Selective pressure on the HIF pathway in renal cancer



Virginia Schmid

Brasenose College,

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy (DPhil)

Trinity, 2019

Abstract

The activation of large transcriptional pathways with heterogeneous effects on tumour biology is a common feature of many types of cancer. This is exemplified by the induction of hypoxia pathways, orchestrated by the transcription factor HIF, which occur as a result of loss of the VHL tumour suppressor in clear cell renal cell cancer (RCC). This entrains multiple responses that both promote and inhibit tumour formation with the overall consequence of HIF activation being a subtle balance of these numerous (often small) effects. I postulate that the development of kidney cancer provides a Darwinian pressure that selects for variants (either germline or somatic) that alter specific components of the HIF transcriptional output and re-balance its ability to promote tumourigenesis. The aim of this thesis was to examine for evidence of this selective pressure by studying germline variants that pre-dispose to kidney cancer.

My work has revealed a significant excess overlap between RCC-associated susceptibility loci and individual components of the HIF transcriptional pathway. This overlap also extended to include RCC-susceptibility loci with levels of association that fall below conventional thresholds of genome-wide significance and, despite the importance of HIF in other tumour types, was unique to kidney cancer. In addition, susceptibility loci associated with breast cancer and with prostate cancer were enriched for oestrogen receptor binding sites and for androgen receptor binding sites respectively in the relevant cell types.

This analysis provides evidence that aspects of the HIF-pathway are under selective pressure in kidney cancer. It helps distinguish those aspects that drive tumour formation from those that are simply co-activated as a consequence of large-pathway activation. It provides a paradigm for studying genetic variants associated with other diseases, in particular with other cancer types. Finally, this work supports a ‘pathway tuning’ model of cancer and suggests that subtle re-balancing of cancer-specific activated pathways is important in oncogenesis.

Declaration

Except for where acknowledgements have been made, the work outlined in this thesis was performed by the author in the CCMP and NDM Research Buildings at the University of Oxford.

Experimental data presented in this thesis was processed and analysed by the author who did not contribute to generation of the data. The datasets analysed and integrated throughout this thesis were performed by various previous members of the Ratcliffe/Mole group.

This includes RNA-seq assays presented in chapter 4. Further, ChIP-seq assays of HIF binding integrated in chapter 2-6 were performed by James Platt and James Smythies. ChIP-seq assays of HIF binding presented chapter 7 were planned by the author and Olivia Lombardi and assays were subsequently performed by Olivia Lombardi. Capture-C experiments planned by the author and Veronique Lafleur and assays subsequently performed by Veronique Lafleur. Interacting regions were determined by Ran Li.

The work presented in this thesis has not been submitted for any other degree in this university, or any other education institute.

Acknowledgements

I would first like to thank Prof. David Mole for the supervision of this work, for supporting me in exploring my interests and the technical training I have received within the lab. I would like to thank Prof. Sir Peter Ratcliffe for the great care he has taken in supervising me and this work. The time in the lab has been very formative and enabling and also just a really great time and I will carry all of it forward to where I go professionally and in life.

I would also like to thank the people who have taught and more closely mentored me on technical aspects over the years. This includes Dr. Rafik Salama, Dr. Sarah Geisler and Dr. Federico Comoglio. I would here also like to thank all members of Ratcliffe group, especially also Cathrine King and further Veronique Lafleur, Olivia Lombardi and Ran Li for their important contributions.

Finally, my thanks go to my parents Franziska Peterhans, and Gregory Schmid and my two brothers Alexander Schmid and Tim Schmid. And my friends in Switzerland, South Africa and the UK for enriching my life, support and fun nights out, basketball games and discussions. Especially, I would also like to thank, Nada Kubikova for always having my back and Zina Zokouri for always coming to visit.

I dedicate this work to my lifelong friend Anna Mueller. Thank you for being my friend and for first opening up the world of maths and science to me.

Publications arising from this thesis

1. Grampp, S, V. Schmid, R. Salama, V. Lauer, F. Kranz, J. L. Platt, J. Smythies, H. Choudhry, M. Goppelt-Struebe, P. J. Ratcliffe, D. R. Mole, and J. Schodel. 2017. 'Multiple renal cancer susceptibility polymorphisms modulate the HIF pathway', PLoS Genet, 13: e1006872. https://doi.org/10.1371/journal.pgen.1006872, shared first authorship between Grampp, S. and V. Schmid

2. Virginia Schmid, Veronique Lafleur, Olivia Lombardi, Ran Li, Rafik Salama, Leandro Colli, Hani Choudhry, Stephen Chanock, Peter J Ratcliffe, David R Mole 'Co-incidence of RCC-susceptibility polymorphisms with HIF cis-acting sequences supports a pathway tuning model of cancer', Sci Rep 9, 18768 (2019) doi:10.1038/s41598-019-55098-7

Table of contents

Declaration **3**

Acknowledgements **4**

Publications arising from this thesis **5**

Table of contents **6**

Chapter 1: Introduction **12**

1.1 Introduction **13**

1.2 Large transcriptional-pathway activation in cancer 14

1.3 Activation of HIF-transcriptional pathway in cancer 18

1.4 Kidney cancer and HIF 23

1.5 The HIF-pathway – a double-edged sword in ccRCC. 26

1.6 Genome-wide association studies 30

1.7 Interpreting GWAS in the light of the epigenetic and transcriptional landscape 34

1.8 Aim of this study 38

Chapter 2: RCC susceptibility loci target HIF binding in RCC cells **40**

2.1 Introduction **41**

2.2 Results **44**

2.2.1 Definition of RCC susceptibility loci from independent GWA studies 44

2.2.2 Definition of high confidence HIF binding sites and enhancers in RCC cancer cell lines 46

2.2.3 RCC susceptibility loci overlap with binding sites of HIF 58

2.2.4 RCC susceptibility loci identified by independent GWA studies coincide HIF binding sites more than expected 63

2.3 Discussion **67**

Chapter 3: A meta-analysis-defined set of RCC susceptibility loci target HIF binding sites in RCC cancer cell lines. **71**

3.1 Introduction **72**

3.2 Results **75**

3.2.1 Definition of RCC susceptibility loci from a meta-analysis of GWA studies 75

3.2.2 RCC susceptibility loci identified by a meta-analysis of GWA studies directly target HIF binding sites 79

3.2.3 RCC susceptibility loci defined by a meta-analysis of GWA studies overlap with HIF binding sites more than expected 82

3.2.4 Educational attainment susceptibility loci coincide HIF binding sites at expected frequency 89

3.3 Discussion **91**

Chapter 4: Identification of target genes of RCC susceptibility loci **95**

4.1 Introduction **96**

4.2 Results **99**

4.2.1 Genes in genomic proximity to RCC susceptibility loci are enriched for genes regulated by the VHL/HIF pathway 99

4.2.2 Identification of distant sites that physically interact with regulatory elements at RCC susceptibility loci by Capture-C 106

4.2.3 Genes with promoters physically interacting with regulatory elements overlapping RCC susceptibility loci are enriched for HIF regulable genes 112

4.2.4 Induction of HIF-target genes in ccRCC samples from the TCGA KIRC cohort 118

4.2.5 The effect of genotype at RCC-susceptibility loci on the expression of direct HIF-target genes 121

4.2.6 The association between RCC-associated polymorphism at the 12p12.1 locus and expression of BHLHE41 is specific to tumours with high levels of HIF-1 124

4.3 Discussion **127**

Chapter 5: Additional mechanisms by which RCC susceptibility loci alter the HIF pathway in kidney cancer **131**

5.1 Introduction **132**

5.2. Results 134

5.2.1 Identification of RCC susceptibility loci in close genomic proximity to HIF binding sites 134

5.2.2 Statistical significance for the overlap and proximity of RCC-susceptibility loci with HIF binding sites 141

5.2.3 Expression of HIF-target genes in ccRCC samples from the TCGA KIRC cohort versus paired normal kidney 143

5.2.4 The effect of genotype at RCC-susceptibility loci on the expression of direct HIF-target genes 145

5.2.5 Identification of more distant HIF binding sites in physical contact with RCC-susceptibility loci using Capture-C 147

5.3 Discussion 153

Chapter 6: Identification of sub-threshold RCC susceptibility loci 155

6.1 Introduction 156

6.1 Results 160

6.2.1 Identification of sub-threshold RCC-susceptibility loci 160

6.2.2 Sub-threshold RCC-susceptibility loci are enriched for proximity to HIF regulated genes 164

6.2.3 Sub-threshold RCC susceptibility loci directly overlap HIF-binding sites 167

6.2.4 Sub-threshold RCC susceptibility loci coincide with HIF binding sites more than expected 168

6.2.5 Defining potential target genes of the HIF-binding site that overlap sub-threshold RCC-susceptibility loci. 172

6.2.6 Analysis of sub-threshold RCC-susceptibility loci that overlap HIF-binding sites using TCGA-KIRC datasets. 174

6.2.7 The 7p21.1 locus and AHR 179

6.3 Discussion 183

Chapter 7: Analysis of cancer-associated GWAS loci in the light of transcriptional pathway activation 186

7.1 Introduction 187

7.2 Results 190

7.2.1 Definition of high confidence HIF-1β binding sites in different cancer cell lines 190

7.2.2 Overlap between RCC susceptibility loci and HIF-binding sites identified in different cancer cell types 201

7.2.3 Overlap between cancer-susceptibility loci and HIF-1β binding sites identified in corresponding cancer cell lines. 208

7.2.4 Analysis of cancer susceptibility loci in the light of transcriptional pathway activation can be extended beyond the HIF pathway 213

7.3 Discussion 217

Chapter 8: Final discussion 219

8.1 Discussion 220

Chapter 9: Materials and Methods 228

9.1 Generation of Next Generation Sequencing datasets 229

9.1.1 Chromatin Immunoprecipitation 229

9.1.2 RNA isolation method 232

9.1.3 Capture-C 232

9.2 Processing and analysis of NGS datasets 236

9.2.1 Sequencing 236

9.2.2 Alignment of ChIP-seq data 236

9.2.3 Bigwig generation for ChIP-seq, RNA-seq and Capture-C datasets 236

9.2.4 Peak calling for ChIP-seq data 237

9.2.5 Motif analysis and MEME-suite on ChIP-seq peaks 237

9.2.6 Visualisation of ChIP-seq data by heatmaps and ngs.plot 238

9.2.7 Combining datasets using bedtools 238

9.2.8 RNA-seq analysis 239

9.2.9 Identification of nearest genes to a feature (e.g. index SNPs) 239
9.2.10 Pre-ranked Gene set enrichment analysis (GSEA) 240
9.2.11 Processing of Capture-C data by applying pipeline 241
9.2.12 Identification of significantly interacting regions by applying pipeline 241
9.3 Processing of GWAS datasets 242
9.3.1 1000 Genomes data 242
9.3.2 Plink 242
9.3.3 Manhattan plot 243
9.3.4 Definition of susceptibility loci associated haplotype blocks from summary statistics 243
9.3.5 Definition of susceptibility loci associated haplotype blocks from database entries 244
9.3.6 Expected frequency distribution of overlap of GWAS locus and HIF binding site .. 244
9.3.7 Expected frequency distribution of overlap of HIF binding site with HIF binding site 245
9.4 TCGA data 245
9.4.1 Datasets included 245
9.4.2 eQTL 246
9.4.3 meta-gene eQTL 246
9.4.4 Differential expression in tumour and normal tissue 246
Chapter 10: Appendix 247
10.1 IGV tracks RCC susceptibility loci 248
10.2 IGV tracks sub-threshold RCC susceptibility loci 274
10.3 Supplemental Tables 294
10.3.1 Breast cancer associated GWAS loci that overlap oestrogen receptor binding sites 294
10.3.2 Prostate cancer associated GWAS loci that overlap androgen receptor binding sites 294

10.3.3 Genes set containing three closest genes to sub-threshold RCC susceptibility loci 296

10.3.4 Genomic coordinates of Capture-C oligos 300

10.4 Code 303

10.4.1 LD pruning of summary statistics 303

10.4.2 Genome-wide shuffling expected frequency distribution 307

10.4.3 Enhancer-wide shuffling expected frequency distribution 308

10.4.4 eQTL 309

10.4.5 Gene expression tissue from tumour and normal kidney 310

Chapter 11: Bibliography 312

Chapter 1: Introduction

1.1 Introduction

Cancer is the collective term given to diseases in which abnormal cells divide without control and may invade nearby tissues or spread to distant parts of the body. With a lifetime incidence of approximately 1 in 2, cancer now accounts for about 28% of all UK deaths. While, breast, lung, prostate and bowel cancers currently account for just over half of all new cancer diagnoses in the UK, the incidence of other types of cancer such as kidney cancer, cutaneous melanoma, and head and neck cancers has increased markedly over the last decade (Cancer Research UK. https://www.cancerresearchuk.org/health-professional/cancer-statistics/incidence#heading-One. August 2019). However, cancer is not one single disease. The National Cancer Institute lists nearly 200 cancer types, and many of these (e.g. kidney cancer) can be sub-divided further (The National Cancer Institute https://www.cancer.gov/types August 2019). Whilst these cancers are highly heterogeneous, general molecular and cellular "hallmarks" that characterise cancer have been described and these are generally reflected by the activation of large-scale transcriptional programmes (Hanahan and Weinberg 2011; Hanahan 2014). In turn, many of these cancer-associated transcriptional responses are constitutively activated or repressed as a result of somatic mutation in key oncogenic or tumour suppressor pathways (Greenman et al. 2007).

1.2 Large transcriptional-pathway activation in cancer

The transcription factor P53 acts as a tumour suppressor and is the most frequently mutated gene in cancer (Lawrence et al. 2014). It is activated in response to stress signals such as DNA damage, oncogene activation, ribosomal stress, and tissue hypoxia, and its main function is the transcriptional regulation of target genes that are involved in numerous cellular processes, including cell cycle control and apoptosis (Beckerman and Prives 2010). It binds the consensus DNA sequence, RRRCWWGYYY (R = A/G, W = A/T, Y = C/T), as a tetramer and pangenomic analyses of p53 binding and gene regulation have identified approximately 3,500 p53 target genes, many of which were specific to a single cell type (Fischer 2017). While the focus has largely been on those genes that support the tumour suppressor role of p53, it seems inconceivable, given the extent of the transcriptional response, that all p53 target genes have a suppressive effect on tumourigenesis.

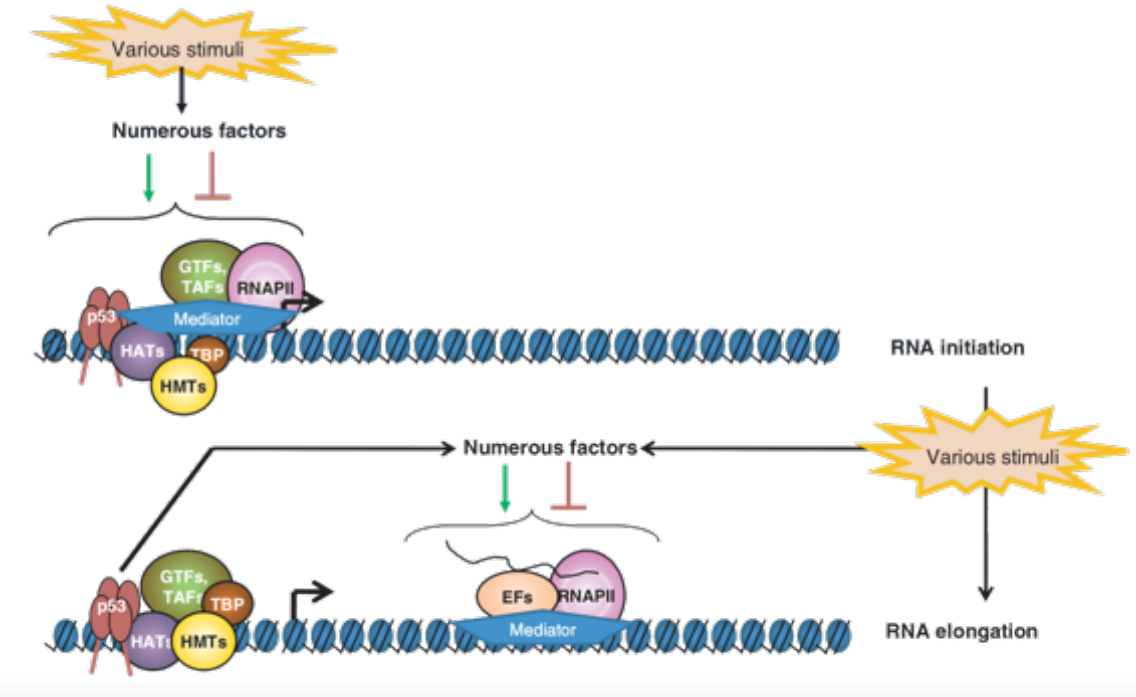


Figure 1.1: The p53 tumour suppressor as an example of a large pathway switch in cancer. (Top) p53 can direct assembly of the preinitiation complex (PIC) through ordered recruitment of histone methyltransferases (HMTs), histone acetyltransferase (HATs), and other coregulators to open up chromatin to allow RNAPII and its associated general transcription factors (GTFs) to bind to the transcription start site. (Bottom) p53 can also modulate transcription elongation via functional and physical interactions with various elongation factors. Adapted from Beckerman and Prives (Beckerman and Prives 2010).

Similarly, the MYC proto-oncogene is one of the most frequently activated oncogenes and is thought to be involved in the pathogenesis of 1 in 5 of all cancers (Nesbit et al. 1999; Dang 1999). Its gene product is a basic helix-loop-helix transcription factor that binds DNA as a heterodimer with MAX. Its canonical DNA-binding motif (5'-CACGTG-3'), known as an E-box is among the most frequently occurring in the human genome and it is estimated that approximately 15% of all genes are regulated by MYC (Park et al. 2004; Xie et al. 2005; Dang et al. 2006 ; Li et al. 2003). Whilst MYC-responsive genes have distinct functions, such as metabolism, protein synthesis, cell-cycle control, DNA replication, and cell adhesion, distinguishing between the physiologic and tumourigenic functions of MYC is, again, not straightforward. Indeed, in non-malignant cells, MYC may even induce checkpoint genes such as ARF or p53, which will restrict cell growth and, as a consequence, many MYC-induced lymphomas have inactivated these genes (Eischen et al. 1999; Zindy et al. 1998).

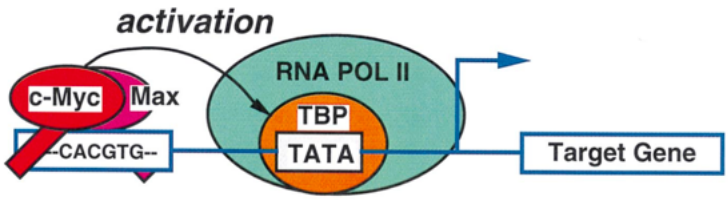


Figure 1.2: The MYC oncogene acts as a transcriptional regulator. (A) The Myc/Max heterodimer binds to the E box 5'-CACGTG-3' and mediates transcription by RNA-polymerase 2. Adapted from Dang et al. (Dang 1999).

It is estimated that approximately 20% of all human cancers have activating mutations in one of the Ras oncogenes (Rajalingam et al. 2007). Ras proteins are a large family of small GTPases that are activated in response to extracellular stimuli such as growth factors acting through receptor tyrosine kinases. They, in turn, activate several signalling cascades including widespread transcriptional responses mediated by the ETS1/2, JUN, FOS and ELK1 transcription factors, which regulate many hundreds of genes (Simanshu et al. 2017).

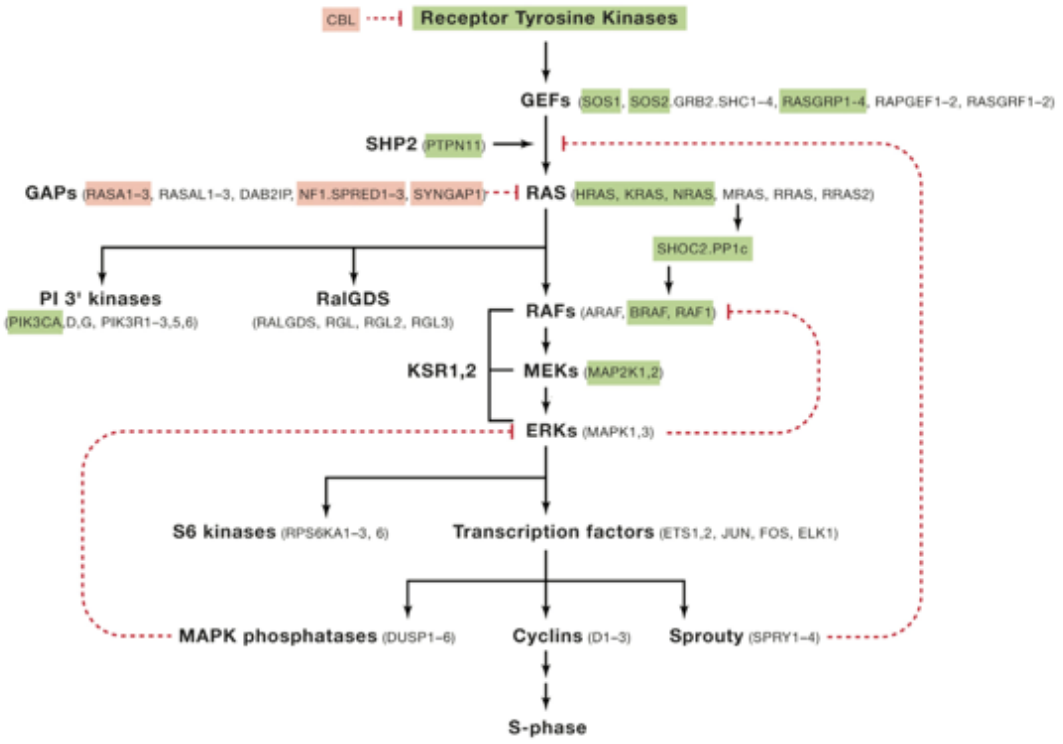


Figure 1.3: The RAS oncogene activates transcriptional responses mediated by multiple transcription factors. Adapted from Simanshu et al. (Simanshu et al. 2017).

Other transcription factors that are not classical oncogenes or tumour suppressor genes may be central to the pathogenesis of different types of cancer cancers. Key amongst these are the sex hormone receptors, which play a central role in the pathogenesis of two of the commonest types of cancer, breast cancer and prostate cancer, and antagonism of these receptors has a central role in the treatment of these tumours. Oestrogen-, progesterone- and androgen-receptor are nuclear receptors, that once bound by their respective hormone ligands bind to DNA and activate broad transcriptional repertoires (Groner and Brown 2017).

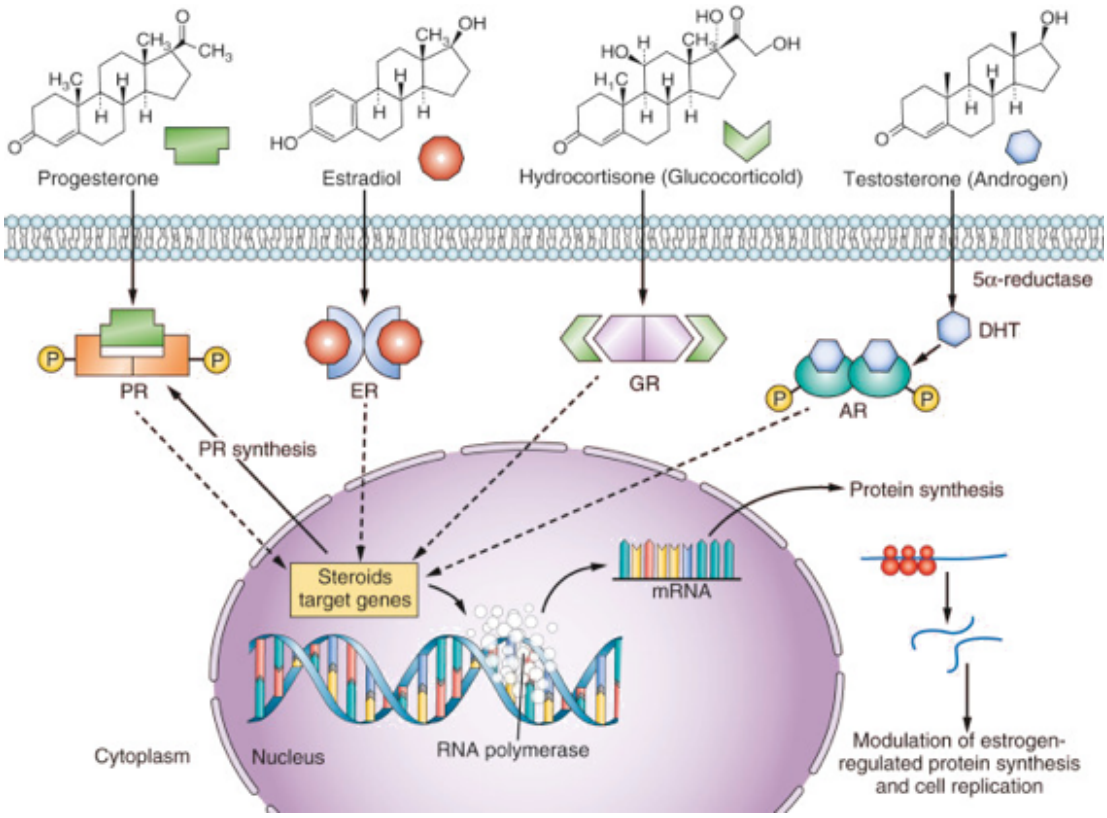


Figure 1.4: Steroid receptors drive transcriptional responses in breast and prostate cancer. Adapted from Abderrahman et al (Abderrahman and Jordan 2017)

1.3 Activation of HIF-transcriptional pathway in cancer

In addition to the transcriptional responses described above, transcriptional pathways that respond to altered levels of oxygen are also frequently activated in many types of cancer. The main regulator of this cellular transcriptional response to altered oxygen levels is the hypoxia-inducible factor (HIF) (Kaelin and Ratcliffe 2008; Semenza 2012). Activation of HIF contributes directly to many of the “hallmarks of cancer” described by Hanahan (see Figure 1.5) - most

classically angiogenesis and deregulation of cellular energetics/metabolic reprogramming.

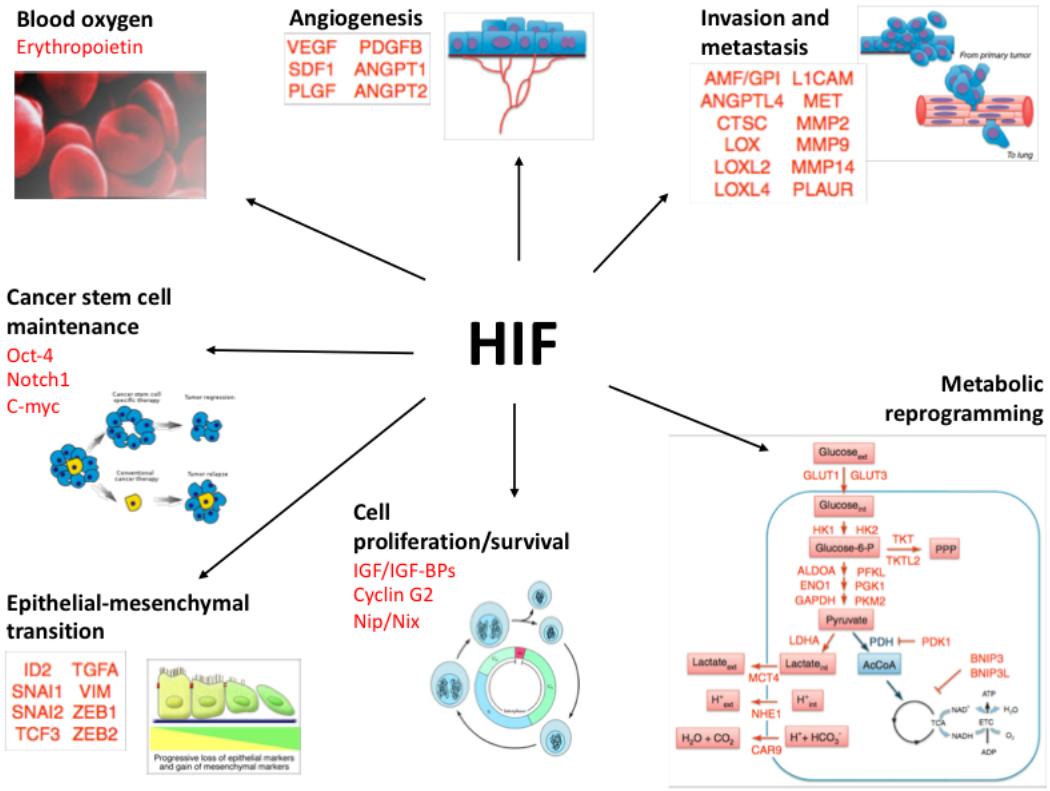


Figure 1.5: The HIF transcription factor activates genes involved in many of the "hallmarks" of cancer.

HIF has two widely expressed isoforms, HIF-1 and HIF-2, each composed of a heterodimer of an α- and β- subunit. The β-subunit, termed HIF-1β or ARNT (aryl hydrocarbon nuclear translocator), is common to both isoforms and dimerises with either HIF-1α or HIF-2α, also known as EPAS1 (endothelial PAS protein 1) (Wang and Semenza 1995; Jiang et al. 1996; Ema et al. 1997; Tian et al. 1997). HIF-1 is almost ubiquitously expressed, whilst HIF-2 has a more tissue-restricted pattern of expression (Maxwell et al. 1993; Tian et al. 1997). A third isoform, HIF-

3α may also be expressed in numerous splice variants, the functions of which are poorly understood (Gu et al. 1998).

All three proteins are members of the bHLH- (basic helix-loop-helix) PAS (PER-AHR/ARNT-SIM) family of transcription factors. The basic and helix-loop-helix regions form the DNA-binding and dimerization domains respectively. The C-terminal portion of both HIF-1α and HIF-2α mediates oxygen-dependent regulation, largely through effects on protein stability and transcriptional activity, whilst HIF-1β is constitutively expressed (Epstein et al. 2001; Bruick and McKnight 2001). At normal oxygen levels (normoxia), hydroxylation of prolyl residues in the HIF-α subunits enables interaction with an E3-ubiquitin ligase that targets these subunits for rapid degradation (Maxwell et al. 1999; Cockman et al. 2000; Ohh et al. 2000). This prolyl hydroxylation is catalysed by three enzymes called PHD (prolyl hydroxylase domain) 1, 2 and 3 (also known as EGLN 2, 1 and 3 respectively) for which oxygen is a rate-limiting substrate (Epstein et al. 2001; Bruick and McKnight 2001). At low levels of oxygen (hypoxia) hydroxylation is inhibited and the protein stabilised, thereby increasing its abundance. A further asparaginyl hydroxylation, catalysed by FIH (factor inhibiting HIF), regulates interaction with p300/CBP and gene transactivation (Mahon et al. 2001; Lando et al. 2000; Hewitson et al. 2002).

Both HIF-1 and HIF-2 bind to the same core consensus DNA motif (5'-RCGTG-3', where R = A or G) called a hypoxia-response element (HRE) (Wenger et al. 2005). Pangenomic analyses of HIF-1 and HIF-2 DNA-binding by ChIP-seq, and

of gene regulation by RNA-seq or microarray assay have identified thousands of genes that are both up- and down-regulated by HIF (Park et al. 2003; Mole et al. 2009; Wykoff et al. 2000; Schoedel et al. 2011). However, those bound directly by HIF are largely upregulated, indicating that HIF itself acts as a transcriptional enhancer (by releasing promoter paused RNA polymerase 2) and that repressed genes are mainly regulated indirectly through intermediary factors (Schoedel et al. 2011). In addition, HIF may bind up to several hundred kilobases away from the promoter of the gene being regulated, indicating that it can exert its transcriptional activity over long genomic distances. Furthermore, whilst HIF-1 and HIF-2 recognise the same consensus DNA-binding motif they have distinct and only partially overlapping transcriptional outputs that are only partly explained by differences in DNA binding (Lau et al. 2007). Finally, comparison of binding patterns and gene regulation between cell lines, reveals that these vary greatly according to cell type with a relatively small set of highly conserved target genes (Smythies et al. 2019).

There are now multiple observational studies in which increased levels of HIF-1α and/or HIF-2α in diagnostic cancer biopsies have been associated with adverse outcomes across many different tumour types (Semenza 2010). However, it may simply be that aggressive tumours, more readily outgrow their blood supply and become more hypoxic. Evidence that increased levels of HIF may be causally related to adverse outcome comes from an increasing number of interventional studies in largely xenograft studies in animal models. These studies have generally shown that suppression or total inactivation of HIF-1α, 2α or the common HIF-1β subunit in various human cancer cell lines inhibits their growth

when they are implanted into immunodeficient mice (Semenza 2010). However, a few have shown the opposite effect when the expression of HIF-1α or HIF-2α is increased. In addition, tumour hypoxia is also important in rendering tumours resistant to specific anti-cancer treatments, most notably radiotherapy (Dewhirst et al. 2007).

HIF is commonly activated in cancer in response to intra-tumour hypoxia. Cells closest to a perfused blood vessel are exposed to relatively high O_2 concentrations, which decline as distance from the vessel increases. As tumour cells proliferate, the oxygen demand increases and the distance over which the oxygen has to diffuse from the blood vessels to reach them also increases. Unless the formation of new blood vessels is able to keep pace with this growth, the tumour becomes increasingly hypoxic (Talks et al. 2000). Indeed, measured oxygen levels in many cancers are much lower than those encountered in normal tissues (Hockel and Vaupel 2001).

In addition to hypoxia, many tumour suppressor and oncogenic pathways have been linked to increased levels of HIF (Hockel and Vaupel 2001). These may operate at multiple levels. For example, defects in TSC2 affect the mTOR pathway leading to alteration in the translation of HIF-1α and HIF-2α mRNA to protein (Brugarolas et al. 2003). Inactivating mutation in TCA cycle enzymes can lead to the accumulation of metabolites that can inhibit the HIF-prolyl hydroxylases (Selak et al. 2005). However, inactivation of the VHL tumour suppressor, which is disrupted in the majority of clear cell kidney cancers,

mediates the most direct and profound activation of the HIF transcription factors in any type of cancer.

1.4 Kidney cancer and HIF

Kidney cancer is the 7th commonest type of cancer in the UK accounting for 4% of cancer diagnoses (roughly 12,900 new cases per year) and the incidence is currently increasing. Tumours most commonly arise from renal tubular cells (renal cell carcinoma- RCC) but may also develop from cells of the urothelial lining (transitional cell carcinoma - TCC), interstitial cells (sarcoma) or lymphoid tissue (lymphoma). Rarely, children may also develop an embryonal tumour or nephroblastoma, called Wilm's tumour. Renal cell carcinoma is further sub-classified according to histological appearance. By far the commonest form in adults is clear cell renal cell cancer (ccRCC), accounting for more than ¾ of all RCCs (Kaelin 2007). ccRCC is characterised by cytoplasmic deposition of glycogen and lipid, which leaches out of cells when they are fixed, leaving behind clear, unstained areas (Pescador et al. 2010).

Each histological sub-type of RCC is associated with a rare familial syndrome, which although accounting for only a small proportion of cases has in each case provided fundamental insight into the pathogenesis of sporadic RCCs. The familial syndrome associated with ccRCC is the von Hippel-Lindau (VHL) Syndrome (Linehan et al. 2009).

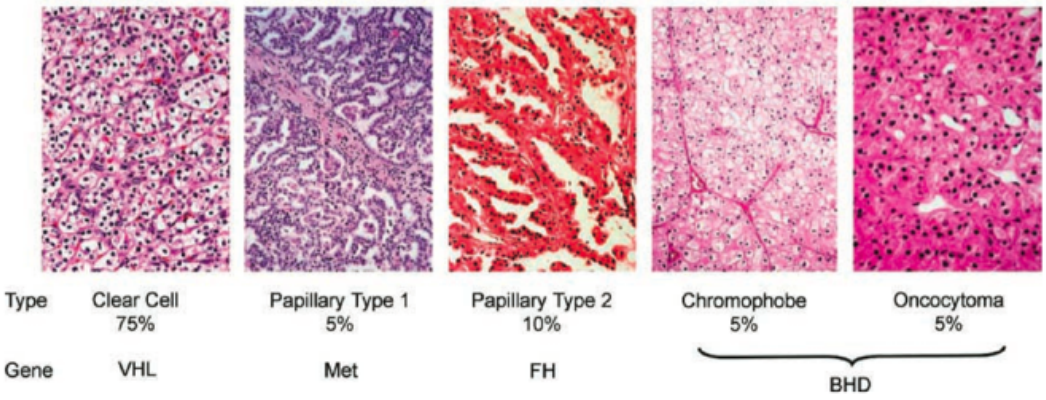


Figure 1.6. Histological sub-types of human epithelial renal cell cancers. Mutation of the von Hippel-Lindau gene (VHL) is linked to clear cell kidney cancer, whilst the Met gene is linked to the papillary type I form, the gene encoding fumarate hydratase (FH) is linked to papillary type I form and inactivation of the Birt-Hogg-Dube tumoursupressor (BHD) is linked to chromophobe and oncocytic, hybrid RCC. Adapted from et al. (Linehan et al. 2009).

VHL syndrome is an autosomal dominant familial cancer syndrome that affects 1 in 36,000 individuals (Lonser et al. 2003). In addition to ccRCC, affected individuals may also develop blood vessel tumours called haemangioblastomas that affect the CNS (predominantly cerebellum and spinal cord) and retina, phaeochromocytomas, pancreatic cysts, endolymphatic sac tumours or epididymal cysts. Classical linkage studies identified the gene responsible (the VHL gene) on the short arm of chromosome 3 (Latif et al. 1993). Affected individuals are heterozygous for hypomorphic mutations in the VHL gene. Tumour formation is associated with subsequent loss of chromosome 3p and over their lifetime, such individuals often develop multiple tumours. The

overwhelming majority of sporadic (non-familial) ccRCCs are also associated with biallelic loss of VHL (or more rarely with other members of the VHL E3-ubiquitin ligase complex). However, since (in accordance with the Knudson two-hit hypothesis) individuals need to experience two somatic events in the same cell, the incidence of tumours is much less than in individuals who already have a germline mutation on one allele.

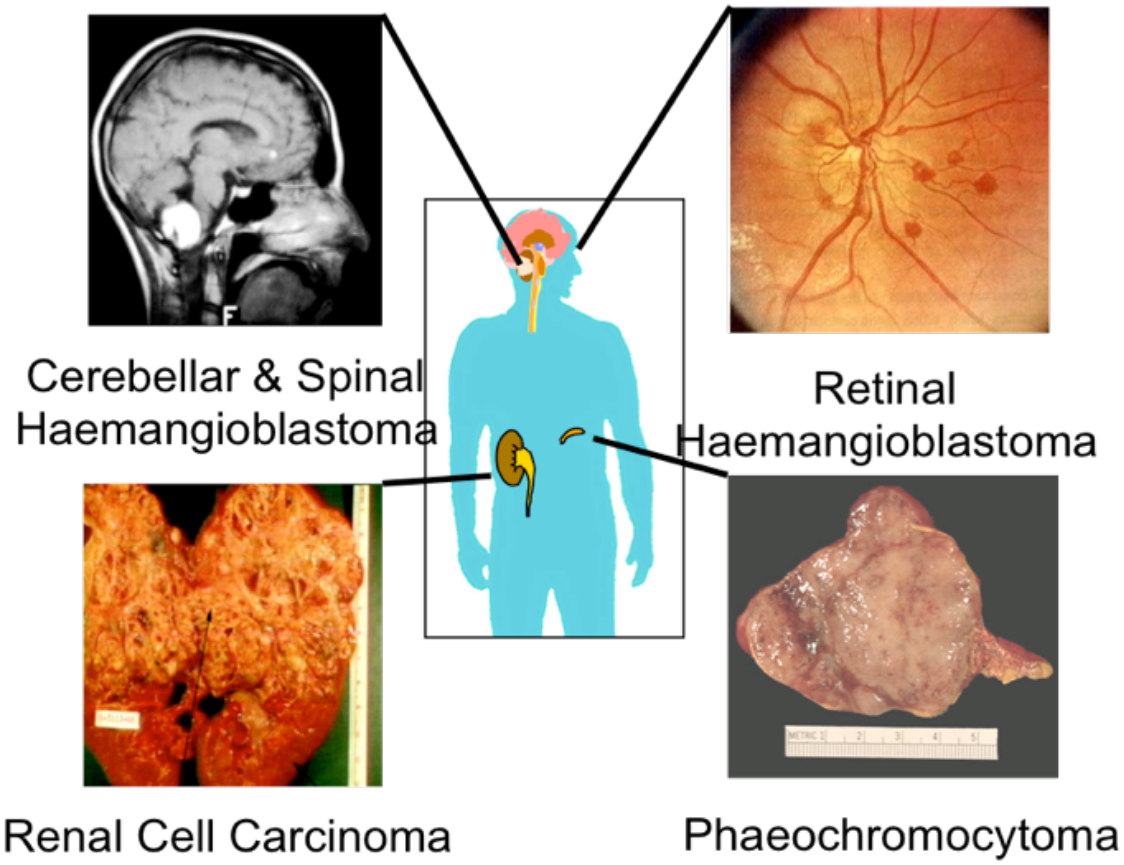


Figure 1.7. The von Hippel Lindau (VHL) Syndrome is characterised by multiple cerebellar, spinal and retinal haemagioblsatomas, ccRCCs and phaeochromocytomas in addition to other rarer tumours.

The VHL gene encodes for a ubiquitin E3-ligase (VHL) that specifically recognises HIF-1α and HIF-2α molecules, which have been hydroxylated by the HIF prolyl hydroxylases in the presence of oxygen and targets them for rapid destruction by the proteasome (Maxwell et al. 1999). Therefore, biallelic inactivation of VHL, in both familial and sporadic ccRCC, leads to constitutive activation of the HIF transcriptional pathway. Furthermore, in VHL-defective ccRCC cells, inactivation of HIF-2α was sufficient to slow tumour growth, whilst expression of a HIF variant that escapes VHL-mediated degradation was able to override the tumour suppressor effect of VHL re-introduction (Kondo et al. 2002; Kondo et al. 2003). Taken together, this strongly implicates the HIF transcriptional response in the pathogenesis of ccRCC.

1.5 The HIF-pathway – a double-edged sword in ccRCC.

Inactivation of VHL and the consequent HIF induction does not automatically progress to cancer. Firstly, the HIF pathway (and VHL) operates almost universally in nucleated cells within the human body. By definition, in VHL syndrome patients, germline inactivation of one VHL allele is also universal. However, it is inconceivable that stochastic somatic inactivating 3p loss in the wild-type allele will be limited to those cell-types involved in VHL-associated tumours. This strongly suggests that VHL inactivation can only lead to tumour formation when it occurs in specific cellular settings (Smythies et al. 2019; Ortiz-Barahona et al. 2010). Furthermore, comparison of HIF-binding sites and HIF-responsive genes in different cell types indicates that the downstream consequences of HIF activation are also highly dependent on the cellular setting.

This may at least in part, help restrict the cell types that are able to form tumours in response to VHL inactivation.

Secondly, acute inactivation of VHL (at least in VHLfl/fl mouse embryonic fibroblasts infected with a Cre-encoding retrovirus) leads to a senescent-like phenotype rather than the unrestricted cell proliferation that is typical of cancer (Young et al. 2008; Mack et al. 2005). It is unclear from these studies whether this senescent-like phenotype is mediated by HIF (Giaccia et al. 2004). However, they strongly suggest that VHL inactivation comes with a potential penalty that needs to be overcome in order for tumours to form.

This is further emphasised in immunostaining studies of human kidneys from patients with VHL Syndrome for VHL, HIF and HIF-target genes (Mandriota et al. 2002; Schietke et al. 2012). Examination of “normal” regions of the kidneys from these patients reveals the presence of numerous foci of VHL inactivation and HIF activation within the tubules. This includes single cells that have not yet started to proliferate as well as small nests of cells that have multiplied but not yet begun to invade the surrounding tissue. The number of such lesions vastly outnumbers the number of tumours that these patients develop. This suggests that only a small proportion of cells in which VHL is inactivated progress to form overt tumours.

The above findings strongly suggest that VHL inactivation and HIF activation are necessary, but not sufficient to induce tumour formation and that additional

events are required in order to form cancers. This is borne out in high-throughput sequencing studies in human ccRCCs, in which the combination of whole-genome sequencing analysis and multi-region sampling provides insights into the nature and timing of key oncogenic events (Gerlinger et al. 2014; Turajlic, Xu, Litchfield, Rowan, Horswell, et al. 2018; Turajlic, Xu, Litchfield, Rowan, Chambers, et al. 2018). These studies show inactivation of VHL in a precise way by driver mutations and 3p-5q copy number variation as a truncal event, occurring in all clones and predating subsequent mutational events often by several decades during the evolution of the ccRCC tumours which subsequently lead to sub-clonal expansion and development of overt cancers (Mitchell et al. 2018).

Other lines of evidence also illustrate the penalty associated with loss of VHL and activation of HIF pathways in renal cancer. In xenograft models, overexpression of HIF-1α paradoxically slows tumour progression, while increasing HIF-2α has the opposite effect and accelerates tumour growth (Raval et al. 2005). Interestingly, normal proximal renal tubule cells, from which ccRCCs are believed to develop, only express the HIF-1α isoform and lack HIF-2α mRNA (Rosenberger et al. 2002). Conversely, renal cancer cells almost all express HIF-2α and many have downregulated or inactivated HIF-1α expression (Schietke et al. 2012; Shen et al. 2011; Mandriota et al. 2002).

Within renal cancer cell lines, HIF-1 and HIF-2 target genes have now been extensively mapped using ChIP-seq analysis and transcriptional profiling (Mole et al. 2009; Smythies et al. 2019; Schodel et al. 2013; Xia and Kung 2009;

Tausendschon et al. 2015). Using publicly available RNA-sequencing datasets from the TCGA, it has been possible to examine each of these genes for association between its expression level and patient prognosis (Salama et al. 2015). Despite an overall association between HIF-1 target genes and a favourable prognosis and between HIF-2 target genes and a poor patient prognosis, the individual gene associations were heterogeneous in sign and magnitude, suggesting that activation of each HIF-α isoform contributes a highly complex mix of pro- and anti-tumorigenic effects (Salama et al. 2015).

Taken together, these findings indicate a non-linear relationship between VHL loss, HIF activation and tumour formation. The resulting effect is a balance of both the pro- and anti-tumourigenic activities of HIF-pathway activation with each individual HIF-target gene conferring only a small effect on the outcome. Indeed, it is likely that many HIF-target genes are simply co-activated as part of the overall pathway switch and are neutral “bystander” events that play no direct role in the pathogenesis of ccRCC. Overall my hypothesis is, that initial loss of VHL confers a net penalty on cells that has to be overcome by subsequent adaptation. The ability of somatic mutational events to overcome this penalty depends on the cellular setting in which the VHL loss occurs, which in turn helps to shape the HIF transcriptional response. This is manifest by the highly restricted tissue types in which VHL-associated tumours develop. However, germline variants might also alter specific aspects of the HIF response in kidney cancer cells and similarly reduce the barrier that cells need to overcome in order to form tumours. This would manifest as an association between the germline variant and the incidence

of kidney cancer. Such associations can be studied using genome-wide association studies (GWAS).

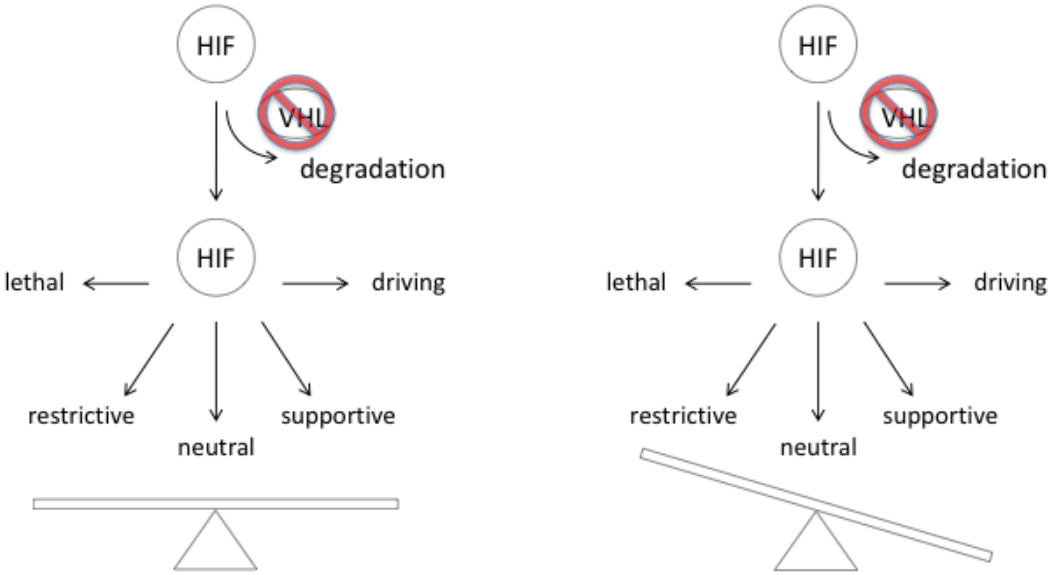


Figure 1.8. The net effect of VHL loss and HIF pathway activation is a balance between multiple differing effects. In order for tumours to form and progress, this balance has to be re-modelled.

1.6 Genome-wide association studies

Genome-wide association studies (GWAS) are population-based genetic studies that test for the association of common germline variants (observed in a minimum of 1% of the population) with development of a common trait. The studies identify single nucleotide polymorphisms (SNPs) for which the allelic balance is significantly discordant between a group of individuals affected by a certain trait or disease and an unaffected control group. In instances where a significant

association can be demonstrated- they are commonly referred to as predisposing genetic loci.

In 2006, the international Hapmap Consortia was established, cataloguing common human sequence variations, and soon after the first GWA studies were performed by the The Wellcome Trust Case Control Consortium identifying risk loci predisposing to bipolar disorder, coronary artery disease, Crohn's disease, rheumatoid arthritis and type 1 and 2 diabetes (Wellcome Trust Case Control 2007) demonstrating the general feasibility of the approach. Since, several thousands of GWA studies have been performed, to identify genetic polymorphisms associated with a range of traits including body-mass-index or height and more recently, GWAS has also found application in as diverse fields as anthropology or behavioural psychology.

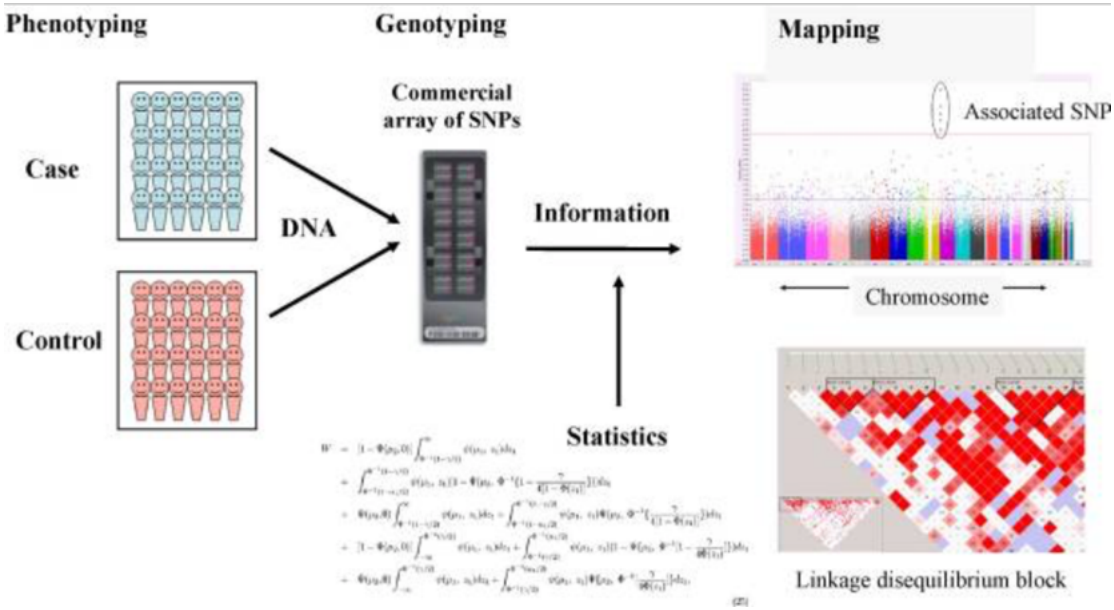


Figure 1.9: Detection of GWAS-defined susceptibility loci. The workflow of GWA studies are separated into 3 steps: Patients and control group are recruited, and a predefined set of SNPs is genotyped for each of the individuals. Finally, the statistical association of each of the genotyped SNPs with the studied trait is defined. This figure is adapted from Ikegawa et al. (Ikegawa 2012).

Whilst these first studies only incorporated information of about 2000 individual cases per disease, modern GWA studies integrate SNP information from hundred thousands of individuals. Large sample sizes are necessary for statistical reasons as, SNPs are present in large numbers across the human genome, which in turn results in a high number of comparisons and multiple hypotheses testing. They are typically found every 120 bp and organised in haplotype blocks, which are entities of coinheritance. In case-control GWAS significance of the allelic difference is determined for every SNP by a logistic regression model. It is estimated that the human genome is organised into 1 million independent haplotype blocks (The_1000_Genomes_Project_Consortium 2015). And hence to correct for a large number of comparisons, Bonferroni correction is applied. The critical P-value is adjusted by the number of comparisons made. For a critical P-value of 0.05, with an estimated number of comparisons of 1 million made across the genome, this results in a statistical significance cut-off at P-value = $5*10^{-8}$. To reach these levels of statistical association, several ten- to hundred-thousand individuals are profiled. This is why GWAS is predominantly applied to the study of common traits and diseases.

A recent landmark study in kidney cancer has performed a meta-analysis of kidney cancer GWAS integrating data from 10,784 kidney cancer patients and 20,406 controls (Scelo et al. 2017). The original genotyping of the individual studies contributing to the meta-analysis had been performed on several different genotyping platforms. The data was imputed onto the reference panel of the 1000 Genome Project for integrated analysis. The statistical significance of association with kidney cancer predisposition of the SNPs was determined by logistic regression by accounting for the number of comparisons made across the genome. Several putative independent loci, alongside newly discovered loci were identified (Gudmundsson et al. 2013; Purdue et al. 2011; Wu et al. 2011; Henrion et al. 2013; Henrion et al. 2015). Independent, replication studies were performed on a smaller case/control dataset to verify the newly discovered loci.

An alternative means to replication studies to verify the validity of susceptibility loci is to determine their biological function. Interestingly, GWAS-identified susceptibility loci not only affect annotated genetic regions but frequently are found in the intergenic space. The intergenic space has important functions in regulatory control of gene expression. Susceptibility loci have been found to be enriched at regulatory genetic regions such as enhancers and promoters (Encode_Project_Consortium et al. 2012). This creates great challenges for investigating the function of these intergenic loci, as it is often unclear what the targets or these intergenic loci are. However, ’next-generation’ sequencing

approaches provide a powerful approach to investigate the role of non-coding variants and their implications in gene regulation.

1.7 Interpreting GWAS in the light of the epigenetic and transcriptional landscape

Non-coding variants often affect regulatory regions of the genome such as promoters and enhancers (Wang et al. 2016; Encode_Project_Consortium et al. 2012; Ernst et al. 2011). These highly tissue-specific regulatory elements can be determined by pan-genomic chromatin immunoprecipitation (ChIP-seq). The method provides powerful means to determine the localization of proteins at chromatin genome-wide, for example, binding sites of transcription factors and profiling of histone modifications marking enhancers, promoters and their activity (Orlando and Paro 1993; Robertson et al. 2007).

The method constitutes three main steps, preparation of chromatin complexes, immunoprecipitation and next-generation sequencing of the recovered DNA fragments. In the chromatin preparation step interactions between all protein in contact with to the DNA are stabilised by formaldehyde crosslinking. Subsequently, the chromatin is fragmented into smaller pieces allowing for antibody-based enrichment for the protein of interest. Enrichment for the protein of interest results in enrichment for protein-associated DNA sequences which then are dissociated by reversing the formaldehyde link. These DNA fragments are purified and subjected to deep sequencing. Mapping of the DNA identification

of enrichment of DNA reads above background levels finally determined regions of around 200-2000 bp length within which one or multiple the transcription factors bind DNA (Zhang et al. 2008). The method is hence a powerful approach to determine transcription factor localisation pan-genomically.

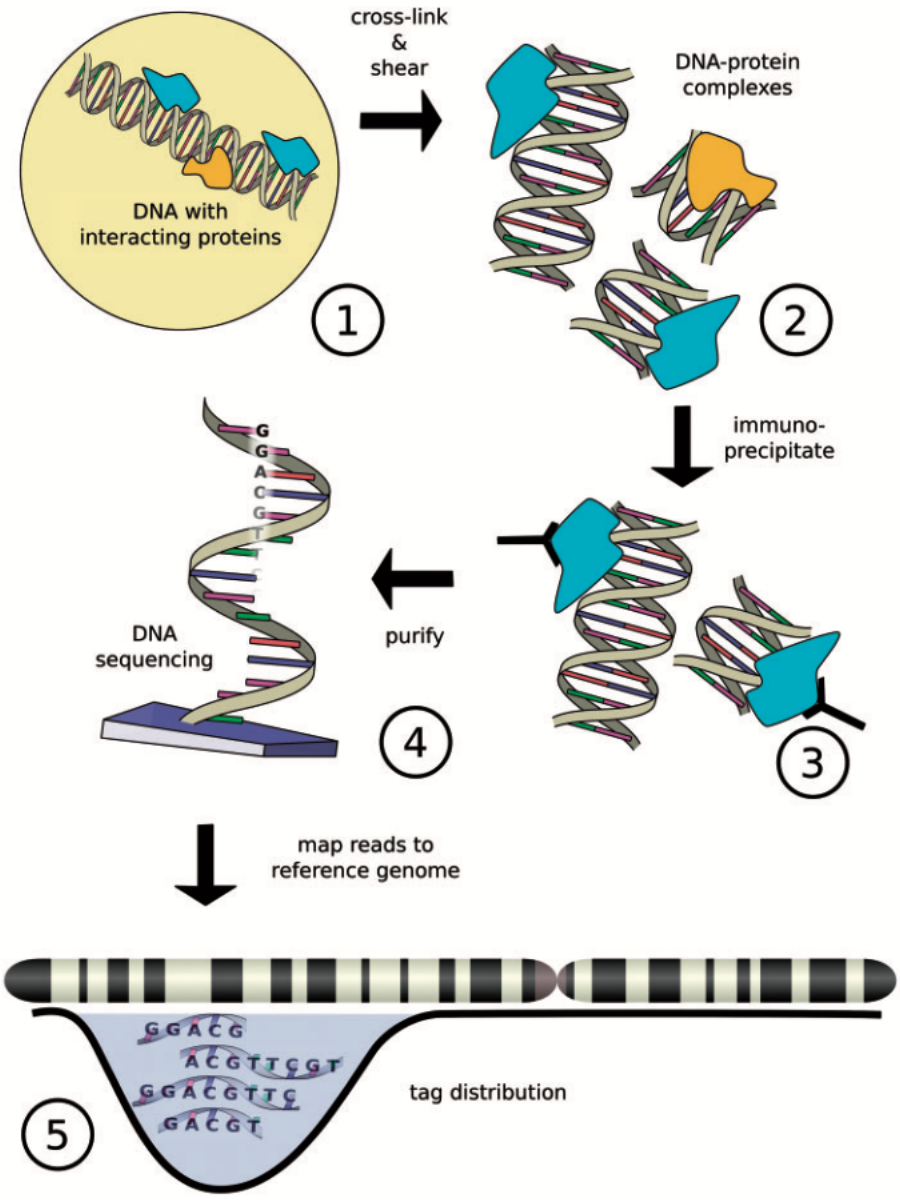


Figure 1.10: Chromatin Immunoprecipitation determines pan-genomic transcription factor binding sites. Chromatin present in the nucleus of cells is linked to the associated protein. A target protein of interest is enriched by immunoprecipitation resulting in an enrichment of the associated DNA which is then subject to next-generation sequencing. Finally, enrichment of these short reads at specific locations of the genome determines the location of transcription factors or histone marks. This figure is adapted from (Szalkowski and Schmid 2011)

Biological effects associated with susceptibility loci may be propagated by impacting on transcription factor binding sites (Schödel et al. 2012). Transcription factor binding sites are present at promoters of genes, where the susceptibility locus may be affecting the gene associated with the promoter. Often enhancers may also be found in intergenic regions, where they may regulate expression of non-coding RNAs or regulate gene expression via long-range chromatin looping (Hughes et al. 2014). Enhancers do not necessarily regulate the nearest gene, and this can make it difficult to identify the gene that is affected by any given polymorphism that may be associated with kidney cancer or another disease or trait.

One way to assess the 3-dimensional structure of chromatin and identify long-range chromatin interactions is by Capture-C assay (Davies et al. 2016). The

method entails fixation of proximal chromatin elements by formaldehyde mediated cross-linking followed by digestion and re-ligation of DNA resulting in DNA fragments representing the 3-dimensional proximity instead of linear proximity. Finally, capture of regions of interest by oligonucleotide pull-down of the DNA fragments results in enrichment of DNA fragments that are in close proximity to the viewpoint and long-range chromatin interactions are determined within a 600 kb window of an enhancer or promoter. The technique therefore greatly facilitates identification of target genes of non-coding susceptibility loci as the basis that functional effects are reflected by these physical interactions (Hughes et al. 2014).

Finally, RNA-seq facilitates assessment of expression levels of genes by assessing the cellular transcriptome. Importantly, the method uses reverse transcription to determine expression levels and often is applied to assess differential expression of genes under two different conditions (Mortazavi et al. 2008). Despite the methods available to describe the epigenome and transcriptome, interpreting the biological effect of susceptibility loci in the light of the epigenetic and transcriptional landscape bares several challenges.

Importantly, there are limitations linked to these descriptive methods. All three methods are best defined in immortalised cell line models which are linked to chromosome abnormalities and only reflect one specific stage and location of one tumour and are therefore an incomplete and not exact representation of a tumour. Further, it is impossible to exactly define the location of the actual binding

site of the transcription factor by ChIP-seq or interacting regions by Capture-C. Whilst most transcription factors only bind several base pairs, regions defined by ChIP-seq range between hundreds up to a few kilobase pairs in length. Similarly interacting regions defined by Capture-C vary greatly in length and may represent an over- or underrepresentation of the interacting regions. Both methods, ChIP-seq and Capture-C, alongside RNA-seq determined differential gene expression analysis, further rely on statistical methods to determine binding sites, interacting regions or differentially expressed genes. The sensitivity and specificity of the models underlying the detecting algorithms may further contribute to the imperfections associated with the description of the epigenetic and transcriptomic landscape. Further, cis-acting sequences display a high level of cell-type specificity (Visel et al. 2009). It is unknown to what extent these the activity and state of cis-acting sequences differ between stages of tumour development. Some of these limitations can be overcome, as will be illustrated within this work, by investigating large scale patient sequencing dataset as provided by The Cancer Genome Atlas (TCGA) or similar efforts.

1.8 Aim of this study

To test the hypothesis that aspects of the HIF pathway are under selective pressure in kidney cancer, within chapter 2 and 3 I explore the relationship between germline variants associated with RCC in GWAS analyses and cis-acting elements of the HIF pathway mapped in cancer cell lines corresponding to RCC. I examine for direct overlap of RCC susceptibility loci, as defined by

GWAS, with cis-acting elements of the HIF-pathway as opposed HIF target genes themselves.

Chapter 4 focuses on different approaches for target gene identification of germline variants associated with RCC in GWAS analyses and I test for their association with HIF target genes.

Apart from direct overlap of RCC susceptibility loci with HIF binding sites, alternative mechanisms of how RCC susceptibility loci could affect the HIF pathway are explored in the context of chapter 5.

Chapter 6 proposes a way of how effects of susceptibility loci of intermediate significance can be identified and studied in the context of a specific pathway activation as exemplified by activation of the HIF pathway in RCC.

Finally, in chapter 7 I test generalisability of the approach to investigate the effect of non-coding susceptibility loci in the light of transcriptional pathway activation proposed in previous chapters in the context of common cancer types.

Chapter 2: RCC susceptibility loci target HIF binding in RCC cells

2.1 Introduction

Activation of the HIF-transcriptional programme is a large-scale physiological response, involving many hundreds of genes, that has evolved to orchestrate the cellular response to hypoxia. As such, it has not evolved simply to promote cancer and many of the genes activated by HIF will, therefore, have neutral or even detrimental effects on cancer cells. Furthermore, oncogenic activation of the HIF pathway by loss of the VHL tumour suppressor gene in clear cell RCC will lead to the inappropriate induction of hypoxic-response genes in normoxic regions of the tumour, which may lead to further deleterious effects. Indeed, in the setting of kidney cancer, HIF-1 and HIF-2 have opposing effects on tumour progression and the specific genes activated by each transcriptional programme have heterogeneous associations with renal cancer prognosis (Raval et al. 2005; Salama et al. 2015). This suggests that there is a penalty associated with VHL loss that needs to be overcome by the tumour cell.

Multiple lines of evidence point to this possibility. Firstly, the acute inactivation of VHL in explanted mouse embryonic fibroblasts using cre-recombinase leads to cellular senescence rather than proliferation (Young et al. 2008). Secondly, renal cancer cells seem to switch from a predominantly anti-tumourigenic HIF-1 driven hypoxic response in normal proximal renal tubule cells to a more pro-tumourigenic HIF-2 driven response in mature renal cancer cells (Raval et al. 2005; Shen et al. 2011).

