression, fewer genes would be determined and more confidence would be placed in the fewer pathways identified by GSEA and DAVID, which would also be further validated by qPCR.

To summarise, the gender comparison of expression profiles was largely inconsistent between NHUC models, likely due to differences that derive from growing cells in culture compared to uncultured primary cells. Furthermore, the high number of immunogenic markers that were expressed in the uncultured female cells also calls into question the urothelial purity of these samples. The few differentially-expressed genes that were consistently identified pertain to the sex chromosomes, but a set of 5 seemingly unrelated autosomal genes were increased in females. A literature review of these genes allowed the inference of a possible hypoxic and inflammatory state in female NHUC that may promote the aggressive bias observed in female bladder cancer, and may in part explain the infiltration of immune cells in the uncultured samples. The upregulation of many snoRNAs and rRNA was also observed in male NHUC, coinciding with the enrichment of gene-sets relating to ribosome biogenesis and translational control. It may be speculated then that male NHUC show differential translational processing of mRNAs even when under similar conditions to female NHUC, or that the male NHUC may have increased protein biosynthesis. The increased ribosome biogenesis signature in males and immunogenic markers in females persisted in the differential expression analysis of NMIBC samples. Furthermore, each gender has a distinct set of differentially expressed genes when comparing the same gene mutations in NMIBC.

Chapter 4
Optimisation and preparation of an assay for transposase-accessible chromatin followed by sequencing (ATAC-seq) in TERT-NHUC

4.1 Introduction

Bladder cancer frequently shows mutations in chromatin modifying proteins (Gui *et al.*, 2011b; Weinstein *et al.*, 2014; Hurst *et al.*, 2017). These are often proteins that regulate enhancer activity, such as those in the COMPASS-like complexes, including KDM6A, KMT2C, and KMT2D (Hu *et al.*, 2013), the histone acetylases EP300 and CREBBP (Garcia-Carpizo *et al.*, 2018), the essential Cohesin complex component STAG2 (Ing-Simmons *et al.*, 2012; Dorsett and Merkenschlager, 2013), and the SWI/SNF chromatin remodelling complex component ARID1A (Lu *et al.*, 2017). Such mutations indicate common perturbations in the regulation of *cis*-regulatory (or enhancer) regions in bladder cancer.

Multicellular organisms contain multiple cell types that are the result of distinct transcriptional programmes arising from the same genome. This is largely attributed to the regulation of gene transcription by non-coding enhancer elements that act in a cell-type specific manner. These enhancers contain DNA motifs that act as binding sites for transcription factors (TFs), and regulate gene activation by bringing DNA-bound TFs into close proximity to promoters through the process of chromatin looping (Shlyueva *et al.*, 2014). Enhancers can therefore act on the promoters of target genes from distal intergenic regions or within the intronic region of genes, in a manner independent of distance or orientation. Active enhancer regions are devoid of nucleosomes and therefore allow the binding of TFs to accessible chromatin, with nearby nucleosomes also undergoing the post-translational modification of histone tails such as H3K4me1 and H3K27ac. The cell-type-specific identification of enhancers can therefore be determined by assaying for regions of chromatin accessibility (Thurman *et al.*, 2012), or the histone modifications H3K27ac (Creyghton *et al.*, 2010; Nord *et al.*, 2013) or H3K4me1 (Heintzman *et al.*, 2007). However, the most reliable results for enhancer identification are obtained when combining such assays (Fu *et al.*, 2018).

To date, the epigenetic landscape of normal urothelium remains poorly characterised and includes minor studies for genome-wide histone modification status, and no studies pertaining to chromatin accessibility (Dudziec *et al.*, 2012; Davis *et al.*, 2018). An aim of this project was to characterise the epigenetic landscape of telomerase-immortalised normal

human urothelial cells (TERT-NHUC) and identify possible gender-associated epigenetic predispositions to bladder cancer. This was to be done by using a combination of ChIP-seq for histone modifications and ATAC-seq for chromatin accessibility, and identifying differential *cis*-regulatory regions between genders. Only subtle differences were expected to be found between male and female TERT-NHUC for *cis*-regulatory regions located on autosomes, with the majority of differences expected to be seen on chrX and chrY (Yen and Kellis, 2015). Moreover, *cis*-regulatory regions shared between each gender will constitute many of the mis-regulated targets of the aforementioned chromatin modifiers that are commonly mutated in bladder cancer.

This project aimed to characterise the chromatin accessibility landscape of TERT-NHUC using the assay for transposase-accessible chromatin coupled to next-generation sequencing (ATAC-seq). ATAC-seq is a relatively new technique that utilises a hyperactive Tn5 transposase that simultaneously cuts DNA and incorporates sequencing adapters (Buenrostro *et al.*, 2013). By carrying out a partial digestion of chromatin with Tn5, regions easily accessible to the transposase are preferentially cut and incorporated with sequencing adapters, whereas steric hindrance by proteins that decorate DNA (e.g. nucleosomes and transcription factors) prevents digestion and the incorporation of adapters. NGS is then used to assay enrichment of accessible chromatin at a global level. As regions of chromatin accessibility often occur where TFs are bound to DNA, motif enrichment analysis of accessible regions can also be carried out on ATAC-seq data to infer the TFs bound to *cis*-regulatory regions (Buenrostro *et al.*, 2013; Li *et al.*, 2019). Given the reduced preparation time and lower cell number required for the assay, ATAC-seq has superseded its predecessors (FAIRE-seq, MNase-seq and DNase-seq) in becoming the preferred method for assaying chromatin accessibility.

As ATAC-seq had not previously been carried out in bladder, the protocol needed to be optimised for TERT-NHUC. A comparison of chromatin accessibility was then made between two male and two female TERT-NHUC lines, with the expectation of only subtle differences being observed on autosomes. Finally, by carrying out ATAC-seq in TERT-NHUC, a "normal" standard is set to which future studies regarding chromatin accessibility in bladder cancer cell lines can be compared.

4.2 Results

4.2.1 Optimisation

To date, ATAC-seq has not been carried out on bladder cell lines, including those used in this study, and as such the technique required optimisation prior to sequencing. The optimised protocol used is based on the original protocol outlined by Buenrostro *et al*, and the development of the final protocol will be discussed in the following sections (Buenrostro *et al*., 2013). Given that chromatin architecture consists of ~147bp DNA wrapped around nucleosomes, which themselves are separated by linker DNA (Ackermann *et al*., 2016), a periodic banding pattern reflecting sequential nucleosomal occupation across fragments of increasing lengths is expected when partially transposed chromatin is separated by gel-electrophoresis. For this reason, quality assessment (QA) of ATAC libraries was carried out using TapeStation, with later quality checks prior to sequencing also requiring qPCR and quantification. It must be noted that QA by TapeStation relies on the use of gel images and is therefore a subjective metric. Nevertheless, quality is determined by the banding pattern shows by Tapestation, where an effective digestion for ATAC-seq is seen as bands separated by a distance of ~150bp, resembling a ladder-like pattern. Moreover, due to time, material and costing constraints, optimisation was restricted to only C-TERT cells and used only the Ad2.1 indexing primer. It is acknowledged that unique optimal conditions may be required for each cell line and indexing primer, however QA prior to sequencing on all libraries would assess if this was appropriate.

4.2.1.1 Trial ATAC protocol with 25,000 and 50,000 whole cells or extracted nuclei

An initial trial ATAC was carried out using the protocol described by Buenrostro *et al* in their original paper (Buenrostro *et al*., 2013). This was carried out on 50,000 C-TERT cells. However, 25,000 C-TERT cells were also included to determine if cell number influences efficiency of transposition. The hypothesis was that if under-digestion was observed with 50,000 cells, then decreasing cell number would increase the ratio of Tn5 to chromatin and increase transposition. The protocol recommends a partial lysis of cells with a gentle lysis buffer (see Methods Chapter 2.22). Therefore in an attempt to increase chromatin accessibility to Tn5, a hypotonic lysis buffer that isolates cell nuclei and removes cellular debris prior to Tn5 transposition was also included in the experiment (Shechter *et al*., 2007; see Chapter 1.7). If under-digestion was observed for both 50,000 and 25,000 cells, the use of cell nuclei rather than partially lysed cells would increase Tn5 accessibility and promote transposition. This initial experiment therefore included ATAC on 25,000 and 50,000 C-

TERT cells prepared according to the Buenrostro *et al* protocol, and 25,000 and 50,000 C-TERT nuclei prepared using a hypotonic lysis buffer (Figure 4.1).

The protocol recommends visualising libraries using either gel electrophoresis with a 5% polyacrylamide gel or on a BioAnalyser. In this experiment libraries were visualised using electrophoresis with 5% TBE polyacrylamide gel, and with a TapeStation instead of a BioAnalyser (Figure 4.1A & C). An agarose gel electrophoresis with SYBR-green was also included (Figure 4.1B). Comparison of visualisation techniques immediately eliminated the use of electrophoresis with a 5% TBE polyacrylamide gel (Figure 4.1A), as DNA appeared only as smears across the gel rather than as concise bands as seen by TapeStation and agarose gel. This result from the 5% TBE polyacrylamide gel was repeated on three occasions (data not shown). In contrast, agarose gel electrophoresis and TapeStation both showed clear DNA bands. However the high-definition of the gel image and the availability of fragment-size distribution histograms make TapeStation the preferred technique (Figure 4.1C). Therefore, all further ATAC experiments used TapeStation as the primary library validation method.

With regard to chromatin transposition in this experiment, it was observed that both 25,000 cells and nuclei were over-digested (Figure 4.1B&C). This was particularly apparent for 25,000 nuclei in Figure 4.1C where a large and intense band was present at ~200bp (also observed as a large spike on the histogram) as well as another large band at ~30bp, compared to its experimental counterpart with 50,000 nuclei (Figure 4.1C, bottom). These large bands at ~200bp and ~30bp likely represent single nucleosomal DNA and linker DNA respectively, which in a partial digestion should not be the dominant bands. As previously stated, partially digested chromatin will produce a spread of DNA fragments encompassing multiple nucleosomes that is observed as a banding pattern by gel electrophoresis. A hint of such a banding pattern was observed for both 50,000 cells and nuclei, although it was weak and therefore indicated the need for further optimisation. Comparison of the cell and nuclei number showed that the use of 25,000 cells or nuclei was too few and led to over-digestion. Increased digestion was observed in nuclei compared to cells when numbers are matched. This was particularly apparent by the increased DNA at ~200bp in 25,000 nuclei vs 25,000 cells (Figure 4.1C). Due to increased digestion in nuclei and the prolonged preparation required for nuclei isolation, the previously established Buenrostro et al protocol for cell lysis was used in subsequent experiments

.

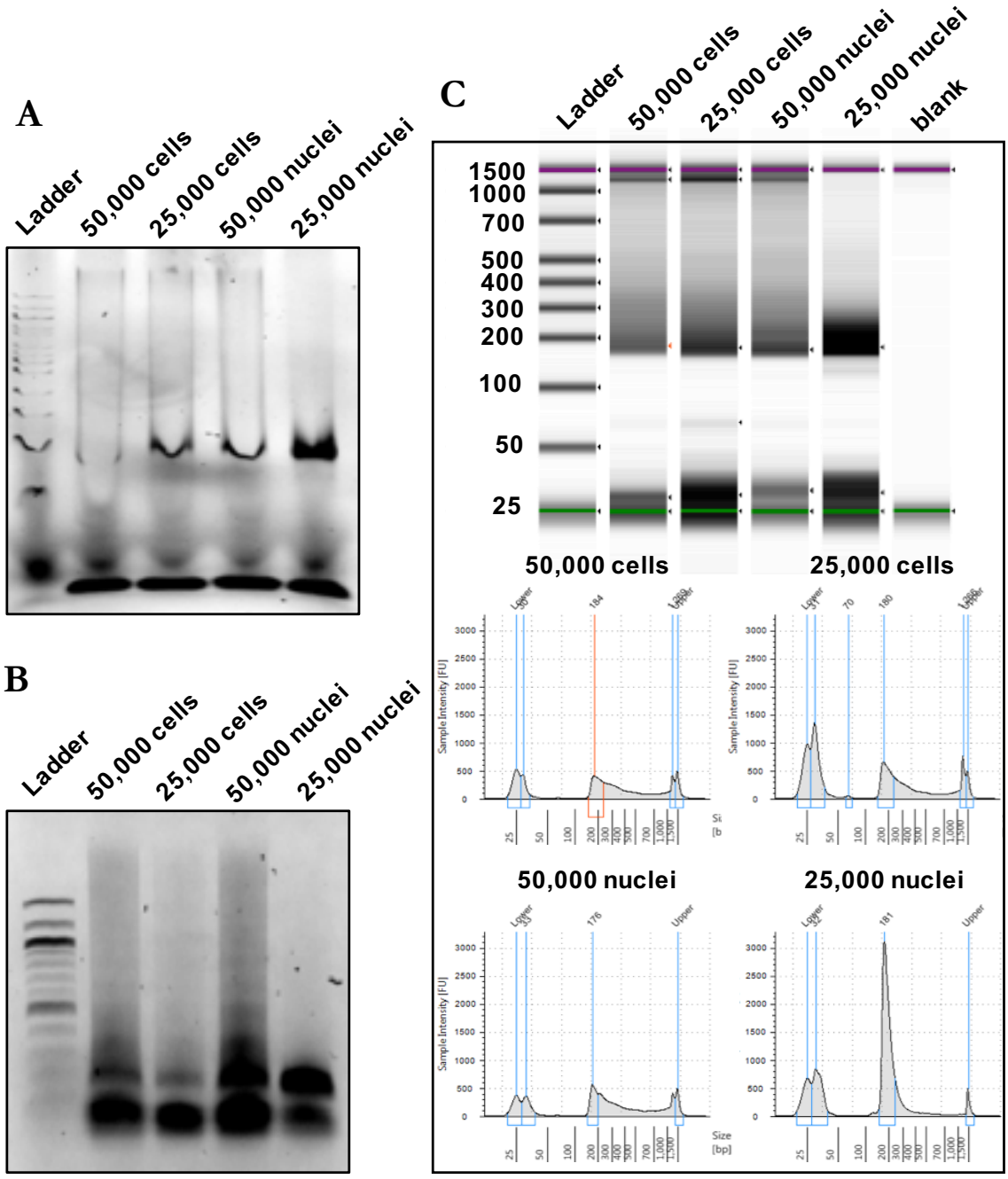


Figure 4.1 Initial ATAC optimisation experiment following the original Buenrostro *et al* protocol.

ATAC was carried out according to the original Buenrostro *et al* protocol on 50,000 and 25,000 cells prepared using the Buenrostro *et al* lysis buffer, and 50,000 and 25,000 nuclei prepared using the Shechter *et al* hypotonic lysis buffer and nuclei isolation protocol. Libraries were visualised using **A)** 5% polyacrylamide TBE gel electrophoresis with SYBR-Green, **B)** 1% agarose gel electrophoresis with SYBR-Green, and **C)** TapeStation**.** TapeStation displays results as a gel image (top) and fragment-size distribution histograms (bottom).

4.2.1.2 ATAC with size selection and 50,000, 75,000, and 100,000 cells

The next ATAC optimisation experiment concerned increasing the number of cells used for transposition. As the previous experiment showed over-digestion with decreasing cell number, it was expected that increasing cell number would promote a partial digestion of chromatin. The previous experiment also showed many fragments below 50bp that may have resulted from over-digestion or were artefacts carried over from PCR amplification of libraries. Consequently, the following experiment included fragment size selection and purification using magnetic beads following library amplification (Figure 4.2). Therefore, the following experiment was carried out: ATAC using 50,000, 75,000 and 100,000 C-TERT cells and including library purification and size-selection using magnetic beads. The hypothesis was that a more partial digestion would be achieved with increasing cell numbers, and only fragments between 150-1500bp would be retained following PCR.

In order to evaluate the efficacy of size-selection using magnetic beads, a fraction of each library sample was kept aside following amplification by PCR (Figure 4.2A), and compared with purified library samples using magnetic beads (Figure 4.2B). The comparison showed effective size selection, most obvious by the absence of the intense band at ~30bp in all libraries following size-selection. Larger fragments were also removed, as is most apparent with 75,000 cells, where fragment sizes over 1500bp were removed following size selection (Figure 4.2B).

TapeStation results prior to size-selection and purification showed that chromatin digestion was too effective, and one may conclude that libraries were over-digested. However, size selection and the removal of smaller fragments revealed that a more partial digestion had occurred, especially with 100,000 cells which showed a faint periodic banding. This disparity is likely due to a technical error whereby the TapeStation over-exposes the smaller bands at the expense of masking the banding pattern displayed by the less ubiquitous larger fragments. The TapeStation histograms support this theory as the abundant peak at ~30bp prior to size selection dwarfs the larger fragments that become more apparent following size selection. When comparing cell number and chromatin digestion, the periodic banding pattern typical of partial digestion intensifies with increasing cell number (Figure 4.2B). However, there was still a considerable amount of mono-nucleosomal DNA present with 100,000 cells, indicating that more optimisation was necessary to attain a more partial digestion by Tn5 transposition.

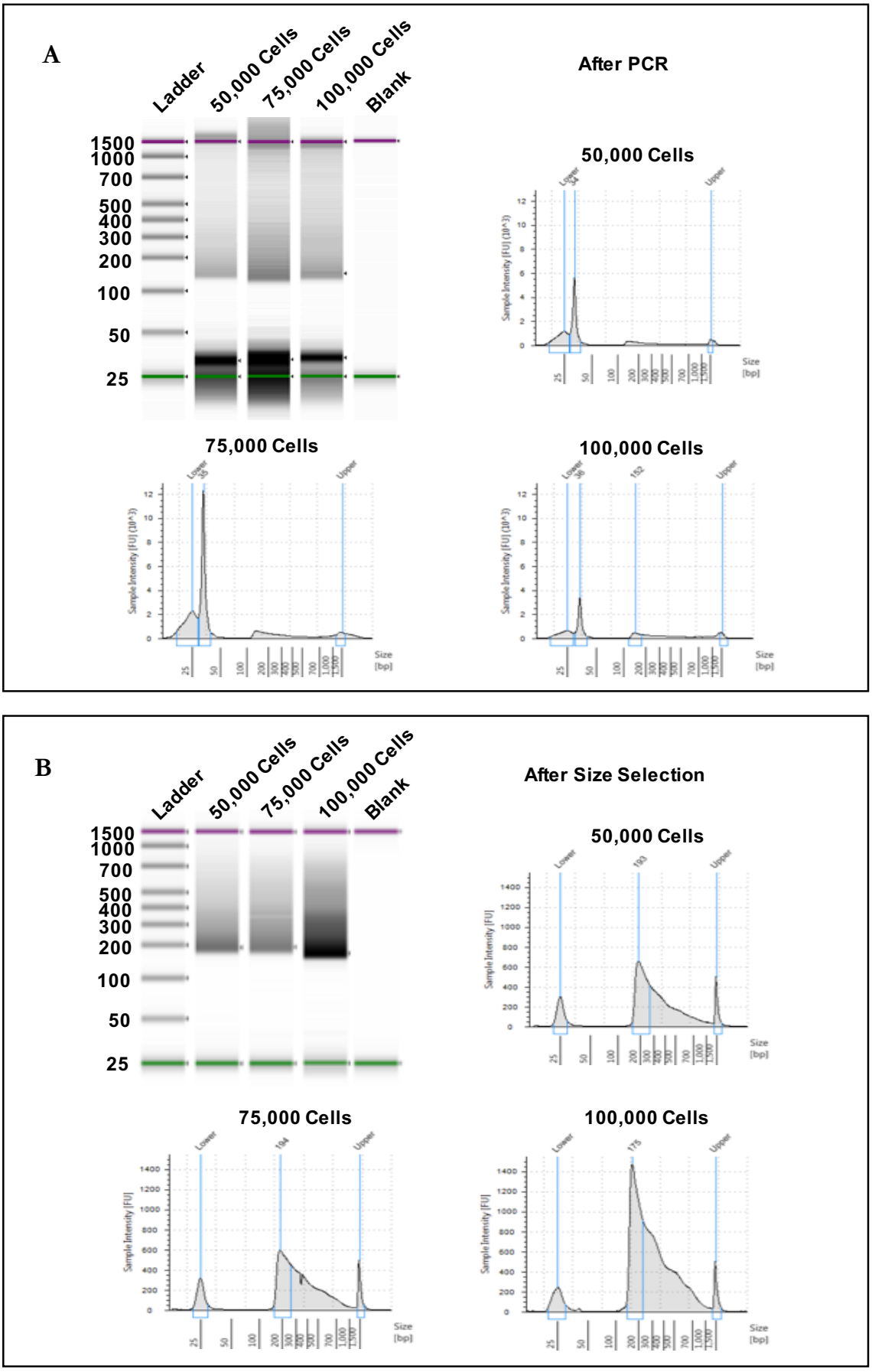


Figure 4.2 ATAC with 50,000, 75,000, and 100,000 C-TERT cells.

A) ATAC was carried out according to Buenrostro *et al* and fragment size distribution was assessed on 2µl of library sample following PCR amplification. **B)** Following PCR amplification, libraries were purified using magnetic beads to select fragment sizes between 100-1500bp, and visualised by TapeStation.

It should be noted that a direct comparison cannot be made between the pre-size-selected libraries in Figure 4.2A and those of Figure 4.1C, as a PCR clean-up kit was used for libraries in the previous experiment whereas the libraries in Figure 4.2A did not undergo purification. This may therefore explain the increased number of smaller fragments seen in this experiment, as the PCR-clean kit up used in the previous experiment removes a large portion of these fragments.

This experiment demonstrated that size selection of ATAC libraries using magnetic beads removes fragments below 100bp and above 1500bp, and assists in visualising the periodic banding pattern expected of partial digestion. Furthermore, increasing cell number improved partial digestion, but the upper limit of 100,000 cells used in this experiment did not show digestion satisfactory for sequencing. Future experiments therefore included size selection, and further optimisation aimed at improving digestion by increasing cell number.

4.2.1.3 ATAC with 100,000 and 200,000 cells

The previous experiment showed that increasing cell number resulted in a more partial digestion. This next experiment aimed to further improve the partial digestion attained using 100,000 cells by carrying out ATAC on 200,000 cells (Figure 4.3).

The use of 200,000 cells for ATAC resulted in a marked difference in digestion compared to 100,000 cells (Figure 4.3). Although the periodic banding pattern that indicated partial digestion was observed for both cell numbers, it was more prominent for 200,000. The TapeStation histograms also showed three distinct peaks that were more prominent in 200,000 cells compared to 100,000 cells, indicating the amplification of fragments encompassing 1-3 nucleosomes. As with Figure 4.2, size selection was carried out to remove large and small fragments. The pre-size-selected TapeStation results can be found in Appendix B-1 and confirm that size selection was necessary. The results from this experiment showed that using 200,000 cells was the optimal condition for partial digestion tested so far.

Throughout the previous set of experiments, as well as others not discussed, the lysis procedure was also optimised. This included standardising the number and precipitancy of pipette aspirations, as well as the centrifugation speed whilst extracting nuclei. This was judged by assessing the quality of nuclei prior to lysis and ensuring that single nucleus suspension was attained (data not shown). It was also noted that throughout the procedure, around half of all cells counted prior to lysis would be lost by the time of transposition ie. 100,000 cells would provide ~50,000 nuclei. Regardless, counting cell number at the start of procedure was maintained throughout all ATAC protocols.

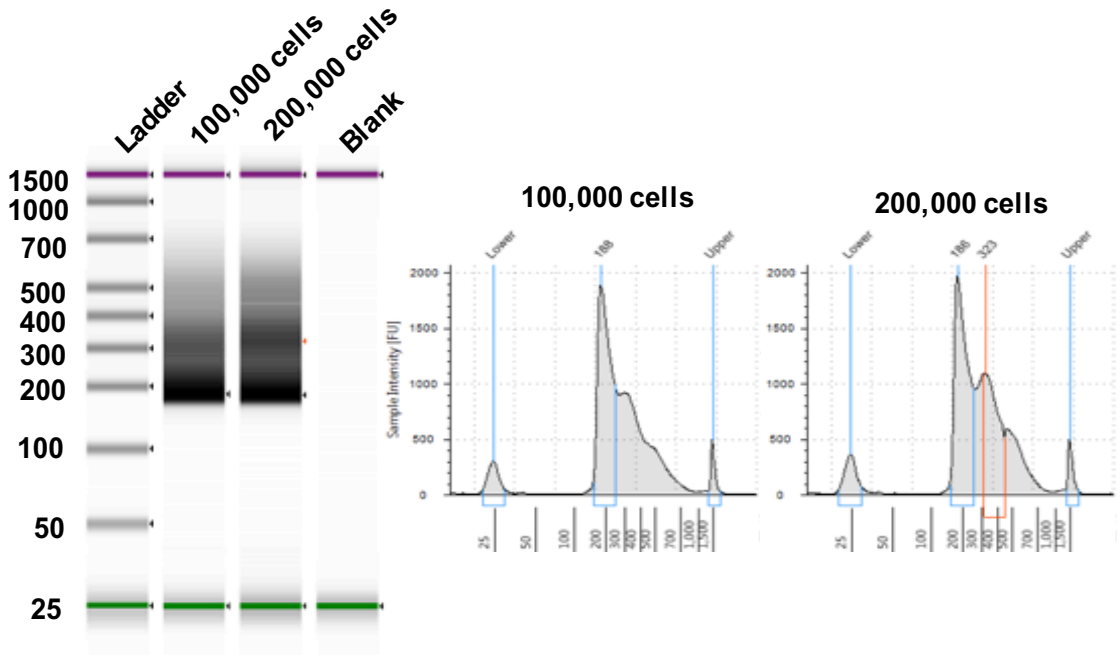


Figure 4.3 ATAC on 100,000 and 200,000 C-TERT cells.

100,00 and 200,000 cells were counted using a haemocytometer and used for ATAC. Results obtained using TapeStation.

4.2.1.4 Optimisation of incubation time

Although the chromatin digestion attained from 200,000 C-TERT cells showed the desired periodic banding pattern, only fragments encompassing three nucleosomes were retained and were not particularly abundant. Two more experiments were therefore carried out with the aim of enhancing the banding pattern for partial digestion. One included decreasing the incubation time with Tn5 to restrict digestion (Figure 4.4), and the other was altering the magnetic bead concentration so as to include larger fragments that may have been eliminated (Figure 4.5, see next section). Both experiments used 200,000 cells and included the standardised lysis procedure.

The time-course ATAC experiment included transposition by Tn5 for 0, 10, 20, and 30 minutes. It was expected that decreasing the incubation time would result in more limited chromatin digestion and larger fragments would be observed. Conversely, Figure 4.4 shows that reducing incubation time resulted in a trend towards over-digestion, with an increased abundance of DNA fragments between 150-350bp, whereas at 30 min incubation the periodic banding pattern emerges. This apparent over-digestion may again be a result of the technical issue previously seen in Figure 4.2, whereby over-exposure of the band between 150-350bp may be masking the banding pattern of that lane. As Tn5 digests large chromatin, the number of nucleosomes encompassed by each fragment moves from a large number towards a mono-nucleosome fragment of 147bp. Therefore, the last fragments to be digested are those between two nucleosomes to produce mono-nucleosomal fragments. This is seen in Figure 4.5, where 10-20 minutes of incubation is enough to convert the majority of the chromatin into mono/di-nucleosomal fragments resulting in the over-exposed band between 150-350bp. By 30 min these fragments are digested to reveal the expected periodic banding pattern.

The 0 min negative control included chromatin which was not subjected to transposition by Tn5, and so it may be expected that a large fragment of undigested genomic DNA would appear at the top of the gel. However, this was not observed. This may be due to these large fragments being removed during size selection, yet it is still not apparent on the gel prior to size selection (see Appendix B-2). An alternative explanation may be that as Tn5 incorporates sequencing adapters onto DNA during digestion, the undigested DNA cannot bind to the indexing primers which therefore prohibits PCR amplification. It should also be noted that following PCR, and before size-selection, there was an abundance of 30-50bp fragments at 0 min incubation that were also seen in the previous experiment. This suggests that these small fragments are possible artefacts from PCR and not over-digested fragments from transposition (see Appendix B-2).

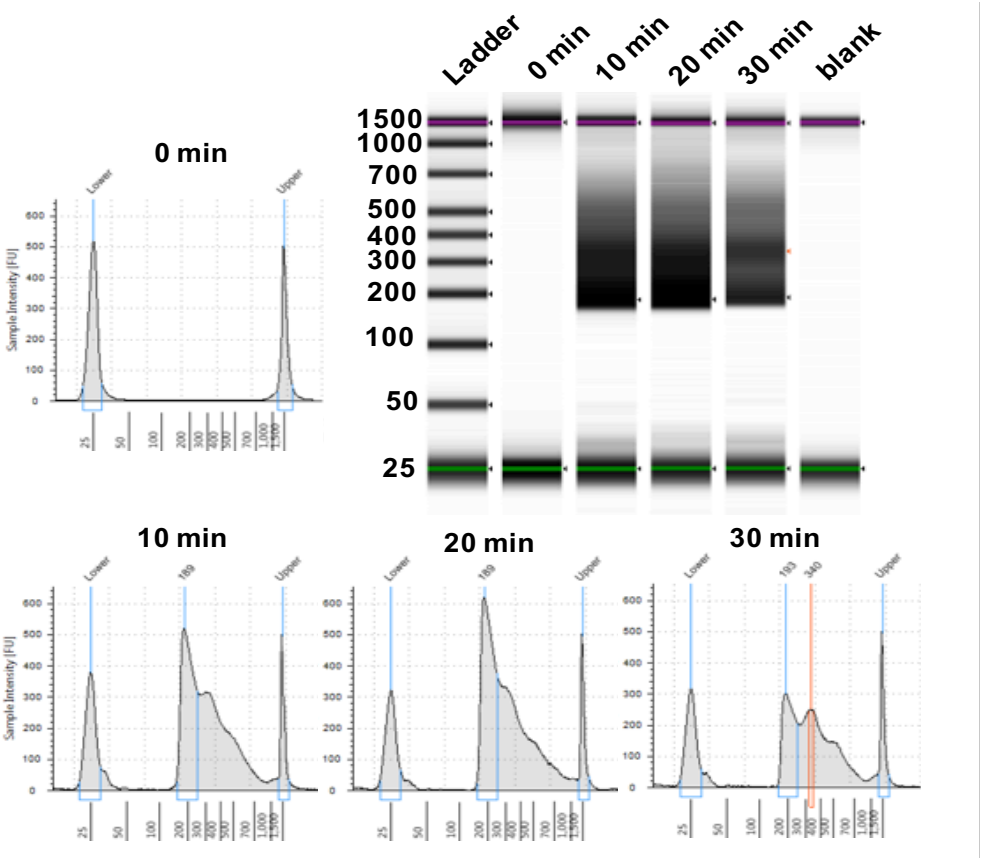


Figure 4.4 Time-course digestion ATAC on C-TERT cells.

200,000 C-TERT cells were lysed and subject to incubation with Tn5 for 0 min, 10 min, 20 min, and 30 min. Libraries were visualised by TapeStation analysis.

This experiment demonstrated that reducing incubation time of chromatin with Tn5 did not enhance partial digestion. Furthermore, the 30 min incubation replicated that of the previous experiment, thus demonstrating reproducibility of this protocol.

4.2.1.5 Optimisation of size selection

The final optimisation experiment concerned size selection. As previously shown in Figure 4.2, as well as throughout all optimisation experiments (see Appendix B), size selection removes smaller fragments and enhances the periodic banding typical of partial digestion. However, the ATAC protocol only showed three bands on the TapeStation which may be due to the size selection procedure removing fragments above 600bp. Following PCR, the size selection procedure had been carried out by first incubating each ATAC library with 0.5x the PCR volume of magnetic beads to remove smaller fragments, followed by a second incubation with 1.2x the PCR volume of magnetic beads to retain the larger fragments of interest and leave behind large undigested chromatin. This experiment therefore used 1.0x, 1.4x, 1.8x, and 2.2x the PCR volume of magnetic beads for the secondary incubation, with the hypothesis that larger fragments would be retained with increasing bead concentration, which would be visualised by an increased number of gel bands on TapeStation (Figure 4.5).

The results shown in Figure 4.5 largely agree with the stated hypothesis. As the secondary bead concentration increases, more bands resembling periodic nucleosomal occupation become apparent. At [1.0x] only two clear bands are seen, whereas at [2.2x] up to five bands resembling nucleosomal banding can be distinguished. However with [2.2x], the fifth band merges with larger fragments, reaching the limit detected by TapeStation and making it hard to distinguish between nucleosomal fragments and undigested genomic DNA. Although the number of bands observed with [1.8x] bead concentration is fewer than that observed with [2.2x] (four clear bands), the fragment-smear into the upper limit detected by TapeStation is much less. For this reason, [1.8x] bead concentration was considered the most appropriate for size selection of ATAC libraires.

4.2.1.6 Optimisation summary

Following an extensive optimisation investigation for a reliable ATAC procedure in C-TERT cells, a final protocol was established and can be read in full in the Methods chapter. Briefly, the protocol largely follows that of Buenrostro *et al* (Buenrostro *et al*., 2013), but with minor modifications including: the use of 200,000 cells rather than 50,000 cells; size selection and purification by magnetic beads rather than the use of a PCR-clean-up kit; and a more detailed lysis procedure concerning pipetting technique and centrifugation.

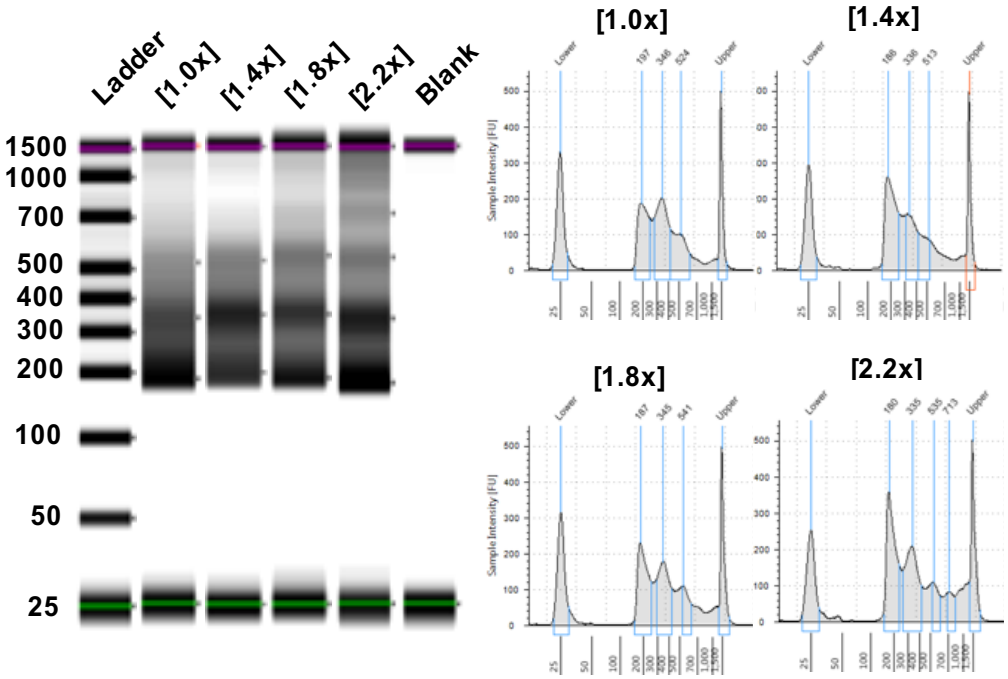


Figure 4.5 ATAC on C-TERT cells with increasing volume of magnetic beads for size selection.

200,000 C-TERT cells underwent ATAC. Following PCR-amplification, libraries were size selected by including increased amounts of magnetic beads for the second incubation (1.0x, 1.4x, 1.8x, and 2.2x post-PCR volume). Libraries were visualised using TapeStation.

4.2.2 Preparation and validation of ATAC-seq libraries from TERT-NHUC cells prior to sequencing

After an optimised ATAC protocol had been established in C-TERT cells, ATAC was carried out on the two male (B & C) and two female (H & K) TERT-NHUC lines in biological duplicate, with each replicate being assigned different indexing primers (see Methods). The following section outlines how libraries were prepared and the quality checks they were subjected to prior to sequencing. This included determining the number of PCR cycles necessary for each library during amplification, qPCR quantification of total DNA in libraries, quality check by TapeStation analysis, and an additional quality check of promoter vs exon enrichment of ATAC by qPCR.

Each library was prepared using 200,000 TERT-NHUC that were lysed and subjected to transposition by Tn5 for 30 min. Following transposition, samples were purified, then amplified and indexed by PCR. During the amplification stage, libraries were tagged with indexing primers (Ad2.1 – Ad2.8) to allow pooling of libraries for sequencing. Indexing primer sequences can be found in the Methods chapter. In order to avoid PCR biases such as reduced amplification of complex regions and over-amplification of GC-rich regions, the number of PCR cycles for each library was kept to a minimum (Aird et al., 2011). Therefore, the number of cycles for each library was determined by using a fraction of each library after 5 cycles for qPCR (Figure 4.6A), then taking the intercept at which one-third of the maximum fluorescence is attained for that library. For instance, the maximum fluorescence attained by the K-TERT Ad2.4 library was 2.5×10^6, thus 1/3 of this maximum fluorescence (8.5×10^5) requires 5 additional PCR cycles (Figure 4.6A). Each library therefore required its own additional number of PCR cycles which ranged from 4 cycles for B-TERT Ad2.5 and K-TERT Ad2.8, up to 9 cycles for C-TERT Ad2.6. The difference in the number of cycles may be due to differences in DNA concentrations preceding PCR, and is later corrected for by quantification of libraries. This qPCR step also provided the first QA test by plotting a melt curve (Figure 4.6B). Melt curves are commonly used in qPCR to show that assays have produced a single, specific product. Although many differing products are amplified in this PCR (due to the nature of ATAC), a single peak resembling a gradient of amplicon sizes should still be seen on a melt curve. Furthermore, as the fragment size distribution should be roughly the same between libraries, a single peak will occur at the same temperature. This can indeed be seen in Figure 4.6B where a single melt curve is observed at ~81°C. Furthermore, amplification was not observed for the no-template negative control for any of the indexing primers, indicating no primer dimerisation or contamination (data not shown).

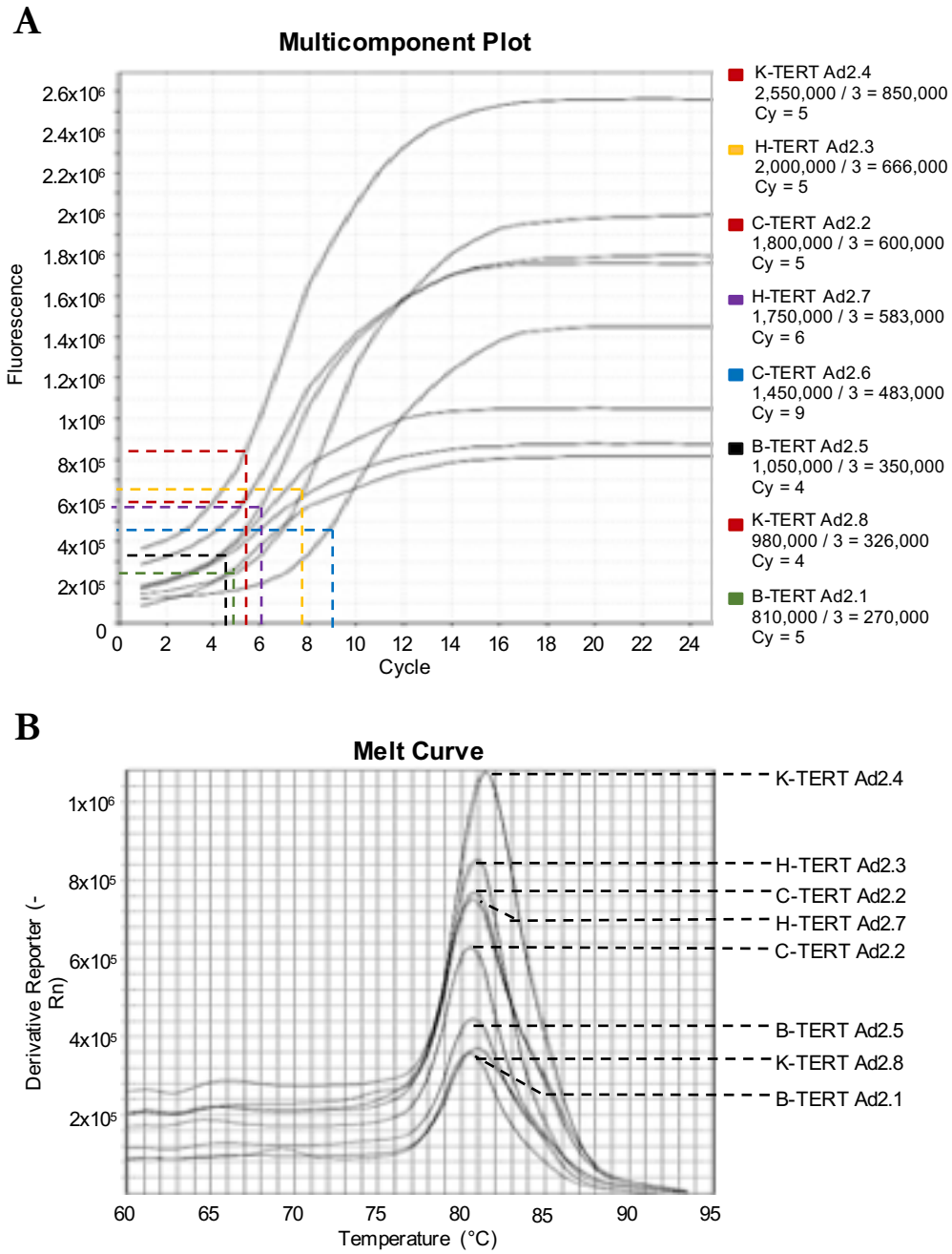


Figure 4.6 PCR amplification of ATAC libraries.

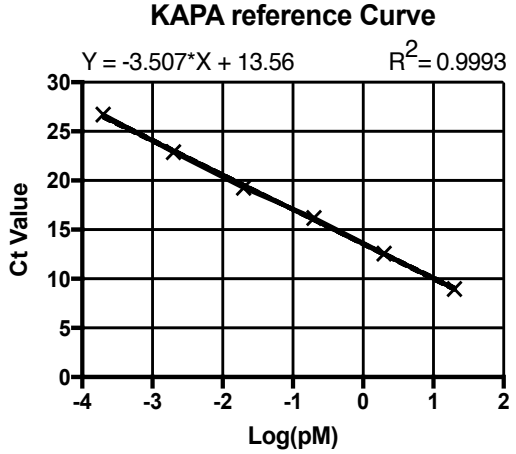
A) qPCR was used to determine the number of additional cycles required for library amplification in order to avoid PCR biases. The additional number of PCR cycles required is determined by taking the intercept at which one-third maximum fluorescence is attained for each library. Each library is colour-coordinated with the intercepts on the graph. **B)** A melt curve is created following qPCR amplification.

After amplification and indexing, libraries were purified and size-selected. In the previous optimisation experiments samples were then assessed by TapeStation but as these libraries were to be used for sequencing, quantification by qPCR was carried out using the KAPA Biosystems library quantification assay (Figure 4.7). This assay uses known concentrations of DNA standards to produce a standard curve from which unknown concentrations of DNA can be quantified. Quantification of ATAC libraries ranged from 2.06ng/µl in C-TERT Ad2.2 to 7.65ng/µl in H-TERT Ad2.7, with a mean concentration of 5.12ng/µl (median 5.09ng/µl) and a standard deviation of 1.84ng/µl. This quantification allowed for equal loading when pooling libraries together for sequencing, therefore decreasing the likelihood of any one library consuming the majority of sequencing reads.

Following quantification, libraries were diluted down to an equal concentration and pooled together for sequencing. QA was carried out on each library and pooled libraries by TapeStation and qPCR (Figure 4.8). TapeStation results show that each library showed a periodic banding pattern with at least 4 bands present for each library, and no fragments smaller than ~150bp (Figure 4.8A). Although libraries were supposedly diluted down to equal concentrations, differences in intensity of the bands on the gel image between libraries were still apparent, possibly arising from pipetting errors which are exacerbated by the sensitivity of TapeStation. Nevertheless, these differences were not large and the concentration values provided by the TapeStation also did not indicate major differences between libraries (data not shown).

A qPCR-based QA was also used in this study to validate ATAC libraries (Figure 4.8B). It is known that expressed genes show higher chromatin accessibility at their promoter regions relative to exonic regions (Thurman *et al.*, 2012). Using this, it is expected that ATAC libraries will show enrichment at promoter regions of active genes relative to their respective exonic regions. Therefore, qPCR targeting these regions for the housekeeping genes *SDHA* and *GAPDH* was used as a QA for the TERT-NHUC ATAC libraries (Figure 4.8B). The qPCR confirms the hypothesis and shows that, for all libraries and for both genes, exonic enrichment relative to promoters is significantly less. This additional QA confirmed chromatin accessibility enrichment at a localised level, and correlated with the TapeStation results to demonstrate that partial transposition by Tn5 was successful for all libraries. With the successful QA and pooling of libraries, sequencing was carried out on a NextSeq-500 with 75bp paired-end reads.

A



B

	Concentration (nM)	Concentration (ng/µl)
B-TERT Ad2.1	23.87	5.16
B-TERT Ad2.5	15.62	3.38
C-TERT Ad2.2	14.14	2.06
C-TERT Ad2.6	20.77	4.49
H-TERT Ad2.3	28.03	6.06
H-TERT Ad2.7	35.35	7.65
K-TERT Ad2.4	32.57	7.04
K-TERT Ad2.8	23.17	5.01

Mean = 5.12
Median = 5.09
SD = 1.84

Figure 4.7 Library quantification by qPCR.

A) A KAPA Biosystems quantification assay was used to produce a standard curve of known DNA concentrations. Standards were used in triplicate. **B)** The standard curve was used to quantify unknown library concentrations. Libraries were diluted 1:1,000, 1:10,000, and 1:100,000, and used for qPCR in triplicate. An average Ct value was taken across all dilutions for each library, and mapped onto the standard curve for quantification.

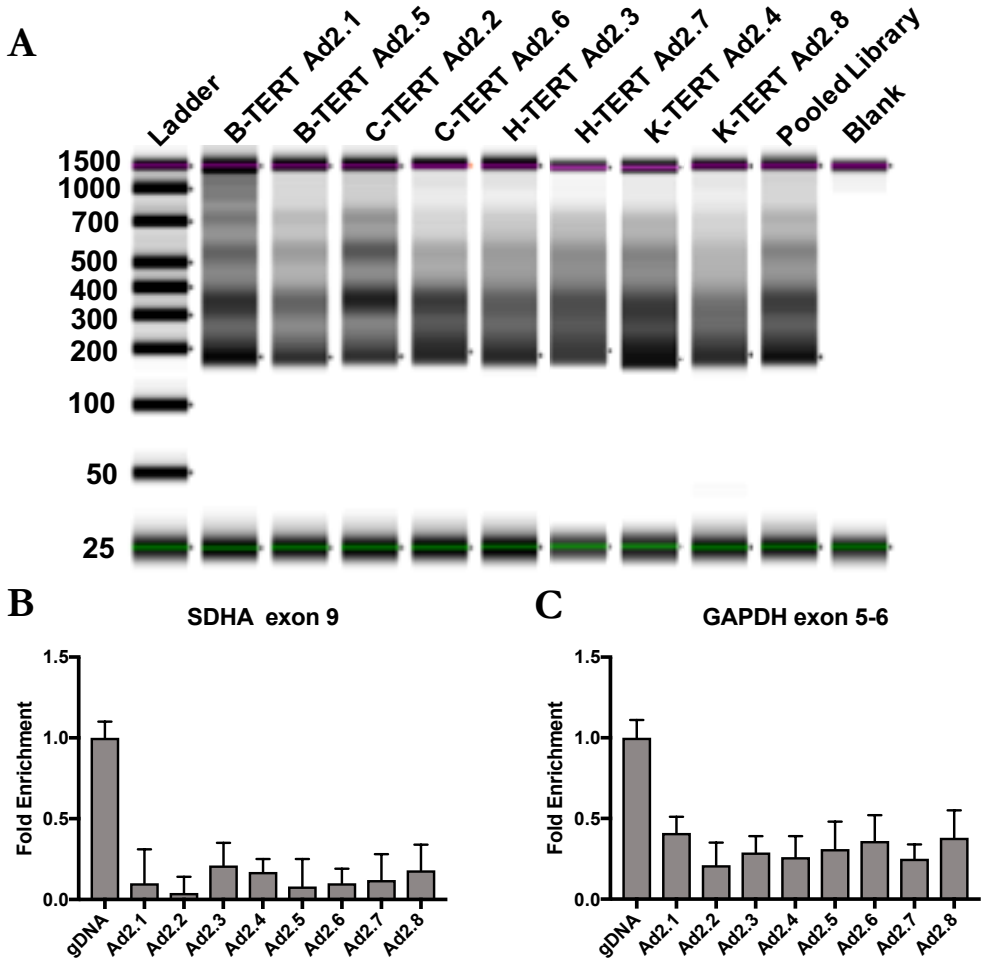


Figure 4.8 Quality assessment of libraries prior to sequencing.

A) Individual and pooled ATAC libraries were run on TapeStation. **B-C)** qPCR was carried out on ATAC libraries to quantify exons against their respective promoter regions for **B)** SDHA, and **C)** GAPDH.

4.2.3 Quality Assessment of ATAC-seq libraries following sequencing by FastQC

QA of ATAC-seq data is carried out using multiple tests and throughout the data analysis. QA can be both subjective and objective. However, when appropriate, objective QA can allow for efficient filtering and is incorporated into the analysis pipeline. The QA includes a series of tests such as FastQC, assessing mappability, plotting replicate correlation, and assessing fragment length distribution. As this chapter concerns the optimisation and initial preparation of ATAC-seq libraries, this section will only concern FastQC analysis to determine quality of sequencing and reads.

FastQC is a QA tool developed by the Babraham Institute that checks raw sequencing data for problems and biases that may have originated from the sequencing run or the starting library material (Andrews, 2010). FastQC was used on the raw fastq files and produced a report documenting how each library performed in a series of tests (Figure 4.9). Sequencing reads are then trimmed and filtered using Cutadapt (Martin, 2011) to remove adapters and low-quality reads, and FastQC is then used again. The output of the FastQC results for pre- and post-trimming of reads is summarised in Figure 4.9 and Table 4.1, and will be discussed below.

The first output from FastQC summarises basic statistics of the sequencing run (Table 4.1). This showed a total of 588,518,132 sequencing reads with a mean of 73,564,767 (median 72,115,277) reads per library, ranging from 60,579,256 reads in B-TERT Ad2.1 to 90,424,867 reads in K-TERT Ad2.4. All reads were 76bp in length and had a GC content of 45-46%. No reads were considered of poor quality, although an average of 61.15% of reads contained adapters. Following trimming and filtering by Cutadapt only 2.96% of reads were lost. This was a mean loss of 2,180,115 (median 2,203,365), ranging from 1,380,274 reads for C-TERT Ad2.6 to 3,226,056 reads for H-TERT Ad2.7, resulting in a total of 571,077,216 reads and an average of 71,384,652 (median 70,680,409) reads per library.

FastQC also generates a series of plots to assess different quality metrics, using a traffic-light system to indicate how each test performs (Figure 4.9, Appendix D). A plot assesses per base sequencing quality of all reads, as degradation of read quality tends to occur with longer sequencing runs due to chemistry of sequencing reactions. The per base sequencing quality plot here shows a high per base sequence quality throughout reads for all libraries (Andrews, 2010). Sequencing quality can also be hindered through technical problems with the run such as bubbles, smudges, and debris in the flow cell. This can be observed in the FastQC plot for per tile sequence quality, which analyses the quality score of

each tile across all bases to determine if loss of quality can be associated with a particular part of the flow cell (Andrews, 2010). For the TERT-NHUC libraries, per base sequencing quality was good for all libraries. Although a warning was given for the per tile sequencing quality for three libraries (B-TERT Ad2.1, C-TERT Ad2.6, and H-TERT Ad2.4), a closer inspection of these plots shows only minor aberrations in single bases at single tiles, and the anomaly is seen in only one of each pair of reads for these libraries. There is therefore little cause for concern regarding per tile sequencing quality. The per sequence quality score tests whether a subset of reads have a universally low-quality score and is usually derived from systematic errors in the sequencing process. All ATAC libraries here passed this test. The Per base N evaluates if the sequencer is unable to determine a particular base within reads for which it assigns an N. This usually indicates poor quality reads or biases within the library, and again all ATAC libraries passed this module. Sequence length distribution shows how varied reads in the library are, and may show over-digested libraries, dimerization and contamination. Following sequencing, all ATAC libraries had a read length of 76bp. However, after processing by Cutadapt, read length varied between 36-76bp (with the majority of reads still 76bp) and displayed an amber warning by FastQC. This is expected given that the reads had been trimmed (Figure 4.9 and Table 4.1). Sequence duplication can occur as a result of high coverage, enrichment bias, over-amplification by PCR, artefacts carried over from PCR, or low complexity in the library. Only H-TERT Ad2.7 produced a warning for this module, and this was corrected following processing by Cutadapt.

The final modules to be discussed are those which change most from pre-trimming to post-trimming and include per base sequence content, per sequence GC content, over-represented regions, and adapter content (Figure 4.9). Many of the ATAC libraries produced a warning for the over-represented regions module and all libraries failed the adapter content test. However following adapter trimming all libraries passed these modules, indicating that the issues pre-trimming are likely to be the result of high adapter content in the reads and therefore of little concern. However, from pre- to post- trimming, nearly all libraries gained a warning for GC content. This may be due to natural biases arising from the ATAC assay itself, as GC regions are known to be enriched at open chromatin (Thomson *et al.*, 2010; Blackledge *et al.*, 2010). Closer inspection showed that this bias was only very slight in all libraries (data not shown). The per base sequence content produced a warning for all libraries pre- and post- trimming. This often arises when there is a high adapter content (as shown by the adapter content module), through dimerization of adapters, and through biased composition of libraries such as GC content. The FastQC manual outlines how fragmentation of libraries using transposases produces inherent and intrinsic biases in the position at which the reads start, and insists that, although this issue cannot be resolved by

Table 4.1 Basic read statistics from FastQC pre- and post- adapter trimming.

	No. Reads	No. Poor Quality Reads	Read Length	GC (%)	% Reads with Adapters	No. Reads Post-Trimming	Read Length Post-Trimming	GC (%) Post-Trimming
B-TERT Ad2.1 R1	60,579,256	0	76	45	63.8	58,370,902	36-76	44
B-TERT Ad2.1 R2	60,579,256	0	76	45	63.8	58,370,902	36-76	44
B-TERT Ad2.5 R1	66,629,968	0	76	45	63.6	64,167,274	36-76	44
B-TERT Ad2.5 R2	66,629,968	0	76	45	63.5	64,167,274	36-76	44
C-TERT Ad2.2 R1	74,958,448	0	76	45	57.2	73,468,985	36-76	45
C-TERT Ad2.2 R2	74,958,448	0	76	45	57.2	73,468,985	36-76	45
C-TERT Ad2.6 R1	69,272,106	0	76	46	58.6	67,891,832	36-76	46
C-TERT Ad2.6 R2	69,272,106	0	76	46	58.9	67,891,832	36-76	46
H-TERT Ad2.3 R1	79,132,427	0	76	45	60.7	76,430,719	36-76	44
H-TERT Ad2.3 R2	79,132,427	0	76	45	60.7	76,430,719	36-76	44
H-TERT Ad2.7 R1	85,378,632	0	76	45	57.6	83,180,256	36-76	45
H-TERT Ad2.7 R2	85,378,632	0	76	45	57.6	83,180,256	36-76	45
K-TERT Ad2.4 R1	90,424,867	0	76	46	65.5	87,198,811	36-76	45
K-TERT Ad2.4 R2	90,424,867	0	76	46	65.5	87,198,811	36-76	45
K-TERT Ad2.8 R1	62,142,428	0	76	45	62.1	60,368,437	36-76	44
K-TERT Ad2.8 R2	62,142,428	0	76	45	62.1	60,368,437	36-76	44

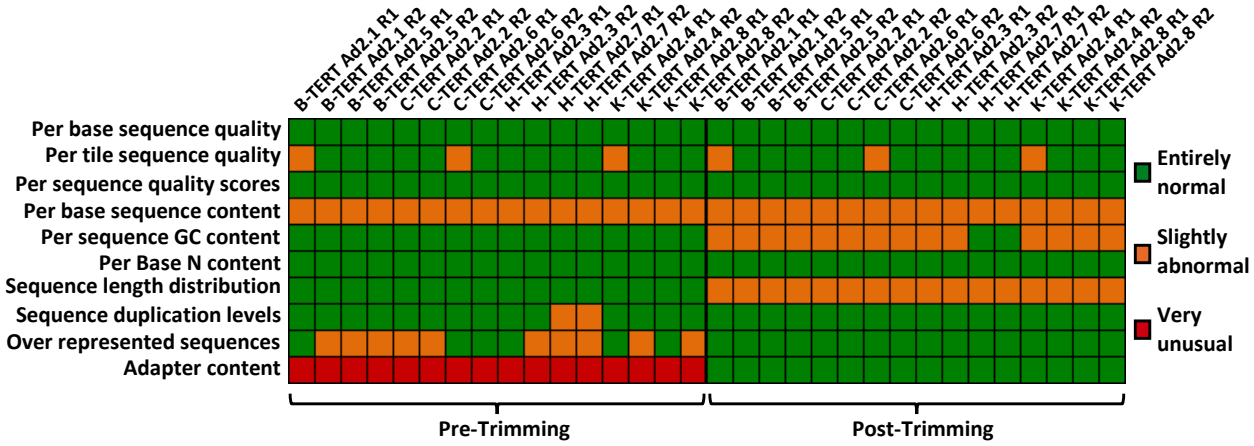


Figure 4.9 FastQC output pre- and post- adapter trimming.

Immediately following sequencing, fastq files are processed (Pre-Trimming) using the FastQC tool which carries out quality assessment on reads. FastQC is used for each pair of library reads and results presented as green, amber, and red score to signal pass, pass with warning, and failed tests respectively. Following the initial QC, reads are processed using the Cutadapt tool to remove adapters and poor-quality reads, and again assessed using FastQC (Post-Trimming).

trimming, there are no adverse effects in downstream analysis (Andrews, 2010). A closer inspection of the module in each library post-trimming showed that this was the case, with the first 12bp of reads displaying the unusual per base sequence content.

4.3 Discussion

This project aimed to characterise the epigenetic landscape of TERT-NHUC, and included mapping the chromatin accessibility landscape of these cells using ATAC-seq. As ATAC-seq had not previously been carried out in normal human urothelial cells, a protocol needed to be optimised for TERT-NHUC. A recent study, which used ATAC-seq to characterise the chromatin accessibility landscape of 20 healthy tissues in mouse, inadvertently showed the necessity of optimising the protocol when their post-sequencing QA produced varied results when using the same protocol for each tissue type (Liu *et al.*, 2019). This chapter has shown that the protocol for ATAC-seq preparation was optimised in NHUC-TERT, largely following the protocol established by Buenrostro *et al* (Buenrostro *et al.*, 2013), but using a greater number of cells and incorporating the size-selection of DNA fragments using magnetic beads. ATAC was then carried out on the two male and two female TERT-NHUC lines in biological duplicate. QA using TapeStation and qPCR respectively showed a periodic banding pattern of DNA fragments and increased chromatin-accessibility at the promoters of two housekeeping genes relative to their respective exonic regions. This therefore provided confidence to carry out next-generation sequencing on the TERT-NHUC ATAC libraries. QA using on the ATAC-seq data showed a good quality of sequencing, although adapter trimming was necessary which resulted in a slight GC bias in the reads. Nevertheless, this is typical of ATAC-seq and therefore further analysis was carried out on the TERT-NHUC ATAC-seq data, and will be discussed in the following chapter.

Chapter 5

Analysis of ATAC-seq data in male and female TERT-NHUC

5.1 Introduction

The previous chapter described the optimisation and preparation of an ATAC-seq protocol for TERT-NHUC. QA of ATAC-libraries from two male (B & C) and two female (H & K) TERT-NHUC lines showed that a sufficient quality standard in preparation and sequencing had been attained. This chapter describes the data analysis carried out on the ATAC-seq in TERT-NHUC, and compares the chromatin accessibility landscape between male and female cells.

Previous studies have shown that there are minimal differences in genome-wide chromatin state and accessibility between genders, and that the majority of differences pertain to the sex chromosomes (Qu *et al.*, 2015; Yen and Kellis, 2015). Therefore, it was hypothesised that only subtle differences in autosomal chromatin accessibility would be seen between genders in TERT-NHUC. Previous microarray analysis (Chapter 3) identified 441 differentially expressed genes between male and female TERT-NHUC. Therefore, it was also hypothesised that differential enrichment of chromatin-accessible peaks would be identified between genders around these gene loci.

Bladder cancer has the highest rate of mutations in chromatin modifying genes compared to any other cancer type (Gui *et al.*, 2011b; Weinstein *et al.*, 2014), and this is likely to result in changes to the chromatin landscape which promote tumorigenesis. The ATAC-seq data acquired from TERT-NHUC will provide a resource to which future studies regarding chromatin accessibility in bladder cancer cell lines can be compared.

5.2 Results

Analysis of ATAC-seq data was carried out in-house and utilised common bioinformatic tools in python and R. This included Bowtie2 for the alignment of reads, MACS for peak calling, and others such as Deeptools, Samtools, and Bedtools for general manipulation and visualisation of deep sequencing data. Parameters for analysis were determined by taking the consensus of multiple publications (Buenrostro *et al.*, 2013; Schep *et al.*, 2015; Ackermann *et al.*, 2016; Wang *et al.*, 2018; Corces *et al.*, 2018) as well as using guidelines set by the Cancer Research UK Cambridge Institute (Dunning *et al.*, 2019) (see Methods and Appendix-C).

5.2.1 Further QA and Basic Statistics on ATAC-analysis

As mentioned in the previous chapter, NGS QA includes a series of tests such as FastQC, assessing mappability, plotting replicate correlation, and assessing fragment length distribution. QA can also be incorporated into the analysis to allow filtering out of poor quality reads and undesired mapping to the genome.

5.2.1.1 Basic Statistics

Following FastQC and filtering of poor quality reads, reads were aligned to the genome and peaks for accessible chromatin regions were called. Analysis of the ATAC-seq data was carried out during and following this process, and will be discussed throughout the chapter. However, basic statistics that were generated throughout the alignment and peak calling process will be discussed in this section and can be seen in Table 5.1.

Raw reads were aligned to the reference genome (hg19) using Bowtie2. Hg19 was chosen as the reference genome for this study as it was more highly annotated than the more recently released hg38 reference, and the majority of publications pertained to hg19 at the time of analysis. During the alignment process, reads were mapped to the genome and then filtered for fragments that map to mitochondrial DNA (chrM), duplicated fragments, and fragments that are mapped to blacklisted regions of the genome (ENCODE Consortium, 2012). Blacklisted regions were those defined by the ENCODE consortium and include anomalous, unstructured or high signal regions seen in NGS experiments independent of cell line of experimental conditions. Table 5.1 shows that a high overall alignment rate of reads was attained for all samples, ranging from 97.2% in C-TERT Ad2.2 to 98.9% in H-TERT Ad2.7. Reads that mapped to chrM range from 1.76% in C-TERT Ad2.6 to 2.93% in H-TERT Ad2.7, and reads that were in duplicated fragments or aligned to blacklisted regions ranged from 17.65% in K-TERT Ad2.4 to 30.45% in H-TERT Ad2.4. A mean of 76.36% (median 77.60%) mapped reads were retained after filtering. This resulted in a total of 435,087,182 filtered mapped reads with a mean of 54,385,898 (median 54,403,795) reads per sample ranging from 45,639,865 mapped reads in B-TERT Ad2.1 to 68,999,586 mapped reads in K-TERT Ad2.4. The overall alignment quality was therefore considered acceptable for all samples.

A total of 390,488 peaks were called with a mean of 48,811 (median 51,172) peaks per sample ranging from 12,670 peaks in K-TERT Ad2.8 to 78,979 peaks in C-TERT Ad2.3. This is an unusually large range in the number of peaks as it is expected that a similar number should be observed between samples. An exceptionally low number of peaks was seen in K-TERT Ad2.8 and is nearly 3.5 times less than the number of peaks observed in its biological

replicate (K-TERT Ad2.4). Within gender groups, males had a mean of 66,098 peaks per sample, whereas females had a mean of 31,524 (37,808 excluding K-TERT Ad2.8) therefore showing that males had >2-fold more peaks per sample compared to females.

Table 5.1 Basic alignment and peak calling statistics.

	No. Total Raw Reads	Overall Alignment Rate (%)	chrM (%)	Duplicate & Blacklisted (%)	No. Final Mapped Reads	Final Mapped Reads (%)	FRiP	No. of Peaks
B-TERT Ad2.1	58,370,902	97.85	2.23	18.27	45,639,865	78.19	13	63843
B-TERT Ad2.5	64,167,274	98.14	2.19	18.83	49,996,367	77.92	13	63835
C-TERT Ad2.2	73,468,985	97.2	1.78	19.04	56,785,929	77.29	19	78979
C-TERT Ad2.6	67,891,832	98.05	1.73	18.57	53,268,506	78.46	11	57736
H-TERT Ad2.3	76,430,719	98.12	1.99	19.73	58,999,609	77.19	6.3	26763
H-TERT Ad2.7	83,180,256	98.9	2.93	30.45	55,539,083	66.77	7.5	44608
K-TERT Ad2.4	87,198,811	97.85	1.8	17.65	68,999,586	79.13	7.2	42054
K-TERT Ad2.8	60,368,437	98.04	2.77	20.31	45,858,236	75.96	5.1	12670

The decreased number of peaks in females compared to males was also represented in the fraction of reads in peaks (FRiP) scores which were 6.5% and 14% in females and males respectively. Given that the number of reads between male and female samples was similar (57,349,129 reads per male sample; 51,422,667 reads per female sample), the low FRiP scores and number of peaks in females indicate that the distribution of peaks is more likely to be spread throughout the genome rather than pertain to concise regions of accessibility. It was unknown whether the low number of peaks observed in females is an indicator of a poor-quality assay or an interesting finding related to gender differences. However, it was noted that K-TERT Ad2.8 showed an unusually low number of peaks even compared to its biological replicate, and despite a good alignment quality that was comparable to all other samples.

5.2.1.2 Fragment-size density plot of ATAC-seq libraries

The study took advantage of paired-end sequencing, which allows for a more accurate read alignment and also provides information on the length of each DNA fragment sequenced (insert-size) (Turner, 2014). When represented as an insert-size density plot, ATAC-seq libraries are expected to show a peak representing open chromatin at <100bp, followed by a peak at 200bp for mono-nucleosome fragments and sequential peaks separated by 200bp for sequential nucleosome fragments (Buenrostro *et al.*, 2013; Schep *et al.*, 2015;

Ackermann *et al.*, 2016). Insert-size density plots were therefore generated on both linear and logarithmic scales for each TERT-NHUC ATAC-seq library and can be seen in Figure 5.1.

For all samples, the fragment density plots (Figure 5.1) showed a high number of fragments <100bp followed by a peak at ~200bp representing open chromatin and mono-nucleosomal fragments respectively. However, the density of mono-nucleosome fragments was not consistent between samples. Male TERT-NHUC showed a distinct peak at 200bp, with the clearest observed in C-TERT Ad2.4, whereas female samples showed a peak at 200bp that merges with the accessible chromatin fragments at <100bp as is particularly apparent with K-TERT Ad2.8. When viewing the density plots on a logarithmic scale the periodic nucleosomal pattern can be observed, with three clear peaks shown for all male samples, and 2-3 peaks seen for the female samples. Again C-TERT Ad2.2 displays the clearest banding pattern with three distinct peaks, whereas K-TERT Ad2.8 showed a gradual decline in fragment size with only two mild peaks observed.

In general, the fragment density plots showed a distribution expected from the partial digestion of genomic DNA by Tn5. The results coincide with the TapeStation results of Figure 4.8 where the open-chromatin peaks at 100bp followed by the 2-3 nucleosomal peaks match the 4 bands observed on the gel. It is noted that nucleosomal fragments were not as distinguishable in female samples compared to the males, a pattern not seen on the TapeStation.

5.2.1.3 Genome-wide signal correlation between replicates

An initial assessment of reproducibility between replicates was carried out by correlating genome-wide ATAC-seq signal following alignment (Figure 5.2). The genome was divided into bins of 1000 bp and an alignment signal (number of aligned reads) was obtained for each bin for each library. Scatter plots correlating the number of reads in bins between biological replicates were then produced (Figure 5.2A). For all samples, a high degree of correlation for genome-wide alignment signal was observed between replicates, with Pearson correlation coefficients ranging from 0.89 for K-TERT to 0.99 for B-TERT Figure 5.2A). This therefore indicated good reproducibility of genome-wide ATAC-seq signal between biological replicates of the ATAC assay.

A heatmap with hierarchical clustering was also generated to show genome-wide signal correlation between all samples (Figure 5.2B). The clustering showed that male and female samples fall into two distinct groups. Biological replicates then clustered together showing that the strongest correlation was between biological replicates of the same cell line. An exception to this was K-TERT, where K-TERT Ad2.8 falls into its own separate cluster

within the female group. Therefore, with regard to genome-wide signal, male and female TERT-NHUC separated into distinct groups. However, within the female group, K-TERT Ad2.8 did not correlate well with the other female samples, including its biological replicate.

The QA post-sequencing demonstrated that the reads generated by the sequencing itself were of high quality, with biases typical of the assay. The mapping of reads to the genome was also of high quality with the majority of reads in all samples uniquely mapping to the reference genome, and demonstrated fragment-size distribution typical of ATAC. Cross-correlation between samples of ATAC signal throughout the genome demonstrated high reproducibility and produced distinct male and female groups by hierarchical clustering. Female samples showed a lower number of peaks for chromatin accessibility as well as decreased clarity of nucleosome periodicity for fragment-size distribution. This may be attributed to a poor-quality assay or the result of biological gender differences. Finally, K-TERT Ad2.8 appeared as an anomaly throughout the QA with a much lower number of accessible peaks, a less pronounced fragment-size curve, and a lower correlation with its biological replicate compared to other samples. Despite this, K-TERT Ad2.8 was still included in further analysis due to the limited number of samples used in this study.

