Molecular heterogeneity of invasive penile cancer

Dr Simon Nicholas Rodney

University College London (UCL)

A thesis submitted for the degree of
PhD: Cancer Genomics

PhD Supervisors: Professor John Kelly and Dr Andrew Feber
Clinical supervisor: Mr Asif Muneer

December 2019

Declaration

I, Dr Simon Nicholas Rodney, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Penile cancer is a rare and mutilating disease. Due to the paucity of basic, molecular and translational work, new treatment options have not been forthcoming and the disease has arguably been neglected, and patients have poor outcomes.

This thesis explores the molecular biology of advanced squamous cell penile carcinoma by assessing its genetic and epigenetic aberrations, and transcriptomic changes. For each patient, five tumour regions were profiled in detail and compared with a matched control sample.

When compared with other cancers, penile cancer appears to have a high tumour mutational load with high intra-tumour heterogeneity. Evidence for the clonal integration of HPV into the human genome was found. HPV positive samples are associated with APOBEC mutational changes and increased expression of *DNMT1* and *DNMT3A* methyltransferases.

TP53 was found to be an early clonal driver in the HPV negative samples, whereas mutations in *mTOR* or *PIK3CA* were found to be early clonal drivers in HPV positive samples. Potentially targetable mutations, such as *EGFR,* were only ever found to be subclonal in this small cohort. Other targetable mutations that were found to be early and shared throughout the primary tumour included *DDR2* and *cMET*.

Increased expression of immune checkpoint inhibitory proteins such as *CTLA4* were found throughout all samples, providing preliminary evidence that checkpoint blockade could be effective in penile cancer.

These findings suggest that penile cancer is a heterogeneous disease with remarkably different genetic and epigenetic profiles for HPV positive and HPV negative disease. These tumours display large amounts of intra-tumour heterogeneity and so may prove difficult to successfully treat with more traditional targeted therapies against tyrosine kinases. However, there is evidence that immune checkpoint blockade may prove to be efficacious in these patients and further work should be undertaken to examine this in more depth.

Impact statement

Penile cancer is a rare and mutilating disease. The outcomes are especially poor for patients who have lymph node metastatic disease. Due to the paucity of basic, molecular and translational work undertaken in penile cancer, new treatment options have not been forthcoming and the disease has arguably been very neglected. Currently patients with metastatic disease are treated with platinum-based compounds with very poor outcomes.

This is the first analysis of intra-tumour heterogeneity where both genetic and epigenetic factors are considered, together with changes in gene expression. The findings presented here will have far reaching impact in improving our understanding of penile cancer development and progression, and they may therefore lead to improved treatment options, which are so desperately needed.

I have demonstrated that there is significant heterogeneity between patients, which appears to be primarily driven by the presence or absence of oncogenic human papillomavirus (HPV) infection. This will inform the future management of the disease, as care should be taken when generalising a treatment in an HPV positive cohort compared with a negative cohort.

I have demonstrated that penile cancer is likely to be immunogenic with infiltrating immune cells within each tumour. In addition, there is increased expression of immune checkpoint proteins, particularly *CTLA4*. This finding is likely to have a significant impact both in academia and in the commercial pharmaceutical world, as it indicates that immunotherapy with immune checkpoint blockade should be considered as a viable treatment modality for penile cancer, after confirmatory work is undertaken.

Care should be taken when initiating targeted therapies to ensure that the chance of success is maximised. This can be achieved by focusing on targetable mutations which are present through the entire tumour. In this thesis, mutations in DDR2 and cMET were found to be shared throughout all tumour regions, whereas mutation and expression of *EGFR* appears to be subclonal in origin. This means that treatment with cetuximab (an EGFR inhibitor currently being trialled for penile cancer) may not achieve an enduring response, as only a subsection of the tumour will be targeted when treatment is initiated. This is likely to lead to tumour relapse and treatment failure.

A large number of potential further drivers have been uncovered, providing the basis for further confirmatory work and functional validation. This will help to inform basic science work in this field and may aid in identifying novel therapeutic strategies. In addition, the method used to assess whether a molecular change is clonal/subclonal may prove to be useful not only therapeutically but also in terms of biomarker development. Molecular changes that were shared were more likely to be corroborated in larger cohorts, providing a further method of biomarker selection.

Acknowledgements

I would like to thank my inspirational supervisors Professor John Kelly and Dr Andrew Feber, who have supported and mentored me for many years throughout my academic clinical fellowship and subsequent PhD. I have thoroughly enjoyed all my time spent in the lab, exploring ideas and tangents to my work as well as learning from you both.

My thanks go to the UCL Cancer Institute and Division of Surgery and all my colleagues who have worked with me throughout this journey. Thank you also to Professor Abhijit Patel for mentoring me during my semester at Yale School of Medicine, and to my colleagues at Yale. Thank you to Mr Asif Muneer for first inspiring me to explore the relatively under-researched field of male cancers, and for providing the invaluable clinical samples. Thank you also to Professor Stephan Beck for advising me and welcoming me into his medical genomics group at the UCL Cancer Institute. Finally, thank you to all my mentors, including Mr Muhamed Al-Dubaisi, Professor Jonathan Waxman and Professor Avihu Boneh.

My research would not have been possible without financial support from the following charitable foundations: Orchid, The Urology Foundation and St Peter's Trust.

I am grateful and indebted to the patients who have donated their time and consented for their tissue samples to be donated for research used in my PhD and clinical trials. Their contributions are invaluable.

Thank you to my family and friends for all the love and support they have shown me, enabling me to complete this PhD thesis. To my parents, who have always calmly encouraged and supported me in everything I do – I am eternally grateful. To my children Maya, Olivia and Rafi, for providing constant entertainment, support and distractions throughout this project. Most importantly, I would like to acknowledge the astounding patience and compassion of my wife Sheli, who is my partner in all my endeavours.

I look forward to the day when cancer research is no longer a necessity.

Table of Contents

Table of figures

Table of tables

Abbreviations

aCGH – array comparative genomic hybridisation

APOBEC – apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like

CCF – cancer cell fraction

cfDNA – cell free DNA

CNA – copy number aberration

COSMIC database – Catalogue of Somatic Mutations in Cancer database

CpG – cytosine-phosphate-guanine within DNA sequences

ctDNA – cell tumour DNA

DMP – differentially methylated position

DMR – differentially methylated region

DNA – deoxyribonucleic acid

EGFR – epidermal growth factor receptor

EMMPRIN – extracellular matrix metalloproteinase inducer

ESMO – European Society for Medical Oncology

GEO – gene expression omnibus

GO – gene ontology

HPV – Human papillomavirus

ITH – intra-tumour heterogeneity

KEGG – Kyoto Encyclopedia of Genes and Genomes

LOH – loss of heterozygosity

MMP – matrix metalloproteinase

mRNA – messenger RNA

NSCLC – non-small cell lung cancer

PenHet – cohort of patients with invasive penile squamous cell carcinoma, the subject of this thesis

PenOld – cohort of patients with invasive penile squamous cell cancer, referenced in this thesis

R – programming language and statistical software

RNA – ribonucleic acid

SCC – squamous cell carcinoma

SCCA – squamous cell carcinoma antigens

SNV – single nucleotide variant

TSS – transcription start site

UCL – University College London

UTR – untranslated region

WES – whole exome sequencing

1 Introduction

Sections of the following introduction have been previously published as first author in *The Textbook of Penile Cancer*[1], *Journal review article on the molecular markers in penile cancer*[2] and *Journal commentary piece on HPV and penile cancer*[3].

1.1 Penile cancer

1.1.1 Epidemiology

Penile cancer is a rare disease in Europe and North America but represents a significant global health problem due to the devastating consequences of treatment and the mortality associated with metastatic disease[4]. The most important prognostic factor is the presence of inguinal lymph node metastases. Surgical resection of the inguinal lymph nodes is the only method to accurately and reliably determine lymph node status. However, in the 75%-80% of cases where no metastases are found, patients have undergone extensive surgery with no survival benefit[5,6]. There is therefore a great clinical need to develop molecular biomarkers to accurately predict lymph node status, alleviating many patients from the harmful effects of lymph node dissection.

The major risk factors for the development of penile cancer include smoking, human papillomavirus (HPV) infection, phimosis, immunodeficiency and age[4]. HPV infection has been demonstrated to be the necessary carcinogenic entity in cervical cancer, as well as highly implicated in head and neck[7], anal and penile squamous cell carcinomas[8].

Penile squamous cell cancer is a heterogeneous disease both morphologically and molecularly. Molecular aberrations causing penile cancer can be classified into two groups depending on the presence of HPV in the samples examined. This allows the construction of both HPV mediated and HPV independent pathways leading to the oncogenesis of penile cancer.

Over the past 30 years, steady work has been undertaken to improve the understanding of the molecular biology of penile cancer. The main limiting factor with regard to researching this field has been the rarity of the disease in the developed world, resulting in a paucity of high-quality, well-powered investigations. Nevertheless, significant studies have been undertaken that examine mutations, chromosomal abnormalities and epigenetic changes affecting the major stages of tumour development. The majority of these studies have been based on a candidate gene approach, but the most recent studies adopt non-a priori approaches to better understand the depth and breadth of molecular aberrations and drivers of this disease. This chapter presents the latest molecular model for penile carcinoma and highlights where further work is needed to increase overall understanding. The model is also further enhanced by the greater number of molecular studies performed in cervical and head and neck squamous cell carcinomas, which bear some similarities to penile squamous cell carcinoma. A greater understanding of these molecular events can drive a more precise management of this disease and may help to elucidate targets for new drug therapies with the ultimate aim of improving morbidity and mortality associated with this mutilating disease.

1.1.2 Human papillomavirus (HPV)

HPV is a DNA virus that can invade and replicate in epithelium of the urogenital tract, the upper respiratory tract and the skin[8]. Many subtypes exist but they can be broadly classified as either high-risk (most commonly types 16 and 18) and low-risk HPV subtypes (most commonly types 6 and 11). High-risk subtypes are responsible for carcinogenesis in almost all cervical carcinomas, and a large proportion of head and neck, anal and penile carcinomas. In the majority of people, HPV infection is cleared within 18 months and can be considered subclinical and benign. However, in certain individuals the infection is not cleared and they become chronically infected. Low-risk subtypes are responsible for benign diseases such as ano-genital warts[9]. High-risk HPV exerts its effects by encoding two oncogenic proteins, E6 and E7, which can cause disruption of both the p14^{ARK}/MDM2/p53 and p16^{INK4a}/cyclin D/Rb pathways, resulting in genetic instability and oncogenesis[9]. The different functions of these oncogenic proteins are shown in Figure 1.

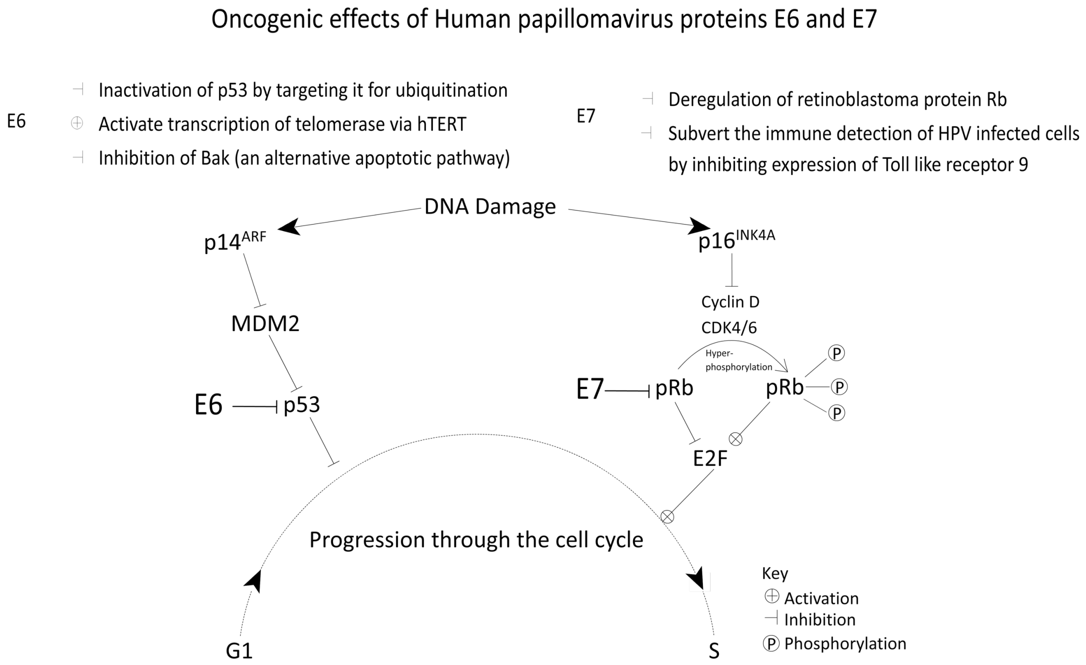

Oncogenic effects of Human papillomavirus proteins E6 and E7
E6
⊣ Inactivation of p53 by targeting it for ubiquitination
⊕ Activate transcription of telomerase via hTERT
⊣ Inhibition of Bak (an alternative apoptotic pathway)
E7
⊣ Deregulation of retinoblastoma protein Rb
⊣ Subvert the immune detection of HPV infected cells by inhibiting expression of Toll like receptor 9
DNA Damage
p14ARF
p16INK4A
MDM2
Cyclin D
CDK4/6
Hyper-phosphorylation
E6
p53
E7
pRb
pRb
Ⓟ
Ⓟ
Ⓟ
E2F
Progression through the cell cycle
G1
S
Key
⊕ Activation
⊣ Inhibition
Ⓟ Phosphorylation

Figure 1: Illustration of the oncogenic effects of HPV proteins E6 and E7[9].

One of the main oncogenic methods E6 utilises is to target p53 for ubiquitination and degradation by the proteasome[9]. E6 has also been shown to inactivate Bak, which is an alternative pathway for apoptosis. E6 is also able to activate transcription of telomerase via hTERT (human telomerase reverse transcriptase)[10]. Telomerase enables maintenance of the telomeres at the end of chromatids, enabling unlimited cell divisions. Telomeres protect the end of chromosomes and gradually get shortened during replication. This 'shortening' results in replicative senescence, giving rise to a protective mechanism that limits uncontrolled cell division. Telomerase activation enables uncontrolled proliferation of a cell to take place far more easily.

E7 exerts its major oncogenic effects by deregulating the cell cycle. Differentiated quiescent cells utilise the tumour suppressor gene retinoblastoma (pRb) to associate with and thus prevent activation of transcription factors (E2F) for progression through the G1 phase of the cell cycle. pRb thus negatively regulates the cell cycle. In a healthy cell, once the exact requirements are in place for a cell to progress through the cell cycle and divide, cyclin D – responding to mitogenic signals – binds and activates cyclin-dependent kinases 4 and 6. These can then phosphorylate and deactivate pRb, thereby activating the transcription factors (E2F) for cell cycle progression. However, E7 binds to pRb, mimicking its phosphorylation, and thus deregulates this cell cycle control[9].

Furthermore, E7 plays a role in subverting immune detection of HPV infected cells by inhibiting the expression of Toll Like Receptor 9 (TLR9)[11], which is a key sensor of the innate immune system responsible for recognising double stranded DNA.

In penile cancer patients the prevalence of HPV infection in Europe, North America, South America and Asia is approximately 40%-50%. The most common subtype is the high-risk subtype HPV 16 at 60%[12]. Other subtypes occur as follows: HPV 18 (13.4%), HPV 6/11 (8.13%), HPV 31 (1.16%), HPV 45 (1.16%), HPV 33 (0.97%), HPV 52 (0.58%), other types (2.47%)[12]. It has been suggested that the molecular carcinogenesis of HPV-related penile carcinoma resembles that of cervical carcinoma. There are epidemiological differences, however. Despite the prevalence of HPV, penile cancer is a very rare disease compared with cervical carcinoma. In addition, it generally affects men 30 years later than the majority of cervical cancer cases affect women. This is likely due to both increased susceptibility of the cervix to malignant transformation as compared with the penile epithelium, and the effect of female hormones on the susceptibility to cervical carcinoma[13].

1.1.3 Molecular biology

1.1.3.1 Stages of tumorigenesis

Tumours develop by accumulating genetic and epigenetic alterations, which result in the cell gaining new malignant functions and losing protective mechanisms. Genetic change refers to either mutations in the DNA sequence or larger scale chromosomal aberrations, whereas epigenetic change refers to alterations in gene expression. The major stages of tumour development include:

- Loss of DNA repair and cell cycle control mechanisms
- Subversion of growth signalling pathways
- Angiogenesis
- Invasion and metastasis

Each of these stages are discussed in turn with evidence for molecular aberrations within penile cancer.

1.1.3.1.1 Loss of DNA repair and cell cycle control mechanisms

One of the early steps in the development of any cancer is disruption of the normal anti-proliferative cell cycle control mechanisms. The replication of a cell is highly regulated to ensure genetic stability. A healthy cell will therefore only replicate after receiving appropriate external growth factors, called mitogens, in the presence of no DNA damage. Mitogens activate mitogen receptors which then signal through transduction proteins, called tyrosine kinases, to activate the G1 and G1/S cyclin-dependent kinases to begin the cell cycle. Multiple control points guard entry into the cell cycle before replication can occur.

The tumour suppressor gene *CDKN2A* encodes proteins that control two of these well-known pathways. The p16^{INK4a}/cyclin D/Rb pathway and the p14^{ARF}/MDM2/p53 pathway have been well examined in a large number of tumours and are frequently found to be disrupted in early oncogenesis. Both of these pathways can arrest the cell cycle in the G1 phase before progression to the S phase of the cell cycle. These two pathways are illustrated in Figure 2.

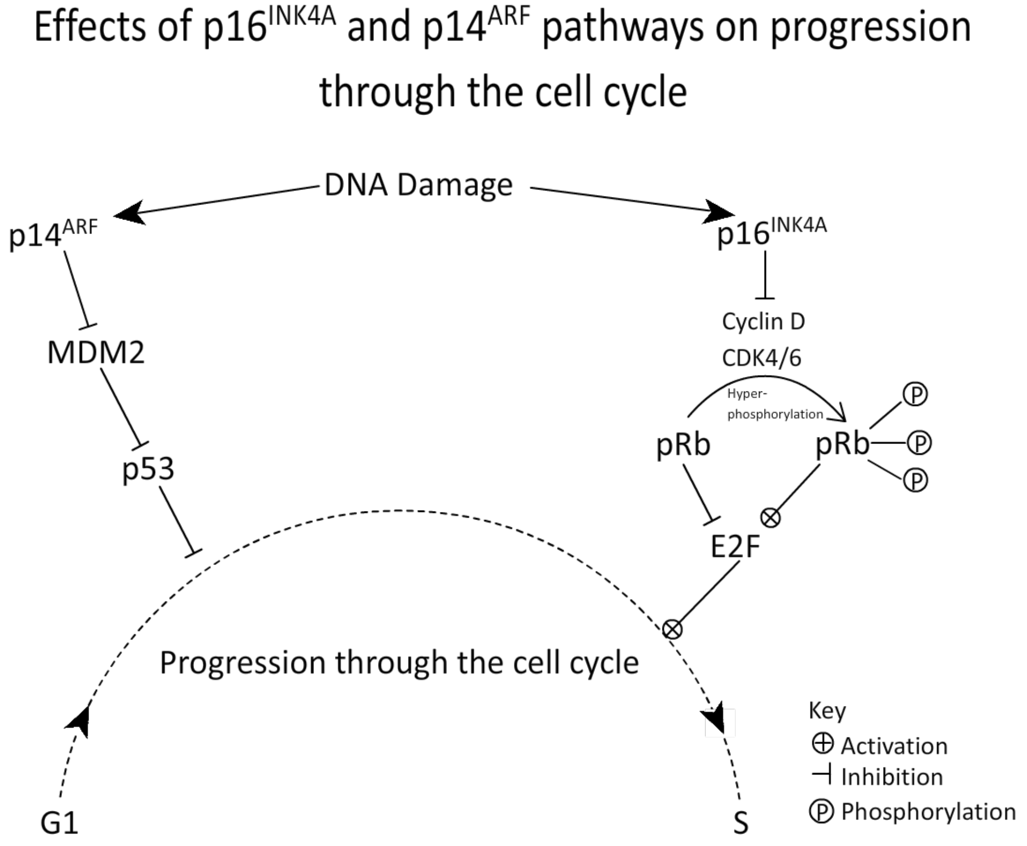


Figure 2: Illustration depicting the effects of the p16^{INK4a}/cyclin d/Rb and p14^{ARF}/MDM2/p53 pathways on cell cycle control [14].

p16^{INK4A} inhibits the G1 cyclin-dependent kinases 4 and 6 (the cyclin D-dependent kinases). These kinases normally initiate phosphorylation of the retinoblastoma tumour suppressor protein Rb, thereby signalling for its degradation. The Rb protein restricts the transcription factors of the E2F family, thereby restricting the cell's ability to replicate DNA, preventing its progression from the G1 phase of the cell cycle. Thus, p16^{INK4a} acts as a tumour suppressor gene with the capacity to arrest the cell in the G1 phase of the cell cycle in response to specific circumstances such as DNA

damage. This makes it an important protective mechanism against genetic instability and it is therefore dysregulated in many cancers.
The p14^{ARF}/MDM2/p53 pathway is also vitally important in ensuring that the cell will not replicate in the presence of DNA damage. It has the ability to direct the cell into senescence, programmed cell death or repair, depending on the level of DNA damage. In response to aberrant growth signalling, p14^{ARF} can form stable complexes with MDM2, thereby promoting p53. p53 is a transcription factor that can promote p21, itself a cyclin-dependent kinase inhibitor. p21 binds and inactivates cyclin-dependent kinase complexes, causing cell cycle arrest at the G1/S checkpoint. This demonstrates that p14^{ARF} acts to prevent tumour development. Mutations in both of these tumour suppressor genes are among those most prevalent in mammalian cancers.

These two pathways are disrupted in penile cancer by multiple mechanisms, including:

- Chromosomal aberrations resulting in loss of heterozygosity
- Promoter hypermethylation, causing downregulation of tumour suppressors
- Point mutations
- Antagonism of tumour suppressors by HPV oncogenic proteins E6 and E7

Analysis of microsatellite markers in penile cancer revealed loss of heterozygosity in the p16^{INK4a} locus in 64% of cases and loss of heterozygosity in the p53 gene in a further 63% of penile squamous cell carcinoma cases[15]. Interestingly when examining the p14^{ARF}/MDM2/p53 pathway, none of the samples had combined loss of heterozygosity of both p53 and MDM2. None of the patients who had loss of heterozygosity, with subsequent loss of expression of p16, had the presence of HPV DNA[15]. This leads to a prediction that an alternative pathway for disruption of these tumour suppressor genes exists in the cases of HPV infection. Furthermore, disruption of both these pathways is heavily implicated in penile cancer oncogenesis.

CpG islands are regions with high levels of CpG sites, many of which are situated at the start of promoter sites. Increased DNA methylation (hypermethylation) of these regions is associated with downstream gene silencing[16,17]. The CpG island status of *CDKN2A* which encodes the two tumour suppressor proteins p16^{INK4A} and p14^{ARF} was examined with methylation levels varying between 0% and 42%[18]. The large range may have been caused in part by the small samples used in many of these studies. However, the frequency of *CDKN2A* promoter hypermethylation was higher in HPV negative tumours than positive cases. Hypermethylation of the *CDKN2A* CpG island was correlated with weak expression of p16.

Expression and immunoreactivity of p53 was considered as a potential biomarker in multiple studies[19-23]. Tumours that stained positive for p53 were associated with a worse 10 year survival (26.4%) than those that stained negative (54.6% p = 0.009)[23]. The same study found an increased relative risk of 4.8 (95% CI = 1.6-14.9) for lymph node metastases[23]. However, cyclin D1 and p21 were not significantly associated with disease-specific mortality. p16^{INK4A} was identified as a marker for favourable prognosis with a hazard ratio of 0.44 (95% CI = 0.23-0.84) [24] with an increase in five-year cancer-specific survival from 57% to 85%. These results were also confirmed in head and neck squamous cell carcinomas[25].

The DNA replication licensing pathway controls the proliferative state of the cell and ensures that the DNA is only replicated once per cell cycle. *MCM2* is downregulated during quiescent states but upregulated during progression through the cell cycle. Its malfunction can result in DNA ploidy. Univariate analysis demonstrated that aneuploidy is a strong prognosticator for overall survival, with a hazard ratio of 4.19 (95% CI = 1.17-14.95, p = 0.03) [26]. However, no association was found on multivariate analysis of *MCM2* expression levels with lymph node metastases[27].

Ki-67 is used as a marker of tumour cell proliferation. It is a nuclear matrix protein that is expressed in all phases of the cell cycle besides G0. Five studies[19,28-31] have examined Ki-67 expression and its association with disease specific mortality and lymph node metastases. Only one of these found an association with lymph node metastases, with a relative risk of 3.73 (95% CI = 1.4-9.7)[31].

1.1.3.1.2 Subversion of growth signalling pathways

One of the most common events in oncogenesis across many cancers is subversion of the growth factor receptor signalling pathways. These tyrosine kinases play a prominent role in the growth and survival of cancer cells. Two important growth signalling pathways – PI3K and Ras – have been implicated in the development of penile cancer[32]. Both these tyrosine kinase pathways activate a cascade of downstream processes. PI3K exerts its effects on downstream targets including cell proliferation, adhesion, motility and intracellular trafficking. PTEN acts as a negative regulator functioning as a tumour suppressor within this pathway. Mutations in *PIK3CA* were found in 29% of penile cancer samples[32]. The Ras pathway is also activated by a receptor tyrosine kinase and consists of HRAS, KRAS and NRAS. These activate ERK which in turn regulates transcription factors controlling cell growth, differentiation and survival. Mutations in *HRAS* and *KRAS* were only found at low frequencies (<10%)[32].

Both the PI3K and Ras pathways can be activated by epidermal growth factor receptors (EGFR). *EGFR* has been seen to be over-expressed in up to 93% of penile cancer cases irrespective of grade, stage or HPV status[33]. This high frequency of overexpression suggests *EGFR* plays an important role in the pathogenesis of SCC. The almost ubiquitous over-expression of *EGFR* in penile cancer cases suggests that anti-EGFR monoclonal antibodies such as cetuximab may be efficacious as targeted therapies.

KAI1 is a cell membrane protein with roles in signal transduction, proliferation and motility. It was originally described as a metastasis suppressor gene in prostate cancer and has since been associated with poor differentiation in cervical carcinomas. As a metastasis suppressor gene its downregulation is associated with metastases in several malignancies. There was no indication of loss of heterozygosity of *KAI1*, so alternative methods of reduced expression need to be considered. In a small sample of 30 patients, *KAI1* expression was associated with increased lymph node metastases ($p = 0.0042$) and disease mortality ($p = 0.0002$) [34].

1.1.3.1.3 Angiogenesis

For a tumour to grow and spread to the inguinal lymph nodes, both angiogenesis and lymphangiogensis are required. Angiogenesis is a multi-step process, involving:

1. Tissue destruction and hypoxia
2. Migration of endothelial cells in response to hypoxia and angiogenic factors
3. Proliferation and stabilisation of endothelial cells

The tyrosine kinase pathways mentioned earlier are responsible for this cascade of new protein production and response to angiogenic factors. Vascular endothelial growth factor (VEGF) is one of the growth factors responsible for angiogenesis. Growth of a tumour outstrips the local blood supply, which creates a hypoxic environment, inducing expression of VEGF. Other growth factors include fibroblast growth factor, transforming growth factor and tumour necrosis factor. Podoplanin is a transmembrane glycoprotein specific to the lymphatic endothelial cells[35]. It is upregulated in many squamous cell carcinomas and associated with lymphangiogenesis[35,36]. It could therefore play a role in penile cancer oncogenesis, and could be used in identification of lymph node metastases. The antibody D2-40 can be used to detect podoplanin. The antibody, together with an intra-tumoural lymphatic vessel density greater than 2, can detect the presence of inguinal lymph node metastases with a specificity of 78% and a sensitivity of 83.3%[37].

1.1.3.1.4 Invasion and metastasis

Invasion involves the tumour adhering, degrading the surrounding extracellular matrix, migrating, and then proliferating at a secondary site[38].

Cadherins are transmembrane proteins that play a key role in regulating cell-to-cell adhesion, which makes them essential to maintaining epithelial integrity and limiting the invasive potential of cells. In a study of 125 patients, 45% of penile cancers were demonstrated to have low E-cadherin expression, which was associated with lymph node metastases[39]. This indicates that E-cadherin may play an important role in the oncogenesis of penile carcinoma. Periostin is a secreted cell adhesion molecule that has been implicated in invasive ovarian, lung and head and neck carcinomas[40]. Cells that overexpress periostin frequently metastasise. In the case of penile carcinoma, high expression was associated with a reduced cancer-specific survival with a hazard ratio of 1.44 (95% CI = 1.14-1.81, p = 0.002)[41].

Matrix metalloproteinases (MMPs) are essential enzymes involved in stromal remodelling. They play a vital role in physiological wound healing but are also utilised by a tumour for malignant invasion. MMPs are not produced by the tumour itself but are induced by CD147/Extracellular matrix metalloproteinase inducer (EMMPRIN) acting on the surrounding stromal cells. High levels of EMMPRIN expression have been found in breast, lung and also penile carcinomas. In one small study of 17 penile cancer patients, high EMMPRIN expression was evident in 41% of patients and was associated with a worse five-year survival with a high relative risk of 420 (95% CI = 51-3460) [42]. The extremely large confidence interval reflects the small number of samples, but the significant result suggests EMMPRIN expression should be considered for future biomarker work. High levels of MMP expression were found in a large cohort of 101 penile cancer patients, in which 72% of samples contained high MMP-2 expression and 25.9% displayed MMP-9 expression[39,43]. The expression of MMP-9 was associated with penile carcinoma recurrence with a relative risk of 3.2 (95% CI = 1.28-8.3)[39]. However, there was no indication of an association with lymph node metastases or prognosis.

1.1.3.2 Penile cancer methylome studies

Two recent studies have undertaken genome-wide methylation analysis[44,45]. As previously explained, methylation of cytosine DNA residues is one epigenetic mechanism of controlling gene expression. Aberrant methylation can result in or be a strong marker for genetic instability, which makes these analyses a useful source of tumour biomarkers. They can also help probe which genes are involved in the oncogenesis of the disease. Both methylation analyses demonstrated overall hypomethylation, which is characteristic of malignant disease. Using 50 penile cancer patient samples, Feber et al found 6,993 positions where differential methylation had occurred when comparing tumours with control samples[44]. Hypermethylation was found in 997 regions in these tumours. These regions corresponded to many tumour suppressor genes, including *CDO1*, *AR1* and *WT1*. A four-gene epigenetic methylation signature (*HMX3*, *IRF4*, *FLI1* and *PPP2R5C*) was then used to accurately predict lymph node status in an independent cohort, with a sensitivity of 93% and a specificity of 80%. Multivariate analysis revealed that this signature was an independent predictor of lymph node metastases ($p = 0.0053$). Aberrant methylation was also identified at several potential therapeutic targets, including *V-EGFR* tyrosine kinases as well as the androgen receptor and programmed cell death receptor 1. Further work is needed to confirm and validate these results, but they represent an exciting opportunity to identify a new generation of clinically useful biomarkers and therapeutic targets. A second analysis by Kuasne et al found CpG hypermethylation and confirmed gene under-expression for a panel of genes including *TWIST1, RSOP2, SOX3, SOX17, PROM1, OTX2, HOXA3* and *MEIS1*[45].

1.1.3.3 Molecular pathways leading to penile tumorigenesis

Although there is a paucity of large molecular studies on penile carcinoma, a preliminary molecular model can be surmised. As already seen, aberrations in many of the major molecular oncogenic pathways have been discovered in penile cancer. The heterogeneity of findings can at least in part be explained as evidence for multiple originating carcinogenic pathways in the development of this disease. One of the initial events in the oncogenesis of penile cancer is disruption of the main pathways that control cell cycle progression. These include the p14^{ARF}/MDM2/p53 and the p16^{INK4a}/cyclin D/Rb pathways. These pathways can be disrupted by HPV infection due to oncogenic E6 and E7, respectively.

Alternatively, they can be disrupted by genetic and epigenetic silencing of tumour suppressors as part of an HPV independent process. Most non-HPV cases are caused by chronic inflammation in the presence of tobacco usage. There is evidence of increased genetic and epigenetic aberrations of the p16^{INK4a}/cyclin D/Rb in non-HPV-associated penile cancer, with mixed evidence for an inverse relationship of p53 mutations and presence of HPV infection. This pathway can also be disturbed in non-HPV infections due to over-expression of *BMI1*, which has been reported in 10% of high-risk HPV negative cases. *BMI1* targets the *CDKN2A* locus, which encodes both p16^{INK4a} and p14^{ARK} [46].

Once these cell cycle checkpoints have been disrupted – by either an HPV or non-HPV dependent process – increasing genetic instability will result in an accumulation of genetic and epigenetic changes. This will disrupt common oncogenic pathways irrespective of HPV status. Disruption will subvert growth signalling pathways, as evidenced by subversion of both PI3K and Ras pathways within penile carcinoma. Eventually angiogenesis and invasive potential will cause aberrations in MMPs, EMMPRIN, cadherins and telomerase, immortalising a tumour with metastatic potential.

1.1.3.4 Molecular biomarkers

DNA copy number variants have been associated with clinical outcome. DNA copy number variation within penile cancer was assessed by Busso-Lopes et al by performing global array comparative genomic hybridization (aCGH) on 46 penile cancer samples[47]. The most frequent alterations were found in chromosomes 3p and 8p and these were also the regions most associated with poor clinical outcome. The gain on chromosome 8 encompasses the gene *MYC.* MYC is a transcription factor responsible for the regulation of a large number of cellular processes relating to tumorigenesis including cell-cycle progression, differentiation and apoptosis[48]. Losses of 3p21.1–p14.3 and gains of 3q25.31–q29 were associated with reduced cancer-specific survival (p = 0.006 and p = 0.023, respectively). These two regions were also associated with a reduced disease free survival (p = 0.023 and p = 0.042, respectively). These regions map to genes *DLC1*, *PPARG*, and *TNFSF10.* In addition to the potential utilisation of *DLC1* and *TNFSF10* as biomarkers, they may also represent future therapeutic targets[48,49].

Other biomarker studies in penile cancer analyse squamous cell carcinoma antigen (SCCA), CD44 and epigenetic methylation signatures. SCCA is a serine protease inhibitor originally identified in the serum of patients with squamous cell carcinoma of the uterine cervix[50]. Elevated levels have also been found in lung, liver and penile cancer patients. Levels of SCCA above 1.5mcg/l are associated with reduced disease-free survival with an odds ratio of 0.13 (95% CI = 0.034-0.55) [51].

CD44 is a cell membrane protein that can be used as a biomarker for cancer stem cells. Its expression has been associated with lymph node metastases. High levels of CD44 expression were found in 73% of patients with lymph node metastases, compared with 44% of patients with no lymph node metastases[52]. However, this level of specificity is not high enough to be used clinically.

Currently no biomarker has been sufficiently validated to be incorporated into the clinical environment. The most significant individual biomarkers associated with mortality include:

- p53: hazard ratio of 3.2 ($p = 0.041$)[20]
- p16^{INK4a} : 27% increased survival at five years[24]
- CD147 ($n = 17$); relative risk 420 (95% CI = 51-3460)[42]
- KAI1: worse five-year prognosis ($p = 0.0042$)[34]

The most promising biomarkers for lymph node metastases consist of the epigenetic methylation panel based on four genes examined by Feber et al[44]. This panel managed to accurately predict lymph node status with a sensitivity of 93% and specificity of 80%[44]. One option would be to introduce additional methylation markers into the panel to further increase these parameters. However, further work is needed to validate these markers and determine the significance with multivariate regression models. It is likely that multiple biomarkers will need to be combined to achieve sufficient sensitivity and specificity. The current gold standard investigation for determining lymph node status is sentinel lymph node biopsy. This has an estimated sensitivity of 89.2%[53].

1.1.4 Current management

The current management of patients with penile cancer in Europe is advised through the European Association of Urology guidelines[54] (last major update in 2014) and the European Society for Medical Oncology penile carcinoma guidelines[55] (last produced in 2013). These guidelines for patients with lymph node metastases are synthesised below:

Surgical options for T2 disease include total glansectomy or partial penile amputation for those unfit for reconstructive surgery. For T3 disease, glansectomy with distal corporectomy and reconstruction or partial amputation are recommended. For T4 disease, extensive partial or total penectomy with perineal urethrostomy are recommended. In addition, neoadjuvant chemotherapy may be considered. All patients with stage and grade of disease greater than T1 and G1 should have lymph node sampling to exclude the presence of lymph node metastases. This is because patient survival is more than 90% with early lymphadenectomy, and below 40% with lymphadenectomy for regional recurrence. At the time of diagnosis, the only way to determine the presence of lymph node micrometastatic disease is by surgical sampling, which is commonly carried out via sentinel lymph node biopsy in non-palpable disease.

In patients with N2 or N3 disease adjuvant chemotherapy is recommended as part of the ESMO 2013 guidelines. There is a paucity of high quality research evaluating adjuvant chemotherapy in penile cancer. The current guideline recommendations are based on generally small retrospective studies with no valid controls. Within these retrospective studies adjuvant chemotherapy appears to improve the long-term disease-free survival from 39% to 84%[56,57]. There is currently a lack of evidence demonstrating any efficacy to adjuvant radiotherapy.

For patients with stage 4, locally advanced or metastatic disease there is a wide variability in regimens. Patients are most commonly treated with cisplatin plus a taxane or 5-fluorouracil. Other options include replacing 5-fluorouracil with gemcitabine or irinotecan.

A summary of management guidelines from ESMO 2013 is reproduced below for the primary tumour and inguinal lymph nodes in Figure 3 and Figure 4. I am grateful to Oxford University Press for granting permission to reproduce these two figures.

[redacted]

Figure 3: Schematic of management guidelines for the management of primary penile tumour. Reproduced with permission from ESMO 2013 Clinical Practice Guidelines for diagnosis, treatment and follow-up[55].

[redacted]

Figure 4: Schematic of management guidelines for management of inguinal lymph nodes. Reproduced with permission from ESMO 2013 Clinical Practice Guidelines for diagnosis, treatment and follow-up[55].

1.1.5 Current clinical trials and future

A list of the most recent clinical trials was downloaded from the clinical trials database at clinicaltrials.gov. As demonstrated in Figure 5, there has been a promising increase in the number of new trials open to penile cancer patients over the past decade (2008-18).

Prior to the commencement of research for this thesis, there were almost no clinical trial registrations for penile cancer that sought to assess the viability of new chemotherapy agents, targeted therapies, immunotherapies or vaccinations. Table 1 displays the list of trials for penile cancer therapies that have commenced since 2013. It is pleasing to see both the acceleration in the number of studies and the breadth of therapies under consideration. However, despite an increase in the number of trials open to penile cancer patients, there is still a lack of trials specifically dedicated to penile cancer. There is a danger that pan-cancer trials may ultimately fail to recruit a significant number of penile cancer patients, thus precluding meaningful conclusions about treatment efficacy for penile cancer. Many of the new trials registrations are testing targeted therapies and immunotherapies that are discussed throughout this thesis.

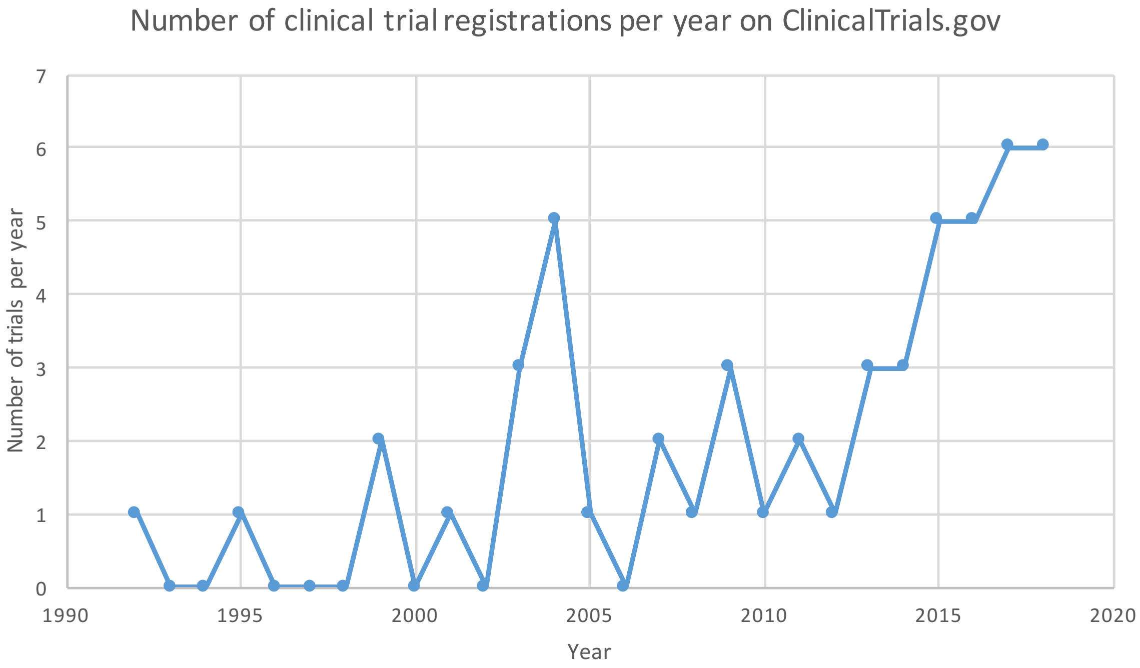


Figure 5: Number of new clinical trial registrations involving penile cancer per year found on ClinicalTrials.gov

Table 1: Table of the therapeutics being tested in new clinical trial registrations over the past five years in patients with penile squamous cell carcinoma. The final column of the table asks whether the trial has specific recruitment target for penile cancer.

Drug	Type	Subtype	Year	Penile cancer focused trial?
Dacomitinib	EGFR inhibitor		2013	Yes
Vinflunine	vinca alkaloid		2014	Yes
Cabazitaxel	Taxane		2015	Yes
Cabozantinib	Receptor typrosine kinase inhibitor	MET inhibitor, RET inhibitor, VEGFR inhibitor, KIT inhibitor, FLT-3 inhibitor	2015	No
Paclitaxel	Taxane		2015	Yes
Pazopanib	Receptor typrosine kinase inhibitor	Greatest inhibition of : c-KIT, FGFR, PDGFR and VEGFR	2015	Yes
PDE-5 inhibitors	PDE5 inhibitor		2015	No
HPV specific T Cells	T cell immunotherapy		2015	Joint with other HPV driven malignancies
Cetuximab	EGFR inhibitor		2016	Yes
Pembrolizumab	Immune checkpoint inhibitor	PDL-1 inhibitor	2016	Yes
Ipilumab	Immune checkpoint inhibitor	CTLA-4 inhibitor	2017	No
Nivolumab	Immune checkpoint inhibitor	PD-1 inhibitor	2017	No
HPV vaccination	HPV immunotherapy		2017	Joint with other HPV driven malignancies
Avelumab	Immune checkpoint inhibitor	PD-L1	2018	Yes

1.1.6 Future perspectives

Due to the rarity of penile carcinoma, relatively slow progress has been made to uncover its genetic and epigenetic drivers. Nonetheless several studies in the last five years have made significant contributions to this field of work. A shift in approach across all cancers away from examining individual candidate loci in favour of whole genome/epigenome analysis provides a much greater understanding of the breadth and depth of oncogenic mechanisms. Further research is needed in the form of whole exome sequencing and expression analysis to tie in with the recent methylome projects. These studies will be vitally important in detecting biomarkers in order to determine the lymph node status of these patients. A biomarker with sufficient negative predictive value could reduce the need for surgery, which is currently of small benefit to the majority of patients. There is increasing evidence for the use of a panel of biomarkers, which could combine molecular markers, radiological and epidemiological factors.

Additional molecular work will also be immensely useful in determining which chemotherapeutic, targeted therapies and immunotherapy agents will be efficacious. These new therapeutic agents have the potential to dramatically improve morbidity and mortality of patients with metastatic penile cancer.

There are now a range of clinical trials ongoing for these new types of therapies which are open to penile cancer patients. Despite many of these not being focused on penile cancer they still represent an excellent opportunity for patients who have either failed to respond to current platinum-based therapies or are unfit for systemic chemotherapy. They also provide an excellent opportunity to gain valuable phase 2 trial data on the response rate of these new agents for penile cancer. It is hoped this will provide the necessary evidence to undertake larger dedicated penile cancer trials in the future.

1.1.7 Oncogenesis as an evolutionary model

Cancer has long been thought of as an evolutionary process since the early works of Nowell in 1976[58]. The basic premise is that some mutations confer a fitness advantage over the surrounding normal cells, allowing it to divide more rapidly. Mutations can be accumulated over time, with each one persisting if proven to confer an advantage, as the cells outcompete surrounding cells without the new mutation. These mutations confer advantages such as disruption of cell cycle control, inhibition of cell death, loss of DNA repair, inhibition of surrounding immune surveillance, promotion of angiogenesis, or gaining of new cell motility functions.

Mutations here can mean any heritable molecular change which can be passed from one cell to daughter cells. This therefore includes genetic changes such as single nucleotide variants (SNVs), insertions, deletions, structural variants, copy number aberrations and epigenetic changes, with the potential to change gene expression, such as methylation aberrations, histone modifications or small RNAs.

This model predicts that cancers form over a stepwise evolutionary process consisting of mutations and subsequent clonal expansion. Each new driving mutation can therefore be traced back to a last common ancestor. Assuming these changes are irreversible – to be discussed in more detail below – a cancer cell will therefore contain not only a current blueprint into the workings and susceptibility of the cell at the current time, but also a history of previous drivers and whether these are genetic or epigenetic. The exact makeup of the proportion of cells with any given driver will depend on the evolutionary model followed by the cancer, as explained further in Section 1.1.7.1.

In addition to positive selection in evolutionary process, other evolutionary processes such as negative selection and neutral evolutionary drift will occur. Therefore, the genetic and epigenetic aberrations can be classified simplistically into drivers, which confer a fitness advantage, and passengers which currently do not confer an advantage but are nevertheless heritable between generations of cell divisions. This classification is not concrete, as an aberration that was initially a passenger may confer a fitness advantage under differing conditions in other tissues, microenvironments or points in time.

The assessment of drivers can be accomplished in the following ways:

- Estimation as to whether a gene is recurrently mutated beyond what would be expected by chance
- Observation of functional effects of mutated genes in cell lines, animal models or even tumouroids.

Cancers originating across a variety of tissues nonetheless share many of the same oncogenic processes that disrupt normal processes and gain malignant functions. Therefore, mutations found to be functional drivers in one cancer are likely to be present in other cancers. As there is a paucity of genomics and functional data in penile cancer, this thesis will highlight genes found in this research to also be recurrently mutated in other cancers, as they may represent candidate drivers within penile cancer.

1.1.7.1 Models of tumour evolution

Competing models exist to explain the genetic and epigenetic diversity reported in next generation sequencing studies. These models are based on a range of assumptions relating to the strength of selection on a group of cells and the timings of aberrations during oncogenesis. Figure 6, reproduced with kind permission from Davis et al, depicts these four competing models. However, there is also evidence that tumours could switch from one model to another depending on the selective pressure and tumour microenvironment[59].

The four widely discussed models are:

1. Linear evolution – each additional driving mutation confers such a strong competitive advantage that it outcompetes all previous clones as part of a selective sweep.
2. Branched evolution – subclones diverge from common ancestors and evolve in parallel. In this model selective sweeps are rare and result in multiple subclonal lineages simultaneously.
3. Neutral evolution – there is no selective advantage to each additional mutation. Random mutations therefore accumulate over time, leading to genetic drift and large numbers of subclones with unique genotypes.
4. Punctuated evolution – periods of sequential bursts of genetic/epigenetic mutations alternating with periods of neutral evolution.

These models of tumour evolution are demonstrated in different graphical forms in Figure 7 and Figure 8.

The majority of evidence from published series thus far points to the branched evolutionary models. Indeed, there is even some evidence for co-operation between subclones in branched evolution. An example of this has been reported by Inda et al[60], demonstrating that in an in-vivo model of glioblastoma, a small population of cells with mutant *EGFR* could drive the growth of surrounding cells by a paracrine cytokine reaction.

There is also evidence in specific cancers for punctuated evolutionary models. Examples of this include pancreatic cancer[61] and prostate cancer[62]. Baca et al demonstrated evidence for punctuated evolution in 88% of samples from a cohort of 57 prostate adenocarcinoma samples. These samples demonstrated co-ordinated DNA translocations and deletions, inducing extensive dysregulation of prostate cancer genes[62].

The status of which model best fits with advanced squamous cell carcinoma will be assessed in Chapter 3.

[redacted]

Figure 6: Illustration of tumour evolution models showing dynamic changes in clonal frequencies over time. This figure is based on the original publication by Marusyk and Polyak[63]. (A) Linear Evolution (B) Branching Evolution (C) Neutral Evolution (D) Punctuated Evolution. Colours indicate clones with different genotypes. Figure and caption reproduced with permission from Davis et al[64].

[redacted]

Figure 7: Progression of ITH in tumour evolution models: Changes in intra-tumour heterogeneity during tumour progression in the context of different tumour evolution models. (A) Linear evolution (B) Branching Evolution (C) Punctuated Evolution (D) Neutral Evolution. Colours indicate different genotypes of clones. Figure and caption reproduced with permission from Davis et al[64].

[redacted]

Figure 8: Clonal lineages and phylogenetic trees: Phylogenetic trees expected from different models of tumour evolution (A) Linear Evolution (B) Branching Evolution (C) Neutral Evolution (D) Punctuated Evolution. Colours indicate clones with different genotypes. Figure and caption reproduced with permission from Davis et al[64].

1.1.8 Intra-tumour heterogeneity (ITH)

Cancers are heterogeneous by their very nature and can be classified into different types of heterogeneity including

- Inter-patient
- Intra-patient
- Intra-tumour

Inter-patient heterogeneity refers to the differences between the same tumour types in different individuals. Each mutation will only ever be present in a subsection of patients' tumours for each solid cancer type. Some drivers appear to be more recurrent in a patient cohort than others. To illustrate: a study using whole exome sequencing of invasive bladder cancer patients found mutations in *TP53* in 34.8% of the patient population and mutations in KDM6A in 16.3% of the patient population[65]. The heterogeneity between patients may partly explain why certain therapies work better on some cohort of patients compared with others.

Intra-patient heterogeneity refers to the differences between tumours within a single patient. For instance, this can refer to the differences between multiple metastatic deposits. These differences are important clinically, as a personalised targeted treatment approach based on the biopsy of one tumour in a patient may not be representative of the tumours in other locations within the same patient.

Inter-tumour heterogeneity simply signifies differences between cells or groups of cells across different tumours. This could therefore refer to inter-patient or intra-patient heterogeneity.

Intra-tumour heterogeneity (abbreviated in this thesis as ITH) refers to the differences between cells or groups of cells within an individual tumour in a specific patient. These differences can be important in terms of diagnostics, prognostics and response to treatments. Certain tumour types appear to harbour larger amounts of ITH than others[66]. This is likely due to the early oncogenic pathways particular to a specific cancer, as well as the smoking status and mean age of onset of a particular tumour type. This is important clinically as high levels of ITH have been associated with a poor prognosis[67], treatment failure and susceptibility to checkpoint immunotherapies[68,69]. ITH has been a characterised since differences between tumour cells were observed when examining tumours microscopically[70]. However, this field has greatly expanded over the past decade with the introduction of next-generation sequencing technologies. This has enabled molecular aberrations to be detected at high sensitivities and has enabled the computation of the resulting tumour structure consisting of clones and subclones[71,72] (Section 1.1.8.1 below) together with phylogenetic trees[72] (Section 1.1.8.4 below).

The majority of heterogeneity analysis in this thesis will be concerned with intra-tumour heterogeneity. However, I will touch on the heterogeneity between patients, specifically the differences between those with HPV positive disease and those with HPV negative disease. In addition, there will be an analysis of intra-patient heterogeneity comparing the primary tumour to that of the lymph node metastasis.

1.1.8.1 Assessment of clones and subclones

At any point in time a tumour consists of a dynamic population of cells that can be grouped according to the specific set mutations and aberrations. As long as a population of similarly mutated cells has an evolutionary advantage, it will continue to grow to take up a progressively larger proportion of the tumour. New driver mutations may develop within one of these cells and if this confers an additional advantage then this cell will more rapidly divide, producing a new group of cells. This new colony of cells will be dividing more quickly, competing with its direct ancestors, as well as groups of cells from previous ancestors and other colonies. This concept results in the idea of colonies being classified as clones and subclones.

A clone can be defined in the contexts of cancer evolutionary biology as a set of cells that share a common genotype due to their descent from a common ancestor.

As the cancer evolves it accumulates colonies of cells as described above with differing driver mutations. These new groups of cells, now divergent from their last common ancestor, can be considered subclones.

The importance of this distinction is that it enables a particular aberration to be classified as clonal, if it was likely present at the time of the last common ancestor, or subclonal if it developed at any time point after that. Clonal mutations are therefore present in all cancer cells. The proportion of cancer cells that contain a specific mutation is defined as the cancer cell fraction (CCF), see Methods, Section 2. A CCF of 100% therefore implies that the mutation in question was clonal in origin. This logic assumes irreversibility of mutations, non-parallel evolution, and that the mutation cannot be lost due to subsequent deletions or copy losses.

Although this does not hold true in all cases, this logic allows a model to be constructed to evaluate the clonal and subclonal structure of the tumour.