Taken together, this suggests that tumour cells need to adapt/evolve the HIF response to overcome this penalty. This may occur as a result of somatic events

occurring within the tumour cells after VHL is inactivated and the HIF response is turned on. Alternatively, germline variants that alter specific aspects of the HIF response in kidney cancer cells may reduce this penalty, thereby reducing the barrier that the cells need to overcome in order to form tumours. Since fewer somatic events are now required to overcome the barrier, it is more likely that sufficient will occur, thereby increasing the chance that tumours will develop. Ultimately, this would manifest as an association between the germline variant and the incidence of kidney cancer, as identified in GWAS analyses.

Therefore, to test the hypothesis that aspects of the HIF pathway are under selective pressure in kidney cancer, I decided to examine the relationship between germline variants associated with RCC in GWAS analyses and cis-acting elements of the HIF pathway. Although these analyses generally do not distinguish the histological sub-type, clear cell RCC will constitute the vast majority of cases. I initially examined for overlap with cis-acting elements of the HIF-pathway for several reasons. Firstly, GWAS loci are frequently intergenic, a long way from any transcribed region of the genome suggesting that they frequently affect enhancers (Schödel et al. 2012). Secondly, these enhancers do not necessarily regulate the nearest gene, and this can make it difficult to identify the gene that is affected by any given polymorphism that may be associated with kidney cancer. Finally, since HIF-binding sites vary from cell-type to cell-type, I made an assumption that at least some of the germline polymorphisms that are associated with renal cancer have a biological effect within the VHL-defective clear cell RCC cells themselves and chose to examine ChIP-seq datasets from these cell types (Smythies et al. 2019). However, this assumption is itself also

being tested in this chapter, since if it were not the case, I would not expect to see an association between loci associated with RCC in GWAS analyses and HIF-binding sites defined in kidney cancer cells.

2.2 Results

2.2.1 Definition of RCC susceptibility loci from independent GWA studies

I first collated a comprehensive list of all common germline variants that had been associated with renal cancer (RCC) in GWAS analyses to date. Studies were identified using a database search of the National Human Genome Research Institute – European Bioinformatics Institute (NHGRI-EBI) GWAS catalogue, combined with a literature search using PubMed (Buniello et al. 2019). At the onset of this work, 5 independent RCC GWA studies conducted in populations of Caucasian origin had been published and had reported a total of 9 distinct RCC predisposing genomic loci (Table 2.1). The studies have been performed between 2011 and 2015 and rely on individuals of European ancestry to account for population stratification. However, the studies differ in cohort size and on the number of SNPs tested in the original scan. Despite these differences, there was a wide-spread concordance between the studies in the loci discovered.

Locus	*Reported SNP*	*Reference*	*Observed loss/gain of Heterozygosity*
2p21	rs7579899	(Purdue et al. 2011)	no
12q24.31	rs4765623	(Purdue et al. 2011)	no
11q13.3	rs7105934	(Purdue et al. 2011)	no
12p12.1	rs718314	(Wu et al. 2011)	no
6q22.31	rs25422	(Wu et al. 2011)	no
8q24.21	rs6470588	(Gudmundsson et al. 2013)	gain
2q22.3	rs12105918	(Henrion et al. 2013)	no
5p13.3	rs10054504	(Henrion et al. 2013)	no
1q24.1	rs3845536	(Henrion et al. 2015)	no

Table 2.1: Summary of 9 renal cancer-predisposing loci identified by independent genome-wide association studies (GWAS) conducted in populations of Caucasian origin by June 2016. The table further indicates if somatic alterations of heterozygosity was observed at the locus (Turajlic, Xu, Litchfield, Rowan, Chambers, et al. 2018). For each of the loci, the study discovering the locus first is indicated. This table adapted from Grampp, Schmid et al. (Grampp et al. 2017).

In most cases, databases and publications had reported the single most significant single nucleotide polymorphism (index SNP) at each susceptibility

locus, to represent the haplotype block. However, the statistical association between RCC susceptibility and the reported SNP may not be due to an effect of that specific SNP, but rather could be due to an effect of any variant that is commonly co-inherited with the index SNP. In other words, the statistical association of the reported SNP may be due to coinheritance rather than a direct effect on RCC. Therefore, to include all SNPs, which have the potential to be causal, I next identified those that are frequently co-inherited with the index SNP as defined by strong linkage disequilibrium with the RCC susceptibility loci. Specifically, I included SNPs in minimum linkage disequilibrium (LD) of r^2=0.8 with the reported SNPs by reference to the 1000 Genomes EUR population (The_1000_Genomes_Project_Consortium 2015). The minimum LD threshold of 0.8 is commonly used and denotes a very high probability of the SNPs being co-inherited with the index SNP (Wang et al. 2016).

2.2.2 Definition of high confidence HIF binding sites and enhancers in RCC cancer cell lines

I next attempted to define the cis-acting elements of the HIF-transcriptional pathway in RCC cells by reference to ChIP-seq analyses of HIF-binding sites in two commonly used VHL-defective clear cell RCC cell lines, 786-O and RCC4. Given the widely known heterogeneity both within and between different kidney cancers, it is unlikely that a sample size of two will represent the entirety of the HIF response in kidney cancer (Gerlinger et al. 2014; Gerlinger et al. 2012). Furthermore, it will only represent the cis-acting elements of the HIF response in

late-stage kidney cancers since to date, it has not proven possible to isolate and study early-stage lesions. In addition, tissue culture conditions differ from those in the tumour itself and may affect binding. Nevertheless, ChIP-seq analyses of RCC cell lines does provide a first approximation of the HIF-transcriptional output in kidney cancer and it is likely that a proportion of these binding sites will be present across multiple tumours and at different stages in tumourigenesis.

Datasets of chromatin immunoprecipitation for HIF-1α, HIF-2α and HIF-1β followed by deep-sequencing have been generated by previous members of the group. A total of 7 ChIP-seq datasets in the two RCC cell lines 786-O and RCC4 were integrated to define the high confidence HIF binding sites. Peaks were identified using the MACS and TPIC peak callers and only peaks identified by both algorithms were used (Zhang et al. 2008; Hower et al. 2011). Input chromatin was used as the negative control when identifying peaks and peaks that coincided with ”blacklisted” regions provided by ENCODE were excluded. For the RCC4 cell line, HIF-1α and HIF-2α chromatin immunoprecipitation had been performed in duplicate and were combined so that only peaks identified in both replicates were taken forward. Only one dataset was available for HIF-1β in RCC4 cells and HIF-2α and HIF-1β in 786-O cells (note 786-O cells are deficient for HIF-1α and so only HIF-2α and HIF-1β chromatin immunoprecipitations were performed).

As a first check on the quality of the ChIP-seq data, I examined the signal-to-noise ratio in the vicinity of the MACS and TPIC identified peaks. First, the summit

(chromosomal location corresponding to maximum ChIP-seq signal) was identified for each peak. Each was then arranged one above the other on the vertical scale such that their summits aligned horizontally. The ChIP-seq signal was then plotted on a colour scale (higher signal = deeper red) for each chromosomal position (± 2kb) flanking the summit on the horizontal scale to give the corresponding heatmaps (Figure 2.1). The ChIP-seq signal at each position flanking the summit was then averaged across all peaks to give the corresponding line graphs. In all cases, there was a strong, narrow spike of ChIP-seq signal extending ± 200-300 bp (the average fragment size of the chromatin following sonication) with low background signal.

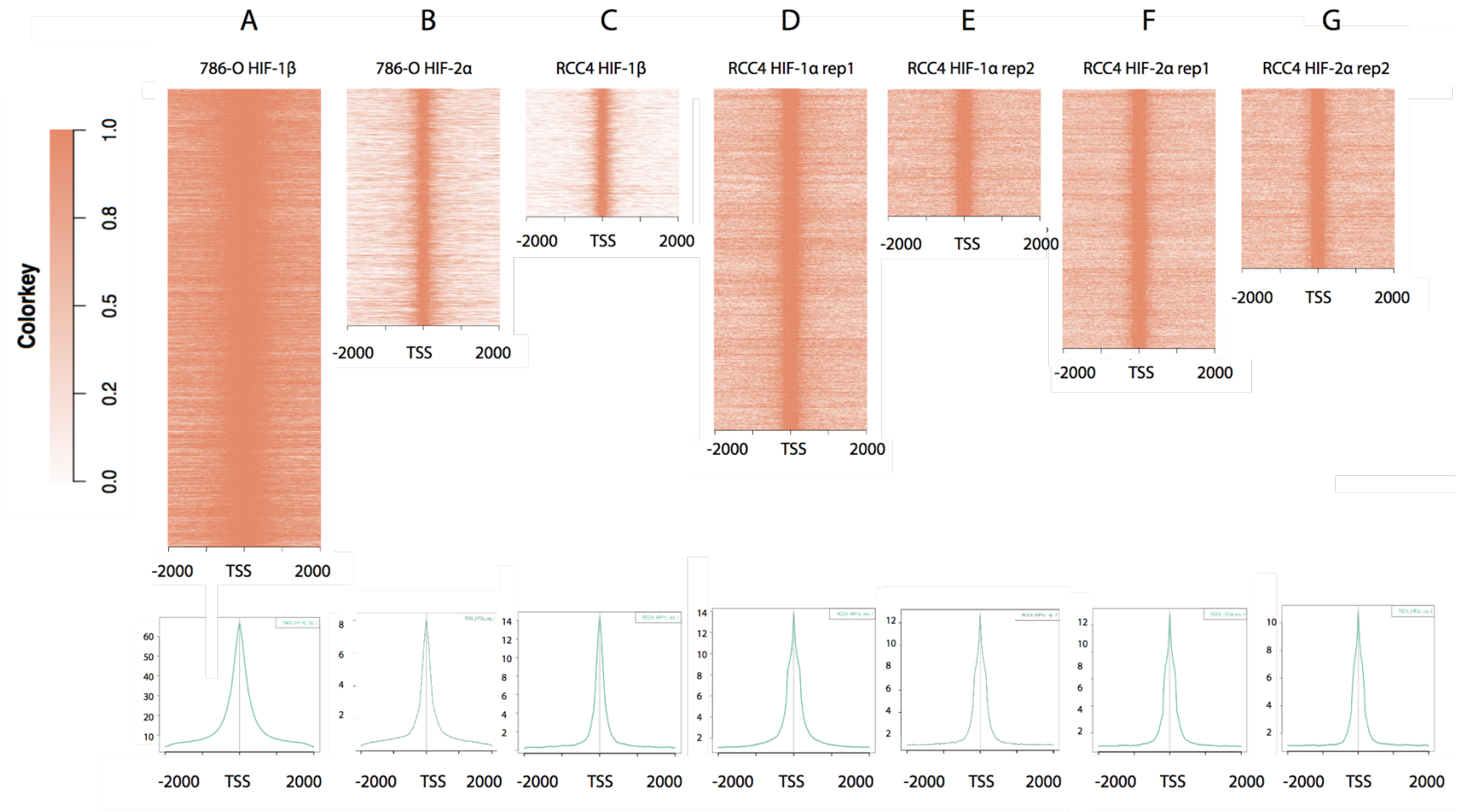

A
B
C
D
E
F
G
786-O HIF-1β
786-O HIF-2α
RCC4 HIF-1β
RCC4 HIF-1α rep1
RCC4 HIF-1α rep2
RCC4 HIF-2α rep1
RCC4 HIF-2α rep2
Colorkey
0.0
0.2
0.5
0.8
1.0
-2000
TSS
2000

Figure 2.1: Heatmaps and average peaks profile of HIF-α and HIF-1β binding sites mapped by pan-genomic chromatin immunoprecipitation (ChIP-seq) in RCC cell lines of RCC4 and 786-O. In the heatmaps, each row corresponds to one peak called in the corresponding ChIP-seq dataset. ChIP-seq signal correlates with colour intensity and is measured in read counts per million mapped reads (CPM) in the region of +/- 2 kb from the summit of the peaks called for each of the 7 datasets. The line diagrams below each heatmap show the average signal intensity (CPM) of all peaks against the distance from the summit of the peaks for a +/- 2 kb window centred at the peak summits. 786-O HIF-1β (A), 786-O HIF-2α (B), RCC4 HIF-1β (C), RCC4 HIF-1α replicate 1 (D), RCC4 HIF-1α replicate 2 (E), RCC4 HIF-2α replicate 1 (F), RCC4 HIF-2α replicate 2 (G).

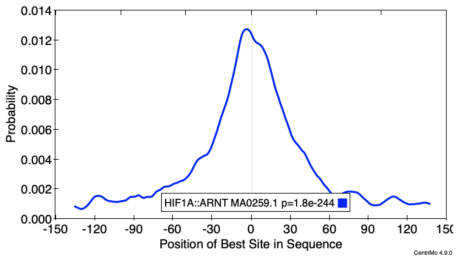
As a second check of the fidelity with which HIF-binding sites were identified, I determined whether the anticipated consensus HRE motif was enriched over a randomized sequence within each of the defined set of peaks. For each set of peaks, the MEME suite of motif-based sequence analysis tools was used to interrogate the JASPAR core database of motif matrices for enrichment (Bailey et al. 2009; Sandelin and Wasserman 2004). In each case, the core consensus HRE motif, MA0259.1 (Figure 2.2), or a closely related variant, was the most highly significantly enriched within the defined peaks. Furthermore, when analysed using the CentriMo tool, this motif was centrally enriched within the

defined peak regions in all 7 datasets (Figure 2.2 B–H), providing further evidence for direct binding of each of the immunoprecipitated HIF isoforms to the core consensus HRE motif (Bailey and Machanick 2012).

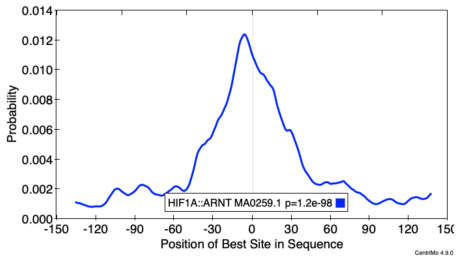
A


Bits
2.0
1.5
1.0
0.5
0.0
1
2
3
4
5
6
7
8

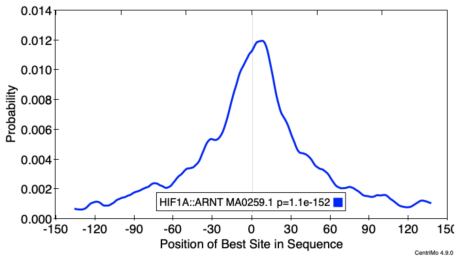
B


Probability
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
-150 -120 -90 -60 -30 0 30 60 90 120 150
Position of Best Site in Sequence
HIF1A::ARNT MA0259.1 p=1.8e-244
CentriMo 4.9.0

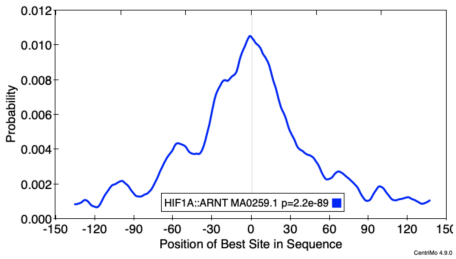
C


Probability
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
-150 -120 -90 -60 -30 0 30 60 90 120 150
Position of Best Site in Sequence
HIF1A::ARNT MA0259.1 p=1.2e-98
CentriMo 4.9.0

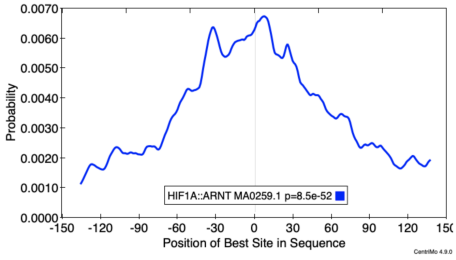
D


Probability
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
-150 -120 -90 -60 -30 0 30 60 90 120 150
Position of Best Site in Sequence
HIF1A::ARNT MA0259.1 p=1.1e-152
CentriMo 4.9.0

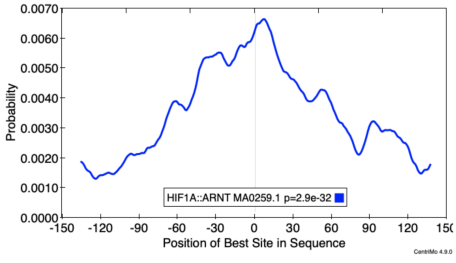
E


Probability
0.012
0.010
0.008
0.006
0.004
0.002
0.000
-150 -120 -90 -60 -30 0 30 60 90 120 150
Position of Best Site in Sequence
HIF1A::ARNT MA0259.1 p=2.2e-89
CentriMo 4.9.0

F


Probability
0.0070
0.0060
0.0050
0.0040
0.0030
0.0020
0.0010
0.0000
-150 -120 -90 -60 -30 0 30 60 90 120 150
Position of Best Site in Sequence
HIF1A::ARNT MA0259.1 p=8.5e-52
CentriMo 4.9.0

G


Probability
0.0070
0.0060
0.0050
0.0040
0.0030
0.0020
0.0010
0.0000
-150 -120 -90 -60 -30 0 30 60 90 120 150
Position of Best Site in Sequence
HIF1A::ARNT MA0259.1 p=2.9e-32
CentriMo 4.9.0

H

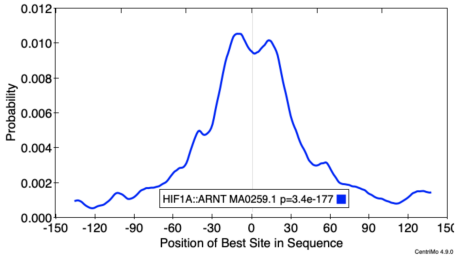

Probability
0.012
0.010
0.008
0.006
0.004
0.002
0.000
-150 -120 -90 -60 -30 0 30 60 90 120 150
Position of Best Site in Sequence
HIF1A::ARNT MA0259.1 p=3.4e-177
CentriMo 4.9.0

Figure 2.2: Motif enrichment analysis of 7 HIF ChIP-seq datasets mapped in cancer cell lines of RCC. (A) MEME suite motif-based sequence analysis of identified peaks demonstrates enrichment of the core HRE motif (MA0259.1) in all 7 ChIP-seq datasets. CentriMo analysis of the frequency distribution of this HRE motif within the defined peaks shows central enrichment in (B) 786-O HIF-2α, (C) 786-O HIF-1β, (D) RCC4 HIF-1β, (E) RCC4 HIF-1α replicate 1, (F) RCC4 HIF-1α replicate 2, (G) RCC4 HIF-2α replicate 1, and (H) RCC4 HIF-2α replicate 2 ChIP-seq datasets.

Finally, I compared the 7 datasets from the two RCC cell lines with each other. For each dataset, the number of called peaks is listed in Table 2.2.

ID	*Cell line*	*Factor*	*Number of peaks*
RCC4_HIF1a_rep1	RCC4	HIF-1α	5680
RCC4_HIF1a_rep2	RCC4	HIF-1α	2120
RCC4_HIF2a_rep1	RCC4	HIF-2α	4317
RCC4_HIF2a_rep2	RCC4	HIF-2α	2990
RCC4_HIF1b	RCC4	HIF-1β	2137
786O_HIF2a	786O	HIF-2α	3939
786O_HIF1b	786O	HIF-1β	7621

Table 2.2: Summary of the number of peaks identified in 7 HIF ChIP-seq datasets mapped in cancer cell lines of RCC. Peaks considered were identified by both peak callers MACS and TPIC. The number of peaks identified for the 7 datasets is shown.

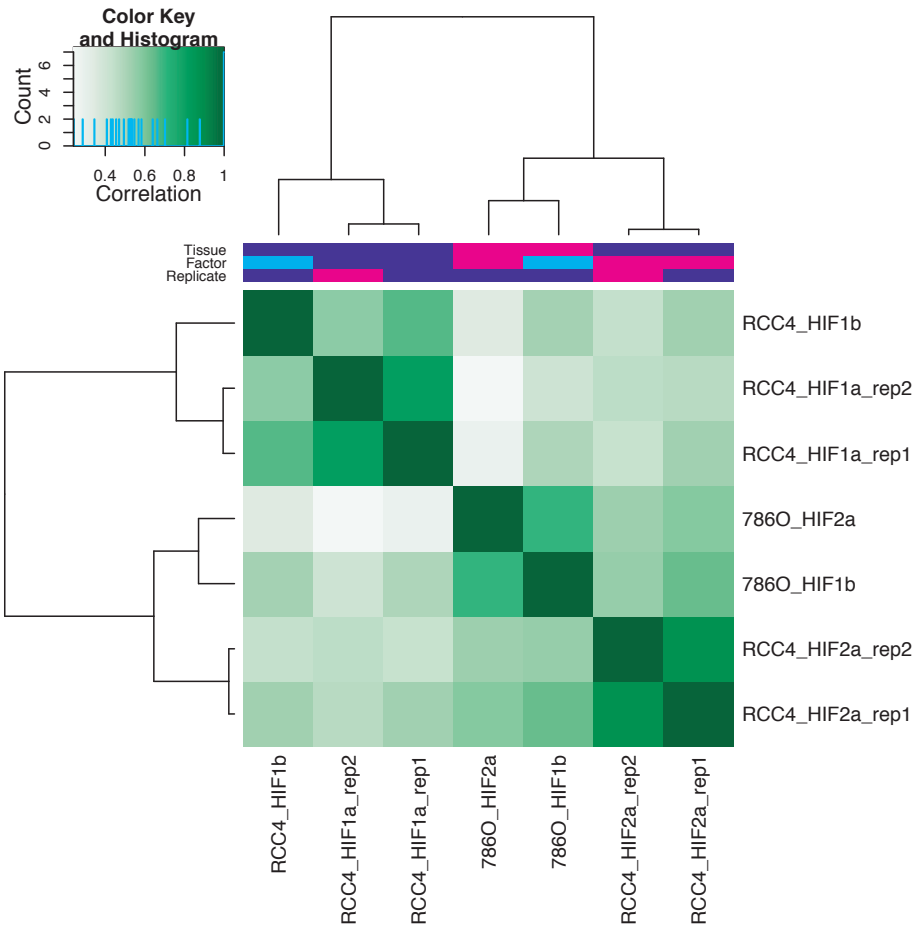
As a first step in the comparison of the 7 datasets, I used DiffBind to correlation of signal and overlap of binding sites of the different ChIP-seq datasets (Stark and Brown 2018; Ross-Innes et al. 2012). For each peak identified, the MACS peak caller assigns a probability score. For peaks present in two or more datasets, it is possible to correlate these probability scores across all common

peaks to generate a correlation coefficient between each pair of datasets. Each dataset can then be clustered according to these correlation coefficients and the overall result plotted as a heatmap in which perfect correlation (i.e. the correlation of a dataset with itself) is represented in dark green and zero correlation is represented in white. (Figure 2.3A). Firstly, as expected, the two replicates of the HIFα ChIP-seq analyses in RCC4 cells correlate most closely with each other. Therefore, in future analyses, these datasets were combined and only peaks present in both replicates for each ChIP were used. Secondly, the two 786-O datasets also correlated closely with each other as did the HIF-1α and HIF-1β datasets in RCC4 cells. However, in this analysis, the HIF-2α ChIP-seq datasets in RCC4 cells clustered most closely with the datasets in 786-O cells (in which HIF-1α is not expressed) and this may represent previously described isoform intrinsic differences in binding of the HIF-1 and HIF-2 isoforms which are central in supporting tumouriginisis in RCC (Yao et al. 2017).

Secondly, I used DiffBind to define the number of datasets in which each peak present in at least one dataset was identified. Binding sites that overlapped were first combined, giving a total of 17,041 non-overlapping binding sites that were identified. Of these 5,097 were identified in two or more datasets, 2,655 in three or more and so on (Figure 2.3 B). It is likely that these differences represent both technical noise as well as true biological differences between the binding sites of the different HIF isoforms in the different cell lines. Therefore, to increase the level of confidence in the called peaks, whilst still allowing for the possibility of biological differences between the different isoforms and cell lines, in future analyses, I only used binding sites that were present in immunoprecipitations for

two or more isoforms or cell lines (i.e. 2 out of 5 datasets after replicates were combined). A total of 5054 sites met these criteria. This is only a small decrease in the number of sites compared to the 5097 sites ignoring replicas, indicating that the concordance of the datasets is high.

A


Color Key
and Histogram
Count
0 2 4 6
0.4 0.6 0.8 1
Correlation
Tissue
Factor
Replicate
RCC4_HIF1b
RCC4_HIF1a_rep2
RCC4_HIF1a_rep1
786O_HIF2a
786O_HIF1b
RCC4_HIF2a_rep2
RCC4_HIF2a_rep1
RCC4_HIF1b
RCC4_HIF1a_rep2
RCC4_HIF1a_rep1
786O_HIF2a
786O_HIF1b
RCC4_HIF2a_rep2
RCC4_HIF2a_rep1

B

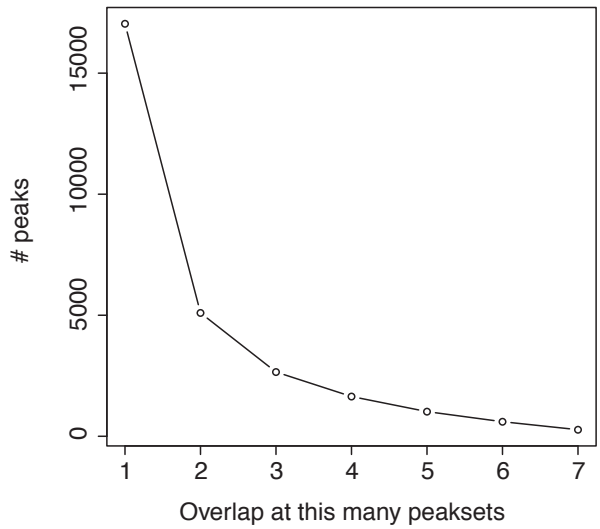

peaks
0
5000
10000
15000
1 2 3 4 5 6 7
Overlap at this many peaksets

Figure 2.3: Assessment of binding site differences of 7 HIF ChIP-seq datasets in cancer cell lines of RCC by DiffBind. (A) For each pair of datasets, in turn, the MACS probability scores, for each peak, were correlated and the correlation coefficients plotted as a heatmap using a colour scale in which perfect correlation (r^2=1) is plotted in dark green and no correlation (r^2=0) is plotted in white. Unsupervised hierarchical clustering was then performed using these correlation coefficients and is represented by the dendrograms. (B) The number of peaks (y-axis) shared between the designated number of datasets (x-axis) is shown as a line graph.

2.2.3 RCC susceptibility loci overlap with binding sites of HIF

I next aligned the haplotype blocks corresponding to the 9 renal cancer-predisposing SNPs defined in section 2.2.1 of this chapter with the set of high confidence HIF binding sites mapped by ChIP-seq in cancer cell lines of RCC defined in section 2.2.2. This showed that the haplotype blocks, at 4 out of the 9 published RCC GWAS loci, overlap with at least one high confidence HIF binding site mapped the RCC cell lines. The loci on 11q13.3, 8q24.21 and 12p12.1 coincided with intergenic HIF binding sites (Figure 2.4 A-C). The locus on 6q22.31 overlaps a HIF binding site close to the promoter of the CEP85L gene (Figure 2.4 D).

Figure 2.4 A: 11q13.3 locus

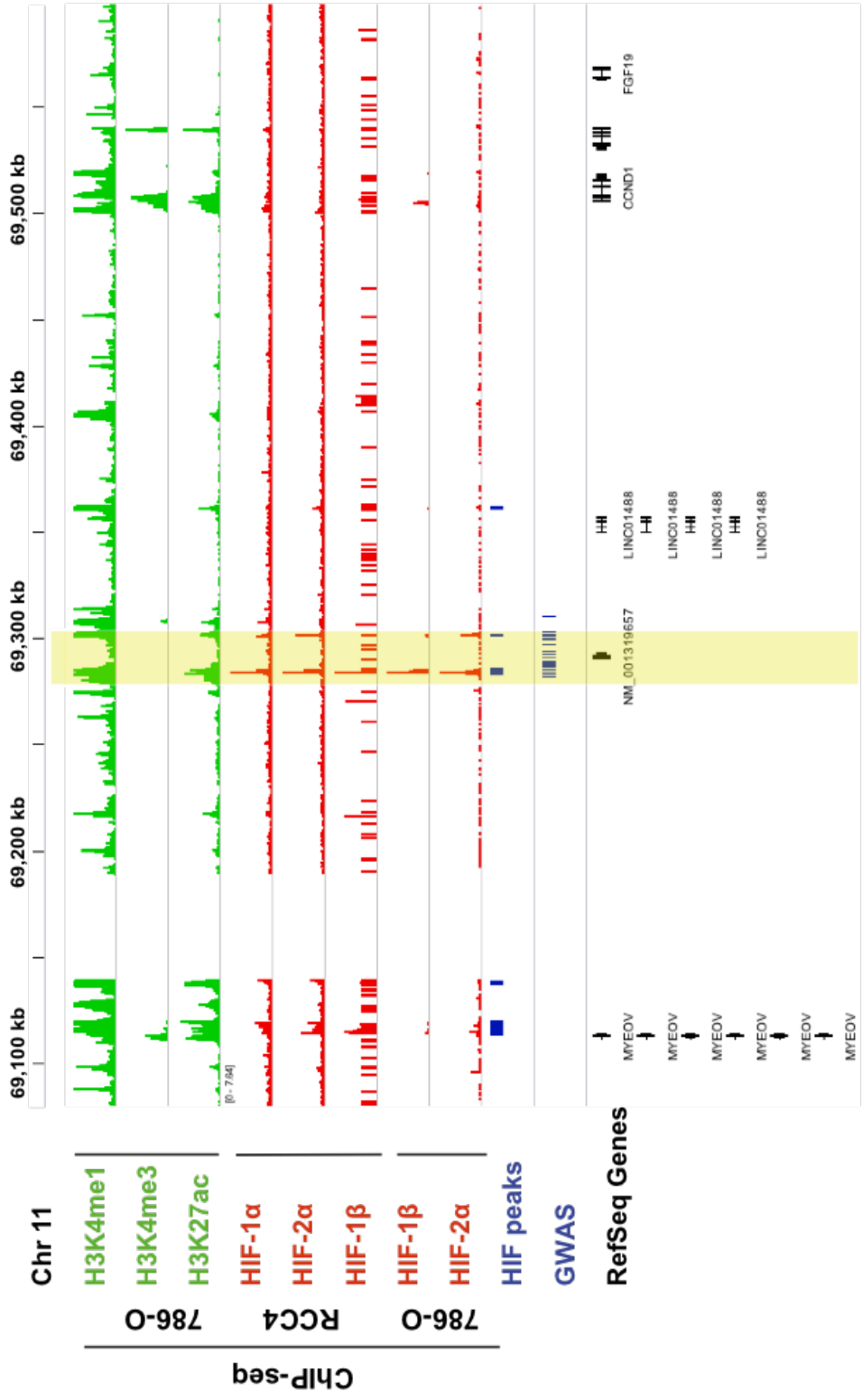


Figure 2.4 B: 8q24.21 locus

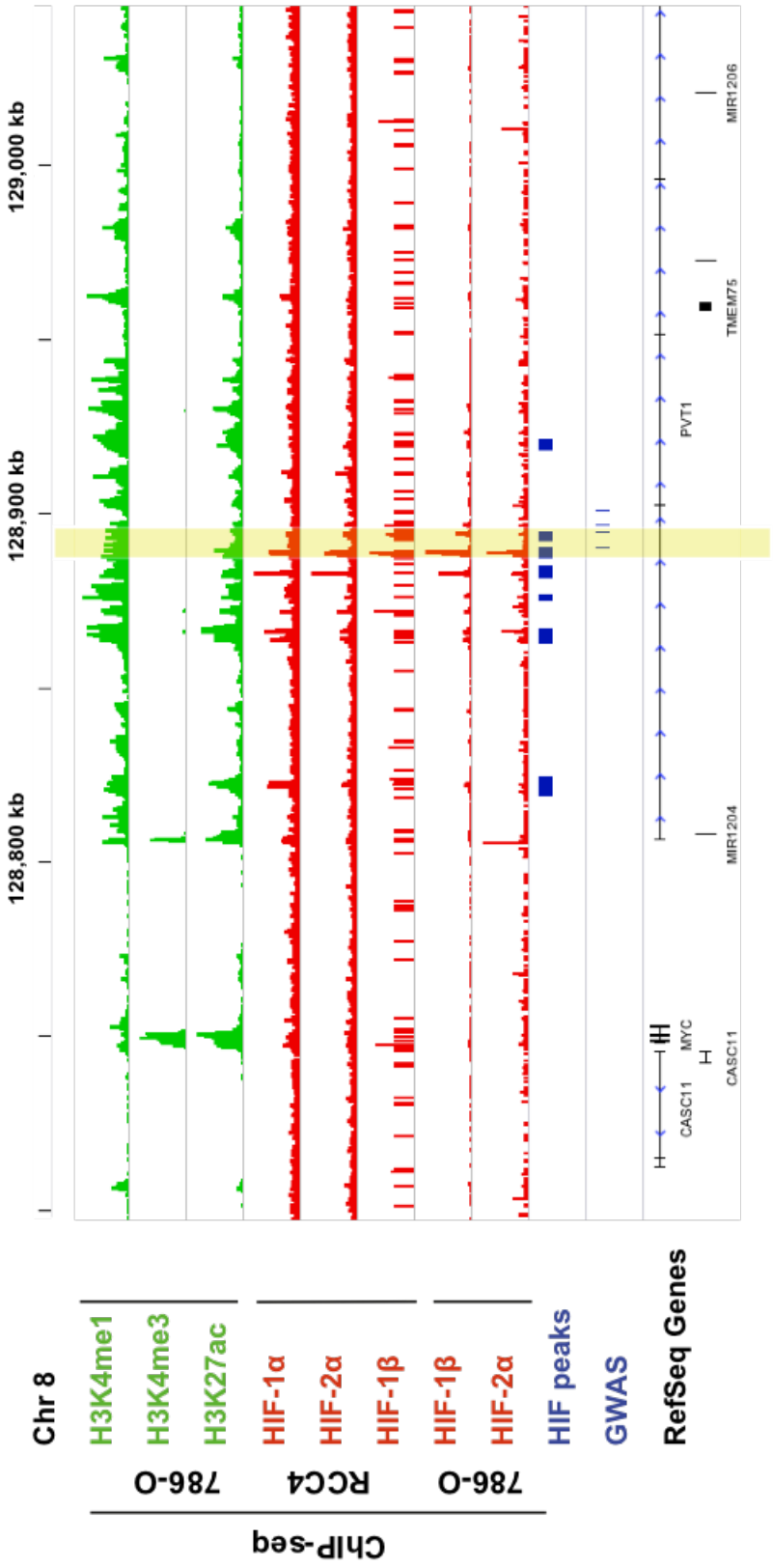


Figure 2.4 C: 12p12.1 locus

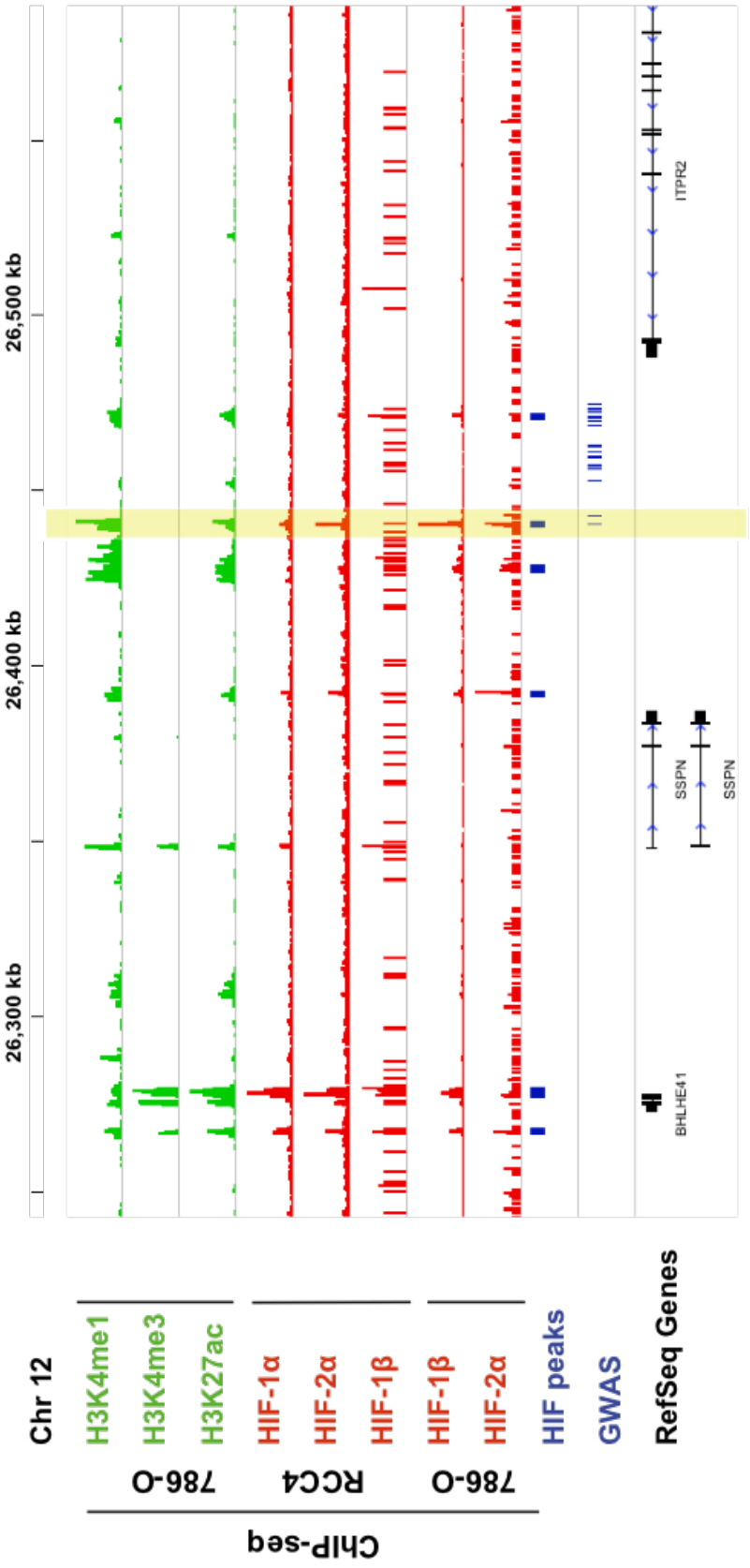


Figure 2.4 D: 6q22.31 locus

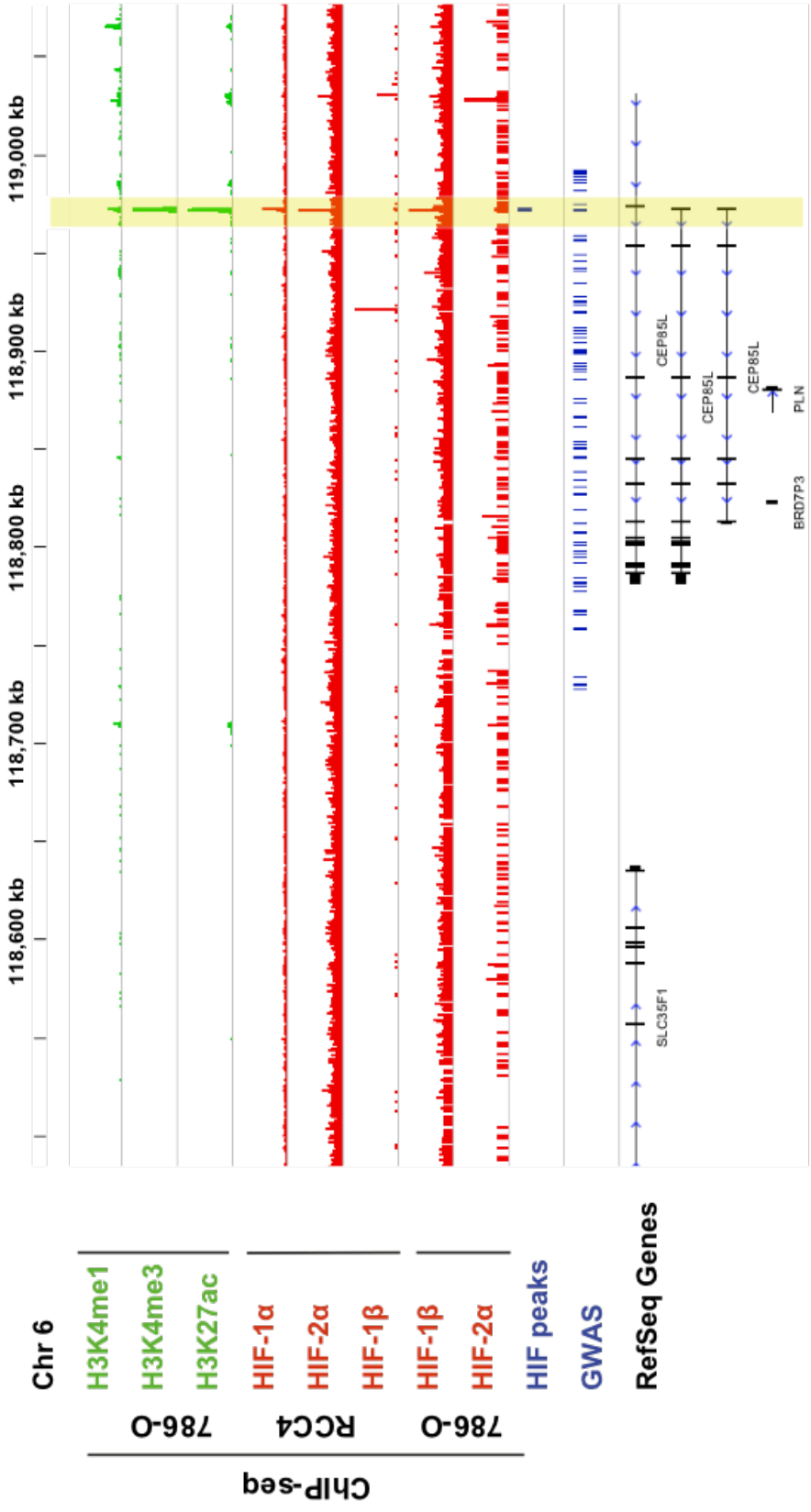


Figure 2.4: Integrative Genomics Viewer (IGV) of 4 RCC-susceptibility loci that overlap with HIF binding sites. (A) 11q13.3, (B) 8q24.21 (C) 12.p.12.1 and (D) 6q22). Track 1-3 from top (in green) display ChIP-seq signal (normalised read counts per million total reads at each chromosomal position) for histone H3K4me1, H3K4me3 and H3K27ac chromatin marks mapped in the 786-O kidney cancer cell line. Track 4-8 (in red) display binding of HIF-1α, HIF-2α and HIF1-β mapped in the two cancer cell lines, RCC4 and 786-O. Track 9 (in blue) indicates coordinates of HIF binding sites identified in at least two out of the 5 HIF ChIP-seq datasets. Track 10 (in blue) indicates the haplotype block corresponding to the LD region of the RCC susceptibility locus (index SNP and SNPs LD≥0.8). Track 11 (in black) displays gene annotation from the RefSeq hg19 (GRCh37) build.

2.2.4 RCC susceptibility loci identified by independent GWA studies coincide HIF binding sites more than expected

The high proportion of renal cancer polymorphisms that overlap with HIF binding sites (4 loci overlap with 6 HIF binding sites) raised the question of whether this proportion is higher than would be expected by random chance. To address this, I developed a novel statistical method to assess whether the observed overlap is more than expected. Essentially, I adopted a bootstrapping approach to determine the frequency distribution of the expected overlap between GWAS-identified RCC-susceptibility loci and HIF binding sites. The expected frequency distribution was modelled by randomly moving the susceptibility loci across the

genome 100,000 times, whilst keeping the LD structure constant. For each round of iteration, the number of loci coinciding with a HIF binding sites was registered. I observed that 4 or more susceptibility loci directly coincided with at least one HIF binding site in only 10 out of the 100,000 iterations. In other words, the probability of observing 4 or more loci overlapping a HIF-binding site by random chance alone is 10/100,000, or 0.0001. This indicates that the observed number of 4 overlaps is highly statistically significant with a P-value of 10^{-4} (Figure 2.5).

It has been reported, that like transcription factor binding sites, susceptibility loci are enriched at regulatory regions such as enhancers and promoters of genes (Parker, 2013). It is therefore important to exclude the possibility that the observed statistically significant overlap between RCC susceptibility loci and HIF binding sites is the result of a systematic error resulting from a mutual bias of both, RCC susceptibility loci and HIF binding sites, towards enhancers. To test this, enrichment of HIF binding sites at susceptibility loci was assessed using a second approach in which the HIF binding sites were randomly shuffled but constrained to fall within sites marked as active regulatory regions based on chromatin marks. Three ChIP-seq datasets (without replicates) were available mapping the histone marks H3K4me1, H3K4me3 and H3K27ac in 786-O cells. H3K4me3 is a marker of gene promoters whereas H3K4me1 is largely present at enhancers and H3K27ac is present at both active promoters and enhancers (Encode_Project_Consortium et al. 2012). To determine the sites marked with each histone modification, peaks were first identified in each dataset using the same analysis pipeline used for HIF ChIP-seq. Sites co-bound by H3K27ac and either H3K4me1 or H3K4me3 or both were defined as active regulatory regions.

In total 61,480 regulatory regions were defined from the three datasets using this approach.

The number of “pseudo” HIF-binding sites that overlapped an RCC-susceptibility locus when randomly shuffled within these regulatory regions was again assessed in 100,000 iterations to generate the expected frequency distribution. This second approach accounts for the fact that HIF can only bind to a specific subset of the genome, such as enhancers and active promoters which are also preferentially targeted by susceptibility loci.

In this analysis, an overlap of 6 or more HIF binding sites with RCC susceptibility loci was only observed in 9 out of 100,000 iterations and hence the observed overlap is again highly statistically significant (P-value = $9*10^{-5}$, Figure 2.5). This confirms that RCC susceptibility loci are enriched at HIF binding sites and that this enrichment is higher at HIF binding sites compared to the set of sites marking active promoters and enhancers.

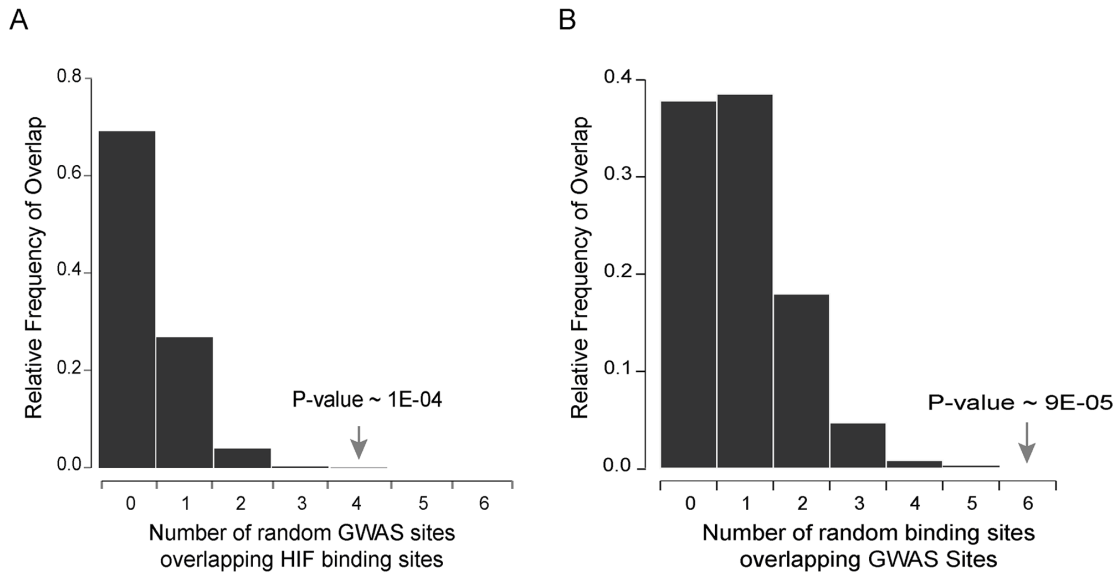


Figure 2.5: RCC-associated polymorphic loci overlap HIF-binding sites in RCC cells. A total of 4 RCC susceptibility loci coincide with a total of 6 HIF binding sites mapped in cancer cell lines of RCC. A) The expected frequency distribution of the overlap of RCC susceptibility loci with HIF binding sites was assessed by moving the index SNPs and corresponding LD SNPs (all SNPs LD≥0.8) randomly along the genome. The overlap with HIF binding sites was assessed for every of the iteration. B) Conversely, the expected overlap of HIF binding sites with the LD regions corresponding to the RCC susceptibility loci was assessed by moving the HIF binding sites within regulatory regions (defined by H3K27ac, H3K4me3, H3K4me1 in 786-O cells). This figure is adapted from Grampp, Schmid et al. (Grampp et al. 2017).

2.3 Discussion

In chapter 2, I have shown that 4 out of 9 RCC susceptibility loci overlap with a total of 6 HIF binding sites defined in RCC cell lines. This is statistically more than would be expected by random chance alone. Furthermore, it is more than would be expected if RCC-susceptibility polymorphisms were targeting active enhancers more widely and suggests a specific association between loci that confer genetic predisposition to RCC and cis-acting components of the HIF pathway. In turn, this suggests that kidney cancer is more prevalent in individuals who are genetically predisposed to have an altered HIF response in the kidney cancer cells themselves.

However, this overlap does not necessarily imply that the index polymorphisms are having a direct effect on HIF binding. Firstly, the index SNP is just a marker for a haplotype block of closely linked SNPs that are commonly co-inherited in the general population and the statistical association is influenced by different parameters such as allele frequencies (Wang et al. 2010). Any one (or more) of these SNPs could be causative and it is therefore impossible to define the precise base substitution responsible for the association with RCC. Secondly, there are both experimental and biological limitations to the precision with which ChIP-seq can identify the base positions of the functional HIF binding element. Experimentally, the resolution of ChIP-seq is limited by the mean size of the chromatin fragments used in the assay which is typically several hundreds of bases. Whilst it may be possible to infer the position more precisely by reference to the point of maximum signal (summit), the position of the known consensus HIF-binding motif, or by using higher precision techniques such as ChIP-exo,

such an approach may be too restrictive (Rhee and Pugh 2011). As illustrated below, the RCC-associated allele may introduce a second HIF-binding motif (not present in the reference genome) that alters the strength of HIF binding. Furthermore, base substitutions that are distant from the HIF-binding site itself may alter the chromatin environment and/or binding of co-factors that are important to the HIF response. Nevertheless, multiple overlap of RCC susceptibility loci with HIF binding sites and the significant statistical association of the observed overlap, do provide a strong indication that RCC-associated polymorphisms are having a functional effect on the HIF response at these loci.

Further evidence for this comes from two sources. Firstly, *in vitro* experiments, performed by co-workers, in RCC cell lines that are heterozygous at the 11q13.3, 8q24.21, or 12p12.1 loci have shown increased HIF binding to the RCC-susceptibility allele as well as allele-specific expression of the HIF-target gene at each locus (Schödel et al. 2012; Grampp et al. 2016; Grampp et al. 2017). Mechanistically, polymorphisms at the 11q13.3 and 8q24.21 loci, alter DNA accessibility at the HIF-binding site while those at the 12p12.1 locus generate a second HIF-binding motif that quantitatively alters HIF binding. Secondly, as I will show in chapter 4, using large-scale public datasets, the genotype at each locus influences the expression level of the HIF target gene *in vivo* in the renal tumours themselves.

In each case, RCC-susceptibility polymorphisms that overlap HIF-binding sites (i.e. cis-acting components of the HIF pathway) affect specific HIF target genes. They therefore potentially implicate these individual genes in the pathogenesis

of kidney cancer and in turn help distinguish diver genes from bystander or passenger genes that are co-activated as a consequence of activation of the large pathway. However, RCC-susceptibility polymorphisms may also reduce the expression of genes that have an inhibitory effect on the development of RCC. In addition, it has been estimated that the RCC-susceptibility polymorphisms identified in GWAS analyses to date account for only a small proportion of the heritability of RCC (Scelo et al. 2017). Therefore, it is likely that additional loci and affected genes would be identified in larger studies. Thus, whilst the positive predictive value of this approach in identifying driver and repressor genes is likely to be high, a proportion of genes currently classified as neutral passengers, may turn out to have an association with kidney cancer.

Whilst RCC-susceptibility polymorphisms may exert their effect at any stage in the development of the cancer, from the precursor cell right through to advanced stages in tumorigenesis, the early activation of the HIF response by loss of VHL suggests that those acting on the HIF pathway likely act after this major pathway switch. Furthermore, given that polymorphisms at 11q13.3, 8q24.21 and 12p12.1 all affect HIF binding and target gene expression in tumour cells derived from advanced tumours, it would be interesting to know whether the genotype affects tumour progression or prognosis in patients with kidney cancer. Unfortunately, datasets with which to address this question are not yet available at sufficient scale.

Finally, there are a number of limitations imposed by the nature of the individual GWAS analyses available for RCC. Firstly, studies do not generally distinguish

between the histological sub-types of kidney cancer. Whilst the vast majority (approximately 85%) will be clear cell RCC a number will be of other histological sub-types (mainly papillary) in which the HIF pathway is not constitutively activated by VHL loss. However, this would tend to dilute any association between RCC-susceptibility loci and HIF-binding sites. Secondly, each individual GWAS is relatively small, typically comprising a few thousand individuals and controls and this raises questions about the sensitivity of the studies in identifying weak associations between susceptibility loci and RCC. Furthermore, each describes different loci raising questions about the reproducibility of these studies. Generally, the studies however have either confirmed previously described loci associated with RCC or have excluded these loci from analysis. Nevertheless, in general terms, the more patients and controls included in GWAS analyses, the more confidently RCC-susceptibility loci can be identified.

Chapter 3: A meta-analysis-defined set of RCC susceptibility loci target HIF binding sites in RCC cancer cell lines.

3.1 Introduction

The statistical overlap observed between HIF-binding sites in RCC cell lines and RCC-susceptibility loci, defined across a number of relatively small individual GWAS analyses, as described in chapter 2, is convincing. However, there are a number of limitations with using small studies such as these.

Firstly, in each study, the reported P-values for the significance of the association of the index SNP at the described loci with RCC only just reach conventional levels of genome-wide significance, even when the discovery and validation cohorts are combined. Furthermore, each study describes between just 1 and 3 loci associated with RCC. In addition, the Odds Ratio for each risk allele is very small, indicating that the described loci account for only a small proportion of the known heritability in RCC. Taken together, this suggests that the sensitivity of these relatively small GWAS analyses for detecting genetic associations with RCC is poor.

Secondly, it is difficult to assess the concordance in the genetic association with renal cancer across the various studies. In particular, it is difficult to assess the extent to which the loci described in one study replicated in the other studies. In some papers, the authors have referred to the significance level of previously described associations in their own dataset. Whilst these generally reach significance (i.e. P-value < 0.05) the associations are generally weaker than in the original study and usually fail to reach genome-wide significance. Furthermore, several papers that are focussed on identifying new genetic

associations with RCC, deliberately exclude previously described loci from their analysis, although it is clear from some of the figures that a number of previously described loci did reach genome-wide significance in later GWAS analyses. This apparent heterogeneity in the association between genetic polymorphism and RCC, may also arise partly from technical issues (the different studies used differing combinations of genotyping microarray, some imputed the genotype, whilst others did not, and the populations studied varied – e.g. Gudmundsson examined cases in an Icelandic population (Gudmundsson et al. 2013)). However, it is likely that the low power to detect these associations accurately also contributed to the differences between the studies. Overall this raises questions about the specificity as well as the sensitivity of using a number of small GWAS analyses to determine genetic associations with RCC.

The limitations associated with small GWAS analyses may be overcome by undertaking large-scale GWAS, with many more cases and controls. However, this is both costly and time-consuming. An alternative approach is to pool all available data into one much larger (meta)-analysis. In this type of analysis, the genotypes determined from the diverse microarray genotyping platforms used in each study are first imputed onto a common set of SNPs before being re-analysed as a single dataset. The increased power afforded by the much larger number of cases and controls generally results in true genetic associations achieving a higher level of statistical significance, whilst false associations will tend not to repeat.

In December 2016, I was granted prepublication access to SNP-level summary statistics from a large meta-analysis of RCC performed by collaborators at the National Institute of Health (NIH) in the US (Scelo et al. 2017). Such detailed, standardised datasets were not available from the individual studies used in chapter 2.

As well as providing the most sensitive and comprehensive analysis of the genetic associations with RCC, this dataset also provided me with several other opportunities for further analyses. Firstly, imputation onto a standard set of SNPs across the genome allows for the comparison with other GWAS datasets for traits unrelated to kidney cancer, which enables another level of control to my work that is important from a methodological perspective. In this chapter, I have used educational attainment, for which summary statistics are publicly available from the (NHGRI-EBI) GWAS catalogue to better act as a negative control (Buniello et al. 2019; Okbay et al. 2016). Secondly, as I will describe later, in chapter 6, it has allowed me to explore genetic associations that fall below conventional levels of genome-wide significance.

3.2 Results

3.2.1 Definition of RCC susceptibility loci from a meta-analysis of GWA studies

Mentioned in the introduction to this chapter, during the course of this work, our collaborators at the National Institute of Health (NIH) performed a meta-analysis of GWA studies in RCC. The study integrated genetic variant information from four previous studies as well as two new genetic scans. Genotype information on a total of 10,784 cases of RCC and 20,406 controls of European ancestry was integrated. Genotyping in these studies had been performed on different platforms, which differed in the genetic markers tested. For the combined meta-analysis, the data was first imputed using the 1,000 Genomes Phase 1 release 3 onto a common set of 7,437,092 SNPs. Statistical significance of any association with RCC was determined by the Cochran's q test assuming a log additive model (Figure 3.1). The study also tested for replication of genome-wide significant SNPs detected in an independent dataset (Scelo et al. 2017). However, for consistency and to be able to systematically compare my results with other datasets, I relied solely on the summary statistics for further analysis.

-log10 p-value
0
5
10
15
20
chromosome
1
2
3
4
5
6
7
8
9
10
12
14
16
18
21

Figure 3.1: Manhattan plot of RCC GWA study. The Manhattan plot shows the statistical significance of RCC association of the SNPs with P-value ≤ 10^{-2} as determined by a meta-analysis of RCC GWA studies against the chromosomal location of the SNP (Scelo et al. 2017).

In order to test whether GWAS-defined RCC susceptibility loci were enriched at HIF-binding sites using this new meta-analysis of GWA studies, SNP-level summary statistics from the study were used to define independent loci that reached genome-wide significance (P-value ≤ $5*10^{-8}$). Independent loci were first identified, beginning with the most significantly associated SNP in the genome, which was defined as the index SNP for that locus. All SNPs in close genetic or physical proximity to the index SNP ($r^2 \geq 0.2$ with reference to the 1000 Genomes EUR population data or distance < 500 Mb) were then excluded. SNPs in genetic linkage with the index SNP were excluded since they would show a significant association with RCC simply by co-inheritance with the causative polymorphism. A threshold of $r^2 \geq 0.2$ was chosen to exclude SNPs that had even a weak correlation with the index SNP. Loci that were in close proximity (500 Mb) were also excluded even if they were not in genetic linkage as they might affect the same set of genes. Having excluded all SNPs that might be associated with the first locus, the next most significant SNP defined the second locus and the process was repeated until there were no further SNPs that reached genome-wide significance (P-value ≤ $5*10^{-8}$).

Having defined the index SNP at each independent locus, I next defined the haplotype blocks corresponding to each index SNP based on all SNPs in strong linkage ($r^2 \geq 0.8$ with reference to the 1000 Genomes EUR population data) with the index SNP. A high level of genetic linkage ($r^2 \geq 0.8$) was used in this analysis since I only wanted to include SNPs that were highly likely to be co-inherited with the index SNP.

This analysis defined 13 index SNPs and their corresponding haplotype blocks as independent RCC susceptibility loci that reached genome-wide significance (P-value $\leq 5 \times 10^{-8}$) (Table 3.1). 6 loci identified in the previous small-scale GWAS analyses referred to in chapter 2 were identified using this approach. The meta-analysis also identified 7 new loci (14q24.2 (does not include HIF-1α), 11q22.3, 1p32.3, 4q23, 3p22.1, 15q22.31, 3q26.31), which had not been identified by the previous smaller separate studies. However, 3 loci identified by the separate studies (6q22.31, 5p13.3, 1q24.1) were not present at genome-wide significance in this larger study (p=0.01, p=0.08 and p=0.006 respectively). These loci might previously have been erroneously identified as a result of population stratification in the prior studies.

Locus	*rs number index SNP*	*P-value*
11q13.3	rs11263654	1.65E-23
2p21	rs2121267	6.95E-19
12p12.1	rs11534749	6.19E-17
14q24.2	rs28840762	1.08E-15
12q24.31	rs10846748	1.70E-12
8q24.21	rs6470588	2.82E-12
11q22.3	rs117706999	6.60E-10
2q22.3	rs11888238	8.15E-10
1p32.3	rs6676515	3.04E-09
4q23	rs7697932	4.70E-09
3p22.1	rs9821249	1.12E-08
15q22.31	rs12905354	2.28E-08
3q26.31	rs234043	4.10E-08

Table 3.1: Independent RCC susceptibility loci (P-value ≤ $5*10^{-8}$) defined from summary statistics of an RCC GWAS meta-analysis. Table 3.1 indicates chromosomal location, the most-significant (index) and the significance level corresponding to the index SNP of each RCC susceptibility locus.

3.2.2 RCC susceptibility loci identified by a meta-analysis of GWA studies directly target HIF binding sites

Next, the haplotype blocks for the 13 RCC-susceptibility loci identified in the meta-analysis of GWAS were aligned with the same set of high confidence HIF binding sites mapped in RCC cancer cell lines as defined in chapter 2.2.2. Four out of the 13 loci directly overlapped with a total of 5 HIF binding sites.

Importantly, three of these loci (those on 11q13.3, 8q24.21 and 12p12.1) were present in the previous analysis and directly overlapped the same HIF binding sites despite the small differences in the index SNPs and consequent haplotype blocks identified. The locus on 6q22.31 (identified in the previous analysis) was not present in the meta-analysis (i.e. was not significantly associated with RCC). However, an additional locus on 3q26.31, which was identified in this larger study did show direct coincidence with a HIF binding site (Figure 3.2).

Figure 3.2: 3q26.31 locus

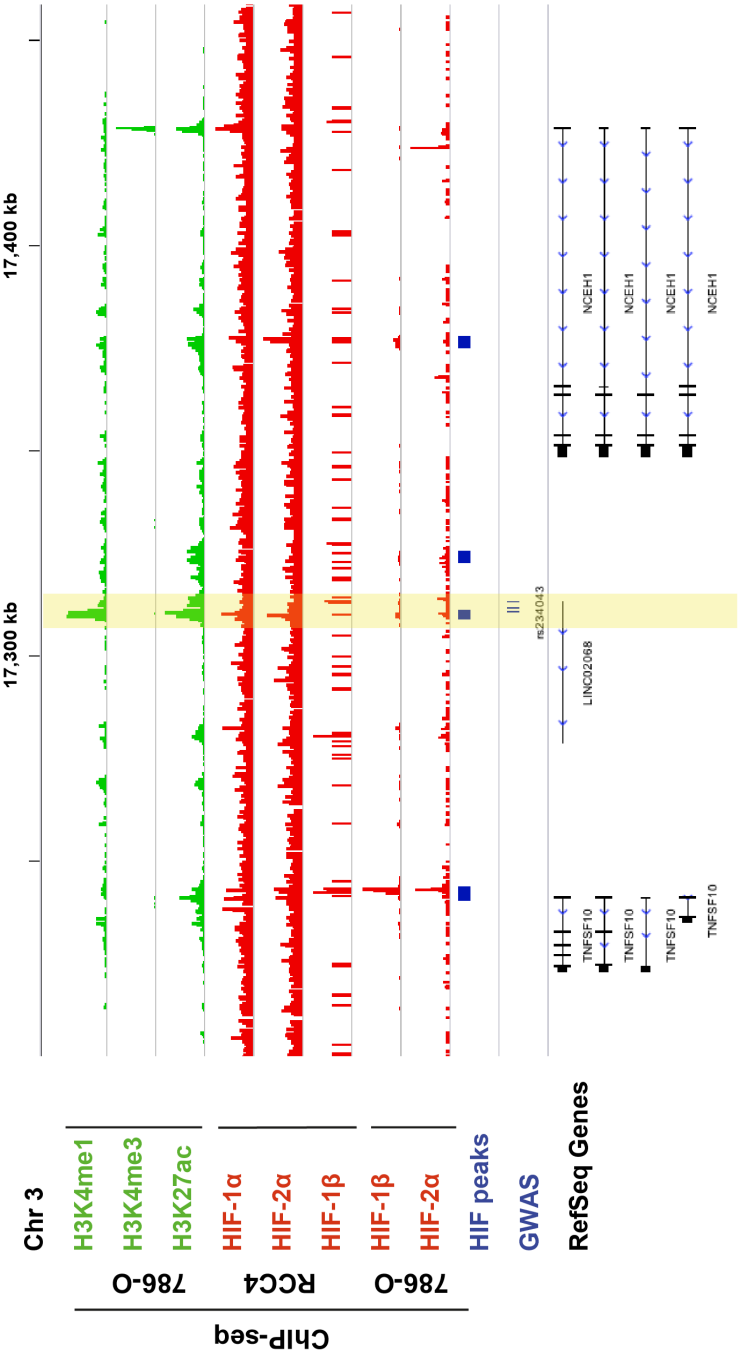


Figure 3.2: Next-Generation Sequencing tracks of the 3q26.31 RCC susceptibility locus. Tracks 1-3 from top (in green) display ChIP-seq analyses of the chromatin marks, H3K4me1, H3K4me3 and H3K27ac mapped in RCC 786-O cell lines. Tracks 4-8 (in red) display ChIP-seq analysis of HIF-1α, HIF-2α and HIF-1β binding mapped in the two cancer RCC cell lines, RCC4 and 786-O. Track 9 (in blue) indicated coordinates of HIF binding sites present in at least two out of the 5 HIF ChIP-seq datasets. Track 10 (in blue) indicates the haplotype block corresponding to the LD region of the RCC susceptibility locus (index SNP and SNPs in high LD with an $r^2 \geq 0.8$). Track 11 (in black) displays chromosomal coordinates and gene annotation from the RefSeq hg19 (GRCh37) build.