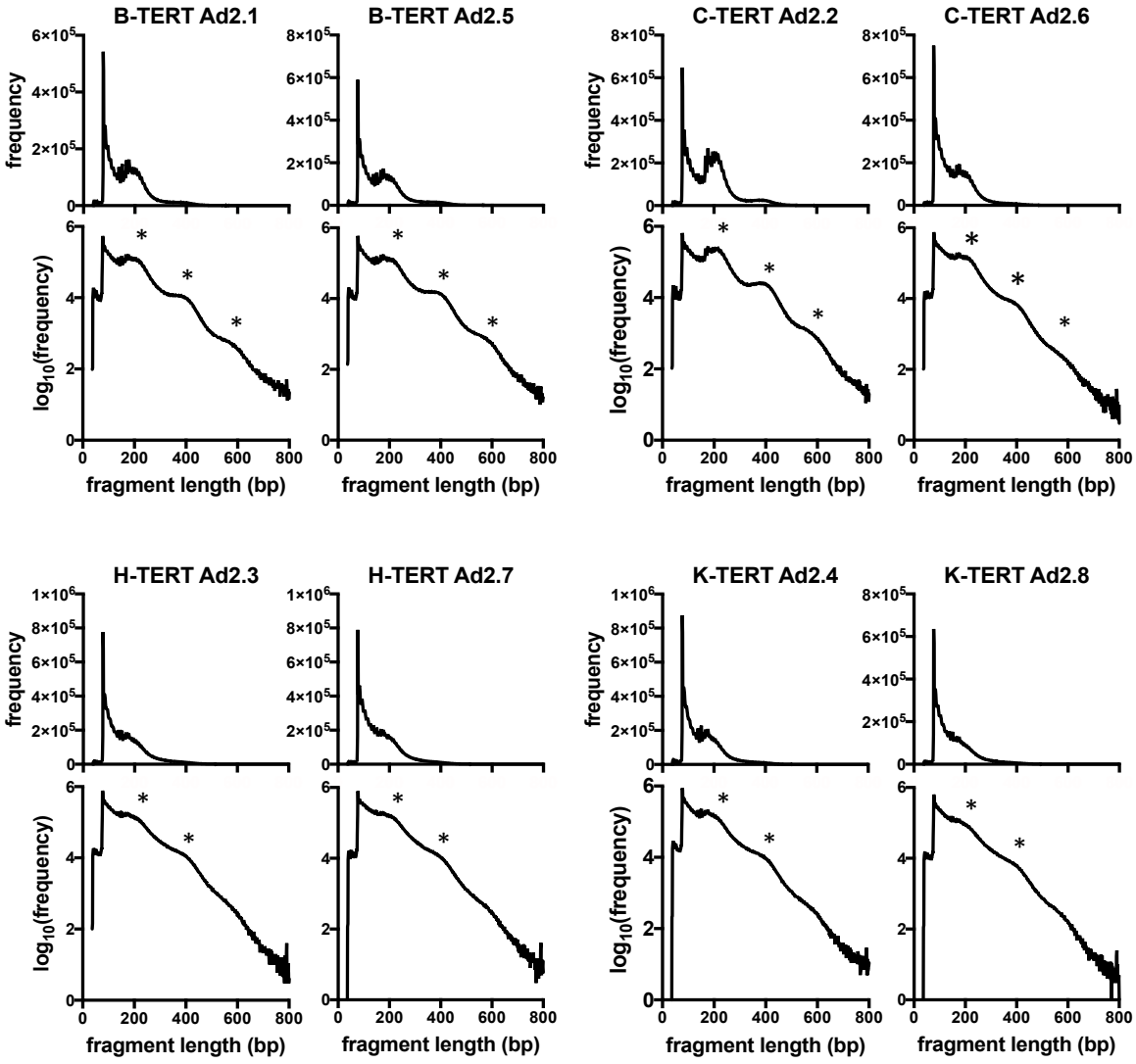


Figure 5.1 ATAC-seq insert size density plots.

Paired-end sequencing allows for sequenced fragment sizes to be computed. The frequency of fragment lengths in each sample was plotted for each sample on both a linear (top) and logarithmic (bottom) scale for fragments ranging from 0bp to 800bp. Asterisk (*) symbols on the logarithmic plots are placed above peaks representing nucleosomal fragments.

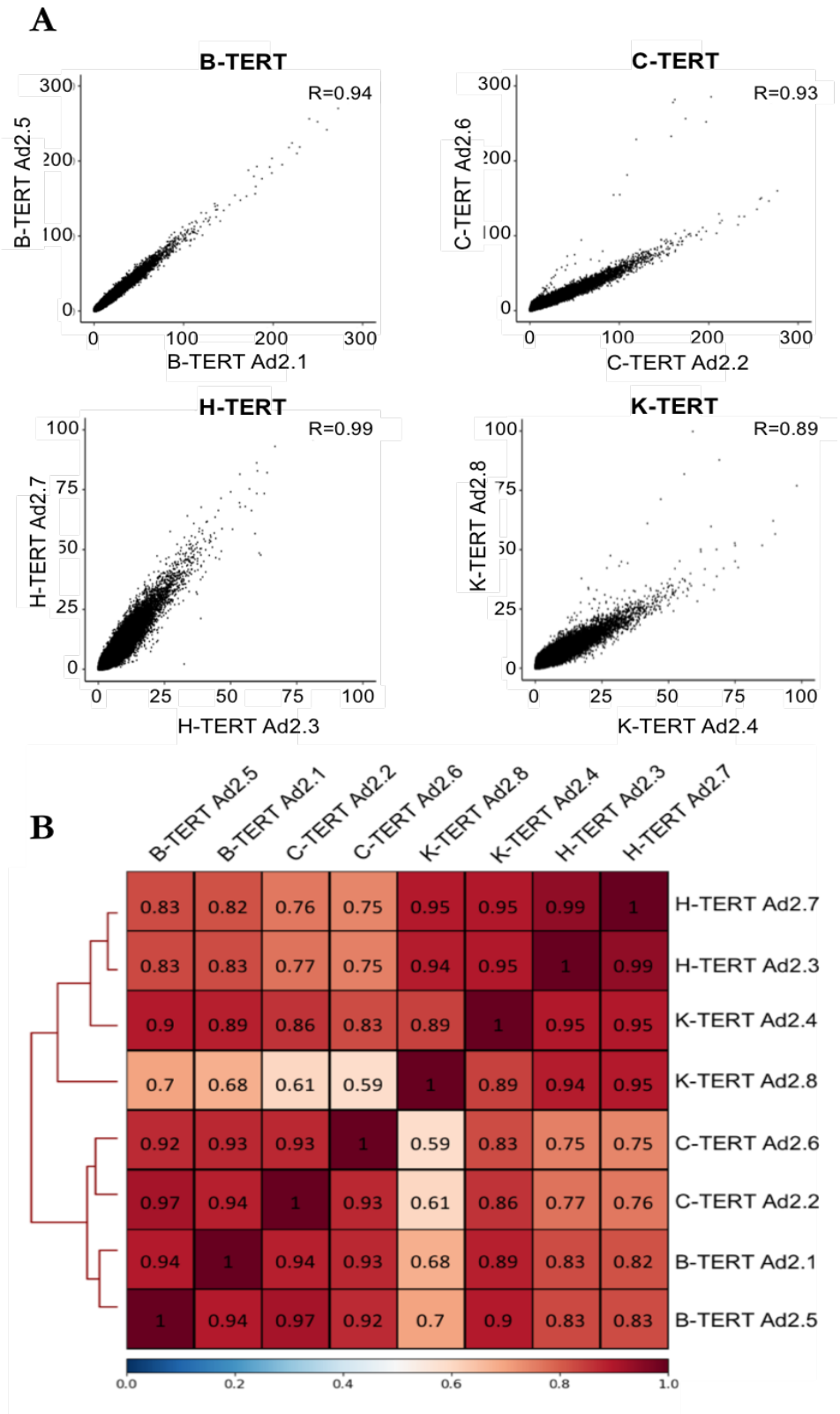


Figure 5.2 Correlation of ATAC-seq signal between samples.

The genome was split into bins of 1000bp and an alignment signal for each bin was assigned for each sample. **A)** Scatter plots of alignment signal within each bin was plotted for each biological replicate. R values indicate Pearson correlation coefficient. **B)** A heatmap of signal was generated to correlate genome-wide signal between all samples. Values indicate the Pearson correlation coefficients.

5.2.2 ATAC-seq shows decreased signal in female TERT-NHUC compared to male TERT-NHUC

Following alignment of reads to the genome, samples can be visualised using tools such as the UCSC Genome Browser and the Integrative Genomics Viewer (IGV; provided by the Broad Institute). This section aims to compare ATAC-seq signal between genders at individual loci of interest and at the genome-wide level. The hypothesis was that ATAC-seq signal throughout the genome would be comparable between genders, as changes in chromatin accessibility were expected to be subtle and at the level of individual peaks at specific loci. For each cell line, mapped reads for each biological duplicates were combined to show tracks.

5.2.2.1 Visualising ATAC-seq tracks using IGV

The integrated genomics viewer (IGV) is a high-performance visualisation tool for interactive exploration of large genomic data sets (Robinson, 2012). By loading alignment files into IGV, ATAC-seq tracks can be visualised at any region of interest, thereby enabling a visual comparison of chromatin accessibility between samples. Alignment files were generated using deepTools (Ramírez *et al.*, 2014) to obtain a signal in 10bp bins across the genome, and normalised using the reads per kilobase of transcript per million mapped reads method (RPKM). IGV was used to view the ATAC-seq signal for each TERT-NHUC by combing tracks of biological duplicated at large genomic loci surrounding *SDHA*, *GAPDH*, *UCA1*, and *MMP1* (Figure 5.3). *SDHA* and *GAPDH* were previously used to validate the ATAC assay prior to sequencing by qPCR (Figure 4.7) which demonstrated chromatin accessibility at the TSS of these genes followed by less accessible chromatin within exons. *UCA1* and *MMP1* have also been displayed here as they represent differentially expressed genes in female and male TERT-NHUC respectively (shown in previous microarray chapter; section 3.1.2.3), and therefore would likely exhibit differential peaks of chromatin accessibility.

The chromatin accessibility landscape surrounding active genes typically consists of accessible regions at the TSS and at *cis*-regulatory sites within introns and intergenic regions, with exons generally exhibiting less chromatin accessibility (Thurman *et al.*, 2012). The IGV tracks of Figure 5.3 conformed to this expected pattern of chromatin accessibility, where spikes in the tracks could be seen at TSSs, introns and intergenic regions of the presented genes.

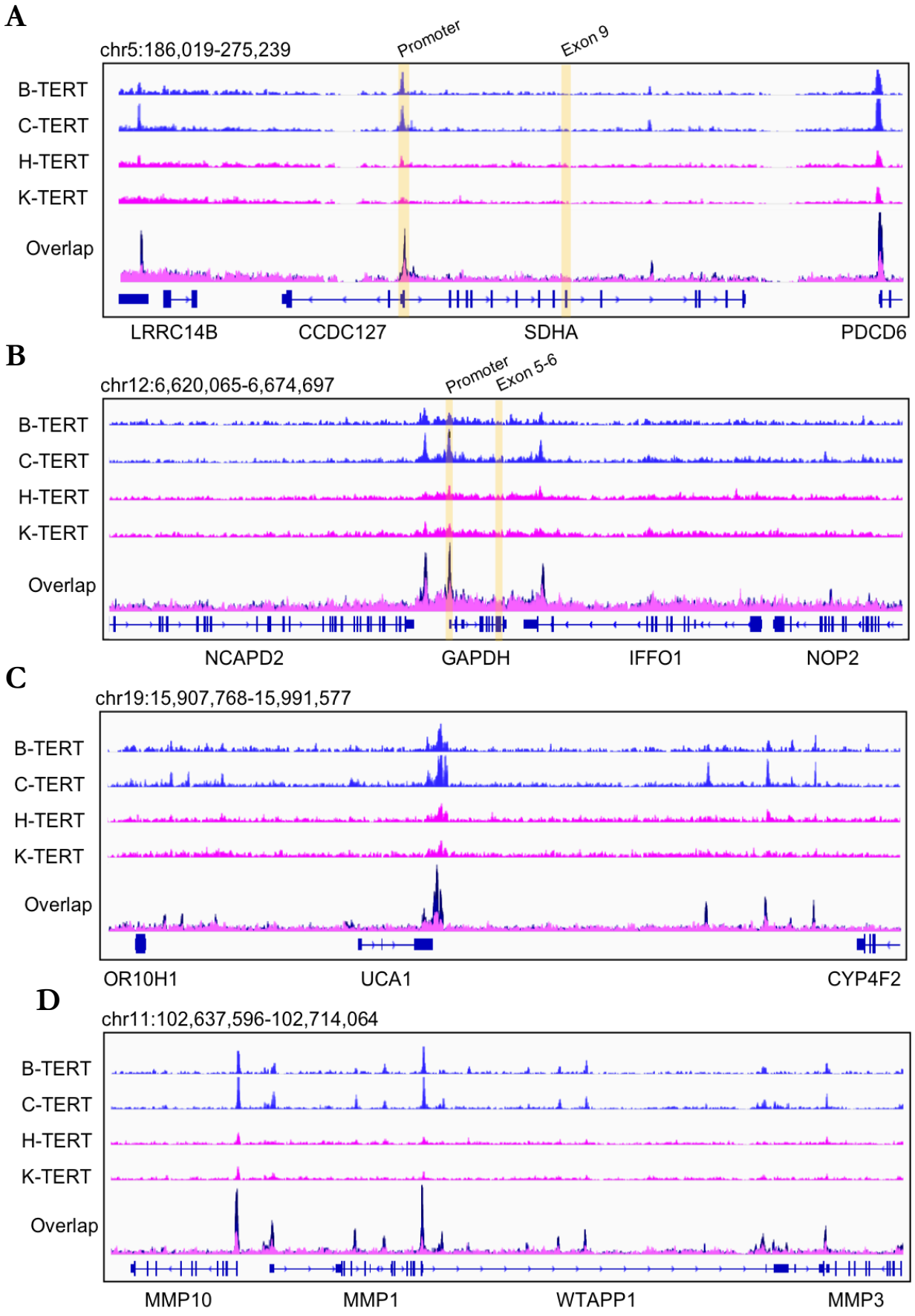


Figure 5.3 IGV tracks of ATAC-seq signal.

ATAC-seq tracks were visualised using IGV at loci surrounding **A)** *SDHA*, **B)** *GAPDH*, **C)** *UCA1* and **D)** *MMP1*. Cell lines are represented by overlapping tracks of biological replicates, and the Overlap track represents the overlay of all samples. Blue tracks represent male TERT-NHUC and pink tracks represent female TERT-NHUC. **A-B)** Regions highlighted in yellow indicate qPCR products in Figure 4.7. All tracks are visualised on the same scale.

Figure 5.3A & B show that the constitutively expressed housekeeping genes SDHA and GAPDH both displayed peaks at their TSS which were followed by less accessible exonic regions devoid of peaks. This can clearly be seen at the highlighted regions indicating the qPCR target loci used to validate the ATAC-assay prior to sequencing (Figure 4.8). The ATAC-seq tracks around SDHA and GAPDH therefore conform to the typical pattern of chromatin accessibility expected of active genes and are analogous to the results of Figure 4.8. Further peaks could also be seen at these loci within TSSs of LRRC14B and PDCD6 in Figure 5.3A and IFFO1 in Figure 5.3B, and were present in both the male and female tracks. However, for all the aforementioned peaks the signal was stronger in male TERT-NHUC compared to the female TERT-NHUC, as is particularly apparent in the overlay tracks of Figure 5.3A & B.

Previous microarray results showed that UCA1 and MMP1 represent the most differentially expressed autosomal genes in female and male TERT-NHUC respectively. It was therefore expected that large peaks of chromatin accessibility would be apparent at the TSS of UCA1 in female TERT-NHUC and at the TSS of MMP1 in the male TERT-NHUC. However, this was not the case, as spikes of chromatin accessibility were seen at the TSSs of these genes in both genders (Figure 5.3A & B). As with the housekeeping genes, the peaks represented across these loci had increased signal in male TERT-NHUC compared to females, including at the TSS of UCA1.

This pattern of IGV tracks observed in Figure 5.3 was also seen across nearly all regions of the genome inspected, with peaks that were present in both male and female TERT-NHUC displaying greater signal in males. Furthermore, it was noted that the background signal (the signal between peaks) was generally greater in the female TERT-NHUC, although only marginally. Given the apparent disparity of chromatin accessibility tracks between the genders, it is important to note that the profiles within each gender were comparable, i.e. B-TERT and C-TERT tracks were alike, and H-TERT and K-TERT tracks were alike.

The visualisation approach thus far only considered specific loci. Therefore, a Circos plot encompassing the entire genome was generated to display ATAC-seq signal for each cell line at a global level (Figure 5.4). These results coincided with what was seen at individual loci, whereby enrichment of ATAC-seq signal was greater at a genome-wide level in male TERT-NHUC compared to female cells. The Circos plot also showed broadly similar profiles of chromatin accessibility between B-TERT and C-TERT, and between H-TERT and K-TERT cells.

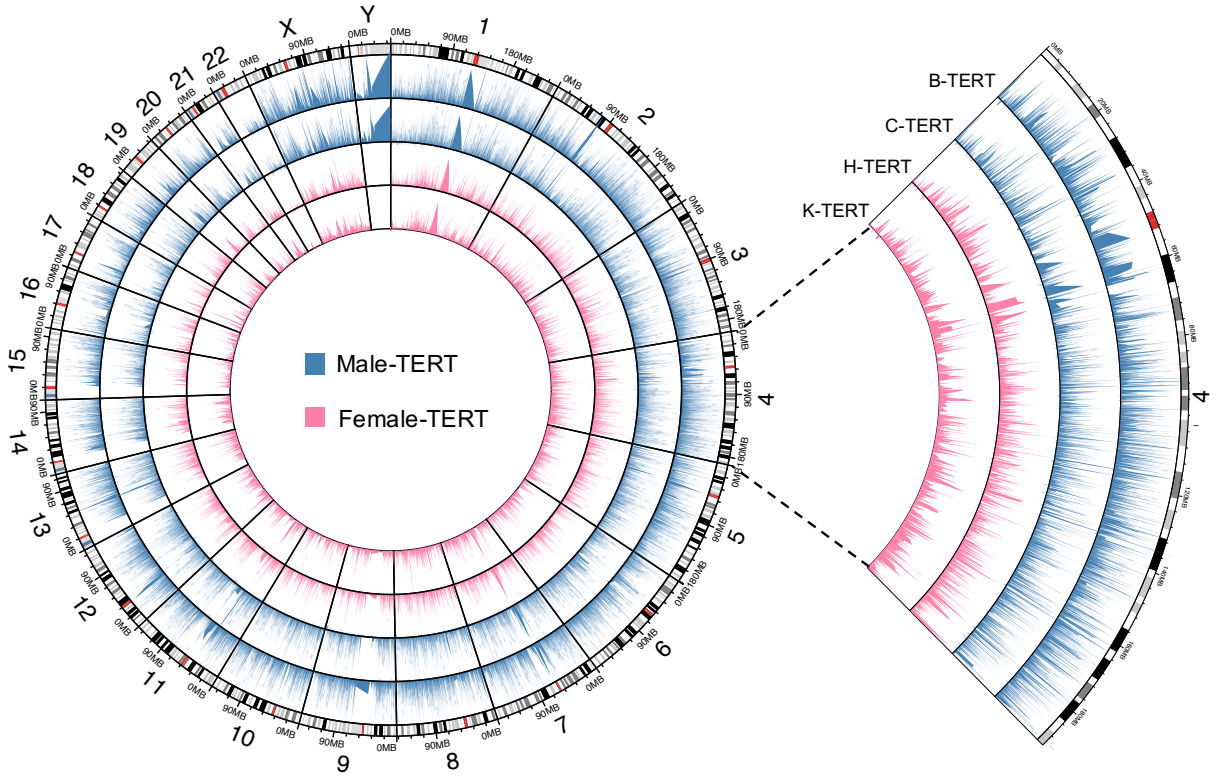


Figure 5.4 Circos plot of genome-wide chromatin accessibility.

Biological replicates were combined to produce an average signal track of chromatin accessibility throughout the genome. The Circos plot shows the ATAC-seq track for each male and female TERT-NHUC across the entire genome. The genome is represented as a circle divided into 24 segments for all chromosomes, with centromeres coloured red. Genome annotation is indicated on the outside of the circle. For clarity, chr4 has been enlarged. From outside to inside the plot shows B-TERT, C-TERT, H-TERT, and K-TERT combined tracks respectively.

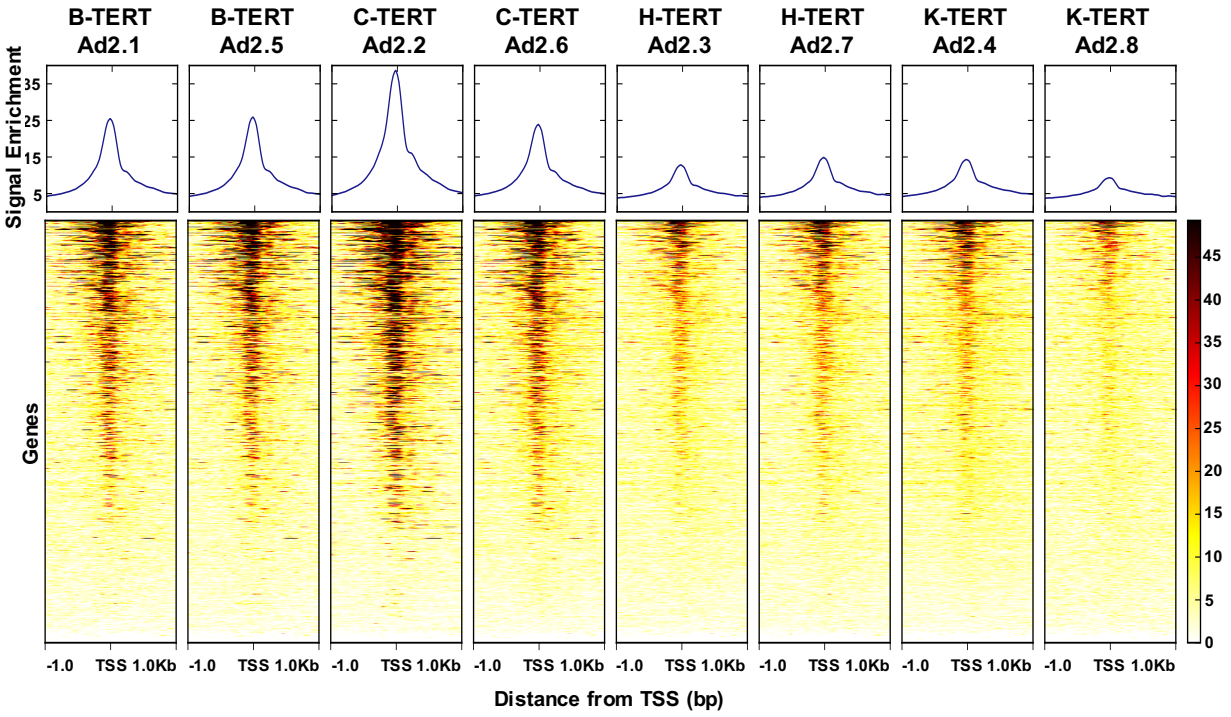


Figure 5.5 Heatmap of ATAC-seq signal around TSSs.

Heatmaps for signal enrichment around all transcription start sites (TSSs) at +/- 1kb were generated for all ATAC-seq samples (bottom) and the average signal enrichment at each base was plotted (above).

Visualising ATAC-seq signal by using IGV showed that peaks of chromatin accessibility were generally shared between male and female TERT-NHUC cells, but exhibited a higher signal in male cells whereas background signal was marginally higher in females. Although male and female TERT-NHUC showed distinct chromatin accessibility profiles, tracks between samples within gender groups remained largely comparable. These observations were made both at the level of individual loci and at the level of the entire genome (Figure 5.3 and Figure 5.4 respectively).

5.2.2.2 Heatmap of signal intensity around TSSs

Transcriptional start sites (TSS) are the centre of transcriptional activation and are immediately flanked by promoter and *cis*-regulatory regions bound by transcription factors, polymerases and a multitude of distal enhancer regions through chromatin looping. This results in a high degree of chromatin accessibility around the TSSs, and especially at active genes (Thurman *et al.*, 2012). ATAC-seq signal is therefore often enriched around TSSs (as seen in Figure 5.3) and can be easily viewed on a heatmap of signal enrichment.

Heatmaps for ATAC-seq signal enrichment were computed for each of the TERT-NHUC samples at all known hg19 TSSs +/-1kb and can be seen in Figure 5.5. These results show that ATAC-seq signal enrichment was indeed centred around TSSs for all samples. However, signal intensity was stronger in male TERT-NHUC compared to the female cells, with a mean peak enrichment at TSSs >2-fold greater in male samples. There was also disparity between the C-TERT and K-TERT duplicates where C-TERT Ad2.2 mean signal enrichment at TSSs was ~1.5-fold greater than C-TERT Ad2.6, where TSS signal enrichment coincided with the two B-TERT samples. K-TERT Ad2.8 was ~1.5-fold less than K-TERT Ad2.4 where TSS enrichment coincided with the two H-TERT samples.

The results in Figure 5.5 compare each of the TERT-NHUC samples on the same scale. However, when computed individually to their own scale, each sample presents a concise TSS signal enrichment across the majority of genes (Appendix E-1). These heatmaps show that an equal proportion of TSSs showed signal enrichment in most samples. The exception, K-TERT Ad2.8, not only showed considerably weaker signal across these regions, but also a high amount of background signal in the region surrounding the TSS.

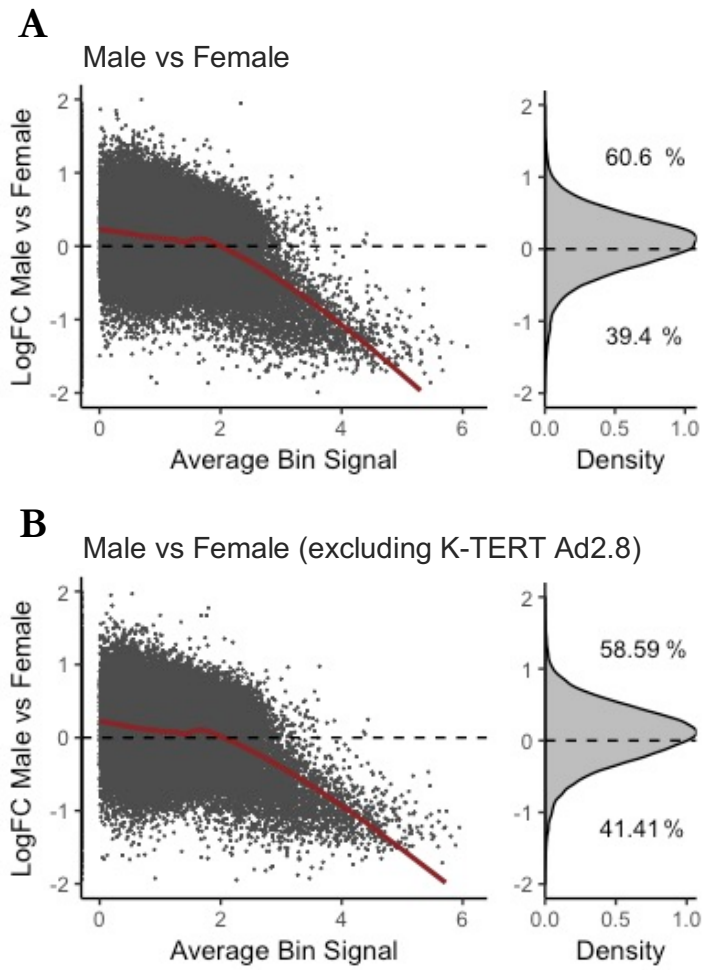


Figure 5.6 MA plot of genome-wide ATAC-seq signal in Male vs Female TERT-NHUC.

The genome was split into bins of 1000bp and the FC of signal enrichment at each bin was determined for male vs female libraries. (Left) MA plot of logFC for each bin was plotted against its average signal across all libraries, (Right) density plot of logFC. **A)** All male and female libraries were used for analysis. **B)** All male and female libraries excluding K-TERT Ad2.8 were used for analysis.

These heatmaps demonstrated that each sample produced signal enrichment at TSSs and conformed to the expected increase of chromatin accessibility at these regions (Thurman *et al.*, 2012). However, the greater signal in males compared to females was again of particular interest and correlated with the previous results in Figure 5.3 and Figure 5.4. The low signal observed in K-TERT Ad2.8 along with the high background signal surrounding the TSS again indicated abnormalities with this sample. This correlated with K-TERT Ad2.8 not grouping with its biological replicate in the hierarchical clustering, the abnormally low peak count and signal, and a low FrIP score (Figure 5.2 and Table 5.1).

5.2.2.3 MA plot of genome-wide ATAC-seq signal between genders

To further represent differences in chromatin accessibility between genders, an M (log ratio) vs A (mean average) plot (MA plot) was generated (Robinson *et al.*, 2009). An MA plot is a modification of the Bland-Altman plot that transforms NGS data onto M (log ratio) and A (mean average) scales (Robinson *et al.*, 2009). To do this, the genome was divided into bins of 1000bp and an alignment signal was obtained for each bin for each library. The average bin signal was then taken across all libraries, and between libraries of the same gender. A LIMMA analysis was then carried out to obtain the fold change (FC) difference for each bin between genders, and was plotted against the average signal for each bin across all libraries (Figure 5.6). The majority of bins should have a low signal and represent regions of the genome devoid of chromatin accessible peaks, whereas bins with increased signal will include regions of the genome containing peaks of chromatin accessibility. When genome-wide ATAC-seq signal is similar between groups, the MA plot centres around a FC of 0.

The MA plot of all male vs all female samples (Figure 5.6A) shows that fold change for bins with an average peak signal <2 is marginally more in favour of female TERT-NHUC. The accompanying density plot further shows that the majority of bins (58.59%) had increased signal in female cells. However, as average bin signal exceeds 2.5, FC increasingly moves in favour of male TERT-NHUC.

The MA and accompanying density plot therefore show that female TERT-NHUC generally had increased signal in bins representing background regions devoid of chromatin accessibility, whereas high-signal bins, that include peaks of chromatin accessibility, were almost exclusively in favour of male cells. These results mirror the IGV tracks of Figure 5.3 that showed that background signal was marginally greater in female-TERT-NHUC, but that the smaller, concise peaks were greater in males.

K-TERT Ad2.8 had previously shown abnormalities in previous results, including a high background signal with indistinguishable peaks that coincided with a low peak number

and FrIP score. A second MA plot was therefore generated to compare male and female TERT-NHUC ATAC-signal, but excluding K-TERT Ad2.8 from the female group (Figure 5.6B). The hypothesis was that removing this sample would reduce the background bias in female samples and reduce the high signal bins exclusive to males. However, the previous results persisted where low-signal bins were still in marginally in favour of female TERT-NHUC and male TERT-NHUC still exclusively harboured bins with higher signal. The density plot does show that the total number of peaks with a FC greater in female TERT-NHUC dropped by 2.01% to 58.59% of bins. This is likely due to female TERT-NHUC having a lower average background signal with K-TERT Ad2.8 removed. The results of Figure 5.6B therefore suggest that K-TERT Ad2.8 only marginally contributed to the overall pattern of increased background signal observed in female TERT-NHUC.

This section has shown that the ATAC-seq results from TERT-NHUC conform to the expected patterns of chromatin accessibility, where active genes present with peaks at the TSS and at *cis*-regulatory regions, as shown using IGV and heatmaps. However, the ATAC-seq signal distribution between male and female TERT-NHUC was shown to be distinctly different. IGV and Circos plots showed that peaks of chromatin accessibility shared between all samples had greater signal in male samples. Heatmaps centred at TSSs of all genes further showed that chromatin enrichment was greater in male TERT-NHUC at these regions. Furthermore, IGV and MA plots showed that the ATAC-seq background signal was marginally greater in female TERT-NHUC. Cumulatively, these results suggest that either female TERT-NHUC have a genome-wide decrease in chromatin accessibility, or reveal an issue with the ATAC assay itself for these samples.

5.2.3 Post-ATAC experiments

5.2.3.1 MNase Digestion Assays

The key finding from the ATAC-seq results was the indication that male TERT-NHUC chromatin might be in a more relaxed state than female TERT-NHUC chromatin. Therefore, a series of wet-lab experiments to confirm or disprove these findings was carried out. The first experiment concerned a time-course digestion assay using micrococcal nuclease (MNase) on each of the TERT-NHUC cell lines used for ATAC-seq was done as a pilot study (Figure 5.7). MNase preferentially cuts linker-DNA between nucleosomes whilst avoiding partially protected nucleosomal DNA. Partial digestion, similar to that obtained by Tn5 transposition, can therefore be achieved using MNase, where increasing digestion time with MNase increases chromatin digestion, with partial digestion being observed as a periodic banding by gel electrophoresis (Axel, 1975; Mieczkowski *et al.*, 2016).

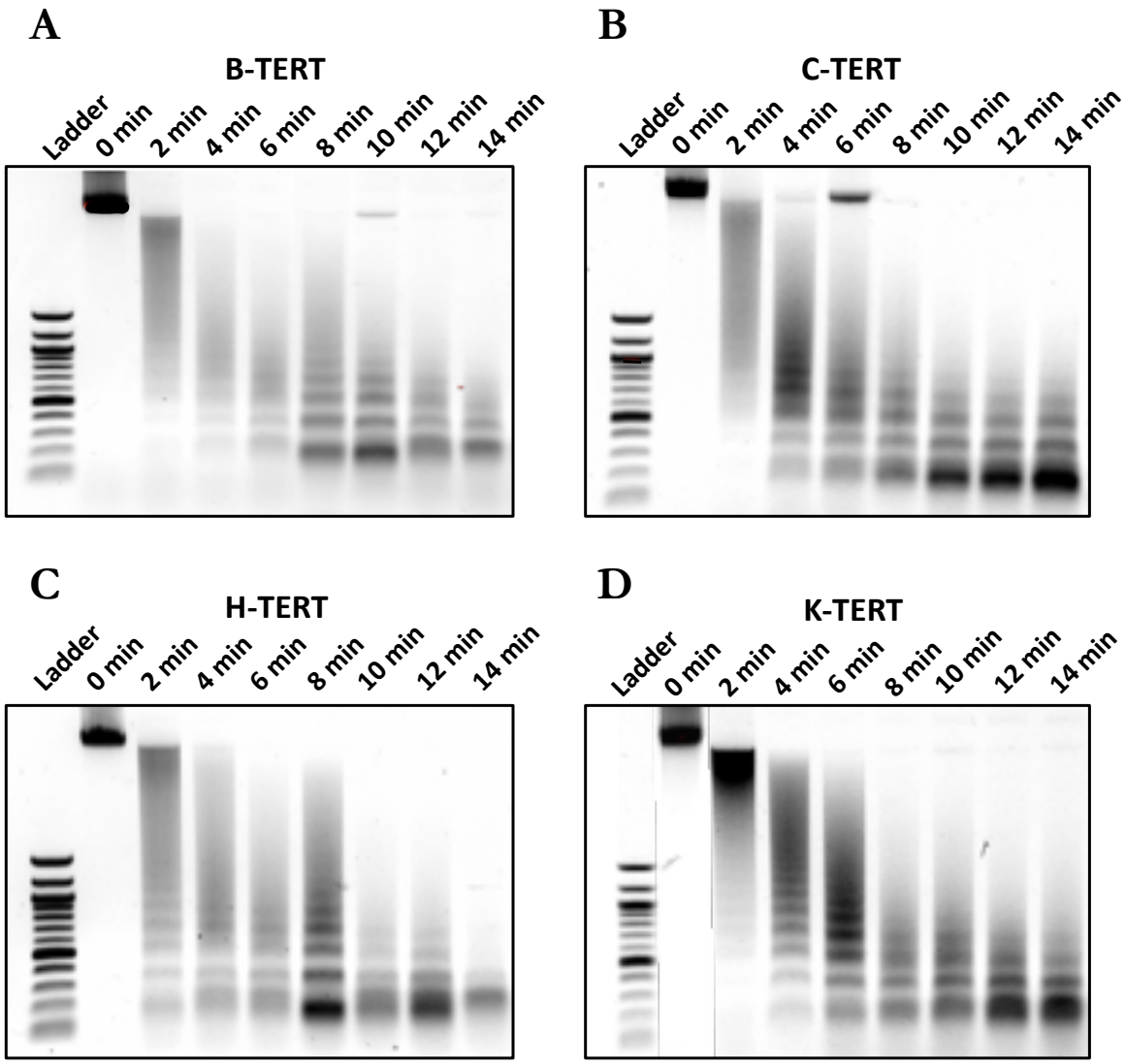


Figure 5.7 TERT-NHUC MNase digestion assays.

Chromatin from TERT-NHUC cells was subjected to digestion by Micrococcal Nuclease (MNase) over 8 time points ranging from 0 min to 14 min, then visualised using agarose gel electrophoresis. MNase digestion assays were carried out on **(A)** B-TERT, **(B)** C-TERT, **(C)** H-TERT, and **(D)** K-TERT cells. This experiment was carried out as a single pilot study.

As MNase preferentially cuts linker DNA, more condensed chromatin takes longer to reach full digestion than more euchromatic chromatin. Given this principle, it was hypothesised that female TERT chromatin would take longer to reach full digestion than male TERT chromatin based on the ATAC-seq data. However, the time-course MNase digestion assay on the TERT-NHUC cells did not confirm this hypothesis (Figure 5.7), where in general B, C, and H-TERT chromatin reached maximum-digestion between 10-12 min. It should be noted that the maximum digestion here may not be equivalent to full-digestion as chromatin still shows nucleosomal banding (with the exception of H-TERT at 14 min). For B-TERT and H-TERT chromatin, where more material was available for the digestion assay, digestion did not progress further from 14 min to 20 min (Appendix). This is likely due to MNase exhausting necessary reagents (such as Ca^{2+}) in the reaction which then prohibits further digestion from 14 min onwards. Nevertheless, the digestion progression was similar between TERT-NHUC cells between 0-14 min with the exception of K-TERT which showed very little digestion at 2 min and persisted as a 2 min lag in digestion compared to the other cell lines. This indicates that K-TERT chromatin is slightly more condensed than the other TERT-NHUC cells, but the lack of this pattern in H-TERT discredits this as a female TERT-specific observation. Furthermore, H-TERT chromatin was the only sample to achieve full chromatin digestion at 14 min, and B-TERT chromatin took 12 min to reach the maximum digestion achieved by K-TERT, further demonstrating a lack of gender-associated MNase digestion dynamics. Overall, the MNase time-course digestion assay largely disagreed with the ATAC-seq results and did not indicate increased heterochromatin for either gender. The results also complement the library preparation samples in Figure 4.8, which also did not display gender-associated nucleosomal differences derived from Tn5 transposition over time.

5.2.3.2 Global histone mark levels

Open and closed chromatin are commonly associated with different histone modifications. For instance, H3K4me3 and H3K27ac are often localised in regions of open chromatin and mark active promoters and enhancers, while H3K27me3 and H3K9me3 are often found at heterochromatin and the promoters of silent genes (Kouzarides, 2007; Kundaje *et al.*, 2015). Given the genome-wide differences in chromatin accessibility found by ATAC-seq in male and female TERT-NHUC cells, it can be hypothesised that global gender-associated differences would be observed for such histone marks. The global histone modification levels for H3K4me3, H3K27ac, and H3K27me3 were therefore assessed by western blot (WB) on purified histone extractions from B, C, H, and K-TERTs, and using unmodified H3 as a loading control (Figure 5.8).

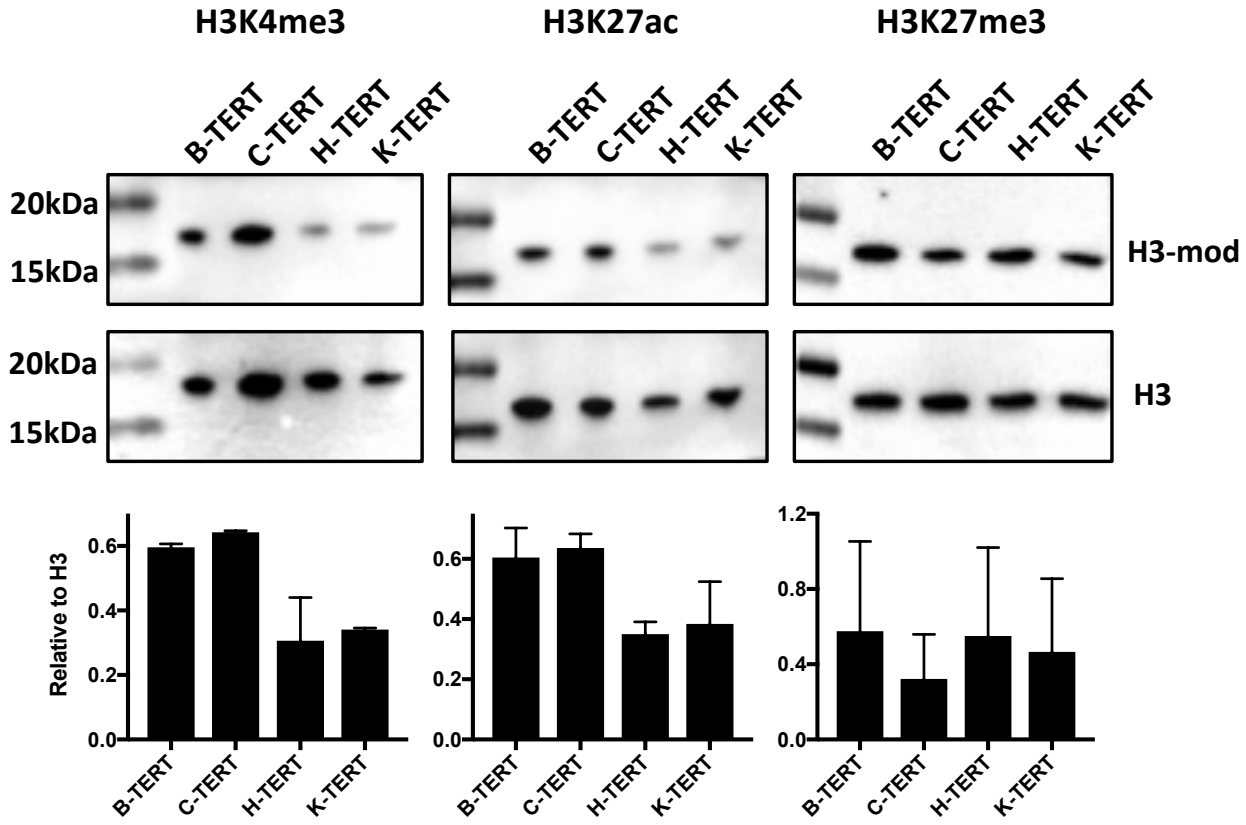


Figure 5.8 Global histone modification levels in TERT-NHUC cells.

Purified histones from B, C, H, and K-TERT cells were used to carry out western blots (top) for the activating histone marks H3K4me3 and H3K27ac, and the inactivating histone mark H3K27me3. H3 was used as a control for loading. Histone modifications were also quantified and normalised to unmodified H3 (bottom). Bar charts are the mean of three biological replicates with error bars indicating SD.

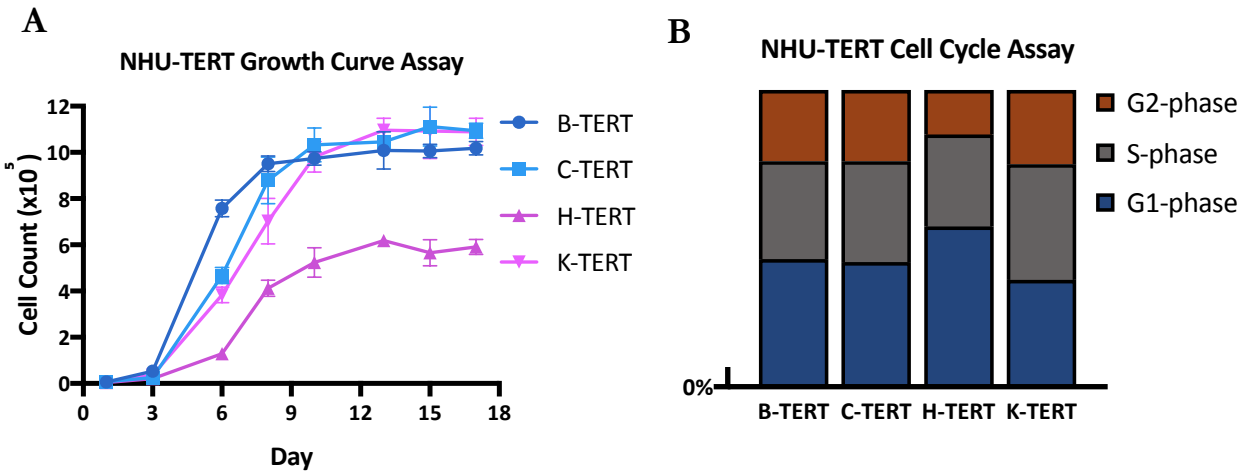


Figure 5.9 TERT-NHUC growth and cell-cycle assays.

(A) 5x10^4 B, C, H and K-TERT cells were seeded in 6-well plates and counted every 2-3 days over an 18-day period. Counting was carried out in biological and technical triplicate. **(B)** 5 x 10^5 B, C, H and K-TERT cells were assayed for cell cycle status using the Guava cell cycle kit and data was acquired using the Guava flow-cytometer to determine distribution of cells in Gap-phase 1 (G1, blue), Synthesis-phase (S, grey), Gap-phase 2 (G2, orange).

Given that ATAC-seq indicated a more condensed chromatin state in female TERT-NHUC, it was hypothesised that the heterochromatin marker H3K27me3 would be higher in H-TERT and K-TERT cells, and/or the active markers H3K4me3 and H2K27ac would be enriched in B-TERT and C-TERT cells. Figure 5.8 agrees with the second part of this hypothesis with WB showing that H3K4me3 and H2K27ac were indeed higher in both male TERT-NHUC compared to female cells. After normalising each modification to H3, a semi-quantitative approach to measuring global histone modification levels showed that H3K4me3 and H2K27ac were ~2-fold greater in both male TERT-NHUC compared to both female lines. No differences were observed in global H3K27me3 levels between the cell lines.

These results indicate that the differences in chromatin accessibility observed by ATAC-seq may be related to euchromatic activation at individual loci instead of widespread formation of heterochromatin. As the MNase digestion assay and WB for H3K27me3 failed to show gender-associated differences, the chromatin accessibility disparity observed in the ATAC-seq data is unlikely to be a result of increased heterochromatin in female TERT-NHUC.

5.2.3.3 TERT-NHUC growth curve assay and cell cycle analysis

It is well known that chromatin state is heavily influenced by the cell cycle (Ma *et al.*, 2015). The earlier cell-cycle stages (G1/S) have a more relaxed chromatin state, allowing for increased gene expression, histone synthesis and the binding of transcription factors. In contrast, chromatin condensation and the dissociation of transcription factors and other DNA/chromatin-binding proteins from the chromatin are seen in late-stage cell cycle and mitosis (G2/M) (Ma *et al.*, 2015). Given the gender-associated chromatin-state differences observed in the ATAC-seq data, it is reasonable to suspect that these may relate to differences in the cell cycle status and growth rates between male and female TERT-NHUC lines. Therefore growth-curves and cell-cycle assays were determined for each of the TERT-NHUC lines (Figure 5.9). The hypothesis was that the increased chromatin-condensation state in female TERT-NHUC indicated by ATAC-seq may be due to cell cycle stage, and that female TERT-NHUC would show a higher proportion of cells in late-stage cell cycle and increased proliferation compared to male TERT-NHUC.

TERT-NHUC reached full confluence by 9 days, with similar cell counts observed for B, C and K-TERT, and a lower cell count for H-TERT (Figure 5.9A). Growth-rate was initially fastest in B-TERT, reaching ~75% confluence by day 6, whereas C and K-TERT

were ~50% and H-TERT ~15% of full confluence at the same time point. Overall the growth curves showed minimal differences in proliferation rates between B, C and K-TERT lines, but a slower proliferation rate with a lower confluent cell number for H-TERT. The cell cycle assay complemented the results of the growth curve assay by showing minimal differences in cell-cycle stage between B, C, and K TERT, and a marginal increase in the proportion of H-TERT in G1-phase, with a lower proportion of cells in G2-phase (Figure 5.9B). Together, these results do not reveal a gender-associated difference in growth-rate or cell-cycle stage in these cell lines, and do not support the aforementioned hypothesis.

In summary, the results from the cell-cycle, growth-curve, MNase digestion assays, and WB for the heterochromatin marker H3K27me3 do not support the hypothesis of global chromatin condensation in female-TERT. Instead, the ATAC-seq results may be a result of global activation of regulatory regions at individual loci throughout the genome in male TERT-NHUC, as supported by the increase in the narrow-peak activating histone marks H3K4me3 and H3K27ac as shown by WB.

5.2.4 Analysis of chromatin-accessible peaks

Once reads are aligned to the genome, concise regions that have significantly higher signal enrichment compared to the background can be identified, and were previously seen in the IGV tracks of Figure 5.3. These regions are referred to as chromatin-accessible peaks and are identified using MACS2 centred around Tn5 cleavage sites by using –shift -100 –extsize 200. A total of 390,488 peaks were identified with a mean of 48,811 (median 51,172) peaks per sample. However, when considering gender, 264,393 peaks (67.7%) were found in male TERT-NHUC with a mean of 66,098 (median 63,839) peaks per male sample, whereas 126,095 peaks (32.3%) were found in female TERT-NHUC with a mean of 31,524 (median 34,409) peaks per female sample. This next section concerns the analysis carried out on these chromatin-accessible peaks and includes correlating peaks between samples, differential expression of peaks between genders, and peak annotation.

5.2.4.1 Correlation of chromatin-accessible peaks

A heatmap of correlation between chromatin-accessible peaks shows both a strong correlation between biological replicates and between samples from the same gender. Males and females cluster into distinct groups and biological replicates show the strongest correlation (Figure 5.10A). This is an improvement on genome-wide signal correlation between samples which showed K-TERT Ad2.8 not falling within either male or female groups (Figure 5.10B). PCA analysis on chromatin-accessible peaks further demonstrated clustering of samples from the same gender, again with biological replicates associating most

closely with each other (Figure 5.10C). Again K-TERT Ad2.8 appears as an outlier (although still within the female group), although C-TERT Ad2.6 also falls away from its biological replicate and the male group by nearly the same margin. A heatmap of "binding affinity" shows signal enrichment of all peaks across all samples (Figure 5.10D). Although the samples themselves cluster with distinct male and female groups as in Figure 5.2, it is clear from this plot that the signal enrichment at male peaks is generally greater than at female peaks, coinciding with the aforementioned reduced signal pattern in the previous section. Differential signal enrichment analysis of peaks was carried out using DiffBind and will be discussed later in this chapter.

The plots of Figure 5.10 show a strong correlation of peaks within male and female groups and between biological replicates. This may reflect the disparate number of peaks between genders and/or the lower signal-enrichment of peaks generally observed in female TERT-NHUC. Unlike with the genome-wide signal correlation, K-TERT Ad2.8 is better correlated with its biological replicate and falls within the female group, although PCA analysis still shows it as distinct from its biological replicate as also seen with C-TERT Ad2.6.

5.2.4.2 Identifying gender-associated chromatin-accessible peaks

The initial hypothesis of this project was that chromatin-accessible regions between male and female normal bladder cell lines would be largely similar, with only subtle differences that may promote gender biases seen in bladder cancer incidence and genomic profile. So far, the results have suggested a widespread decrease of chromatin accessibility in female cells. With respect to chromatin-accessible peaks this is seen as a smaller number of overall peaks called and a lower enrichment-signal at called-peaks in female TERT-NHUC. This section aims to identify gender-associated chromatin-accessible regions using a combination of occupancy-based analysis that considers peaks at given loci that are shared between samples, and an affinity-based analysis that carries out differential peak-enrichment analysis between genders. It should be noted that from here, the term "peaks" refers to chromatin-accessible regions called using MACS2 during analysis.

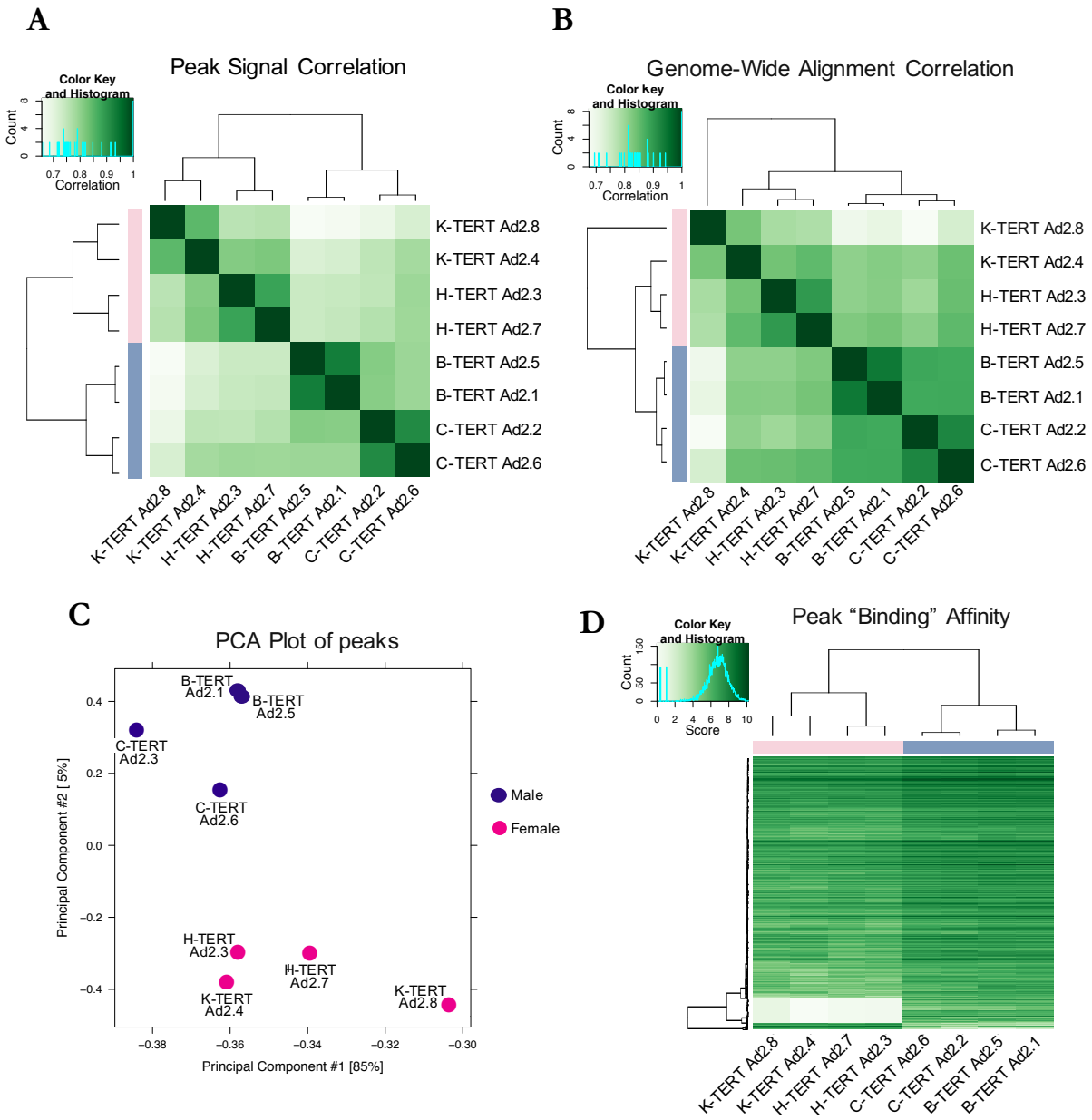


Figure 5.10 Correlation of genome-wide signal and chromatin-accessible peaks between TERT-NHUC cells using DiffBind.

A) Heatmap of signal correlation of chromatin-accessible peaks identified using MACS2. **B)** Heatmap for the correlation of genome-wide alignment signal (as seen in Figure 5.2B) where the genome was split into bins of 1000bp and alignment signals were correlated between sample bins. **C)** Principal component analysis (PCA) plot of chromatin-accessible peaks. **D)** Heatmap of peak "binding" affinity for each sample where each line corresponds to a peak and peak signal is represented by heatmap colour in each sample. For each plot blue bars/points represent male samples and pink bars/points represent female samples.

5.2.4.2.1 Occupancy-based analysis to identify gender-associated peaks

An overlap-rate plot of peaks between samples shows the number of peaks shared between TERT-NHUC peak-sets of increasing number. These plots were used to determine the number of overlapping peaks between all samples, male samples, and female samples (Figure 5.11A, B, & C) and show that out of the 390,488 peaks called by MACS2, 128,555 peaks occupied unique sites, with 82,234 peaks shared between at least two samples, and only 10,547 peaks found in all samples (Figure 5.11A). For male TERT-NHUC, 81,395 peaks, from a total of 264,393 called peaks, occupied unique sites, with 78,541 shared by at least two male samples and 44,574 peaks shared between all male samples (Figure 5.11B). For female TERT-NHUC, 50,417 peaks from a total of 126,095 called peaks occupied unique sites, with 33,334 peaks shared by at least two female samples and 10,879 peaks shared between all female samples (Figure 5.11D).

To better visualise how peak occupancy is shared between samples, Venn diagrams were produced for male and female TERT-NHUC separately. These Venn diagrams show that in female TERT-NHUC 17,083 peaks were not shared with any other female sample, the majority of which can be found in K-TERT Ad2.4 where 10,760 peaks were unique to this sample. In contrast, only 2,854 peaks were found in only one sample of male TERT-NHUC, 2,451 of which were unique only to C-TERT Ad2.2. This shows a high reproducibility rate for peaks in male TERT-NHUC, and also in female TERT-NHUC when excluding K-TERT Ad2.8. When considering the overlap of peaks between biological replicates, B-TERT and H-TERT had roughly equal numbers of peaks unique to each individual replicate, with the majority of peaks shared between biological replicates. However, only 369 of 57,736 peaks in C-TERT Ad2.6 and 186 of 12,670 peaks in K-TERT Ad2.8 were unique to those replicates, with 99.4% and 98.5% of all peaks shared with their biological replicates respectively. This may therefore indicate that an increased sequencing depth may have been required in these libraries to allow identification of the peaks found in their biological replicates.

Two occupancy-based analyses for identifying gender-associated peaks were carried out (Welch *et al.*, 2014). The first measures overlaps in peaks shared between all male TERT-NHUC with peaks shared between all female TERT-NHUC (Figure 5.11F). The overlap consisted of the 10,547 peaks that are present in all samples, leaving 34,027 peaks exclusive to male TERT-NHUC (male-specific peaks) and 332 peaks exclusive to female TERT-NHUC (female-specific peaks). For male-specific peaks, 33,159 (97.45%) peaks were located in autosomes and 864 (2.54%) peaks were located on sex chromosomes (810 (2.38%) chrX; 54 (0.16%) chrY), whereas for female-specific peaks 310 (93.38%) peaks were located on

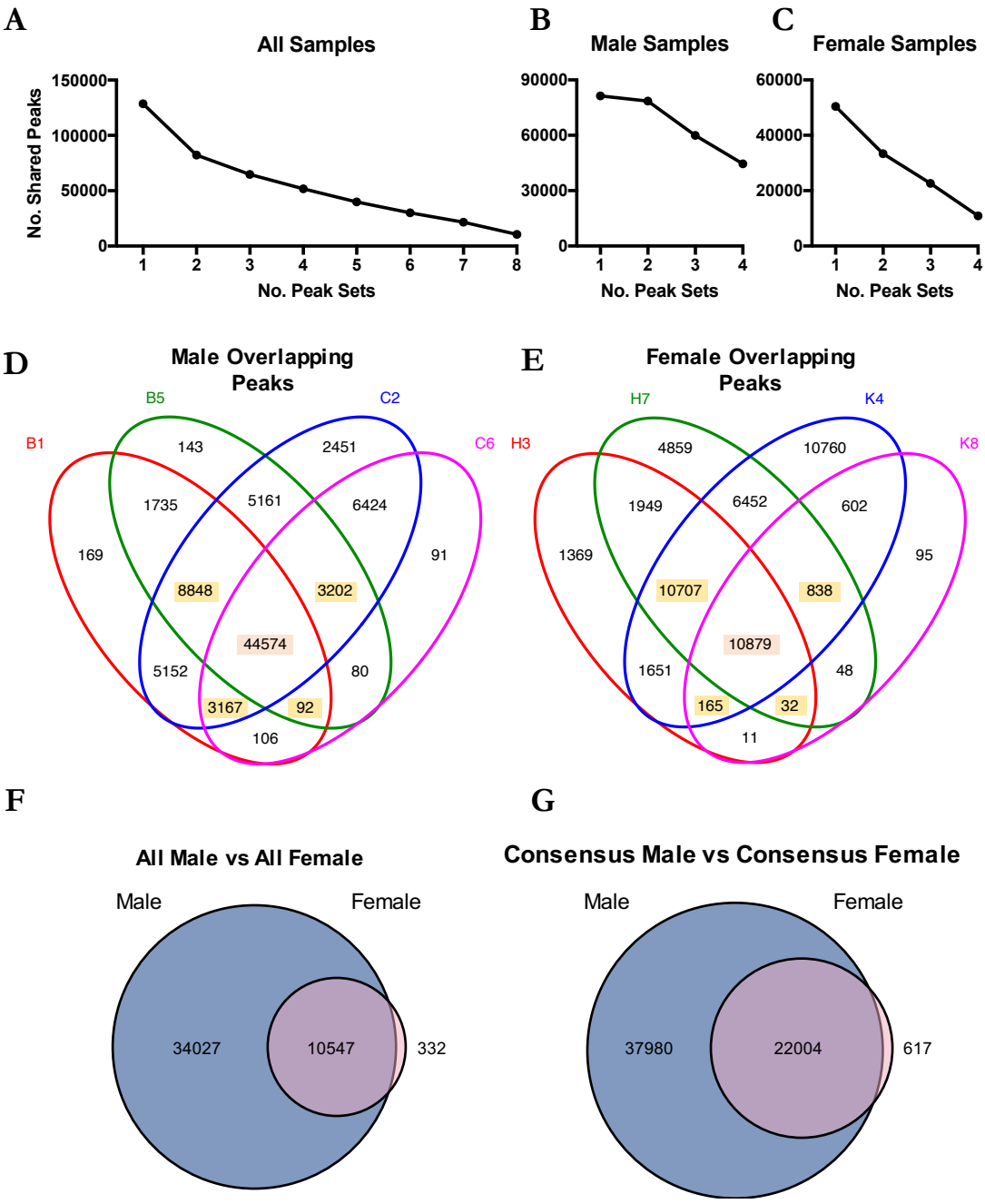


Figure 5.11 Occupancy-based identification of gender-associated peaks.

A-C) Overlap rate plots showing the number of peaks shared over an increasing number of peak sets for **A)** all samples, **B)**, male samples, and **C)** female samples. **D-E)** Venn diagrams showing overlapping peaks in **D)** male samples, and **E)** female samples. **F)** Venn diagram of peaks found in all males and peaks found in all females (highlighted orange in D) and E)). **G)** Venn diagram of consensus peaks found in 3 out of 4 male samples and 3 out of 4 female samples (highlighted yellow and orange in D) and E)). Blue circles are male peaks and pink circles are female peaks.

autosomes and 22 (6.72%) on chrX. The second occupancy-based approach used consensus calling to identify peaks that were shared in at least three out of the four male samples, and overlapping these with peaks that were found in three out of the four female samples (Figure 5.11G). Using this less stringent approach, 37,980 peaks were exclusive to male TERT-NHUC (male-consensus-specific peaks) whereas 617 were exclusive to female TERT-NHUC (female-consensus-specific peaks). For male-consensus-specific peaks, 37,000 (97.42%) of peaks were located on autosomes and 980 (2.58%) of peaks were located on sex chromosomes (915 (2.41%) chrX; 65 (0.17%) chrY), whereas for female-consensus-specific there were 578 (93.62%) peaks located on autosomes and 39 (6.38%) on chrX. Both approaches therefore show that the majority of gender-associated peaks are found throughout the genome and not located exclusively on the sex chromosomes.

5.2.4.2.2 Affinity-based analysis to identify gender-specific peaks

The occupancy-based analysis for identifying gender-associated peaks relies on a peak being present within gender groups and does not take into consideration signal intensity of peaks. Therefore, an affinity-based analysis for identifying gender-associated peaks was also carried out (Figure 5.12). Differential-enrichment analysis of peaks between genders was carried out using the DESeq2 analysis incorporated into DiffBind (Stark and Brown., 2011; Love *et al.*, 2014), where differentially-enriched (DE) peaks were considered as those with a FC $\geq$ 1.5 and an FDR $\leq$ 0.05.

Using this approach, a total of 63,316 DE peaks were identified between genders, of which 62,945 were enriched in males and only 371 enriched in females (Figure 5.12A). A mean enrichment FC of 2.11 (median 2.15) in favour of males was observed for all DE peaks. In male cells 61,762 (98.12%) DE peaks were located on autosomes, although all 96 chrY peaks (0.16% of total) were found in the top 100 DE peaks. Of the 371 female-DE peaks, 338 (91.11%) were located on autosomes, with 33 (8.89%) DE peaks located on chrX. An MA plot shows that peaks were predominantly enriched in males regardless of the average peak signal across all samples (Figure 5.12B). This means that peaks with a low average signal enrichment between all samples showed an increased FC in males, as was the case with peaks with a high average signal enrichment between all samples. IGV was then used to visualise the top DE peaks for each gender, as well as typical peaks for the mean FC and median FC (Figure 5.12C).

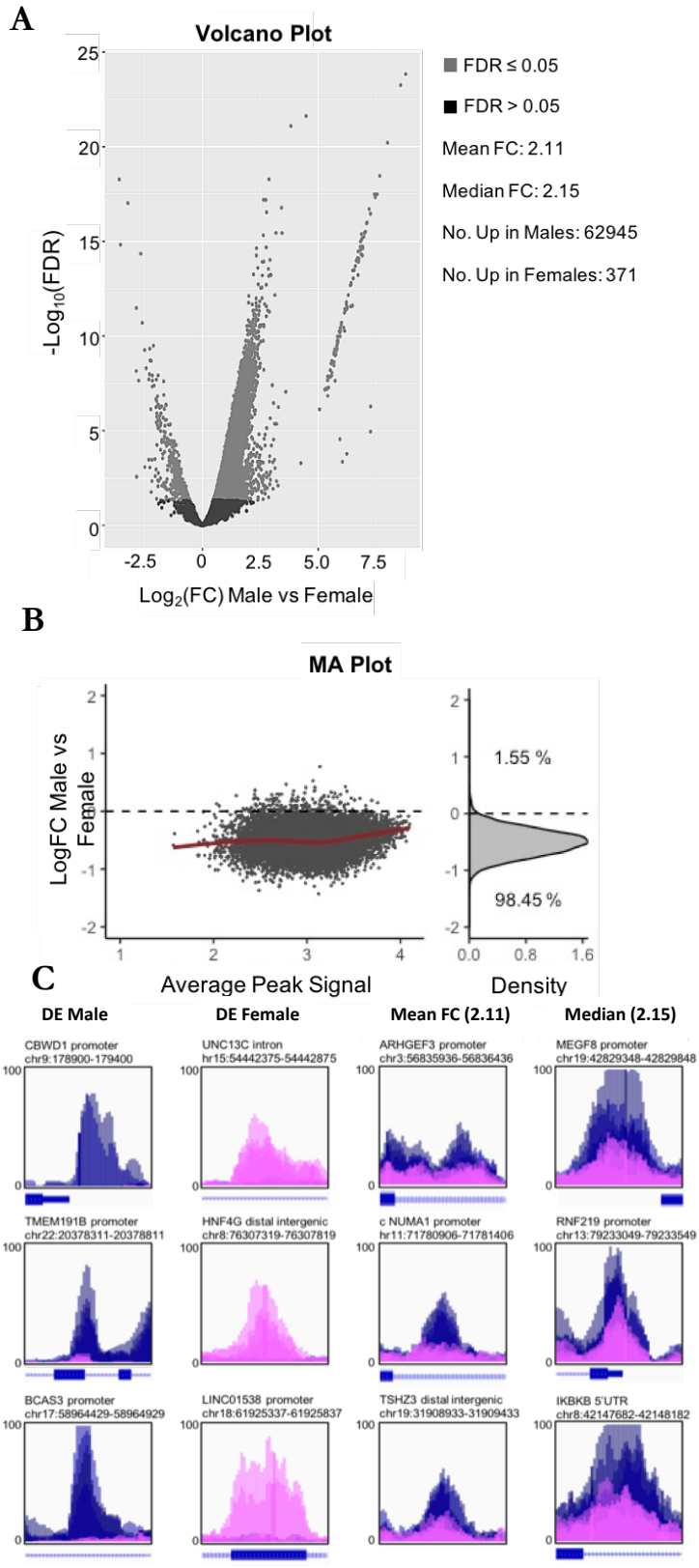


Figure 5.12 Affinity-based identification of gender-associated peaks.

DESeq2 was used to carry out differential enrichment analysis of ATAC-seq peaks between gender. This is visualised as **A)** a Volcano plot of LogFDR over LogFC, and **B)** an MA plot of LogFC over average peak signal across all samples. Light grey points in both plots are DE peaks with an FDR ≤ 0.05. **C)** IGV was used to visualize DE peaks. All ATAC-seq tracks were overlaid, with male tracks coloured blue and female tracks coloured pink. Representative peaks are shown for Male DE peaks, Female DE peaks, Mean FC and Median FC. The peaks for Mean and Median FC shown here were "non-gender-associated peaks" identified in the overlap of the Venn-diagram Figure 5.11F.

The affinity-based analysis identifies >30,000 more peaks as male-associated compared to the occupancy-based analysis. This is likely derived from the ~22,000 peaks that are found in both male and female samples as well as those that were not filtered from the consensus overlapping (Figure 5.11). Given the average FC enrichment of 2.11 in favour of males, these aforementioned peaks are likely to have increased enrichment in males and contribute to the increased number of male-DE peaks. Indeed, many of the non-gender-associated peaks from the consensus calling in Figure 5.11G have a FC in male cells equal to that of the mean and median FC, representative examples of which can be seen in Figure 5.12C.

5.2.4.2.3 Combining occupancy- and affinity-based analyses to identify gender-associated peaks

The final approach to identifying gender-associated peaks was to combine the previous occupancy- (using peaks identified by consensus occupancy analysis) and affinity-based methods. This was done by filtering the DE peaks identified in Figure 5.12 for peaks that were found in at least three of the four samples for the respective gender as in Figure 5.11G. This approach identified a total of 54,070 gender-associated peaks, of which 53,849 were male-specific and 221 were female-specific. 52,725 (97.91%) of male-specific peaks were located in autosomes, with chrX and chrY accounting for 1,061 (1.97%) and 63 (0.12%) peaks respectively. For female-specific peaks, 202 (91.4%) were located on autosomes, with 19 (8.6%) located on chrX.

5.2.4.2.4 Comparison of the gender-associated peaks identified using different methods

This section has applied four different approaches to identify gender-associated peaks, with each approach obtaining differing numbers of peaks. Table 5.2 compares the number of peaks obtained by each approach as well as the proportion of peaks located on autosomes and sex chromosomes. The comparison shows that the two consensus-based approaches identified nearly two-fold fewer male-associated peaks than approaches incorporating differential enrichment analysis. The same was not seen for female-associated peaks, and instead the largest difference was observed within the occupancy-based approaches where consensus calling resulted in ~two-fold greater increase in female-associated peaks compared to the specific-occupancy based approach. The proportion of peaks located on autosomes was comparable between approaches within genders. However, between approaches, males showed a mean proportion of 97.73% of peaks located on autosomes whereas in females the mean proportion was 92.38%. The mean proportion of peaks located on chrX in females was 7.65% compared to 2.12% in males.