1.1.8.2 Clones and subclones applied to targeted therapies

The distinction between clones and subclones is clinically important for the fields of targeted therapies and biomarkers. Targeted therapies are pharmacotherapeutic agents, usually monoclonal antibodies or small inhibitory molecules that target a specific protein or enzyme that is aberrantly over-expressed in a particular cancer. One of the first targeted therapies for solid malignancies was Trastuzumab (Herceptin), which targeted the Her2/neu(ERBB2) receptor tyrosine kinase[73]. Since then many have been approved, the majority of which target tyrosine kinases and their receptors such as EGFR and VEGF. These types of targeted therapies have the highest chance of success if the protein they are targeting is present in all the cells of the tumour – in other words, the protein is clonally over-expressed with a CCF of 100%. If the mutation is present in a subclone then it is likely that the targeted therapy will only be targeting a portion of the tumour. One can hypothesise that is would be less likely to be successful.

1.1.8.3 Neoantigens and immunotherapy

As tumours become more genetically unstable they accrue mutations, which eventually affects all the normal control mechanisms for ensuring DNA replication is performed accurately. This results in accelerated production of mutations, epigenetic aberrations, structural variants and copy number changes. Tumour neoantigens are antigenic peptides that are entirely absent from the normal human genome[74]. These can be created by tumour specific DNA alterations, causing the formation of novel protein sequences[74]. As the ITH increases within a cancer, the chances of one subclone of the tumour not being susceptible to a particular targeted therapy increases[75,76]. Despite this, increased ITH may also be the tumour's downfall. This is because tumours with large amounts of ITH will have large numbers of novel genomic sequences, which when transcribed and translated lead to the production of tumour neoantigens[77]. Therefore, although patients with large amounts of ITH may be less susceptible to individual targeted therapies, they may be more susceptible to immunotherapies as there is a larger pool of tumour-associated antigens to be targeted by T cells.

Over the last decade, a new class of therapeutics involving immune checkpoint blockade has been introduced for several cancers, including melanoma[78], NSCLC (non-small cell lung cancer)[79] and bladder cancer[80], amongst others. As the tumour grows it appears to create a microenvironment of localised immune suppression by activating inhibitory immune molecules such as CTLA4, PD1 and PDL1. These suppress the T cell response, which is normally responsible for the immune surveillance a healthy individual has against malignancies[81]. Targeting these immune checkpoints with inhibitors of CTLA4, PD1, PDL1 and others is possible with some remarkable complete responses generated in a portion of patients[81].

Previously published work has proposed that patients are more likely to respond to immunotherapies if the tumours in question have a high mutational load and a large number of neoantigens[77]. In addition, expression of these inhibitory immune checkpoints can be evaluated, which may also influence treatment response[81]. In Chapter 3, the tumour mutational load of the tumours in my cohort of patients will be assessed, while in Chapter 5, the expression of key immune checkpoints will be assessed.

1.1.8.4 Phylogenetic trees

A phylogenetic tree or clone tree represents the evolutionary relationships among different cell lineages identified[82]. It is useful for modelling the ITH by depicting the relative structure of the tumour. In addition, it can be annotated with drivers and therapeutic targets to give understanding of which drivers would make the best potential targets and the relative proportion of clonal versus subclonal drivers. It can also be used to help ascertain the relationship between the primary and metastatic samples.

In previously published work there are two types of ‘phylogenetic trees’ used to display this tumour structure. The first is regional/sample trees and the second is clonal trees.

1.1.8.4.1 Regional/sample phylogenetic trees

Regional/sample phylogenetic trees are not true phylogenetic trees as they do not attempt to resolve the individual clones or genetic lineages. Instead, they are a type of similarity tree used to assess the relationship between different tumour samples[82]. Despite this failing they can easily be produced by creating a distance matrix based on a specific matrix of comparable values between samples/regions of a tumour. They can consist of a trunk (representing the molecular aberrations that are found in every region), with shared branches (representing aberrations that are present in several but not all regions) and unique branches (representing aberrations that are present in only one region) belonging to each sample or region sequenced. The methods involved in producing these trees can be found in the Methods in Chapter 2.

1.1.8.4.2 Clonal phylogenetic trees

Clonal phylogenetic trees take into account the frequency of each aberration and cluster each of these into a clone and potential subclones depending on the CCF of each colony[83]. When produced as part of multi-regional sequencing it takes these into account for each sample and then produces a consensus. Copy number aberration data is frequently integrated to provide further information to accurately obtain the true CCF. The methods for producing these trees are discussed in the Methods in Chapter 2.

Once the mutations are split into clusters with attached CCF values, a phylogenetic tree can be produced by following some baseline principles, namely:

- Clones can always be ordered linearly[84]
- Pigeon hole rule/Sum rule: clusters can be arranged in a fork unless the sum of the children is greater than the CCF of the parent[85]
- Crossing rule: the same cluster of mutations cannot appear simultaneously on different branches[86]

In addition, there are several assumptions to assist in the modelling, namely:

- The assumption of irreversibility of mutations: once an SNV has taken place it will not revert back to normal via another mutation at a later time. This is a safer assumption to make for DNA mutations than for more dynamic aberrations such as methylation changes at CpG level.
- The assumption of non-parallel evolution: a particular mutation cannot happen independently in differing tumour clones.

Although these assumptions may be violated at an individual mutation level - for example, an individual mutation in a gene may occur more than once - this is unlikely to occur for all mutations within a subclone cluster[87]. It therefore remains a useful tool to compute phylogenetic trees as demonstrated in a range of other studies[85-86].

1.1.8.5 Circulating tumour DNA (ctDNA)

During the expected turnover of cells in any organ, genomic DNA content is released into the blood. This genomic content is fragmented into lengths of DNA, wrapped around a nucleosome of lengths approximately 160 base pairs, and is known as circulating free DNA (cfDNA). Any solid tumour undergoes rapid growth and turnover of cells resulting in tumour DNA being released into the blood in the same way. This DNA is referred to as circulating tumour DNA (ctDNA). ctDNA has the potential to revolutionise the diagnosis and surveillance of cancer patients. ctDNA provides a minimally invasive liquid biopsy capable of being sequenced to provide the genetic landscape of a tumour. Some of the challenges of this technology are: the fragmented nature of the DNA; the low concentration of DNA in the plasma; background noise from other circulating free DNA; and the possibility that some zones of the tumour do not excrete ctDNA, resulting in biased tumour profiling.

One advantage of ctDNA is being able to obtain molecular diagnostic information from a tumour in patients where a traditional surgical biopsy is not possible. In addition, due to the minimally invasive process of performing a blood test, ctDNA can be collected at frequent intervals, producing almost real-time updates on the molecular behaviour and resistance patterns of a tumour.

ctDNA also provides one potential solution to some of the problems faced in treating cancers that display large amounts of ITH. ctDNA can theoretically provide an overview of the entire cancer not restricted to the precise location of a physical biopsy. ctDNA can therefore provide genetic resistance information of the subclones that may become prevalent post treatment but may not have been considered with traditional surgical biopsy. Furthermore, in patients with metastatic disease, ctDNA provides molecular information on all cancer deposits throughout the patient.

As technologies develop there will be tangible clinical advantages to providing molecular clonal and subclonal information about a patient’s cancer.

1.2 Aims and objectives of this thesis

As demonstrated in the first section of this chapter there is a great need to improve our understanding of the oncogenesis of penile cancer and enable new treatments to be developed. In response to this need and with the capabilities of my lab in the UCL Cancer Institute, the following aims were instigated:

1. Uncover key drivers in penile squamous cell carcinoma with the potential for therapeutic intervention
2. Characterise the intra-tumour heterogeneity (ITH) of penile cancer to classify the importance of drivers in terms of early versus late or clonal versus subclonal changes.

These aims were fulfilled by undertaking three major experimental projects using the same matched samples from a cohort of patients with invasive penile cancer, as described in Chapter 2.

All three projects involved multi-region sampling from each primary tumour together with tissue adjacent normal and lymph node metastasis to investigate different molecular aberrations that may be driving penile cancer. In addition, these changes were used to characterise the drivers in terms of the intra-tumour heterogeneity.

In the first project, in Chapter 3, I assess DNA mutations and copy number aberrations. In the second project, in Chapter 4, I assess methylation aberrations (a type of epigenetic change). In the third project, in Chapter 5, I assess mRNA expression changes and integrate these data with the data from Chapters 3 and 4.

2 Methods

All experimental methods were completed by me unless otherwise stated.

All bioinformatics pipelines were created by me for the use of this project unless stated and cited.

Where significance tests were performed, unless otherwise stated, adjusted p values were assigned based on Benjamini-Hochberg correction of a Wald Chi-Squared test applied for each variant.

The majority of statistical analyses were performed using R version 3.4.0. The following R programming packages in Table 2 were utilised in the analysis explained in Section 2.3.

Table 2: All statistical packages utilised in R for the analysis of experiments listed in Chapters 3–5.

AnnotationDbi 1.39.1	Ape 5.1
Biobase 2.37.22	biomaRT 2.33.3
Biostrings 2.45.1	Bsseq 1.13.2
Bumphunter 1.16.0	CNAmet 1.2
CNATools 1.33.0	Copynumber 1.17.0
Cultevo 1.0.2	Data.table 1.10.4
DESeq2 1.17.8	DMRcate 1.14.0
Dplyr 0.7.2	FlowSorted.Blood.450k
Foreach 1.4.3	Fuzzyjoin 0.1.4
Gage 2.28.2	geneFilter 1.58.1
GenomicRanges 1.29.4	GenomicFeatures 1.29.1
GenVisR 1.7.0	GEOquery 2.43.0
Ggplot2 2.2.1	Ggpubr 0.1.5
Gviz 1.21.1	iGC 1.8.0
IlluminaHumanMehtylation450kanno.ilmn12.hg19 0.6.0	IlluminaHumanMehtylationEPICanno.ilmn10b2.hg19 0.6.0
Limma 3.33.3	Maftools 1.2.0
Magrittr 1.5	MethylMix 2.8.0

Minfi 1.22.1	missMethyl 1.12.0
mixOmics 6.2.0	MNF 0.20.6
Org.Hs.eg.db 3.4.1	Parallel 3.4.0
RColorBrewer 1.1-2	Readr 1.1.1
Reshape2 1.4.2	Rphylip 0.1-23
Session 1.0.3	Sqldf 0.4-11
Stringr 1.2.0	SummarizedExperiemnt 1.7.2
VennDiagram 1.6.17	Xseq 0.2.1

Tables were created using Microsoft Office 2016. Illustrations were formatted to improve readability using Adobe Illustrator CS6 and Sketch version 50.

2.1 Patient selection

2.1.1 Penile cancer heterogeneity cohort (PenHet)

The majority of cancer genomics and epigenomics undertaken internationally has primarily involved sequencing of primary tumour samples. The most likely reason is that the primary tumour site is most frequently accessible for translational tissue work, as biopsies are currently undertaken as part of the normal standard of care. Fresh tissue surplus to diagnostic need may therefore be used for research, or alternatively FFPE blocks may be fairly easily obtained.

In order to improve our understanding of the oncogenesis of the disease, multiple samples would ideally be used both spatially and chronologically. The normal standard of care when managing patients with penile cancer is that any patient with the minimum stage of T1G2 disease requires sampling of the inguinal lymph nodes[88]. It was therefore fairly easy to obtain lymph node metastatic tissue to carry out research in this cancer. The cohort of patients used throughout this study, herein referred to as the PenHet cohort, were recruited to the joint University College London/University College London Hospital onco-urology biobank. Ethical approval for this biobank was obtained on 5th August 2010 with REC reference number 10/H1306/42.

The inclusion criteria for the PenHet cohort were:

- Minimum of four samples available from each primary penile cancer
- Minimum of one lymph node metastasis
- Presence of adjacent 'normal' tissue, taken from the closest macroscopically 'normal' tissue (either penile glans or shaft) as assessed by the urologist performing the primary surgery.
- Adjacent normal tissue assessed as being microscopically 'normal' by a trained uro-histopathologist
- Whole blood for germline sequencing
- Primary tumour cellular purity of primary samples as assessed by H&E staining and verified by a consultant uro-histopathologist of > 80%

The exclusion criteria for the PenHet cohort were:

- Degraded DNA post DNA purification
- Degraded RNA post RNA purification

Ten patients were initially recruited under the joint UCL/UCLH uro-oncology biobank ethical approval (REC: 10/H1306/42). Two patients were excluded as the resulting RNA extracted was severely degraded across all samples, to the point of not being usable for RNA sequencing. Eight patients met the selection criteria, with a full set of four primary tumour samples, one lymph node metastasis, one adjacent normal and one whole blood sample.

In the case of four of the patients, the tissue-adjacent normal provided insufficient RNA for RNA sequencing and they were therefore supplemented with a panel of four further normal samples from further patients recruited to the uro-oncology biobank.

All eight patients had tumours of the common squamous cell carcinoma subtype.

2.1.1.1 Clinical and histopathological characteristics:

The baseline clinical and histopathological characteristics are displayed in Table 3.

Table 3: Table displaying clinic-pathological characteristics of the eight patients with penile cancer included in this study.

Patient identifier	Age	Grade	Stage - T	Stage - N	Stage - M	Smoker
39	51	G3	T2	N2	M0	Ex heavy
45	78	G3	T3	N3	M1	No
49	84	G3	T2	N3	M1	Unknown
51	88	G2-3	T2	N3	M0	Unknown
63	49	G2	T1	N2	M0	Ex heavy
64	53	G2	T3-4	N3	M1	Unknown
66	56	G2-3	T2	N3	M0	Unknown
79	59	G3	T3	N3	M1	Unknown

2.1.1.2 Histopathological cellularity

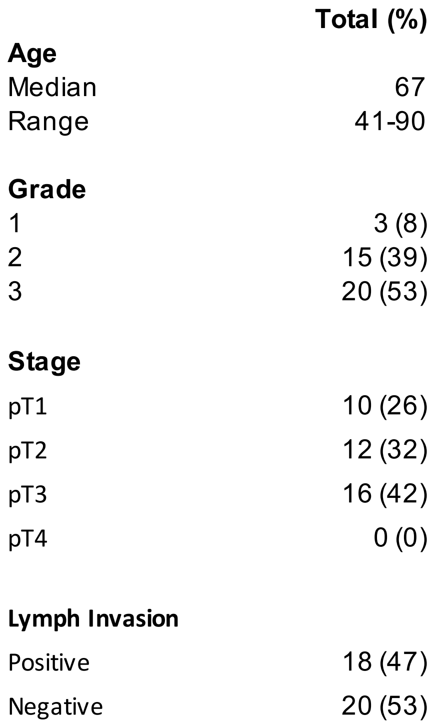
All samples were processed as described in Section 2.2.2. Tumour cellularity of each sample based on haematoxylin and eosin staining was assessed in conjunction with Dr Alex Freeman, consultant histopathologist at University College London Hospital, who has more than ten years' experience in reviewing penile carcinoma specimens. All primary tumour samples had a tumour purity of greater than 80%. One lymph node metastasis contained a reduced tumour purity consisting 50% necrotic tissue, 30% tumour cells and 20% normal lymphoid tissue.

2.1.2 Methylation corroboration cohort

Previously published data[44] from a second cohort of patients previously recruited to the joint UCL/UCLH uro-oncology biobank was used to determine if the significant differentially methylated positions discovered in the PenHet cohort in Chapter 4 could be found in a larger cohort of penile cancer patients.

Identical methods were used for this cohort and the PenHet cohort in terms of sample, processing and Illumina methylation array bioinformatics analysis. This dataset was kindly provided by the research scientist in our laboratory, Dr Andrew Feber.

Histopathological data for this additional cohort can be found in Figure 9.



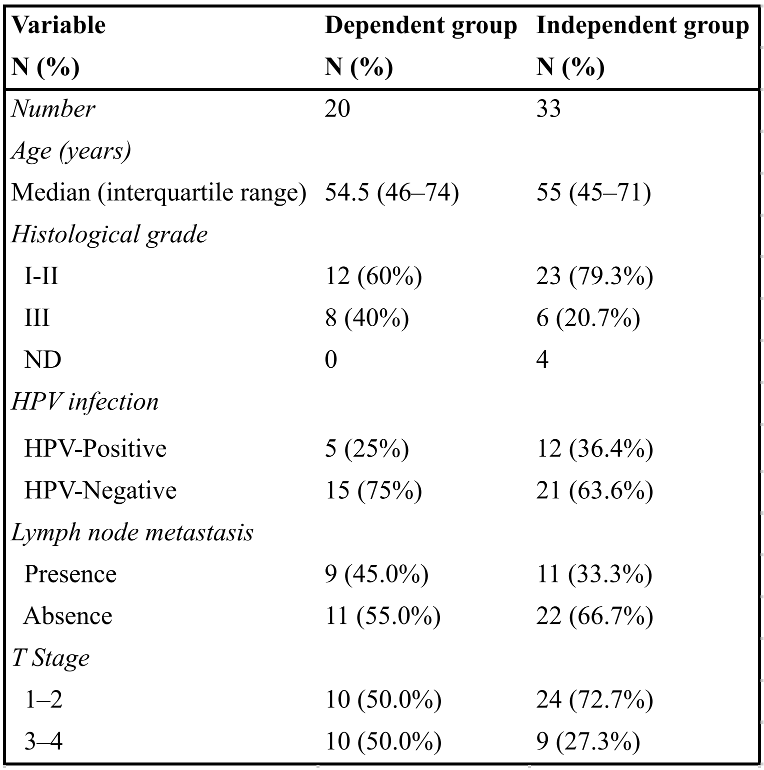
	Total (%)
Age	
Median	67
Range	41-90
Grade	
1	3 (8)
2	15 (39)
3	20 (53)
Stage	
pT1	10 (26)
pT2	12 (32)
pT3	16 (42)
pT4	0 (0)
Lymph Invasion	
Positive	18 (47)
Negative	20 (53)

Figure 9: Histopathological baseline patient characteristics of the previously published dataset from 38 samples patients.

2.1.3 RNA expression corroboration independent external cohort

An external previously published independent data set from Marchi et al[89] was used as a means of corroborating the differentially expressed genes discovered in the PenHet cohort in chapter 5.

This dataset was downloaded from GEO accession number GSE57955. Clinico-pathological information for this dataset is displayed in Figure 10.



Variable	Dependent group	Independent group
N (%)	N (%)	N (%)
Number	20	33
Age (years)		
Median (interquartile range)	54.5 (46–74)	55 (45–71)
Histological grade		
I-II	12 (60%)	23 (79.3%)
III	8 (40%)	6 (20.7%)
ND	0	4
HPV infection		
HPV-Positive	5 (25%)	12 (36.4%)
HPV-Negative	15 (75%)	21 (63.6%)
Lymph node metastasis		
Presence	9 (45.0%)	11 (33.3%)
Absence	11 (55.0%)	22 (66.7%)
T Stage		
1–2	10 (50.0%)	24 (72.7%)
3–4	10 (50.0%)	9 (27.3%)

Figure 10: Clinico-pathological information from independent cohort of patients previously published by Marchi et al [89]

The dataset was processed using the same methods to those of the published paper by Marchi et al[89]. In summary, differentially expressed genes were evaluated using R package 'limma' as described in the published methods. Genes were filtered and deemed significant if the adjusted p value was < 0.01.

The results from this Marchi cohort of patients have very different baseline characteristics. The Marchi patients are more heterogeneous, with only 27% having lymph node metastasis and 20% having low grade 1 disease, compared with 100% of patients in the PenHet cohort having aggressive grade 2-3 disease and lymph node metastases. These results were therefore used cautiously when compared with the results from the PenHet cohort.

2.2 Sample processing

2.2.1 Sample collection

Five samples were excised from each primary tumour in a spatially similar manner as depicted in Figure 11, with five samples taken around a 'clock face'. In addition, one further sample was excised from each predicted metastatic lymph node. ‘Normal’ tissue was taken from tissue adjacent to the tumour on the glans or shaft of the penis. Each sample was approximately 5mm x 3mm x 3mm.

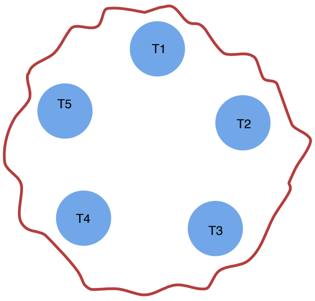


Figure 11: Diagram of a tumour with a ‘clock face’ depicted where five tumour samples are taken sequentially spatially around the tumour. Diagram not to scale. The red outline depicts the border of the tumour whilst each blue circle depicts a biopsy taken for analysis.

Fresh samples were collected intra-operatively and immediately submerged into a 1.5 mL cryotube containing RNAlater (ThermoFisher product no AM7020). Each tube was labelled according to the 12-hour clock position at which the sample was taken from the primary tumour. The samples were then stored at 4°C overnight before being transferred to a –80°C freezer, following the RNAlater tissue storage guidelines.

Whole blood was also collected to be used as a germline control at the time of anaesthetic induction, via a cannula into a purple 5mL EDTA tube. The entire tube was then frozen at –80°C.

2.2.2 Tissue processing

2.2.2.1 Tissue sectioning

Each sample was cut in half, with one half kept back at –80°C and the other thawed at room temperature. RNAlater lowers the freezing temperature of its contained sample and so could not be cut on a standard cryostat, as it would not have remained frozen. It was therefore washed with PBS twice to remove excess RNAlater.

The thawed washed sample was then stored on dry ice until ready to be sectioned using a cryostat.

A cryostat was used following standard operating procedures to section the tissue. Each tissue sample was sectioned to produce a slide for histopathological analysis either side of 200µm of tissue to be used for DNA and RNA extraction. Up to three 4µm sections were cut and placed onto a cooled glass slide. Then 200µm of tissue was sectioned using the cutter placed at 20µm and placed into a 1.5mL Eppendorf tube. Three further 4µm sections were then cut and placed onto a further slide. A further 200µm was then sectioned, with a final three 4µm sectioned for a final slide. These slides were created to ensure that the tumour content could be checked prior to nucleic acid extraction and sequencing. The interspersed slides were created to ensure that similarly high tumour cell content (> 80% was contained between these sections).

2.2.2.2 Tissue staining

The slides were fixed by submerging in acetone for 20 minutes at –20°C. Slides were then air dried while taking great care not to over-dry the fixed tissue specimens. Samples were then hydrated in tap water for 5 minutes. Slides were then immersed in haematoxylin for 3 minutes and checked for the required intensity. Slides were then gently rinsed in tap water until tissue stain turned blue. This was followed by immersion in eosin for one minute and then gentle rinsing with tap water. The slides were then dehydrated sequentially in 70% and then 100% IMS for 2 and then 4 minutes respectively. The slides were then immersed in histoclear three times for 2 minutes each time. Finally, the slides were mounted with DPX. The fixed mounted slides were then left to dry for 48 hours before use.

2.2.2.3 Histopathological confirmation of malignant tissue

For each region of tumour at least two slides of frozen tissue were stained for histopathological review. The tissue cut for each slide was cut either side of the tissue set aside for nucleic acid purification. Each slide was then reviewed by me and then in a blinded manner, reviewed by consultant uro-histopathologist Dr Alex Freeman, who has had many years' experience in reviewing pathological slides for penile squamous cell carcinomas. Each slide was scored based on the proportion of tumour tissue present in the following categories; normal, < 10%, 10%-50%, 50%-80%, 80%-90% and > 90% tumour content.

2.2.2.4 DNA purification – Extraction from fresh frozen tissue and whole blood

DNA was purified using the Qiagen DNeasy purification kit following the standard protocol. One aliquot of frozen 200µm tissue section was used for the purification. DNA was also purified from the whole blood collected as per the Qiagen protocol. The optional RNase step was carried out as per the standard Qiagen protocol.

2.2.2.5 RNA purification

RNA was extracted using the Qiagen microRNeasy kit following the standard protocol. Tissue was macerated after first re-submerging in RNAlater to reduce the chance of tissue degradation. A DNase step was utilised following the standard Qiagen protocol.

Adjacent histopathologically 'normal' skin was used as control samples in this experiment. Difficulties were encountered in obtaining high-quality RNA from these skin samples, as they required extensive mechanical and biochemical homogenisation to enable purification of mRNA. Although methods were employed to minimise the action of endogenous nucleases by homogenising at low temperatures and using RNAlater for storage of tissue samples, many of the samples were too degraded with RIN scores below 7 to proceed with capture and sequencing. This was not a problem for the tumour samples, as the homogenisation was much quicker, with high RNA content resulting in less time available for the endogenous nucleases to act. In only four of the eight PenHet cohort patients, was RNA able to be extracted at sufficient quality from the tissue-adjacent normal samples. In the case of the remaining four patients, the decision was therefore taken to substitute the tissue-adjacent normal samples – just for the RNA sequencing experiment – with samples collected from other Biobank patients with similarly advanced penile cancer.

2.2.2.6 Nucleic acid quality control

Nucleic acids were quantified using the Qubit RNA and DNA broad range kits. Quality of RNA was assessed by automatically calculating the RIN scores by running each sample on an Agilent RNA Nano 6000 Bioanalyzer kit. The quality of the purified DNA was assessed by running 10ng of DNA on a 0.7% agarose gel to check for a distinct band of high molecular weight DNA. DNA samples were also run on a DNA 1000 Agilent Bioanalyzer chip. Purity was also assessed using the 260/280nm and 260/230nm ratios calculated using the NanoDrop. For DNA, the purity requirements involved an absorption ratio of 260/280 between 1.8 and 2.0. For RNA, the minimum RIN score to pass the quality control was greater than or equal to 8.

2.2.3 EPIC methylation array

EPIC methylation 850k arrays were chosen as the method of choice for the intra-tumour methylome assessment for the PenHet cohort. Following a comparison of the different methylome assessments undertaken prior to the start of this project, the Illumina 850k methylation arrays were chosen as the method of choice. It was chosen because there were already established standardised processing and bioinformatic pathways to enable the comparative analysis in this cohort. Developing new methods to undertake an analysis of intra-tumour heterogeneity would have been beyond the scope of this study. Using the more easily analysable methylation arrays also meant that existing algorithms for detecting immune cell content could be performed. The costs involved in utilising whole genome or captured methylation sequencing were also beyond the scope of this project and the reduction in depth of sequencing would have reduced sensitivities of these methods. If there were no financial considerations when making this decision then the ideal method would have been whole genome bisulfite sequencing. The reason for this is that whole genome sequencing would have enabled better modelling of the methylation clonal and subclonal structure of individual tumours.

2.2.3.1 Bisulfite conversion

500ng of DNA was bisulfite converted as recommended by Illumina for the EPIC methylation Arrays. The DNA was converted using the EZ-96 DNA MagPrep Methylation kit from Zymo Research. The manufacturer’s instructions were precisely followed and consisted of denaturing and bisulfite converting the DNA overnight for 15 hours using the CT conversion reagent. DNA was then bound to the magnetic MagPrep beads. With the DNA bound the beads were then washed, desulphonated and washed again. They DNA was then eluted using 12mcl M-elution buffer.

The eluted DNA was loaded onto a 96 well plate for preparation of Illumina EPIC 850k methylation array. Samples belonging to a single patient were kept together on each slide. The array samples were further processed and loaded onto the methylation arrays by Mark Kristiansen at UCL Genomics, who processed them further as per the Illumina EPIC array processing guidelines. In summary, this process involves: whole genome amplifying, fragmentation, precipitation, resuspension, hybridisation, washing, staining and imaging the

DNA samples. The final output from this process were iDat files for each colour channel for each sample.

2.2.4 Whole exome sequencing

Whole exome sequencing was chosen to assess the mutation status of each sample as a compromise between cost and depth of sequencing. By limiting the analysis to the more thoroughly characterised exomes, a higher depth of sequencing could be achieved, increasing the sensitivity for detecting rare subclonal mutations. Leaving aside the increased costs of performing whole genome sequencing at sufficient depth, there still would have been a further problem in that the computational requirements to analyse the data would have been far higher, which would have significantly lengthened the time taken to perform the bioinformatics. Whole exome sequencing was therefore chosen as a compromise between depth of sequencing, coverage, cost and computational processing time.

Libraries were prepared for whole exome sequencing by using the Nimblegen seq-cap ez v3 capture kit. 1mcg of purified DNA was required for the Nimblegen capture kit at a concentration of 18ng/mcl in a volume of 55mcl. Following this the library preparation was undertaken externally by the Wellcome Oxford Genomics Centre, as this was the most cost and time efficient option available at the time. Quality control was undertaken prior to sequencing by assessing the quality of the prepared library using an Agilent high sensitivity DNA bioanalyzer chip. Both the quantity and average size were assessed.

Each library had an individual index and was sequenced across eight HiSeq 4000 75bp paired end lanes with a target coverage above 100x across the captured regions, as recommended by Illumina for clinical whole exome sequencing samples.

2.2.5 RNA sequencing

Multiple library preparation and RNA species selection methods exist to sequence RNA for expression studies. These include poly A reduction, ribo-depletion, and size selection. Size selection can be used to specifically select small RNA species such as micro RNAs, or small interfering piwi-interacting RNAs. Ribo-depletion and poly A selection can both be used for gene expression studies. Poly A selection selects only for transcripts with poly A tails, which include messenger RNA, pre-messenger RNA and ncRNAs. Ribo-depletion removes ribosomal RNA and therefore enriches for mRNA, pre-mRNA and ncRNA. Ribo-depletion is therefore useful in situations of RNA degradation as it will still be possible to select and sequence degraded fragments. However, this comes with an increased cost of library preparation and increased depth of sequencing required. As only pure non-degraded samples with a RIN score above 7 were selected the increased costs and potentially lower depth of sequencing using ribo-depletion was not deemed worthwhile, and so poly A selection method was used.

A choice between short and long reads was also made. Short reads of 50-100 base pairs in length have the advantage of increased read coverage and reduced error rate. On the other hand, long reads can be used for de novo transcriptome assembly, which is advantageous for resolving splice junctions and repetitive regions. For this project, short reads of 75 base pairs length were used.

Libraries were prepared for mRNA sequencing by using a poly A preparation directional capture kit. 1.5mcg of purified RNA was required for the capture kit at a concentration of 50ng/mcl in a volume of 30mcl. Following this, the library preparation was undertaken externally by the Wellcome Oxford Genomics Centre.

Quality control was undertaken prior to sequencing by assessing the quality of the prepared library using an Agilent high sensitivity RNA bioanalyzer chip. Both the quantity and average size was assessed.

Each library had an individual index and was sequenced across six HiSeq 4000 75bp paired end lanes with a minimum output of 30 million reads per sample.

2.3 Bioinformatics pipelines and data analysis

All bioinformatics was performed on the following three computing systems:

1. UCL Medical Genomics Server – 70 cores with 500GB of RAM
2. UCL Legion Computing Cluster
3. MacBook Pro 2GHz Intel Core i5. 16 GB 1867 MHz LPDDR3. Intel Iris Graphics 540 1536 MB

Default settings for all computational packages were used unless specified.

2.3.1 Sequencing quality control

2.3.1.1 FASTQC

Sequencing reads require a minimum base quality > 39 to pass the filter at the end of a sequencing run. The resulting fastq files are then assessed using FASTQC version 0.11.5[90]. The per base sequence quality scores are assessed across all reads for each sample on each lane to identify any problems with the sequencing chemistry or library quality. The base quality scores were then assessed across each tile to identify any spatial defects or biases. The mean sequence quality scores were also assessed to check what proportion of sequencing reads were at a high Phred score, e.g. 37. The proportion of each base and average QC for each read was assessed to check for any anomalies. An inspection was made for adaptor content to calculate where in the reads adaptors would be expected and observed. Finally, the proportion of duplicated reads was also checked. Duplicated reads were checked again downstream after alignment to the reference genome.

2.3.1.2 Qualimap

Depth and coverage was determined using the Linux programme Qualimap[91]. This programme can be downloaded from http://qualimap.bioinfo.cipf.es/.

2.3.2 Whole exome sequencing

This pipeline, from mapping/alignment onwards until variant calling, conforms to the respected and most cited standards and best practices published by the Broad Institute – Genome Analysis Toolkit Best Practice Guidelines, current as of January 2017. The guidelines are for the use of Mutect2 to call somatic variants on whole sequence DNA data where normal controls are available and sequenced under the same conditions. The general workflow is visualised in Figure 12.

[redacted]

Figure 12: Genome Analysis Toolkit Best Practices for somatic single nucleotide variant and insertion/deletion calling for whole genome and whole exome DNA sequencing.

2.3.2.1 Sequence alignment

After running through FASTQC and assessing that each sequencing run was successful, the reads were aligned using BWA mem to the human reference genome hs37d5. Hs37d5 is a combination of the human genome version b37 together with a list of 'decoy' genomic sequences derived from HuRef, human BAC and Fosmid clones, and NA12878 (named 'hs37d5'). In addition, the pseudo-autosomal regions (PAR) of chromosome Y have been masked out (replaced with 'N'), so that the respective regions in chromosome X may be treated as diploid. This is used for alignment of genomes within the 1000 genomes project[92]. The decoy genome was assembled and released by bioinformatician Heng Li. This new decoy reference has the beneficial properties of optimising read mapping and variant calling while still being compatible with other applications such as GATK b37 and IGV. This is because the read aligner can now easily align DNA which is not in the human reference genome, such as EBV, HSV 1 and 2 to its appropriate viral references instead of wasting computing resources trying to align to the human reference or even worse mapping incorrectly to human sequences.

BWA mem version 0.6.2 was utilised to map all the sequenced reads to the hs37d5 genome, using the default parameters. The unaligned reads were saved and used for further processing to discover whether any reads belong to viral genomes such as human papillomaviruses.

2.3.2.1.1 Alignment to viral human papillomavirus (HPV) genomes

2.3.2.1.1.1 Sequenced viral reads

BWA mem version 0.6.2 was utilised to map all the unmapped reads, from the alignment to the human genome, to the following HPV viral genomes 1, 2, 4, 5, 6, 7, 9, 10, 11, 16, 18, 26, 32, 34, 41, 48, 49, 50, 53, 60, 63, 88, 90, 92, 96, 101, 103 and 108. Of these the commonly occurring HPV species are the low-risk benign subtypes, 6 and 11, and the high risk oncogenic subtypes, 16 and 18. These viral genomes were obtained from the NCBI Entrez utilities after looking up each genome in the NCBI Nucleotide database https://www.ncbi.nlm.nih.gov/nucleotide/. The number of reads mapping to each of these viral genomes were counted. To give an approximate comparative qualitative estimation of viral load, the following calculation was undertaken:

Surrogate marker of viral load = Viral genome length in base pairs / (number of reads mapping x read length in base pairs)

This detection method relies on off-target capturing of human exome regions and is therefore a vast underestimate of the total viral load. It can, however, be a useful indicator for the presence of each HPV species, particularly across each primary tumour. However, the caveat is that it will also be influenced by the differing coverage, sequencing depth and integration status of the HPV species.

2.3.2.1.1.2 Sequenced concatemer HPV reads

HPV can be present either episomally, or integrated into the host human DNA, or both. Concatemer reads are reads which align on one end to one genome and on the other to an alternative genome. This indicates that the specific length of DNA being sequenced contains the presence of both genomes. In the context of HPV, a concatemer read indicates that there is viral integration into the host genome. As well as detecting sequenced viral reads it may also be useful to detect concatemer reads, which represent the point of integration of the HPV. The detection of concatemer reads gives an indication that the virus has integrated into the host DNA and presumably has a higher oncogenic potential. These reads can also be used to determine whether the virus preferentially integrates into specific sites within the genome. Furthermore, these sites can be compared with those from other studies in HPV derived cervical and oropharyngeal squamous cell carcinomas. In addition, each concatemer can be compared with other concatemers found in alternative regions of the same primary tumour. If the same integration site is found, this indicates that integration of the HPV viral was an early event consistent with taking place in an ancestor or clone preceding the formation of each region sampled. This is illustrated in Chapter 3.

Concatemers were discovered for each sample by creating a combined reference genome of human and HPV viral subtypes. Where one read of a pair mapped to the human genome and the other to a viral one, the discordant pair was selected and analysed further. The exact locations of both alignments were noted to compute the integration sites. This process was automated using the programme HPVdetector version 1.0[93].

2.3.2.2 Assessment of microsatellite instability

Microsatellite instability is a marker of genetic instability. The presence of microsatellite instability was determined by MSISensor [94]. MSI can be scored based on the proportion of microsatellites at each site shared across the tumour and normal. A score of more than 3.5 indicates high levels of MSI and in the context of colorectal cancer would constitute eligibility for treatment with a PDL-1 blocker, as recently approved by the FDA[95,96].

2.3.2.3 Genome Analysis Toolkit Mutect 2 Pipeline

Following mapping of the sequenced reads to the hs37d5 genome, each BAM file belonging to a specific sample was merged by concatenating the files. Duplicates were detected and marked by running each samples file through PicardTools, version 2.8.1 released by the Broad Institute[97]. BAM files belonging to the same biological sample were merged as the output. Each sample was then sorted and indexed in parallel using SamBamba v0.6.5[98].

PicardTools was used to add read groups specific to each sample as recommended by GATK using the function AddOrReplaceReadGroups.

Systemic biases apply when assigning the quality scores to each base during sequencing. If the quality scores can be improved to parallel the true empirical quality scores then the adjusted quality scores should be used. This is accomplished by a two-step process in the GATK toolkit called Base Quality Score Recalibration. The first step is to use machine learning to create a model of the variability of the quality scores taking into account the position along each read and the read groups. The second step is to apply this model to each quality score to recalibrate them to approach their true empirical value. This was accomplished by undertaking the

BaseRecalibrator function in GATK to create the model and the PrintReads function to apply the model and recalibrate the quality scores for each base.

2.3.2.3.1 Contamination estimation

Single nucleotide polymorphisms were then called together with Insertions and Deletions by running Mutect2 on the merged sorted base quality score recalibrated BAM files. Mutect2 was run with additional white list input from COSMIC (version 70) and recognised human SNP information from the Single Nucleotide Polymorphism Database (dbSNP build 132). The calling of variants was limited to 100bp surrounding each capture region. The capture regions were supplied by SeqCap EZ Exome v3 Nimblegen[99].

2.3.2.3.2 Variant calling

All variants were filtered using the default options in Mutect2 SelectVariants function to select all variants classified with 'PASS'. Variants were annotated using Annovar[100] version build (2016Feb01) including options for refGene, cytoBand, genomicSuperDups, esp6500siv2_all, 1000g2015aug_all, snp138, dbnsfp30a, cosmic70, exac03 and clinvar_20160302.

2.3.2.3.3 Variant filtering

Variants were further filtered to ensure each region of the primary tumour had a similar chance of detecting each variant. The options chosen were purposeful in conservatively estimating heterogeneity. The aim was to lower the threshold for a variant to be called in one region if also called in another. Each tumour region contains a mixture of 'normal' host cells, tumour cells, infiltrating immune cells and potentially genomes from infective agents/pathogens. Therefore, for the same amount of sequencing depth, across the different tumour regions, there is potentially an imbalance in the amount of sequencing of the specific target tumour cells. This has the potential to result in variant calls at low frequencies to be more easily called in 'purer' tumour samples with no contaminating normal, germline or pathogenic genetic material. It was therefore decided to create a dynamic cut-off for calling a variant depending on the predicted number of reads being sequenced from each pure tumour cell in each region.

The following criteria were used to create cut-offs to filter variants with the aim of conservatively calling a variant as shared. Similar cut-offs have previously been used by other authors in successfully undertaking multi-region DNA sequencing with validation of variants in the region of 93.5%[101]:

- Minimum variant allele frequency of ≥ 5% in at least one tumour region
- Minimum predicted number of tumour reads of ≥ 5 across each tumour region
- Not detected in artefact blacklist

Therefore, if the minimum predicted number of tumour reads in one region was < 5 then this variant was removed from all regions of this patient as it would not be possible to determine if the lack of variant detected was due to the lack of sequencing depth or from a true lack of biological variant in that region.

2.3.2.4 Annotation

Samples were annotated with Annovar[100] to annotate all SNVs, insertions and deletions, produced from the output of Mutect2. VCFs from Mutect2 were converted into Annovar compatible inputs by using the convert2annovar function. Annovar was run through the perl interface on the UCL medical genomics server. The following databases were utilised in the annotation:

- refGene
- cytoBand
- genomicSuperDups
- esp6500siv2_all
- 1000g2015aug
- snp138
- dbnsfp30a
- exac03
- clinvar_20160302

The following function was run to achieve the annotation:

```
perl ~/Scratch/progs/annovar/table_annovar.pl
~/Scratch/WES/annovar/merged/all.pseudohg19.avinput ~/Scratch/progs/annovar/humandb/ -
buildver hg19 -out ~/Scratch/WES/annovar/merged/all.avoutput -remove -protocol
refGene,cytoBand,genomicSuperDups,esp6500siv2_all,1000g2015aug_all,snp138,dbnsfp30a,cos
mic70,exac03,clinvar_20160302 -operation g,r,r,f,f,f,f,f,f,f –otherinfo
```

Due to the lack of significant evidence in previously published studies for DNA genetic drivers of penile cancer, additional methods were utilised to predict driver status and narrow down some of the outputs of all DNA, methylation and RNA aberrations. One method deployed involved determining whether the gene involved had previously been found to be aberrantly mutated, to have a change in copy number, change in methylation status or change in expression. The presence of the gene in COSMIC database[102] version 83 was used for this purpose.

2.3.2.5 Copy number calling

Sequenza[103] was used to calculate allele specific copy number events, and to estimate ploidy and tumour cellularity across all of the samples by using the mapped whole exome BAM files.

2.3.2.6 Mutation signatures

Alexandrov et al[104] demonstrated and validated 21 different mutational signatures based on the distinct proportions of each of these different mutation types. Using the package deconstructSigs in R, written by Rosenthal et al[105], the proportion and statistical significance for the presence of each signature was assessed throughout the PenHet cohort.

Mutation signatures were assessed for all filtered mutations in each sample across the PenHet cohort.

2.3.2.7 Determining HPV status

HPV status of each sample was determined by, firstly, the presence of HPV viral reads in sequencing data and, secondly, the presence of HPV integrated viral concatemer reads.
In addition, further evidence for HPV status was that HPV samples were the only ones to display mutational signature 2. It has been proposed that signature 2 and signature 13 are caused by activation of APOBEC enzymes, particularly APOBEC1, APOBEC3A and APOBEC3B, which are induced by HPV viral oncogenes E6 and E7[106].

2.3.2.8 Visualisations

The R package Maftools was used to convert the annovar output into a maf file and to calculate the combined frequency of all mutations across the cohort. Visualisations were also generated of the clusters of variant allele frequencies of all mutations per sample.

2.3.2.8.1 Heatmaps

Heatmaps were created using the ‘aheatmap’ function of the ‘NMF’ R package[107].

2.3.2.9 Cancer cell fraction and mutational copy number

The variant allele frequency is the proportion of alleles that carry a specific mutation. It can be easily calculated by number of sequenced reads with mutant/total reads sequenced.

However, as explained in Chapter 3, this calculation is not useful in determining the relative portion of cells that contain a given mutation. Instead, the cancer cell fraction (CCF) can be determined. This refers to the fraction of cancer cells that harbour a mutation. In order to calculate the CCF, one needs to take into account the purity of the cells being sequenced, in other words, cancer purity as well as the relative copy numbers of the normal and mutant cells. This was accomplished by providing the output from Sequenza of the mutant and normal copy number. The formula for calculating the CCF is as follows[108,109]:

$$\textit{Expected VAF} = p^{*}CCF\ /\ CPNnorm\ (1\text{-}p) + p^{*}CPNmut.$$

Where p = tumour cellularity, CCF = cancer cell fraction, CPNnorm = local copy number normal cells, CPNmut = local copy number mutant cells

The formula for calculating the mutant copy number is as follows:

$$\textit{Mutational copy number} = (VAF/p)*((p*CPNmut)+CPNnorm*(1\text{-}p))$$

The probability of a given CCF and thus of clonal status was determined using the same method as Landau et al [108]:

> “For a given mutation with ‘a’ alternative reads, and a depth of ‘N’, the probability of a given CCF can be estimated using a binomial distribution P(CCF) = binom(aIN, VAF(CCF)). CCF values can then be calculated over a uniform grid of 100 CCF values (0.01, 1) and subsequently normalized to obtain a posterior distribution. To avoid overestimating the number of clonal alterations, we classified mutations as clonal if there was >0.5 probability that the cancer cell fraction was >0.95.”[108]

2.3.2.10 Phylogenetic tree construction

2.3.2.10.1 Regional phylogenetic tree construction

Regional phylogenetic trees were constructed using the ratchet algorithm in the RPhylip package[110], considering all filtered mutations across the primary and metastatic samples for each patient individually. All mutations were converted to binary matrixes for the input to RPhylip.

Regional phylogenetic trees were also constructed using the Euclidian distance between all tumour and normal samples based on each sample's binary genomic aberrations. These trees were constructed using the fastme.bal function utilising the Euclidian distances calculated per sample found in the ape R package[111].

2.3.2.10.2 Clonal phylogenetic/riverplot construction

Clonal phylogenetic trees and riverplots were reconstructed using the R packages canopy[83], superFreq[112] and fishplot[113]. SuperFreq and canopy both model clones and subclones taking into account the cancer cell fraction independently for SNV, short INDELs and CNAs. Default settings were used throughout. The inputs for SuperFreq included the following:

- Capture regions file
- Aligned BAM files
- Variant calling from Mutect2
- dbSNP, COSMIC databases

The resulting plots were annotated with potential driver genes found by cross reference to the COSMIC database. This labelling and improvement to the graphical appearance of the plots was achieved through manual editing by using Sketch and Adobe Illustrator CS6.

2.3.3 Copy number aberration analysis

2.3.3.1 Allele specific copy number calling

Allele specific copy number calling was undertaking using the R package Sequenza[103]. Sequenza has comparable operator statistics in calling ploidy and cellularity as well established methods using SNP arrays and ASCAT algorithms[114] (r = 0.94 and r = 0.9 respectively). Sequenza uses matched tumour and normal whole exome sequencing data to calculate estimated cellularity and allele specific copy number information. Sequenza can also calculate the mutational copy number, which is useful in estimating relative timings of mutations and copy number changes, as discussed in Chapter 3.

2.3.3.2 Clustering of samples

Samples were clustered for the visualisation in heatmaps using the NMF R package. Samples were clustered based on the Euclidian distances between files when converting all aberrations to a binary input.

For clustering of copy number aberrations, the function from CNTools R package was used to convert multiple DNA copy number sequences into a reduced segment matrix of overlapping chromosomal regions. This ensures that all segments can be compared across samples to allow for further downstream analysis such as clustering.

The copy number frequency plots and circular aberration plots produced in Chapter 3 were produced by using the copynumber R package.

2.3.4 Methylome analysis

As described in Section 2.2.3 above, the initial bisulfite converted DNA was kindly processed by Mark Kristiansen at UCL Genomics, where he ran the Illumina EPIC methylation arrays on my behalf. He returned the output from the image processing, which contained the Illumina idat files. These files were processed primary using the functions in the R package called 'minfi'[115].

2.3.4.1 Normalisation

Samples were first normalised using the function preprocessNoob from the minfi package. This normalisation function uses 'the noob background subtraction method with dye-bias normalisation'. This method of normalisation was preferred as there was a danger that other forms of inter- or intra-array or slide-based normalisation would reduce the variance between or within primary tumour samples, thereby reducing the biological intra-tumour heterogeneity that exists.

2.3.4.2 Quality control

The minfi function densityPlot was utilised to visualise the raw and normalised density plots of beta methylation values of all CpGs within all samples. This was to ensure that all samples displayed the expected binomial distribution of methylation values.

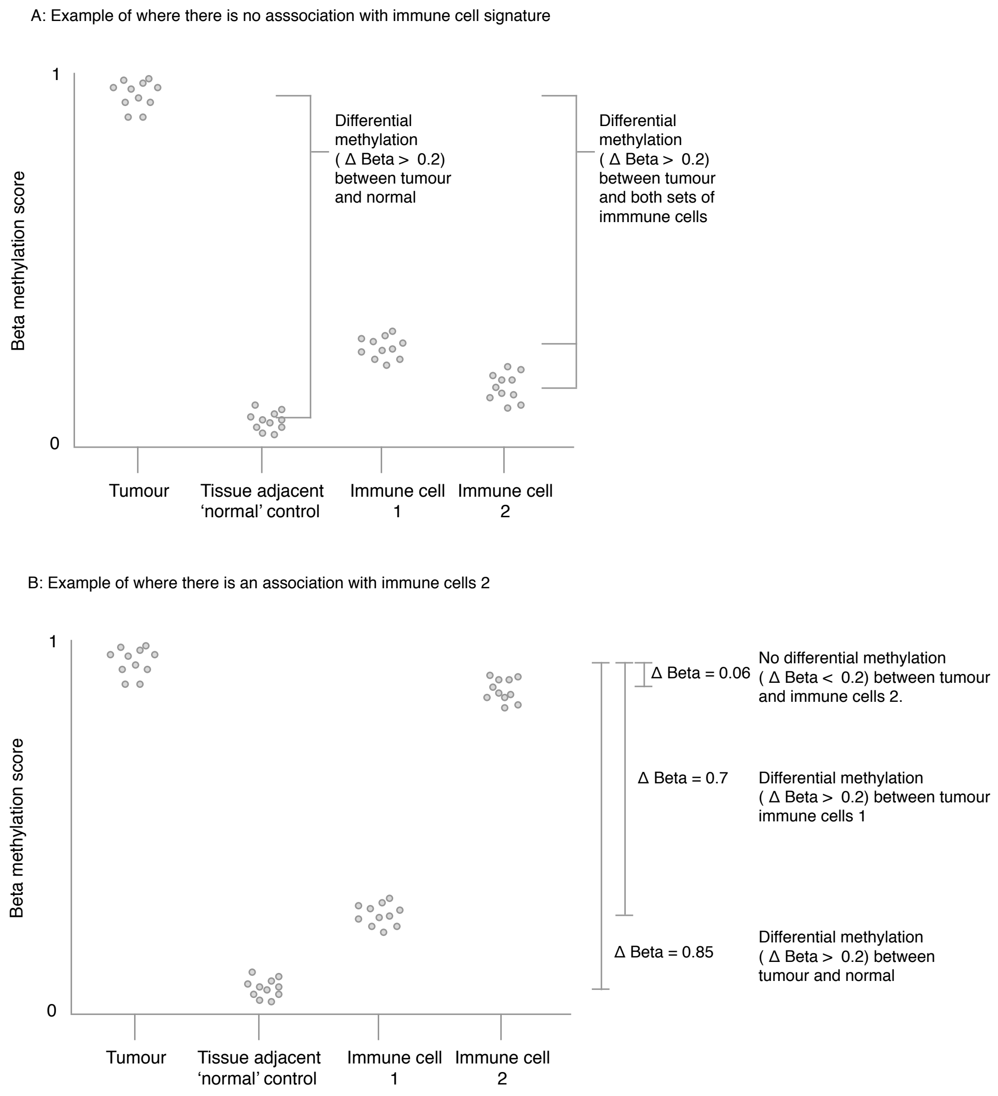
Each raw data file resulted in a beta methylation call at 866,838 CpG loci for each sample. These were then filtered in several steps to remove poorly performing probes. The first step involved removing all probes with a poor detection p value, which totalled 11,290 (1.3%), leaving 855,548 probes for downstream analysis. There is significant evidence that any CpG within five bases of a single nucleotide polymorphism (SNP) on a probe can result in a non-specific and thus inaccurate methylation signal[116]. Therefore, all instances of these as previously investigated by Zhou et al[116] were removed. These probes totalled 79,976 (9.3%) leaving 775,572 probes left for analysis after removing for poor detection p values and closeness to SNPs.

2.3.4.3 Immune contamination

To assess the extent of immune cell infiltration the methylation signatures of immune cells can be compared to that of tumour samples obtained across the PenHet cohort. This was achieved by combining the PenHet datasets with those of purified immune cells previously published. The methylation profile of these purified immune cells was assessed by undertaking 450k Illumina Methylation Array analysis. The resultant iDat files were obtained from GEO accession number GSE35069 and R package 'FlowSorted.Blood.450k'[117] and processed together using the same pipeline created for the EPIC arrays used in the PenHet cohort. Arrays from 450k and EPIC were combined using the minfi function 'combineArrays'.

The immune status of a DMP found in the PenHet cohort was assessed to determine if there was differential methylation between the tumour sample and all immune cell sorted samples. A DMP was labelled as associated with immune signature if there was not a minimum of 0.2 difference in beta methylation score between the tumour and all the immune cells. This is displayed diagrammatically in Figure 13.

All DMPs associated with the immune signature were removed from the filtered list of significant DMPs found between all of primary versus normal, tumour versus normal, lymph node metastasis and HPV datasets.



A: Example of where there is no asssociation with immune cell signature

B: Example of where there is an association with immune cells 2

Figure 13: Illustration of DMP filtering by immune signature status.

A: Depicts example where there is a tumour versus normal hypermethylation with a change in beta methylation of the tumour samples of at least 0.2. In this example, the tumour samples also have a change in beta methylation score of greater than 0.2 against both example immune cells. This demonstrates hypermethylation significantly above both tumour adjacent normal and immune cell populations.

B: Depicts example where there is a tumour versus normal hypermethylation with a change in beta methylation of the tumour samples of least 0.2. In this example, the tumour samples also have a change in beta methylation score of greater than 0.2 against the immune cell 1. However, there is no significant change (Δ Beta > 0.2) in methylation between the tumour cells and immune cell 2. This demonstrates that although hypermethylation exists between tumour and normal cells, there is a risk that at least part of this signal is driven by the immune cell 2 signature which may be infiltrating the bulk tissue sample.

2.3.4.4 MDS plots

Multi-dimensional scaling (MDS) plots are used to visually represent complex multidimensional distances in two dimensions[118]. This is accomplished by creating a matrix of differences of the M methylation value for each probe across every sample. The shortest distance between the matrix of values for each sample can be calculated. Using the most variable probes between samples. An MDS plot was created for the 1,000 most variable probes across the primary and normal control samples by using the 'mdsPlot' function within the minfi R software package[115].