3.2.3 RCC susceptibility loci defined by a meta-analysis of GWA studies overlap with HIF binding sites more than expected

Having determined that 4 (11q13.3, 8q24.21, 12p12.1 and 3q26.31) out of the 13 meta-analysis defined RCC susceptibility loci had at least one RCC-associated SNP that directly overlapped with a total of 6 HIF binding peaks, I again examined whether this overlap is more than expected by chance. Applying the same methods, described in chapter 2, I first estimated the frequency distribution for the expected overlap between the 13 RCC susceptibility loci identified in the meta-analysis of GWAS and the HIF binding sites. Again, it was confirmed that the observed overlap of 4 RCC susceptibility loci coinciding with HIF binding sites was significantly more than expected (P-value = $7*10^{-5}$). Despite the lower fraction of 4 out 13 loci overlapping based on the meta-analysis compared to the

4 out of 9 loci from the independent GWA studies, the significance here is higher. This likely reflects differences in the linkage disequilibrium structure and sizes of the haplotype blocks.

To confirm that this overlap represents a specific concordance between RCC-associated susceptibility loci and HIF binding sites, as opposed to general regulatory elements such as enhancers and promoters, the reverse analysis was again performed. HIF binding sites were randomly shuffled within active regulatory regions (as defined in chapter 2.2.2) and the number of HIF binding sites directly overlapping at least one RCC-associated SNP corresponding to the 13 RCC susceptibility loci was registered. This analysis estimated that the overlap of 6 HIF binding sites with 4 RCC susceptibility loci is significantly more than expected (P-value = $6*10^{-5}$).

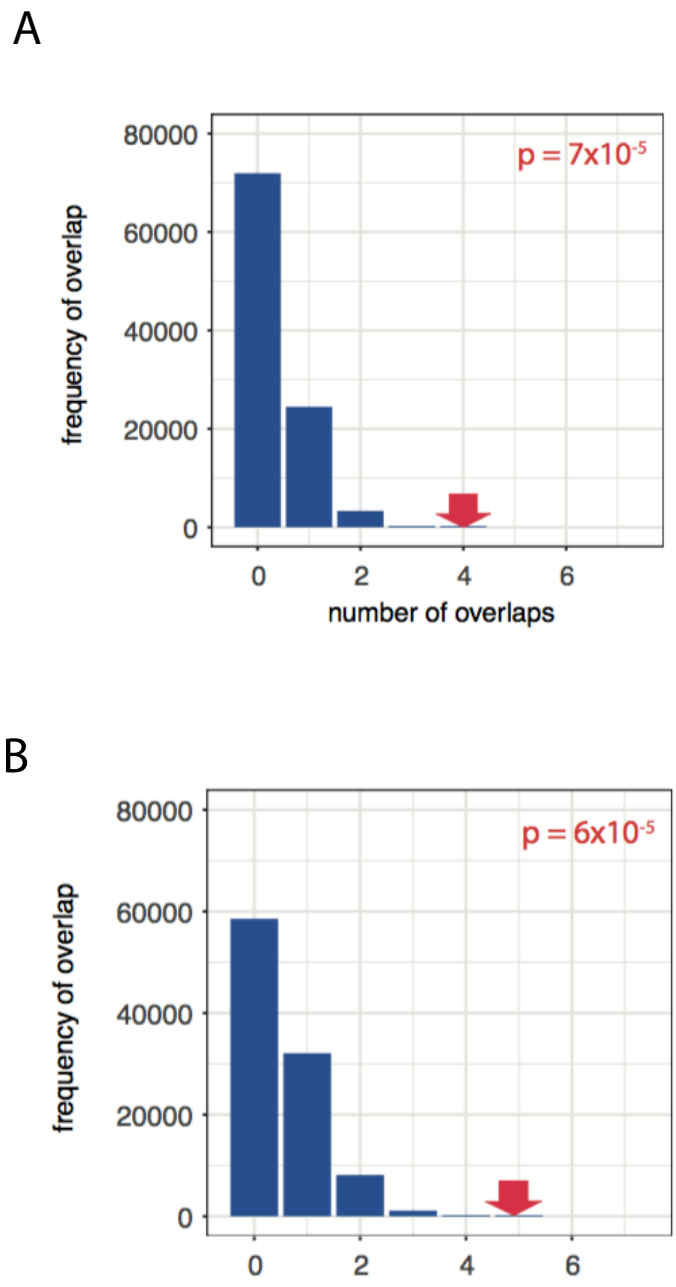


Figure 3.3: A total of 4 RCC susceptibility loci directly overlap with a total of 5 HIF bindings sites mapped in renal cancer 786-O and RCC4 cell lines. A) The expected frequency distribution (blue bars) for the overlap of GWAS-identified susceptibility loci with HIF binding sites was assessed by moving the

index SNPs and corresponding LD SNPs (all SNPs $r^2 \geq 0.8$) randomly along the chromosome of the locus. The overlap with HIF binding sites was assessed for each iteration. For comparison, the observed overlap is indicated by the red arrow. B) Conversely, the expected overlap of HIF binding sites with LD regions of RCC susceptibility loci was assessed by moving the HIF binding sites within the enhancer space (defined by H3K27ac, H3K4me3, H3K4me1 in 786-O cells).

To provide further evidence for the excess overlap between RCC susceptibility loci and HIF binding sites, an independent method was applied to the analysis of the two datasets (Wang et al. 2016). This method calculates the proportion of SNPs above a given statistical significance level that overlap a HIF-binding peak, normalised to the overall proportion of SNPs that overlap such a peak genome-wide. For comparison, the same analysis is performed for SNPs overlapping the regulatory regions defined by histone H3K4me1, H3K4me3 and H3K27ac modifications. This type of analysis is only possible where SNP level summary statistics are available and makes no assumptions about the haplotype structure or linkage between SNPs at each locus. SNPs are first ranked in order of descending statistical association with the trait and plotted on the x-axis on a -log10 scale (Figure 3.4). Starting from the top ranked (most significant) SNPs, an enrichment score is then computed (plotted on the y-axis) comprising the proportion of SNPs with the same or more significant P-values that are present

at the functional elements in question compared to the same proportion across the entire genome, according to the equation:

$$ES = (Xo/Xt)/(G_o/G_t)$$

ES = enrichment score

X_o = Number of SNPs above threshold that overlap HIF-binding site

X_t = Total number of SNPs above threshold

G_o = number of SNPs genome-wide that overlap HIF-binding site

G_t = total number of SNPs genome-wide

Applying this method (Figure 3.4), I tested whether SNPs stratified by P-value (for their association with RCC) show enrichment specifically at HIF binding sites (blue line) or at regulatory elements in general (red line), as mapped in ChIP-seq analysis of RCC cell lines (described in chapter 2.2.2.). The index SNP at the 11q13.3 locus does not overlap a HIF binding site, which manifests as no enrichment for HIF binding sites and the blue line dropping to zero for the most significant SNPs in this dataset. However, several highly significant SNPs at the 11q13.3 and additional loci do overlap with HIF binding sites, leading to a high level of enrichment at the highest significance levels. Hence, as the significance threshold is reduced, additional SNPs that overlap a HIF-biding site are uncovered and their enrichment is maintained. Indeed, SNPs with P-values of

10^{-15} to 10^{-20} show the highest levels of enrichment at HIF binding sites, being roughly 80-times more likely to overlap a HIF-binding site than expected by chance alone (blue line).

SNPs associated with RCC at above genome-wide significance levels are also enriched at regulatory elements in general (red line). However, this is consistently much lower (up to 10-fold enrichment) and is in contrast to the enrichment levels observed for HIF binding sites. Furthermore, whilst the enrichment for HIF binding sites diminishes at lower levels of significance, it remains above that for general regulatory elements even for SNPs below the genome-wide level of significance. This observation is in line with the previous result that HIF binding sites directly overlap RCC susceptibility loci more than expected (Figure 3.3 B). Enrichment for HIF binding sites is also observed for SNPs below genome-wide significance and this will be addressed in chapter 6.

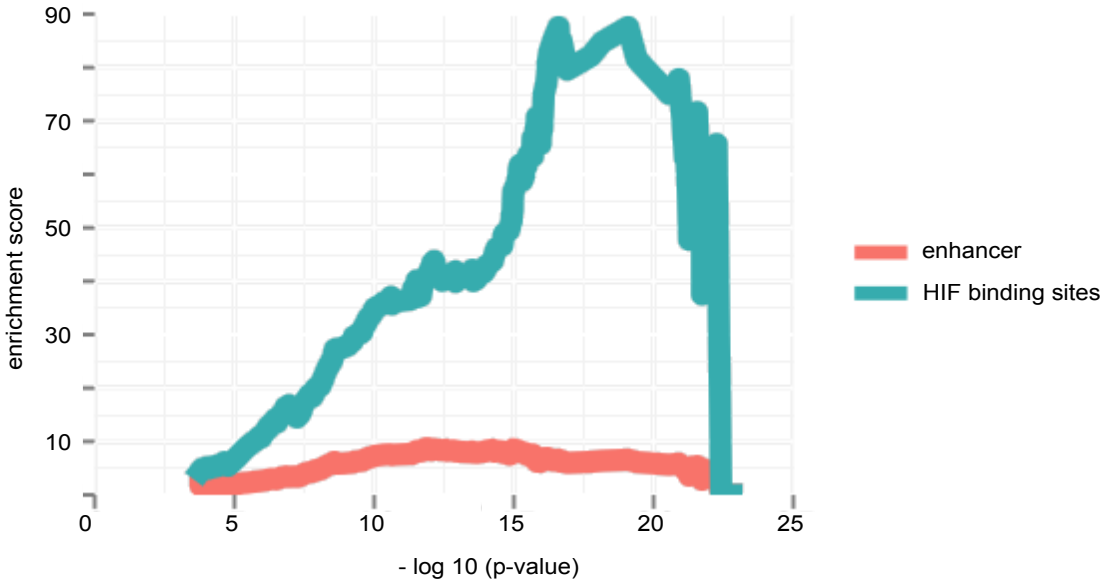


Figure 3.4: Fold enrichment of SNPs, ranked by their association with RCC, for HIF binding sites and general regulatory elements. As the significance threshold, for the SNP's association with RCC, was progressively reduced, the enrichment score for all SNPs above the threshold was calculated for HIF-binding sites (blue line) and general regulatory elements (red line). This was then plotted on the y-axis against the threshold used (-log10 scale) on the x-axis.

3.2.4 Educational attainment susceptibility loci coincide HIF binding sites at expected frequency

In the above analyses, I have used two independent methods to show that GWAS-defined RCC-susceptibility loci overlap with HIF-binding sites more than expected by random chance. As a negative control for these experiments, I next applied the same two methods to examine the overlap between these HIF-binding sites and GWAS-defined loci for a hypoxia-pathway unrelated trait.

I chose to use educational attainment, which is a trait that has not been linked to the hypoxia pathway as verified by an extensive literature search using Pubmed. Summary statistics of the educational attainment GWA study, imputed on the 1000 Genomes EUR population, were obtained from the National Human Genome Research Institute – European Bioinformatics Institute (NHGRI-EBI) GWAS catalogue (Okbay et al. 2016; Buniello et al. 2019). Based on the summary statistics, independent loci and haplotype blocks were defined in the same way as for the RCC GWAS meta-analysis earlier in this chapter. A total of 102 loci reached genome-wide significance (P-value $\leq 5*10^{-8}$).

7 out of the 102 educational attainment-associated susceptibility loci (index SNP and all SNPs in strong linkage, $r^2 \geq 0.8$) directly overlap 9 HIF-binding sites mapped in RCC cell lines. The frequency distribution for the expected overlap of educational attainment GWAS-defined loci with binding sites of HIF mapped in cancer cell lines of RCC was generated using the same iterative, boot-strapping approach as for RCC. The observed overlap of 7 loci is within the expected

distribution (P-value= 0.27). Conversely, I assessed the number overlaps of HIF binding sites mapped in RCC cell lines to determine the frequency distribution of the expected number of HIF binding sites to overlap with at least one SNP of educational attainment associated loci again using the same iterative, boot-strapping approach as for RCC. I generated the frequency distribution by randomly moving the HIF binding sites within active regulatory regions mapped in RCC cell lines (as defined in section 2.2.2) and assessing the overlap of the binding sites with the set of education attainment loci after each iteration. This showed that the observed overlap of 9 in cancer cell lines of RCC mapped HIF binding sites with GWAS-defined loci of educational attainment falls within the expected distribution (P-value=0.45).

3.3 Discussion

Analysis of SNP-level summary statistics from the discovery stage of the RCC GWAS meta-analysis identified 13 independent RCC-susceptibility loci. Six of these loci had been identified in the previous individual GWAS analyses, while seven were previously unreported. This is to be anticipated, since the meta-analysis includes data from 4 of these previous studies, alongside 2 new scans. However, the larger number of cases and controls in the pooled dataset increases the power of the study and provides a greater level of certainty regarding the association of the polymorphic loci with RCC. Indeed, the total number of loci associated with RCC is now increased and the significance levels for many of the previously established loci are now considerably above conventional levels of genome-wide significance.

Of the six loci that had previously been reported, three overlapped with a HIF-binding site in RCC cell lines. One of the new loci also overlapped a HIF-binding site. Using the same statistical methods that I developed in chapter 2, the overall number of RCC-associated loci that overlapped a HIF-binding site was again significantly more than would be expected by chance, confirming my findings in the previous chapter. Furthermore, imputation of the individual GWA studies onto a standard set of SNPs using the 1,000 Genomes whole-genome sequencing data, facilitated the use of other GWAS analyses of unrelated traits as a negative control, further increasing the confidence in my findings.

The availability of SNP level summary statistics also facilitated an alternative approach to looking for enrichment of these loci at HIF-binding sites (Figure 3.4). Using this independent methodology showed that SNPs that were highly significantly associated with RCC, were enriched up to 80-fold at HIF-binding sites, whilst their enrichment at active regulatory elements more generally was much lower at around 10-fold. Thus, two independent methodologies have confirmed enrichment of RCC-associated genetic variants specifically at HIF-binding sites.

Although the greater number of patients and controls in the combined meta-analysis increases the statistical certainty of the association between genetic polymorphism and RCC, systematic errors may affect any GWAS analysis and lead to false positive associations with RCC. Chief amongst these are errors resulting from population stratification. To exclude population stratification as a possible cause of the observed association at a GWAS locus, replication studies are performed in independent populations. If the association can be confirmed on this independent disease/control cohort this is strong evidence for a true disease association. Importantly, SNPs at all 4 loci from the discovery stage of the meta-analysis that overlapped a HIF-binding site were significantly associated with RCC in the independent replication cohort confirming their validity (Scelo et al. 2017).

Whilst SNPs that are significantly associated with RCC are enriched approximately 10-fold at general regulatory elements, this cannot account for the

more than 80-fold enrichment observed specifically at HIF-binding sites. However, it is possible that their enrichment at HIF-binding sites may account for the enrichment at sites marked by active chromatin modifications (H3K4me1, H3K4me3 and H3K27ac). Indeed, roughly 1 in 12 active regulatory regions bind HIF in my analyses, whilst the enrichment at active regulatory regions is approximately 1/10th that at HIF-binding sites.

Although the meta-analysis of RCC GWAS represents by far the largest and most powerful assessment of genetic variants associated with renal cancer, Scelo et al. estimate that the loci identified account for only about 10% of the heritability of RCC. Generally, the power to detect such statistical associations depends strictly on the number of alleles of each type rather than the number of individuals in the study. Thus, although the study was adequately powered to detect associations with common alleles, as the minor allele frequency falls, the ability of the study to detect even major effects on disease pre-disposition will also fall. This will be further compounded when the effect of a polymorphism on disease susceptibility is small. Taken together, this suggests that a large proportion of the genetic variation that pre-disposes to RCC fails to reach conventional levels of genome-wide significance. However, it is possible that such signals are enriched amongst SNPs with intermediate levels of association with RCC. If this was the case, SNPs below conventional levels of genome-wide significance may be enriched for HIF binding sites in the same way as those above this threshold are, albeit at a lower level. The alternate analysis illustrated in Figure 3.4, shows enrichment at significance levels between 10^{-4} and 10^{-8} and hints at this possibility. However, in this analysis, the apparent enrichment may be driven by

the sum of all SNPs more significant than the level shown and not necessarily by the SNPs at this lower P-value range and so this is something that I will explore further in chapter 6.

Finally, as observed in the case of RCC, disease-susceptibility loci as defined in GWAS analyses are frequently intergenic, a long way from any gene, and this can make identification of the biological effect of the polymorphism difficult. This can be confounded in complex traits, by uncertainty around the cell type in which the disease-associated polymorphism may be acting. By demonstrating that these polymorphisms are specifically targeting the HIF pathway in RCC cells, my work provides a basis for studying the effect of RCC-associated genetic variants by determining their association with HIF-regulated genes in RCC. This forms the basis of chapter 4.

Chapter 4: Identification of target genes of RCC susceptibility loci

4.1 Introduction

In chapters 2 and 3, I have shown a statistically significant overlap between RCC-susceptibility loci and cis-acting components of the HIF pathway as defined in RCC cells. However, as is often observed in GWAS, the majority of the RCC susceptibility loci are intergenic, a long way from any transcribed region. In addition, at some loci, the extended haplotype block overlaps with several annotated genetic regions. In either case, it can be difficult to define which genes are affected by the susceptibility polymorphisms at a given locus and therefore to link genetic susceptibility to biological mechanisms of tumourigenesis. However, the overlap between RCC-susceptibility polymorphisms and the HIF pathway provides a potential mechanism for doing this, since if the HIF-target genes at these loci can be identified, they will provide strong candidates for targets of the polymorphisms.

Again, the transcriptional targets of HIF-binding sites are not always straightforward to identify. As a first approximation, the target gene is often considered to be the closest to the HIF-binding site. However, the regulated gene may be hundreds of kb away from the HIF-binding site and is not necessarily the nearest gene (Schödel et al. 2012; Vernimmen et al. 2011). This is thought to occur through looping of chromatin in a 3-D structure that allows distant elements to contact one another. Nevertheless, it is possible to study gene expression in the presence or absence of the HIF transcription factor in RCC cells and therefore to define HIF-regulated genes relatively easily. These can then in turn be related both to HIF-binding sites and to RCC-susceptibility loci.

Indeed, relating HIF-regulated genes, as well as HIF-binding sites to RCC-susceptibility polymorphisms, is important for a number of reasons. Firstly, it confirms the functional relevance of the HIF-binding sites that overlap RCC-susceptibility loci as identified in chapters 2 and 3. Secondly, it may identify effects on other HIF target genes that are affected by RCC-susceptibility polymorphisms that do not directly overlap the HIF-binding site (e.g. by influencing another enhancer of the same gene). Thirdly, it may identify effects on indirect/downstream targets of HIF that are regulated indirectly by a HIF-target gene rather than by HIF itself.

This chapter primarily utilises a pan-genomic assessment, by RNA-seq, of gene regulation by HIF/VHL axis in 786-O RCC cells, the setting in which overlap between RCC-susceptibility loci and HIF-binding sites was studied, to determine the functional effects of the HIF pathway. However, my findings are then cross-referenced to gene expression patterns in renal tumours themselves using RNA-seq data from the TCGA database. In the early parts of this chapter, I employ a simplified model, using genomic proximity, to associate HIF-binding sites and RCC-susceptibility polymorphisms to gene expression. I then use Capture-C methodologies that measure physical interacts between DNA sequences lying at a distance from each other to better define gene promoters that interact with polymorphic HIF-binding sites and enhancers, at each of the RCC-susceptibility loci. In brief, the Capture-C methodology uses restriction enzyme digestion and proximity ligation followed by deep-sequencing to analyse chromatin interactions in high resolution and high throughput. The methodology is able to detect chromatin looping from a defined viewpoint with distant enhancer elements in the

vicinity of the viewpoint. Finally, I demonstrate the effect of the RCC-susceptibility allele on gene expression in renal tumours, for each of the polymorphic HIF-binding sites associated with RCC, using expression quantitative trait locus (eQTL) analysis on patient data from The Cancer Genome Atlas Kidney Renal Clear Cell Carcinoma (TCGA-KIRC) data collection (Cancer_Genome_Atlas_Research_Network 2013).

4.2 Results

4.2.1 Genes in genomic proximity to RCC susceptibility loci are enriched for genes regulated by the VHL/HIF pathway

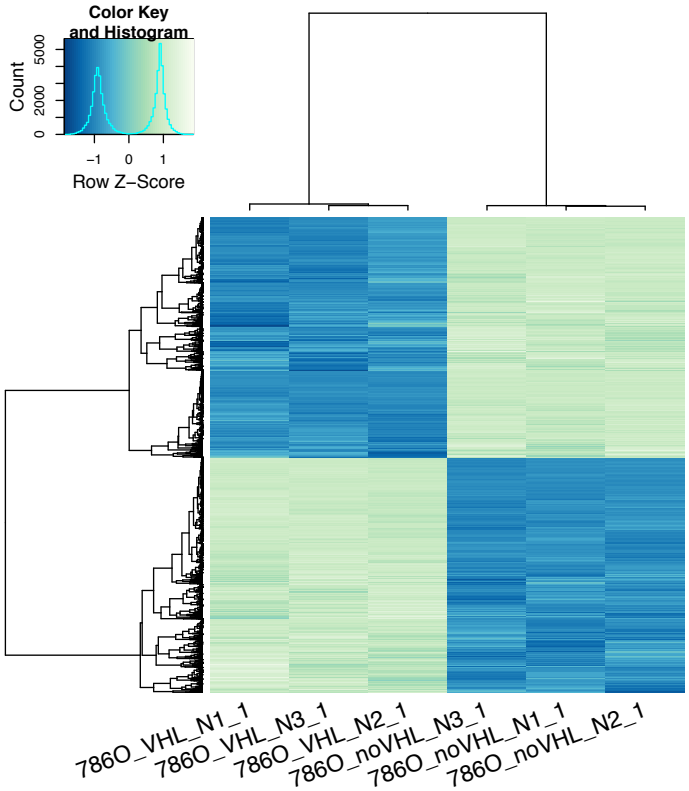
As a first check on the functional relevance of the overlap between RCC-susceptibility loci and binding sites of the HIF transcription factor in RCC cells, I examined for genomic proximity of these loci to HIF-regulated genes in 786-O RCC cells, one of the cell lines used to define the HIF-binding sites. This cell line lacks functional VHL and so does not regulate HIF in response to hypoxia. Instead, HIF is constitutively activated, even in normoxic culture. However, re-introduction of VHL by stable transfection restores normal regulation of HIF, which is almost undetectable in cells cultured in normoxia. I therefore compared gene expression profiles from RNA-seq analysis of polyadenylated transcripts in 786-O cells stably transfected with either wild-type VHL or an empty expression vector to determine genes regulated by HIF in renal cancer cells.

Cell lines were stably transfected and subsequently RNA-seq experiments were performed by one of my co-workers in triplicate. Reads were then mapped as outlined in the methods and feature counts were determined using htseq. Differential gene expression and the statistical significance of these changes was then determined using DEseq.

Hierarchical clustering of the 6 samples based on differentially expressed genes (P-value<0.001) confirmed close correlation between the three replicates performed in each condition (Figure 4.1A). A volcano plot shows -log10(P-value)

against log2(fold change – VHL defective/VHL wt) for each of the 10,990 genes detected in this analysis (Figure 4.1B) confirmed highly statistically significant differences in large numbers of genes. Genes were then ranked according to the recommended combined ranking metric of fold change and significance (-log10sig*log2fc), to test for enrichment amongst VHL-regulated genes using gene set enrichment analysis (GSEA) (Subramanian et al. 2005; Xiao et al. 2014).

A


Color Key and Histogram
Count
0
2000
5000
−1
0
1
Row Z−Score
786O_VHL_N1_1
786O_VHL_N3_1
786O_VHL_N2_1
786O_noVHL_N3_1
786O_noVHL_N1_1
786O_noVHL_N2_1

B

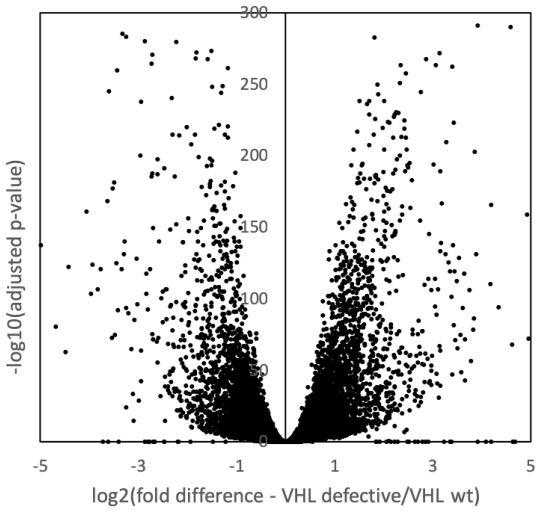

300
250
200
150
100
50
0
-log10(adjusted p-value)
-5
-3
-1
1
3
5
log2(fold difference - VHL defective/VHL wt)

Figure 4.1: Polyadenylated RNA-seq analysis of 786-O cells stably transfected with wild type VHL or empty vector. (A) Heatmap showing row normalised RNA-seq expression analysis of differentially expressed genes. Hierarchical clustering based on correlation coefficient confirms close correlation between the three replicates for each condition. (B) Volcano plot for RNA-seq analysis of VHL/HIF regulated genes in 786-O cells. X-axis denotes the log2(fold change – VHL defective/VHL wt) and Y-axis denotes -log10(adjusted P-value) for all 10,990 genes detected.

To determine whether genes neighbouring RCC-susceptibility loci were enriched for HIF regulated genes, I next identified the three closest genes to the index SNP at each locus by reference to the Ensembl GRCh37 (hg19) reference annotation. This is an arbitrary choice designed to include at least a proportion of genes that are targeting by HIF at these loci, since the RCC-susceptibility polymorphism may not be affecting one of the three closest genes, but the inclusion of too many genes would add in noise and dilute the gene set enrichment analysis. This set of genes is listed in Table 4.1.

This set of genes was then tested for enrichment amongst genes regulated by HIF in renal cancers cells, using the pre-ranked list from the RNA-seq data above, and the Gene Set Enrichment Analysis (GSEA) package from the Broad Institute (Figure 4.2).

To identify genes regulated by the VHL/HIF pathway, poly-adenylated RNA-seq analysis of 786-O cells stably transfected either with wild-type VHL or with control vector was performed. Genes were ranked genes according to their regulation differential expression in VHL deficient versus VHL competent cells by the VHL/HIF pathway, defined as a combined metric of fold change and significance (-log10sig*log2fc).

locus	*index SNP*	*P-values*	*3 closest genes to index SNP*
11q13.3	rs11263654	1.65E-23	CCND1, MYEOV, ORAOV1
2p21	rs2121267	6.95E-19	EPAS1, PRKCE, TMEM247
12p12.1	rs11534749	6.19E-17	BHLHE41, ITPR2, SSPN
14q24.2	rs28840762	1.08E-15	DPF3, DCAF4, ZFYVE1
12q24.31	rs10846748	1.70E-12	SCARB1, DHX3, UBC
8q24.21	rs6470588	2.82E-12	MYC, POU5F1B, TMEM75
11q22.3	rs117706999	6.60E-10	EXPH5, C11orf65, KDELC2
2q22.3	rs11888238	8.15E-10	ZEB2, ARHGAP15, GTDC1
1p32.3	rs6676515	3.04E-09	FAF1, DMRTA2, ELAVL4
4q23	rs7697932	4.70E-09	DDIT4L, DNAJB14, H2AFZ
3p22.1	rs9821249	1.12E-08	RPL14, ZNF619, ZNF620
15q22.31	rs12905354	2.28E-08	SNAPC5, ZWILCH, RPL4
3q26.31	rs234043	4.10E-08	NCEH1, TNFSF10, AC007919.2

Table 4.1: The 3 closest genes to each RCC susceptibility loci. Chromosomal location, P-value and 3 closest genes to each of the RCC susceptibility loci, as defined in chapter 3.

This GSEA showed significant enrichment (nominal P-value =0.022) for the 39-genes that neighboured the RCC-susceptibility loci amongst genes upregulated by HIF in the 786-O renal cancer cell line (listed in table 4.1). Previous work has shown that direct transcriptional targets of HIF (i.e. those associated with direct HIF binding) are induced, but not repressed by HIF-binding (Schoedel et al. 2011). Therefore, this result suggests that genes neighbouring RCC-susceptibility loci are enriched at direct transcriptional targets of HIF. This is consistent with the enrichment for HIF-binding sites observed in chapters 2 and 3.

Conversely, if RCC-susceptibility loci were enriched at both direct and indirect targets of HIF, I would have expected to see some enrichment at downregulated as well as up-regulated genes, since overall these constitute relatively equal numbers in the RNA-seq analysis. However, no enrichment of RCC-susceptibility neighbouring genes was seen amongst genes down-regulated by HIF.

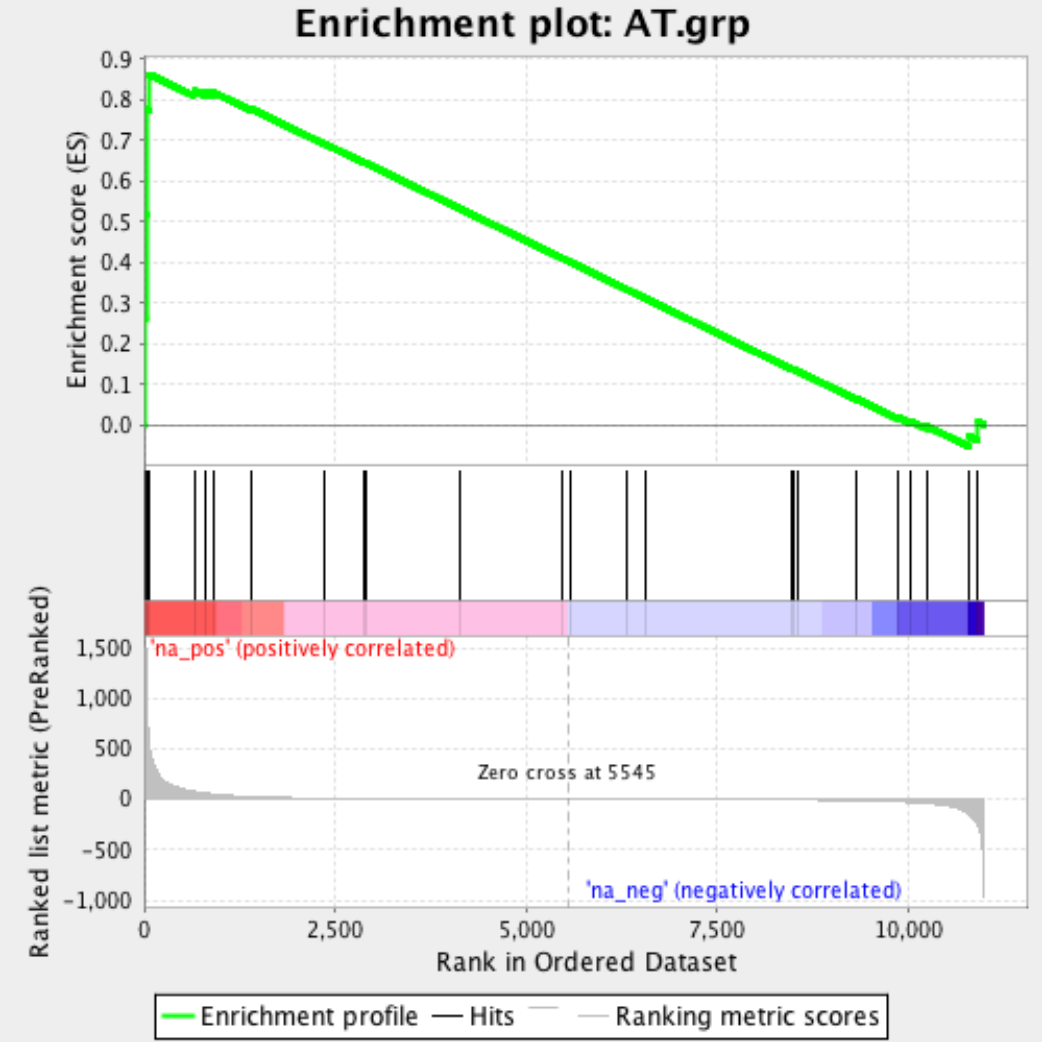


Figure 4.2: Enrichment of RCC-susceptibility loci in close proximity with HIF regulated genes. The 3 closest genes to the index SNPs at each of the 13 genome-wide significant RCC susceptibility loci (P-value ≤ $5*10^{-8}$) were identified. Enrichment for regulation by VHL within the 39-gene set was shown by weighted, pre-ranked GSEA (nominal P-value=0.022). Differential expression by HIF was assessed between 786-O cells stably transfected either with wild-type VHL or with a control vector. Genes were ranked using a combined metric of log fold change and significance level (-log10sig*log2fc).

4.2.2 Identification of distant sites that physically interact with regulatory elements at RCC susceptibility loci by Capture-C

Distant enhancers and regulatory elements control transcription through physical interaction with the promoters of their target genes, which is brought about by 3-dimensional looping of the linear DNA molecule. Therefore, rather than relying on genomic proximity to define the target genes (as above), I next examined for interactions between the HIF-binding sites and other regulatory elements, at each RCC-susceptibility locus, and the promoters of annotated genes (using Ensembl hg19 annotation).

HIF-binding sites, regulatory elements and RCC-susceptibility loci were first defined as in chapter 3. At each locus, SNPs significantly associated (P-value ≤ $5*10^{-8}$) with RCC or in high linkage disequilibrium ($r^2 \geq 0.8$) with the index SNP were then identified. HIF binding sites were mapped by ChIP-seq in 786-O and RCC4 cell lines and as previously binding sites in at least 2 out of 5 data sets considered for this analysis (786-O HIF-1β and HIF-2α, RCC4 HIF-1β, HIF-1α and HIF-2α). The regulatory elements consist of H3K4me1, H3K4me3 and H3K27ac chromatin marks mapped in the 786-O kidney cancer cell line. All HIF-binding sites and regulatory elements that overlay one or more of these SNPs were then determined. The locus on 2q22.3 does not overlap any HIF binding site or regulatory element and so was not studied further.

A Capture-C assay (see methods 9.1.3) was then used to identify distant regions of the genome that physically interacted with each of this defined set of sites. The Capture-C assay is a modified form of a chromatin conformation capture (3C) assay coupled to high-throughput sequencing that allows distant interactions with multiple sites of interest (viewpoint sites) to be identified in parallel. It uses a restriction enzyme (Dpn2), which recognises a 4-bp motif that occurs frequently in the genome, thereby giving it a relatively high resolution for this type of methodology. Biotinylated oligonucleotides complementary to each viewpoint site are used to enrich the signal by formaldehyde crosslinking the interacting chromatin to the capture site. In order to design these oligonucleotides, I used the CapSequm tool developed by Hughes et al (Davies et al. 2016). Essentially, this first identifies the Dpn2 restriction fragment that encompasses the point of maximum ChIP-seq signal at each site. Repetitive sequences within these Dpn2 fragments are then excluded using Repeatmasker. Next, multiple 120 bp oligonucleotides complementary to each end of the dpn2 fragment are designed. Finally, to ensure that the designed oligonucleotides align uniquely within the genome, they are assessed and filtered using a BLAT based quantification of repetitiveness.

Using these oligonucleotides, Capture-C experiments were then performed in duplicate by Veronique Lafleur in 786-O renal cancer cell lines. Paired-end sequencing was performed and read-pairs, where one end mapped to a viewpoint Dpn2 fragment, were extracted. The opposite end of each extracted read-pair was then mapped to the genome and used to generate bigwig files for display (Figure 4.3).

Using these data, Dpn2 fragments that interacted with viewpoint sites at frequencies significantly above “background” level were identified by Ran Li using a method developed by a previous member of the group Rafik Salama (Platt et al. 2016). Essentially, this employed a negative binomial model to estimate the expected “background” read count as a function of the distance from the viewpoint oligo and used this to calculate the statistical significance of the observed signal for the set of reads generated from a given viewpoint. This method estimated high levels of “background” for regions within 25-kb of the viewpoint site. However, while this signal does not necessarily represent discrete looping of distant sites with the viewpoint, it does still represent physical proximity and therefore the potential for functional interaction. For this reason, gene promoters and regulatory elements lying within 25-kb of a viewpoint site were included as well as more sites for which statistically significant interactions were identified. The method has limitations to identify interacting regions, including determining interaction points close to the viewpoint site and interactions of low frequency. Whilst for either there are potential experimental challenges contributing, relaying on the totality of all capture sites to construct the null model might be an additional limiting factor. Alternative approaches to define interacting site have been now been described and could be explored (Servant et al. 2015, Cairns et al. 2016).

Figure 4.3 A: 3q26.31 locus

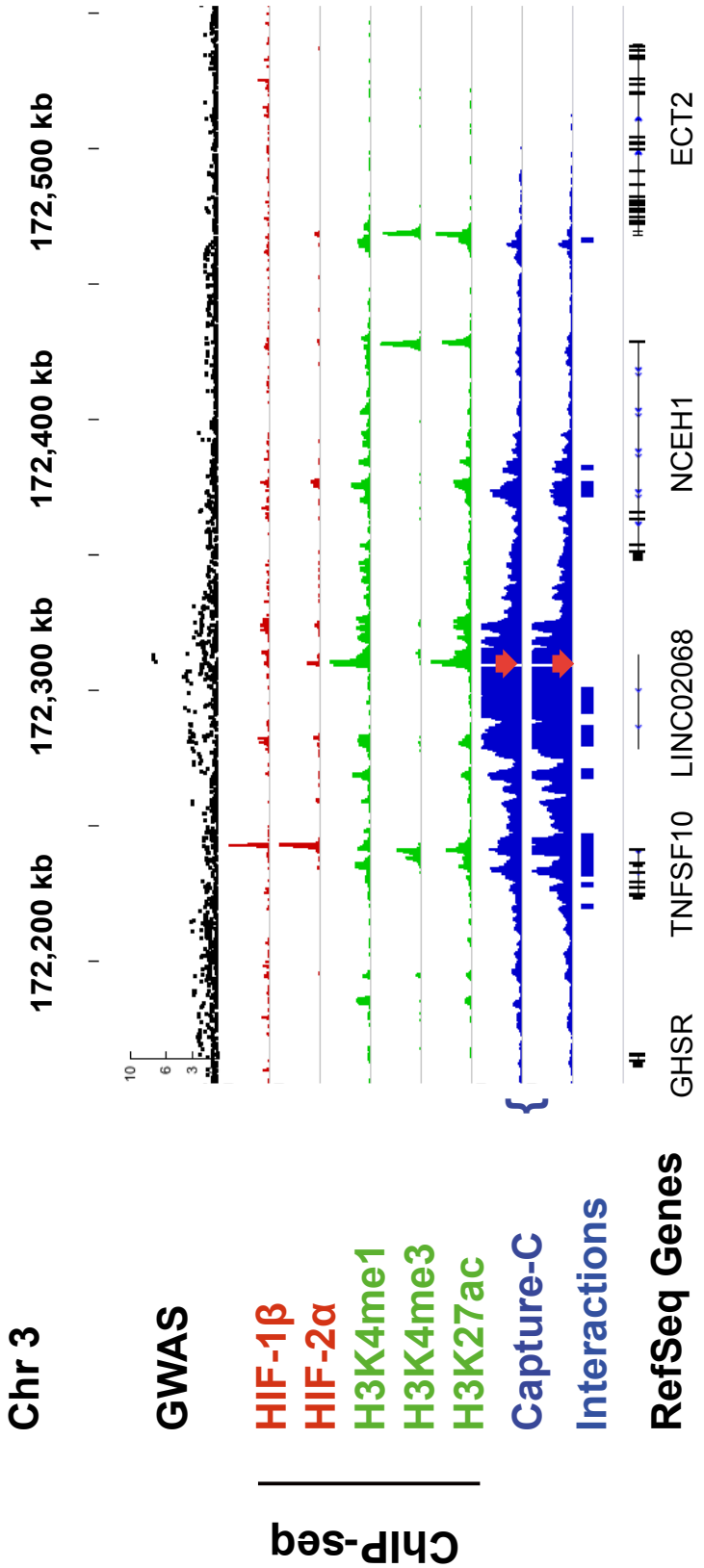


Figure 4.3 B: 12p12.1 locus

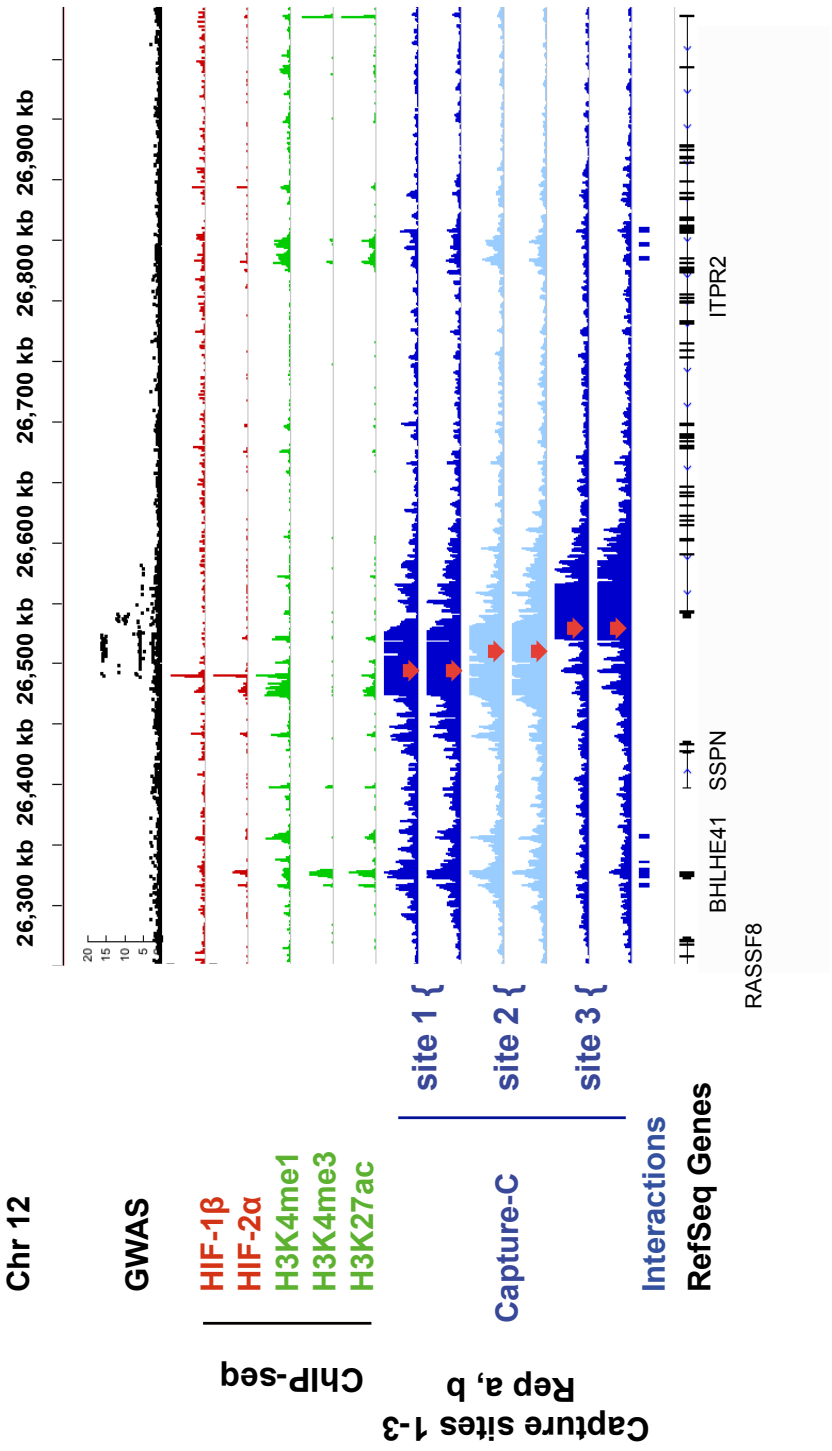


Figure 4.3: Integrative Genomics Viewer (IGV) tracks of the 3q26.31 (A) and 12p12.1 (B) RCC susceptibility loci identified by a recent meta-analysis of RCC GWAS. Track 1 displays GWAS SNP-level P-values from the RCC GWAS meta-analysis. Tracks 2-3 (in red) display binding of HIF-2α and HIF1-β mapped in 786-O cancer cell lines of RCC. Tracks 4-6 from top (in green) display ChIP-seq signal (normalised read counts per million total reads at each chromosomal position) for histone H3K4me1, H3K4me3 and H3K27ac chromatin marks mapped in the 786-O kidney cancer cell line. Tracks in blue show Capture-C data with assays performed in 786-O cells using different "viewpoint" (bait site) oligonucleotides at each active enhancer that coincided with the linked SNPs at the susceptibility locus ($r^2 \geq 0.8$ with the index SNP) and are indicated by the red arrows. Regions with Capture-C interaction (blue bars) and in black gene annotation from the RefSeq hg19 (GRCh37) build is displayed are displayed in the bottom 2 tracks. HIF binding data of RCC4 used to determine the set of high confidence HIF binding sites are not shown here due to space restrictions.

4.2.3 Genes with promoters physically interacting with regulatory elements overlapping RCC susceptibility loci are enriched for HIF regulable genes

Having defined a set of sites that physically interact with regulatory elements at each of the 13 RCC-susceptibility loci identified in chapter 3, I next determined which of these sites corresponded to gene promoters. Promoters were defined as the transcription start site +/- 2 kb with reference to the genome annotation corresponding to the hg19 assembly annotation provided by Ensembl and genes with promoters overlapping the set of interacting regions identified. Additionally, Capture-C can not detect interactions in close proximity to the viewpoint oligo. I hence identified further all genes which fall within 25 kB from the RCC loci associated LD regions ($r^2 \geq 0.8$ with reference to the 1000 Genomes EUR population data). A total of 35 gene promoters overlapped sites that physically interact with regulatory elements at RCC-susceptibility loci or overlapped the LD region of the locus itself (Table 4.2).

locus	rs number	p-value	proximity to HIF BS	genes in proximity	regulation by VHL
11q13.3	rs11263654	1.65E-23	overlap	CCND1 MYEOV ORAOV1	↑ n.e. ↓
2p21	rs2121267	6.95E-19		EPAS1	↓
12p12.1	rs11534749	6.19E-17	overlap	BHLHE41 ITPR2 SSPN	- ↑ n.e.
14q24.2	rs28840762	1.08E-15		DPF3	n.e.
12q24.31	rs10846748	1.70E-12	< 25kb	SCARB1	↑
8q24.21	rs6470588	2.82E-12	overlap	MYC	↑
11q22.3	rs117706999	6.60E-10		DDX10 EXPH5 ATM NPAT C11orf65 KDELC2	↑ n.e.. ↑ - n.e. ↓
2q22.3	rs11888238	8.15E-10		ZEB2	↑
1p32.3	rs6676515	3.04E-09		FAF1 DMRTA2	- n.e.
4q23	rs7697932	4.70E-09	< 25kb		
3p22.1	rs9821249	1.12E-08		EIF1B ENTPD3 RPL14 ZNF619 ZNF620 ZNF621	↓ n.e. ↓ ↑ n.e. ↑
15q22.31	rs12905354	2.28E-08	< 25kb	DIS3L MAP2K1 MEGF11 SMAD6 SNAPC5 TIPIN ZWILCH RPL4	- ↑ n.e. - ↓ ↑ ↑ -
3q26.31	rs234043	4.10E-08	overlap	ECT2 TNFSF10	↑ ↑

Table 4.2: Promoters of genes looping to RCC susceptibility loci. The table summarises the genetic locus, index SNPs and index SNP P-value of all independent genome-wide significant (P-value ≤ $5*10^{-8}$) RCC-susceptibility loci alongside their proximity to HIF- binding sites. Further, genes either directly overlapping the locus or whose promoters interact with the loci are listed. Differential regulation between VHL-deficient and VHL-competent cells 786-O is indicated in the last column (↑= upregulated in VHL-deficient cells, ↓ = downregulated in VHL-deficient cells, - = not regulated, n.e. = not expressed).

Of these 35 genes, 23 had been identified in the previous approach, which used the three closest genes in genomic proximity to the index SNP at each RCC-susceptibility locus (Figure 4.4). However, 12 genes were not identified in the earlier more simplistic approach. Comparison of the two gene sets suggests that detailed analysis of chromatin looping at susceptibility loci offers a refined approach to target gene identification.

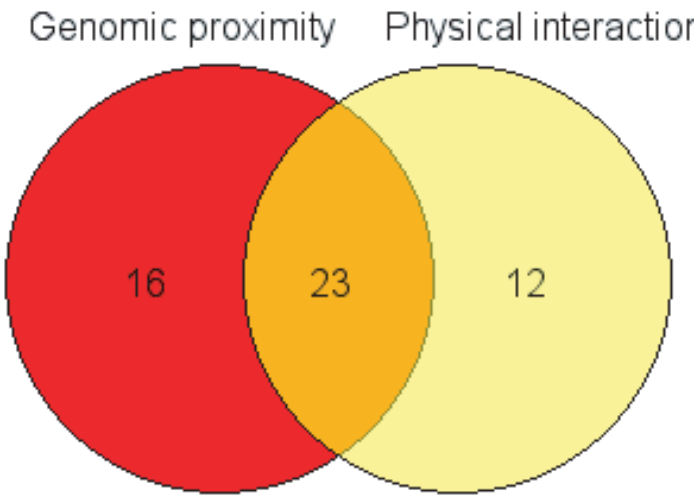


Figure 4.4: Venn diagram showing the number of genes associated with RCC-susceptibility loci that were identified using the genomic proximity (nearest three) and physical interaction (Capture-C) approaches.

This new set of 35 genes that physically interacted with regulatory elements at the RCC-susceptibility loci was then tested for enrichment amongst genes regulated by HIF in renal cancers cells using the pre-ranked list from the RNA-seq, and the Gene Set Enrichment Analysis (GSEA) package from the Broad Institute as for genes in genomic proximity (Figure 4.5).

Gene Set Enrichment Analysis (GSEA) of this the 35-gene set showed enrichment for this gene set amongst genes upregulated in VHL defective cells (Figure 4.5). Again, no enrichment was observed amongst genes downregulated in the presence of HIF. Importantly, HIF-inducible genes were identified at all 4 loci (11q13.3, 12p12.1, 8q24.21 and 3q26.31) that directly overlapped with a HIF-binding site (Table 4.2).

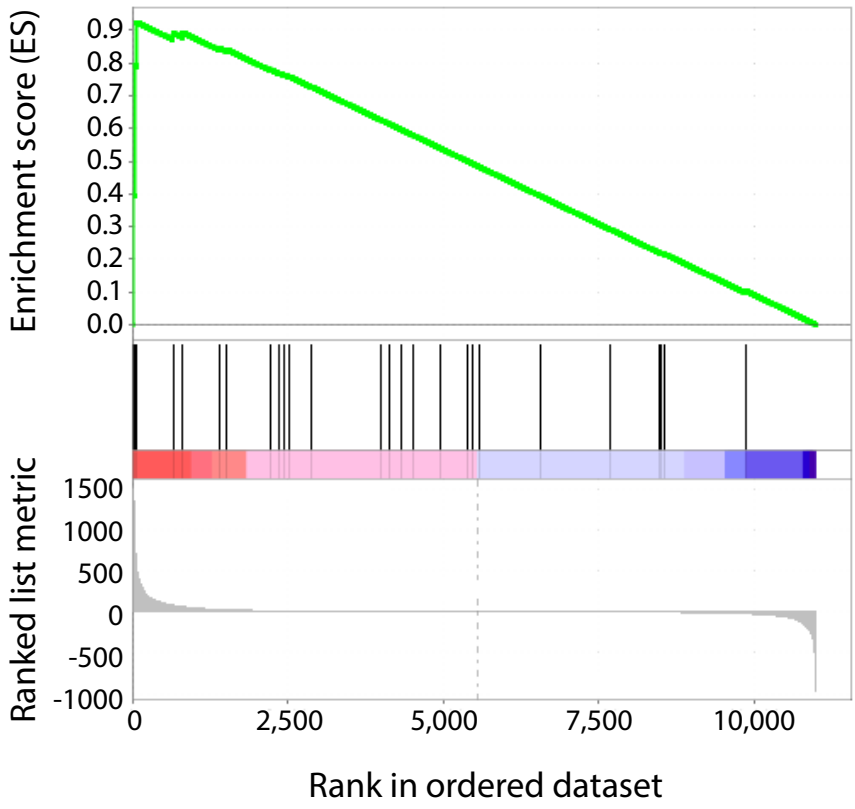


Figure 4.5: Genes that are physically interacting with RCC-susceptibility loci are enriched amongst genes regulated by the HIF/VHL pathway. The gene set enrichment analysis (GSEA) was performed with reference to RNA-seq analysis (n = 3) of 786-O (VHL-defective) cells and 786-O cells re-transfected with wild-type VHL. Genes were ranked according to their regulation (combined fold-change and P-value) by VHL (x-axis). The set tested consisted of genes with lying within 25-kb of each RCC-susceptibility locus and distant genes with promoters detected to loop to these loci as identified by Capture-C analyses in 786-O cells. The vertical bars represent the individual genes analysed in the weighted gene set enrichment analysis (green line). The association tested

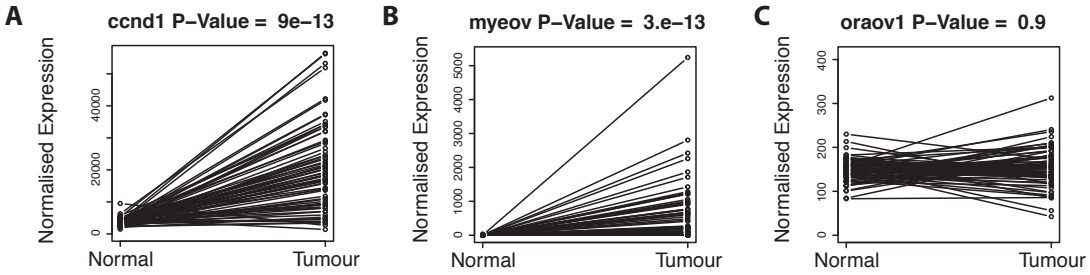
showed significant enrichment of the gene set amongst those upregulated in VHL-defective cells (P-value = 0.017).

4.2.4 Induction of HIF-target genes in ccRCC samples from the TCGA KIRC cohort

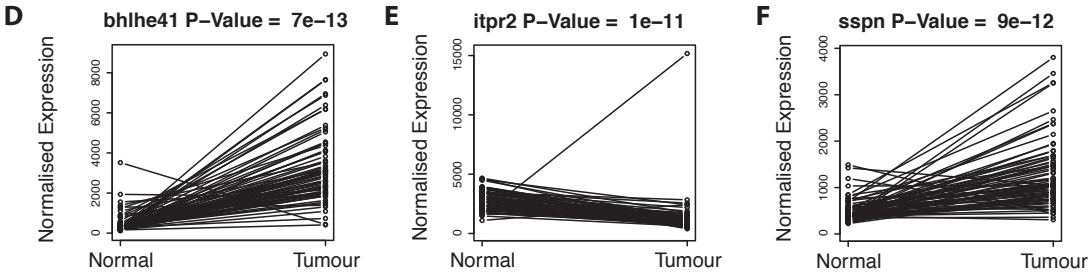
I was particularly interested to examine regulation of HIF-inducible genes identified at all 4 HIF binding site overlapping loci (11q13.3, 12p12.1, 8q24.21 and 3q26.31) in renal cancer more generally, rather than just in 786-O cells. Hence, I next examined their expression in renal tumours themselves. Since these clear cell renal cancers mostly lack functional VHL, they generally express constitutively high levels of HIF protein and HIF target gene mRNA. While it is not possible to compare renal cancer cells with VHL re-complemented cells in these tumours it is possible to compare expression levels for these genes in the renal tumours with expression in the VHL wild-type normal kidney surrounding the tumour.

To do this I used paired RNA-seq analyses of gene expression of the 72 cases in the TCGA clear cell renal cancer (KIRC) cohort for which expression data of clear cell renal cancers tumours together with from matched normal kidney from the same individuals was available at the time of analysis. The cases were not stratified by tumour stage. Gene expression FPKM) was normalised to total reads for each sample and transcript size (Figure 4.6). Differential expression was tested using a paired Mann-Whitney-Wilcoxon Test.

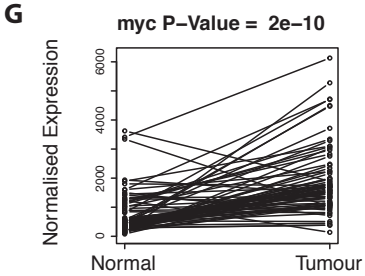
11q13.3



12p12.1



8q24.21



3q26.31

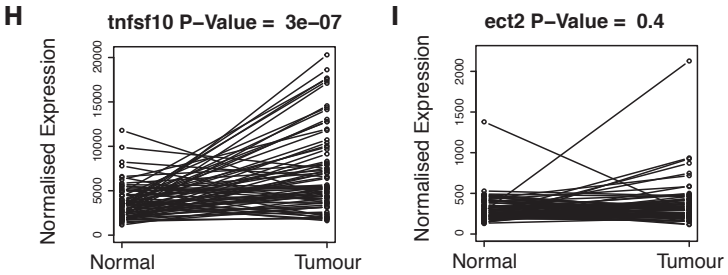


Figure 4.6: Gene expression in ccRCC compared to normal kidney. Gene expression in 72 ccRCC tumours and matched surrounding normal renal tissue was assessed using RNA-seq data from the TCGA KIRC cohort. Expression was normalised to the number of reads in each sample and the transcript length. Differential expression was assessed using a Mann-Whitney-Wilcoxon Test.

To a large extent, genes induced by the presence of HIF in 786-O RCC cells were also increased in ccRCC tumours compared to VHL-competent normal renal tissue. However, several genes (MYEOV and SSPN) that were not detected in 786-O cells were found to be activated in ccRCC tumours. In addition, BHLHE41, which has previously been shown to be specifically induced by HIF-1 was not induced in 786-O cells, which do not express functional HIF-1 but was induced in ccRCC tumours more generally (Grampp et al. 2017). Conversely, ORAOV1, ITPR2 and ECT2 were not activated in ccRCC tumours compared to normal kidney. Nevertheless, at each of the 11q13.3, 12p12.1 and 8q24.21 loci, the previously reported transcriptional targets of the polymorphic HIF-binding site, associated with kidney cancer, were shown to be induced in ccRCC tumours. Likewise, at the 3q26.31 locus, consistent with its increased expression in the VHL-defective 786-O cell line expression of TNFSF10 was increased in tumour tissues. It is important to note that this analysis does not control for alterations in expression potentially arising from somatic copy number alterations or alternative genetic or epigenetic alterations occurring in these patients.

4.2.5 The effect of genotype at RCC-susceptibility loci on the expression of direct HIF-target genes

My work to this point has shown a statistically significant association between renal cancer-susceptibility loci and both HIF-binding sites and HIF-induced genes. At 4 loci, RCC-susceptibility polymorphisms directly overlapped a HIF-binding site and each of these interacted with the promoter of at least one HIF-regulated gene. Using allele-specific assays of HIF-binding in cell lines that were heterozygous for these polymorphisms, co-workers have been able to show an effect of the genotype on HIF-binding at 3 of these loci, suggesting that the polymorphisms might affect the expression of one of more HIF target genes at that locus by such a mechanism (Schödel et al. 2012; Grampp et al. 2016; Grampp et al. 2017). To assess evidence for this I examined for association between these polymorphisms and the level of expression of transcripts at the relevant loci by reference to transcript expression data from the TCGA KIRC cohort.

The TCGA KIRC cohort contains RNA-seq expression profiling data for ccRCCs from 534 individuals of whom 452 also have germline SNP-genotyping data, using the Affymetrix 6 array, available. Using the 1000 Genomes data, I was able to identify the genotyped SNP that was most closely linked to the SNP at the HIF-binding site (highest r^2 value) at each of the 4 RCC-susceptibility loci that directly overlapped a HIF-binding site. I then extracted genotype information for these SNPs, together with gene expression data for the HIF-inducible genes whose promoters interacted with these sites. The expression level for each gene was then plotted as a box-and-whisker plot for each genotype at the typed SNP

(Figure 4.7). To determine the statistical significance, the expression level is first fitted to a negative binomial generalized linear model (GLM) against the genotype status of the patients at the locus of interest. The significance of the eQTL is then determined using a Chi-Square test on the likelihood ratio of the genotype-based GLM against a null hypothesis that ignores the genotype.

Germline polymorphisms at RCC-associated SNPs overlapping HIF-binding sites at the 12p12.1 and 8q24.21 loci had statistically significant associations with the expression of BHLHE41 and MYC expression in renal tumours respectively. Expression levels of other genes at the 12p12.1 locus whose promoters physically interacted with this site did not show an association with the genotype at the HIF-binding site. Although polymorphisms at the 11q13.3 locus appeared to affect the expression of CCND1, this did not reach statistical significance. Potentially this might be a result of the low minor allele frequency (0.07) at this locus, which would result in small numbers of individuals with this allele. At the 3q26.31 locus, neither TNFSF10 nor ECT2 mRNA levels correlated with the genotype. However, for all genes, expression levels were determined using RNA-seq data from advanced tumours and it is possible the genotype might affect expression at an earlier stage in tumourigenesis.

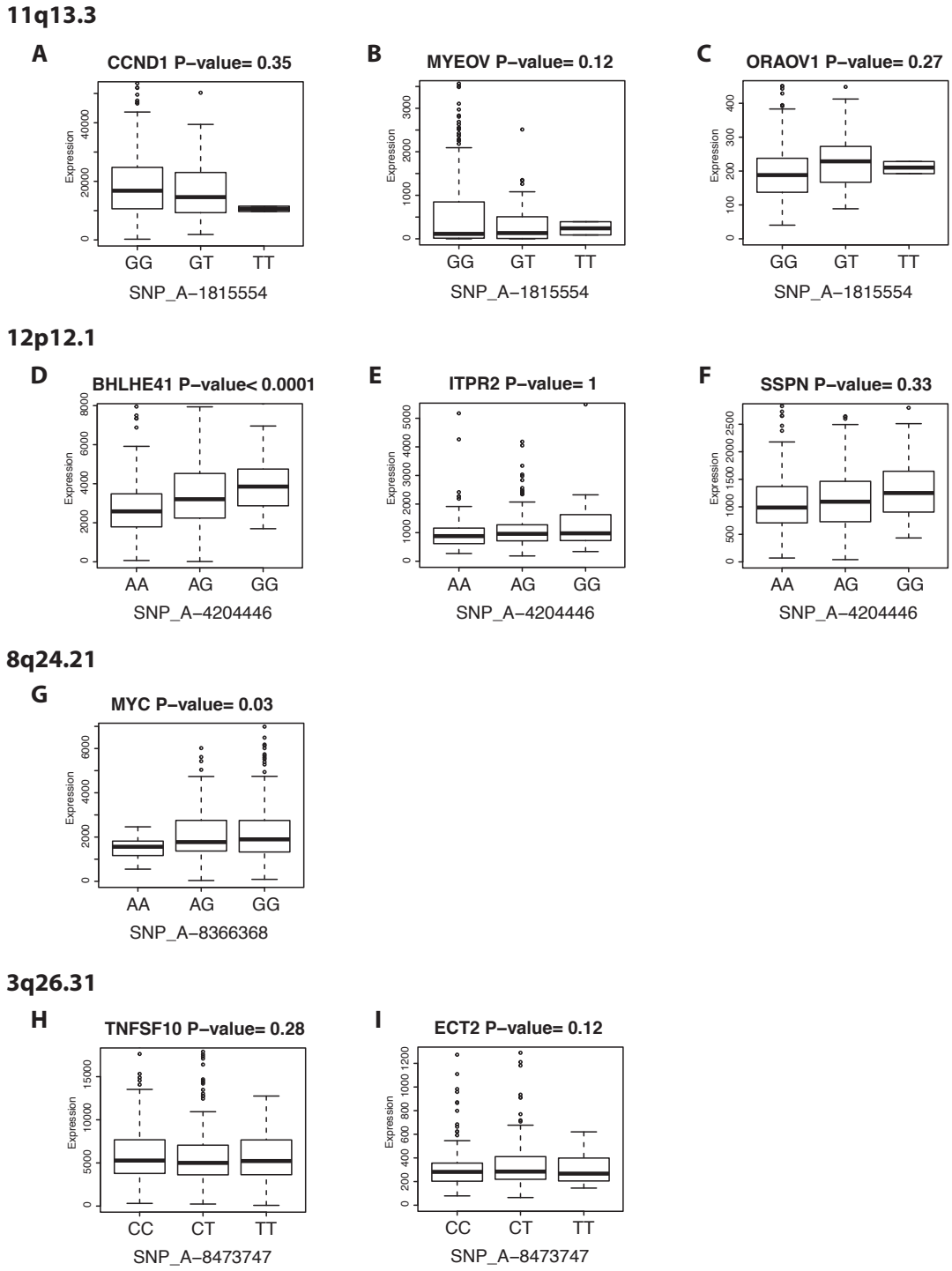


Figure 4.7: Box-and-whisker plots showing eQTL analysis for the 4 RCC susceptibility loci that directly overlap a HIF-binding site.