Table 5.2 Number and proportion of gender-associated peaks identified using each of the discussed methods; in total, on autosomes, on chrX, and on chrY

		Total	Autosomes	chrX	chrY
Occupancy: Specific	**Male**	34027	33159 (97.45%)	810 (2.38%)	54 (0.16%)
	Female	332	310 (93.38%)	22 (6.72%)	-
Occupancy: Consensus	**Male**	37980	37000 (97.42%)	915 (2.41%)	65 (0.17%)
	Female	617	578 (93.62%)	39 (6.38%)	-
Affinity	**Male**	62945	61762 (98.12%)	1083 (1.72%)	96 (0.16%)
	Female	371	338 (91.12%)	33 (8.89%)	-
Affinity + Occupancy	**Male**	53849	52725 (97.91%)	1061 (1.97%)	63 (0.12%)
	Female	221	202 (91.4%)	19 (8.6%)	-

A further comparison regarding distribution and position of gender-associated peaks identified by each of the approaches was also carried out and showed comparable distributions within each gender between each of the approaches (Figure 5.13). These comparisons also showed that male-associated peaks were commonly located at distal intergenic, intronic, and promoter regions, whereas the majority of female-associated peaks were located at distal intergenic regions, with very few located within promoter regions close to the TSS (Figure 5.13A). Furthermore, the majority of female-associated peaks (~60%) were located >100kb away from the TSS, whereas male-associated peaks were most commonly found 10-100kb away from the TSS (~45%). Female peaks also had a very low proportion located close to the TSS and within promoters compared to male-associated peaks. The proportion of peaks found within intronic regions of genes was comparable between genders (~17%).

Finally, non-gender-associated peaks (those identified in the overlap of Figure 5.11G) were more commonly located at promoter regions within 1kb of the TSS compared to gender-associated peaks. However, the distribution of non-gender-associated peaks was more similar to male-associated peaks than female-associated peaks.

Four different approaches to identify gender-associated peaks were carried out and discussed throughout the previous sections. This included two occupancy-based approaches (specific and consensus), an affinity-based approach, and a combination of affinity and occupancy-based approaches. The approaches showed comparable peak-feature distributions and distance from TSS for gender-associated peaks within each gender, but differed in the number of peaks called owing to the stringency/facility of the approach used. For these reasons, the gender-associated peaks identified using the combined occupancy/affinity approach (section 5.2.4.2.3) will be considered as gender-associated peaks throughout the remainder of this chapter, and were used for further analysis.

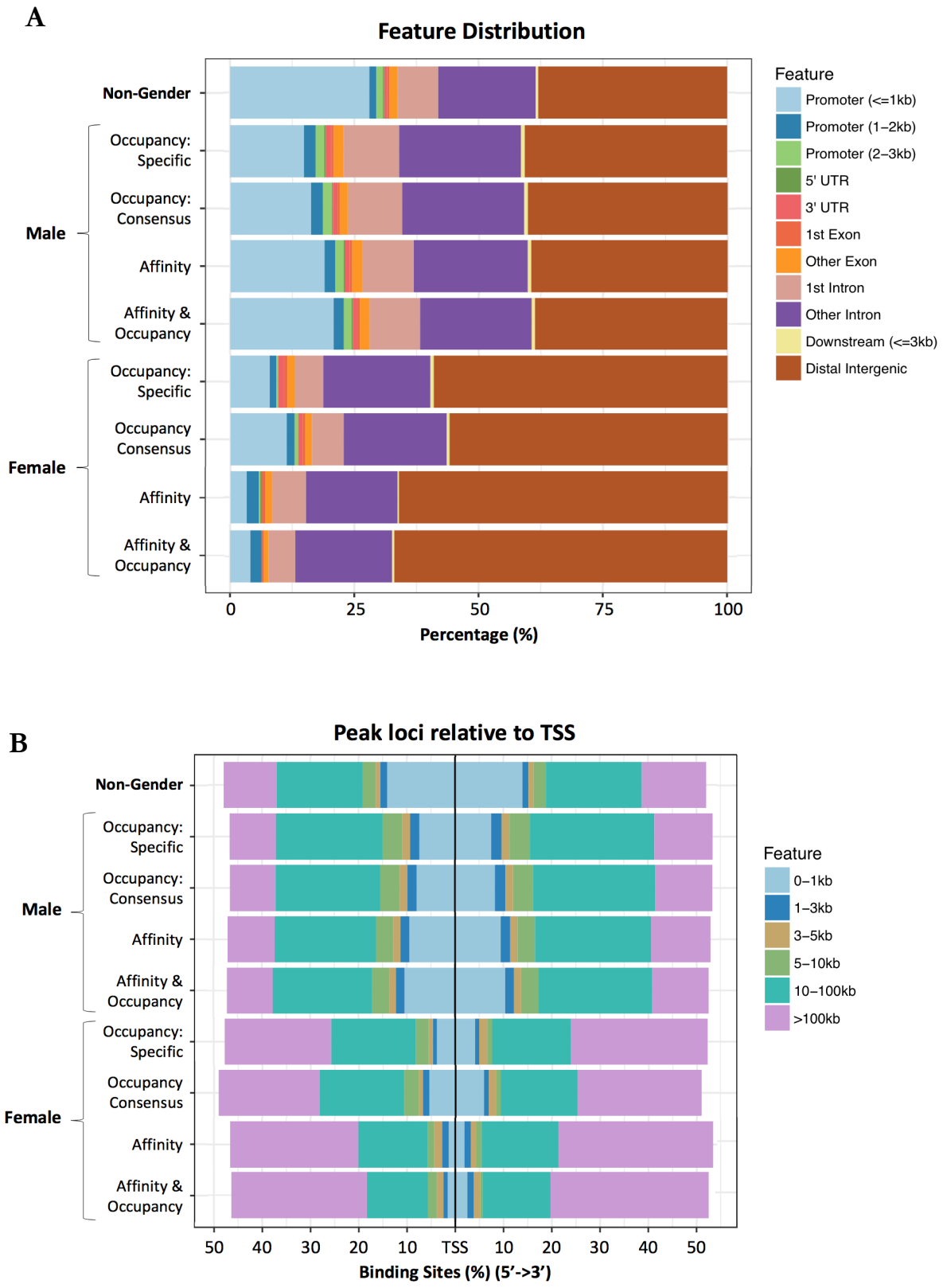


Figure 5.13 Gender-associated peak distribution and position.

Gender-associated peaks identified by each of the techniques as well as non-gender-associated peaks identified in Figure 5.11G were **A)** annotated for genomic feature (Promoter, 5'UTR, 3'UTR, exonic, intronic, and intergenic) and **B)** distance from nearest TSS.

5.2.4.3 Functional analysis of gender-associated peaks

The gender-associated peaks obtained in Section 5.2.4.2.3 were in part determined using differential enrichment analysis. Therefore, peaks could be ranked according to logFC to show those most strongly associated with each gender (Tables 5.3 and 5.4, and Appendix E-2 and E-3). Perhaps unsurprisingly, many of the most differentially enriched peaks were located on the sex chromosomes, with 64 of the top 100 peaks in males located on chromosome Y at 20 different loci, and females showing many peaks on chromosome X, including at the ubiquitously female-expressed RNA gene XIST (Appendix E-2 and E-3 respectively). Tables 5.3 and 5.4 show the top 25 autosomal peaks for males and females respectively, with the top 100 gender-associated peaks across all chromosomes shown in Appendix E-2 and E-3.

The 52,725 male-associated peaks were annotated to a total of 15,067 unique genes, where each gene had a mean of 3.6 (median 2) male-associated peaks. Therefore, nearly three quarters of all genes in the genome were linked to at least one male-associated peak. Genes with the greatest number of male-associated peaks include TENM4, EFNA5, CSMD3, and PCDH7, all of which had over 40 male-associated peaks. However, over 100 gene-loci are annotated with at least 25 male-associated peaks.

Females had only 146 unique genes linked to female-associated peaks with a mean of 1.5 peaks per gene (median 1). Fifty-six of these genes were not linked to any male-associated peaks and therefore were exclusively female-associated peaks (Appendix E-11). These 56 genes included a high number of non-coding loci, including 7 miRNAs, 10 lncRNAs, and 7 pseudogenes, leaving only 32 protein-coding genes. Genes with the greatest number of female-associated peaks included AMOT (6 peaks), PCDH20 (6 peaks), LOC100133050 (4 peaks) and LRTM1 (4 peaks), located on chrX, chr13, chr5 and chr3 respectively. Three peaks linked to the AMOT locus were amongst the top 25 most differentially enriched peaks in females, including 1 peak more differentially-enriched than the peak located at the XIST promoter.

Functional enrichment analysis of gender-associated peaks was carried out using a many-to-many annotation of peaks followed by an over-representation functional enrichment analysis using the following biological ontologies: Gene Ontology (GO; for biological processes, molecular functions and cellular components), Reactome (for pathways), and the Kyoto Encyclopaedia of Genes and Genomes (KEGG; for pathways and reactions) (Ashburner *et al.*, 2000; Kanehisa *et al.*, 2004; Croft *et al.*, 2014).

Table 5.3 Top 25 male-associated differentially enriched autosomal peaks

Peak Location	logFC	Annotation	DistToTSS	Symbol	Gene
chr22: 20378561	4.48	Promoter	642	TMEM191B	transmembrane protein 191B
chr17: 58964679	3.82	Promoter	18	BCAS3	BCAS3, microtubule associated cell migration factor
chr13: 109963464	3.44	Distal Intergenic	-143563	MYO16-AS1	MYO16 antisense RNA 1
chr1: 192753125	3.41	Distal Intergenic	-24794	RGS2	regulator of G protein signaling 2
chr4: 96212025	3.28	Intron (3 of 15)	186269	BMPR1B	bone morphogenetic protein receptor type 1B
chr17: 3056564	3.2	Distal Intergenic	-25469	OR1G1	olfactory receptor family 1 subfamily G member 1
chr2: 1711733	3.18	Intron (1 of 22)	36308	PXDN	peroxidasin
chr7: 118484977	3.13	Distal Intergenic	620015	ANKRD7	ankyrin repeat domain 7
chr7: 119993447	3.12	Intron (1 of 5)	79475	KCND2	potassium voltage-gated channel subfamily D member 2
chr11: 107243907	3.11	Intron (9 of 15)	69618	CWF19L2	CWF19 like 2, cell cycle control (S. pombe)
chr5: 35047103	3.07	Promoter	887	AGXT2	alanine--glyoxylate aminotransferase 2
chr1: 166459613	3.05	Distal Intergenic	-113290	FMO9P	flavin containing monooxygenase 9 pseudogene
chr4: 121663057	2.98	Intron (14 of 15)	180706	PRDM5	PR/SET domain 5
chr12: 116472327	2.97	Intron (4 of 30)	113882	MIR620	microRNA 620
chr18: 54890527	2.93	Distal Intergenic	75984	BOD1L2	biorientation of chromosomes in cell division 1-like 2
chr13: 95767989	2.92	Exon (20 of 31)	-52626	ABCC4	ATP binding cassette subfamily C member 4
chr18: 39772996	2.92	Intron (1 of 4)	6113	LINC00907	long intergenic non-protein coding RNA 907
chr4: 138966379	2.89	Intron (8 of 10)	-43539	SLC7A11-AS1	SLC7A11 antisense RNA 1
chr10: 109812928	2.87	Distal Intergenic	-888212	SORCS1	sortilin related VPS10 domain containing receptor 1
chr15: 97311108	2.87	Distal Intergenic	13526	SPATA8-AS1	SPATA8 antisense RNA 1 (head to head)
chr8: 133520088	2.85	Distal Intergenic	-26834	KCNQ3	potassium voltage-gated channel subfamily Q member 3
chr6: 155594787	2.82	Intron (1 of 1)	9390	CLDN20	claudin 20
chr2: 5390810	2.78	Distal Intergenic	-441739	SOX11	SRY-box 11
chr12: 16536053	2.77	Distal Intergenic	29452	MGST1	microsomal glutathione S-transferase 1
chr21: 42308833	2.77	Distal Intergenic	-89544	DSCAM	DS cell adhesion molecule

Table 5.4 Top 25 female-associated differentially enriched autosomal peaks

Peak Location	logFC	Annotation	DistToTSS	Symbol	Gene
chr15: 54442625	3.21	Intron (3 of 30)	-113468	UNC13C	unc-13 homolog C
chr8: 76307569	2.84	Distal Intergenic	-12364	HNF4G	hepatocyte nuclear factor 4 gamma
chr16: 59937574	2.66	Distal Intergenic	-148229	APOOP5	apolipoprotein O pseudogene 5
chr22: 33776953	2.59	Intron (5 of 10)	55452	MIR4764	microRNA 4764
chr22: 33747139	2.49	Intron (5 of 10)	85266	MIR4764	microRNA 4764
chr2: 59928579	2.48	Distal Intergenic	685751	MIR4432	microRNA 4432
chr11: 39012572	2.31	Distal Intergenic	1301858	LRRC4C	leucine rich repeat containing 4C
chr15: 26129181	2.28	Distal Intergenic	-18076	LINC02346	long intergenic non-protein coding RNA 2346
chr10: 66954086	2.27	Distal Intergenic	368551	ANXA2P3	annexin A2 pseudogene 3
chr4: 64987730	2.26	Distal Intergenic	159297	TECRL	trans-2,3-enoyl-CoA reductase-like
chr1: 95608584	2.21	Intron (1 of 6)	24855	TMEM56-RWDD3	TMEM56-RWDD3 readthrough
chr14: 27311316	2.16	Distal Intergenic	-66282	MIR4307	microRNA 4307
chr15: 26271536	2.16	Intron (2 of 4)	-89174	LINC00929	long intergenic non-protein coding RNA 929
chr13: 93162830	2.11	Intron (7 of 7)	209147	GPC5-AS1	GPC5 antisense RNA 1
chr2: 107954105	2.08	Distal Intergenic	-450292	ST6GAL2	ST6 beta-galactoside alpha-2,6-sialyltransferase 2
chr8: 52498726	2.05	Intron (3 of 22)	-176355	PXDNL	Peroxidasin-like
chr2: 107985004	2.03	Distal Intergenic	457328	RGPD4-AS1	RGPD4 antisense RNA 1 (head to head)
chr16: 59939681	-2	Distal Intergenic	-150336	APOOP5	apolipoprotein O pseudogene 5
chr5: 155829571	-2	Intron (6 of 8)	75554	SGCD	sarcoglycan delta
chr3: 75334867	-1.99	Distal Intergenic	70990	MIR4444-1	microRNA 4444-1
chr11: 91425283	-1.97	Distal Intergenic	-659729	FAT3	FAT atypical cadherin 3
chr11: 25445516	-1.96	Distal Intergenic	-765063	ANO3	anoctamin 3
chr3: 55095882	-1.96	Intron (35 of 37)	-133560	LRTM1	Leucine-rich repeats and transmembrane domains 1
chr5: 99038477	-1.96	Distal Intergenic	685231	LOC100133050	glucuronidase beta pseudogene

Given the restricted gene list obtained from female-associated peaks, very few terms were identified by the functional enrichment analysis (e.g. 16 GO terms in females compared to 1,187 in males). However, those that were enriched pertained to a neuronal phenotype, and the regulation of glycosylation (Table 5.6, Appendix E 8-10). For example, enrichment analysis using GO derived synaptic and post-synaptic membrane components (GO:0097060 & GO:004521), post-synaptic specialization components (GO:0099572) and axon components (GO:0033267) in the top 10 enriched GO terms, with the glycoprotein complex (GO:0016010, GO:0090665) as the most enriched term (Table 5.6). Reactome ontology further demonstrated enrichment of pathways related to diseases of glycosylation (R-HSA-3781865), and the regulation of heparin/heparan sulphate (R-HSA-2022928, R-HSA-3560782) (Appendix E–9).

Conversely, male-associated peaks were annotated to over 75% of all protein-coding genes, therefore resulting in over 1,000 enriched GO terms. These lack a common theme and often contradict each other. For example, cell-cycle related processes within the top 100 most enriched GO terms included regulation of G1/S phase transition (GO:0044843), regulation of G2/M phase transition (GO:0044839), positive and negative control of mitotic cell-cycle phase transition (GO:1901990 & GO:0045930 respectively) and cell cycle arrest GO:0007050). These are also accompanied by processes that occur during different points of the cell cycle such as spindle formation (GO:0005819), cilium assembly (GO:006027), DNA damage repair and replication (GO:0006260 & GO:0006282) and chromosomal organisation and segregation (GO:0033044 & GO:0000819) (Appendix E–4). For all terms discussed above, a $p<1x10\text{-}8$ for pathway enrichment was observed.

The top 10 male GO terms include enrichment of late-stage cell-cycle components and processes such as G2/M phase transition (GO:0044839), centrosome (GO:0005813), spindle (GO:0005819) and cilium regulation (GO:0044782, GO:0060271) (Table 5.5). The Reactome ontology also shows enrichment of cell-cycle regulation in the top 10 most enriched pathways; including cell cycle checkpoints (R-HSA-69620), mitotic phases (R-HSA-453279) and the biogenesis and assembly of organelles and cilium (R-HSA-1852241 & R-HSA-5617833).

Table 5.5 Top 10 enriched pathways from male-associated peaks according to GO; considering ontologies for biological processes (BP), cellular component (CC) and molecular function.

ID (ontology)	Description (no. genes in set)	P value	geneID (top 50)	No. peaks
GO:0005813 (CC)	Centrosome (512)	6.84E-21	FBXL7, DCAF13, GNAI1, KIF20B, PIBF1, MDM1, HNRNPU, MASTL, ORC2, CEP55, SPICE1, NUP107, CEP85L, ERCC6L2, CEP57L1, APC, CDC14B, TTC8, CCNB1, SSX2IP, CEP350, RAPGEF6, MACROD2, CEP120, CCDC15, RNF19A, PKHD1, TBC1D31, STIL, PPP4R3B, GEN1, TTLL5, TTC12, POC1B, CHEK1, TXNDC9, AKNA, ATP6V1D, ZNF322, CEP70, CEP295, OLA1, DTL, AKAP9, WDR35, PCGF5, PROCR, CEP152…	407
GO:0005759 (CC)	mitochondrial matrix (265)	9.70E-15	BTD, HIBCH, MRPS36, DHTKD1, MCCC2, CCNB1, GLRX2, MTRF1L, DARS2, LYRM7, CREB1, ETFDH, PARS2, CBR4, NUDT9, MTERF2, ATXN3, ISCA2, DLD, IARS2, MRPL18, DBT, NUDT2, HYKK, ATP5E, PDK1, ARG2, GCSH, GSR, GLS, HADH, TFAM, PRIMPOL, SIRT5, NARS2, PDHX, FDX1, MALSU1, MCCC1, HSPE1, ALDH6A1, MRPS18C, NADK2, ETFA, ALDH4A1, ABCE1, GARS, OAT, MRPL32, AASS, CDK1…	380
GO:0010256 (BP)	endomembrane system organization (384)	1.58E-14	BCAS3, TPR, OSBPL8, NUP107, VTA1, ANK2, CCNB1, TRIP11, VPS36, SYNE1, DYNC2H1, GOLGA5, TRAPPC11, CREB1, ARV1, CLCN3, PI4K2A, SH3TC2, AKAP9, CAV1, SH3GLB1, RAB33B, CCDC47, NDRG1, RAB3GAP2, VPS4B, TOR1AIP1, VMP1, VAPB, LEMD3, GOLGB1, GORASP2, GBF1, CHMP5, RAB5B, DNAJC13, ALS2, SEC23IP, OPTN, CRB1, PPP2R1A, ANK3, MPP5, PLSCR1, WHAMM, EHD3, IST1…	339
GO:0005819 (CC)	Spindle (274)	4.13E-14	TPR, KIF20B, HNRNPU, SPICE1, CDC14B, HECW2, CCNB1, BRCC3, CEP350, PKHD1, NEDD9, KATNA1, GEM, SPIN1, POC1B, APP, INVS, ATM, ACOT13, NEK7, RIF1, VPS4B, FAM83D, KIF20A, SEPT7, DCTN4, PRC1, HSPA2, CYLD, MAK, ASPM, KIF3A, TNKS, PKP4, CDC27, KIF14, STAG1, KATNB1, KIFC1, CKAP2L, CDK5RAP2, KIF11, CENPF, CDCA8, MMS19, PPP2CA, ECT2, FBXO5, VRK1, SPAST, CLASP2, PKD2…	275
GO:0051052 (BP)	regulation of DNA metabolic process (365)	1.25E-13	USP1, HNRNPU, CACYBP, IL2, MLH1, HELB, BRCC3, ESCO2, IGF1R, CCT2, FBXW7, CHEK1, HMBOX1, WAPL, SIRT1, UBR5, ATM, NEK7, RIF1, KDM1B, DCP2, ATR, HMGB1, TNKS2, JUN, TICRR, KDM4D, SLF2, WRNIP1, UIMC1, OGG1, PRKCQ, THOC1, STN1, FBXO18, BMPR2, PPP2R1A, TNKS, UBE2N, RAD17, RAD52, EYA4, PPP2CA, MSH3, SETMAR, ERCC4, MGMT, UNG, CDK1, GMNN, BLM, NEK2, SLF1…	333
GO:0060271 (BP)	cilium assembly (296)	2.64E-13	ABCC4, SPAG1, PIBF1, CDC14B, TTC8, SSX2IP, TBC1D7, OCRL, INTU, SPAG16, TRIP11, CEP120, TROVE2, DYNC2H1, PKHD1, TTLL5, POC1B, ATP6V1D, CEP70, AKAP9, WDR35, TMEM237, CEP152, TTC21B, PLK4, GORAB, MNS1, WDR19, SDCCAG8, SEPT7, BBS4, IFT74, CYLD, MAK, TNPO1, RAB3IP, NEK1, CEP97, TTC30B, GALNT11, KIF3A, CNTRL, WDR5B, IQUB, PPP2R1A, IFT80…	298
GO:0044839 (BP)	cell cycle G2/M phase transition (131)	3.18E-12	FBXL7, MASTL, CLSPN, TAOK3, CCNB1, FOXN3, CDC25C, CDK7, PSMD14, APP, CEP70, DTL, AKAP9, CCNH, ATM, CEP152, BORA, VPS4B, RPS27A, PLK4, BACH1, PSMA6, TICRR, ABCB1, SDCCAG8, HSPA2, PPM1D, MTA3, BTRC, TAF2, PPP1R12B, CNTRL, OPTN, PPP2R1A, RAD17, PSMA5, PSMC6, KIF14, CEP41, PSMA3, PPME1, ENSA, CDK5RAP2, CENPF, MIIP, PSMB1, CCNA2, SKP2, CDK1, CEP131, PSME4…	225
GO:0044782 (BP)	cilium organization (323)	3.20E-12	ABCC4, SPAG1, PIBF1, CDC14B, TTC8, SSX2IP, TBC1D7, OCRL, INTU, SPAG16, TRIP11, CEP120, TROVE2, DYNC2H1, PKHD1, TTLL5, POC1B, ATP6V1D, CEP70, AKAP9, WDR35, TMEM237, CEP152, TTC21B, PLK4, GORAB, MNS1, WDR19, SDCCAG8, SEPT7, BBS4, BBS12, IFT74, CYLD, MAK, TNPO1, RAB3IP, NEK1, CEP97, TTC30B, GALNT11, KIF3A, CNTRL, WDR5B, IQUB, PPP2R1A, IFT80, BBS10, KIF27…	304
GO:0000226 (BP)	microtubule cytoskeleton organization (569)	5.20E-12	BCAS3, ULK4, SPAG1, TPR, GNAI1, PIBF1, MDM1, HNRNPU, SPICE1, MLH1, SNCA, APC, CDC14B, FER, CCNB1, SSX2IP, RGS2, CEP350, SPAG16, CEP120, RNF19A, MAP7, PKHD1, STIL, KATNA1, GEN1, TTLL5, WASHC5, CHEK1, SPRY1, ATXN3, AKAP9, TBCE, SPC25, NEK7, CEP152, BORA, VPS4B, CENPA, PLK4, GADD45A, KIF20A, SDCCAG8, BBS4, PRC1, SON, CYLD, CHMP5, EFNA5..	391
GO:0010498 (BP)	proteasomal protein catabolic process (456)	6.47E-12	FBXL7, RHBDD1, EDEM3, ARIH1, RNF38, HSP90B1, APC, HECW2, TLK2, CCNB1, TMTC3, UBR3, BIRC2, RNF19A, PSMD14, GNA12, FBXW7, DNAJC10, SIRT1, ATXN3, CAV1, OS9, CCDC47, CDC23, RAD23B, RPS27A, FBXO31, UBE2C, RFFL, PSMA6, BUB3, ZNRF2, KCTD2, MDM2, NEDD4L, TRIM9, USP14, RNF103, WAC, RNF7, UBE2W, KLHL20, TRIB1, BTRC, RNF138, UGGT1…	364

Table 5.6 Top 10 enriched pathways from female-associated peaks according to GO; considering ontologies for biological processes (BP), cellular component (CC) and molecular function

ID (ontology)	Description (no. genes in set)	P value	geneID (top 50)	No. peaks
GO:0016010 (CC)	dystrophin-associated glycoprotein complex (20)	2.68E-05	SGCD, SNTG2, SNTG1	3
GO:0090665 (CC)	glycoprotein complex (20)	2.68E-05	SGCD, SNTG2, SNTG1	3
GO:0033267 (CC)	axon part	0.00015149	UNC13C, EPHA4, PTPRN2, DLG2, COBL	5
GO:0097060 (CC)	synaptic membrane (509)	0.0001712	UNC13C, LRRC4C, CLSTN2, EPHA4, TENM2, DLG2	6
GO:0045211 (CC)	postsynaptic membrane (374)	0.0004143	LRRC4C, CLSTN2, EPHA4, TENM2, DLG2	5
GO:0098794 (CC)	Postsynapse (731)	0.0011645	LRRC4C, CLSTN2, EPHA4, TENM2, MAP2, DLG2	6
GO:0044295 (CC)	axonal growth cone (44)	0.00157075	EPHA4, COBL	2
GO:0014069 (CC)	postsynaptic density (391)	0.00256465	CLSTN2, EPHA4, MAP2, DLG2	4
GO:0099572 (CC)	postsynaptic specialization (430)	0.00261031	CLSTN2, EPHA4, MAP2, DLG2	4
GO:0032279 (CC)	asymmetric synapse (394)	0.00279856	CLSTN2, EPHA4, MAP2, DLG2	4

5.2.4.4 Motif-enrichment analysis of gender-associated peaks

Providing there is a sufficient read depth from sequencing, ATAC-seq can be used to determine common motifs at chromatin accessible regions and infer binding of transcription factors (TFs) with matching DNA-binding motifs (Heinz *et al.*, 2010; Li *et al.*, 2019). The Hypergeometric Optimization of Motif Enrichment suite (HOMER) was used to identify and compare common motif themes of known TFs for gender-associated peaks located at promoter, intronic and distal intergenic sites (Figure 5.14, Appendix E-12). These results show that the most ubiquitous motifs found at distal intergenic and intronic regions were bZIP-motif TFs, whereas Zinc-Finger domain TFs were most common at promoter regions.

Due to the limited number of female-associated peaks located at promoter regions, only a single TF (HLF; bZIP-motif) was identified with a p-value of less than 0.05, and therefore a comparison between males and females was not made. However, males showed a high rate of ZF-motifs (including Sp2/4 and Kruppel-like factors), as well as ETS-TF family motifs (including ETV1/4, ERG and ELK1) for peaks at promoter regions (Appendix E-12).

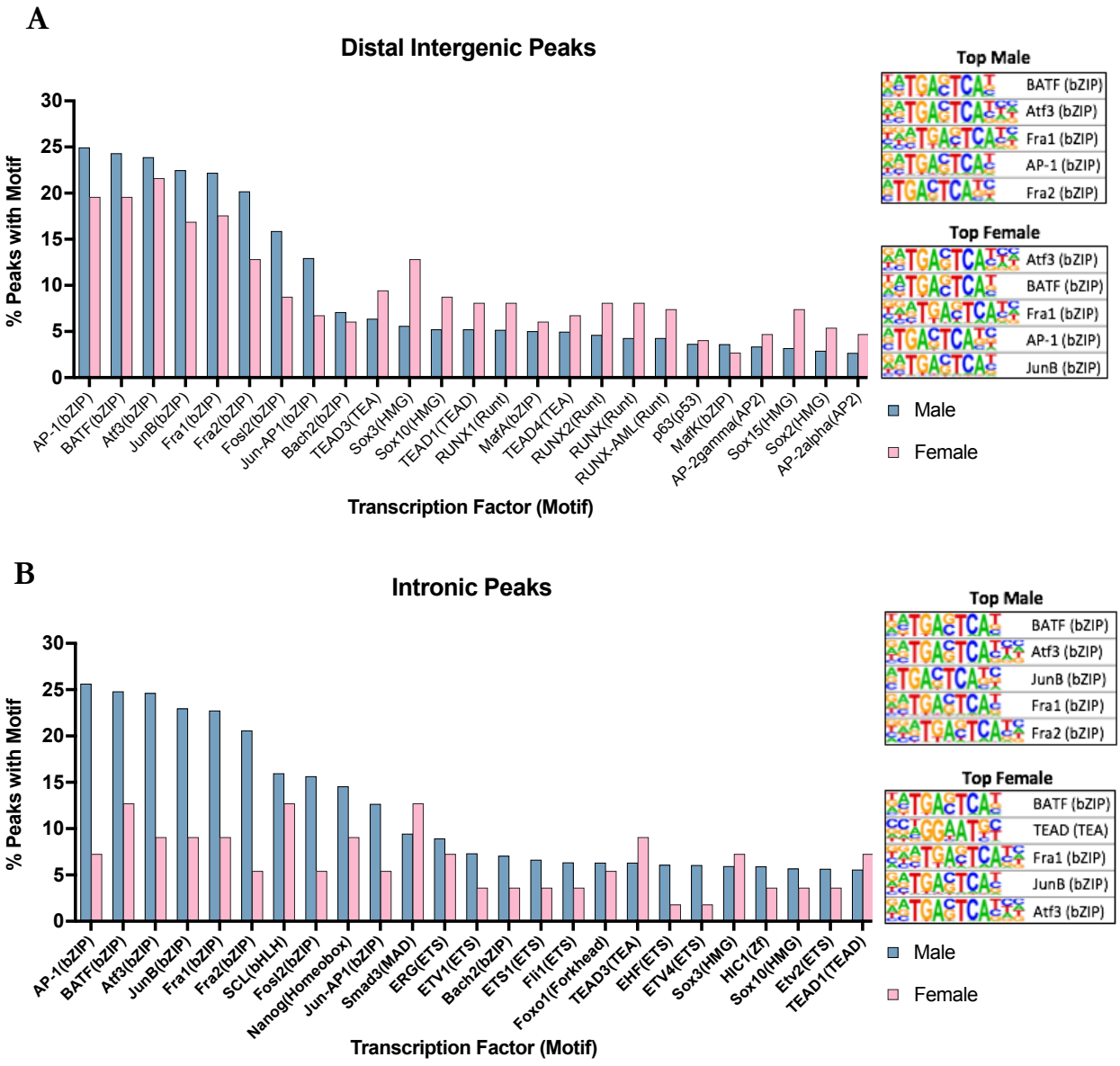


Figure 5.14 Motif-enrichment analysis of gender-associated peaks.

HOMER was used to determine motif-enrichment of known transcription factors for **A)** peaks located at distal intergenic regions and **B)** peaks located at intronic regions. Motif enrichment of male-associated peaks at promoter regions can found in Appendix E-12. For all motif discoveries a p-value of <0.05 was attained. Graphs show top 25 discoveries ranked by males.

A comparison of male and female gender-associated peak motif enrichment at distal intergenic regions showed a marginal increase in males for bZIP-motif TFs including AP-1, BATF and ATF3. Conversely, the majority of transcription factors with motifs other than bZIP-motifs, such as SOX TFs with HMG-motifs, RUNX TFs with Runt-motifs and TEAD TFs with TEA-motifs were marginally enriched at female-associated peaks (Figure 5.14A).

In intronic regions, male-associated peaks showed a large increase for bZIP-motif TF's, such as AP1, BATF, ATF3 and JunB, which were over 2-fold enriched in male-associated peaks compared to female-associated peaks. For other TFs such as SCL (bHLH -motif), SMAD3 (MAD-motif) and ETS-TFs, only marginal differences were observed between genders, although usually in favour of male-associated peaks (Figure 5.14B). It is unknown whether these results pertain to increased regulation of female-associated intronic *cis*-regulatory regions by multiple TFs or a single TF that has a bZIP-motif conserved between TFs.

5.3 Discussion

This chapter and the preceding one set out to establish and carry out an assay for transposase-accessible chromatin following by sequencing (ATAC-seq) in TERT-NHUC, and determine gender-related differences in chromatin-accessibility between these cells. Previous studies have shown that sex differences in various tissues are primarily driven by the inequality in expression of genes located on sex chromosomes (chrX and chrY), and that the majority of gender differences pertaining to chromatin state concern heterochromatin on chrX driven by the polycomb repressive complex (Yen and Kellis, 2015; Arnold, 2017). Therefore, it was hypothesised that the majority of gender-related differences in chromatin accessibility for TERT-NHUC would be located on the sex chromosomes. However, differences were also found within autosomes that may be promoting the gender bias that is seen in bladder cancer.

Although this study is probably the first to carry out ATAC-seq on NHUC, recent publications have used the technique on male bladder cancer samples, as part of larger efforts to characterise the chromatin-accessibility landscape of primary human tumours (Corces *et al.*, 2018) and healthy mouse tissues (Liu *et al.*, 2019a). Other large-scale attempts have also been made across many cancer types, including bladder, aiming to identify *cis*-regulatory regions in the absence of chromatin accessibility data (Chen *et al.*, 2018).

In the absence of chromatin accessibility data, Chen *et al* used RNA-seq data provided by the TCGA to compare RNA transcribed from enhancers (eRNA) in 8,928

tumour samples across 33 cancer types (Chen *et al.*, 2018). The authors found that enhancer activation in cancer was positively associated with aneuploidy but not point mutations, and suggested that chromatin state is a key contributor to genomic alterations in cancer, whereby closed chromatin favours point mutations and open chromatin favours structural rearrangements. Using this RNA-seq method to identify active enhancers, the authors showed that bladder cancer had increased activation of ~12.5% of enhancers, and identified 4,102 active enhancers in bladder of which 87 were potentially prognostic (Chen *et al.*, 2018). However, the approach is limited to only active and "leaky" enhancers and therefore misses poised enhancers and many active but non-leaky enhancers that do not transcribe eRNA (de Santa *et al.*, 2010).

Corces *et al* profiled the chromatin accessibility landscape by ATAC-seq in 410 tumour samples across 23 cancer types (including 9 male MIBC samples)(Corces *et al.*, 2018). The study found that the majority of chromatin-accessible regions across cancers were located at distal intergenic and intronic regions, implicating them as active/poised enhancers (Corces *et al.*, 2018), and it is therefore interesting that a higher proportion of female TERT-NHUC enhancers are also located at distal intergenic regions. The number of chromatin accessible regions varied by cancer type, ranging from 50,000 peaks in cervical squamous cell carcinoma to over 200,000 peaks in breast invasive carcinoma, with bladder urothelial carcinoma having ~100,000 chromatin-accessible peaks. Interestingly, the number of unique peaks identified in male TERT-NHUC was 81,395, which allowing for ~12.5% enhancer activation in cancer (de Santa *et al.*, 2010), would give nearly 100,000 peaks. The study also showed that, although interactions between peaks and their target genes decay rapidly with distance, only 24% of ATAC-seq peaks target the nearest gene and therefore the majority of interactions skip over at least one gene. Furthermore, the authors predict that the expression of most genes is correlated to the activity of ~5 peaks, whereas each peak is suspected to interact with only 1 gene (Corces *et al.*, 2018).

One study in mouse aimed to provide a comprehensive reference of ATAC-seq datasets across 20 healthy tissue types, and included 2 male bladder samples. (Liu *et al.*, 2019). The authors used the same omni-ATAC-seq protocol for all tissues and reported varied QA metrics across samples, therefore demonstrating the necessity of optimising the protocol for each tissue type (Liu *et al.*, 2019). Nevertheless, bladder tissue grouped by itself by PCA when plotted against other tissue types, although only a small proportion of peaks were specific to bladder and were predominately Forkhead TFs. Although gender comparisons in these three aforementioned studies were not possible for bladder (Chen *et al.*, 2018; Corces *et al.*, 2018; C. Liu *et al.*, 2019), they demonstrated distinct clustering of bladder samples based on

chromatin-accessibility, and identified chromatin-accessible peaks unique to bladder and bladder tumours.

Here, the results from the ATAC-seq in TERT-NHUC did not conform to the proposed hypothesis of only subtle differences in chromatin accessibility between genders. Instead, the results showed a widespread decrease in chromatin accessibility in female TERT-NHUC compared to male TERT-NHUC. This was apparent at the level of individual loci and throughout the genome, and it is possibly why far fewer chromatin-accessible peaks were identified in female TERT-NHUC. The majority of studies which compare chromatin accessibility between healthy and diseased states often describe differential peak enrichment but not at a genome-wide level. Indeed, genome-wide changes in chromatin state are typically restricted to developmental biology (Zhu *et al.*, 2013), with recent ATAC-seq results showing how heterochromatic oocytes become euchromatic in the zygote, and continuously open through each progressive stage of early embryonic development (Liu *et al.*, 2019b; Wu *et al.*, 2016).

Nevertheless, recent studies have shown widespread changes in chromatin accessibility in disease, and may support the ATAC-seq results in TERT-NHUC. In mouse, metastatic small cell lung cancer (SCLC) separated into a distinct group from primary SCLC when sorted on ATAC-seq signal, with the first principal component of variation explaining 58% of variance. Furthermore, ~24% of chromatin accessible peaks found in both cohorts were over 2-fold more accessible in metastatic SCLC compared to less than 0.5% in the primary SCLC. This increased chromatin accessibility was predominately driven by overexpression of the Nfib transcription factor, which increased chromatin accessibility at distal intergenic regions and promoted a neuronal gene expression programme and metastasis (Denny *et al.*, 2016). Motif enrichment analysis for gender-associated TERT-NHUC chromatin accessible peaks showed that promoter regions predominantly contained binding sites for TFs with zinc finger motifs, whereas distal intergenic and intronic peaks contained mainly TFs with bZip motifs and were more common in male TERT-NHUC. Perhaps more interesting is that the neuronal gene expression programme seen in metastatic SCLC is mimicked in TERT-NHUC, where female-associated peaks are enriched at genes related to neuronal pathways. However, it is noted that this link is weak given the limited number of enriched gene peaks contributing to these pathways in female TERT-NHUC.

An even greater widespread decrease in chromatin accessibility, that was more akin to the results in TERT-NHUC, was found for patients with age-related macular degeneration (AMD) (Wang *et al.*, 2018). Using ATAC-seq, this study found that the retina and the retinal pigmented epithelium underwent a genome-wide decrease in chromatin accessibility in

AMD, which decreased progressively from normal to early-stage and then to late-stage AMD. Furthermore, iPSC-derived RPE cells exposed to cigarette smoke also showed a widespread decrease in chromatin accessibility in a manner that correlated with AMD, as did the exogenous expression of HDAC11 which was overexpressed in AMD patients (Wang *et al.*, 2018). The association of smoking with decreased chromatin accessibility in AMD may be relevant for the observation in TERT-NHUC, especially considering that smoking is a leading cause of bladder cancer (Sanli *et al.*, 2017). However, it is noted that increased expression of HDAC, which was associated with smoking in AMD, was not shown in female TERT-NHUC by microarray analysis, although increased H3K27ac was shown by western blot.

In the pancreas, α- and β- islet cells share a common developmental origin and have highly similar transcriptomic profiles. However, the two cells have opposing roles in regulating blood glucose levels through glucagon and insulin secretion respectively. ATAC-seq showed that the majority (78%) of chromatin accessible peaks in β-cells were also found in α-cells, owing in part to the much greater number of chromatin accessible peaks identified in α-cells (Ackermann *et al.*, 2016). The number of β-cell-specific peaks reduced further when only endocrine-specific peaks were considered. In this comparison, over 95% of β-cell endocrine peaks were shared with α-cells, identifying 26,952 α-cell-specific endocrine peaks compared to only 1,850 β-cell-specific endocrine peaks. The authors also found that 78% of genes that had increased expression in α-cells had at least one α-cell-specific chromatin accessible peak, compared to only 41% in β-cells. However, only 5% of α-cell- and 12% β-cell-specific ATAC-seq peaks mapped to any differentially expressed genes in these cells (Ackermann *et al.*, 2016). A later study showed that 1,078 chromatin accessible peaks were differentially enriched between type 2 diabetic and non-diabetic islet donors, with the majority (1,044) enriched in the diabetic group (Bysani *et al.*, 2019). Genomic distribution of chromatin accessible peaks was similar between donor groups, which is in contrast to the results of TERT-NHUC where a higher proportion of peaks in female TERT-NHUC were found at distal intergenic regions. Unlike the previous study comparing α- and β- islet cells, there was a stronger relationship between chromatin accessibility and gene expression in the diabetic study (Bysani *et al.*, 2019). Interestingly, the paper also showed TF occupancy of AP-1, BATF, ATF3, FRA1, and FRA1, which were also enriched in male TERT-NHUC (Bysani *et al.*, 2019).

The results from ATAC-seq in TERT-NHUC showed differential chromatin accessibility across the entire genome between each gender. Follow-up MNase digestions assays, growth curve assays and cell cycle analysis did not support these findings and did not

show noticeable differences between male and female TERT-NHUC. However, global levels of the activating histone marks H3K4me3 and H3K27ac were increased in male TERT-NHUC, although no differences were shown for the heterochromatin marker H3K27me3. Together, these results suggest that the differences observed in chromatin accessibility by ATAC-seq are not due to a large-scale heterochromatin event in female TERT-NHUC, but are likely due to an increased activation of TSS and *cis*-regulatory regions in male TERT-NHUC. ChIP-seq for these histone marks as well as H3K4me1 will help to determine if this is indeed the case, and will show whether the global increase of H3K4me3 and H3K27ac observed by western blot is correlated with the ATAC-seq data of chromatin accessibility at individual loci throughout the genome.

An experiment that involves the use of histone deacetylase inhibitors (HDACi) may also provide more support for the ATAC-seq results in TERT-NHUC, with the hypothesis that the use of HDACi in female TERT-NHUC would promote increased chromatin accessibility, comparable to what is observed in untreated male TERT-NHUC. However, studies have shown that cell response to HDACi can be varied, along with the rate of chromatin decompaction (Li and Sun, 2019). For instance, only about a third of cutaneous T-cell lymphoma (CTCL) patients responded to HDACi, although patients that did respond showed altered chromatin accessibility (Qu *et al.*, 2017). Nevertheless, HDACi only re-opened accessible sites that were lost in CTCL, suggesting that HDACi may not resolve widespread decrease in chromatin accessibility seen in female TERT-NHUC (Qu *et al.*, 2017). However, a series of studies, all from the same research group, have demonstrated differential responses in bladder cancer cell lines when treated with family-specific HDACi and pan-HDACi, and shown that, although individual HDACs regulate distinct cellular processes, effective therapeutic treatment in bladder would require targeting multiple/all HDAC families (Rosik *et al.*, 2014; Lehmann *et al.*, 2014; Pinkerneil *et al.*, 2016; Kaletsch *et al.*, 2018; Vasudevan *et al.*, 2019). Nevertheless, these studies show the potential of TERT-NHUC to respond to HDACi, which may increase H3K27ac levels and chromatin accessibility in these cells and help support the findings from ATAC-seq.

It may be expected that a widespread increase in chromatin accessibility would result in a similar widespread increase in gene expression. Indeed, the aforementioned study of chromatin accessibility throughout mouse embryonic development used a novel technique that separates the nuclei from the cytoplasm to carry out low-input chromatin accessibility and transcriptome sequencing (liCAT-seq), and showed that the widespread increase of chromatin accessibility throughout embryonic development is highly correlated with a global increase in gene activity (L. Liu *et al.*, 2019). However, global transcriptional changes were

not reported with widespread decrease of chromatin accessibility in AMD (Wang *et al.*, 2018), and only 5% of α-cell- and 12% of β-cell-specific ATAC-seq peaks mapped to any differentially expressed genes between these cells (Ackermann *et al.*, 2016). Such results therefore demonstrate the lack of predictive ability for chromatin accessibility on gene expression. This may be a result of difficulties in predicting the gene which a given chromatin accessible region may be acting upon, as only 24% of ATAC-seq peaks target the nearest gene and therefore the majority of interactions skip over at least one gene (Corces *et al.*, 2018). The true targets of chromatin accessible regions may only be determined with the use of powerful but expensive chromatin conformation capture technologies such as Hi-C, and even then should be coupled with ChIP-seq to determine how they affect their targets (Schmitt *et al.*, 2016; Schoenfelder and Fraser, 2019). Nevertheless, there is very little correlation between the results from ATAC-seq and those from microarray in TERT-NHUC. Very few gender-associated chromatin-accessible peaks were associated with differentially expressed genes between genders, and there was no global increase of male gene expression in male TERT-NHUC.

The results of this study suggest that male TERT-NHUC have a widespread increase in chromatin accessibility relative to female TERT-NHUC, that correlates with a global increase in H3K4me3 and H3K27ac, but does not correlate with observed changes at the transcriptional level. The mechanisms through which this may be driven are also unknown. Microarray results do not show differential expression of histone modifying genes that may be promoting the altered chromatin state. The mild hypoxic state hypothesis in female TERT-NHUC that was presented at the end of Chapter 3, is also not supported here, given that chromatin accessibility changes do not appear to be due to an increased heterochromatin state in female TERT-NHUC. Studies have shown that different concentrations and combinations of salts and cations commonly found in the cellular environment, such as Na^+, K^+, Mg^{2+}, Ca^{2+}, are able to alter chromatin state by promoting or abrogating chromatin condensation (Korolev *et al.*, 2012; Allahverdi *et al.*, 2015). However, as with chromatin modifier proteins, modulators of cellular ion concentration do not show differential expression between genders, and therefore differential ion concentrations in the nucleus are also unlikely. Another explanation could be impeded accessibility of transposase into the nucleus during transposition. Studies have shown that molecules above 19kDa do not freely cross the nuclear membrane and permeability is modulated, and that this modulation can involve G_q protein-coupled hormone receptors and the influx of Ca^{2+} (O'Brien *et al.*, 2007). This may therefore impede accessibility of transposase (53.3kDa) into the nucleus depending on the permeability status of the nuclear membrane that may be gender-dependent.

However, at the time of transposition nuclei are in a partially lysed state and impeded transposase entry into the nucleus is therefore unlikely. Furthermore, differential expression of proteins that may regulate nuclear permeability, such as G_q protein-coupled hormone receptors, is also not observed by microarray analysis.

Throughout this study the quality of the ATAC-seq was continuously assessed. The metrics measured, and discussed throughout, predominately coincided with those shown in the literature as well as what is recommended by the ENCODE consortium. The ATAC-seq data was analysed using an in-house pipeline, but data was also tested against the recently published PEPATAC and kundajelab pipelines (Corces *et al.*, 2018; Lee *et al.*, 2019). These pipelines confirmed the QA metrics displayed throughout this chapter and included others that assessed PCR bottlenecking and library complexity, both of which also attained the standards recommended by the ENCODE consortium and were not dissimilar between genders. However, two metrics tested in the PEPATAC pipeline that did not meet the standards set by the ENCODE consortium include the fraction of reads in peaks (FRiP) and TSS-enrichment scores, and both metrics are used to test background enrichment of ATAC-seq signal. All ATAC-seq libraries in this study attained scores below what is expected for FRiP and TSS-enrichment, particularly in female TERT-NHUC. The MA plots of genome-wide ATAC-signal showed low signal throughout the majority of the genome for all samples, although this was marginally greater in female TERT-NHUC indicating a higher background signal. This was also apparent when visualising signal using IGV. Furthermore, although fragment size distribution curves did show a distribution typical of ATAC-seq for all samples, the nucleosomal periodicity was less defined in female TERT-NHUC. These results therefore indicate that the majority of the reads generated in female TERT-NHUC are more evenly distributed throughout the genome than at distinct loci, and may be the result of inefficient transposition. This is unusual given that the TapeStation results confirming efficacy of ATAC prior to sequencing did not show gender-related differences in transposition, and qPCR for ATAC enrichment of promoters at housekeeping genes also did not display a gender bias. Nevertheless, inefficient transposition is still a major cause of concern for the results attained in this study.

The biggest limitation to this study is that of sample size. Due to the limitations of available cell lines and their *in vitro* growth characteristics, only two male and two female TERT-NHUC could be used in this study if ATAC-seq were to be complemented with ChIP-seq experiments. ChIP-seq for the histone marks, H3K4me1, H3K4me3 H3K27ac, and H3K27me3 would help support the findings from ATAC-seq, and it is hypothesised that widespread increase of H3K4me1, H3K4me3, and H3K27ac will be observed in male-

TERT-NHUC that will largely overlap with peaks identified by ATAC-seq. A more powerful study that would further support the ATAC-seq in TERT-NHUC would be to carry out ATAC-seq on UHUC, although this should be complemented with more recent techniques such ChIPmentation or CUT&RUN, which require a far fewer number of cells compared to traditional ChIP-seq (Schmidl *et al.*, 2015; Skene and Henikoff, 2017)

To summarise, the ATAC-seq results in TERT-NHUC showed a widespread increase of chromatin accessibility in male TERT-NHUC, which was correlated with a global increase in H3K4me3 and H3K27ac. However, the results were not supported by MNase digestion assays, growth curve assays, or cell cycle analysis, and did not correlate with the transcriptional profiles of these cells as shown by microarray analysis. QA for the ATAC-seq predominantly showed an effective assay with a high quality of sequencing. The exceptions, FrIP and TSS-enrichment scores, were lower than should be expected for ATAC-seq, and also correlated with an MA plot of genome-wide ATAC-signal to show that background enrichment in female TERT-NHUC was marginally greater than in male TERT-NHUC. This may be the result of inefficient transposition, although gender differences in the preparation of libraries were not seen by TapeStation or qPCR analysis. A repeat of the study in UHUC would support these results, as well as future ChIP-seq for histone marks in TERT-NHUC.

Chapter 6
Optimisation of Chromatin Immunoprecipitation (ChIP) in TERT-NHUC

6.1 Introduction

The previous results from ATAC-seq and microarray analyses showed that male TERT-NHUC have a widespread increase in chromatin accessibility relative to female TERT-NHUC that correlates with a global increase in H3K4me3 and H3K27ac, but that this does not correlate with changes at the transcriptional level. Therefore, these results therefore do not support the initial hypothesis of only subtle epigenetic differences on autosomes between genders. However, ChIP-seq on TERT-NHUC could be used to support the findings of widespread epigenomic differences between genders, and demonstrate whether the globally increased activating histone marks shown by western blot are localised at the chromatin-accessible loci that showed increased signal strength by ATAC-seq and which are more numerous in male-TERT NHUC.

Chromatin Immunoprecipitation (ChIP) followed by NGS (ChIP-seq) was first used to map 20 different histone modifications, as well as RNA Polymerase II and CTCF in $CD4^+$ T-cells, and immediately superseded ChIP-on-chip technologies due to increased resolution and the ability to interrogate entire genomes for protein binding (Barski *et al.*, 2007). ChIP-seq has since been applied across many cell/tissue types under different conditions and between species, and large collaborations such as ENCODE, modENCODE and the Roadmap Epigenomics Mapping Consortium have been established with the aim of providing reference epigenomes for humans and other species (Marinov and Kundaje, 2018). Despite these efforts, few studies have included the use of ChIP-seq in bladder. Establishment of a reproducible ChIP protocol in TERT-NHUC is paramount for future studies regarding epigenetic perturbations in bladder cancer.

A standard ChIP protocol begins with the fixation of cells by formaldehyde which reversibly cross-links DNA and associated proteins. Fixed cells are then lysed, and chromatin is either enzymatically digested (often by MNase) or physically sheared using sonication. For histone ChIP, DNA fragments between 100-400bp are desirable as this increases the resolution of the assay. The fragmented DNA is then immunoprecipitated (IP) using antibodies that target the protein of interest. Prior to IP a fraction of the sonicated DNA is kept aside and used as an input control to which IP samples are later normalised.

Immunocomplexes are then pulled-down using secondary antibodies bound to magnetic beads, and then reverse-crosslinked and purified. Purified ChIP samples therefore consist of fragmented DNA that was previously bound by the protein of interest, and can be visualised at the level of individual loci using qPCR, or at a genome-wide level using NGS following appropriate library preparation.

As mentioned in previous chapters, bladder cancer shows high rates of mutations in chromatin modifying proteins. Mutations are frequently found in genes that encode components of the COMPASS-like, Cohesin, and SWI/SNF complexes, and therefore indicate common perturbations of enhancer regions in bladder cancer (Gui *et al.*, 2011b; Weinstein *et al.*, 2014; Hurst *et al.*, 2017). The use of ChIP-seq for histone markers in TERT-NHUC will provide a "normal" standard to which future studies regarding chromatin perturbations in bladder cancer can be compared.

The identification of enhancer regions has a long history, where the enrichment of markers such as the acetyltransferase p300, RNA polymerase II, H2A.Z, H3K4me1, H3K4me2, H3K27ac, chromatin accessibility and more, has been used to identify enhancer regions (Barski *et al.*, 2007; Ernst *et al.*, 2011; Zentner and Scacheri, 2012; ENCODE Consortium, 2012). The current consensus is that active enhancer regions are devoid of nucleosomes and allow the binding of TFs to accessible chromatin, with nearby nucleosomes also undergoing post-translational modification of histone tails, including H3K4me1 and H3K27ac. The cell-type-specific identification of enhancers can therefore be determined by assaying for regions of chromatin accessibility (Thurman *et al.*, 2012), or the histone modifications H3K27ac (Creyghton *et al.*, 2010; Nord *et al.*, 2013) or H3K4me1 (Heintzman *et al.*, 2007). However, the most reliable results for enhancer identification are obtained when combining such assays (Fu *et al.*, 2018). Furthermore, different combinations of histone marks at *cis*-regulatory regions also indicate enhancer activation status, where active enhancers have enrichment of H3K4me1, H3K27ac, and high chromatin accessibility, and primed/silent enhancers have enrichment of H3K4me1 with/without H3K27me3, and low chromatin accessibility (Rivera and Ren, 2013; Klemm *et al.*, 2019). The histone marks of interest for this study were therefore H3K4me1, H3K27ac, and H3K27me3 for enhancer identification, and H3K4me3 for the identification of active/primed promoters.

This chapter describes the establishment of a reliable ChIP protocol in TERT-NHUC for H3K4me1, H3K4me3, H3K27ac and H3K27me3. Once ChIP is established, future ChIP-seq analysis can be carried out in TERT-NHUC and will complement the ATAC-seq and microarray results from the previous chapters.

6.2 Results

To establish and optimise a ChIP protocol in bladder, the bladder cancer cell line RT112 was used due to ease of culture and lower cost of materials. An optimised ChIP protocol would then be used on TERT-NHUC with the expectation that the protocols would be largely similar between TERT-NHUC and RT112. The ChIP protocol optimised in this chapter is based on two previously established protocols in leukaemia and liver cell lines, and was chosen due to the available experience of individuals at this Institute (Follows *et al.*, 2003; Wederell *et al.*, 2008). Quality of optimisation determined by gel-electrophoresis to assess sonication, where a smeared band 100-500bp indicative of partial but intact fragmentation of DNA is seen, and by qPCR following ChIP to assess enrichment of known regions occupied by known histone modifications.

6.2.1 Identifying control loci for the enrichment of histone marks

Following a ChIP-assay, qPCR is commonly used as a QA metric to validate successful enrichment/depletion of the ChIP protein of interest at control loci where occupation is known. However, as no publicly available ChIP-seq data is available for bladder tissue or TERT-NHUC for the histone marks H3K4me1, H3K4me3, H3K27ac or H3K27me3, such control loci are unknown. Therefore, control loci in bladder were inferred by using the University of California Santa Cruz (UCSC) genome browser to identify common enrichment/depletion of the histone marks at different loci in six distinct and unrelated cell lines (NHO, HeLa, HepG2, K562, NHEK, NHLF) with publicly available ChIP-seq data (Kent *et al.*, 2002).

To identify loci with appropriate histone modification patterns that could act as reasonable controls, three types of loci were considered: the promoter regions of common housekeeping genes, tissue-specific genes that are silent in the majority of tissues (in this case neuronal-specific genes were considered), and enhancer regions. Three loci were identified to act as potential control regions for the enrichment of histone marks by ChIP in TERT-NHUC (Figure 6.1). The first included the promoter region of the *SDHA* gene, which is positive for enrichment of H3K4me3 and H3K27ac and negative for enrichment of H3K4me1 and H3K27me3. The second was an intronic region of *MYT-1* (myelin transcription factor 1), which is positive for H3K27me3 and negative for enrichment of H3K4me1, H3K4me3, and H3K27ac. The final region included an enhancer of *ACTBL2* (β-actin-like protein 2), which is positive for enrichment of H3K4me1 and negative for enrichment of H3K27me3. As enhancer activation is tissue-dependent and can vary between individuals, enhancers cannot act as positive or negative controls for H3K27ac.

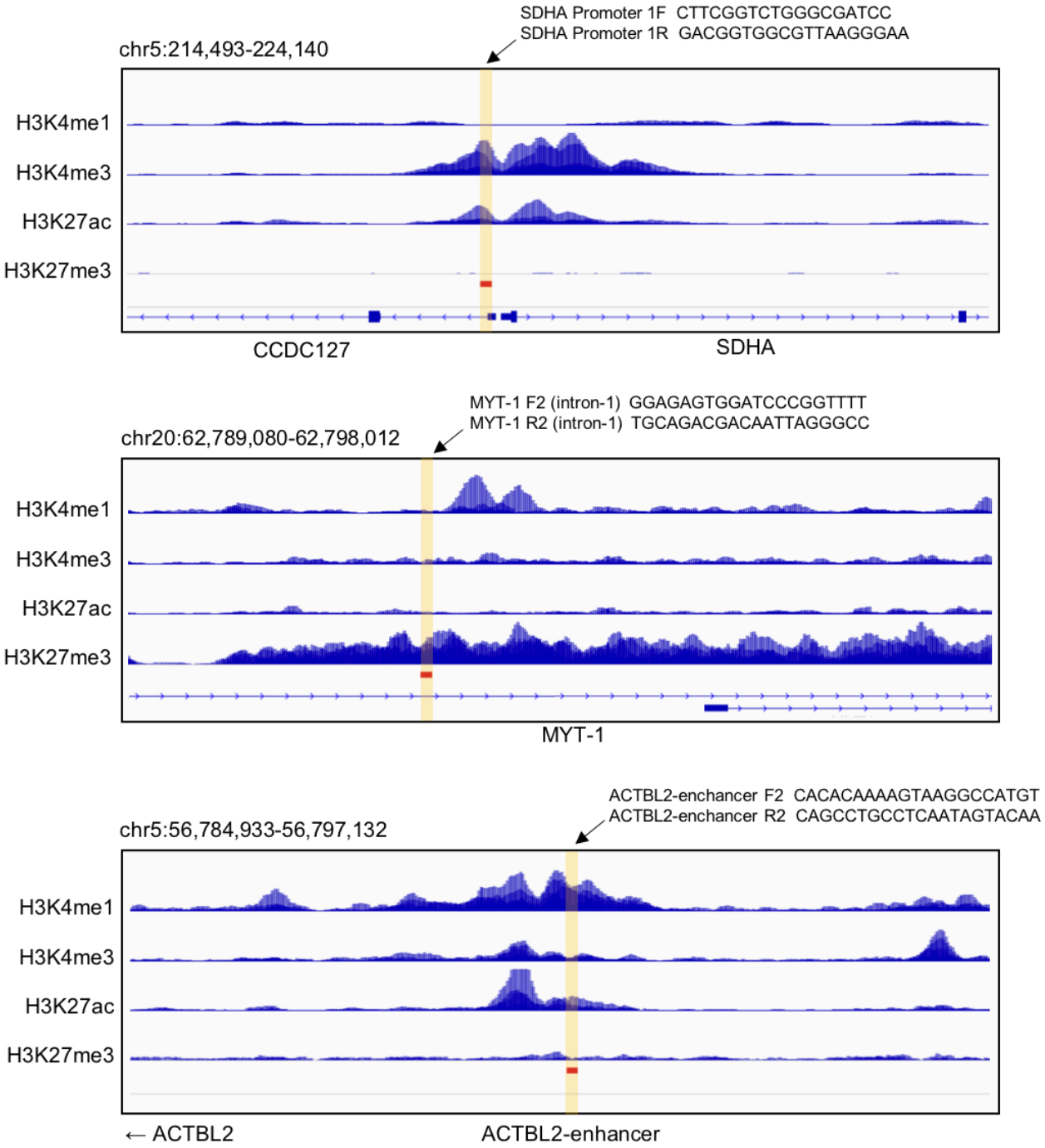


Figure 6.1 Control loci for the enrichment of histone modifications by ChIP

Loci for the enrichment of histone modifications determined by ChIP-seq in 6 cell lines (Osteoblast, HeLa, HepG2, K562, NHEK, and NHLF) were visualised using the UCSC genome browser (hg19). Histone tracks for H3K4me1, H3K4me3, H3K27ac, and H3K27me3 are visualised at the promoter of SDHA (top), an intron of MYT-1 (middle), and an enhancer upstream of ACTBL2 (bottom). Regions targeted for amplification by qPCR following ChIP are highlighted in yellow and annotated with primer sequences.

6.2.2 Optimising sonication and lysis buffer for chromatin preparation in RT112

The ChIP protocol begins with suspending and washing $5x10^6$ cells in PBS then cross-linking with formaldehyde. Following cross-linking, cells are lysed and sonicated using a Bioruptor® ultrasonicator. The sonication step is used to break genomic DNA (gDNA) into 100-400bp fragments, which includes DNA spanning a maximum of 2 nucleosomes and therefore increases resolution of ChIP compared to longer fragment sizes. As proteins can be damaged during the sonication process, the number of sonication cycles is kept to a minimum to prevent dissociation of protein from the DNA and prevent damage to epitopes recognised by IP antibodies. A sonication cycle in this context includes 30 seconds of sonication followed by 30 seconds of no sonication. Nuclear lysis also requires SDS, which precipitates in solution at low temperatures and is also able to disrupt DNA-protein binding. The concentration of SDS buffers for ChIP typically varies by protocol, such as 0.5% and 2% SDS for the lysis buffers in the reference protocols (Follows *et al.*, 2003; Wederell *et al.*, 2008). Therefore, sonication and lysis for ChIP requires a lysis buffer with a moderate but not high concentration of SDS and as few sonication cycles as possible to produce DNA fragments 100-400bp in length. An optimisation experiment was therefore carried out on RT112 using an increasing number of sonication cycles (6, 8, 12, and 16 cycles) and lysis buffers with increasing concentrations of SDS (0.5%, 1%, 1.5%, and 2%) (Figure 6.2). An optimised sonication would be seen as a smear of DNA 100-500bp by gel-electrophoresis.

With regard to SDS concentration there were very few differences in the pattern of fragmentation between each of the lysis conditions. With regard to the number of sonication cycles, 6 and 8 cycles in all buffers showed large smears spanning from ~300bp to undigested DNA. For 10 and 12 cycles, smears spanning 100bp-500bp were seen. As 10 cycles showed roughly the same fragmentation as 12 cycles, and spanned the desired fragment lengths for ChIP, 10 cycles was deemed the optimal number for RT112. In 0.5% SDS, 12 cycles showed over-fragmentation of chromatin, unlike 1% SDS, indicating possible protection of DNA by SDS in the higher concentration buffer. A similar pattern was also seen for 1.5% and 2.0% SDS buffers at 6 and 8 cycles where the minimum fragment length was ~500bp whereas lower concentrations had slightly smaller fragments.

The optimal cell lysis and sonication conditions for RT112 from this experiment included 1.0% SDS lysis buffer and 10 cycles of sonication, and produced fragment sizes spanning 100-500bp. Further experiments therefore used these conditions in the chromatin preparation for ChIP.

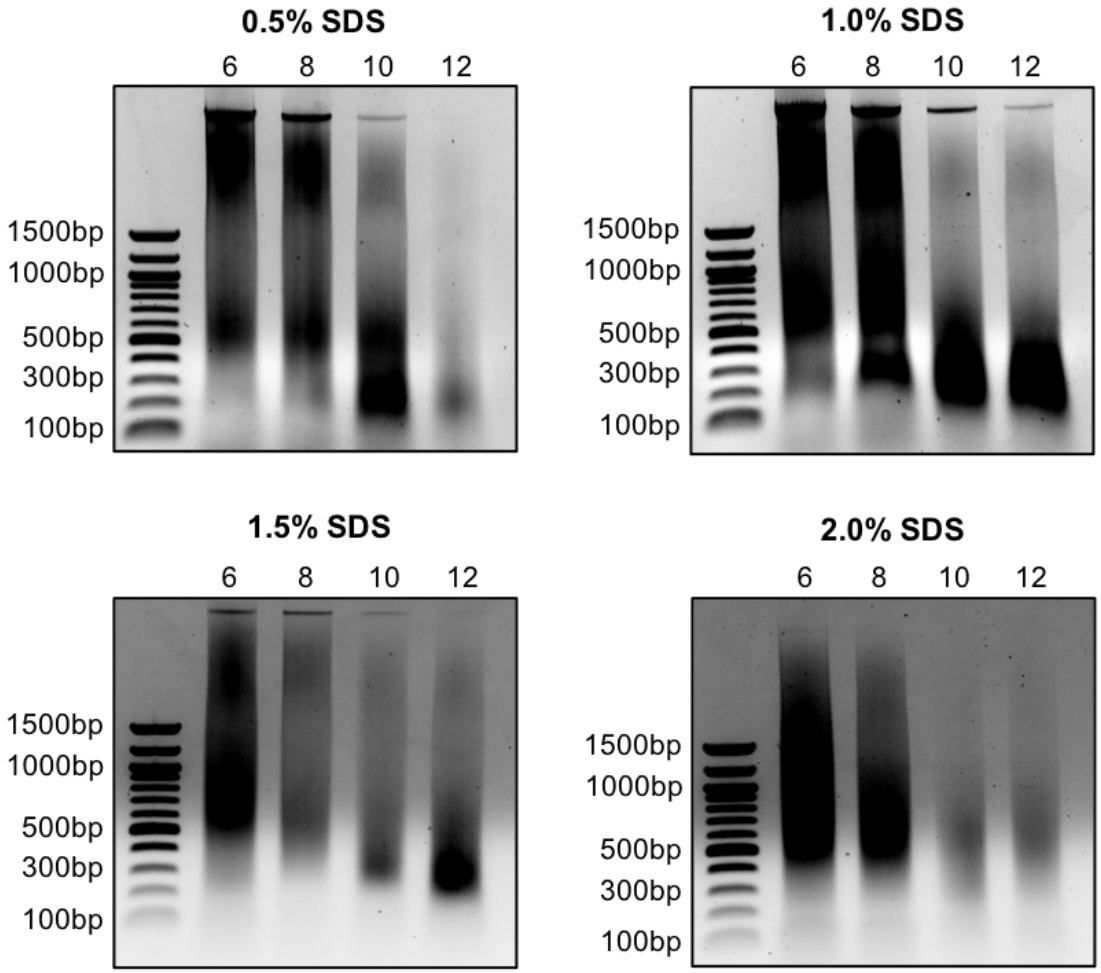


Figure 6.2 Optimisation of SDS concentration and number of sonication cycles in RT112

$5x10^6$ RT112 cells were resuspended in PBS and fixed using a formaldehyde cross-linking buffer. Fixed cells were then lysed using lysis buffers with 0.5% (top left), 1.0% (top right), 1.5%, (bottom left), or 2.0% SDS and sonicated using a Bioruptor® Ultrasonicator for 6, 8, 10 , or 12 cycles of 30s on/off. Samples were then reverse-crosslinked, purified into 0.1X TE buffer following phenol:chloroform extraction, then visualised using 1.0% agarose gel electrophoresis.

6.2.3 Testing histone-targeting antibodies for ChIP

In order to select appropriate antibodies be used in a ChIP protocol for histone modifications in TERT-NHUC, the interactive database for the assessment of histone antibody specificity was used (Rothbart *et al.*, 2015). The database includes over 100 of the most commonly used and cited histone-targeting antibodies, with specificity determined using an array platform of over 250 purified biotinylated histone peptides with different combinations of post-translational modifications (Rothbart *et al.*, 2015). Using the database, 8 histone-targeting antibodies were selected for their specificity to the histone modifications of interest in this study (see Methods and Appendix F-1) and tested for ChIP in RT112.

Of the 8 histone-targeting antibodies, 5 showed the pattern of enrichment expected for each control locus, and were distinct from the background enrichment shown by the negative IgG control (Figure 6.3A-D). The first antibody was an anti-H3K4me1 polyclonal antibody (pAb) from Abcam which showed enrichment at the ACTBL2-enhancer and no enrichment at MYT-1 or SDHA (Figure 6.3A). The second was an anti-H3K4me3 monoclonal antibody (mAb) from Cell Signaling Technologies (CST) which showed enrichment at SDHA, and minimal or no enrichment at the ACTBL2 enhancer or MYT-1 (Figure 6.3B). The third was an anti-H3K27me3 mAb, also from CST, which only showed enrichment at MYT-1 (Figure 6.3C). The last was an anti-H3K27ac mAb from Active Motif which showed the expected enrichment at SDHA, no enrichment at MYT-1, and marginal enrichment at the ACTBL2 enhancer, which indicated possible activation of this enhancer in RT112 (Figure 6.3D). An additional anti-H3K27ac pAb from Abcam also showed a similar pattern of enrichment across the three loci (Figure 6.3E). However due to the increased benefits in reproducibility that come with using monoclonal antibodies, the anti-H3K27ac from Active Motif was considered more appropriate for ChIP in this study. The three remaining antibodies did not show enrichment at the expected loci or greater enrichment than that of the IgG negative control, and included an anti-H3K4me3 pAb from Millipore (Figure 6.3F) and two anti-H3K27me3 mAbs from Diagenode and Millipore (Figure 6.3G & H).

This experiment showed the expected pattern of enrichment for H3K4me1, H3K4me3, H3K27ac, and H3K27me3 at the three loci identified in Figure 6.1 for four anti-histone antibodies. However, for H3K27ac and H3K27me3 only a low level of enrichment over input was shown. Therefore, an additional ChIP was carried out that increased the amount of anti-H3K27ac and anti-H3Kme3 antibodies from 4µl to 10µl per ChIP (Figure 6.4). By increasing the amount of antibody, enrichment was also increased and the histone enrichment pattern for the control loci remained the same.