2.3.4.5 DMP finder

Differentially methylated positions (DMPs) were initially assessed by using the DMPfinder function as part of the Minfi package[115]. This function performs an F-test to test for a statistically significant difference between the beta methylation values of tumour and control samples across a cohort.

As well as testing for a statistically significant difference, a filter with a minimum difference in methylation value of 20% was also utilised. A minimum difference filter was introduced to ensure that the methylation differences found were of a sufficient amplitude to conceivably result in a biological difference and furthermore increasing the chance that the methylation call was a true positive result[119], when taking into account sources of biological and technical variation. 20% was specifically chosen to improve the true positive detection rate at the expense of reducing the overall detection rate of DMPs. A study by Du et al[119] demonstrated that the true positive rate appears to hit a maximum level at around 20%. A Bonferroni multiple testing adjustment was made with an adjusted p value threshold of 0.01.

2.3.4.6 DMR finder

Differentially methylated regions (DMRs) can be defined as contiguous regions that differ between phenotypes[120]. DMRs were assessed using the DMRcate R package[121]. Significant differentially methylated CPGs found using the dmpfinder function, as described in Section 2.3.4.5, were used as the input to the 'dmrcate' function. The following settings were used:

- Minimum methylation difference = 0.2
- Number of cores = 2
- Lambda (minimum number of nucleotides for DMR separation) = 1,000
- Adjusted p value cut-off = 0.05

The DMR.plot function was utilised from the DMRcate package to visualise significant DMRs.

2.3.4.7 Methylation corroboration cohort

An additional cohort of 23 tumour samples with 15 controls was obtained from previously published work by Feber et al[44], as explained in Section 2.1.2. The raw iDAT 450k data files were re-analysed using the same methodology (including normalisation) as used for the PenHet cohort. The normalisation method 'preprocessNoob' as described above was used for all samples as it provides single sample normalisation based on background correction method with dye-bias normalisation. This enables comparison of samples from different batches and even different Illumina array types[122]. In addition, it reduced technical variation and improved classification across Illumina array types[122].

2.3.4.8 Lymph node normal external controls

The dataset of previously published normal pelvis lymph nodes was utilised as a control for the lymph node metastases. The iDAT files for these three samples were extracted from GEO accession GSE73549 and analysed in an identical manner to the internally produced datasets as explained in Section 2.3.4. In addition, the method minfi method 'preprocessNoob' was utilised for reasons explained in Section 2.3.4.7.

2.3.4.9 Methylation intra-tumour heterogeneity

2.3.4.9.1 Regional phylogenetic trees

A data table was created for each patient with rows signifying each CpG and columns signifying each sample belonging to that patient. Each cell contained a binary value 0 or 1 depending on whether there was differential methylation in either direction of at least 0.3.

Regional phylogenetic trees were produced by using the Euclidian distance between all samples for each patient based on each sample's binary methylation aberrations. These trees were constructed using the fastme.bal function utilising the Euclidian distances calculated per sample found in the 'ape' R package[111].

2.3.4.9.2 Assignment of trunk, shared and branch regions

In the context of methylation intra-tumour heterogeneity, a DMP can be considered a methylation aberration or epiMutant. In terms of phylogenetics for a DMP to exist in all current tumour cells it must have occurred at a time at or before the last common ancestor. In this study, I assessed the bulk analysis of four primary tumour regions. The fact that the DMP was found in four areas means that, even though it may not be present in the complete tumour, it is likely present in a large portion of cells and was therefore an earlier event in the history of the tumour's evolution. This is discussed in further detail along with potential exceptions to this logic in the Introduction (Chapter 1).

DMPs that were shared between all tumour samples were assigned to the trunk of the tree and can be considered truncal events. Truncal events are therefore also classified as likely 'early events'. DMPs that were shared among some of the cancer samples but not all, were classified as shared events. The remaining DMPs or methylation aberrations that were only found in one tumour sample were classified as 'branch' events. The branch and shared events can be considered later events in the phylogenetic oncogenic history of an individual patient's cancer.

2.3.4.9.3 Scoring of intra-tumour methylation heterogeneity

The extent of intra-tumour heterogeneity (ITH) can be scored by calculating the inverse of the fraction of DMPs which are present in the trunk of the phylogenetic trees created. This is displayed in the following equation:

$$ITH = 1 - n$$

Where n = the proportion of methylation changes that are truncal in origin.

ITH scores were compared across samples and groups of patients depending on HPV status.

2.3.4.10 Visualisations

2.3.4.10.1 Canonical gene plots

Canonical gene plots were created by annotating each CpG to both genomic feature (promoter, 5'UTR, 3'UTR, gene body, intergenic), and also CpG location eg, CpG island, shore and shelves. The beta methylation value of each tumour sample and adjacent control sample was then added to each superimposed ideogram. This was generated in R using code kindly shared by Dr Andrew Feber (UCL Medical Genomics). This code was further adapted to allow the differentiation between HPV positive and negative samples to be displayed.

2.3.4.10.2 Heatmaps

Heatmaps were created using the ‘aheatmap’ function of the ‘NMF’ R package[107].

2.3.5 RNA sequencing

All eight patients in the PenHet cohort, together with the additional four tissue-adjacent control samples, were included for RNA sequencing. Poly A selection RNA sequencing was chosen as the preferred method to balance the cost, depth and coverage of sequencing with a target of 100x across captured regions.

RNA-seq was performed on four regions of each primary tumour as well as a matched lymph node metastasis. Each library had an individual index and was sequenced across six HiSeq 4000 75bp paired end lanes with a target coverage above 100x across the captured regions by The Oxford Genomics Centre. The resulting fastq files were processed by me using a custom generated pipeline as described below.

FASTQC was utilised for basic quality control of sequencing as explained in Section 2.3.1.1.

2.3.5.1 Pseudo-alignment

RNA transcripts were quantified using the programme Kallisto[123]. Kallisto uses a pseudoalignment method to match sequenced reads to targets. Kallisto was used to align the reads to a ‘transcriptome’ i.e. a set of cDNA transcripts. A cDNA transcriptome for GRCh37 was obtained from Ensemble. The index of cDNA transcripts was created using the Kallisto function Kallisto index. The abundancies were then calculated using the Kallisto quant function using default parameters.

In order to bridge Kallisto output into an acceptable input into DESeq2, the tximport function was utilised from the ‘tximport’ R package [124], as recommended by DESeq2[125]. At this point transcripts were also combined at the gene level.

2.3.5.2 Differential expression

A differential expression analysis was subsequently performed using the recommended pipeline in the DESeq2 R package[125], comparing the relative expression levels, once normalised to length, between tumour versus normal, primary versus normal, primary versus metastatic tissue and HPV positive versus HPV negative samples.

Results were filtered by both a minimum log2 fold change of 2 and adjusted p value of 0.05. Adjusted p values were corrected using the Benjamini-Hochberg correction of a Wald Chi-Squared test for each gene. Additional analyses were also performed with a log2 fold change of 1, as specified in Chapter 5.

A Volcano plot was produced to visualise the proportion of genes that are significantly differentially expressed using the function in DESeq 2 package. Genes also present in the COSMIC database were highlighted on the plot.

2.3.5.3 MDS plot

Multi-dimensional scaling (MDS) plots are used to visually represent complex multi-dimensional distances in 2 dimensions[118]. This is accomplished by creating a matrix of differences of the log2 fold changes in expression for every sample. The shortest distance between the matrix of values for each sample can be calculated. Using the 1,000 most variably expressed genes between samples, an MDS plot was created across the primary and normal control samples by using the 'mdsPlot' function within the minfi R software package[115] in a similar way as for the methylation analysis in Section 2.3.4.4.

2.3.5.4 Calculation of immune infiltration by xCell

xCell[126] was used to estimate the relative proportions of immune and stroma content using default parameters. Spearman's rank was used to associate xCell scores with methylation immune contamination scores as well as tumour cellularity scores derived from the Sequenza package during whole exome sequencing.

xCell scores were broken down by immune cell type to ascertain the relative infiltration of differing immune cells within the PenHet RNA sequencing samples.

2.3.5.5 Intra-tumour heterogeneity (ITH)

2.3.5.5.1 Assignment of trunk, shared and branch expression events

In the context of expression intra-tumour heterogeneity, a differentially expressed gene can be considered an expression mutant. In terms of phylogenetics, for a differentially expressed gene to exist in all current tumour cells it must have occurred at a time at or before the last common ancestor. In this study, I assessed the bulk analysis of four primary tumour regions. The fact that the differentially expressed gene was found in four areas means that, even though it may not be present in the complete tumour, it is likely to be present in a large portion of cells and therefore was an earlier event in the history of the tumour's evolution. This is discussed in further detail along with potential exceptions to this logic in the introduction, Chapter 1.

Differentially expressed genes that were shared between all tumour samples were assigned to the trunk of the tree and can be considered truncal events. Truncal events are therefore also classified as likely 'early events'. Differentially expressed genes that were shared amongst some of the cancer samples but not all, were classified as 'shared events'. The remaining differentially expressed genes which were only found in one tumour sample were classified as 'branch' events. The branch and shared events can be considered later events in the phylogenetic oncogenic history of an individual patient's cancer.

2.3.5.5.2 Regional phylogenetic tree creation

A data table was created for each patient, with rows signifying each differentially expressed gene and columns signifying each sample belonging to that patient. Each cell contained a binary value 0 or 1 depending on whether there was significant differential gene expression (defined by log2 fold change comparing tumour to normal of at least positive or negative 1).

Regional phylogenetic trees were produced by using the Euclidian distance between all samples for each patient based on each sample's binary methylation aberrations. These trees were constructed using the fastme.bal function utilising the Euclidian distances calculated per sample found in the 'ape' R package[111].

2.3.5.5.3 Scoring

The extent of ITH can be scored by calculating the inverse of the fraction of differentially expressed genes which are present in the trunk of the phylogenetic trees created. This is displayed in the following equation:

$$ITH = 1 - n$$

Where n = the proportion of differentially expressed genes that are truncal in origin.

ITH scores were compared across samples and groups of patients depending on HPV status.

2.3.5.6 RNA expression corroboration in an independent external cohort

Significant genes uncovered during the differential expression analysis with DESeq2 were compared with an independently published external cohort of penile squamous cell carcinoma samples[89]. This cohort is described in Section 2.1.3. The core dataset was downloaded from GEO accession number GSE57955. The study by Marchi et al utilised Agilent aCGH expression arrays and so were analysed using a different pipeline, as published in the Marchi et al paper[89]. In summary, differentially expressed genes were evaluated using R package 'limma' as described in the published methods. Genes were filtered and deemed significant if the adjusted p value of was < 0.01.

2.3.5.7 Integration of RNA and DNA mutations

Datasets from the RNAseq and whole exome studies were integrated using the R package xseq[127]. Xseq utilises a hierarchical Bayes statistical model to systematically quantify the impact of somatic mutations on expression profiles. The model was run in 'cis' mode assessing the impact of mutations and copy number changes within a specific gene on the expression of that gene itself. Matrices of annotated mutation data were derived from the output of Annovar (Section 2.3.2.4) together with a gene expression matrix derived from Sequenza (Section 2.3.3). The model was run with default parameters. Potential genetic drivers were ranked by probabilities and analysed in Chapter 5.

2.3.5.8 Integration of RNA and copy number aberrations

The copy number aberrations generated in Chapter 3 using Sequenza (Section 2.3.3) were integrated with the RNA sequencing data by the use of R package 'iGC'[128]. Each gene overlapping a region of copy number aberration was paired with cis gene expression data. Student's t-test with unequal variance was used to identify differentially expressed genes ($p < 0.001$) that were significantly associated with CNA.

2.3.5.9 Integration of RNA and differential methylation

All statistically significant differentially methylated positions were integrated with the RNA sequencing data in this chapter to evaluate any associations between methylation changes and changes in gene expression. For this analysis data from primary versus tissue-adjacent normal samples were utilised.

A comparison of log2 fold changes at different differentially methylated positions (DMPs) was undertaken. All DMPs were grouped via the CpG location into promoters, 5'UTR, Body, 1st exon and 3'UTR.

Potential methylation gene drivers were assessed using the 'MethylMix'[129,130] R package. MethylMix identifies differential and functional DNA methylation by using a beta mixture model to identify subpopulations of samples with different DNA methylation compared with normal tissue. Functional DNA methylation is predicted based on correlations with matched gene expression data. The input for MethylMix included a matrix of tumour beta methylation values, a matrix of normal control methylation values and an expression matrix of log2 fold changes for each cancer sample.

2.3.5.10 Comparison of genetic, methylation and expression phylogenetic trees

Regional phylogenetic trees were compared using the Mantel test of significance of correlation between distance matrices[131] for each of mutation, copy number, methylation and expression aberrations. The level of significance was displayed below each phylogenetic tree with a colour coded p value. The Mantel test and level of significance for each tree comparison calculated by the Mantel test function in the R package 'cultevo' using Spearman's method. The function was run with 10,000 permutations. The significance testing is accomplished by generating a null distribution by randomly shuffling the matrix rows and columns and comparing with the supplied dataset.

2.3.6 Gene set enrichment analysis

2.3.6.1 Methylation arrays

Gene set enrichment was undertaken based on the filtered list of significantly differentially methylated CpG probes. The 'gometh' function from the missMethyl R package[132] was used, which modifies the previously published goseq method by Young et al[133]. The gometh function solves the problem of accounting for the inherent bias in the wide variance in the number of CpG probes per gene. The gometh function takes as input a vector of differentially methylated CpG probes and calculates the probability of an associated gene being selected taking into account the number of probes available for each gene. It then performs Wallenius' non-central hypergeometric distribution for each GO and KEGG category.

2.3.6.2 RNA expression

Gene set enrichment was undertaken based on the filtered list of differentially expressed genes generated by DESeq2. The R package 'gage'[134] was used to determine enriched GO and KEGG pathways using log2 fold changes of each significantly differentially expressed gene and calculates significance based on parametric gene randomisation[134].

2.3.6.3 Visualisation of KEGG pathways

The resulting enriched KEGG categories were visualised using the pathway figures generated from the 'pathview' R library[135].

3 DNA mutation and copy number

3.1 Introduction

In this chapter, the genetic landscape of penile squamous cell carcinomas is described in the context of recurrently mutated genes, mutational signatures, copy number changes and the presence of human papillomavirus. Following this, an analysis of intra-tumour heterogeneity is presented elucidating the genes and pathways which are mutated across all regions sequenced in the tumour, compared with those defects which are not shared or are even unique to a particular region. Variants can be classified in terms of whether they are conserved across all regions; present in multiple regions or are solely found in one region examined. These categories have been previously named[136] truncal, shared and private respectively. This information can be combined with copy number to predict the phylogenetic history of the tumour and predict the clonal versus subclonal origin of each variant. Clonal/truncal variants are likely to have existed at the time of the last common ancestor. This type of analysis is therefore powerful in providing evidence for the variants present at the time of this last common ancestor, which are thought to be vital in driving the initial oncogenesis. Subclonal variants, which are more likely to be of the shared and private type, only exist in a portion of cancer cells, meaning they came into existence at a later date. In addition, this information is clinically relevant, as discussed in Chapter 1, because oncological targeted therapies (e.g. cetuximab an EGFR inhibitor) that target specific genomic aberrations may have the greatest efficacy when the targeted aberration is clonal and therefore exists in all cells.

In order to assess the genetic heterogeneity associated with penile cancer, whole exome sequencing (WES) was performed. This allowed me to gain a broad understanding of the genetic substitutions, inclusions, deletions and copy number events that may be driving penile cancer. These events were examined both across the patient cohort and within the multiple surgical samples extracted from each patient at a significantly higher depth (approximately 100x) of sequencing than previously performed in penile cancer (60x)[137].

WES was performed on four regions of each primary tumour as well as a matched lymph node metastasis. All samples from each patient were surgically removed and processed simultaneously as described in the methods in Chapter 2. A matched germline blood sample was also taken at the same time and used as a germline reference.

All eight patients in the PenHet cohort, described in the methods, were included for whole exome sequencing. WES was chosen as the preferred method to balance the cost, depth and coverage of sequencing with a targeted depth of 100x. Further details on the depth and breadth of sequencing can be found in Chapter 2.

3.2 Sequencing output

The quality of sequencing output determines the substrate for all further bioinformatics analysis discussed below. Library preparation and sequencing methods can be found in Chapter 2.

High-quality sequencing output was obtained with a mean sequencing depth of 107x from a mean of 128.3 million 75 base pair reads per sample. 100% of these passed filter (base quality > 39). These reads were mapped to hs37D5 with a mean mapping percentage of 99.0%. Further information regarding quality control and mapping can be found in Chapter 2. Full sequencing output statistics per sample can be found in the Appendix.

3.3 Tumour cellularity

Tumour cellularity or purity is the percentage of cells in a bulk tissue sample that are tumour in origin as opposed to surrounding non-transformed ‘normal cells’ or infiltrating immune cells. This calculation is important as it determines the sensitivity of detecting tumour derived molecular aberrations. The lower the tumour purity, the more sequencing depth is required to accurately detect a molecular aberration such as an SNV.

Tumour cellularity can be determined by both histopathological assessment by using haematoxylin and eosin staining or by the sequencing data itself as explained in Chapter 2. Tumour cellularity was first assessed histopathologically to determine if sufficient tumour was available for sequencing and then by sequencing based methods as demonstrated below.

3.3.1 Histopathological tumour cellularity

All samples were processed as described in Chapter 2, with sectioning on a cryostat and staining with haematoxylin and eosin either side of sections cut for nucleic acid purification. Tumour cellularity of each sample based on haematoxylin and eosin staining was assessed in conjunction with Dr Alex Freeman, consultant histopathologist at University College London Hospital, who has more than ten years' experience of reviewing penile carcinoma specimens. All primary tumour samples had a tumour purity of > 80%. One lymph node metastasis contained a reduced tumour purity consisting of 50% necrotic tissue, 30% tumour cells, and 20% normal lymphoid tissue. Individual section histopathological data is displayed in Chapter 2.

3.3.2 Sequencing based estimation of tumour purity

Accurate estimation of tumour purity is vitally important in determining the cancer cell fraction of a mutation and therefore its clonality, as discussed in the Methods (Chapter 2). The relative number of reads mapping to a specific genomic position (depth ratio) was calculated together with the proportion of each group of sequencing reads at germline heterozygous positions. This information was calculated using the Sequenza algorithm[103] to elucidate the allele specific copy number aberrations, ploidy and cellularity across the multiple regions sequenced for each patient. Table 4 displays the tumour purities for all primary and metastatic lymph node samples as calculated using this sequencing based method in Sequenza. The tumour cellularity as determined by this sequencing based method was on average 41% lower ($p < 0.0001$ Wilcoxon Signed-Rank Test) than the cellularity confirmed on histopathological review. This was despite stringent inclusion criteria of more than 80% tumour content required for tissue section selection. This has previously been demonstrated in many other tumour sequencing experiments and is likely due to the difficulty and subjective nature of calling the cellularity based on one slide of tissue, as well as the difficulty of subtracting the proportion of infiltrating immune cells[138]. The tumour purity of each tumour region is shown in Table 4. Despite the resulting purity being significantly lower than predicted by histopathology, the tumour content is still sufficient to reliably call variants at allele frequencies of 5% with a true positive rate of > 93%[101]. This is accomplished by combining the variant calls from all tumour regions to improve the true positive calling rate as discussed in Chapter 2 and previously published by Kim et al[101]. Based on an analysis of TCGA tumour purities for lung adenocarcinoma, using the same methods as used in this thesis, the average reported tumour purities of submitted samples was 42% compared with 48% achieved in this cohort[114]. If this experiment were to be expanded upon in the future, laser dissection of tumour cells or single cell sequencing could be utilised to minimise the contamination from stromal and immune cells. This is discussed further in the discussion, Chapter 6.

Table 4: Table of percentage tumour purities for each sample of each patient calculated using Sequenza. Each tumour section is depicted by the prefix T. LN1 refers to the metastatic lymph node. Further demographic information on each of these patients is given in the Methods section in Chapter 2.

Patient number	**T1**	**T2**	**T3**	**T4**	**LN1**
39	0.42	0.44	0.50	0.40	0.16
45	0.46	0.61	0.26	0.39	0.21
49	0.40	0.43	0.60	0.51	0.43
51	0.70	0.80	0.74	0.83	0.47
63	0.81	0.82	0.30	0.65	0.32
64	0.36	0.80	0.60	0.54	0.66
66	0.41	0.51	0.53	0.58	0.16
79	0.20	0.46	0.41	0.17	0.37

3.4 Variants identified by whole exome sequencing

Tumour mutational load has been demonstrated as a potential biomarker in immunotherapy treatments, particularly immune checkpoint blockade (for example, in ipilimumab[139]). It is also associated with neoantigen expression, which itself has also been associated with immunotherapy response[140].

The total number of single nucleotide variants were therefore quantified for each tumour sample and compared to give an overall appreciation for inter-patient and intra-tumour differences in mutation rates. This was accomplished by simple quantification of all non-synonymous variants after applying the default parameters and filters in Mutect2 (see Chapter 2). All variants assigned by Mutect2 filter as 'PASSED' were used, producing a median number of variants of 90.5 with a range of 40-471 variants across all samples. The distribution of variants is displayed in Figure 14.

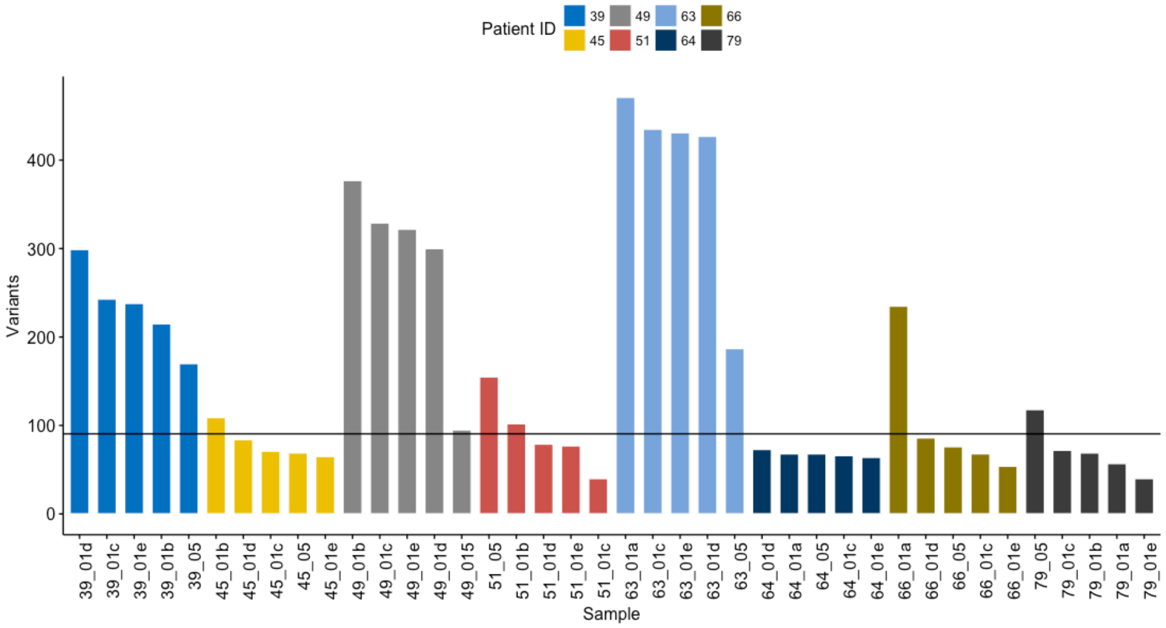


Figure 14: Number of variants per tumour sample sorted by patient ID. Samples ending in 05 indicate a lymph node metastasis. Horizontal reference line is at the median number of variants across the entire cohort.

These results were also used to determine the tumour mutational load. The mutational load was calculated to be 1.4/megabase(Mb) with a range of 0.625-7.36 mutations/Mb[141]. Although this mutational load is higher than that seen for many cancers it is lower than melanoma, lung and bladder cancer, where high mutational loads are often seen and correlated with response to immunotherapies. This data suggests that, assuming other factors are equal, immunotherapy should be considered as a new treatment modality for penile cancer. Chapter 5 will evaluate the relative expression levels of immune checkpoint inhibitors to give further evidence for the consideration of immunotherapy usage in penile cancer. A comparison of total mutations found across a range of 34 cancers can be seen in Figure 15. Although the two patients with the highest number of variants both come from samples infected with HPV (patients 49 and 63 as demonstrated below) there is not enough evidence here to demonstrate a correlation with mutational load based on so few patients. Indeed, the patient with the next highest mutational burden is 39, who is HPV negative. This data may suggest that there is a trend that HPV positive samples may have very high mutational loads of > 5 mutations/Mb. Further work is needed to assess this across a large cohort of patients. Later in this chapter an association between HPV status and APOBEC mutational signatures is presented in Figure 30.

The high variability between patients in the number of SNVs per sample was investigated further to determine if this was associated with differences in tumour cellularity. This has previously been explored by Aran et al[142], where no correlation was found between tumour purity and mutational load. The results from the PenHet cohort analysed in this chapter corroborate this with no correlation found using Pearson's correlation (R2 = 0.0022, p = 0.36).

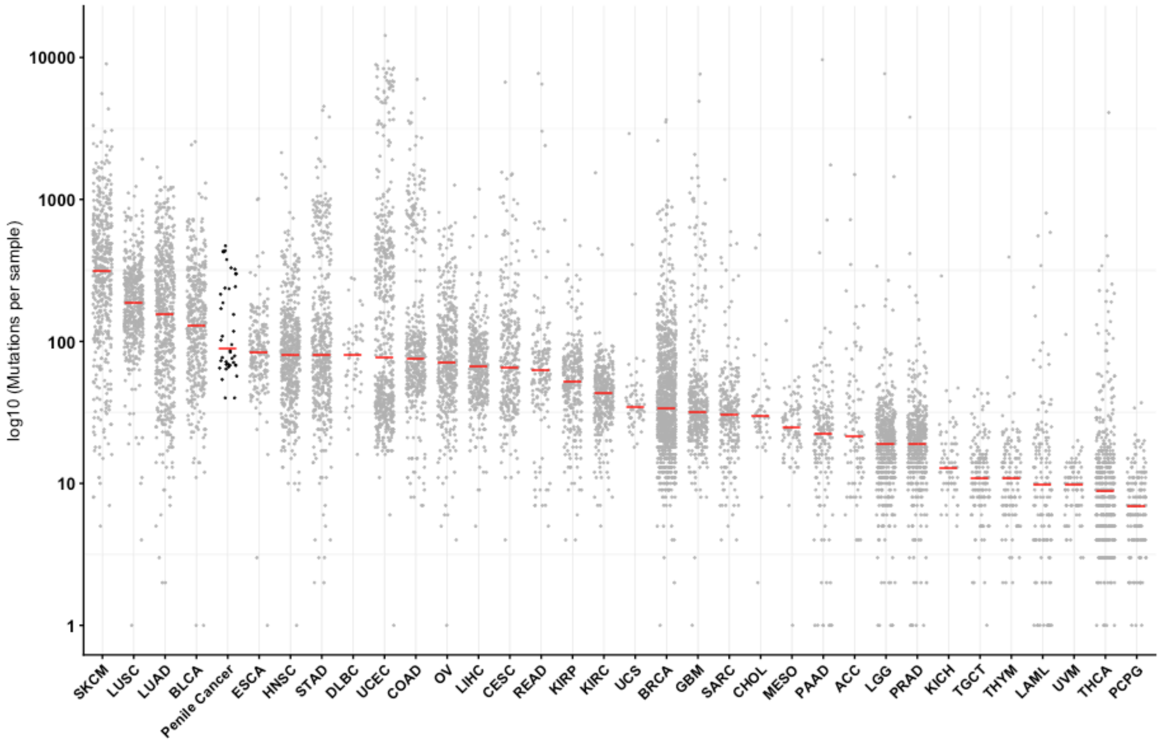


Figure 15: Total number of mutations found across 34 cancers extracted from TCGA data. The penile cancer cohort represents data solely from the results of this experiment.

Previously, Feber et al 2016[137] performed single sample tumour sequencing from a cohort of 27 patients at a depth of 60x revealing a paucity of mutations with a mean somatic mutation rate of 30 per tumour. The likely cause for the difference in mutational load can be explained by the relative lack of power in detecting mutations at only 60x compared with 100x, where tumour purity may have a large effect on the sensitivity of mutation detection.

3.5 Microsatellite instability

Microsatellite instability (MSI) is a further biomarker for response to immunotherapy[143] and potential cause of high mutational loads[143]. MSIsensor[94] was used to assess MSI status within the PenHet cohort. MSI can be scored based on the proportion of microsatellites at each site shared across the tumour and normal. This has become important clinically, as a high proportion of MSI would constitute eligibility for treatment with a PD-1 inhibitor (Pembrolizumab), as approved by the FDA in 2017[96]. The FDA approval in this context was the first pan-cancer approval of a molecular oncological therapy. It is hoped that this will enable new therapeutic medicines to be available for rare cancers, such as penile caner, on the basis of molecular aberrations alone despite the lack of a dedicated clinical trial for a particular rare cancer type.

None of the samples in the PenHet cohort displayed microsatellite instability. This is scored by MSIsensor with a proportion of microsatellite less than 10% being considered microsatellite stable. Microsatellites were not detected in 15 out of 40 samples and were detected at low levels in 25 out of 40 patients. One sample belonging to patient 39 had microsatellite percentage of 1.59%, which at less than 10% is still considered stable. The remaining 24 all had percentages of less than 1%. The median percentage of somatic microsatellite instability was 0.04 IQR (0-0.18). These results are displayed in Table 5.

Table 5: Table displaying the percentage of microsatellites detected for each sample. Percentages < 10% indicate microsatellite stability. Columns headed T1 to T4 represent primary tumour regions. LN represents the matched lymph node metastasis.

Patient ID	T1 (%)	T2 (%)	T3 (%)	T4 (%)	LN (%)
39	0.73	0.70	1.59	0.31	0.04
45	0.24	0.25	0.00	0.15	0.00
49	0.00	0.00	0.04	0.13	0.00
51	0.03	0.00	0.00	0.14	0.00
63	0.45	0.53	0.00	0.16	0.00
64	0.02	0.26	0.16	0.07	0.28
66	0.00	0.00	0.00	0.06	0.00
79	0.04	0.10	0.04	0.02	0.00

3.6 Human papillomavirus (HPV)

The strong association between penile squamous cell carcinoma and human papillomavirus (HPV) has long been recognised. There is a clear cohort of patients who progress from pre-malignant penile carcinoma in situ or penile intraepithelial neoplasia (PeIN) to penile squamous cell carcinoma. Chronic infection with HPV 16 appears to drive the formation of HPV associated Penile Intraepithelial Neoplasia (PeIN)[144]. Previous work from Feber et al[44] demonstrated that patients' methylomes cluster clearly into two distinct groups depending on HPV status. This provides evidence that HPV is at minimum a disruptive passenger and is likely a key driver for approximately half of all penile cancer patients.

In order to assess the presence and potential integration of HPV and identify which HPV subtypes are present, sequencing reads which were not mapped to the human genome assembly (version b37) were re-mapped to a large range of HPV genomes including HPV subtypes 1, 2, 4, 5, 6, 7, 9, 10, 11, 16, 18, 26, 32, 34, 41, 48, 49, 50, 53, 60, 63, 88, 90, 92, 96, 101, 103 and 108.

Only the presence of HPV 16 was detected in this cohort. The number of viral sequencing reads mapping to HPV 16 for each sample is shown in Table 6. As whole exome capture was performed, viral DNA would only have been captured if either it had formed concatemers with human DNA fragments, due to integrating into the human genome, or as part of an off-target effect of the capture probes. This means that whatever is captured will potentially only represent a subset of the total HPV burden in these cancer genomes. To investigate this further whole genome sequencing of these samples would have to be performed to determine all integrations within the genome.

Table 6: Number of sequencing reads aligning to HPV 16 genome for each sample of the PenHet cohort.

Patient number	**T1**	**T2**	**T3**	**T4**	**LN1**	**WB**
39	0	0	0	0	0	0
45	0	0	0	0	0	0
49	40	12	62	39	66	2
51	42	21	34	117	53	2
63	1627	5454	1910	936	924	18
64	28	19	20	14	4	28
66	29	13	24	8	8	14
79	1059	700	736	1360	1621	10

The total number of viral genome equivalents captured was calculated by multiplying the number of reads by read length and then dividing by the length of HPV 16 viral genome (7909 bp). The distribution is displayed in Figure 16.

The presence of concatemeric sequencing reads, which map to both human and viral genomes, was also evaluated and identified in 50% (four out of eight) of the patients investigated. These four patients have the highest number of reads mapping to the HPV 16 genome. As expected these concatemeric reads only mapped to the oncogenic 16 subtype. Figure 17 displays the number of concatemer reads and Figure 18 the most common integration locations. The quantity of concatemers is theoretically based on location of capture probes and affinity of capture probes to HPV. Therefore there could be instances in which HPV would integrate with tandem copies. As such, when HPV genomes are in tandem, the concatemeric reads would be less frequent and the viral reads may appear episomal in origin, when they are in fact fully integrated. Therefore, the total quantity of concatemer reads is a vast under-representation of reality but is still useful to assess relative and qualitative differences. The presence of concatemers is a much more significant finding than that of viral DNA itself, as in those patients with concatemers present the oncogenic HPV 16 has integrated into the genome as opposed to being solely episomal in origin. There is significant evidence for the oncogenic effect only taking place when viral integration has occurred[145].

The location of HPV viral integration was compared with a previous meta-analysis of HPV integration sites in cervical and head and neck cancer[93,146]. Thirty-two out of 43 HPV 16 integration sites (Figure 8) in the PenHet cohort were also recurrent integration sites in cervical and oro-pharyngeal cancer. The expected probability of a random integration event falling within one of these previously reported sites of the genome is 0.52, whereas the observed proportion of events was 0.744. A binomial test was therefore calculated to assess whether this was greater than expected taking into account the number of observations. The p value of this test was 0.003, confirming that HPV 16 in the PenHet cohort integrated more frequently than expected into the same locations as in cervical and oro-pharyngeal cancers. This result corroborates the findings in cervical and oro-pharyngeal cancer that specific sites in the human genome are more susceptible to viral integration than others. Although preferential integration sites – such as enhancers, promotors, fragile sites – appear to exist for HPV within cancer genomes, this may represent a selection bias[147]. To uncover the full repertoire, timing and preferential sites for integration it would be beneficial to undertake whole genome sequencing at multiple time points during normal, dysplastic/pre-malignant and then malignant tissue.

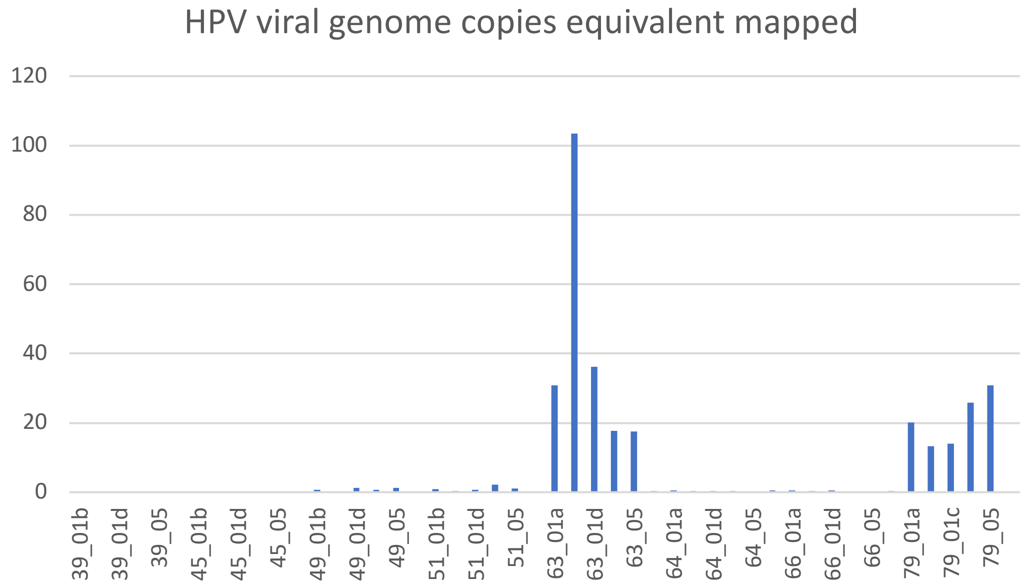


Figure 16: Equivalent HPV viral genome copies mapped from whole exome sequencing reads.

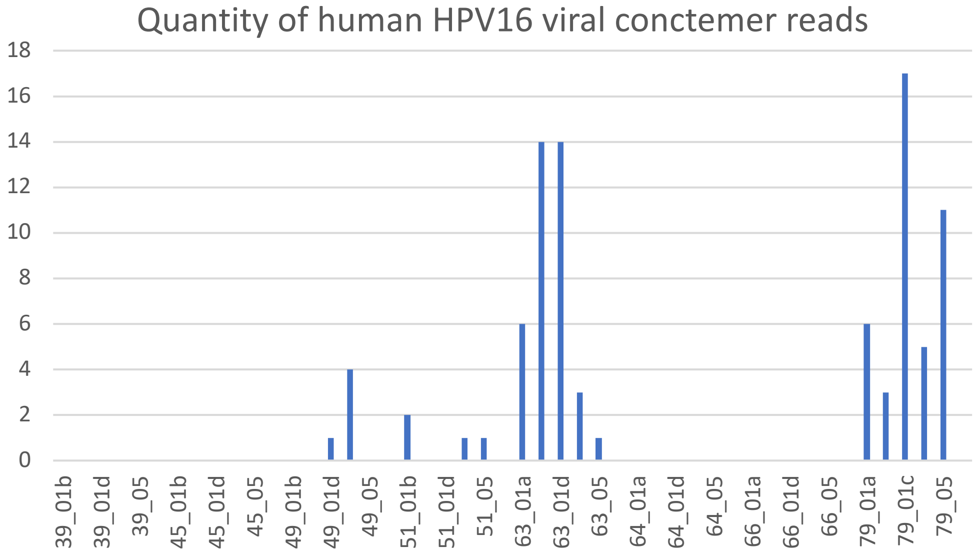


Figure 17: Quantity of concatemers sequencing reads found which map on one part to the human genome and for the other part to the HPV 16 viral genome.

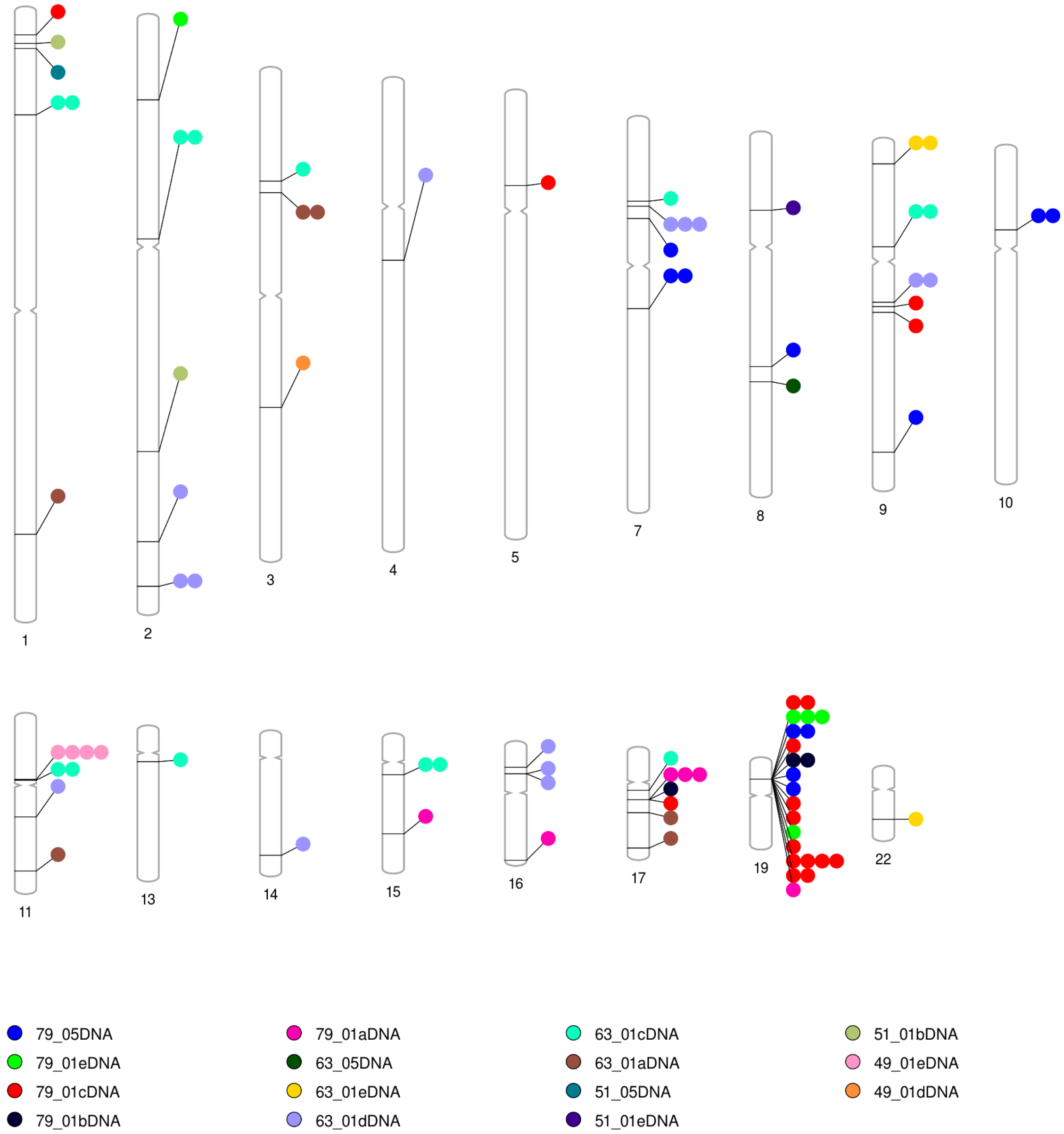


Figure 18: Chromosomal sites of HPV integration based on whole exome sequencing of captured DNA fragments, mapping to HPV 16 genome.

This division between the patients where HPV 16 has integrated into the genome (patients 49, 51, 63, 79) and those where HPV does not integrate (patients 39, 45, 64, 66) can clearly be seen throughout this thesis. This division is apparent when assessing and clustering by mutation signature as well as by differential expression. I will keep referring back to these two groups (HPV positive and HPV negative) as there appears to be very distinct onco-genomic profiles, which may represent a different disease process warranting personalised management and treatment options.

It has previously been suggested that HPV integration is an initiator/driver event in the development of penile cancer. We therefore sought to assess whether HPV integration, at least at the sites identified, is a clonal event. Only a single patient (patient 79), harboured a large number of concatemer reads, which appear to show integration at the exact same loci on chromosome 19 throughout the primary and lymph node metastasis. Therefore, in this case there is evidence to suggest that HPV integration was an early, shared, truncal event which occurred prior to the formation of the later clones. This phenomenon is not detected in the other three HPV positive patients. One clear potential explanation for this is that this experiment is based on capturing of specific genomic locations (exons) utilising capture probes of specific targets of interest. Therefore, the HPV genomic integration sites displayed in Figure 18, are dependent upon HPV integrating into an exon that is sufficiently captured and sequenced. The other three patients with HPV infection may have an initial clonal integration event just like in patient 79 into a location that is not captured by this whole exome sequencing capture experiment. Patient 79 had the highest viral load detected in this capture experiments but it is not the extent of viral load integrated into other locations of the genome cannot be determined in this experiment. This may be explored at a later date utilising whole genome sequencing to better understand the timings and location of HPV infection and integration into the human genome.

The specific integration sites were assessed in further detail by examining which genes they integrated into. These results are displayed in Table 7. The common integration site used in all sequenced samples of patient 79 is within the gene *NFIX*. Interestingly, this site has also been previously reported as a site of HPV 16 integration in a cervical squamous cell carcinoma cell line [148]. The precise relevance of this is beyond the scope of this thesis but I note that *NFIX* has been previously described as a regulator in both oesophageal[149], lung carcinomas[150] and hepatocellular carcinomas[151].

Table 7: Table displaying all genomic integration sites from concatemer HPV reads found using whole exome sequencing

Chromosome	Genomic position	Sample	Gene symbol
10	32974965	79_05	CCDC7
10	32974965	79_05	CCDC7
7	40134318	79_05	CDK13
7	77041598	79_05	GSAP
19	13122455	79_05	NFIX
19	13122531	79_05	NFIX
19	13122492	79_05	NFIX
r9	127361725	79_05	NR6A1
8	94797596	79_05	TMEM67
2	33141569	79_01e	LINC00486
19	13122569	79_01e	NFIX
19	13122450	79_01e	NFIX
5	37246032	79_01c	C5orf42
9	67330767	79_01c	Intergenic
9	69721144	79_01c	Intergenic
19	13125264	79_01c	NFIX
19	13125275	79_01c	NFIX
19	13125251	79_01c	NFIX
19	13122551	79_01c	NFIX
19	13122456	79_01c	NFIX
19	13122535	79_01c	NFIX
17	37828404	79_01c	PGAP3
19	13060249	79_01c	RAD23A
1	9642265	79_01c	SLC25A33
19	13122466	79_01b	NFIX
17	37828377	79_01b	PGAP3
15	75119298	79_01a	CPLX3
16	90162198	79_01a	Intergenic
19	13125312	79_01a	NFIX
17	37828137	79_01a	PGAP3
8	100855686	63_05	VPS13B
9	8516895	63_01e	PTPRD
22	38150819	63_01e	TRIOBP
4	73148194	63_01d	ADAMTS3
7	35057523	63_01d	DPY19L1
2	233445835	63_01d	EIF4E2
16	16484535	63_01d	Intergenic
14	94836263	63_01d	Intergenic
16	21851769	63_01d	Intergenic
9	65662704	63_01d	LOC286297
11	78277315	63_01d	NARS2
16	21419318	63_01d	NPIPB4
16	21419318	63_01d	SMG1P3
2	215176333	63_01d	SPAG16
7	32981972	63_01c	AVL9
1	42414562	63_01c	HIVEP3
2	90499347	63_01c	Intergenic
13	24515641	63_01c	Intergenic
15	28777993	63_01c	Intergenic
9	42676612	63_01c	LINC01189
17	30486091	63_01c	RHOT1
7	32981972	63_01c	RP9P
3	44943040	63_01c	TGM4
11	49056353	63_01c	TRIM49B
3	49688432	63_01a	BSN
11	120732628	63_01a	GRIK4
17	47905002	63_01a	KAT7
1	215256574	63_01a	KCNK2
17	76045632	63_01a	TNRC6C
1	15120732	51_05	KAZN
8	30209516	51_01e	Intergenic
2	178083964	51_01b	HNRNPA3
1	12954449	51_01b	PRAMEF10
11	48791891	49_01e	Intergenic
3	138009389	49_01d	ARMC8
3	138009389	49_01d	NME9

3.7 DNA ploidy and copy number changes

Copy number aberrations frequently occur in cancers[152]. Where there is an increase in the normal diploid (two) copies of each DNA segment, an amplification has occurred, and where there is a loss of one or both copies, then a copy number reduction can occur. These two types of aberrations provide important causes of oncogenesis. Amplifications of oncogenes and loss of tumour suppressor genes can result in a malignant phenotype[153]. Recurrent copy number aberrations over genes with characterised roles in cancer are more likely to be driving events. The study of allele specific copy number changes can also be used to model the specific timings of genomic mutations as discussed in Chapter 1 and below in Section 3.14.

3.7.1 DNA ploidy and genome duplications

Allele specific copy number and ploidy aberrations were detected using Sequenza[103]. The mean ploidy of each sample is displayed in Figure 19 and Figure 20 below. Fifteen out of 40 samples displayed significant ploidy changes, with evidence of genome duplication in all fifteen of these samples. Three (patients 39, 51 and 66) out of eight patients displayed almost no changes in overall ploidy, whereas patients 45 and 79 demonstrated ploidy changes across all samples. In patients 63, 64 and 79 there was one sample within the multiple regions sequenced where there was an almost doubling of ploidy. These ploidy changes were apparent across both HPV positive and negative samples with no trend seen for over representation based on HPV or metastatic lymph node status. The relationship between samples within each tumour is assessed in the analysis of regional intra-tumour copy number heterogeneity plots in Section 3.7.3 below. Individual allele-specific copy number aberration plots are displayed in the Appendix.

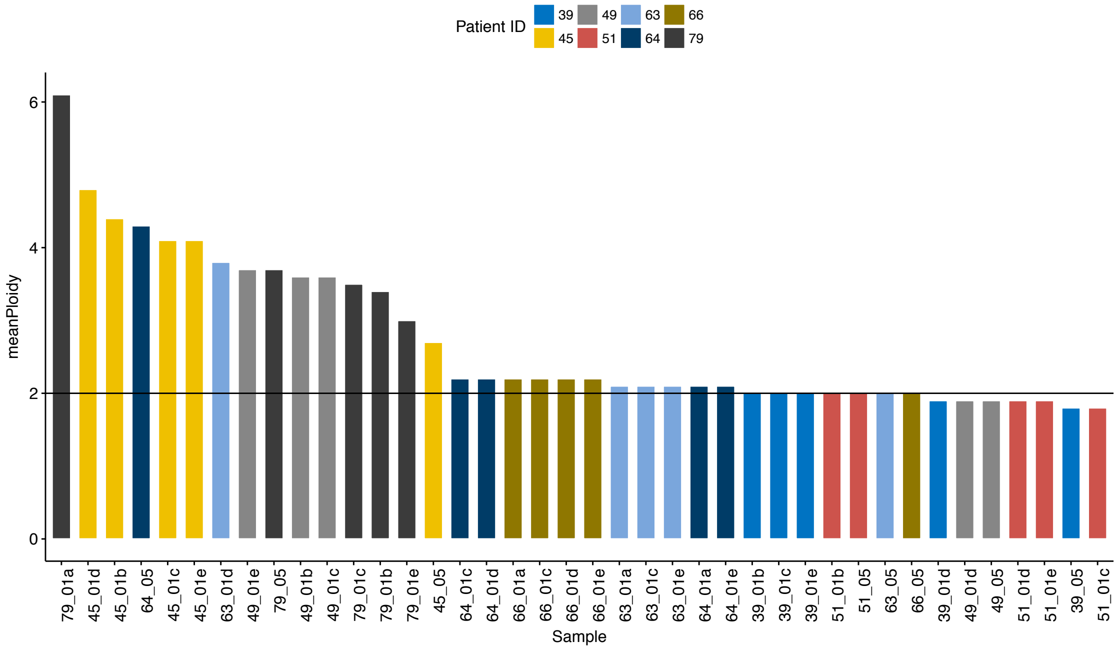


Figure 19: Bar chart of the mean ploidy across all tumour sample in the PenHet cohort , arranged by mean ploidy. Calculated using Sequenza.

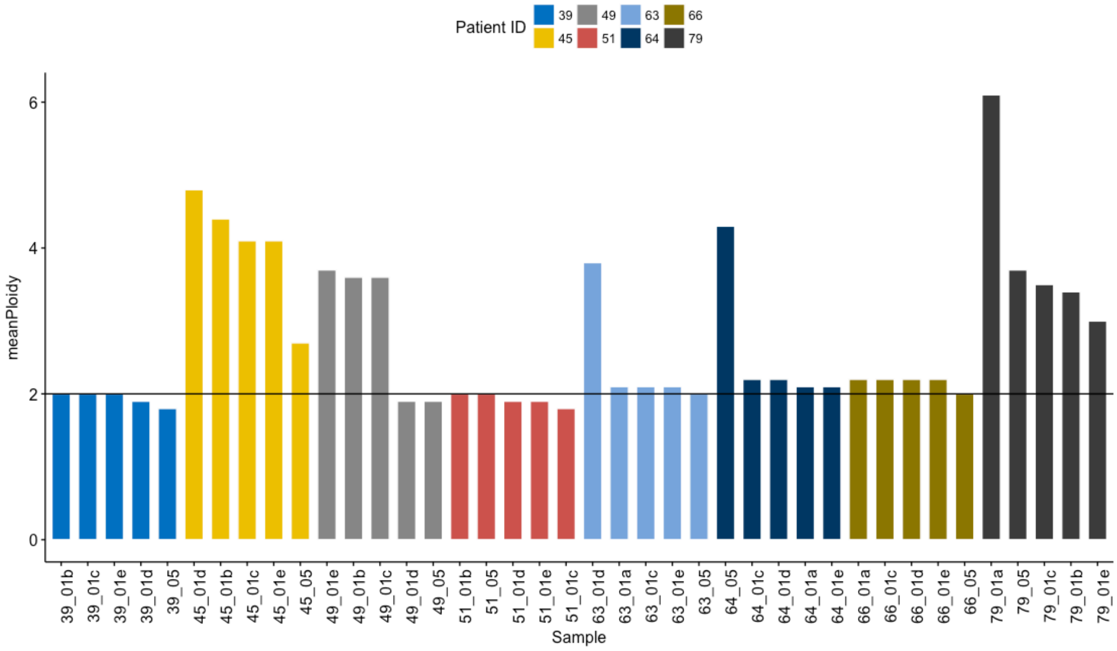


Figure 20: Bar chart of the mean ploidy across each tumour sample. Calculated using Sequenza.

3.7.2 Copy number gains and losses

Copy number gains and losses were calculated relative to genome ploidy specific to each sample. A gain was defined as a relative doubling of copy number and a loss was defined as a relative halving of copy number. Recurrent copy number losses were found in up to 42.5% of samples and gains in up to 30% of samples. The frequency of all significant copy number aberrations is plotted in Figure 21.