4.2.6 The association between RCC-associated polymorphism at the 12p12.1 locus and expression of BHLHE41 is specific to tumours with high levels of HIF-1

Taken together, the analysis above provides strong evidence that germline RCC-susceptibility polymorphisms alter the expression of HIF transcriptional targets in the renal cancers themselves - in each case the RCC-susceptibility allele acts to increase expression of the gene. However, it is possible that these polymorphisms alter expression of HIF target genes through mechanisms that do not involve HIF itself. I therefore examined the dependence (or otherwise), of the association between the genotype and gene expression, on HIF expression levels. Given the relatively small numbers of cases and the weak statistical associations, such sub-group analysis was only possible for the association between the polymorphism at 12p12.1 and the expression of BHLHE41, which showed the strongest association.

BHLHE41 is specifically activated by the HIF-1α isoform of HIF in renal cancer cell lines. Whilst it is possible to determine mRNA levels for HIF-1α in the TCGA KIRC dataset, these correlate poorly with HIF-1α protein levels and transcriptional activity. This is due, at least in part, to the presence of a HIF-1α antisense transcript, which is itself induced by HIF and suppresses expression of the sense transcript (Uchida et al. 2004). Therefore, as a surrogate, I used the expression of a previously defined HIF-1 metagene to first classify tumours in the KIRC cohort according to HIF-1 activity. Tumours were divided into tertiles according to HIF-1 metagene expression and the association between genotype at this locus and BHLHE41 examined in the upper and lower tertiles as above

(Figure 4.8). In tumours with high expression of the HIF-1 metagene, a statistically significant association between the genotype and gene expression was again observed (even though the number of cases was much reduced). Conversely, in tumours with low HIF-1 metagene expression, the association between the genotype and gene expression was lost (even though the number of cases in the upper and lower tertiles was the same). This dependence of the association on the activity of HIF-1 is strong evidence that the effect of the genotype on gene expression is mediated through HIF.

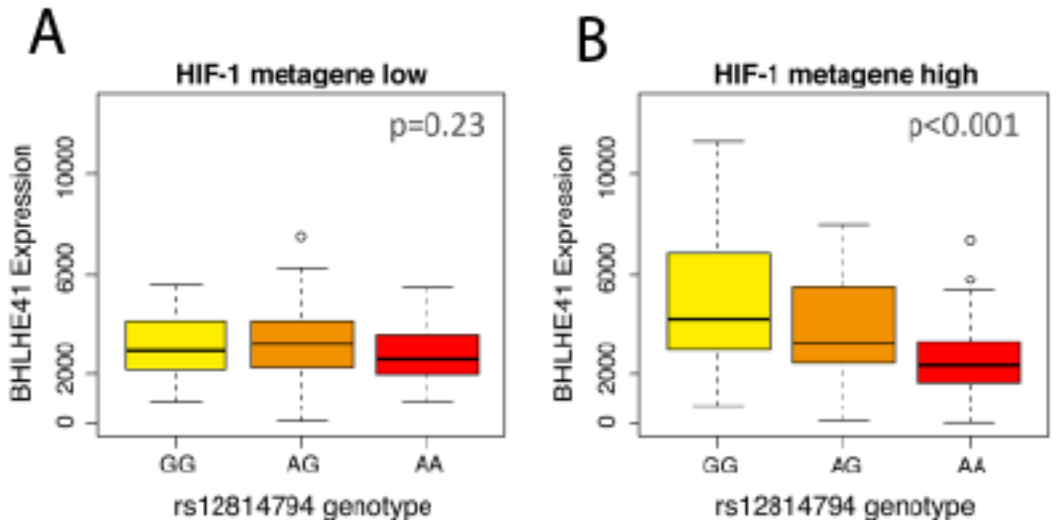


Figure 4.8: Meta-gene eQTL of the genotype at rs12814794 and mRNA expression levels of BHLHE41 based on the TCGA analysis of ccRCC tumors (KIRC) data. Tumours were separated into tertiles by their respective expression levels of previously defined HIF-1 or HIF-2 metagenes. A) lower tertile for HIF-1 metagene expression, B) upper tertile for HIF-1 metagene expression. χ2-test assessed the significance of the difference in expression between carriers of a GG allele compared to carriers of the AA allele. This figure is adapted from Grampp, Schmid et al. (Grampp et al. 2017).

4.3 Discussion

Chapter 4 is focused on identification of target genes of RCC susceptibility loci. Long-range physical interactions were identified between the HIF-binding site overlapping with RCC susceptibility loci and the promoters of HIF-induced genes, at all 4 RCC-susceptibility loci that directly overlap a HIF-binding site.

Many of these genes have well-defined biological roles that are important to propagate growth of cancer cells. For example, the HIF-binding site at the 11q13.3 locus interacts with the promoter of the CCND1 gene that encodes for cyclin D1. This is a key cell-cycle protein that is important in the progression of cells from the G1-phase to the S-phase of the cell cycle. At the 8q24.21 locus, the HIF-binding site interacts with and activates the MYC proto-oncogene. Myc forms a heterodimeric transcription factor with Max that activates gene programs involved in cell-cycle progression, apoptosis and cellular transformation. The typical MYC gene amplification seen in many cancer types is not observed in ccRCC, perhaps because it is transcriptionally activated by HIF instead. HIF-binding sites at the 12p12.1 locus interact with the promoter of BHLHE41 another basic helix-loop-helix DNA-binding protein (the same class as HIF and Myc) that acts as a transcriptional repressor with possible roles in circadian rhythm and cellular proliferation. At the 3q26.31 locus, TNFSF10 is a member of the TNF family of cytokines that induces apoptosis.

In general, GWAS defined susceptibility loci are largely intergenic, a long way from the nearest gene and this can make linking susceptibility loci to gene

function rather difficult. Conventionally, many works have simply screened genes neighbouring the loci for those with known or plausible roles in tumourigenesis. However, this approach has limited potential to discover new disease-related pathways, since it is predicated on known gene functions. Other strategies have included examining for neighbouring genes whose expression level is associated with the disease-susceptibility genotype at the locus (i.e. for eQTLs). However, effects on expression can be subtle requiring large numbers of patients (particularly when the minor allele-frequency is low), may be specific to particular cell-types (which may be difficult to determine for complex disease) or may depend on the stage in development of the disease. Therefore, whilst the presence of an eQTL may be helpful in linking disease-susceptibility loci to genes, the absence of an eQTL is harder to interpret.

Examining for gene promoters that interact with disease-susceptibility loci provides an alternative paradigm for identifying the functional targets of loci identified in GWAS analyses. However, GWAS defined susceptibility loci are generally too large and diffuse to interrogate in such assays. Nevertheless, this limitation can be largely overcome, by focussing on enhancers and other regulatory elements that overlap the loci, which generally provide much more discrete and functionally relevant regions from which to examine long-range chromatin interactions. This approach can be particularly powerful when these sites are enriched for a functionally relevant transcription factor, as in the case of RCC, since it allows these physical interactions to be cross-referenced to transcriptional regulation of the gene.

For example in this study, since in chapters 2 and 3 I demonstrated a statistically significant overlap between RCC-susceptibility loci and binding sites of the HIF transcription factor, which is constitutively activated in kidney cancer cells as a result of inactivation of the VHL tumour suppressor the functional correlate of regulation by the hypoxia pathway can be examined. In this chapter, I show that RCC-susceptibility loci are also enriched for both genomic and physical proximity to HIF-regulated genes confirming the functional relevance of this association. HIF-binding sites and other regulatory elements at these loci form long-distance interactions (loops) with the promoters of HIF-target genes at these loci.

When HIF is activated, roughly equal numbers of genes are induced or suppressed. However, RCC-susceptibility loci are specifically associated with genes that are induced by HIF, but not with suppressed genes. Induction is a hallmark of direct transcriptional regulation by HIF, which acts as an enhancer but not as a repressor. This suggests that RCC-susceptibility loci are targeting direct transcriptional targets of HIF rather than indirect targets that might be mediated through intermediate factors. Indeed, all 4 loci directly overlapping a HIF-binding site are associated with a HIF-induced gene. However, other loci that do not directly overlap a HIF-binding site are also associated with HIF-induced genes. One possible explanation for this is that these genes are driven by binding of HIF to a nearby enhancer, while the polymorphism affects a different enhancer that co-regulates the same gene. I will examine this possibility further in chapter 5. As well as HIF-dependent induction in the 786-O RCC cell line, expression of

these genes is also enhanced in ccRCC tumours, in which HIF is constitutively activated, providing independent corroboration of their importance in RCC.

Overall, formal proof that the potential target genes play a role in the development of ccRCC is much harder to obtain. Simple deletion or overexpression of the genes may have profound effects on cellular function that do not represent those of more subtle HIF-dependent induction of the genes. Modern cell-based editing technologies now allow the HRE-recognition motif at each binding site to be disrupted and would facilitate the effects of HIF-dependent induction of individual genes to be studied. However, cell-based models of tumourigenesis generally examine late-stage properties of tumours such as proliferation, invasion, migration, metastasis etc. and may not accurately reflect tumour initiation.

Chapter 5: Additional mechanisms by which RCC susceptibility loci alter the HIF pathway in kidney cancer

5.1 Introduction

In the previous chapter, I have shown that RCC-susceptibility loci are enriched for both genomic proximity and physical interaction with genes regulated by the VHL/HIF pathway. Importantly, enrichment was observed for upregulated, but not downregulated genes, which is consistent with these being direct, rather than indirect (mediated through an intermediate factor) transcriptional targets of HIF. However, direct overlap with a HIF-binding site was observed at only 4 loci. One potential explanation for this is that the RCC-associated polymorphisms at other loci might alter the expression of direct transcriptional targets of HIF through mechanisms that do not affect the HIF-binding site itself.

Previous work by the ENCODE Consortium and others has indicated that on average each individual gene is regulated by an average of 5-6 distinct enhancer and repressor regions (Hughes et al. 2013; Encode_Project_Consortium et al. 2012). These may be hundreds of kb distant from the gene promoter and from each other. They influence gene transcription co-operatively, by forming physical interactions, over long genomic distances, with the gene promoter and with each other, through chromatin looping. Disruption of transcription factor binding at any one of these sites in RCC cells could affect gene expression and could have effects on HIF target gene expression alongside those directly mediated by modulation of HIF binding itself. Furthermore, transcription factors binding at these remote enhancers will be in physical proximity with the HIF-binding site within the 3-dimensional transcriptional complex. Therefore, it is possible that they may themselves interact with HIF and strengthen HIF binding at remote genomic sites. Furthermore, genetic variants in regions flanking these enhancers

and which alter the epigenetic structure of the enhancer/promoter complex may also affect gene transcription. Thus, RCC susceptibility loci that overlap or lie close to any enhancer of a HIF-target gene may impact on the HIF-transcriptional output. Therefore, in an attempt to test the possibility that RCC-polymorphisms might also be enriched at these sites I first relaxed the criteria for overlap to look for excess overlap of RCC susceptibility loci with HIF-binding site proximal regions. I then made further use of the Capture-C analysis to look for more distant HIF-binding sites that exhibited looping interactions with enhancers at each of the sites covered by the viewpoint oligonucleotides.

5.2. Results

5.2.1 Identification of RCC susceptibility loci in close genomic proximity to HIF binding sites

As a first step, I examined for genomic proximity between the RCC-susceptibility loci defined from the GWAS meta-analysis in chapter 3 and HIF-binding sites in 786-O and RCC4 renal cancer cell lines. Owing to high levels of “background” interactions, Capture-C analysis is not able to reliably identify chromatin looping within 25 kb of the viewpoint of the capture site. Therefore, I initially decided to examine for RCC-susceptibility loci and HIF binding sites that lie within 25 kb of each other. More distant 3-dimensional looping interactions were then assessed using Capture-C analyses, as described in the next section.

Firstly, the HIF-binding regions identified in 786-O and RCC4 renal cancer cell lines in chapter 2 were extended by 25-kb upstream and downstream from the original coordinates. These new regions were aligned with the RCC-susceptibility loci identified in the GWAS meta-analysis in chapter 3. In addition to the 4 loci that directly overlapped a HIF-binding site a further three loci were found to lie within 25-kb of a HIF-binding site (Table 5.1 and Figure 5.2). This gives a total of 7 out of the 13 loci that either directly overlap or are in very close proximity to a HIF binding site.

Locus	*Index SNP*	*P-value*	*Proximity*
11q13.3	rs11263654	1.65E-23	**direct overlap**
2p21	rs2121267	6.95E-19	
12p12.1	rs11534749	6.19E-17	**direct overlap**
14q24.2	rs28840762	1.08E-15	
12q24.31	rs10846748	1.70E-12	< 25kb
8q24.21	rs6470588	2.82E-12	**direct overlap**
11q22.3	rs117706999	6.60E-10	
2q22.3	rs11888238	8.15E-10	
1p32.3	rs6676515	3.04E-09	
4q23	rs7697932	4.70E-09	< 25kb
3p22.1	rs9821249	1.12E-08	
15q22.31	rs12905354	2.28E-08	< 25kb
3q26.31	rs234043	4.10E-08	**direct overlap**

Table 5.1: Summary of the genetic locus, index SNPs, corresponding P-value of the index SNP for all independent genome-wide significant (P-value ≤ $5*10^{-8}$) RCC susceptibility loci and their proximity to HIF-binding sites.

Figure 5.2 A: 12q24.31 locus

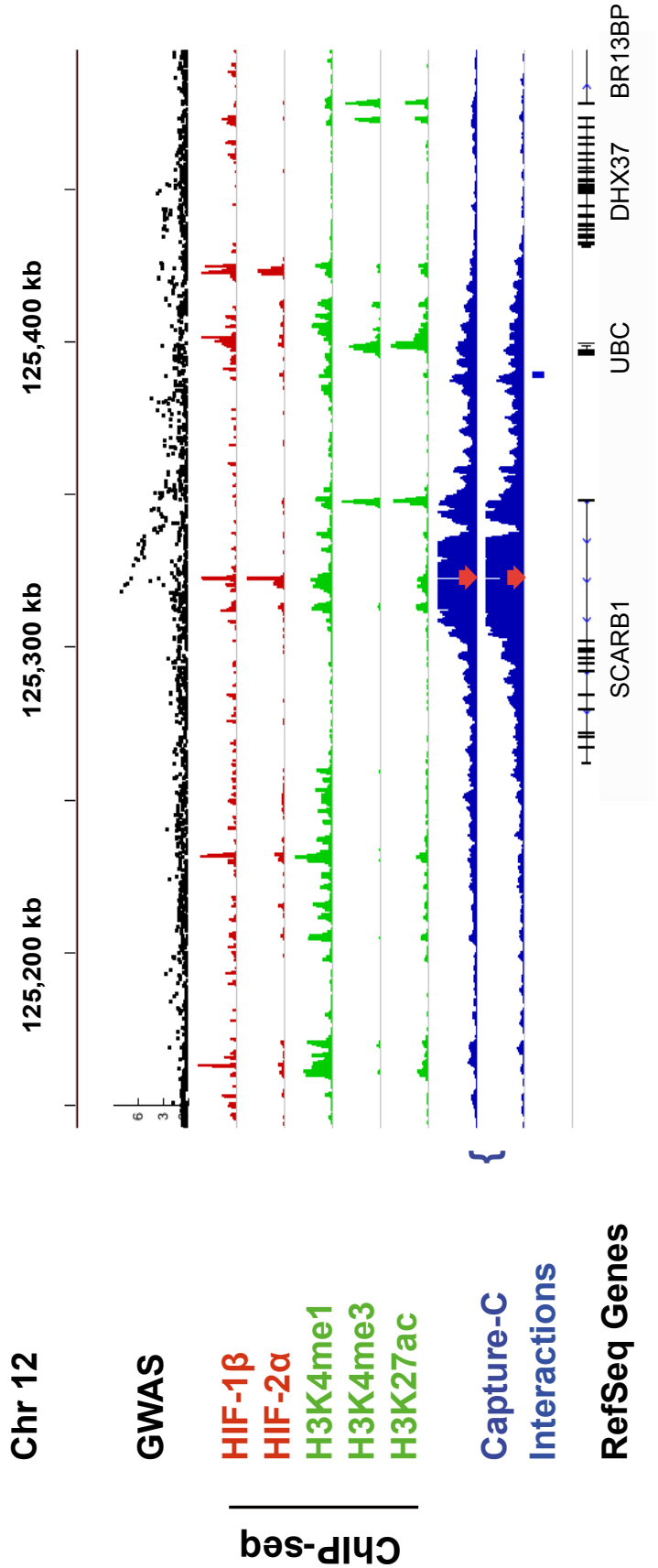


Figure 5.2 B: 4q23 locus

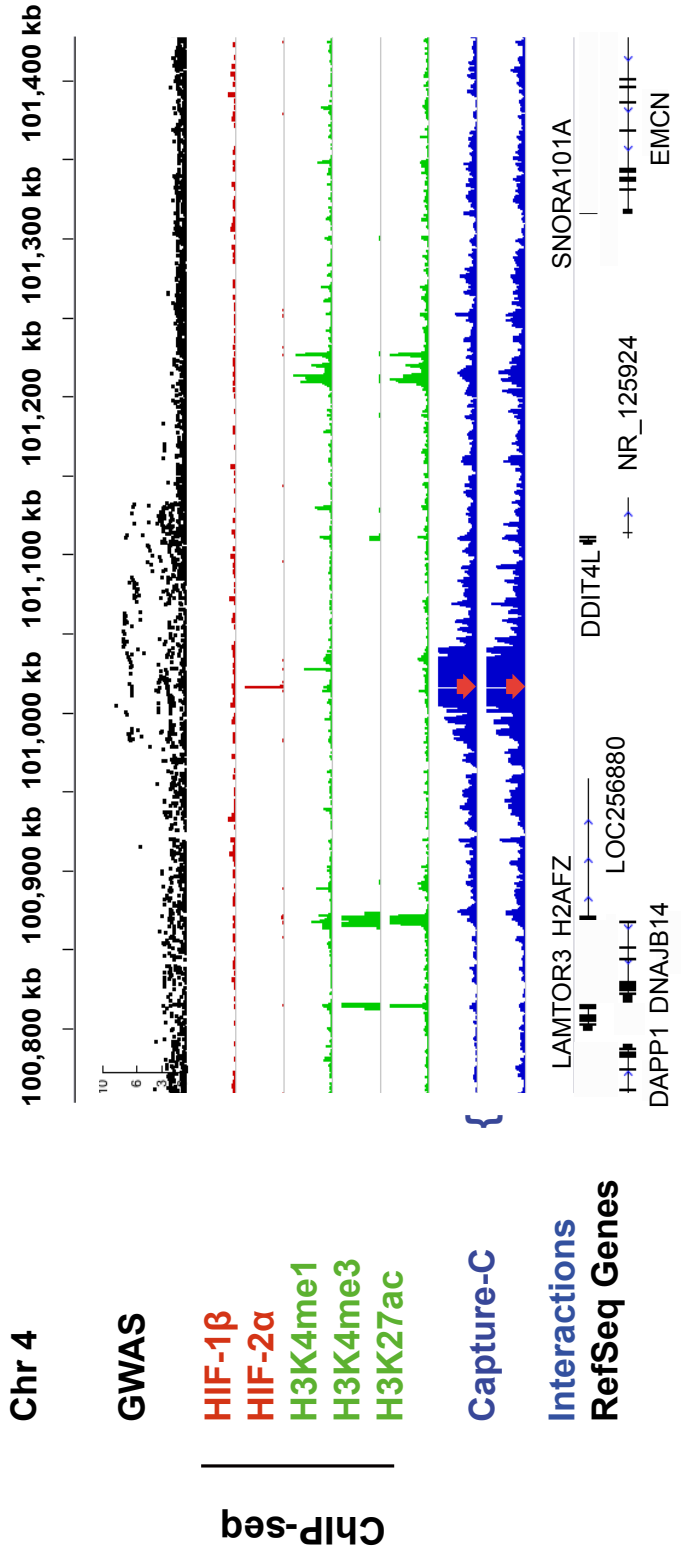


Figure 5.2 C: 15q22.31 locus

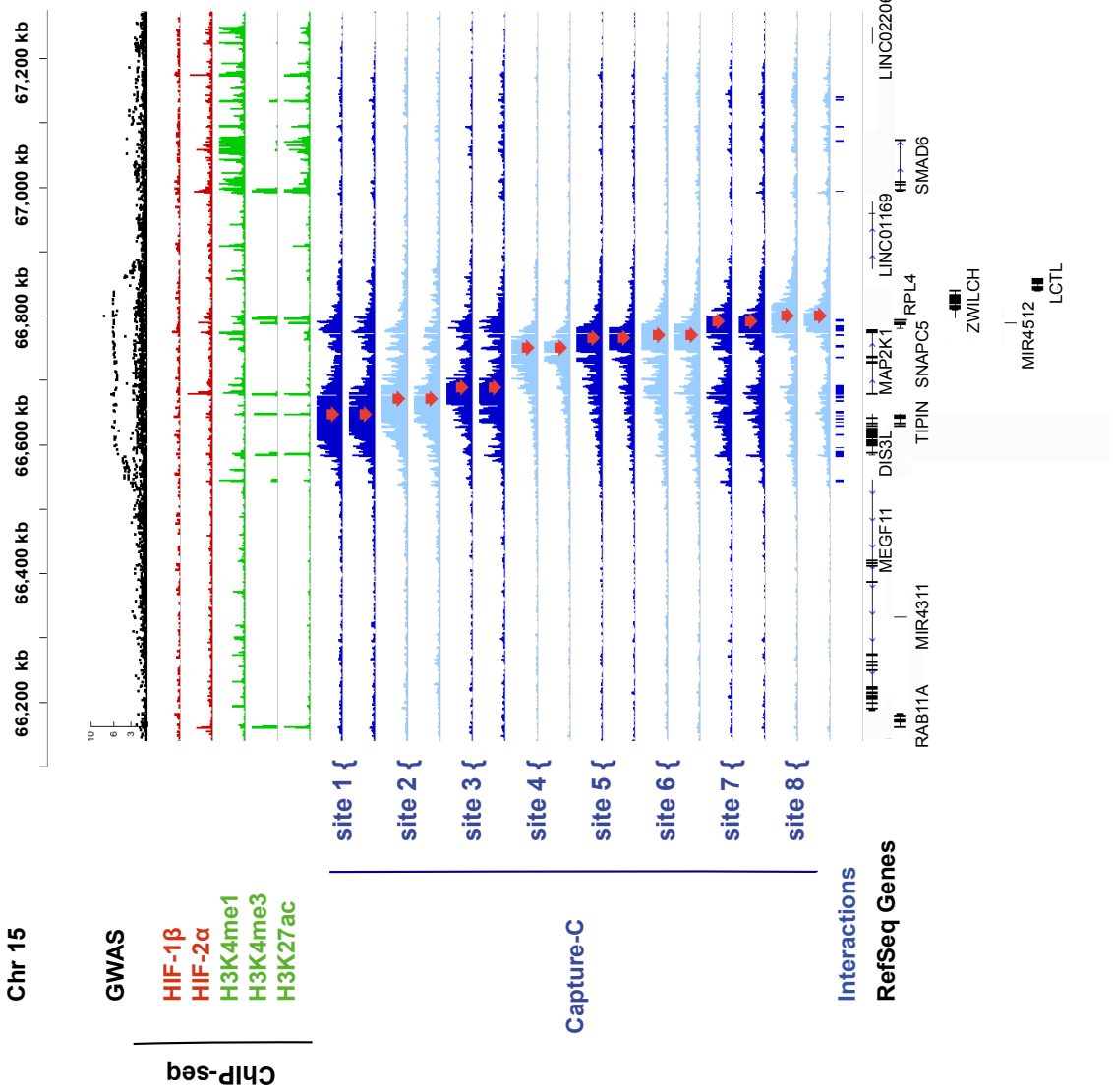


Figure 5.2: Integrative Genomics Viewer (IGV) of 3 HIF-proximal RCC susceptibility loci on 12q24.31, 4q23 and 15q22.31. The three loci are not directly coinciding but are located within a maximum distance of 25 kb from a HIF binding site. The closest SNP associated with the 12q24.31 is found 1.2 kb, for the 4q23 11.3 kb and 15q22.31 8kb away from the HIF binding site. Track 1 displays GWAS SNP-level P-values from the RCC GWAS meta-analysis. Track 2-3 (in red) display binding of HIF-2α and HIF1-β mapped in 786-O cancer cell lines of RCC. 4-6 from top (in green) display ChIP-seq signal (normalised read counts per million total reads at each chromosomal position) for histone H3K4me1, H3K4me3 and H3K27ac chromatin marks mapped in the 786-O kidney cancer cell line. The bottom tracks show Capture-C data with assays performed in 786-O cells using "viewpoint" (bait site) oligonucleotides at each active enhancer that coincided with the linked SNPs at the susceptibility locus (r^2 ≥ 0.8 with the index SNP) and is indicated by the red arrows. Regions with Capture-C interaction (blue bars) and in black gene annotation from the RefSeq hg19 (GRCh37) build is displayed. HIF binding data of RCC4 used to determine the set of high confidence HIF binding sites is not shown here due to space restrictions.

The 3 additional loci that lie within 25 kb of a HIF-binding site are located on chromosomes 12q24.31, 4q23 and 15q22.31. The 12q24.31 RCC-susceptibility locus lies 1.2 kb from a HIF-binding site that lies within the first intron of the gene encoding for the Scavenger Receptor Class B Member 1 (SCARB1) (Figure 5.1). SCARB1 is a plasma membrane receptor for high-density lipoproteins. As such, it may contribute to the high lipid content of RCC cells, which in turn partially accounts for the clear cell phenotype (glycogen and lipid are eluted during the fixation of cells giving clear areas on histology). SCARB1 has also been linked to progression and prognosis in RCC (Xu et al. 2018). Furthermore, analysis of RNA-seq and Capture-C data in 786-O cells from chapter 4 (Table 4.2) showed HIF-dependent induction of SCARB1 in this cell line and did not show interaction with the promoter of any other gene.

The 4q23 and 15q22.31 RCC-susceptibility loci both lie in intergenic regions. Capture-C analysis in chapter 4 did not identify any annotated gene promoters that interacted with the 4q23 locus (Table 4.2). However, 8 gene promoters potentially interact with the 15q22.31 locus (i.e. lie within 25 kb or exhibit long-distance 3-D interactions in Capture-C analysis). Of these, 3 (DDIT4L, DNAJB14, H2AFZ), were induced by HIF in the RNA-seq analysis of 786-O cells in chapter 4.

5.2.2 Statistical significance for the overlap and proximity of RCC-susceptibility loci with HIF binding sites

To assess the statistical significance of the observed overlap of 7 RCC-susceptibility loci that either directly overlapped or lay within 25-kb of a HIF-binding site I repeated the same bootstrap approach that I applied in chapter 2 and 3. The set of RCC susceptibility loci was again randomly moved around the genome 100,000 times and the overlap with the 25 kB extended set of HIF binding sites, described in the previous section, was assessed after each iteration (Figure 5.3). The observed number (7) of RCC-susceptibility loci that overlapped the extended set of HIF binding sites, was seen in only one out of the 100,000 iterations. Thus, the statistical significance of observing 7 RCC-susceptibility loci overlapping with a 25 kb extended HIF-binding interval corresponds to a P-value of $1*10^{-5}$ (Figure 5.3). Importantly, the significance of the observed overlap between RCC-susceptibility loci and the 25-kb extended set of HIF binding sites is comparable to that which was computed for the directly overlapping loci alone and has not been diluted by the addition of the new loci. Since the expanded HIF-binding sites extended beyond the regulatory regions defined in chapter 3, I was unable to randomly shuffle the extended HIF-binding sites within these regions of the genome in order to perform the reverse analysis as before.

As before, for comparison, the same analysis was then performed to assess the overlap between educational attainment susceptibility loci and the 25 kb extended set of HIF binding sites. 14 loci overlapped one of the extended regions, but this was not statistically significant (P-value = 0.32).

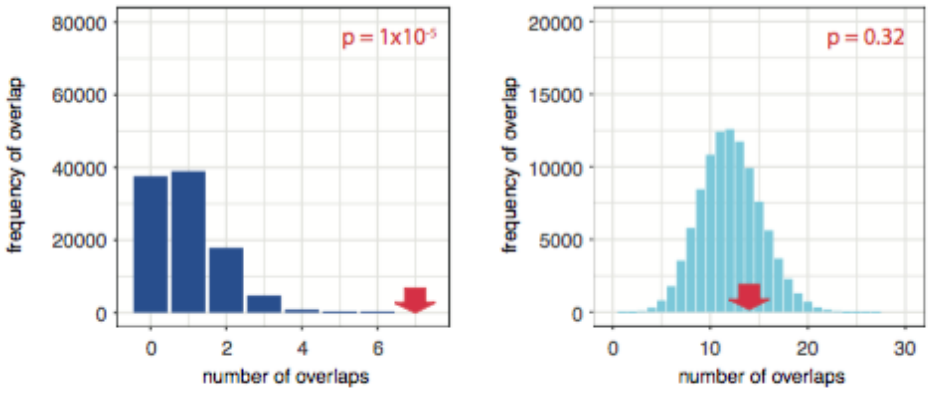


Figure 5.3: RCC-associated polymorphic loci overlap or lie within 25-kb of a HIF-binding site in RCC cells. A) The expected frequency distribution of the overlap of GWAS-identified RCC susceptibility loci with HIF binding sites extended by 25 kb on either side of each peak, was assessed by moving the index SNPs and corresponding LD SNPs (all SNPs $r^2 \geq 0.8$) randomly along the corresponding chromosome of each of the loci. The overlap with the set regions extended by 25 kb on each side of the HIF binding sites was assessed for each iteration (P-value = $1*10^{-5}$). B) The same analysis was performed on a set of control loci predisposing to educational attainment (P-value = 0.45).

5.2.3 Expression of HIF-target genes in ccRCC samples from the TCGA KIRC cohort versus paired normal kidney

As for RCC-susceptibility loci that directly overlapped a HIF-binding site, I examined the expression of putative HIF-target genes in ccRCC tumours themselves using RNA-seq expression analysis of renal tumours and surrounding kidney in the KIRC cohort of patients from the TCGA (Figure 5.4) as in chapter 4. At the 12q24.31 locus, SCARB1 was much increased in ccRCC tumours when compared to the surrounding kidney. At the 15q22.31 locus, only H2AFZ was highly significantly induced in the ccRCC tumours.

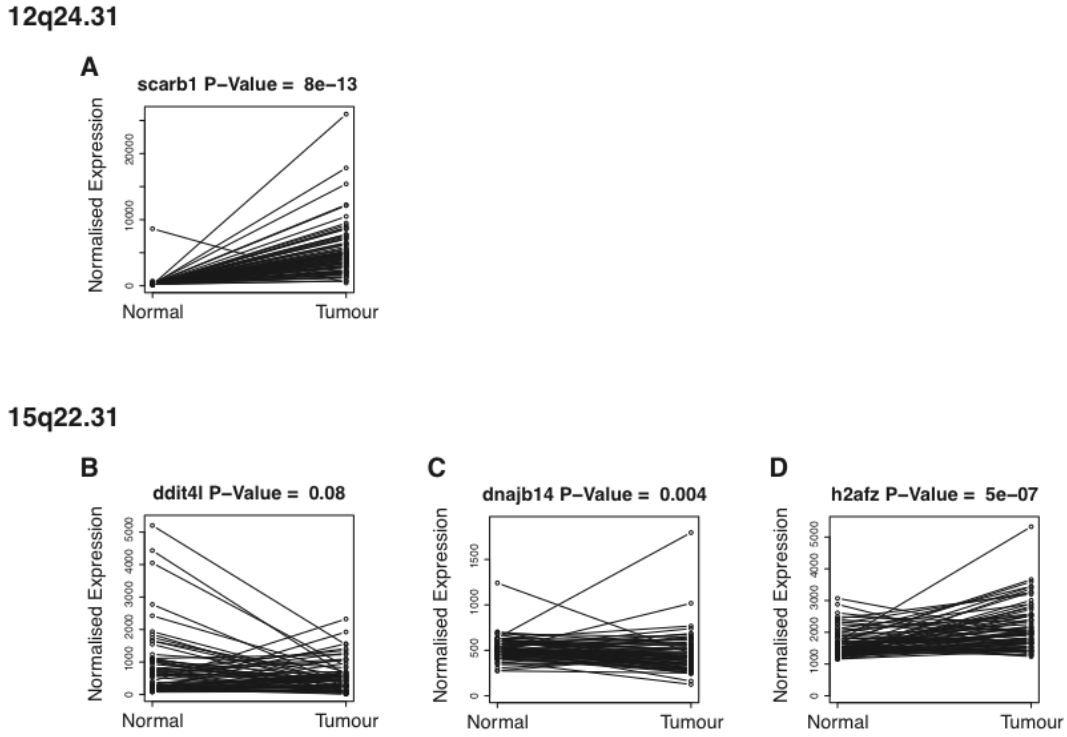


Figure 5.4: Gene expression in ccRCC compared to normal kidney. Gene expression in 72 ccRCC tumours and matched surrounding normal renal tissue was assessed using RNA-seq data from the TCGA KIRC cohort. Expression was normalised to the number of reads in each sample and the transcript length. Differential expression was assessed using a Mann-Whitney-Wilcoxon Test.

5.2.4 The effect of genotype at RCC-susceptibility loci on the expression of direct HIF-target genes

I next examined whether germline polymorphisms at these 2 loci had an effect on expression of the putative HIF-target gene in ccRCC tumours, again using data from the KIRC cohort of patients in the TCGA as in chapter 4. Germline RCC-susceptibility variants, at the 12q24.31 locus, were highly statistically associated with altered expression of SCARB1 (P-value = $3*10^{-4}$) in ccRCC tumours (Figure 5.5), the risk allele being associated with higher expression. RCC-susceptibility variants at the 15q22.31 locus were not significantly associated with expression levels of any of the three putative HIF-target genes. However, it is difficult to know how to interpret the lack of a significant eQTL, since the number of cases in the TCGA KIRC cohort that have both genotype and expression data available is relatively low, and the statistical power will therefore also be small. Furthermore, the stage of tumourigenesis at which the polymorphisms have their effect is unclear and the TCGA expression data all refer to late-stage disease.

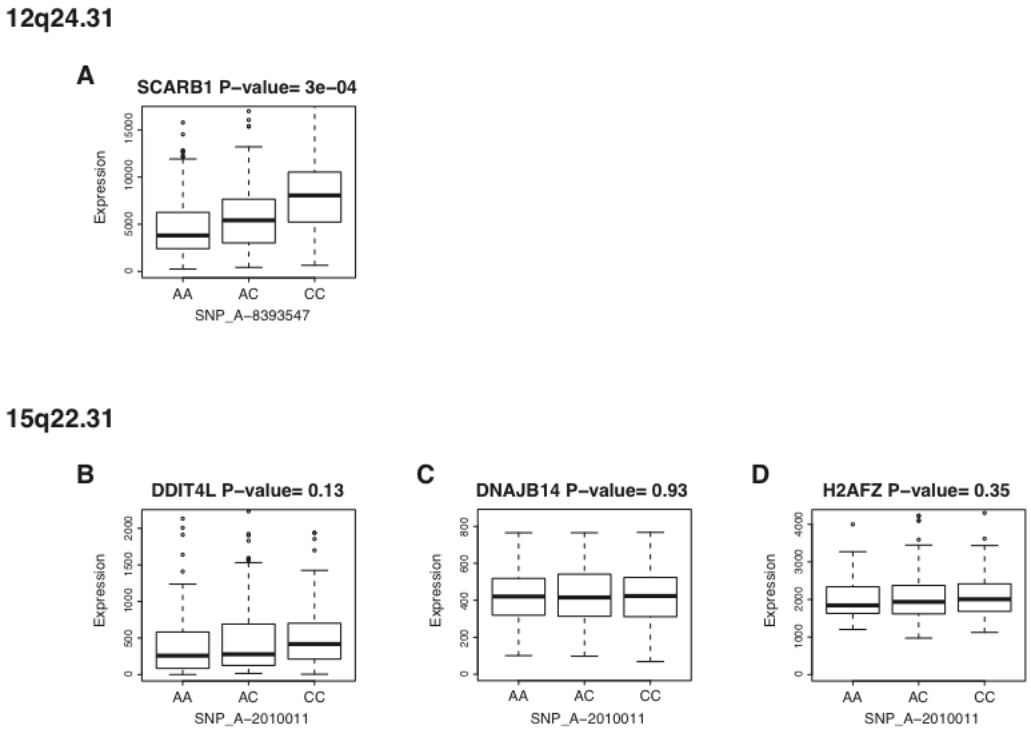


Figure 5.5 Box-and-whisker plots showing eQTL analysis for RCC-susceptibility loci that directly that lie within 25-kb of a HIF-binding site. Box-and-whisker plots showing the correlation, between (A) SCARB1, (B) DDIT4L, (C) DNAJB14 and (D) H2AFZ mRNA levels and the genotyped SNP in highest LD with the index SNP at each locus, in TCGA analysis of ccRCC tumors (KIRC).

5.2.5 Identification of more distant HIF binding sites in physical contact with RCC-susceptibility loci using Capture-C

Having identified additional HIF-binding sites in the regions immediately flanking, as well as directly overlapping, the RCC-susceptibility loci identified in the GWAS meta-analysis, I next examined for HIF-binding sites in more distant regions of the genome that interact with these loci in Capture-C analyses. The Capture-C analysis was performed by Veronique LaFleur, and Ran Li performed the initial bioinformatic analysis to identify statistically significant interacting regions.

Examined the regions considered to physical interact with RCC susceptibility loci via chromatin looping, as assessed by significant signal in the Capture-C assays, I then aligned the significantly interacting regions with the set of HIF binding sites defined in chapter 2. Interestingly, each of the 4 loci that overlapped directly with HIF-binding sites (11q13.3, 8q24.21, 12p12.1, 3q26.31) showed long-range interactions with other HIF ChIP-seq peaks. At most of these sites, the interacting HIF peaks had weak ChIP-seq signal compared to the directly overlapping site. Therefore, it is possible that this ChIP-seq signal arises as a result of co-immunoprecipitation of interacting regions with the main HIF-binding site, which will all be present in the same complex. However, the 3q26.31 locus which directly overlaps a weak HIF binding site was observed to interact with another, much stronger HIF-binding site lying close to the promoter of the TNFSF10 gene. Thus, it may be that polymorphisms at this locus are affecting an accessory site that is co-regulating TNFSF10, rather than the main site of HIF-binding.

In addition, the alignment also revealed that enhancers at the 2p21 (EPAS1 alias HIF-2α) locus have long-range interactions with a distant, HIF-bound enhancer, approximately 200 kb telomeric to the capture viewpoint (Figure 5.6). Interestingly, a second HIF binding site is present approximately 40 kb away from the capture viewpoint, although this region was not determined as significantly interacting above background levels based on the applied statistical model. However visual inspection of the Capture-C tracks suggests a second interaction point of the locus with this more proximal HIF binding site.

Analysis of EPAS1 mRNA levels in clear cell renal tumours in the TCGA KIRC cohort also showed increased expression in the tumours compared to the surrounding kidney, consistent with EPAS1 itself being a HIF target gene and subject to positive feedback (Figure 5.7A). At this locus, the genotyped SNP in highest LD with the index SNP (rs2121267) was rs4953345 (SNP_A-2299844). However, the genotype of this SNP did not correlate significantly with expression of EPAS1 mRNA (Figure 5.7B).

In total, 8 out of the 13 RCC-susceptibility loci, either directly overlap, are in close genomic proximity with, or physically interact with a HIF-binding site mapped in RCC cell lines. Furthermore, EPAS1 is the gene that encodes for the HIF-2α subunit of HIF itself. Thus, examining for direct overlap alone appears to underestimates the extent of the interaction between RCC-susceptibility loci and the HIF pathway.

Figure 5.6: 2p21 locus

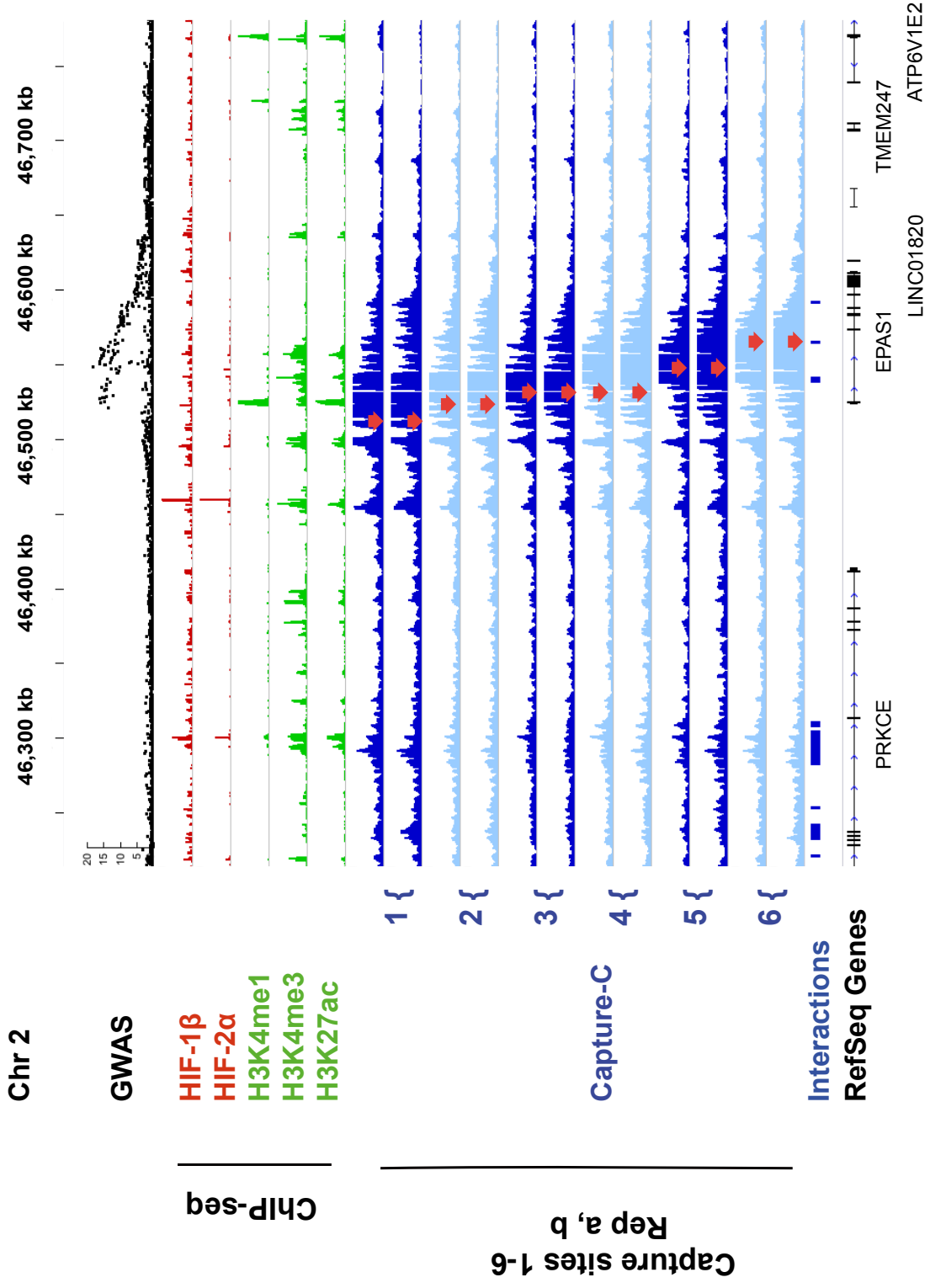


Figure 5.6: Integrative Genomics Viewer (IGV) tracks of the 2p21 locus. Layout is as for Figure 5.2. Interactions with a distant HIF binding site are detected with Capture-C. Additionally, interactions with an additional strong HIF-binding site close to the viewpoint (red arrow) might be indicated by the Capture-C data. The proximity to the viewpoint might prevent the detection of this site above background level.

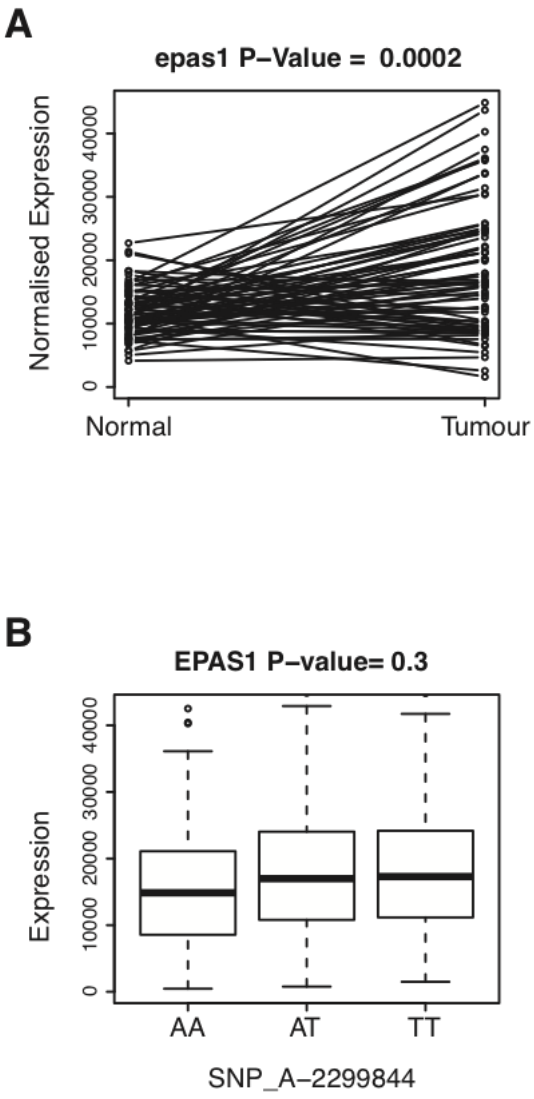

A
epas1 P−Value = 0.0002
Normalised Expression
0
10000
20000
30000
40000
Normal
Tumour
B
EPAS1 P−value= 0.3
Expression
0
10000
20000
30000
40000
AA
AT
TT
SNP_A−2299844

Figure 5.7: Analysis of EPAS1 expression levels in TCGA RNA-seq datasets. A) EPAS1 mRNA expression in 72 ccRCC tumours and matched surrounding normal renal tissue was assessed using RNA-seq data from the TCGA KIRC cohort. Expression was normalised to the number of reads in each sample and the transcript length. Differential expression was assessed using a Mann-Whitney-Wilcoxon Test. B) Box-and-whisker plots showing the correlation between EPAS1 mRNA levels and the genotyped SNP (SNP_A-2299844) in highest LD with the index SNP (rs2121267) at the 2p21 locus.

5.3 Discussion

Work in previous chapters has shown a significant overlap between RCC-susceptibility loci and HIF-binding sites. These physically interact with and regulate the expression of a number of HIF-target genes. However, it is difficult to define the full extent of the cis-acting elements that co-operate in regulating the expression of a HIF-target gene and therefore the potential overlap with RCC-susceptibility loci. Such elements will include, not only the HIF-binding site itself but also the binding sites of other transcription factors that co-regulate the same gene and/or help stabilise HIF-binding. These may lie close to the HIF-binding site, for instance within the same region defined by ChIP-seq analysis, or they may appear apparently remote from the HIF-binding site but interact with it through long-distance chromatin-looping interactions. Furthermore, the full cis-acting apparatus will include additional elements that help define both the local and long-range chromatin environment, but which may themselves lie distant from recognisable regulatory elements. Genetic variants at any one of these elements could therefore alter the expression of a HIF-target gene.

In chapter 5, I have explored this possibility by examining for RCC-susceptibility loci that lie adjacent to HIF-binding sites and/or interact with such sites through chromatin looping. In this context, the definition of “adjacent” is an arbitrary one. Too strict a definition, in terms of genomic distance, risks missing true associations, whilst a definition that encompasses very large genomic regions will almost invariably identify a HIF-binding site. As a compromise, I have settled on a distance of 25-kb, since analysis of Capture-C data, indicates a high level of “background” chromatin interactions at distances shorter than this.

This analysis has revealed 3 additional loci (12q24.31, 4q23 and 15q22.31) that are associated with HIF-binding sites, through proximity, to give a total of 7 loci that either directly overlap or are close to a HIF-binding site. Again, this is significantly more than would be expected by random chance alone, with a P-value comparable to that for seen in chapter 4 for direct overlap. In addition, examination of Capture-C data has identified an 8th RCC-susceptibility locus (2p21) that shows statistically significant long-distance chromatin-looping interactions with a distant HIF-binding site that may therefore regulate the same gene. Therefore, in total, 8 out of the 13 RCC-susceptibility loci potentially influence the cis-acting elements of a HIF-target gene. Thus, the association between RCC-susceptibility polymorphisms and cis-acting elements of the HIF pathway may be as much as double that estimated by simple overlap alone. It also strongly suggests that RCC-susceptibility polymorphisms may be affecting the expression of HIF-target genes through mechanisms other than direct disruption of the HIF-binding site.

Chapter 6: Identification of sub-threshold RCC susceptibility loci

6.1 Introduction

In the previous chapters, I have shown that RCC-susceptibility loci, which have been defined in GWAS analyses, significantly overlap HIF-binding sites in RCC cell lines and are enriched for proximity to HIF-induced genes. However, these GWAS analyses have defined relatively small numbers of RCC-susceptibility loci, each of which has a relatively minor effect on the risk of developing RCC. In line with previous studies, Scelo et al. have estimated that the loci that they identify in their meta-analysis account for only about 10% of the familial risk of RCC (Gudbjartsson et al. 2002). Indeed, an increasing number of studies have demonstrated that for a variety of common traits, susceptibility loci reaching genome-wide significance only account for a small proportion of heritability to the respective trait. This suggests that there are other genetic variants that pre-dispose to RCC that are yet to be discovered.

The significance thresholds used in GWA studies to indicate genetic association with a specific trait are set at conservative significance levels – so called genome-wide significance (typically P-value = $5*10^{-8}$ - 10^{-8}). This is stricter than conventional significance thresholds used commonly in scientific studies in which single hypotheses are tested (typically P-value = 0.05-0.01). In GWAS however, multiple comparisons are undertaken, and hence multiple hypothesis are tested simultaneously on one dataset. Indeed, it is estimated that the several million SNPs present in the human genome are organised in about 1 million independent haplotype blocks (Gabriel et al. 2002). Hence, in a GWA study, roughly 1 million independent statistical tests are performed on the same dataset. If a significance threshold of 0.05 was used, then 5% (or 50,000) of these comparisons would be

expected to reach significance by random chance alone, even if there was no association.

The simplest method to account for these multiple comparisons is the Bonferroni correction in which the desired significance threshold, usually a P-value of 0.05, is divided by the number of tests performed on the dataset. In the case of GWAS analyses, the standard P-value of 0.05-0.01 is divided by the 1 million tests performed on the dataset giving a "corrected" threshold P-value of $5*10^{-8}$ – 10^{-8} for genome-wide significance.

To achieve these high levels of statistical association, especially with the weak effects typical of common genetic variants, large sample sizes are required in GWA studies. Dependent on the trait studied, these numbers might be difficult to achieve and therefore studies are often underpowered to detect all common susceptibility variants.

This problem is further compounded by low frequency variants. Technically, the power of a study to detect an association depends not only on the number of individuals, but also on the number of times a specific allele is observed. These will be lower for rarer frequency alleles than for alleles that approach a minor allele-frequency (MAF) of 50%. For example, a minor allele frequency of 10% means that approximately 90% of individuals will be homozygous for the major allele approximately 10% will be heterozygous, whilst only 1% of the patients in the study will be homozygous for the minor allele. This can make it even harder

and requires much larger total number of individuals to reach P-values of $5*10^{-8}$ for these rarer alleles. Indeed, Scelo et al. estimated that while the power of their meta-analysis to detect a relative risk of RCC of 1.2 was about 80% for polymorphisms where the MAF was greater than 20%, it was very poor where the MAF was less than 10% (Scelo et al. 2017).

To increase the potential information that can be obtained from current GWA studies, Wang et al. have suggested an alternative way to prioritise disease-predisposing loci that fail to reach genome-wide significance (Wang et al. 2016). Essentially, they combined SNP-level summary statistics from GWAS of cardiac QT interval and QRS duration with epigenetic maps of active enhancers. They ranked all measured SNPs according to the significance of their association with cardiac QT interval and QRS duration and examined for enrichment of these SNPs within regions marked as active enhancers. They found that SNPs associated with the trait at intermediate levels of significance (10^{-4} > P-value ≥ $5*10^{-8}$), as well as SNPs that achieved conventional genome-wide levels of significance (P-value < $5*10^{-8}$), were enriched at active enhancers. Interpreting this result, they suggested that a higher than expected proportion of these SNPs, with a P-value between 10^{-4} and $5*10^{-8}$, represented true association between the genotype and the trait. However, it is unclear at this point if the observed enrichment of RCC SNPs of intermediate levels of significance is a reflection of SNPs which are in linkage disequilibrium with genome-wide significant RCC susceptibility loci are if they are independent sub-threshold RCC susceptibility loci.

Applying this method to the meta-analysis of RCC GWAS in chapter 3, suggested that RCC-susceptibility loci of intermediate statistical significance were further enriched for HIF-binding sites, beyond the enrichment observed for active enhancers. And in chapter 6, I examine this possibility further (Figure 3.4).

6.1 Results

6.2.1 Identification of sub-threshold RCC-susceptibility loci

To test the for an association between sub-threshold RCC-associated GWAS signal and the HIF pathway, I first defined sub-threshold RCC susceptibility loci. SNP-level summary statistics from the study were used to sequentially define independent loci, as in chapter 3. Essentially, independent loci were identified, beginning with the most significantly associated SNP in the genome, which was defined as the index SNP for that locus. All SNPs in close genetic or physical proximity to the index SNP ($r^2 \geq 0.2$ with reference to the 1000 Genomes EUR population data or distance < 500 Mb) were then excluded. SNPs in genetic linkage with the index SNP were excluded since they would show a significant association with RCC simply by co-inheritance with the causative polymorphism. A threshold of $r^2 \geq 0.2$ was chosen to exclude SNPs that hade even a weak correlation with the index SNP. Loci that were in close proximity (500 Mb) were also excluded even if they were not in genetic linkage as they might affect the regulation of the same set of genes. Having excluded all SNPs that might be associated with the first locus, the next most significant SNP defined the second locus and the process was repeated. However, as opposed to chapter 3 this time, the process of defining loci continued until a threshold of 10^{-4} was reached. This level of significance was chosen to define sub-threshold loci since it was the same threshold used by Wu et al. to define sub-threshold loci associated with QT interval duration.

In total 3796 out of the 7 million SNPs genome-wide had an association with RCC with a P-value of 10^{-4} or less (Figure 6.1). These defined a total of 232 loci (Figure

6.2), of which 13 reached genome-wide significance (P-value < $5*10^{-8}$) which are described in chapter 3. A further 219 loci were classified as sub-threshold (index SNP between 10^{-4} and $5*10^{-8}$) and are the subject of this chapter. For each index SNP identified, the corresponding haplotype block was then determined by identifying SNPs in high LD ($r^2 \geq 0.8$) with the identified index SNP as before.

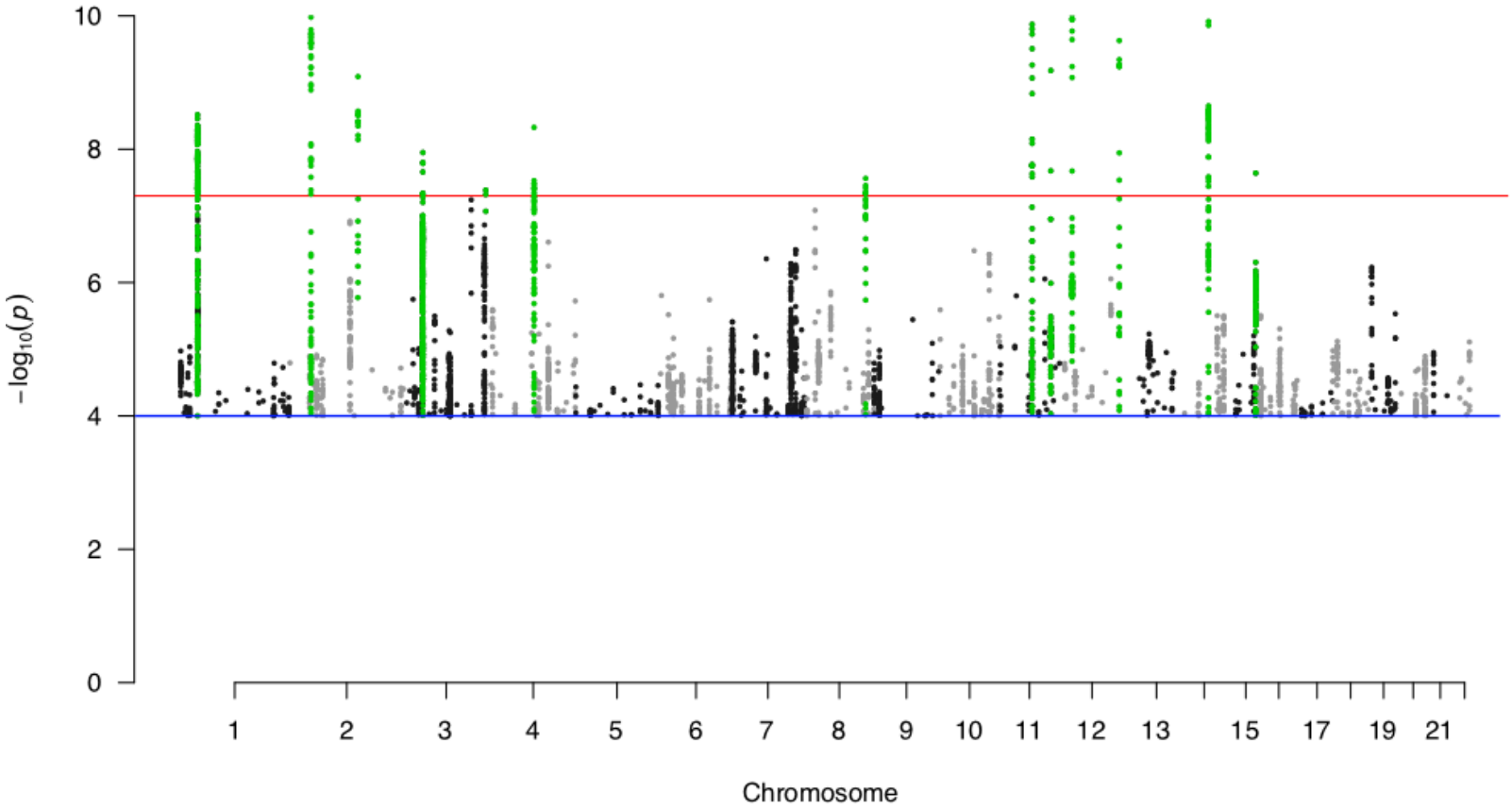

10
8
6
4
2
0
$-\log_{10}(p)$
1
2
3
4
5
6
7
8
9
10
11
12
13
15
17
19
21
Chromosome

Figure 6.1: Manhattan plot showing SNP-level P-values for all SNPs with significance levels between 10^{-4} and 10^{-10}. SNPs associated with loci that reach conventional genome-wide significance are coloured in green. The red line denotes a P-value of $5*10^{-8}$. The blue line denotes a P-value of 10^{-4}.

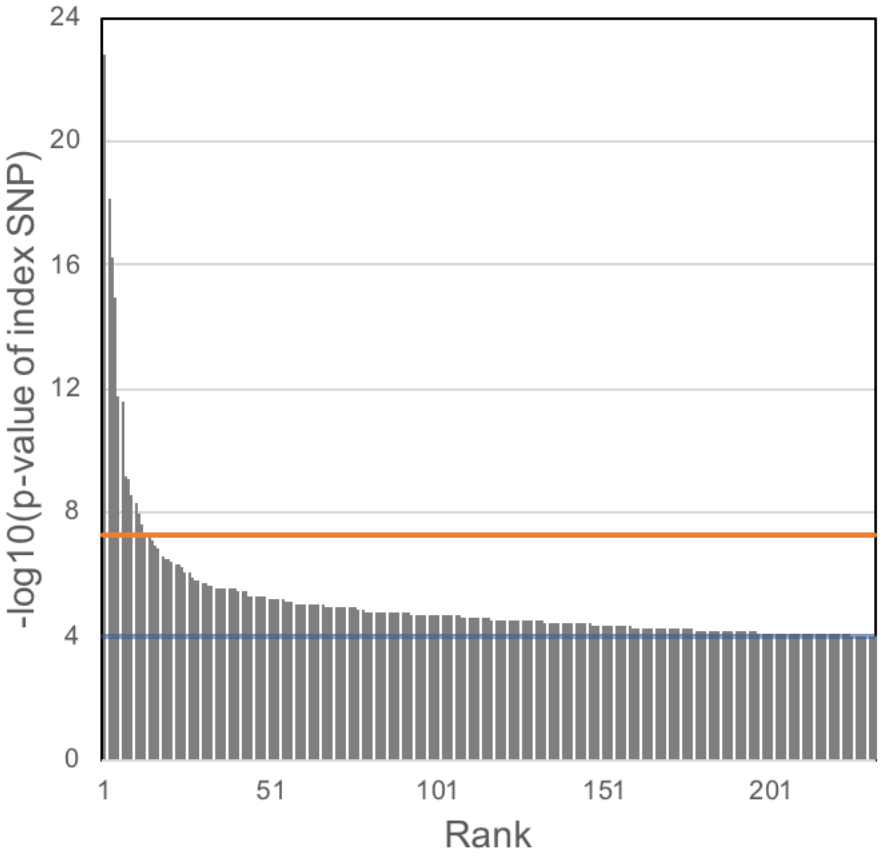


Figure 6.2: Independent RCC-associated loci and the significance level of the index SNP. The red line denotes a P-value of $5*10^{-8}$. The blue line denotes a P-value of 10^{-4}.

6.2.2 Sub-threshold RCC-susceptibility loci are enriched for proximity to HIF regulated genes

I initially examined whether the sub-threshold RCC-susceptibility loci were enriched for proximity to HIF regulated genes. Genes were first ranked according to their differential expression (-log10 *log2FC) in 786-O cells stably transfected

either with control vector or with wild-type VHL using the same poly-adenylated RNA-seq analysis as in Figure 4.2. The 3 nearest genes to each sub-threshold locus were then identified and a weighted, pre-ranked GSEA was performed using this set of genes (Figure 6.3). This showed that genes proximal to the sub-threshold loci are significantly enriched (ES = 0.64, NES = 1.34, P-value= 0.01) for genes upregulated by VHL in the 786-O renal cancer cell line. This result is comparable to that observed for genes that lie in close proximity to the RCC-susceptibility loci that reach conventional levels of genome-wide significance (P-value$\leq 5*10^{-8}$). However, the enrichment score was lower than that observed for the above-threshold loci as seen in Figure 4.2. Indeed, this is to be expected, since relaxing the statistical threshold for identifying susceptibility loci will likely lead to the inclusion of an increased number of loci with a false-positive association with RCC. Nevertheless, the observation that the enrichment for HIF-regulated genes extends from the above-threshold into the sub-threshold loci strongly suggests that the latter includes a significant proportion of loci that are truly associated with RCC.

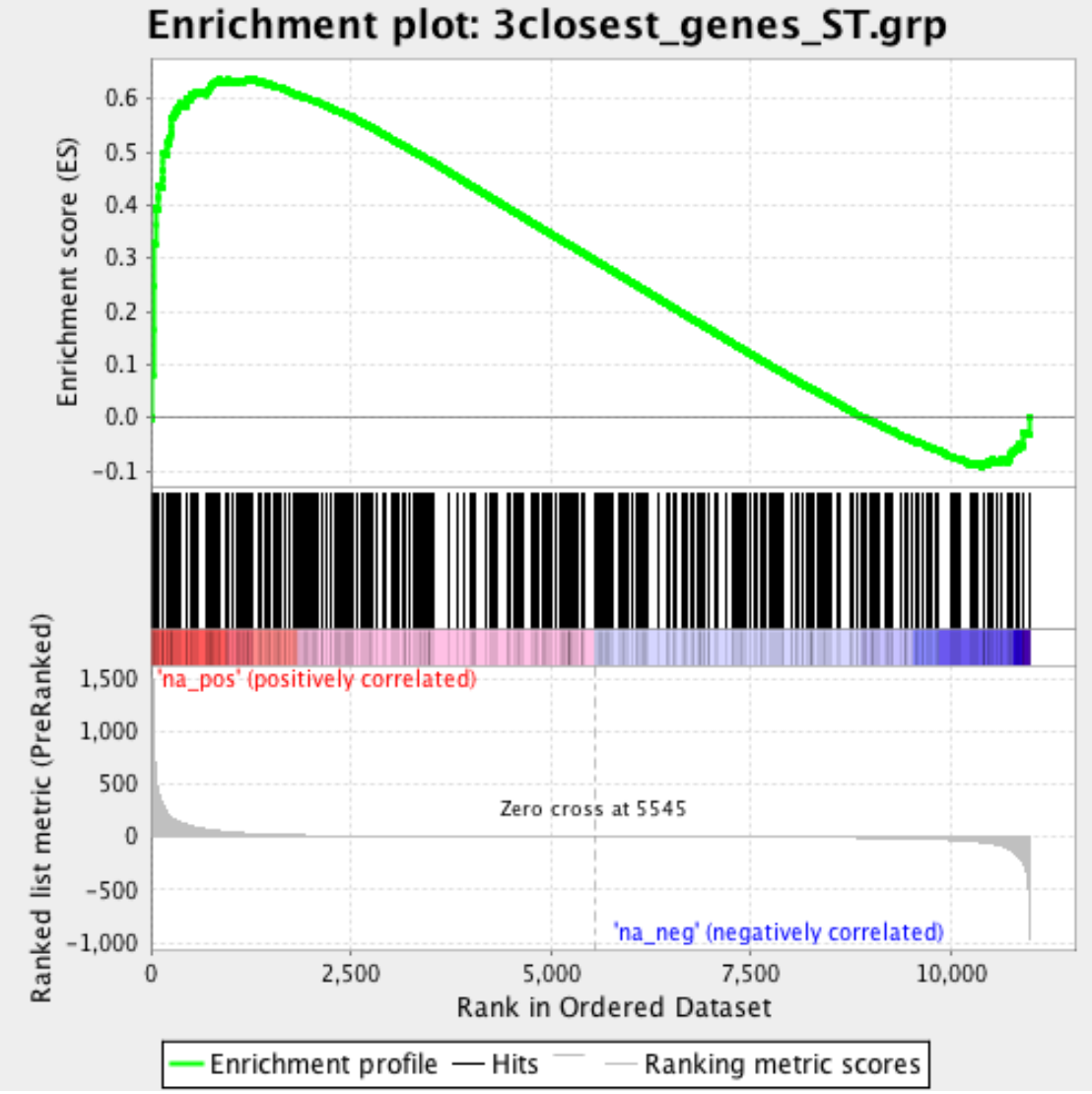


Figure 6.3: Enrichment of sub-threshold RCC susceptibility loci in close proximity to HIF regulated genes. The 3 closest genes to the index SNPs at each of the sub-threshold significant RCC susceptibility loci ($5*10^{-8}$≤ P-value < 10^{-4}) were identified. Enrichment for VHL regulation of the 657-gene set was shown by weighted, pre-ranked GSEA (P-value = 0.01). Differential expression by HIF was assessed between 786-O cells stably transfected either with wild-type VHL or with a control vector. Genes were ranked using a combined metric of log fold change and significance level (-log10sig*log2fc).