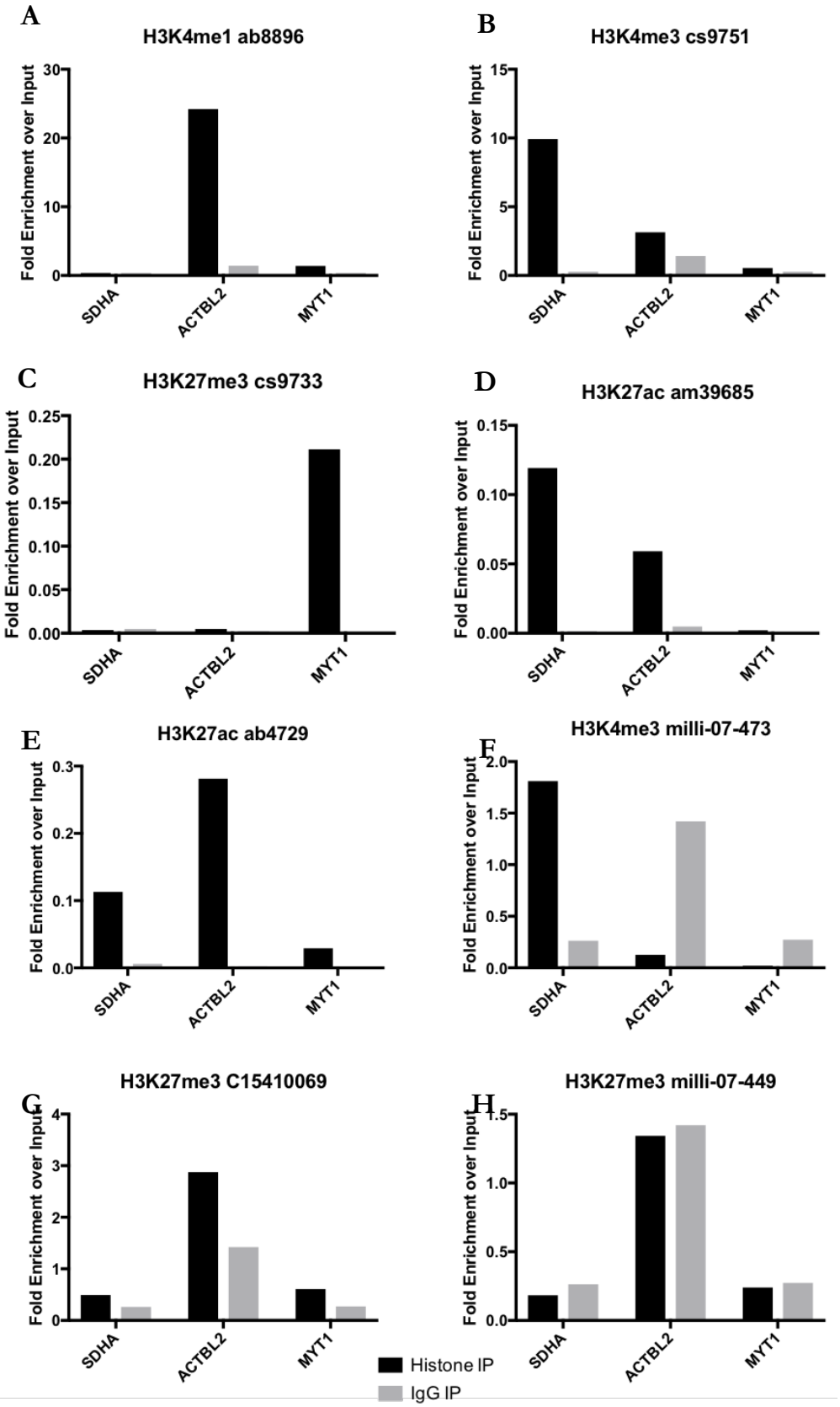


Figure 6.3 Testing anti-histone antibodies for ChIP in RT112

ChIP was carried out in RT112 cells using **A)** anti-H3K4me1 Abcam 8895, **B)** anti-H3K4me3 CST 9751, **C)** anti H3K27me3 CST 9733, **D)** anti-H3K27ac Active Motif 39685, **E)** anti-H3K27ac Abcam 4729, **F)** anti-H3K4me3 Millipore 07-473, **G**) anti-H3K27me3 Diagenode C15410069, and **G)** anti-H3K27me3 Millipore 07-449. Following ChIP, qPCR was carried out to show histone enrichment at *SDHA, ACTBL2,* and *MYT-1.* Bar plots show fold-change enrichment of histone marks (black bars) or IgG (grey bars) over input.

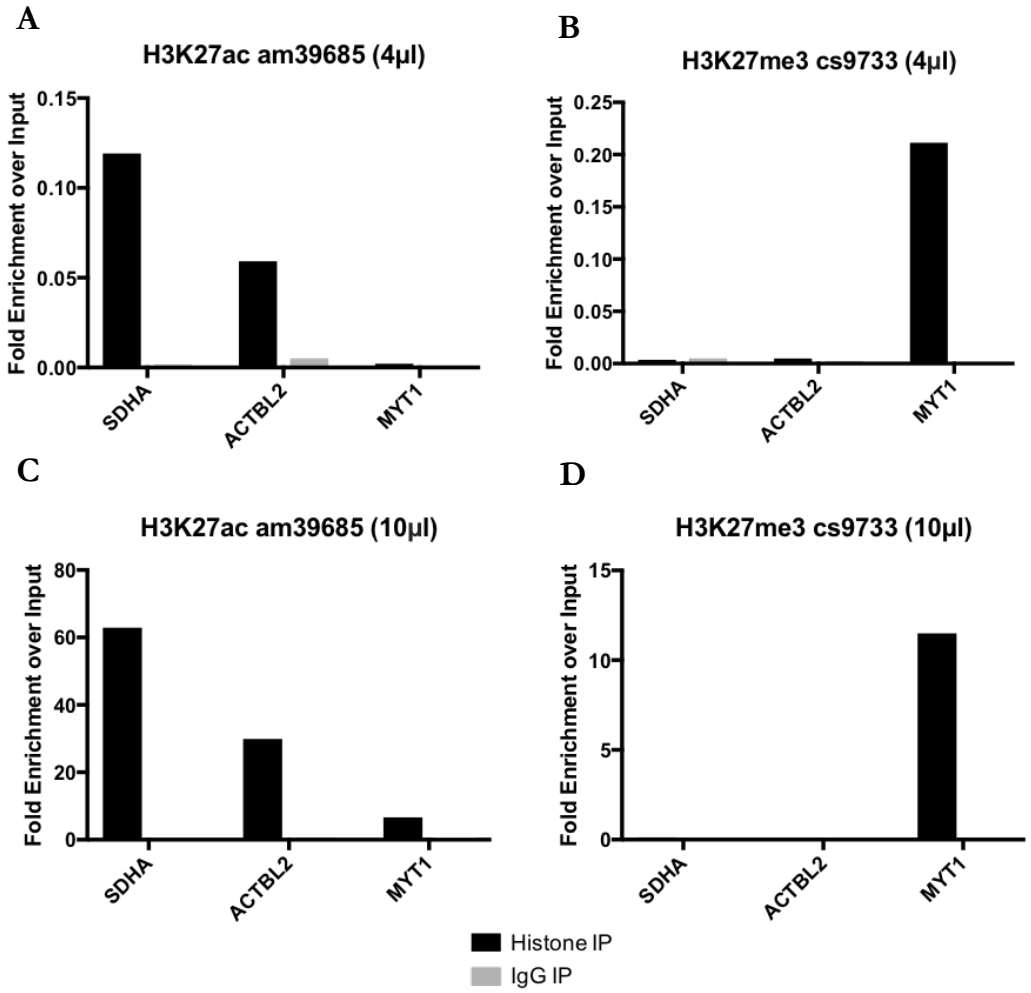


Figure 6.4 Optimising volume of anti-H3K27 antibodies for ChIP in RT112

A) Same as Figure 6.3D, ChIP on RT112 using 4µl of anti-H3K27ac antibody. **B)** Same as Figure 6.3C, ChIP on RT112 using 4µl of anti-H3K27me3 antibody. **C-D)** ChIP was repeated on RT112 using 10µl of **C)** anti-H3K27ac and **D)** anti-H3K27ac antibodies. Following ChIP, qPCR was carried out to show histone enrichment at *SDHA, ACTBL2,* and *MYT-1.* Bar plots show fold-change enrichment of histone marks (black bars) or IgG (grey bars) over input.

6.2.4 Sonication requires 1.5ml Diagenode Bioruptor® tubes

Having established an optimised ChIP protocol in RT112, ChIP was carried out on B-TERT with the intention of using samples for NGS. ChIP was carried out in B-TERT using a 1% SDS lysis buffer, 10 cycles of sonication, 4µl of anti-H3K4me1 and H3K4me3 antibodies, and 10µl of anti-H3K27ac and anti-H3K27me3 antibodies. Enrichment of histone marks at control loci was determined by qPCR (Figure 6.5A-D).

The results from qPCR showed enrichment of all histone marks at ACTBL2, low enrichment of all histone marks at SDHA and MYT-1, and negligible enrichment of IgG across all loci. Although the pattern of enrichment attained for H3K4me1 was as expected (Figure 6.5A), overall the enrichment of histone marks across all loci indicated a failure of ChIP in B-TERT. Following ChIP and qPCR, the input control of sonicated DNA that did not undergo ChIP was visualised using gel electrophoresis (Figure 6.5E). The electrophoresis results showed a concise band of large and unfragmented DNA, typical of gDNA that has not undergone sonication. This was unexpected given the spread of fragment sizes observed when sonicating RT112 over the same number of cycles (Figure 6.2), but explained the unusual enrichment pattern across the control loci in Figure 6.5A-D. Optimisation of sonication was therefore carried out in B-TERT, where chromatin was sonicated for between 0 and 25 cycles with 5-cycle increments (Figure 6.5F). Surprisingly, all samples showed minimal fragmentation of chromatin for each sonication. 25 cycles of sonication was comparable to 5 cycles of sonication, and all samples showed large concise bands typical of unfragmented gDNA that was also seen in the 0 cycle negative control. These results therefore suggested that B-TERT required an even greater number of sonication cycles of sonication to produced DNA fragments between 100-500bp, or that the sonication was not working as efficiently as in RT112.

To test if the Bioruptor® itself was still working as efficiently as during optimisation, the sonication experiment in B-TERT (Figure 6.5F) was repeated with RT112. As the protocol was established in RT112, it was expected that chromatin would be fully fragmented by 25 cycles of sonication (Figure 6.6A). However, in contrast to expectation, the results of sonication over 0 to 25 cycles were identical between RT112 and B-TERT, with minimal fragmentation of DNA being seen even at 25 cycles. An additional experiment which increased the number of cycles in RT112 from 30 to 70 with 10-cycle increments showed that fragmentation was not seen until about 50 cycles of sonication, although 70 cycles of sonication only showed a smear of DNA fragment sizes in excess of 600bp (Figure 6.6B). These results therefore showed that sonication was no longer working as efficiently as during optimisation, and was not appropriate for ChIP.

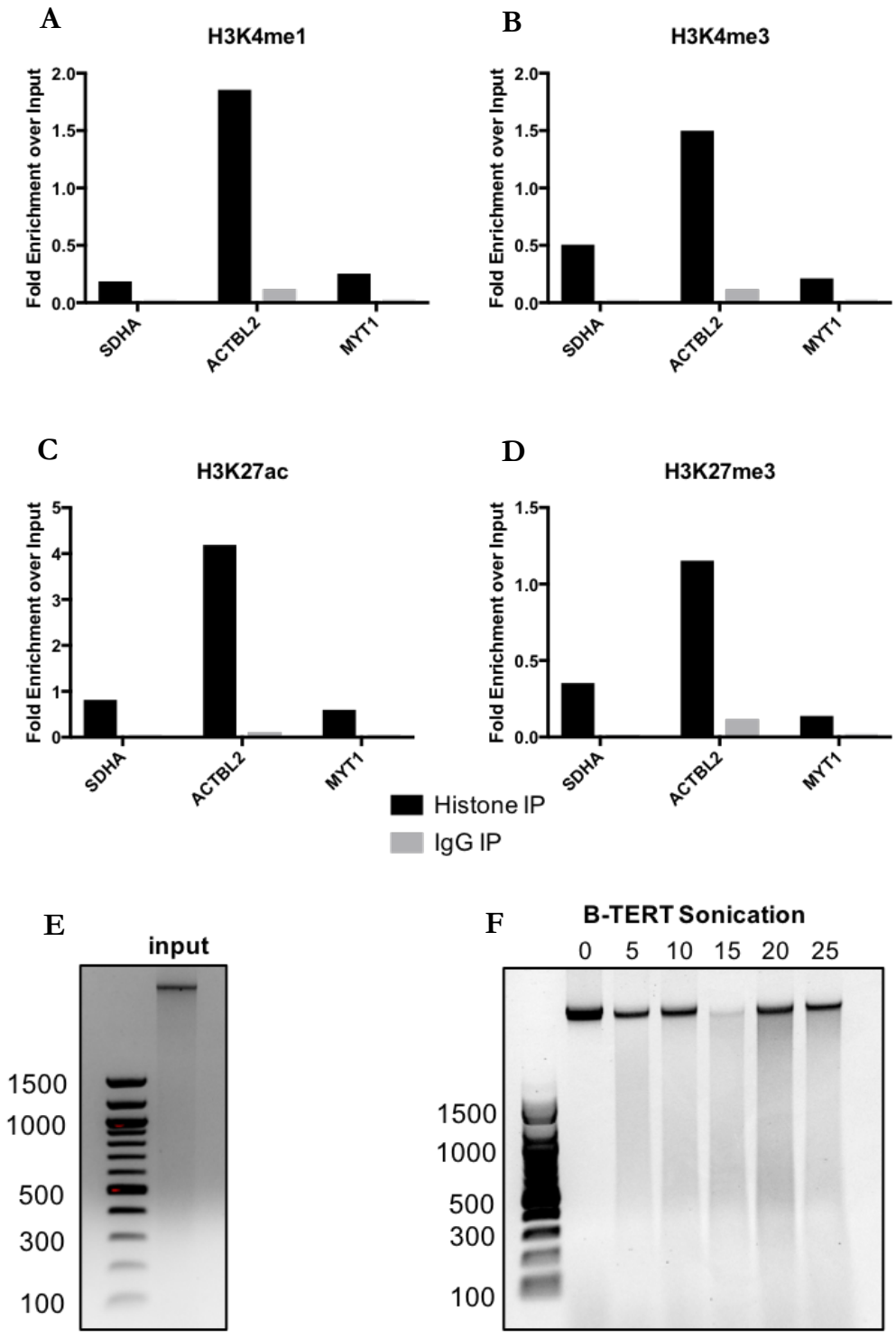


Figure 6.5 Failed ChIP for histone marks in B-TERT

Using the protocol established in RT112, ChIP was carried out in B-TERT cells using **A)** anti-H3K4me1, **B)** anti-H3K4me3, **C)** anti-H3K27ac, and **D)** anti-H3K27me3 antibodies. Following ChIP, qPCR was carried out to show histone enrichment at *SDHA, ACTBL2,* and *MYT-1*. Bar plots show fold-change enrichment of histone marks (black bars) or IgG (grey bars) over input. **E)** The input DNA from the ChIP in A-D was analysed by 1.0% agarose gel electrophoresis. **F)** $5x10^6$ B-TERT cells were harvested and cross-linked in 1% formaldehyde, then sonicated for 0, 5, 10, 15, 20, and 25, cycles. Samples were then reverse-crosslinked then purified into 0.1X TE buffer using phenol:chloroform purification, and visualised using 1.0% agarose gel electrophoresis.

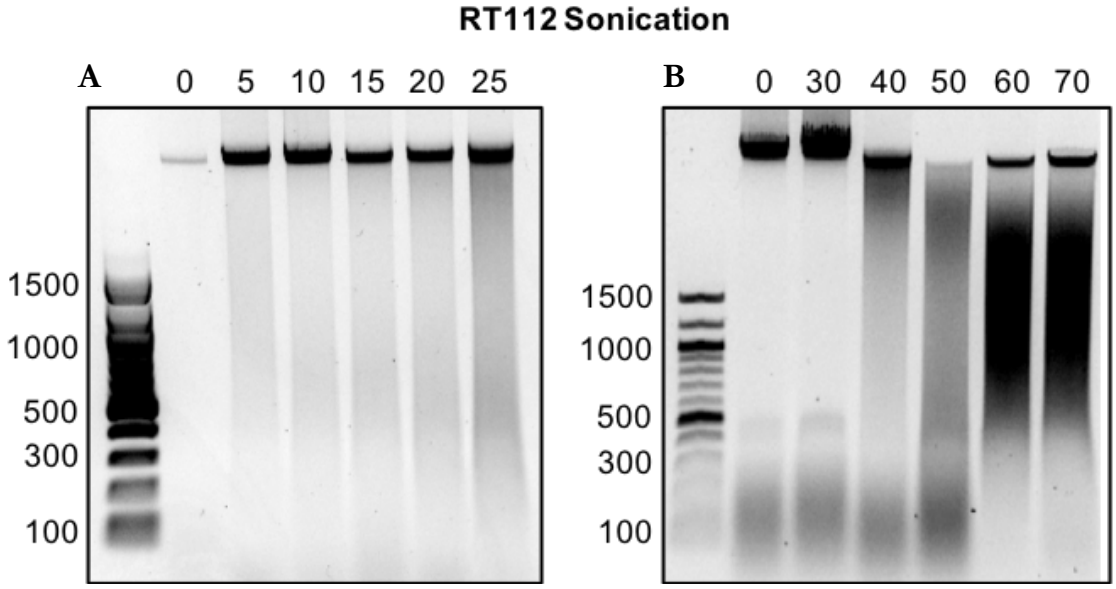


Figure 6.6 Sonication in RT112

$5x10^6$ RT112 cells were harvested and cross-linked in 1% formaldehyde, then sonicated for either **A)** 0, 5, 10, 15, 20, or 25 cycles, or **B)** 0, 30, 40, 50, 60, of 70 cycles. Samples were then reverse-crosslinked and purified into 0.1X TE buffer using phenol:chloroform purification and visualised using 1.0% agarose gel electrophoresis.

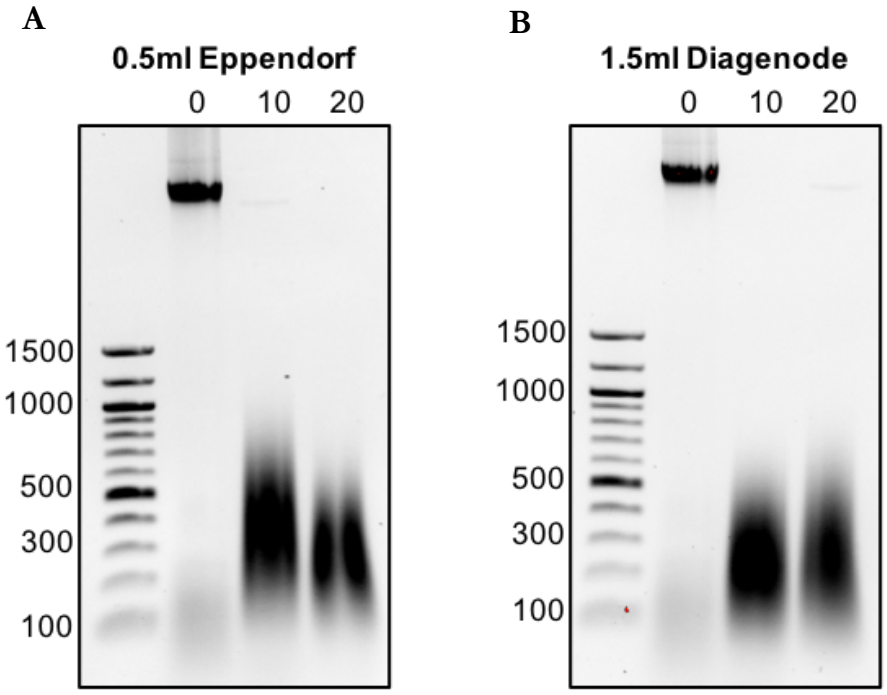


Figure 6.7 Using 0.5ml Eppendorf and 1.5ml Diagenode tube for sonication

$5x10^6$ RT112 cells were harvested and cross-linked in 1% formaldehyde, then sonicated for 0, 10, or 20 cycles in either **A)** 0.5ml Eppendorf tubes, or **B)** 1.5ml Diagenode Bioruptor® tubes. Samples were then reverse-crosslinked and purified into 0.1X TE buffer using phenol:chloroform purification and visualised using 1.0% agarose gel electrophoresis.

After extensive troubleshooting regarding why the Bioruptor® was not working as efficiently in the previous experiments as it had during optimisation, a proposed solution considered the vessel used for sonication. In the previous experiments with B-TERT and RT112, sonication was carried out using 1.5ml Eppendorf tubes, but it was not documented what vessel was used for sonication during optimisation. Therefore, sonication was carried out in RT112 over 10 and 20 cycles, using 0.5ml Eppendorf or 1.5ml Diagenode Bioruptor® tubes (Figure 6.7). For both vessels, RT112 chromatin was fragmented to produce sizes between 100-400bp at both 10 and 20 cycles, similar to the optimisation experiment. Therefore, the likely cause for the differences in sonication efficiency between the optimisation in RT112 and the ChIP in B-TERT was the vessel used for sonication. Further chromatin preparations for ChIP in bladder cells therefore used 1.5ml Diagenode Bioruptor tubes during sonication.

6.2.5 ChIP in TERT-NHUC

The established ChIP protocol in RT112 was carried out in duplicate for each TERT-NHUC line (B-, C-, H-, and K-TERT) with the intention of using the products for NGS. Prior to ChIP, each TERT-NHUC was subjected to a series of sonications to determine the optimal number of sonication cycles to be used (Figure 6.8). This would ensure that DNA fragment sizes were consistent between cell lines and eliminate variation in ChIP resolution. Although sonication was optimised for each TERT-NHUC, the number of cycles that could be tested was restricted by the number of cells available, which was fewer for C-TERT and H-TERT. For B-TERT, sonication was carried out from 0 to 10 cycles with 1-cycle increments, and showed optimal fragmentation at 8 cycles (Figure 6.8A). For C-TERT, sonication was carried out from 7 to 13 cycles with 1-cycle increments, and showed optimal fragmentation at 10 cycles (Figure 6.8B). For H-TERT, sonication was carried out from 0 to 6 cycles with 1-cycle increments, and showed optimal fragmentation at 6 cycles (Figure 6.8C). For K-TERT, sonication was carried out from 0 to 10 cycles with 1-cycle increments, and showed optimal fragmentation at 9 cycles (Figure 6.8D).

ChIP was then carried out in biological duplicate for each TERT-NHUC, using the respective number of sonication cycles outlined above, and the protocol established in RT112 (Figure 6.9). The pattern of histone enrichment across all loci for all TERT-NHUC was as expected; enrichment of H3K4me1 was seen at ACTBL2 (Figure 6.9A), H3K4me3 at SDHA (Figure 6.9B), H3K27me3 at MYT-1 (Figure 6.9C), and H3K27ac at SDHA and ACTBL2 (Figure 6.9D). Enrichment of H3K27ac was notably higher at ACTBL2 for C-TERT. However, this may just indicate activation of the enhancer in this cell line.

Enrichment of the IgG negative control was negligible across all loci for all TERT-NHUC (Figure 6.9E). Given these results, ChIP samples were considered to be suitable for NGS.

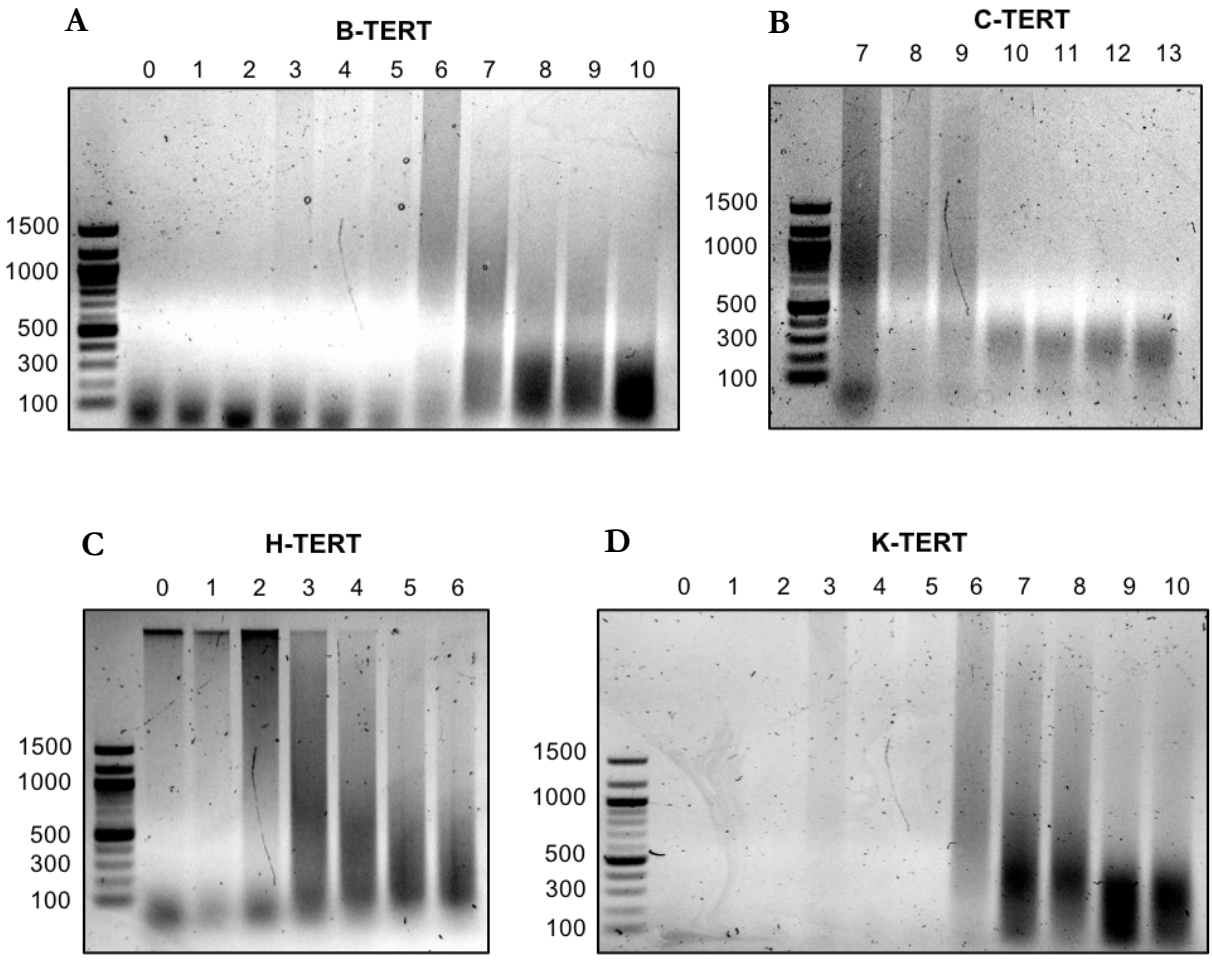


Figure 6.8 Optimisation of sonication in TERT-NHUC

$5x10^6$ TERT-NHUC cells were harvested and cross-linked in 1% formaldehyde, then sonicated with an increasing number of sonication cycles. Optimisation of sonication was carried out for **A)** B-TERT, **B)** C-TERT, **C)** H-TERT, and **D)** K-TERT. Samples were then reverse-crosslinked, purified into 0.1X TE buffer using phenol:chloroform purification and visualised using 1.0% agarose gel electrophoresis.

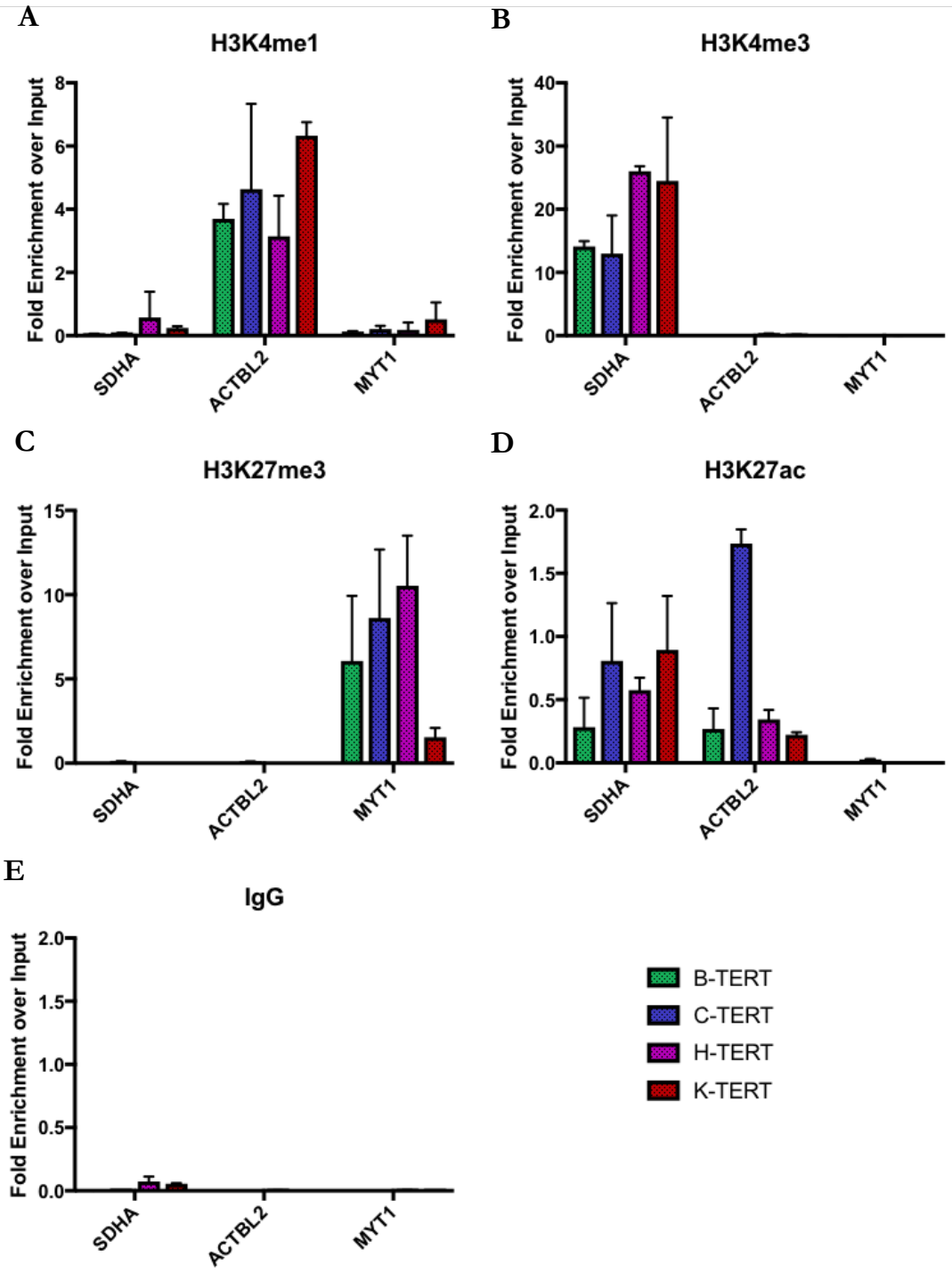


Figure 6.9 qPCR for enrichment of histones at control loci in TERT-NHUC

ChIP was carried out in TERT-NHUC using **A)** anti-H3K4me1, **B)** anti-H3K4me3, **C)** anti H3K27me3, **D)** anti-H3K27ac, **E)** anti-IgG. Following ChIP, qPCR was carried out to show histone enrichment at *SDHA, ACTBL2,* and *MYT-1.* Bar plots show average fold-change enrichment between biological duplicates for histone marks/IgG over input. Bars are coloured for B-TERT (green), C-TERT (blue), H-TERT (purple), and K-TERT (red).

6.3 Discussion

Despite the mass utilisation of ChIP-seq throughout the literature, few studies have carried out effective ChIP-seq for histone marks in bladder. For instance, of over 15,000 entries into the ENCODE consortium encompassing 147 cell/tissue types, only 6 entries include ChIP-seq in bladder, all of which have insufficient or extremely low read depth, short read lengths, and low library complexity (ENCODE Consortium, 2012). A similar lack of bladder samples is also seen with the Roadmap Epigenomics Project, where their integrative analysis of 111 primary human tissues and cells did not include bladder samples (Kundaje *et al.*, 2015). Nevertheless, some of examples of ChIP-seq in bladder samples can be found in smaller studies.

One such study utilised ChIP-seq to map H2K27me3 and H3K9me3 in the bladder cancer cell lines RT112♀ and T24♀, and in NHUC. The authors showed that H3K27me3, but not H3K9me3, was associated with only ~20% of genes that had low expression, and identified only 33 genes that were silenced in cancer cells and associated with H3K27me3. However, the authors neglected to include essential input and IgG controls, which prohibits effective normalisation and useful interpretation of their data (Dudziec *et al.*, 2012). This may in part explain the low number of silenced genes associated with H3K27me3 in the cancer cells in their analysis, and the minimal overlap of these epigenetic marks between each of the cell lines.

A recent study that demonstrated a novel approach for carrying out ChIP on formalin-fixed, paraffin-embedded tissue samples (FiT-seq), included FiT-seq for H3K4me2 on 6 male NMIBC samples as a demonstration of their protocol (Cejas *et al.*, 2016). However the antibody used to target H3K4me2 (Millipore 07-030) had previously shown preferential binding to many other histone marks in a separate study, including increased affinity for H3K4me1 and H3K4me3 (Rothbart *et al.*, 2015). Cejas *et al* also carried out FiT-seq for H3K4me2, H3K4me3, H3K9ac, H3K36me3, and H3K27me3 in a single male NMIBC sample. However, their choice of antibodies is again questionable given that non-specific binding and batch-to-batch variability had previously been noted for these antibodies in the histone antibody database (Rothbart *et al.*, 2015). The authors concluded that bladder cancer contains tumour-type-specific enhancers (Cejas *et al.*, 2016). However, the limited number of samples, lack of normal control, use of irrelevant histone markers, absence of chromatin accessibility data and poor choice of antibodies, suggest that these conclusions are exaggerated (Cejas *et al.*, 2016).

One study also carried out ChIP-seq for FOXA1 in the NMIBC cell line, RT4♂ (Warrick *et al.*, 2016). However, despite the study concluding that GATA3 and PPARγ cooperate with FOXA1 to promote a luminal phenotype in bladder cancer, DNA that was bound by FOXA1 did not show motif enrichment for GATA3 and PPARγ. Furthermore, a poor correlation was shown between biological ChIP-seq replicates, indicating the use of an ineffective ChIP protocol (Warrick *et al.*, 2016).

This chapter has demonstrated an optimised ChIP protocol for the enrichment of H3K4me1, H3K4me3, H3K27ac, and H3K27me3 in two male and two female TERT-NHUC and in RT112♀. The robustness of this protocol was demonstrated with both time and location, as the initial optimisation in RT112♀ was carried out two years prior to ChIP in TERT-NHUC and in a separate laboratory with separately prepared buffers and antibodies. However, it is noted that a distinct number of sonication cycles is required for each TERT-NHUC and for RT112♀. The protocol uses three control loci for the enrichment of these histone marks that were identified using the UCSC genome browser: the *SDHA* promoter, an intronic region of *MYT-1*, and intergenic enhancer of *ACTBL2*. Interestingly, validating enrichment of the histone marks across these loci in each TERT-NHUC by ChIP-qPCR correlated well with the ATAC-seq results. The *MYT-1* locus was enriched for the heterochromatin marker H3K27me3 and lacked ATAC-seq signal in all samples. The promoter of *SDHA* was enriched for H3K4me3 and H3K27ac and correlated with a high signal enrichment by ATAC-seq. The enhancer of *ACTBL2* was enriched for H3K4me1 and H3K27ac and also correlated with high ATAC-seq signal. Furthermore, C-TERT showed a stronger enrichment of H3K27ac at the *ACTBL2* enhancer compared to the other TERT-NHUC, and this was also seen by ATAC-seq where C-TERT had increased chromatin accessibility at this locus. These results followed the expected pattern of enrichment that was demonstrated by the use of the UCSC genome browser in 6 distinct cell types, and showed high correlation with the previous ATAC-seq results. Therefore, at the time of writing, library preparation is being carried out on these TERT-NHUC ChIP samples with the intention of carrying out NGS using the Illumina NextSeq 500 platform.

Similarly to the limited number of studies regarding epigenetics in bladder, there is also a lack of studies comparing the epigenomes of males and females in other tissues. One study highlighted this lack of consideration for gender in the epigenetics of cardiovascular diseases, where only 75 out of 3071 papers included both male and female samples, and 86% of papers exclusively used male samples (Hartman *et al.*, 2018). Only 13 papers considered stratifying some of their data for gender, all of which concerned DNA modifications and

showed conflicting results as to whether differences were observed between genders (Hartman *et al.*, 2018).

The majority of studies that do compare sex differences in epigenetics pertain to neurobiology, and again are predominantly focused on DNA modifications (McCarthy *et al.*, 2017). However, one study has reported histone modification biases in the bed nucleus of the terminal stria and preoptic area of mice (Shen *et al.*, 2015). This previously described sexually dimorphic region of the brain was shown by ChIP-seq to have differential enrichment of H3K4me3 at 248 loci that were associated with synaptic function, mainly at TSS in females. The differential enrichment was not correlated with increased expression of these genes when tested by RT-qPCR, and it was therefore hypothesised that these genes are in a primed state of activation to allow for quick changes in gene expression (Shen *et al.*, 2015).

Two studies by Waxman *et al* did report differences in the chromatin state of male and female mouse livers, particularly at distal intergenic chromatin-accessible regions (Ling *et al.*, 2010; Sugathan and Waxman, 2013). These regions primarily had a bias for H3K4me1 and H3K27ac in male mouse liver, and were associated with FOXA pioneer factors that were proposed to facilitate male-enriched STAT5 binding. Furthermore, a previously identified set of 1000 gender-related differentially expressed genes did not display differences in chromatin state at their TSS or within their gene bodies, which indicated that their differential regulation was driven by gender-related activation of distal enhancer regions (Sugathan and Waxman, 2013). However, genes that were only expressed in males were repressed by H3K27me3 across the gene body in females, although this was not reciprocated at the loci of female-expressed genes in males (Sugathan and Waxman, 2013).

One study integrated the ChromHMM model of chromatin states into a new bioinformatic tool (ChromDiff) to compare genome-wide chromatin states between biological groups (Ernst and Kellis, 2012; Yen and Kellis, 2015). As part of their demonstration, the authors used ChromDiff to compare chromatin states between male and female epigenome data obtained from the Roadmap Epigenomics Project. The authors found 536 different epigenomic features corresponding to 369 genes, 70% of which pertain to Polycomb-repressed heterochromatic states on the X-chromosome. Furthermore, only 2 out of the 368 genes with differential chromatin states between genders also showed differential expression, although this is consistent with what is expected of XCI (Yen and Kellis, 2015). This study showed that the majority of epigenetic differences between genders are located on the sex chromosomes; however, the authors restricted their analysis to gene-

bodies and therefore would have missed epigenetic differences at intergenic regions, including enhancers.

Given that the ATAC-seq results showed a widespread increase of chromatin accessibility in male TERT-NHUC that correlated with a global increase in H3K4me3 and H3K27ac, one may hypothesise that this is due to a widespread activation of *cis*-regulatory regions in males, whereas female TERT-NHUC are only primed at these same regions. Therefore, it may be expected that a similar distribution of enrichment of the key enhancer marker H3K4me1 should be seen in both male and female TERT-NHUC. However, there will be a large overlap for H3K4me1 enrichment with the chromatin-accessible regions identified by ATAC-seq in male TERT-NHUC. Furthermore, chromatin accessibility and H3K4me1 should be correlated with H3K27ac in male TERT-NHUC but not in female TERT-NHUC. Very few differences should be seen in regard to H3K27me3 enrichment throughout the genome; however, H3K4me3 should be enriched at the target promoter regions of the active enhancers in male TERT-NHUC, but will not be enriched in female TERT-NHUC.

This chapter has demonstrated that a reliable ChIP protocol for the enrichment of histone marks has been established for normal and cancerous bladder cells. This protocol is currently being combined with library preparation for NGS, and the results will complement the ATAC-seq and microarray results in this study. ChIP-seq for TERT-NHUC may confirm a hypothesis of a primed epigenetic state in female TERT-NHUC compared to an active epigenetic state in male TERT-NHUC. Furthermore, ChIP-seq in TERT-NHUC will provide a foundation for future studies of epigenetic perturbations in bladder cancer.

Chapter 7
Final Discussion

This exploratory project aimed to carry out transcriptional and chromatin accessibility profiling of normal human urothelial cells and determine gender-associated differences. Previous studies which have compared transcriptomes and epigenomes between genders have shown that gender-related differences are predominately located on the sex chromosomes, with minimal differences on autosomes (Ernst and Kellis, 2012; Deluca et al., 2015; Yen and Kellis, 2015; Singmann et al., 2015; Gershoni and Pietrokovski, 2017). The results of this project only partially agree with these previous findings. Gender-associated gene expression changes were mainly chrY genes and genes that commonly escape X-chromosome inactivation in females, and only 5 autosomal genes were identified as differentially expressed in females across three healthy urothelial models. However, the results from ATAC-seq indicated a wide-spread increase in chromatin accessibility across all chromosomes in male TERT-NHUC.

Although widespread chromatin accessibility changes are associated with embryonic development (Liu *et al.*, 2019; Wu *et al.*, 2016), similar observations have also been made in somatic tissues such as in metastatic SCLC and in macular degeneration (Denny et al., 2016; Wang et al., 2018). Such widespread increases are expected to result in similar widespread changes in gene expression (Liu *et al.*, 2019). However, in male TERT-NHUC the global increase in chromatin accessibility was not correlated with a widespread increase in gene expression, or by differential expression of histone modifiers. Indeed, there was very little correlation between the transcriptome data and chromatin accessibility data in this study. A similar lack of correlation between differential chromatin accessibility and gene expression data has been seen in other tissues (Ackermann et al., 2016; Fu et al., 2018). This discrepancy between chromatin accessibility and gene expression may simply result from ineffective mapping of a *cis*-regulatory chromatin accessible peak to its target gene, as the majority of studies, including this one, simply map peaks to their closest gene. However, in a large study of over 410 tumours samples, only 24% of chromatin accessible peaks were found to target their closest gene, suggesting that over 75% of chromatin accessible peaks in this study may not be annotated to the correct target gene (Corces *et al.*, 2018). Nevertheless, this does not explain the lack of widespread increase of gene expression in males. Results from Hi-C studies have shown that loss of chromatin looping, to which chromatin accessibility is

correlated, does not dramatically alter global gene expression (Rao *et al.*, 2017), and similar regulatory dynamics may be at play in male and female TERT-NHUC.

The widespread chromatin accessibility differences in TERT-NHUC were only supported by global increases in activating the histone marks H3K4me3 and H3K27ac, but not the heterochromatin mark H3K27me3 or by MNase digestion assays in this study. It is speculated therefore that the difference in chromatin accessibility is not the result of a large-scale heterochromatin event in female TERT-NHUC, but increased activation of *cis*-regulatory regions in male TERT-NHUC. Future ChIP-seq experiments will determine if this is the case. It is hypothesised that H3K4me1 will co-occupy sites with H3K27ac more often in males than females to indicated enhancer activation, and that females will have similar distribution of H3K4me1 to males but not colocalised with H3K27ac. H3K4me1 would also be expected to correlate with distal-intergenic and intronic peaks identified by ATAC-seq. This study has therefore also established a robust ChIP protocol in TERT-NHUC for H3K4me1, H3K4me3, H3K27ac, and H3K3me3, and ChIP-seq is currently underway.

The main limitation of this study is one intrinsic to the bladder research community, that is the lack of available urothelial models for *in vitro* research (Crallan et al., 2006; Mullenders et al., 2019). Moreover, current *in vitro* urothelial models are often phenotypically and genetically distinct from their *in vivo* counterparts. Indeed, this study showed distinct transcriptional profiles between cultured and uncultured urothelial cells, indicating a failure to faithfully recapitulate *in vivo* characteristics of the urothelium. The primary model used in this study was TERT-NHUC, and was chosen due to the high input of DNA that is required for ChIP, which could not be attained by UHUC or NHUC. Although low-input ChIP-like protocols such as ChIP-nexus and CUT&RUN have been developed in recent years (He et al., 2015; Skene and Henikoff, 2017), extensive optimisation of these protocols would have been required which was not feasible given the limited number of available UHUC samples. Indeed, the ATAC-seq protocol optimised in TERT-NHUC was applied to UHUC in a preliminary experiment, but over-digestion of chromatin indicated that further optimisation would have been required. Providing a sufficient number of samples for optimisation and experimental research is available, a combination of low-input techniques such as ATAC-seq and ChIP-nexus in UHUC is feasible, and would provide epigenomic data more relevant to the biology of the human urothelium. Nevertheless, such an approach would not be appropriate for further downstream i*n vitro* experiments such as drug assays and genetic manipulation, where longer-term cell culture is required.

In recent years, *in vitro* studies have seen a large shift towards organoid models. Compared to traditional 2D cell culture models, organoids more phenotypically and genetically resemble the tissues from which they are derived (Amiri et al., 2018). Furthermore, organoids can be cultured in a similar manner to 2D culture methods, therefore enabling further downstream *in vitro* experiments. Organoids have been established from many tissue types such as colon, stomach, brain, liver, and more (Drost and Clevers, 2018). Organoid models for bladder tumours are still in their infancy (Lee et al., 2018; Mullenders et al., 2019), though 3D culture of normal urothelial cells is possible and is currently being optimised (J Burns, unpublished data). It was therefore not appropriate for this project to have used bladder organoids, particularly when ATAC and ChIP also required optimising. Nevertheless, future *in vitro* studies for epigenetics in bladder should consider the use of organoids to increase the relevance to urothelium in patients, as was demonstrated for brain organoids where RNA-seq, ChIP-seq, and ATAC-seq in cortical organoids showed that the transcriptome and epigenetic landscape of iPSC-derived cortical organoids closely resemble primary cortical tissue (Amiri et al., 2018)

The limited number of available samples in this study reduces confidence in the findings. ATAC-seq was carried out using only two male and two female TERT-NHUC, and showed genome-wide chromatin accessibility differences that were consistent within gender groups and within biological duplicates. By including just one extra male and female TERT-NHUC in this study, confidence would have greatly increased. This is particularly necessary given that transcriptome analysis in 102 TaG2 samples showed that variation between individuals was greater than variation between genders. Indeed, previous ATAC-seq results in blood have shown that variation in chromatin accessibility is considerably greater between individuals than it is between genders (Qu *et al.*, 2015). It is therefore difficult to determine if the widespread chromatin accessibility differences shown by ATAC-seq in this study can be attributed to variation between genders, or to variation between individuals that by chance fitted into gender groups.

Throughout the entire study, one cell line was consistently an outlier compared to the others: K-TERT. In the transcriptome analysis, K-TERT clustered separately from other TERT-NHUC, and was closer to NHUC by PCA. Furthermore, the heatmaps of enriched gene-sets consistently showed that K-TERT had a distinct enrichment profile compared to TERT-NHUC, and again was more similar to NHUC. Interestingly, two K-TERT RNA samples failed QC following microarray, requiring more samples to be prepared for effective analysis. In the ATAC-seq analysis, K-TERT had the lowest number of chromatin accessible peaks, and identified peaks generally had lower signal enrichment. However, proliferation

rate and cell cycle distribution of K-TERT was similar to the other TERT-NHUC lines. Given that the transcriptional profile of K-TERT was more similar to that of NHUC, it could be speculated that these cells are more "normal" than their TERT-immortalised counterparts.

Although it was shown that retroviral transduction of hTERT immortalises NHUC without inactivation of the p16/Rb pathway, it does result in the differential expression of many Polycomb-group target genes involved in differentiation and tumorigenesis (Chapman *et al.*, 2006; Chapman *et al.*, 2008). It can be speculated that K-TERT does not have the same differential regulation in response to hTERT expression, and therefore these cells are more like their non-immortalised counterparts. It is possible that the diminished chromatin accessibility seen in K-TERT, and even H-TERT, is more akin to what is observed in non-immortalised cells, which would promote the hypothesis that increased chromatin accessibility in males is the result of hTERT expression. This would require testing chromatin accessibility in matched pairs of hTERT-immortalised and non-immortalised NHUC, although again such an experiment is limited by sample availability.

As epigenetic changes are reversible, therapies aimed at reversing epigenetic aberrations in cancer are being considered. Only two types of epigenetic therapies have been approved for clinical use to date; DNMT inhibitors for reversing DNA hypermethylation, and HDAC inhibitors for increasing histone acetylation (Bennett and Licht, 2018). The DNMT inhibitors azacytidine and decitabine are used in the treatment of myelodysplastic syndrome and acute myeloid leukaemia, and result in the demethylation of promoters and reactivation of genes (Kantarjian et al., 2007). Ongoing trials for the use of both of these compounds are being carried out in a number of solid tumours and hematologic malignancies (Lee and Song, 2017). There are currently four HDAC inhibitors for clinical use and include: vorinostat, romidepsin, and belinostat (for T-cell lymphoma), and panobinostat (for multiple myeloma), although over 20 clinical trials are currently underway for other HDAC inhibitors for both solid tumours and hematologic malignancies (Bennett and Licht, 2018). Many other drugs targeting HATs and HMTs are also in development.

Given the high rate of mutations in *KDM6A* in bladder cancer, *in vitro* studies have considered inhibition of its antagonist EZH2 (EZH2i), which results in reduced cancer cell proliferation and survival, and increased apoptosis (Ler et al., 2017; Chen et al., 2019). As EZH2 is upregulated in many tumour types, inhibitors such as EPZ-6438, GSK2816126, and CPI-1205 are being used in clinical trials to treat patients with B-cell lymphomas, non-Hodgkin lymphoma, and malignant rhabdoid tumour, and may therefore be appropriate for treating bladder cancers (Knutson et al., 2014; Kim and Roberts, 2016). Previous studies

have also shown that bladder cancer cell lines display varying responses to HDAC inhibition (HDACi), which is also relevant given that KDM6A also constitutes part of the KMT2C/KMT2D COMPASS-like complexes, components of which are commonly mutated in bladder cancer. (Rosik *et al.*, 2014; Lehmann *et al.*, 2014; Pinkerneil *et al.*, 2016; Kaletsch *et al.*, 2018; Vasudevan *et al.*, 2019). Given the results of this project, such drug-related research in bladder cancer should also take into consideration gender. For instance, female KDM6A^{mut} TaG2 tumours showed differential expression of genes that regulate chromatin architecture whereas their male counterparts did not. Therefore, HDACi and EZH2i may show greater therapeutic potential in female bladder cancers than in males.

The widespread increase of chromatin accessibility and global increase of activating histone marks in male TERT-NHUC indicates that normal urothelium may also respond differently to such epigenetic-based drug treatments. An interesting experiment to build upon the work of this study would be to treat male and female TERT-NHUC with HDACi and EZH2i. It is hypothesised that male but not female TERT-NHUC will respond to HDACi which may result in a chromatin accessibility profile similar to untreated TERT-NHUC. EZH2i on the other hand would not be expected to have gender-associated differences in response, as the chromatin differences in TERT-NHUC are hypothesised not to be the result of widespread heterochromatin in females.

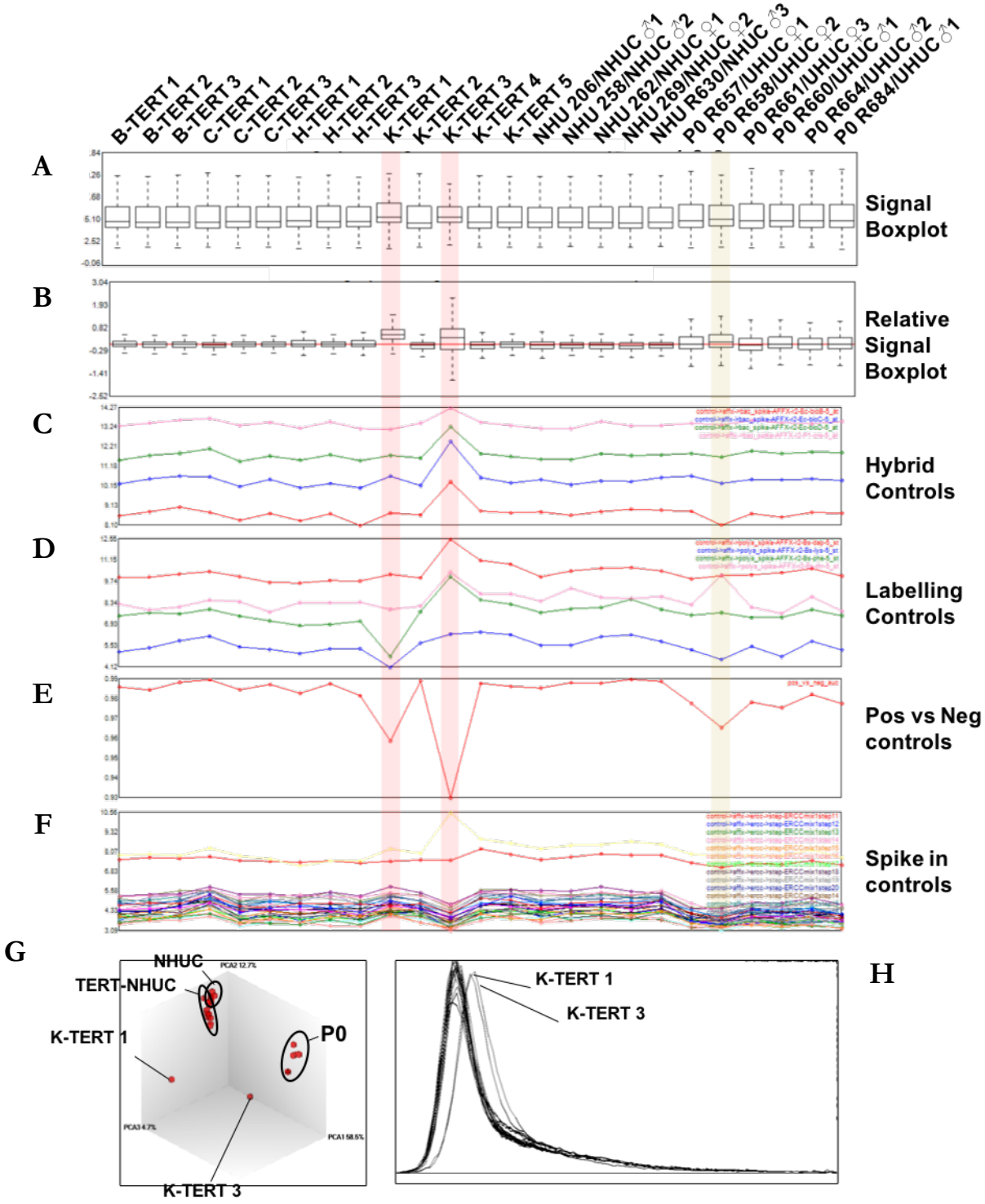
Despite documented differences in pharmacokinetics and pharmacodynamics of drug response between genders, the majority of biological research, from *in vitro* and *in vivo* studies through to clinical trials, is predominately carried out in males (Klein et al., 2015; Tannenbaum and Day, 2017). For instance, around 80% of all rodent studies are carried out only in males (Hughes, 2007). It is therefore expected that this over-reliance on male models in preclinical research masks intrinsic biological differences between genders, and may partially explain why women have higher rates of adverse effects to drug usage than men (Franconi et al., 2007). Indeed, in an effort to correct for male-biased research, the NIH recently set out policies that require applicants to report plans for the balance of male and female cells and animals in all preclinical studies (Clayton and Collins, 2014). The results of this study are possibly the first *in vitro* comparison between genders in urothelial cells, and accompany similar studies in hypothalamus, kidney, liver, and heart, as well as larger studies across multiple tissues (Rinn et al., 2004; Isensee et al., 2008; Deluca et al., 2015; Gershoni and Pietrokovski, 2017). These studies continue to demonstrate the need for each gender to be considered separately in future research.

Similar to the absence of female cells and animal models in preclinical studies is the neglect of bladder samples in larger scale transcriptome and epigenome projects such as such

as ENCODE, modENCODE, Roadmap Epigenetics Consortium, and GTEx to be included in their data banks (ENCODE Consortium, 2012; Contrino et al., 2012; Forrest et al., 2014; Kundaje et al., 2015; Deluca et al., 2015). Therefore, the bladder research community must continue to carry out smaller scale NGS studies, to further understanding of epigenetic mechanisms in bladder cancer. The generation of such data is of the upmost importance, given that bladder cancer has the highest mutation rate of chromatin modifying genes compared to any other cancer type. The results of this project go some way to filling this gap by providing chromatin accessibility data in commonly used normal urothelial cell lines, and a robust ChIP protocol for common histone marks.

To summarise, this project has carried out RNA microarray analysis on three models of normal human urothelium, and 102 TaG2 bladder tumours. Gender-associated gene expression changes were predominately located on the sex chromosomes, although five autosomal genes involved in inflammatory response and hypoxia were upregulated in females. Although transcriptional variation between individuals was greater than variation between genders, each gender showed different gene expression profiles in relation to the same gene mutation. ATAC-seq in immortalised normal urothelial cells showed a genome-wide increase of chromatin accessibility in males that was correlated with a global increase in activating histone marks, but not gene expression. An effective and robust ChIP protocol was established for common histone marks in immortalised normal urothelial cells.

Appendix A Microarray Data



Appendix A - 1 QA analysis on NHU-TERT, NHU, and P0 microarrays

QA on microarray data was carried out using the ThermoFisher Transcription Analysis Console. QA plot include **(A)** a boxplot of signal intensity, **(B)** a boxplot of normalised signal intensity, **(C)** a line graph of hybrid controls, **(D)** a line graph of poly-A labelling controls, **(E)** a receiver for operating characteristics (ROC) plot, **(F)** a line graph of ERCC spike controls, **(G)** a principal component analysis (PCA) plot, and **(H)** a histogram of log signal intensities. Samples highlighted in red have failed the QA.

Appendix A - 2 Top 100 DE male gene probes (male vs female TERT-NHUC)

Symbol	Gene name	Chr	FC	P.Value
ANXA6	annexin A6	chr5	29.92	0.00
MMP1	matrix metallopeptidase 1	chr11	23.06	0.00
MMP1	matrix metallopeptidase 1	chr11	19.70	0.00
MMP1	matrix metallopeptidase 1	chr11	17.78	0.00
RPS4Y1	ribosomal protein S4, Y-linked 1	chrY	16.55	0.00
TTTY15	testis-specific transcript, Y-linked 15 (non-protein coding)	chrY	16.16	0.00
IL33	interleukin 33	chr9	12.87	0.00
NLGN4Y	neuroligin 4, Y-linked	chrY	11.32	0.00
USP9Y	ubiquitin specific peptidase 9, Y-linked	chrY	10.72	0.00
DDX3Y	DEAD (Asp-Glu-Ala-Asp) box helicase 3, Y-linked	chrY	10.61	0.00
UTY	ubiquitously transcribed tetratricopeptide repeat containing, Y-linked	chrY	10.16	0.00
TXLNGY	taxilin gamma pseudogene, Y-linked	chrY	7.95	0.00
EYA4	EYA transcriptional coactivator and phosphatase 4	chr6	7.51	0.00
ZFY	zinc finger protein, Y-linked	chrY	7.35	0.00
GBP3	guanylate binding protein 3	chr1	7.30	0.00
SLC16A4	solute carrier family 16, member 4	chr1	7.28	0.00
FKBP10	FK506 binding protein 10	chr17	6.56	0.00
DSG3	desmoglein 3	chr18	6.33	0.03
EIF1AY	eukaryotic translation initiation factor 1A, Y-linked	chrY	6.31	0.00
EYA4	EYA transcriptional coactivator and phosphatase 4	chr6	6.11	0.00
TXLNGY	taxilin gamma pseudogene, Y-linked	chrY	6.03	0.00
KRTAP2-3	keratin associated protein 2-3	chr17	4.71	0.00
PLCB4	phospholipase C, beta 4	chr20	4.47	0.04
KRT6A	keratin 6A, type II	chr12	4.27	0.00
SAA1	serum amyloid A1	chr11	4.16	0.05
CSF3	colony stimulating factor 3	chr17	4.10	0.04
KRT6C	keratin 6C, type II	chr12	3.84	0.00
MRGPRX3	MAS-related GPR, member X3	chr11	3.59	0.00
KCCAT198	renal clear cell carcinoma-associated transcript 198	chr12	3.53	0.00
SORL1	sortilin-related receptor, L(DLR class) A repeats containing	chr11	3.38	0.00
CSF3	colony stimulating factor 3	chr17	3.34	0.03
UTY	ubiquitously transcribed tetratricopeptide repeat containing, Y-linked	chrY	3.34	0.00
KRT6B	keratin 6B, type II	chr12	3.31	0.00
SLFN11	schlafen family member 11	chr17	3.28	0.01
TAGLN	transgelin	chr11	3.26	0.01
NLGN4Y	neuroligin 4, Y-linked	chrY	3.24	0.00
CCDC144B	coiled-coil domain containing 144B (pseudogene)	chr17	3.24	0.01
SHISA2	shisa family member 2	chr13	3.20	0.00
CDH11	cadherin 11, type 2, OB-cadherin (osteoblast)	chr16	3.18	0.04
TNC	tenascin C	chr9	3.15	0.00
SNORA38B	small nucleolar RNA, H/ACA box 38B	chr17	3.15	0.01
RPS4Y2	ribosomal protein S4, Y-linked 2	chrY	3.08	0.00
MYOCD	myocardin	chr17	3.06	0.02
CCAT1	colon cancer associated transcript 1 (non-protein coding)	chr8	3.02	0.01
DKK1	dickkopf WNT signaling pathway inhibitor 1	chr10	2.91	0.00
LOC100134868	uncharacterized LOC100134868	chr20	2.90	0.00
VEPH1	ventricular zone expressed PH domain containing 1	chr3	2.89	0.00
KDM5D	lysine (K)-specific demethylase 5D	chrY	2.88	0.00
BMI1	BMI1 proto-oncogene, polycomb ring finger	chr10	2.84	0.02
FOXD1	forkhead box D1	chr5	2.83	0.00
C20orf197	chromosome 20 open reading frame 197	chr20	2.80	0.00
IL7R	interleukin 7 receptor	chr5	2.80	0.00
SAA2	serum amyloid A2	chr11	2.72	0.04
ZFY	zinc finger protein, Y-linked	chrY	2.68	0.00
LPXN	leupaxin	chr11	2.67	0.00
CXCL8	chemokine (C-X-C motif) ligand 8	chr4	2.66	0.00
C6orf223	chromosome 6 open reading frame 223	chr6	2.65	0.00

ERVK-7	endogenous retrovirus group K, member 7	chr1	2.59	0.00
SORL1	sortilin-related receptor, L(DLR class) A repeats containing	chr11	2.57	0.00
LOC400043	uncharacterized LOC400043	chr12	2.52	0.04
CLGN	calmegin	chr4	2.49	0.01
PTHLH	parathyroid hormone-like hormone	chr12	2.49	0.05
SEMA3A	sema domain, immunoglobulin domain (Ig), short basic domain, secreted,3A	chr7	2.49	0.00
TPM2	tropomyosin 2 (beta)	chr9	2.46	0.00
LOC105379362	uncharacterized LOC105379362	chr8	2.46	0.04
LOC400043	uncharacterized LOC400043	chr12	2.45	0.04
RP11-727M10.2	---	chr4	2.45	0.05
F3	coagulation factor III (thromboplastin, tissue factor)	chr1	2.42	0.00
IFITM2	interferon induced transmembrane protein 2	chr11	2.42	0.00
CCDC144B	coiled-coil domain containing 144B (pseudogene)	chr17	2.40	0.02
MIR1299	microRNA 1299	chr9	2.40	0.00
MXRA7	matrix-remodelling associated 7	chr2	2.32	0.00
PNPLA3	patatin-like phospholipase domain containing 3	chr22	2.26	0.00
CLEC2B	C-type lectin domain family 2, member B	chr12	2.24	0.02
MAP1B	microtubule associated protein 1B	chr5	2.24	0.00
LINC01002	long intergenic non-protein coding RNA 1002	chr19	2.22	0.00
EYA4	EYA transcriptional coactivator and phosphatase 4	chr6	2.21	0.00
GALNT5	polypeptide N-acetylgalactosaminyltransferase 5	chr2	2.20	0.00
FOXD1	forkhead box D1	chr5	2.19	0.00
CXCL8	chemokine (C-X-C motif) ligand 8	chr4	2.18	0.00
LOC729737	uncharacterized LOC729737	chr1	2.18	0.00
CCDC144CP	coiled-coil domain containing 144C, pseudogene	chr17	2.17	0.02
MXRA7	matrix-remodelling associated 7	chr2	2.16	0.00
PLLP	plasmolipin	chr16	2.13	0.00
IFITM2	interferon induced transmembrane protein 2	chr11	2.10	0.00
LINC01002	long intergenic non-protein coding RNA 1002	chr19	2.09	0.00
ABI3BP	ABI family, member 3 (NESH) binding protein	chr3	2.07	0.00
CCBE1	collagen and calcium binding EGF domains 1	chr18	2.07	0.00
CLGN	calmegin	chr4	2.05	0.01
PRKY	protein kinase, Y-linked, pseudogene	chrY	2.04	0.00
CCDC144CP	coiled-coil domain containing 144C, pseudogene	chr17	2.04	0.03
BBOX1-AS1	BBOX1 antisense RNA 1	chr11	2.03	0.00
SLC22A3	solute carrier family 22 (organic cation transporter), member 3	chr6	2.02	0.00
LTB	lymphotoxin beta (TNF superfamily, member 3)	chr6	2.00	0.03
CXCL5	chemokine (C-X-C motif) ligand 5	chr4	1.99	0.04
FKBP11	FK506 binding protein 11	chr12	1.98	0.00
MIR503HG	MIR503 host gene	chrX	1.95	0.01
HEG1	heart development protein with EGF-like domains 1	chr3	1.95	0.00
ALOX5AP	arachidonate 5-lipoxygenase-activating protein	chr13	1.95	0.01
PRKY	protein kinase, Y-linked, pseudogene	chrY	1.95	0.00

Appendix A - 3 Top 100 DE female gene probes (male vs female TERT-NHUC)

Symbol	Gene name	Chr	FC	P.Value
XIST	X inactive specific transcript (non-protein coding)	chrX	35.45	0.00
XIST	X inactive specific transcript (non-protein coding)	chrX	26.66	0.00
MIR4273	microRNA 4273	chr3	14.06	0.00
FRG2C	FSHD region gene 2 family, member C	chr3	13.12	0.00
ALOX15B	arachidonate 15-lipoxygenase, type B	chr17	10.83	0.01
UCA1	urothelial cancer associated 1 (non-protein coding)	chr19	10.59	0.00
UCA1	urothelial cancer associated 1 (non-protein coding)	chr19	9.40	0.00
LINC00960	long intergenic non-protein coding RNA 960	chr3	8.83	0.00
TMEM47	transmembrane protein 47	chrX	7.04	0.00
ADGRL4	adhesion G protein-coupled receptor L4	chr1	7.00	0.00
NLRP2	NLR family, pyrin domain containing 2	chr19	6.75	0.00
PRKAA2	protein kinase, AMP-activated, alpha 2 catalytic subunit	chr1	5.82	0.00
XG	Xg blood group	chrX	4.86	0.03
LINC00960	long intergenic non-protein coding RNA 960	chr3	4.50	0.00
VGLL1	vestigial-like family member 1	chrX	4.33	0.03
AKT3	v-akt murine thymoma viral oncogene homolog 3	chr1	4.08	0.01
PAQR5	progestin and adipoQ receptor family member V	chr15	4.05	0.00
GJB6	gap junction protein beta 6	chr13	3.71	0.00
SORT1	sortilin 1	chr1	3.39	0.01
HCAR3	hydroxycarboxylic acid receptor 3	chr12	3.27	0.00
RPS6KA6	ribosomal protein S6 kinase, 90kDa, polypeptide 6	chrX	3.17	0.00
TXNIP	thioredoxin interacting protein	chr1	3.16	0.00
DPY19L2P1	DPY19L2 pseudogene 1	chr7	3.11	0.00
LAMA1	laminin, alpha 1	chr18	3.07	0.00
LINC01296	long intergenic non-protein coding RNA 1296	chr22	2.98	0.00
TMOD2	tropomodulin 2 (neuronal)	chr15	2.98	0.00
VCAN	versican	chr5	2.97	0.00
KLF8	Kruppel-like factor 8	chrX	2.92	0.01
PEG10	paternally expressed 10	chr7	2.90	0.00
RPS6KA6	ribosomal protein S6 kinase, 90kDa, polypeptide 6	chrX	2.78	0.00
LOC102723854	uncharacterized LOC102723854	chr2	2.75	0.04
GPX3	glutathione peroxidase 3	chr5	2.75	0.00
LAMA1	laminin, alpha 1	chr18	2.67	0.00
PLXDC2	plexin domain containing 2	chr10	2.66	0.00
ACAA2	acetyl-CoA acyltransferase 2	chr18	2.64	0.00
ERAP2	endoplasmic reticulum aminopeptidase 2	chr5	2.60	0.05
TC2N	tandem C2 domains, nuclear	chr14	2.59	0.03
CHRNB1	cholinergic receptor, nicotinic beta 1	chr17	2.56	0.00
GIPC2	GIPC PDZ domain containing family, member 2	chr1	2.55	0.00
PYGO1	pygopus family PHD finger 1	chr15	2.55	0.01
LOC105372321	uncharacterized LOC105372321	chr19	2.49	0.00
SVIP	small VCP/p97-interacting protein	chr11	2.47	0.00
GSTM3	glutathione S-transferase mu 3 (brain)	chr1	2.47	0.00
SGCE	sarcoglycan epsilon	chr7	2.44	0.00
AKT3	v-akt murine thymoma viral oncogene homolog 3	chr1	2.43	0.01
RHOB	ras homolog family member B	chr2	2.42	0.00
GJB6	gap junction protein beta 6	chr13	2.40	0.00
DPY19L2P1	DPY19L2 pseudogene 1	chr7	2.38	0.00
SLC38A5	solute carrier family 38, member 5	chrX	2.37	0.03
HOXD10	homeobox D10	chr2	2.36	0.00
HNRNPLL	heterogeneous nuclear ribonucleoprotein L-like	chr2	2.35	0.00
HTRA1	HtrA serine peptidase 1	chr10	2.35	0.00
LOC100132111	uncharacterized LOC100132111	chr1	2.33	0.00
XYLT1	xylosyltransferase I	chr16	2.31	0.05
ENPP5	ectonucleotide pyrophosphatase/phosphodiesterase 5 (putative)	chr6	2.25	0.00
GSTM3	glutathione S-transferase mu 3 (brain)	chr1	2.24	0.00
DUXAP10	double homeobox A pseudogene 10	chr14	2.24	0.02
IL12RB2	interleukin 12 receptor, beta 2	chr1	2.22	0.01