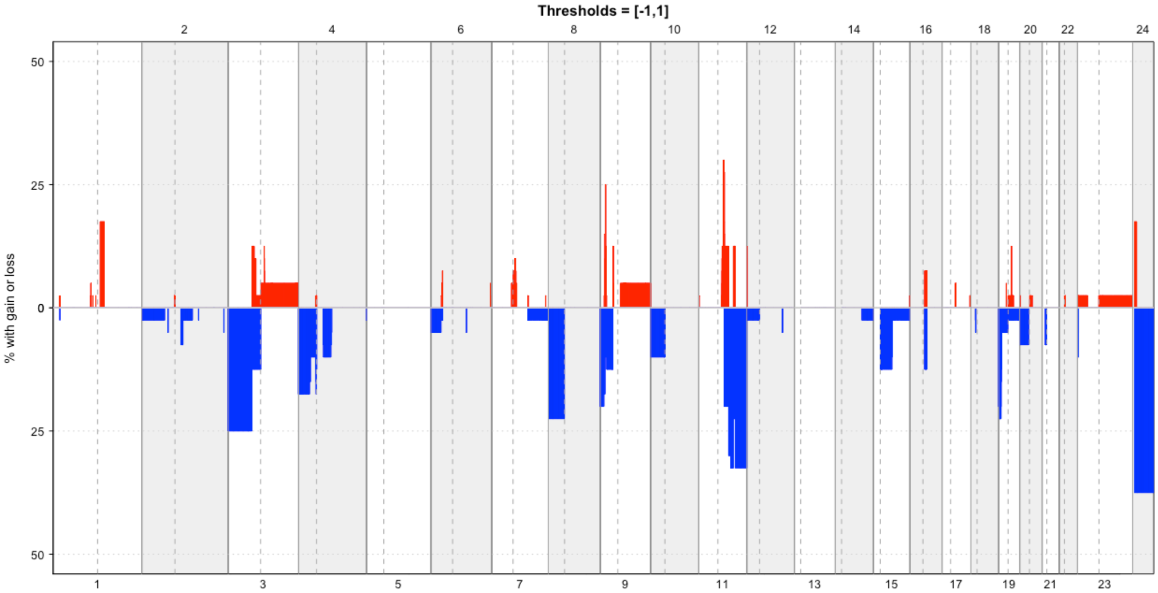


Figure 21: Frequency plot of all copy number gains and losses with a cut-off relative copy number change of +1 or -1 respectively. Gains are depicted as positive deflections in red and losses as negative deflections in blue.

At the gene level, 4,245 genes were in regions of significant gain and 9,781 genes were in regions of significant loss. When an assessment was made of just the genes in recurrent CNAs above a frequency of 10%, 624 and 4,495 genes displayed gains and losses respectively. Aberrations were more closely analysed in genes previously designated as drivers and those with potential actionable drug targets. For copy gains, nine genes were previously classified as drivers, and for copy losses 78 were previously classified as drivers. Comparing the recurrent and non-recurrent aberrations, 12 actionable targets were copy gained and 29 exhibited copy losses. These specific genetic drivers and drug targets are displayed with their frequency of aberrations in Figure 22 and Figure 23.

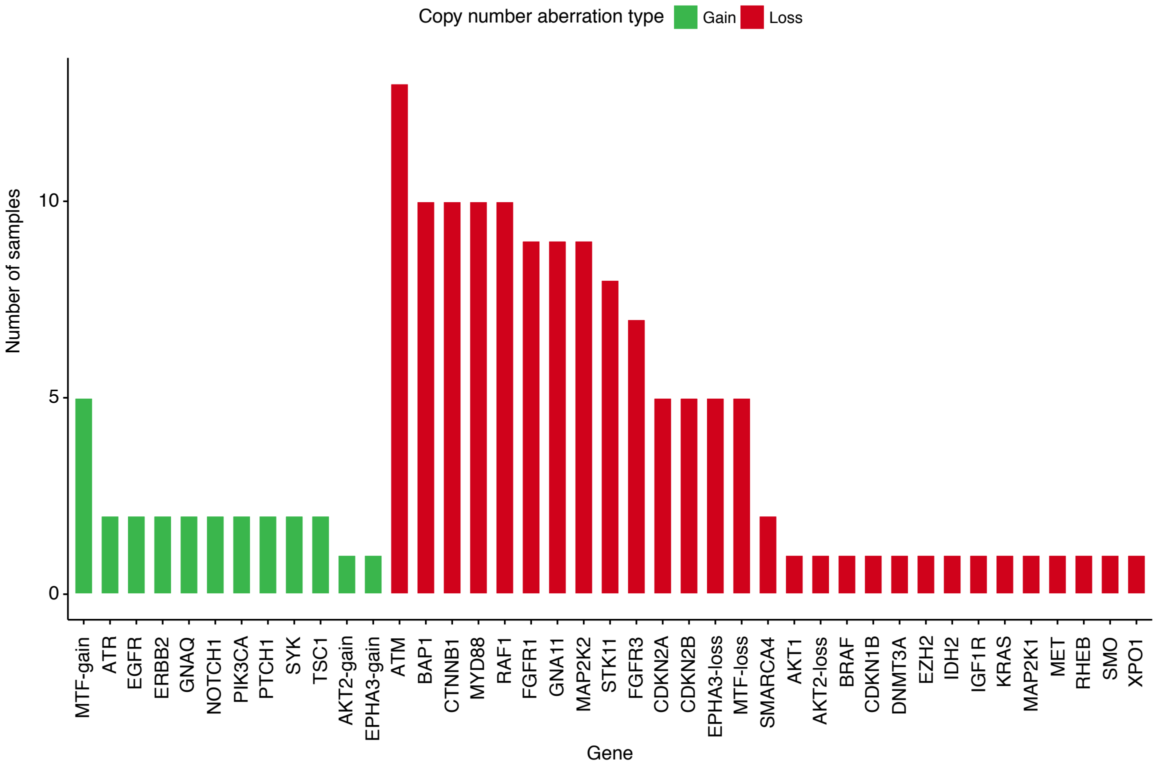


Figure 22: Frequency of copy number aberrations for genes with known drug targets. Blue bars indicate copy gains and yellow bars indicate copy losses.

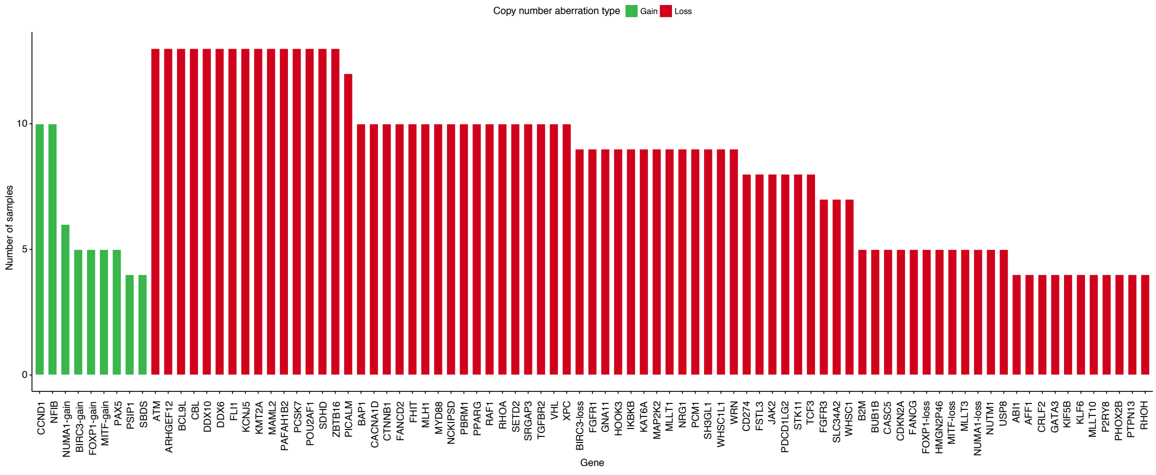


Figure 23: Frequency of copy number aberrations for genes previously described as drivers. Blue bars indicate copy gains and yellow bars indicate copy losses.

All regions were also assessed in an unsupervised manner to determine the relationship between samples using Minkowski distance measures. The samples generally clustered by patient but not by HPV or mutation signature status, Figure 24.

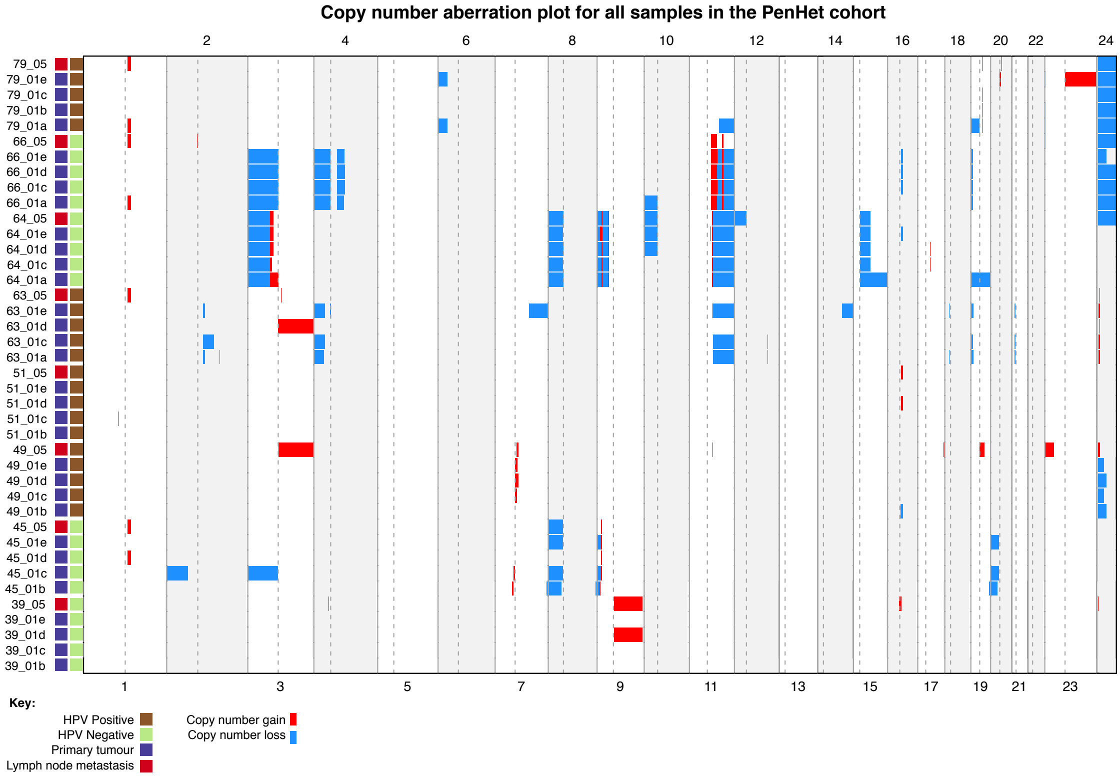


Figure 24: Heatmap of copy number aberration events across the PenHet cohort. Gains and losses are defined as copy changes of +1 or -1 respectively with gains depicted in red and losses in blue.

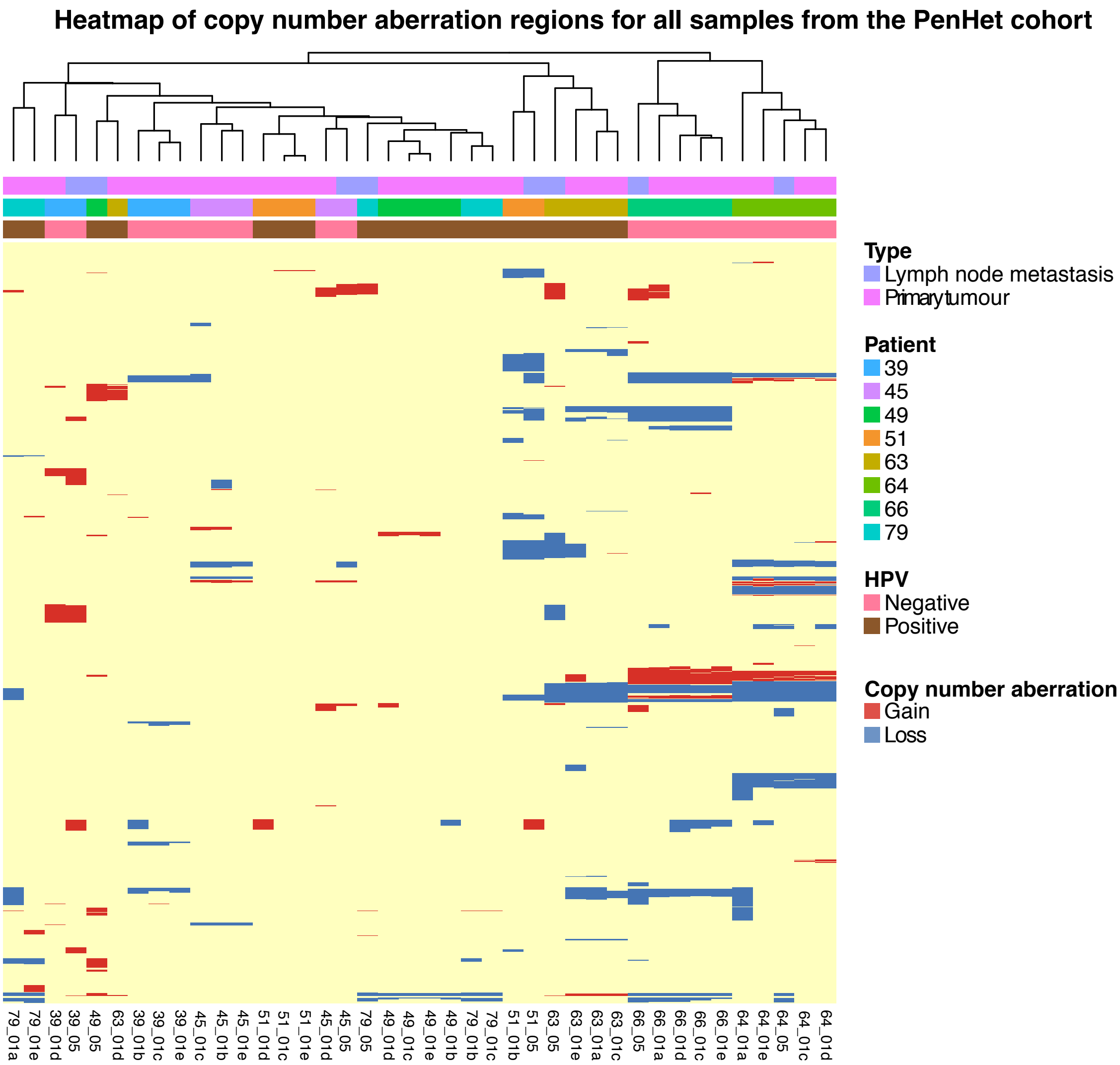


Figure 25: Heatmap and unsupervised hierarchical clustering of samples based on significant copy number aberration events. Significant gains are depicted in red, losses in blue. Copy number regions are segmented using the Copynumber R package as discussed in Chapter 2 (Methods).

3.7.2.1 Focal copy number aberrations

Focal copy number aberrations have previously been demonstrated to confer selective growth advantages in a range of cancers. Unlike in large scale CNAs where it can be difficult to determine which genes are driving a growth advantage, focal CNAs provide a clearer insight into potential driver genes. For this experiment a focal CNA was defined as a CNA that occurred within a gene. Focal CNAs were detected by assessing all recurrent focal CNAs present from the individual genomic copy number profiles within the PenHet cohort. As demonstrated in Figure 24 and Figure 25 only a very small minority of CNAs within this cohort are focal. Of particular interest are four focal CNAs which have been previously characterised as established oncogenic drivers including *EGFR, CCND1, ERBB2, NFIB, FOXP1, MITF* and *MDM2.* A selection of these copy number profiles can be found in Figure 26. The remaining copy number profiles not displayed here can be found in the Appendix.

Amplifications of *CCND1* have been repeatedly found in other squamous cell carcinomas such as oropharyngeal[154 155], as well as other many other cancers[152]. *CCND1* was found amplified in all samples from both patients 64 (HPV negative) and 66 (HPV positive). Cyclin D1 inactivates retinoblastoma protein and thereby promotes progression through G1 to S phase of the cell cycle. Amplification of *CCND1* has been demonstrated to promote increased expression of this oncogene.

ERBB2 encodes HER2 protein, which is an oncogenic receptor tyrosine kinase. It is well characterised in breast cancer where it is amplified and overexpressed in up to 20% of cancers[156]. This protein can be inhibited by the approved targeted monoclonal antibody Herceptin (trastizumab). Within the PenHet cohort *ERBB2* was found amplified in three out of four primary tumour samples from patient 79 (HPV positive) (all except for sample 79_01e) and in two samples (64_01c and 64_01d) from patient 64 (HPV negative).

EGFR is also a potentially oncogenic receptor tyrosine kinase found amplified as well as over-expressed in a wide range of cancers[152]. Amplification of *EGFR* was found amplified in two samples from patient 45 (45_01c and 45_01b) (HPV negative) and two samples from patient 64 (64_01c and 64_01d) (HPV negative). This may be clinically relevant as EGFR inhibitors are one of the targeted therapies currently being investigated and considered as a treatment option for penile cancer[157].

There was also one focal amplification of *MDM2* found solely in the metastasis of patient 79. *MDM2* is an oncogene that appears to exert its oncogenic activity by downregulating *TP53* activity[158]. This was only found in the metastasis of one out of the eight patients, so further work is needed to ascertain the role of *MDM2* in penile cancer.

FOXP1 and *MITF* were both amplified in all primary and metastatic tumour samples from patient 64 (HPV negative). *MITF* is an oncogene previously described in malignant melanoma[159] but never before in penile cancer. Further work will be needed to assess whether it is recurrently amplified and overexpressed in penile caner. *FOXP1* encodes the Fox transcription factors that appear to have a diverse range of functions and critical roles in immune responses, cell proliferation and oncogenesis[160]. Amplification of *FOXP1* may be important clinically as it has also been considered as a potential therapeutic target[160].

NFIB is a previously characterised oncogene and was found amplified in all primary and metastatic samples from patient 45 (HPV negative). *NFIB* encodes a transcription factor that can regulate the expression of over 1400 genes. It has been associated with metastatic aggressive phenotypes in small cell[161] and oesophageal squamous cell carcinomas[162].

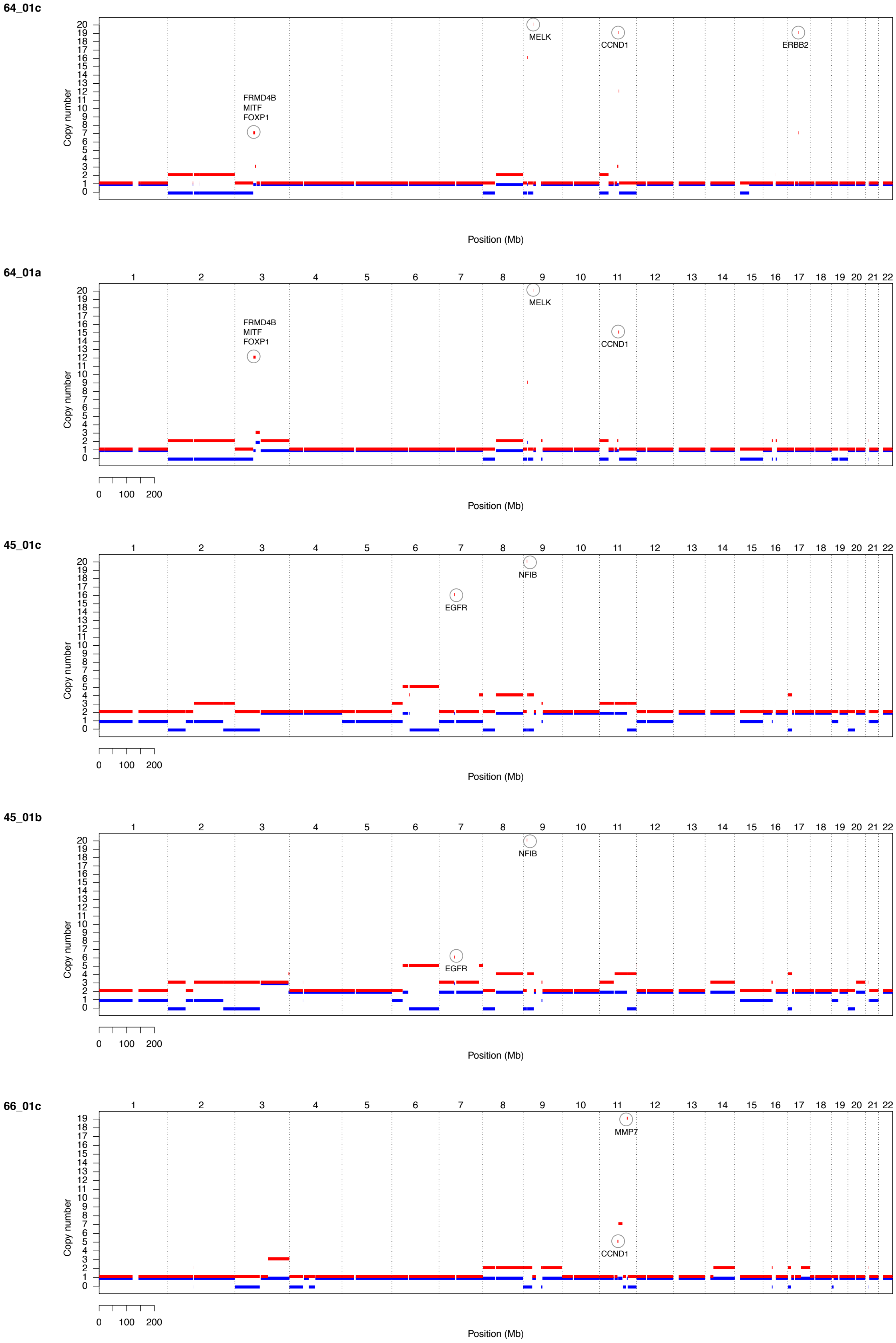


Figure 26: Genome wide copy number profiles for selected tumour samples based on positive findings of focal copy number aberrations. Red horizontal line indicates copy number of A – allele, blue lines indicates copy number of B –

allele. Circled vertical red lines indicate focal copy number amplifications of potential oncogenic genes. From the top: the first two profiles belong to primary tumour samples from patient 64. The next two profiles are from primary tumour samples from patient 45 and the last is from a primary tumour sample from patient 66.

3.7.3 Regional phylogenetic trees

The relationship between samples belonging to the same primary tumour is important to improve our understanding of how the tumour develops and what its key drivers are that may form the basis of therapeutic targets. Section 3.11 below on intra-tumour heterogeneity and the Introduction (Chapter 1) explain in further detail the concepts and use of evaluating intra-tumour heterogeneity.

Molecular aberrations that are present throughout all regions sampled within a primary tumour are likely to have developed at an early time point at a time of the last common ancestor. Copy number aberrations can be assessed in this way to determine what events are more likely to have occurred at an earlier or later time point. In addition, molecular aberrations which only affect a small sample of the tumour are unlikely to make successful therapeutic targets.

Regional phylogenetic trees to demonstrate the relationship between individual samples belonging to each patient were created using the fastme.bal method from the R package ‘ape’, as explained in the Methods in Chapter 2. These trees are split into Figure 27 for HPV negative patients and Figure 28 for HPV positive patients.

The copy number profiles from each sample are integrated with the SNV and INDEL information below, in Sections 3.11 to 3.14, to improve our understanding of the relative timings and importance of potential drivers in penile cancer and produce clonal phylogenetic river plots to analyse this information further.

Focal copy number amplification of *EGFR* was only found in two samples from one patient, patient 45. In this patient, this amplification was a relatively later event as it was not shared throughout the primary tumour. Further work should be undertaken on a larger cohort of penile cancer patients, because if this finding is replicated then it suggests that *EGFR* amplification is a later or subclonal event in penile cancer. Therefore, therapeutic targeting of *EGFR* may undertreat the entire cancer.

Although there is only a small sample size of HPV positive and negative patients, it is interesting to note that focal amplifications of oncogenic drivers were only found shared across tumour samples in the HPV negative patients. This suggests that focal copy number aberrations of *CCND1*, *NFIB*, *MITF* and *FOXP1* may play an important early role in oncogenesis in HPV negative patients. The only previously characterised focal copy number aberration in HPV patients was of *ERBB2*, which occurred in only a subset of two samples from patient 79. This suggests that amplification of *ERBB2* occurred as a later event than the SNVs which are shared throughout all tumour regions.

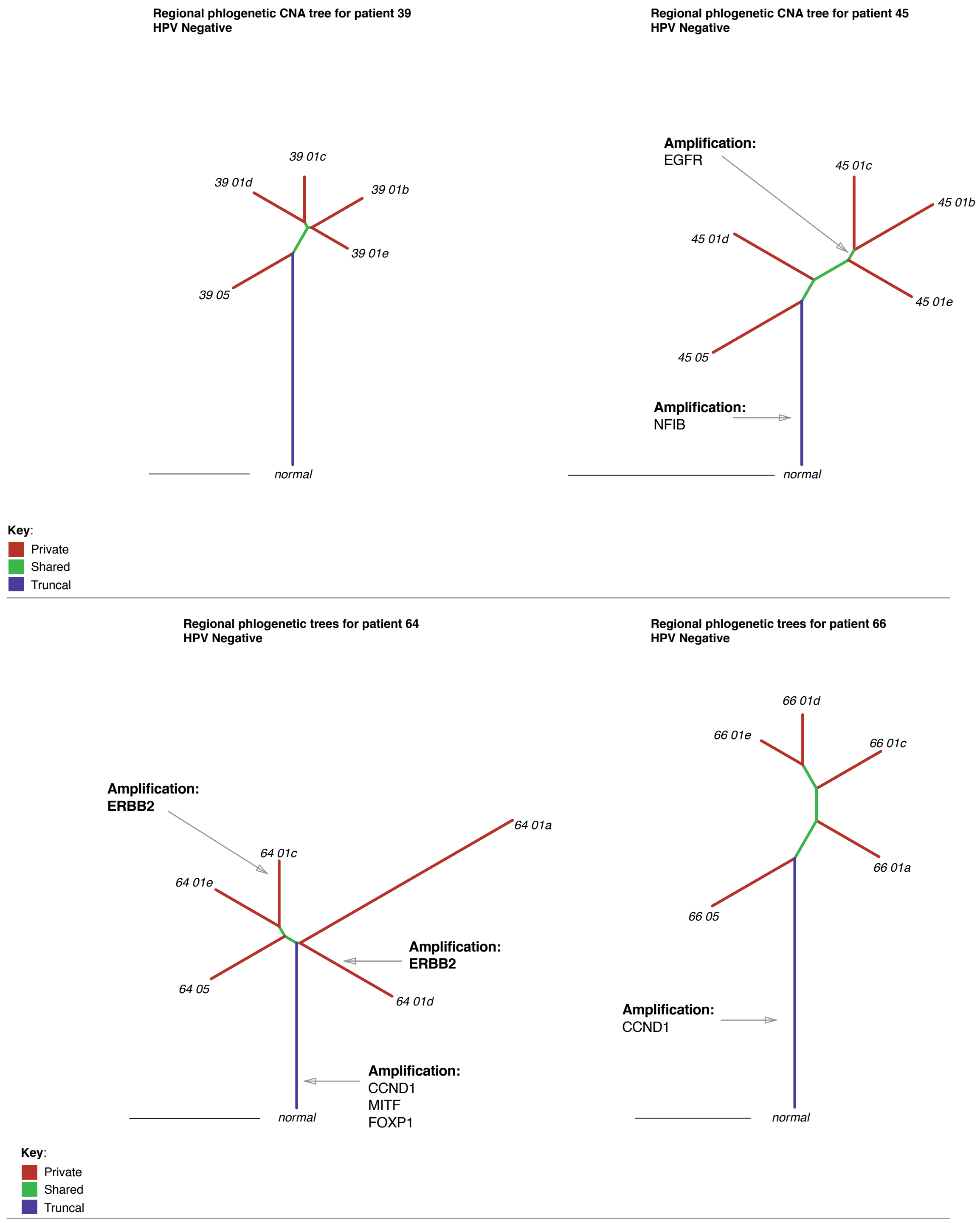


Figure 27: Regional CNA phylogenetic trees for HPV negative patients in the PenHet cohort . Focal copy number aberrations previously described as oncogenic drivers in other cancers are annotated and discussed in the text. The regional phylogenetic trees are generated using the fastme.bal function of the 'ape' R package as described in the methods (Chapter 2). This function creates a tree based on a minimum evolution algorithm. Sample labels with the suffix 05 represent lymph node metastases. Sample labels ending in a letter are primary tumour samples.

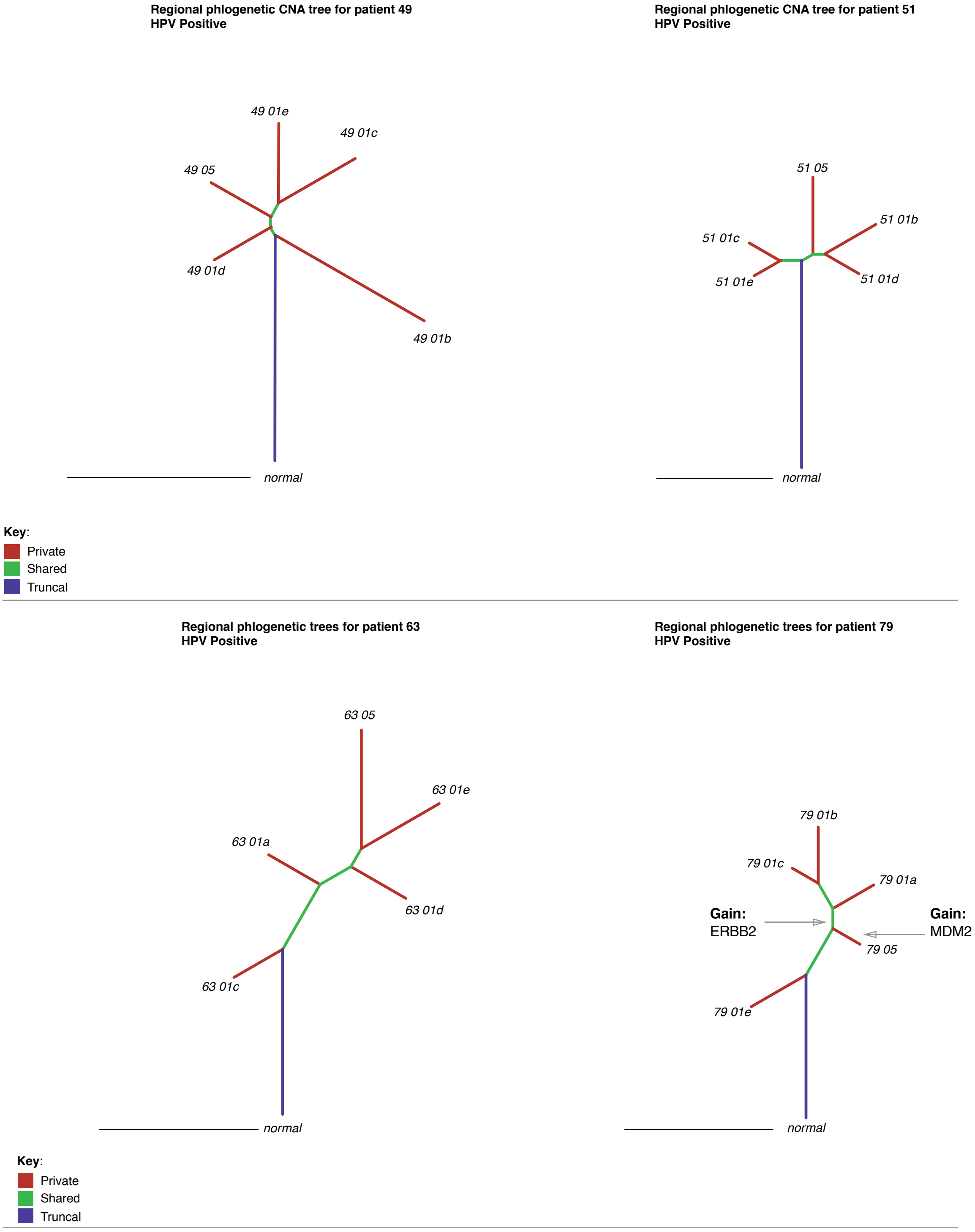


Figure 28: Regional CNA phylogenetic trees for HPV positive patients in the PenHet cohort. Focal copy number aberrations previously described as oncogenic drivers in other cancers are annotated and discussed in the text. The regional phylogenetic trees are generated using the fastme.bal function of the 'ape' R package as described in the methods (Chapter 2). This function creates a tree based on a minimum evolution algorithm. Sample labels with the suffix 05 represent lymph node metastases. Sample labels ending in a letter are primary tumour samples.

3.8 Whole genome doubling

Whole genome doubling can be determined by undertaking allele specific copy number profiling as demonstrated above in Section 3.7. In order to assess whether whole genome duplication has occurred, one needs to determine whether either the whole genome has duplicated at a specific time point or if individual chromosomal segments have progressively gained in copy over time. This can be determined by examining the allele specific copy number profiles to understand whether there is a major copy number of greater than or equal to 2 throughout the genome. Examples of this can be demonstrated in Figure 29, where the whole genome was doubled in all three primary tumour regions. It is also clearly demonstrated how the majority of copy number events are conserved, with little evidence for copy number heterogeneity between these three samples.

Whole genome doubling was observed in 15 samples from five patients. This can be seen from the overall ploidy in Figure 19. It was an early truncal event in two of these, patient 49 and patient 79. The remaining samples where genome doubling was observed either took place in solitary regions (patient 63 and patient 64) or only 75% of the regions sampled (patient 49). However, due to the small sample sizes, there is insufficient evidence to determine if whole genome doubling is predominantly an early or late event in the development of penile cancer. It would also not be appropriate to predict the whole genome doubling rate based on this small cohort of patients. In addition, it is not currently clear whether there is any association with HPV infection. Further work is needed on a larger cohort to determine the frequency and timing of whole genome doubling within penile cancer. The required sample size to power a future study to investigate these findings can be estimated by predicting the minimum prevalence, significance level and allele frequency. Based on the above findings for genome doubling, approximately 100 tumour samples would be required based on 80% power to detect the prevalence of genome doubling, with a significance of 0.05[163].

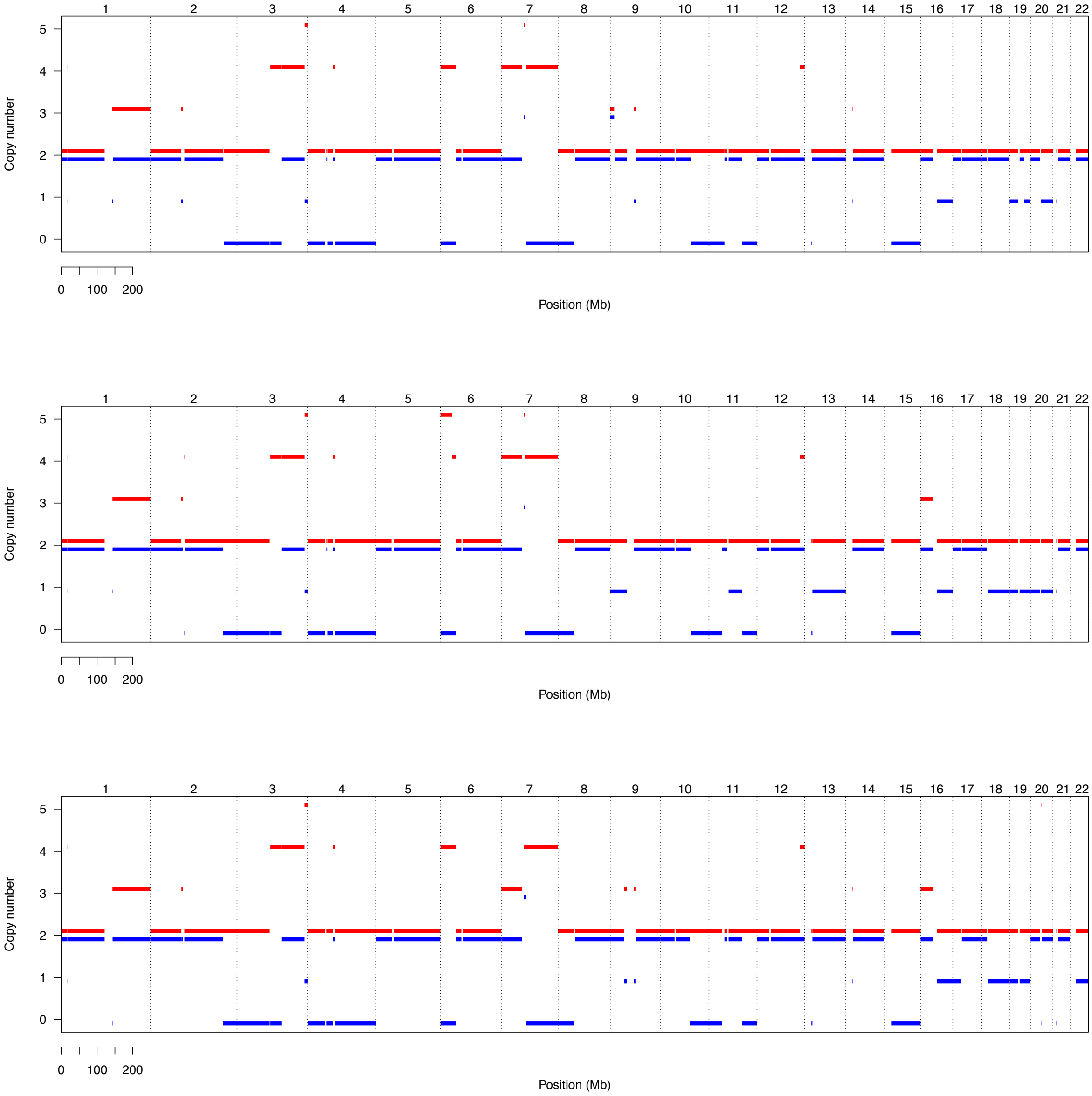


Figure 29: Allele specific copy number profiles demonstrating copy number duplication for patient 49. Top profile is for primary tumour section 01e, middle for section for primary tumour section 01c, and bottom for primary tumour section 01b. Key: Red horizontal line = A – allele, Blue horizontal line = B - allele

3.8.1 Copy number analysis conclusions

In summary, in general there is a paucity of focal recurrent copy number aberrations of known drivers within penile cancer. Furthermore, there does not appear to be any evidence that HPV is associated with any global increased or decreased rates of large-scale CNAs.

Focal drivers of known oncogenes were found in 17 out of 40 samples. These included putative oncogenes *CCND1*, *EGFR*, *ERBB2, FOXP1, NFIB* and *MDM2*. Unlike when assessing SNVs and INDELs, most of the copy number aberrations were shared throughout all the primary tumour samples. This will be discussed in the section below on intra-tumour heterogeneity, but essentially indicates that these major broad and focal copy number aberrations occur relatively earlier than most of the other mutation types discovered in this cohort of penile cancer patients. There does appear to be a trend that shared focal amplifications of oncogenes only appear in the HPV negative samples. Perhaps these potential oncogenic drivers are necessary for early HPV negative tumorigenesis, whereas in HPV positive disease these may not be necessary as alternative pathways exist due to the viral oncogenes, human viral defence or other reactive pathways. Focal copy number aberrations that were not shared and therefore may represent later events were found in HPV positive samples.

Amplifications in *EGFR* were found in only a small subsection of samples and were not shared throughout the tumour samples belonging to one patient. This raises the possibility that targeted EGFR inhibition in penile cancer may be sub-optimal. It will not be targeting the entire tumour, as *EGFR* is likely subclonal in origin.

3.9 Mutational signatures

Specific mutational substitutions are associated with different cancer phenotypes and characteristics, including age, smoking history, APOBEC activity (particularly seen in HPV driven cancers), defective DNA double stranded breaks, defective DNA mismatch repair, ultraviolet radiation exposure and alkylating agent exposure. An analysis of mutational signatures was therefore undertaken to understand what the key classes of mutations are within the PenHet cohort and penile cancer in general. All potential mutational substitutions can be classified according to six classes: C>A, C>G, C>T, T>A, T>C, T>G. In addition, the bases immediately 5' and 3' from the mutated base can also be classified. This produces a large grid of 96 potential mutation subtypes with distinctive patterns of clusters of each type of mutations seen as previously characterised. Alexandrov et al[104] demonstrated and validated 21 different mutational signatures based on the distinct proportions of each of these different mutation types.

Using the package deconstructSigs in R, written by Rosenthal et al[105], the proportion and statistical significance for the presence of each signature was assessed throughout the PenHet cohort. The following mutational signatures were found based on a minimum detection limit of 6%: signature 1A, 1B (age related), 2 (APOBEC), 13 (APOBEC), 16 (unknown) and 17 (unknown) demonstrated in the bar charts in Figure 30. As previously shown in other cancers[164], oncogenic HPV integration is associated with APOBEC activity. One can hypothesise that APOBEC is part of the immune viral defence system with hypermutations rendering many viruses ineffective. It is therefore not surprising that in many cancers oncogenic HPV integration is associated with APOBEC activity. Bar charts of mutational signatures across the entire cohort is displayed in Figure 30. Signatures 2 and 13, which are associated with APOBEC activity, are more prevalent than one would expect by chance ($p < 0.0E-8$) in the four patients with integrated HPV16 (patients 49, 51, 63 and 79). The APOBEC associated mutational profile can be found in all HPV positive samples and in small amounts in three HPV negative samples. The HPV viral load data in Figure 16 was compared to the APOBEC (signature 2) weights from Figure 30. There was an association between mutational APOBEC signature 2 weights and HPV viral load (Rs = 0.67888, p (2-tailed) < 0.001). It is possible that this association is an underrepresentation due to the inherent inaccuracy of estimating viral load based on a bias capture of exomes.

Signatures 16 and 17 have an unknown aetiology and are found in liver cancer (in the case of signature 16) or oesophageal, breast, liver, lung, lymphoma, stomach and melanoma (in the case of signature 17). Mutational signature 16 is overrepresented in 25 out of 40 samples encompassing all patients, with the exception of patient 63. Mutational signature 17 is highly overrepresented in patient 63. Overrepresentation was assessed using Fishers exact test with a minimum normalised signature weight of more than 6%, as previously demonstrated as an optimum to maintain sensitivity whilst reducing false positives[105].

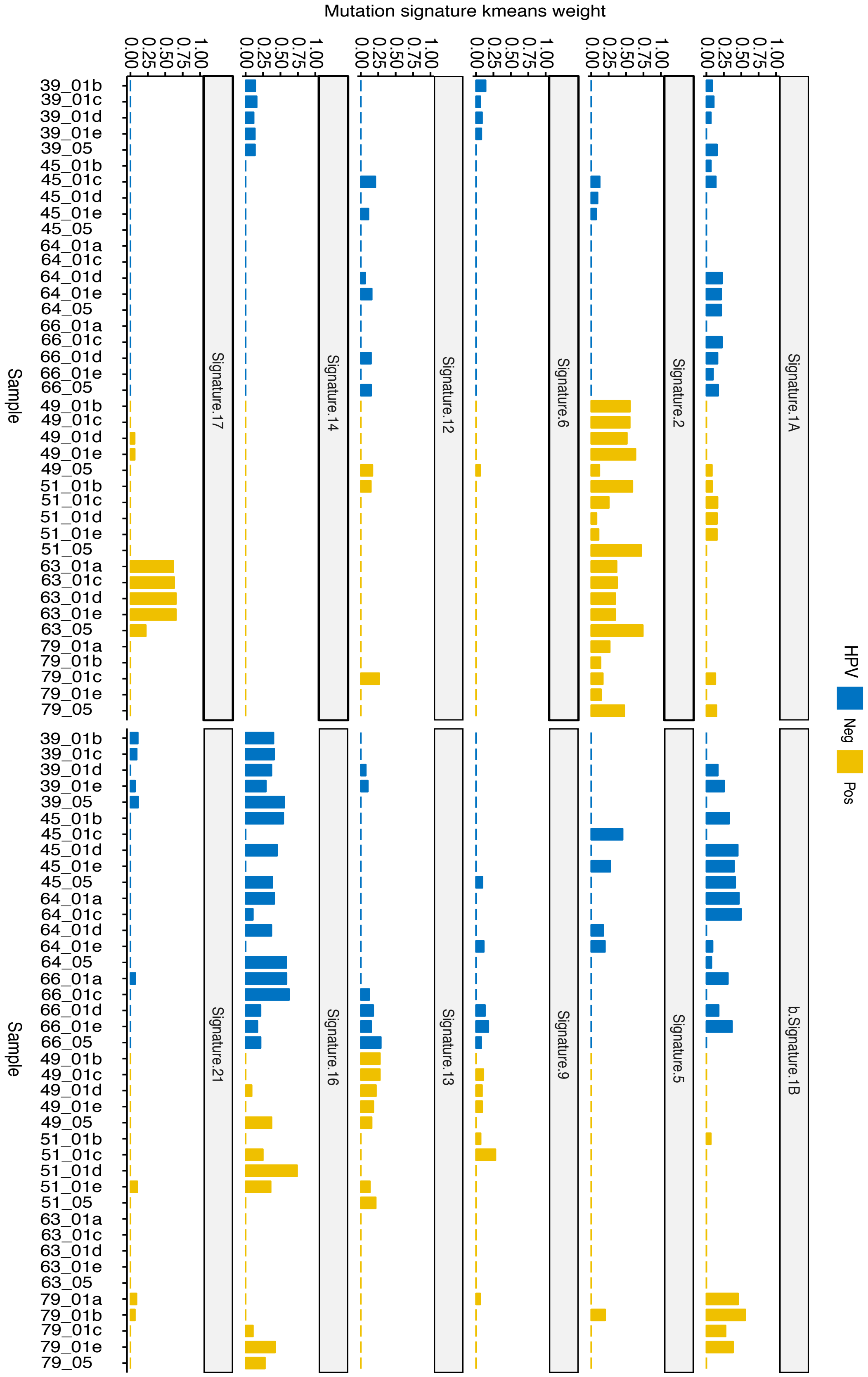


Figure 30: Bar chart of normalised weights for each mutational signature for all samples in the PenHet cohort, grouped by patient and HPV status (positive samples in yellow). A significant signature has a weight of above 0.06 as this cut-off was previously shown to limit the amount of false positives whilst maintaining sensitivity[105].

3.10 Summary statistics of somatic mutations

Summary statistics for all annotated non-synonymous mutations are displayed in Figure 31. 6,004 single nucleotide variants (SNVs) (mean 150 SNVs per sample) were identified, followed by 237 frame shift deletions (5.9 per sample) and 89 frame shift insertions (2.2 per sample). The mean transition/transversion ratio was 2.10. One class of SNV transversion stood out as being significantly more frequent than others: that of cytosine to thymine transversion. This has been found in most sequencing of other tumour types and is due to cytosine being the least stable of nucleotide bases. This is most likely caused by the observed spontaneous deamination of cytosine to uracil[165].

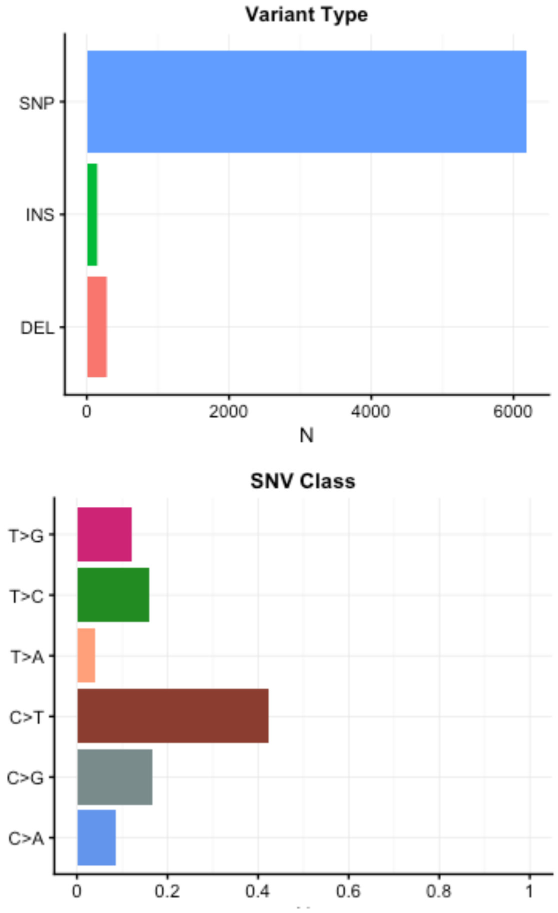


Figure 31: Proportion of mutation types (top), proportion of mutation transversions (bottom)

There were 2,683 uniquely mutated genes in the PenHet cohort. Of these, 105 were found in the COSMIC driver database. The following seventeen were found as potentially actionable: *PIK3CA, BRCA1, BRCA2, CDKN2A, RB1, MTOR, NF1, NOTCH1, NOTCH2, DNMT3A, APA, EPHA3, GNAQ, PIK3R1, SMARCA4, STK11.* However, mutations in these genes where rarely recurrent across the patient cohort with only one mutation, *PIK3CA*, recurrent at a frequency greater than 12.5%, as seen in Figure 32.

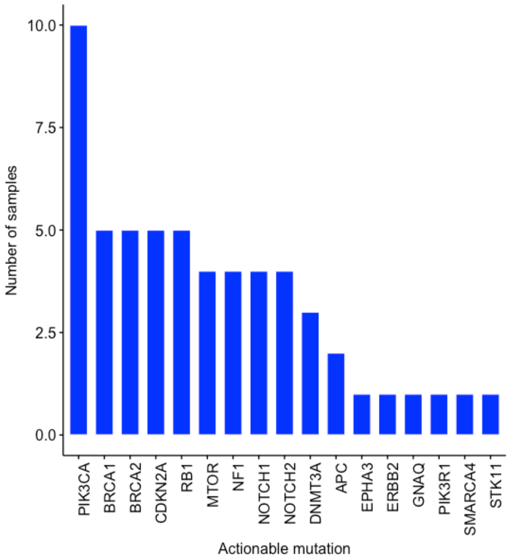


Figure 32: Bar chart displaying the number of samples mutated for each actionable mutation discovered.

There are 29 mutations that are recurrent in more than 20% of samples, as demonstrated in Figure 33. Of these mutations only *TP53, PIK3CA, FAT1* and *HNRNPA2B1* are classified as drivers, with only *PIK3CA* potentially actionable. Over half of all samples harbour fewer than six of these mutations. Interestingly, *TP53* was only found to be mutated in HPV negative samples, which was mutually exclusive to *PIK3CA* mutations, which were only found in HPV positive samples. This phenomenon has been previously alluded to in a small study of head and neck cancers[166,167]. When all mutations were clustered in an unsupervised manor, samples did not cluster by HPV status. However, when only considering driver mutations, there was some incomplete clustering by HPV status (Figure 34).

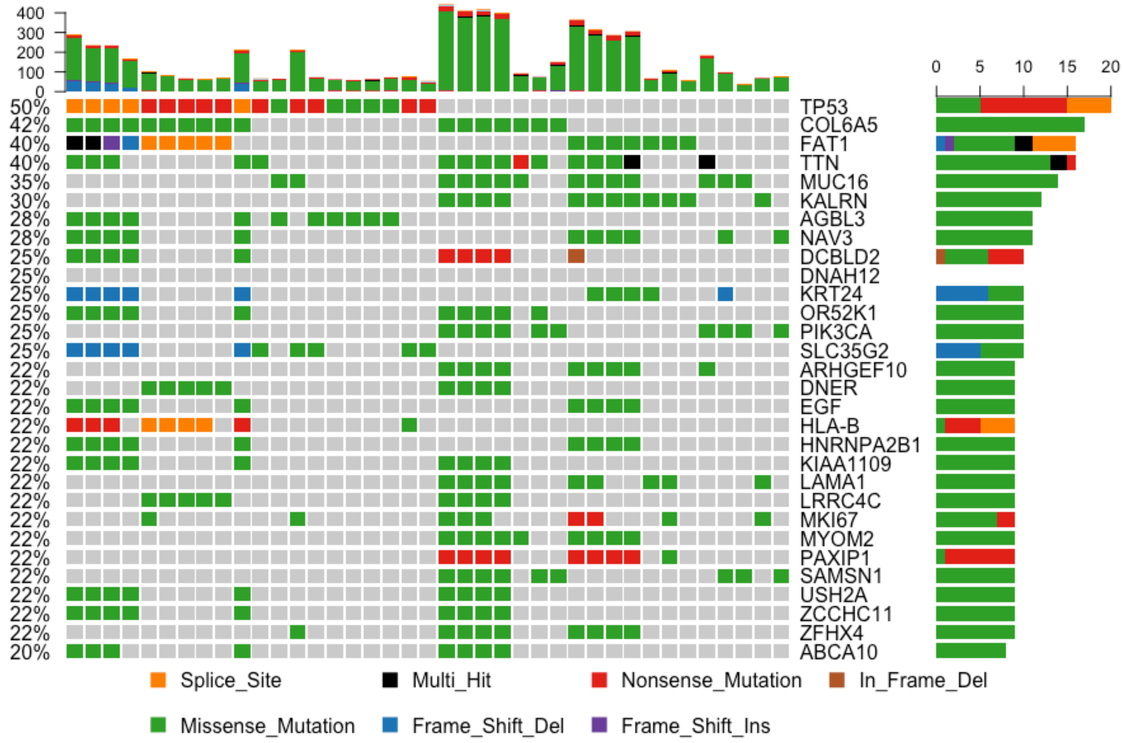


Figure 33: Recurrent mutations found in at least 20% of samples. The first 20 samples from the left are HPV negative, followed by 20 HPV positive samples.