6.2.3 Sub-threshold RCC susceptibility loci directly overlap HIF-binding sites

To identify sub-threshold RCC susceptibility loci that directly overlap with a HIF binding site, the sub-threshold loci identified above were then aligned with the same set of 5,054 HIF-binding sites as defined in the ChIP-seq analyses of RCC cell lines in chapter 2. This identified 10 sub-threshold RCC-susceptibility loci that directly overlapped with a total of 13 HIF binding sites (Table 6.1, see appendix for IGV tracks).

locus	*index SNP*	*P-value*
10q22.1	rs6480601	3.33E-07
15q22.2	rs11071676	6.26E-06
6p22.2	rs13191296	1.94E-05
7p15.3	rs2285944	2.13E-05
12q23.1	rs151008469	2.23E-05
6q25.1	rs2145744	3.77E-05
21q22.3	rs112658743	4.98E-05
7p21.1	rs11764350	5.03E-05
19q13.32	rs6509303	5.98E-05
18q21.2	rs488155	6.11E-05

Table 6.1: Sub-threshold RCC-susceptibility loci that overlap HIF-binding sites. Chromosomal location, rs number of the most-significant (index) SNP and the significance level corresponding to the index SNP are shown for the 10 independent sub-threshold RCC-susceptibility loci (P-value 10^{-4} > P-value ≥ $5*10^{-8}$), defined from summary statistics of a meta-analysis of RCC GWAS performed by Scelo et al., which directly overlap a HIF binding site mapped in RCC cancer cell lines (Scelo et al. 2017).

6.2.4 Sub-threshold RCC susceptibility loci coincide with HIF binding sites more than expected

Given the striking enrichment of RCC-susceptibility loci that meet conventional levels of genome-wide significance at HIF binding sites (as observed in chapter 2 and 3), I next tested whether this enrichment also extended to sub-threshold RCC-susceptibility loci. The same bootstrapping approach, used previously, was applied to determine the likelihood of observing 10 sub-threshold RCC-susceptibility loci overlapping a HIF-binding site by random chance alone (Figure 6.4).

When the sub-threshold loci were randomly shuffled across the genome, the observed overlap of 10 RCC-susceptibility loci or more with a HIF-binding site occurred in 3877 out of the 100,000 iterations, giving a statistically significant P-value = 0.04 (Figure 6.4 A). Thus, just as for loci that reach conventional levels

of genome-wide significance, the sub-threshold loci are also enriched for HIF-binding sites in RCC cells. The level of statistical significance for this overlap is less for the sub-threshold loci, which is again consistent with a higher proportion of these loci having a false positive association with RCC. However, the number observed (10) is roughly double that expected from the frequency distribution (point of maximum and median are both 5 overlapping loci). This might suggest that as many as half of these sub-threshold RCC-susceptibility loci are truly associated with RCC.

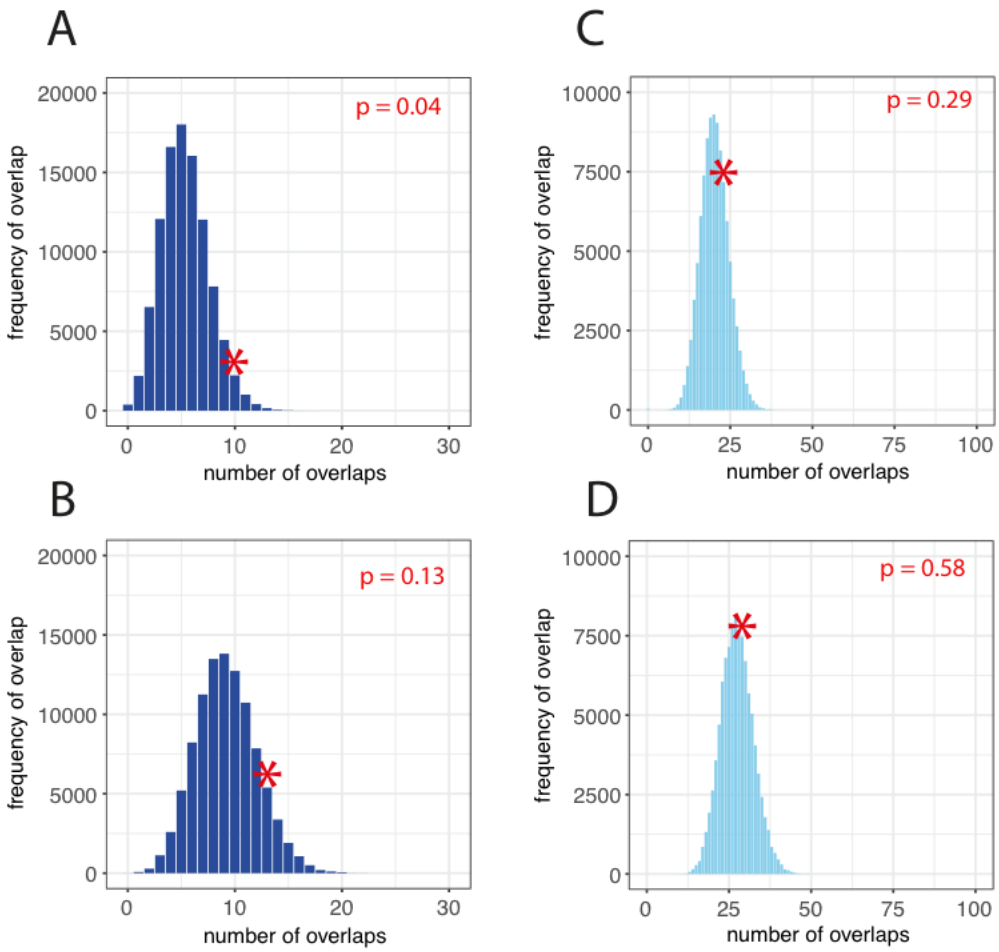


Figure 6.4: 10 RCC sub-threshold GWAS loci ($5*10^{-8}$ ≤ P-value <10^{-4}) coincide with a total of 13 HIF bindings sites mapped in 786-O and RCC4 cell lines. A) The expected frequency distribution of the overlap of sub-threshold RCC susceptibility loci with HIF binding sites. B) Inversely, the expected overlap of HIF binding sites with LD regions of sub-threshold RCC susceptibility loci assessed by moving the HIF binding sites along the enhancer space (defined by H3K27ac, H3K4me3, H3K4me1 in 786-O cells). Corresponding control experiments were performed based on education attainment sub-threshold loci (C, D).

I next tested whether the probability of a sub-threshold locus overlapping a HIF bound enhancer was greater than that observed with enhancers in general. HIF binding sites were randomly shuffled within regions marked as active enhancers defined in chapter 2 (defined by H3K27ac, H3K4me3, H3K4me1 in 786-O cells). The frequency distribution for the overlap between these randomly assigned sites and sub-threshold RCC-susceptibility loci is shown in Figure 6.4 B. When shuffled in this way, the expected overlap was approximately 8-10 sites. However, the probability of 13 HIF-binding sites overlapping a sub-threshold RCC-susceptibility locus did not reach statistical significance (P-value=0.13). Therefore, while sub-threshold RCC-susceptibility loci are enriched for HIF-binding sites, this may, at least in part, be accounted for by enrichment for enhancers more generally. This might indicate that the enrichment of sub-threshold RCC susceptibility loci towards HIF binding sites is at similar levels as towards enhancer sites. This may indicate effects on alternative non-HIF bound enhancers, however, enrichment for VHL regulable genes in proximity of sub-threshold loci observed earlier in this chapter, corroborates the evidence for an effect on the VHL/HIF axis.

As a comparator, I then defined sub-threshold susceptibility loci for educational attainment in the same way as for RCC, using the SNP-level summary statistics from the (NHGRI-EBI) GWAS catalogue used in chapter 3. 595 sub-threshold loci were identified, of which 23 overlapped with 27 HIF-binding sites defined in the ChIP-seq analysis of RCC cell lines. The same bootstrapping approaches determined the frequency distributions for the expected overlap (Figure 6.4C and 6.4D). No statistically significant overlap between sub-threshold educational

attainment loci and HIF-binding sites was observed for either method (P-value=0.29 and P-value=0.58, respectively).

6.2.5 Defining potential target genes of the HIF-binding site that overlap sub-threshold RCC-susceptibility loci.

Having identified 10 sub-threshold RCC-susceptibility loci that overlap with 13 HIF-binding sites, I next wished to determine the potential transcriptional targets of each HIF-binding site. Capture-C experiments were performed in duplicates by Veronique Lafleur in 786-O renal cancer cell lines, using oligonucleotides targeting viewpoints at each of the 13 HIF-binding sites. Analysis was performed, as previously, to define chromatin regions that showed interactions with these viewpoints that were significantly above the estimated background signal. Gene promoters that overlapped these interacting regions, or that lie within 25 kb of the locus were then identified (Table 6.2). Promoters were defined as the transcription start site +/- 2 kb by reference to the hg19 assembly annotation provided by Ensembl 19 gene promoters (plus the histone gene cluster) were found to interact with the 13 HIF-binding sites across the 10 sub-threshold loci. Of these, 9 were induced by the presence of HIF (introduction of VHL) in 786-O RCC cells, 6 were downregulated, 5 were not significantly altered and 5 were not expressed.

locus	index SNP	P-value	interacting genes	Regulation by VHL
10q22.1	rs6480601	3.33E-07	CHST3	↑
			PSAP	↑
			DDIT4	-
			ANAPC16	↓
			ASCC1	↓
15q22.2	rs11071676	6.26E-06	TPM1	↑
			TLN2	↑
6p22.2	rs13191296	1.94E-05	HIST gene cluster	
			SLC17 gene cluster HFE	
			HMGN4	↓
			ABT1	-
			TRIM38	↓
			SCGN	n.e.
7p15.3	rs2285944	2.13E-05	DNAH11	n.e.
12q23.1	rs151008469	2.23E-05	ELK3	↑
6q25.1	rs2145744	3.77E-05	MTHFD1L	-
			ZBTB2	↑
21q22.3	rs112658743	4.98E-05	TRPM2	n.e.
7p21.1	rs11764350	5.03E-05		
19q13.32	rs6509303	5.98E-05	ARHGAP35	↑
			SLC1A5	↓
			NPAS1	n.e.
			AP2S1	↓
			BBC3	-
18q21.2	rs488155	6.11E-05	POLI	-
			MBD2	↑
			STARD6	n.e.
			C18orf54	↑

Table 6.2: Genes within 25 kb and with promoters looping to sub-threshold RCC-susceptibility loci overlapping with a HIF binding site. Table showing sub-threshold loci that overlap a HIF-binding site together with the index SNP, P-value of index SNP. Genes whose promoters lie within 25-kb of a HIF binding site, or whose promoter loops to a HIF binding site in 786-O RCC cells are listed. Regulation of each gene by VHL in 786-O cells is indicated in the last column (↑= upregulated in VHL-deficient cells, ↓ = downregulated in VHL-deficient cells, - = not regulated, ne = not expressed).

6.2.6 Analysis of sub-threshold RCC-susceptibility loci that overlap HIF-binding sites using TCGA-KIRC datasets.

I next examined whether the 9 genes induced by HIF in 786-O cells were also increased in ccRCCs more generally. In these tumours HIF will be constitutively activated and therefore will potentially drive expression of these genes. Using RNA-seq data from 72 individuals in the TCGA KIRC cohort, I compared gene expression in matched ccRCC and surrounding kidney samples (Figure 6.5). Only 3 genes showed statistically significant higher mRNA levels in the tumours compared to normal tissues. However, if expression levels are low in tubular cells, from which ccRCC originate, any induction in the cancer cells may be overshadowed by constitutive mRNA levels in other cell types. Total tumour RNA contains transcripts from stromal and immune cells as well as the tumour cells

themselves and hence expression levels are effected by the purity of the samples which was not estimated here.

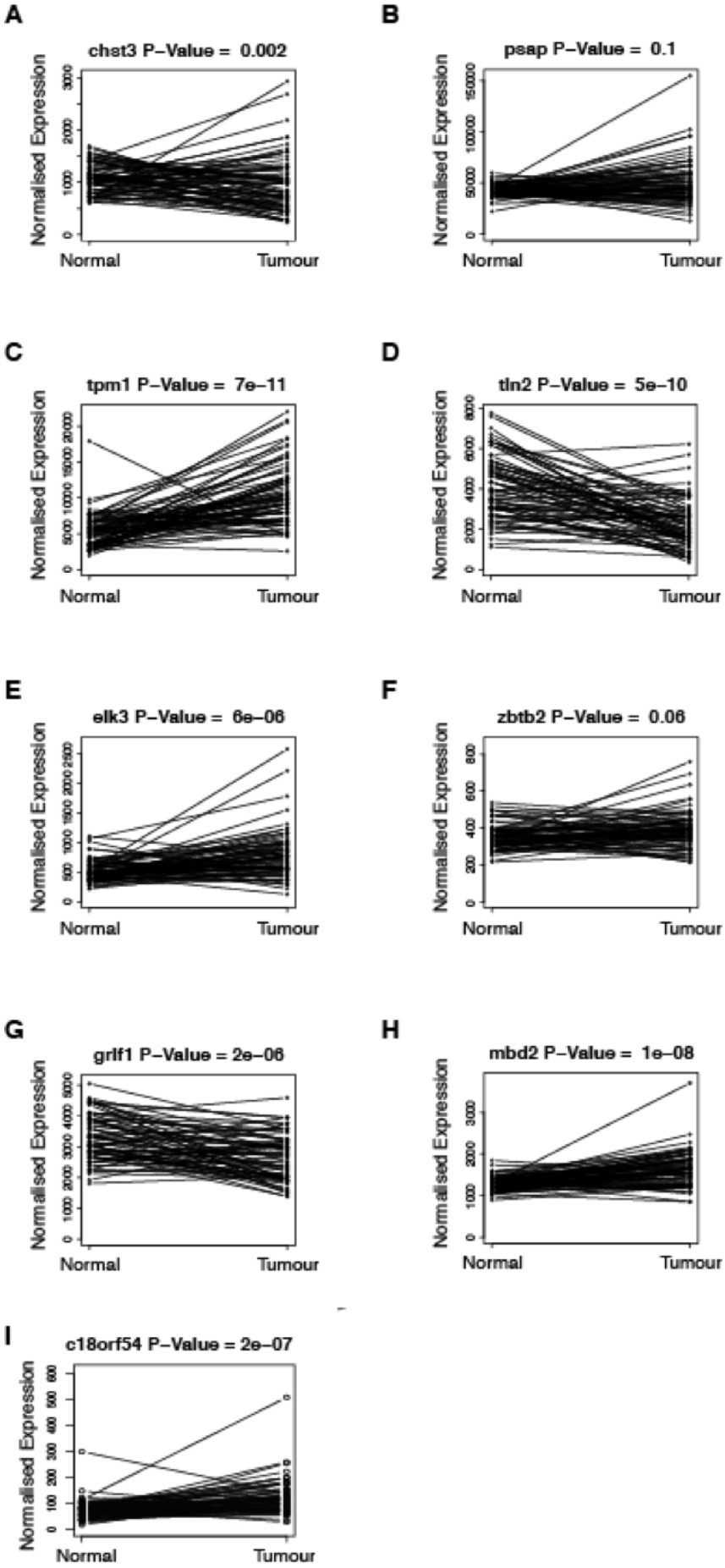

A
chst3 P−Value = 0.002
B
psap P−Value = 0.1
C
tpm1 P−Value = 7e−11
D
tln2 P−Value = 5e−10
E
elk3 P−Value = 6e−06
F
zbtb2 P−Value = 0.06
G
grlf1 P−Value = 2e−06
H
mbd2 P−Value = 1e−08
I
c18orf54 P−Value = 2e−07
Normalised Expression
Normal
Tumour

Figure 6.5: Gene expression in ccRCC compared to normal kidney. Gene expression in 72 ccRCC tumours and matched surrounding normal renal tissue was assessed using RNA-seq data from the TCGA KIRC cohort. Expression was normalised to the number of reads in each sample and the transcript length. Differential expression was assessed using a Mann-Whitney-Wilcoxon Test. Note GRLF1 is an alias for AHRGAP35.

Finally, I examined whether the genotype at these loci affected the expression of the putative HIF-target genes. Since both HIF-binding sites and HIF-inducible genes had been defined in ccRCC cells (and since both vary widely between cell types), I specifically examined the effect of the genotype on gene expression in clear cell kidney cancers, using SNP profiling in RCC patients in the TCGA KIRC cohort together with RNA-seq datasets from their respective kidney cancers. Only CHST3 and TLN2 expression was significantly associated with the genotype at the sub-threshold RCC-susceptibility loci, although for CHST3 expression was highest in the heterozygotes. Taken together this analysis of the TCGA datasets does not provide compelling evidence for the authenticity of the sub-threshold loci, and therefore further work will be required to explore their effects.

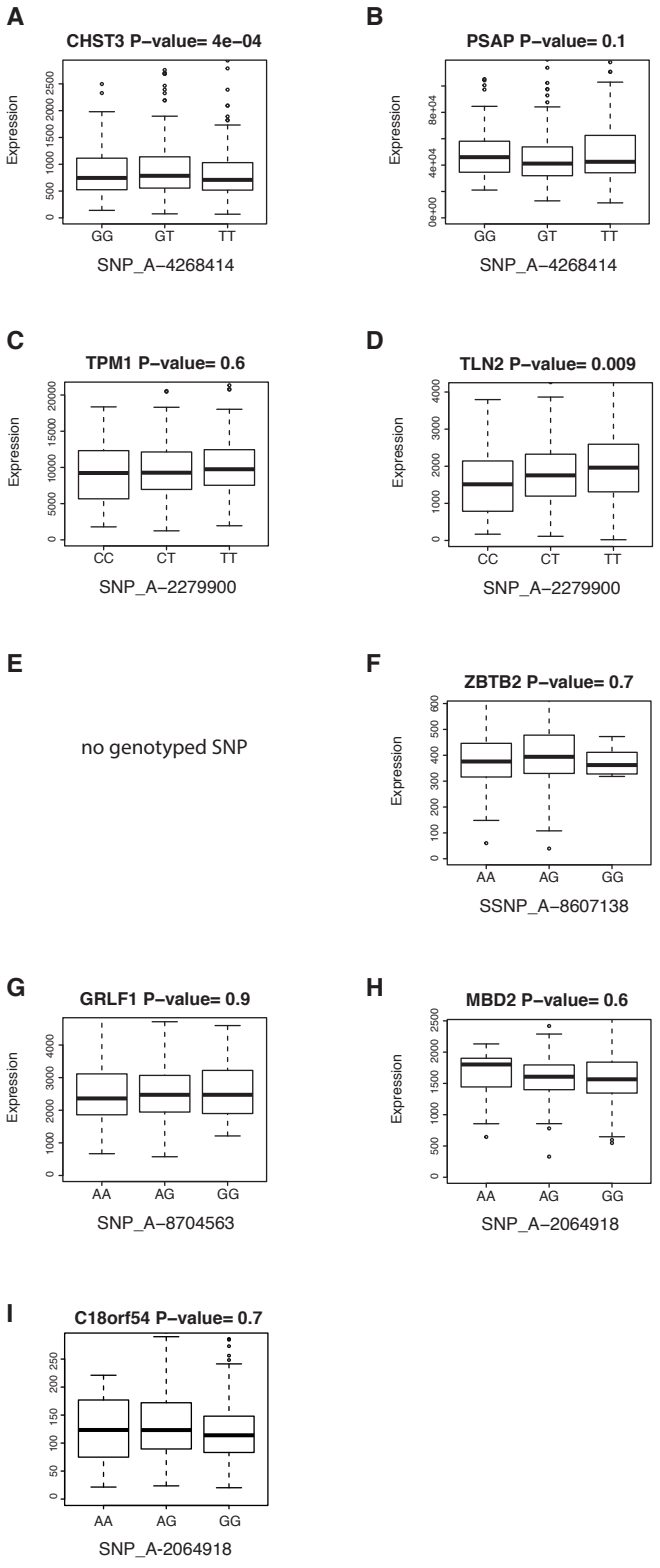


Figure 6.6: Box-and-whisker plots showing eQTL analysis for HIF-induced genes at the sub-threshold RCC-susceptibility loci that directly overlap a HIF-binding site.

6.2.7 The 7p21.1 locus and AHR

Finally, on manual inspection of the sub-threshold RCC-susceptibility loci that overlap with HIF-binding sites, it was noted that the 7p21.1 locus was adjacent to the AHR gene. AHR was further detected as one of the strongest VHL-regulated gene candidates in the GSEA (Figure 6.1). The gene encodes for the aryl hydrocarbon receptor protein, which is another basic helix-loop-helix transcription factor in the same family as HIF. Indeed, AHR shares the same dimerization partner ARNT (aryl hydrocarbon nuclear translocator, also known as HIF-1β) as HIF-1α and HIF-2α (Kobayashi et al. 1996; Wang et al. 1995). Previous work by others has shown that HIF-α and AHR may compete for ARNT/HIF-1β, thereby generating cross-talk between the two transcriptional pathways (Chan et al. 1999). Although there were weak interactions observed by Capture-C between the HIF-binding site at the 7p21.1 locus and the AHR promoter, these were not significantly above background signal. Furthermore, no other promoters of coding genes interacted with this site. However, expression of AHR was induced by HIF when 786-O cells stably transfected with control vector were compared with those transfected with wild-type VHL in RNA-seq analyses. Furthermore, the HIF-binding site at this locus is closely associated with the promoter of the long non-coding RNA transcript KCCAT333 (Renal Clear Cell Carcinoma-Associated Transcript 333). Regulation of coding transcripts in cis through HIF-dependent regulation of a neighbouring non-coding transcript has recently been described (Brown et al. 1991) (Lauer et al – in review). Therefore, it is possible that a similar mechanism may act to regulate AHR.

As a further test of HIF-dependent AHR induction in RCC, I examined expression levels of AHR in ccRCCs and compared them to levels in the surrounding kidney using RNA-seq data from the TCGA KIRC cohort (Figure 6.7 B). AHR expression was significantly higher in the ccRCC tumours compared to the surrounding kidney (P-value = $3*10^{-9}$) consistent with HIF-driven expression in ccRCC.

I next examined whether the germline genotype at the 7p21.1 locus was associated with tumour expression levels of AHR mRNA in the RNA-seq datasets from the TCGA KIRC cohort. The genotyped SNP in highest LD with the index SNP rs11764350 at the 7p21.1 locus was SNP_A_1803312. Importantly, the genotype of this SNP showed a highly significant correlation (P-value = $7*10^{-6}$) with AHR expression. Taken together, these data strongly suggest that the HIF-binding site is regulating expression of AHR (possibly through indirect regulation of mRNA levels by a clear-cell carcinoma-associated transcript) and that this induction is altered at by the RCC-associated polymorphism (Figure 6.7 C).

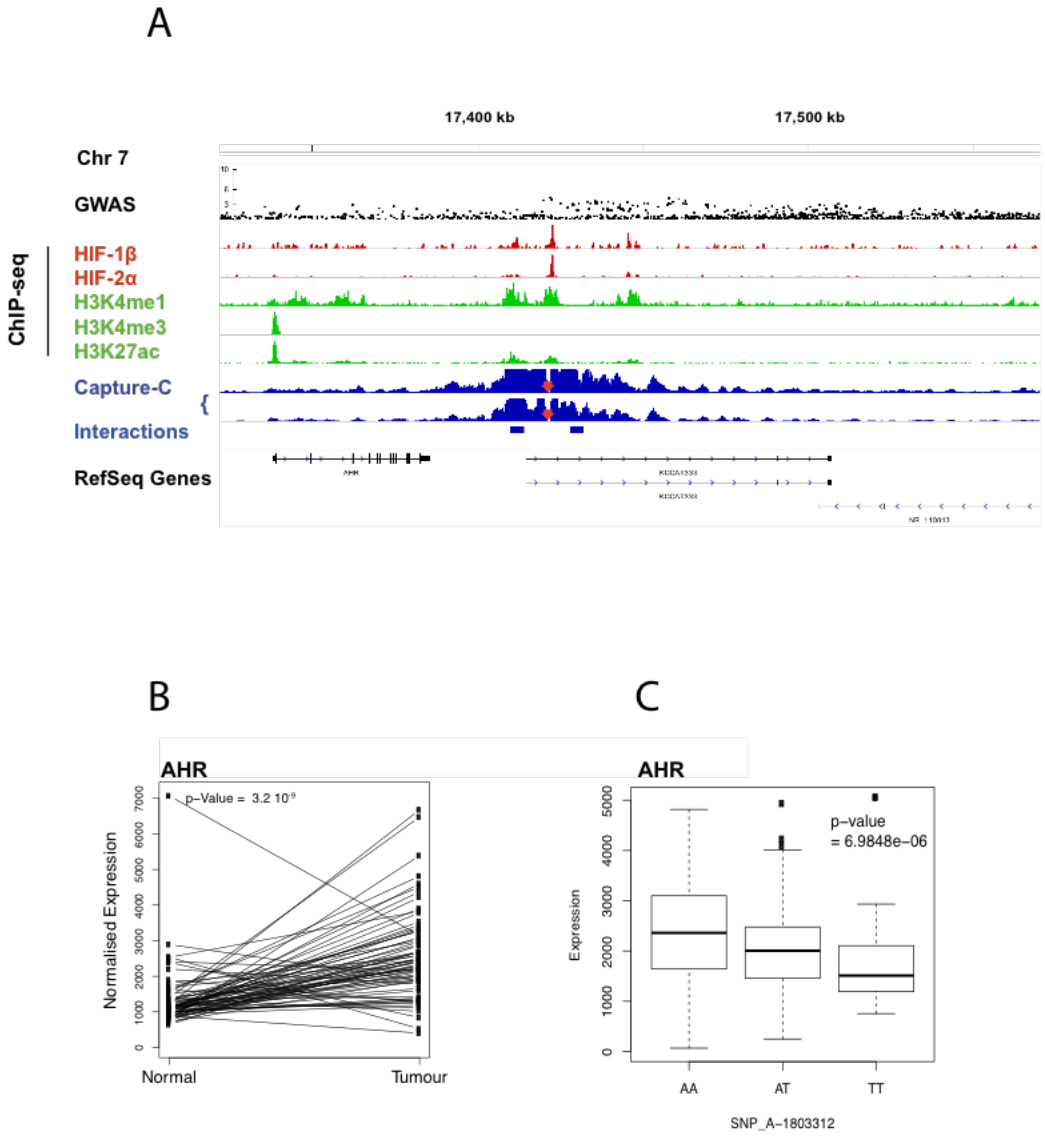

A
17,400 kb
17,500 kb
Chr 7
GWAS
ChIP-seq
HIF-1β
HIF-2α
H3K4me1
H3K4me3
H3K27ac
Capture-C
Interactions
RefSeq Genes
B
AHR
p-Value = 3.2 10-9
Normalised Expression
Normal
Tumour
C
AHR
p-value
= 6.9848e-06
Expression
AA
AT
TT
SNP_A-1803312

Figure 6.7: Integrative Genomics Viewer (IGV) of the 7p21.1 sub-threshold RCC GWAS locus. (A) Track 1 displays GWAS SNP-level P-values from the RCC GWAS meta-analysis. Track 2-3 (in red) display binding of HIF-2α and HIF1-β mapped in 786-O cancer cell lines of RCC. 4-6 from top (in green) display ChIP-seq signal (normalised read counts per million total reads at each chromosomal position) for histone H3K4me1, H3K4me3 and H3K27ac chromatin marks mapped in the 786-O kidney cancer cell line. The bottom tracks show Capture-C data with assays performed in 786-O cells using "viewpoint" (bait site) oligonucleotides at each active enhancer that coincided with the linked SNPs at the sub-threshold susceptibility locus ($r^2 \geq 0.8$ with the index SNP) and is indicated by the red arrows. Regions with Capture-C interaction (blue bars) and in black gene annotation from the RefSeq hg19 (GRCh37) build is displayed. (B) AHR mRNA expression in 72 paired ccRCC tumours and surrounding kidney samples in the TCGA-KIRC database. (C) Box-and-whisker plots showing eQTL analysis for the genotyped SNP in highest LD with the index SNP at 7p21.1 and the AHR gene.

6.3 Discussion

In chapter 6, I observe an enrichment of both HIF-binding sites and HIF-induced genes at and near to RCC-susceptibility loci that are associated with RCC at levels that fall below conventional thresholds of genome-wide significance. This is similar to that observed for loci that surpass these conventional significance levels and strongly suggests that at least a sub-set of these sub-threshold loci are true predisposing loci. If the GWAS cohort size were increased further, this set of loci may reach genome-wide significance. However, the level and statistical significance associated with the overlap between these sub-threshold loci and the HIF pathway is less than for RCC-susceptibility loci that surpass conventional levels of genome-wide significance. This is to be expected, since as the level of statistical significance decreases, an increasing proportion of false positive associations will be included, which will reduce the overall enrichment for biologically relevant features.

Importantly, approximately twice as many sub-threshold RCC-susceptibility loci coincide with a HIF binding site than expected. By extrapolation, this might suggest that roughly half of sub-threshold overlapping with HIF binding sites loci are true RCC predisposing loci. These loci may account partially for the missing heritability in GWAS, although it is unlikely that this will fully explain the deficit estimated by Scelo et al. However, GWA studies, by definition only integrate information from common variants and therefore low frequency variants are not examined in the majority of these studies. It is therefore likely that these rare undetected variants, as well as sub-threshold signal from common variants, will also account for some of the missing heritability. While the advent of whole-

genome-sequencing (WGS) makes it possible (if expensive) to detect low frequency variants, without reference to known common polymorphisms, it does not help with interpreting sub-threshold signals. However, the observation that roughly half of the sub-threshold RCC-susceptibility loci that overlap a HIF-binding site are likely to be true, provides a powerful mechanism for prioritising sub-threshold loci. Indeed, if only half of the loci prioritised in this way prove ultimately to be truly associated with RCC, this still represents a considerable advance. Further, genes in proximity to sub-threshold loci are enriched amongst those regulated by VHL which supports the on a functional level that RCC-susceptibility polymorphisms may be present beyond those currently defined as significant in GWAS, and that at least part of this set of susceptibility loci mediates their effect via the VHL/HIF pathway.

However, the specific enrichment of RCC-susceptibility loci at cis-acting components of the HIF-pathway raises another possibility to examine for genotypic associations with RCC. In GWAS analyses, the high level of statistical stringency required (P-value < $5*10^{-8}$) is determined by the large number of comparisons performed (approximately 1 million for a full GWAS). At the other extreme, simple association analyses that examine the link between a single polymorphism in a candidate gene and a disease do not need to allow for multiple comparisons and would use a threshold of 0.05. Therefore, if the number of polymorphisms being studied was limited by focussing on those that overlap with a known pathway (e.g. HIF-binding sites), then an intermediate statistical threshold would apply, which would increase the power of the study. This would essentially be a HIF-wide association study (HWAS) rather than a genome-wide

association study (GWAS). Such a methodology may also be applicable to other pathways that may be activated in different cancer-types if these too were enriched at susceptibility loci for those tumours.

Lastly, identification of potential transcriptional target of sub-threshold loci by eQTL analysis and differential expression analysis between tumour and normal kidney has failed to provide compelling evidence for target gene candidacy. It may be that sub-threshold loci may have smaller effects on their targets which would make the detection of an eQTL particularly difficult. Especially, since tumour expression data is infiltrated by multiple different cell-types. And since my findings in the previous chapters suggest that RCC susceptibility loci reveal their effect within the tumours cells, the infiltration with non-tumour cells may compromise this type of analysis. Further, expression of tumour DNA was measured in late-stage RCC whilst the loci may propagate their effect at different stages during cancer development.

Chapter 7: Analysis of cancer-associated GWAS loci in the light of transcriptional pathway activation

7.1 Introduction

Disease-associated polymorphisms are germline variants. By definition, as such they are present in all cells of the human body. Similarly, the VHL-HIF system is also ubiquitous and HIF can be activated in a vast variety of different cell types (Maxwell et al. 1993; Firth et al. 1995; Semenza et al. 1994). However, despite the general operation of both, cancer-susceptibility loci and the VHL-HIF mediated transcriptional response exhibit a high degree of tissue specificity.

Cancer-associated polymorphisms, whilst present in all tissues of the human body, are often private in their predisposition to a single cancer type. For example, GWA studies have identified a susceptibility locus at the SMAD7 gene locus which appears to be private to colorectal cancer, as no other cancer types have been associated with this specific susceptibility locus to date (Broderick et al. 2007). However, other susceptibility loci may be shared between cancer types or affect the same target gene. Indeed, the 8q24 susceptibility locus is common to several cancer types and can influence expression of the MYC gene in renal, prostate, breast and colon cancers (Sotelo et al. 2010).

Similarly, the action of VHL as a tumour suppressor appears to be highly tissue-specific (Seizinger et al. 1988). Further, both HIF activated transcriptional targets and the sites to which HIF binds differ greatly between different cell types (Smythies et al. 2019). This is true not only for genes that are uniquely expressed in one cell type but not the other, but it is also conceivable that a gene which is expressed in many cell types might only be a HIF target gene in a small subset

of these cells. Such specificity in gene regulation is often the result of cell-type specificity in the enhancers that regulate the gene.

It is highly likely that these cell-type specific transcriptional responses play a role in determining the effects of often intergenic ubiquitously present germline polymorphisms. In complex phenotypes, like cancer where multiple effects of different cell types accumulate to promote the phenotype, it can hence be difficult to know in what cell-type the interplay between susceptibility locus and transcriptional system occurs.

However, the excess overlap between RCC-susceptibility polymorphisms and cis-acting elements of the HIF transcriptional pathway in RCC cells observed in chapter 2 and 3 might indicate that a proportion have their effect in the cancer cells themselves as opposed to affect a cell type associated with systematic importance in promoting cancer. To test this, in this chapter, I compare the interplay between renal cancer-associated polymorphisms and the binding sites of HIF in renal cancer cells to the interplay between renal cancer-associated polymorphisms and cancer cells of non-renal origin. Specifically, I explore if HIF-binding sites affected by RCC-susceptibility loci exhibit a degree of specificity for renal cancer cells.

Further, since HIF has an established role in the pathogenesis of many cancer types associated with pathway activation by micro-environmental tumour hypoxia, I test if susceptibility loci of non-renal cancers are enriched for HIF-

binding in the corresponding tissue of origin of these cancer types. This pan-cancer analysis assesses the association of cancer-susceptibility loci with binding sites of HIF across further 4 common cancer types of breast, prostate, lung and colon and hence tests, if the association observed in Chapter 2 and 3 is a unique feature of RCC.

Finally, I explore ways that the approach introduced in chapter 2 and 3 can be extended to other cancer types associated with transcriptional pathways other than HIF, such as estrogen- and androgen-receptor signalling in breast and prostate cancer. This final part also assesses the generalisability of the approach proposed in the context of this thesis to study pre-disposing non-coding variants in cancer.

7.2 Results

7.2.1 Definition of high confidence HIF-1β binding sites in different cancer cell lines

I first attempted to define HIF binding sites in A549 lung cancer cells, HCT-116 colorectal carcinoma cells, PC3 prostate cancer cells and T47D oestrogen receptor positive (ER+) breast cancer cells by reference to HIF-1β ChIP-seq datasets performed following 16 hours hypoxic incubation. ChIP-seq assays were performed in duplicate by Olivia Lombardi and, at the time this work was undertaken, only HIF-1β datasets were available. However, HIF-1β is the common dimerization partner of both HIF-1α and HIF-2α and hence will identify both HIF-1 and HIF-2 binding sites. To allow a more direct comparison with data in RCC, ChIP-seq analysis of HIF-1β sites in 786-O cells alone were used to define HIF-binding in RCC cells in this chapter. All datasets were handled using the same analysis pipeline outlined in chapter 2.

To assess the quality of the datasets, I first examined the signal-to-noise ratio in the vicinity of the identified peaks as in chapter 2. The summit of the peak (chromosomal location corresponding to maximum ChIP-seq signal) is first identified and arranged on the vertical scale such that their summits aligned horizontally. The ChIP-seq signal was then plotted on a colour scale (higher signal = deeper red) for each chromosomal position (± 2kb) flanking the summit on the horizontal scale to give the corresponding heatmaps (Figure 7.1). The ChIP-seq signal at each position flanking the summit was then averaged across all peaks to give the corresponding line graphs. In all cases, there was a strong,

narrow spike of ChIP-seq signal extending ± 200-300 bp (the average fragment size of the chromatin following sonication) with low background signal.

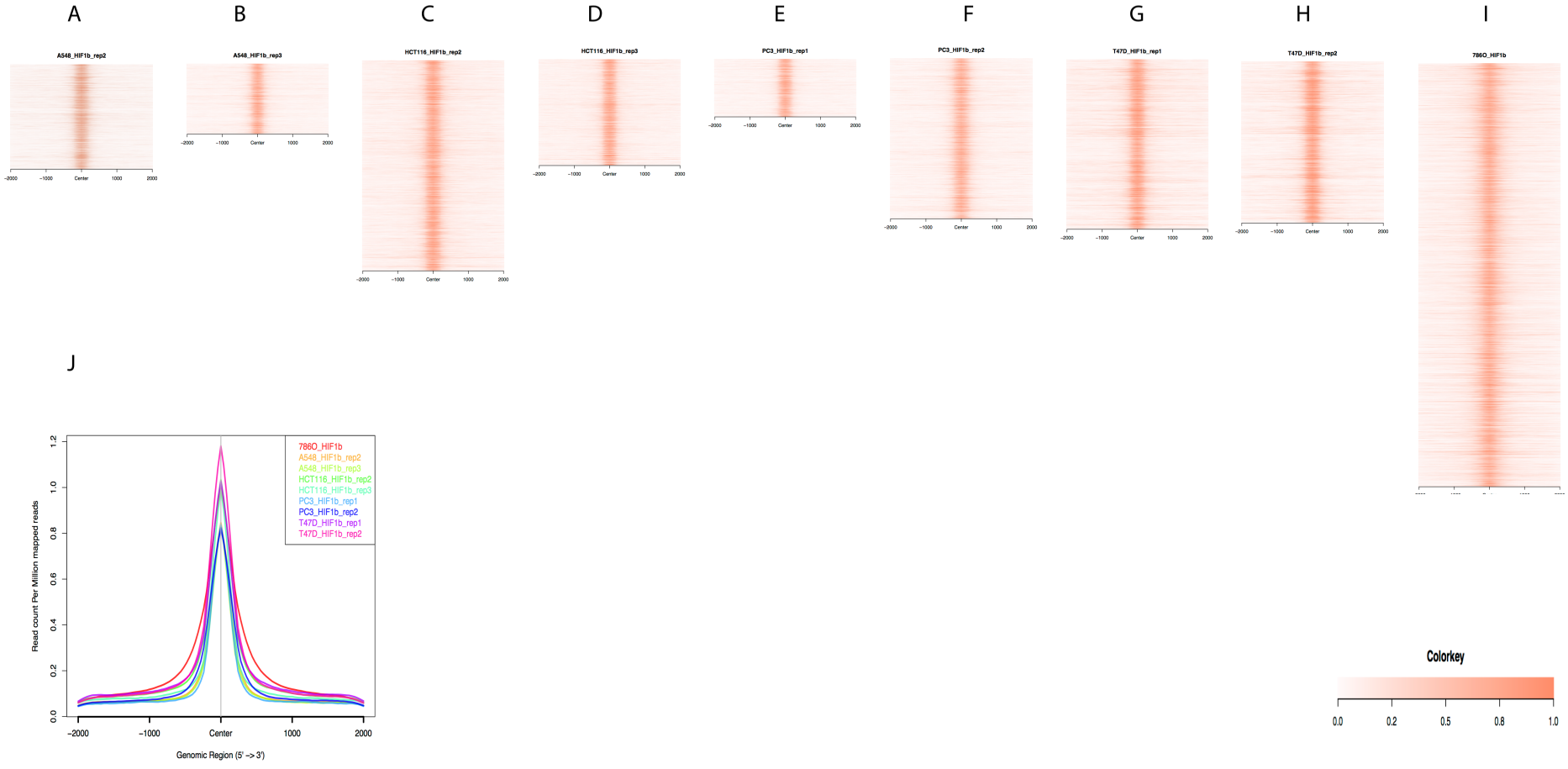

A
A548_HIF1b_rep2
B
A548_HIF1b_rep3
C
HCT116_HIF1b_rep2
D
HCT116_HIF1b_rep3
E
PC3_HIF1b_rep1
F
PC3_HIF1b_rep2
G
T47D_HIF1b_rep1
H
T47D_HIF1b_rep2
I
786O_HIF1b
−2000
−1000
Center
1000
2000
J
786O_HIF1b
A548_HIF1b_rep2
A548_HIF1b_rep3
HCT116_HIF1b_rep2
HCT116_HIF1b_rep3
PC3_HIF1b_rep1
PC3_HIF1b_rep2
T47D_HIF1b_rep1
T47D_HIF1b_rep2
Read count Per Million mapped reads
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Genomic Region (5' -> 3')
Colorkey
0.0
0.2
0.5
0.8
1.0

Figure 7.1: Heatmaps and average peaks profile of HIF-1β binding sites mapped by pan-genomic Chromatin Immunoprecipitation (ChIP-seq).

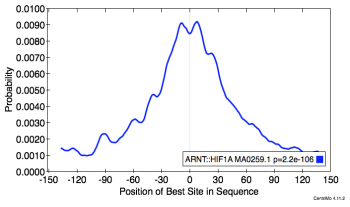
ChIP-seq assays were performed in (A, B) A549 lung cancer cells, (C, D) HCT-116 colorectal carcinoma cells, (E, F) PC3 prostate cancer cells, (G, H) T47D oestrogen receptor positive (ER+) breast cancer cells, and (I) 786-O cells. In the heatmaps, each row corresponds to one peak called in the corresponding ChIP-seq dataset. ChIP-seq signal correlates with colour intensity and is measured in read counts per million mapped reads (CPM) in the region of +/- 2 kb from the summit of the peaks called for each of the 7 datasets. (J) The summary graphs below the heatmaps show the average signal intensity (CPM) for all peaks in each dataset against the distance from the summit of the peaks over a (+/- 2 kb) window centred at the peak summits. Signal to noise and peak characteristics were similar in different cells and similar to the RCC datasets used in previous chapters.

Secondly, I examined whether the anticipated consensus HRE motif was enriched within each of the defined sets of peaks. For each set of peaks, the MEME suite of motif-based sequence analysis tools was used to interrogate the JASPAR core database of motif matrices for enrichment. In each case, the core consensus HRE motif, MA0259.1 (Figure 7.2 A), or a closely related variant, was the most highly significantly enriched, within the defined peaks. Furthermore, when analysed using the CentriMo tool, this motif was centrally enriched within the defined peak regions in all 7 datasets (Figure 7.2 B–H), providing further evidence for direct binding of each of the immunoprecipitated HIF-1β isoforms to the core consensus HRE motif.

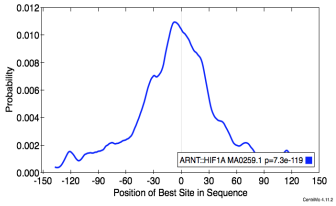
A



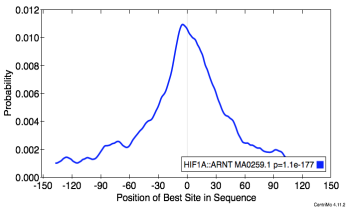
B


Probability
Position of Best Site in Sequence
ARNT::HIF1A MA0259.1 p=2.2e-106

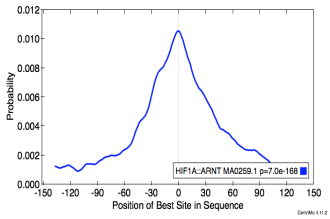
C


Probability
Position of Best Site in Sequence
ARNT::HIF1A MA0259.1 p=7.3e-119

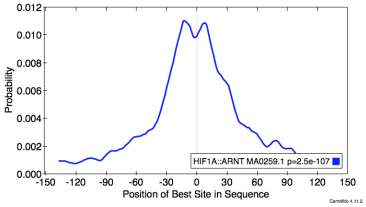
D


Probability
Position of Best Site in Sequence
HIF1A::ARNT MA0259.1 p=1.1e-177

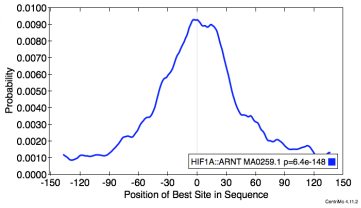
E


Probability
Position of Best Site in Sequence
HIF1A::ARNT MA0259.1 p=7.0e-168

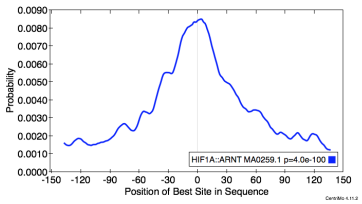
F


Probability
Position of Best Site in Sequence
HIF1A::ARNT MA0259.1 p=2.5e-107

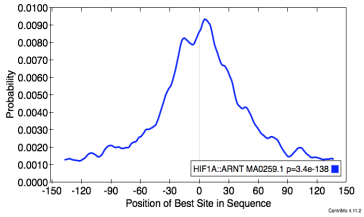
G


Probability
Position of Best Site in Sequence
HIF1A::ARNT MA0259.1 p=6.4e-148

H


Probability
Position of Best Site in Sequence
HIF1A::ARNT MA0259.1 p=4.0e-100

I


Probability
Position of Best Site in Sequence
HIF1A::ARNT MA0259.1 p=3.4e-138

J

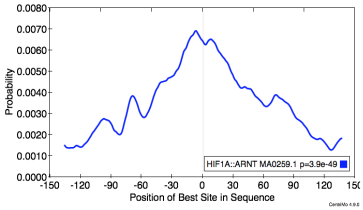

Probability
Position of Best Site in Sequence
HIF1A::ARNT MA0259.1 p=3.9e-49

Figure 7.2: Motif Enrichment Analysis of 7 HIF ChIP-seq datasets mapped in cancer cell lines. (A) MEME suite motif-based sequence analysis of identified peaks demonstrates enrichment of the core HRE motif (MA0259.1) in all datasets. CentriMo analysis of the frequency distribution of this HRE motif within the defined peaks shows central enrichment in (B,C) A549 lung cancer cells, (D,E) HCT-116 colorectal carcinoma cells, (F,G) PC3 prostate cancer cells, (H,I) T47D oestrogen receptor positive (ER+) breast cancer cells, and (J) 786-O RCC cells.

Finally, I compared the 9 HIF-1β ChIP-seq datasets from the 5 cancer types with each other. For each dataset the number of called peaks is listed in Table 7.1.

ID	*Tissue*	*Intervals*
A549_HIF1b_rep1	lung	1768
A549_HIF1b_rep2	lung	1270
HCT116_HIF1b_rep1	colon	3790
HCT116_HIF1b_rep2	colon	1913
PC3_HIF1b_rep1	prostate	1042
PC3_HIF1b_rep2	prostate	2891
T47D_HIF1b_rep1	breast	3061
T47D_HIF1b_rep2	breast	2910
786O_HIF1b	kidney	7621

Table 7.1: Summary of the number of peaks identified in 9 HIF-1 ChIP-seq datasets mapped in cancer cell lines. Peaks were identified using the MACS and TPIC peak callers and the number identified for each of the 9 datasets is shown.

DiffBind was then used to define the number of datasets in which each peak was identified. Binding sites were first combined and where peaks from more than one dataset overlapped, the total region was merged into one peak, giving a total of 13,008 non-overlapping binding sites that were identified in one or more of the datasets. Of these 5,198 were identified in two or more datasets, 2,893 in three or more and so on (Figure 7.3A).

I then compared the different ChIP-seq datasets. For peaks present in two or more datasets, I used the MACS defined probability score to generate a correlation coefficient between each pair of datasets using DiffBind. Each dataset was then clustered according to these correlation coefficients and the overall result plotted as a heatmap in which perfect correlation (i.e. the correlation of a dataset with itself) is represented in dark green and zero correlation is represented in white. (Figure 7.3B). As expected, the two replicates of the HIF-1β ChIP-seq analyses for each cell type correlate most closely with each other. Interestingly the RCC dataset clustered separately from the other cell lines in this analysis.

Next, for each of the 5,198 binding sites that were present in more than one dataset the signal intensity in each of the 9 HIF-1β ChIP-seq datasets was determined. Using the resulting 5198 x 9 matrix, the 9 ChIP-seq datasets were then analysed using principal component analysis (Figure 7.3B). Again, each of the duplicate datasets was most closely related to each other than to datasets from the other cell lines. Although 786-O cells were less closely related to A549, PC3 and HCT-116 cells on the 1st principal component, T47D breast cancer cells were the least closely related to other cell types. Unsupervised hierarchical clustering was also applied to the same data matrix confirming similar relationships between the different cell lines (Figure 7.4D).

Given the close association between the duplicates for each cell line, only peaks present in both replicas were considered for further analysis. The number of peaks present in both replicas was 970 for PC3 cells, 1,044 for A549 cells, 1,767 for HCT-116 cells and 2,081 for T47D cells (Table 7.2).

cancer	*GWAS loci*	*Cell-type*	*HIF-1β peaks*
RCC	13	786-O	7621
Breast	124	T47D	2081
Prostate	99	PC3	970
Lung	18	A549	1044
Colon	66	HTC116	1767

Table 7.2: Number of independent cancer-susceptibility loci and number of HIF-1β binding sites mapped in the cancer cell lines corresponding to the tissue of origin of the cancer type.

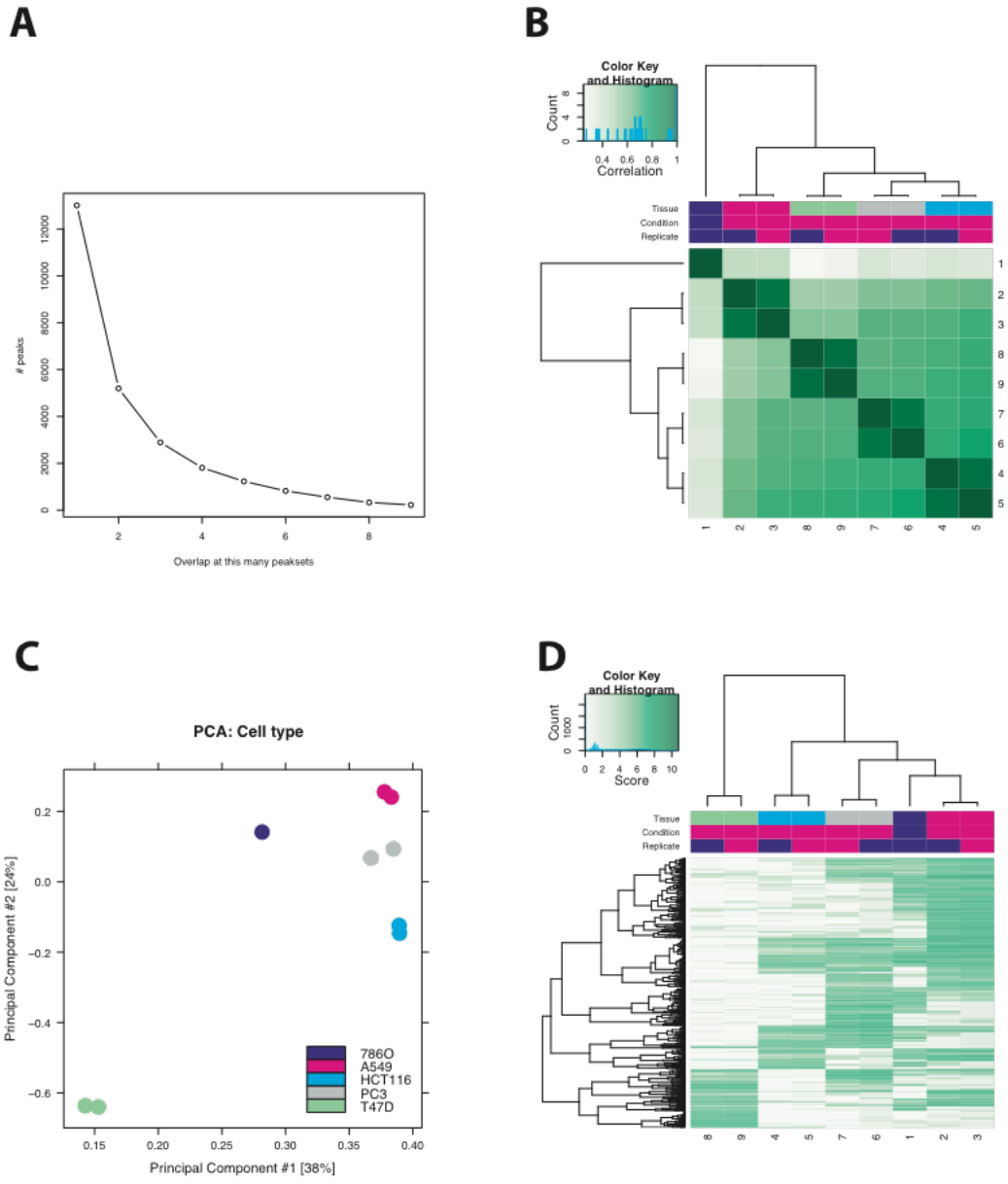


Figure 7.3: Comparison of HIF-1β ChIP-seq analyses across different cancer cell lines. (A) The number of peaks (y-axis) shared between the designated number of datasets (x-axis) is shown as a line graph. (B) For each pair of datasets in turn, the MACS probability scores, for each peak, were correlated and the correlation coefficients plotted as a heatmap using a colour scale in which perfect correlation (r^2=1) is plotted in dark green and no correlation (r^2=0) is plotted in white. Unsupervised hierarchical clustering was then

performed using these correlation coefficients and is represented by the dendrograms. For each of the 5,198 binding sites that were present in 2 or more datasets, the ChIP-seq signal intensity was then determined. Tissue indicates here the cell line whilst condition refers to the RCC vs non-RCC state. (C) Principal component analysis of ChIP-seq signal intensity across these sites. (D) Unsupervised hierarchical clustering of the same sites. In each case the duplicates from each cell line were more closely related to each other than they were to the datasets form other cell lines. Tissue indicates here the cell line whilst condition refers to the RCC vs non-RCC state.

7.2.2 Overlap between RCC susceptibility loci and HIF-binding sites identified in different cancer cell types

By definition, germline polymorphisms are present in all cell types within the body. However, those polymorphisms associated with RCC are generally not associated with pre-disposition to cancer in these other tissue types. This raises questions as to how a variant that is present in all cells can pre-dispose to cancer in one cell type alone. Possible explanations for this are that these loci do not have a biological role in other cell types or that they play a different role. Therefore, given the overlap between RCC-polymorphisms and the biological output of the HIF pathway in RCC-cells, I next examined whether the HIF-binding sites that overlapped RCC-susceptibility loci were unique to RCC or observed in other cancer types as well.

The 13 RCC susceptibility loci identified in chapter 3, were aligned with HIF-1β binding sites defined in each of non-RCC cell lines above (Table 7.3). For comparison, HIF-1β binding sites in 786-O RCC cancer cells was also performed. Five RCC loci overlap HIF-1β binding sites in 786-O RCC cell lines. Four of these (11q13.3, 12p12.1, 8q24.21 and 3q26.31) had also overlapped the more strictly defined set of RCC HIF-binding sites used in chapter 3 and the fifth locus (12q24.31) was adjacent to one of the sites in the earlier more stringent HIF-ChIP dataset. Only two of these loci (12q24.31 and 8q24.21) overlapped with HIF-1β binding sites identified in non-RCC cell lines. Furthermore, in each of the non-RCC cell lines, no more than one RCC-susceptibility locus overlapped with a HIF-1β binding site.

locus	*rs numbers*	*P-values*	*HIF-1β binding*				
			786-O	A549	HCT116	PC3	T47D
11q13.3	rs11263654	1.65E-23	x				
2p21	rs2121267	6.95E-19					
12p12.1	rs11534749	6.19E-17	x				
14q24.2	rs28840762	1.08E-15					
12q24.31	rs10846748	1.70E-12	x		x		
8q24.21	rs6470588	2.82E-12	x	x			x
11q22.3	rs117706999	6.60E-10					
2q22.3	rs11888238	8.15E-10					
1p32.3	rs6676515	3.04E-09					
4q23	rs7697932	4.70E-09					
3p22.1	rs9821249	1.12E-08					
15q22.31	rs12905354	2.28E-08					
3q26.31	rs234043	4.10E-08	x				

Table 7.3: RCC cancer susceptibility loci and their coincidence with HIF-1β binding sites mapped in 786-O RCC cells, PC3 prostate cancer cells, T47D oestrogen receptor positive (ER+) breast cancer cells, A549 lung cancer cells, and HCT-116 colorectal carcinoma cells. Genome-wide significant RCC susceptibility locus, rs number of the index SNP, P-value of the index SNP and presence (+) or absence (-) of an overlapping HIF-1β binding site in each of the cancer cell lines is indicated.

To determine the statistical significance of this overlap, I again generated frequency distributions for the expected overlap in each of the four non-RCC cancer as before (Figure 7.4). This confirmed that RCC-susceptibility loci coincide with HIF-1β binding sites in 786-O RCC cancer cell lines significantly more than expected, as previously observed for the more strictly defined HIF-binding sites in chapter 3. Conversely, the overlap between RCC-susceptibility loci and HIF-1β binding sites in non-RCC cell lines did not differ significantly from that expected by random chance.

Taken together, this suggests that RCC-susceptibility loci preferentially affect HIF-binding sites that exhibit a degree of specificity towards RCC cells. However, this is not the case for all of the RCC-susceptibility loci, although interestingly, the 8q24.21 locus, which is the least specific is also associated with pre-disposition to other cancer types as well. Furthermore, the survey of HIF-binding site in different cancer cell lines is not an exhaustive survey across the whole of human biology and apparently specific binding sites may turn out to be occupied in as yet untested cell lines. However, such a survey would be almost impossible, since relatively rare cell types are poorly defined, and patterns of HIF-binding may also be influenced by developmental stage and other environmental factors.

Furthermore, binding in non-RCC cell lines was tested under laboratory conditions that were designed to maximise the HIF response. It is likely that physiological levels of HIF in these tissues (and in early tumour lesions) will be much lower than this and also much lower than those observed in VHL-defective

clear cell kidney cancer lesions. Thus, specificity of binding may be provided not only by the ability to bind, but also by the levels of HIF present.

Overall, the relative specificity of some HIF-binding sites that overlap RCC-associated polymorphisms is consistent with the prior hypothesis that a significant proportion of these variants have a direct impact on the biology of the renal cancer cell itself. It also provides a plausible mechanism by which RCC-associated polymorphisms may be specific to kidney cancer and not associated with other types of cancer since they will likely only have a biological effect in cell-types in which they are bound by an active transcription factor.

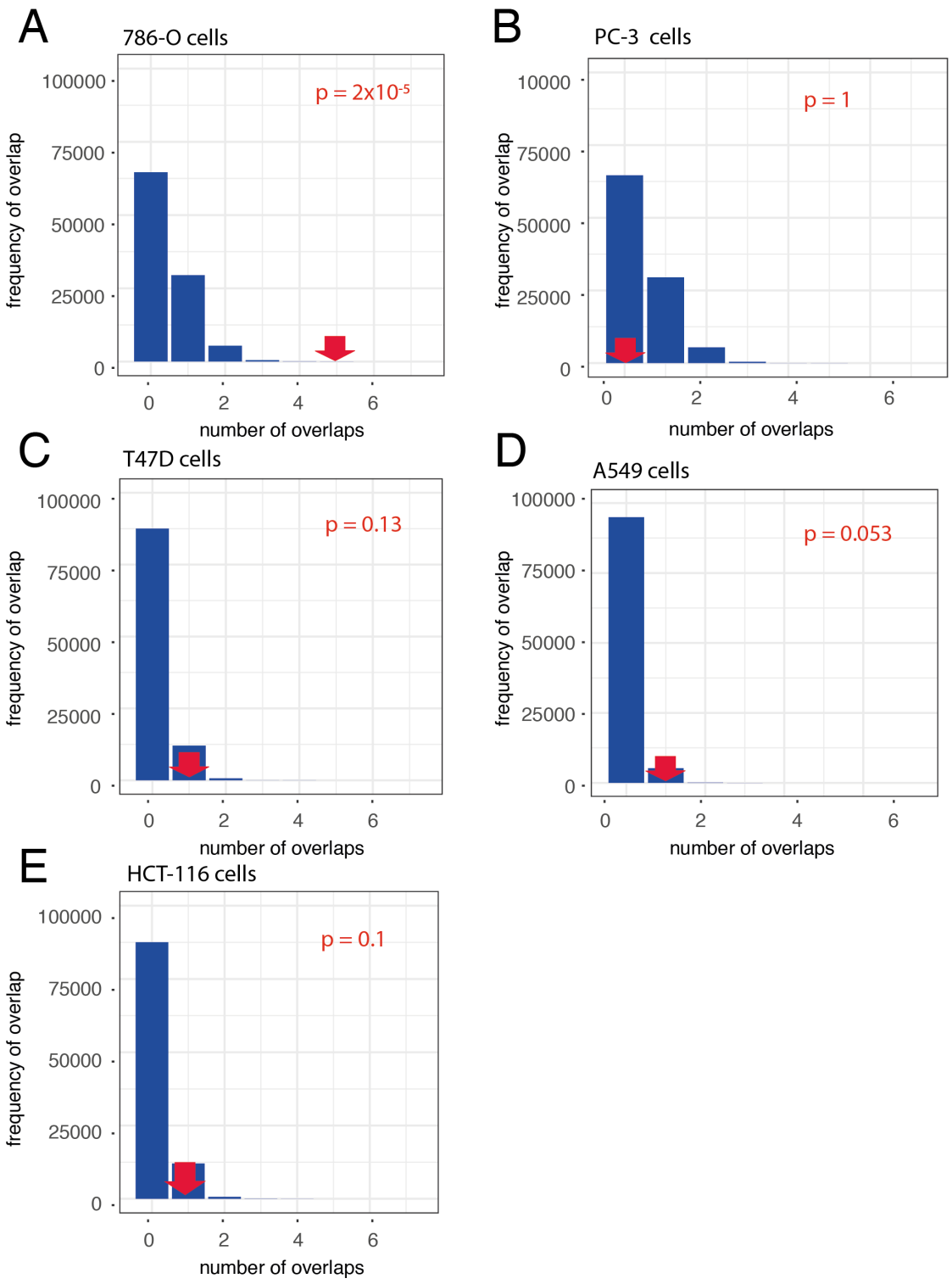


Figure 7.4: Overlap of RCC susceptibility loci with HIF-1β binding mapped in cancer cell lines of RCC, prostate, breast, lung and colorectal carcinoma.

RCC susceptibility loci were defined from GWAS analyses. HIF-1β binding sites were identified by ChIP-seq analysis for the 5 types of cancer cell lines. The expected overlap (blue bars) was generated and the observed overlap it indicated (red arrow) for: (A) RCC susceptibility loci and HIF-1β binding sites in 786-O cells, (B) RCC susceptibility loci and HIF-1β binding sites in PC-3 cells, (C) RCC susceptibility loci and HIF-1β binding sites in T47D cells, (D) RCC susceptibility loci and HIF-1β binding sites in A549 cells, and (E) RCC susceptibility loci and HIF-1β binding sites in HCT-116 cells. The observed overlap of RCC susceptibility with HIF-1β was more than expected (P-value≤0.5) for HIF-1β sites mapped in RCC cell lines 786-O.

7.2.3 Overlap between cancer-susceptibility loci and HIF-1β binding sites identified in corresponding cancer cell lines.

Although clear cell renal cancers manifest the most direct and profound activation of HIF as a consequence of inactivation of the HIF E3 ubiquitin ligase, VHL, activation of HIF is also a feature of other tumour types. In these cases, induction of HIF is largely attributed to the stabilisation of HIF-α subunits by intra-tumour hypoxia, although other mechanisms (e.g. regulation of HIF translation by mTOR pathways) may also operate. Activation of HIF in these tumours is almost universally associated with adverse prognosis (higher levels of HIF being associated with more rapidly progressive disease) (Zhong et al. 1999; Talks et al. 2000). A causal contribution of HIF to these outcomes is indicated by interventional studies in animal models in which suppression of HIF slows tumour growth, whilst its overexpression has been shown to accelerates it in a variety of contexts (Raval et al. 2005). Furthermore, examination of HIF-target genes in many types of cancer frequently reveals direct transcriptional targets of HIF with key roles in cancer pathways

Given the importance of HIF biology to non-renal as well as renal cancers, I next examined whether polymorphisms that pre-dispose to other cancer types might impact on HIF-binding sites as defined in a cell line relevant to each of those cancers. Specifically, I tested, whether cancer susceptibility loci associated with breast, lung, prostate and colon cancer overlapped HIF-1β binding sites, profiled in a cell line reflecting the corresponding tissue of origin, more than expected.