HCAR2	hydroxycarboxylic acid receptor 2	chr12	2.21	0.00
CYP24A1	cytochrome P450, family 24, subfamily A, polypeptide 1	chr20	2.18	0.00
EML1	echinoderm microtubule associated protein like 1	chr14	2.17	0.00
PYGO1	pygopus family PHD finger 1	chr15	2.16	0.03
SLCO4C1	solute carrier organic anion transporter family, member 4C1	chr5	2.14	0.00
H1F0	H1 histone family, member 0	chr22	2.14	0.01
SHC3	SHC (Src homology 2 domain containing) transforming protein 3	chr9	2.14	0.00
CORO2B	coronin, actin binding protein, 2B	chr15	2.13	0.00
SNX9	sorting nexin 9	chr6	2.12	0.00
ZYG11A	zyg-11 family member A, cell cycle regulator	chr1	2.09	0.00
ARHGAP32	Rho GTPase activating protein 32	chr11	2.05	0.01
IRAK2	interleukin 1 receptor associated kinase 2	chr3	2.04	0.00
DPY19L2	dpy-19-like 2 (C. elegans)	chr12	2.04	0.00
MFSD6	major facilitator superfamily domain containing 6	chr2	2.04	0.01
SCN9A	sodium channel, voltage gated, type IX alpha subunit	chr2	2.02	0.00
CAT	catalase	chr11	2.02	0.00
LGALS9C	lectin, galactoside-binding, soluble, 9C	chr17	2.01	0.00
PRKD1	protein kinase D1	chr14	2.01	0.00
VGLL3	vestigial-like family member 3	chr3	2.01	0.00
LINC01296	long intergenic non-protein coding RNA 1296	chr14	2.00	0.03
TMOD2	tropomodulin 2 (neuronal)	chr15	2.00	0.00
EIF1AX	eukaryotic translation initiation factor 1A, X-linked	chrX	1.99	0.00
KIAA1324L	KIAA1324-like	chr7	1.96	0.04
PI3	peptidase inhibitor 3, skin-derived	chr20	1.96	0.00
FZD3	frizzled class receptor 3	chr8	1.96	0.03
SPOCK1	sparc/osteonectin, cwcv and kazal-like domains proteoglycan 1	chr5	1.95	0.00
NCOA1	nuclear receptor coactivator 1	chr2	1.95	0.03
SMC1A	structural maintenance of chromosomes 1A	chrX	1.94	0.00
SCCPDH	saccharopine dehydrogenase (putative)	chr1	1.93	0.00
MT1E	metallothionein 1E	chr16	1.93	0.03
GJB2	gap junction protein beta 2	chr13	1.91	0.00
MCTP1	multiple C2 domains, transmembrane 1	chr5	1.91	0.00
SCARA3	scavenger receptor class A, member 3	chr8	1.90	0.00
UGT8	UDP glycosyltransferase 8	chr4	1.89	0.00
BST2	bone marrow stromal cell antigen 2	chr19	1.88	0.05
LGALS9C	lectin, galactoside-binding, soluble, 9C	chr17	1.88	0.00
MMP2	matrix metallopeptidase 2	chr16	1.88	0.00
SHC3	SHC (Src homology 2 domain containing) transforming protein 3	chr9	1.87	0.00
DHRS2	dehydrogenase/reductase (SDR family) member 2	chr14	1.87	0.00
SCD5	stearoyl-CoA desaturase 5	chr4	1.86	0.00
SPATA22	spermatogenesis associated 22	chr17	1.86	0.04
APBA2	amyloid beta (A4) precursor protein-binding, family A, member 2	chr15	1.86	0.00

Appendix A - 4 able or snoRNAs and rRNAs that were upregulated in males (male vs female TERT NHUC)

Symbol	Gene name	Chr	FC	P.Value
RNA5S1	RNA, 5S ribosomal 1	chr1	1.70	0.04
RNA5S10	RNA, 5S ribosomal 10	chr1	1.70	0.04
RNA5S11	RNA, 5S ribosomal 11	chr1	1.70	0.04
RNA5S12	RNA, 5S ribosomal 12	chr1	1.70	0.04
RNA5S13	RNA, 5S ribosomal 13	chr1	1.70	0.04
RNA5S14	RNA, 5S ribosomal 14	chr1	1.70	0.04
RNA5S15	RNA, 5S ribosomal 15	chr1	1.70	0.04
RNA5S16	RNA, 5S ribosomal 16	chr1	1.70	0.04
RNA5S17	RNA, 5S ribosomal 17	chr1	1.70	0.04
RNA5S2	RNA, 5S ribosomal 2	chr1	1.70	0.04
RNA5S3	RNA, 5S ribosomal 3	chr1	1.70	0.04
RNA5S4	RNA, 5S ribosomal 4	chr1	1.70	0.04
RNA5S5	RNA, 5S ribosomal 5	chr1	1.70	0.04
RNA5S6	RNA, 5S ribosomal 6	chr1	1.70	0.04
RNA5S7	RNA, 5S ribosomal 7	chr1	1.70	0.04
RNA5S8	RNA, 5S ribosomal 8	chr1	1.70	0.04
RPS4Y1	ribosomal protein S4, Y-linked 1	chrY	16.55	0.00
RPS4Y2	ribosomal protein S4, Y-linked 2	chrY	3.08	0.00
SCARNA12	small Cajal body-specific RNA 12	chr12	1.56	0.01
SCARNA12	small Cajal body-specific RNA 12	chr12	1.56	0.00
SNORA38B	small nucleolar RNA, H/ACA box 38B	chr17	3.15	0.01
SNORD111	small nucleolar RNA, C/D box 111	chr16	1.56	0.02
SNORD14D	small nucleolar RNA, C/D box 14D	chr11	1.65	0.00
SNORD3A	small nucleolar RNA, C/D box 3A	chr17	1.81	0.04
SNORD3B-1	small nucleolar RNA, C/D box 3B-1	chr17	1.83	0.04
SNORD3B-2	small nucleolar RNA, C/D box 3B-2	chr17	1.83	0.04
SNORD3C	small nucleolar RNA, C/D box 3C	chr17	1.87	0.03
SNORD3D	small nucleolar RNA, C/D box 3D	chr17	1.76	0.05
SNORD58C	small nucleolar RNA, C/D box 58C	chr18	1.57	0.05
SNORD93	small nucleolar RNA, C/D box 93	chr7	1.64	0.03

Appendix A - 5 Male enriched gene-sets from GSEA between male vs female TERT-NHUC

Gene-Set (male enriched)	Set-Size	Score	p-value	q-value
KEGG_VASCULAR_SMOOTH_MUSCLE_CONTRACTION	108	0.33	0.00	0.57
REACTOME_CELL_CELL_COMMUNICATION	109	0.41	0.00	0.28
GO_PROTEIN_MODIFICATION_BY_SMALL_PROTEIN_REMOVAL	119	0.38	0.00	1.00
GO_UBIQUITIN_LIKE_PROTEIN_SPECIFIC_PROTEASE_ACTIVITY	103	0.42	0.00	0.66
GO_PROTEIN_FOLDING	207	0.40	0.03	0.55
GO_COLUMNAR_CUBOIDAL_EPITHELIAL_CELL_DIFFERENTIATION	106	0.41	0.00	0.69
GO_REGULATION_OF_G_PROTEIN_COUPLED_RECEPTOR_PROTEIN_SIGNALING_PATHWAY	121	0.40	0.00	0.65
GO_TRANSLATIONAL_INITIATION	107	0.60	0.01	0.75
GO_RESPONSE_TO_RETINOIC_ACID	102	0.39	0.00	0.89
GO_ENDOPLASMIC_RETICULUM_LUMEN	188	0.36	0.00	0.83
GO_SYNAPSE_ORGANIZATION	139	0.39	0.02	0.81
GO_POSITIVE_REGULATION_OF_BINDING	118	0.39	0.04	0.75
GO_STEM_CELL_DIFFERENTIATION	184	0.38	0.01	0.71
GO_RESPIRATORY_SYSTEM_DEVELOPMENT	188	0.35	0.01	0.68
GO_RESPONSE_TO_MECHANICAL_STIMULUS	200	0.33	0.00	0.69
GO_REGULATION_OF_MUSCLE_CONTRACTION	141	0.37	0.00	0.66
GO_NEGATIVE_REGULATION_OF_IMMUNE_SYSTEM_PROCESS	336	0.37	0.01	0.65
GO_DEVELOPMENTAL_GROWTH_INVOLVED_IN_MORPHOGENESIS	101	0.37	0.00	0.64
GO_DEFENSE_RESPONSE_TO_VIRUS	151	0.41	0.00	0.63
GO_DRUG_BINDING	103	0.37	0.05	0.68
GO_MESENCHYME_DEVELOPMENT	182	0.37	0.05	0.64
GO_ISOMERASE_ACTIVITY	144	0.34	0.04	0.63
GO_CELL_GROWTH	131	0.37	0.02	0.61
GO_REGULATION_OF_ADAPTIVE_IMMUNE_RESPONSE	121	0.40	0.02	0.60
GO_RESPONSE_TO_FIBROBLAST_GROWTH_FACTOR	111	0.37	0.02	0.58
GO_NEGATIVE_REGULATION_OF_IMMUNE_RESPONSE	113	0.44	0.02	0.57
GO_NEGATIVE_REGULATION_OF_CELL_ADHESION	215	0.35	0.01	0.55
GO_REGULATION_OF_MUSCLE_SYSTEM_PROCESS	185	0.34	0.01	0.55
GO_BRANCHING_MORPHOGENESIS_OF_AN_EPITHELIAL_TUBE	128	0.37	0.04	0.54
GO_REGIONALIZATION	295	0.33	0.04	0.54
GO_SARCOLEMMA	120	0.33	0.01	0.53
GO_REGULATION_OF_MUSCLE_CELL_DIFFERENTIATION	146	0.34	0.00	0.58
GO_NOTCH_SIGNALING_PATHWAY	108	0.33	0.01	0.60
GO_REGULATION_OF_LEUKOCYTE_MIGRATION	141	0.42	0.01	0.60
GO_TRANSCRIPTION_FACTOR_ACTIVITY_RNA_POLYMERASE_II_CORE_PROMOTER_PROXIMAL_REGION_SEQUENCE_SPECIFIC_BINDING	321	0.28	0.02	0.59
GO_POSITIVE_REGULATION_OF_GROWTH	224	0.27	0.03	0.60
GO_MORPHOGENESIS_OF_A_BRANCHING_STRUCTURE	164	0.35	0.03	0.60
GO_NEPHRON_DEVELOPMENT	113	0.34	0.01	0.59
GO_DEVELOPMENTAL_GROWTH	320	0.28	0.01	0.59
GO_REGULATION_OF_CHEMOTAXIS	174	0.35	0.03	0.59
GO_DIGESTIVE_SYSTEM_DEVELOPMENT	143	0.35	0.02	0.57
GO_POSITIVE_REGULATION_OF_CHEMOTAXIS	117	0.38	0.03	0.57
GO_EXCITATORY_SYNAPSE	190	0.28	0.04	0.61
GO_INTERACTION_WITH_HOST	129	0.31	0.04	0.61
GO_RESPONSE_TO_VIRUS	228	0.32	0.01	0.60
GO_CYTOKINE_MEDIATED_SIGNALING_PATHWAY	427	0.34	0.03	0.59
GO_POSITIVE_REGULATION_OF_DEVELOPMENTAL_GROWTH	150	0.29	0.05	0.59
GO_ORGANIC_ACID_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	138	0.31	0.04	0.58
GO_ADAPTIVE_IMMUNE_RESPONSE_BASED_ON_SOMATIC_RECOMBINATION_OF_IMMUNE_RECEPTORS_BUILT_FROM_IMMUNOGLOBULIN_SUPERFAMILY_DOMAINS	122	0.37	0.05	0.57
GO_HOMOPHILIC_CELL_ADHESION_VIA_PLASMA_MEMBRANE_ADHESION_MOLECULES	119	0.38	0.01	0.57
GO_GROWTH	394	0.27	0.04	0.58
GO_APOPTOTIC_SIGNALING_PATHWAY	272	0.29	0.05	0.60
GO_GLYCEROLIPID_BIOSYNTHETIC_PROCESS	200	0.23	0.04	0.62
GO_IMMUNE_EFFECTOR_PROCESS	430	0.29	0.00	0.61
GO_REGULATION_OF_CELL_ACTIVATION	447	0.29	0.03	0.62
GO_POSITIVE_REGULATION_OF_CYTOKINE_PRODUCTION	349	0.29	0.05	0.60
GO_REGULATION_OF_OSSIFICATION	168	0.29	0.04	0.60
GO_CELL_MORPHOGENESIS_INVOLVED_IN_DIFFERENTIATION	490	0.24	0.02	0.61

Appendix A - 6 Female enriched gene-sets from GSEA between male vs female TERT-NHUC

Gene-Set (female enriched)	Set-Size	Score	p-value	q-value
HALLMARK_UV_RESPONSE_UP	157	-0.43	0.01	0.56
KEGG_T_CELL_RECEPTOR_SIGNALING_PATHWAY	105	-0.37	0.03	0.31
REACTOME_SIGNALING_BY_FGFR_IN_DISEASE	116	-0.30	0.02	1.00
GO_NEGATIVE_REGULATION_OF_INTRACELLULAR_TRANSPORT	134	-0.41	0.00	1.00
GO_ACTIN_FILAMENT_BINDING	114	-0.38	0.01	1.00
GO_INTRACELLULAR_RECEPTOR_SIGNALING_PATHWAY	159	-0.42	0.00	0.73
GO_REGULATION_OF_CIRCADIAN_RHYTHM	101	-0.38	0.00	0.65
GO_STEROID_HORMONE_MEDIATED_SIGNALING_PATHWAY	119	-0.43	0.00	0.54
GO_VACUOLAR_LUMEN	106	-0.41	0.01	0.62
GO_NEGATIVE_REGULATION_OF_CYTOPLASMIC_TRANSPORT	111	-0.37	0.01	0.61
GO_GERM_CELL_DEVELOPMENT	203	-0.41	0.00	0.67
GO_REGULATION_OF_PHOSPHATASE_ACTIVITY	121	-0.34	0.02	0.64
GO_FATTY_ACID_METABOLIC_PROCESS	278	-0.37	0.00	0.67
GO_HYDROLASE_ACTIVITY_ACTING_ON_GLYCOSYL_BONDS	112	-0.31	0.02	0.72
GO_CELLULAR_PROCESS_INVOLVED_IN_REPRODUCTION_IN_MULTICELLULAR_ORGANISM	243	-0.38	0.00	0.68
GO_CELL_MATRIX_ADHESION	111	-0.42	0.01	0.63
GO_MONOCARBOXYLIC_ACID_BIOSYNTHETIC_PROCESS	161	-0.40	0.00	0.59
GO_RAS_GUANYL_NUCLEOTIDE_EXCHANGE_FACTOR_ACTIVITY	212	-0.35	0.04	0.59
GO_NEGATIVE_REGULATION_OF_CELLULAR_PROTEIN_LOCALIZATION	131	-0.33	0.03	0.56
GO_LOCOMOTORY_BEHAVIOR	177	-0.39	0.00	0.61
GO_PRIMARY_CILIUM	191	-0.35	0.04	0.61
GO_SPINAL_CORD_DEVELOPMENT	104	-0.38	0.02	0.62
GO_PEPTIDYL_SERINE_MODIFICATION	144	-0.39	0.02	0.60
GO_GUANYL_NUCLEOTIDE_EXCHANGE_FACTOR_ACTIVITY	282	-0.30	0.02	0.59
GO_HORMONE_MEDIATED_SIGNALING_PATHWAY	152	-0.36	0.03	0.61
GO_REGULATION_OF_DEPHOSPHORYLATION	151	-0.33	0.04	0.63
GO_NEGATIVE_REGULATION_OF_PROTEIN_COMPLEX_DISASSEMBLY	158	-0.33	0.04	0.62
GO_TETRAPYRROLE_BINDING	123	-0.42	0.02	0.61
GO_AMINO_ACID_TRANSPORT	121	-0.36	0.01	0.59
GO_NEGATIVE_REGULATION_OF_SEQUENCE_SPECIFIC_DNA_BINDING_TRANSCRIPTION_FACTOR_ACTIVITY	131	-0.37	0.04	0.59
GO_MONOCARBOXYLIC_ACID_METABOLIC_PROCESS	469	-0.36	0.01	0.60
GO_HORMONE_RECEPTOR_BINDING	155	-0.29	0.00	0.59
GO_SEX_DIFFERENTIATION	249	-0.32	0.00	0.61
GO_FATTY_ACID_BIOSYNTHETIC_PROCESS	104	-0.40	0.02	0.59
GO_CELL_SUBSTRATE_ADHESION	155	-0.37	0.04	0.59
GO_CELLULAR_RESPONSE_TO_STEROID_HORMONE_STIMULUS	206	-0.33	0.02	0.58
GO_ORGANIC_ACID_BIOSYNTHETIC_PROCESS	255	-0.38	0.01	0.57
GO_CELLULAR_RESPONSE_TO_OXIDATIVE_STRESS	178	-0.37	0.04	0.56
GO_OXIDOREDUCTASE_ACTIVITY_ACTING_ON_PAIRED_DONORS_WITH_INCORPORATION_OR_REDUCTION_OF_MOLECULAR_OXYGEN	146	-0.37	0.00	0.54
GO_GROWTH_FACTOR_BINDING	118	-0.37	0.00	0.54
GO_ORGANIC_HYDROXY_COMPOUND_BIOSYNTHETIC_PROCESS	171	-0.35	0.03	0.53
GO_RESPONSE_TO_TOXIC_SUBSTANCE	227	-0.40	0.02	0.52
GO_VESICLE_LOCALIZATION	209	-0.28	0.05	0.54
GO_POSITIVE_REGULATION_OF_NF_KAPPAB_TRANSCRIPTION_FACTOR_ACTIVITY	119	-0.38	0.02	0.53
GO_POSITIVE_REGULATION_OF_I_KAPPAB_KINASE_NF_KAPPAB_SIGNALING	165	-0.34	0.03	0.52
GO_RESPONSE_TO_KETONE	172	-0.40	0.00	0.53
GO_IRON_ION_BINDING	150	-0.37	0.00	0.53
GO_REGULATION_OF_REACTIVE_OXYGEN_SPECIES_METABOLIC_PROCESS	146	-0.34	0.01	0.53
GO_CARBOHYDRATE_DERIVATIVE_CATABOLIC_PROCESS	169	-0.31	0.00	0.51
GO_STEROID_METABOLIC_PROCESS	221	-0.36	0.03	0.50
GO_REGULATION_OF_CARBOHYDRATE_METABOLIC_PROCESS	158	-0.35	0.03	0.51
GO_REGULATION_OF_PEPTIDASE_ACTIVITY	367	-0.34	0.00	0.50

GO_REGULATION_OF_SEQUENCE_SPECIFIC_DNA_BINDING_TRANSCRIPTION_FACTOR_ACTIVITY	340	-0.31	0.02	0.50
GO_LYTIC_VACUOLE	492	-0.27	0.01	0.49
GO_APPENDAGE_DEVELOPMENT	161	-0.30	0.03	0.48
GO_REGULATION_OF_I_KAPPAB_KINASE_NF_KAPPAB_SIGNALING	217	-0.32	0.02	0.48
GO_REGULATION_OF_CYSTEINE_TYPE_ENDOPEPTIDASE_ACTIVITY	201	-0.34	0.02	0.47
GO_SMALL_MOLECULE_BIOSYNTHETIC_PROCESS	424	-0.35	0.05	0.47
GO_STEROID_BIOSYNTHETIC_PROCESS	110	-0.36	0.03	0.47
GO_CARBOHYDRATE_HOMEOSTASIS	161	-0.32	0.02	0.47
GO_EXOCYTOSIS	293	-0.30	0.04	0.46
GO_NITROGEN_COMPOUND_TRANSPORT	474	-0.26	0.03	0.46
GO_MALE_SEX_DIFFERENTIATION	142	-0.34	0.03	0.51
GO_EAR_DEVELOPMENT	190	-0.32	0.02	0.50
GO_RESPONSE_TO_STEROID_HORMONE	473	-0.30	0.03	0.50
GO_STEROL_METABOLIC_PROCESS	117	-0.32	0.04	0.50
GO_RHYTHMIC_PROCESS	284	-0.28	0.03	0.51
GO_CELLULAR_RESPONSE_TO_EXTRACELLULAR_STIMULUS	183	-0.32	0.05	0.52
GO_DEVELOPMENT_OF_PRIMARY_SEXUAL_CHARACTERISTICS	202	-0.29	0.02	0.53
GO_CELL_ADHESION_MOLECULE_BINDING	178	-0.37	0.05	0.52
GO_ORGANIC_HYDROXY_COMPOUND_METABOLIC_PROCESS	460	-0.27	0.04	0.52
GO_REPRODUCTIVE_SYSTEM_DEVELOPMENT	390	-0.30	0.04	0.51
GO_REGULATION_OF_LIPID_METABOLIC_PROCESS	261	-0.26	0.03	0.51
GO_ALCOHOL_METABOLIC_PROCESS	333	-0.27	0.04	0.51
GO_RECEPTOR_COMPLEX	315	-0.26	0.05	0.50

Appendix A - 7 Top 100 (of 103) upregulated genes in male NHUC/TERT-NHUC

Symbol	Gne Name	Chr	FC	P-value
MMP1	matrix metallopeptidase 1	chr11	28.02	0.01
MMP1	matrix metallopeptidase 1	chr11	26.58	0.01
MMP1	matrix metallopeptidase 1	chr11	23.19	0.01
TTTY15	testis-specific transcript, Y-linked 15 (non-protein coding)	chrY	16.82	0.00
ANXA6	annexin A6	chr5	15.67	0.00
RPS4Y1	ribosomal protein S4, Y-linked 1	chrY	14.41	0.00
NLGN4Y	neuroligin 4, Y-linked	chrY	10.88	0.00
DDX3Y	DEAD (Asp-Glu-Ala-Asp) box helicase 3, Y-linked	chrY	10.5	0.00
USP9Y	ubiquitin specific peptidase 9, Y-linked	chrY	10.04	0.00
GBP3	guanylate binding protein 3	chr1	7.69	0.01
UTY	ubiquitously transcribed tetratricopeptide repeat containing, Y-linked	chrY	7.34	0.00
DSG3	desmoglein 3	chr18	7.25	0.05
TXLNGY	taxilin gamma pseudogene, Y-linked	chrY	7.2	0.00
SLC16A4	solute carrier family 16, member 4	chr1	6.68	0.00
ZFY	zinc finger protein, Y-linked	chrY	6.28	0.00
EIF1AY	eukaryotic translation initiation factor 1A, Y-linked	chrY	6.14	0.00
TXLNGY	taxilin gamma pseudogene, Y-linked	chrY	6.11	0.00
MYOCD	myocardin	chr17	5.93	0.00
KRTAP2-3	keratin associated protein 2-3	chr17	5.2	0.01
FKBP10	FK506 binding protein 10	chr17	4.81	0.00
EYA4	EYA transcriptional coactivator and phosphatase 4	chr6	4.29	0.00
EYA4	EYA transcriptional coactivator and phosphatase 4	chr6	3.64	0.00
UTY	ubiquitously transcribed tetratricopeptide repeat containing, Y-linked	chrY	3.21	0.00
FOXD1	forkhead box D1	chr5	3.2	0.01
NLGN4Y	neuroligin 4, Y-linked	chrY	3	0.00
SLFN11	schlafen family member 11	chr17	2.81	0.02
KDM5D	lysine (K)-specific demethylase 5D	chrY	2.78	0.00
LOC100134868	uncharacterized LOC100134868	chr20	2.77	0.00
MRGPRX3	MAS-related GPR, member X3	chr11	2.71	0.03
ZFY	zinc finger protein, Y-linked	chrY	2.69	0.00
MYOCD	myocardin	chr17	2.69	0.00
RPS4Y2	ribosomal protein S4, Y-linked 2	chrY	2.62	0.00
TNC	tenascin C	chr9	2.4	0.00
IL7R	interleukin 7 receptor	chr5	2.36	0.01
AOX1	aldehyde oxidase 1	chr2	2.35	0.01
CCDC144B	coiled-coil domain containing 144B (pseudogene)	chr17	2.34	0.02
KCCAT198	renal clear cell carcinoma-associated transcript 198	chr12	2.33	0.01
F3	coagulation factor III (thromboplastin, tissue factor)	chr1	2.18	0.01
SHISA2	shisa family member 2	chr13	2.16	0.00
BBOX1-AS1	BBOX1 antisense RNA 1	chr11	2.04	0.00
C6orf223	chromosome 6 open reading frame 223	chr6	2	0.00
PRKY	protein kinase, Y-linked, pseudogene	chrY	1.98	0.00
TAGLN	transgelin	chr11	1.95	0.02
PRKY	protein kinase, Y-linked, pseudogene	chrY	1.93	0.00
EYA4	EYA transcriptional coactivator and phosphatase 4	chr6	1.93	0.03
FOXD1	forkhead box D1	chr5	1.91	0.02
CCBE1	collagen and calcium binding EGF domains 1	chr18	1.9	0.00
TPM2	tropomyosin 2 (beta)	chr9	1.88	0.01
XRRA1	X-ray radiation resistance associated 1	chr11	1.84	0.05
MXRA7	matrix-remodelling associated 7	chr2	1.84	0.00
CKMT1A	creatine kinase, mitochondrial 1A	chr15	1.83	0.04
CCDC144B	coiled-coil domain containing 144B (pseudogene)	chr17	1.83	0.02
ALDH1A1	aldehyde dehydrogenase 1 family, member A1	chr9	1.83	0.04
AC006370.2	---	chrY	1.83	0.00

SLC22A3	solute carrier family 22 (organic cation transporter), member 3	chr6	1.82	0.00
MXRA7	matrix-remodelling associated 7	chr2	1.81	0.01
PNPLA3	patatin-like phospholipase domain containing 3	chr22	1.8	0.00
SNORD111	small nucleolar RNA, C/D box 111	chr16	1.78	0.05
SLFN12	schlafen family member 12	chr17	1.78	0.01
ERVK-7	endogenous retrovirus group K, member 7	chr1	1.74	0.00
SNORD14D	small nucleolar RNA, C/D box 14D	chr11	1.72	0.01
LOC729737	uncharacterized LOC729737	chr1	1.71	0.01
LINC01151	long intergenic non-protein coding RNA 1151	chr8	1.71	0.00
CD274	CD274 molecule	chr9	1.69	0.04
INSC	inscuteable homolog (Drosophila)	chr11	1.67	0.00
PLLP	plasmolipin	chr16	1.66	0.00
NAIP	NLR family, apoptosis inhibitory protein	chr5	1.66	0.02
TGM2	transglutaminase 2	chr20	1.62	0.04
POLR2J4	polymerase (RNA) II (DNA directed) polypeptide J4, pseudogene	chr7	1.6	0.00
LINC01186	long intergenic non-protein coding RNA 1186	chrX	1.59	0.00
CCDC144CP	coiled-coil domain containing 144C, pseudogene	chr17	1.59	0.02
MICB	MHC class I polypeptide-related sequence B	chr6_cox_hap2	1.58	0.00
MIR4677	microRNA 4677	chr1	1.57	0.00
GYG2P1	glycogenin 2 pseudogene 1	chrY	1.57	0.00
ZNF697	zinc finger protein 697	chr1	1.56	0.04
LINC01002	long intergenic non-protein coding RNA 1002	chr19	1.56	0.00
LINC01002	long intergenic non-protein coding RNA 1002	chr19	1.56	0.00
FKBP11	FK506 binding protein 11	chr12	1.56	0.00
EFNB2	ephrin-B2	chr13	1.56	0.02
NAV3	neuron navigator 3	chr12	1.55	0.01
38047	membrane associated ring finger 4	chr2	1.55	0.00
MAB21L1	mab-21-like 1 (C. elegans)	chr13	1.55	0.04
INSC	inscuteable homolog (Drosophila)	chr11	1.55	0.00
MICB	MHC class I polypeptide-related sequence B	chr6_mann_hap4	1.54	0.00
MT1X	metallothionein 1X	chr16	1.54	0.00
HEG1	heart development protein with EGF-like domains 1	chr3	1.54	0.01
CTAGE9	CTAGE family, member 9	chr6	1.54	0.01
FAM167A	family with sequence similarity 167, member A	chr8	1.53	0.03
CLDN11	claudin 11	chr3	1.53	0.00
MICB	MHC class I polypeptide-related sequence B	chr6_dbb_hap3	1.52	0.00
MICB	MHC class I polypeptide-related sequence B	chr6_mcf_hap5	1.52	0.00
MXRA7	matrix-remodelling associated 7	chr17	1.52	0.01
LINC01001	long intergenic non-protein coding RNA 1001	chr1	1.52	0.01
AKAP12	A kinase (PRKA) anchor protein 12	chr6	1.52	0.04
FAT1	FAT atypical cadherin 1	chr4	1.51	0.01
CTAGE15	CTAGE family, member 15	chr7	1.51	0.01
AGMAT	agmatinase	chr1	1.51	0.01
HMGCS1	3-hydroxy-3-methylglutaryl-CoA synthase 1 (soluble)	chr5	1.51	0.01
SLC39A8	solute carrier family 39 (zinc transporter), member 8	chr4	1.5	0.01
MICB	MHC class I polypeptide-related sequence B	chr6_apd_hap1	1.5	0.00

Appendix A - 8 Top 100 (of 107) upregulated genes in female NHUC/TERT-NHUC

Symbol	Gne Name	Chr	FC	P-value
XIST	X inactive specific transcript (non-protein coding)	chrX	39.95	0.00
XIST	X inactive specific transcript (non-protein coding)	chrX	34.41	0.00
LINC00960	long intergenic non-protein coding RNA 960	chr3	12.86	0.00
LINC00960	long intergenic non-protein coding RNA 960	chr3	8.26	0.00
GJB6	gap junction protein beta 6	chr13	7.74	0.00
CYP4F11	cytochrome P450, family 4, subfamily F, polypeptide 11	chr19	6.97	0.04
UCA1	urothelial cancer associated 1 (non-protein coding)	chr19	6.44	0.04
PRKAA2	protein kinase, AMP-activated, alpha 2 catalytic subunit	chr1	5.54	0.00
UCA1	urothelial cancer associated 1 (non-protein coding)	chr19	5.04	0.05
LINC00960	long intergenic non-protein coding RNA 960	chr3	4.33	0.00
TMEM47	transmembrane protein 47	chrX	4.2	0.00
LINC01296	long intergenic non-protein coding RNA 1296	chr22	3.45	0.04
HCAR3	hydroxycarboxylic acid receptor 3	chr12	3.39	0.00
NLRP2	NLR family, pyrin domain containing 2	chr19	3.31	0.00
CADM1	cell adhesion molecule 1	chr11	3.3	0.02
PAQR5	progestin and adipoQ receptor family member V	chr15	3.22	0.02
VCAN	versican	chr5	3.08	0.00
GJB6	gap junction protein beta 6	chr13	3.02	0.00
ALOX15B	arachidonate 15-lipoxygenase, type B	chr17	2.66	0.01
HTRA1	HtrA serine peptidase 1	chr10	2.64	0.00
TXNIP	thioredoxin interacting protein	chr1	2.49	0.01
SVIP	small VCP/p97-interacting protein	chr11	2.45	0.02
LOC105372321	uncharacterized LOC105372321	chr19	2.28	0.00
DPY19L2P1	DPY19L2 pseudogene 1	chr7	2.24	0.00
TMOD2	tropomodulin 2 (neuronal)	chr15	2.24	0.00
ASS1	argininosuccinate synthase 1	chr9	2.15	0.00
PLXDC2	plexin domain containing 2	chr10	2.02	0.01
TMPRSS4	transmembrane protease, serine 4	chr11	2.02	0.03
EML1	echinoderm microtubule associated protein like 1	chr14	1.99	0.01
ADGRL4	adhesion G protein-coupled receptor L4	chr1	1.95	0.00
HOXD10	homeobox D10	chr2	1.93	0.00
SPOCK1	sparc/osteonectin, cwcv and kazal-like domains proteoglycan (testican) 1	chr5	1.92	0.01
GIPC2	GIPC PDZ domain containing family, member 2	chr1	1.88	0.00
LOC100132111	uncharacterized LOC100132111	chr1	1.88	0.00
MCTP1	multiple C2 domains, transmembrane 1	chr5	1.86	0.01
ENPP5	ectonucleotide pyrophosphatase/phosphodiesterase 5 (putative)	chr6	1.85	0.05
PRKD1	protein kinase D1	chr14	1.84	0.00
EIF1AX	eukaryotic translation initiation factor 1A, X-linked	chrX	1.84	0.00
EIF2S3	eukaryotic translation initiation factor 2, subunit 3 gamma, 52kDa	chrX	1.81	0.00
DPY19L2P1	DPY19L2 pseudogene 1	chr7	1.79	0.00
KDM5C	lysine (K)-specific demethylase 5C	chrX	1.79	0.00
RIMS2	regulating synaptic membrane exocytosis 2	chr8	1.76	0.00
CORO2B	coronin, actin binding protein, 2B	chr15	1.75	0.01
FAM210B	family with sequence similarity 210, member B	chr20	1.75	0.02
CYP24A1	cytochrome P450, family 24, subfamily A, polypeptide 1	chr20	1.74	0.02
HCAR2	hydroxycarboxylic acid receptor 2	chr12	1.74	0.02
PSAT1P4	phosphoserine aminotransferase 1 pseudogene 4	chr3	1.73	0.02
ZYG11A	zyg-11 family member A, cell cycle regulator	chr1	1.73	0.00
CLU	clusterin	chr8	1.72	0.01
SLCO4C1	solute carrier organic anion transporter family, member 4C1	chr5	1.72	0.01
CAT	catalase	chr11	1.71	0.00
IRAK2	interleukin 1 receptor associated kinase 2	chr3	1.71	0.00
SHC3	SHC (Src homology 2 domain containing) transforming protein 3	chr9	1.71	0.03
EIF2S3	eukaryotic translation initiation factor 2, subunit 3 gamma, 52kDa	chrX	1.7	0.00
LINC00662	long intergenic non-protein coding RNA 662	chr19	1.7	0.02

F2RL1	coagulation factor II (thrombin) receptor-like 1	chr5	1.69	0.00
MRC2	mannose receptor, C type 2	chr17	1.69	0.00
HNRNPLL	heterogeneous nuclear ribonucleoprotein L-like	chr2	1.68	0.02
LAMA1	laminin, alpha 1	chr18	1.68	0.00
LOC100132111	uncharacterized LOC100132111	chr1	1.68	0.00
NPNT	nephronectin	chr4	1.66	0.01
ITGB8	integrin beta 8	chr7	1.65	0.00
SCCPDH	saccharopine dehydrogenase (putative)	chr1	1.65	0.01
DPYSL2	dihydropyrimidinase-like 2	chr8	1.64	0.00
CA5B	carbonic anhydrase VB, mitochondrial	chrX	1.63	0.00
APBA2	amyloid beta (A4) precursor protein-binding, family A, member 2	chr15	1.63	0.00
GPX3	glutathione peroxidase 3	chr5	1.63	0.00
CECR1	cat eye syndrome chromosome region, candidate 1	chr22	1.62	0.00
CPVL	carboxypeptidase, vitellogenic-like	chr7	1.62	0.00
RHOB	ras homolog family member B	chr2	1.61	0.04
KIZ	kizuna centrosomal protein	chr20	1.6	0.00
DDIT4	DNA damage inducible transcript 4	chr10	1.59	0.01
ICA1	islet cell autoantigen 1	chr7	1.59	0.03
MT1E	metallothionein 1E	chr16	1.59	0.04
SMC1A	structural maintenance of chromosomes 1A	chrX	1.59	0.00
C10orf10	chromosome 10 open reading frame 10	chr10	1.58	0.03
DLL1	delta-like 1 (Drosophila)	chr6	1.58	0.02
NTSR1	neurotensin receptor 1 (high affinity)	chr20	1.58	0.01
GAA	glucosidase, alpha	chr17	1.57	0.05
GJB2	gap junction protein beta 2	chr13	1.57	0.01
NET1	neuroepithelial cell transforming 1	chr10	1.57	0.05
ADGRF4	adhesion G protein-coupled receptor F4	chr6	1.56	0.02
LAMA1	laminin, alpha 1	chr18	1.56	0.00
CA5BP1	carbonic anhydrase VB pseudogene 1	chrX	1.55	0.00
DDX3X	DEAD (Asp-Glu-Ala-Asp) box helicase 3, X-linked	chrX	1.55	0.00
ALDH1A3	aldehyde dehydrogenase 1 family, member A3	chr15	1.55	0.01
PARP4	poly(ADP-ribose) polymerase family member 4	chr13	1.55	0.02
PCDHGB5	protocadherin gamma subfamily B, 5	chr5	1.55	0.00
PLAG1	pleiomorphic adenoma gene 1	chr8	1.55	0.01
LOC283922	pyruvate dehydrogenase phosphatase regulatory subunit pseudogene	chr16	1.54	0.01
PCDHGA10	protocadherin gamma subfamily A, 10	chr5	1.54	0.00
PDPR	pyruvate dehydrogenase phosphatase regulatory subunit	chr16	1.54	0.01
USP9X	ubiquitin specific peptidase 9, X-linked	chrX	1.54	0.00
CA5BP1	carbonic anhydrase VB pseudogene 1	chrX	1.53	0.00
ACAA2	acetyl-CoA acyltransferase 2	chr18	1.53	0.00
MAGI2-AS3	MAGI2 antisense RNA 3	chr7	1.53	0.05
SHC3	SHC (Src homology 2 domain containing) transforming protein 3	chr9	1.53	0.03
ULBP1	UL16 binding protein 1	chr6	1.53	0.00
H3F3AP5	H3 histone, family 3A, pseudogene 5	chrX	1.52	0.02
RPS6KA6	ribosomal protein S6 kinase, 90kDa, polypeptide 6	chrX	1.51	0.00
ASNS	asparagine synthetase (glutamine-hydrolyzing)	chr7	1.51	0.01
ITGA4	integrin alpha 4	chr2	1.51	0.03

Appendix A - 9 Female enriched gene-sets from GSEA between male vs female NHU/TERT-NHUC

Gene-Set (female enriched)	Size	Score	p-value	q-value
GO_SPINDLE_POLE	117	-0.54	0.03	0.88
GO_ACTIN_BASED_CELL_PROJECTION	169	-0.42	0.01	0.32
GO_EARLY_ENDOSOME_MEMBRANE	103	-0.30	0.04	0.33
GO_TRANSCRIPTION_FACTOR_COMPLEX	279	-0.34	0.05	0.37
GO_SITE_OF_POLARIZED_GROWTH	136	-0.35	0.02	0.37
GO_BASOLATERAL_PLASMA_MEMBRANE	196	-0.39	0.02	0.34
GO_CELL_BODY	472	-0.27	0.02	0.32
GO_APICAL_PART_OF_CELL	348	-0.27	0.01	0.31
GO_APICAL_PLASMA_MEMBRANE	279	-0.27	0.03	0.31
HALLMARK_PEROXISOME	100	-0.51	0.00	0.39
HALLMARK_FATTY_ACID_METABOLISM	156	-0.40	0.02	0.34
HALLMARK_KRAS_SIGNALING_DN	192	-0.42	0.03	0.28
HALLMARK_UV_RESPONSE_UP	157	-0.42	0.02	0.25
KEGG_OOCYTE_MEIOSIS	106	-0.40	0.04	0.33
REACTOME_METABOLISM_OF_CARBOHYDRATES	222	-0.39	0.01	1.00
REACTOME_SLC_MEDIATED_TRANSMEMBRANE_TRANSPORT	234	-0.31	0.01	0.50
REACTOME_TRANSMEMBRANE_TRANSPORT_OF_SMALL_MOLECULES	397	-0.27	0.02	0.35

Appendix A - 10 Top 100 DE male gene probes (male vs female UHUC)

Symbol	Gene name	Chr	FC	p-value
UTY	ubiquitously transcribed tetratricopeptide repeat containing, Y-linked	chrY	67.35	0.00
TXLNGY	taxilin gamma pseudogene, Y-linked	chrY	47.57	0.00
UTY	ubiquitously transcribed tetratricopeptide repeat containing, Y-linked	chrY	33.25	0.00
RPS4Y1	ribosomal protein S4, Y-linked 1	chrY	26.58	0.00
USP9Y	ubiquitin specific peptidase 9, Y-linked	chrY	24.62	0.00
TXLNGY	taxilin gamma pseudogene, Y-linked	chrY	19.58	0.00
DDX3Y	DEAD (Asp-Glu-Ala-Asp) box helicase 3, Y-linked	chrY	17.12	0.00
TTTY15	testis-specific transcript, Y-linked 15 (non-protein coding)	chrY	13.49	0.00
LINC00278	long intergenic non-protein coding RNA 278	chrY	11.90	0.00
ZFY	zinc finger protein, Y-linked	chrY	10.51	0.00
NLGN4Y	neuroligin 4, Y-linked	chrY	9.76	0.00
SNORD91B	small nucleolar RNA, C/D box 91B	chr17	9.43	0.02
GPNMB	glycoprotein (transmembrane) nmb	chr7	9.01	0.01
KDM5D	lysine (K)-specific demethylase 5D	chrY	8.97	0.00
TBX2	T-box 2	chr17	6.23	0.01
EIF1AY	eukaryotic translation initiation factor 1A, Y-linked	chrY	5.80	0.00
SNORD116-27	small nucleolar RNA, C/D box 116-27	chr15	5.10	0.05
PRKY	protein kinase, Y-linked, pseudogene	chrY	4.94	0.00
CAPNS2	calpain, small subunit 2	chr16	4.73	0.04
NLGN4Y	neuroligin 4, Y-linked	chrY	4.09	0.00
SNORA22	small nucleolar RNA, H/ACA box 22	chr7	4.09	0.05
SNORD93	small nucleolar RNA, C/D box 93	chr7	4.04	0.02
PRKY	protein kinase, Y-linked, pseudogene	chrY	3.87	0.00
UGT2B7	UDP glucuronosyltransferase 2 family, polypeptide B7	chr4	3.74	0.00
ZFY	zinc finger protein, Y-linked	chrY	3.72	0.00
SNORD115-15	small nucleolar RNA, C/D box 115-15	chr15	3.66	0.01
TAC3	tachykinin 3	chr12	3.54	0.04
SNORA59B	small nucleolar RNA, H/ACA box 59B	chr1	3.47	0.04
SNORA59B	small nucleolar RNA, H/ACA box 59B	chr17	3.47	0.04
CEL	carboxyl ester lipase	chr9	3.38	0.05
PYGO1	pygopus family PHD finger 1	chr15	3.31	0.00
SNORD115-21	small nucleolar RNA, C/D box 115-21	chr15	3.30	0.01
SNORD115-32	small nucleolar RNA, C/D box 115-32	chr15	3.28	0.01
SNORD115-6	small nucleolar RNA, C/D box 115-6	chr15	3.17	0.01
SNORD115-15	small nucleolar RNA, C/D box 115-15	chr15	3.14	0.03
LOC100129046	uncharacterized LOC100129046	chr1	3.10	0.01
SNORD105B	small nucleolar RNA, C/D box 105B	chr19	3.04	0.01
RPS4Y2	ribosomal protein S4, Y-linked 2	chrY	3.03	0.00
SNORD115-42	small nucleolar RNA, C/D box 115-42	chr15	3.02	0.01
SNORD61	small nucleolar RNA, C/D box 61	chrX	2.98	0.02
TTTY14	testis-specific transcript, Y-linked 14 (non-protein coding)	chrY	2.93	0.00
TBX2	T-box 2	chr17	2.88	0.00
LOC100129046	uncharacterized LOC100129046	chr1	2.85	0.01
LONRF2	LON peptidase N-terminal domain and ring finger 2	chr2	2.79	0.00
SNORD99	small nucleolar RNA, C/D box 99	chr1	2.79	0.01
GYG2P1	glycogenin 2 pseudogene 1	chrY	2.77	0.00
AC006370.2	---	chrY	2.76	0.00
SNORD115-10	small nucleolar RNA, C/D box 115-10	chr15	2.75	0.03
LOC389906	zinc finger protein 839 pseudogene	chrX	2.71	0.04
SNORD115-11	small nucleolar RNA, C/D box 115-11	chr15	2.64	0.05
SNORD115-29	small nucleolar RNA, C/D box 115-29	chr15	2.64	0.05
SNORD115-43	small nucleolar RNA, C/D box 115-43	chr15	2.64	0.05
SNORD115-43	small nucleolar RNA, C/D box 115-43	chr15	2.64	0.05
SNORD59A	small nucleolar RNA, C/D box 59A	chr12	2.54	0.05
GYG2P1	glycogenin 2 pseudogene 1	chrY	2.53	0.00
PER3	period circadian clock 3	chr1	2.52	0.03
SNORD116-14	small nucleolar RNA, C/D box 116-14	chr15	2.51	0.03
C14orf105	chromosome 14 open reading frame 105	chr14	2.44	0.02
GLDC	glycine dehydrogenase (decarboxylating)	chr9	2.43	0.02
CPNE1	copine I	chr20	2.41	0.01

KLHDC7A	kelch domain containing 7A	chr1	2.40	0.01
LINC01296	long intergenic non-protein coding RNA 1296	chr22	2.39	0.05
MIR944	microRNA 944	chr3	2.37	0.04
SNORD116-26	small nucleolar RNA, C/D box 116-26	chr15	2.36	0.03
MMRN2	multimerin 2	chr10	2.34	0.00
SNORD116-18	small nucleolar RNA, C/D box 116-18	chr15	2.33	0.02
RNY4	RNA, Ro-associated Y4	chr6	2.32	0.04
SNORD115-1	small nucleolar RNA, C/D box 115-1	chr15	2.32	0.04
TRPC6	transient receptor potential cation channel, subfamily C, member 6	chr11	2.32	0.01
PYGO1	pygopus family PHD finger 1	chr15	2.30	0.01
LINC01296	long intergenic non-protein coding RNA 1296	chr14	2.29	0.04
SNORD116-16	small nucleolar RNA, C/D box 116-16	chr15	2.29	0.02
FOXQ1	forkhead box Q1	chr6	2.18	0.02
LOC101929378	uncharacterized LOC101929378	chr2	2.18	0.01
FAM118A	family with sequence similarity 118, member A	chr22	2.16	0.03
SNORD20	small nucleolar RNA, C/D box 20	chr2	2.16	0.01
SNORD115-16	small nucleolar RNA, C/D box 115-16	chr15	2.14	0.04
LOC101929378	uncharacterized LOC101929378	chr2	2.13	0.02
TMSB4Y	thymosin beta 4, Y-linked	chrY	2.13	0.00
ELF5	E74-like factor 5 (ets domain transcription factor)	chr11	2.12	0.03
SCD	stearoyl-CoA desaturase (delta-9-desaturase)	chr10	2.09	0.03
CHMP1B2P	charged multivesicular body protein 1B2, pseudogene	chrX	2.08	0.04
FAHD2CP	fumarylacetoacetate hydrolase domain containing 2C, pseudogene	chr2	2.08	0.03
FASN	fatty acid synthase	chr17	2.08	0.00
SCNN1B	sodium channel, non voltage gated 1 beta subunit	chr16	2.07	0.02
SNORD105	small nucleolar RNA, C/D box 105	chr19	2.07	0.02
CELP	carboxyl ester lipase pseudogene	chr9	2.05	0.03
CELP	carboxyl ester lipase pseudogene	chr9	2.02	0.04
HIST1H2AB	histone cluster 1, H2ab	chr6	2.02	0.01
PTPRU	protein tyrosine phosphatase, receptor type, U	chr1	2.02	0.04
SNORD116-17	small nucleolar RNA, C/D box 116-17	chr15	2.01	0.03
GATM	glycine amidinotransferase (L-arginine:glycine amidinotransferase)	chr15	1.99	0.00
BAMBI	BMP and activin membrane-bound inhibitor	chr10	1.98	0.03
MYEF2	myelin expression factor 2	chr15	1.98	0.01
DUXAP10	double homeobox A pseudogene 10	chr14	1.97	0.04
SCNN1B	sodium channel, non voltage gated 1 beta subunit	chr16	1.96	0.01
NOB1	NIN1/RPN12 binding protein 1 homolog	chr16	1.95	0.03
MTRNR2L7	MT-RNR2-like 7	chr10	1.92	0.01
SDK1	sidekick cell adhesion molecule 1	chr7	1.91	0.02
SCARNA23	small Cajal body-specific RNA 23	chrX	1.90	0.04

Appendix A - 11 Top 100 DE female gene probes (male vs female UHUC)

Symbol	Gene name	Chr	FC	p-value
XIST	X inactive specific transcript (non-protein coding)	chrX	2568	0.00
XIST	X inactive specific transcript (non-protein coding)	chrX	1659	0.00
HLA-DQA1	major histocompatibility complex, class II, DQ alpha 1	chr6	128	0.00
HLA-DQA1	major histocompatibility complex, class II, DQ alpha 1	chr6	128	0.00
RGS1	regulator of G-protein signaling 1	chr1	14.82	0.00
LYZ	lysozyme	chr12	14.06	0.00
GPR183	G protein-coupled receptor 183	chr13	12.76	0.00
SGK1	serum/glucocorticoid regulated kinase 1	chr6	12.30	0.03
CD74	CD74 molecule, major histocompatibility complex, class II invariant chain	chr5	11.43	0.00
FGL2	fibrinogen-like 2	chr7	11.16	0.00
HLA-DRB1	major histocompatibility complex, class II, DR beta 1	chr6	11.08	0.00
HLA-DRB1	major histocompatibility complex, class II, DR beta 1	chr6	11.08	0.00
HLA-DRB1	major histocompatibility complex, class II, DR beta 1	chr6	11.08	0.00
HLA-DRB1	major histocompatibility complex, class II, DR beta 1	chr6	11.00	0.00
SRGN	serglycin	chr10	8.91	0.00
FGL2	fibrinogen-like 2	chr7	8.53	0.00
HLA-DRB1	major histocompatibility complex, class II, DR beta 1	chr6	7.39	0.00
HLA-DRB1	major histocompatibility complex, class II, DR beta 1	chr6	7.23	0.00
C8orf4	chromosome 8 open reading frame 4	chr8	6.93	0.03
SRGN	serglycin	chr10	6.79	0.00
PFKFB3	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 3	chr10	6.74	0.05
HLA-DRB4	major histocompatibility complex, class II, DR beta 4	chr6	6.32	0.03
CXCR4	chemokine (C-X-C motif) receptor 4	chr2	6.22	0.00
HLA-DRA	major histocompatibility complex, class II, DR alpha	chr6	6.21	0.00
TYROBP	TYRO protein tyrosine kinase binding protein	chr19	6.14	0.00
FCGR2C	Fc fragment of IgG, low affinity IIc, receptor for (CD32) (gene/pseudogene)	chr1	6.11	0.00
HLA-DRA	major histocompatibility complex, class II, DR alpha	chr6	6.09	0.00
HLA-DRA	major histocompatibility complex, class II, DR alpha	chr6	6.09	0.00
HLA-DRA	major histocompatibility complex, class II, DR alpha	chr6	6.09	0.00
HLA-DRA	major histocompatibility complex, class II, DR alpha	chr6	6.06	0.00
HLA-DRA	major histocompatibility complex, class II, DR alpha	chr6	6.05	0.00
NLRP2	NLR family, pyrin domain containing 2	chr19	5.99	0.04
LAPTM5	lysosomal protein transmembrane 5	chr1	5.87	0.00
CXCL8	chemokine (C-X-C motif) ligand 8	chr4	5.83	0.02
HLA-DRB1	major histocompatibility complex, class II, DR beta 1	chr6	5.72	0.00
TM4SF1	transmembrane 4 L six family member 1	chr3	5.66	0.05
CYP24A1	cytochrome P450, family 24, subfamily A, polypeptide 1	chr20	5.19	0.00
RGS2	regulator of G-protein signaling 2	chr1	5.06	0.01
TYROBP	TYRO protein tyrosine kinase binding protein	chr19	5.06	0.00
HLA-DRA	major histocompatibility complex, class II, DR alpha	chr6	4.92	0.00
OLR1	oxidized low density lipoprotein (lectin-like) receptor 1	chr12	4.82	0.00
VTCN1	V-set domain containing T cell activation inhibitor 1	chr1	4.75	0.01
MPEG1	macrophage expressed 1	chr11	4.39	0.00
EVI2B	ecotropic viral integration site 2B	chr17	4.33	0.00
HLA-DRB1	major histocompatibility complex, class II, DR beta 1	chr6	4.32	0.00
CD69	CD69 molecule	chr12	4.31	0.00
HLA-DRB1	major histocompatibility complex, class II, DR beta 1	chr6	4.23	0.00
HLA-DRB4	major histocompatibility complex, class II, DR beta 4	chr6	4.23	0.05
CYBB	cytochrome b-245, beta polypeptide	chrX	4.01	0.00
CD69	CD69 molecule	chr12	3.94	0.01
NCCRP1	non-specific cytotoxic cell receptor protein 1 homolog (zebrafish)	chr19	3.91	0.01
SAMSN1	SAM domain, SH3 domain and nuclear localization signals 1	chr21	3.90	0.00
CLCA4	chloride channel accessory 4	chr1	3.82	0.04
FCER1A	Fc fragment of IgE, high affinity I, receptor for	chr1	3.75	0.01
FCGR2A	Fc fragment of IgG, low affinity IIa, receptor (CD32)	chr1	3.75	0.00
HLA-DQA1	major histocompatibility complex, class II, DQ alpha 1	chr6	3.67	0.04
BIRC3	baculoviral IAP repeat containing 3	chr11	3.64	0.03
AGR3	anterior gradient 3, protein disulphide isomerase family member	chr7	3.60	0.01

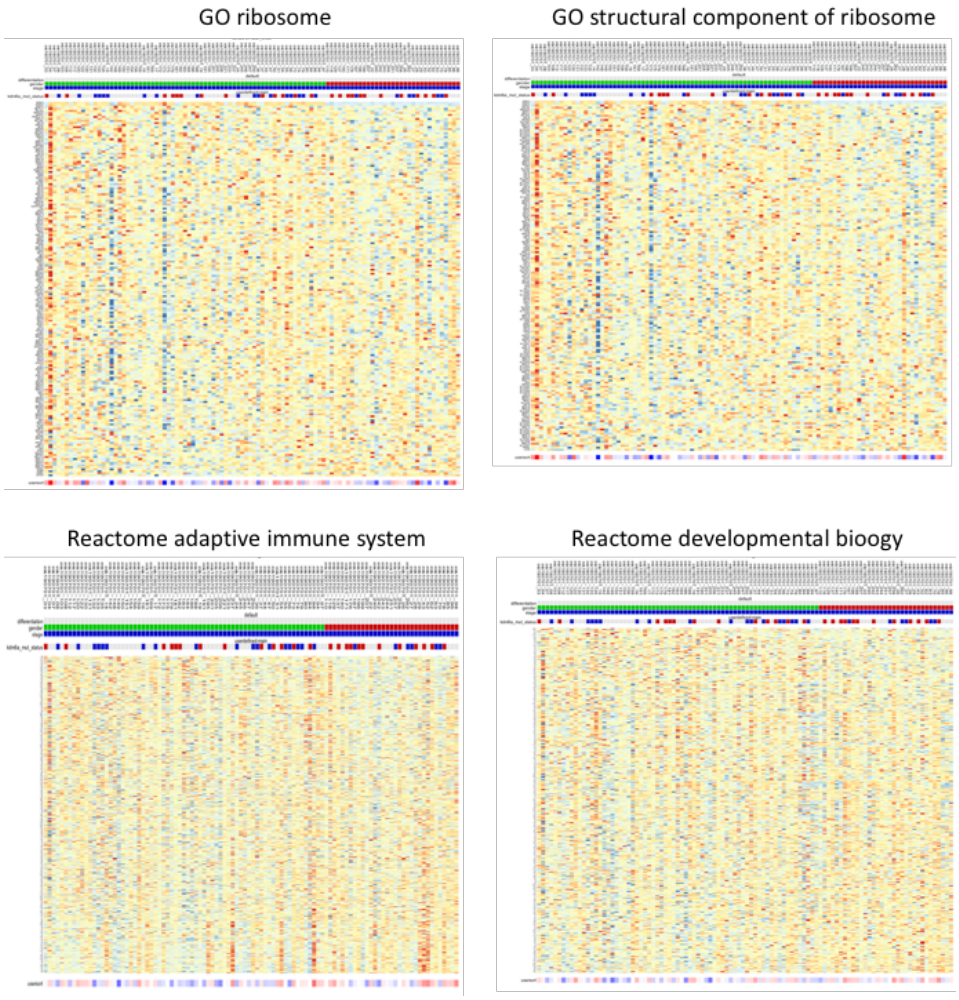
PTPRC	protein tyrosine phosphatase, receptor type, C	chr1	3.56	0.00
HLA-DQA2	major histocompatibility complex, class II, DQ alpha 2	chr6	3.54	0.00
CD74	CD74 molecule, major histocompatibility complex, class II invariant chain	chr5	3.51	0.00
CLCA4	chloride channel accessory 4	chr1	3.51	0.04
HLA-B	major histocompatibility complex, class I, B	chr6	3.51	0.00
EYA4	EYA transcriptional coactivator and phosphatase 4	chr6	3.49	0.02
TFPI	tissue factor pathway inhibitor (lipoprotein-associated coagulation inhibitor)	chr2	3.49	0.01
VTCN1	V-set domain containing T cell activation inhibitor 1	chr1	3.46	0.02
PIGR	polymeric immunoglobulin receptor	chr1	3.44	0.01
TNFAIP3	tumor necrosis factor, alpha-induced protein 3	chr6	3.44	0.02
SPRR3	small proline-rich protein 3	chr1	3.43	0.01
MS4A6A	membrane-spanning 4-domains, subfamily A, member 6A	chr11	3.40	0.00
HLA-DMA	major histocompatibility complex, class II, DM alpha	chr6	3.38	0.00
HLA-DMA	major histocompatibility complex, class II, DM alpha	chr6	3.36	0.00
A2M	alpha-2-macroglobulin	chr12	3.32	0.00
C1QC	complement component 1, q subcomponent, C chain	chr1	3.32	0.00
IL7R	interleukin 7 receptor	chr5	3.29	0.03
DAB2	Dab, mitogen-responsive phosphoprotein, homolog 2 (Drosophila)	chr5	3.28	0.02
CD53	CD53 molecule	chr1	3.26	0.00
IFITM2	interferon induced transmembrane protein 2	chr11	3.26	0.02
HLA-DPA1	major histocompatibility complex, class II, DP alpha 1	chr6	3.23	0.02
RNASE6	ribonuclease, RNase A family, k6	chr14	3.23	0.01
MIR4802	microRNA 4802	chr4	3.20	0.03
HLA-DPA1	major histocompatibility complex, class II, DP alpha 1	chr6	3.18	0.02
HLA-DPA1	major histocompatibility complex, class II, DP alpha 1	chr6	3.18	0.02
HLA-DPA1	major histocompatibility complex, class II, DP alpha 1	chr6	3.18	0.02
CD163	CD163 molecule	chr12	3.17	0.02
EGLN3	egl-9 family hypoxia-inducible factor 3	chr14	3.11	0.00
HLA-DRB3	major histocompatibility complex, class II, DR beta 3	chr6	3.10	0.00
CCL4	chemokine (C-C motif) ligand 4	chr17	3.09	0.00
MUC1	mucin 1, cell surface associated	chr1	3.09	0.03
TMC5	transmembrane channel like 5	chr16	3.08	0.04
C3	complement component 3	chr19	3.07	0.00
HLA-DPA1	major histocompatibility complex, class II, DP alpha 1	chr6	3.07	0.02
HLA-DPA1	major histocompatibility complex, class II, DP alpha 1	chr6	3.07	0.02
HLA-DRB6	major histocompatibility complex, class II, DR beta 6 (pseudogene)	chr6	3.06	0.00
TNFAIP3	tumor necrosis factor, alpha-induced protein 3	chr6	3.06	0.03
DOCK10	dedicator of cytokinesis 10	chr2	3.04	0.00
AOAH	acyloxyacyl hydrolase (neutrophil)	chr7	3.01	0.00
MUC1	mucin 1, cell surface associated	chr1	3.01	0.01
HLA-DPA1	major histocompatibility complex, class II, DP alpha 1	chr6	2.99	0.02
HLA-DPA1	major histocompatibility complex, class II, DP alpha 1	chr6	2.99	0.02

Appendix A - 12 Upregulated male gene probes (male vs female TaG2)

Symbol	Gene name	Chr	FC	p-value
FRG1BP	FSHD region gene 1 family member B, pseudogene	chr20	-1.5	0.00
UTY	ubiquitously transcribed tetratricopeptide repeat containing, Y-linked	chrY	-27.63	0.00
RPS4Y1	ribosomal protein S4, Y-linked 1	chrY	-20.51	0.00
LOC283788	FSHD region gene 1 pseudogene	chrUn	-1.53	0.00
TMEM97	transmembrane protein 97	chr17	-1.57	0.04
USP9Y	ubiquitin specific peptidase 9, Y-linked	chrY	-18.52	0.00
ZFY	zinc finger protein, Y-linked	chrY	-11.54	0.00
UTY	ubiquitously transcribed tetratricopeptide repeat containing, Y-linked	chrY	-11.38	0.00
DHCR24	24-dehydrocholesterol reductase	chr1	-2.14	0.04
LINC00278	long intergenic non-protein coding RNA 278	chrY	-10.49	0.00
DAB1	Dab, reelin signal transducer, homolog 1 (Drosophila)	chr1	-2.48	0.05
STAG2	stromal antigen 2	chrX	-2.6	0.01
TXLNGY	taxilin gamma pseudogene, Y-linked	chrY	-9.62	0.00
DDX3Y	DEAD (Asp-Glu-Ala-Asp) box helicase 3, Y-linked	chrY	-8.71	0.00
STAG2	stromal antigen 2	chrX	-3.38	0.01
TTTY15	testis-specific transcript, Y-linked 15 (non-protein coding)	chrY	-8.57	0.00
TXLNGY	taxilin gamma pseudogene, Y-linked	chrY	-8.5	0.00
EIF1AY	eukaryotic translation initiation factor 1A, Y-linked	chrY	-8.26	0.00
KDM5D	lysine (K)-specific demethylase 5D	chrY	-5.54	0.00
NLGN4Y	neuroligin 4, Y-linked	chrY	-4.94	0.00
ZFY	zinc finger protein, Y-linked	chrY	-4.28	0.00
RPS4Y2	ribosomal protein S4, Y-linked 2	chrY	-2.91	0.00
NLGN4Y	neuroligin 4, Y-linked	chrY	-2.84	0.00
PRKY	protein kinase, Y-linked, pseudogene	chrY	-2.14	0.00
PRKY	protein kinase, Y-linked, pseudogene	chrY	-2.03	0.00
AC006370.2	---	chrY	-1.89	0.00
TTTY14	testis-specific transcript, Y-linked 14 (non-protein coding)	chrY	-1.62	0.00
GYG2P1	glycogenin 2 pseudogene 1	chrY	-1.51	0.00
TMSB4Y	thymosin beta 4, Y-linked	chrY	-1.51	0.00

Appendix A - 13 Upregulated female gene probes (male vs female TaG2)

Symbol	Gene name	Chr	FC	p-value
XIST	X inactive specific transcript (non-protein coding)	chrX	502.5	0.00
XIST	X inactive specific transcript (non-protein coding)	chrX	337.44	0.00
H19	H19, imprinted maternally expressed transcript (non-protein coding)	chr11	6.43	0.03
H19	H19, imprinted maternally expressed transcript (non-protein coding)	chr11	6.41	0.04
NLRP2	NLR family, pyrin domain containing 2	chr19	6.14	0.01
JCHAIN	joining chain of multimeric IgA and IgM	chr4	3.01	0.04
NCCRP1	non-specific cytotoxic cell receptor protein 1 homolog (zebrafish)	chr19	2.71	0.03
GLDC	glycine dehydrogenase (decarboxylating)	chr9	2.55	0.02
CYP24A1	cytochrome P450, family 24, subfamily A, polypeptide 1	chr20	2.44	0.01
PAX8	paired box 8	chr2	2.4	0.00
TRPA1	transient receptor potential cation channel, subfamily A, member 1	chr8	2.3	0.00
KRT5	keratin 5, type II	chr12	2.3	0.04
HS6ST3	heparan sulphate 6-O-sulfotransferase 3	chr13	2.18	0.04
IGHV3-23	immunoglobulin heavy variable 3-23	chr14	2.06	0.02
PTPRS	protein tyrosine phosphatase, receptor type, S	chr19	2.03	0.00
LOC389906	zinc finger protein 839 pseudogene	chrX	1.94	0.01
LOC389906	zinc finger protein 839 pseudogene	chrX	1.89	0.01
BMP5	bone morphogenetic protein 5	chr6	1.86	0.04
COL6A3	collagen, type VI, alpha 3	chr2	1.81	0.01
AQP3	aquaporin 3 (Gill blood group)	chr9	1.79	0.01
EFNB2	ephrin-B2	chr13	1.79	0.01
PTPRS	protein tyrosine phosphatase, receptor type, S	chr19	1.79	0.00
STS	steroid sulfatase (microsomal), isozyme S	chrX	1.78	0.00
CYP1B1	cytochrome P450, family 1, subfamily B, polypeptide 1	chr2	1.76	0.03
MUC2	mucin 2, oligomeric mucus/gel-forming	chr11	1.76	0.02
MOGAT2	monoacylglycerol O-acyltransferase 2	chr11	1.75	0.01
PTPRS	protein tyrosine phosphatase, receptor type, S	chr19	1.74	0.00
ZDBF2	zinc finger, DBF-type containing 2	chr2	1.72	0.01
STS	steroid sulfatase (microsomal), isozyme S	chrX	1.69	0.00
ABCC4	ATP binding cassette subfamily C member 4	chr13	1.67	0.00
MXRA5	matrix-remodelling associated 5	chrX	1.66	0.02
MUC2	mucin 2, oligomeric mucus/gel-forming	chr11	1.66	0.04
ARSD	arylsulfatase D	chrX	1.65	0.00
PNPLA4	patatin-like phospholipase domain containing 4	chrX	1.62	0.00
UCP2	uncoupling protein 2 (mitochondrial, proton carrier)	chr11	1.61	0.04
LOC389906	zinc finger protein 839 pseudogene	chrX	1.59	0.00
GPR183	G protein-coupled receptor 183	chr13	1.56	0.01
LOXL1	lysyl oxidase-like 1	chr15	1.56	0.02
LOC389906	zinc finger protein 839 pseudogene	chrX	1.55	0.00
DDR2	discoidin domain receptor tyrosine kinase 2	chr1	1.53	0.05
PUDP	pseudouridine 5-phosphatase	chrX	1.51	0.00
IGHA1	immunoglobulin heavy constant alpha 1	chr14	1.51	0.04
A2M	alpha-2-macroglobulin	chr12	1.5	0.04

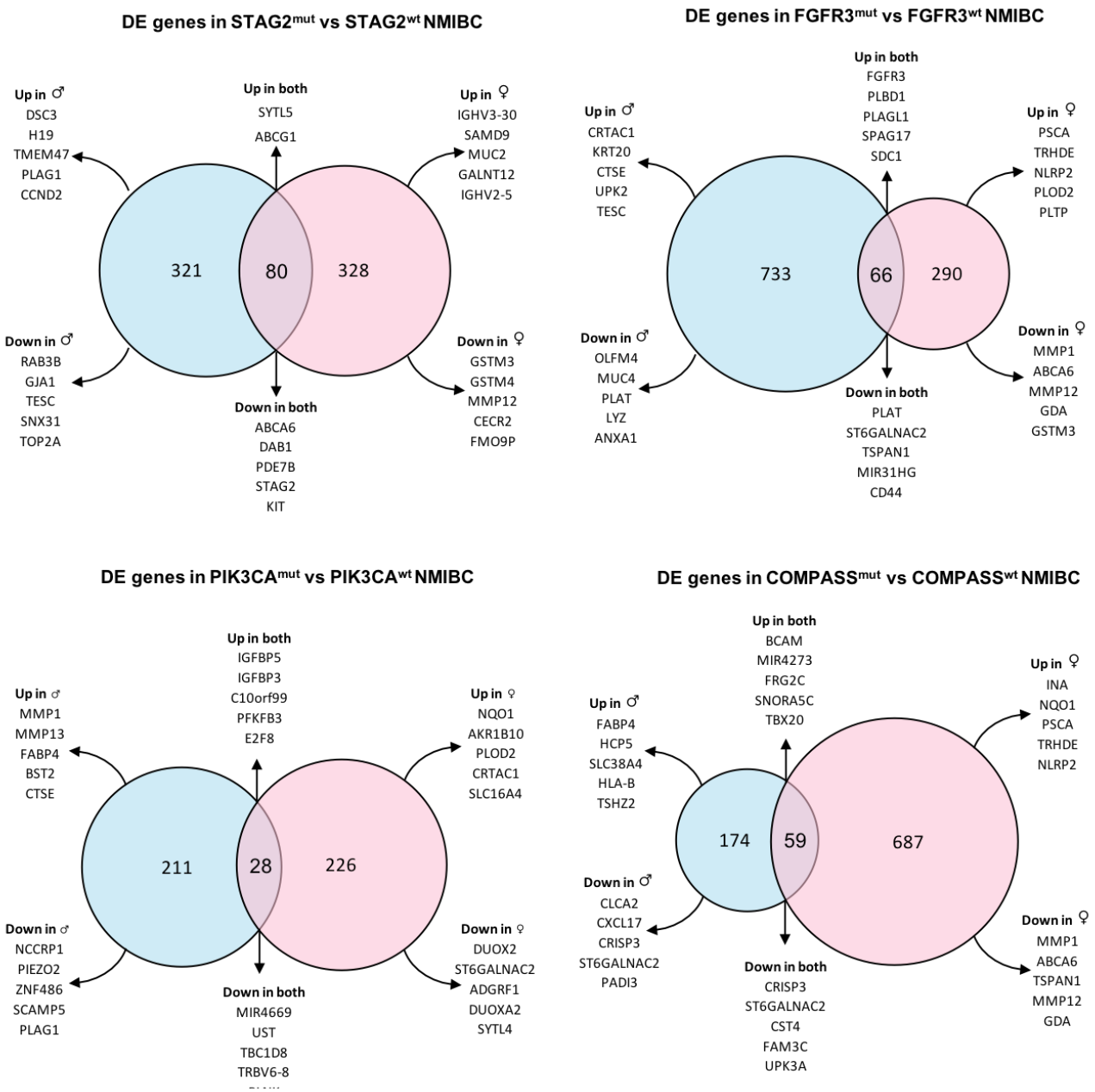


Appendix A - 14 Heatmaps in TaG2 for using GSEA enriched gene lists. Samples are ordered male to female. Supplementary to Fig 3.11.