3.11 Intra-tumour heterogeneity (ITH)

3.11.1 Calculation and scoring of intra-tumour heterogeneity

Comparison of the somatic alterations revealed heterogeneity in samples encompassing all eight patients in the PenHet cohort, regardless of grade, stage or smoking status, as demonstrated in Figure 24, Figure 25, Figure 33 and Figure 34. The relationship between each primary tissue sample from each patient can be examined by assessing which mutations are shared amongst all sequenced regions, which are shared only through a portion of regions and which are unique to each region.

Regional phylogenetic trees where produced by assessing the mutations that fall into each of these three categories of mutations. Mutations that are conserved across all regions make up the trunk of each tree, mutations that are unique to each region make up the terminal branches, and all other mutations make up the shared branches. Several methods are used throughout this thesis to produce this topological configuration and assess the relatedness of each sample to another. The two most commonly used methods involve ‘binarising’ the data and either assessing the Euclidian distance between each sample or by using a maximum parsimony ratchet method detailed in the methods section above. Extensive inter-patient and intra-sample heterogeneity was found, where the mutation profiles of samples from within a patient cluster together, yet still display remarkable intra-tumour heterogeneity as shown in Figure 34.

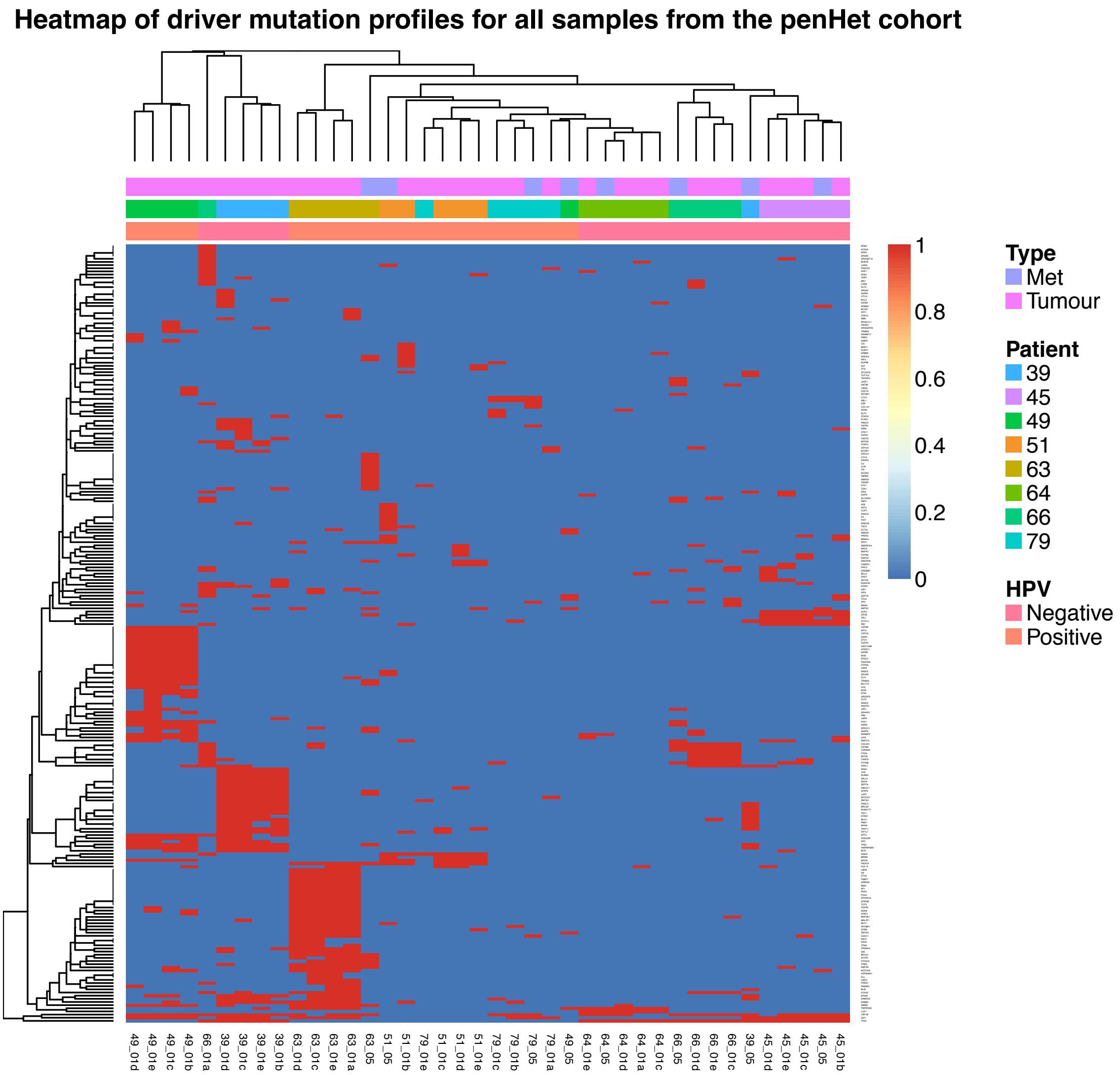

Heatmap of driver mutation profiles for all samples from the penHet cohort

Figure 34: Heatmap of all mutations, previously classified in other cancers as driver mutations in the COSMIC database. Mutations are clustered in a un supervised manner. Tissue type, patient and HPV status is labelled across the top of the heatmap.

The extent of intra-tumour heterogeneity (ITH) can be scored by several methods. Most simply the following can be calculated:

$$ITH = 1 - n$$

Where n = the proportion of mutations that are truncal in origin

Using this scoring system, the ITH ranged from 36%-96% with a mean of 69%, as displayed in Table 8.

One potential problem with this method of scoring ITH is that by this definition if one region branches off the main trunk at an early time point, then all later mutations are considered heterogeneous. Therefore, the ITH scores are very sensitive to the position of first regional branch. This may overestimate the amount of ITH, as the remaining regions may share the remaining mutations but would still be considered heterogeneous. Other methods of scoring ITH can be surmised that are less sensitive to the effect of one region sampled.

One previous study in lung cancer used a scoring system that was developed to disregard the trunk and instead calculate the average mutation distances between each region on the tree[168]. An alternative approach would be to disregard the trunk as before and instead calculate a 'unique branch mutational spread' by summing the length of the unique terminal branches and divide them by the total non-truncal mutations as follows:

Mutational branch spread ITH = Σ(Unique terminal branch lengths) / Σ(all non-truncal mutations)

When using this 'unique branch mutational spread' measure, the ITH ranged from 23%-85% with a mean of 60% as displayed in Table 8. The advantage of this measure is that it is not affected by trunk length. When one sample emanates from the main trunk, it can give the impression of extensive heterogeneity when the other samples may in fact be relatively homogeneous.

Table 8: Intra-tumour heterogeneity scores for each patient in the PenHet cohort. Two different methods of calculating ITH are used: a simple proportion of unique mutations in the branches of the phylogenetic trees, and a measure of branch spread.

Patient identifier	HPV status	ITH score: Proportion of mutations in branches	ITH score: mutational branch spread
39	Negative	0.63	0.35
45	Negative	0.74	0.70
49	Positive	0.61	0.54
51	Positive	0.93	0.80
63	Positive	0.51	0.23
64	Negative	0.66	0.65
66	Negative	0.89	0.85
79	Positive	0.95	0.67
HPV positive		0.75	0.56
HPV negative		0.73	0.64
Combined		0.70	0.66

3.11.2 Regional phylogenetic trees

To explore the relationship between individual tumour regions and the mutational drivers across a tumour, regional phylogenetic trees were calculated using the maximum parsimony method. The extent of the ITH as exemplified by the different ITH scores can be seen in Table 8. Using the mutational branch spread score, globally, HPV negative tumours showed a trend for higher ITH score compared with HPV positives, albeit not statistically significant (Mann-Whitney U, p = 0.486). Further work on a larger number of samples would be needed to demonstrate this, as there was only a limited power to statistically discriminate between two groups each containing only four patients. Within the HPV positive tumours there was no association between viral load (as calculated using WES) and ITH score.

The phylogenetic trees generated for each tumour are shown in Figure 35 and Figure 36 along with the key drivers and actionable mutations present in each trunk and branch of the tree. The figures also show, for each tumour, the proportion of truncal, shared and private SNVs between each region. Assuming the exact same mutation does not develop in two regions of a tumour in parallel, and that once a mutation is present it does not revert back to the wild type, one can hypothesise that mutations which are present in all areas of a tumour must have occurred prior to mutations present in only a proportion of regions. Thus, mutations which are in the trunk are more likely to be early events, and those in terminal branches are more likely to be later events. Furthermore, the trunk is likely to contain the clone, which represents the last common ancestor. This will be looked at in further detail by assessing the clonal status of each mutation by combining the mutation, copy number and tumour purity data.

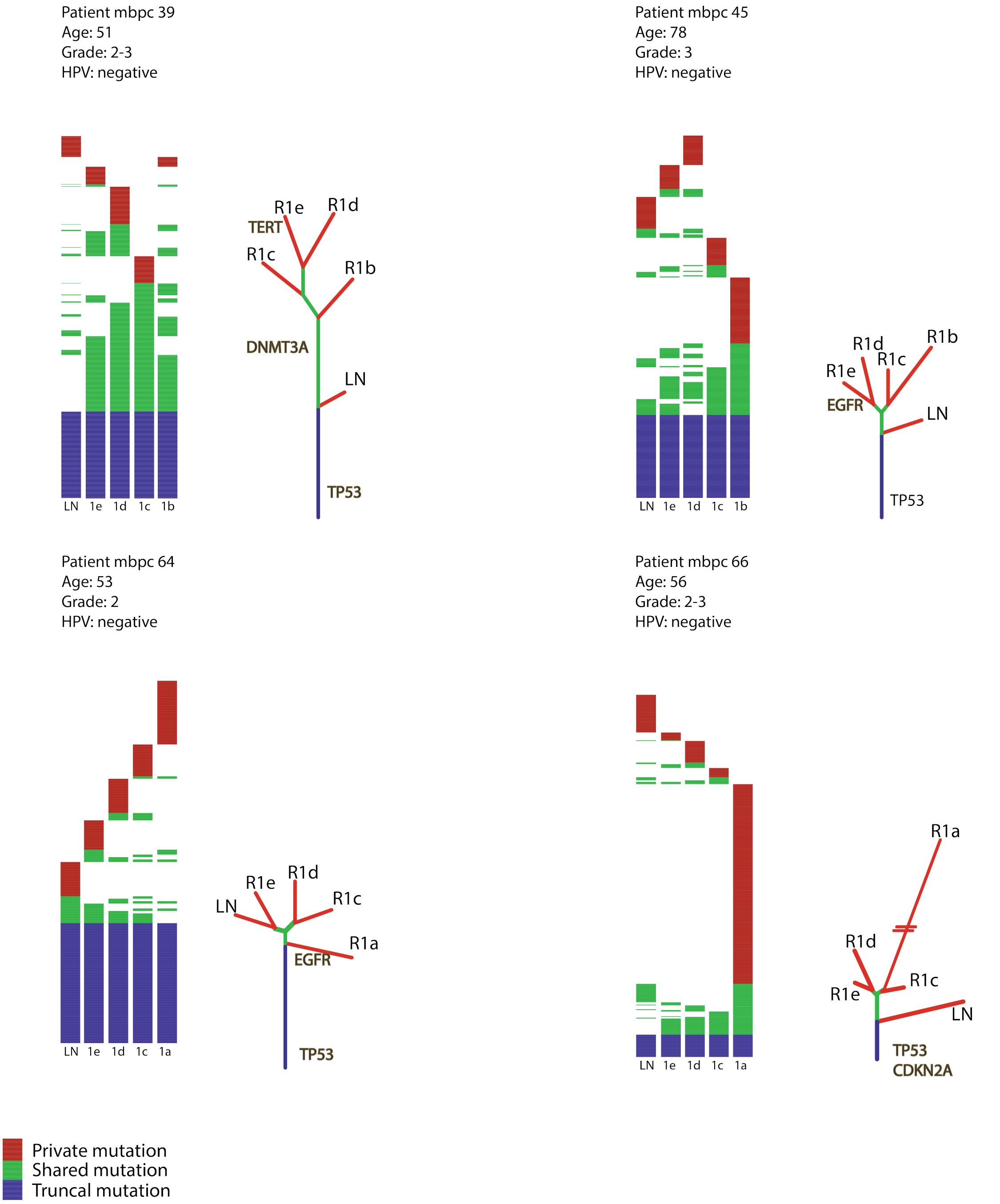


Figure 35: Regional phylogenetic trees depicting the relationship between samples for HPV negative patients. Major potential driver mutations listed in the COSMIC database, discussed in the text and demonstrated in the clonal phylogenetic riverplots are annotated. Key: LN = lymph node metastasis, Sample suffix a-e = primary tumour sample.

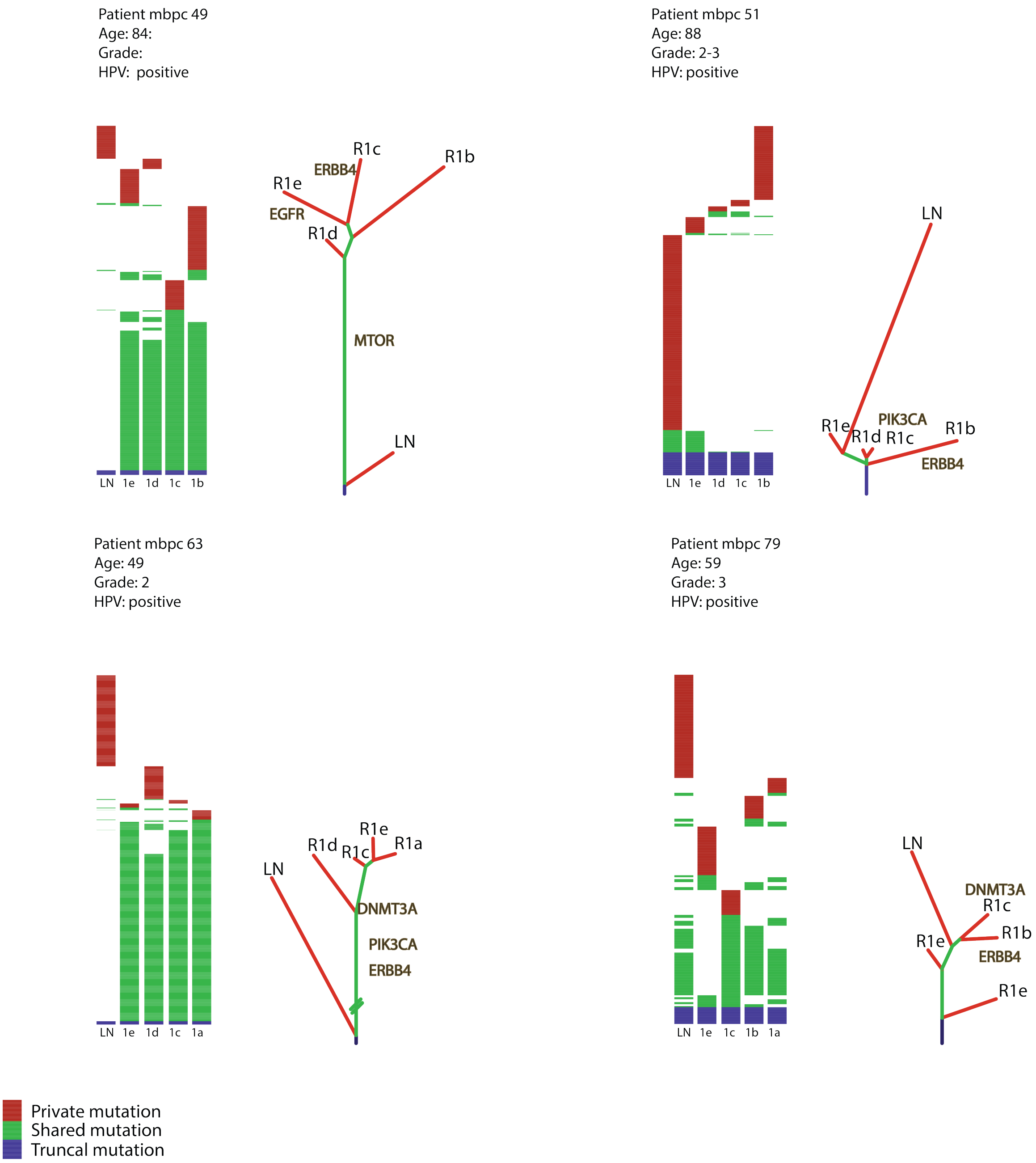


Figure 36: Regional phylogenetic trees depicting the relationship between samples for HPV negative patients. Major potential driver mutations listed in the COSMIC database, discussed in the text and demonstrated in the clonal phylogenetic riverplots are annotated. Key: LN = lymph node metastasis, Sample suffix a-e = primary tumour sample.

3.11.3 Clonal phylogenetic trees

Regional phylogenetic trees allow the mutations which are shared among all regions to be classified as early events and unique mutations to be classified as late events. As discussed in the Introduction (Section 1.1.8) the concept of tumour clones and subclones refers to the grouping of cells which are molecularly similar and have a similar phylogenetic history. A mutation which is clonal in origin refers to the idea that the mutation was present in the last common ancestral clone. A subclonal mutation occurred after this time point and will only be present in a fraction of tumour cells. The cancer cell fraction (CCF) refers to the proportion of cells within a cancer which harbour a given mutation. This can be calculated, as explained in the Methods (Chapter 2), by taking into account the variant allele frequency, the tumour content fraction and the copy number. The CCF is a useful measure in determining the stage at which a subclone developed. A CCF of 1 demonstrates a mutation that is present throughout all cancer cells and is therefore clonal in origin. As the CCF drops progressively lower it represents a mutation in a subclone that developed at a relatively later time point. The CCF was calculated for all mutations across all regions sequenced. Mutations can be clustered according to the CCF to mathematically predict the clonal and subclonal structure of a tumour. Each mutation can then be assigned a cluster based on its CCF and these clusters can be compared across each region sampled within each tumour. A visual representation of these clusters can be seen in the riverplots created for each patient in the PenHet cohort (Figure 37 to Figure 44). Each colour represents a distinct clone or subclone. The height of each of these clusters represents the proportion of cancer cells of a particular region sampled. The x-axis contains a column for each region sampled. If the coloured area of a particular subclone traverses horizontally across the columns of the x-axis then this demonstrates that the subclone is shared throughout the regions identified on the traversed columns.

These clonal phylogenetic figures reveal several insights into the heterogeneity and oncogenesis of metastatic penile cancer. Firstly, in all eight patients there is evidence of ITH with multiple clusters and clones/subclones of mutations. This reflects similar findings in other cancers as discussed in Chapter 1. Secondly, ITH is clearly present irrespective of HPV status and there does not appear to be any relationship between number of subclones and HPV status. Thirdly, there appear to be distinct oncogenic pathways when comparing HPV positive with HPV negative patients. In all cases, HPV negative patients have a clonal mutation in *TP53*, thereby indicating that *TP53* is a likely important early driver of HPV negative penile cancer. In contrast to this, *MTOR* and *PIK3CA* mutations are present in a clonal fashion in HPV positive samples. Fourthly, mutations in specific actionable mutations such as *EGFR*, *ERBB2* and *ERBB4* were only ever subclonal in nature. This indicates that targeted therapy against these actionable mutations may not be optimal in penile cancer as these therapies may only target a proportion of a patient's cancer. Fifthly, the lymph node metastases all appear to share the same initial clone as the primary tumour sample. This indicates that the lymph node metastasis developed from the primary and that if a targetable mutation was present in the initial clone then it would also be present in the lymph node metastasis.

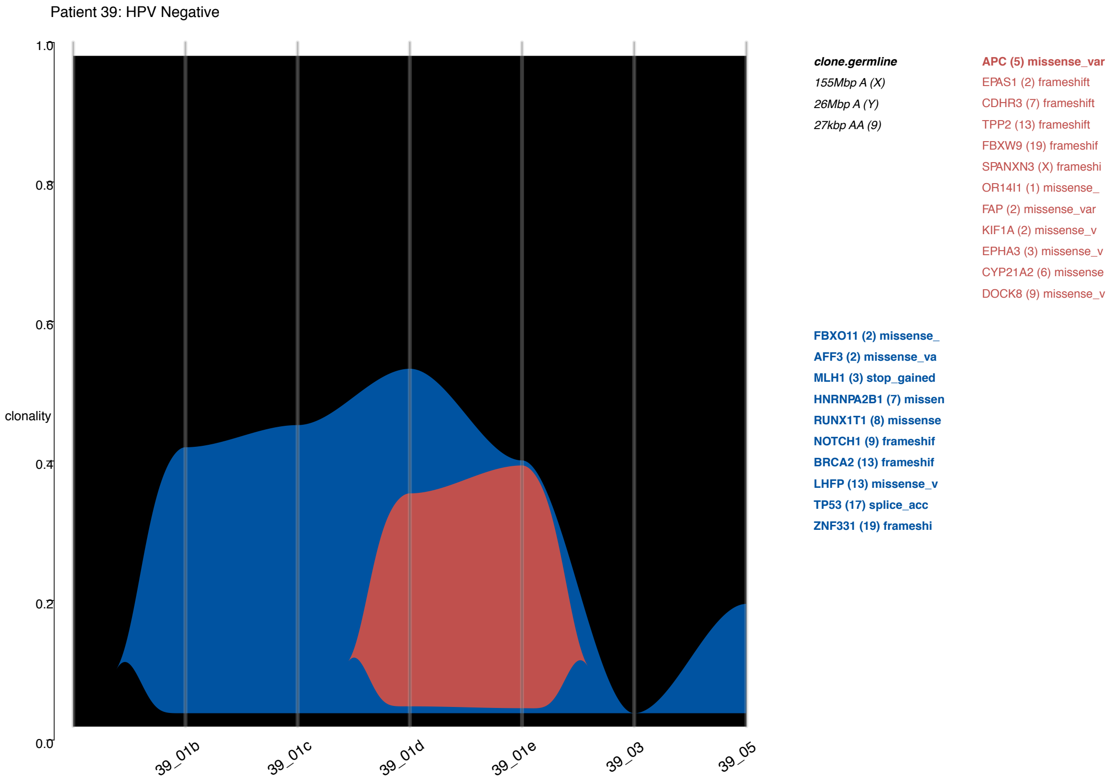


Figure 37: Riverplot for patient 39 (HPV negative) demonstrating the major clone and subclone detected across all regions sampled. Tissue samples are displayed across the x-axis. Sample names ending in a letter indicate a primary tumour sample. Samples ending in 03 indicate the whole blood control sample. Samples ending in 05 represent the lymph node metastasis.

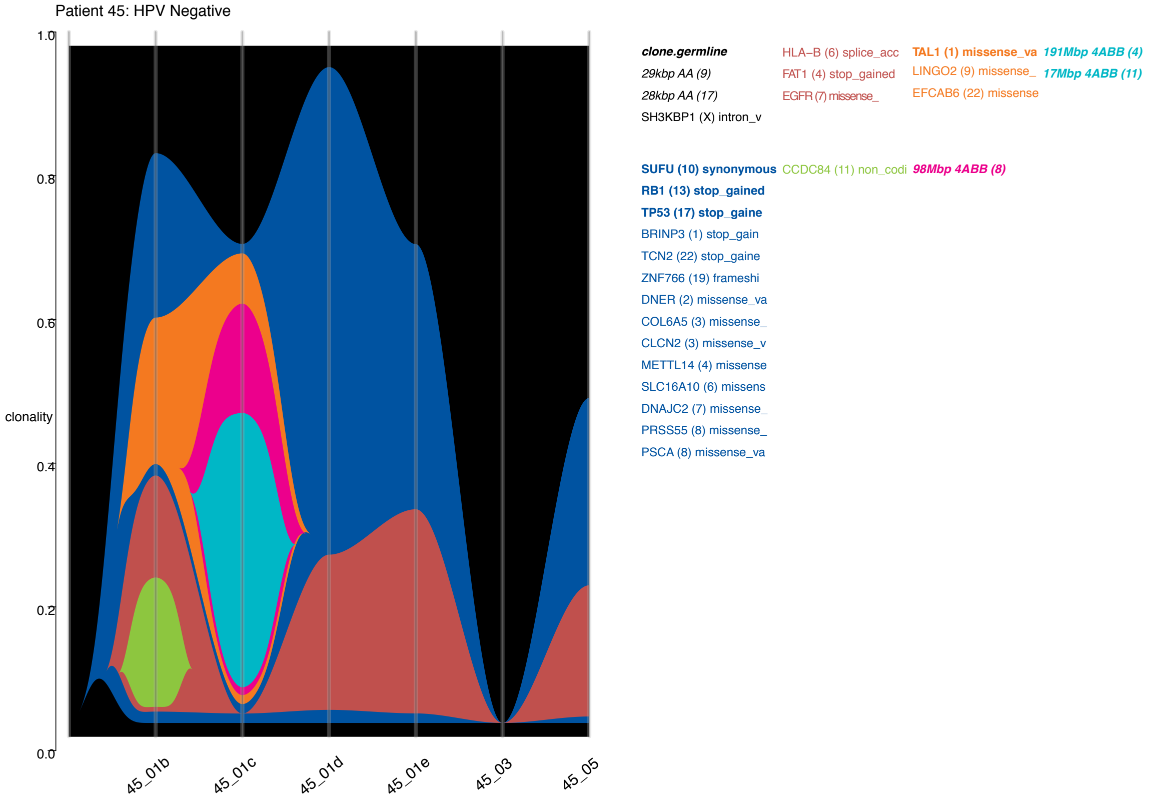


Figure 38: Riverplot for patient 45 (HPV negative) demonstrating the major clone and subclones detected across all regions sampled. Tissue samples are displayed across the x-axis. Sample names ending in a letter indicate a primary tumour sample. Samples ending in 03 indicate the whole blood control sample. Samples ending in 05 represent the lymph node metastasis.

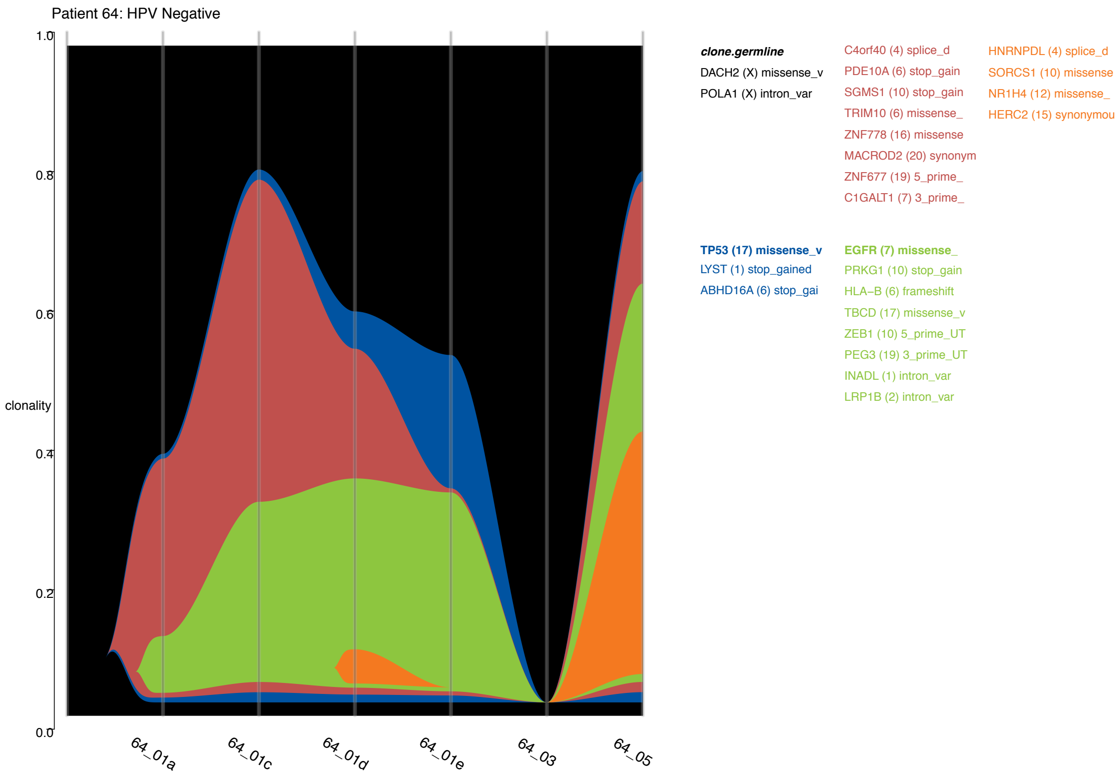


Figure 39: Riverplot for patient 64 (HPV negative) demonstrating the major clone and subclones detected across all regions sampled. Tissue samples are displayed across the x-axis. Sample names ending in a letter indicate a primary tumour sample. Samples ending in 03 indicate the whole blood control sample. Samples ending in 05 represent the lymph node metastasis.

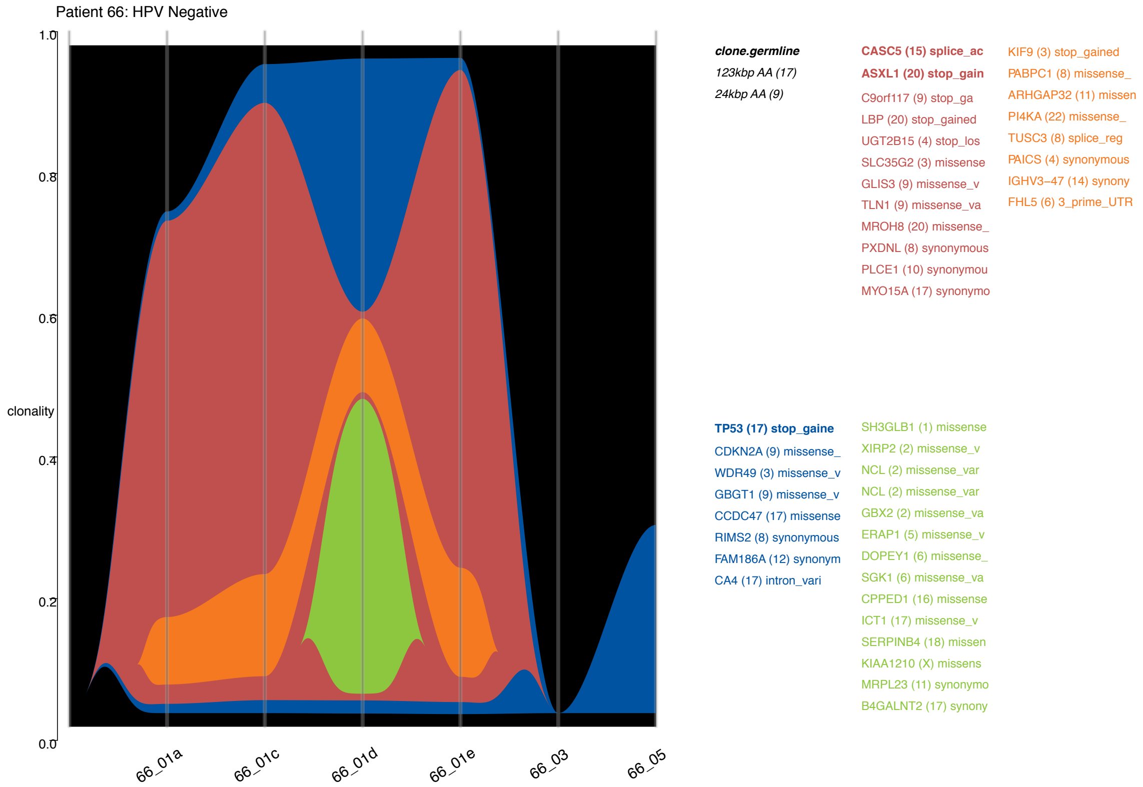


Figure 40: Riverplot for patient 66 (HPV negative) demonstrating the major clone and subclones detected across all regions sampled. Tissue samples are displayed across the x-axis. Sample names ending in a letter indicate a primary tumour sample. Samples ending in 03 indicate the whole blood control sample. Samples ending in 05 represent the lymph node metastasis.

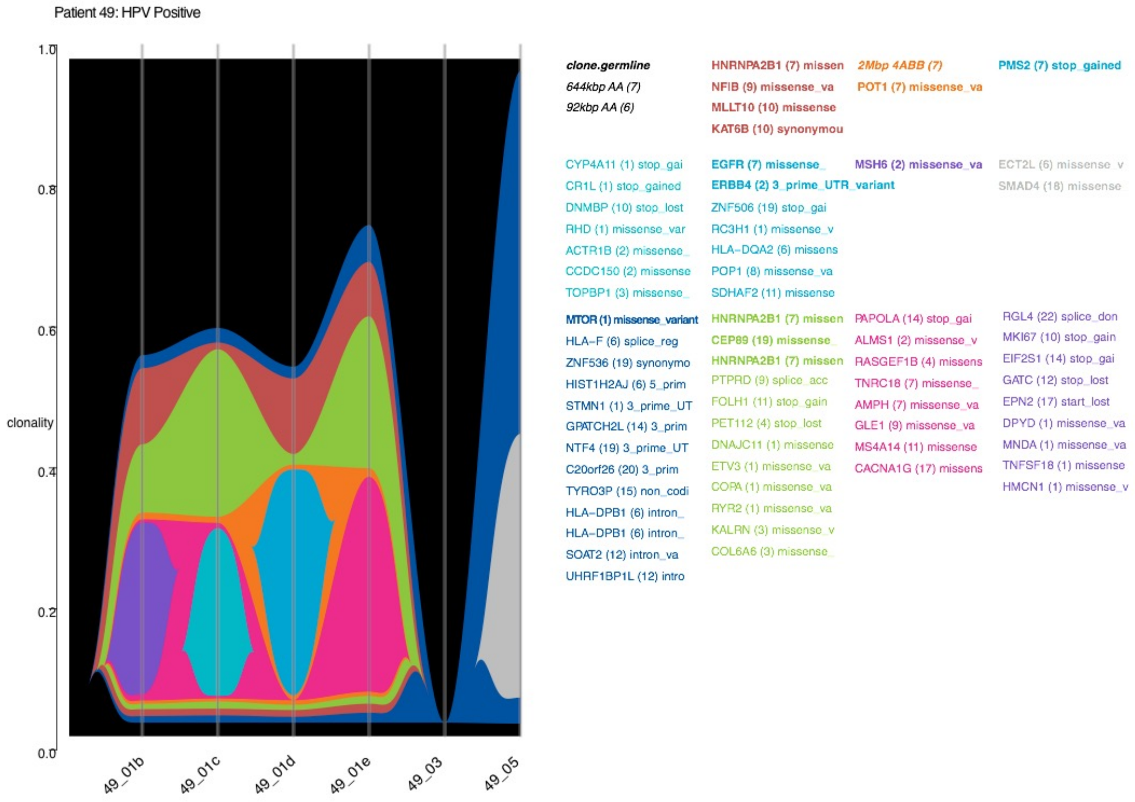


Figure 41: Riverplot for patient 49 (HPV positive) demonstrating the major clone and subclones detected across all regions sampled. Tissue samples are displayed across the x-axis. Sample names ending in a letter indicate a primary tumour sample. Samples ending in 03 indicate the whole blood control sample. Samples ending in 05 represent the lymph node metastasis.

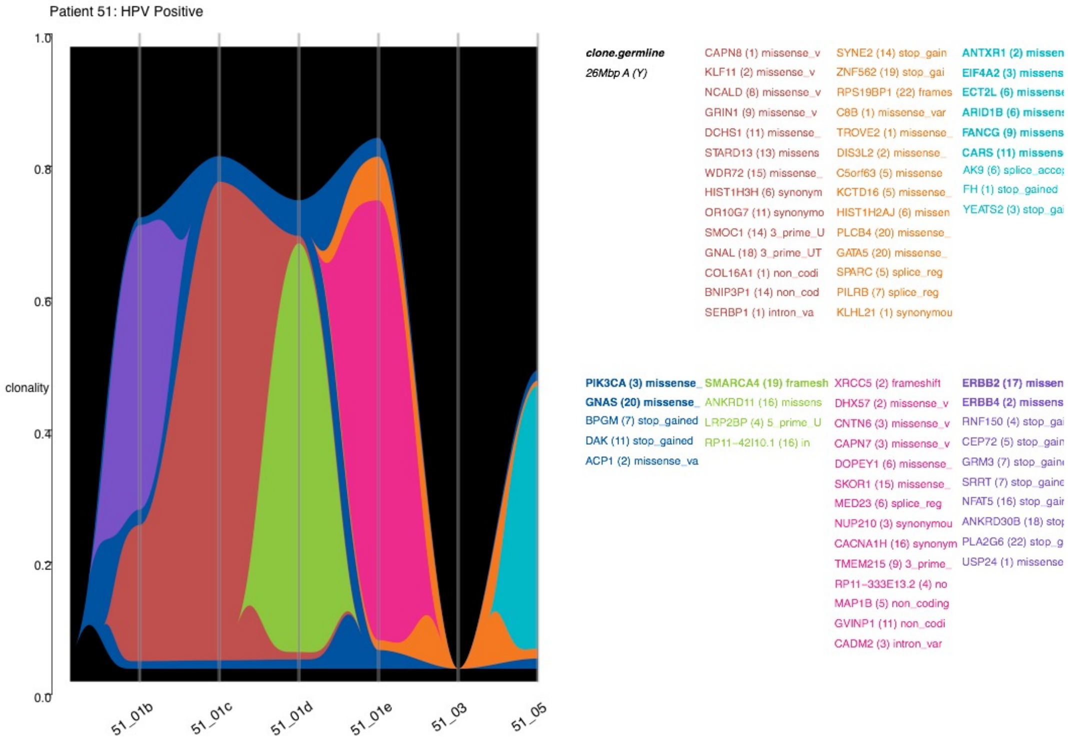


Figure 42: Riverplot for patient 51 (HPV positive) demonstrating the major clone and subclones detected across all regions sampled. Tissue samples are displayed across the x-axis. Sample names ending in a letter indicate a primary tumour sample. Samples ending in 03 indicate the whole blood control sample. Samples ending in 05 represent the lymph node metastasis.

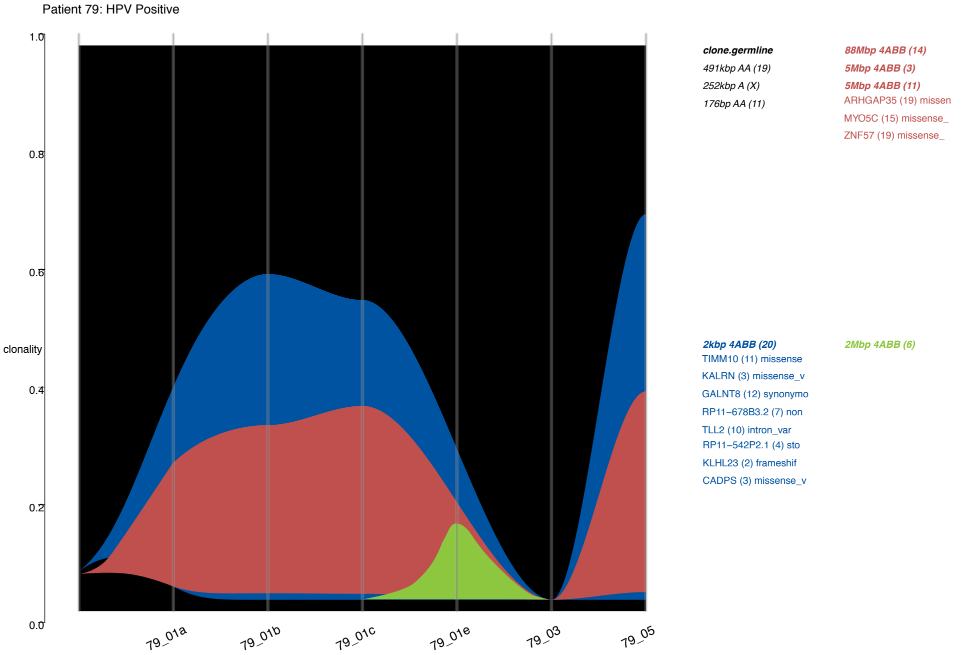


Figure 43: Riverplot for patient 79 (HPV positive) demonstrating the major clone and subclones detected across all regions sampled. Tissue samples are displayed across the x-axis. Sample names ending in a letter indicate a primary tumour sample. Samples ending in 03 indicate the whole blood control sample. Samples ending in 05 represent the lymph node metastasis.

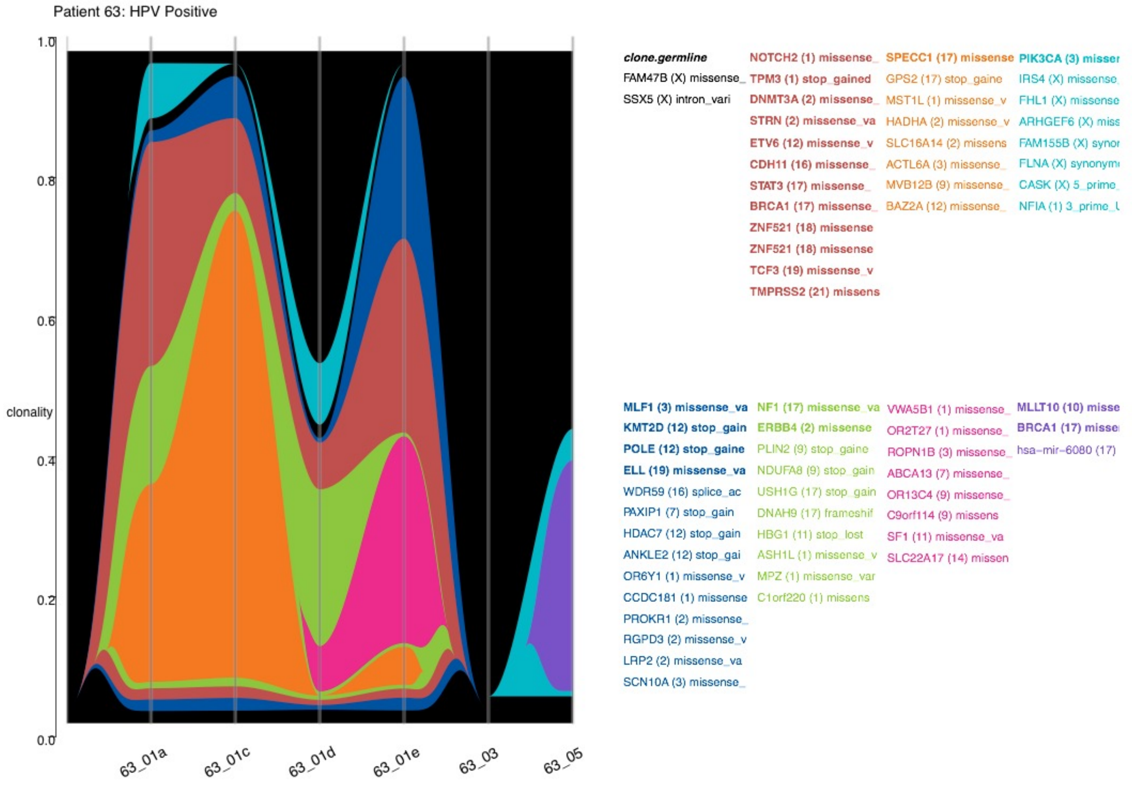


Figure 44: Riverplot for patient 63 (HPV positive) demonstrating the major clone and subclones detected across all regions sampled. Tissue samples are displayed across the x-axis. Sample names ending in a letter indicate a primary tumour sample. Samples ending in 03 indicate the whole blood control sample. Samples ending in 05 represent the lymph node metastasis.

3.12 Potential driver mutations

The definition of what mutation constitutes a driver or is actionable is contentious. The contention arises from both the inadequate binary classification of any molecular lesion, as well as the amount of evidence needed to prove 'driver' status in a particular cancer. In a disease such as penile cancer, where there is a paucity of previous basic experimental work and molecular sequencing performed on large numbers of patients, predictions have to be made in light of the lack of experimentally validated drivers. Some information may be gleaned from other squamous cell carcinomas, which may behave in an oncogenically similar way. The two most likely examples of this would be head and neck squamous cell carcinomas, and cervical squamous carcinomas, which are also driven by HPV in 50% and almost 100% respectively. For the sake of analysing the PenHet cohort, it is useful to highlight genes that are recurrently disrupted in other cancers. For this reason, a cohort of genes from the latest COSMIC[169] collection has been used to narrow down the results to focus on specific genes that may be oncogenically involved, as opposed to passenger mutations. This full curated list is displayed in the Appendix.

Drivers are overrepresented (p = 0.003) as shared mutations in the trunk of each phylogenetic tree. The proportion of driver mutations present in the trunk of each tree and the expected amount based on the number of mutations in the trunk and branches are displayed in Table 9. A significantly higher proportion of driver mutations were found in the trunk than expected based on the total number of mutations, 38% versus 33% respectively (X^2, p=0.003). When performing subgroup analysis based on HPV typing, there was less power to detect a statistical difference. There was only a significantly higher proportion of drivers in the phylogenetic trunks from the patients who were HPV negative (p = 0.003), not in those patients who were HPV positive (p = 0.064). Thus, a significantly higher proportion of drivers are found in the phylogenetic trunks from patients who were HPV negative compared with those who were HPV positive. One potential biologically explanation is that early oncogenesis in HPV-driven cancers may be more dependent on the initiating activity of the HPV oncogenes E6 and E7, which may be sufficient to drive early disruption of cell cycle controls without needing the loss of function mutations in key tumour suppressor genes and the gain of function mutations.

Table 9: Comparison of proportions of mutations in the phylogenetic trunk for drivers and non-drivers in each patient and by HPV status. X^2 test performed.

Patient identifier	Proportion of mutations in trunk	Proportion of driver mutations in trunk	p -value
mbpc39	0.37	0.41	0.10
mbpc45	0.26	0.40	0.02
mbpc49	0.39	0.46	0.01
mbpc51	0.07	0.09	0.35
mbpc63	0.49	0.49	0.63
mbpc64	0.34	0.25	0.73
mbpc66	0.11	0.26	<0.002
mbpc79	0.05	0.00	0.77
HPV positive	0.36	0.39	0.06
HPV negative	0.28	0.36	0.003
Combined	0.33	0.38	0.003

These results demonstrate that the early mutations are more likely to be drivers than later mutations. This significance level is increased when only examining the patients who are HPV negative (patients 39, 45, 64 and 66) and is no longer significant when examining those HPV positive patients. One hypothesis which this work generates is that the oncogenic HPV proteins are vital for the early stage of penile squamous cell cancer development, but a greater proportion of mutated drivers are required for the later stages of heterogeneous development.

The mutational copy number is the number of alleles harbouring a specific mutation. The mutational copy number can be used to time events with higher resolution than the truncal versus branch timings mentioned previously. If a mutation occurs in the trunk of the phylogenetic tree, the molecular event can be pinpointed to have occurred either before or after copy number and loss of heterozygosity changes. If a mutation occurs prior to a copy number gain, then the mutational copy number will increase in proportion to the copy number gain, as exemplified in Figure 45. The same is seen during whole genome doubling events. Methods to estimate this are explained in the Methods (Chapter 2).

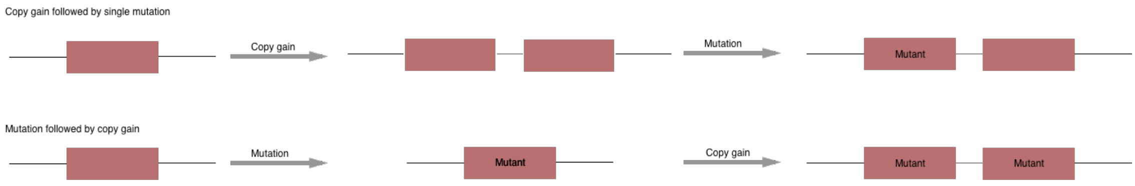

Copy gain followed by single mutation
Copy gain
Mutation
Mutant
Mutation followed by copy gain
Mutation
Mutant
Copy gain
Mutant
Mutant

Figure 45: Illustration to show the mutational copy number is dependent upon the order in which the copy gain and mutation occurred. Brown boxes represents a segment of the genome.

3.13 Timings of discovered drivers and therapeutic actionable mutations

Very few recurrent drivers were found throughout the PenHet cohort. Only 29 genes harboured mutations at a frequency above 20% across the PenHet cohort. Twenty per cent was chosen as it would require a mutation to be present in samples from more than one patient and would also enable the analysis to focus on a smaller number of most recurrent mutations. Of these 29 only four are classified as potential drivers: *TP53, PIK3CA, FAT1* and *HNRNPA2B1*. Other important actionable mutations, of which treatments are currently being considered at a very early phase, include *EGFR, cMET* and *cMYC*. Although one current registered study is looking at the use of cMET and cMYC inhibition in penile cancer, no mutations were identified in these genes in the PenHet cohort. Low frequency mutations were also found in the following 'actionable' mutations: *ERBB2*, *ERBB4*, *MTOR*, *PIK3CA*, *BRCA1*, *BRCA2*, *CDKN2A*, *RB1*, *MTOR*, *NF1*, *NOTCH1*, *NOTCH2*, *DNMT3A*, *APA*, *EPHA3*, *GNAQ*, *PIK3R1*, *SMARCA4* and *STK11*.

The use of a biological therapy, which targets a molecular pathway and oncogenic driving mutation, will only be successful if the mutation is present in a large number of tumour cells. As explained in the Introduction (Chapter 1), if an agent is only targeting some of the cancer cells then it is very likely that the cancer cells not harbouring the particular mutation will be unperturbed by the biological agent, resulting in relapse and progression of the disease. It is therefore ideal that the mutation being targeted is present in all cells of the cancer. For this to be the case it would have had to have come about during the last ancestral clone. It will therefore be present in the trunk of a regional phylogenetic tree. Any mutation only present in a terminal branch will belong to a cancer subclone and would have come about at a later time point in the evolution of the cancer.

All driver and actionable mutations listed above were interrogated to assess whether the mutation was shared across regions, and if so, whether this increased the likelihood that it was an early clonal event. These drivers and actionable mutations are annotated on the phylogenetic trees in Figure 35 to Figure 44.

Only two mutations from this group of driver and actionable genes, *TP53* and *PIK3CA*, were found to be early truncal events in the PenHet cohort. Not only were they truncal events but they were also found in a mutually exclusive manner with *TP53* present in the trunks of all HPV negative patients and *PIK3CA* found only in the trunks of HPV positive patients. As seen in the riverplots these early truncal events are also predicted to have taken place in the first ancestral clone.

TP53

Patient 39 has a truncal mutation of *TP53*. Sample 39_05 has a mutation copy number of 2, with loss of heterozygosity and subsequent copy gain at that position. This indicates that the mutation was present before the loss of heterozygosity event. If it had occurred after the copy number aberration then it would only have had a mutational copy number of 1. For patient 45 there has been a loss of heterozygosity, a copy gain and genome doubling event across the samples. In these samples the mutational copy number is 4, indicating that once again the mutation of *TP53* predated all these other copy number aberrations. Sample 64_05 also has a raised mutation copy number for *TP53* of 2. In this case, there was also a genome doubling event, indicating that a *TP53* mutation occurred before the genome doubling event. For patient 66 there was a shared loss of heterozygosity and copy number gain of *TP53*. The mutational copy number of *TP53* in these cases was also 2, indicating the mutation occurred before the copy number aberration and loss of heterozygosity events.

PIK3CA

Patients 51 and 63 both had shared early clonal mutations in *PIK3CA*, present throughout the tumour samples. However, in the case of patient 51 it could be demonstrated that this occurred prior to the copy number aberrations over the same genomic location. The mutational copy number of this mutation was raised to 2 in the two samples where a copy number gain was demonstrated over this mutation. This indicated that the *PIK3CA* mutation occurred prior to the copy number aberration, reaffirming that this is an early event in HPV positive patients. Previous analysis of HPV positive head and neck squamous cell carcinoma reveals that APOBEC mutational signature is associated with HPV positivity, just as in the PenHet cohort[170]. The authors also found an association between *PIK3CA* hotspot mutation and APOBEC mutagenesis, implicating APOBEC activity as a key driver of *PIK3CA* mutagenesis and HPV transformation[170].

EGFR

Mutations in the tyrosine kinase receptor *EGFR* are the first mutations targeted by a biological agent in a new phase two trial testing daconitinib in penile cancer[171]. However, in the PenHet dataset mutations in *EGFR* were only found in single regions of the primary tumours of two patients. Therefore, in our small dataset there is no evidence of mutations in *EGFR* being an early clonal/truncal event. This is in contrast to the recently published TRACERx non-small cell lung cancer trial that found mutation in *EGFR* to almost always be a clonal event[172]. The cancer cell fraction for the *EGFR* mutants in sample 64_01e was 19% compared with 100% for the mutant in sample 45_01e.

MTOR

The PI3K-AKT-mTOR pathway is frequently mutated in cancer and interacts with *PIK3CA*, discussed above. Mutations in *MTOR* were only found in two out of four patients who were HPV positive. In patient 51 the mutations were present throughout all regions including the lymph node metastasis. This therefore indicates that it is was an early event with a cancer cell fraction of approximately 100% throughout all regions. However, in patient 49 the mutation was shared throughout the primary tumour but not in the lymph node metastasis. As can be seen from the regional phylogenetic tree, the lymph node for this patient seems to have formed at a very early stage with very few mutations shared with the rest of the trunk of the primary tumour. The cancer cell fraction within the primary tumour was 100%. Upon closer inspection of the lymph node metastasis there is loss of heterozygosity of *MTOR*. This may explain why the mutation was not found in the lymph node metastasis. There is also evidence that the mutation occurred prior to the whole genome doubling of sample 49_01b as the mutational copy number was 2. All this evidence suggests that mutations in *MTOR* in both patients were early events, occurring before genome doubling and potentially before the progression of the lymph node metastasis.