GWAS-defined susceptibility loci predisposing to breast, lung, prostate and colon cancer were defined by examining publicly available datasets provided by the National Human Genome Research Institute – European Bioinformatics Institute (NHGRI-EBI) GWAS catalogue (June 2018). For breast and lung cancer, full summary statistics were available, whilst for prostate and colon cancer, only genome-wide significant loci were reported (Buniello et al. 2019; McKay et al. 2017; Michailidou et al. 2017; Scelo et al. 2017).

In each case, index SNPs and their corresponding LD regions ($r^2 \geq 0.8$) were identified as described previously and examined for overlap with the HIF-1β binding sites. One prostate cancer susceptibility locus directly overlapped a HIF-1β binding site in PC-3 cells, three breast cancer susceptibility loci overlapped a HIF-1β binding site in T47D cells, none of the lung cancer susceptibility loci overlapped a HIF-1β binding site in A549 cells and none of the colorectal cancer susceptibility loci overlap HIF-1β binding sites in HCT-116 cells.

To assess if the observed overlap of the susceptibility loci with HIF binding sites was more than expected the same bootstrapping approach described in chapter 3 was applied (Figure 7.5). Apart from RCC, no significant overlap was identified between cancer-susceptibility loci and HIF-1β binding sites in any of the other 4 cancer types.

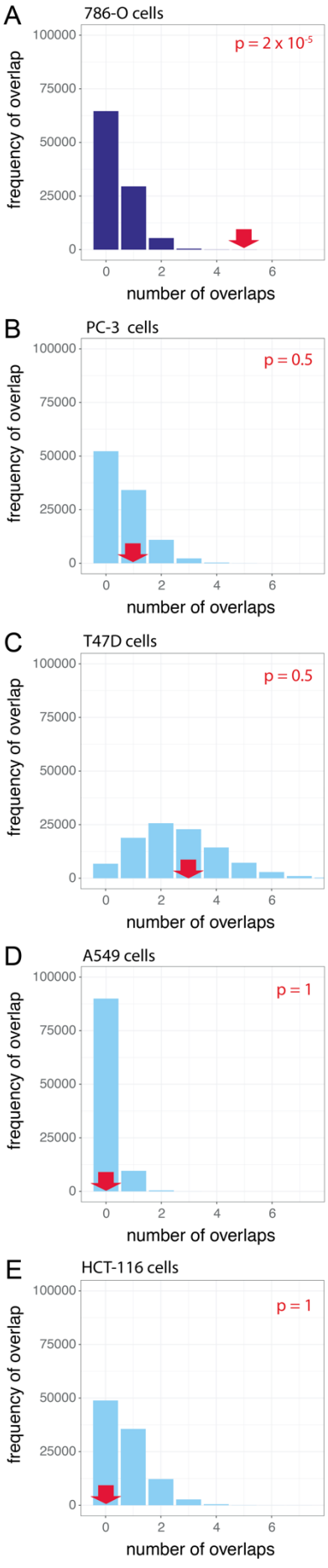

A
786-O cells
p = 2 x 10-5
B
PC-3 cells
p = 0.5
C
T47D cells
p = 0.5
D
A549 cells
p = 1
E
HCT-116 cells
p = 1
frequency of overlap
number of overlaps
100000
75000
50000
25000
0
0
2
4
6

Figure 7.5: Enrichment of cancer susceptibility loci at cell-type corresponding HIF -1β binding sites. Cancer susceptibility loci were defined from GWAS analyses performed for prostate, breast, lung or colorectal cancer as for RCC. HIF-1β binding sites were identified by ChIP-seq analysis in cell lines corresponding to the tissue of origin of each cancer type. The frequency distribution for the expected overlap (blue bars) was generated and the observed overlap is indicated (red arrow) for: (A) RCC susceptibility loci and HIF-1β binding sites in 786-O cells, (B) prostate cancer susceptibility loci and HIF-1β binding sites in PC-3 cells, (C) breast cancer susceptibility loci and HIF-1β binding sites in T47D cells, (D) lung cancer susceptibility loci and HIF-1β binding sites in A549 cells, and (E) colorectal cancer susceptibility loci and HIF-1β binding sites in HCT-116 cells. Renal cancer was the only cancer type for which a significant overlap of susceptibility loci with HIF-1β binding sites was observed (P-value = $2*10^{-5}$).

This contrast between renal cancer and non-renal cancers is striking. Given the multiple lines of evidence linking HIF to prognosis in these other tumour types (outlined above), the importance of HIF in the progression of late-stage non-renal tumours seems well established. However, there are major differences in both the degree and timing of HIF activation in (clear cell) kidney cancer compared to other cancer types. Inactivation of VHL (and consequent activation of HIF) is an early and truncal event in the pathogenesis of clear cell renal cancer and may

occur years or even decades before clinically detectable disease. In non-renal cancer types, HIF activation is more often a consequence of intra-tumour hypoxia, becoming more prevalent in larger, later stage tumours. Secondly, loss of functional VHL in clear cell renal cancers leads to full stabilisation of HIF in all tumour cells. In non-renal cancers oxygen gradients within the tumour will mean that in many tumour cells HIF is only partially activated. Therefore, there is a quantitative as well as temporal difference in HIF activation between clear cell renal cancer and other tumour types. This may, at least in part, explain why variants that are associated with the development rather than progression of the different cancer types are variably related to the HIF pathway in renal compared to non-renal cancers.

It is further possible, that in kidney cancer, the inappropriate activation of HIF and specific target genes in normoxic cells carries a penalty that needs to be overcome through the modification of the HIF response. The appropriate activation of HIF in hypoxic non-renal cancer cells may not carry the same penalty and therefore may not be under the same selective pressure.

Nevertheless, given the importance of HIF to the progression of non-renal tumour types, it might be interesting to speculate whether germline polymorphisms that pre-dispose to more rapid progression of non-renal tumours might be linked to the HIF-pathway. However, at present data necessary to study this possibility are lacking.

Finally, it might also be interesting to compare clear cell renal cancer (in which VHL is usually inactivated) with other kidney cancers such as papillary cell renal cancer, which do not manifest the same profound disruption of HIF regulation. However, GWAS analyses of kidney cancer do not distinguish according to histological sub-type and such analyses will be difficult to undertake at the appropriate scale, since these rarer types of tumour only account for about 15% of renal cancers.

7.2.4 Analysis of cancer susceptibility loci in the light of transcriptional pathway activation can be extended beyond the HIF pathway

So far, I have limited my analysis to the HIF-pathway, which shows a statistically significant overlap with cancer-susceptibility loci that is restricted to kidney cancer. In kidney cancer, the HIF pathway is constitutively activated as a truncal event in the pathogenesis of the cancer. This raises important questions about whether other large transcriptional networks that are active in different types of cancer and are important to their pathogenesis, might also be under a similar genetic selection. To test this hypothesis, I examined the overlap between GWAS defined prostate cancer-associated loci and androgen receptor (AR) binding sites defined in the prostate cancer cell line, PC3 and between breast cancer-associated loci and estrogen receptor (ER) binding sites defined in the ER+ T47D breast cancer cell line. In both types of cancer, hormonal signalling through androgens and estrogens, respectively, are important in the pathogenesis and progression of the cancer and act as the basis for important therapeutic interventions (Burd et al. 2006; Holst et al. 2007).

In each case, I used the same cancer-susceptibility loci, identified in cancer-specific GWAS analyses, as I used for the overlap with HIF-binding sites. The binding sites for AR in PC3 prostate cell lines and for ER in T47D cells were determined from publicly available datasets provided by Gene Expression Omnibus (GEO). Specifically, the ER ChIP-seq dataset was from cells stimulated by the oestrogen receptor agonist bisphenol A (BPA) (GSE54110). The raw data (fastq file) was processed using the same pipeline as used for the HIF ChIP-seq datasets and identified 1,878 binding sites genome-wide. 5 out of 124 breast cancer susceptibility loci directly overlapped an ER-binding site. The expected frequency distribution of the overlap was generated as previously to determine the significance of this observed overlap at P-value = $3*10^{-4}$ (Figure 7.6A).

Similarly, the AR ChIP-seq dataset was from cells stimulated with the androgen receptor agonist R1881 (GSM1010823). Analysis using the same pipeline identified 34,777 sites genome-wide. 26 of the 99 prostate cancer susceptibility loci overlapped an AR binding site. Applying the same bootstrapping approach as above, the statistical significance of this overlap was P-value= $5*10^{-5}$ (Figure 7.6B).

In each case the overlap between cancer-associated polymorphisms and cis-acting elements of the transcriptional network activated by the hormone receptor related to that cancer type was significantly enriched. This is in contra-distinction to the overlap observed with the HIF transcriptional pathway, which did not show statistically significant enrichment for either tumour type. This contrast is again

striking and further emphasises the lack of association between the cancer-associated loci and HIF-binding sites for these types of cancer.

Furthermore, the statistically significant overlap between prostate-cancer associated loci and AR-binding sites and between breast-cancer associated polymorphisms and ER-binding sites again suggests that specific aspects of these hormonally driven transcriptional pathways are under selection during the formation of these types of tumour. Since polymorphisms coinciding with a transcription factor binding site will only affect transcription of individual target genes, it implicates induction of these genes as key events in the pathogenesis of the cancer rather than simply as neutral by-stander events resulting from their co-activation alongside that of driver genes.

Having shown a statistically significant enrichment for AR- and ER-binding sites at prostate and breast cancer-susceptibility loci, it will also be important to again show that genes regulated by these hormone receptors are also enriched at these loci and that the cancer-susceptibility loci affect expression of these genes. Nevertheless, the current work provides a “proof-of-principle” that the approach to examining for statistically significant overlap, between cancer-susceptibility loci and transcription factor binding sites, can be used to examine for selective pressures on other transcriptional cascades that are activated in different types of cancer.

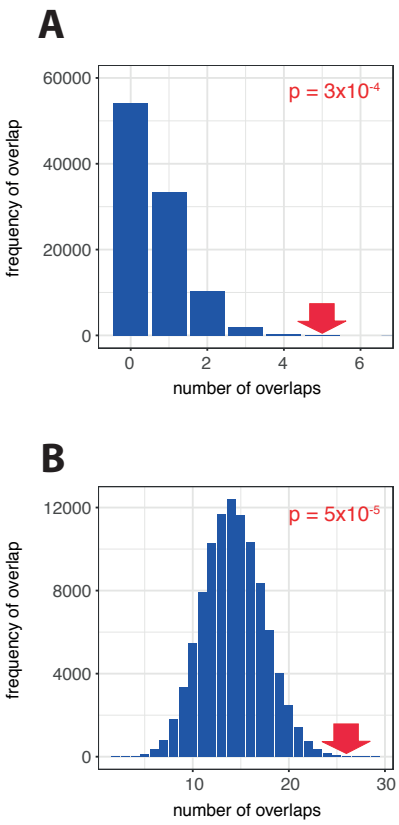


Figure 7.6: Overlap of breast and prostate cancer susceptibility loci with cis-acting elements in activated transcriptional pathways. Breast and Prostate cancer susceptibility loci were defined as index SNP and SNPs in high linkage ($r^2 \geq 0.8$) and AR transcription factor defined by publicly available datasets and analysed as described in this chapter. (A) analysis of oestrogen receptor 1 binding in T47D cells stimulated with the oestrogen analogue bisphenol A (GSM1010823), and (B) androgen receptor binding in PC3 cells stimulated with the androgen analogue R1881 (GSE54110). The blue bars represent the expected frequency distribution of Cancer susceptibility loci and the relevant transcription factor binding sites, the red bar indicated the observed overlap.

7.3 Discussion

In previous chapters, I have shown that the overlap between RCC-susceptibility loci and HIF-binding sites is highly enriched over that which would be expected by random chance. In this chapter, I have shown that this overlap appears highly constrained. In particular, HIF-binding sites that overlap RCC-susceptibility loci exhibit a degree of specificity towards RCC cancer cells themselves. This may in turn account, at least in part, for the specificity of these loci for renal cancer, but not other cancer types. However, other HIF-binding sites are observed in other cell types such as those at 8q24.1, which is also associated with other cancer types. However, it remains unclear whether HIF activation and binding at these sites, in non-RCC cells, is sufficient to generate a selective pressure on the HIF pathway in these cancer types. Indeed, the lack of significant overlap between cancer-susceptibility for lung, colon, breast and prostate cancer, and HIF-binding sites in cell lines derived from these cancers would suggest that this is not the case. This is in direct contrast to the overlap seen between breast and prostate cancer-associated polymorphisms and estrogen- and androgen-receptor binding in these cancer types. Importantly, both estrogen- and androgen-receptor signalling will be operating at physiological levels comparable to physiological activation of the HIF-pathway induced by tumour hypoxia in these cancer types. However, both of these signalling pathways will be operating at early and late stages in cancer development, whereas tumour hypoxia will only occur after the cancer cells have outgrown their blood supply. This suggests that the different interactions between cancer-susceptibility loci and the HIF-pathway in RCC and non-RCC cancers may result from the timing of HIF activation rather than its degree.

Finally, the significant overlap between breast and prostate cancer-associated polymorphisms and estrogen- and androgen-receptor binding sites suggests that this approach might be used to examine for overlap between cancer-susceptibility loci and other transcriptional pathways in different cancer types. This may help in identifying the targets of intergenic cancer-associated polymorphisms as well as in defining driver genes within large activated pathways. Alternately, it may be used to interrogate publicly accessible datasets of transcription factor binding sites, in different cancer cell types. Significant overlap might highlight novel transcriptional pathways that are important in the development of these cancers.

Chapter 8: Final discussion

8.1 Discussion

Although cancer is a heterogenous group of diseases, they share a number of common hallmarks. One such feature is the activation of large transcriptional programmes, often as a consequence of mutation of genes controlling oncogenic or tumour suppressor pathways. These pathways and their transcriptional outputs have largely evolved to mediate cellular adaptation to environmental stimuli and stresses rather than to drive the unrestricted cellular proliferation that leads to cancer. Recent advances in high-throughput technologies in molecular and cellular biology have made it possible to characterise both direct and indirect components of these pathways and have revealed a hitherto unanticipated level of complexity. Indeed, activation of such pathways may lead to the dysregulation of many hundreds of genes that might have diverse actions on tumourigenesis, both promoting and inhibiting tumour development. The overall consequence of pathway activation, therefore, would then entail a subtle balance of multiple (often small) effects. Furthermore, the transcriptional consequences of activating a specific pathway are highly dependent upon the cellular context and the setting in which the cell finds itself. This has several important implications.

Firstly, since gene regulation is highly cell-type specific, the net balance of the effects of pathway activation is likely to vary between tissues, and this may account, to some extent, for the tissue specificity of activation or inactivation of oncogenic or tumour suppressor pathways. Secondly, the genetic landscape in which cells operate may affect specific components of the transcriptional output of the dysregulated pathway and subtly alter the balance of effects and its ability to promote tumourigenesis. This may arise through somatic mutation, occurring

after the pathway becomes activated and conferring a survival advantage to subsequent generations of cells in a classical Darwinian fashion. However, germline variants selected over generations and manifested as cancer predisposing loci might alter the balance of effects in a way that promotes tumourigensis and may reduce the requirement for subsequent re-balancing somatic mutations, thereby increasing the chance of that individual developing cancer.

Activation of the HIF-transcriptional pathway in clear cell kidney cancer as a consequence of inactivation of the VHL tumour suppressor gene provides a good example of large pathway activation in cancer. Whilst HIF is frequently activated in many types of solid tumour, this is typically an appropriate, adaptive, physiological response to intra-tumour hypoxia. Conversely, in clear cell kidney cancer, activation of HIF is a maladaptive, un-physiological response that is agnostic to levels of oxygenation. Many of the responses engendered by this activation will be inappropriate in the absence of hypoxia and so will be likely to have heterogeneous consequences for the cell. Indeed, the two main isoforms of HIF have opposing effects in tumour models of clear cell kidney cancer and individual target genes have diverse associations with disease prognosis in patients. This in turn means that individual components of the HIF transcriptional pathway are more likely to be under selective pressure in clear cell kidney cancer than in other cancer types.

This, therefore, provided the starting point of my thesis, the aim of which was to test for evidence of selective pressure on large pathway activation in cancer and specifically upon the HIF pathway in kidney cancer (the vast majority of which are clear cell kidney cancer). In particular, I examined whether there was evidence that germline variants that pre-dispose to kidney cancer, as identified in GWAS analyses, specifically targeted components of the HIF pathway in kidney cancer cells themselves.

My work has revealed a significant excess overlap between RCC-associated susceptibility loci and individual components of the HIF transcriptional pathway. This remained the case when susceptibility loci were defined based upon individual GWAS as well as for loci identified in a pooled meta-analysis. The association was also robust to different definitions of the HIF-pathway, including direct overlap with HIF-binding sites defined by ChIP-seq analyses, proximity to such sites and to HIF-induced genes and direct overlap with enhancers that looped to the promoters of HIF-target genes. The overlap also extended to include RCC-susceptibility loci with levels of association that fall below conventional thresholds of genome-wide significance. Importantly, the overlap between RCC-susceptibility loci and HIF-binding sites was unique to HIF-binding sites defined in renal cancer cells. Furthermore, in contra-distinction to kidney cancer, susceptibility loci associated with other types of cancer did not specifically overlap with components of the HIF pathway in cell lines derived from those cancers. Nevertheless, susceptibility loci associated with breast cancer and with prostate cancer were enriched for oestrogen receptor binding sites and for androgen receptor binding sites respectively in the relevant cell types.

Taken together, my findings provide compelling evidence that specific components of the HIF-pathway are under genetic selection in kidney cancer. Since each RCC-susceptibility locus affects expression of only one or two HIF-target genes, the analysis strongly implicates these genes in the pathogenesis of kidney cancer and distinguishes them from neutral passenger genes that are simply co-induced as a result of large-pathway activation. Nominally, my analysis cannot distinguish driver genes from repressors, since it is possible that the RCC-susceptibility allele could reduce the expression of a repressive gene. However, at loci those that have been studied in more detail, the RCC-susceptibility allele generally increases binding of HIF and the expression of its target gene.

In addition to direct overlap with HIF-binding sites, I also observed enrichment of RCC-susceptibility loci in the local vicinity of these sites as well as in the proximity of HIF-target genes. On average, each gene is regulated by a combination of separate enhancer regions. Variants at any one of these sites, as well as in the gene body itself, might affect expression of the HIF-target gene and phenocopy loss of HIF binding and provide an explanation for this observation.

In total 8 out of the 13 RCC-susceptibility loci are associated with components of the HIF pathway as defined in RCC cell lines. Furthermore, many of the HIF-binding sites that overlap with these loci are not seen in a number of other cell types. This strongly suggests that many RCC-susceptibility loci directly impact upon the biology of the cancer cells themselves. Whilst some loci might also

directly influence the cell biology of these cells by non-HIF-dependent mechanisms, it is also possible that other loci might have indirect effects (e.g. through altering metabolism of carcinogens, affecting other confounding risk factors such as obesity, or by influencing immune surveillance).

Another implication of the overlap between multiple RCC-susceptibility loci and components of the HIF pathway is that small effects matter in the development of kidney cancer. Whilst highly statistically significant, the effect size (relative risk) conferred by each susceptibility locus is relatively small and the effects on the HIF target gene are largely quantitative rather than qualitative. If only one or two HIF-target genes were responsible for driving the pathogenesis of kidney cancer, then one might expect to observe only one or two susceptibility loci, each of which had a major impact upon the risk of developing kidney cancer. This model of cancer, in which multiple small effects matter is also consistent with the heterogeneous or sub-clonal nature of somatic mutations and with the highly constrained patterns of mutation in major tumour suppressors and oncogenes that have been well described in kidney cancer. If such events had major effects on tumour growth, then they would lead to a clonal sweep and would not be present as sub-clones within the tumour.

Whilst a high proportion of RCC-susceptibility loci overlap components of the HIF-pathway, this accounts for only a very small proportion of HIF-binding sites and HIF-target genes in RCC cells. Chanock et al have estimated that the RCC-susceptibility loci identified in their study account for only a small proportion of

the heritability in kidney cancer, so it is possible that there are many more loci that overlap components of the HIF pathway that are still to be discovered. This hypothesis is supported by my finding that RCC-susceptibility loci that fall below conventional genome-wide significance thresholds are also enriched for HIF-binding sites and proximity to HIF-induced genes. However, even if the number of HIF-binding sites that overlap RCC-susceptibility loci increases 10-fold it will still only account for a minority of HIF-binding sites and HIF-target genes. The implication of this is that only a small proportion of the HIF pathway is under selection in kidney cancer and that therefore, most HIF-target genes are simply neutral bystanders in the pathogenesis of the cancer. It may be that this is a simple consequence of lack of power in current RCC GWA studies and that with increase in cohort size this proportion also increases. Further, RCC susceptibility loci that do not appear to impact on the HIF pathway based on my analysis may still target the pathway in an alternative setting in kidney cancer to the one studied here.

Besides, whilst components of the HIF pathway appear to be under selection in kidney cancer, this does not appear to be the case in other types of cancer. Superficially, this might seem surprising, since HIF is often activated in solid tumours as a consequence of intra-tumour hypoxia as the tumour outgrows its blood supply. Furthermore, in these tumours, high levels of HIF are often associated with adverse patient prognosis. However, the activation of HIF in clear cell kidney cancer, is different from that in other tumours in a number of ways that may account for this difference. Firstly, in clear cell kidney cancer, HIF induction is a maladaptive response that occurs irrespective of oxygenation and may

engender responses that are inappropriate and deleterious to cells that are subjected to normal levels of oxygenation. It is therefore, much more likely that components of the pathway will be under selection in this setting as opposed to other tumour types in which the induction of HIF target genes is appropriate to the ambient level of oxygenation. Secondly, intra-tumour hypoxia is a relatively late phenomenon in the pathogenesis of solid tumours, whereas VHL loss and HIF activation occurs very early in the pathogenesis of kidney cancer and it is possible that these differences in timing also play a role in selecting specific genotypes.

Relating RCC-susceptibility loci to epigenetic markers of active chromatin and pangenomic maps of HIF binding, it has been possible to begin to understand their functional effects and to identify candidate genes that are influenced by the RCC-associated variants. This provides an important paradigm for determining the consequences of susceptibility loci in other diseases, particularly those in which there is a strong candidate cell type, relevant to the biology of the disease, that can be studied. In this respect, the overlap between breast cancer susceptibility loci and oestrogen receptor binding sites in breast cancer cells and between prostate cancer susceptibility loci and androgen receptor binding sites in prostate cancer cells re-inforces the generalisability of my approach.

Finally, given the strong evidence from germline variants that components of the HIF-pathway are under genetic selection in kidney cancer, it is tempting to speculate whether somatic mutations that occur in kidney cancer after VHL

inactivation might also preferentially affect specific components of the HIF pathway. Whilst this might include new sites that have not been identified in germline studies, those loci identified in my analysis would provide a good starting point for such analyses. However, preliminary analysis of whole-genome sequencing data from approximately 130 kidney cancers from the TCGA/ICGC has so far not identified significant enrichment of somatic mutations at these sites. It is possible that these analyses are underpowered and as new datasets become available it will be interesting to see whether variants are identified at these sites. It is also conceivable that the somatic mutation frequency in kidney cancer is too low to ever observe significant numbers of mutations affecting small regions of the genome such as transcription factor binding sites. Alternately, it is also possible that variants at these sites confer relatively weak effects on growth of the cells that do not lead to a clonal sweep and which are therefore present at low frequency below that which can be detected at the conventional 25x sequencing depth used by the TCGA/ICGC. However, whilst for germline variants their presence is ubiquitous in any one individual across all cell-types and time points and their effect is defined by epigenetic states, somatic alterations of small effects are much more difficult to capture and may manifest themselves at different time points during cancer evolution.

Chapter 9: Materials and Methods

9.1 Generation of Next Generation Sequencing datasets

9.1.1 Chromatin Immunoprecipitation

Chromatin immunoprecipitation experiments were performed by various other members of the Ratcliffe/Mole group according to the following protocol.

Cell culture:

786-O and PC3 cell lines were obtained from the American Type Culture Collection (ATCC). RCC4, T47D, A549 and HCT-116 were obtained from the European Collection of Authenticated Cell Cultures (ECACC). The identity of the cell lines was confirmed by Short Tandem Repeats (STR) genotyping and cells were screened to exclude mycoplasma infection. Additionally, RNA-seq analysis of RCC4 and 786-O cells also confirmed the presence of unique VHL gene coding mutations (chr3:10,183,725 C>G and chr3:10,183,841 G>del, respectively). Dulbecco's modified Eagle's Medium, supplemented with 100 U/ml penicillin, 100 µg/ml streptomycin and 10% fetal bovine serum (Sigma Aldrich) was used as growth medium. Hypoxic incubations were performed, in an In Vivo2 400 Hypoxia Work Station (Ruskinn Technology).

Antibodies:

Name	**Source**	**Cat**	**Host**
HIF-1α	In-house	PM-14	Rabbit antiserum
HIF-2α	In-house	PM-9	Rabbit antiserum
HIF-1β	Novus Biologicals	NB 100-110	Rabbit antiserum
H3K4me1	Millipore	07–436	rabbit polyclonal

H3K4me3	Cell Signaling Technology	9751	rabbit monoclonal
H3K27ac	Abcam	ab4729	rabbit polyclonal

Cell fixation and harvest:

Formaldehyde was added to a final concentration of 1%. After incubation on ice for 10-12 min, glycine was then added to a final concentration of 0.125 mM in order to quench the formaldehyde. The cells were incubated for an additional 10 minutes. Cells were kept on ice to maintain protein and nucleic acid integrity.

Chromatin shearing:

Cells were centrifuged 800 rpm for 10 min at 4°C and washed twice with Phosphate buffered saline (PBS). The PBS was then removed, and the cells were resuspended in SDS lysis buffer (1% SDS, 10 mM EDTA, 50mM Tris pH 8.1, 2x complete PI cocktail. miliQ) and incubated for a further 10 minutes at 4°C. Cell debris was then removed by centrifugation at 13,000 rpm and 4°C for 10 mins.

Immunoprecipitation:

Millipore Protein A agarose 50% beads were washed twice with 800 µl dilution buffer (0.01% SDS, 1.1% triton X-100, 1.2mM EDTA, 16.7mM Tris pH 8.1, 167 mM NaCl, miliQ). Samples were incubated with beads for 1 h. Antibodies were added (15 µl of HIF-1α antibody (PM14), 15 µl, of HIF-2α antibody (PM9), 10 µl of HIF-1β antibody (NB 100-110), 10 µl of rabbit pre-immunisation sera) and

incubated on an end-over-end rotator. Samples were resuspended and washed for 5 min in low salt buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris pH 8.1, 150 mM NaCl, MilliQ). Next, samples were resuspended and washed for a further 5 min in high salt buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris pH 8.1, 500 mM NaCl, MilliQ). This was followed by two 5 min washes in LiCl wash buffer (1% Igepal, 1 mM EDTA, 10 mM Tris pH 8.1, 250 mM LiCl, 1% sodium deoxycholate, MilliQ) and a final wash in TE wash buffer (10 mM Tris pH 8.0, 1 mM EDTA, MilliQ). Samples were then eluted in 120 µl elution buffer (0.1M $NaHCO_3$, 1% SDS, Autoclaved water).

Recovery of immunoprecipitated DNA:

The eluted DNA-protein complex was incubated at 65 °C overnight to reverse the crosslinking. The samples were subjected to Proteinase K treatment for approx. 4h and subsequent RNaseA digestion for 30 min. The DNA was recovered by phenol-chloroform extraction.

For each experiment, an aliquot of non-immunoprecipitated chromatin was processed in parallel to act as a control when calling peaks in the bioinformatic analysis.

9.1.2 RNA isolation method

RNA was isolated by a previous member of the group using the RNeasy Micro Kit (Qiagen, U.K.) kit.

9.1.3 Capture-C

Capture-C libraries were generated by Veronique Lafleur in line with the protocol first described by Davis et al. (Figure 7.1). The protocol consists of several steps starting with fixation of cells followed by digestion and ligation, decrosslinking, library preparation and the actual capture assay. The assay was performed using 786-O cancer cell lines obtained from the American Type Culture Collection (ATCC).

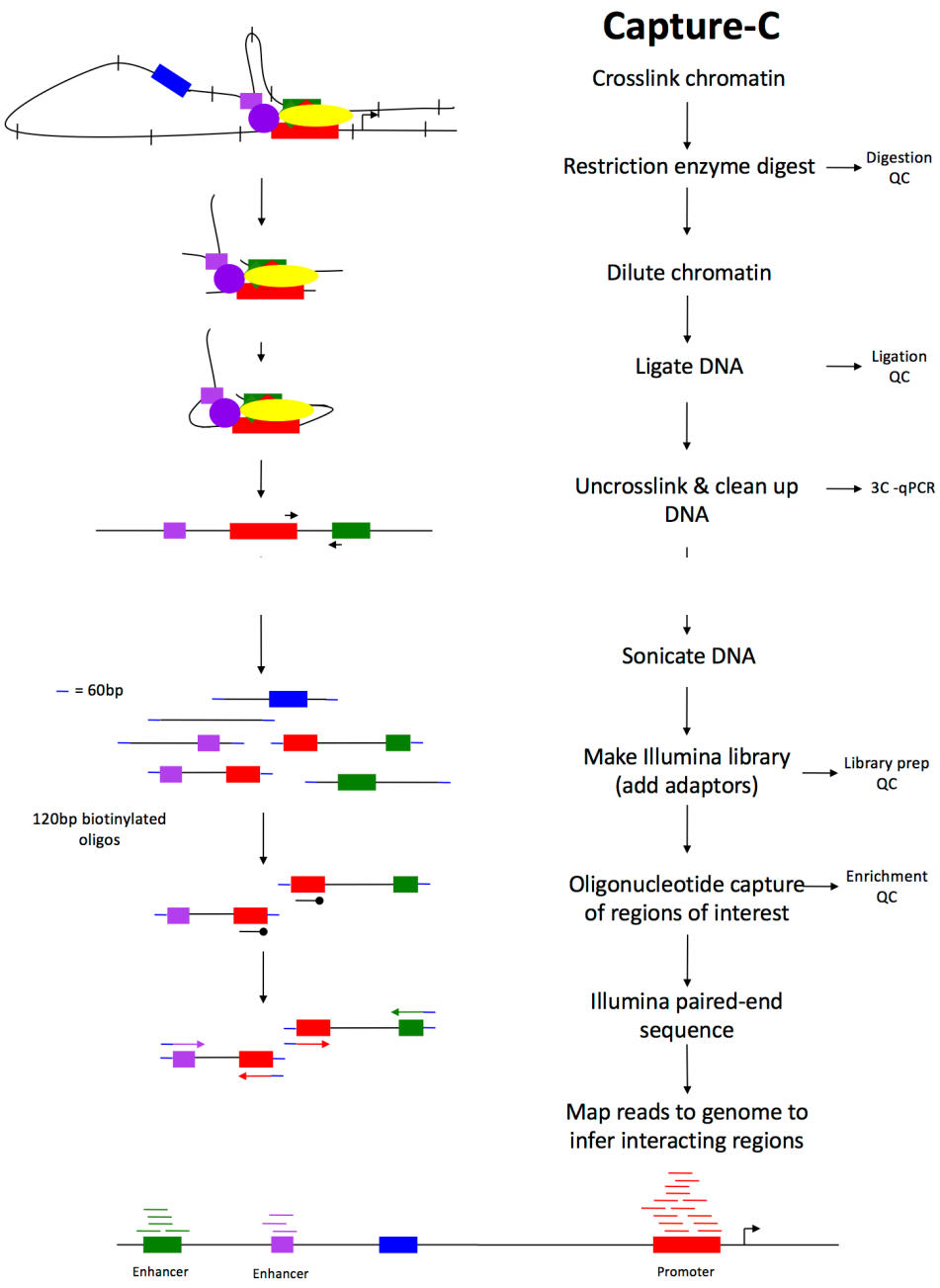


Figure 9.1 Outline of the Capture-C workflow

Fixation of cells:

Cells were trypsinised and re-suspended in aliquots of $1x10^7$ cells in 10ml media and then fixed for 10 min with formaldehyde in a final concentration of 2% (vol/vol). The formaldehyde was then quenched with 125 mM glycine in PBS and lysed in cold (4°C) lysis buffer (10 mM Tris–HCl, pH 8, 10 mM NaCl, 0.2% IGEPAL, 1 x complete protease inhibitor cocktail (Roche)). Finally, cells were resuspended in 1ml lysis buffer and snap frozen in liquid nitrogen.

Digestion and Re-ligation:

Thawed cells were homogenized with a Dounce homogeniser and cell debris removed by centrifugation. Chromatin was then digested overnight with 20 µl DpnII (High concentration 50,000 U/ml, New England Biolabs) at 37°C. This was followed by a second 2-hour digestion step using a further 5 µl DpnII. The restriction enzyme was then inactivated by heating to 65°C. The efficiency of the digestion step was determined by assessing fragment sizes on gel electrophoresis. Digested fragments were then diluted in ligation buffer and re-ligated using 8 µl T4 DNA ligase (HC T4 DNA HC ligase 30U/µl EL0013 Thermoscientific) at 16°C overnight. Next the chromatin was de-crosslinked and protein was digested by incubating overnight at 60°C in the presence of Proteinase K (Bioline).

Sonication:

The re-ligated DNA fragments were sheared using a Covaris S220 Focussed Ultrasonicator (duty factor = 10%, peak power = 175, cycles per burst = 200, 6 cycles each of 60s = 6mins, set mode = frequency sweeping) aiming for a fragment size of 150-200 bp.

3C library preparation:

The sonicate, re-ligated DNA fragments were then purified in a two-step phenol-chloroform extraction followed by overnight precipitation in ethanol at -80°C. Next, sequencing adapters were added using the NEBNext DNA library Prep Reagent Set (E6040S or E6000S) together with NEBNext Multiplex Oligos for Illumina (E7335 or E7500S) in line with the manufacturer's protocol.

Oligonucleotide Capture:

Two capture rounds were performed using Nimblegen SeqCap oligonucleotide complementary to bait regions (see Appendix 10.3.4) and according to the manufacturer's protocol. The first hybridisation step was for 64-72h, whilst the second lasted 24h. The effectiveness of the capture was assessed using real-time PCR. Finally, captured multiplex DNA sample are amplified using PCR.

9.2 Processing and analysis of NGS datasets

9.2.1 Sequencing

All libraries were sequenced by the High Throughput Genomics Core facility in the Wellcome Trust Centre for Human Genetics using HiSeq2500 and HiSeq4000 (Illumina, U.K.) instruments according to the manufacturer's protocol.

9.2.2 Alignment of ChIP-seq data

Next, DNA sequences in fastq format were aligned to the human hg19 reference genome (Genome Reference Consortium, GRCh37) using BWA (0.5.9) (https://github.com/lh3/bwa). The output binary Sequence Alignment/Map BAM files were converted to SAM format using samtools (0.1.19) (https://github.com/samtools/samtools). Reads with low mapping quality (MapQ < 15) were excluded using SAMtools (0.1.19). Next, reads which mapped to the Duke Encode black list regions (http://hgwdev.cse.ucsc.edu/cgibin/hgFileUi?db=hg19&g=wgEncodeMapability) were excluded with BEDTools (2.17.0) (https://github.com/arq5x/bedtools2). Duplicated reads were filtered with Picard tools (1.106) (https://github.com/broadinstitute/picard).

9.2.3 Bigwig generation for ChIP-seq, RNA-seq and Capture-C datasets

To generate bigwig tracks for display in IGV Browser (https://github.com/igvteam/igv), genome-wide coverage was first assessed using the genomeCoverageBed function within BEDTools (2.17.0). The resulting bedGraph file was then converted to a BigWig file using the bedGraphToBigWig

function.

9.2.4 Peak calling for ChIP-seq data

ChIP-seq peaks were detected by applying two different peak callers, T-PIC (Tree shape Peak Identification for ChIP-Seq) (Hower et al. 2011) and MACS 1.4 (Model-based analysis of ChIP-Seq) (Zhang et al. 2008) and using overlapping regions from these two methods.

The MACS peaks caller was run with default options, except that the minimum read length was increased from 25 to 36 bp (macs14 -t <treatment file> -c <control file> -n name -s 36 -f BAM -m 10,30) for more stringent peaks calling. For each ChIP-seq experiment, peaks were called by reference to a control sample consisting of non-immunoprecipitated chromatin from the same experiment. TIPC was also run applying default options. Finally, MACS peaks that directly overlapped with a TPIC-identified peak were identified using the intersect function of BEDTools.

9.2.5 Motif analysis and MEME-suite on ChIP-seq peaks

Motif-based sequence analysis was performed using MEME Suite (http://meme-suite.org) (Bailey et al. 2009). Regions flanking (±150 bp) MACS-defined summits (location of maximum ChIP-seq signal intensity) at each of the ChIP-seq peaks were used for this analysis. Fasta files were generated for each of these regions using the fastafromBED function of BEDTools. RepeatMasker 4.0.3 (www.repeatmasker.org) was then used to remove repetitive and low

complexity DNA sequences within these regions. The Analysis of Motif Enrichment (AME) function within MEME Suite was then used to identify motifs from the JASPAR Core (2016) Vertebrates library (http://jaspar.genereg.net/) that were enriched in these regions compared to a randomised version of the same sequences. The same function also generated a frequency distribution for each enriched motif according to distance from the summit of the ChIP-seq peak, which could be used to check that motif enrichment co-localised with the position of maximal ChIP-seq signal.

9.2.6 Visualisation of ChIP-seq data by heatmaps and ngs.plot

ChIP-seq datasets were also visualised by generating heatmaps and average profile line plots using ngs.plot (Shen et al., 2014). The following parameters were used: -GO none (no particular ranking applied along y-axis), -SE 0 (standard error), -FL 50 (fragment length), -MQ 15 (read mapping quality cut-off according to Phred-scale), -L 5000 (5,000 bp flanking regions size either side of the ChIP-seq summit).

9.2.7 Combining datasets using bedtools

BED files of all 7 BED files were concatenated and sorted by chromosomal location. Next peaks overlapping by a minimum of 1 bp were merged with BEDtools (*bedtools merge -i*). Overlap of this merged set of peaks with each of the original BED files was determined. Peaks present in at least two data sets present were considered, where if replicates were present it only was counted as 'present' when the peak was present in both replicas.

9.2.8 RNA-seq analysis

As for the ChIP-seq data sequencing adaptors are trimmed by Trimgalore (0.3.3). Tophat was used to recognise identify exon-exon splice junctions (Trapnell et al. 2009). Reads were aligned using the same version of bowtie 1.0.0 (http://bowtie-bio.sourceforge.net/index.shtml) and reads were then aligned to the same reference as in the previous section (Genome Reference Consortium GRCh37 (hg19)) GRCh37 using Tophat 2.0.8b (http://ccb.jhu.edu/software/tophat/index.shtml). Non-uniquely mapping fragments excluded using SAMtools (0.1.19). The total read counts for each UCSC-defined genes computed with HTSeq (0.5.4p3) 48 with "intersection-strict" mode (Anders et al. 2015). DESeq2 was used to determine significantly differentially regulated genes (Love et al. 2014).

9.2.9 Identification of nearest genes to a feature (e.g. index SNPs)

Identification of the 3 closest genes was performed with bedtools (*bedtools closest -k 3 -a <bed file containing coordinates of regions of interest> -b <bed file of the hg19 assembly annotation provided by Ensembl* >). This identified a 39 gene set for RCC loci defined in chapter 3 and a further 657 gene set for the sub-threshold loci defined in chapter 6.

Identification of interacting genes of genome-wide significant loci defined based on a recent meta-analysis of RCC and described in chapter 3 was performed by identifying genes within 25 kb from the LD region associated with the index SNP of all 13 genome-wide significant RCC susceptibility loci ($r^2 \geq 0.8$ with reference to the 1000 Genomes EUR population data). Further, significantly interacting

regions, defined by analysis of Capture-C data by Ran Li and locating at the same genomic locus were analysed for overlap with promoters (TSS +/- 2000 kb) with reference to the Ensembl GRCh37 (hg19) genome annotation.

Identification of interacting genes of sub-threshold loci (defined in chapter 6) overlapping HIF binding sites was done by first identifying loci bound by HIF. Genes within 25 kb from the associated LD regions ($r^2 \geq 0.8$ with reference to the 1000 Genomes EUR population data) were identified and alongside genes with promoters interacting with this set of sub-threshold loci, identified with reference to the same Capture-C dataset.

The pre-ranked GSEA was performed with reference of a gene set pre-ranked by a weighted enrichment score of significance and fold change (x = -log10 (adjusted P-value)*log2(fold change)) (Subramanian et al., 2005), where fold change and P-value was determined by DESeq2 (Xiao et al., 2014).

9.2.10 Pre-ranked Gene set enrichment analysis (GSEA)

Gene sets for testing in Gene Set Enrichment Analysis (GSEA) were defined in various ways as outlined in the relevant sections (Subramanian et al. 2005). Gene expression data were derived from RNA-seq analyses, each of which was performed in triplicate. Datasets were processed as described in chapter 9.2.8 and the differential expression, together with the associated P-value for each expressed gene, was derived using DESeq2. Genes were then pre-ranked using

a combined metric of these two parameters (Subramanian et al., 2005) (Xiao et al. 2014).

$$x = -\log10[\text{adjusted p-value}] \times \log2[\text{fold change}]$$

GSEA was then performed in pre-ranked mode using a weighted enrichment score.

9.2.11 Processing of Capture-C data by applying pipeline

Significantly interacting regions were identified using an analysis pipeline established by Rafik Salama (Platt et al. 2016). The pipeline integrates multiple steps. First, the sequencing reads are in silico digested, by reference to the Dpn2 cleavage motif (CATG), before alignment to the hg19 human genome using Bowtie 1.0.0 (http://bowtie-bio.sourceforge.net/ index.shtml).

9.2.12 Identification of significantly interacting regions by applying pipeline

Sequences at each bait/capture site acted as a barcode that was used to focus the analysis on each captured region in turn. Sequences at the other end of the paired-reads then defined regions that interact with each of these sites. CCanalyser2.pl (https://github.com/telenius/captureC/releases) was used to determine the interaction frequencies between each bait site and flanking Dpn2 restriction fragments.

To determine the statistical significance of these interaction frequencies, signal intensity was first normalized to the total number of interactions with the respective capture probe. Random background interactions at a given distance from the capture probe were then empirically estimated by reference to the average normalized interaction signal across all experiments. The significance of the interaction at each Dpn2 fragment was finally computed by reference to this estimated average background signal and its modelled variance according to a negative binomial distribution.

9.3 Processing of GWAS datasets

9.3.1 1000 Genomes data

Vcf files of all SNPs identified in the 20130502 release of the 1000 Genomes Project (http://www.internationalgenome.org/data/) were downloaded and converted into ped format. Data belonging to populations of European ancestry (CEU, FIN, GBR, IBS, TSI) was used for all analysis, since GWAS datasets were predominantly from these populations. Data was filtered to exclude SNPs with allele frequency ≤ 0.01 or with a genotyping rate below 90%.

9.3.2 Plink

Regions of linkage disequilibrium (LD) within the 1000 Genomes data were determined using plink 1.9 (http://zzz.bwh.harvard.edu/plink/ld.shtml). The LD window was set to 2000 (only the closest 2,000 SNPs are analysed) and the LD window was set to 500,000 kb (only SNPs within 500,000 kb are analysed). The r^2 parameter was set to either 0.8 or 0.2 depending was trying to identify SNPs

in strong linkage or exclude SNPs in weak linkage. For all other parameters, the default values were used.

9.3.3 Manhattan plot

The 'qqman' R package (https://CRAN.R-project.org/package=qqman) was used to generate Manhattan plots of the summary statistics from the GWAS dataset provided by the National Institute of Health.

9.3.4 Definition of susceptibility loci associated haplotype blocks from summary statistics

To define haplotype blocks of SNPs that are in high Linkage disequilibrium (LD) with the index SNP at each of the susceptibility loci, from the GWAS summary statistics, the index SNPs (most significant SNP) at each of the loci were first defined. Essentially, the summary statistics were ranked by decreasing P-value. Starting with the most significant SNP (which constitutes the first index SNP), all SNPs within 500 Mb or in close genetic linkage with an $r^2 \leq 0.2$ (analysed as above), were then excluded. The next most significant SNP after this then constituted the second index SNP and the process was then repeated until there were no more SNPs with a P-value $\leq 5*10^{-8}$. For the RCC GWAS meta-analysis this defined 13 SNPs that each represented an independent locus that was associated with RCC. The haplotype block associated with each of the index SNPs was then defined as all SNPs in high linkage disequilibrium ($r^2 \geq 0.8$) with each SNP, again with reference to the 1000 Genomes EUR population. Summary

statistics for RCC, breast, lung cancer and educational attainment were all processed in the same manner.

9.3.5 Definition of susceptibility loci associated haplotype blocks from database entries

For two cancer types (prostate and colon cancer), no summary statistics were available and so loci reported in the NHGRI-EBI GWAS catalogue (accessed June 2018) (Buniello et al., 2019) were used instead. Using these reported SNPs, haplotype blocks were defined as in chapter 9.3.4.

9.3.6 Expected frequency distribution of overlap of GWAS locus and HIF binding site

To determine the statistical significance of the overlap between the RCC-susceptibility loci and HIF binding sites, I applied a bootstrapping approach. The expected frequency distribution was determined by randomly replacing the 13 RCC-associated haplotype blocks loci on the genome 100,000 times and assessing the number of blocks that directly direct overlapped with a HIF binding sites on each occasion. Each time the 13 haplotype blocks were placed on the genome *en bloc*, whereby the SNPs belonging to each haplotype bock remained separated by the same distance as at the original locus. The proportion of iterations in which a given number of overlaps was observed was then displayed as a frequency histogram. The proportion of iterations in which the number of overlapping blocks was the same as or greater than the number observed with

the actual haplotype blocks (i.e. 4) then gave the statistical significance for the finding. Similar analyses were performed for other datasets.

9.3.7 Expected frequency distribution of overlap of HIF binding site with HIF binding site

To determine whether HIF-binding sites were enriched compared to non-HIF bound enhancers, the converse analysis, in which the ChIP-seq peaks were randomly moved around regions of the genome defined as active enhancers based on H3K4me1, H3K4me3 and H3K27ac analyses in 786-O RCC cells (Barski et al., 2007), was also performed. This process was again repeated 100,000 times to examine the overlap between GWAS-defined susceptibility loci and HIF-binding sites in RCC cells, and the corresponding frequency distribution and P-value evaluated.

9.4 TCGA data

9.4.1 Datasets included

Data from the Cancer Genome Atlas (TCGA) kidney clear cell renal cell cancer (KIRC) dataset was downloaded, stored as a MySQL database and analysed using RMySQL. Specifically, level 3 data comprising RNA-Seq expression data for 450 clear cell Renal Cell Carcinoma patients was associated with level 2 Affymetrix Genome-Wide Human SNP Array 6.0 data from the same patients.

9.4.2 eQTL

Potential eQTLs were analysed by sorting patients according to the relevant genotype using the Affymetrix Genome-Wide Human SNP Array 6.0 data and assessing gene expression using the RNA-Seq expression data of 450 clear cell Renal Cell Carcinomas in the same patients. RNA-seq expression data was modelled with a negative binomial Generalized Linear Model (GLM) against the genotype status and likelihood was computed by reference to a model that ignored the genotype status. The significance was determined by a Chi-Squared test determining the significance of the genotype in the stratification of the patients.

9.4.3 meta-gene eQTL

The meta-gene set for HIF signatures was previously defined with reference to HIF-1α and HIF-2α re-expressing 786-O cell lines.

9.4.4 Differential expression in tumour and normal tissue

The TCGA-KIRC cohort includes matched tumour and normal RNA expression data from 72 cases.

Chapter 10: Appendix

10.1 IGV tracks RCC susceptibility loci

11q13.3 locus:

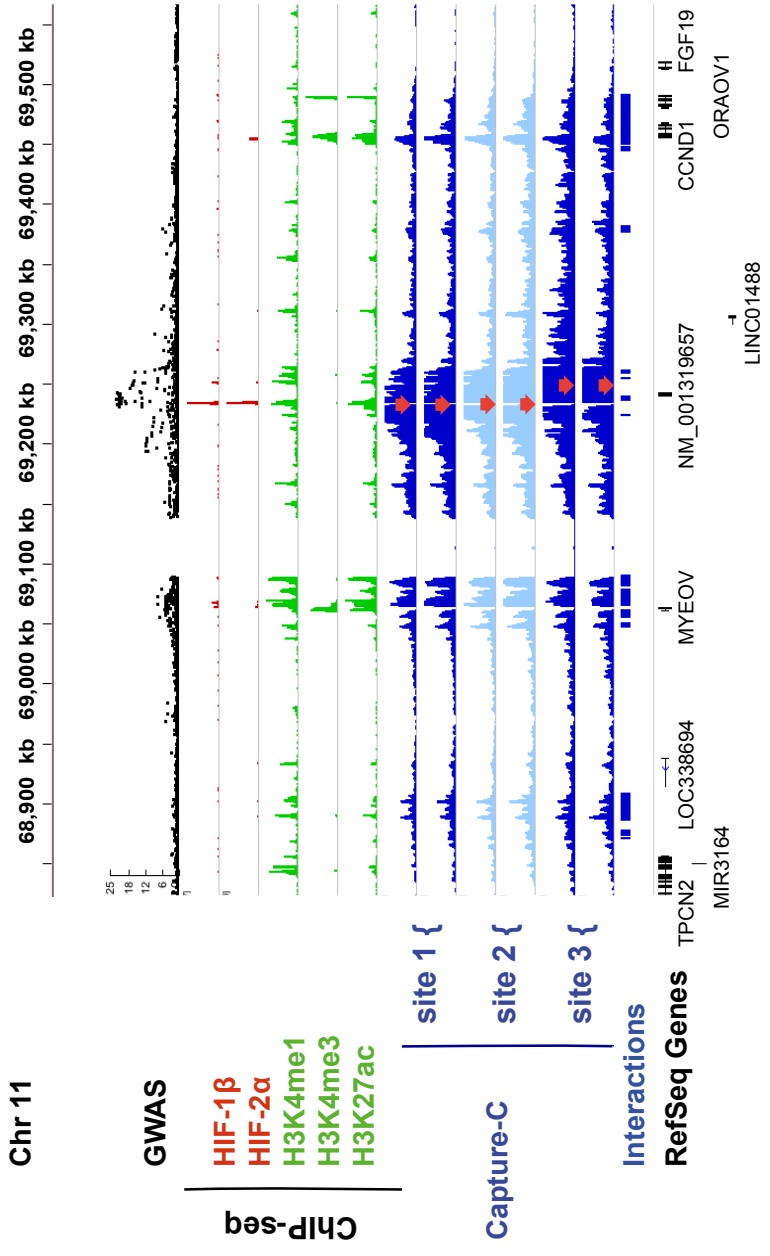


Figure 10.1 A: 11q13.3 centred Integrative Genomics Viewer (IGV) tracks displaying SNP P-values identified by RCC GWAS meta-analysis (black track), with ChIP-seq signal of HIF binding (red tracks), histone modifications (green tracks), and Capture-C tracks performed in duplicates (in blue) indicating chromatin looping. The red arrows are indicating the bait site designed towards each active enhancer that overlaps with the linked SNPs at the susceptibility locus ($r^2 \geq 0.8$ with the index SNP). All data displayed here was assayed in 786-O cells. Promoters of the *CCND1*, *ORAOV1* and *MYEOV* genes overlap with sites where Capture-C interaction signals significantly greater than local background was detected (indicated by blue bars). The linkage disequilibrium region further overlaps with the putative gene, *NM_001319657.* RefSeq hg19 (GRCh37) build is displayed for chromosomal coordinates and gene annotations (in black).

2p21.3 locus:

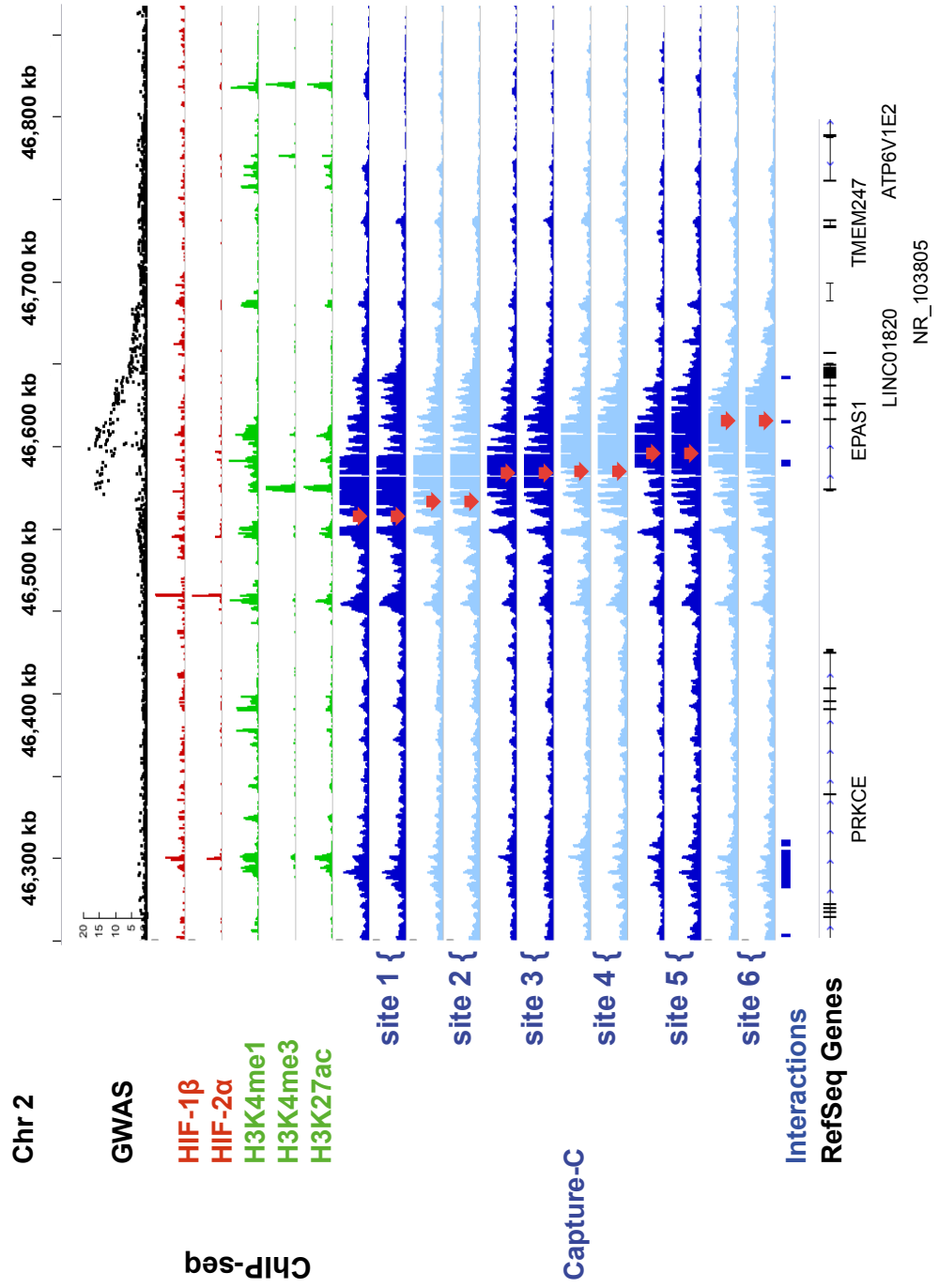

Chr 2
46,300 kb
46,400 kb
46,500 kb
46,600 kb
46,700 kb
46,800 kb
GWAS
ChIP-seq
HIF-1β
HIF-2α
H3K4me1
H3K4me3
H3K27ac
Capture-C
site 1
site 2
site 3
site 4
site 5
site 6
Interactions
RefSeq Genes
PRKCE
EPAS1
LINC01820
TMEM247
ATP6V1E2
NR_103805

Figure 10.1 B: 2p21.3 IGV tracks centred at the 2p21.3 locus. Individual tracks are described in Figure 10.1 A. The EPAS1 (alias HIF-2α) encoding region locate within +/- 25 kb of the LD region ($r^2 \geq 0.8$) of the locus. Interaction with more distant promoters or enhancers of the susceptibility locus was not detected by Capture-C assays. However, HIF-1β and HIF-2α binding sites are present in a region that was shown to loop to an enhancer in proximity to the EPAS1 (alias HIF-2α) promoter by the same Capture-C analysis.

12p12.1 locus:

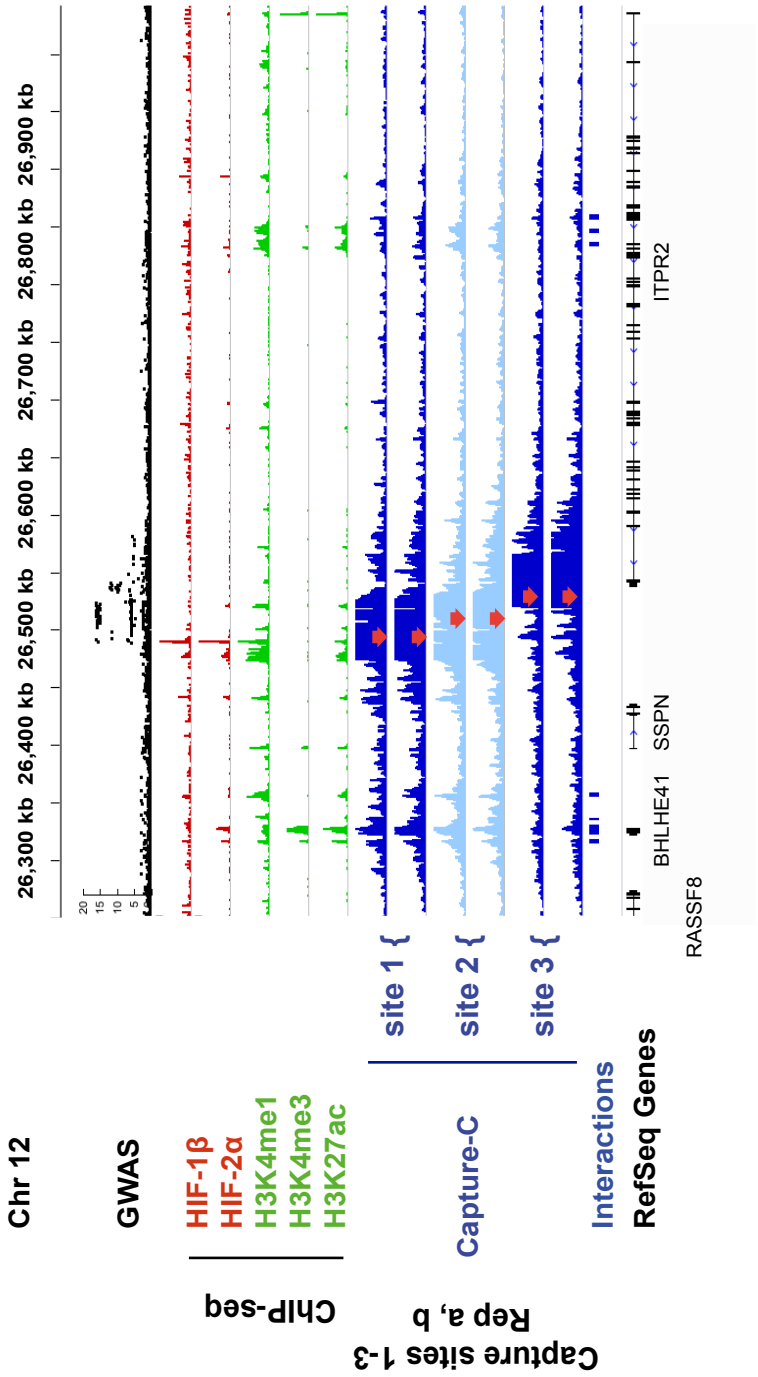


Figure 10.1 C: IGV tracks at centred at the 12p12.1 locus. Individual tracks are described in Figure 10.1 A. The SSPN and ITPR2 encoding regions locate within +/- 25 kb of the LD region ($r^2 \geq 0.8$) of the locus, whilst the promoter of BHLHE41 and an alternate promoter of ITPR2 (data from ENSEMBL) both overlap with Capture-C detected interaction regions of the viewpoint oligonucleotides.

14q24.2 locus:

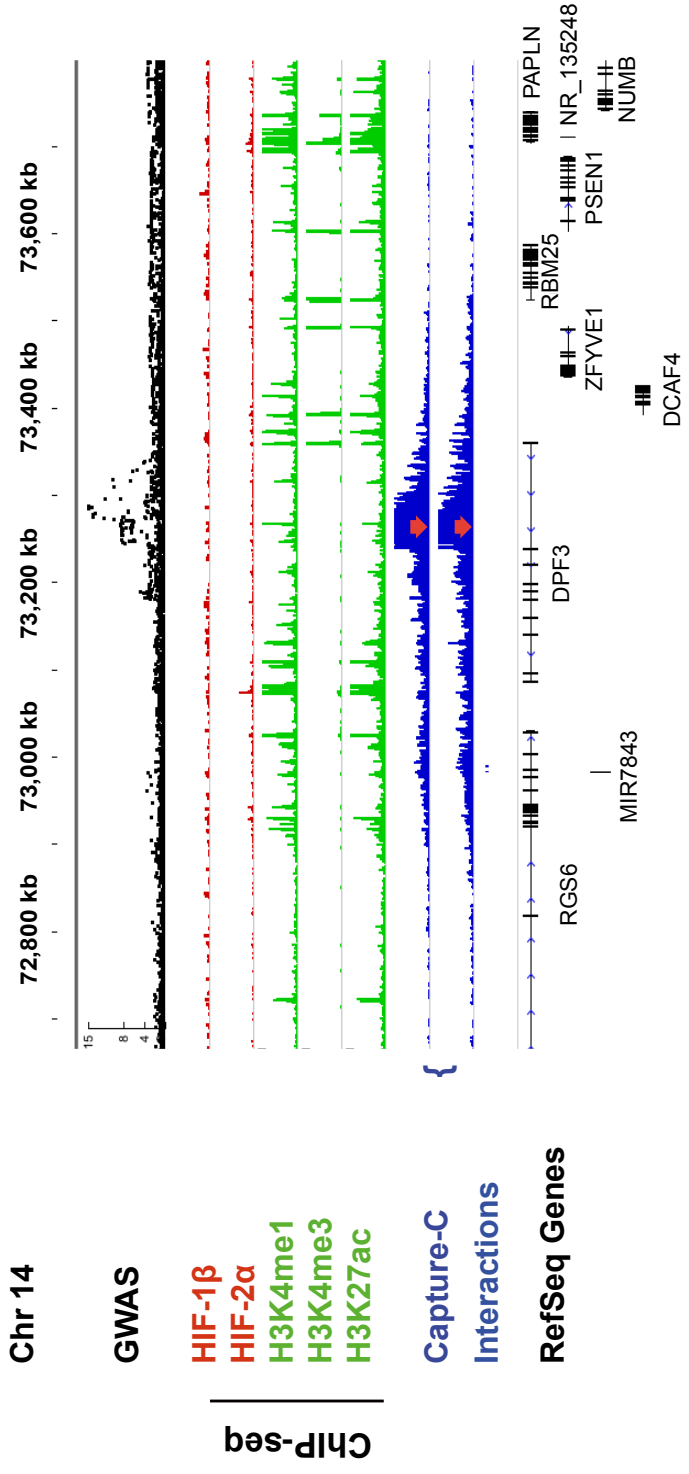


Figure 10.1 D: IGV tracks centred at the 14q24.2 locus. Individual tracks are described in Figure 10.1 A. The DPF3 gene locates within +/- 25 kb of the LD region ($r^2 \geq 0.8$) of the locus. No statistically significant interactions were detected by Capture-C indicating looping to more distant gene promoters.

12q24.31 locus:

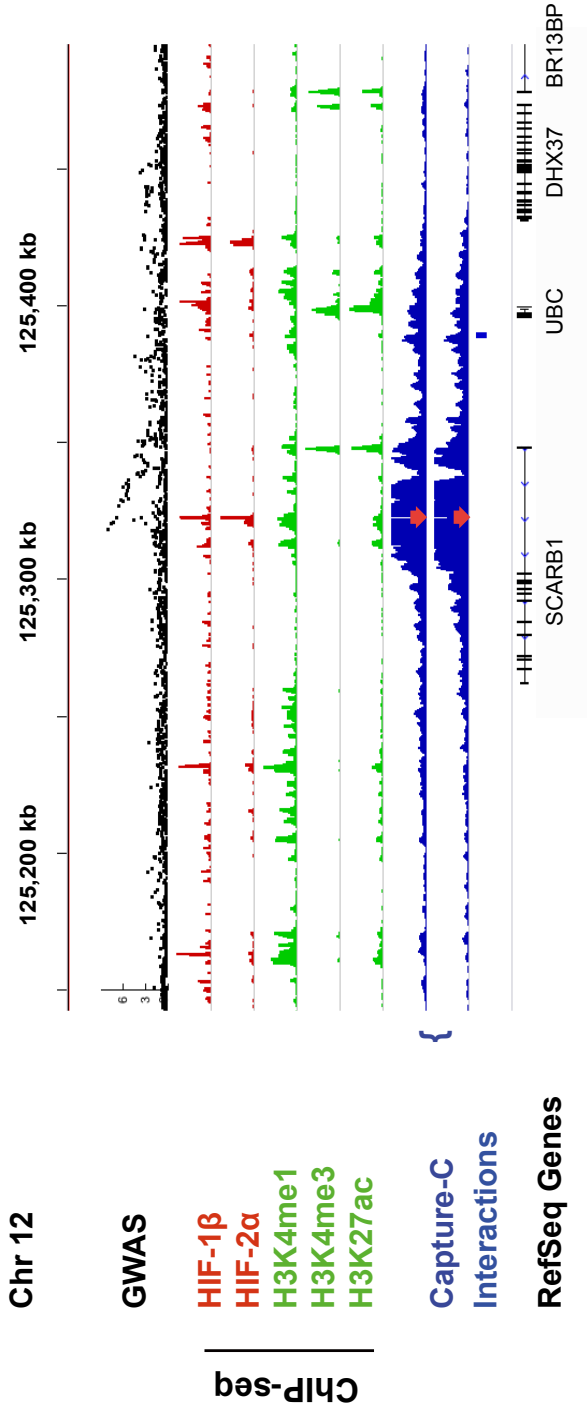


Figure 10.1 E: IGV tracks centred at the 12.q24.31 locus. Individual tracks are described in Figure 10.1 A. The *SCARB1* locates within +/- 25 kb of the LD region ($r^2 \geq 0.8$) of the locus. SNPs associated with the LD region of the locus are in proximity with a HIF binding site that overlaps an enhancer in the first intron of the *SCARB1* gene. Interactions by chromatin looping were detected by Capture-C with the promoter of *SCARB1*. However, these observed interactions were not identified as statistically above the local background (this is likely due to the proximity of the region with the Capture-C "viewpoint"). No statistically significant interactions were detected by Capture-C, indicating looping to more distant gene promoters.