	Transcription Factor	Actual	DE targerts of TF	n	R	N	Expected	Ratio	p-value	z-score
Male TERT-NHUC	GATA-6	7	Claudin-11, PDEF, PIB4, Tenascin-C, DPP4, DKK-1,SARG	202	201	40313	1.007	6.95	0.000	6.001
Male TERT-NHUC/NHUC	Myocardin	1	Transgelin	83	5	40313	0.01029	97.14	0.010	9.765
Male UHUC	SREBP2 (nuclear)	4	FASN, SCD, DHCR7, SREBP2 precursor	196	71	40313	0.3452	11.59	0.000	6.241
Female UHUC	E4BP4	5	EPSTI1, GIMAP7, PORIMIN, THAS, TIM-3	539	66	40313	0.8824	5.666	0.002	4.416
	BLIMP1 (PRDI-BF1)	9	IRF8, HLA-DMA, CXCR6, IFNGR1, CD30L, TLR3, IL10RA, EHMT1, TAP2	539	156	40313	2.086	4.315	0.000	4.829
	RUNX3	7	P2Y14, PORIMIN, FAM105A, RUNX, microRNA 29b-1, microRNA 18S, ITGA4	539	158	40313	2.113	3.314	0.006	3.392

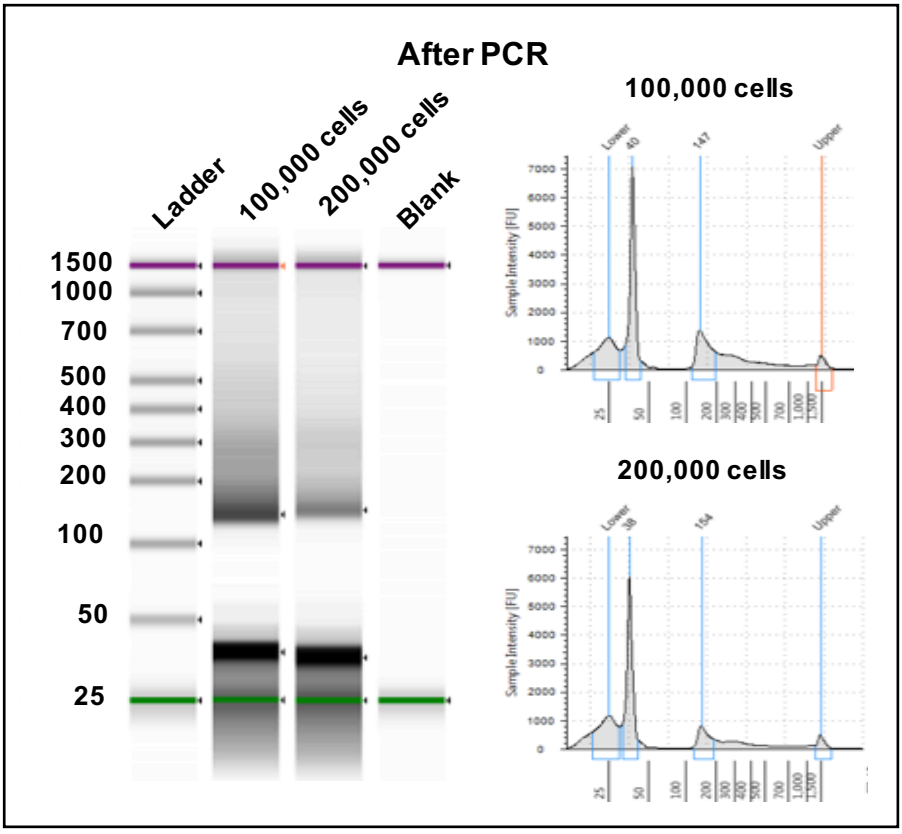
Actual	number of objects in DE gene list regulated by the TF
n	size of DE gene list
R	number of targets regulated by TF
N	number of targets regulated by TF
Expected	mean value for hypergeometric distribution (n*R/N)
Ratio	connectivity ratio(Actual/Expected)
z-score	z-score ((Actual-Expected)/sqrt(variance))
p-value	probability to have the given value of Actual or higher (or lover for negative z-score)

Appendix A - 16 Metacore analysis of enriched transcription factors in DE gene lists determined by LIMMA

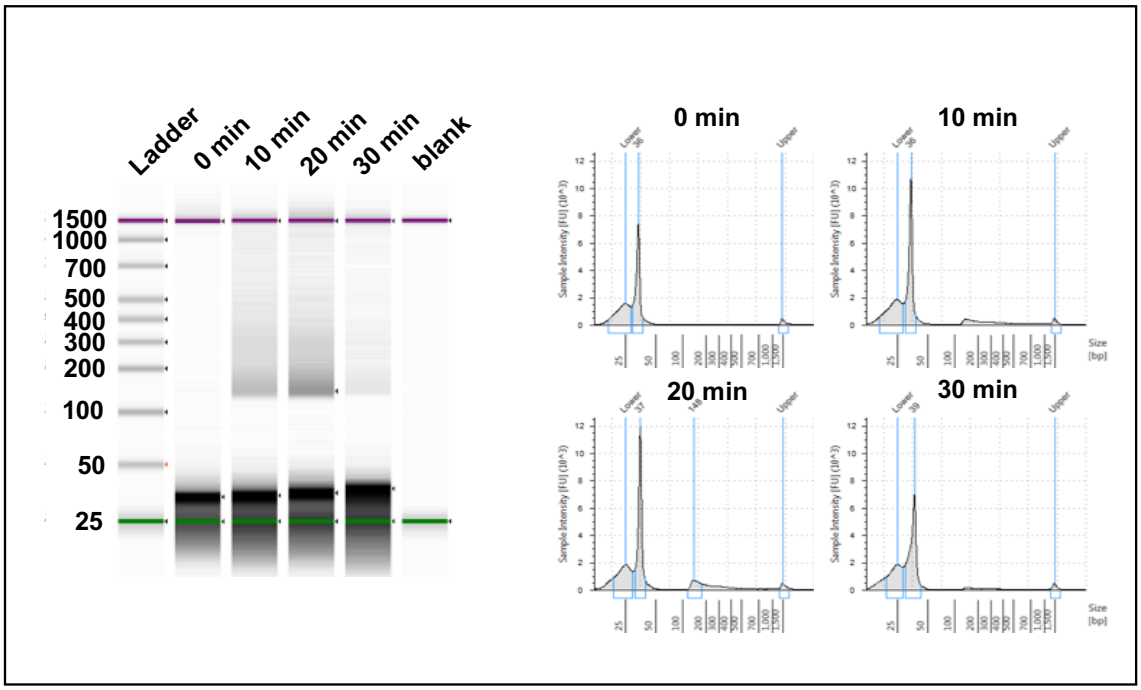


Appendix A - 17 Differential expression analysis for commonly mutated genes in male and female NMIBC.

Appendix B ATAC-seq optimisation



Appendix B - 1 TapeStation results following PCR amplification of for ATAC libraries (without size selection) with 100,000 and 200,000 C-TERT cells



Appendix B - 2 Time course transposition assay on C-TERT cells. TapeStation carried out directly after PCR amplification (without size selection).

Appendix C Codes used for analysis of ATAC-seq data

```
#!/bin/bash
# OPTIONS FOR GRID ENGINE ===================================================================
#
#$ -S /bin/bash
#$ -l h_rt=48:00:00
#$ -l h_vmem=16G
#$ -pe smp 4
#$ -cwd
#$ -j y
#$ -V
# OPTIONS FOR GRID ENGINE=====================================================================
echo "Running on `hostname`"

cd ..
mkdir SAMfiles
mkdir BAMfiles
mkdir BigWigs
mkdir MACS
mkdir cutadapt
mkdir fastqc

# Perform quality control
echo "Running fastqc"
cd fastqc
mkdir pre-trimming
cd pre-trimming
fastqc /nobackup/umbch/ATAC_raw_fastq/B_Tert_Ad2_1_S1_R1.fastq.gz -o .
fastqc /nobackup/umbch/ATAC_raw_fastq/B_Tert_Ad2_1_S1_R2.fastq.gz -o .

#Trim adaptors with cutadapt then QC with FastQC
echo "Running on `hostname`"
cd ../../cutadapt
echo "Running cutadapt"
cutadapt -a CTGTCTCTTATACACATCT -A CTGTCTCTTATACACATCT -q 30 --minimum-length 36 -o
B_Tert_Ad2_1_S1_R1_cutadapt.fastq.gz -p B_Tert_Ad2_1_S1_R2_cutadapt.fastq.gz
/nobackup/umbch/ATAC_raw_fastq/B_Tert_Ad2_1_S1_R1.fastq.gz
/nobackup/umbch/ATAC_raw_fastq/B_Tert_Ad2_1_S1_R2.fastq.gz
echo "Running fastqc"
cd ../fastqc
mkdir post-trimming
cd post-trimming
fastqc B_Tert_Ad2_1_S1_R1_cutadapt.fastq.gz -o .
fastqc B_Tert_Ad2_1_S1_R2_cutadapt.fastq.gz -o .
cd ../../

#Align to Genome
echo "Align to genome"
bowtie2 --threads 4 -x /nobackup/umbch/reference-genomes/GRCh37_26-bowtie2/GRCh37_26 -X 2000 -t --very-sensitive -1
/nobackup/umbch/postCRUK/cutadapt/quality20/B_Tert_Ad2_1_S1_R1_cutadapt.fastq.gz -2
/nobackup/umbch/postCRUK/cutadapt/quality20/B_Tert_Ad2_1_S1_R2_cutadapt.fastq.gz -S
SAMfiles/B_Tert_Ad2_1_ATAC_CRUK.sam
echo "Creating BAM file"
samtools view --threads 4 -q 30 -bt /nobackup/umbch/reference-genomes/GRCh37_26-bowtie2/GRCh37_26 -o
BAMfiles/B_Tert_Ad2_1_ATAC_CRUK.nonSorted.bam SAMfiles/B_Tert_Ad2_1_ATAC_CRUK.sam
echo "Sorting BAM file"
samtools sort --threads 4 BAMfiles/B_Tert_Ad2_1_ATAC_CRUK.nonSorted.bam -o BAMfiles/B_Tert_Ad2_1_ATAC_CRUK.bam
rm BAMfiles/B_Tert_Ad2_1_ATAC_CRUK.nonSorted.bam
echo "Create BAM index"
samtools index -@ 4 BAMfiles/B_Tert_Ad2_1_ATAC_CRUK.bam

# Remove duplicates
export _JAVA_OPTIONS=-Xmx16000M
picard MarkDuplicates M=B_Tert_Ad2_1_ATAC_dupstats.txt REMOVE_DUPLICATES=TRUE I=BAMfiles/B_Tert_Ad2_1_ATAC.bam
O=BAMfiles/B_Tert_Ad2_1_ATAC_NODUPS.nonSorted.bam TMP_DIR=$TMPDIR
echo "Sorting BAM file"
samtools sort --threads 4 BAMfiles/B_Tert_Ad2_1_ATAC_NODUPS.nonSorted.bam -o
BAMfiles/B_Tert_Ad2_1_ATAC_NODUPS.bam
rm BAMfiles/B_Tert_Ad2_1_ATAC_NODUPS.nonSorted.bam
echo "Create BAM index"
samtools index -@ 4 BAMfiles/B_Tert_Ad2_1_ATAC_NODUPS.bam
```

```
samtools idxstats /nobackup/umbch/postCRUK/BAMfiles/B_Tert_Ad2_1_ATAC_NODUPS_CRUK.bam|  cut -f 1 | grep -v MT |
xargs samtools view -b /nobackup/umbch/postCRUK/BAMfiles/B_Tert_Ad2_1_ATAC_NODUPS_CRUK.bam  >
B_Tert_Ad2_1_ATAC_NODUPS_NOMT_CRUK.bam
echo "Create BAM index"
samtools index -@ 4 BAMfiles/B_Tert_Ad2_1_ATAC_NODUPS.bam

# Create Coverage using deepTools
cd ../BigWigs
bamCoverage --binSize 10 -b BAMfiles/B_Tert_Ad2_1_ATAC_NODUPS_CRUK.bam --normalizeUsingRPKM  -o
BigWigs/B_Tert_Ad2_1_ATAC_NODUPS_CRUK.bw

#Peak calling
echo "Cut site Peak calling using MACS"
cd ../MACS
macs2 callpeak -g hs --nomodel --shift -100 --extsize 200 -f BAM -t ../BAMfiles/B_Tert_Ad2_1_ATAC_NODUPS_NOMT_CRUK.bam
-n B_Tert_Ad2_1_ATAC_CutSite_MACS2_NOMT_CRUK
echo "Nucleosome Peak calling using MACS"
cd ../MACS
macs2 callpeak -g hs --nomodel --shift -37 --extsize 73 -f BAM -t ../BAMfiles/B_Tert_Ad2_1_ATAC_NODUPS_NOMT_CRUK.bam -n
B_Tert_Ad2_1_ATAC_Nucleosome_MACS2_NOMT_CRUK

#Remove blacklisted regions from peak list
bedtools intersect -a
/nobackup/umbch/postCRUK/MACS/NODUPS/KB_Tert_Ad2_1_ATAC_CutSite_MACS2_CRUK_peaks.narrowPeak  -b
/nobackup/umbch/postCRUK/MACS_blacklisted/blacklist_ENCFF001TDO.bed  -v >
/nobackup/umbch/postCRUK/MACS_blacklisted/B_Tert_Ad2_1_ATAC_CutSite_blacklisted_peaks.narrowPeak

#Peak QC
Rscript ChIPQC

#IDR for confident peaks
Rscript IDR

#Annotation
Rscript Annotation

echo "end now"

#################################################################################################################

# Generate Matrix to use for heatmap and plots
computeMatrix reference-point -S /nobackup/umbch/postCRUK/BigWigs/B_Tert_Ad2_1_ATAC_NODUPS_CRUK.bw
/nobackup/umbch/postCRUK/BigWigs/B_Tert_Ad2_5_ATAC_NODUPS_CRUK.bw
/nobackup/umbch/postCRUK/BigWigs/C_Tert_Ad2_2_ATAC_NODUPS_CRUK.bw
/nobackup/umbch/postCRUK/BigWigs/C_Tert_Ad2_6_ATAC_NODUPS_CRUK.bw
/nobackup/umbch/postCRUK/BigWigs/H_Tert_Ad2_3_ATAC_NODUPS_CRUK.bw
/nobackup/umbch/postCRUK/BigWigs/H_Tert_Ad2_7_ATAC_NODUPS_CRUK.bw
/nobackup/umbch/postCRUK/BigWigs/K_Tert_Ad2_4_ATAC_NODUPS_CRUK.bw
/nobackup/umbch/postCRUK/BigWigs/K_Tert_Ad2_8_ATAC_NODUPS_CRUK.bw  -R /nobackup/umbch/reference-
genomes/ensembl_tss2kb.bed  -a 1000 -b 1000 --skipZeros -o matrix1000_Tert_ATAC_TSS.gz

#################################################################################################################

# Generate heatmap
plotHeatmap -m matrix1000_Tert_ATAC_TSS.gz -out Tert_heatmap1000.pdf  --heatmapHeight 10 --colorMap hot_r --yAxisLabel
"Signal Enrichment" --xAxisLabel "Distance from TSS (bp)"

#################################################################################################################

# Fragment size distribution
picard CollectInsertSizeMetrics W=1000 I=../BAMfiles/B_Tert_Ad2_1_ATAC_NODUPS_NOMT_CRUK.bam
O=B_Tert_Ad2_1_ATAC_insert_size_metrics.txt H=B_Tert_Ad2_1_ATAC_insert_size_histogram.pdf

#################################################################################################################
# Correlation of signal (heatmap & scatter correlation)
multiBamSummary bins -b ../BAMfiles/B_Tert_Ad2_1_ATAC_NODUPS_CRUK.bam
../BAMfiles/B_Tert_Ad2_5_ATAC_NODUPS_CRUK.bam ../BAMfiles/C_Tert_Ad2_2_ATAC_NODUPS_CRUK.bam
../BAMfiles/C_Tert_Ad2_6_ATAC_NODUPS_CRUK.bam ../BAMfiles/H_Tert_Ad2_3_ATAC_NODUPS_CRUK.bam
../BAMfiles/H_Tert_Ad2_7_ATAC_NODUPS_CRUK.bam ../BAMfiles/K_Tert_Ad2_4_ATAC_NODUPS_CRUK.bam
../BAMfiles/K_Tert_Ad2_8_ATAC_NODUPS_CRUK.bam -bs 1000 -o BAM1000bpresults.npz --outRawCounts
BAM1000bpreadCounts.tab
```

```
deepTools2.0/bin/plotCorrelation \
   -in BAM1000bpreadCounts.npz \
   --corMethod spearman --skipZeros \
   --plotTitle "Spearman Correlation of Read Counts" \
   --whatToPlot heatmap --colorMap RdYlBu --plotNumbers \
   --whatToPlot scatterplot\
   -o heatmap_SpearmanCorr_readCounts.png   \
   --outFileCorMatrix SpearmanCorr_BAM1000bpreadCounts.tab

#########################################################################################################################

# MA plot of bin signal
Rscript MAplot.r
library(edgeR)
setwd("~/OneDrive - University of Leeds/Google_Drive/_PhD/ATAC 1/B0070_ATAC_all_Terts/postCRUK/MAplot")

Counts <- read.delim("Bigwig1000bpreadCounts.tab")
Counts <- Counts[1:100000,]
colnames(Counts) <- c("chr", "start", "end", "B_TERT_Ad2.1", "B_TERT_Ad2.5", "C_TERT_Ad2.2", "C_TERT_Ad2.6",
"H_TERT_Ad2.3", "H_TERT_Ad2.7", "K_TERT_Ad2.4", "K_TERT_Ad2.8")
CountsnoXY <- subset(Counts, chr != "chrX")
CountsnoXY <- subset(CountsnoXY, chr != "chrY")
CountsnoXY <- CountsnoXY[order(CountsnoXY$chr,   CountsnoXY$start),]
CountstoDElist <- CountsnoXY[,4:11]
y <- DGEList(counts=CountstoDElist)
y$samples$group <- c(1,1,1,1,2,2,2,2)
dim(y)
Gender <- c("male", "male","male","male","female", "female", "female", "female")

head(CountsnoXY)
eset <- CountsnoXY[,4:11]
eset <- log2(eset)
name <- c("B1", "B5", "C2", "C6", "H3", "H7", "K4","K8")
gender <- c("Male","Male","Male","Male","Female","Female", "Female", "Female")
targets <- data.frame(name,gender)
targets
Gender <- factor(targets$gender, levels = c("Male","Female"))
design3 <- model.matrix(~Gender)
colnames(design3)
fit <- lmFit(eset, design3)
fit <- eBayes(fit)
peaksFC <- data.frame(topTable(fit, number = 900000000))
peaksFC <- subset(peaksFC, logFC != "NA")
peaksFC <- peaksFC[which(peaksFC$P.Value <= 0.05),]
peaksFC <- peaksFC[which(peaksFC$logFC != "-Inf"),]
tail(peaksFC)
nrow(peaksFC)

a <- nrow(peaksFC[(peaksFC$logFC)>0,])
b <- nrow(peaksFC[(peaksFC$logFC)<0,])
nrow(peaksFC)

percentMale <- round((a/(a+b))*100,2)
percentFemale <- round((b/(a+b))*100,2)

plot1 <- ggplot(peaksFC, aes(AveExpr, logFC)) +
 labs(title = "MA Plot ATAC LIMMA") + xlab("Average Bin Signal") + ylab("LogFC Male vs Female") +
 geom_point(size = 0.02, colour = "gray30") +
 geom_smooth(colour = "brown4",se=FALSE) +
 ylim(-2,2) + xlim(0,6.1) +
 geom_hline(yintercept = 0, linetype = "dashed") +
 theme_classic()

plot2 <-
 ggplot(peaksFC, aes(logFC)) +
 geom_density(colour = "black", fill = "gray") +
 labs(title = "Density Plot") + xlab(NULL) + ylab("Density") +
 annotate("text", y = 0.6, x = -1.2, label = percentFemale, cex=3.5) +
 annotate("text", y = 0.91, x = -1.2, label = "%", cex=3.5) +
 annotate("text", y = 0.6, x = 1.3, label= percentMale, cex=3.5) +
 annotate("text", y = 0.93, x = 1.3, label= "%", cex=3.5)
```

```
 coord_flip() +
 geom_vline(xintercept = 0, linetype = "dashed") +
 scale_y_continuous(expand = c(0,0), breaks =c(0,0.5,1)) +
 xlim(-2,2) +
 theme_classic()

grid.arrange(plot1, plot2, ncol=2, widths = c(3,1.25))

#####################################################################################################

# Circos plot with all samples
Rscript circos.r
library(circlize)
setwd("/nobackup/umbch/postCRUK/bamtobed/bedgraph/ba mCoverage")
circos.clear()
pdf(file= "test_circos3.pdf" , width=6, height=6)
circos.par("start.degree" = 90, "track.height" = 0.08, "track.margin" = c(0,0), cell.padding = c(0, 0, 0, 0), "gap.degree" = 0)
circos.initializeWithIdeogram(species = "hg19")
circos.genomicTrack(c(MACS_blacklisted/B_Tert_Ad2_1_ATAC_CutSite_blacklisted_peaks.narrowPeak,MACS_blacklisted/B_Tert_
Ad2_5_ATAC_CutSite_blacklisted_peaks.narrowPeak),
                panel.fun = function(region, value, ...) {
                  circos.genomicLines(region, value, border = NA, area = T, col = "steelblue")})
circos.genomicTrack(c(MACS_blacklisted/C_Tert_Ad2_2_ATAC_CutSite_blacklisted_peaks.narrowPeak,
MACS_blacklisted/C_Tert_Ad2_6_ATAC_CutSite_blacklisted_peaks.narrowPeak),
                panel.fun = function(region, value, ...) {
                  circos.genomicLines(region, value, border = NA, area = TRUE, col = "lightskyblue3")})
circos.genomicTrack(c(MACS_blacklisted/H_Tert_Ad2_3_ATAC_CutSite_blacklisted_peaks.narrowPeak,
MACS_blacklisted/H_Tert_Ad2_7_ATAC_CutSite_blacklisted_peaks.narrowPeak),
                panel.fun = function(region, value, ...) {
                  circos.genomicLines(region, value, border = NA, area = TRUE, col = "palevioletred1")})
circos.genomicTrack(c(MACS_blacklisted/K_Tert_Ad2_4_ATAC_CutSite_blacklisted_peaks.narrowPeak,MACS_blacklisted/K_Tert_
Ad2_8_ATAC_CutSite_blacklisted_peaks.narrowPeak),
                panel.fun = function(region, value, ...) {
                  circos.genomicLines(region, value, border = NA, area = TRUE, col = "pink2")})
dev.off()
q()

#####################################################################################################

# Peak annotation and functional enrichment analysis
Rscript

library("GenomicRanges")
library("TxDb.Hsapiens.UCSC.hg19.knownGene")
library("EnsDb.Hsapiens.v86")
library("org.Hs.eg.db")
library("ChIPseeker")
library("ChIPpeakAnno")
library ("rtracklayer")
library("clusterProfiler")
#import replicate peakfile into R
peakfile1 <- "MalesSpecificPeaks2602.bed"
peaks_DF1 <-read.delim2(FemaleDEConsensus, comment.char = "#", header = T)
colnames(peaks_DF1) <- c("chr","start","end","width",  "strand","score")
head(peaks_DF1)
#Create GRanges object made of chr names and intervals stored as IRanges
peaks_GR1 <- GRanges(
  seqnames=peaks_DF1[,"chr"],
  IRanges(peaks_DF1[,"start"],
          peaks_DF1[,"end"]),
  mcols = peaks_DF1[,c("width",  "strand","score")])
head(peaks_GR1)
#Writedata Frame for peaksfiles
df1 <- data.frame(seqnames=seqnames(peaks_GR1),
                  starts=start(peaks_GR1)-1,
                  ends=end(peaks_GR1),
                  names=c(rep(".", length(peaks_GR1))),
                  scores=c(rep(".", length(peaks_GR1))),
                  strands=strand(peaks_GR1))
head(df1)
#ChIPseeker 3000bp
```

```
txdb <- TxDb.Hsapiens.UCSC.hg19.knownGene
promoter <- getPromoters(TxDb=txdb, upstream=3000, downstream=3000)
tagMatrixGR1 <- getTagMatrix(peaks_GR1, windows=promoter)
peakAnnoGR1 <- annotatePeak(peaks_GR1, tssRegion=c(-3000,  3000), TxDb=txdb, annoDb="org.Hs.eg.db") #peak annotation by
genomic features
peakAnnoGR1df <- data.frame(peakAnnoGR1)
head(peakAnnoGR1df)
dim(peakAnnoGR1df)
write.table(peakAnnoGR1df, file="Female_ConsensusSpecific_peaks_annotated010419.bed",  quote=F, sep="\t", row.names=F,
col.names=T)

plotAnnoPie(peakAnnoGR1)# pie chart of features
plotAnnoBar(peakAnnoGR1) # bar plot of features
upsetplot(peakAnnoGR1, vennpie=TRUE, text.scale = c(1.5, 1.5, 1.3, 1.3, 1.6, 1.3))  # Venn Plot
plotDistToTSS(peakAnnoGR1, title="Distribution of transcription factor-binding loci relative to TSS")

#Functional Enrichment
gene <- seq2gene(peaks_GR1, tssRegion = c(-1000, 1000), flankDistance = 50000, TxDb=txdb)
pathwayGO <- enrichGO(gene,OrgDb = org.Hs.eg.db, ont = "All", readable = T )
pathwayReactome <- enrichPathway(gene, readable = T )
pathwayKEGG <- enrichKEGG(gene)
write.table(pathwayGO, file="GO_List.txt",quote=F, sep="\t", row.names=F, col.names=T)
write.table(pathwayReactome, file="GO_List.txt",quote=F, sep="\t", row.names=F, col.names=T)
write.table(pathwayKEGG, file="GO_List.txt",quote=F, sep="\t", row.names=F, col.names=T)
barplot(pathwayGO,showCategory = 10, colorBy="pvalue", title = "GO Female Peaks")
barplot(pathwayReactome,showCategory = 10, colorBy="pvalue", title = "Reactome Associated Peaks ")
barplot(pathwayKEGG,showCategory = 10, colorBy="pvalue", title = "KEGG Associated Peaks ")

#########################################################################################################################

# Finding Gender-associated consensus peaks, differential Enrichment & analysis on peaks
Rscript

library(DiffBind)

samples <- read.delim("samples.txt")
peaks <- dba(sampleSheet=samples)
peaks <- dba.count(peaks, summits=250)
peaks <- dba.contrast(peaks, categories=DBA_FACTOR)
peaks <- dba.analyze(peaks)
peaks.DB <- dba.report(peaks)
tamoxifen.OL <- dba.overlap(peaks, peaks$masks$Male)
samples <- read.delim("samples.txt")
peaks <- dba(sampleSheet=samples)

pdf("heatmap.pdf")
plot(peaks) #hierarchical clustering
dev.off()

peaks <- dba.count(peaks, summits=250)
pdf("heatmap_centeredAtPeaks.pdf")
plot(peaks) #hierarchical clustering
dev.off()

peaks <- dba.contrast(peaks, categories=DBA_FACTOR)
peaks <- dba.analyze(peaks)

pdf("heatmap-postDEanalysis")
plot(peaks, contrast=1)
dev.off()

peaks.DB <- dba.report(peaks)

pdf("PCA_plot2")
dba.plotPCA(peaks,DBA_TISSUE,label=DBA_FACTOR)
dev.off()

pdf("PCA_plot  PCA plot using affinity data for only differentially bound sites")
dba.plotPCA(peaks, contrast=1,label=DBA_TISSUE)
dev.off()
```

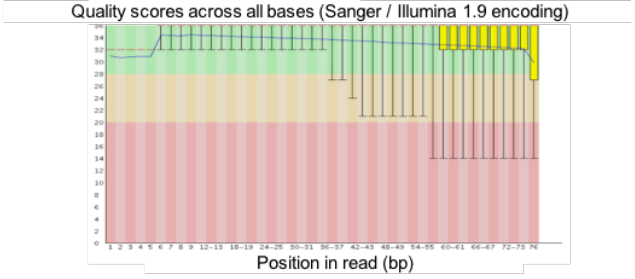
```
pdf("MA  plot 2fold")
dba.plotMA(peaks, fold = 2)
dev.off()
pdf("MA  plot 200fold2")
dba.plotMA(peaks, fold = 200)
dev.off()
pdf("volcano  plot")
dba.plotVolcano(peaks) + geom_point(size = 0.02, colour  = "gray30") + theme_classic()
dev.off()

labs(title = "MA Plot ATAC LIMMA") + xlab("Average Bin Signal") + ylab("LogFC Male vs Female") +
  geom_point(size = 0.02, colour  = "gray30") + geom_smooth(colour  = "brown4",se=FALSE) +
  ylim(-2,2) + xlim(0,6.1) + geom_hline(yintercept = 0, linetype = "dashed") +
  theme_classic()

pdf("box  plot")
dba.plotBox(peaks)
dev.off()

pdf("binding  affinity heatmap")
dba.plotHeatmap(peaks, contrast=1, correlations=FALSE)
dev.off()

olap.rate <- dba.overlap(peaks,mode=DBA_OLAP_RATE)
pdf("overlap  rate plot")
plot(olap.rate,type='b',ylab='# peaks', xlab='Overlap at least this many peaksets')
dev.off()

pdf("overlap  Male Venn")
dba.plotVenn(peaks,peaks$masks$Male)
dev.off()

pdf("overlap  Female Venn")
dba.plotVenn(peaks,peaks$masks$Consensus)
dev.off()

#Consensus peak calling
dba.overlap(peaks,peaks$masks$Male,mode=DBA_OLAP_RATE)
dba.overlap(peaks,peaks$masks$Female,mode=DBA_OLAP_RATE)
peaks <- dba.peakset(peaks, consensus=DBA_FACTOR,  minOverlap=0.75)
dba.peakset(tamoxifen, consensus=-c(DBA_REPLICATE,DBA_FACTOR)
dba.plotVenn(peaks,peaks$masks$Consensus)
peaks.OL <- dba.overlap(peaks, peaks$masks$Consensus)
peaks.OL$onlyA
peaks.OL$onlyB

#############################################################################################################
# Peak Motif enrichment analysis
HOMER
mkdir DistalIntergenicPeaks
mkdir IntragenicIntronPeaks
mkdir PromoterPeaks
cd DistalIntergenicPeaks
findMotifsGenome.pl  ../peakfiles/Annotated_DistalIntergenic_Female_Unique_Peaks.txt hg19
DistalIntergenicPeaks_Female_Unique -size 50
findMotifsGenome.pl  ../peakfiles/Annotated_DistalIntergenic_Male_Unique_Peaks.txt hg19 DistalIntergenicPeaks_Male_Unique
-size 50
findMotifsGenome.pl  ../peakfiles/Annotated_DistalIntergenic_SharedMaleFemale_Peaks.txt hg19 DistalIntergenicPeaks_Shared -
size 50
cd ../IntragenicIntronPeaks
findMotifsGenome.pl  ../peakfiles/Annotated_IntragenicIntron_Female_Unique_Peaks.txt hg19 IntragenicIntron_Female_Unique -
size 50
findMotifsGenome.pl  ../peakfiles/Annotated_IntragenicIntron_Male_Unique_Peaks.txt hg19 IntragenicIntron_Male_Unique -size
50
findMotifsGenome.pl  ../peakfiles/Annotated_IntragenicIntron_SharedMaleFemale_Peaks.txt hg19 IntragenicIntron_Shared -size
50
cd ../PromoterPeaks
findMotifsGenome.pl  ../peakfiles/Annotated_Promoter_Female_Unique_Peaks.txt hg19 PromoterPeaks_Female_Unique -size 50
findMotifsGenome.pl  ../peakfiles/Annotated_Promoter_Male_Unique_Peaks.txt hg19 PromoterPeaks_Male_Unique -size 50
findMotifsGenome.pl  ../peakfiles/Annotated_Promoter_SharedMaleFemale_Peaks.txt hg19 PromoterPeaks_Shared -size 50
```

Appendix D ATAC-seq QA

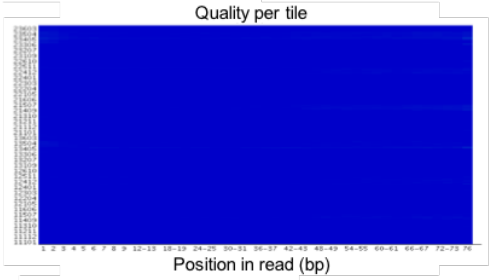
Basic Statistics

Measure	Value
Filename	B_Tert_Ad2_1_S1_R2.fastq.qz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	60579256
Sequences flagged as poor quality	0
Sequence length	76
%GC	45

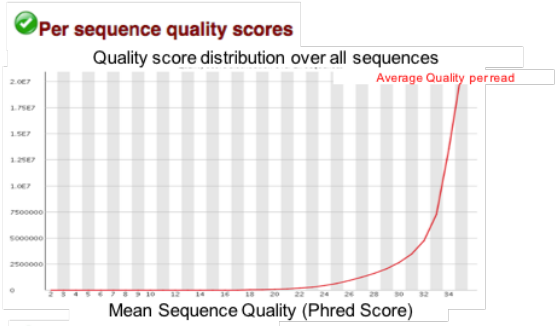
Per base sequence quality



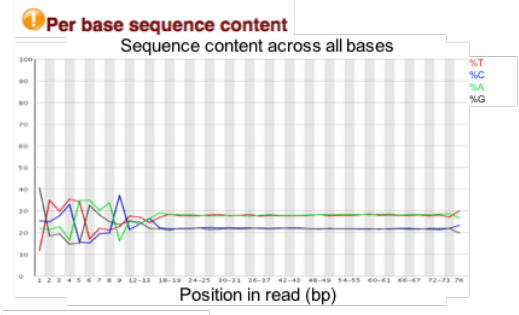
Per tile sequence quality



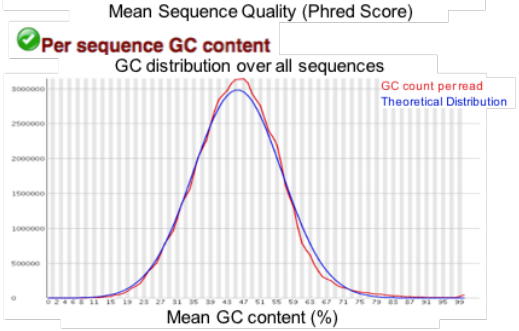
Per sequence quality scores



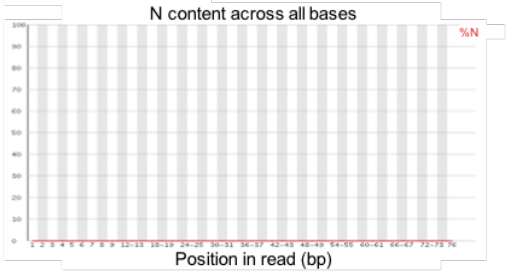
Per base sequence content



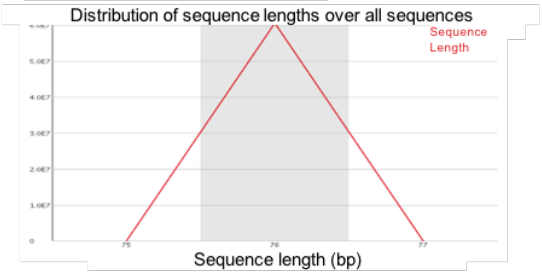
Per sequence GC content



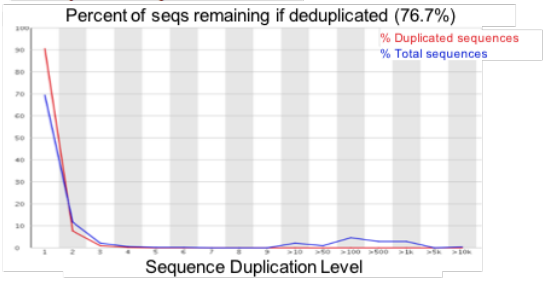
Per base N content



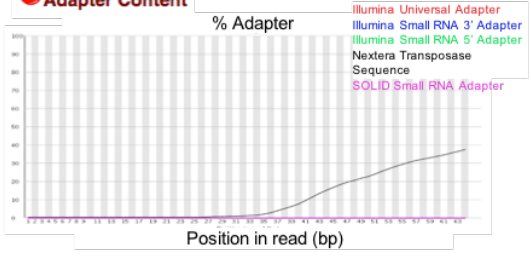
Sequence Length Distribution



Sequence Duplication Levels



Adapter Content

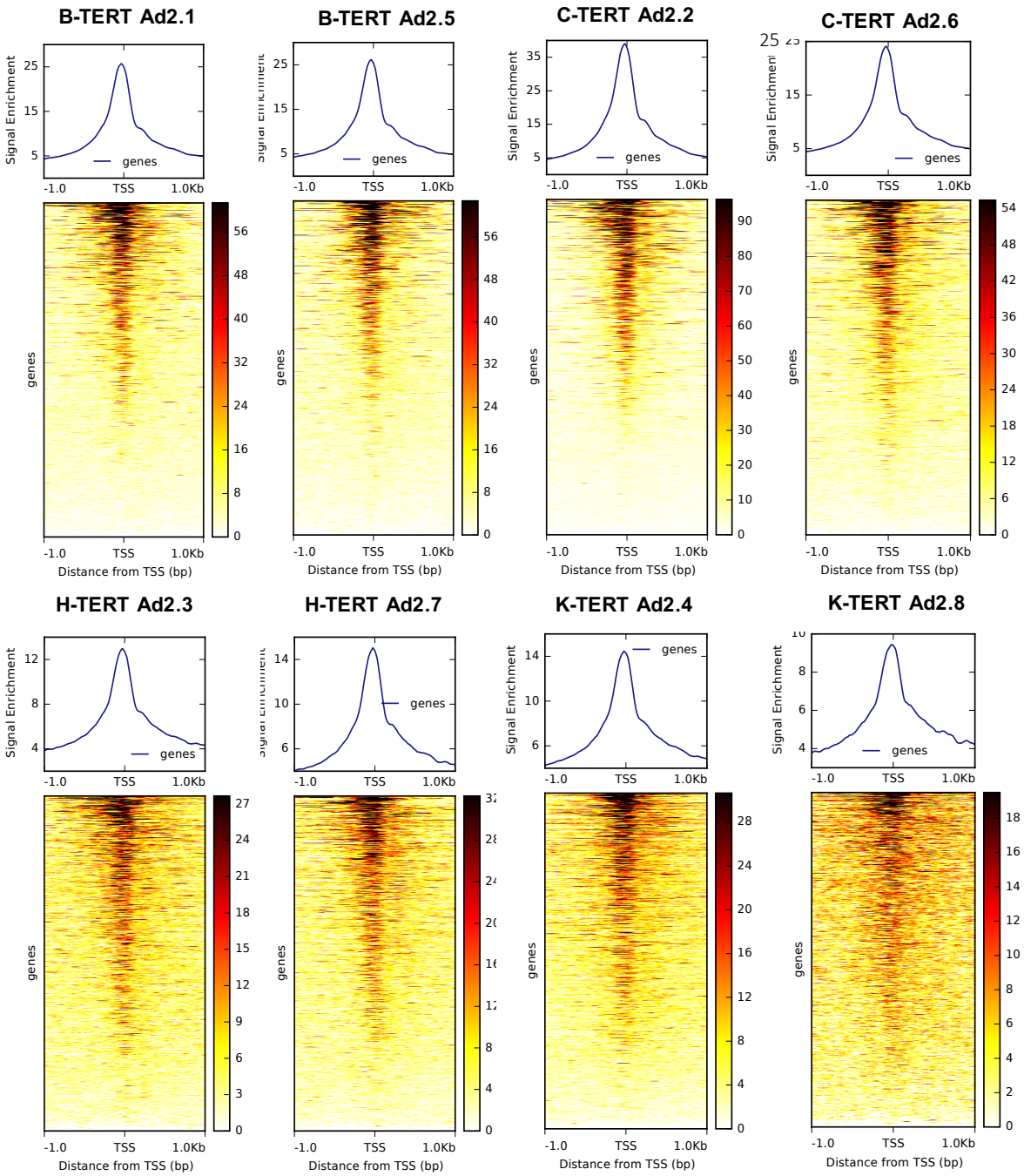


Overrepresented sequences

Sequence	Count	Percentage	Possible Source
CTGTCTCTTATACACATCTGACGCT GCCGACGAGTGTAGATCTCGGTGGT	238040	0.3929397878376057	Illumina Single End PCR Primer 1 (95% over 21bp)

Appendix D - 1 Output plots from FastQC for B-TERT Ad2.1 R2 pre-trimming

Appendix E ATAC-seq results



Appendix E - 1 Individual heatmaps of NHU-TERT ATAC-samples centred around all hg19 TSSs +/-1kb. Each heatmap is scaled to the individual sample.

Appendix E - 2 : Top 100 (of 53849) Male DE-consensus peaks

Peak Location	logFC	Annotation	DistToTSS	Symbol	Gene
chrY:21729047	8.77	Promoter	0	TXLNGY	taxilin gamma pseudogene, Y-linked
chrY:22737529	8.55	Promoter	0	EIF1AY	eukaryotic translation initiation factor 1A, Y-linked
chrY:7649764	7.99	Intron	-22952	TTTY12	testis-specific transcript, Y-linked 12 (non-protein coding)
chrY:17861585	7.65	Distal Intergenic	-945421	NLGN4Y-AS1	NLGN4Y antisense RNA 1
chrY:16946746	7.53	Intron	-30582	NLGN4Y-AS1	NLGN4Y antisense RNA 1
chrY:15591971	7.46	Promoter	330	UTY	ubiquitously transcribed tetratricopeptide repeat containing, Y-linked
chrY:21573027	7.42	Distal Intergenic	91763	BCORP1	BCL6 corepressor pseudogene 1
chrY:21329948	7.25	Distal Intergenic	-90395	TTTY14	testis-specific transcript, Y-linked 14 (non-protein coding)
chrY:15016715	7.19	Promoter	0	DDX3Y	DEAD-box helicase 3, Y-linked
chrY:15863335	7.1	Distal Intergenic	47637	TMSB4Y	thymosin beta 4, Y-linked
chrY:2804057	7.08	Promoter	288	ZFY	zinc finger protein, Y-linked
chrY:21516155	7.03	Distal Intergenic	148635	BCORP1	BCL6 corepressor pseudogene 1
chrY:15605786	7	Distal Intergenic	-12985	UTY	ubiquitously transcribed tetratricopeptide repeat containing, Y-linked
chrY:15815660	6.96	Promoter	0	TMSB4Y	thymosin beta 4, Y-linked
chrY:16743752	6.92	Intron	9600	NLGN4Y	neuroligin 4, Y-linked
chrY:2756187	6.91	Distal Intergenic	46409	RPS4Y1	ribosomal protein S4, Y-linked 1
chrY:7318125	6.91	Distal Intergenic	175861	PRKY	protein kinase, Y-linked, pseudogene
chrY:14763686	6.9	Distal Intergenic	-10363	TTTY15	testis-specific transcript, Y-linked 15 (non-protein coding)
chrY:7302043	6.88	Distal Intergenic	159779	PRKY	protein kinase, Y-linked, pseudogene
chrY:7142110	6.86	Promoter	0	PRKY	protein kinase, Y-linked, pseudogene
chrY:2755647	6.85	Distal Intergenic	45869	RPS4Y1	ribosomal protein S4, Y-linked 1
chrY:2872121	6.83	Promoter	833	LINC00278	long intergenic non-protein coding RNA 278
chrY:21906751	6.82	Promoter	0	KDM5D	lysine demethylase 5D
chrY:21483772	6.82	Distal Intergenic	181018	BCORP1	BCL6 corepressor pseudogene 1
chrY:2870439	6.75	Promoter	-349	LINC00278	long intergenic non-protein coding RNA 278
chrY:18780497	6.74	Distal Intergenic	-832092	FAM41AY2	family with sequence similarity 41 member A, Y-linked 2
chrY:15280385	6.73	Distal Intergenic	190409	UTY	ubiquitously transcribed tetratricopeptide repeat containing, Y-linked
chrY:16598012	6.72	Distal Intergenic	-36227	NLGN4Y	neuroligin 4, Y-linked
chrY:7994417	6.71	Distal Intergenic	321201	TTTY12	testis-specific transcript, Y-linked 12 (non-protein coding)
chrY:21430506	6.69	Distal Intergenic	-190953	TTTY14	testis-specific transcript, Y-linked 14 (non-protein coding)
chrY:16452810	6.64	Distal Intergenic	-181429	NLGN4Y	neuroligin 4, Y-linked
chrY:21238048	6.6	Promoter	1005	TTTY14	testis-specific transcript, Y-linked 14 (non-protein coding)
chrY:14756063	6.56	Distal Intergenic	-17986	TTTY15	testis-specific transcript, Y-linked 15 (non-protein coding)
chrY:15727428	6.51	Distal Intergenic	-87770	TMSB4Y	thymosin beta 4, Y-linked
chrY:17647852	6.48	Distal Intergenic	-731688	NLGN4Y-AS1	NLGN4Y antisense RNA 1
chrY:7990065	6.48	Distal Intergenic	316849	TTTY12	testis-specific transcript, Y-linked 12 (non-protein coding)
chrY:18953593	6.42	Distal Intergenic	-658996	FAM41AY2	family with sequence similarity 41 member A, Y-linked 2
chrY:19156367	6.41	Distal Intergenic	-456222	FAM41AY2	family with sequence similarity 41 member A, Y-linked 2
chrY:21068289	6.4	Distal Intergenic	-27924	NA	NA
chrY:16440742	6.38	Distal Intergenic	-193497	NLGN4Y	neuroligin 4, Y-linked
chrY:19431997	6.29	Distal Intergenic	-180592	FAM41AY2	family with sequence similarity 41 member A, Y-linked 2
chrY:16358987	6.25	Distal Intergenic	190638	VCY	variable charge, Y-linked
chrY:14595407	6.25	Distal Intergenic	-61767	GYG2P1	glycogenin 2 pseudogene 1
chrY:18728673	6.23	Distal Intergenic	-883916	FAM41AY2	family with sequence similarity 41 member A, Y-linked 2
chrY:18946223	6.21	Distal Intergenic	-666366	FAM41AY2	family with sequence similarity 41 member A, Y-linked 2
chrY:18946223	6.21	Distal Intergenic	-666366	FAM41AY2	family with sequence similarity 41 member A, Y-linked 2
chrY:17705821	6.21	Distal Intergenic	-789657	NLGN4Y-AS1	NLGN4Y antisense RNA 1
chrY:16549745	6.18	Distal Intergenic	-84494	NLGN4Y	neuroligin 4, Y-linked
chrY:14845748	6.15	Intron	24176	USP9Y	ubiquitin specific peptidase 9, Y-linked
chrY:16584500	6.13	Distal Intergenic	-49739	NLGN4Y	neuroligin 4, Y-linked

chrY:14775217	6	Promoter	668	TTTY15	testis-specific transcript, Y-linked 15 (non-protein coding)
chrY:7662662	5.99	Intron	-10054	TTTY12	testis-specific transcript, Y-linked 12 (non-protein coding)
chrY:15017597	5.93	Promoter	0	DDX3Y	DEAD-box helicase 3, Y-linked
chrY:17018030	5.92	Distal Intergenic	-101866	NLGN4Y-AS1	NLGN4Y antisense RNA 1
chrY:7703663	5.84	Distal Intergenic	30447	TTTY12	testis-specific transcript, Y-linked 12 (non-protein coding)
chrY:19278987	5.83	Distal Intergenic	-333602	FAM41AY2	family with sequence similarity 41 member A, Y-linked 2
chrY:7325829	5.83	Distal Intergenic	183565	PRKY	protein kinase, Y-linked, pseudogene
chrY:21221819	5.8	Intron	17234	TTTY14	testis-specific transcript, Y-linked 14 (non-protein coding)
chrY:19431412	5.72	Distal Intergenic	-181177	FAM41AY2	family with sequence similarity 41 member A, Y-linked 2
chrY:14576174	5.68	Distal Intergenic	-42534	GYG2P1	glycogenin 2 pseudogene 1
chrY:17926342	5.67	Distal Intergenic	-1010178	NLGN4Y-AS1	NLGN4Y antisense RNA 1
chrY:21071898	5.53	Distal Intergenic	-31533	NA	NA
chrY:15009122	5.52	Distal Intergenic	-6648	DDX3Y	DEAD-box helicase 3, Y-linked
chr22:20378562	4.48	Promoter	642	TMEM191B	transmembrane protein 191B
chr17:58964680	3.82	Promoter	18	BCAS3	BCAS3, microtubule associated cell migration factor
chr17:58964680	3.82	Promoter	18	BCAS3	BCAS3, microtubule associated cell migration factor
chrX:26280406	3.6	Distal Intergenic	45869	MAGEB5	MAGE family member B5
chr13:109963465	3.44	Distal Intergenic	-143563	MYO16-AS1	MYO16 antisense RNA 1
chr1:192753126	3.41	Distal Intergenic	-24794	RGS2	regulator of G protein signaling 2
chr4:96212026	3.28	Intron	186269	BMPR1B	bone morphogenetic protein receptor type 1B
chr17:3056565	3.2	Distal Intergenic	-25469	OR1G1	olfactory receptor family 1 subfamily G member 1
chr2:1711734	3.18	Intron	36308	PXDN	peroxidasin
chr7:118484978	3.13	Distal Intergenic	620015	ANKRD7	ankyrin repeat domain 7
chr7:119993448	3.12	Intron	79475	KCND2	potassium voltage-gated channel subfamily D member 2
chr11:107243908	3.11	Intron	69618	CWF19L2	CWF19 like 2, cell cycle control (S. pombe)
chr5:35047104	3.07	Promoter	887	AGXT2	alanine--glyoxylate aminotransferase 2
chr1:166459614	3.05	Distal Intergenic	-113290	FMO9P	flavin containing monooxygenase 9 pseudogene
chr4:121663058	2.98	Intron	180706	PRDM5	PR/SET domain 5
chr12:116472328	2.97	Intron	113882	MIR620	microRNA 620
chr18:54890528	2.93	Distal Intergenic	75984	BOD1L2	biorientation of chromosomes in cell division 1 like 2
chr18:39772997	2.92	Intron	6113	LINC00907	long intergenic non-protein coding RNA 907
chr13:95767990	2.92	Exon	-52626	ABCC4	ATP binding cassette subfamily C member 4
chr4:138966380	2.89	Intron	-43539	SLC7A11-AS1	SLC7A11 antisense RNA 1
chr15:97311109	2.87	Distal Intergenic	13526	SPATA8-AS1	SPATA8 antisense RNA 1 (head to head)
chr10:109812929	2.87	Distal Intergenic	-888212	SORCS1	sortilin related VPS10 domain containing receptor 1
chr8:133520089	2.85	Distal Intergenic	-26834	KCNQ3	potassium voltage-gated channel subfamily Q member 3
chr6:155594788	2.82	Intron	9390	CLDN20	claudin 20
chr2:5390811	2.78	Distal Intergenic	-441739	SOX11	SRY-box 11
chr21:42308834	2.77	Distal Intergenic	-89544	DSCAM	DS cell adhesion molecule
chr12:16536054	2.77	Distal Intergenic	29452	MGST1	microsomal glutathione S-transferase 1
chr9:1198244	2.76	Distal Intergenic	146458	DMRT2	doublesex and mab-3 related transcription factor 2
chr18:35633053	2.76	Distal Intergenic	395704	MIR4318	microRNA 4318
chr6:9140455	2.76	Distal Intergenic	487762	HULC	hepatocellular carcinoma up-regulated long non-coding RNA
chr18:1711834	2.76	Distal Intergenic	-304402	LINC00470	long intergenic non-protein coding RNA 470
chr8:25548009	2.73	Intron	197230	EBF2	early B-cell factor 2
chr9:32424783	2.69	Exon	39931	ACO1	aconitase 1
chr6:161730981	2.67	Distal Intergenic	-35623	AGPAT4	1-acylglycerol-3-phosphate O-acyltransferase 4
chr1:237922845	2.66	Exon	-32052	RYR2	ryanodine receptor 2
chr14:37130997	2.66	5' UTR	-3952	PAX9	paired box 9
chr6:130774590	2.66	Distal Intergenic	16077	TMEM200A	transmembrane protein 200A

Appendix E - 3 : Top 100 (of 221) Female DE-consensus peaks

Peak Location	logFC	Annotation	DistToTSS	Symbol	Gene
chrX:112100886	-3.58	Distal Intergenic	-16592	AMOT	angiomotin
chrX:73070997	-3.53	Promoter	1342	XIST	X inactive specific transcript (non-protein coding)
chr15:54442626	-3.21	Intron	-113468	UNC13C	unc-13 homolog C
chrX:112211333	-2.85	Distal Intergenic	-127039	AMOT	angiomotin
chr8:76307570	-2.84	Distal Intergenic	-12364	HNF4G	hepatocyte nuclear factor 4 gamma
chr16:59937575	-2.66	Distal Intergenic	-148229	APOOP5	apolipoprotein O pseudogene 5
chr22:33776954	-2.59	Intron	55452	MIR4764	microRNA 4764
chr22:33747140	-2.49	Intron	85266	MIR4764	microRNA 4764
chr2:59928580	-2.48	Distal Intergenic	685751	MIR4432	microRNA 4432
chrX:56055792	-2.42	Distal Intergenic	-202781	KLF8	Kruppel like factor 8
chr11:39012573	-2.31	Distal Intergenic	1301858	LRRC4C	leucine rich repeat containing 4C
chr15:26129182	-2.28	Distal Intergenic	-18076	LINC02346	long intergenic non-protein coding RNA 2346
chr10:66954087	-2.27	Distal Intergenic	368551	ANXA2P3	annexin A2 pseudogene 3
chr4:64987731	-2.26	Distal Intergenic	159297	TECRL	trans-2,3-enoyl-CoA reductase like
chr1:95608585	-2.21	Intron	24855	TMEM56-RWDD3	TMEM56-RWDD3 readthrough
chr14:27311317	-2.16	Distal Intergenic	-66282	MIR4307	microRNA 4307
chr15:26271537	-2.16	Intron	-89174	LINC00929	long intergenic non-protein coding RNA 929
chrX:112277377	-2.16	Distal Intergenic	-193083	AMOT	angiomotin
chr13:93162831	-2.11	Intron	209147	GPC5-AS1	GPC5 antisense RNA 1
chr2:107954106	-2.08	Distal Intergenic	-450292	ST6GAL2	ST6 beta-galactoside alpha-2,6-sialyltransferase 2
chr8:52498727	-2.05	Intron	-176355	PXDNL	peroxidasin like
chr2:107985005	-2.03	Distal Intergenic	457328	RGPD4-AS1	RGPD4 antisense RNA 1 (head to head)
chr16:59939682	-2.00	Distal Intergenic	-150336	APOOP5	apolipoprotein O pseudogene 5
chr5:155829572	-2.00	Intron	75554	SGCD	sarcoglycan delta
chr3:75334868	-1.99	Distal Intergenic	70990	MIR4444-1	microRNA 4444-1
chrX:91502250	-1.98	Intron	411539	PCDH11X	protocadherin 11 X-linked
chr11:91425284	-1.97	Distal Intergenic	-659729	FAT3	FAT atypical cadherin 3
chr5:99038478	-1.96	Distal Intergenic	685231	LOC100133050	glucuronidase beta pseudogene
chr3:55095883	-1.96	Intron	-133560	LRTM1	leucine rich repeats and transmembrane domains 1
chr11:25445517	-1.96	Distal Intergenic	-765063	ANO3	anoctamin 3
chr8:79027165	-1.93	Distal Intergenic	-400922	PKIA	cAMP-dependent protein kinase inhibitor alpha
chrX:79590837	-1.92	Promoter	0	FAM46D	family with sequence similarity 46 member D
chr16:59912580	-1.91	Distal Intergenic	-123234	APOOP5	apolipoprotein O pseudogene 5
chr1:228756789	-1.90	Distal Intergenic	-23356	DUSP5P1	dual specificity phosphatase 5 pseudogene 1
chr22:33834396	-1.90	Promoter	-1490	MIR4764	microRNA 4764
chr5:24804445	-1.90	Distal Intergenic	35998	LINC02239	long intergenic non-protein coding RNA 2239
chr3:467963	-1.90	Distal Intergenic	106346	CHL1	cell adhesion molecule L1 like
chr5:24789096	-1.86	Distal Intergenic	51347	LINC02239	long intergenic non-protein coding RNA 2239
chr11:38836470	-1.85	Distal Intergenic	1477961	LRRC4C	leucine rich repeat containing 4C
chr8:78562131	-1.84	Distal Intergenic	-648600	PEX2	peroxisomal biogenesis factor 2
chr13:93226979	-1.81	Intron	144999	GPC5-AS1	GPC5 antisense RNA 1
chr11:41075592	-1.81	Intron	405345	LRRC4C	leucine rich repeat containing 4C
chr7:154559130	-1.80	Intron	-160848	PAXIP1-AS2	PAXIP1 antisense RNA 2
chr21:17960888	-1.79	Promoter	-1420	MIR125B2	microRNA 125b-2
chr2:130457527	-1.79	Distal Intergenic	234114	PLAC9P1	placenta specific 9 pseudogene 1
chr15:54423804	-1.76	Intron	118452	UNC13C	unc-13 homolog C
chr9:115825348	-1.75	Distal Intergenic	-6101	ZFP37	ZFP37 zinc finger protein
chrX:112076167	-1.75	Intron	-7560	AMOT	angiomotin
chr10:131105529	-1.71	Distal Intergenic	-159676	MGMT	O-6-methylguanine-DNA methyltransferase
chrX:95259580	-1.71	Distal Intergenic	333072	BRDTP1	bromodomain testis associated pseudogene 1
chr8:62146421	-1.69	Distal Intergenic	-53855	CLVS1	clavesin 1
chr4:171147690	-1.69	Distal Intergenic	-136067	AADAT	aminoadipate aminotransferase
chr14:38873423	-1.67	Distal Intergenic	-147597	CLEC14A	C-type lectin domain containing 14A
chr3:55226234	-1.67	Distal Intergenic	-263911	LRTM1	leucine rich repeats and transmembrane domains 1

chr16:79001302	-1.66	Intron	633071	MAF	MAF bZIP transcription factor
chr7:154568898	-1.65	Intron	-151080	PAXIP1-AS2	PAXIP1 antisense RNA 2
chr11:91864860	-1.64	Distal Intergenic	-220153	FAT3	FAT atypical cadherin 3
chr7:154561451	-1.63	Exon	-158527	PAXIP1-AS2	PAXIP1 antisense RNA 2
chr8:52091404	-1.62	Distal Intergenic	230468	PXDNL	peroxidasin like
chr3:55281876	-1.62	Distal Intergenic	232836	WNT5A	Wnt family member 5A
chr2:108010180	-1.62	Distal Intergenic	432153	RGPD4-AS1	RGPD4 antisense RNA 1 (head to head)
chr7:154596296	-1.60	Intron	-123682	PAXIP1-AS2	PAXIP1 antisense RNA 2
chr1:242589778	-1.60	Intron	22757	PLD5	phospholipase D family member 5
chr16:79084090	-1.59	Intron	550283	MAF	MAF bZIP transcription factor
chr4:97570904	-1.57	Distal Intergenic	-716924	STPG2-AS1	STPG2 antisense RNA 1
chr3:140011736	-1.57	Intron	357458	CLSTN2	calsyntenin 2
chr11:59936911	-1.56	Downstream	13644	MS4A6A	membrane spanning 4-domains A6A
chr2:107975110	-1.56	Distal Intergenic	467223	RGPD4-AS1	RGPD4 antisense RNA 1 (head to head)
chr5:24823906	-1.55	Distal Intergenic	16537	LINC02239	long intergenic non-protein coding RNA 2239
chr5:155806420	-1.54	Intron	52402	SGCD	sarcoglycan delta
chr13:70278632	-1.51	Intron	-402464	ATXN8OS	ATXN8 opposite strand (non-protein coding)
chrX:36627888	-1.51	Distal Intergenic	380741	NA	NA
chr2:130469535	-1.50	Distal Intergenic	222106	PLAC9P1	placenta specific 9 pseudogene 1
chr2:118900434	-1.50	Distal Intergenic	39410	INSIG2	insulin induced gene 2
chr13:48054465	-1.50	Distal Intergenic	520396	SUCLA2	succinate-CoA ligase ADP-forming beta subunit
chr13:93803685	-1.45	Distal Intergenic	-75144	GPC6	glypican 6
chr19:21646777	-1.45	Exon	-19150	LINC00664	long intergenic non-protein coding RNA 664
chr6:145434147	-1.45	Distal Intergenic	558071	EPM2A	EPM2A, laforin glucan phosphatase
chr3:55033435	-1.45	Intron	-71112	LRTM1	leucine rich repeats and transmembrane domains 1
chr3:5275383	-1.43	Distal Intergenic	16308	MIR4790	microRNA 4790
chr10:44796738	-1.43	Distal Intergenic	8289	C10orf142	chromosome 10 open reading frame 142
chr9:7434150	-1.41	Distal Intergenic	365368	DMAC1	distal membrane arm assembly complex 1
chr12:71270936	-1.40	Intron	43399	PTPRR	protein tyrosine phosphatase, receptor type R
chr3:55194647	-1.39	Distal Intergenic	-232324	LRTM1	leucine rich repeats and transmembrane domains 1
chr13:54785019	-1.38	Distal Intergenic	100915	MIR1297	microRNA 1297
chr13:61989355	-1.37	Promoter	51	PCDH20	protocadherin 20
chr11:92689285	-1.36	Distal Intergenic	-13255	MTNR1B	melatonin receptor 1B
chr15:20563113	-1.35	Distal Intergenic	74865	CHEK2P2	checkpoint kinase 2 pseudogene 2
chr6:96683966	-1.35	Distal Intergenic	219870	FUT9	fucosyltransferase 9
chr1:248020782	-1.34	Promoter	30	TRIM58	tripartite motif containing 58
chr13:93199739	-1.33	Intron	172239	GPC5-AS1	GPC5 antisense RNA 1
chr14:38812215	-1.33	Distal Intergenic	-86389	CLEC14A	C-type lectin domain containing 14A
chr10:42971097	-1.32	Promoter	0	LINC00839	long intergenic non-protein coding RNA 839
chr10:42971097	-1.32	Promoter	0	LINC00839	long intergenic non-protein coding RNA 839
chr10:66716329	-1.32	Distal Intergenic	130793	ANXA2P3	annexin A2 pseudogene 3
chr5:178012698	-1.30	Intron	4609	COL23A1	collagen type XXIII alpha 1 chain
chr6:117690660	-1.29	Intron	56109	ROS1	ROS proto-oncogene 1, receptor tyrosine kinase
chr16:79406567	-1.28	Distal Intergenic	227806	MAF	MAF bZIP transcription factor
chr7:51408416	-1.27	Distal Intergenic	-23650	COBL	cordon-bleu WH2 repeat protein

Appendix E - 4 : Top 100 male-associated peaks GO terms (P-value < $1x10^{-7}$ for all terms shown)

	ID	Description		ID	Description
1	GO:0005813	centrosome	**51**	GO:0072331	signal transduction by p53 class mediator
2	GO:0005759	mitochondrial matrix	**52**	GO:0010506	regulation of autophagy
3	GO:0010256	endomembrane system organization	**53**	GO:0042623	ATPase activity, coupled
4	GO:0005819	spindle	**54**	GO:1903320	regulation of protein modification by small protein conjugation or removal
5	GO:0051052	regulation of DNA metabolic process	**55**	GO:0006397	mRNA processing
6	GO:0060271	cilium assembly	**56**	GO:0050839	cell adhesion molecule binding
7	GO:0044839	cell cycle G2/M phase transition	**57**	GO:1902749	regulation of cell cycle G2/M phase transition
8	GO:0044782	cilium organization	**58**	GO:0033044	regulation of chromosome organization
9	GO:0000226	microtubule cytoskeleton organization	**59**	GO:0007093	mitotic cell cycle checkpoint
10	GO:0010498	proteasomal protein catabolic process	**60**	GO:0006644	phospholipid metabolic process
11	GO:0000086	G2/M transition of mitotic cell cycle	**61**	GO:0006732	coenzyme metabolic process
12	GO:0005925	focal adhesion	**62**	GO:0042470	melanosome
13	GO:0030055	cell-substrate junction	**63**	GO:0048770	pigment granule
14	GO:0019787	ubiquitin-like protein transferase activity	**64**	GO:0010389	regulation of G2/M transition of mitotic cell cycle
15	GO:0016607	nuclear speck	**65**	GO:0051169	nuclear transport
16	GO:0005924	cell-substrate adherens junction	**66**	GO:0051054	positive regulation of DNA metabolic process
17	GO:0004674	protein serine/threonine kinase activity	**67**	GO:0044389	ubiquitin-like protein ligase binding
18	GO:0004842	ubiquitin-protein transferase activity	**68**	GO:0043393	regulation of protein binding
19	GO:0005743	mitochondrial inner membrane	**69**	GO:0021537	telencephalon development
20	GO:0061659	ubiquitin-like protein ligase activity	**70**	GO:0006900	vesicle budding from membrane
21	GO:0007030	Golgi organization	**71**	GO:0097711	ciliary basal body-plasma membrane docking
22	GO:0022406	membrane docking	**72**	GO:2001020	regulation of response to DNA damage stimulus
23	GO:0048193	Golgi vesicle transport	**73**	GO:0005635	nuclear envelope
24	GO:0061630	ubiquitin protein ligase activity	**74**	GO:0140014	mitotic nuclear division
25	GO:0006914	autophagy	**75**	GO:0006260	DNA replication
26	GO:0061919	process utilizing autophagic mechanism	**76**	GO:0071897	DNA biosynthetic process
27	GO:0000775	chromosome, centromeric region	**77**	GO:0051648	vesicle localization
28	GO:0000819	sister chromatid segregation	**78**	GO:0006913	nucleocytoplasmic transport
29	GO:0000075	cell cycle checkpoint	**79**	GO:0006979	response to oxidative stress
30	GO:0009896	positive regulation of catabolic process	**80**	GO:0048194	Golgi vesicle budding
31	GO:0140056	organelle localization by membrane tethering	**81**	GO:0051650	establishment of vesicle localization
32	GO:0043161	proteasome-mediated ubiquitin-dependent protein catabolic process	**82**	GO:0036064	ciliary basal body
33	GO:0016887	ATPase activity	**83**	GO:0031625	ubiquitin protein ligase binding
34	GO:0045296	cadherin binding	**84**	GO:1903362	regulation of cellular protein catabolic process
35	GO:0031252	cell leading edge	**85**	GO:0030496	midbody
36	GO:0031331	positive regulation of cellular catabolic process	**86**	GO:0034470	ncRNA processing
37	GO:0098687	chromosomal region	**87**	GO:0000922	spindle pole
38	GO:0042176	regulation of protein catabolic process	**88**	GO:0007265	Ras protein signal transduction
39	GO:0005911	cell-cell junction	**89**	GO:0006353	DNA-templated transcription, termination
40	GO:0198738	cell-cell signaling by wnt	**90**	GO:0006650	glycerophospholipid metabolic process
41	GO:1901990	regulation of mitotic cell cycle phase transition	**91**	GO:1903322	positive regulation of protein modification by small protein conjugation or removal
42	GO:0016055	Wnt signaling pathway	**92**	GO:0034976	response to endoplasmic reticulum stress
43	GO:0061695	transferase complex, transferring phosphorus-containing groups	**93**	GO:0031396	regulation of protein ubiquitination
44	GO:0000209	protein polyubiquitination	**94**	GO:0035091	phosphatidylinositol binding
45	GO:0045930	negative regulation of mitotic cell cycle	**95**	GO:0008380	RNA splicing
46	GO:1901987	regulation of cell cycle phase transition	**96**	GO:0006605	protein targeting
47	GO:0051186	cofactor metabolic process	**97**	GO:0007050	cell cycle arrest
48	GO:0005874	microtubule	**98**	GO:0045732	positive regulation of protein catabolic process
49	GO:0099568	cytoplasmic region	**99**	GO:0000151	ubiquitin ligase complex
50	GO:0051098	regulation of binding	**100**	GO:0001890	placenta development