APOBEC mutational signatures

The proportion of mutations with the APOBEC mutational signatures were compared for clonal versus subclonal clusters of mutations. Although there was only APOBEC mutational data for four patients, there was a trend ($p = 0.093$) for increasing APOBEC mutations in subclonal versus clonal mutations. There was a mean proportion of APOBEC signature mutations of 0.077 versus 0.298 in clonal versus subclonal mutations respectively. This indicates that over time APOBEC activity may increase and be responsible for a larger proportion of mutagenesis, providing a substrate for further cancer evolution. This would have to be validated initially in a larger cohort and potentially functionally, with samples taken longitudinally during penile cancer development. Furthermore, presence of APOBEC mutation signatures 2 and 13 in clonal mutations indicates that the APOBEC mutations may represent an important driver of mutagenesis early on in penile cancer development.

3.14 Hypothesised relative timings of key molecular aberrations

Although all these experiments are based on a sample of size of just 48 samples from eight patients, specific trends and correlations are observed, which can be used to produce a working hypothesis of the relative timings of events during oncogenesis.

This chapter provides evidence that in HPV patients, infection and integration of HPV is an early clonal event, which one can hypothesise may result in early mutational signatures for APOBEC activity. Other early events include clonal mutations in *MTOR* and *PIK3CA*. Clonal early mutation in *PIK3CA* appear to occur before CNA events in the same genomic region. Later events include amplification and mutation of *ERBB2* as well as mutations in *DNMT3A* and *EGFR*. Genome doubling was found as both an early and late event.

Regarding HPV negative patients, early clonal mutation of *TP53* was a universal event. Where genome doubling occurred, clonal mutation of *TP53* appeared to take place before the genome doubling or CNA events. In addition, there were instances of early clonal amplification of *CCND1*, *MITF* and *NFIB*. Later events include amplification and mutation of *EGFR*. Genome doubling was found both as an early and late event.

3.15 Discussion

Intra-tumour heterogeneity has previously been demonstrated in a range of cancers, including lung, renal cell, head and neck, breast, and colorectal carcinomas. This body of work is the first time that intra-tumour heterogeneity has been evaluated in penile squamous cell carcinoma. Furthermore, it is also the first time that it has been evaluated in the context of HPV infection. The following two chapters will expand upon these results further by integrating these data sets with methylation and expression data.

3.15.1 Mutational load

Whole exome sequencing had been performed previously for this disease. However, the previous analysis was performed at a depth of only 60x, which when considering the high proportion of normal and immune cell contamination, the sensitivity for detecting single nucleotide variations is low. With a depth of 60x and an estimated 50% contamination, there would only be enough depth of sequencing to detect a minimum variant allele frequency of 20%. Furthermore, the mutations that are detected will disproportionately be those that have a high variant allele frequency, and thus truncal/clonal events are more likely to be detected than branch/subclonal events. The analyses completed in this thesis were performed at a median depth of 100x, enabling more sensitive detection of both clonal and subclonal events. Indeed, the number of mutations detected per tumour increased from 38 to 90.

Compared with many other cancers, there were only a few recurrent mutations across the cohort. This is despite the mutational load of 1.4 mutations/Mb. Although, this mutational load is not as high as that of melanoma, lung or bladder it still places on the higher mutational load end of the spectrum of cancers (Figure 15). One explanation for this is that a large proportion of the mutations encountered are passenger mutations created by the mutagenesis driven by APOBEC and the overall increase in genomic instability due to the loss of *TP53* and activation of *PIK3CA/MTOR*.

There is an association between mutational load and efficacy of immunotherapies in treating solid cancers[173]. For the very first time, the results presented above provide evidence that a proportion of penile squamous cell carcinoma carries a relatively high mutational load (>5 mutations / Mb). Furthermore, the sequencing data suggests that there is a large immune cell component to the cancer, which can also be used as a biomarker for immunotherapy success[174]. The extent of microsatellite instability (MSI) was also assessed and found to have no level or very low levels of microsatellite instability. MSI is very strongly associated with immunotherapy success and has therefore been the first biomarker approved by the FDA across solid cancers as being sufficient evidence to start immunotherapy. This new way of approving medications across cancers negates the need for individual cancer-specific clinical trials, and this could potentially revolutionise the availability of treatments for rare cancers with a lack of associated clinical trials. Considering the lack of efficacious systemic therapy for this cancer, this new knowledge should act as the preliminary evidence necessary to warrant the first small scale phase 2 trials of immunotherapy in this cancer.

3.15.2 Human papillomavirus (HPV)

Despite this experiment being designed to only capture human genome exons, large numbers of reads mapping to the HPV 16 genome were discovered. Furthermore, concatemer reads were found spanning both viral and human genomes, indicating points of viral integration. Subsequently the presence of concatemers was shown to have a complete positive association with APOBEC activity and the resulting mutational signature number 2. In addition, for the patient (79) which had the largest number of concatemer reads, the same integration site (Chromosome 19 13122530 *NFIX*) was found for all regions of the cancer including the lymph node metastasis. This indicates that the integration of HPV is an early/truncal event, occurring well before the formation of the metastasis. In addition, the two regions that are most closely related phylogenetically (79_01b and 79_01c) both share a branch on the phylogenetic tree as well as a further second shared integration site on chromosome 17, as seen in Figure 18. This suggests that HPV integration was an early clonal event in patient 79. In the remaining HPV positive tumours, HPV integration did not appear to be clonal. This raises the question as to whether HPV needs to integrate into the host genome to drive oncogenesis, or whether exosomal expression is sufficient. However, these data should be caveated by the fact that the sequencing performed for this thesis was limited to exons only, and therefore only encompasses approximately 2% of the genome. It is therefore possible that clonal HPV integration sites exist within the other tumours' regions, outside sites captured during these experiments.

3.15.3 Intra-tumour heterogeneity (ITH)

Assessment of ITH is a useful endeavour as it can be a prognostic indicator[175], a biomarker for treatment response, it can be modelled to demonstrate the clonal status of actionable mutations, and used as a method for the relative timing of molecular events throughout the oncogenesis of a cancer[66]. As demonstrated in this chapter, and again in the next two, a large proportion of mutations found are heterogeneous in nature. This is a negative prognostic indicator[175]. It has previously been hypothesised that the presence of ITH may result in increased mutational diversity and the generation of neoantigens, which could give rise to susceptibility to immunotherapy treatment{JamalHanjani:2013jy}. Further work needs to be undertaken to evaluate the presence of neoantigens and neoantigen ITH.

Evidence from this chapter can be used to hypothesise the timing of molecular events which encompass the oncogenesis and development of penile squamous cell carcinoma. Each SNV can be categorised into an early or late event depending on whether it occurs in the trunk or branch of the phylogenetic tree, displayed above (Section 3.11). It can also be categorised as clonal or subclonal with clonal events originating at the point of the last common ancestor. Within the trunk of these trees, relative timings were also calculated for some of the key drivers by integrating copy number and genome doubling data with the calculated mutant copy number. Genome doubling and copy number events occurred early and late, but these individual events can be used to time SNVs. Before integrating this data with those from the next two chapters, my current hypothesis is that chronic HPV infection results in integration of HPV and expression of HPV oncogenes. These result in the disruption of cell cycle regulators and the activation of APOBEC enzymes, which in evolutionary terms were likely a viral defence mechanism. APOBEC enzymes result in extensive mutagenesis, particularly in key oncogenes of the *PIK3CA/MTOR*, where mutations took place in an early/clonal manner prior to CNA and genome doubling events. Further disruption of cell cycle regulators and increasing accumulation of APOBEC mutagenesis results in a ballooning of mutations, many of them passenger and non-functional initially. Ultimately, further driver mutations accumulate in a subclonal branched evolution manner in genes such as *EGFR*. For HPV negative disease it is not HPV/APOBEC that is driving the mutagenesis, but may be the early clonal mutation of *TP53*. The next two chapters will assess how methylation and gene expression data can reinforce or change this hypothesis.

3.16 Conclusions

In summary, penile squamous cell carcinomas are heterogeneous with distinct pathways driving the disease depending on whether they have been driven by HPV or not. HPV appears to be associated with APOBEC activity and the presence of mutations in *PIK3CA*/*MTOR* pathway, while HPV negative disease is associated with *TP53* mutations in a mutually exclusive manner. Both HPV positive and negative disease contain a relatively high tumour mutational load. Extensive intra-tumour heterogeneity is seen, with predicted branched evolution, and this has now been modelled to produce a clonal and subclonal structure. Early clonal drivers appear to be *TP53*, *PIK3CA* and *MTOR*, while mutations in *EGFR* appear as late subclonal mutations. Furthermore, the subclonal nature of mutations in *EGFR* predicts that any initial therapeutic benefit to EGFR/tyrosine kinase inhibitors may be short lived.

4 Methylation

4.1 Introduction

Very few recurrent drivers were found when undertaking deep whole exome sequencing across the PenHet cohort. This was especially the case when only considering clonal early driver mutations. An alternative method of subverting the expression of tumour suppressor and oncogenes includes epigenetic dysregulation. One epigenetic method, previously demonstrated to be heavily disrupted in penile cancer, is the methylation of CpG[44]. Differentially methylated positions (DMPs) are single CpG sites that are hypermethylated (show a gain in methylation) or hypomethylated (show a loss of methylation) compared with a reference. These positions can be used to hypothesise which genes may be regulated or controlled by DNA methylation changes during oncogenesis. Furthermore, the quantity of changes gives an indication of the extent of methylation changes present in a particular cancer.

Methylation changes are more likely to have a biological effect if multiple DMPs are present in close proximity. Typically, hypermethylation changes in the promoters of genes can turn off or reduce gene expression, while the opposite is true in gene bodies. DMPs which are in close proximity to one another can be coalesced into regions termed differentially methylated regions (DMRs). Therefore, the presence of a DMR is potentially more likely to be significant biologically than the presence of a single DMP. Ultimately, the presence of DMPs and DMRs can be associated with gene expression data to give further evidence as to which methylation changes may have a biological effect. This will be accomplished in the next chapter (Chapter 5) utilising RNA-seq data to integrate both the methylation and whole exome sequencing data.

Our previously published work found distinct cancer specific methylation profiles comparing penile tumour with adjacent normal tissue[44]. To investigate the penile cancer methylation heterogeneity, methylation arrays were used to detect methylation changes in more than 850,000 individual CpG sites across the genome. This was undertaken across all patients and samples to determine the inter- and intra-tumour heterogeneity in the PenHet cohort.

Illumina EPIC methylation arrays were used to assess the methylomes of all four regions of each primary tumour as well as a matched lymph node metastasis. Samples from each patient were surgically removed and processed simultaneously as described in the Methods (Chapter 2). Matched adjacent normal tissue was used as a control for each patient.

These methylation arrays were used as an alternative to reduced representation or whole genome bisulfite sequencing. The reasons behind this are expanded upon in the Methods (Chapter 2). In summary, the methylation arrays were a cost-effective way of gaining data of the methylation status across a very large cross section of the methylome covering CpG islands, enhancers, promotors, shores, shelves and open sea. This technology has also previously been technically validated, ensuring that the results are reproducible and obviating the need for technical replicates.

Solid tumours constitute a mixture of cancer cells, normal cells, infiltrating immune cells and potential viral pathogens. Unlike in whole exome sequencing, where all somatic mutations can generally be attributable to the cancer cells, in methylation analysis, distinct profiles will be present in cancer cells, normal cells, lymphoid cells and infiltrating tumour cells. This means that a specific methylation change in a bulk tumour sample may be attributable to either a cancer specific change in a tumour squamous epithelial cell or the presence of other cells such infiltrating immune cells, which will also not be present in the ‘normal’ control tissue. Therefore, it can be difficult to detect the tumour specific methylation changes through the noise of these other mixed methylation profiles.

Changes in methylation state of an individual CpG site are inherently unstable. They can be short lived, flipping backwards and forwards in the two binary states at an individual locus. There is evidence that some of these changes at an individual CpG may in fact be stochastic in nature[176], and only after several methylation changes, occurring within close proximity, can a functional effect be produced. For this reason, there is more noise in the methylation profiles than in the genetic profiles generated from whole exome DNA sequencing. However, although there is more noise, there is the possibility of generating a signal more easily, and potentially at an earlier stage of onco-phenotypic change in a cell.

The inherent noise in the methylome analysis has the potential to limit the sensitivity for detecting small methylation changes. To reduce the potential for false positive values, attempts were made to remove all methylation changes that could be attributed to the presence of immune cells. This was achieved by creating immune methylation profile controls from previously published datasets of cell sorted immune cells and normal pelvic lymph nodes as discussed later in this chapter.

Strict criteria were used to filter methylation changes to find the most significant changes and to conservatively estimate intra-tumour heterogeneity across the regions from each individual. Further details on the criteria chosen can be found in the Methods (Chapter 2). The criteria chosen included:

- Minimum beta methylation difference between tumour and normal was 0.20
- Minimum beta methylation difference between tumour and immune controls was 0.10
- Minimum beta methylation difference between tumour and normal lymph nodes was 0.10

In this chapter, the epigenomic landscape of penile squamous cell carcinomas is described in the context of DNA methylation changes: at the level of CpG loci, DMRs (coalesced CpGs), and genomic features (CpG island). Furthermore, distinct methylation profiles of infiltrating immune cells as well as those infected with HPV are detected. Following this, an analysis of methylation intra-tumour heterogeneity is undertaken, elucidating the genes and pathways that are differentially methylated across all regions in the tumour compared with those defects which are not shared, or may even be unique to a particular region. Methylation variants which are conserved across all regions sampled are termed truncal, whereas those which are not, are termed branch methylation variants. Truncal variants are likely to have existed at the time of the last common ancestor and hence likely to be clonal in origin. It is for this reason that intra-tumour heterogeneity analysis can provide evidence for the variants present at the time of this last common ancestor, which are thought to be vital in driving the oncogenesis. In the next chapter, the methylation changes will be integrated further with expression data to determine what changes are likely to contribute to a change in expression and potential functional impact.

4.2 Quality control and array output

All samples passed the quality control steps discussed in the Methods (Chapter 2). Each raw data file resulted in a beta methylation ‘call’ at 866,838 CpG loci for each sample. These were then filtered in several steps to remove poorly performing probes. At each step if a poorly or potentially poorly performing probe was found in one sample then it was removed from all samples as follows: the first step involved removing all probes with a poor detection p value, which totalled 11,290 (1.3%) probes, leaving 855,548 probes for downstream analysis. There is significant evidence that any CpG within five bases of a single nucleotide polymorphism (SNP) on a probe can result in a non-specific and thus inaccurate methylation signal. The second step, therefore, involved removing all instances of CpGs within five base pairs of an SNP as previously investigated by Zhou et al[116]. These probes totalled 79,976 (9.3%) leaving 775,572 probes per sample left for analysis after removing for poor detection p values and closeness to SNPs.

4.3 A suitable control for lymph node metastases

No normal lymph nodes were resected from patients in the PenHet cohort. Therefore, when conducting the analysis above, all tumour samples irrespective of whether they originated form the primary or metastasis were compared with the adjacent normal sample taken from the penis. This can result in the loss of methylation signal from the lymph node metastases, so an external cohort of normal lymph node samples were needed to act as a surrogate control for the lymph node metastases. A control set of three normal lymph nodes was obtained from a study of prostate cancer where normal pelvic lymph nodes were removed as part of surgery. These lymph nodes were subjected to methylation analysis by using 450k Illumina methylation arrays. The resultant iDat files were obtained from GEO accession GSE73549 and processed together using the same pipeline created for the EPIC arrays used in the PenHet cohort. Please see the Methods (Chapter 2) for detailed methods and batch correction. This enabled metastatic lymph node methylation beta values to be compared with the pelvic normal lymph nodes and processed using otherwise identical methods.

4.4 Global methylation analysis

Global methylation profiles of all tumour and 'normal' control samples from the PenHet cohort were compared to determine if there were any differences in overall distribution of methylation values. The 'normal' control samples consisted of tissue adjacent histopathologically normal samples and histopathologically normal pelvic lymph nodes. This was accomplished by comparing the density plot of each tissue type by averaging the plots for each tissue type using the mean and plotting using the function sm.density.compare from the sm R package as described in the Methods (Chapter 2). Statistical testing to assess whether the density plots were significantly different was undertaken using the bootstrap hypothesis permutation function from the same R package. As demonstrated in Figure 46 there is a significant difference ($p < 0.001$) between the density plots of cancer samples and control samples. This can be explained by a reduction in the number of methylated loci (approx. 0.8-1) and the concomitant increase in the amount of intermediate methylation present in the cancer samples (approx. 0.2-0.8). Methylation at a single CpG is a binary event. However, as bulk tissue samples were used, intermediate methylation can come about due to both impurities between tissues samples – for example, tumour versus normal or tumour versus immune cell mix – as well as intra-tumour heterogeneity. As assessed in Section 4.10.3 there is intra-tumour heterogeneity found within all tumour samples. In addition, there is also contaminating immune and normal stroma within the tumour samples as assessed in Section 4.6.

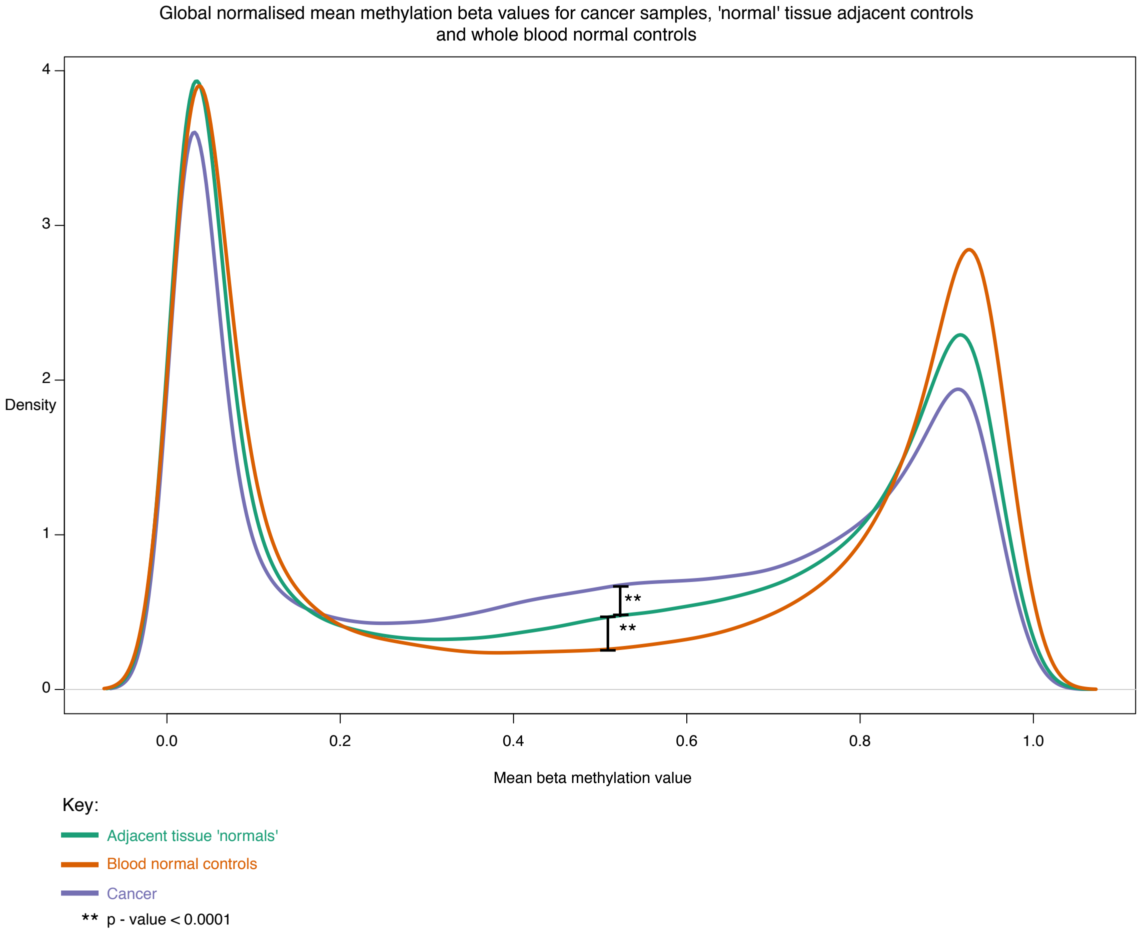


Figure 46: Density plot comparison of global methylation beta values between cancer and normal control samples.

For the above data, it appears there are distinct differences in global DNA methylation between penile cancer and adjacent normal tissue. In order to determine if these global methylation differences could accurately differentiate disease state, I compared the methylation profiles of all the samples in the PenHet cohort. Multi-dimensional scaling (MDS) plots were generated as described in the methods (Chapter 2).

Figure 47 displays an MDS plot of the 1,000 most variable probes across the primary and normal control samples. The top 1,000 most variable probes were chosen to maximise the differences between samples in a computationally effective manner. This enables the samples to be viewed spatially based on how similar one sample is to another. Further explanation of this method can be found in Chapter 2. Three distinct clusters are formed, all 'normal' control samples cluster together, as expected, with relatively limited variability (Figure 2). However, the primary tumour samples appear less homogeneous than normal, as they are dispersed more widely in the MDS plot, clustering into two broad groups of samples. These groups reflect the status of HPV infection, and 100% of samples cluster according to HPV infection/APOBEC mutational signature status (as defined by presence of HPV 16 DNA sequences discovered in whole exome sequencing as well as APOBEC mutation signatures). Figure 48 demonstrates an MDS plot of just the primary tumour samples to demonstrate more clearly how the greatest variability of methylation values between primary penile samples can be attributed to HPV status or APOBEC activity. Interestingly, the external normal lymph nodes appear to cluster closely with normal foreskin, suggesting that the differences between tumour and normal are greater than differing normal tissues.

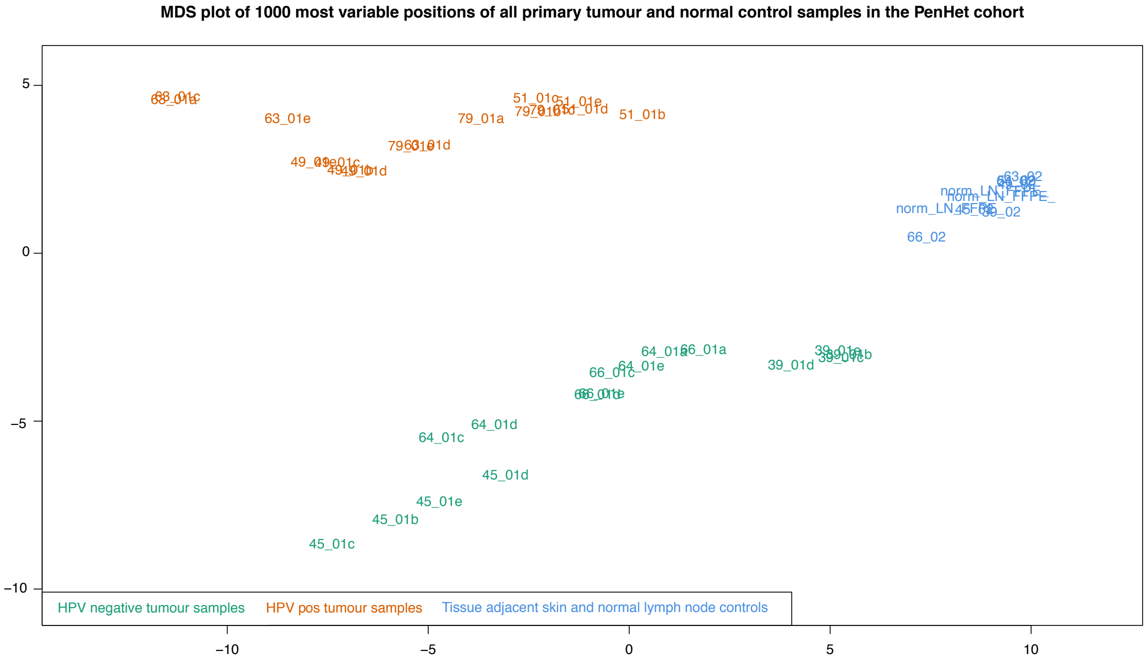


Figure 47: MDS plot of primary penile squamous cell carcinoma and control samples using the 1,000 most variable CpG positions. Sample numbers refer to patient identifiers in the PenHet cohort.

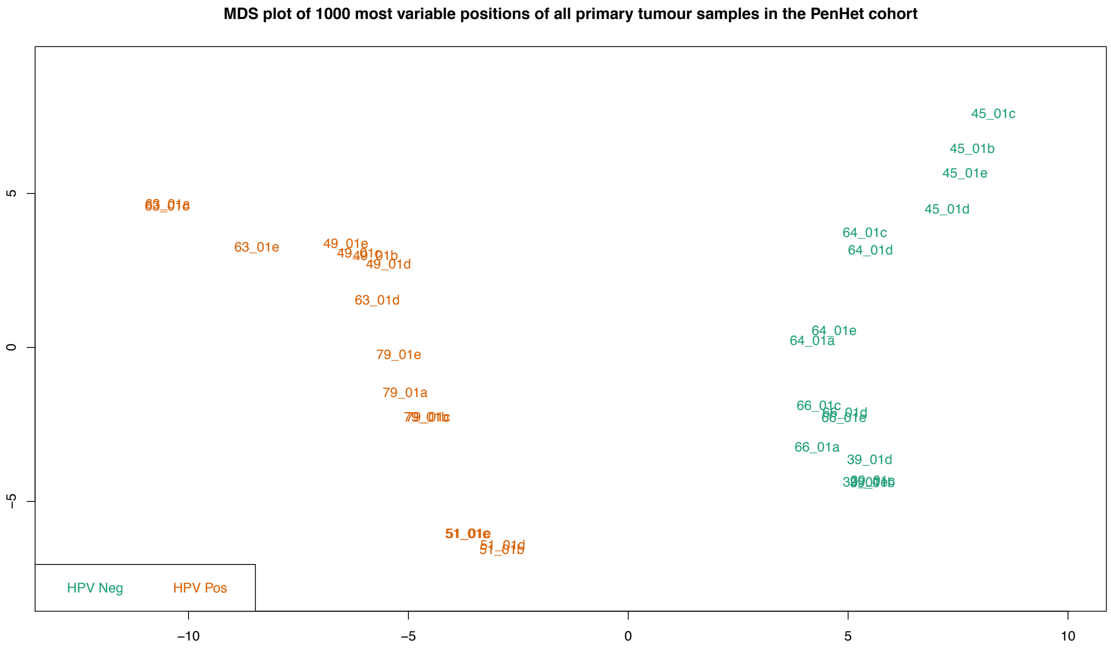


Figure 48: MDS plot of primary penile squamous cell carcinoma samples using the 1,000 most variable CpG positions. Sample numbers indicate patient identifiers.

4.5 Differentially methylated positions (DMPs)

Differential methylation at CpG sites can drive changes in gene expression[177]. These changes in expression can cause loss of expression of tumour suppressors and over-expression of oncogenes[178]. These genes, which are differentially methylated at the CpG locus, may therefore operate as methylation gene drivers. Furthermore, the DMPs assessed at every position can be compared for all regions of the tumour to assess the extent of intra-tumour heterogeneity. Differential methylation was therefore assessed between tumour and normal for every CpG on the Illumina methylation arrays, enabling the assessment of DMPs that may be involved in the oncogenesis of disease. DMPs were assessed by using the DMPfinder function as part of the 'Minfi' package[115]. This function performs an F-test to test for a statistically significant difference between the beta methylation values of tumour and control samples across a cohort. Further details on the methods for assessment of DMPs are provided in the Methods (Chapter 2).

151,597 CpGs were significantly differentially methylated before p value multiple testing adjustment. After Bonferroni adjustment, 15,076 CpGs remained significant. This set was further reduced after a filter was set that required a minimum mean change in methylation of 20%. This resulted in a set of 11,617 DMPs, which were differentially methylated between tumour and normal. Of these, 7,685 (66%) were hypermethylated and 3,932 (34%) were hypomethylated. A heatmap demonstrating these initial DMPs and proportion of hyper- and hypomethylated probes can be seen in Figure 49. Before further analysis of these DMPs was commenced, an assessment of immune cell infiltration was undertaken; if present this would introduce a major source of bias and complexity into the analysis.

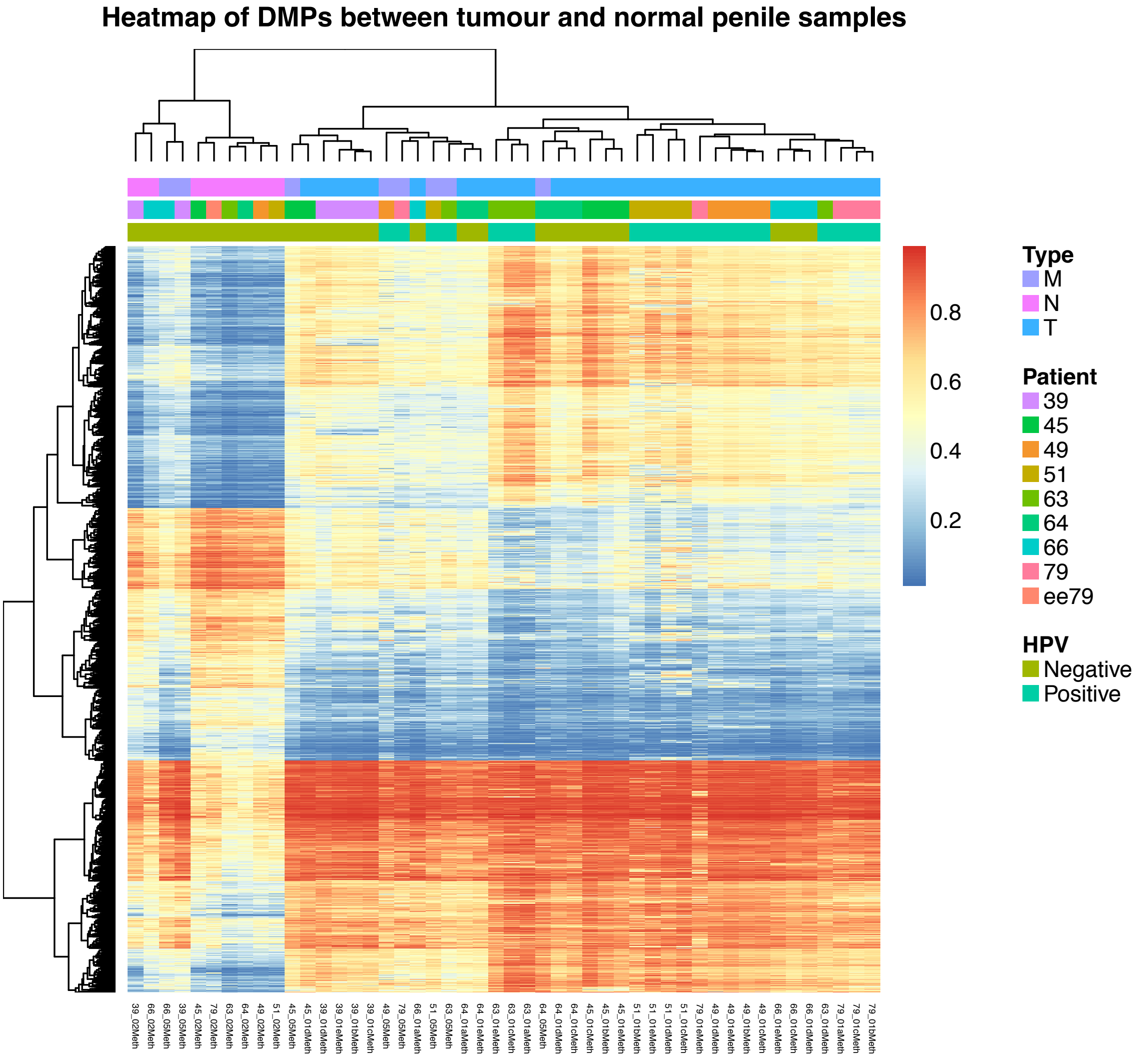


Figure 49: Heatmap of all DMPs between tumour and tissue adjacent normal samples, before any immune cell filtering for the full EPIC array dataset. Heatmap cell colours depict beta methylation scores for each DMP ranging from 0 (blue) to 1(red). Samples key: M = lymph node metastasis, T = primary tumour sample, N = tissue adjacent normal sample.

4.6 Immune cell contamination

The tumour samples used in this study represent bulk populations, and therefore contain a heterogeneous mix of cells. As a result, the mean beta value at each CpG site is an average signal encompassing the heterogeneity of cell types sampled. These cells include tumour cells, ‘normal’ epithelial cells, HPV infected non-cancer cells and an immune cell component.

Several methods exist to predict the proportion of immune cell contamination within tissue samples. However, these generally rely on a ‘homogeneous’ reference sample such as a cancer cell line, in order to define the ‘true’ cancer methylation profile. Robust well characterised penile cancer cell lines are not available. These methods also only estimate the potential proportion of immune cell contamination, and do not allow the identification/removal of specific loci at which the methylation signal is driven by the presence of contaminating immune cells[179,180]. To assess the immune cell contamination and identify the presence of differentially methylated CpGs, which are due to immune cell contamination, I identified those loci which were differentially methylated in immune cells compared with normal. The immune signature was calculated using CpGs from a 450k Illumina methylation array by comparing ‘normal’ epithelium to a panel of immune cell methylomes previously described, GEO accession number GSE35069. As the PenHet methylation data was generated using the Illumina EPIC array, the number of DMPs was reduced to only include those probes overlapping with the 450k array. The combined arrays were then processed together using the same pipeline created for the EPIC arrays used in the PenHet cohort. Detailed methods and batch correction are listed in Chapter 2. This immune signature was then compared with DMPs discovered in the PenHet cohort.

To determine if this methylation immune signature reflected true biological immune processes, the immune-related CpGs discovered were evaluated for significant (adjusted p value < 0.001) overrepresentation in GO biological pathways as described in the Methods (Chapter 2). The top 15 GO terms significantly overrepresented in this signature are displayed in Table 10. The full list of overrepresented GO terms can be found in the Appendix.

Table 10: Top 15 immune signature CpGs, significantly overrepresented in GO pathways, as assessed using the missMethyl R package.

GO Term	FDR
immune system process	7.01E-28
immune response	1.15E-24
cell activation	9.60E-20
leukocyte activation	9.60E-20
regulation of response to stimulus	9.60E-20
positive regulation of response to stimulus	3.69E-18
regulation of signaling	1.62E-17
single-organism cellular process	1.71E-17
regulation of cell communication	6.08E-17
single-organism localization	2.77E-16
immune effector process	5.47E-16
cell surface receptor signaling pathway	6.05E-16
positive regulation of biological process	7.09E-16
regulation of signal transduction	1.04E-15
regulation of immune system process	1.42E-15

There was a significant (p = 0.01, Fisher’s exact test) overlap between DMPs discovered in the PenHet cohort, which could be attributed to an immune cell contamination, Figure 50, with more than 20% (1,507 out of 7,482) of penile cancer DMPs representing potential false positives. In Chapter 5 the transcription profiling obtained through RNA-sequencing was also assessed for the presence of immune cell expression signatures. The expression data also demonstrated significant immune cell infiltration, confirming the results identified in the methylome analysis.

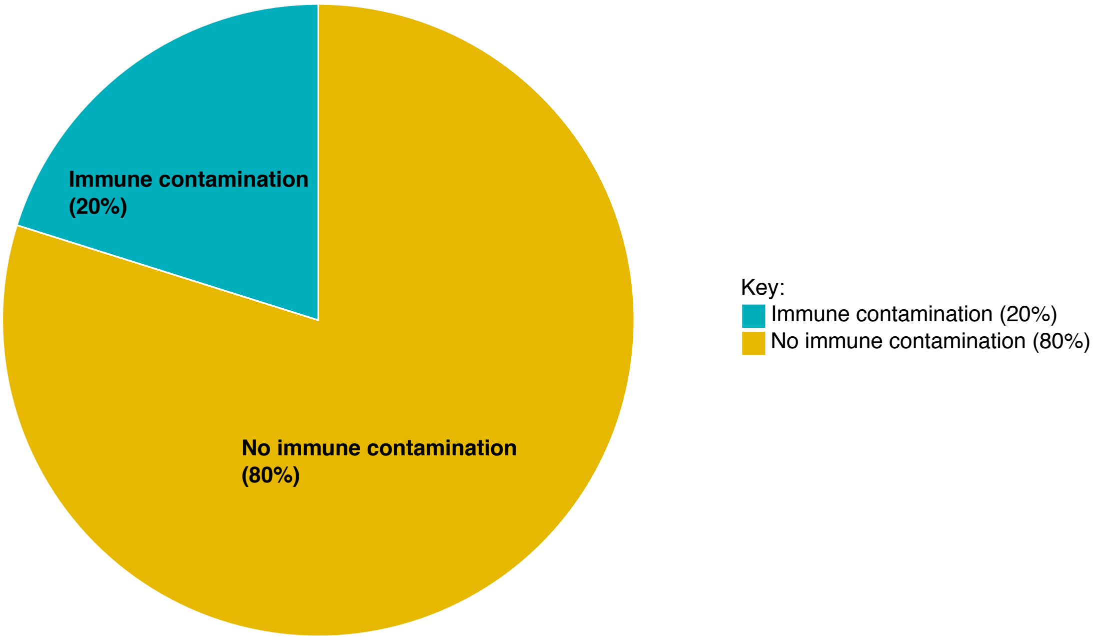


Figure 50: Pie chart displaying the proportion of significantly (p = 0.01, Fisher's exact test) differentially methylated CpGs which overlap with the immune signature.

The presence of immune cell signatures within the DMPs found in the PenHet cohort could potentially bias the results of all further analyses. Without attempting any sort of correction there is a danger that the positive signal across the cohort may in fact be attributed to differential immune cell infiltration rather than cancer cell methylation driven changes. Furthermore, there is also a danger of misattributing methylation heterogeneity differences within a patient's samples to oncological processes rather than differing immune cell infiltration.

The caveat here is that the epigenetic changes are dynamic and can be induced by the tumour microenvironment[181]. Therefore, the presence of immune cells may directly influence epigenetic changes in the tumour cells (or normal) and together may be driving part of the oncogenesis.

To define the true penile cancer methylation events on an individual sample level the methylation profiles of the immune cells were compared with those of the individual tumour samples. Only CpGs which were differentially methylated between tumour and control as well as between tumour and all immune cells were kept for further analysis.

A new F-Test was performed for each of these scenarios. This method resulted in a very large reduction, of between 23% and 84%, in the number of differentially methylated CpGs, but simultaneously improved the robustness of the results by limiting the contamination of potential immune associated DMPs. Unsurprisingly the samples with the greatest proportion of infiltrating immune cells were the lymph node metastases. However, the extent of immune cell signatures within primary tumour samples was surprising with a minimum of 20% of DMPs within all primary tumour samples likely attributable to immune cell contamination (Figure 51).

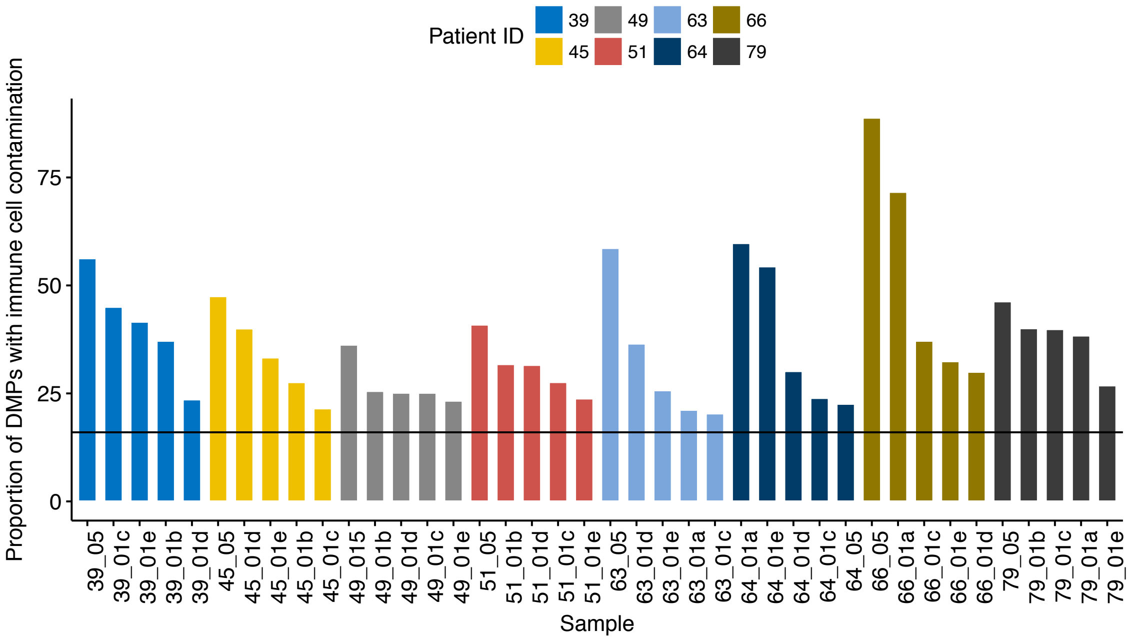


Figure 51: Percentage of DMPs with methylation immune signatures discovered for each sample, ordered by patient. Horizontal line depicts the 20% minimum level of immune contaminated DMPs discovered. Sample names with a suffix ending in a letter reflect primary tumour samples, sample names ending with the number 5 reflect lymph node metastases.

4.7 Immune corrected cancer DMPs

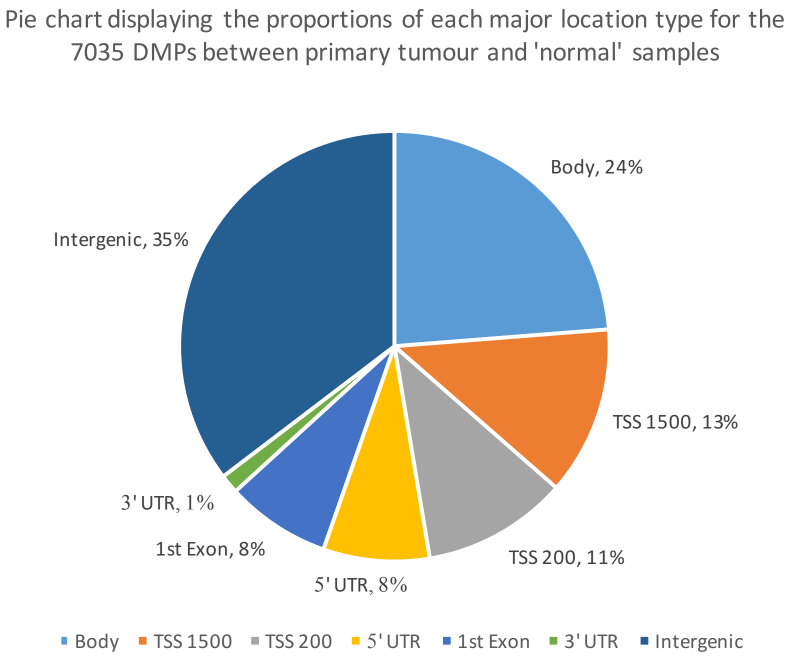
In order to ascertain which methylation changes detected were truly present in the cancer cells and were not solely caused by the presence of immune cell infiltration, immune filtered lists of penile cancer DMPs were computed between a range of scenarios, as displayed in Table 11. As demonstrated in previous studies, hypermethylation changes predominate the penile cancer methylome with 84.5% of all DMPs (5,946 out of 7,035) found to be hypermethylated between primary and tissue adjacent controls[44]. Hypermethylation remained predominant in other comparisons, as displayed in Table 11. Furthermore, there was more than six times the number of DMPs detected in HPV positive samples (3,390) compared with HPV negative samples (590), p < 0.0001). This finding is investigated further in Chapter 5, where the expression of methyltransferases is assessed in the context of HPV status.

Only two DMPs were discovered when comparing primary versus metastatic lymph node samples. This is due to the relative similarity of methylation profiles of primary and metastatic lymph node samples, where metastatic lymph nodes tended to cluster towards their matched primary sample rather than with other lymph node metastases. Despite the small sample sizes of eight lymph node metastases and three lymph node controls, 41 DMPs were discovered when comparing lymph node metastases with lymph node 'normal' control samples.

The immune filtered list of penile cancer DMPs were annotated by: genomic location, gene name, functional location, previous identification as a gene driver, and whether the gene is potentially actionable as a therapeutic target. The first analysis undertaken was the assessment of DMPs between primary and tissue adjacent control samples. The distribution of CpGs within differing genomic locations and features are displayed in Figure 52 and Figure 53.

Table 11: Immune filtered DMPs for a range of comparisons with total and percentage CpGs that were hypermethylated.

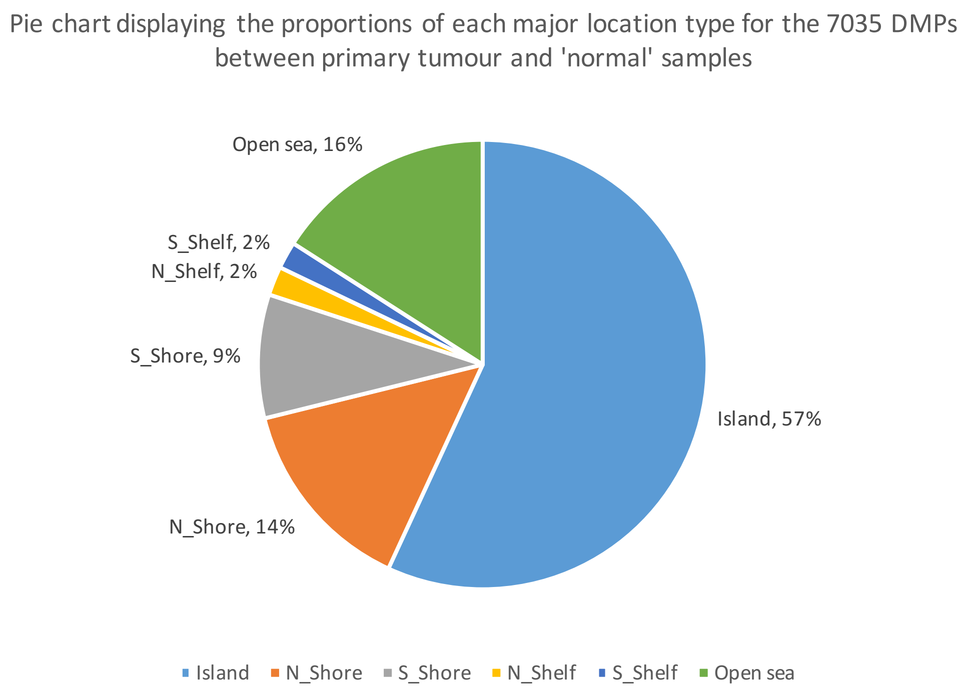
Comparison	Total DMPs	Hypermethylated CpGs	Hypomethylated CpGs	Hypermethylated (%)
Primary versus adjacent skin controls	7035	5946	1089	84.52%
Primary versus metastatic lymph node	2	1	1	50.00%
Lymph node metastases versus lymph node controls	41	11	30	26.83%
HPV positive primary samples versus matched tissue adjacent skin controls	3390	2971	419	87.64%
HPV negative primary samples versus matched tissue adjacent skin controls	590	473	117	80.17%



Pie chart displaying the proportions of each major location type for the 7035 DMPs between primary tumour and 'normal' samples

CpG Location	Distribution of significant CpG DMPs %	Expected distribution CpG %	Z-test (p value)
Body	24%	33%	< 0.00001
TSS 1500	13%	14%	0.008
TSS 200	11%	11%	0.766
5' UTR	8%	9%	0.094
1st Exon	8%	5%	< 0.00001
3' UTR	1%	4%	< 0.00001
Intergenic	35%	25%	< 0.00001

Figure 52: Distribution of DMP locations across a gene between primary tumour and normal samples.



CpG Location	Distribution of significant CpG DMPs %	Expected distribution CpG %	Z-test (p value)
Island	57%	31%	< 0.00001
N_Shore	14%	13%	0.025
S_Shore	9%	10%	0.013
N_Shelf	2%	5%	< 0.00001
S_Shelf	2%	5%	< 0.00001
Open sea	16%	36%	< 0.00001

Figure 53: Distribution of DMP locations within CpG islands, shores, shelves and open sea between primary tumour and normal samples.

Methylation changes that occur at specific gene locations appear to exert a greater influence on gene expression[182]. For instance, hypermethylation of gene promoters (consisting of TSS 1500 and TSS 200) have been previously shown to cause a loss of expression and gene silencing. Furthermore, previously published methylomes of other cancers reveal an overrepresentation of DMPs in CpG islands in comparison to shores and shelves[183]. A z-test was performed between the expected and observed locations of DMPs in the PenHet cohort to determine whether there was an overrepresentation of DMPs within islands, shores, shelves, open sea, promoters, gene bodies or intergenic regions (Figure 52 and Figure 53). Overall, a significantly higher proportion of DMPs were found in CpG islands (57% versus 31%, p < 0.001) and a corresponding significantly lower proportion of DMPs were found in open sea (16% versus 36%, $p < 0.001$) and shelves (2% versus 5%, $p < 0.001$). A similar proportion of DMPs were found in shores as expected (23% versus 23%). In terms of locations within a gene there was no significant difference in the observed or expected proportion of DMPs within promoters. However, there was significant underrepresentation of DMPs within gene bodies (24% versus 33%, $p < 0.001$) and an overrepresentation of DMPs in intergenic regions (35% versus 25%, $p < 0.001$). However, despite no overall increased representation of CpGs in promoters when assessing all DMPs, there was an overrepresentation of hypermethylated DMPs at TSS 200 (observed 14% versus expected 11%, $p < 0.001$). Conversely, when assessing hypomethylated DMPs separately, the only significant change was the overrepresentation of DMPs falling within gene bodies (observed 46% versus expected 33%, $p < 0.001$) (Figure 53).

To visualise the relationship between individual samples and their methylation state, all 7,035 DMPs were further analysed by performing supervised hierarchical clustering, using the Minkowski distances between each CpG for each sample. The majority of samples clustered by patient as demonstrated in Figure 54. At least three out of four of each clock face region clustered by patient. In addition, tumour samples clustered into two main groups independent of HPV status.

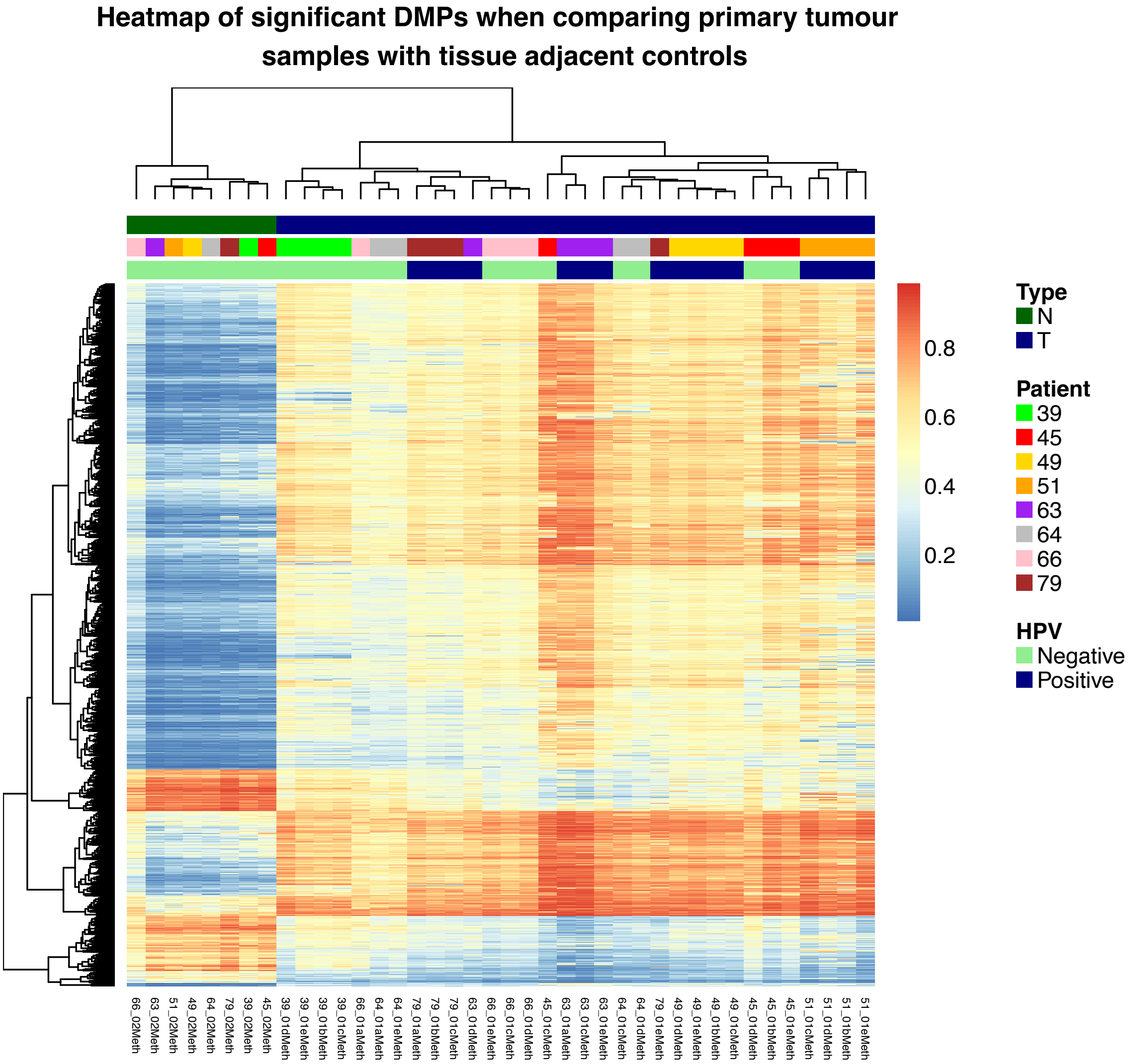


Figure 54: Hierarchical clustering of samples comparing beta methylation values of all DMPs found for the primary and matched control samples. Heatmap cell colours depict beta methylation scores for each DMP ranging from 0 (blue) to 1 (red). Samples key: T = primary tumour sample, N = tissue adjacent normal sample.