8q24.21 locus:

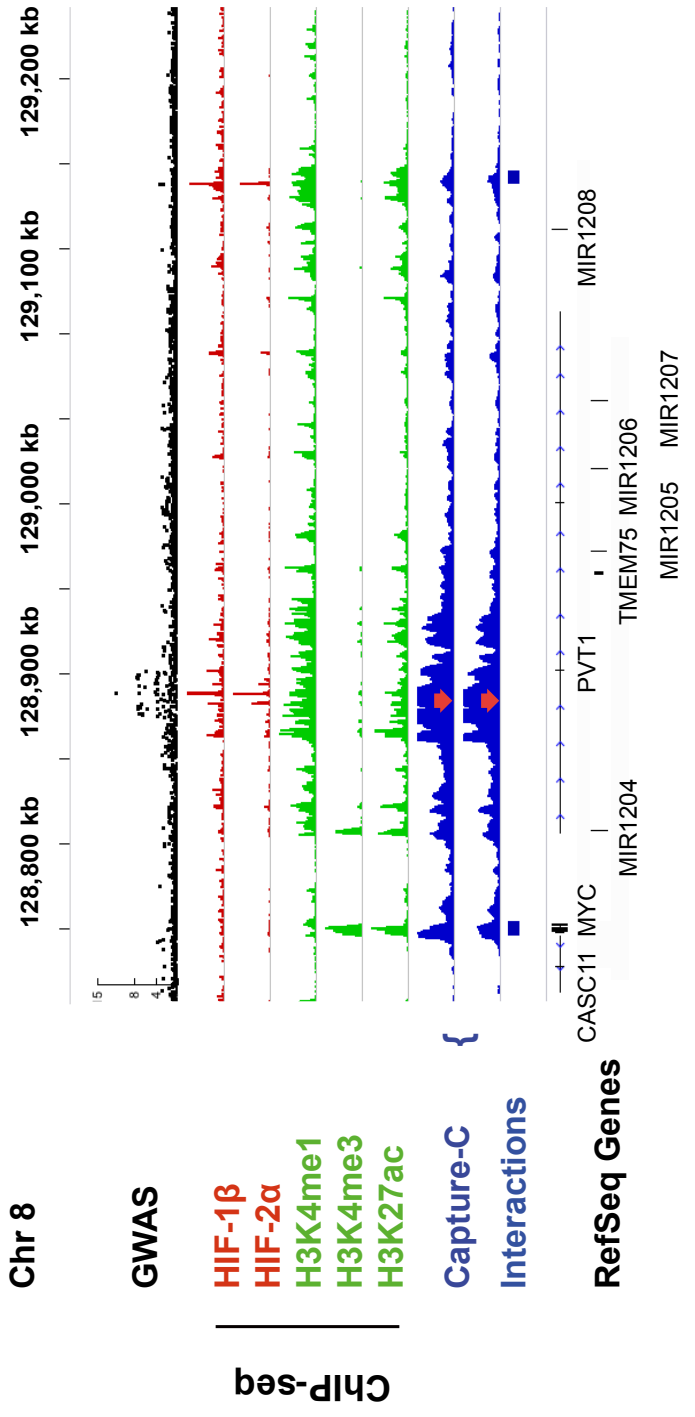


Figure 10.1 F: IGV tracks centred at the 8q24.12 locus. Individual tracks are described in Figure 10.1 A. The HIF-binding site that overlaps with the LD region ($r^2 \geq 0.8$) of the locus was detected to have a statistically significant interaction detected by Capture-C assays of chromatin looping with the MYC promoter alongside an additional, distant HIF-binding site. This additional HIF binding site may also loop to the MYC promoter. Further the non-coding RNA, PVT1, also overlaps with the LD region ($r^2 > 0.8$) of the locus.

11q22.3 locus:

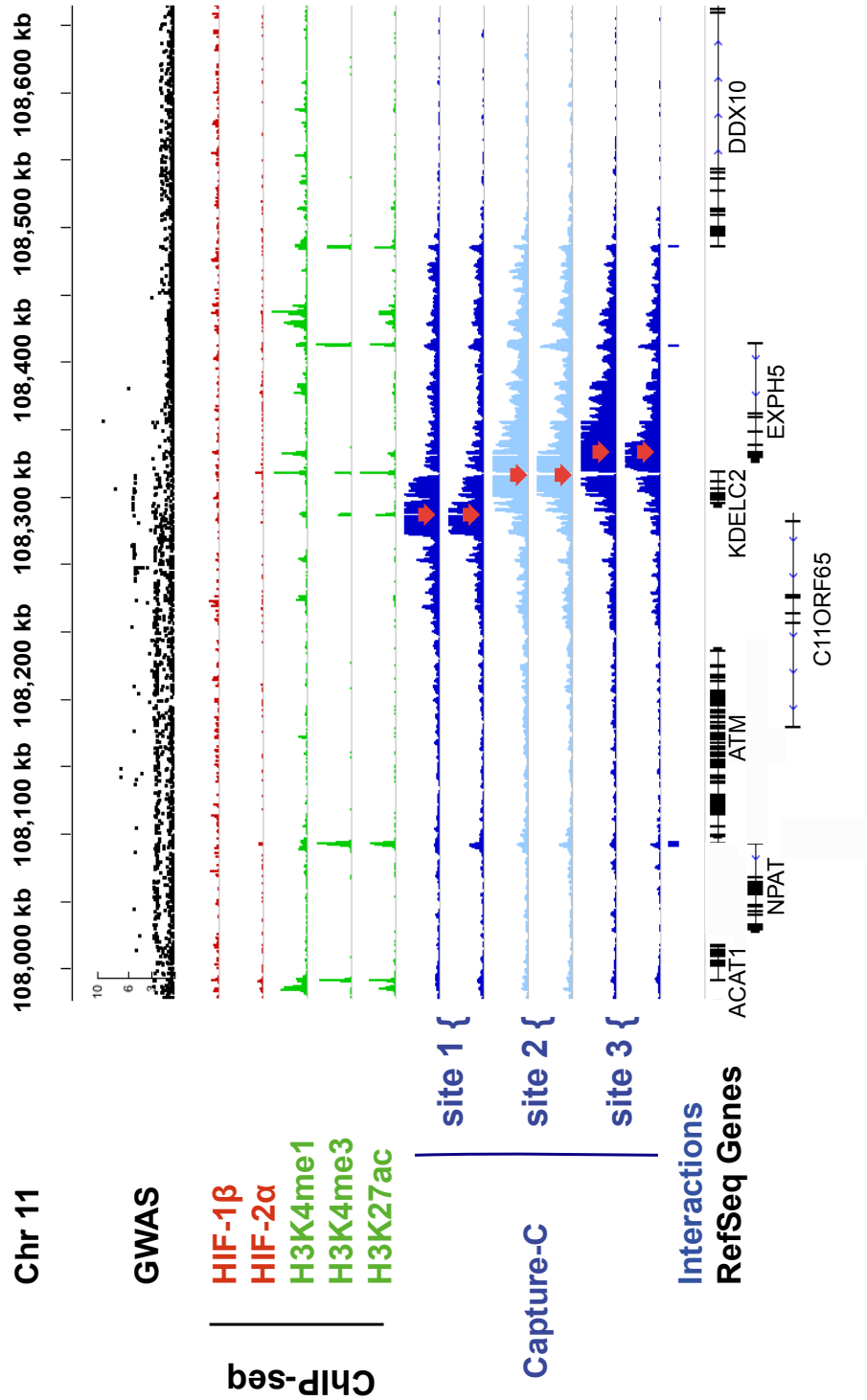


Figure 10.1 G: IGV tracks centred at the 11q22.3 locus. Individual tracks are described in Figure 10.1 A. The EXPH5, C11orf65 KDELC2 genes locate within +/- 25 kb of the LD region ($r^2 \geq 0.8$) of the locus. The promoters of DDX10, EXPH5, NPAT and ATM overlap with Capture-C detected interaction regions of the viewpoint oligonucleotides.

2q22.3 locus:

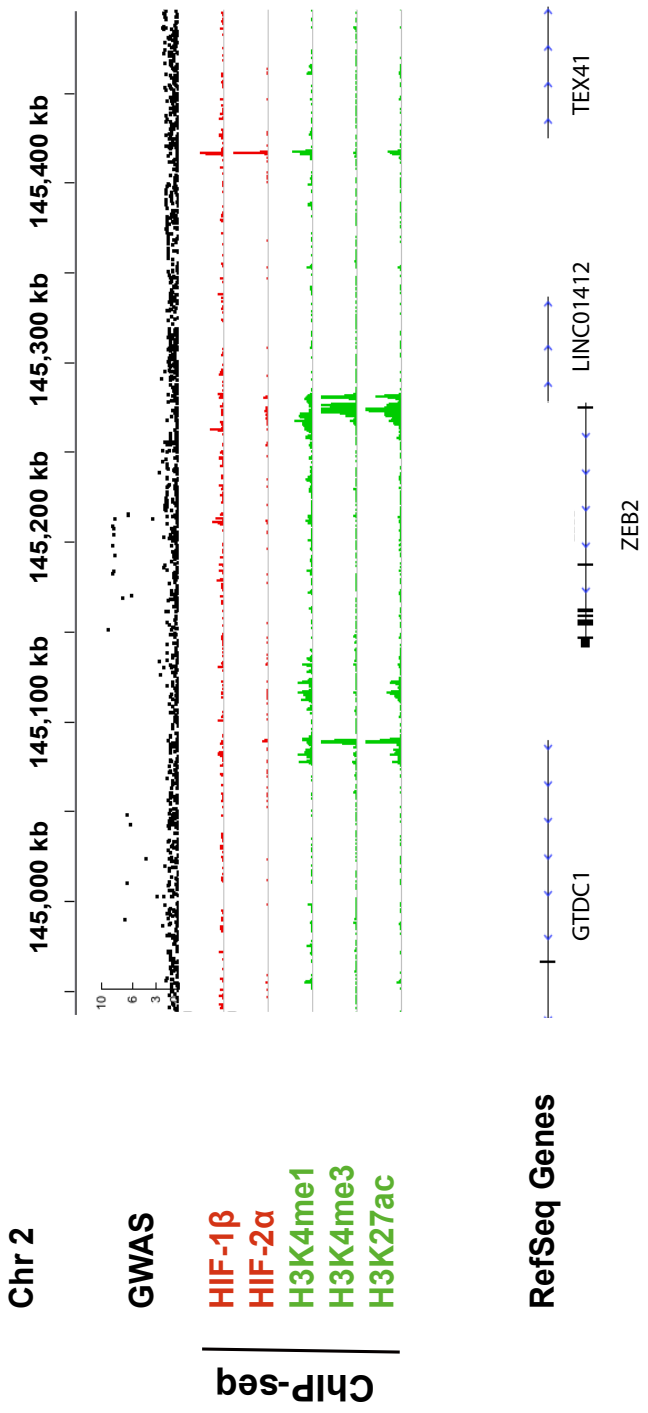

Chr 2
145,000 kb
145,100 kb
145,200 kb
145,300 kb
145,400 kb
GWAS
10
6
3
ChIP-seq
HIF-1β
HIF-2α
H3K4me1
H3K4me3
H3K27ac
RefSeq Genes
GTDC1
LINC01412
ZEB2
TEX41

Figure 10.1 H: IGV tracks at centred at the 2q22.3 locus. Individual tracks are described in Figure 10.1 A. The LD region ($r^2 \geq 0.8$) of the locus overlaps with the ZEB2 gene. However, it was not possible to assess chromatin looping was not since no enhancers overlapped with the LD region.

1p32.3 locus:

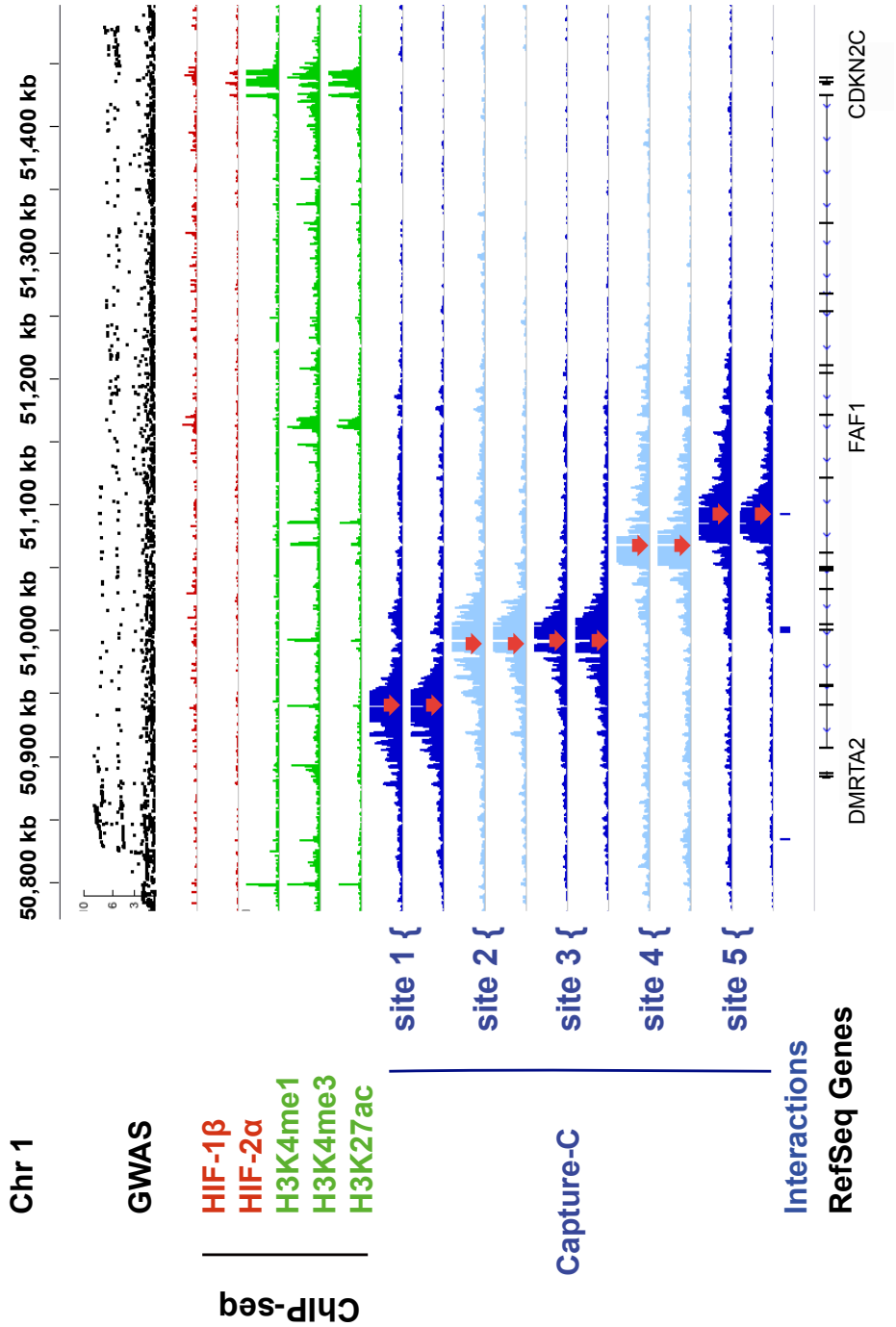

Chr 1
50,800 kb
50,900 kb
51,000 kb
51,100 kb
51,200 kb
51,300 kb
51,400 kb
GWAS
ChIP-seq
HIF-1β
HIF-2α
H3K4me1
H3K4me3
H3K27ac
Capture-C
site 1
site 2
site 3
site 4
site 5
Interactions
RefSeq Genes
DMRTA2
FAF1
CDKN2C

Figure 10.1 I: IGV tracks at cantered the 1p32.3 locus. Individual tracks are described in Figure 10.1 A. *FAF1* and *DMRTA2* genes locate within +/- 25 kb of the LD region ($r^2 \geq 0.8$) of the locus. No statistically significant Capture-C interactions with more distant gene promoters were identified.

4q23 locus:

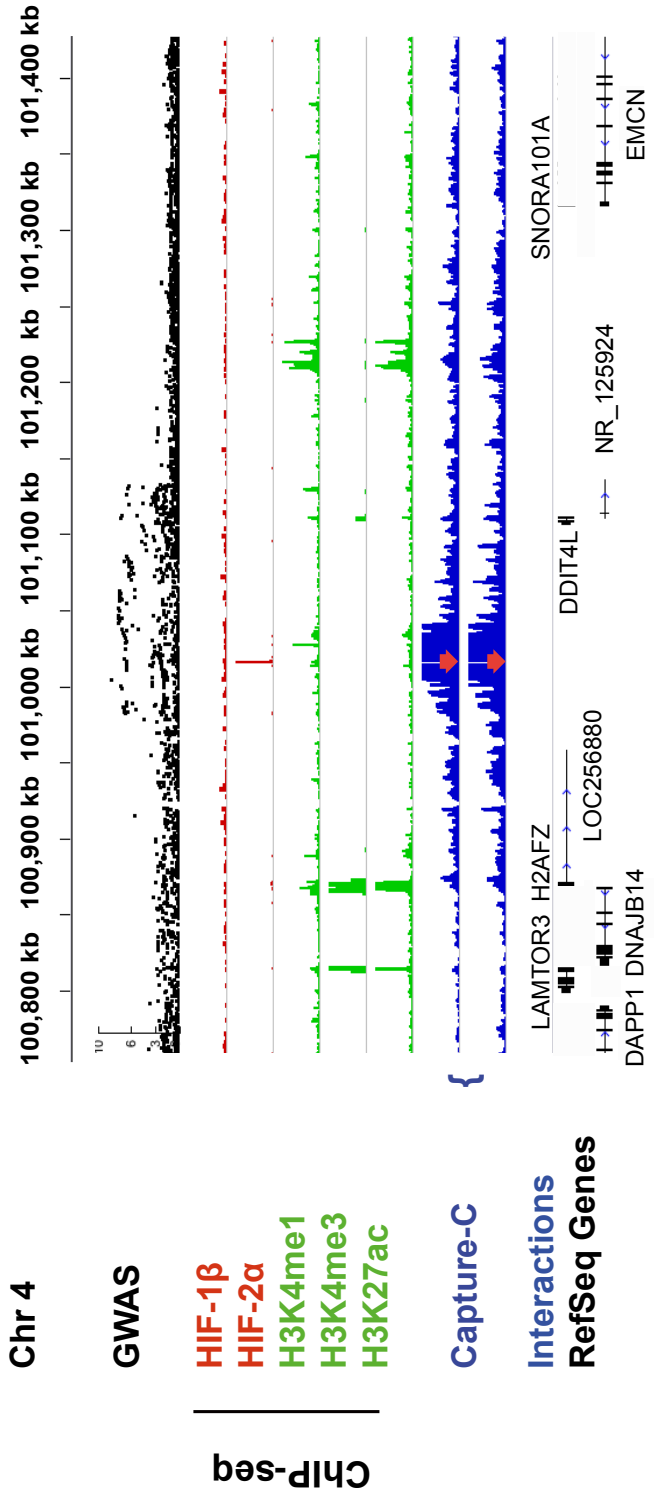

Chr 4
100,800 kb
100,900 kb
101,000 kb
101,100 kb
101,200 kb
101,300 kb
101,400 kb
GWAS
ChIP-seq
HIF-1β
HIF-2α
H3K4me1
H3K4me3
H3K27ac
Capture-C
Interactions
RefSeq Genes
LAMTOR3
H2AFZ
LOC256880
DDIT4L
NR_125924
SNORA101A
EMCN
DAPP1
DNAJB14

Figure 10.1 J: IGV tracks centred at the 4q23 locus. Individual tracks are described in Figure 10.1 A. No genes fall within 25 kb of the LD region at this locus and no statistically significant interactions were detected by Capture-C, indicating looping to more distant gene promoters.

3p22.1 locus:

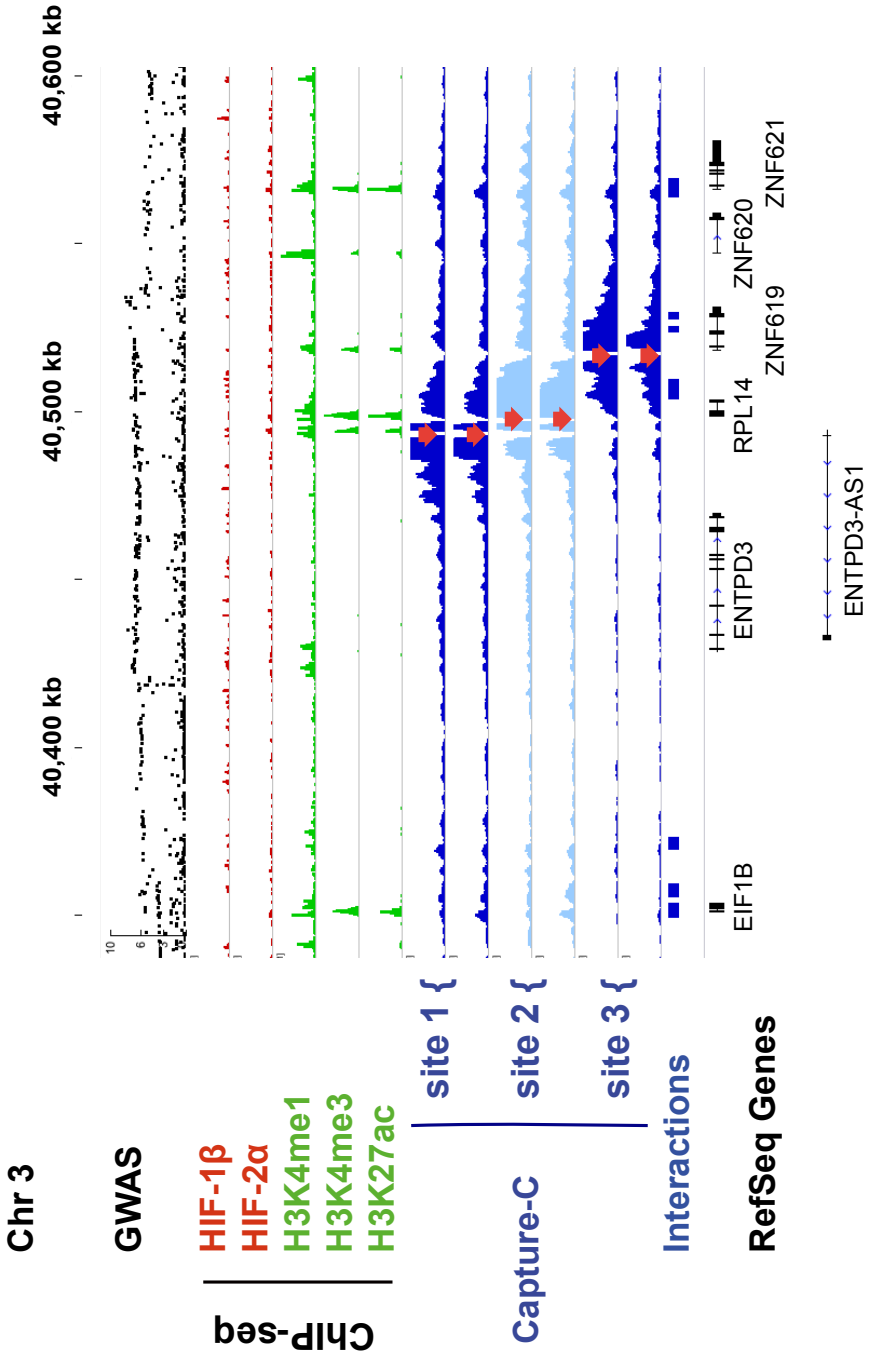

Chr 3
40,400 kb
40,500 kb
40,600 kb
GWAS
ChIP-seq
HIF-1β
HIF-2α
H3K4me1
H3K4me3
H3K27ac
Capture-C
site 1
site 2
site 3
Interactions
RefSeq Genes
EIF1B
ENTPD3
RPL14
ZNF619
ZNF620
ZNF621
ENTPD3-AS1

Figure 10.1 K: IGV tracks centred at the 3p22.1 locus. Individual tracks are described in Figure 10.1 A. ENTPD3, RPL14 ZNF619, ZNF620 locate within +/- 25 kb of the LD region ($r^2 \geq 0.8$) of the locus. EIF1B and ZNF621 gene promoters overlap with statistically significant interacting regions detected by Capture-C.

15q22.31 locus:

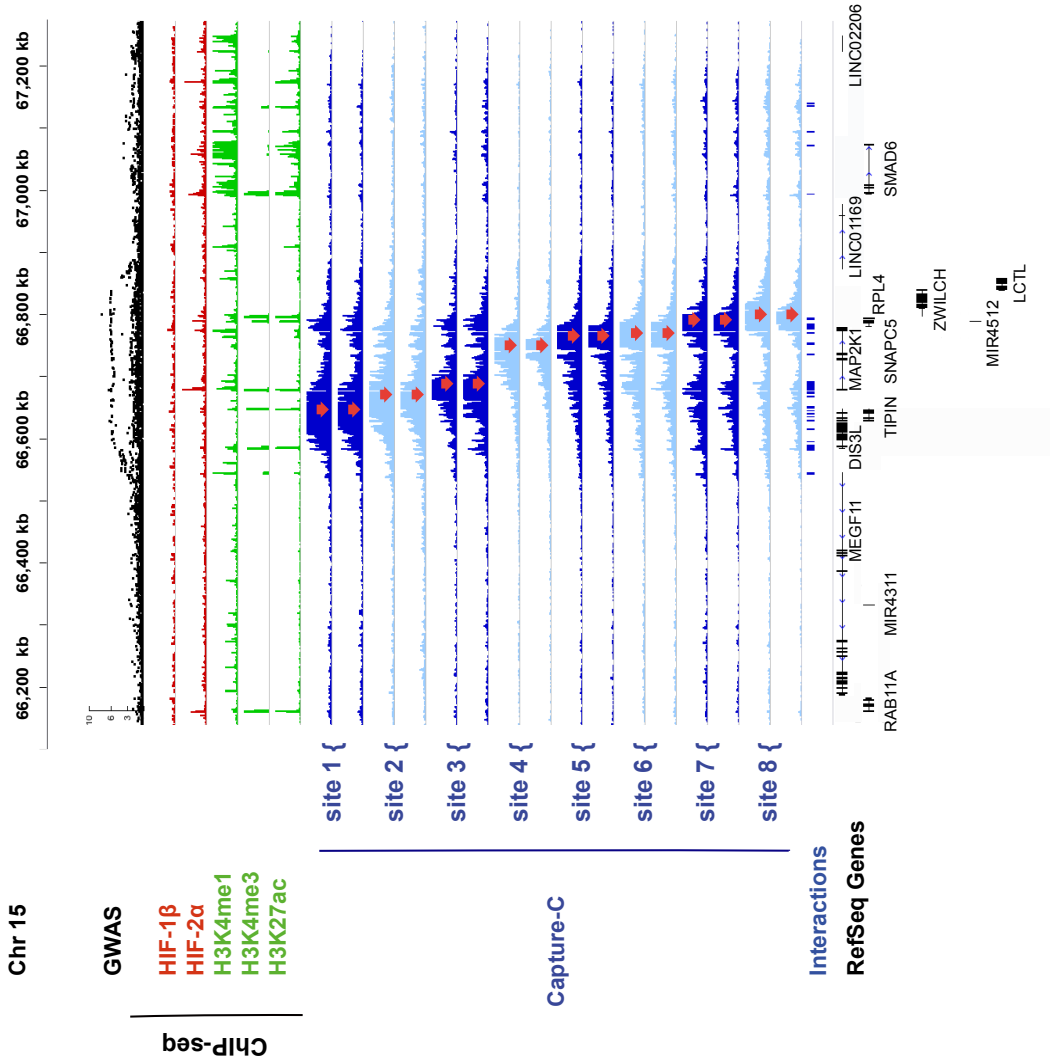


Figure 10.1 L: IGV tracks centred at the 15q22.31 locus. Individual tracks are described in Figure 10.1 A. DIS3L, TIPIN, MAP2K1, SNAPC5, RPL4 and ZWILCH all locate within +/- 25 kb of the LD region ($r^2 \geq 0.8$) of the locus. MEGF11 and SMAD6 gene promoters overlap with statistically significant interacting regions detected by Capture-C.

3q26.31 locus:

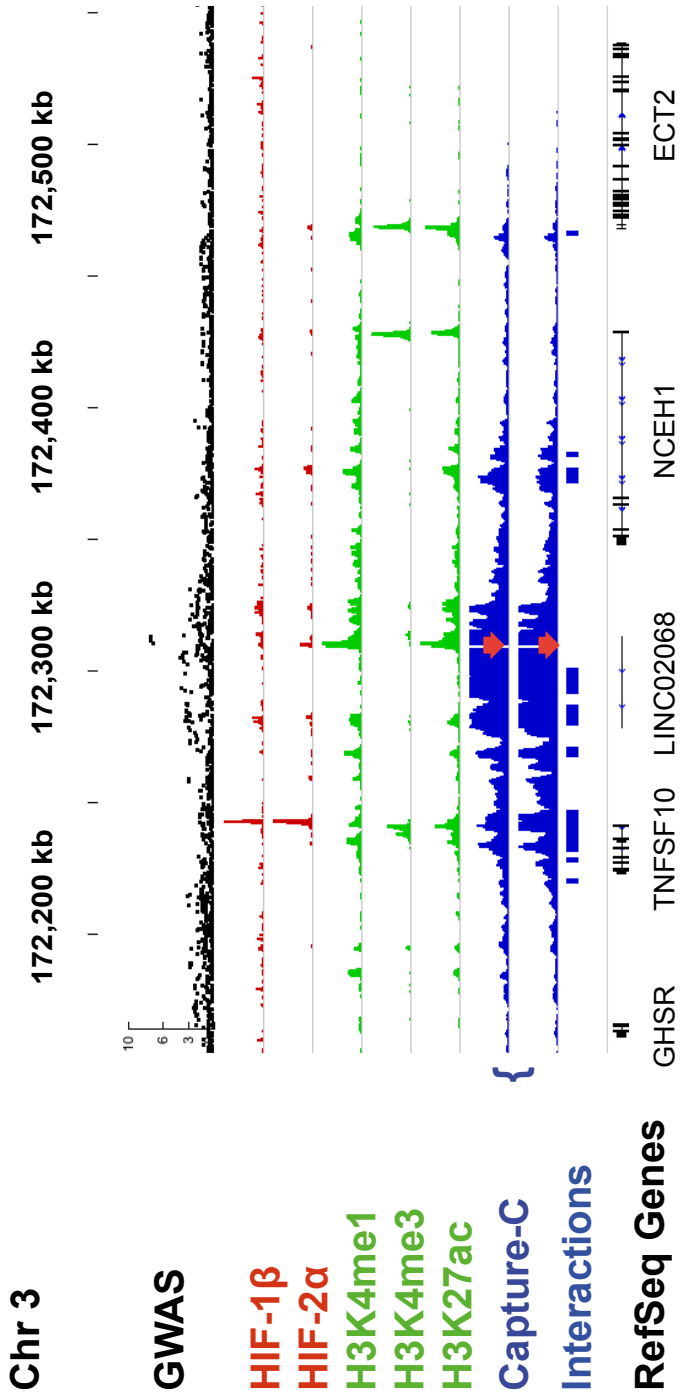


Figure 10.1 M: IGV tracks centred at the 31.26.31 locus. Individual tracks are described in Figure 10.1 A. The non-coding RNA, LINC02068, locate within +/- 25 kb of the LD region ($r^2 \geq 0.8$) of the locus, whilst ECT2 and TNFSF10 gene promoters overlap with statistically significant interacting regions detected by Capture-C.

10.2 IGV tracks sub-threshold RCC susceptibility loci

This section includes IGV tracks of sub-threshold loci that directly overlap with a HIF binding site, as described in chapter 6.2.3.

10q22.1 locus:

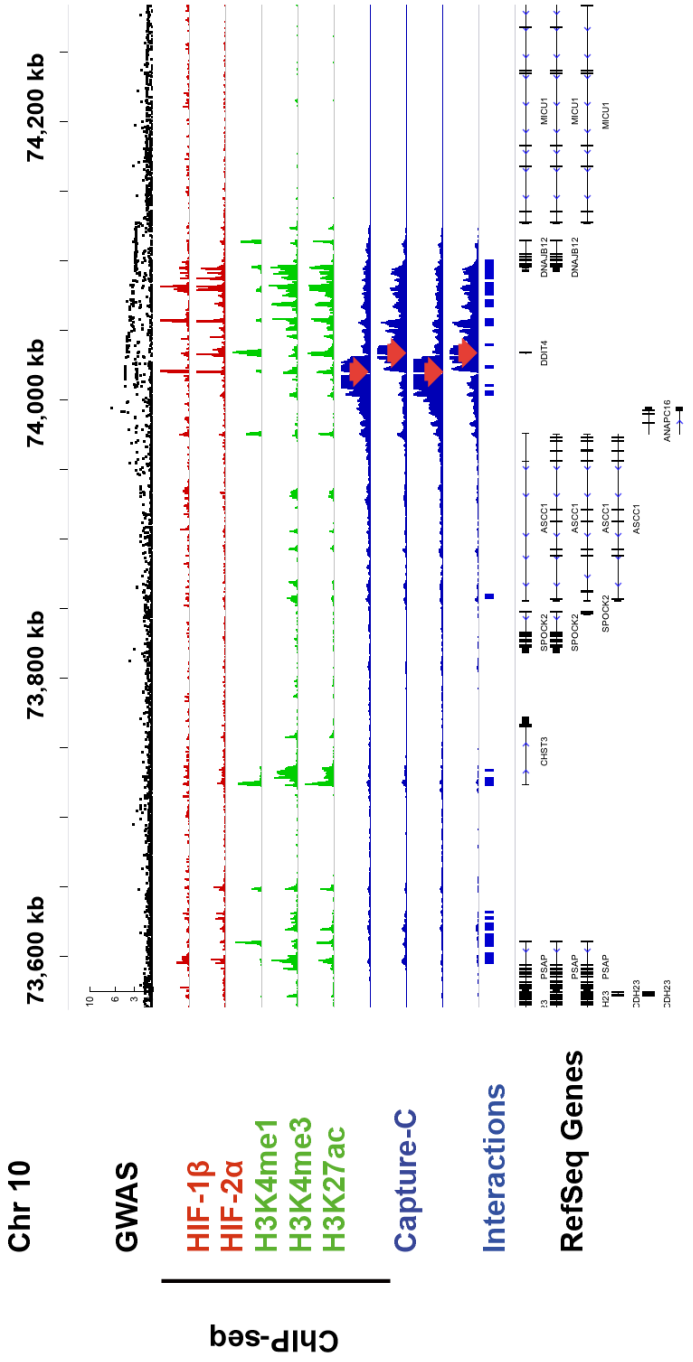


Figure 10.2 A: IGV tracks centred at the 10q22.1 locus. Individual tracks are described in Figure 10.1 A.

15q22.2 locus:

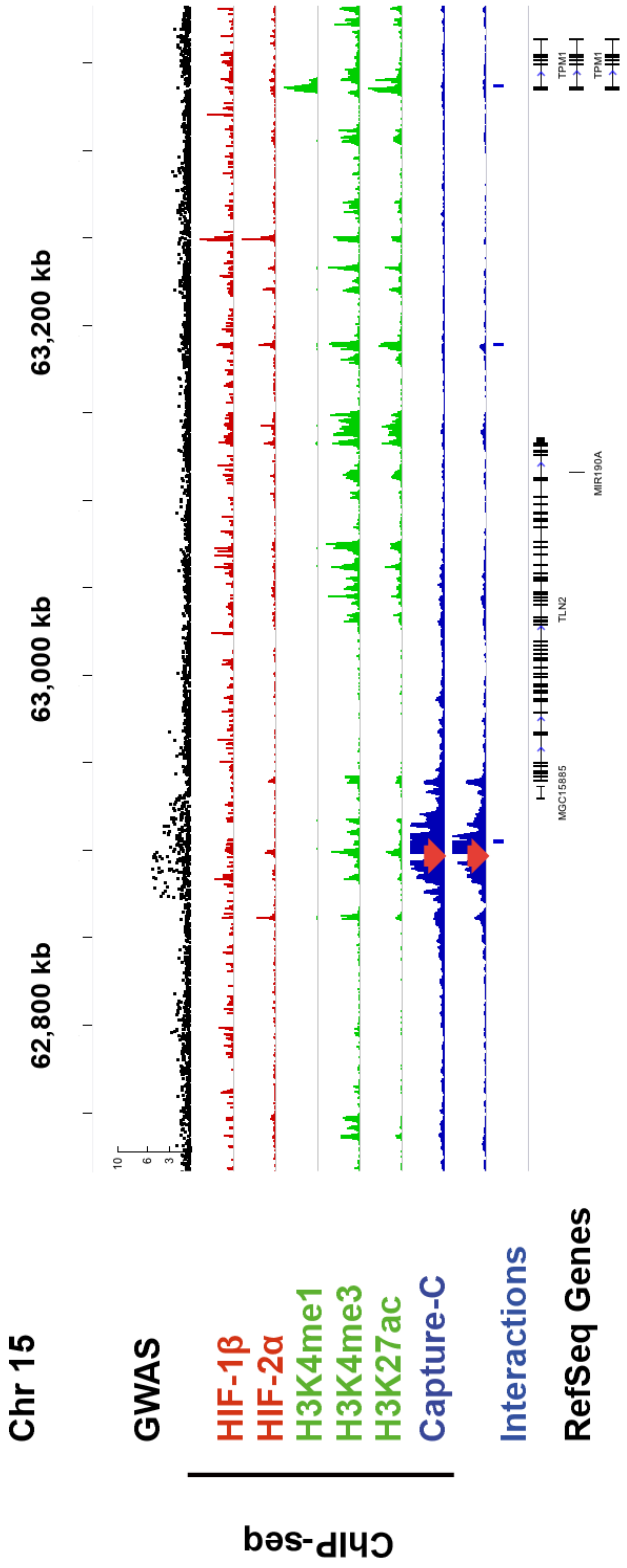


Figure 10.2 B: IGV tracks centred at the 15q22.2 locus. Individual tracks are described in Figure 10.1 A.

6p22.2 locus:

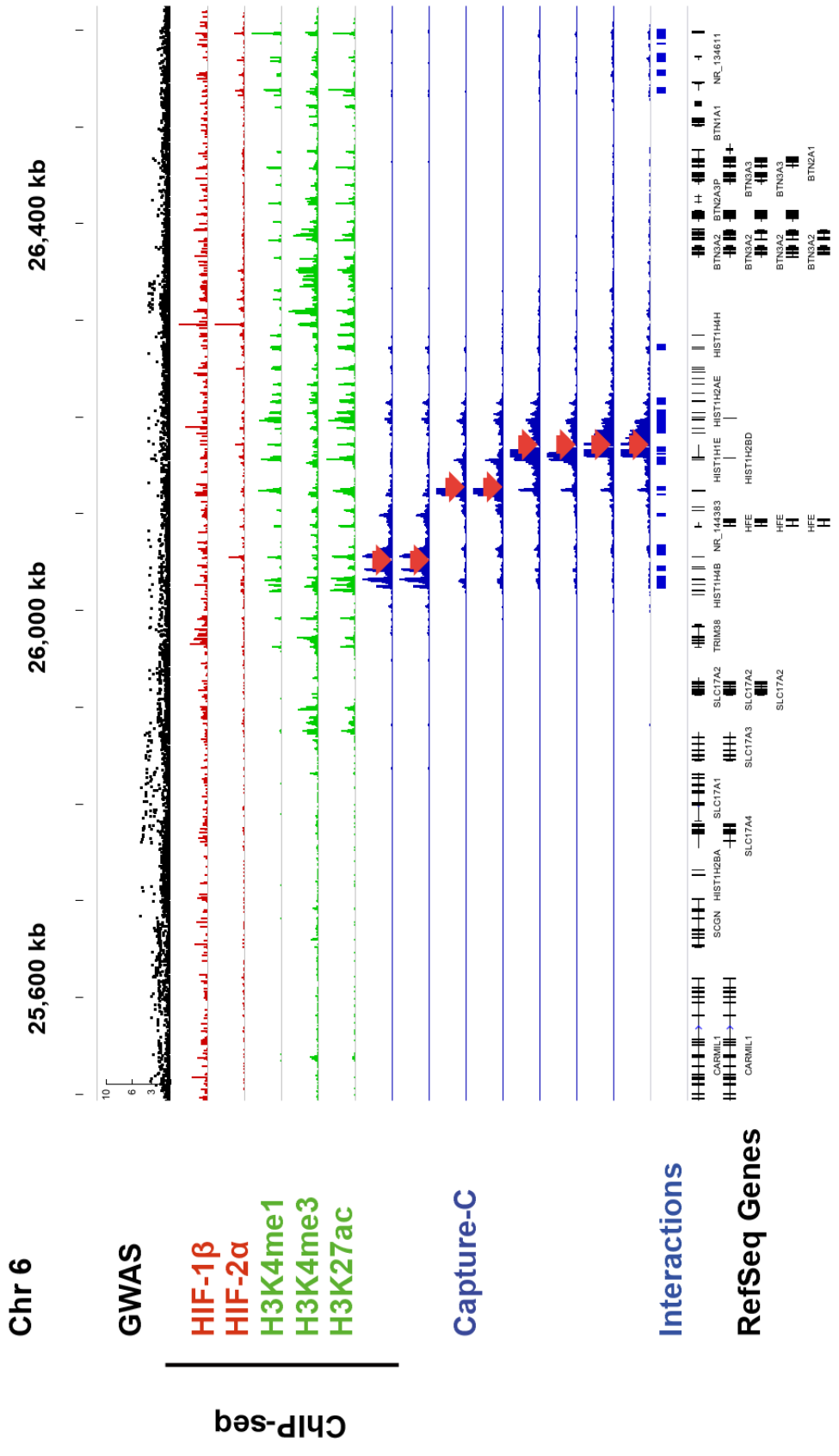

Chr 6
25,600 kb
26,000 kb
26,400 kb
GWAS
ChIP-seq
HIF-1β
HIF-2α
H3K4me1
H3K4me3
H3K27ac
Capture-C
Interactions
RefSeq Genes

Figure 10.2 C: IGV tracks centred at the 6p22.2 locus. Individual tracks are described in Figure 10.1 A.

7p15.3 locus:

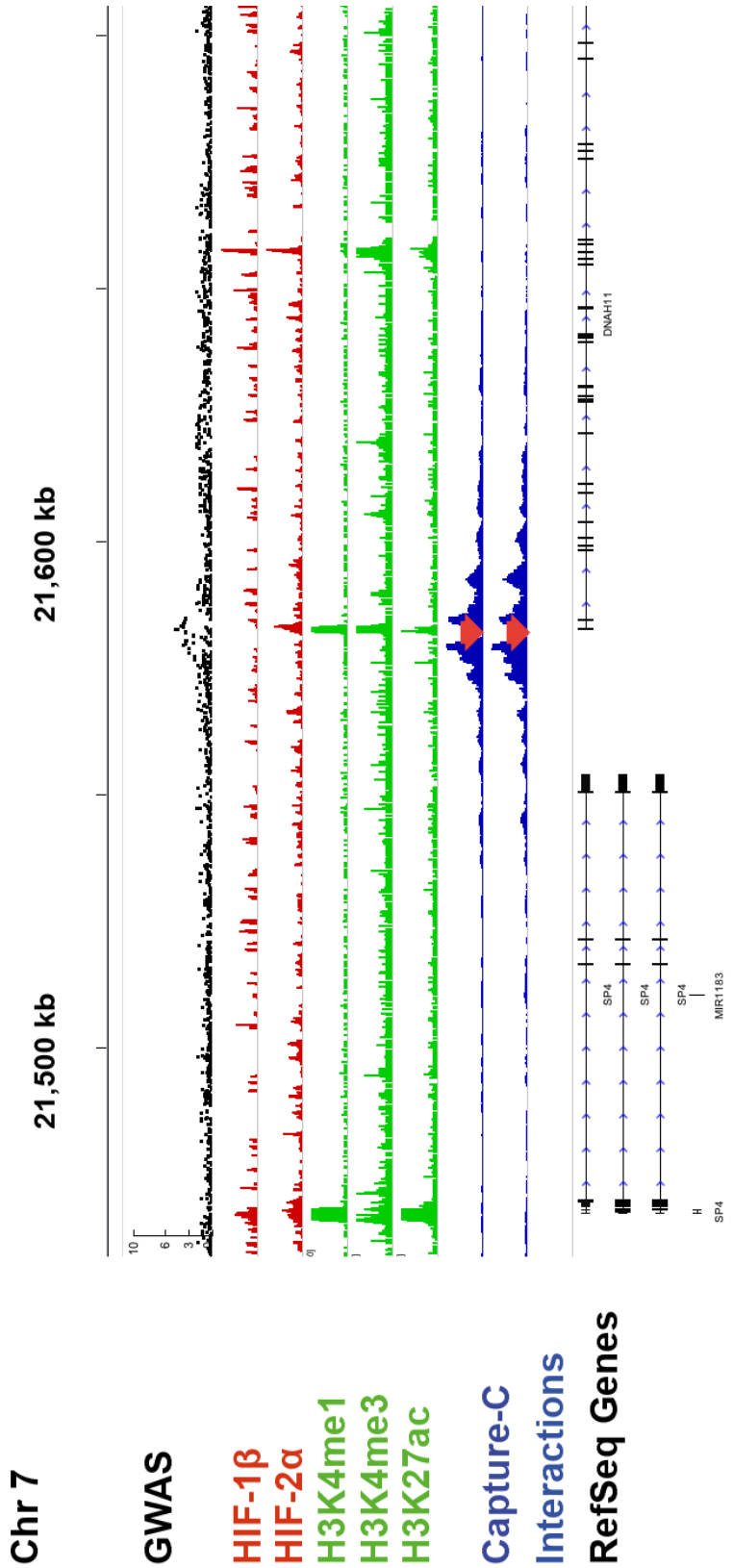

Chr 7
21,500 kb
21,600 kb
GWAS
ChIP-seq
HIF-1β
HIF-2α
H3K4me1
H3K4me3
H3K27ac
Capture-C
Interactions
RefSeq Genes
DNAH11
SP4
SP4
SP4
SP4
MIR1183

Figure 10.2 D: IGV tracks centred at the 7p15.3 locus. Individual tracks are described in Figure 10.1 A.

12q23.1 locus:

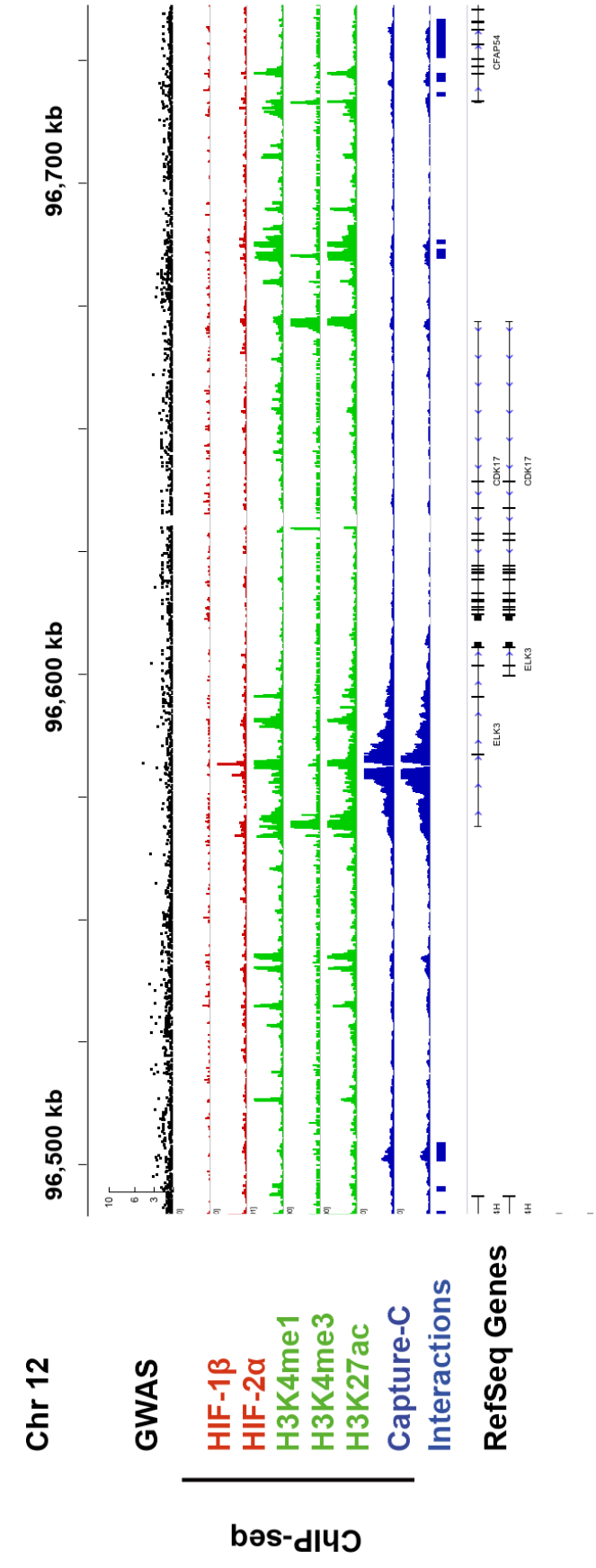


Figure 10.2 E: IGV tracks centred at the 12q23.1 locus. Individual tracks are described in Figure 10.1 A.

6q25.1 locus:

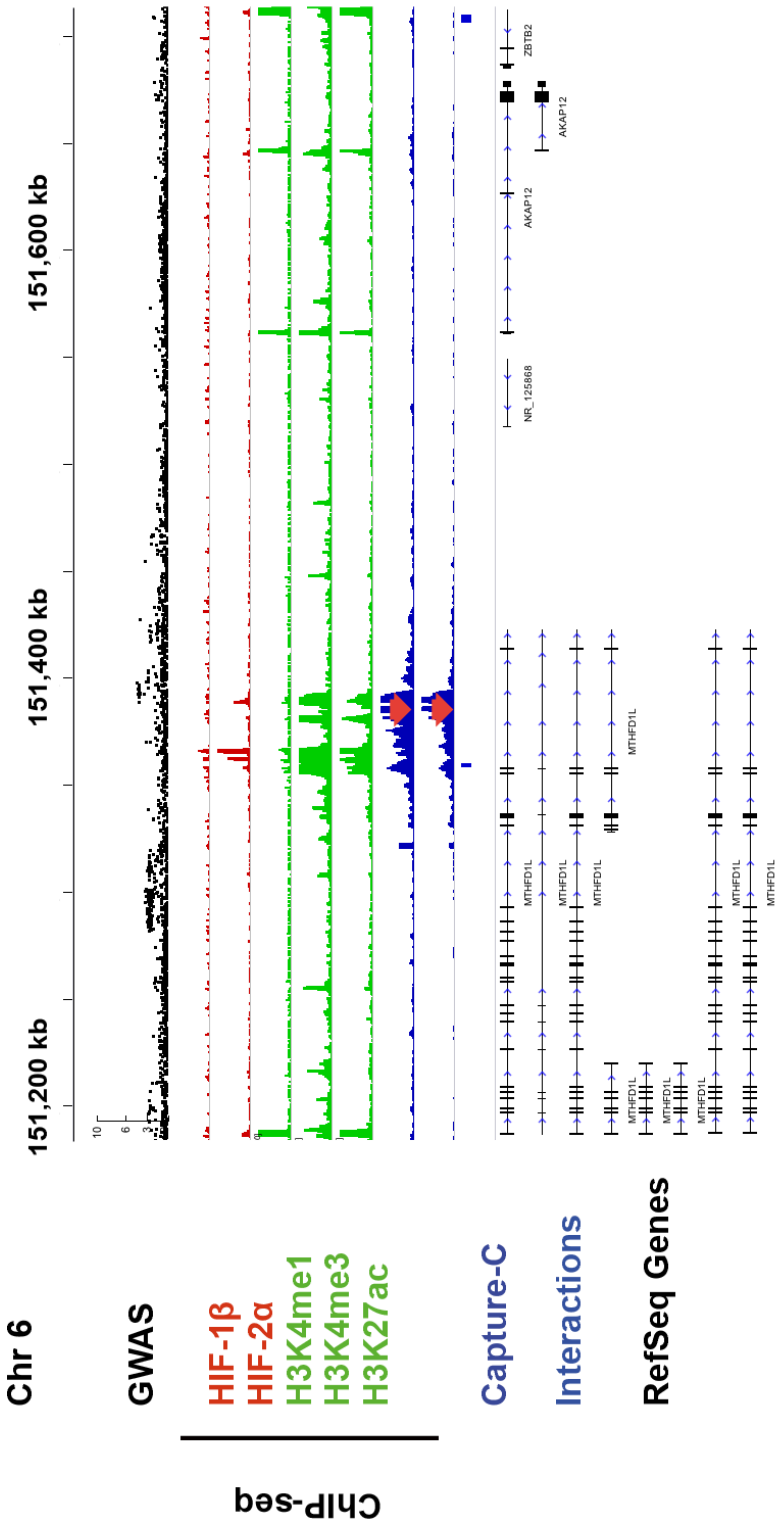

Chr 6
151,200 kb
151,400 kb
151,600 kb
GWAS
ChIP-seq
HIF-1β
HIF-2α
H3K4me1
H3K4me3
H3K27ac
Capture-C
Interactions
RefSeq Genes
MTHFD1L
AKAP12
ZBTB2
NR_125868

Figure 10.2 F: IGV tracks centred at the 6q25.1 locus. Individual tracks are described in Figure 10.1 A.

21q22.3 locus:

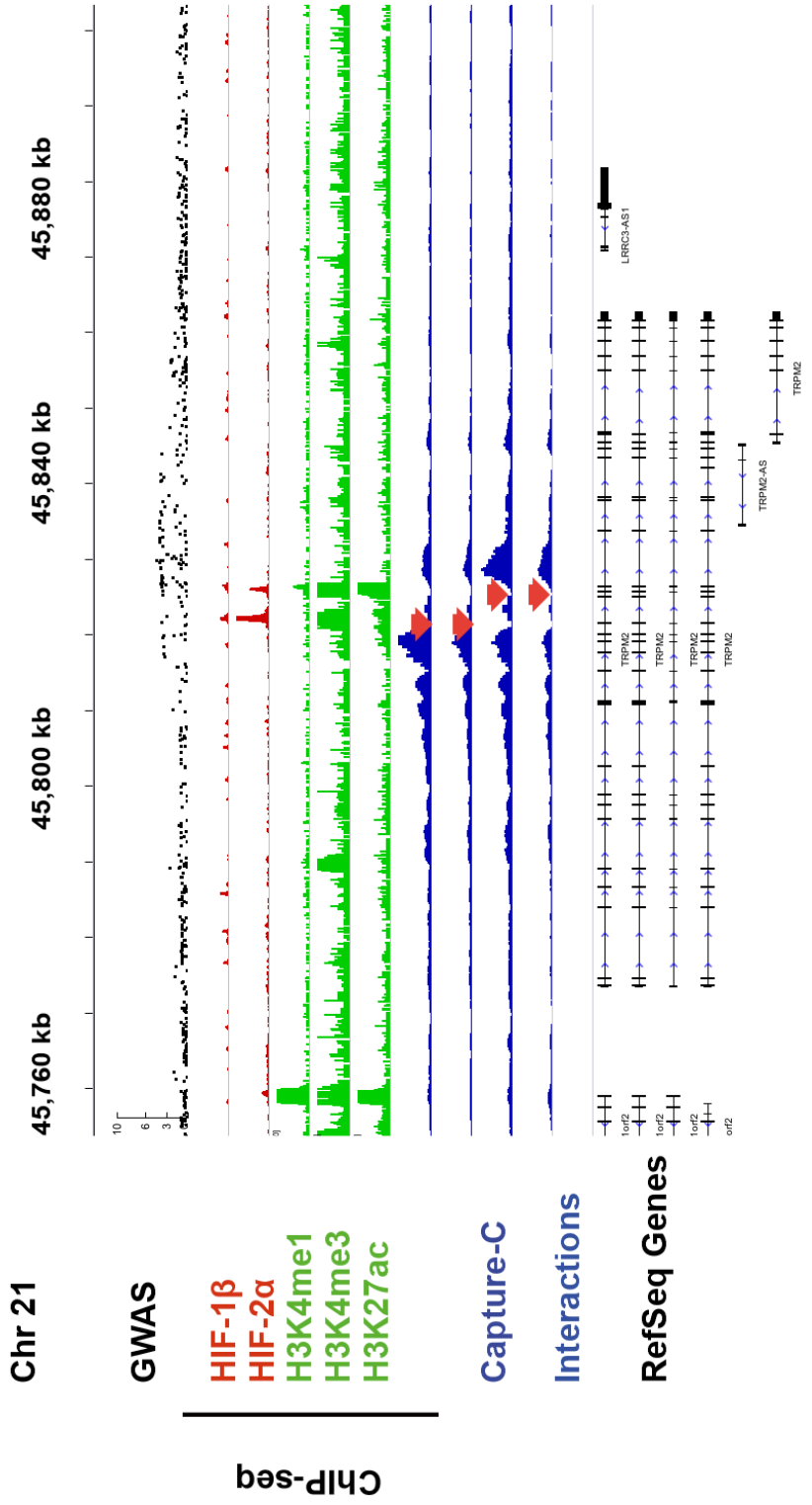


Figure 10.2 G: IGV tracks centred at the 21q22.3 locus. Individual tracks are described in Figure 10.1 A.

7p21.1 locus:

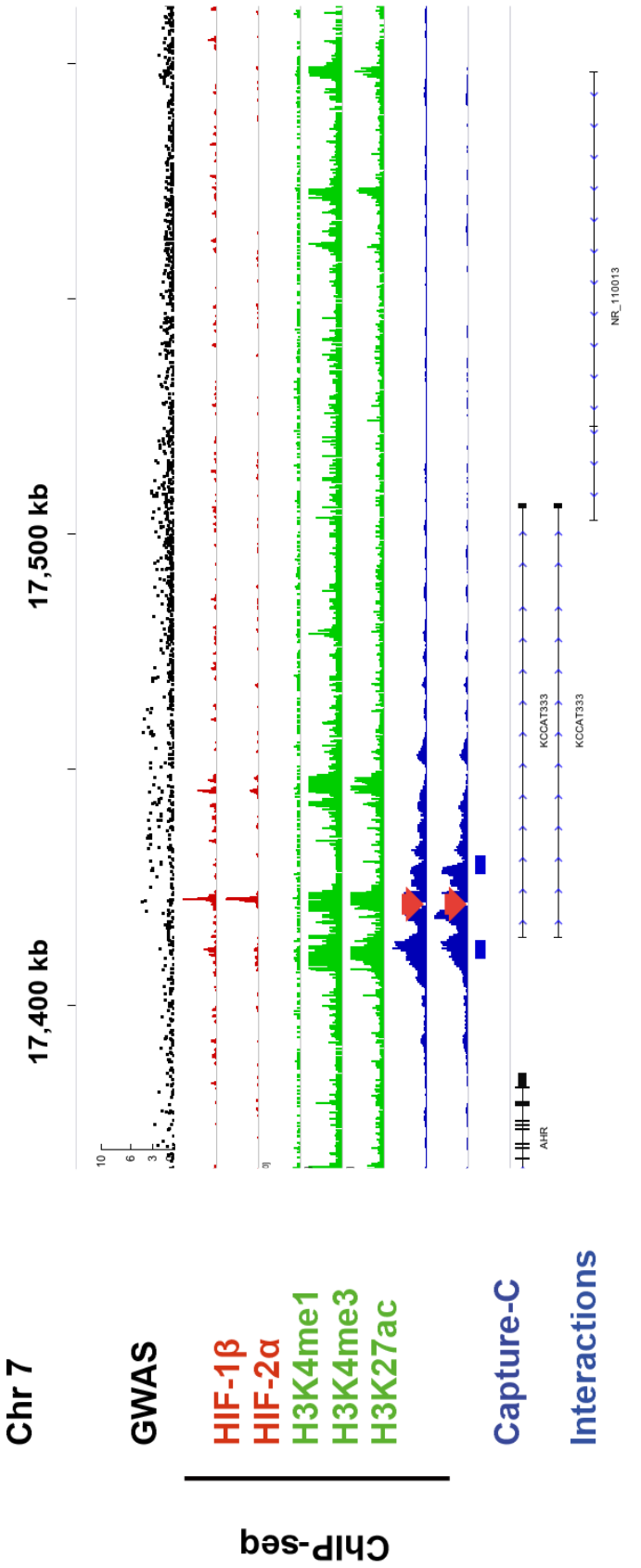

Chr 7
17,400 kb
17,500 kb
GWAS
ChIP-seq
HIF-1β
HIF-2α
H3K4me1
H3K4me3
H3K27ac
Capture-C
Interactions
RefSeq Genes
AHR
KCCAT333
KCCAT333
NR_110013

Figure 10.2 H: IGV tracks centred at the 7p21.1 locus. Individual tracks are described in Figure 10.1 A.

19q13.32 locus:

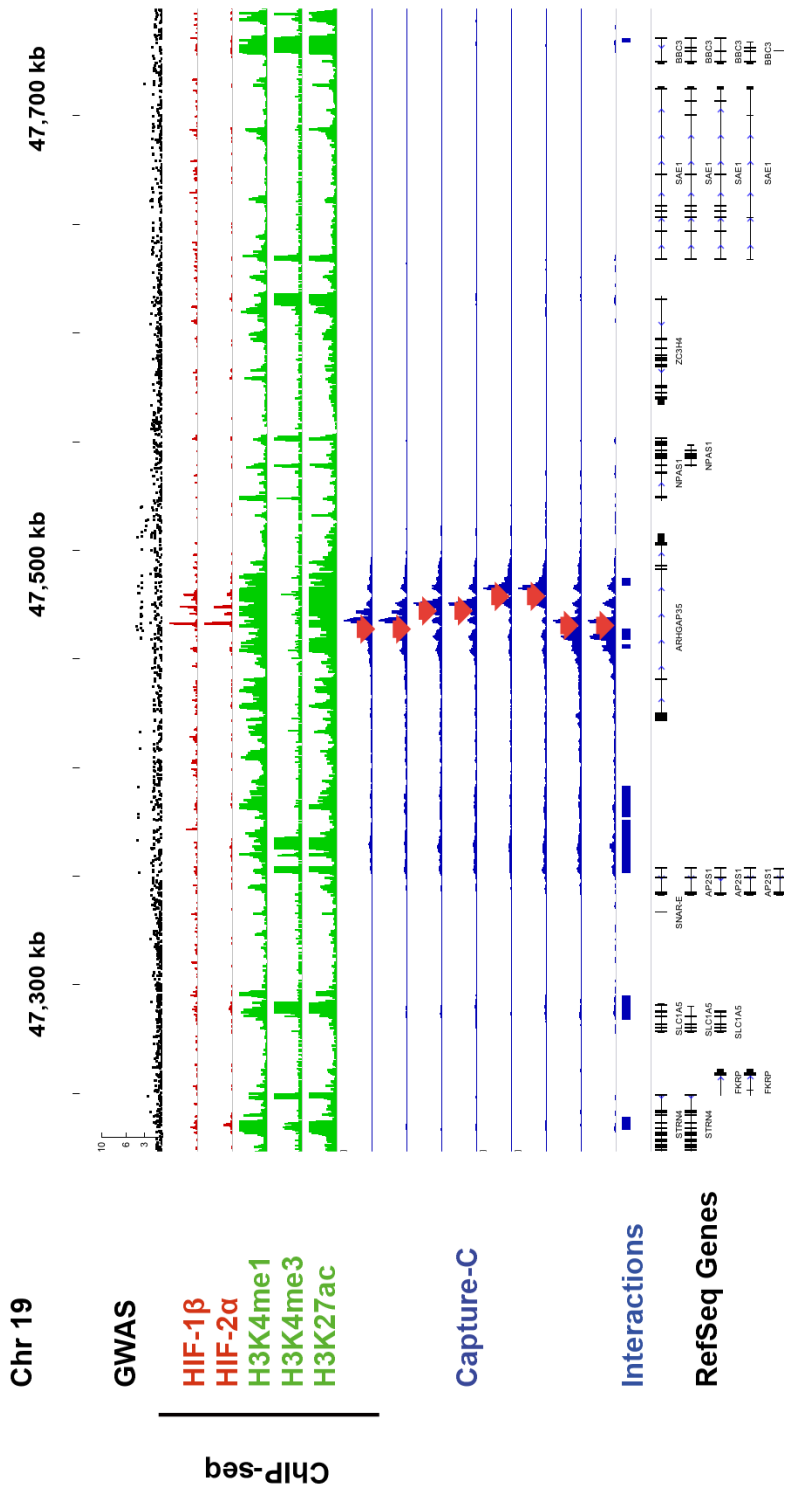


Figure 10.2 I: IGV tracks centred at the 19q13.32 locus. Individual tracks are described in Figure 10.1 A.