ID (ontology)	Description	geneID (top ~80)	Count
GO:0005813 (CC)	centrosome	FBXL7, DCAF13, GNAI1, KIF20B, PIBF1, MDM1, HNRNPU, MASTL, ORC2, CEP55, SPICE1, NUP107, CEP85L, ERCC6L2, CEP57L1, APC, CDC14B, TTC8, CCNB1, SSX2IP, CEP350, RAPGEF6, MACROD2, CEP120, CCDC15, RNF19A, PKHD1, TBC1D31, STIL, PPP4R3B, GEN1, TTLL5, TTC12, POC1B, CHEK1, TXNDC9, AKNA, ATP6V1D, ZNF322, CEP70, CEP295, OLA1, DTL, AKAP9, WDR35, PCGF5, PROCR, CEP152, NDRG1, VPS4B, FBXO31, PLK4, UBR4, TNKS2, MBNL1, SDCCAG8, DCTN4, BBS4, IFT74, RABGAP1, TFAP2A, CYLD, PLEKHA7, MAK, NEK1, CEP97, ALS2, DCLRE1B, KIF3A, CNTRL, PPP1R42, SORBS1, IFT80, TNKS, CDC27, CLIC4, KATNB1, CEP41, PATJ, CKAP2L, CDK5RAP2, TTC26, IST1, CENPF, MCM3…	407
GO:0005759 (CC)	mitochondrial matrix	BTD, HIBCH, MRPS36, DHTKD1, MCCC2, CCNB1, GLRX2, MTRF1L, DARS2, LYRM7, CREB1, ETFDH, PARS2, CBR4, NUDT9, MTERF2, ATXN3, ISCA2, DLD, IARS2, MRPL18, DBT, NUDT2, HYKK, ATP5E, PDK1, ARG2, GCSH, GSR, GLS, HADH, TFAM, PRIMPOL, SIRT5, NARS2, PDHX, FDX1, MALSU1, MCCC1, HSPE1, ALDH6A1, MRPS18C, NADK2, ETFA, ALDH4A1, ABCE1, GARS, OAT, MRPL32, AASS, CDK1, SSBP1, ACSM6, ACAD10, MRPL42, ACAD8, BCKDHB, IBA57, MMAA, ATP5F1, MRPS31, ACAT1, COQ3, FH, MTHFS, LIPT1, PDK3, PARK7, PDHB, MRPS14, MRPL30, MRPL17, GPT2, ATP5C1, FDXR, PPA2, PITRM1, NAXD, VDAC2, NDUFAF7, BDH1, GRPEL1, MLYCD, MRPS15, GLRX5, DIMT1, PCCB…	380
GO:0010256 (BP)	endomembrane system organization	BCAS3, TPR, OSBPL8, NUP107, VTA1, ANK2, CCNB1, TRIP11, VPS36, SYNE1, DYNC2H1, GOLGA5, TRAPPC11, CREB1, ARV1, CLCN3, PI4K2A, SH3TC2, AKAP9, CAV1, SH3GLB1, RAB33B, CCDC47, NDRG1, RAB3GAP2, VPS4B, TOR1AIP1, VMP1, VAPB, LEMD3, GOLGB1, GORASP2, GBF1, CHMP5, RAB5B, DNAJC13, ALS2, SEC23IP, OPTN, CRB1, PPP2R1A, ANK3, MPP5, PLSCR1, WHAMM, EHD3, IST1, RAB18, UBXN2A, USO1, PPP2CA, TSG101, HACE1, RAB38, STX6, SGIP1, VRK1, RTN4, SPAST, SNX3, CLASP2, HOOK1, CDK1, PIK3C3, VTI1A, CSNK1A1, USP8, GOLPH3L, LYST, HIKESHI, NUP133, TMEM43, SPAG4, DOPEY2, TRAPPC8, BLZF1, LMAN1, TBC1D20, RAB3GAP1, ARL6IP1, SNX19, RAB10, NUPL2…	339
GO:0005819 (CC)	spindle	TPR, KIF20B, HNRNPU, SPICE1, CDC14B, HECW2, CCNB1, BRCC3, CEP350, PKHD1, NEDD9, KATNA1, GEM, SPIN1, POC1B, APP, INVS, ATM, ACOT13, NEK7, RIF1, VPS4B, FAM83D, KIF20A, SEPT7, DCTN4, PRC1, HSPA2, CYLD, MAK, ASPM, KIF3A, TNKS, PKP4, CDC27, KIF14, STAG1, KATNB1, KIFC1, CKAP2L, CDK5RAP2, KIF11, CENPF, CDCA8, MMS19, PPP2CA, ECT2, FBXO5, VRK1, SPAST, CLASP2, PKD2, CDK1, CENPE, CDC20, GPSM2, NEK2, CKAP2, MAP7D1, SPDL1, NPM1, RTRAF, KIF2A, SKA1, TTK, PTP4A1, RANGAP1, FAM110A, DYNLT1, ALMS1, KATNBL1, RASSF10, CSPP1, RB1, KNTC1, KIF23, CEP128, CEP85, MAEA, MYH9, MZT2B, MAP4, POC1A, ABRAXAS2, IK, KLHL21…	275
GO:0051052 (BP)	regulation of DNA metabolic process	USP1, HNRNPU, CACYBP, IL2, MLH1, HELB, BRCC3, ESCO2, IGF1R, CCT2, FBXW7, CHEK1, HMBOX1, WAPL, SIRT1, UBR5, ATM, NEK7, RIF1, KDM1B, DCP2, ATR, HMGB1, TNKS2, JUN, TICRR, KDM4D, SLF2, WRNIP1, UIMC1, OGG1, PRKCQ, THOC1, STN1, FBXO18, BMPR2, PPP2R1A, TNKS, UBE2N, RAD17, RAD52, EYA4, PPP2CA, MSH3, SETMAR, ERCC4, MGMT, UNG, CDK1, GMNN, BLM, NEK2, SLF1, PTK2B, NUCKS1, NPM1, NUDT16, HNRNPD, E2F8, TIGAR, XRCC5, E2F7, DSCC1, LIG3, TERF2IP, HNRNPC, NVL, HNRNPA1, ANKRD17, DFFA, PKIB, CCT8, WRN, CXorf57, GTPBP4, BMP4, PARPBP, RPA2, CST3, RAC1, BCL6, CBX8, TRIP12, MBD2, PDGFC, PTGES3, ABRAXAS1, MBD1, CUL4A, EYA3…	333
GO:0060271 (BP)	cilium assembly	ABCC4, SPAG1, PIBF1, CDC14B, TTC8, SSX2IP, TBC1D7, OCRL, INTU, SPAG16, TRIP11, CEP120, TROVE2, DYNC2H1, PKHD1, TTLL5, POC1B, ATP6V1D, CEP70, AKAP9, WDR35, TMEM237, CEP152, TTC21B, PLK4, GORAB, MNS1, WDR19, SDCCAG8, SEPT7, BBS4, IFT74, CYLD, MAK, TNPO1, RAB3IP, NEK1, CEP97, TTC30B, GALNT11, KIF3A, CNTRL, WDR5B, IQUB, PPP2R1A, IFT80, BBS10, KIF27, CEP41, RFX3, CDK5RAP2, EHD3, TTC26, ACTR2, TCTN3, BBS9, KIAA0586, CEP126, CDK1, IFT122, CEP131, RP1L1, NEK2, DNAI1, C5orf30, RAB8A, TMEM67, ACTR1A, CFAP206, ATG5, ABLIM1, PCNT, IFT52, TBC1D32, CFAP20, SPAG17, CEP164, DYNC2LI1, ATMIN, LRGUK, ALMS1, FOPNL, VANGL2…	298
GO:0044839 (BP)	cell cycle G2/M phase transition	FBXL7, MASTL, CLSPN, TAOK3, CCNB1, FOXN3, CDC25C, CDK7, PSMD14, APP, CEP70, DTL, AKAP9, CCNH, ATM, CEP152, BORA, VPS4B, RPS27A, PLK4, BACH1, PSMA6, TICRR, ABCB1, SDCCAG8, HSPA2, PPM1D, MTA3, BTRC, TAF2, PPP1R12B, CNTRL, OPTN, PPP2R1A, RAD17, PSMA5, PSMC6, KIF14, CEP41, PSMA3, PPME1, ENSA, CDK5RAP2, CENPF, MIIP, PSMB1, CCNA2, SKP2, CDK1, CEP131, PSME4, BLM, NEK2, RAB8A, PSMD7, ARPP19, NPM1, ACTR1A, PCNT, CDK4, PKIA, PSMA1, PPP1CB, AKAP8L, CEP164, PSMD12, ALMS1, RAD51B, PPP1R12A, FBXW11, CENPJ, PSMB2, PSMC5, WEE1, PSMD1, CIT, CUL1, PCM1, H2AFY, CEP63, RAD51C, CDK5RAP3, ODF2, LATS1, HAUS1, PHLDA1, PSMC2…	225
GO:0044782 (BP)	cilium organization	ABCC4, SPAG1, PIBF1, CDC14B, TTC8, SSX2IP, TBC1D7, OCRL, INTU, SPAG16, TRIP11, CEP120, TROVE2, DYNC2H1, PKHD1, TTLL5, POC1B, ATP6V1D, CEP70, AKAP9, WDR35, TMEM237, CEP152, TTC21B, PLK4, GORAB, MNS1, WDR19, SDCCAG8, SEPT7, BBS4, BBS12, IFT74, CYLD, MAK, TNPO1, RAB3IP, NEK1, CEP97, TTC30B, GALNT11, KIF3A, CNTRL, WDR5B, IQUB, PPP2R1A, IFT80, BBS10, KIF27, CEP41, RFX3, CDK5RAP2, EHD3, TTC26, ACTR2, TCTN3, BBS9, KIAA0586, CEP126, CDK1, IFT122, CEP131, CFAP61, RP1L1, NEK2, DNAI1, C5orf30, RAB8A, TMEM67, ACTR1A, CFAP206, ATG5, ABLIM1, PCNT, RTTN, IFT52, TBC1D32, CFAP20, SPAG17, CEP164, DYNC2LI1, ATMIN, LRGUK, ALMS1, FOPNL…	304
GO:0000226 (BP)	microtubule cytoskeleton organization	BCAS3, ULK4, SPAG1, TPR, GNAI1, PIBF1, MDM1, HNRNPU, SPICE1, MLH1, SNCA, APC, CDC14B, FER, CCNB1, SSX2IP, RGS2, CEP350, SPAG16, CEP120, RNF19A, MAP7, PKHD1, STIL, KATNA1, GEN1, TTLL5, WASHC5, CHEK1, SPRY1, ATXN3, AKAP9, TBCE, SPC25, NEK7, CEP152, BORA, VPS4B, CENPA, PLK4, GADD45A, KIF20A, SDCCAG8, BBS4, PRC1, SON, CYLD, CHMP5, EFNA5, ASPM, STARD9, NAV3, CEP97, DIXDC1, KIF3A, PPP2R1A, TNKS, XPO1, STAG1, KATNB1, KIFC1, CDK5RAP2, KIF11, MARK1, TACC2, USP33, FBXO5, GAS2L3, SPAST, CLASP2, PKD2, HOOK1, CEP126, CDK1, CEP131, CENPE, CDC20, GPSM2, KPNB1, RP1L1, UVRAG, NEK2, CKAP2, DNAI1, CRIPT, MAP7D1, TMEM67, SPDL1…	391
GO:0010498 (BP)	proteasomal protein catabolic process	FBXL7, RHBDD1, EDEM3, ARIH1, RNF38, HSP90B1, APC, HECW2, TLK2, CCNB1, TMTC3, UBR3, BIRC2, RNF19A, PSMD14, GNA12, FBXW7, DNAJC10, SIRT1, ATXN3, CAV1, OS9, CCDC47, CDC23, RAD23B, RPS27A, FBXO31, UBE2C, RFFL, PSMA6, BUB3, ZNRF2, KCTD2, MDM2, NEDD4L, TRIM9, USP14, RNF103, WAC, RNF7, UBE2W, KLHL20, TRIB1, BTRC, RNF138, UGGT1, PSMA5, CDC27, PSMC6, GCLC, ERLIN1, KIF14, UBXN4, NUDT15, PSMA3, FAF1, RNF139, EDEM1, FBXL3, UBXN2A, HSPA5, UBAC2, PSMB1, PTTG1, SKP2, ARNTL, CUL4B, CDK1, CDC20, PSME4, CSNK1A1, CHFR, ALAD, TMEM67, SOCS5, PSMD7, PPP2R5C, ZNRF1, MTM1, DET1, PARK7, RNF5, UBE4B, SELENOS, PSMA1, USP5…	364

Appendix E - 5 : Top 10 (out of 1187) Seq2Gene GO analysis in Male Associated ATAC peaks (P-value < $1x10^{-12}$ for all terms shown)

IID	Description	geneID (top ~80)	Count
R-HSA-3700989	Transcriptional Regulation by TP53	FANCC, MLH1, RRM2B, CCNB1, CDC25C, CDK7, ELOC, MTOR, CCNK, CHEK1, TP53I3, GTF2H1, COX20, TAF10, CCNH, ATM, NDRG1, COX7A2L, FAS, RPS27A, TXN, RFFL, TP53RK, ATR, GADD45A, JUN, CNOT2, CNOT3, YWHAQ, TBP, MDM2, JMY, EXO1, GSR, GLS, TP53BP2, CASP1, ATF2, CNOT4, AGO3, TAF2, PPP2R1A, RAD17, POLR2B, BRIP1, RBBP4, RMI2, RHEB, PPP2CA, CYCS, CCNA2, COX14, ING2, CSNK2A1, KMT5A, CDK1, CNOT10, NELFCD, BLM, TAF13, MAPKAP1, PPP2R5C, NPM1, CNOT1, SMYD2, PRDX1, PRKAA2, E2F8, TIGAR, RRAGC, POLR2K, PLAGL1, E2F7, COX5B, WRN, RBL2, POLR2H, RPA2, BCL6, COX18, CNOT6, TAF3, TAF5, GTF2H5, TAF1, TP63, GTF2H4,…	311
R-HSA-1852241	Organelle biogenesis and maintenance	TTC8, TRIP11, DYNC2H1, CREB1, IMMT, CCT2, CEP70, AKAP9, EXOC4, WDR35, CAMK4, CEP152, TTC21B, CHCHD3, PLK4, CYS1, WDR19, SDCCAG8, BBS4, BBS12, IFT74, ATP5G1, GBF1, TNPO1, RAB3IP, ATP5E, CEP97, ATF2, TTC30B, KIF3A, CNTRL, TFAM, PPP2R1A, IFT80, SIRT5, EXOC1, BBS10, CEP41, EHHADH, CDK5RAP2, TTC26, ATP5G3, ARF4, NRF1, CYCS, TCTN3, BBS9, PKD2, CDK1, SSBP1, IFT122, CEP131, NEK2, CHD9, ATP5F1, CNGA4, RAB8A, TMEM67, ACTR1A, ATP5G2, PCNT, MTX2, MINOS1, PRKAA2, IFT52, ATP5C1, CEP164, DYNC2LI1, CCT8, ALMS1, PDE6D, MED1, GABPA, DYNLRB2, CENPJ, TTC30A, EXOC2, SOD2, PCM1, CHCHD6, HDAC6, CEP63, TGS1, DYNLRB1…	255
R-HSA-69620	Cell Cycle Checkpoints	ORC2, CLSPN, NUP107, MCM10, CENPO, CCNB1, BRCC3, CDC25C, PSMD14, CHEK1, CENPI, ATM, CDC23, SPC25, MCM8, CENPN, CENPA, RPS27A, UBE2C, BUB1, ATR, PSMA6, BUB3, SGO2, UIMC1, YWHAQ, CENPU, MDM2, EXO1, PPP2R5E, PPP2R1A, XPO1, UBE2N, RAD17, PSMA5, CDC27, PSMC6, BRIP1, RMI2, ORC3, PSMA3, CENPF, MCM3, CDCA8, PPP2CA, CENPH, PSMB1, CCNA2, HIST1H4A, CLASP2, CDK1, CENPE, CDC20, PSME4, BLM, SPDL1, PSMD7, NUP133, PPP2R5C, PPP2R5D, KIF2A, SKA1, CENPC, RANGAP1, CENPP, DBF4, PSMA1, PSMD12, WRN, RPA2, KNTC1, NUF2, PSMB2, PSMC5, ABRAXAS1, WEE1, PSMD1, UBE2E1, MDC1, ANAPC15, RNF168, RFC2,…	251
R-HSA-166520	Signalling by NGF	FRS2, RPS6KA5, PRKCI, CREB1, ARHGEF37, PIK3R1, RPS27A, DUSP6, GNA13, ADAM17, PPP2R1A, VAV3, PPP2CA, TRIO, ECT2, RTN4, PPP2R5D, SOS1, MAPK7, DNM3, KRAS, RIPK2, RAC1, VRK3, AKAP13, MAPKAPK3, PRDM4, CASP2, ARHGEF26, AP2B1, KALRN, STAT3, AP2M1, RALA, ARHGEF10L, ARHGEF10, NFKB1, RALB, ARHGEF2, NET1, CLTA, HDAC1, CRK, MAPK8, MAPK14, NTRK2, PPP2CB, RAP1A, ARHGEF3, IRS2, IKBKB, SMPD2, PPP2R1B, NRAS, MYD88, BCL2L11, TRAF6, PLEKHG5, HDAC2, CRKL, GRB2, PSEN1, FGD4, AP2S1, ARHGEF7, ARHGEF38, PSEN2, RPS6KA3, CLTC, PIK3R2, PLCG1, RPS6KA1, RALGDS, NFKBIA, BAD, KIDINS220, AATF, PCSK5, BRAF, DNM1, HRAS, ARHGEF1,…	154
R-HSA-5617833	Cilium Assembly	TTC8, TRIP11, DYNC2H1, CCT2, CEP70, AKAP9, EXOC4, WDR35, CEP152, TTC21B, PLK4, CYS1, WDR19, SDCCAG8, BBS4, BBS12, IFT74, GBF1, TNPO1, RAB3IP, CEP97, TTC30B, KIF3A, CNTRL, PPP2R1A, IFT80, EXOC1, BBS10, CEP41, EHHADH, CDK5RAP2, TTC26, ARF4, TCTN3, BBS9, PKD2, CDK1, IFT122, CEP131, NEK2, CNGA4, RAB8A, TMEM67, ACTR1A, PCNT, IFT52, CEP164, DYNC2LI1, CCT8, ALMS1, PDE6D, DYNLRB2, CENPJ, TTC30A, EXOC2, PCM1, HDAC6, CEP63, DYNLRB1, MKS1, ODF2, HAUS1, DYNLL2, TRAF3IP1, FBF1, PLK1, RAB11A, DCTN2, C2CD3, HAUS6, CEP192, FGFR1OP, CEP78, TUBG1, IFT88, BBS7, TTBK2, ASAP1, IFT57, CLASP1, ARL13B, WDR60, BBS2, SEPT2, SSNA1, PAFAH1B1, HSP90AA1, CCT5, DYNC1I2, DYNLL1, NEDD1, CNGB1, IFT140, KIFAP3, EXOC6, TUBB4B, CCP110, CCT4, BBS5, RP2, TCTN1, PRKACA, IFT172…	175
R-HSA-204005	Coat Protein 2 Mediated Vesicle Transport	SEC24A, SEC23IP, TRAPPC3, USO1, TRAPPC9, CTSC, ANKRD28, F5, SEC31A, GOSR2, SEC22C, LMAN1, BET1, TBC1D20, TRAPPC4, RAB1A, CNIH1, SEC24D, SEC16B, STX17, SCFD1, NAPA, SEC24B, SEC13, TRAPPC6A, TRAPPC6B, TFG, SEC22A, TGFA, RAB1B, PPP6C, CD59, NAPB, YKT6, LMAN2, SEC24C, SAR1B, SEC23A, TMED2, SERPINA1, LMAN2L, PREB, NAPG, TRAPPC10, STX5, LMAN1L, TRAPPC2L, AREG, F8, CNIH3, GRIA1, NSF, GOLGA2, PPP6R3, TRAPPC1, MCFD2, TRAPPC2, CSNK1D, TMED10, PPP6R1, GORASP1, SEC16A, COL7A1, SEC22B, CTSZ	65
R-HSA-5620912	Anchoring basal body to plasma membrane	CEP70, AKAP9, CEP152, PLK4, SDCCAG8, RAB3IP, CEP97, CNTRL, PPP2R1A, CEP41, CDK5RAP2, TCTN3, CDK1, CEP131, NEK2, RAB8A, TMEM67, ACTR1A, PCNT, CEP164, ALMS1, CENPJ, PCM1, CEP63, MKS1, ODF2, HAUS1, FBF1, PLK1, RAB11A, DCTN2, C2CD3, HAUS6, CEP192, FGFR1OP, CEP78, TUBG1, TTBK2, CLASP1, SEPT2, SSNA1, PAFAH1B1, HSP90AA1, DYNC1I2, DYNLL1, NEDD1, TUBB4B, CCP110, TCTN1, PRKACA, HAUS7, DCTN3, MAPRE1, CEP72, TUBB, B9D1, HAUS4, CEP250, TCTN2, DYNC1H1, KIF24, CEP76, CC2D2A, AHI1, NPHP1, PRKAR2B, OFD1, CSNK1E, NINL, SFI1, NDE1, YWHAE, SCLT1, CEP162, CETN2, CEP135, IQCB1, NPHP4, HAUS8, CSNK1D, CEP57..	90
R-HSA 453279	Mitotic G1-G1/S phases	ORC2, MCM10, CCNB1, JAK2, CDK7, PSMD14, E2F3, LIN54, CCNH, MCM8, RPS27A, PSMA6, MAX, PPP2R1A, PSMA5, PSMC6, CDKN2C, RBBP4, ORC3, PSMA3, MCM3, PPP2CA, PSMB1, CCNA2, FBXO5, SKP2, CDK1, PSMD7, CDK4, DBF4, PSMA1, PSMD12, RBL2, RPA2, RB1, LIN9, E2F6, PSMB2, PSMC5, WEE1, PSMD1, CUL1, AKT2, POLA1, POLE4, PSMC2, PSMB7, TFDP2, MCM2, PSMD9, CKS1B, TOP2A, RRM2, CDC7, PSMD11, HDAC1, CDC45, PSMC4, CCNE1, SEM1, PSMD3, MCM4, PSMB3, PPP2CB, LIN37, PPP2R1B, MYC, ORC5, MCM6, PPP2R2A, PSMB9, PSMD2, CDKN2B, TFDP1, CDC6, PSMD13, E2F1, PSMF1, PSMB10, CDC25A, SKP1, MCM5, CDK6, CDKN1A, PSME3, LYN,…	130
R-HSA-73894	DNA Repair	USP1, CLSPN, FANCC, ACTL6A, MLH1, BRCC3, COPS4, YY1, TDP1, CDK7, GEN1, CHEK1, SPRTN, GTF2H1, COPS5, DTL, CCNH, ATM, RIF1, RAD23B, RPS27A, ATR, UIMC1, OGG1, DCLRE1C, EXO1, XRCC4, DCLRE1B, POLI, ASCC2, UBE2N, RAD17, POLR2B, ASCC3, BRIP1, RAD52, RMI2, USP45, RUVBL1, EYA4, CCNA2, MSH3, HIST1H4A, ERCC4, MGMT, TDG, UNG, INO80C, CUL4B, COPS8, BLM, ERCC6, XPA, SMARCA5, POLR2K, XRCC5, POLK, POLN, LIG3, FANCF, RNF111, ERCC5, WRN, CHD1L, RAD51B, POLR2H, RPA2, PIAS1, GTF2H5, GTF2H4, ELL, ABRAXAS1, KDM4A, CUL4A, EYA3, PCLAF, MDC1, RNF168, RFC2, POLD1, RAD51C, FTO, XRCC2, POLM, POLE4, VCP, FAAP24, AQR,…	263
R-HSA-141424	Amplification of signal from the kinetochores	NUP107, CENPO, CENPI, SPC25, CENPN, CENPA, BUB1, BUB3, SGO2, CENPU, PPP2R5E, PPP2R1A, XPO1, CENPF, CDCA8, PPP2CA, CENPH, CLASP2, CENPE, CDC20, SPDL1, NUP133, PPP2R5C, PPP2R5D, KIF2A, SKA1, CENPC, RANGAP1, CENPP, KNTC1, NUF2, DYNC1LI1, DYNLL2, SEC13, NDEL1, PLK1, MAD2L1, MIS12, ZWINT, MAD1L1, NDC80, ZWILCH, RANBP2, ZW10, SEH1L, PPP2CB, NUP43, CLASP1, INCENP, SPC24, ERCC6L, PPP2R1B, PAFAH1B1, DYNC1I2, DYNLL1, KIF2C, DYNC1I1, CENPQ, CENPT, NUDC, MAPRE1, NUP98, CENPM, AURKB, DYNC1H1, DYNC1LI2, NUP85, PPP1CC, BIRC5, DSN1, NDE1, ITGB3BP, KIF18A, CENPK, SKA2, PMF1, CLIP1, NSL1, PPP2R5A, NUP37…	88

Appendix E - 7 : Top 10 (out of 166) Seq2Gene Reactome analysis in Male Associated ATAC peaks (P-value < $1x10^{-7}$ for all terms shown)

ID	Description	geneID (top ~80)	Count
hsa05205	Proteoglycans in cancer	FRS2, ANK2, ITGAV, MTOR, IGF1R, FZD6, PIK3R1, CAMK2D, HIF1A, CAV1, FAS, COL21A1, ITGB1, MDM2, PPP1R12B, GAB1, RPS6, ANK3, ITPR1, CD44, PIK3R3, IQGAP1, CTSL, DDX5, PLCE1, RDX, SOS1, ESR1, ROCK2, KRAS, PLCG2, PPP1CB, CAV2, FZD1, SDC2, TGFB2, RAC1, PPP1R12A, PDCD4, SMAD2, AKT2, PAK1, STAT3, EIF4B, PTK2, WNT8B, HPSE, TLR2, ITPR2, FGF2, RPS6KB1, MAPK14, ITGA5, PRKCA, PTPN6, MYC, NRAS, RAF1, FN1, WNT2B, CTNNB1, CDC42, ERBB2, GRB2, MRAS, SDC1, PPP1R12C, PDPK1, MMP2, PRKACA, ROCK1, MAP2K1, PIK3R2, PLCG1, EZR, DROSHA, HSPB2, ITGB5, BRAF, HRAS, SLC9A1, FZD2, ERBB4, ARHGEF1, CDKN1A, ELK1, FZD8, NUDT16L1, TWIST2, ARAF...	174
hsa04141	Protein processing in endoplasmic reticulum	EDEM3, DNAJB11, HSP90B1, HSPA4L, DNAJC10, STT3A, ATXN3, DNAJB2, OS9, PDIA4, RAD23B, CANX, SSR3, HSPA2, ERO1A, ERO1B, HSPA8, SEC24A, SEC23B, HSPH1, MAN1A2, SSR1, UGGT1, SEC61A2, EDEM1, HSPA5, HSPA6, P4HB, EIF2AK4, DNAJC3, SEC62, RRBP1, MBTPS1, SEC31A, RNF5, UBE4B, SELENOS, LMAN1, CAPN2, DERL1, NFE2L2, SEL1L, UGGT2, SEC24D, SEC63, ERP29, SAR1A, UBQLN1, RPN2, CUL1, PDIA6, UBE2J1, DAD1, UBE2D1, ATF6B, DNAJA1, SEC24B, TUSC3, SEC13, VCP, UBE2G2, MAPK9, MAN1B1, HSPBP1, UBE2D2, WFS1, NGLY1, MAPK8, DERL3, TRAF2, SEC61A1, DNAJA2, BCAP31, DDOST, SIL1, Mar-06, HSP90AA1, ERN1, XBP1, TXNDC5, HYOU1, AMFR, EIF2AK1...	144
hsa04012	ErbB signaling pathway	ABL2, MTOR, PIK3R1, CAMK2D, JUN, GAB1, PIK3R3, SOS1, PAK2, KRAS, PLCG2, NCK1, CBLB, AKT2, PAK1, EGF, PTK2, MAPK9, BTC, NRG1, RPS6KB1, CRK, MAPK8, TGFA, PRKCA, MYC, NRAS, RAF1, CRKL, ERBB2, GRB2, PAK4, MAP2K4, MAP2K1, PIK3R2, PLCG1, BAD, MAP2K7, BRAF, HRAS, ERBB4, CDKN1A, ELK1, PAK3, NRG2, PAK6, ARAF, CAMK2G, MAP2K2, HBEGF, EGFR, SHC3, AKT3, NRG3, NRG4, AREG, PIK3CA, SHC4, MAPK10, GSK3B, CBL, EIF4EBP1, ABL1, NCK2, CAMK2A, MAPK1, PIK3CD, EREG, CAMK2B, SHC1, SRC, CDKN1B, SOS2, ERBB3, PIK3CB, PRKCG, STAT5A, PRKCB	78
hsa04722	Neurotrophin signaling pathway	FRS2, RPS6KA5, PIK3R1, CAMK2D, CAMK4, JUN, GAB1, PIK3R3, SOS1, MAPK7, IRAK3, KRAS, PLCG2, RIPK2, RAP1B, RAC1, FOXO3, PRDM4, AKT2, IRAK4, MAPK9, NFKB1, BDNF, CRK, MAPK8, NFKBIB, MAPK14, NTRK2, RAP1A, IKBKB, NRAS, RAF1, SH2B1, ZNF274, TRAF6, CALML3, CRKL, CDC42, GRB2, PSEN1, PDPK1, PSEN2, RPS6KA3, MAP2K1, PIK3R2, PLCG1, RPS6KA1, NFKBIA, BAD, KIDINS220, MAP2K7, BRAF, IRAK2, HRAS, TP73, CAMK2G, MAGED1, MAP3K1, NGF, MAP2K2, MAP2K5, RELA, BAX, YWHAE, SHC3, AKT3, MAP3K5, RPS6KA2, FASLG, PIK3CA, SHC4, MAPK10, GSK3B, ARHGDIB, IRS1, ABL1, RAPGEF1, BCL2, MAPKAPK2, CAMK2A, MAPK1, PIK3CD, NGFR, NTRK3, CAMK2B, SH2B3...	105
hsa05222	Small cell lung cancer	ITGAV, LAMB4, PTGS2, BIRC2, E2F3, PIK3R1, GADD45G, XIAP, GADD45A, LAMA3, ITGB1, BIRC3, MAX, LAMA2, CYCS, PIK3R3, SKP2, CDK4, POLK, LAMC2, RB1, APAF1, LAMB1, AKT2, CKS1B, PTK2, NFKB1, LAMA4, CCNE1, TRAF2, IKBKB, CHUK, MYC, FN1, COL4A6, TRAF6, LAMC1, CDKN2B, COL4A5, E2F1, ITGA6, NOS2, IKBKG, DDB2, ZBTB17, PIK3R2, FHIT, NFKBIA, LAMA1, CDK6, COL4A4, COL4A1, RXRA, CDKN1A, TRAF5, ITGA2, ITGA3, COL4A2, CDK2, RELA, BAK1, BAX, AKT3, PIK3CA, CASP3, TRAF3, RARB, BCL2, PIK3CD, BCL2L1, TRAF1, E2F2, LAMC3, LAMA5, COL4A3, CKS2, CCNE2, TP53, CDKN1B, LAMB3, PIK3CB, RXRG, ITGA2B, CCND1	84
hsa04110	Cell cycle	ORC2, CDC14B, CCNB1, CDC25C, CDK7, E2F3, CHEK1, GADD45G, CCNH, ATM, CDC23, BUB1, ATR, GADD45A, BUB3, YWHAQ, MDM2, CDC27, STAG1, CDKN2C, ORC3, MCM3, CCNA2, PTTG1, SKP2, CDK1, CDC20, CDK4, SMAD3, TTK, DBF4, RBL2, TGFB2, RB1, SMAD2, WEE1, CUL1, CDC14A, ANAPC4, PLK1, TFDP2, SMAD4, MCM2, MAD2L1, MAD1L1, CDC7, HDAC1, CDC45, CCNE1, MCM4, ANAPC5, CCNB2, MYC, ORC5, MCM6, HDAC2, CDKN2B, TFDP1, YWHAZ, CDC6, ANAPC11, E2F1, RBX1, MAD2L2, CDC25A, ZBTB17, SKP1, MCM5, CDK6, CDKN1A, FZR1, CDKN2D, ANAPC7, CDC16, E2F5, RBL1, CCND3, SFN, CDKN1C, YWHAH, YWHAB, MCM7, CCNA1, CDK2, CCNB3, YWHAE, ANAPC10, GSK3B,...	108
hsa04390	Hippo signaling pathway	APC, PRKCI, BIRC2, MOB1A, BMPR1B, FZD6, SMAD1, STK3, TGFBR1, TCF7L2, YWHAQ, TGFBR2, BMPR2, TP53BP2, BTRC, CRB1, PPP2R1A, MPP5, LLGL2, YAP1, RASSF6, PPP2CA, BMPR1A, SMAD3, TEAD1, PPP1CB, BMP4, FZD1, TGFB2, SMAD2, FBXW11, LATS1, ID2, NF2, SMAD4, WWTR1, TCF7, WNT8B, CTGF, MOB1B, WWC1, PPP2CB, PPP2R1B, MYC, SMAD7, PPP2R2A, CTNNA1, WNT2B, PRKCZ, YWHAZ, CTNNB1, FRMD6, NKD1, DLG3, BMP5, RASSF1, CDH1, FZD2, AXIN2, BBC3, BBC3, FZD8, TEAD3, TP73, ID1, PPP1CC, CCND3, LLGL1, PPP2R2B, BIRC5, CSNK1E, FZD10, FZD3, YWHAH, YWHAB, SAV1, PARD6A, LATS2, YWHAE, WNT16, AREG, FGF1, GSK3B, PPP2R2C, CTNNA3, PARD3, CTNNA2,...	131
hsa05215	Prostate cancer	HSP90B1, CREB5, CREB1, MTOR, IGF1R, CREB3L2, E2F3, ETV5, PIK3R1, FGFR2, TCF7L2, MDM2, PIK3R3, SOS1, KRAS, FOXO1, RB1, PDGFC, CREB3, AKT2, EGF, TCF7, NFKB1, PDGFRA, CCNE1, TGFA, IKBKB, CHUK, NRAS, RAF1, HSP90AA1, ZEB1, CTNNB1, ERBB2, GRB2, E2F1, NKX3-1, PDPK1, IKBKG, CREB3L4, PDGFD, MAP2K1, PIK3R2, NFKBIA, BAD, BRAF, HRAS, CDKN1A, ARAF, GSTP1, MAP2K2, PLAT, CDK2, RELA, HSP90AB1, EGFR, AKT3, IL1R2, PIK3CA, GSK3B, BCL2, ERG, AR, TCF7L1, MAPK1, IGF1, PIK3CD, LEF1, SRD5A2, SPINT1, E2F2, CCNE2, TP53, CDKN1B, SOS2, ATF4, CREBBP, MMP3, PIK3CB, MMP9, CREB3L1, PLAU, TMPRSS2, PDGFB, CCND1, PDGFA	86
hsa05214	Glioma	MTOR, IGF1R, E2F3, PIK3R1, CAMK2D, GADD45G, CAMK4, GADD45A, MDM2, PIK3R3, SOS1, CDK4, POLK, KRAS, PLCG2, RB1, AKT2, EGF, PDGFRA, TGFA, PRKCA, NRAS, RAF1, CALML3, GRB2, E2F1, DDB2, MAP2K1, PIK3R2, PLCG1, CDK6, BRAF, HRAS, CDKN1A, ARAF, CAMK2G, MAP2K2, BAK1, EGFR, BAX, CAMK1, SHC3, AKT3, PIK3CA, SHC4, CAMK1D, CAMK2A, MAPK1, IGF1, PIK3CD, CAMK2B, E2F2, SHC1, CALM1, CALM2, CALM3, TP53, SOS2, PIK3CB, CALML4, CAMK1G, CALML6, PRKCG, CALML5, PDGFB, PRKCB, CCND1, PDGFA	68
hsa04070	Phosphatidylinositol signaling system	INPP4B, SYNJ2, OCRL, PIKFYVE, MTMR4, PI4K2A, PIK3R1, SACM1L, MTMR7, IPPK, PIK3C2A, INPP5B, ITPR1, IMPA1, PPIP5K2, PIK3R3, PLCE1, PIK3C3, DGKE, IPMK, MTMR14, MTM1, CDS2, INPP5F, PLCG2, PLCD4, IP6K1, PIK3C2B, MTMR2, PIP4P1, DGKH, IP6K2, ITPR2, DGKG, PI4KB, ITPK1, PIP4P2, PRKCA, PIP4K2B, IMPAD1, CALML3, PIP4K2C, DGKZ, INPP1, PLCB4, ITPKC, PIK3R2, PLCG1, PIK3C2G, PIP4K2A, DGKA, DGKI, INPP4A, PIP5K1B, PLCB1, MTMR6, MTMR3, PLCD1, DGKB, PIK3CA, PLCZ1, DGKD, CDS1, ITPKB, SYNJ1, PI4K2B, INPP5A, PIK3CD, MTMR1, INPP5D, IMPA2, PLCD3, CALM1, CALM2, CALM3, INPP5K, PI4KA, PIK3CB, CALML4, ITPKA, CALML6, PRKCG, PLCB2, CALML5, PIP5K1A, PRKCB, INPPL1	87

Appendix E - 9 : Top 10 (out of 115) Seq2Gene KEGG pathway analysis in Male Associated ATAC peaks (P-value < $1x10^{-5}$ for all terms shown)

Appendix E - 11 : Top 10 (out of 16) Seq2Gene GO analysis in Female Associated ATAC peaks

ID (ontology)	Description	pvalue	geneID	Count
GO:0016010 (CC)	dystrophin-associated glycoprotein complex	2.68E-05	SGCD, SNTG2, SNTG1	3
GO:0090665 (CC)	glycoprotein complex	2.68E-05	SGCD, SNTG2, SNTG1	3
GO:0033267 (CC)	axon part	0.00015149	UNC13C, EPHA4, PTPRN2, DLG2, COBL	5
GO:0097060 (CC)	synaptic membrane	0.0001712	UNC13C, LRRC4C, CLSTN2, EPHA4, TENM2, DLG2	6
GO:0045211 (CC)	postsynaptic membrane	0.0004143	LRRC4C, CLSTN2, EPHA4, TENM2, DLG2	5
GO:0098794 (CC)	postsynapse	0.0011645	LRRC4C, CLSTN2, EPHA4, TENM2, MAP2, DLG2	6
GO:0044295 (CC)	axonal growth cone	0.00157075	EPHA4, COBL	2
GO:0014069 (CC)	postsynaptic density	0.00256465	CLSTN2, EPHA4, MAP2, DLG2	4
GO:0099572 (CC)	postsynaptic specialization	0.00261031	CLSTN2, EPHA4, MAP2, DLG2	4
GO:0032279 (CC)	asymmetric synapse	0.00279856	CLSTN2, EPHA4, MAP2, DLG2	4

Appendix E - 12 : Top 10 (out of 14) Seq2Gene Reactome pathway analysis in Female Associated ATAC peaks

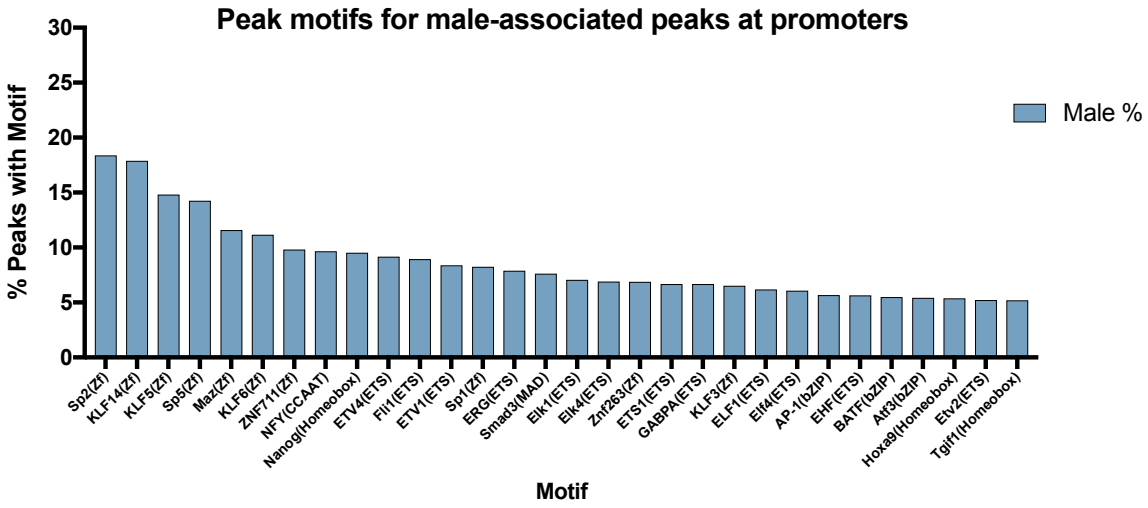
ID	Description	pvalue	geneID	Count
R-HSA-3781865	Diseases of glycosylation	0.00019052	LARGE1, GPC5, GPC6, MUC12	4
R-HSA-3656237	Defective EXT2 causes exostoses 2	0.00035194	GPC5, GPC6	2
R-HSA-3656253	Defective EXT1 causes exostoses 1, TRPS2 and CHDS	0.00035194	GPC5, GPC6	2
R-HSA-3560783	Defective B4GALT7 causes EDS, progeroid type	0.00084934	GPC5, GPC6	2
R-HSA-3560801	Defective B3GAT3 causes JDSSDHD	0.00084934	GPC5, GPC6	2
R-HSA-4420332	Defective B3GALT6 causes EDSP2 and SEMDJL1	0.00084934	GPC5, GPC6	2
R-HSA-2024096	HS-GAG degradation	0.00102987	GPC5, GPC6	2
R-HSA-1971475	A tetrasaccharide linker sequence is required for GAG synthesis	0.00144127	GPC5, GPC6	2
R-HSA-2022928	HS-GAG biosynthesis	0.00204846	GPC5, GPC6	2
R-HSA-3560782	Diseases associated with glycosaminoglycan metabolism	0.00306825	GPC5, GPC6	2

Appendix E - 13 : Only result from Seq2Gene KEGG pathway analysis in Female Associated ATAC peaks

ID	Description	pvalue	geneID	Count
hsa05412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	0.00075356	CTNNA3, CACNA2D3, SGCD	3

Appendix E - 14 : Table of 56 genes that have female-associated peaks but no male peaks

Symbol	Gene	Peaks
AMOT	angiomotin	chrX:112076166,chrX:112084125,chrX:112084839, chrX:112100885, chrX:112211332, chrX:112277376
PCDH20	protocadherin 20	chr13:61989354,chr13:62106068, chr13:62269525, chr13:62302045, chr13:62487523, chr13:62868553
LOC100133050	glucuronidase beta pseudogene	chr5:99038477, chr5:99236318, chr5:99236318, chr5:99340870
LRTM1	leucine rich repeats and transmembrane domains 1	chr3:55033434, chr3:55095882, chr3:55194646, chr3:55226233
CHL1	cell adhesion molecule L1 like	chr3:467962, chr3:477056, chr3:482764
CLEC14A	C-type lectin domain containing 14A	chr14:38724738, chr14:38812214, chr14:38873422
MAFB	MAF bZIP transcription factor B	chr20:39118768, chr20:39122400, chr20:39193856
MIR4764	microRNA 4764	chr22:33747139, chr22:33776953, chr22:33834395
ANXA2P3	annexin A2 pseudogene 3	chr10:66716328, chr10:66954086
C10orf142	chromosome 10 open reading frame 142	chr10:44691110, chr10:44796737
CHEK2P2	checkpoint kinase 2 pseudogene 2	chr15:20546535, chr15:20563112
IZUMO1R	IZUMO1 receptor, JUNO	chr11:93971643, chr11:93971643
LINC00839	long intergenic non-protein coding RNA 839	chr10:42971096, chr10:42971096
LSP1P3	lymphocyte-specific protein 1 pseudogene 3	chr5:28408506, chr5:28543347
MIR595	microRNA 595	chr7:158224891, chr7:158225732
PLAC9P1	placenta specific 9 pseudogene 1	chr2:130457526, chr2:130469534
PLD5	phospholipase D family member 5	chr1:242589777, chr1:242694269
PTPRR	protein tyrosine phosphatase, receptor type R	chr12:71153204, chr12:71270935
MIR125B2	microRNA 125b-2	chr21:17960887
MIR1297	microRNA 1297	chr13:54785018
MIR4307	microRNA 4307	chr14:27311316
MIR4444-1	microRNA 4444-1	chr3:75334867
MIR4675	microRNA 4675	chr10:20887725
CNTN6	contactin 6	chr3:1129136
CPXCR1	CPX chromosome region, candidate 1	chrX:87991025
CRNDE	colorectal neoplasia differentially expressed	chr16:54963170
CYSLTR1	cysteinyl leukotriene receptor 1	chrX:77613391
DUSP5P1	dual specificity phosphatase 5 pseudogene 1	chr1:228756788
ESPNP	espin pseudogene	chr1:17045397
FAM46D	family with sequence similarity 46 member D	chrX:79590836
FUT9	fucosyltransferase 9	chr6:96683965
GAS7	growth arrest specific 7	chr17:9989578
ITM2A	integral membrane protein 2A	chrX:78885932
KDM6A	lysine demethylase 6A	chrX:44795101
KLF8	Kruppel like factor 8	chrX:56055791
LINC00648	long intergenic non-protein coding RNA 648	chr14:48820361
LINC00664	long intergenic non-protein coding RNA 664	chr19:21646776
LINC00692	long intergenic non-protein coding RNA 692	chr3:25937152
LINC00841	long intergenic non-protein coding RNA 841	chr10:44409344
LINC00929	long intergenic non-protein coding RNA 929	chr15:26271536
LINC00939	long intergenic non-protein coding RNA 939	chr12:126540260
LINC01822	long intergenic non-protein coding RNA 1822	chr2:21691011
LINC02346	long intergenic non-protein coding RNA 2346	chr15:26129181
LOC100128239	uncharacterized LOC100128239	chr11:133904083
MS4A6A	membrane spanning 4-domains A6A	chr11:59936910
MUC12	mucin 12, cell surface associated	chr7:100607835
OR52J3	olfactory receptor family 52 subfamily J member 3	chr11:5063479
PCDH11X	protocadherin 11 X-linked	chrX:91502249
SI	sucrase-isomaltase	chr3:164176940
ST6GAL2	ST6 beta-galactoside alpha-2,6-sialyltransferase 2	chr2:107954105
TECRL	trans-2,3-enoyl-CoA reductase like	chr4:64987730
TOX3	TOX high mobility group box family member 3	chr16:52439560
TPTE	transmembrane phosphatase with tensin homology	chr21:10623446
TRIM58	tripartite motif containing 58	chr1:248020781
XIST	X inactive specific transcript (non-protein coding)	chrX:73070996



Appendix E - 15 : TF peak motifs found in male-associated peaks located at promoter regions

References

Ackermann, A.M., Wang, Z., Schug, J., Naji, A. and Kaestner, K.H. 2016. Integration of ATAC-seq and RNA-seq identifies human alpha cell and beta cell signature genes. *Molecular Metabolism*. 5, pp.233–244.

Agger, K., Cloos, P.A.C., Christensen, J., Pasini, D., Rose, S., Rappsilber, J., Issaeva, I., Canaani, E., Salcini, A.E. and Helin, K. 2007. UTX and JMJD3 are histone H3K27 demethylases involved in HOX gene regulation and development. *Nature*. 449, pp.731–734.

Ahn, J., Kim, K.H., Park, S., Ahn, Y., Kim, H.Y., Yoon, H., Lee, J.H., Bang, D. and Lee, D.H. 2016. Target sequencing and CRISPR / Cas editing reveal simultaneous loss of UTX and UTY in urothelial bladder cancer. *Oncotarget*. 7, pp.63252–63260.

Aird, D., Ross, M.G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D.B., Nusbaum, C. and Gnirke, A. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*. 12, pp.1-7.

Allahverdi, A., Chen, Q., Korolev, N. and Nordenskio, L. 2015. Chromatin compaction under mixed salt conditions : Opposite effects of sodium and potassium ions on nucleosome array folding. *Scientific Reports*. 5, pp1-6

Allfrey, V., Faulkner, R. and Mirsky, A. 1964. Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. *Proceedings of the National Academy of Sciences*. 315, pp.786–794.

Allory, Y., Beukers, W., Sagrera, A., Flández, M., Marqués, M., Márquez, M., Van Der Keur, K.A., Dyrskjot, L., Lurkin, I., Vermeij, M., Carrato, A., Lloreta, J., Lorente, J.A., Carrillo-De Santa Pau, E., Masius, R.G., Kogevinas, M., Steyerberg, E.W., Van Tilborg, A.A.G., Abas, C., Orntoft, T.F., Zuiverloon, T.C.M., Malats, N., Zwarthoff, E.C. and Real, F.X. 2014. Telomerase reverse transcriptase promoter mutations in bladder cancer: High frequency across stages, detection in urine, and lack of association with outcome. *European Urology*. 65, pp.360–366.

Amiri, A., Coppola, G., Scuderi, S., Wu, F., Roychowdhury, T., Liu, F., Pochareddy, S., Shin, Y., Safi, A., Song, L., Zhu, Y., Sousa, A.M.M., Gerstein, M., Crawford, G.E., Sestan, N., Abyzov, A. and Vaccarino, F.M. 2018. Transcriptome and epigenome landscape of human cortical development modeled in organoids. *Science*. 362. pp.1-8

An, Y., Li, H., Wang, K.J., Liu, X.H., Qiu, M.X., Liao, Y. and Huang, J.L. 2015. Meta-analysis of the relationship between slow acetylation of N-acetyl transferase 2 and the risk of bladder cancer. *Genetics and Molecular Research*. 14, pp.16896–16904.

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

Arnold, A.P. 2017. A general theory of sexual differentiation. *Journal of Neuroscience Research*. 95, pp.291–300.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. 2000. Gene Ontology: tool for the unification of biology The Gene Ontology Consortium*. *Nature Genetics*. 25, pp.25–29.

Axel, R. 1975. Cleavage of DNA in nuclei and chromatin with staphylococcal nuclease. *Biochemistry*. 14, pp.2921–2925.

Bannister, A.J. and Kouzarides, T. 2011. Regulation of chromatin by histone modifications. *Cell Research*. 21, pp.381–395.

Bannister, A.J., Schneider, R., Myers, F.A., Thorne, A.W., Crane-Robinson, C. and Kouzarides, T. 2005. Spatial distribution of di- and tri-methyl lysine 36 of histone H3 at active genes. *Journal of Biological Chemistry*. 280, pp.17732–17736.

Barna, M., Pusic, A., Zollo, O., Costa, M., Kondrashov, N., Rego, E., Rao, P.H. and Ruggero, D. 2008. Suppression of Myc oncogenic activity by ribosomal protein haploinsufficiency. *Nature*. 456, pp.971–975.

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I. and Zhao, K. 2007. High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*. 129, pp.823–837.

Bayne, C.E., Farah, D., Herbst, K.W. and Hsieh, M.H. 2018. Role of urinary tract infection in bladder cancer: a systematic review and meta-analysis. *World Journal of Urology*. 36, pp.1181–1190.

Bennett, R.L. and Licht, J.D. 2018. Targeting Epigenetics in Cancer. *Annual Review of Pharmacology and Toxicology*. 58, pp.187–207.

Berletch, J.B., Deng, X., Nguyen, D.K. and Disteche, C.M. 2013. Female Bias in Rhox6 and 9 Regulation by the Histone. *PLoS Genetics*. 9, pp.1–12.

Bertram, J.S. and Craig, A.W. 1972. Specific Induction of Bladder Cancer in Mice by Butyl-(4-hydroxybutyl)-nitrosamine and the Effects of Hormonal Modifications on the Sex Difference in Response. *European Journal of Cancer*. 8, pp.587–594.

Blackledge, N.P., Zhou, J.C., Tolstorukov, M.Y., Farcas, A.M., Park, P.J. and Klose, R.J. 2010. CpG islands recruit a histone H3 lysine 36 demethylase. *Molecular Cell*. 38, pp.179–90.

Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S. and Crawford, G.E. 2008. High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell*. 132, pp.311–322.

Brierley, J.D., Gospodarowicz, M.K. and Witterkind, C. 2017. *TNM Classification of Malignant Tumours 8th Edition*.

Brożyna, A.A., Jochymski, C., Janjetovic, Z., Jóźwicki, W., Tuckey, R.C. and Slominski, A.T. 2014. CYP24A1 expression inversely correlates with melanoma progression: Clinic-pathological studies. *International Journal of Molecular Sciences*. 15, pp.19000–19017.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. and Greenleaf, W.J. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods*. 10, pp.1213–8.

Bustelo, X.R. and Dosil, M. 2018. Ribosome biogenesis and cancer: basic and translational challenges. *Current Opinion in Genetics and Development*. 48, pp.22–29.

Bysani, M., Agren, R., Davegårdh, C., Volkov, P., Rönn, T., Unneberg, P., Bacos, K. and Ling, C. 2019. ATAC-seq reveals alterations in open chromatin in pancreatic islets from subjects with type 2 diabetes. *Scientific Reports*. 9, pp.7785.

Caiafa, P. and Zampieri, M. 2005. DNA methylation and chromatin structure: the puzzling CpG islands. *Journal of Cellular Biochemistry*. 94, pp.257–65.

Casadevall, D., Kilian, A.Y. and Bellmunt, J. 2017. The prognostic role of epigenetic dysregulation in bladder cancer: A systematic review. *Cancer Treatment Reviews*. 61, pp.82–93.

Cedar, H. and Bergman, Y. 2009. Linking DNA methylation and histone modification: patterns and paradigms. *Nature Reviews Genetics*. 10, pp.295–304.

Cejas, P., Li, L., O'Neill, N.K., Duarte, M., Rao, P., Bowden, M., Zhou, C.W., Mendiola, M., Burgos, E., Feliu, J., Moreno-Rubio, J., Guadalajara, H., Moreno, V., García-Olmo, D., Bellmunt, J., Mullane, S., Hirsch, M., Sweeney, C.J., Richardson, A., Liu, X.S., Brown, M., Shivdasani, R.A. and Long, H.W. 2016. Chromatin immunoprecipitation from fixed clinical tissues reveals tumor-specific enhancer profiles. *Nature Medicine*. 22, pp.1–8.

Chakraborty, A.A., Laukka, T., Myllykoski, M., Ringel, A.E., Booker, M.A., Tolstorukov, M.Y., Meng, Y.J., Meier, S.R., Jennings, R.B., Creech, A.L., Herbert, Z.T., McBrayer, S.K., Olenchock, B.A., Jaffe, J.D., Haigis, M.C., Beroukhim, R., Signoretti, S., Koivunen, P. and Kaelin, W.G. 2019. Histone demethylase KDM6A directly senses oxygen to control chromatin and cell fate. *Science*. 363, pp.1217–1222.

Chapman, E.J., Hurst, C.D., Pitt, E., Chambers, P., Aveyard, J.S. and Knowles, M. a 2006. Expression of hTERT immortalises normal human urothelial cells without inactivation of the p16/Rb pathway. *Oncogene*. 25, pp.5037–5045.

Chapman, E.J., Kelly, G. and Knowles, M.A. 2008. Genes Involved in Differentiation, Stem Cell Renewal, and Tumorigenesis Are Modulated in Telomerase-Immortalized Human Urothelial Cells. *Molecular Cancer Research*. 6, pp.1154–1168.

Chavarría-Smith, J. and Vance, R.E. 2015. The NLRP1 inflammasomes. *Immunological Reviews*. 265, pp.22–34.

Chen, H., Li, C., Peng, X., Zhou, Z., Weinstein, J.N., Caesar-Johnson, S.J., Demchok, J.A., Felau, I., Kasapi, M., Ferguson, M.L., Hutter, C.M., Liang, H., et al. 2018. A Pan-Cancer Analysis of Enhancer Expression in Nearly 9000 Patient Samples. *Cell*. 173, pp.386-399.e12.

Chen, X., Xie, W., Gu, P., Cai, Q., Wang, B., Xie, Y., Dong, W., He, W., Zhong, G., Lin, T. and Huang, J. 2015. Upregulated WDR5 promotes proliferation, self-renewal and chemoresistance in bladder cancer via mediating H3K4 trimethylation. *Scientific Reports*. 5, pp.1–12.

Chen, Z., Du, Y., Liu, X., Chen, H., Weng, X., Guo, J., Wang, M., Wang, X. and Wang, L. 2019. EZH2 inhibition suppresses bladder cancer cell growth and metastasis via the JAK2/STAT3 signaling pathway. *Oncology Letters*., pp.907–915.

Chi, J.T., Wang, Z., Nuyten, D.S.A., Rodriguez, E.H., Schaner, M.E., Salim, A., Wang, Y., Kristensen, G.B., Helland, Å., Børresen-Dale, A.L., Giaccia, A., Longaker, M.T., Hastie, T., Yang, G.P., Van De Vijver, M.J. and Brown, P.O. 2006. Gene expression programs in response to hypoxia: Cell type specificity and prognostic significance in human cancers. *PLoS Medicine*. 3, pp.395–409.

Clayton, J.A. and Collins, F.S. 2014. Policy: NIH to balance sex in cell and animal studies : Nature News & Comment. *Nature*. 509, pp.282–283.

Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M. and Jacobsen, S.E. 2008. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*. 452, pp.215–219.

Contrino, S., Smith, R.N., Butano, D., Carr, A., Hu, F., Lyne, R., Rutherford, K., Kalderimis, A., Sullivan, J., Carbon, S., Kephart, E.T., Lloyd, P., Stinson, E.O., Washington, N.L., Perry, M.D., Ruzanov, P., Zha, Z., Lewis, S.E., Stein, L.D. and Micklem, G. 2012. modMine: Flexible access to modENCODE data. *Nucleic Acids Research*. 40, pp.1082–1088.

Corces, M.R., Granja, J.M., Shams, S., Louie, B.H., Seoane, J.A., Zhou, W., Silva, T.C., Groeneveld, C., Wong, C.K., Cho, W., Satpathy, A.T., Mumbach, M.R., Hoadley, K.A., Robertson, A.G., Sheffield, N.C., Felau, I., Castro, M.A.A., Berman, B.P., Staudt, L.M., Zenklusen, J.C. and Laird, P.W. 2018. The chromatin accessibility landscape of primary human cancers. *Science*. 362. pp.1-8

Crallan, R.A., Georgopoulos, N.T. and Southgate, J. 2006. Experimental models of human bladder carcinogenesis. *Carcinogenesis*. 27, pp.374–381.

Crawford, G.E., Davis, S., Scacheri, P.C., Renaud, G., Halawi, M.J., Erdos, M.R., Green, R., Meltzer, P.S., Wolfsberg, T.G. and Collins, F.S. 2006. DNase-chip: A high-resolution

method to identify DNase I hypersensitive sites using tiled microarrays. *Nature Methods*. 3, pp.503–509.

Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A., Boyer, L.A., Young, R.A. and Jaenisch, R. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America*. 107, pp.21931–21936.

Croft, D., Fabregat Mundo, A., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M.R., Jassal, B., Jupe, S., Matthews, L., May, B., Palatnik, S., Rothfels, K., Shamovsky, V., Song, H., Williams, M., Birney, E., Hermjakob, H., Stein, L. and D'eustachio, P. 2014. The Reactome pathway knowledgebase. *Nucleic Acids Research*. 42, pp.D472–D477.

Croft, P.R., Lathrop, S.L., Feddersen, R.M. and Joste, N.E. 2005. Estrogen Receptor Expression in Papillary Urothelial Carcinoma of the Bladder and Ovarian Transitional Cell Carcinoma. *Archives of Pathology and Laboratory Medicine*. 129, pp.194–199.

Cumberbatch, M., Cox, A., Teare, D. and Catto, J. 2015. Contemporary Occupational Carcinogen Exposure and Bladder Cancer A Systematic Review and Meta-analysis. *JAMA Oncology*. 1, pp.1282–1290.

Cumberbatch, M.G., Rota, M., Catto, J.W.F. and La Vecchia, C. 2016. The Role of Tobacco Smoke in Bladder and Kidney Carcinogenesis: A Comparison of Exposures and Meta-analysis of Incidence and Mortality Risks. *European Urology*. 70, pp.458–466.

Cumberbatch, M.G.K., Jubber, I., Black, P.C., Esperto, F., Figueroa, J.D., Kamat, A.M., Kiemeney, L., Lotan, Y., Pang, K., Silverman, D.T., Znaor, A. and Catto, J.W.F. 2018. Epidemiology of Bladder Cancer: A Systematic Review and Contemporary Update of Risk Factors in 2018. *European Urology*. 74, pp.784–795.

Dalbagni, G., Vora, K., Kaag, M., Cronin, A., Bochner, B., Donat, S.M. and Herr, H.W. 2009. Clinical outcome in a contemporary series of restaged patients with clinical T1 bladder cancer. *European Urology*. 56, pp.903–910.

Davies, J.O.J., Oudelaar, A.M., Higgs, D.R. and Hughes, J.R. 2017. How best to identify chromosomal interactions: A comparison of approaches. *Nature Methods*. 14, pp.125–134.

Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K., Onate, K.C., Graham, K., Miyasato, S.R., Dreszer, T.R., Strattan, J.S., Jolanki, O., Tanaka, F.Y. and Cherry, J.M. 2018. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Research*. 46.

DeGraff, D.J., Clark, P.E., Cates, J.M., Yamashita, H., Robinson, V.L., Yu, X., Smolkin, M.E., Chang, S.S., Cookson, M.S., Herrick, M.K., Shariat, S.F., Steinberg, G.D.,

Frierson, H.F., Wu, X.-R., Theodorescu, D. and Matusik, R.J. 2012. Loss of the Urothelial Differentiation Marker FOXA1 Is Associated with High Grade, Late Stage Bladder Cancer and Increased Tumor Proliferation. *PLoS ONE*. 7, p.e36669.

Dekker, J. 2002. Capturing Chromosome Conformation. *Science*. 295, pp.1306–1311.

Deluca, D.S., Segrè, A. V, Sullivan, T.J., Young, T.R., Gelfand, E.T., Trowbridge, C.A., Maller, J.B., Tukiainen, T., Lek, M., Ward, L.D., Kheradpour, P., Iriarte, B., Meng, Y., Palmer, C.D., Esko, T., Winckler, W., Hirschhorn, J.N., Kellis, M., Macarthur, D.G., Getz, G., Ncsu, U.N.C., Shabalin, A.A., Li, G., Choi, C. and Foster, B.A. 2015. The human transcriptome across tissues and individuals. *Science*. 348, pp.660–665.

Denny, S.K., Yang, D., Chuang, C.H., Brady, J.J., Lim, J.S.S., Grüner, B.M., Chiou, S.H., Schep, A.N., Baral, J., Hamard, C., Antoine, M., Wislez, M., Kong, C.S., Connolly, A.J., Park, K.S., Sage, J., Greenleaf, W.J. and Winslow, M.M. 2016. Nfib Promotes Metastasis through a Widespread Increase in Chromatin Accessibility. *Cell*. 166, pp.328–342.

Denzinger, S., Fritsche, H.M., Otto, W., Blana, A., Wieland, W.F. and Burger, M. 2008. Early Versus Deferred Cystectomy for Initial High-Risk pT1G3 Urothelial Carcinoma of the Bladder: Do Risk Factors Define Feasibility of Bladder-Sparing Approach? *European Urology*. 53, pp.146–152.

Dietrich, B. and Srinivas, S. 2018. Urothelial carcinoma: The evolving landscape of immunotherapy for patients with advanced disease. *Research and Reports in Urology*. 10, pp.7–16.

Dion, M.F., Altschuler, S.J., Wu, L.F. and Rando, O.J. 2005. Genomic characterization reveals a simple histone H4 acetylation code. *Proceedings of the National Academy of Sciences*. 102, pp.5501–5506.

Dobruch, J., Daneshmand, S., Fisch, M., Lotan, Y., Noon, A.P., Resnick, M.J., Shariat, S.F., Zlotta, A.R., Boorjian, S.A. and Catto, J. 2016. Gender and Bladder Cancer : A Collaborative Review of Etiology , Biology , and Outcomes. *European Urology*. 69, pp.300–310.

Dorsett, D. and Merkenschlager, M. 2013. Cohesin at active genes: A unifying theme for cohesin and gene expression from model organisms to humans. *Current Opinion in Cell Biology*. 25, pp.327–333.

Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C., Green, R.D. and Dekker, J. 2006. Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Research*. 16, pp.1299–1309.

Down, T.A., Rakyan, V.K., Turner, D.J., Flicek, P., Li, H., Kulesha, E., Gräf, S., Johnson, N., Herrero, J., Tomazou, E.M., Thorne, N.P., Bäckdahl, L., Herberth, M., Howe, K.L., Jackson, D.K., Miretti, M.M., Marioni, J.C., Birney, E., Hubbard, T.J.P., Durbin, R.,

Tavaré, S. and Beck, S. 2008. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nature Biotechnology*. 26, pp.779–85.

Drost, J. and Clevers, H. 2018. Organoids in cancer research. *Nature Reviews Cancer*. 18, pp.407–418.

Du, X., Wang, Q.R., Chan, E., Merchant, M., Liu, J., French, D., Ashkenazi, A. and Qing, J. 2012. FGFR3 stimulates stearoyl CoA desaturase 1 activity to promote bladder tumor growth. *Cancer Research*. 72, pp.5843–5855.

Dudziec, E., Gogol-Döring, A., Cookson, V., Chen, W. and Catto, J. 2012. Integrated epigenome profiling of repressive histone modifications, DNA methylation and gene expression in normal and malignant urothelial cells. *PloS ONE*. 7, p.e32750.

Dunning, M., McCarthy, D. and Carroll, T. 2019. bioinformatic core shared training CRUK.

Dupuis-Sandoval, F., Poirier, M. and Scott, M.S. 2015. The emerging landscape of small nucleolar RNAs in cell biology. *Wiley Interdisciplinary Reviews: RNA*. 6, pp.381–397.

Dyrskjøt, L., Kruhøffer, M., Thykjaer, T., Marcussen, N., Jensen, J.L., Møller, K. and Ørntoft, T.F. 2004. Gene expression in the urinary bladder: A common carcinoma in situ gene expression signature exists disregarding histopathological classification. *Cancer Research*. 64, pp.4040–4048.

Eagen, K.P. 2018. Principles of Chromosome Architecture Revealed by Hi-C. *Trends in Biochemical Sciences*. 43, pp.469–478.

Ellinger, J., Schneider, A.C., Bachmann, A., Kristiansen, G., Müller, S.C. and Rogenhofer, S. 2016. Evaluation of global histone acetylation levels in bladder cancer patients. *Anticancer Research*. 36, pp.3961–3964.

Emil, S., Noyes, K., Feng, C. and Messing, E. 2009. Sex and Racial Differences in Bladder Cancer Presentation and Mortality in the US. *Cancer*. 115, pp.68–74.

ENCODE Consortium 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 489, pp.57–74.

Ernst, J. and Kellis, M. 2012. ChromHMM: Automating chromatin-state discovery and characterization. *Nature Methods*. 9, pp.215–216.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M. and Bernstein, B.E. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*. 473, pp.43–49.

Fajkovic, H., Halpern, J.A., Cha, E.K., Bahadori, A., Chromecki, T.F., Karakiewicz, P.I., Breinl, E., Merseburger, A.S. and Shariat, S.F. 2011. Impact of gender on bladder cancer

incidence, staging, and prognosis. *World Journal of Urology*. 29, pp.457–463.

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D.M., Forman, D. and Bray, F. 2015. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *International Journal of Cancer*. 136.

Finger, J.N., Lich, J.D., Dare, L.C., Cook, M.N., Brown, K.K., Duraiswamis, C., Bertin, J.J. and Gough, P.J. 2012. Autolytic proteolysis within the function to find domain (FIIND) is required for NLRP1 inflammasome activity. *Journal of Biological Chemistry*. 287, pp.25030–25037.

Fishwick, C., Higgins, J., Percival-Alwyn, L., Hustler, A., Pearson, J., Bastkowski, S., Moxon, S., Swarbreck, D., Greenman, C.D. and Southgate, J. 2017. Heterarchy of transcription factors driving basal and luminal cell phenotypes in human urothelium. *Cell Death and Differentiation*. 24, pp.809–818.

Flori, A.R., Löwer, R., Schmitz-Drager, B.J. and Schulz, W.A. 1999. DNA methylation and expression of LINE-1 and HERV-K provirus sequences in urothelial and renal cell carcinomas. *British Journal of Cancer*. 80, pp.1312–1321.

Follows, G.A., Tagoh, H., Lefevre, P., Hodge, D., Morgan, G.J. and Bonifer, C. 2003. Epigenetic consequences of AML1 ± ETO action at the human c-FMS locus. *The EMBO journal*. 22.pp1-8

Forrest, A.R.R., Kawaji, H., Rehli, M., Kenneth Baillie, J., de Hoon, M.J.L., Haberle, V., Lassmann, T., Kulakovskiy, I. V, Lizio, M., Itoh, M., Andersson, R., Hayashizaki, Y., et al. 2014. A promoter-level mammalian expression atlas. *Nature*. 507, p.462.

Foxman, B. 2010. The epidemiology of urinary tract infection. *Nature Reviews Urology*. 7, pp.653–660.

Franconi, F., Brunelleschi, S., Steardo, L. and Cuomo, V. 2007. Gender differences in drug responses. *Pharmacological Research*. 55, pp.81–95.

Fresno Vara, J.Á., Casado, E., de Castro, J., Cejas, P., Belda-Iniesta, C. and González-Barón, M. 2004. P13K/Akt signalling pathway and cancer. *Cancer Treatment Reviews*. 30, pp.193–204.

Fry, C. 2005. Role of the bladder in storage and micturition. *Surgery (Oxford)*. 23, pp.93–96.

Fu, S., Wang, Q., Moore, J.E., Purcaro, M.J., Pratt, H.E., Fan, K., Gu, C., Jiang, C., Zhu, R., Kundaje, A., Lu, A. and Weng, Z. 2018. Differential analysis of chromatin accessibility and histone modifications for predicting mouse developmental enhancers. *Nucleic Acids Research*. 46, pp.11184–11201.

Galupa, R. and Heard, E. 2018. X-Chromosome Inactivation: A Crossroads Between Chromosome Architecture and Gene Regulation. *Annual Review of Genetics*. 52, pp.535–566.

Gao, L., Ma, J., Mannoor, K., Guarnera, M.A., Shetty, A., Zhan, M., Xing, L., Stass, S.A. and Jiang, F. 2015. Genome-wide small nucleolar RNA expression analysis of lung cancer by next-generation deep sequencing. *International Journal of Cancer*. 136, pp.E623–E629.

Garcia-Carpizo, V., Ruiz-Llorente, S., Sarmentero, J., Graña-Castro, O., Pisano, D.G. and Barrero, M.J. 2018. CREBBP/EP300 bromodomains are critical to sustain the GATA1/MYC regulatory axis in proliferation. *Epigenetics & Chromatin*. 11, pp112-122

Gershoni, M. and Pietrokovski, S. 2017. The landscape of sex-differential transcriptome and its consequent selection in human adults. *BMC Biology*. 15, pp.1–15.

Gilmour, D.S. and Lis, J.T. 1984. Detecting protein-DNA interactions in vivo: Distribution of RNA polymerase on specific bacterial genes. *Proceedings of the National Academy of Sciences*. 81, pp.4275–4279.

Giresi, P.G., Kim, J., McDaniell, R.M., Iyer, V.R. and Lieb, J.D. 2007. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Research*. 17, pp.877–885.

Godoy, G., Gakis, G., Smith, C.L. and Fahmy, O. 2016. Effects of androgen and estrogen receptor signaling pathways on bladder cancer initiation and progression. *Bladder Cancer*. 2, pp.127–137.

Greenfield, A., Carrel, L., Pennisi, D., Philippe, C., Quaderi, N., Siggers, P., Steiner, K., Tam, P.P.L., Willard, H.F. and Koopman, P. 1998. The UTX gene escapes X inactivation in mice and humans. *Human Molecular Genetics*. 7, pp.737–742.

Greenwald, W.W., Chiou, J., Yan, J., Qiu, Y., Dai, N., Wang, A., Nariai, N., Aylward, A., Han, J.Y., Kadakia, N., Regue, L., Okino, M.L., Drees, F., Kramer, D., Vinckier, N., Minichiello, L., Gorkin, D., Avruch, J., Frazer, K.A., Sander, M., Ren, B. and Gaulton, K.J. 2019. Pancreatic islet chromatin accessibility and conformation reveals distal enhancer networks of type 2 diabetes risk. *Nature Communications*. 10, pp1-12

Greer, E.L. and Shi, Y. 2012. Histone methylation: a dynamic mark in health, disease and inheritance. *Nature Reviews Genetics*. 13, pp.343–57.

Guey, L.T., García-Closas, M., Murta-Nascimento, C., Lloreta, J., Palencia, L., Kogevinas, M., Rothman, N., Vellalta, G., Calle, M.L., Marenne, G., Tardón, A., Carrato, A., García-Closas, R., Serra, C., Silverman, D.T., Chanock, S., Real, F.X. and Malats, N. 2010. Genetic Susceptibility to Distinct Bladder Cancer Subphenotypes. *European Urology*. 57, pp.283–292.

Gui, Y., Guo, G., Huang, Y., Hu, X., Tang, A., Gao, S., Wu, R., Chen, C., Li, X., Zhou, L., He, M., Cai, Z., et al. 2011b. Frequent mutations of chromatin remodeling genes in transitional cell carcinoma of the bladder. *Nature Genetics*. 43, pp.875–878.

Guo, G., Sun, X., Chen, C., Wu, S., Huang, P., Li, Z., Dean, M., Huang, Y., Jia, W., Zhou, Q., Tang, A., Cai, Z., et al. 2013. Whole-genome and whole-exome sequencing of

bladder cancer identifies frequent alterations in genes involved in sister chromatid cohesion and segregation. *Nature Genetics*. 45, pp.1459–63.

van Haaften, G., Dalgliesh, G.L., Davies, H., Chen, L., Bignell, G., Greenman, C., Edkins, S., Hardy, C., O'Meara, S., Teague, J., Butler, A., Futreal, P.A., et al. 2009. Somatic mutations of the histone H3K27 demethylase gene UTX in human cancer. *Nature Genetics*. 41, pp.521–523.