Differential methylation can cause gene silencing of tumour suppressor genes and over-expression of oncogenes[182]. Candidate methylation drivers can be investigated by determining whether any of the differentially methylated genes were previously described as oncogenic drivers. To define potential methylation drivers, the genetic location of each DMP was annotated with the corresponding gene name and compared with the curated list of putative driver genes (Methods, Chapter 2). 156 DMPs were annotated to 50 'driver' genes. Of these, 44 exhibited hypermethylation and 10 were hypomethylated. Four genes (*CAMTA1, PRDM16, PTPRT* and *ZNF521*) were associated with hyper- and hypomethylated DMPs. These DMP associated driver genes can be visualised in Figure 55. The heterogeneity of these changes will be analysed later in this chapter. However, it is interesting to note, at this stage, that within these 50 ‘driver’ genes there are CpG sites that show low inter-tumour heterogeneity, being consistently differentially methylated in primary penile cancer samples compared with their normal controls. The genomic locations of these DMPs were analysed in further detail, using identical methods to those used above in Figure 52 and Figure 53, comparing the CpG locations with those expected. As demonstrated in Table 12 and Table 13 there is an overrepresentation of potential gene driver associated DMPs within CpG islands (p < 0.0001) and an underrepresentation within open sea (p < 0.0001). Furthermore, there is an overrepresentation of DMPs located within gene bodies (p < 0.0001) and DMPs were underrepresented in intergenic regions (p < 0.0001).

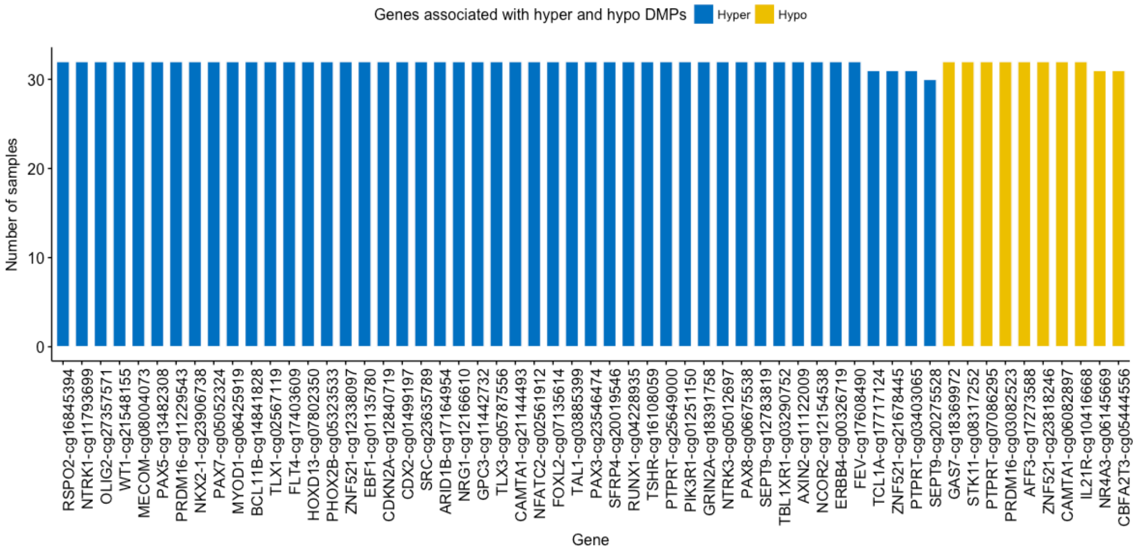


Figure 55: Frequency of CpGs differentially methylated in 'driver' genes in primary penile tumour samples compared to tissue adjacent normal controls. CpGs which are also differentially methylated between primary and cell sorted blood cells were removed as a method of reducing the effect of contaminating immune cells.

Table 12: Comparison of genomic locations features of DMPs associated with COSMIC driver genes

CpG Location	Distribution of significant CpG DMPs %	Expected distribution CpG %	Z-test (p value)
Island	60%	31%	< 0.00001
N_Shore	3%	13%	0.001
S_Shore	19%	10%	0.024
N_Shelf	3%	5%	0.368
S_Shelf	7%	5%	0.459
Open sea	9%	36%	< 0.00001

Table 13: Comparison of genomic locations features of DMPs associated with COSMIC driver genes

CpG Location	Distribution of significant CpG DMPs %	Expected distribution CpG %	Z-test (p value)
Body	57%	33%	< 0.00001
TSS 1500	14%	14%	1
TSS 200	15%	11%	0.294
5' UTR	4%	9%	0.073
1st Exon	7%	5%	0.459
3' UTR	3%	4%	0.631
Intergenic	0%	25%	< 0.00001

To examine the epigenetic changes within the 50 ‘driver’ genes, canonical gene plots were created by annotating each CpG within both genomic features (such as gene body, transcription start sites and exons), and also location within CpG island, CpG shores and CpG shelves. The beta methylation value of each tumour sample and adjacent control sample was then added to each superimposed ideogram. All of these can be visualised in the Appendix.

The methylation gene plots in Figure 56 demonstrate the variability of methylation at each CpG for the cancer samples compared with the normal controls. In each case, there is minimal variability amongst the controls compared with the stark variability amongst the cancer samples. This will be examined in further detail when assessing intra-tumour heterogeneity later in this chapter.

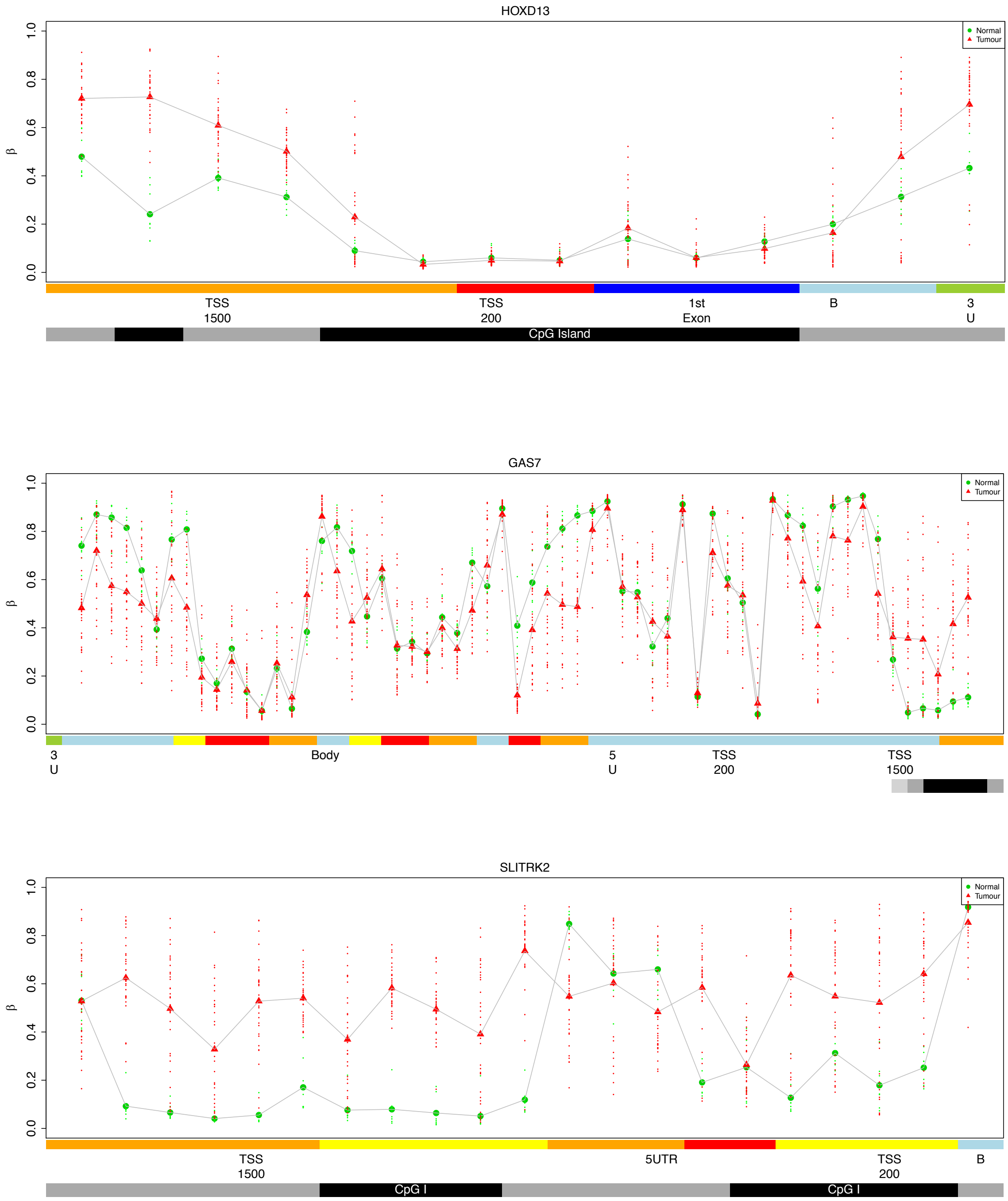


Figure 56: Gene methylation plots for HOXD13, GAS7 and SLITRK2, demonstrating the beta methylation value for each sample across a gene. Genes are annotated with CpG island location (black horizontal bar at the bottom of the figure) as well as transcription start site (TSS), 5'UTR and gene body. Each red point indicates a primary tumour sample beta methylation value at an individual locus. Each green point represents a normal control sample value.

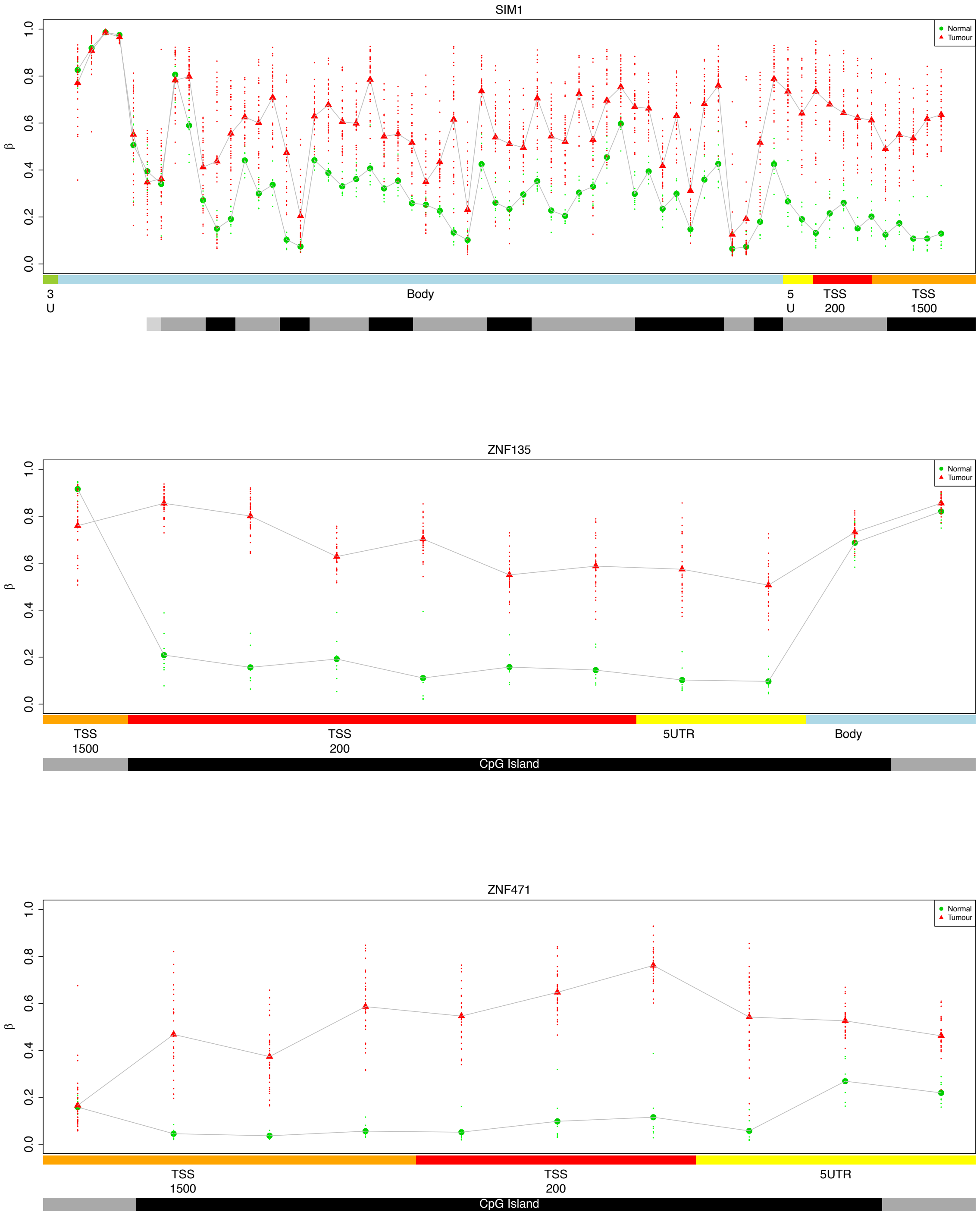


Figure 57: Gene methylation plots for SIM1, ZNF135, and ZNF471, demonstrating the beta methylation value for each sample across a gene. Genes are annotated with CpG island location (black horizontal bar at the bottom of the figure) as well as transcription start site(TSS), 5'UTR and gene body. Each red point indicates a primary tumour sample beta methylation value at an individual locus. Each green point represents a normal control sample value.

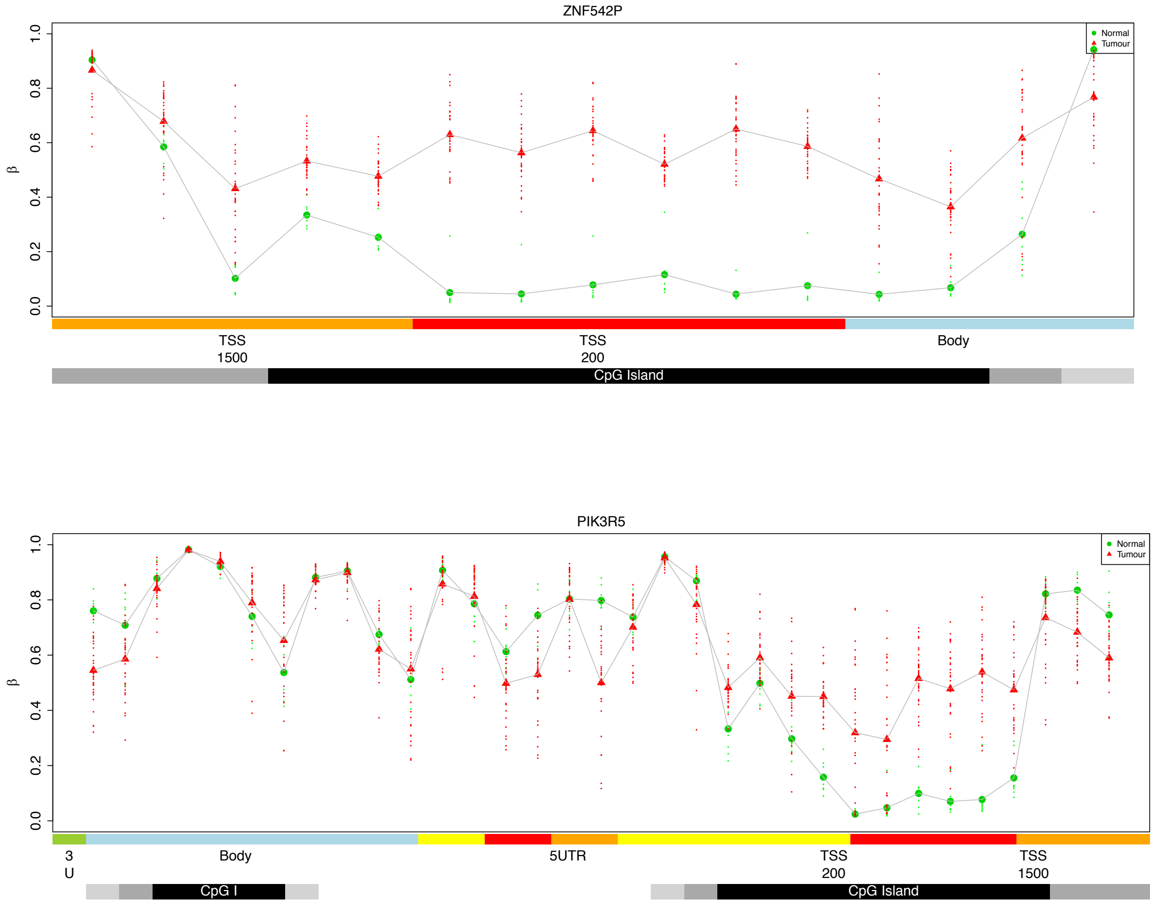


Figure 58: Methylation plot for PIK3R5 and ZNF542P demonstrating the beta methylation value for each sample across a gene. Genes are annotated with CpG island location (black horizontal bar at the bottom of the figure) as well as transcription start site(TSS), 5'UTR and gene body. Each red point indicates a primary tumour sample beta methylation value at an individual locus. Each green point represents a normal control sample value.

Although these DMPs were found to be significantly aberrantly methylated within the PenHet cohort, the significance of these findings would be increased if corroborated in an additional cohort of penile cancer patients. The DMPs identified were therefore compared with an additional cohort of penile cancer patients (PenOld), previously profiled using the 450k Illumina methylation panels[44]. The clinical characteristics of these patients are detailed in the Methods (Chapter 2). The raw iDAT 450k data files were re-analysed using the same methodology as used for the PenHet cohort. The samples collected consisted of 23 tumour samples and 15 tissue adjacent control samples. Of the 7,035 significant DMPs (which overlap on both the EPIC and 450K arrays platforms) from the PenHet cohort, 2,889 (41.1%) were also found to have recurrent changes in methylation in this independent PenOld cohort. Furthermore, when restricting the analysis solely to CpGs within genes previously described as driver genes, 26 out of 45 driver genes (57.8%) were also shown to be differentially methylated in the previously published cohort. The PenOld cohort is based on a different patient cohort in which only 34.8% of patients had higher stage disease with lymph node positive samples. This is in contrast to the PenHet cohort where all patients had lymph node positive disease. The CpGs in common between both cohorts likely represent a selection of loci, which may be epigenetic drivers of penile cancer development, irrespective of stage of disease, within penile cancer. On the other hand, the CpGs only occurring in the PenHet cohort may represent methylation changes associated with a more aggressive version of penile cancer with resulting lymph node metastases. As the comparator PenOld cohort is highly related from the same hospital and biobank, these findings will need to be validated in an independent external cohort. This corroboration of results is analysed further when assessing intra-tumour heterogeneity in the PenHet cohort below.

4.7.1 Gene set enrichment analysis

When assessing the functional impact of a large numbers of DMPs, it can be useful to group these changes into pathways disproportionately affected by methylation changes. When undertaking gene set enrichment analysis at the probe level, the uneven density of probes across different genes must be taken into account. This was accomplished by using the ‘gometh’ function as part of the MissMethyl R package (Methods, Section 2.3.6.1). The results of this analysis are displayed in Figure 59 and Table 14. As demonstrated, there is differential methylation in the generic ‘cancer pathway’ KEGG term together with specific cancer pathways including *RAP1, MAPK, RAS,* PI3K-AKT*, mTOR, ERBB, WNT,* JAK-STAT*, TP53* and *PPAR*. This demonstrates that the penile cancer methylome is perturbed throughout a large number of genes involved in the regulation of many pathways vital for cell growth and cell cycle control. Whether these epigenetic changes result in a functional change in gene expression, will be examined in the following chapter on gene expression (Chapter 5).

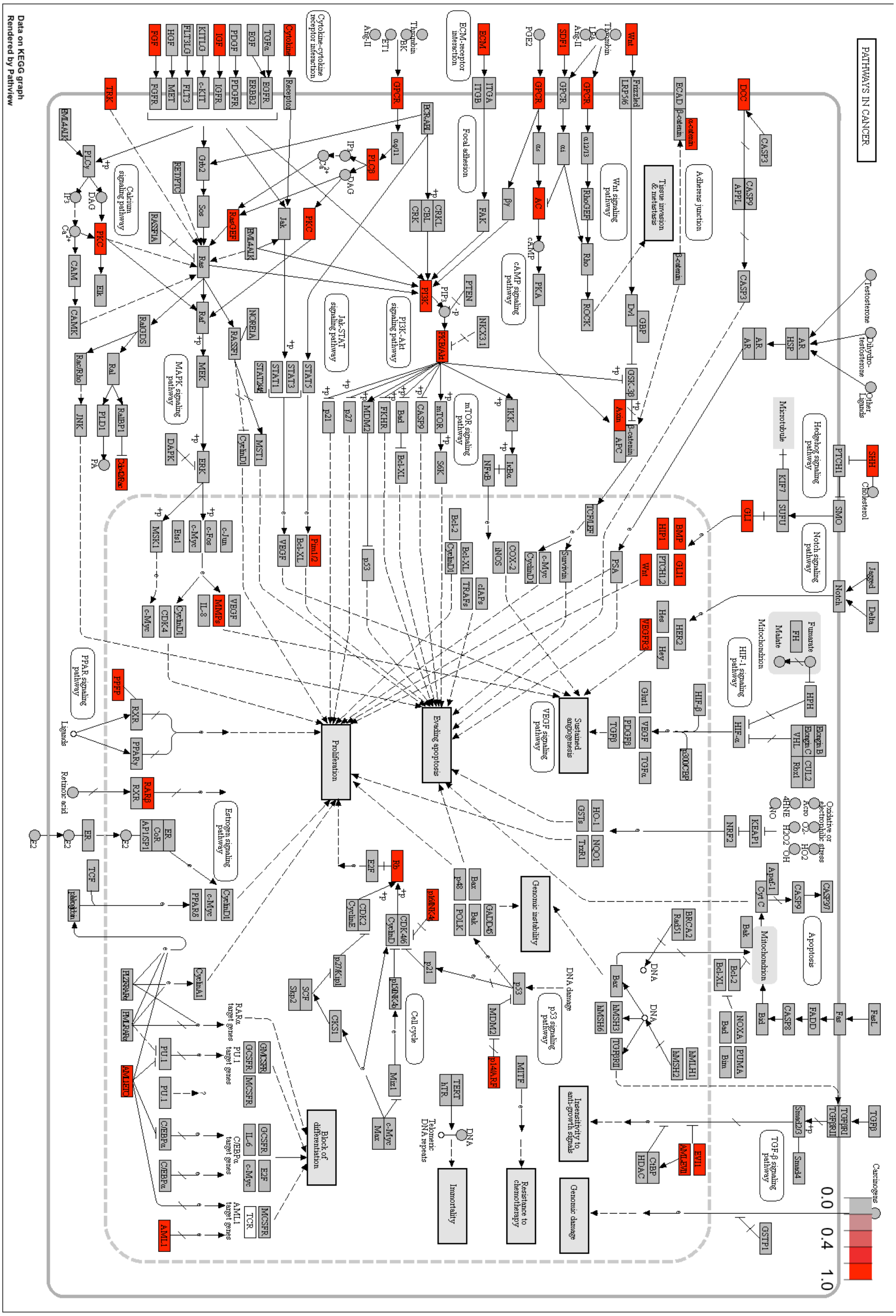


Figure 59: KEGG graph demonstrating which pathways in cancer are differentially methylated in the methylome of the PenHet cohort compares primary tumour samples with tissue adjacent normals. This figure was produced by the PathView r package after first generating the list of differentially methylated CpGs across genes and pathways utilising the 'gometh' function in the MissMethyl R package. Genes in red signal aberrant methylation..

Table 14: Top 50 KEGG pathways differentially methylated.

Pathway	N	DE	P.DE	FDR
Neuroactive ligand-receptor interaction	271	69	2.40E-60	7.76E-58
cAMP signaling pathway	198	48	3.71E-37	5.99E-35
Rap1 signaling pathway	210	41	8.13E-27	8.75E-25
Calcium signaling pathway	180	38	1.84E-26	1.49E-24
MAPK signaling pathway	293	45	6.77E-25	4.38E-23
GABAergic synapse	88	27	1.10E-23	5.94E-22
Morphine addiction	91	28	1.46E-23	6.72E-22
Nicotine addiction	40	19	1.73E-21	7.00E-20
Glutamatergic synapse	114	28	5.92E-21	2.13E-19
Pathways in cancer	515	51	5.54E-20	1.79E-18
Retrograde endocannabinoid signaling	140	27	5.19E-19	1.52E-17
Metabolic pathways	1226	68	7.83E-18	2.11E-16
Ras signaling pathway	233	31	5.33E-16	1.32E-14
Circadian entrainment	96	22	5.96E-16	1.37E-14
PI3K-Akt signaling pathway	335	36	1.04E-15	2.24E-14
Cholinergic synapse	112	23	1.66E-15	3.34E-14
Oxytocin signaling pathway	152	25	5.01E-15	9.52E-14
Inflammatory mediator regulation of TRP channels	97	20	3.23E-14	5.80E-13
Insulin secretion	85	19	3.71E-14	6.31E-13
Serotonergic synapse	112	20	4.05E-14	6.54E-13
Maturity onset diabetes of the young	26	12	4.84E-14	7.45E-13
cGMP-PKG signaling pathway	162	24	7.57E-14	1.11E-12
Dopaminergic synapse	129	22	8.75E-14	1.23E-12
Taste transduction	81	16	1.50E-13	2.02E-12
Adrenergic signaling in cardiomyocytes	143	22	3.17E-13	4.09E-12
Cell adhesion molecules (CAMs)	138	21	4.44E-13	5.52E-12
Dilated cardiomyopathy (DCM)	89	18	8.09E-13	9.68E-12
Hypertrophic cardiomyopathy (HCM)	83	17	1.39E-12	1.60E-11
Olfactory transduction	351	22	4.55E-12	5.07E-11
Amphetamine addiction	68	15	1.91E-11	2.05E-10
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	72	15	8.94E-11	9.31E-10
Long-term potentiation	67	14	1.14E-10	1.15E-09
Melanoma	76	15	1.59E-10	1.56E-09
Chemokine signaling pathway	179	20	2.04E-10	1.94E-09
Gastric cancer	153	20	5.66E-10	5.22E-09
Regulation of actin cytoskeleton	211	22	8.47E-10	7.60E-09
Phospholipase D signaling pathway	145	19	1.23E-09	1.07E-08
Cocaine addiction	49	12	1.27E-09	1.08E-08
Pancreatic secretion	93	14	1.60E-09	1.32E-08
Vascular smooth muscle contraction	120	16	1.78E-09	1.44E-08
Leukocyte transendothelial migration	109	15	2.99E-09	2.36E-08
Axon guidance	174	21	4.47E-09	3.43E-08
Cardiac muscle contraction	74	12	4.82E-09	3.62E-08
Long-term depression	60	12	6.24E-09	4.58E-08
Tight junction	169	18	7.82E-09	5.61E-08
Protein digestion and absorption	87	13	1.10E-08	7.69E-08
Alcoholism	172	17	1.28E-08	8.78E-08
Type II diabetes mellitus	46	11	1.45E-08	9.75E-08
Longevity regulating pathway - multiple species	62	12	1.79E-08	1.18E-07

4.8 Lymph node metastatic DMPs

Having defined the epigenetic changes potentially associated with penile cancer, I subsequently sought to delineate the methylation changes associated with the phenotypic characteristics of the patient. To achieve this methylation profiles were also assessed between:

- primary and lymph node metastatic tissue samples
- lymph node metastatic and lymph node control samples
- HPV positive and HPV negative primary tissue samples

In order to see if a global change in DNA methylation exists between tumour and lymph node disease, a multidimensional scaling plot was created for the 1,000 most variable probes assessed between primary tumour, metastatic lymph node and independent normal lymph node control samples (Figure 60).

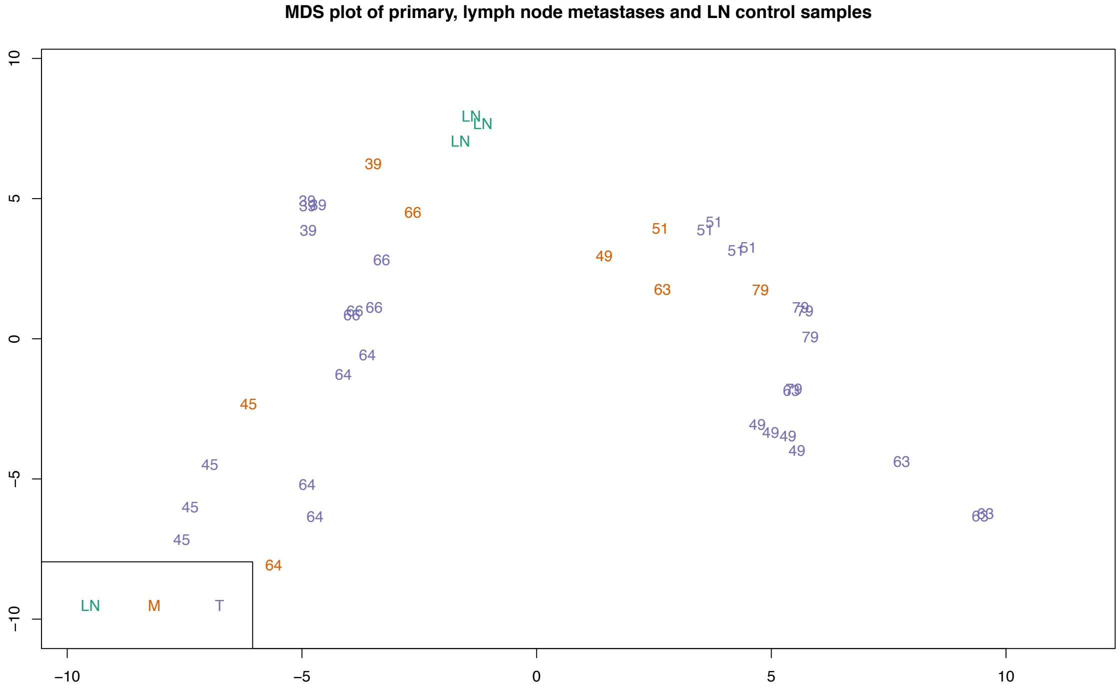


Figure 60: Multi-dimensional scaling plot of the 1,000 most variable CpG positions between primary tumour samples and lymph metastasis samples from the PenHet cohort as well as normal lymph node controls from an independent source. The samples on the left are all negative for HPV 16 infection as opposed to all the samples on the right which are positive for HPV 16 infection. M = lymph node metastasis, T = primary tumour sample, LN = normal lymph node control sample.

Figure 60 demonstrates that the metastatic lymph node samples tend to cluster primarily by patient, but in many cases they appear distinct from the primary tumour. The primary and lymph node metastasis belonging to each patient cluster by HPV status (HPV positive patient numbers included 49, 51, 63 and 79). In addition, there seems to be the propensity for samples with lower tumour cellularity to cluster towards the normal lymph node control samples. The methylation profiles of lymph node metastases, HPV positive and HPV negative patient samples are assessed below in Sections 4.8 and 4.9.

In order to assess whether there were specific methylation profiles that characterise lymph node metastases, DMPs were assessed, firstly between both primary and lymph node metastases and secondly between lymph node metastases and pelvic lymph node control samples.

There were only two recurrent CpGs that appeared to distinguish lymph node metastases from the matched primary cancer samples. Lymph node metastatic samples primarily tend to cluster by patient, therefore the epigenetic profile of the individual lymph node metastasis was more similar to that of their matched primary sample than another lymph node metastasis from an alternative patient (Figure 60). Furthermore, there was a lack of power to distinguish differences from only eight lymph node positive samples. Therefore, very few DMPs were found for this comparison.

Instead of searching for DMPs between primary and lymph node metastases, DMPs were subsequently assessed between lymph node metastases and the independent lymph node control samples. When applying Bonferroni correction and specifying a minimum mean methylation difference of 20% at each probe, 49 DMPs were discovered. The threshold of 20% difference was chosen based on previously published work demonstrating a 99% confidence of detecting a true methylation difference[184]. To reduce the potential bias of contaminating immune cells, all DMPs which were also associated with the whole blood immune controls were also removed, as discussed in the Methods (Chapter 2). This resulted in a filtered list of 41 DMPs. The discovery of 41 DMPs was impressive considering that there were only eight lymph node metastases and three lymph node control samples available to power the analysis. Thirty-eight of these DMPs were unique to the lymph node metastatic samples and not present in the primary tumour samples. The methylation differences and locations of these 41 DMPs can be visualised in a table of values in Table 15. None of these DMPs were associated with a COSMIC driver gene. Only three of these DMPs were within a CpG island and, of these, two were within the same island within the gene *ZNF582*. All three of these DMPs were also found differentially methylated between primary and tissue adjacent normal samples. This CpG island was hypermethylated and located within the predicted gene promotor at the TSS200 (Figure 61). Interestingly this gene has also been hypermethylated and implicated as a biomarker in two other cancers, cervical[185,186] and oral[187] – both of which are also squamous cell carcinomas. Additional CpGs from this gene were also hypermethylated in the primary cancer samples when compared with the tissue adjacent normal. As will be seen in the next chapter (Chapter 5), this gene is found to have reduced expression in the matched primary cancer samples, with significant further loss of gene expression in the matched lymph node metastasis samples.

Table 15: Table of DMPs between lymph node metastases and lymph node control samples. DMPs which also exist between lymph node metastases and lymph node controls have been removed. Bonferroni correction has been used to calculate the adjusted p values.

CpG probe ID	pval	Mean beta methylation Metastasis	Mean beta methylation lymph node control	Change in beta methylation	Adjusted p value	Gene name	In COSMIC database?	Location	CpG island location	DMP in primary tumour samples
cg23215407	5.81E-14	0.010647847	0.386595023	-0.375947176	2.19E-08	AMPD3	FALSE	NA	NA	No
cg04464357	4.43E-13	0.009192512	0.265426649	-0.256234138	1.67E-07	TMEM110	FALSE	NA	NA	No
cg04683551	2.77E-12	0.018666964	0.244282216	-0.225615252	1.04E-06	CDNF	FALSE	NA	NA	No
cg11767392	9.00E-12	0.020317705	0.251261218	-0.230943513	3.39E-06	LAMTOR1	FALSE	NA	NA	No
cg23795893	1.27E-11	0.031998173	0.389621061	-0.357622888	4.79E-06	PGR	FALSE	NA	NA	No
cg07116712	2.20E-11	0.946247923	0.166291799	0.779956124	8.28E-06	RP11-262A16.1	FALSE	NA	NA	No
cg27352063	4.07E-11	0.009379901	0.266565131	-0.25718523	1.53E-05	PPIF	FALSE	NA	NA	No
cg00491963	9.05E-11	0.015357663	0.279869177	-0.264511513	3.41E-05	UBE2L6	FALSE	NA	NA	No
cg02154531	9.30E-11	0.198708966	0.904363536	-0.70565457	3.50E-05	.	FALSE	NA	NA	No
cg26175287	1.09E-10	0.310078573	0.914845342	-0.604766769	4.10E-05	SYNE3	FALSE	NA	NA	No
cg00649606	1.26E-10	0.014556278	0.298051661	-0.283495382	4.75E-05	RP1-206D15.6	FALSE	NA	NA	No
cg09081596	1.54E-10	0.016122101	0.228038116	-0.211916015	5.78E-05	PEX6	FALSE	NA	NA	No
cg13254979	2.04E-10	0.012221589	0.242076284	-0.229854695	7.68E-05	LINC00327	FALSE	NA	NA	No
cg05128056	2.08E-10	0.211128371	0.865912209	-0.654783838	7.83E-05	CERS3-AS1	FALSE	NA	NA	No
cg07808087	2.69E-10	0.681183994	0.053851772	0.627332222	0.00010	CPEB2	FALSE	NA	NA	No
cg18538297	3.04E-10	0.012622119	0.248690298	-0.236068179	0.00011	ZNF837	FALSE	NA	NA	No
cg25418777	4.61E-10	0.014143346	0.227589016	-0.213445669	0.00017	DBNDD2	FALSE	NA	NA	No
cg05935660	5.51E-10	0.188820741	0.732812318	-0.543991577	0.00021	TRAF3IP2	FALSE	NA	NA	No
cg00578039	5.52E-10	0.961292337	0.702026826	0.259265511	0.00021	C6orf25	FALSE	NA	NA	No
cg04825119	9.92E-10	0.015701368	0.242750265	-0.227048896	0.00037	AQP6	FALSE	NA	NA	No
cg09360715	1.37E-09	0.021927439	0.24503716	-0.223109721	0.00052	SEPHS1	FALSE	NA	NA	No
cg18776876	1.90E-09	0.773184689	0.200609208	0.572575482	0.00072	PPP1CA	FALSE	NA	NA	No
cg09999563	1.98E-09	0.013416346	0.216287748	-0.202871403	0.00074	AASDH	FALSE	NA	NA	No
cg18691800	3.37E-09	0.837537059	0.03048881	0.80704825	0.00127	RBM24	FALSE	NA	NA	No
cg09041268	3.63E-09	0.876431606	0.570885068	0.305546537	0.00136	TSPAN9	FALSE	NA	NA	No
cg04173252	4.05E-09	0.044666166	0.324370808	-0.279704642	0.00152	CCDC115	FALSE	NA	NA	No
cg18644543	4.74E-09	0.017039981	0.233054272	-0.216014291	0.00178	HMHA1	FALSE	NA	NA	No
cg02171500	5.02E-09	0.014120504	0.263071502	-0.248950998	0.00189	CHKA	FALSE	NA	NA	No
cg04653776	6.81E-09	0.016603297	0.234863886	-0.218260589	0.00257	CHMP4B	FALSE	NA	NA	No
cg05498379	7.30E-09	0.021688935	0.24081707	-0.219128135	0.00275	WAC	FALSE	NA	NA	No
cg02763101	8.64E-09	0.544956678	0.012205806	0.532750873	0.00325	ZNF582	FALSE	TSS200	Island	Yes
cg08884539	9.26E-09	0.922584042	0.567407152	0.35517689	0.00349	TSPAN9	FALSE	NA	NA	No
cg02366961	1.29E-08	0.479420738	0.83318915	-0.353768412	0.00486	.	FALSE	NA	NA	No
cg13543375	1.44E-08	0.29598783	0.060830208	0.235157622	0.00541	WNT4	FALSE	NA	NA	No
cg03872745	1.49E-08	0.015337881	0.24271055	-0.227372669	0.00561	CORO1C	FALSE	NA	NA	No
cg15427886	1.57E-08	0.629344919	0.126961786	0.502383134	0.00592	RP11-53M11.5	FALSE	NA	Island	Yes
cg02177646	1.64E-08	0.015090671	0.270405448	-0.255314777	0.00618	ATL3	FALSE	NA	NA	No
cg12026095	1.72E-08	0.014087478	0.520568271	-0.506480793	0.00648	FTL	FALSE	NA	NA	No
cg27070372	1.87E-08	0.664891634	0.886782327	-0.221890693	0.00702	AC005262.2	FALSE	NA	NA	No
cg09568464	2.23E-08	0.507956006	0.011914506	0.4960415	0.00841	ZNF582	FALSE	TSS200	Island	Yes
cg12150817	2.25E-08	0.014033921	0.234001707	-0.219967786	0.00846	DHRS1	FALSE	NA	NA	No

A gene plot demonstrating the methylation beta methylation values of the gene *ZNF582* was produced in Figure 61.

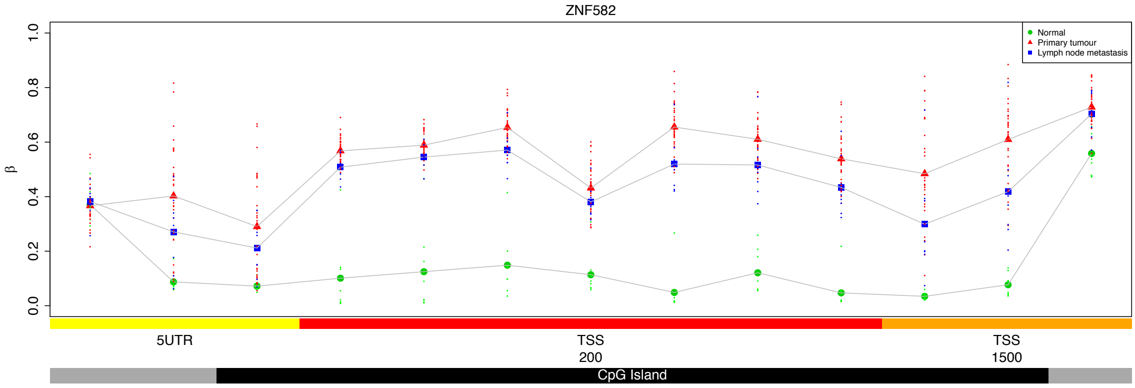


Figure 61: Gene plot of the methylation profile of the DMPs associated with ZNF582 found between lymph node metastases and lymph node normal controls. The same DMPs were found when comparing primary tumour samples and tissue adjacent controls. Genes are annotated with CpG island location (black horizontal bar at the bottom of the figure) as well as the transcription start site (TSS), 5'UTR and gene body. Each red point represents a primary tumour sample beta methylation value at an individual locus. Each green point represents a normal control sample value. Each purple point represents a lymph node metastatic sample.

4.9 Human papillomavirus (HPV)

The presence of oncogenic integrated HPV likely subverts both the host methylome and the regulation of key oncogenic genes[188]. The methylome of primary squamous cell carcinomas has previously been shown above to be highly disrupted in multiple HPV driven cancers[189,190]. The hypothesis that HPV plays a part in this disruption of the methylome in penile cancer was assessed by evaluating the number of DMPs between HPV positive samples and controls, in comparison to the number of HPV negative samples and controls.

The total number of statistically significant recurrent DMPs discovered was an order of magnitude smaller than when utilising all samples of the PenHet cohort. By subgrouping samples by HPV status only 20 tumour samples remained in each group, reducing the statistical power to detect recurrent DMPs compared with when using the complete 40 sample dataset. Due to this reduced statistical power, only a basic hypothesis producing analysis can be undertaken. These results will have to be validated by further analysis of samples typed for HPV in larger cohorts of patients.

After immune correction, 3,390 DMPs were found to be differentially methylated between HPV positive primary penile cancer samples and matched controls, compared with the 590 DMPs when assessing the HPV negative samples. These DMPs were obtained by using the same method employed for discovering the DMPs that existed between all primary squamous cell carcinomas and matched adjacent control samples. When assessing the 590 DMPs of the HPV negative samples, 283 (48%) were found in common with the HPV positive samples (Figure 62). This was further assessed by evaluating which methylation changes occur within genes previously classified as drivers by the COSMIC database. Larger numbers of DMPs found in HPV positive samples compared with HPV negative samples have previously been demonstrated in another HPV driver squamous cell carcinoma, oropharyngeal cancer[191].

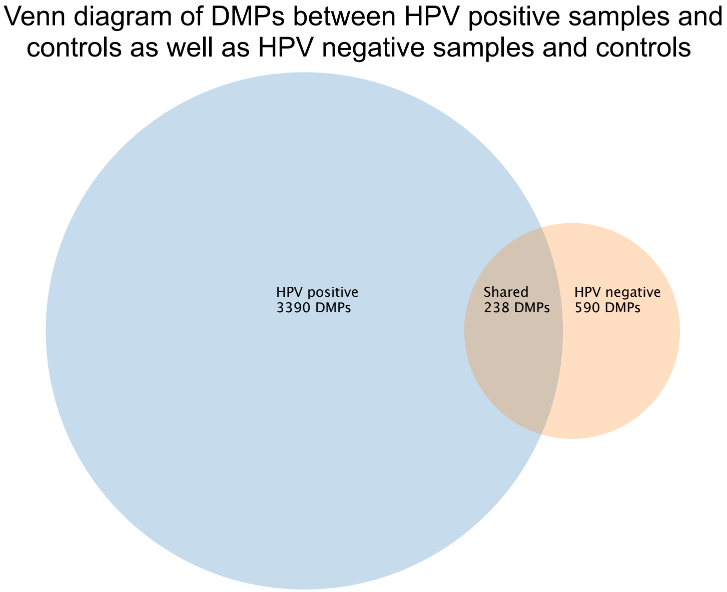
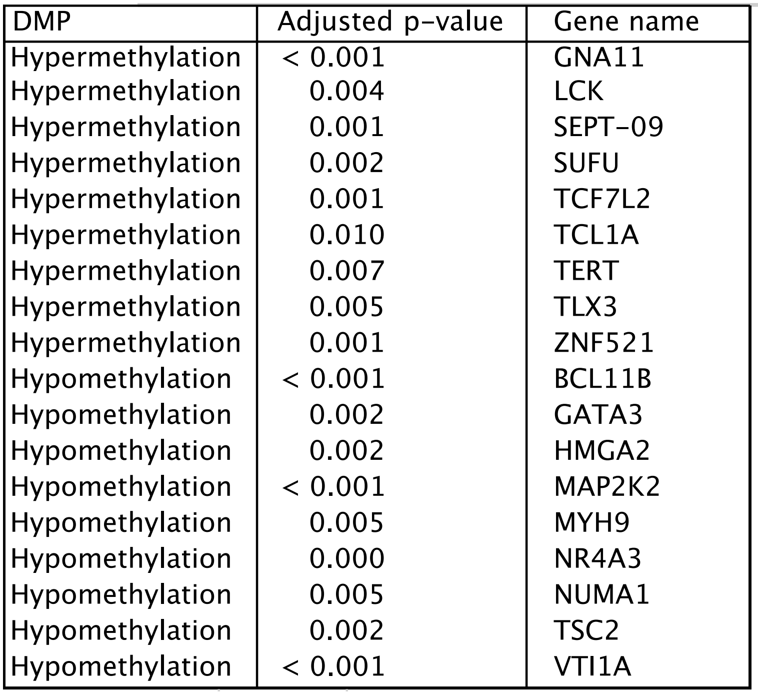


Figure 62: Venn diagram of DMPs between HPV positive samples and tissue adjacent controls as compared with HPV negative samples and controls.

Pathways overrepresented amongst differentially methylated CpG sites were also assessed by utilising the 'gometh' function in the missMethyl R package as explained above. The resulting disrupted pathways were visualised further by constructing a diagram displaying all the major pathways disrupted in cancer, using the 'pathview' package. The previous results in Chapter 3 revealed early clonal mutations in the PIK3CA/MTOR pathways solely within the HPV positive samples. Yet again, PI3K-Akt signalling pathway is significantly overrepresented in terms of the number of hypermethylated CpG sites (18 out of a total of 336 assessed, adjusted p = 1.88e-06), indicating that this pathway is deregulated by both mutation and epigenetic changes. Other pathways overrepresented in the HPV positive samples include *MAPK*, *RAP1*, *RAS* and *ERBB*. Figure 63 and Figure 64 are pictorial representations of these significant pathways disproportionately affected by in HPV positive and negative cancer samples respectively. All recurrent DMPs associated with genes previously characterised in the COSMIC database of potential drivers were assessed for both HPV positive samples (Table 16) and HPV negative samples (Table 17). However, as these tables are only based on samples from within four HPV positive and four HPV negative patients, these gene sets should be interpreted with caution. Further testing on larger groups of patients would improve the reliability of these results.

Table 16: Genes containing recurrent DMPs unique to HPV negative samples.



DMP	Adjusted p-value	Gene name
Hypermethylation	< 0.001	GNA11
Hypermethylation	0.004	LCK
Hypermethylation	0.001	SEPT-09
Hypermethylation	0.002	SUFU
Hypermethylation	0.001	TCF7L2
Hypermethylation	0.010	TCL1A
Hypermethylation	0.007	TERT
Hypermethylation	0.005	TLX3
Hypermethylation	0.001	ZNF521
Hypomethylation	< 0.001	BCL11B
Hypomethylation	0.002	GATA3
Hypomethylation	0.002	HMGA2
Hypomethylation	< 0.001	MAP2K2
Hypomethylation	0.005	MYH9
Hypomethylation	0.000	NR4A3
Hypomethylation	0.005	NUMA1
Hypomethylation	0.002	TSC2
Hypomethylation	< 0.001	VTI1A

Table 17: Recurrent DMPs unique to HPV positive samples.

DMP	Adjusted p-value	Gene name
Hypermethylation	<0.001	ACVR1
Hypermethylation	<0.001	ARID1B
Hypermethylation	<0.001	BRAF
Hypermethylation	<0.001	CAMTA1
Hypermethylation	0.004	CDKN2A
Hypermethylation	<0.001	CRTC3
Hypermethylation	0.007	DNMT3A
Hypermethylation	0.005	EBF1
Hypermethylation	<0.001	ERCC4
Hypermethylation	<0.001	FIP1L1
Hypermethylation	<0.001	GPC3
Hypermethylation	<0.001	IKZF1
Hypermethylation	0.002	JAZF1
Hypermethylation	<0.001	MECOM
Hypermethylation	0.001	MKL1
Hypermethylation	0.005	MNX1
Hypermethylation	<0.001	MSI2
Hypermethylation	0.006	MYOD1
Hypermethylation	0.009	NCOA1
Hypermethylation	0.005	NCOR2
Hypermethylation	0.005	NFIB
Hypermethylation	0.002	NKX2-1
Hypermethylation	<0.001	NRG1
Hypermethylation	0.007	PAX3
Hypermethylation	<0.001	PAX7
Hypermethylation	0.001	PHOX2B
Hypermethylation	0.001	PIK3R1
Hypermethylation	<0.001	PRDM16
Hypermethylation	<0.001	PREX2
Hypermethylation	0.008	PTPRT
Hypermethylation	<0.001	RSPO2
Hypermethylation	0.002	TP63
Hypermethylation	<0.001	ZFHX3
Hypomethylation	0.009	BCL9
Hypomethylation	<0.001	BCR
Hypomethylation	<0.001	CASP8
Hypomethylation	<0.001	CD74
Hypomethylation	<0.001	CD79A
Hypomethylation	0.005	CREB3L2
Hypomethylation	<0.001	CUX1
Hypomethylation	0.004	FHIT
Hypomethylation	<0.001	HRAS
Hypomethylation	0.003	MLLT6
Hypomethylation	0.005	NOTCH1
Hypomethylation	<0.001	TBL1XR1
Hypomethylation	<0.001	VHL

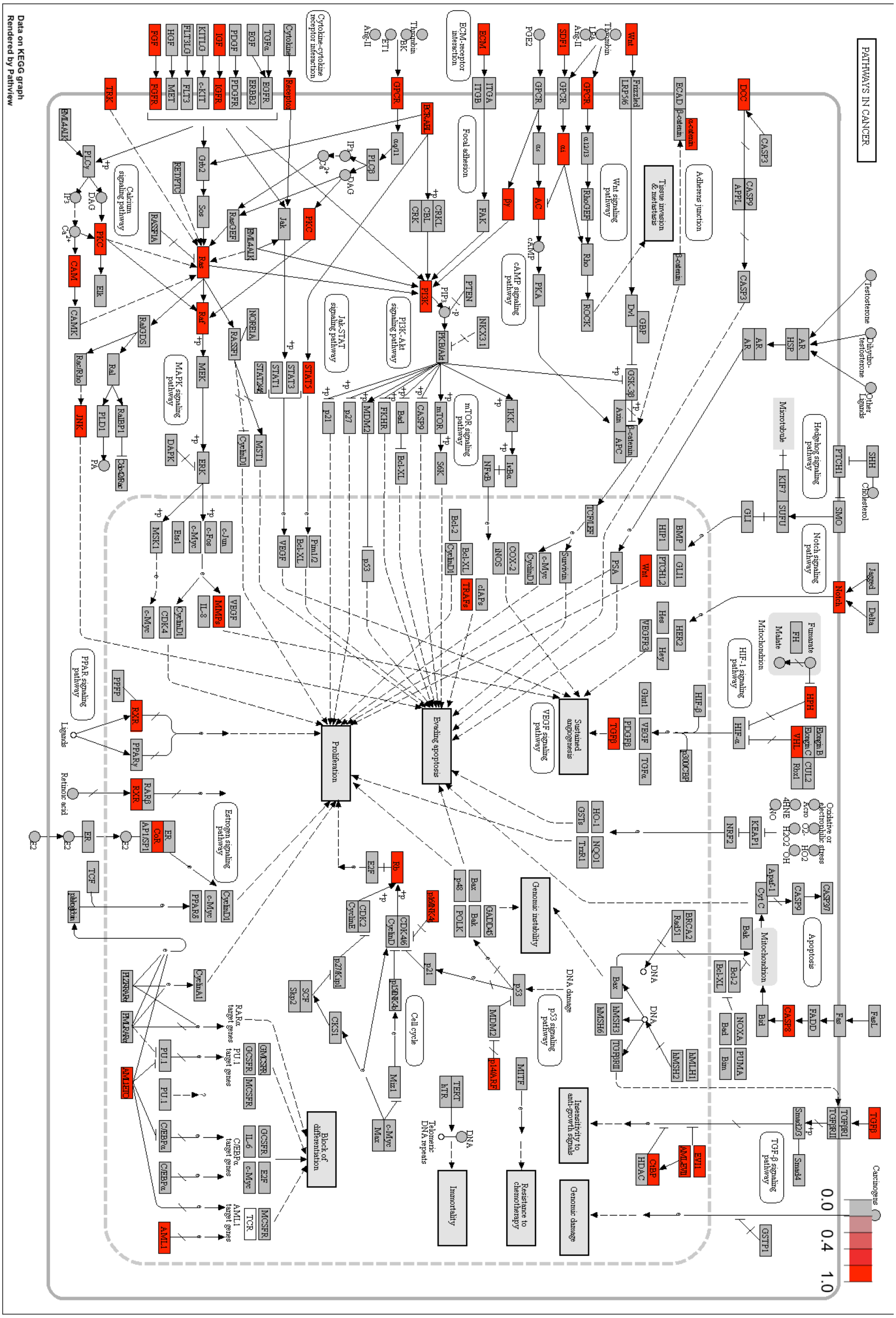


Figure 63: KEGG pathway analysis for pathways with disproportionate DMPs in HPV positive samples in the PenHet cohort.