18q21.2 locus

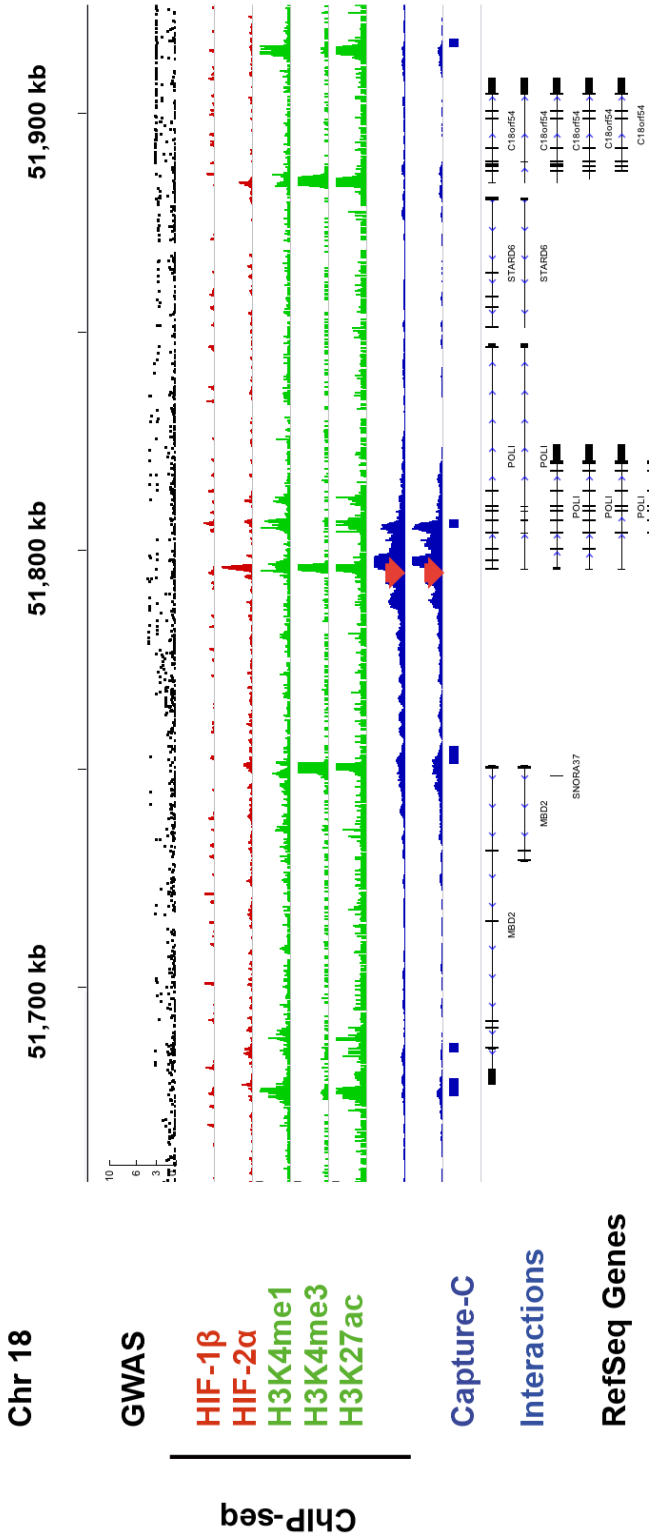


Figure 10.2 J: IGV tracks centred at the 18q21.2 locus. Individual tracks are described in Figure 10.1 A.

10.3 Supplemental Tables

10.3.1 Breast cancer associated GWAS loci that overlap oestrogen receptor binding sites

Locus	*Index SNP*	*Chromosome*	*Position*	*P-value*	*3 nearest genes*
2q35	rs4442975	chr2	217920769	1.1E-95	AC007557.1, TNP1, DIRC3
3p24.1	rs60936670	chr3	27332409	8.0E-64	NEK10, SLC4A7, EOMES
16q23.2	rs7500067	chr16	80648296	4.1E-27	CDYL2, DYNLRB2, CMC2
8q22.3	rs514192	chr8	102478959	5.6E-09	GRHL2, NCALD, ZNF706
16q12.2	rs2432539	chr16	56420987	4.0E-08	AMFR, GNAO1, NUDT21

Table 10.1: Breast cancer suceptability loci that directly overlap with oestrogen receptor binding site.

10.3.2 Prostate cancer associated GWAS loci that overlap androgen receptor binding sites

Locus	*Index SNP*	*Chromosome*	*Position*	*P-value*	*3 nearest genes*
19q13.33	rs11665748	chr19	50851141	3.E-54	NR1H2, NAPSA, KCNC3
8p21.2	rs13272392	chr8	23670998	4.E-34	STC1, NKX2-6, NKX3-1

11q12.3	rs174549	chr11	61803910	2.E-30	AP003733.1, FTH1, BEST1
2q37.1	rs887829	chr2	233759924	9.E-25	NGEF, C2orf82, GIGYF2
2q31.1	rs12621278	chr2	172446825	9.E-23	CYBRD1, DYNC1I2, DCAF17
12q13.13	rs73110464	chr12	52918828	2.E-15	AC055736.1, KRT5, KRT71
7p15.3	rs12155172	chr7	20954872	5.E-13	SP8, ABCB5, ITGB8
17q24.1	rs4351	chr17	63492371	9.E-13	AXIN2, CTD-2535L24.2, CEP112
1q32.1	rs11568818	chr11	102530930	2.E-11	RP11-817J15.3, MMP27, MMP20
20q13.2	rs7582141	chr2	159042977	5.E-11	CCDC148, UPP2, PKP4
2q24.1	rs12480328	chr20	50911385	5.E-11	ZFP64, SALL4, ATP9A
8p21.2	rs11135910	chr8	26034626	8.E-11	PPP2R2A, EBF2, BNIP3L
11p13	rs4378355	chr11	34761870	2.E-10	EHF, APIP, PDHX
3q21.3	rs10934853	chr3	128319530	3.E-10	RPN1, C3orf27, GATA2
10q24.31	rs3850699	chr10	102654464	5.E-10	FAM178A, PAX2, MRPL43
7q22.1	rs6465657	chr7	98187015	1.E-09	NPTX2, BAIAP2L1, BRI3
12q13.11	rs80130819	chr12	48025835	2.E-09	RPAP3, ENDOU, RAPGEF3
18q23	rs7241993	chr18	79013973	2.E-09	PARD6G, AC139100.2, ADNP2
3p24.1	rs481519	chr3	27285723	2.E-09	NEK10, SLC4A7, EOMES

20q11.22	rs10187424	chr2	85567174	3.E-09	RETSAT, TGOLN2, ELMOD3
2p11.2	rs11907546	chr20	34131991	3.E-09	ERGIC3, C20orf173, CEP250
7p15.2	rs10486567	chr7	27936944	3.E-09	JAZF1, TAX1BP1, HIBADH
4q24	rs7679673	chr4	105140377	4.E-09	CXXC4, TACR3, TET2
5q14.3	rs35148638	chr5	87315172	6.E-09	TMEM161B, CCNH, RASA1
6q21	rs2273669	chr6	108963986	8.E-09	FOXO3, LACE1, ARMC2
16q22.2	rs12051443	chr16	71657426	3.E-08	MARVELD3, PHLPP2, TAT

Table 10.2: Prostate cancer susceptibility loci that directly overlap with androgen receptor binding site.

10.3.3 Genes set containing three closest genes to sub-threshold RCC susceptibility loci

ABCA1, ABCB10, ABCB5, ABCC1, ABHD10, ABHD4, AC002310.13, AC002451.1, AC004466.1, AC016885.1, AC026202.1, AC062017.1, AC079612.1, AC110781.3, ACTR3B, ADAM15, ADAMTS9, ADAP1, ADD2, ADD3, ADSSL1, AE000662.92, AGBL3, AGPAT5, AGR3, AHNAK, AHR, AKAP12, AL137002.1, ALG14, ALS2, ANAPC16, ANGPT2, ANKRD42, ANKRD7, APCDD1L, ARHGAP10, ARHGAP22, ARHGAP35, ARHGAP39, ARHGAP42, ARHGEF10, ARHGEF38, ARID5B, ARL13B, ASB3, ASB4, ASCC1, ASCC2, ASPHD2, ASRGL1, ATAD1, ATAD2, ATF6B, ATF7IP2, ATP11A, ATP1B3, ATP6V1F, ATP8B1, AXDND1, BEND3, BMF, BMPR1B, BNC2, BOD1L1, BPGM, BRINP1, BRINP3, BTBD11, BTBD3, BUB1B, C1, f234,

C1, f61, C1, f94, C10, f107, C12, f36, C14, f159, C16, f97, C18, f56, C19, f40, C1D, C21, f37, C22, f23, C2CD4B, C3, f38, C4, f33, C5, f22, C5, f30, C5, f47, C6, f203, C7, f31, C7, f33, C8, f87, C9, C9, f92, CACNA1D, CALD1, CAND1, CASC10, CASP9, CATSPER4, CCBE1, CCBL2, CCDC11, CCDC171, CCDC73, CCDC82, CCDC85A, CCDC90B, CCL24, CCL26, CCNB1, CD36, CDCA7, CDCA7L, CDH13, CDH4, CDK15, CDK17, CDK5RAP2, CDKAL1, CDKN1C, CDYL, CELA2A, CELA2B, CELSR1, CENPQ, CEP78, CEP85, CEP89, CERK, CGGBP1, CHAC2, CLEC19A, CLEC3A, CLEC4F, CLINT1, CLN8, CLNK, CLPTM1L, CNEP1R1, CNGB3, CNKSR1, CNN3, CNTN5, CNTNAP2, COA6, COG1, COL25A1, COL4A1, COMMD9, CPEB2, CPEB4, CPNE3, CPS1, CRY1, CSMD2, CTAGE1, CTB-167G5.5, CTC-490E21.12, CTTNBP2, CUL1, CXADR, CXCR4, CYCS, CYLD, CYP21A2, CYP26A1, CYP26C1, CYP2A6, CYP2A7, DAB2, DACH1, DAD1, DARS, DCLK2, DCP1A, DCPS, DCST1, DCTD, DDC, DDIT4, DEAF1, DIAPH3, DLEU1, DLEU7, DLG2, DLGAP1, DLGAP2, DMRT1, DMRT3, DMXL1, DNAH11, DNAJC30, DNM1L, DPRX, DPYSL3, DRD4, DROSHA, DTWD2, DUSP6, DYRK2, E2F3, EDEM1, EFHB, EFNA4, EIF3L, ELFN1, ELK3, ELOVL6, ELP2, EMP1, ENOSF1, ENPEP, EP300, EPB41L4B, EPHA3, EPHA6, EPHB2, ERBB4, ETAA1, ETNPPL, ETV7, EXO1, F11, F13A1, FABP7, FAM92A1, FAN1, FANCL, FARP1, FBLN2, FBXO32, FBXO5, FECH, FHL5, FHOD3, FIGLA, FIGNL1, FLCN, FOPNL, FOXC2, FOXL1, FRRS1L, FSTL5, FXYD2, FXYD6, FXYD6-FXYD2, FZD7, GALNT4, GALP, GALR1, GATA6, GBE1, GCOM1, GDAP1L1, GET4, GIN1, GIPC1, GK5, GLRX5, GLYATL3, GNAT3, GPR17, GPR37, GPR63, GPR68, GPR75-ASB3, GRAMD4, GRB10, GRIN2A, GRIN2B, GRM7, GSTM2, GSTM3, GSTM5, GTF2B, GTF2E2, HAAO, HDAC11, HDAC4, HDAC7, HEATR3, HHAT, HIST1H2AA, HIST1H2BA,

HJURP, HLF, HMGB4, HPS4, HS3ST1, HSD17B2, HTR3A, HTR3B, IFNG, IFNL1, IFNL2, IFNL3, IGFBP7, INF2, INSR, IPO5, IRF2BP2, IRF5, IRS2, ISM1, ITGB8, JADE1, JPH2, KANK1, KBTBD12, KCNIP4, KCNQ1, KCNRG, KCTD20, KIRREL3, KLF6, KLHL1, KLHL3, KLLN, KMT2C, LACTBL1, LANCL1, LAPTM5, LDLRAD3, LHPP, LIMS2, LPCAT1, LPHN3, LRBA, LRIG3, LRRC18, LRRC3, LRRC31, LRRC34, LRRC37B, LRRC3DN, LRRIQ4, LSM11, LTA4H, LURAP1L, LY86, LYZL2, MAB21L2, MACC1, MAD1L1, MAF, MAGI1, MAML2, MAP1LC3C, MARCH4, MATN1, MBD1, MBD2, MBD3L3, MBP, MC4R, MCF2L, MCPH1, MED22, MEF2D, MEIS1, MGC4294, MGLL, MICALL1, MLLT10, MLXIPL, MMD, MOCOS, MPHOSPH6, MPRIP, MROH2A, MTERFD3, MTHFD1L, MTHFSD, MTMR10, MTMR2, MTMR3, MTNR1A, MTRF1L, MTRNR2L12, MTSS1, MUT, MYH11, MYL12B, MYO16, MYO5B, MYO7B, MYOF, MYT1L, MYZAP, MZT1, NAA16, NAA38, NAF1, NARS, NDFIP2, NDRG1, NDUFB9, NIN, NIPSNAP3A, NIPSNAP3B, NKIRAS1, NKX1-2, NKX3-2, NPAS1, NPHS2, NPVF, NPY1R, NR3C2, NSG2, NSUN3, NT5E, NTN1, NUP210, OAT, OBFC1, OLA1, OR5H1, OR5H14, OR5H15, OSBPL10, OSER1, OSTC, PACRGL, PAK6, PAPD5, PCDH17, PCDH20, PCDH7, PDHA2, PDK4, PDZD2, PEBP4, PFKP, PHLDB2, PIH1, PIK3R1, PITRM1, PITX2, PKIB, PKN1, PKN2, PKP2, PLA2G4A, PLEKHG1, PMAIP1, POC1B, POC1B-GALNT4, POLI, POLR2M, POT1, PP2D1, PPA2, PPIP5K2, PPP1R3G, PPP2R2B, PRMT10, PROS1, PRR20D, PRR20E, PRR5L, PRRG4, PSAT1, PTEN, PTGDR, PTGER1, PTGER2, PTGER4, PTGS2, PTPRD, PXT1, PYGL, QRSL1, QSER1, RAB31, RAB5A, RBBP8, RBM26, RBPMS, RBX1, RD3, RGCC, RGS17, RHBDD2, RHBG, RHOBTB2, RHPN2, RMDN1, RNF113B, RNF219, ROBO1, ROBO2, RP1, RP11-133K1.2, RP11-180C1.1, RP11-215A19.2, RP11-272B17.2, RP11-295D22.1, RP11-

297M9.1, RP11-386G21.1, RP11-41O4.1, RP11-429E11.3, RP11-451M19.3, RP11-45M22.4, RP11-625H11.1, RP4-583P15.14, RPL7A, RPP40, RPS6KA5, RREB1, RTKN2, RTTN, SALL1, SAV1, SCGB1A1, SCGN, SCLT1, SCT, SDC3, SEC61A1, SEMA3C, SERINC5, SERTAD4, SEZ6L, SH3PXD2A, SIVA1, SKIDA1, SLC16A7, SLC22A18AS, SLC25A26, SLC2A4RG, SLC30A1, SLC30A5, SLC39A11, SLC44A3, SLC6A3, SLIT2, SLK, SMIM18, SMPDL3A, SNRPN, SNURF, SNX13, SNX14, SOCS6, SOX4, SP3, SP4, SPOCK1, SPTLC3, SPZ1, SSTR2, ST3GAL1, ST3GAL4, STARD6, STIM2, STK32A, STT3B, STX19, STX8, STXBP4, SUN1, SURF6, SUZ12, SYT14, SYT17, TAF4, TAF5L, TAGLN3, TARBP1, TAS2R7, TAS2R8, TAS2R9, TATDN1, TBX18, TCL1A, TCL1B, TDRD3, TDRD5, TECRL, TENM3, TET2, TFDP2, TGIF1, THADA, THG1L, THSD7B, TKT, TLE4, TLN2, TMC5, TMEM133, TMEM160, TMEM169, TMEM229A, TMEM245, TNFAIP8, TNFRSF10B, TNPO3, TNXB, TRAF5, TRAPPC12, TRPC7, TRPM1, TRPM2, TRPM8, TSSC1, TXNDC16, TXNDC2, TYMS, TYRP1, UBE2E1, UBE2E2, UFL1, UNC5C, UQCR10, URB2, USP25, UTP6, VAPA, VAPB, VAT1L, VDR, VPS37D, VRK2, VWA8, WDFY4, WDR16, WDR64, WDYHV1, WWC2, WWOX, XKR4, XPNPEP1, XPNPEP3, XRCC2, XRCC5, YARS2, ZBTB16, ZBTB46, ZEB1, ZFAT, ZFP36L2, ZFYVE16, ZNF236, ZNF251, ZNF331, ZNF34, ZNF438, ZNF518B, ZNF557, ZNF688, ZNF764, ZNF860, ZSCAN5B, ZSCAN5C

10.3.4 Genomic coordinates of Capture-C oligos

chromosome	*start*	*end*
chr1	26606373	26606493
chr1	50940907	50941027
chr1	50941154	50941274
chr1	50992473	50992593
chr1	50992590	50992710
chr1	50993661	50993781
chr1	51069440	51069560
chr1	51086777	51086897
chr1	51086931	51087051
chr10	105677527	105677647
chr10	105678325	105678445
chr10	74019880	74020000
chr10	74021108	74021228
chr10	74033444	74033564
chr10	74033936	74034056
chr10	89622238	89622358
chr10	89622301	89622421
chr11	108337896	108338016
chr11	108368505	108368625
chr11	108382270	108382390
chr11	108382797	108382917
chr11	69063141	69063261
chr11	69234206	69234326
chr11	69234497	69234617
chr11	69234812	69234932
chr11	69236338	69236458
chr11	69251621	69251741
chr11	69252082	69252202
chr12	125322779	125322899
chr12	26451882	26452002
chr12	26452432	26452552
chr12	26471516	26471636
chr12	26492641	26492761
chr12	26493464	26493584
chr12	96613646	96613766
chr12	96613695	96613815
chr14	73267875	73267995
chr14	73268300	73268420
chr15	62899185	62899305
chr15	66648115	66648235

chr15	66648600	66648720
chr15	66662187	66662307
chr15	66662495	66662615
chr15	66679729	66679849
chr15	66748305	66748425
chr15	66748860	66748980
chr15	66764854	66764974
chr15	66765509	66765629
chr15	66774531	66774651
chr15	66775143	66775263
chr15	66789735	66789855
chr15	66790184	66790304
chr15	66796512	66796632
chr15	66797675	66797795
chr16	10823234	10823354
chr16	10823528	10823648
chr16	10823715	10823835
chr16	31119498	31119618
chr16	31120217	31120337
chr18	51796296	51796416
chr19	47466800	47466920
chr19	47471111	47471231
chr19	47471155	47471275
chr19	47474226	47474346
chr19	47479682	47479802
chr19	47481089	47481209
chr19	7196414	7196534
chr19	7197417	7197537
chr19	7198409	7198529
chr2	46525829	46525949
chr2	46526314	46526434
chr2	46532980	46533100
chr2	46534782	46534902
chr2	46537855	46537975
chr2	46547178	46547298
chr2	46547554	46547674
chr2	46552302	46552422
chr2	46552572	46552692
chr2	46558642	46558762
chr2	46559648	46559768
chr20	42939793	42939913
chr20	42940031	42940151
chr20	48213024	48213144

chr21	45822033	45822153
chr21	45822777	45822897
chr21	45826133	45826253
chr21	45826512	45826632
chr22	41346575	41346695
chr22	41347409	41347529
chr22	41686842	41686962
chr3	172309649	172309769
chr3	172310704	172310824
chr3	40493835	40493955
chr3	40494767	40494887
chr3	40499354	40499474
chr3	40518796	40518916
chr4	101017255	101017375
chr6	151389175	151389295
chr6	26022104	26022224
chr6	26023538	26023658
chr6	26055845	26055965
chr6	26123848	26123968
chr6	26124529	26124649
chr6	26159109	26159229
chr6	26159255	26159375
chr6	26172249	26172369
chr7	134854873	134854993
chr7	17422066	17422186
chr7	17422779	17422899
chr7	2001034	2001154
chr7	2001736	2001856
chr7	21582420	21582540
chr7	21583849	21583969
chr8	22875370	22875490
chr8	23195202	23195322
chr8	23196739	23196859
chr8	23202276	23202396
chr8	30354177	30354297
chr8	94373141	94373261
chr8	94373648	94373768

10.4 Code

10.4.1 LD pruning of summary statistics

SNP pruning (with plink version 1.90b2n)

1. **Matrix indicating relational r^2 between each of the SNPs in summary statistic**

```
plink --bfile ${TG}      merged_filtered_TG/merged       --
extract <rs_numbers> --make-bed --out <r_2_matrix>

/apps/well/plink/1.90b2n/plink --bfile <r_2_matrix> --r2 --
matrix
```

2. **R script called to identify independent loci based r^2 values**

```
#Read in table

plink      <-      read.table("plink.ld",      quote="\"",
comment.char="")

merged <- read.delim("temp/sub.bim", header=FALSE)

rs_numbers  <-  read.table("temp/rs_num.txt",  quote="\"",
comment.char="")


rownames(plink)=merged$V2

colnames(plink)=merged$V2

df=plink


#rank matrix by p-vlaue
```

```
rs_ranked=Henrion_rs_subTH[match(as.vector(Henrion_rs_subT
H[,1]),as.vector(merged$V2) , nomatch=FALSE)>0,]

inx=match(rs_ranked,colnames(plink))

df_ranked=df[inx,inx]

counter=0

names=c()

df_rsnum=data.frame()


#produces index SNPs

while (length((rownames((df_ranked)))>=1) {

  counter = counter+1

  df_rsnum[counter,1]=rownames(df_ranked[1,])

  df_ranked=
df_ranked[df_ranked[1,]<=0.2,df_ranked[1,]<=0.2]


  print(counter)

  print(min(df_ranked))

}


#write file of all index SNPs

write(df_rsnum[,1], file = "temp/pruned_rs.txt",sep = "/n")

#Assignement of each LD SNP to index SNP that is is in highes
LD with

df_ranked=df[inx,inx]


#df shows LD values between index SNPs and SNPs in high LD

df_reduced = df_ranked[df_rsnum[,1], ]


#assign LD SNP to index SNP is is in highest LD with

inx_ldSNP = apply( df_reduced ,2,function(x) which.max(x) )
```

```
ld_ldSNP = apply( df_reduced ,2,function(x) max(x) )


#dataframe with INX SNP und LD SNP

df_sum=data.frame(rownames(df_reduced)[inx_ldSNP],names(in
x_ldSNP), ld_ldSNP )


xargs -I {} grep "\<{}\>" GWAS_<cancertype>.txt <
temp/pruned_rs.txt |awk -F "\t" '{ print
"chr"$12"\t"$13"\t"$13"\t"$22"\t" $28}'>
temp/pruned_rs_summary.tab
```

3. Exclusion of a 1 Mb window around index SNPs

```
awk '{print $1 "\t" $2-500000 "\t" $3+500000 }'
temp/pruned_rs_summary.tab | sort -k1,1 -k2,2n >
temp/exclusionregion_sorted.bed

bedtools merge -i temp/exclusionregion_sorted.bed >
temp/exclusionregion_sortedMerged.bed


bedtools intersect -wb -wa -a temp/pruned_rs_summary.tab -b
temp/exclusionregion_sortedMerged.bed > temp/intersect.bed

awk -F"\t" '{ if(!seen[$7]++) print
$1"\t"$2"\t"$3"\t"$4"\t"$5}' temp/intersect.bed >
temp/pruned_windowex_rs_summary.tab

awk '{print $4}' temp/pruned_windowex_rs_summary.tab >
temp/pruned_windowex_rs.txt
```

4. Assign SNP in $r^2 \geq 0.8$ with index SNPs and assessment of HIF overlap

```
while read p; do

    for n in {1..22};

    do

        echo "Getting LD for chr${n}...";
```

```
          /apps/well/plink/1.90b2n/plink            --bfile
${TG}/filtered_TG/filtered${n} \

               --r2 \

               --ld-window 2000 \

               --ld-window-kb 500000 \

               --ld-window-r2 0.8 \

               --ld-snp ${p} \

               --out temp/LD${n}

     done;

     tail -n +2 temp/LD*.ld | grep rs > temp/LD_sumary.tab;

    awk   < temp/LD_sumary.tab '{ print   "chr" $4 "\t"$5
"\t"$5 +1}' > temp/SNP.bed;

    rm temp/LD${n}*;

    /data/nephron/rafik/tools/bedtools/bin/bedtools
intersect   -wb   -wa   -a   <HIF_binding_sites.bed>   -b
temp/SNP.bed|wc -l >> temp/LDHIFcout.txt;

     rm temp/LD*.ld

done < temp/pruned_windowex_rs.txt


#Add count data to last column of summary

paste --delimiter "\t"

temp/pruned_windowex_rs_summary.tab  temp/LDHIFcout.txt  >
temp/rs_num_TGfiltered_summary_HIF1bcount.tab


# rs numbers of all

cat  temp/rs_num_TGfiltered_summary_HIF1bcount.tab    |awk
'{if ($5 < 5*10^-8) print $4}' > AT_rs.txt
```

10.4.2 Genome-wide shuffling expected frequency distribution

```
import pandas as pd;

import numpy as np;

import sys as os;

import os;

name="AT"


GWAS     =     pd.read_table("temp/"+name+"/LD_counts.bed",
header=None)

chromosome_sizes                                                =
pd.read_table("/data/nephron/rafik/ref_beds/hg19.chrom.siz
es", header=None)

u = list(set(GWAS.loc[:,3]))


count = 0

while (count < 100000):

 Shuffled_GWAS = GWAS.copy()

 for SNP_grp in u:

  filters = GWAS.loc[:,3]==SNP_grp ##index one SNP group

  dist = GWAS[filters].iloc[:,1]-GWAS[filters].iloc[0,1]

#vector to be shuffled 0 to max

  dist_max = dist.max()

### what needs be subtracted from the max position #let us
get a random position in the same chromosome and make sure
it does not exceed the chromosome

  extent                                                        =
np.array(chromosome_sizes.loc[chromosome_sizes.loc[:,0]==G
WAS[filters].iloc[0,0],1])[0]

#length of chromosome

  choice = np.random.random_integers(0,extent)

#random starting  position

  while ((choice+dist_max)>extent):
```

```
#if the chosen position falls out of "range" choose new on

   choice = np.random.random_integers(0,extent)

  factor = choice - GWAS[filters].iloc[0,1]

  Shuffled_GWAS.loc[filters,1:2]                          =
GWAS.loc[filters,1:2]+factor


Shuffled_GWAS.to_csv("temp/"+name+"/foo.bed",sep="\t",head
er=None, index=None)


os.system("/data/nephron/rafik/tools/bedtools/bin/bedtools
intersect -a temp/"+name+"/foo.bed -b <HIF binding sites
.bed file> | awk '{ print $4 }' |uniq | wc -l >>
temp/"+name+"/counts2reps.txt ")

 count          =          count          +          1
```

10.4.3 Enhancer-wide shuffling expected frequency distribution

```
name="edu_AT"

bedfiles="/data/nephron/vischmid/bed_files"

mkdir temp/${name}


/data/nephron/rafik/tools/bedtools/bin/bedtools intersect -
wa -a ${bedfiles}/HIF_peaks.bed -b temp/${name}/LD_08.bed |
awk    '{    print    $2    }'|sort|uniq|wc    -l        >
temp/${name}/data_HIF.txt


for ((x=1;x<=100000;x++))

do

     bedtools  shuffle  -chromFirst  -noOverlapping  -incl
${bedfiles}/all_anno_sorted_1p.bed                       -i
${bedfiles}/HIF_peaks.bed                                -g
/data/nephron/rafik/ref_beds/hg19.chrom.sizes             >
temp/${name}/shuffled_HIF.bed

     bedtools          intersect          -wa          -a
temp/${name}/shuffled_HIF.bed  -b  temp/${name}/LD_08.bed  |
```

```
awk '{ print $2 }'|sort|uniq|wc -l >> temp/${name}/${name}_enhancerwide_count.txt

done
```

10.4.4 eQTL

```
getSNP <- function(snp, gname)

{

  con <- dbConnect(MySQL(), user="root", password="root",dbname="TCGA", host="ndmws3.ndm.ox.ac.uk")

  statistic =p.v = hr = score =  vector()

  mat = as.data.frame(fetch(dbSendQuery(con,paste("select distinct PID, Expression, Haplotype, confidence, Normal, Allele from GermPatientSNP inner join  MetaSNP on SNPID = SNP_Probe inner join Patient_Gene_Expression on barcode = PID and type = '01A' where Gene_Name= '",gname,"' and SNPID = '",snp,"'", sep="")), n = 10000))

  for(con in  dbListConnections(MySQL()))dbDisconnect(con)

  mat[mat[,3]==0,3] = paste(mat[mat[,3]==0,5],mat[mat[,3]==0,5], sep="")

  mat[mat[,3]==1,3] = paste(mat[mat[,3]==1,5],mat[mat[,3]==1,6], sep="")

  mat[mat[,3]==2,3] = paste(mat[mat[,3]==2,6],mat[mat[,3]==2,6], sep="")

  mat[,1:4]

}


testSignificanceofSNP <- function(y,group)

{

  require(MASS)

  tryCatch({

    fit1 <- glm.nb( round(y) ~ 1+as.factor(group))

    fit0 <- glm.nb( round(y) ~ 1)
```

```
    anova( fit0, fit1, test="Chisq")[2,8]
    #min(summary(fit1)$coefficients[,4][3])
  },warning = function(w){
    1
  })


#Plot
set.seed(1)
l <- sample(letters, 500, replace=TRUE)
n <- runif(500)
l <- factor(l, levels = rev(letters))
boxplot(n ~ l)

x = getSNP('<A6 number of SNP>','<gene name>')
#Filter low significance thresholds
x = x[x[,4]<0.01,]
vital = sign(((x[,6]+1)/2)-1)*-1
time = x[,5]
status = x[,3]
```

10.4.5 Gene expression tissue from tumour and normal kidney

```
getTumourNormalMatchedExpression = function(gname)
{
  con       <-       dbConnect(MySQL(),       user="root",
password="root",dbname="TCGA",
host="dhcp276.well.ox.ac.uk")
  fullmat = matrix(ncol=length(gname), nrow=72)
  for(i in 1:length(gname))
```

```
  {
    q = paste("select barcode,type,Expression from
Patient_Gene_Expression inner join Patient on
bcr_patient_barcode = barcode and Gene_Name= '",gname[i],"'
and barcode in (select barcode from Patient_Gene_ID_Type
where type='11A') order by barcode", sep="")
    v = as.data.frame(fetch(dbSendQuery(con,q), n = 10000))
    normal = v[v$type=="11A",]
    tumour = v[v$type!="11A",]
    mat = cbind(normal[normal[,1]%in%tumour[,1],],
tumour[tumour[,1]%in%normal[,1],])
    #fullmat[,i] = log2(mat[,6]+1)/log2(mat[,3]+1)
  }
  for(con in  dbListConnections(MySQL()))dbDisconnect(con)
  plot(rbind(c(1,mat[1,3]),c(2,mat[1,6])), type="n",
ylim=c(0,max(mat[,c(3,6)])+100), xaxt="n",xlab="",
ylab="Normalised Expression", main=paste(gname,"(n = 72), P-
Value = ",wilcox.test(log1p(mat[,3]),
log1p(mat[,6]),paired=T,exact=F)$p.value))
  for(i in 1:length(mat[,1]))
  {
    lines(rbind(c(1,mat[i,3]),c(2,mat[i,6])), type="b")
  }
  axis(side = 1,at = 1:2,labels = c("Normal", "Tumour"))
  mat
}
```

Chapter 11: Bibliography

Abderrahman, Balkees, and V. Craig Jordan. 2017. *The Breast, Comprehensive Management of Benign and Malignant Diseases* (Elsevier Inc.).

Anders, S., P. T. Pyl, and W. Huber. 2015. 'HTSeq-a Python framework to work with high-throughput sequencing data', *Bioinformatics*, 31: 166-9.

Bailey, T. L., M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble. 2009. 'MEME SUITE: tools for motif discovery and searching', *Nucleic Acids Res*, 37: W202-8.

Bailey, TL, and P Machanick. 2012. 'Inferring direct DNA binding from ChIP-seq', *Nucleic Acids Res.*, 40: e128.

Beckerman, R., and C. Prives. 2010. 'Transcriptional Regulation by p53', *Cold Spring Harb Perspect Biol.*, 2.

Broderick, Peter, Luis Carvajal-Carmona, Alan M Pittman, Emily Webb, Kimberley Howarth, Andrew Rowan, Steven Lubbe, et al. 2007. 'A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk', *Nature Genetics*, 39: 1315–17.

Brown, Carolyn J., Andrea Ballabio, James L. Rupert, Ronald G. Lafreniere, Markus Grompe, Rossana Tonlorenzi, and Huntington F. Willard. 1991. 'A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome', *Naturevolume*, 349: 38–44.

Brugarolas, J. B., F. Vazquez, A. Reddy, W. R. Sellers, and W. G. Kaelin. 2003. 'TSC2 regulates VEGF through mTOR-dependent and -independent pathways', *Cancer Cell*, 4: 147-58.

Bruick, R. K., and S. L. McKnight. 2001. 'A conserved family of prolyl-4-hydroxylases that modify HIF', *Science*, 294: 1337-40.

Buniello, A., J. A. L. MacArthur, M. Cerezo, L. W. Harris, J. Hayhurst, C. Malangone, A. McMahon, et al. 2019. 'The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019', *Nucleic Acids Res*, 47: D1005-D12.

Burd, C. J., C. E. Petre, L. M. Morey, Y. Wang, M. P. Revelo, C. A. Haiman, S. Lu, et al. 2006. 'Cyclin D1b variant influences prostate cancer growth through aberrant androgen receptor regulation', *Proc Natl Acad Sci U S A*, 103: 2190-5.

Cancer_Genome_Atlas_Research_Network. 2013. 'Comprehensive molecular characterization of clear cell renal cell carcinoma', *Nature*, 499: 43-9.

Cairns, W. K., G. Yao, Y. Z. Gu, and C. A. Bradfield. 2016. ' CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data', *Gen. Biology*, 17:127.

Chan, W. K., G. Yao, Y. Z. Gu, and C. A. Bradfield. 1999. 'Cross-talk between the aryl hydrocarbon receptor and hypoxia inducible factor signaling pathways. Demonstration of competition and compensation', *J Biol Chem*, 274: 12115-23.

Cockman, M. E., N. Masson, D. R. Mole, P. Jaakkola, G. W. Chang, S. C. Clifford, E. R. Maher, C. W. Pugh, P. J. Ratcliffe, and P. H. Maxwell. 2000. 'Hypoxia inducible factor-alpha binding and ubiquitylation by the von Hippel-Lindau tumor suppressor protein', *J Biol Chem*, 275: 25733-41.

Dang, CV. 1999. 'c-Myc target genes involved in cell growth, apoptosis, and metabolism.', *Mol Cell Biol*, 19: 1-11.
Dang, CV, KA O'Donnell, KI Zeller, T Nguyen, RC Osthus, and F Li. 2006 'The c-Myc target gene network', *Semin Cancer Biol.*, 16: 253-64.
Davies, J. O., J. M. Telenius, S. J. McGowan, N. A. Roberts, S. Taylor, D. R. Higgs, and J. R. Hughes. 2016. 'Multiplexed analysis of chromosome conformation at vastly improved sensitivity', *Nat Methods*, 13: 74-80.
Dewhirst, M. W., Y. Cao, C. Y. Li, and B. Moeller. 2007. 'Exploring the role of HIF-1 in early angiogenesis and response to radiotherapy', *Radiother Oncol*, 83: 249-55.
Eischen, CM, JD Weber, MF Roussel, CJ Sherr, and JL Cleveland. 1999. 'Disruption of the ARF-Mdm2-p53 tumor suppressor pathway in Myc-induced lymphomagenesis', 13: 2658-69.
Ema, M., S. Taya, N. Yokotani, K. Sogawa, Y. Matsuda, and Y. Fujii-Kuriyama. 1997. 'A novel bHLH-PAS factor with close sequence similarity to hypoxia-inducible factor 1alpha regulates the VEGF expression and is potentially involved in lung and vascular development', *Proc Natl Acad Sci U S A*, 94: 4273-8.
Encode_Project_Consortium, I. Dunham, A. Kundaje, S. F. Aldred, P. J. Collins, C. A. Davis, F. Doyle, et al. 2012. 'An integrated encyclopedia of DNA elements in the human genome', *Nature*, 489: 57-74.
Epstein, A. C., J. M. Gleadle, L. A. McNeill, K. S. Hewitson, J. O'Rourke, D. R. Mole, M. Mukherji, et al. 2001. 'C. elegans EGL-9 and mammalian homologs define a family of dioxygenases that regulate HIF by prolyl hydroxylation', *Cell*, 107: 43-54.
Ernst, Jason, Pouya Kheradpour, Tarjei S. Mikkelsen, Noam Shoresh, Lucas D. Ward, Charles B. Epstein, Xiaolan Zhang, et al. 2011. 'Mapping and analysis of chromatin state dynamics in nine human cell types', *Nature*, 473: 43–49.
Firth, J. D., B. L. Ebert, and P. J. Ratcliffe. 1995. 'Hypoxic regulation of lactate dehydrogenase A. Interaction between hypoxia-inducible factor 1 and cAMP response elements', *J Biol Chem*, 270: 21021-7.
Fischer, M. 2017. 'Census and evaluation of p53 target genes', *Oncogene*, 36: 3943-56.
Gabriel, Stacey B., Stephen F. Schaffner, Huy Nguyen, Jamie M. Moore, Jessica Roy, Brendan Blumenstiel, John Higgins, et al. 2002. 'The Structure of Haplotype Blocks in the Human Genome', *Science*, 296: 2225-29.
Gerlinger, M., S. Horswell, J. Larkin, A. J. Rowan, M. P. Salm, I. Varela, R. Fisher, et al. 2014. 'Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing', *Nat Genet*, 46: 225-33.
Gerlinger, M., A. J. Rowan, S. Horswell, J. Larkin, D. Endesfelder, E. Gronroos, P. Martinez, et al. 2012. 'Intratumor heterogeneity and branched evolution revealed by multiregion sequencing', *N Engl J Med*, 366: 883-92.
Giaccia, A. J., M. C. Simon, and R. Johnson. 2004. 'The biology of hypoxia: the role of oxygen sensing in development, normal function, and disease', *Genes Dev*, 18: 2183-94.
Grampp, S., J. L. Platt, V. Lauer, R. Salama, F. Kranz, V. K. Neumann, S. Wach, et al. 2016. 'Genetic variation at the 8q24.21 renal cancer susceptibility locus affects HIF binding to a MYC enhancer', *Nat Commun*, 7: 13183.

Grampp, S., V. Schmid, R. Salama, V. Lauer, F. Kranz, J. L. Platt, J. Smythies, et al. 2017. 'Multiple renal cancer susceptibility polymorphisms modulate the HIF pathway', *PLoS Genet*, 13: e1006872.

Greenman, Christopher, Philip Stephens, Raffaella Smith, Gillian L. Dalgliesh, Christopher Hunter, and Michael R. Stratton. 2007. 'Patterns of somatic mutation in human cancer genomes', *Nature*, 446: 153–58.

Groner, Anna C., and Myles Brown. 2017. 'Role of steroid receptor and coregulator mutations in hormone-dependent cancers', *J Clin Invest.*

Gu, Y. Z., S. M. Moran, J. B. Hogenesch, L. Wartman, and C. A. Bradfield. 1998. 'Molecular characterization and chromosomal localization of a third alpha-class hypoxia inducible factor subunit, HIF3alpha', *Gene Expr*, 7: 205-13.

Gudbjartsson, T., T. J. Jónasdóttir, A. Thoroddsen, G. V. Einarsson, G. M. Jónsdóttir, K. Kristjánsson, S. Hardarson, et al. 2002. 'A population-based familial aggregation analysis indicates genetic contribution in a majority of renal cell carcinomas', *IJC*, 4: 476-79.

Gudmundsson, J., P. Sulem, D. F. Gudbjartsson, G. Masson, V. Petursdottir, S. Hardarson, S. A. Gudjonsson, et al. 2013. 'A common variant at 8q24.21 is associated with renal cell cancer', *Nat Commun*, 4: 2776.

Hanahan, D. 2014. 'Rethinking the war on cancer', *Lancet*, 383: 558-63.

Hanahan, D., and R. A. Weinberg. 2011. 'Hallmarks of cancer: the next generation', *Cell*, 144: 646-74.

Henrion, M., M. Frampton, G. Scelo, M. Purdue, Y. Ye, P. Broderick, A. Ritchie, et al. 2013. 'Common variation at 2q22.3 (ZEB2) influences the risk of renal cancer', *Hum Mol Genet*, 22: 825-31.

Henrion, M. Y., M. P. Purdue, G. Scelo, P. Broderick, M. Frampton, A. Ritchie, A. Meade, et al. 2015. 'Common variation at 1q24.1 (ALDH9A1) is a potential risk factor for renal cancer', *PLoS One*, 10: e0122589.

Hewitson, K. S., L. A. McNeill, M. V. Riordan, Y. M. Tian, A. N. Bullock, R. W. Welford, J. M. Elkins, et al. 2002. 'Hypoxia-inducible factor (HIF) asparagine hydroxylase is identical to factor inhibiting HIF (FIH) and is related to the cupin structural family', *J Biol Chem*, 277: 26351-5.

Hockel, M., and P. Vaupel. 2001. 'Tumor hypoxia: definitions and current clinical, biologic, and molecular aspects', *J Natl Cancer Inst*, 93: 266-76.

Holst, F., P. R. Stahl, C. Ruiz, O. Hellwinkel, Z. Jehan, M. Wendland, A. Lebeau, et al. 2007. 'Estrogen receptor alpha (ESR1) gene amplification is frequent in breast cancer', *Nat Genet*, 39: 655-60.

Hower, V., S. N. Evans, and L. Pachter. 2011. 'Shape-based peak identification for ChIP-Seq', *BMC Bioinformatics*, 12: 15.

Hughes, J. R., K. M. Lower, I. Dunham, S. Taylor, M. De Gobbi, J. A. Sloane-Stanley, S. McGowan, et al. 2013. 'High-resolution analysis of cis-acting regulatory networks at the alpha-globin locus', *Philos Trans R Soc Lond B Biol Sci*, 368: 20120361.

Hughes, J. R., N. Roberts, S. McGowan, D. Hay, E. Giannoulatou, M. Lynch, M. De Gobbi, S. Taylor, R. Gibbons, and D. R. Higgs. 2014. 'Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment', *Nat Genet*, 46: 205-12.

Ikegawa, Shiro. 2012. 'A Short History of the Genome-Wide Association Study: Where We Were and Where We Are Going', *Genomics Inform.*, 10: 220–25.

Jiang, B. H., E. Rue, G. L. Wang, R. Roe, and G. L. Semenza. 1996. 'Dimerization, DNA binding, and transactivation properties of hypoxia-inducible factor 1', *J Biol Chem*, 271: 17771-8.

Kaelin, W. G., Jr. 2007. 'The von Hippel-Lindau tumor suppressor protein and clear cell renal carcinoma', *Clin Cancer Res*, 13: 680s-84s.

Kaelin, W. G., Jr., and P. J. Ratcliffe. 2008. 'Oxygen sensing by metazoans: the central role of the HIF hydroxylase pathway', *Mol Cell*, 30: 393-402.

Kobayashi, Akira, Kazuhiro Sogawa, and Yoshiaki Fujii-Kuriyama. 1996. 'Cooperative Interaction between AhR·Arnt and Sp1 for the Drug-inducible Expression of CYP1A1 Gene', *The Journal of Biological Chemistry*, 271: 12310-16.

Kondo, K., W. Y. Kim, M. Lechpammer, and W. G. Kaelin, Jr. 2003. 'Inhibition of HIF2alpha is sufficient to suppress pVHL-defective tumor growth', *PLoS Biol*, 1: E83.

Kondo, K., J. Klco, E. Nakamura, M. Lechpammer, and W. G. Kaelin, Jr. 2002. 'Inhibition of HIF is necessary for tumor suppression by the von Hippel-Lindau protein', *Cancer Cell*, 1: 237-46.

Lando, D., I. Pongratz, L. Poellinger, and M. L. Whitelaw. 2000. 'A redox mechanism controls differential DNA binding activities of hypoxia-inducible factor (HIF) 1alpha and the HIF-like factor', *J Biol Chem*, 275: 4618-27.

Latif, F., K. Tory, J. Gnarra, M. Yao, F. M. Duh, M. L. Orcutt, T. Stackhouse, et al. 1993. 'Identification of the von Hippel-Lindau disease tumor suppressor gene', *Science*, 260: 1317-20.

Lau, K. W., Y. M. Tian, R. R. Raval, P. J. Ratcliffe, and C. W. Pugh. 2007. 'Target gene selectivity of hypoxia-inducible factor-alpha in renal cancer cells is conveyed by post-DNA-binding mechanisms', *Br J Cancer*, 96: 1284-92.

Lawrence, M. S., P. Stojanov, C. H. Mermel, J. T. Robinson, L. A. Garraway, T. R. Golub, M. Meyerson, S. B. Gabriel, E. S. Lander, and G. Getz. 2014. 'Discovery and saturation analysis of cancer genes across 21 tumour types', *Nature*, 505: 495-501.

Li, Zirong, Sara Van Calcar, Chunxu Qu, Webster K. Cavenee, Michael Q. Zhang, and Bing Ren. 2003. 'A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells', *PNAS*, 100: 8164-69.

Linehan, W. M., P. A. Pinto, G. Bratslavsky, E. Pfaffenroth, M. Merino, C. D. Vocke, J. R. Toro, et al. 2009. 'Hereditary kidney cancer: unique opportunity for disease-based therapy', *Cancer*, 115: 2252-61.

Lonser, R. R., G. M. Glenn, M. Walther, E. Y. Chew, S. K. Libutti, W. M. Linehan, and E. H. Oldfield. 2003. 'von Hippel-Lindau disease', *Lancet*, 361: 2059-67.

Love, M. I., W. Huber, and S. Anders. 2014. 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biol*, 15: 550.

Mack, FA, JH Patel, MP Biju, VH Haase, and MC Simon. 2005. 'Decreased growth of Vhl-/- fibrosarcomas is associated with elevated levels of cyclin kinase inhibitors p21 and p27', *Mol Cell Biol*, 25: 4565-78.

Mahon, P. C., K. Hirota, and G. L. Semenza. 2001. 'FIH-1: a novel protein that interacts with HIF-1alpha and VHL to mediate repression of HIF-1 transcriptional activity', *Genes Dev*, 15: 2675-86.

Mandriota, S. J., K. J. Turner, D. R. Davies, P. G. Murray, N. V. Morgan, H. M. Sowter, C. C. Wykoff, et al. 2002. 'HIF activation identifies early lesions in

VHL kidneys: evidence for site-specific tumor suppressor function in the nephron', *Cancer Cell*, 1: 459-68.
Maxwell, P. H., C. W. Pugh, and P. J. Ratcliffe. 1993. 'Inducible operation of the erythropoietin 3' enhancer in multiple cell lines: evidence for a widespread oxygen-sensing mechanism', *Proc Natl Acad Sci U S A*, 90: 2423-7.
Maxwell, P. H., M. S. Wiesener, G. W. Chang, S. C. Clifford, E. C. Vaux, M. E. Cockman, C. C. Wykoff, C. W. Pugh, E. R. Maher, and P. J. Ratcliffe. 1999. 'The tumour suppressor protein VHL targets hypoxia-inducible factors for oxygen-dependent proteolysis', *Nature*, 399: 271-5.
McKay, J. D., R. J. Hung, Y. Han, X. Zong, R. Carreras-Torres, D. C. Christiani, N. E. Caporaso, et al. 2017. 'Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes', *Nat Genet*, 49: 1126-32.
Michailidou, K., S. Lindstrom, J. Dennis, J. Beesley, S. Hui, S. Kar, A. Lemacon, et al. 2017. 'Association analysis identifies 65 new breast cancer risk loci', *Nature*, 551: 92-94.
Mitchell, T. J., S. Turajlic, A. Rowan, D. Nicol, J. H. R. Farmery, T. O'Brien, I. Martincorena, et al. 2018. 'Timing the Landmark Events in the Evolution of Clear Cell Renal Cell Cancer: TRACERx Renal', *Cell*, 173: 611-23 e17.
Mole, D. R., C. Blancher, R. R. Copley, P. J. Pollard, J. M. Gleadle, J. Ragoussis, and P. J. Ratcliffe. 2009. 'Genome-wide association of hypoxia-inducible factor (HIF)-1alpha and HIF-2alpha DNA binding with expression profiling of hypoxia-inducible transcripts', *J Biol Chem*, 284: 16767-75.
Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. 2008. 'Mapping and quantifying mammalian transcriptomes by RNA-Seq', *Nat Methods*, 5: 621-8.
Nesbit, CE, JM Tersak, and EV Prochownik. 1999. 'MYC oncogenes and human neoplastic disease', *Oncogene.*, 18: 3004-16.
Ohh, M., C. W. Park, M. Ivan, M. A. Hoffman, T. Y. Kim, L. E. Huang, N. Pavletich, V. Chau, and W. G. Kaelin. 2000. 'Ubiquitination of hypoxia-inducible factor requires direct binding to the beta-domain of the von Hippel-Lindau protein', *Nat Cell Biol*, 2: 423-7.
Okbay, A., J. P. Beauchamp, M. A. Fontana, J. J. Lee, T. H. Pers, C. A. Rietveld, P. Turley, et al. 2016. 'Genome-wide association study identifies 74 loci associated with educational attainment', *Nature*, 533: 539-42.
Orlando, V, and R Paro. 1993. 'Mapping Polycomb-repressed domains in the bithorax complex using in vivo formaldehyde ', 75: 1187-98.
Ortiz-Barahona, A., D. Villar, N. Pescador, J. Amigo, and L. del Peso. 2010. 'Genome-wide identification of hypoxia-inducible factor binding sites and target genes by a probabilistic model integrating transcription-profiling data and in silico binding site prediction', *Nucleic Acids Res*, 38: 2332-45.
Park, S, S Chung, KM Kim, KC Jung, C Park, ER Hahm, and CH Yang. 2004. 'Determination of binding constant of transcription factor myc-max/max-max and E-box DNA: the effect of inhibitors on the binding', *Biochim Biophys Acta*, 1670: 217-28.
Park, S. K., A. M. Dadak, V. H. Haase, L. Fontana, A. J. Giaccia, and R. S. Johnson. 2003. 'Hypoxia-induced gene expression occurs solely through the action of hypoxia-inducible factor 1alpha (HIF-1alpha): role of cytoplasmic trapping of HIF-2alpha', *Mol Cell Biol*, 23: 4959-71.

Pescador, N., D. Villar, D. Cifuentes, M. Garcia-Rocha, A. Ortiz-Barahona, S. Vazquez, A. Ordonez, et al. 2010. 'Hypoxia promotes glycogen accumulation through hypoxia inducible factor (HIF)-mediated induction of glycogen synthase 1', *PLoS One*, 5: e9644.

Platt, J. L., R. Salama, J. Smythies, H. Choudhry, J. O. Davies, J. R. Hughes, P. J. Ratcliffe, and D. R. Mole. 2016. 'Capture-C reveals preformed chromatin interactions between HIF-binding sites and distant promoters', *EMBO Rep*, 17: 1410-21.

Purdue, M. P., M. Johansson, D. Zelenika, J. R. Toro, G. Scelo, L. E. Moore, E. Prokhortchouk, et al. 2011. 'Genome-wide association study of renal cell carcinoma identifies two susceptibility loci on 2p21 and 11q13.3', *Nat Genet*, 43: 60-5.

Rajalingam, K, R Schreck, UR Rapp, and S Albert. 2007. 'Ras oncogenes and their downstream targets', *Biochim Biophys Acta.*, 1773: 1177-95.

Raval, R. R., K. W. Lau, M. G. Tran, H. M. Sowter, S. J. Mandriota, J. L. Li, C. W. Pugh, P. H. Maxwell, A. L. Harris, and P. J. Ratcliffe. 2005. 'Contrasting Properties of Hypoxia-Inducible Factor 1 (HIF-1) and HIF-2 in von Hippel-Lindau-Associated Renal Cell Carcinoma', *Mol Cell Biol*, 25: 5675-86.

Rhee, Ho Sung, and B. Franklin Pugh. 2011. 'Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution', *Cell*, 147: 1199-422.

Robertson, G, M Hirst, M Bainbridge, M Bilenky, Y Zhao, T Zeng, G Euskirchen, et al. 2007. 'Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing'.

Rosenberger, C., S. Mandriota, J. S. Jurgensen, M. S. Wiesener, J. H. Horstrup, U. Frei, P. J. Ratcliffe, P. H. Maxwell, S. Bachmann, and K. U. Eckardt. 2002. 'Expression of hypoxia-inducible factor-1alpha and -2alpha in hypoxic and ischemic rat kidneys', *J Am Soc Nephrol*, 13: 1721-32.

Ross-Innes, C. S., R. Stark, A. E. Teschendorff, K. A. Holmes, H. R. Ali, M. J. Dunning, G. D. Brown, et al. 2012. 'Differential oestrogen receptor binding is associated with clinical outcome in breast cancer', *Nature*, 481: 389-93.

Salama, R., N. Masson, P. Simpson, L. K. Sciesielski, M. Sun, Y. M. Tian, P. J. Ratcliffe, and D. R. Mole. 2015. 'Heterogeneous Effects of Direct Hypoxia Pathway Activation in Kidney Cancer', *PLoS One*, 10: e0134645.

Sandelin, A, and WW. Wasserman. 2004. 'Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics.', *J Mol Biol.*, 338: 207-15.

Scelo, G., M. P. Purdue, K. M. Brown, M. Johansson, Z. Wang, J. E. Eckel-Passow, Y. Ye, et al. 2017. 'Genome-wide association study identifies multiple risk loci for renal cell carcinoma', *Nat Commun*, 8: 15724.

Schietke, R. E., T. Hackenbeck, M. Tran, R. Gunther, B. Klanke, C. L. Warnecke, K. X. Knaup, et al. 2012. 'Renal Tubular HIF-2alpha Expression Requires VHL Inactivation and Causes Fibrosis and Cysts', *PLoS One*, 7: e31034.

Schodel, J., D. R. Mole, and P. J. Ratcliffe. 2013. 'Pan-genomic binding of hypoxia-inducible transcription factors', *Biol Chem*, 394: 507-17.

Schödel, Johannes, Chiara Bardella, Lina K Sciesielski, Jill M Brown, Chris W Pugh, Veronica Buckle, Ian P Tomlinson, Peter J Ratcliffe, and David R Mole. 2012. 'Common genetic variants at the 11q13.3 renal cancer susceptibility locus influence binding of HIF to an enhancer of cyclin D1 expression', *Nature Genetics*, 44: 420–25.

Schoedel, J., S. Oikonomopoulos, J. Ragoussis, C. W. Pugh, P. J. Ratcliffe, and D. R. Mole. 2011. 'High-resolution genome-wide mapping of HIF-binding sites by ChIP-seq', *Blood*, 117: e207-17.

Seizinger, B. R., G. A. Rouleau, L. J. Ozelius, A. H. Lane, G. E. Farmer, J. M. Lamiell, J. Haines, et al. 1988. 'Von Hippel-Lindau disease maps to the region of chromosome 3 associated with renal cell carcinoma', *Nature*, 332: 268-9.

Selak, M. A., S. M. Armour, E. D. MacKenzie, H. Boulahbel, D. G. Watson, K. D. Mansfield, Y. Pan, M. C. Simon, C. B. Thompson, and E. Gottlieb. 2005. 'Succinate links TCA cycle dysfunction to oncogenesis by inhibiting HIF-alpha prolyl hydroxylase', *Cancer Cell*, 7: 77-85.

Semenza, G. L. 2010. 'Defining the role of hypoxia-inducible factor 1 in cancer biology and therapeutics', *Oncogene*, 29: 625-34.

———. 2012. 'Hypoxia-inducible factors: mediators of cancer progression and targets for cancer therapy', *Trends Pharmacol Sci*, 33: 207-14.

Semenza, G. L., P. H. Roth, H. M. Fang, and G. L. Wang. 1994. 'Transcriptional regulation of genes encoding glycolytic enzymes by hypoxia-inducible factor 1', *J Biol Chem*, 269: 23757-63.

Servant, N., N. Varoquaux, B. R. Lajoie, E. Viara, C. J. Chen, J. P. Vert, E. Heard, J. Dekker and E. Barillot. 2015. ' HiC-Pro: an optimized and flexible pipeline for Hi-C data processing', *Gen. Biol.*, 16: 259.

Shen, C., R. Beroukhim, S. E. Schumacher, J. Zhou, M. Chang, S. Signoretti, and W. G. Kaelin, Jr. 2011. 'Genetic and Functional Studies Implicate HIF1alpha as a 14q Kidney Cancer Suppressor Gene', *Cancer Discov*, 1: 222-35.

Simanshu, DK, DV Nissley, and F McCormick. 2017. 'RAS Proteins and Their Regulators in Human Disease', *Cell*, 170: 17-33.

Smythies, J. A., M. Sun, N. Masson, R. Salama, P. D. Simpson, E. Murray, V. Neumann, et al. 2019. 'Inherent DNA-binding specificities of the HIF-1alpha and HIF-2alpha transcription factors in chromatin', *EMBO Rep*, 20.

Sotelo, Jose, Dominic Esposito, Maria Ana Duhagon, Kelley Banfield, Jennifer Mehalko, Hongling Liao, Robert M. Stephens, Timothy J. R. Harris, David J. Munroe, and Xiaolin Wua. 2010. 'Long-range enhancers on 8q24 regulate c-Myc', *Proc Natl Acad Sci U S A*, 107: 3001–05.

Stark, Rory, and Gord Brown. 2018. 'DiffBind: Differential binding analysis of ChIP- Seq peak data', *Bioconductor*.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, et al. 2005. 'Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles', *Proc Natl Acad Sci U S A*, 102: 15545-50.

Szalkowski, AM, and CD Schmid. 2011. 'Rapid innovation in ChIP-seq peak-calling algorithms is outdistancing benchmarking efforts', *Brief Bioinform*, 12: 626-33.

Talks, K. L., H. Turley, K. C. Gatter, P. H. Maxwell, C. W. Pugh, P. J. Ratcliffe, and A. L. Harris. 2000. 'The expression and distribution of the hypoxia-inducible factors HIF-1alpha and HIF-2alpha in normal human tissues, cancers, and tumor-associated macrophages', *Am J Pathol*, 157: 411-21.

Tausendschon, M., M. Rehli, N. Dehne, C. Schmidl, C. Doring, M. L. Hansmann, and B. Brune. 2015. 'Genome-wide identification of hypoxia-inducible

factor-1 and -2 binding sites in hypoxic human macrophages alternatively activated by IL-10', *Biochim Biophys Acta*, 1849: 10-22.
TCGA. 'The Cancer Genome Atlas'. http://cancergenome.nih.gov.
The_1000_Genomes_Project_Consortium. 2015. 'A global reference for human genetic variation', *Nature*, 526: 68-74.
Tian, H., S. L. McKnight, and D. W. Russell. 1997. 'Endothelial PAS domain protein 1 (EPAS1), a transcription factor selectively expressed in endothelial cells', *Genes Dev*, 11: 72-82.
Trapnell, C., L. Pachter, and S. L. Salzberg. 2009. 'TopHat: discovering splice junctions with RNA-Seq', *Bioinformatics*, 25: 1105-11.
Turajlic, S., H. Xu, K. Litchfield, A. Rowan, T. Chambers, J. I. Lopez, D. Nicol, et al. 2018. 'Tracking Cancer Evolution Reveals Constrained Routes to Metastases: TRACERx Renal', *Cell*, 173: 581-94 e12.
Turajlic, S., H. Xu, K. Litchfield, A. Rowan, S. Horswell, T. Chambers, T. O'Brien, et al. 2018. 'Deterministic Evolutionary Trajectories Influence Primary Tumor Growth: TRACERx Renal', *Cell*, 173: 595-610 e11.
Uchida, T., F. Rossignol, M. A. Matthay, R. Mounier, S. Couette, E. Clottes, and C. Clerici. 2004. 'Prolonged hypoxia differentially regulates hypoxia-inducible factor (HIF)-1alpha and HIF-2alpha expression in lung epithelial cells: implication of natural antisense HIF-1alpha', *J Biol Chem*, 279: 14871-8.
Vernimmen, Douglas, Magnus D. Lynch, Marco De Gobbi, David Garrick, Jacqueline A. Sharpe, Jacqueline A. Sloane-Stanley, Andrew J.H. Smith, and Douglas R. Higgs. 2011. 'Polycomb eviction as a new distant enhancer function', *Genes & Dev.*, 25: 1583-88.
Visel, A., M. J. Blow, Z. Li, T. Zhang, J. A. Akiyama, A. Holt, I. Plajzer-Frick, et al. 2009. 'ChIP-seq accurately predicts tissue-specific activity of enhancers', *Nature*, 457: 854-8.
Wang, G. L., B. H. Jiang, E. A. Rue, and G. L. Semenza. 1995. 'Hypoxia-inducible factor 1 is a basic-helix-loop-helix-PAS heterodimer regulated by cellular O2 tension', *Proc Natl Acad Sci U S A*, 92: 5510-4.
Wang, G. L., and G. L. Semenza. 1995. 'Purification and characterization of hypoxia-inducible factor 1', *J Biol Chem*, 270: 1230-7.
Wang, K , SP Dickson, CA Stolle, ID Krantz, DB Goldstein, and Hakonarson H. 2010. 'Interpretation of association signals and identification of causal variants from genome-wide association studies', *Am J Hum Genet.*, 86: 730-42.
Wang, X., N. R. Tucker, G. Rizki, R. Mills, P. H. Krijger, E. de Wit, V. Subramanian, et al. 2016. 'Discovery and validation of sub-threshold genome-wide association study loci using epigenomic signatures', *Elife*, 5.
Wellcome Trust Case Control, Consortium. 2007. 'Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls', *Nature*, 447: 661-78.
Wenger, R. H., D. P. Stiehl, and G. Camenisch. 2005. 'Integration of oxygen signaling at the consensus HRE', *Sci STKE*, 2005: re12.
Wu, X., G. Scelo, M. P. Purdue, N. Rothman, M. Johansson, Y. Ye, Z. Wang, et al. 2011. 'A genome-wide association study identifies a novel susceptibility locus for renal cell carcinoma on 12p11.23', *Hum Mol Genet*.

Wykoff, C. C., C. W. Pugh, P. H. Maxwell, A. L. Harris, and P. J. Ratcliffe. 2000. 'Identification of novel hypoxia dependent and independent target genes of the von Hippel-Lindau (VHL) tumour suppressor by mRNA differential expression profiling', *Oncogene*, 19: 6297-305.

Xia, X., and A. L. Kung. 2009. 'Preferential binding of HIF-1 to transcriptionally active loci determines cell-type specific response to hypoxia', *Genome Biol*, 10: R113.

Xiao, Y., T. H. Hsiao, U. Suresh, H. I. Chen, X. Wu, S. E. Wolf, and Y. Chen. 2014. 'A novel significance score for gene selection and ranking', *Bioinformatics*, 30: 801-7.

Xie, X, J Lu, EJ Kulbokas, TR Golub, V Mootha, K Lindblad-Toh, ES Lander, and M Kellis. 2005. 'Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals', *Nature*, 434: 338-45.

Xu, Guang-hua, Ning Lou, Hang-chuan Shi, Yu-chen Xu, Hai-long Ruan, Wen Xiao, Lei Liu, et al. 2018. 'Up-regulation of SR-BI promotes progression and serves as a prognostic biomarker in clear cell renal cell carcinoma', *BMC Cancer*, 18.

Yao, X., J. Tan, K. J. Lim, J. Koh, W. F. Ooi, Z. Li, D. Huang, et al. 2017. ' VHL Deficiency Drives Enhancer Activation of Oncogenes in Clear Cell Renal Cell Carcinoma', *Cancer Discov,* 11: 1284-305.

Young, A. P., S. Schlisio, Y. A. Minamishima, Q. Zhang, L. Li, C. Grisanzio, S. Signoretti, and W. G. Kaelin, Jr. 2008. 'VHL loss actuates a HIF-independent senescence programme mediated by Rb and p400', *Nat Cell Biol*, 10: 361-9.

Zhang, Y., T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, et al. 2008. 'Model-based analysis of ChIP-Seq (MACS)', *Genome Biol*, 9: R137.

Zhong, H., A. M. De Marzo, E. Laughner, M. Lim, D. A. Hilton, D. Zagzag, P. Buechler, W. B. Isaacs, G. L. Semenza, and J. W. Simons. 1999. 'Overexpression of hypoxia-inducible factor 1alpha in common human cancers and their metastases', *Cancer Res*, 59: 5830-5.

Zindy, F, CM Eischen, DH Randle, T Kamijo, JL Cleveland, CJ Sherr, and MF Roussel. 1998. 'Myc signaling via the ARF tumor suppressor regulates p53-dependent apoptosis and immortalization', *Genes Dev.*, 12: 2424-33.