Han, B., Cui, D., Jing, Y., Hong, Y. and Xia, S. 2012. Estrogen receptor β (ERβ) is a novel prognostic marker of recurrence survival in non-muscle-invasive bladder cancer potentially by inhibiting cadherin switch. *World Journal of Urology*. 30, pp.861–867.

Hartge, P., Harvey, E.B., Linehan, W.M., Silverman, D.T., Sullivan, J.W., Hoover, R.N. and Fraumeni, J.F. 1990. Unexplained excess risk of bladder cancer in men. *Journal of the National Cancer Institute*. 82, pp.1636–1640.

Hartman, R.J.G., Huisman, S.E. and Den Ruijter, H.M. 2018. Sex differences in cardiovascular epigenetics - A systematic review. *Biology of Sex Differences*. 9, pp.1–8.

Hayami, S., Kelly, J.D., Cho, H.-S., Yoshimatsu, M., Unoki, M., Tsunoda, T., Field, H.I., Neal, D.E., Yamaue, H., Ponder, B.A.J., Nakamura, Y. and Hamamoto, R. 2011. Overexpression of LSD1 contributes to human carcinogenesis through chromatin regulation in various cancers. *International Journal of Cancer*. 128, pp.574–586.

He, Q., Johnston, J. and Zeitlinger, J. 2015. ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. *Nature Biotechnology*. 33, pp.395–401.

Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A., Wang, W., Weng, Z., Green, R.D., Crawford, G.E. and Ren, B. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics*. 39, pp.311–8.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H. and Glass, C.K. 2010. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell*. 38, pp.576–589.

Heldring, N., Pike, A., Andersson, S., Matthews, J., Cheng, G., Hartman, J., Tujague, M., Ström, A., Treuter, E., Warner, M. and Gustafsson, J. 2007. Estrogen Receptors: How Do They Signal and What Are Their Targets. *Physiological Reviews*. 87, pp.905–931.

Hemelt, M., Yamamoto, H., Cheng, K.K. and Zeegers, M.P.A. 2009. The effect of smoking on the male excess of bladder cancer: A meta-analysis and geographical analyses. *International Journal of Cancer*. 124, pp.412–419.

Hemming, S., Cakouros, D., Isenmann, S., Cooper, L., Menicanin, D., Zannettino, A. and Gronthos, S. 2014. EZH2 and KDM6A act as an epigenetic switch to regulate mesenchymal stem cell lineage specification. *Stem Cells*. 32, pp.802–815.

Hengbin, W., Liangjun, W., Erdjument-Bromage, H., Miguel, V., Paul, T., Richard, S.J. and Yi, Z. 2004. Role of histone H2A ubiquitination in Polycomb silencing. *Nature*. 431, pp.862–868.

Hershko, A. and Aaron, C. 1998. the Ubiquitin System. *Annual Review of Biochemistry*. 67, pp.425–79.

Hertting, O., Holm, Å., Lüthje, P., Brauner, H., Dyrdak, R., Jonasson, A.F., Wiklund, P., Chromek, M. and Brauner, A. 2010. Vitamin D induction of the human antimicrobial peptide cathelicidin in the urinary bladder. *PLoS ONE*. 5, pp.1–9.

Hirota, T., Lipp, J.J., Toh, B.-H. and Peters, J.-M. 2005. Histone H3 serine 10 phosphorylation by Aurora B causes HP1 dissociation from heterochromatin. *Nature*. 438, pp.1176–1180.

Hoffman, K.L., Lerner, S.P. and Smith, C.L. 2013. Raloxifene Inhibits Growth of RT4 Urothelial Carcinoma Cells via Estrogen Receptor-Dependent Induction of Apoptosis and Inhibition of Proliferation. *Hormones and Cancer*. 4, pp.24–35.

Hong, L., Schroth, G.P., Matthews, H.R., Yau, P. and Bradbury, E.M. 1993. Studies of the DNA binding properties of histone H4 amino terminus. Thermal denaturation studies reveal that acetylation markedly reduces the binding constant of the H4 'tail' to DNA. *Journal of Biological Chemistry*. 268, pp.305–314.

Hu, D., Gao, X., Morgan, M. a, Herz, H.-M., Smith, E.R. and Shilatifard, A. 2013. The MLL3/MLL4 branches of the COMPASS family function as major histone H3K4 monomethylases at enhancers. *Molecular and Cellular Biology*. 33, pp.4745–54.

Hu, Z. and Tee, W.W. 2017. Enhancers and chromatin structures: Regulatory hubs in gene expression and diseases. *Bioscience Reports*. 37, pp.1–14.

Huang, D.W., Sherman, B.T. and Lempicki, R.A. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*. 4, pp.44–57.

Hughes, R.N. 2007. Sex does matter: Comments on the prevalence of male-only investigations of drug effects on rodent behaviour. *Behavioural Pharmacology*. 18, pp.583–589.

Hurst, C.D., Alder, O., Platt, F.M., Mott, H.R., Gordenin, D.A., Knowles, M.A., Hurst, C.D., Alder, O., Platt, F.M., Droop, A., Stead, L.F., Burns, J.E. and Burghel, G.J. 2017. Genomic Subtypes of Non-invasive Bladder Cancer with Distinct Metabolic Profile and Female Gender Bias in KDM6A Mutation Frequency. *Cancer Cell*. 32, pp.701-715.e7.

Hurst, C.D., Platt, F.M. and Knowles, M.A. 2014. Comprehensive mutation analysis of the TERT promoter in bladder cancer and detection of mutations in voided urine. *European Urology*. 65, pp.367–369.

Hurst, C.D., Platt, F.M., Taylor, C.F. and Knowles, M.A. 2012. Novel tumor subgroups of urothelial carcinoma of the bladder defined by integrated genomic analysis. *Clinical Cancer Research*. 18, pp.5865–5877.

Hurtado, A., Holmes, K.A., Ross-Innes, C.S., Schmidt, D. and Carroll, J.S. 2011. FOXA1 is a key determinant of estrogen receptor function and endocrine response. *Nature Genetics*. 43, pp.27–33.

Imbeaud, S., Graudens, E., Boulanger, V., Barlet, X., Zaborski, P., Eveno, E., Mueller, O., Schroeder, A. and Auffray, C. 2005. Towards standardization of RNA quality assessment using user-independent classifiers of microcapillary electrophoresis traces. *Nucleic Acids Research*. 33, pp1-8

Inatome, R., Tsujimura, T., Hitomi, T., Mitsui, N., Hermann, P., Kuroda, S., Yamamura, H. and Yanagi, S. 2000. Identification of CRAM, a novel unc-33 gene family protien that associates with CRMP3 and protein-tyrosine kinase(s) in the developing rat brain. *Journal of Biological Chemistry*. 275, pp.27291–27302.

Ing-Simmons, E., Seitan, V.C., Faure, A.J., Flicek, P., Carroll, T., Dekker, J., Fisher, A.G., Lenhard, B. and Merkenschlager, M. 2012. Spatial enhancer clustering and regulation of enhancer-proximal genes by cohesin. *Genome Research*. 25, pp.504–513.

International Human Genome Sequencing (IHGS) Consortium 2004. International Human Genome Sequencing Consortium,Finishing the euchromatic sequence of the human genome. *Nature*. 431, pp.931–945.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U. and Speed, T.P. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 4, pp.249–264.

Isensee, J., Witt, H., Pregla, R., Hetzer, R., Regitz-Zagrosek, V. and Ruiz Noppinger, P. 2008. Sexually dimorphic gene expression in the heart of mice and men. *Journal of Molecular Medicine*. 86, pp.61–74.

Izumi, K., Taguri, M., Miyamoto, H., Hara, Y., Kishida, T., Chiba, K., Murai, T., Hirai, K., Suzuki, K., Fujinami, K., Ueki, T., Udagawa, K., Kitami, K., Moriyama, M., Miyoshi, Y., Tsuchiya, F., Ikeda, I., Kobayashi, K., Sato, M., Morita, S., Noguchi, K. and Uemura, H. 2014. Androgen deprivation therapy prevents bladder cancer recurrence. *Oncotarget*. 5, pp.12665–12674.

Jebar, A.H., Hurst, C.D., Tomlinson, D.C., Johnston, C., Taylor, C.F. and Knowles, M.A. 2005. FGFR3 and Ras gene mutations are mutually exclusive genetic events in urothelial cell carcinoma. *Oncogene*. 24, pp.5218–5225.

Jia, M., Assistant, R., Dahlman-Wright, K. and Gustafsson, J.-Å. 2015. Estrogen receptor alpha and beta in health and disease. *Best Practice & Research Clinical Endocrinology & Metabolism*. 29, pp.557–568.

Jiang, W., Wang, J. and Zhang, Y. 2013. Histone H3K27me3 demethylases KDM6A and KDM6B modulate definitive endoderm differentiation from human ESCs by regulating WNT signaling pathway. *Cell Research*. 23, pp.122–130.

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., Palin, K., Vaquerizas, J.M., Vincentelli, R., Luscombe, N.M., Hughes, T.R., Lemaire, P., Ukkonen, E., Kivioja, T. and Taipale, J. 2013. DNA-binding specificities of human transcription factors. *Cell*. 152, pp.327–339.

Jóźwicki, W., Brożyna, A.A., Siekiera, J. and Slominski, A.T. 2015. Expression of vitamin D receptor (VDR) positively correlates with survival of urothelial bladder cancer patients. *International Journal of Molecular Sciences*. 16, pp.24369–24386.

Kaelin, W.G. 2008. The von Hippel-Lindau tumour suppressor protein: O2 sensing and cancer. *Nature Reviews Cancer*. 8, pp.865–873.

Kaestner, K.H. 2010. The FoxA factors in organogenesis and differentiation. *Current Opinion in Genetics & Development*. 20, pp.527–532.

Kaletsch, A., Pinkerneil, M., Hoffmann, M.J., Jaguva Vasudevan, A.A., Wang, C., Hansen, F.K., Wiek, C., Hanenberg, H., Gertzen, C., Gohlke, H., Kassack, M.U., Kurz, T., Schulz, W.A. and Niegisch, G. 2018. Effects of novel HDAC inhibitors on urothelial carcinoma cells. *Clinical Epigenetics*. 10, p.100.

Kamei, J., Ito, H., Aizawa, N., Hotta, H., Kojima, T., Fujita, Y., Ito, M., Homma, Y. and Igawa, Y. 2018. Age-related changes in function and gene expression of the male and female mouse bladder. *Scientific Reports*. 8, pp.1–9.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. and Hattori, M. 2004. The KEGG resource for deciphering the genome. *Nucleic Acids Research*. 32, pp.D277–D280.

Kaneko, S. and Li, X. 2018. X chromosome protects against bladder cancer in females via a KDM6A-dependent epigenetic mechanism. *Science Advances*. 4, pp1-7

Kantarjian, H., Oki, Y., Garcia-Manero, G., Huang, X., O'Brien, S., Cortes, J., Faderl, S., Bueso-Ramos, C., Ravandi, F., Estrov, Z., Ferrajoli, A., Wierda, W., Shan, J., Davis, J., Giles, F., Saba, H.I. and Issa, J.P.J. 2007. Results of a randomized study of 3 schedules of low-dose decitabine in higher-risk myelodysplastic syndrome and chronic myelomonocytic leukemia. *Blood*. 109, pp.52–57.

Karagas, M.R., Park, S., Warren, A., Hamilton, J., Nelson, H.H., Mott, L.A. and Kelsey, K.T. 2005. Gender, smoking, glutathione-S-transferase variants and bladder cancer incidence: A population-based study. *Cancer Letters*. 219, pp.63–69.

Karan, D. 2018. Inflammasomes: Emerging Central Players in Cancer Immunology and Immunotherapy. *Frontiers in Immunology*. 9, p.3028.

Kauffman, E.C., Robinson, B.D., Downes, M.J., Powell, L.G., Lee, M.M., Scherr, D.S.,

Gudas, L.J. and Mongan, N.P. 2011. Role of androgen receptor and associated lysine-demethylase coregulators, LSD1 and JMJD2A, in localized and advanced human bladder cancer. *Molecular Carcinogenesis*. 50, pp.931–44.

Kauffman, E.D., Robinson, B.D., DDownes, M., Marcinkiewicz, K., Vourganti, S., Scherr, D.S., Gudas, L.J. and Mongan, N.P. 2013. Estrogen receptor-β expression and pharmacological targeting in bladder cancer. *Oncology Reports*. 30, pp.131–138.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. and Haussler, D. 2002. The Human Genome Browser at UCSC. *Genome Research*. 12, pp.996–1006.

Kil Nam, J., Sung, W.P., Don Lee, S. and Moon, K.C. 2014. Prognostic Value of Sex-Hormone Receptor Expression in Non-Muscle-Invasive Bladder Cancer. *Yonsei Medical Journal*. 55, pp.1214–1221.

Kim, K.H. and Roberts, C.W.M. 2016. Targeting EZH2 in cancer. *Nature Medicine*. 22, pp.128–134.

Klein, S.L., Schiebinger, L., Stefanick, M.L., Cahill, L., Danska, J., De Vries, G.J., Kibbe, M.R., McCarthy, M.M., Mogil, J.S., Woodruff, T.K. and Zucker, I. 2015. Opinion: Sex inclusion in basic research drives discovery. *Proceedings of the National Academy of Sciences of the United States of America*. 112, pp.5257–5258.

Klemm, S.L., Shipony, Z. and Greenleaf, W.J. 2019. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*. 20, pp.207–220.

Klose, R.J. and Bird, A.P. 2006. Genomic DNA methylation: the mark and its mediators. *Trends in Biochemical Sciences*. 31, pp.89–97.

Knowles, M.A. and Hurst, C.D. 2015. Molecular biology of bladder cancer: new insights into pathogenesis and clinical diversity. *Nature Reviews Cancer*. 15, pp.25–41.

Knowles, M.A., Platt, F.M., Ross, R.L. and Hurst, C.D. 2009. Phosphatidylinositol 3-kinase (PI3K) pathway activation in bladder cancer. *Cancer and Metastasis Reviews*. 28, pp.305–316.

Knutson, S.K., Kawano, S., Minoshima, Y., Warholic, N.M., Huang, K.-C., Xiao, Y., Kadowaki, T., Uesugi, M., Kuznetsov, G., Kumar, N., Wigle, T.J., Klaus, C.R., Allain, C.J., Raimondi, a., Waters, N.J., Smith, J.J., Porter-Scott, M., Chesworth, R., Moyer, M.P., Copeland, R. a., Richon, V.M., Uenaka, T., Pollock, R.M., Kuntz, K.W., Yokoi, a. and Keilhack, H. 2014. Selective Inhibition of EZH2 by EPZ-6438 Leads to Potent Antitumor Activity in EZH2-Mutant Non-Hodgkin Lymphoma. *Molecular Cancer Therapeutics*. 13, pp.842–854.

Kornberg, R. 1974. Chromatin Structure : A Repeating Unit of Histones and DNA Chromatin. *Science*. 184, pp.868–871.

Korolev, N., Fan, Y., Lyubartsev, A.P. and Nordenskiöld, L. 2012. Modelling chromatin structure and dynamics: Status and prospects. *Current Opinion in Structural Biology*. 22, pp.151–159.

Kouzarides, T. 2007. Chromatin Modifications and Their Function. *Cell*. 128, pp.693–705.

Krabbe, L.M., Svatek, R.S., Shariat, S.F., Messing, E. and Lotan, Y. 2015. Bladder cancer risk: Use of the PLCO and NLST to identify a suitable screening cohort. *Urologic Oncology: Seminars and Original Investigations*. 33, 65.e19-65.e25.

Kufel, J. and Grzechnik, P. 2019. Small Nucleolar RNAs Tell a Different Tale. *Trends in Genetics*. 35, pp.104–117.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., Amin, V., Kellis, M., et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature*. 518, pp.317–329.

Kuroki, S. and Tachibana, M. 2018. Epigenetic regulation of mammalian sex determination. *Molecular and Cellular Endocrinology*. 468, pp.31–38.

Lamb, M.E., Sternberg, K., Wan, C., Lin, G., Jin, Q., Leichtman, M.D., White, S.H., Meidinger, C., Rapoport, B., Kolm, S.C., Feiring, M.C., Laosa, L.M., Sigel, I.E., Blanton, P., Leibbrandt, A., Huang, J. and Preferences, R. 2013. Highly Recurrent TERT Promoter Mutations in Human Melanoma. *Science*. 339, pp.957–959.

Lan, F., Bayliss, P.E., Rinn, J.L., Whetstine, J.R., Wang, J.K., Chen, S., Iwase, S., Alpatov, R., Issaeva, I., Canaani, E., Roberts, T.M., Chang, H.Y. and Shi, Y. 2007. A histone H3 lysine 27 demethylase regulates animal posterior development. *Nature*. 449, pp.689–694.

Lang, A., Yilmaz, M., Hader, C., Murday, S., Kunz, X., Wagner, N., Wiek, C., Petzsch, P., Köhrer, K., Koch, J., Hoffmann, M.J., Greife, A. and Schulz, W.A. 2019. Contingencies of UTX / KDM6A Action in Urothelial Carcinoma. *Cancer*. 11, pp.1–19.

Leal, J., Luengo-Fernandez, R., Sullivan, R. and Witjes, J.A. 2015. Economic Burden of Bladder Cancer Across the European Union. *European Urology*. 6494, pp.1–10.

Lebrun, L., Milowich, D., Le Mercier, M., Allard, J., Van Eycke, Y.R., Roumeguere, T., Decaestecker, C., Salmon, I. and Rorive, S. 2018. UCA1 overexpression is associated with less aggressive subtypes of bladder cancer. *Oncology Reports*. 40, pp.2497–2506.

Lee, J., Daniel, K., Seth, S. and Anshul, K. 2019. kundajelab ATAC-seq pipeline.

Lee, J.S., Shukla, A., Schneider, J., Swanson, S.K., Washburn, M.P., Florens, L., Bhaumik, S.R. and Shilatifard, A. 2007. Histone Crosstalk between H2B Monoubiquitination and H3 Methylation Mediated by COMPASS. *Cell*. 131, pp.1084–1096.

Lee, K. and Song, C.G. 2017. Epigenetic regulation in bladder cancer : development of new prognostic targets and therapeutic implications. *Translational Cancer Research*. 6, pp225

Lee, S.H., Hu, W., Matulay, J.T., Silva, M. V., Owczarek, T.B., Kim, K., Chua, C.W., Barlow, L.M.J., Kandoth, C., Williams, A.B., Bergren, S.K., Pietzak, E.J., Anderson, C.B., Benson, M.C., Coleman, J.A., Taylor, B.S., Abate-Shen, C., McKiernan, J.M., Al-Ahmadie, H., Solit, D.B. and Shen, M.M. 2018. Tumor Evolution and Drug Response in Patient-Derived Organoid Models of Bladder Cancer. *Cell*. 173, pp.515-528.e17.

Lehmann, M., Hoffmann, M.J., Koch, A., Ulrich, S.M., Schulz, W.A. and Niegisch, G. 2014. Histone deacetylase 8 is deregulated in urothelial cancer but not a target for efficient treatment. *Journal of Experimental & Clinical Cancer Research*. 33, p.59.

Ler, L.D., Ghosh, S., Chai, X., Thike, A.A., Heng, H.L., Siew, E.Y., Dey, S., Koh, L.K., Lim, J.Q., Lim, W.K., Myint, S.S., Loh, J.L., Ong, P., Sam, X.X., Huang, D., Lim, T., Tan, P.H., Nagarajan, S., Wai, C., Cheng, S., Ho, H., Ng, L.G., Yuen, J., Lin, P., Chuang, C., Chang, Y., Weng, W., Rozen, S.G., Tan, P., Creasy, C.L., Pang, S. and Mccabe, M.T. 2017. Loss of tumor suppressor KDM6A amplifies PRC2-regulated transcriptional repression in bladder cancer and can be targeted through inhibition of EZH2. *Science Translational Medicine*. 8312, pp.1–14.

Li, L., Ren, F., Qi, C., Xu, L., Fang, Y., Liang, M., Feng, J., Chen, B., Ning, W. and Cao, J. 2018. Intermittent hypoxia promotes melanoma lung metastasis via oxidative stress and inflammation responses in a mouse model of obstructive sleep apnea. *Respiratory Research*. 19, pp.1–9.

Li, P., Chen, J. and Miyamoto, H. 2017. Androgen receptor signaling in bladder cancer. *Cancers*. 9, pp.1–14.

Li, W. and Sun, Z. 2019. Mechanism of Action for HDAC Inhibitors—Insights from Omics Approaches. *International Journal of Molecular Sciences*. 20, p.1616.

Li, Y., Zheng, Y., Izumi, K., Ishiguro, H., Ye, B., Li, F. and Miyamoto, H. 2013. Androgen activates ??-catenin signaling in bladder cancer cells. *Endocrine-Related Cancer*. 20, pp.293–304.

Li, Z., Schulz, M.H., Look, T., Begemann, M., Zenke, M. and Costa, I.G. 2019. Identification of transcription factor binding sites using ATAC-seq. *Genome Biology*. 20, pp1-8

Lieb, J.D., Liu, X., Botstein, D. and Brown, P.O. 2001. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nature Genetics*. 28, pp.327–334.

Lieberman-Aiden, E., Berkum, N.L. Van, Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., Sandstrom, R., Bernstein, B., Bender, M.A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J. and Mirny, L.A. 2009. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*. 326, pp.289–294.

Ling, G., Sugathan, A., Mazor, T., Fraenkel, E. and Waxman, D.J. 2010. Unbiased, Genome-

Wide In Vivo Mapping of Transcriptional Regulatory Elements Reveals Sex Differences in Chromatin Structure Associated with Sex-Specific Liver Gene Expression. *Molecular and Cellular Biology*. 30, pp.5531–5544.

Liu, C., Wang, M., Wei, X., Wu, L., Xu, J., Dai, X., Xia, J., Cheng, M., Yuan, Y., Zhang, P., Li, J., Feng, T., Chen, A., Zhang, W., Chen, F., Shang, Z., Zhang, X., Peters, B.A. and Liu, L. 2019. An ATAC-seq atlas of chromatin accessibility in mouse tissues. *Scientific Data*. 6, p.65.

Liu, L., Leng, L., Liu, C., Lu, C., Yuan, Y., Wu, L., Gong, F., Zhang, S., Wei, X., Wang, M., Zhao, L., Hu, L., Wang, J., Yang, H., Zhu, S., Chen, F., Lu, G., Shang, Z. and Lin, G. 2019. An integrated chromatin accessibility and transcriptome landscape of human pre-implantation embryos. *Nature Communications*. 10, pp.1–11.

Liu, M.-M., Albanese, C., Anderson, C.M., Hilty, K., Webb, P., Uht, R.M., Price, R.H., Pestell, R.G. and Kushner, P.J. 2002. Opposing action of estrogen receptors alpha and beta on cyclin D1 gene expression. *The Journal of Biological Chemistry*. 277, pp.24353–60.

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, p.550.

Lu, C., David Allis, C. and Genet Author manuscript, N. 2017. SWI/SNF Complex in Cancer: 'Remodeling' Mechanisms Uncovered HHS Public Access Author manuscript. *Nature Genetics*. 49, pp.178–179.

Luger, K., Dechassa, M.L. and Tremethick, D.J. 2012. New insights into nucleosome and chromatin structure: An ordered state or a disordered affair? *Nature Reviews Molecular Cell Biology*. 13, pp.436–447.

Luger, K., Mäder, A.W., Richmond, R.K., Sargent, D.F. and Richmond, T.J. 1997. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*. 389, pp.251–260.

Lukacz, E.S., Sampselle, C., Gray, M., MacDiarmid, S., Rosenberg, M., Ellsworth, P. and Palmer, M.H. 2011. A healthy bladder: A consensus statement. *International Journal of Clinical Practice*. 65, pp.1026–1036.

Luo, J., Chen, J., Li, H., Yang, Y., Yun, H., Yang, S. and Mao, X. 2017. LncRNA UCA1 promotes the invasion and EMT of bladder cancer cells by regulating the miR-143/HMGB1 pathway. *Oncology Letters*. 14, pp.5556–5562.

Ma, Y., Kanakousaki, K., Buttitta, L., Schwartz, J., Kovalchuk, I. and Groth, A. 2015. How the cell cycle impacts chromatin architecture and influences cell fate. *Frontiers in Genetics*. 6, pp1-9

Maase, B.H. Von Der, Hansen, S.W., Roberts, J.T., Dogliotti, L., Oliver, T., Moore, M.J., Bodrogi, I., Albers, P., Knuth, A., Lippert, C.M., Kerbrat, P., Rovira, P.S., Wersall, P., Cleall, S.P., Roychowdhury, D.F., Tomlin, I. and Conte, P.F. 2000. Gemcitabine and Cisplatin Versus Methotrexate, Vinblastine, Doxorubicin, and Cisplatin in Advanced or

Metastatic Bladder Cancer: Results of a Large, Randomized, Multinational, Multicenter, Phase III Study. *Society*. 17, pp.3068–3077.

Marinov, G.K. 2018. A decade of ChIP-seq. *Briefings in Functional Genomics*. 17, pp.77–79.

Marinov, G.K. and Kundaje, A. 2018. ChIP-ping the branches of the tree: Functional genomics and the evolution of eukaryotic gene regulation. *Briefings in Functional Genomics*. 17, pp.116–137.

Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*. 17, p.10.

Di Martino, E., L'Ĥôte, C.G., Kennedy, W., Tomlinson, D.C. and Knowles, M.A. 2009. Mutant fibroblast growth factor receptor 3 induces intracellular signaling and cellular transformation in a cell type-and mutation-specific manner. *Oncogene*. 28, pp.4306–4316.

Di Martino, E., Tomlinson, D.C. and Knowles, M.A. 2012. A decade of FGF receptor research in bladder cancer: Past, present, and future challenges. *Advances in Urology*. 2012.

Masaki, S., Keijiro, K., Akira, Y., Ario, T., Eiji, K., Takashi, D., Ryosuke, T., Junichi, I., Katsunori, T. and Masatoshi, E. 2017. Suppressed Recurrent Bladder Cancer after Androgen Suppression with Androgen Deprivation Therapy or 5α-Reductase Inhibitor. *American Urology Association*. 197, pp.308–313.

Mashhadi, R., Pourmand, G., Kosari, F., Mehrsai, A., Salem, S., Pourmand, M.R., Alatab, S., Khonsari, M., Heydari, F., Beladi, L. and Alizadeh, F. 2014. Role of Steroid Hormone Receptors in Formation and Progression of Bladder Carcinoma: A Case-Control Study. *Urology Journal*. 11, pp.1968–1973.

McCarthy, M.M., Nugent, B.M. and Lenz, K.M. 2017. Neuroimmunology and neuroepigenetics in the establishment of sex differences in the brain. *Nature Reviews Neuroscience*. 18, pp.471–484.

McGarry, T., Biniecka, M., Veale, D.J. and Fearon, U. 2018. Hypoxia, oxidative stress and inflammation. *Free Radical Biology and Medicine*. 125, pp.15–24.

McHugh, C.A., Chen, C.K., Chow, A., Surka, C.F., Tran, C., McDonel, P., Pandya-Jones, A., Blanco, M., Burghard, C., Moradian, A., Sweredoski, M.J., Shishkin, A.A., Su, J., Lander, E.S., Hess, S., Plath, K. and Guttman, M. 2015. The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature*. 521, pp.232–236.

McMahon, M., Contreras, A. and Ruggero, D. 2015. Small RNAs with big implications: New insights into H/ACA snoRNA function and their role in human disease. *Wiley Interdisciplinary Reviews: RNA*. 6, pp.173–189.

McNeill, R. V., Mason, A.S., Hodson, M.E., Catto, J.W.F. and Southgate, J. 2019. Specificity of the metallothionein-1 response by cadmium-exposed normal human urothelial cells. *International Journal of Molecular Sciences*. 20, pp1-11

Meeks, J.J. and Shilatifard, A. 2017. Multiple Roles for the MLL/COMPASS Family in the Epigenetic Regulation of Gene Expression and in Cancer. *Annual Review of Cancer Biology*. 1, pp.425–446.

Merrill, L., Gonzalez, E.J., Girard, B.M. and Vizzard, M.A. 2016. Receptors, channels, and signalling in the urothelial sensory system in the bladder. *Nature Reviews Urology*. 13, pp.193–204.

Meulen, J. Van Der, Sanghvi, V., Mavrakis, K., Durinck, K., Fang, F., Matthijssens, F., Rondou, P., Rosen, M., Pieters, T., Vandenberghe, P., Delabesse, E., Lammens, T., Moerloose, B. De, Roy, N. Van, Verhasselt, B., Poppe, B., Benoit, Y., Taghon, T., Melnick, A.M., Speleman, F., Wendel, H. and Vlierberghe, P. Van 2015. The H3K27me3 demethylase UTX is a gender-specific tumor suppressor in T-cell acute lymphoblastic leukemia. *Blood*. 125, pp.13–22.

Mieczkowski, J., Cook, A., Bowman, S.K., Mueller, B., Alver, B.H., Kundu, S., Deaton, A.M., Urban, J.A., Larschan, E., Park, P.J., Kingston, R.E. and Tolstorukov, M.Y. 2016. MNase titration reveals differences between nucleosome occupancy and chromatin accessibility. *Nature Communications*. 7, p.11485.

Minner, S., Kilgué, A., Stahl, P., Weikert, S., Rink, M., Dahlem, R., Fisch, M., Hppner, W., Wagner, W., Bokemeyer, C., Terracciano, L., Simon, R., Sauter, G. and Wilczak, W. 2010. Y chromosome loss is a frequent early event in urothelial bladder cancer. *Pathology*. 42, pp.356–359.

Mir, C., Shariat, S.F., van der Kwast, T.H., Ashfaq, R., Lotan, Y., Evans, A., Skeldon, S., Hanna, S., Vajpeyi, R., Kuk, C., Alkhateeb, S., Morote, J., van Rhijn, B.W.G., Bostrom, P., Yao, J., Miyamoto, H., Jewett, M., Fleshner, N., Messing, E. and Zlotta, A.R. 2011. Loss of androgen receptor expression is not associated with pathological stage, grade, gender or outcome in bladder cancer: a large multi-institutional study. *BJU International*. 108, pp.24–30.

Miyamoto, H., Yang, Z., Chen, Y.-T., Ishiguro, H., Uemura, H., Kubota, Y., Nagashima, Y., Chang, Y.-J., Hu, Y.-C., Tsai, M.-Y., Yeh, S., Messing, E.M. and Chang, C. 2007. Promotion of Bladder Cancer Development and Progression by Androgen Receptor Signals. *Journal of the National Cancer Institute*. 99, pp.558–568.

Miyamoto, H., Yao, J.L., Chaux, A., Zheng, Y., Hsu, I., Izumi, K., Chang, C., Messing, E.M., Netto, G.J. and Yeh, S. 2012. Expression of androgen and oestrogen receptors and its prognostic significance in urothelial neoplasm of the urinary bladder. *BJU International*. 109, pp.1716–1726.

Moch, H., Humphrey, P.., Ulbright, T.. and Reuter, V.. 2016. *WHO Classification of Tumours of the Urinary System and Male Genital Organs. Fourth edition.*

Mullenders, J., de Jongh, E., Brousali, A., Roosen, M., Blom, J.P.A., Begthel, H., Korving, J., Jonges, T., Kranenburg, O., Meijer, R. and Clevers, H.C. 2019. Mouse and human

urothelial cancer organoids: A tool for bladder cancer research. *Proceedings of the National Academy of Sciences of the United States of America*. 116, pp.4567–4574.

Mungan, N.A., Kiemeney, L.A.L.M., Van Dijck, J.A.A.M., Van Der Poel, H.G. and Witjes, J.A. 2000. Gender differences in stage distribution of bladder cancer. *Urology*. 55, pp.368–371.

Muñoz-Culla, M., Irizar, H., Sáenz-Cuesta, M., Castillo-Triviño, T., Osorio-Querejeta, I., Sepúlveda, L., López De Munain, A., Olascoaga, J. and Otaegui, D. 2016. SncRNA (microRNA &snoRNA) opposite expression pattern found in multiple sclerosis relapse and remission is sex dependent. *Scientific Reports*. 6, pp.1–10.

Nagano, T., Lubling, Y., Várnai, C., Dudley, C., Leung, W., Baran, Y., Mendelson Cohen, N., Wingett, S., Fraser, P. and Tanay, A. 2017. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature*. 547, pp.61–67.

NICE 2017. Bladder cancer: diagnosis and management of bladder cancer. *BJU International*. 120, pp.755–765.

Nickerson, M.L., Dancik, G.M., Im, K.M., Edwards, M.G., Turan, S., Brown, J., Ruiz-Rodriguez, C., Owens, C., Costello, J.C., Guo, G., Tsang, S.X., Li, Y., Zhou, Q., Cai, Z., Moore, L.E., Lucia, M.S., Dean, M. and Theodorescu, D. 2014. Concurrent Alterations in TERT, KDM6A, and the BRCA Pathway in Bladder Cancer. *Clinical Cancer Research*. 20, pp.4935–4948.

Nord, A.S., Blow, M.J., Attanasio, C., Akiyama, J.A., Holt, A., Hosseini, R., Phouanenavong, S., Plajzer-Frick, I., Shoukry, M., Afzal, V., Rubenstein, J.L.R., Rubin, E.M., Pennacchio, L.A. and Visel, A. 2013. Rapid and Pervasive Changes in Genome-wide Enhancer Usage during Mammalian Development. *Cell*. 155, pp.1521–1531.

O'Brien, E.M., Gomes, D.A., Sehgal, S. and Nathanson, M.H. 2007. Hormonal regulation of nuclear permeability. *Journal of Biological Chemistry*. 282, pp.4210–4217.

Oh, J.J., Byun, S.S., Lee, S.E., Hong, S.K., Jeong, C.W., Choi, W.S., Kim, D., Kim, H.J. and Myung, S.C. 2014. Genetic variants in the CYP24A1 gene are associated with prostate cancer risk and aggressiveness in a Korean study population. *Prostate Cancer and Prostatic Diseases*. 17, pp.149–156.

Oottamasathien, S., Wang, Y., Williams, K., Franco, O.E., Wills, M.L., Thomas, J.C., Saba, K., Sharif-Afshar, A.-R., Makari, J.H., Bhowmick, N.A., DeMarco, R.T., Hipkens, S., Magnuson, M., Brock, J.W., Hayward, S.W., Pope, J.C. and Matusik, R.J. 2007. Directed differentiation of embryonic stem cells into bladder tissue. *Developmental Biology*. 304, pp.556–566.

Pelava, A., Schneider, C. and Watkins, N.J. 2016. The importance of ribosome production, and the 5S RNP-MDM2 pathway, in health and disease. *Biochemical Society Transactions*. 44, pp.1086–1090.

Phillips, D.M.P. 1963. The Presence of Acetyl Groups in Histones. *Biochemical Journal*. 87, pp.258–263.

Pietzak, E.J., Bagrodia, A., Cha, E.K., Drill, E.N., Iyer, G., Isharwal, S., Ostrovnaya, I., Baez, P., Li, Q., Berger, M.F., Zehir, A., Schultz, N., Rosenberg, J.E., Bajorin, D.F., Dalbagni, G., Al-Ahmadie, H., Solit, D.B. and Bochner, B.H. 2017. Next-generation Sequencing of Nonmuscle Invasive Bladder Cancer Reveals Potential Biomarkers and Rational Therapeutic Targets. *European Urology*. 72, pp.952–959.

Pinkerneil, M., Hoffmann, M.J., Deenen, R., Ko hrer, K., Arent, T., Schulz, W.A. and Niegisch, G. 2016. Inhibition of Class I Histone Deacetylases 1 and 2 Promotes Urothelial Carcinoma Cell Death by Various Mechanisms. *Molecular Cancer Therapeutics*. 15, pp.299–312.

Porten, S.P. 2018. Epigenetic Alterations in Bladder Cancer. *Current Urology Reports*. 19, pp.1–8.

Poyet, C., Jentsch, B., Hermanns, T., Schweckendiek, D., Seifert, H.H., Schmidtpeter, M., Sulser, T., Moch, H., Wild, P.J. and Kristiansen, G. 2014. Expression of histone deacetylases 1, 2 and 3 in urothelial bladder cancer. *BMC Clinical Pathology*. 14, pp.1–9.

Pugh, T.J., Weeraratne, S.D., Archer, T.C., Pomeranz Krummel, D.A., Auclair, D., Bochicchio, J., Carneiro, M.O., Carter, S.L., Cibulskis, K., Erlich, R.L., Greulich, H., Greulich, H., Lennon, N.J., Mc Kenna, A., Meldrim, J., Ramos, A.H., Ross, M.G., Russ, C., Shefler, E., Sivachenko, A., Sogoloff, B., Stojanov, P., Tamayo, P., Mesirov, J.P., Amani, V., Teider, N., Sengupta, S., Francois, J.P., Northcott, P.A., Taylor, M.D., Yu, F., Crabtree, G.R., Kautzman, A.G., Gabriel, S.B., Getz, G., Jäger, N., Jones, D.T.W., Lichter, P., Pfister, S.M., Roberts, T.M., Dangl, J.L., Pomeroy, S.L. and Cho, Y.J. 2012. Medulloblastoma exome sequencing uncovers subtype-specific somatic mutations. *Nature*. 488, pp.106–110.

Qu, K., Zaba, L.C., Giresi, P.G., Li, R., Longmire, M., Kim, Y.H., Greenleaf, W.J. and Chang, H.Y. 2015. Individuality and Variation of Personal Regulomes in Primary Human T Cells. *Cell Systems*. 1, pp.51–61.

Qu, K., Zaba, L.C., Satpathy, A.T., Giresi, P.G., Li, R., Jin, Y., Armstrong, R., Jin, C., Schmitt, N., Rahbar, Z., Ueno, H., Greenleaf, W.J., Kim, Y.H. and Chang, H.Y. 2017. Chromatin Accessibility Landscape of Cutaneous T Cell Lymphoma and Dynamic Response to HDAC Inhibitors. *Cancer Cell*. 32, pp.27-41.e4.

Ramírez, F., Dündar, F., Diehl, S., Grüning, B.A. and Manke, T. 2014. DeepTools: A flexible platform for exploring deep-sequencing data. *Nucleic Acids Research*. 42, pp.187–191.

Rao, S.S.P., Huang, S.C., Glenn St Hilaire, B., Engreitz, J.M., Perez, E.M., Kieffer-Kwon, K.R., Sanborn, A.L., Johnstone, S.E., Bascom, G.D., Bochkov, I.D., Huang, X., Shamim, M.S., Shin, J., Turner, D., Ye, Z., Omer, A.D., Robinson, J.T., Schlick, T., Bernstein, B.E., Casellas, R., Lander, E.S. and Aiden, E.L. 2017. Cohesin Loss

Eliminates All Loop Domains. *Cell*. 171, pp.305-320.e24.

Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S. and Aiden, E.L. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*. 159, pp.1665–1680.

Reddy, O.L., Cates, J.M., Gellert, L.L., Crist, H.S., Yang, Z., Yamashita, H., Taylor, J.A., Smith, J.A., Chang, S.S., Cookson, M.S., You, C., Barocas, D.A., Grabowska, M.M., Ye, F., Wu, X.-R., Yi, Y., Matusik, R.J., Kaestner, K.H., Clark, P.E. and DeGraff, D.J. 2015. Loss of FOXA1 Drives Sexually Dimorphic Changes in Urothelial Differentiation and Is an Independent Predictor of Poor Prognosis in Bladder Cancer. *The American Journal of Pathology*. 185, pp.1385–1395.

Reinert, T., Modin, C., Castano, F.M., Lamy, P., Wojdacz, T.K., Hansen, L.L., Wiuf, C., Borre, M., Dyrskjøt, L. and Ørntoft, T.F. 2011. Comprehensive genome methylation analysis in bladder cancer: Identification and validation of novel methylated genes and application of these as urinary tumor markers. *Clinical Cancer Research*. 17, pp.5582–5592.

Rhee, H.S. and Pugh, B.F. 2011. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*. 147, pp.1408–1419.

Rinn, J.L., Rozowsky, J.S., Laurenzi, I.J., Petersen, P.H., Zou, K., Zhong, W., Gerstein, M. and Snyder, M. 2004. Major molecular differences between mammalian sexes are involved in drug metabolism and renal function. *Developmental Cell*. 6, pp.791–800.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W. and Smyth, G.K. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*. 43, pp.e47–e47.

Rivera, C.M. and Ren, B. 2013. Mapping human epigenomes. *Cell*. 155, pp.39–55.

Robertson, A.G., Kim, J., Al-Ahmadie, H., Bellmunt, J., Guo, G., Cherniack, A.D., Hinoue, T., Laird, P.W., Hoadley, K.A., Akbani, R., Castro, M.A.A., Zwarthoff, E.C., et al. 2017. Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer. *Cell*. 171, pp.540-556.e25.

Robinson, G., Parker, M., Kranenburg, T.A., Lu, C., Chen, X., Ding, L., Phoenix, T.N., Hedlund, E., Wei, L., Zhu, X., Chalhoub, N., Baker, S.J., Huether, R., Kriwacki, R., Curley, N., Thiruvenkatam, R., Wang, J., Wu, G., Rusch, M., Hong, X., Becksfort, J., Gupta, P., Ma, J., Easton, J., Vadodaria, B., Onar-Thomas, A., Lin, T., Li, S., Pounds, S., Paugh, S., Zhao, D., Kawauchi, D., Roussel, M.F., Finkelstein, D., Ellison, D.W., Lau, C.C., Bouffet, E., Hassall, T., Gururangan, S., Cohn, R., Fulton, R.S., Fulton, L.L., Dooling, D.J., Ochoa, K., Gajjar, A., Mardis, E.R., Wilson, R.K., Downing, J.R., Zhang, J. and Gilbertson, R.J. 2012. Novel mutations target distinct subgroups of medulloblastoma. *Nature*. 488, pp.43–48.

Robinson, J. 2012. Integrated genomics viewer. *Nature Biotechnology*. 29, pp.24–26.

Robinson, M.D., McCarthy, D.J. and Smyth, G.K. 2009. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 26, pp.139–140.

Rosik, L., Niegisch, G., Fischer, U., Jung, M., Schulz, W.A. and Hoffmann, M.J. 2014. Limited efficacy of specific HDAC6 inhibition in urothelial cancer cells. *Cancer Biology & Therapy*. 15, 1, pp1-8

Rossetto, D., Avvakumov, N. and C??t??, J. 2012. Histone phosphorylation: A chromatin modification involved in diverse nuclear events. *Epigenetics*. 7, pp.1098–1108.

Rothbart, S.B., Dickson, B.M., Raab, J.R., Grzybowski, A.T., Krajewski, K., Guo, A.H., Shanle, E.K., Josefowicz, S.Z., Fuchs, S.M., Allis, C.D., Magnuson, T.R., Ruthenburg, A.J. and Strahl, B.D. 2015. An Interactive Database for the Assessment of Histone Antibody Specificity. *Molecular Cell*. 59, pp.502–511.

Rowley, M.J. and Corces, V.G. 2018. Organizational principles of 3D genome architecture. *Nature Reviews Genetics*. 19, pp.789–800.

Sahu, B., Laakso, M., Ovaska, K., Mirtti, T., Lundin, J., Rannikko, A., Sankila, A., Turunen, J.P., Lundin, M., Konsti, J., Vesterinen, T., Nordling, S., Kallioniemi, O., Hautaniemi, S. and Jännne, O.A. 2011. Dual role of FoxA1 in androgen receptor binding to chromatin, androgen signalling and prostate cancer. *EMBO Journal*. 30, pp.3962–3976.

Said, N., Sanchez-Carbayo, M., Smith, S.C. and Theodorescu, D. 2012. RhoGDI2 suppresses lung metastasis in mice by reducing tumor versican expression and macrophage infiltration. *Journal of Clinical Investigation*. 122, pp.1503–1518.

Sanger, F., Nicklen, S. and Coulson, R. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*. 74, pp.5463–5467.

Sanli, O., Dobruch, J., Knowles, M.A., Burger, M., Alemozaffar, M., Nielsen, M.E. and Lotan, Y. 2017. Bladder cancer. *Nature Reviews Disease Primers*. 3, p.17022.

de Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B.K., Muller, H., Ragoussis, J., Wei, C.L. and Natoli, G. 2010. A large fraction of extragenic RNA Pol II transcription sites overlap enhancers. *PLoS Biology*. 8.

Sauter, G., Moch, H., P, C., R, K., Mihatsch, M.. and Waldman, F.. 1995. Y Chromosome Loss Detected by FISH in Bladder Cancer. *Cancer Genetics Cytogenetics*. 82, pp.163–169.

Scharff, A.Z., Rousseau, M., Mariano, L.L., Canton, T., Consiglio, C.R., Albert, M.L., Fontes, M., Duffy, D. and Ingersoll, M.A. 2019. Sex differences in IL-17 contribute to chronicity in male versus female urinary tract infection. *The Jounal of Clinical Investigation Insight*. 4, p.e122998.

Schep, A.N., Buenrostro, J.D., Denny, S.K., Schwartz, K., Sherlock, G. and Greenleaf, W.J. 2015. Structured nucleosome fingerprints enable high-resolution mapping of chromatin

architecture within regulatory regions. *Genome Research*. 25, pp.1757–1770.

Schmidl, C., Rendeiro, A.F., Sheffield, N.C. and Bock, C. 2015. ChIPmentation: fast, robust, low-input ChIP-seq for histones and transcription factors. *Nature Methods*. 12, pp.963–5.

Schmitt, A.D., Hu, M. and Ren, B. 2016. Genome-wide mapping and analysis of chromosome architecture. *Nature Reviews Molecular Cell Biology*. 17, pp.743–755.

Schneider, R., Bannister, A.J., Myers, F.A., Thorne, A.W., Crane-Robinson, C. and Kouzarides, T. 2004. Histone H3 lysine 4 methylation patterns in higher eukaryotic genes. *Nature Cell Biology*. 6, pp.73–77.

Schoenfelder, S. and Fraser, P. 2019. Long-range enhancer–promoter contacts in gene expression control. *Nature Reviews Genetics*. 20.

Schones, D.E., Cui, K.R., Cuddapah, S., Roh, T.Y., Barski, A., Wang, Z.B., Wei, G. and Zhao, K.J. 2008. Dynamic regulation of nucleosome positioning in the human genome. *Cell*. 132, pp.887–898.

Sekiya, T. and Zaret, K.S. 2007. Repression by Groucho/TLE/Grg Proteins: Genomic Site Recruitment Generates Compacted Chromatin In Vitro and Impairs Activator Binding In Vivo. *Molecular Cell*. 28, pp.291–303.

Sengoku, T. and Yokoyama, S. 2011. Structural basis for histone H3 Lys 27 demethylation by UTX / KDM6A. *Genes and Development*. 25, pp.2266–2277.

Shan, Q., Zeng, Z., Xing, S., Li, F., Hartwig, S.M., Gullicksrud, J.A., Kurup, S.P., Van Braeckel-Budimir, N., Su, Y., Martin, M.D., Varga, S.M., Taniuchi, I., Harty, J.T., Peng, W., Badovinac, V.P. and Xue, H.H. 2017. The transcription factor Runx3 guards cytotoxic CD8 + effector T cells against deviation towards follicular helper T cell lineage. *Nature Immunology*. 18, pp.931–939.

Shechter, D., Dormann, H.L., Allis, C.D. and Hake, S.B. 2007. Extraction, purification and analysis of histones. *Nature Protocol*. 2, pp.1445–1457.

Shen, E.Y., Ahern, T.H., Cheung, I., Straubhaar, J., Dincer, A., Houston, I., de Vries, G.J., Akbarian, S. and Forger, N.G. 2015. Epigenetics and sex differences in the brain: A genome-wide comparison of histone-3 lysine-4 trimethylation (H3K4me3) in male and female mice. *Experimental Neurology*. 268, pp.21–29.

Shen, S.S., Smith, C.L., Hsieh, J.-T., Yu, J., Kim, I.Y., Jian, W., Sonpavde, G., Ayala, G.E., Younes, M. and Lerner, S.P. 2006. Expression of estrogen receptors-α and -β in bladder cancer cell lines and human bladder tumor tissue. *Cancer*. 106, pp.2610–2616.

Shendure, J., Balasubramanian, S., Church, G.M., Gilbert, W., Rogers, J., Schloss, J.A. and Waterston, R.H. 2017. DNA sequencing at 40: Past, present and future. *Nature*. 550, pp.345–353.

Sherr, C.J. and McCormick, F. 2002. The RB and p53 pathways in cancer. *Cancer Cell*. 2, pp.103–12.

Shimada, K., Ishikawa, T., Nakamura, F., Shimizu, D., Chishima, T., Ichikawa, Y., Sasaki, T., Endo, I., Nagashima, Y. and Goshima, Y. 2014. Collapsin response mediator protein 2 is involved in regulating breast cancer progression. *Breast Cancer*. 21, pp.715–723.

Shlyueva, D., Stampfel, G. and Stark, A. 2014. Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics*. 15, pp.272–86.

Shou, J., Lai, Y., Xu, J. and Huang, J. 2016. Prognostic value of FOXA1 in breast cancer: A systematic review and meta-analysis. *Breast*. 27, pp.35–43.

Shpargel, K.B., Sengoku, T., Yokoyama, S. and Magnuson, T. 2012. UTX and UTY Demonstrate Histone Demethylase-Independent Function in Mouse Embryonic Development. *PLoS Genetics*. 8, pp.1-9

Si, M. and Lang, J. 2018. The roles of metallothioneins in carcinogenesis. *Journal of Hematology and Oncology*. 11, pp.1–20.

Siegel, R.L., Miller, K.D. and Jemal, A. 2017. Cancer Statistics, 2017. *Cancer Journal for Clinicians*. 67, pp.7–30.

Siegsmund, M.J., Marx, C., Seemann, O., Schummer, B., Steidler, A., Toktomambetova, L., Köhrmann, K.U., Rassweiler, J. and Alken, P. 1999. Cisplatin-resistant bladder carcinoma cells: Enhanced expression of metallothioneins. *Urological Research*. 27, pp.157–163.

Singmann, P., Shem-Tov, D., Wahl, S., Grallert, H., Fiorito, G., Shin, S.-Y., Schramm, K., Wolf, P., Kunze, S., Baran, Y., Guarrera, S., Vineis, P., Krogh, V., Panico, S., Tumino, R., Kretschmer, A., Gieger, C., Peters, A., Prokisch, H., Relton, C.L., Matullo, G., Illig, T., Waldenberger, M. and Halperin, E. 2015. Characterization of whole-genome autosomal differences of DNA methylation between men and women. *Epigenetics & Chromatin*. 8, p.43.

Siprashvili, Z., Webster, D.E., Johnston, D., Shenoy, R.M., Ungewickell, A.J., Bhaduri, A., Flockhart, R., Zarnegar, B.J., Che, Y., Meschi, F., Puglisi, J.D. and Khavari, P.A. 2015. The noncoding RNAs SNORD50A and SNORD50B bind K-Ras and are recurrently deleted in human cancer. *Nature Genetics*. 48, pp.53–58.

Skene, P.J. and Henikoff, S. 2017. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *eLife*. 6, pp.1–35.

Snyder, B., Shell, B., Cunningham, J.T. and Cunningham, R.L. 2017. Chronic intermittent hypoxia induces oxidative stress and inflammation in brain regions associated with early-stage neurodegeneration. *Physiological Reports*. 5, pp.1–13.

Sotoodehnejadnematalahi, F. and Burke, B. 2013. Structure, function and regulation of

versican: the most abundant type of proteoglycan in the extracellular matrix. *Acta medica Iranica*. 51, pp.740–50.

Sr, P.J.L., Einhorn, L.H., Elson, P.J., Crawford, E.D. and ... 1992. A Randomized Comparison of Cisplatin Alone or in Combination With Methotrexate, Vinblastine, and Doxorubicin in Patients With Metastatic Urothelial Carcinoma: A Cooperative Group Study. *Journal of Clinical Oncology*. 10, pp.1066–1073.

Stein, J.P., Lieskovsky, G., Cote, R., Groshen, S., Feng, A.-C., Boyd, S., Skinner, E., Bochner, B., Thangathurai, D., Mikhail, M., Raghavan, D. and Skinner, D.G. 2001. Radical cystectomy in the treatment of invasive bladder cancer: Long-term results in 1,054 patients. *Journal of Clinical Oncology*. 19, pp.666–675.

Stern, M.C., Lin, J., Figueroa, J.D., Kelsey, K.T., Kiltie, A.E., Yuan, J.-M., Matullo, G., Fletcher, T., Benhamou, S., Taylor, J.A., Placidi, D., Zhang, Z.-F., Steineck, G., Rothman, N., Kogevinas, M., Silverman, D., Malats, N., Chanock, S., Wu, X., Karagas, M.R., Andrew, A.S., Nelson, H.H., Bishop, D.T., Sak, S.C., Choudhury, A., Barrett, J.H., Elliot, F., Corral, R., Joshi, A.D., Gago-Dominguez, M., Cortessis, V.K., Xiang, Y.-B., Gao, Y.-T., Vineis, P., Sacerdote, C., Guarrera, S., Polidoro, S., Allione, A., Gurzau, E., Koppova, K., Kumar, R., Rudnai, P., Porru, S., Carta, A., Campagna, M., Arici, C., Park, S.S.L. and Garcia-Closas, M. 2009. Polymorphisms in DNA Repair Genes, Smoking, and Bladder Cancer Risk: Findings from the International Consortium of Bladder Cancer. *Cancer Research*. 69, pp.6857–6864.

Stransky, N., Vallot, C., Reyal, F., Bernard-Pierrot, I., De Medina, S.G.D., Segraves, R., De Rycke, Y., Elvin, P., Cassidy, A., Spraggon, C., Graham, A., Southgate, J., Asselain, B., Allory, Y., Abbou, C.C., Albertson, D.G., Thiery, J.P., Chopin, D.K., Pinkel, D. and Radvanyi, F. 2006. Regional copy number-independent deregulation of transcription in cancer. *Nature Genetics*. 38, pp.1386–1396.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M. a, Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. and Mesirov, J.P. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*. 102, pp.15545–50.

Sugathan, A. and Waxman, D.J. 2013. Genome-Wide Analysis of Chromatin States Reveals Distinct Mechanisms of Sex-Dependent Gene Regulation in Male and Female Mouse Liver. *Molecular and Cellular Biology*. 33, pp.3594–3610.

Sze, C.C. and Shilatifard, A. 2016. MLL3/MLL4/COMPASS Family on Epigenetic Regulation of Enhancer Function and Cancer. *Cold Spring Harbor Laboratory Press Research.*, pp.1–16.

Szerlong, H.J. and Hansen, J.C. 2011. Nucleosome distribution and linker DNA: connecting nuclear function to dynamic chromatin structure. *Biochemistry and Cell Biology*. 89, pp.24–34.

Tafani, M., Sansone, L., Limana, F., Arcangeli, T., De Santis, E., Polese, M., Fini, M. and Russo, M.A. 2016. The Interplay of Reactive Oxygen Species, Hypoxia, Inflammation, and Sirtuins in Cancer Initiation and Progression. *Oxidative Medicine and Cellular Longevity*. 2016.

Tan, M.E., Li, J., Xu, H.E., Melcher, K. and Yong, E.L. 2015a. Androgen receptor: Structure, role in prostate cancer and drug discovery. *Acta Pharmacologica Sinica*. 36, pp.3–23.

Tan, W., Boorjian, S., Advani, P., Farmer, S., Lohse, C., Cheville, J., Kwon, E. and Leibovich, B. 2015b. The Estrogen Pathway: Estrogen Receptor-α, Progesterone Receptor, and Estrogen Receptor-β Expression in Radical Cystectomy Urothelial Cell Carcinoma Specimens. *Clinical Genitourinary Cancer*. 13, pp.476–484.

Tannenbaum, C. and Day, D. 2017. Age and sex in drug development and testing for adults. *Pharmacological Research*. 121, pp.83–93.

Tellier, J., Shi, W., Minnich, M., Liao, Y., Crawford, S., Smyth, G.K., Kallies, A., Busslinger, M. and Nutt, S.L. 2016. Blimp-1 controls plasma cell function through the regulation of immunoglobulin secretion and the unfolded protein response. *Nature Immunology*. 17, pp.323–330.

Teng, J., Wang, Z.Y., Jarrard, D.F. and Bjorling, D.E. 2008. Roles of estrogen receptor α and β in modulating urothelial cell proliferation. *Endocrine-Related Cancer*. 15, pp.351–364.

Tessarz, P. and Kouzarides, T. 2014. Histone core modifications regulating nucleosome structure and dynamics. *Nature Reviews Molecular Cell Biology*. 15, pp.703–8.

Thomson, J.P., Skene, P.J., Selfridge, J., Clouaire, T., Guy, J., Webb, S., Kerr, A.R.W., Deaton, A., Andrews, R., James, K.D., Turner, D.J., Illingworth, R. and Bird, A. 2010. CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature*. 464, pp.1082–6.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., Garg, K., Stamatoyannopoulos, J.A., et al. 2012. The accessible chromatin landscape of the human genome. *Nature*. 489, pp.75–82.

Tomlins, S.A., Rhodes, D.R., Perner, S., Dhanasekaran, S.M., Mehra, R., Sun, X.-W., Varambally, S., Cao, X., Tchinda, J., Kuefer, R., Lee, C., Montie, J.E., Shah, R.B., Pienta, K.J., Rubin, M.A. and Chinnaiyan, A.M. 2005. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*. 310, pp.644–8.

Trojer, P. and Reinberg, D. 2007. Facultative Heterochromatin: Is There a Distinctive Molecular Signature? *Molecular Cell*. 28, pp.1–13.

Tukiainen, T., Villani, A., Yen, A., Rivas, M.A., Marshall, J.L., Satija, R., Aguirre, M., Gauthier, L., Fleharty, M., Kirby, A., Cummings, B.B., Castel, S.E., Karczewski, K.J., Aguet, F., Byrnes, A., Consortium, G., Lappalainen, T., Regev, A., Ardlie, K.G.,

Hacohen, N. and Macarthur, D.G. 2017. Landscape of X chromosome inactivation across human tissues. *Nature*. 550, pp.244–248.

Turner, F.S. 2014. Assessment of insert sizes and adapter content in fastq data from NexteraXT libraries. *Frontiers in Genetics*. 5.

Vallot, C., Herault, A., Boyle, S., Bickmore, W.A. and Radvanyi, F. 2015. PRC2-independent chromatin compaction and transcriptional repression in cancer. *Oncogene*. 34, pp.741–751.

Vallot, C., Stransky, N., Bernard-Pierrot, I., H??rault, A., Zucman-Rossi, J., Chapeaublanc, E., Vordos, D., Laplanche, A., Benhamou, S., Lebret, T., Southgate, J., Allory, Y. and Radvanyi, F. 2011. A novel epigenetic phenotype associated with the most aggressive pathway of bladder tumor progression. *Journal of the National Cancer Institute*. 103, pp.47–60.

Vasudevan, A.A.J., Hoffmann, M.J., Beck, M.L.C., Poschmann, G., Petzsch, P., Wiek, C., Stühler, K., Köhrer, K., Schulz, W.A. and Niegisch, G. 2019. HDAC5 expression in urothelial carcinoma cell lines inhibits long-term proliferation but can promote epithelial-to-mesenchymal transition. *International Journal of Molecular Sciences*. 20.

Walport, L.J., Hopkinson, R.J., Vollmar, M., Madden, S.K., Gileadi, C., Oppermann, U., Schofield, C.J. and Johansson, C. 2014. Human UTY (KDM6C) Is a Male-specific N ♬ -Methyl Lysyl. *Journal of Biological Chemistry*. 289, pp.18302–18313.

Wang, J., Zibetti, C., Shang, P., Sripathi, S.R., Zhang, P., Cano, M., Hoang, T., Xia, S., Ji, H., Merbs, S.L., Zack, D.J., Handa, J.T., Sinha, D., Blackshaw, S. and Qian, J. 2018. ATAC-Seq analysis reveals a widespread decrease of chromatin accessibility in age-related macular degeneration. *Nature Communications*. 9, pp.1–13.

Wang, S.P., Tang, Z., Chen, C.W., Shimada, M., Koche, R.P., Wang, L.H., Nakadai, T., Chramiec, A., Krivtsov, A. V., Armstrong, S.A. and Roeder, R.G. 2017. A UTX-MLL4-p300 Transcriptional Regulatory Network Coordinately Shapes Active Enhancer Landscapes for Eliciting Transcription. *Molecular Cell*. 67, pp.308-321.e6.

Warrick, J.I., Walter, V., Yamashita, H., Chung, E., Shuman, L., Amponsa, V.O., Zheng, Z., Chan, W., Whitcomb, T.L., Yue, F., Iyyanki, T., Kawasawa, Y.I., Kaag, M., Guo, W., Raman, J.D., Park, J.S. and Degraff, D.J. 2016. FOXA1, GATA3 and PPARIγ Cooperate to drive luminal subtype in bladder cancer: A molecular analysis of established human cell lines. *Scientific Reports*. 6, pp.1–15.

Wederell, E.D., Bilenky, M., Cullum, R., Thiessen, N., Dagpinar, M., Delaney, A., Varhol, R., Zhao, Y., Zeng, T., Bernier, B., Ingham, M., Hirst, M., Robertson, G., Marra, M.A., Jones, S. and Hoodless, P.A. 2008. Global analysis of in vivo Foxa2-binding sites in mouse adult liver using massively parallel sequencing. *Nucleic Acids Research*. 36, pp.4549–4564.

Wei, C.L., Wu, Q., Vega, V.B., Chiu, K.P., Ng, P., Zhang, T., Shahab, A., Yong, H.C., Fu, Y.T., Weng, Z., Liu, J., Zhao, X.D., Chew, J.L., Lee, Y.L., Kuznetsov, V.A., Sung, W.K., Miller, L.D., Lim, B., Liu, E.T., Yu, Q., Ng, H.H. and Ruan, Y. 2006. A global map of p53 transcription-factor binding sites in the human genome. *Cell*. 124, pp.207–219.

Wei, Z. and Liu, H.T. 2002. MAPK signal pathways in the regulation of cell proliferation in mammalian cells. *Cell Research*. 12, pp.9–18.

Weinstein, J.N., Akbani, R., Broom, B.M., Wang, W., Verhaak, R.G.W., McConkey, D., Lerner, S., Morgan, M., Creighton, C.J., Smith, C., Kwiatkowski, D.J., Eley, G., et al. 2014. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature*. 507, pp.315–322.

Welch, R.P., Lee, C., Imbriano, P.M., Patil, S., Weymouth, T.E., Smith, R.A., Scott, L.J. and Sartor, M.A. 2014. ChIP-Enrich: Gene set enrichment testing for ChIP-seq data. *Nucleic Acids Research*. 42.

Wernig, M., Zhao, J.-P., Pruszak, J., Hedlund, E., Fu, D., Soldner, F., Broccoli, V., Constantine-Paton, M., Isacson, O. and Jaenisch, R. 2008. Neurons derived from reprogrammed fibroblasts functionally integrate into the fetal brain and improve symptoms of rats with Parkinson's disease. *Proceedings of the National Academy of Sciences of the United States of America*. 105, pp.5856–61.

Woldu, S.L., Bagrodia, A. and Lotan, Y. 2017. Guideline of guidelines: non-muscle-invasive bladder cancer. *BJU International*. 119, pp.371–380.

Wolff, E.M., Chihara, Y., Pan, F., Weisenberger, D.J., Siegmund, K.D., Sugano, K., Kawashima, K., Laird, P.W., Jones, P.A. and Liang, G. 2010. Unique DNA methylation patterns distinguish noninvasive and invasive urothelial cancers and establish an epigenetic field defect in premalignant tissue. *Cancer Research*. 70, pp.8169–8178.

Wu, C. 1980. The 5' ends of drosophila heat shock genes in chromatin are hypersensitive to DNase I. *Nature*. 286, pp.854–860.

Wu, C.C., Chen, H.C., Chen, S.J., Liu, H.P., Hsieh, Y.Y., Yu, C.J., Tang, R., Hsieh, L.L., Yu, J.S. and Chang, Y.S. 2008. Identification of collapsin response mediator protein-2 as a potential marker of colorectal carcinoma by comparative analysis of cancer cell secretomes. *Proteomics*. 8, pp.316–332.

Wu, J.-T., Han, B.-M., Yu, S.-Q., Wang, H.-P. and Xia, S.-J. 2010. Androgen Receptor Is a Potential Therapeutic Target for Bladder Cancer. *Urology*. 75, pp.820–827.

Wu, J., Huang, B., Chen, H., Yin, Q., Liu, Y., Xiang, Y., Zhang, B., Liu, B., Wang, Q., Xia, W., Li, W., Li, Y., Ma, J., Peng, X., Zheng, H., Ming, J., Zhang, W., Zhang, J., Tian, G., Xu, F., Chang, Z., Na, J., Yang, X. and Xie, W. 2016. The landscape of accessible chromatin in mammalian preimplantation embryos. *Nature*. 534, pp.652–7.

Wu, S., Yang, Z., Ye, R., An, D., Li, C., Wang, Yitian, Wang, Yongqiang, Huang, Y., Liu, H.,

Li, F., He, L., Sun, D., Yu, Y., Li, Q., Huang, P., Zhang, M., Zhao, X., Bi, T., Zhuang, X., Zhang, L., Lu, J., Sun, X., Zhou, F., Liu, C., Yang, G., Hou, Y., Fan, Z. and Cai, Z. 2016. Novel variants in MLL confer to bladder cancer recurrence identified by whole-exome sequencing. *Oncotarget*. 7, pp.2629–2645.

Wu, Y., Siadaty, M.S., Berens, M.E., Hampton, G.M. and Theodorescu, D. 2008. Overlapping gene expression profiles of cell migration and tumor invasion in human bladder cancer identify metallothionein 1E and nicotinamide N-methyltransferase as novel regulators of cell migration. *Oncogene*. 27, pp.6679–6689.

Wülfing, C., van Ahlen, H., Eltze, E., Piechota, H., Hertle, L. and Schmid, K.W. 2007. Metallothionein in bladder cancer: Correlation of overexpression with poor outcome after chemotherapy. *World Journal of Urology*. 25, pp.199–205.

Xu, J., Deng, X., Watkins, R. and Disteche, C.M. 2008. Sex-Specific Differences in Expression of Histone Demethylases Utx and Uty in Mouse Brain and Neurons. *The Journal of Neuroscience*. 28, pp.4521–4527.

Xu, W., Seok, J., Mindrinos, M.N., Schweitzer, A.C., Jiang, H., Wilhelmy, J., Clark, T.A., Kapur, K., Xing, Y., Faham, M., Storey, J.D., Moldawer, L.L., Maier, R. V., Tompkins, R.G., Wong, W.H., Davis, R.W. and Xiao, W. 2011. Human transcriptome array for high-throughput clinical studies. *Proceedings of the National Academy of Sciences*. 108, pp.3707–3712.

Xu, W., Yang, L., Lee, P., Huang, W.C., Nossa, C., Ma, Y., Deng, F.-M., Zhou, M., Melamed, J. and Pei, Z. 2014. Mini-review: perspective of the microbiome in the pathogenesis of urothelial carcinoma. *American Journal of Clinical and Experimental Urology*. 2, pp.57–61.

Yang, L., Taylor, J., Eustace, A., Irlam, J.J., Denley, H., Hoskin, P.J., Alsner, J., Buffa, F.M., Harris, A.L., Choudhury, A. and West, C.M.L. 2017. A gene signature for selecting benefit from hypoxia modification of radiotherapy for high-risk bladder cancer patients. *Clinical Cancer Research*. 23, pp.4761–4768.

Yates, D.R., Rehman, I., Abbod, M.F., Meuth, M., Cross, S.S., Linkens, D.A., Hamdy, F.C. and Catto, J.W.F. 2007. Promoter hypermethylation identifies progression risk in bladder cancer. *Clinical Cancer Research*. 13, pp.2046–2053.

Yen, A. and Kellis, M. 2015. Systematic chromatin state comparison of epigenomes associated with diverse properties including sex and tissue type. *Nature Communications*. 6, p.7973.

Yong, W.S., Hsu, F.M. and Chen, P.Y. 2016. Profiling genome-wide DNA methylation. *Epigenetics and Chromatin*. 9, pp.1–16.

Zentner, G.E. and Henikoff, S. 2013. Regulation of nucleosome dynamics by histone modifications. *Nature Structural and Molecular Biology*. 20, pp.259–66.

Zentner, G.E. and Scacheri, P.C. 2012. The chromatin fingerprint of gene enhancer

elements. *Journal of Biological Chemistry*. 287, pp.30888–30896.

Zhai, Z., Liu, W., Kaur, M., Luo, Y., Domenico, J., Samson, J.M., Shellman, Y.G., Norris, D.A., Dinarello, C.A., Spritz, R.A. and Fujita, M. 2017. NLRP1 promotes tumor growth by enhancing inflammasome activation and suppressing apoptosis in metastatic melanoma. *Oncogene*. 36, pp.3820–3830.

Zhang, L., Xue, G., Liu, J., Li, Q. and Wang, Y. 2018. Revealing transcription factor and histone modification co-localization and dynamics across cell lines by integrating ChIP-seq and RNA-seq data. *BMC Genomics*. 19.

Zhang, Y. 2013. Understanding the gender disparity in bladder cancer risk: The impact of sex hormones and liver on bladder susceptibility to carcinogens. *Journal of Environmental Science and Health - Part C Environmental Carcinogenesis and Ecotoxicology Reviews*. 31, pp.287–304.

Zhang, Y., Klein, K., Sugathan, A., Nassery, N., Dombkowski, A., Zanger, U.M. and Waxman, D.J. 2011. Transcriptional profiling of human liver identifies sex-biased genes associated with polygenic dyslipidemia and coronary artery disease. *PLoS ONE*. 6.

Zhao, Z., Tavoosidana, G., Sjölinder, M., Göndör, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Singh Sandhu, K., Singh, U., Pant, V., Tiwari, V., Kurukuti, S. and Ohlsson, R. 2006. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature Genetics*. 38, pp.1341–1347.

Zheng, H. and Xie, W. 2019. The role of 3D genome organization in development and cell differentiation. *Nature Reviews Molecular Cell Biology*. 20, pp.535–550.

Zheng, Y., Izumi, K., Yao, J.L. and Miyamoto, H. 2011. Dihydrotestosterone upregulates the expression of epidermal growth factor receptor and ERBB2 in androgen receptor-positive bladder cancer cells. *Endocrine-Related Cancer*. 18, pp.451–464.

Zhu, J., Adli, M., Zou, J.Y., Verstappen, G., Coyne, M., Zhang, X., Durham, T., Miri, M., Deshpande, V., De Jager, P.L., Bennett, D.A., Houmard, J.A., Muoio, D.M., Onder, T.T., Camahort, R., Cowan, C.A., Meissner, A., Epstein, C.B., Shoresh, N. and Bernstein, B.E. 2013. Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell*. 152, pp.642–654.