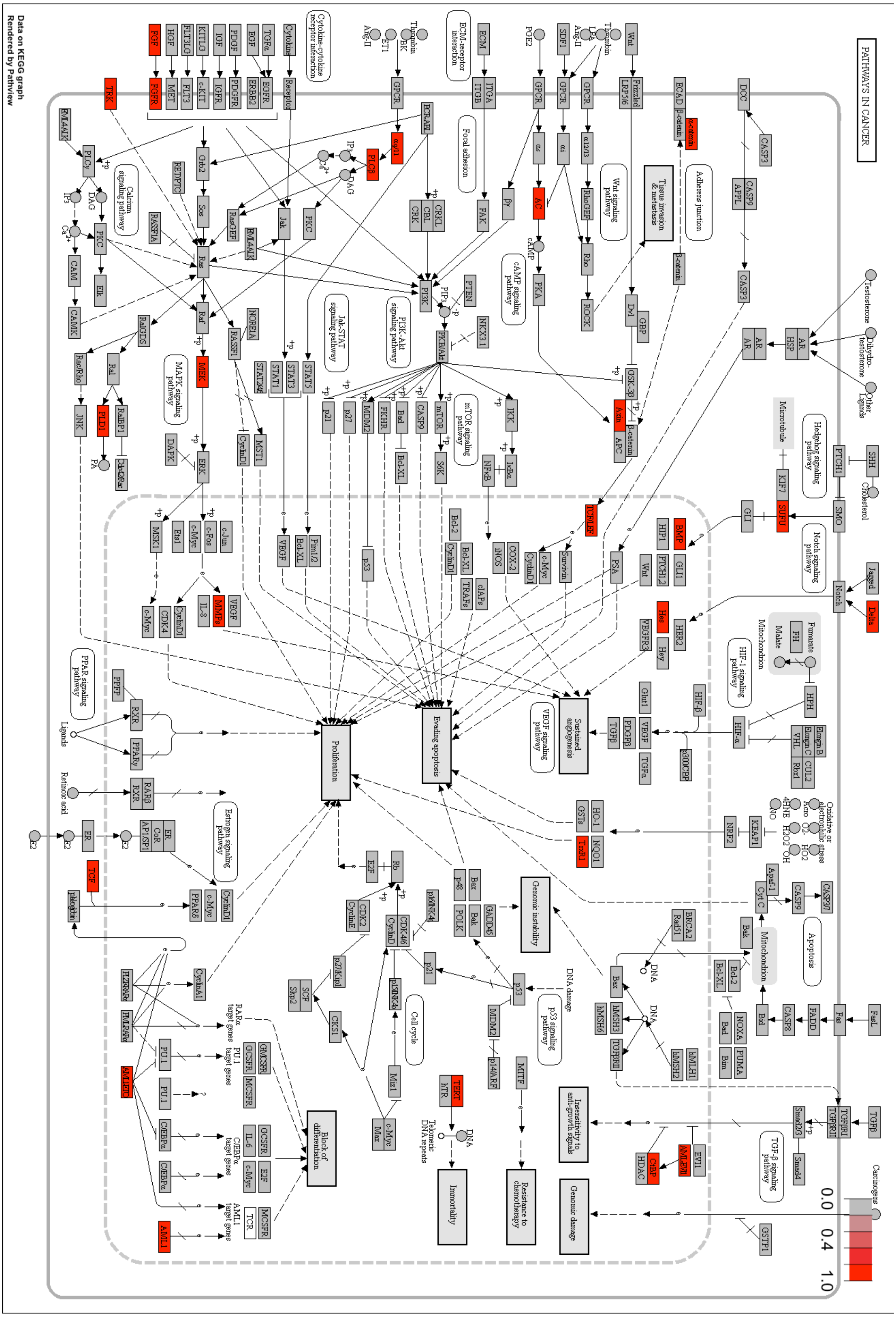


Figure 64: KEGG pathway analysis for pathways with disproportionate DMPs in HPV negative samples in the PenHet cohort.

4.9.1 Confirmatory differentially methylated regions (DMRs) in candidate methylation drivers

DMPs in close proximity to one another have a greater chance of causing a biologically meaningful impact than a single DMP[192]. In addition, DMPs in specific regions of the genome seem to have a greater propensity for regulating transcription[182]. Differentially methylated regions (DMRs) can be defined as contiguous regions that differ between phenotypes[120]. Differentially methylated regions (DMRs) were assessed using the DMRcate R package[121] (See Methods, Chapter 2).

3,347 DMRs were identified, including 1,049 hypomethylated and 2,298 hypermethylated from the 7,035 DMPs between primary cancer and normal. 1,273 DMRs were located within promoters, with these disproportionately consisting of 1,005 hypermethylated promoter DMRs and only 268 hypomethylated promoter DMRs.

The significance of the DMPs associated with driver genes, above in Figure 55, was assessed to determine whether there was just a single CpG reaching the DMP threshold within the gene, or whether a DMR was found in the potential oncogenic driver. DMRs were found in 30 out of the 50 driver genes (60%). Further details of these DMRs can be found in the Appendix.

The significance of recurrent DMPs were also assessed by determining if they could be corroborated in an additional cohort of samples, from Marchi et al[89], as discussed below. The genes associated with these corroborated CpGs are displayed in Table 23 below. From this dataset in Table 23, 22 genes were found previously described in the COSMIC database. The presence of DMRs within these potential drivers was also determined to ensure that the differential methylation at any specific gene was not solely attributable to individual CpGs, that may have little chance of being biologically significant. DMRs were found in 19 out of 22 of these genes (86%). Further details of these DMRs can be found in the Appendix. The presence of confirmatory DMRs in these genes was reassuring as it confirmed that the differential methylation, previously discovered at the individual CpG locus, could be confirmed over significant regions.

In the next chapter these methylation changes will be associated with RNA expression data. This will provide a method of narrowing down the large number of DMPs and DMRs to focus in on the methylation changes that effect gene expression changes.

4.10 Intra-tumour methylation heterogeneity

Intra-tumour methylation heterogeneity is defined by differences in the methylation found within a tumour. The assessment of methylation ITH enables the tumour to be modelled in terms of early/shared and late/unique events. Furthermore, the quantity of ITH can be compared between patients and molecular aberration types – for example, genetic and epigenetic. The methylation profiles of each sample within the primary tumour as well as the lymph node metastasis were compared to assess for intra-tumour heterogeneity. These methylation changes were then grouped into categories depending on what proportion of samples within a patient contained the specific change. Using the same methods as for the regional sample analysis of mutations in Chapter 3, each DMP was classified: as truncal if present in all cancer regions, as shared if absent from one or more regions, or as private if only present in one region.

A DMP was calculated for the primary tumour samples where there was a 20% methylation change between the primary tissue and the matched adjacent normal. A DMP was calculated for the lymph node metastasis where there was a methylation difference of 20% between the lymph node metastasis and the median methylation value of the external, histologically normal, lymph nodes. The methylation status of the lymph nodes was assessed using the same bioinformatics pipeline used for the internal PenHet cohort. The external lymph node profiles were obtained from prostate cancer patients with a normal negative lymph node dissection.

4.10.1 Immune cell contamination

The presence of tumour infiltrating immune cells can profoundly change the methylation profile of the bulk tissue extracted. Unlike when assessing the DNA mutation status of a sample, tumour infiltrating immune cells have the potential to cause apparent intra-tumour methylation heterogeneity. To reduce the chance of this happening, a DMP was only called if it was additionally differentially methylated compared to a comprehensive profile of immune cells (Methods 2.3.4.3). This was carried out in a conservative manner to ensure that the methylation ITH represented changes in cancer cells as opposed to infiltrated immune cells. The proportion of DMPs removed from each sample due to the presence of immune cell methylation signatures can be visualised in Figure 65 below. Despite large numbers of potential immune associated DMPs being removed from the analysis, this only resulted in a minimal change in the following regional phylogenetic trees constructed below (Section 4.11.2). Only in patient 39 did the removal of the immune contaminated DMPs affect the phylogenetic relationship between two regions.

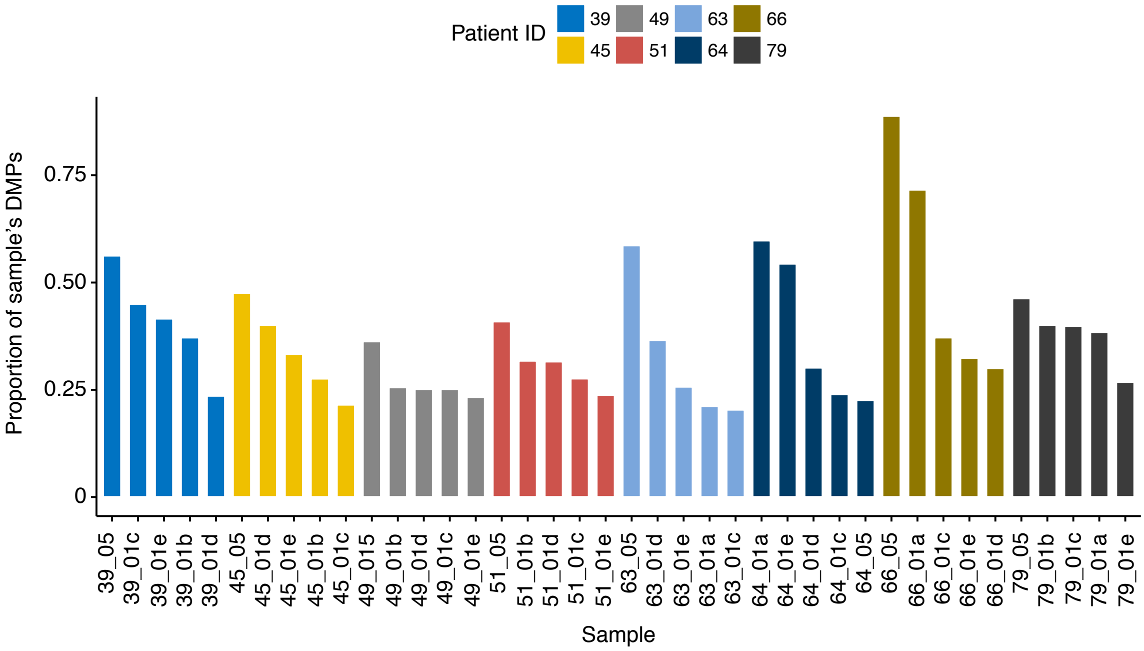


Figure 65: Bar chart depicting the proportion of a sample's DMPs which overlap with immune methylation signatures across all samples in the PenHet cohort. Samples with the suffix _05 are lymph node metastases, samples ending in a letter belong to regions of primary tumour. Samples with the prefix 39, 45, 64 and 66 are HPV negative. Samples with the prefix 49, 51, 63 and 79 are HPV positive.

The presence of immune cell contamination was compared with the calculated values of tumour cell cellularity, as assessed in the previous whole exome sequencing (Chapter 3). As demonstrated in Figure 66 there is an expected negative relationship where a large immune cell contamination is associated with a low tumour purity ($r = –0.62$, $p < 0.00002$). Clearly other factors such as the presence of stroma and other tissue types can also affect the methylation tumour cell purity. This is further assessed in the following chapter on gene expression using an alternative method of assessing sample purity (Chapter 5).

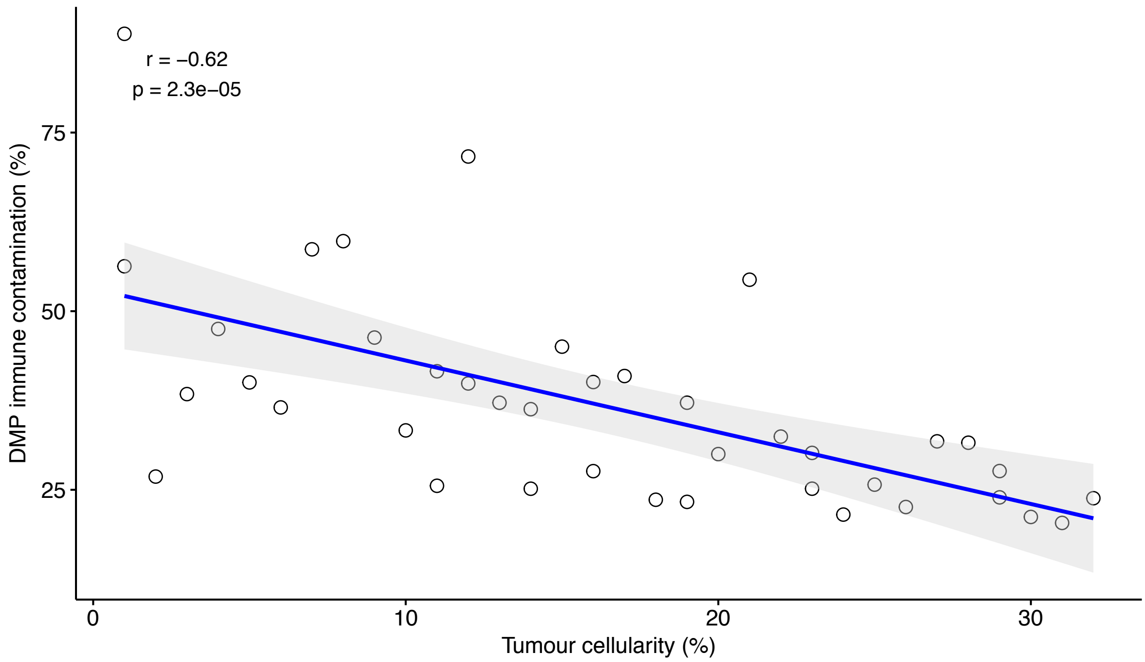


Figure 66: Scatter plot comparing the DMP immune cell contamination with the derived tumour cellularity (previously calculated in Chapter 3).

4.10.2 Regional methylation phylogenetic trees

Regional phylogenetic trees visually demonstrate the heterogeneity between regions sampled for each patient. Regional phylogenetic trees can be produced by assessing the DMPs that fall into each of three intra-tumour heterogeneity categories described above. Unlike in the previous chapter on DNA mutations (Chapter 3), the exact clonal structure of the tumour samples has not been elucidated. This is because there is not currently a standard method of calculating what constitutes a clonal methylation event when using data from methylation arrays. Unresolved challenges exist in attempting to calculate the clonal status of methylation events as discussed in the Introduction (Chapter 1).

There are several challenges when using array based DNA methylation profiles to estimate tumour heterogeneity. This includes the inability to accurately calculate the cancer cell fraction (CCF) of each methylation event. For the CCF to be calculated, the immune content influencing the bulk delta methylation level needs to be deconvoluted. In addition, the copy number at that location also needs to be taken into account. The relationship between copy number and resulting detection of methylation signals on arrays is also not yet clearly understood. Therefore, early clonal changes can only be approximated. One method used by other researchers[193-195] and utilised here is by finding DMPs that are recurrent throughout all the regions of the primary tumour. Strictly speaking these are truncal shared DMPs that are likely to be early events and possibly clonal in origin. However, as demonstrated in the previous chapter it is possible for recurrent genomic events to appear as clonal when not. Therefore, although the terms truncal/early/clonal are utilised interchangeably in many publications, caution should be exercised when interpreting these results. Fortunately, for the purposes of this analysis the exact subclonal structure of the primary cancer is not required to be calculated, as most of the insights can be gleaned from splitting the methylation events into truncal versus non-truncal.

Several methods are used throughout this thesis to produce this topological configuration and assess the molecular relatedness of each sample to another. The two most commonly used methods involve 'binarising' the data and either assessing the Euclidian distance between each sample or by using a maximum parsimony ratchet method detailed in the Methods (Section 2.3.4.9.1). Employing this method, I was able to demonstrate extensive epigenetic heterogeneity across all patients samples in the PenHet cohort, as demonstrated in Figure 54, Figure 67 and Figure 68. Scoring of ITH was undertaken in Section 4.10.3. Methylation changes tended to cluster by patient, but remarkable intra-tumour methylation heterogeneity was still observed. One might have expected the lymph node metastases to feature as the region with the greatest distance from the normal control samples. This was not found, and in seven out of eight of the patients the lymph node metastasis appeared to have formed at an earlier time point containing fewer DMPs than the primary tumour regions. This is similar to the timings of lymph node metastasis implied from the genomic analysis performed in Chapter 3. All eight patients in the PenHet cohort had advanced disease with proven lymph node metastases at the time of diagnosis. One potential biological explanation for the spatial representations of the phylogenetic trees is that, at the time that the initial lymph node metastasis formed, the methylomes of the primary and new metastasis were relatively similar. However, over time there were more constraints on methylome dysregulation at the site of the lymph node than in the primary tumour, therefore limiting its ability to acquire new epigenetic changes. Interestingly this is similar to the pattern observed with genetic rearrangements. The phylogenetic distances between all samples of one patient for one type of aberration (for example, SNV or methylation) were compared with other aberration types in regional phylogenetic trees in Chapter 5. One potential method that could be utilised to investigate this idea further would be to undertake multi-region methylome analysis sequentially throughout the early and later stages of tumour development to more accurately model the timing of these events and associate molecular alterations.

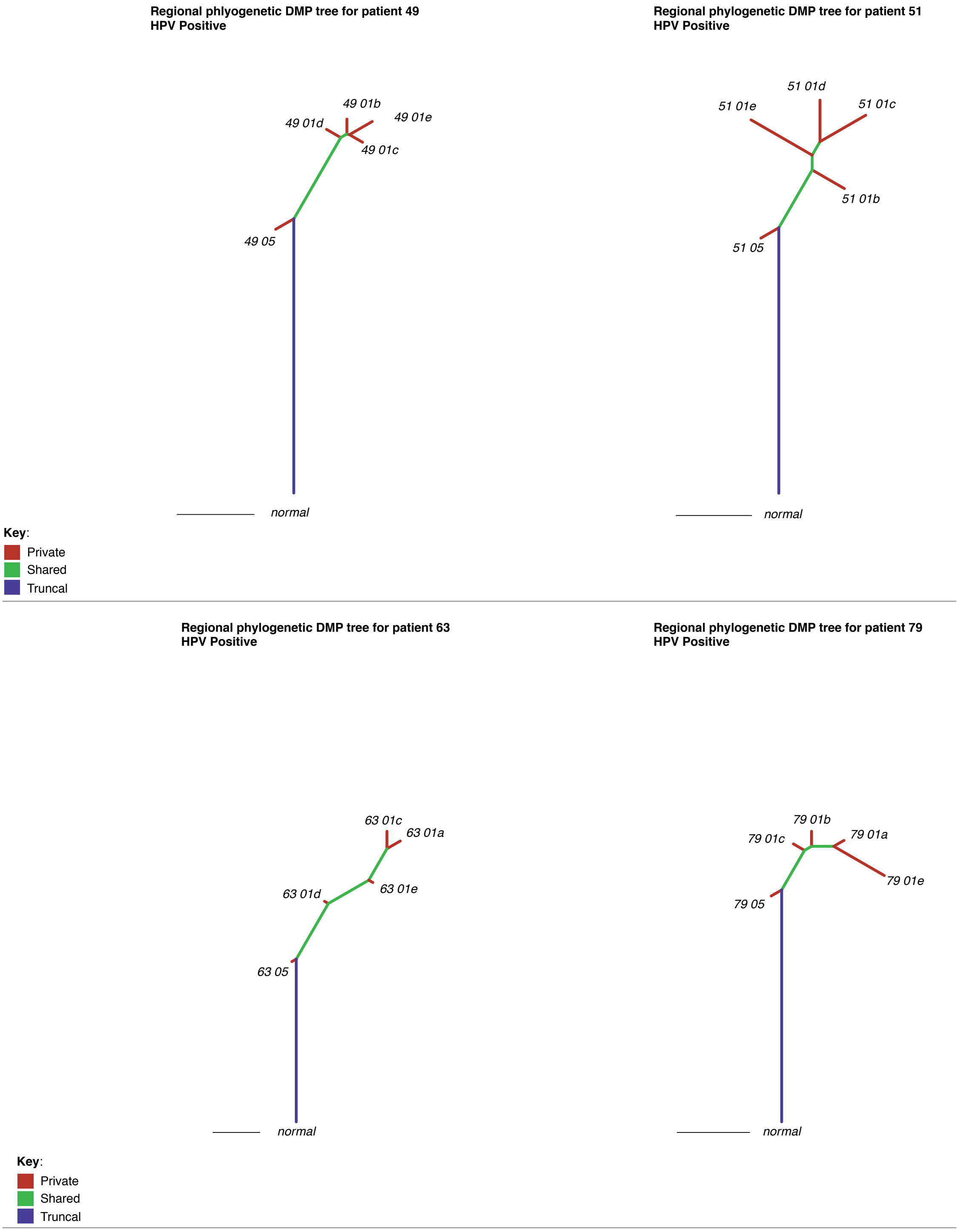


Figure 67: Regional DMP phylogenetic trees for HPV positive patients. Regions with the suffix 05 are lymph node metastases. Regions that end in a letter are primary tumour samples.

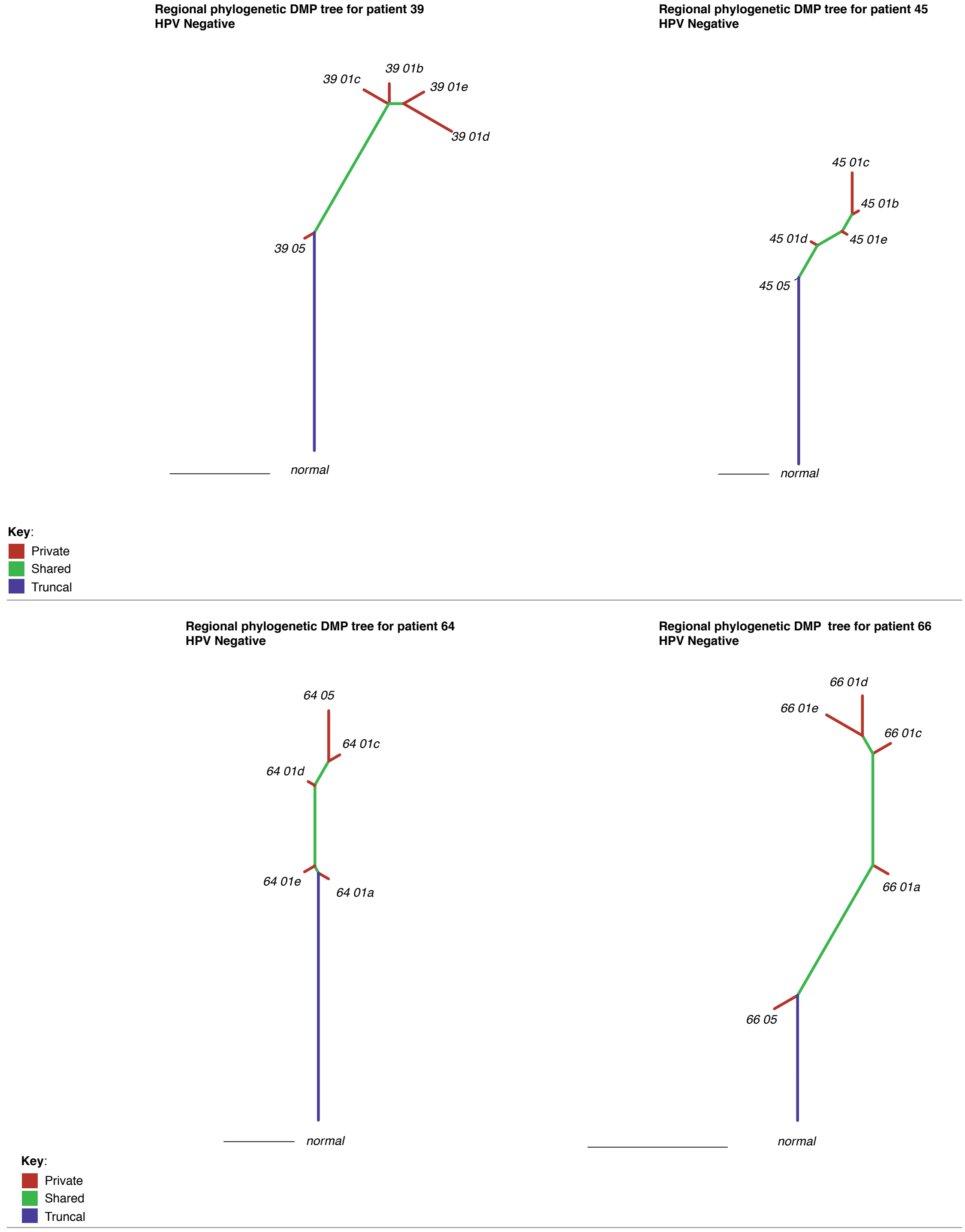


Figure 68: Regional DMP HPV negative phylogenetic trees for HPV negative patients. Regions that end in 05 are lymph node metastases. Regions that end in a letter are primary tumour samples.

4.10.3 Scoring of ITH

In an attempt to accurately define the extent of intra-tumour heterogeneity, as opposed to simply binarising alterations, I defined an ITH score for each tumour region, which was calculated as follows:

$$ITH = 1 - n$$

Where n = the proportion of methylation changes that are truncal in origin.

A truncal methylation change refers to a DMP that is shared throughout all regions of the primary tumour, as explained in Section 4.10.2.

Unlike in the assessment of DNA mutations, the methylation status of the lymph node metastasis was excluded from this calculation. This is because the environment of the tissue being examined plays a significant role in its methylation status and gene expression. For instance, clonal mutations present in the primary cancer will likely be present in lymph node metastasis, irrespective of the surrounding tissue or presence of infiltrating immune cells. Alternatively, the methylation changes, which may be clonal and present in the all the primary tissue, may be affected by the new environment of the lymph node, resulting in a change of signature. It is not currently possible to deconvolute this signal and compensate for the differences in tissue type when comparing methylation changes across tissue types. This is problematic when scoring ITH, as described above, as a methylation change may be described as heterogeneous despite being present throughout all primary cancer regions when not detected in the lymph node. Therefore, when calculating ITH scores for methylation, the lymph node status was excluded so as not to over-estimate the ITH score.

Using this scoring system, the ITH scores ranged from 44%-83% with a mean of 69%. Table 18 displays the proportion of truncal changes for each sample as a proportion of the total number of DMPs. A significantly higher proportion of truncal methylation changes were found in HPV positive samples compared with HPV negative samples (z score, $p < 0.0001$) with corresponding lower ITH demonstrated in the HPV positive samples compared with HPV negative samples (ITH scores of 64% versus 73% for HPV positive and negative samples respectively).

Table 18: Percentage of DMPs shared throughout all regions of each primary tumour, also termed truncal DMPs. A truncal methylation change was defined as a beta methylation score of > 30% conserved across all samples of a primary tumour. 'Combined' refers to the median HPV positive and negative values.

Patient ID	Age	HPV status	Number of DMPs	DMPs shared throughout all regions of primary tumour (%)
39	51	Negative	21715	30.0%
45	78	Negative	64810	35.0%
49	84	Positive	49924	56.0%
51	88	Positive	61027	28.0%
63	49	Positive	72685	34.0%
64	53	Negative	58278	21.0%
66	56	Negative	29354	17.0%
79	59	Positive	60314	30.0%
HPV Positive			60671	36.0%
HPV Negative			43816	27.0%
Combined			59296	31.0%

4.10.4 Early versus late associations

The number and directionality of truncal DMPs differ drastically from non-truncal DMPs. Table 19, Table 20, Table 21 and Table 22 demonstrate that truncal DMPs are statistically more likely to be hypermethylation events within CpG islands, with a far greater proportion of promoter sites than expected ($p < 0.0001$). Although these differences may represent true biological differences in early methylation changes when compared with later non-truncal methylation changes, these findings need to be interpreted cautiously. Truncal DMPs are defined as DMPs that are recurrent throughout the primary tumour samples of a patient. Therefore, by their very definition, truncal DMPs are more likely to be biologically significant and may therefore represent a method of filtering out the stochastic noise seen in methylation profiles. An alternative explanation exists for the differences between truncal versus non-truncal DMPs: these differences may be explained by the different levels of biologically significant methylation profiles in truncal versus branch DMPs. This may be confounded further by the selection of targets on the Illumina arrays, which were primarily designed to detect biologically significant methylation changes at CpG islands. These changes have mostly been found to be driven by hypermethylation events at promoters of driver genes.

Table 19: Table assessing the proportion of hypermethylated DMPs that are truncal versus non-truncal in origin for each patient.

			Hypermethylation		
Patient ID	Age	HPV status	Proportion of clonal DMPs	Proportion of non clonal DMPs	p value
39	51	Negative	86.0	53.7	< 0.0001
45	78	Negative	75.9	59.0	< 0.0001
49	84	Positive	86.0	53.7	< 0.0001
51	88	Positive	75.8	59.0	< 0.0001
63	49	Positive	79.6	30.6	< 0.0001
64	53	Negative	82.3	41.6	< 0.0001
66	56	Negative	76.3	45.6	< 0.0001
79	59	Positive	53.8	31.6	< 0.0001

Table 20: Table assessing the proportion of DMPs located in promoter regions that are truncal versus non-truncal in origin for each patient.

			Promoters		
Patient ID	Age	HPV status	Proportion of clonal DMPs	Proportion of non clonal DMPs	p value
39	51	Negative	24.1	20.4	< 0.0001
45	78	Negative	29.3	19.5	< 0.0001
49	84	Positive	28.1	23.0	< 0.0001
51	88	Positive	23.0	23.6	= 0.2920
63	49	Positive	28.3	18.8	< 0.0001
64	53	Negative	27.6	19.9	< 0.0001
66	56	Negative	30.0	21.8	< 0.0001
79	59	Positive	23.7	19.0	< 0.0001

Table 21: Table assessing the proportion of DMPs located in CpG islands that are truncal versus non-truncal in origin for each patient.

			CpG Islands		
Patient ID	Age	HPV status	Proportion of clonal DMPs	Proportion of non clonal DMPs	p value
39	51	Negative	61.4	30.0	< 0.0001
45	78	Negative	62.9	23.1	< 0.0001
49	84	Positive	53.6	29.6	< 0.0001
51	88	Positive	40.8	27.2	< 0.0001
63	49	Positive	54.3	18.9	< 0.0001
64	53	Negative	65.4	27.6	< 0.0001
66	56	Negative	52.7	27.9	< 0.0001
79	59	Positive	41.6	23.4	< 0.0001

Table 22: Table assessing the proportion of DMPs located in genes previously described as genetic drivers as per the COSMIC database that are truncal versus non-truncal in origin for each patient.

			Driver' DMPs		
Patient ID	Age	HPV status	Proportion of clonal DMPs	Proportion of non clonal DMPs	p value
39	51	Negative	2.6	2.5	0.78
45	78	Negative	2.9	2.8	0.81
49	84	Positive	3.0	3.2	0.31
51	88	Positive	2.8	2.9	0.61
63	49	Positive	2.3	3.2	< 0.0001
64	53	Negative	2.7	2.8	0.96
66	56	Negative	2.7	2.7	1.00
79	59	Positive	2.7	2.9	0.27

4.10.5 Recurrent DMPs across the PenHet cohort

Cancer associated DMPs can equate to DNA mutations and can be referred to as epiMutants. When assessing DNA mutations or epiMutants it is important to consider that there is a background level of genetic/epigenetic stochastic noise which occurs early in the evolution of a tumour and may be carried throughout future cell divisions, giving the impression of a significant early tumour driver. These changes are considered passengers and, although they may have an important role in providing the substrate for future genomic instability or becoming driving factors at future time points, they may not be vital in the early tumour oncogenesis. One method of finding significant necessary drivers in DNA mutation sequencing studies is to assess which epiMutations are recurrent throughout a cohort of patients. The more common a mutation the lower the chance that it is a relic of recurrent genetic/epigenetic noise.

In Chapter 3, no significant mutations previously characterised (in the COSMIC database) as drivers were found recurrently mutated throughout all primary cancer samples. However, there were important mutations – namely *PIK3CA* and *TP53* – that were recurrent in a subset of either HPV positive or HPV negative samples respectively.

In contrast to Chapter 3, this methylation study discovered a cohort of 125 epiMutants that were recurrent in all tumour samples (post immune correction). One can hypothesise that some of these epiMutants may play a role in the oncogenesis of penile cancer. The RNA expression of genes representing this cohort of significant DMPs will be assessed in Chapter 5. An additional method of determining the importance of these changes is to assess the proportion of these changes can be corroborated in a previous cohort comprising 27 independent primary penile cancer samples (PenOld).

In total 94 out of 125 (75.2%) of recurrent DMPs were corroborated in the PenOld cohort compared to 41% of all DMPs irrespective of truncal/branch status (p < 0.001). A far greater proportion of truncal DMPs 89/107 (83.2%) were validated, compared with branch DMPs 5/18 (27.8%) (Table 24). Only three genes, *RSPO2*, *CASP8* and *TERT*, containing DMPs were identified in the COSMIC database of cancer associated genes. *RSPO2* (Figure 69) has previously been identified as a tumour suppressor in gastric[196] and colorectal[197] carcinomas. Furthermore, promoter hypermethylation of *RSPO2* has previously been associated with downregulation/reduced expression of *RSPO2*[197]. The PenHet cohort provides evidence that *RSPO2* may be an important tumour suppressor gene in the development of penile cancer, as it is found early in the trunks of all samples and is also corroborated in the PenOld cohort of penile cancer samples. In Chapter 5, the expression of *RSPO2* will be assessed to determine whether there is an association in penile cancer between promoter hypermethylation and loss of expression.

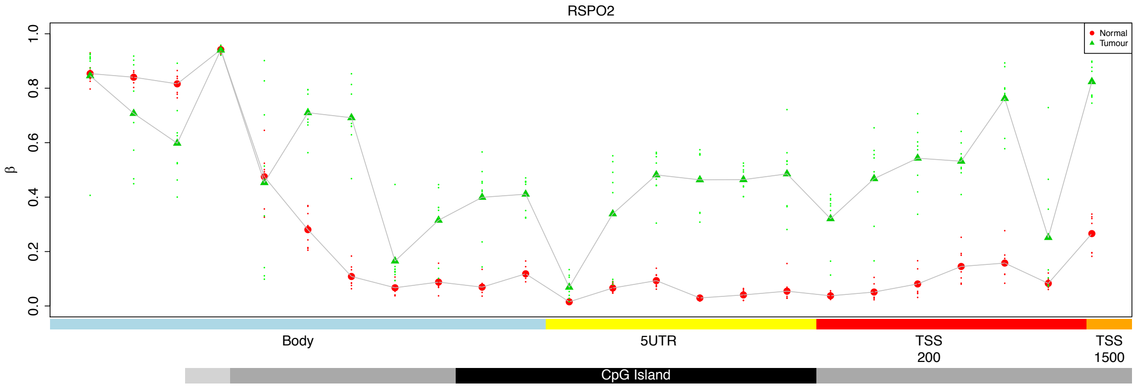


Figure 69: Canonical gene plot for RSPO2 demonstrating hypermethylation of this gene comparing tissue adjacent normal (red) and tumour samples (green).

TERT has been well characterised as a gene potentially subverted in cancer to express telomerase as a method of immortalising cells and promoting cell proliferation[198]. Differential methylation across multiple regions of *TERT* can be seen in Figure 70.

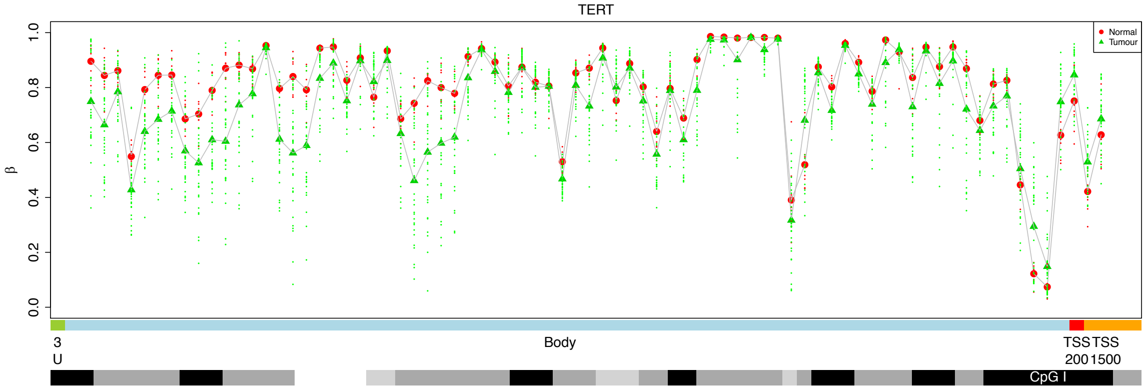


Figure 70: Methylation plot for TERT demonstrating the beta methylation value for each sample across a gene. Genes are annotated with CpG island location (black horizontal bar at the bottom of the figure) as well as transcription start site (TSS), 5'UTR and gene body. Each red point indicates a primary tumour sample beta methylation value at an individual locus. Each green point represents a normal control sample value.

In addition to looking specifically at the cancer associated genes in the COSMIC database, a literature search was performed to assess whether any of the other genes in Table 23 have been reported as demonstrating promoter hypermethylation, and whether any have led to changes in expression or are associated with tumour suppressor or oncogenic activity. The following genes were all found to have recurrent promoter hypermethylation in other cancers: the recurrent hypermethylation of the transcription factor *ZNF135* has previously been found to be one of the most frequently hypermethylated transcription factors in a pan-cancer methylome analysis[199]; *GALNTL6* has previously been found to be hypermethylated in endometrial cancer[200]; epigenetic loss of the putative tumour suppressor *FRZB* by hypermethylation has been described as associated with a poor prognosis in lung adenocarcinoma[201]; *OTX2*, a homeobox gene related to cell differentiation and expression, has been found in a pan-cancer analysis to be recurrently hypermethylated[202]; *MDGA2* is a tumour suppressor that has previously been found to be inactivated and hypermethylated in gastric cancer[203]; *EDNRB* is a candidate tumour suppressor gene and has been found to have loss of expression with promoter hypermethylation in hepatocellular[204], head and neck[205] and colorectal carcinomas[206]; hypermethylation of *SOX17* inhibits its antagonism of Wnt signalling pathway in lung[207], breast[208] and head and neck cancer[209]; and *NID* is hypermethylated in head and neck cancers with evidence that it can inhibit the EGFR/Akt and integrin/FAK/PLCγ metastasis related pathways[210].

A comparison of the proportion of DMPs that are recurrent and truncal in nature was also undertaken when taking into account the HPV status of the patient. Table 25 displays these results, where 8.7% (21,232) of DMPs in HPV positive disease were recurrent throughout the cohort, compared with 3.3% (5,680) in HPV negative disease. In this cohort it is therefore more than 2.5 times as likely that recurrent DMP will be found in HPV positive disease. It could therefore be argued that there is less inter-tumour heterogeneity in HPV positive disease in the PenHet cohort of penile cancer samples. Furthermore, the proportion of DMPs that are truncal in nature was also significantly higher in the HPV positive patients. Therefore, in HPV positive patients there is less inter- and intra-tumour heterogeneity in this cohort, despite there being more DMPs.

Table 23: Table of genes that are recurrently aberrantly methylated compared with normal samples throughout the PenHet cohort. The validation status of each DMP was determined by the presence of that DMP in the independent cohort of samples processed using the same pipeline discussed in the methods (Chapter 2).

Probe ID	Gene	In COSMIC	Gene location	CpG location	Hyper or hypomethylated	Truncal	Validated in external dataset
cg16845394	RSPO2	TRUE	TSS200	S_Shore	Hyper	Trunk	TRUE
cg26799474	CASP8	TRUE	5'UTR		Hypo	Trunk	TRUE
cg13823136	ST6GALNAC5	FALSE	TSS200	Island	Hyper	Trunk	TRUE
cg14794428	ASCL1	FALSE	TSS200	N_Shore	Hyper	Trunk	TRUE
cg07601320	RP11-96H19.1	FALSE	TSS200		Hyper	Trunk	TRUE
cg26394244	NID2	FALSE	TSS200	Island	Hyper	Trunk	TRUE
cg03278146	C18orf42	FALSE	TSS200	Island	Hyper	Trunk	TRUE
cg06433694	CTC-512J12.4	FALSE	TSS200		Hyper	Trunk	TRUE
cg15241920	TTYH1	FALSE	TSS200	N_Shore	Hyper	Trunk	TRUE
cg27477373	AC006116.21	FALSE	TSS200	Island	Hyper	Trunk	TRUE
cg08701621	ZNF135	FALSE	TSS200	Island	Hyper	Trunk	TRUE
cg12919006	AC079154.1	FALSE	TSS200	Island	Hyper	Trunk	TRUE
cg13356896	BOLL	FALSE	TSS200	Island	Hyper	Trunk	TRUE
cg26492446	BHLHE23	FALSE	TSS200	Island	Hyper	Trunk	TRUE
cg14859460	GRM6	FALSE	TSS200	Island	Hyper	Trunk	TRUE
cg02467990	VWC2	FALSE	TSS200	Island	Hyper	Trunk	TRUE
cg24928391	SOX17	FALSE	TSS200	Island	Hyper	Trunk	TRUE
cg14653281	GRIN3A	FALSE	TSS200	Island	Hyper	Trunk	TRUE
cg10636246	AIM2	FALSE	TSS1500		Hypo	Trunk	TRUE
cg21675115	EDNRB	FALSE	TSS1500	S_Shore	Hyper	Trunk	TRUE
cg08217024	MDGA2	FALSE	TSS1500	Island	Hyper	Trunk	TRUE
cg09624466	OTX2-AS1	FALSE	TSS1500	Island	Hyper	Trunk	TRUE
cg04037038	FRZB	FALSE	TSS1500	Island	Hyper	Trunk	TRUE
cg00970325	PAQR9	FALSE	TSS1500	Island	Hyper	Trunk	TRUE
cg03304610	GALNTL6	FALSE	TSS1500	Island	Hyper	Trunk	TRUE
cg18325622	MARCH11	FALSE	TSS1500	Island	Hyper	Trunk	TRUE
cg09591286	ZNF804B	FALSE	TSS1500	N_Shore	Hyper	Trunk	TRUE
cg07792478	MIR124-2	FALSE	TSS1500	Island	Hyper	Trunk	TRUE
cg21578219	IGSF21	FALSE	5'UTR	Island	Hyper	Trunk	TRUE
cg10224098	RNF220	FALSE	5'UTR	Island	Hyper	Trunk	TRUE
cg14780632	GAL3ST3	FALSE	5'UTR	Island	Hyper	Trunk	TRUE
cg14699728	NPAS4	FALSE	5'UTR	Island	Hyper	Trunk	TRUE
cg23989821	C14orf39	FALSE	5'UTR	Island	Hyper	Trunk	TRUE
cg02401399	AC002116.7	FALSE	5'UTR	Island	Hyper	Trunk	TRUE
cg26246807	ZIK1	FALSE	5'UTR	Island	Hyper	Trunk	TRUE
cg06714480	CERKL	FALSE	5'UTR	N_Shelf	Hyper	Trunk	TRUE
cg06818532	BBX	FALSE	5'UTR		Hyper	Trunk	TRUE
cg05663573	CLDN11	FALSE	5'UTR	N_Shore	Hyper	Trunk	TRUE
cg15031661	FMN2	FALSE	1stExon	Island	Hyper	Trunk	TRUE
cg20168230	GRIK3	FALSE	1stExon	Island	Hyper	Trunk	TRUE
cg19839798	FAM155A	FALSE	1stExon	Island	Hyper	Trunk	TRUE
cg04945331	SOX14	FALSE	1stExon	Island	Hyper	Trunk	TRUE
cg09221867	PCDH10	FALSE	1stExon	Island	Hyper	Trunk	TRUE
cg05716166	RALYL	FALSE	1stExon	N_Shore	Hyper	Trunk	TRUE
cg08738570	TMEM240	FALSE	Body	N_Shore	Hyper	Trunk	TRUE
cg09671258	LHX4	FALSE	Body	Island	Hyper	Trunk	TRUE
cg07046369	PAX2	FALSE	Body	Island	Hyper	Trunk	TRUE
cg18419977	SLC22A18	FALSE	Body	N_Shelf	Hypo	Trunk	TRUE
cg17203063	KCNC2	FALSE	Body	Island	Hyper	Trunk	TRUE
cg25317585	FGF14	FALSE	Body	Island	Hyper	Trunk	TRUE
cg13012916	RP11-896J10.3	FALSE	Body	Island	Hyper	Trunk	TRUE
cg18716164	VSTM2B	FALSE	Body	Island	Hyper	Trunk	TRUE
cg06428620	PCDHGA1	FALSE	Body	N_Shore	Hyper	Trunk	TRUE
cg18789958	HCN1	FALSE	Body	N_Shore	Hyper	Trunk	TRUE
cg20348196	HLA-DRA	FALSE	Body		Hypo	Trunk	TRUE
cg21179088	VSTM2A	FALSE	Body	Island	Hyper	Trunk	TRUE
cg11395386	MECP2	FALSE	Body	N_Shelf	Hyper	Trunk	TRUE
cg09755589	RP11-276E17.2	FALSE	Intergenic		Hypo	Trunk	TRUE
cg26986871	KLRC4-KLRK1	FALSE	Intergenic		Hypo	Trunk	TRUE
cg26132774	RNF219-AS1	FALSE	Intergenic	Island	Hyper	Trunk	TRUE
cg01630690	NKX2-1-AS1	FALSE	Intergenic	Island	Hyper	Trunk	TRUE
cg00001747	LINC01158	FALSE	Intergenic	Island	Hyper	Trunk	TRUE
cg25946059	LINC01237	FALSE	Intergenic		Hypo	Trunk	TRUE
cg02421985	FOXO1B	FALSE	Intergenic	Island	Hyper	Trunk	TRUE
cg04349084	RP11-175E9.1	FALSE	Intergenic		Hypo	Trunk	TRUE
cg18249580	RP11-32K4.1	FALSE	Intergenic	N_Shore	Hyper	Trunk	TRUE
cg22001496	C8orf34	FALSE	Intergenic	Island	Hyper	Trunk	TRUE
cg22770135	TRPC7	FALSE	TSS200		Hyper	Trunk	FALSE
cg27136241	RP5-1186N24.3	FALSE	TSS200	S_Shore	Hyper	Trunk	FALSE
cg08063125	ZNF667	FALSE	TSS1500	Island	Hyper	Trunk	FALSE
cg26597242	GABRA1	FALSE	TSS1500		Hyper	Trunk	FALSE
cg15423872	FAM110B	FALSE	TSS1500	N_Shore	Hyper	Trunk	FALSE
cg10109500	GHSR	FALSE	1stExon	Island	Hyper	Trunk	FALSE
cg12880658	CDO1	FALSE	1stExon	Island	Hyper	Trunk	FALSE
cg17627654	SHANK2	FALSE	Body	Island	Hyper	Trunk	FALSE
cg08567279	LECT1	FALSE	Body	Island	Hyper	Trunk	FALSE
cg27555582	ABCC9	FALSE	Intergenic	Island	Hyper	Trunk	FALSE
cg11014373	RP11-13J10.1	FALSE	Intergenic	Island	Hyper	Trunk	FALSE
cg19971388	GATA4	FALSE	Intergenic	Island	Hyper	Trunk	FALSE
cg24931138	TERT	TRUE	Body	S_Shore	Hypo	Branch	FALSE
cg09842118	RNASE3	FALSE	5'UTR		Hypo	Branch	TRUE
cg22459052	SLC6A7	FALSE	Body		Hypo	Branch	TRUE
cg12930338	MS4A6E	FALSE	TSS200		Hypo	Branch	FALSE
cg19828416	OR4D1	FALSE	TSS1500	N_Shelf	Hypo	Branch	FALSE
cg27109600	SATB2	FALSE	TSS1500	N_Shore	Hyper	Branch	FALSE
cg06580419	AC107218.3	FALSE	TSS1500		Hypo	Branch	FALSE
cg14627175	DACT2	FALSE	TSS1500	S_Shore	Hypo	Branch	FALSE
cg19825483	RYR2	FALSE	Body		Hypo	Branch	FALSE
cg06375949	MSX1	FALSE	Body	N_Shore	Hyper	Branch	FALSE
cg11234281	ZNF32-AS3	FALSE	Intergenic		Hypo	Branch	FALSE
cg23253961	CTD-2277K2.1	FALSE	Intergenic		Hypo	Branch	FALSE
cg03116035	RP11-205M3.3	FALSE	Intergenic		Hypo	Branch	FALSE

Table 24: Table displaying the proportion of DMPs validated in an external cohort of penile cancer samples. This is in comparison to the figure of 41% validated across all DMPs discovered irrespective of truncal status.

	Recurrent DMPs present in all PenHet cohort		Recurrent DMPs validated in external dataset	
	Number of recurrent DMPs	**Recurrent DMPs as a percentage of trunk / non-trunk (%)**	**Number of recurrent DMPs validated**	**Percentage of recurrent DMPs validated (%)**
Truncal	107	0.17	89	83.2
Non truncal	18	0.01	5	27.8
Total	125		94	

Table 25: Table comparing the proportion of DMPs that are recurrent as well as truncal in origin for HPV positive and HPV negative tumour samples.

	Number of DMPs	Number of recurrent DMPs	Recurrent DMPs as a percentage of all DMPs	Number of truncal DMPs	Percentage of DMPs that are truncal
HPV Positive	243950	21232	8.7	87822	36.0
HPV Negative	174157	5680	3.3	47022	27.0
p value			< 0.0001		< 0.0001

4.11 Discussion

Changes to the epigenome represent some of the earliest alterations in the tumorigenic process. To better understand the role of aberrant DNA methylation in the development of penile cancer, I performed an epigenome-wide methylation interrogation. There are a large number of recurrent differentially methylated positions in the PenHet cohort of advanced penile cancer patients. These positions have been clustered into differentially methylated regions as a way to narrow down the thousands of DMPs into regions, which are more likely to produce a biological effect. The genes associated with these DMPs and DMRs include many genes reported to be responsible for cell cycle control, cell death, proliferation and differentiation. These include genes such as *RSPO2*, *CASP8* and *TERT*. There is an overlap between the mutated genes and pathways found in the DNA sequencing study and this methylome analysis. This is particularly the case in HPV positive samples, which seem to all have mutations and differential methylation in the PI3K/MTOR pathways.

As in Chapter 3 on whole exome DNA sequencing, inter- and intra-tumour epigenetic heterogeneity was observed throughout, and between all patients in the PenHet cohort. Early methylation changes can be sought by assessing which changes are shared throughout all regions of each primary penile cancer sample. These likely represent an early clone, which may form part of the last common ancestor. These early methylation changes were found to be far more likely to be validated in PenOld – a previous cohort of penile cancer samples. One of the great challenges in analysing methylation signals is determining the biologically meaningful changes from stochastic 'noise'. Truncal DNPs are more likely to be significant than branch DNPs, and could in fact be a driving force in the oncogenesis of metastatic penile cancer. Selecting truncal DNPs, for example, could therefore be a way to overcome this noise. Chapter 5 will determine whether truncal methylation changes are more likely to cause aberrant gene expression than branch methylation changes.

The relative order of early epi-methylation versus genetic mutation events is currently unknown. Answering this question is beyond the scope of this thesis. However, work has already begun in answering this question. In addition to the completed methylome analysis, using methylation microarrays discussed in this chapter, the same have also been subjected to RRBS (Reduced Representation Bisulphite Sequencing). RRBS is a method of undertaking genome-wide methylation profiling at a reduced cost to whole genome-wide bisulfite sequencing. It is undertaken by digesting DNA with restriction enzymes and then bisulfite converting, amplifying and sequencing the DNA. This method enables a reduced sample of the genome to be sequenced that is enriched for CpGs containing the majority of promoters[211]. The analysis of the sequencing data can determine the overlap between early methylation clonal events and mutation clonal events. The results of this analysis are eagerly awaited.

Distinct methylation profiles were observed for both HPV positive and negative samples. These distinct profiles give credence to the hypothesis that the underlying biology and characteristics of penile cancer are distinct in the subsets of HPV positive and HPV negative disease. Further work should be undertaken to evaluate whether oncological outcome differences exist between these two groups of patients and determine whether the next generation of therapeutics for these patients needs to take into account the HPV status. When clustering all primary penile cancer samples in an unsupervised manor, all samples cluster into two distinct groups representing the binary HPV status, despite significant differences in the apparent viral load. As demonstrated in the previous chapter, infection with HPV is directly associated (in these exact samples) with APOBEC enzyme activity with clear APOBEC mutational signatures. It is therefore unclear whether HPV provides a direct methylation oncogenic effect or whether the effect is indirectly mediated by the induced APOBEC mutation pattern. One way this could be investigated further would be to assess the methylation profile and APOBEC characteristic mutations of early or pre-malignant disease. 48% of DMPs in HPV negative patients were shared between both HPV positive and negative cohorts. There were a significantly greater number of recurrent DMPs in the HPV positive samples compared with the HPV negative ones ($p < 0.0001$). This is due to the greater inter-tumour heterogeneity between patients with no HPV. One could hypothesise that HPV positive samples have shared distinct pathways of deregulated methylation contributing to their oncogenesis, while in HPV negative disease there are many potential pathways to genetic and epigenetic instability resulting in their tumour oncogenesis.

Squamous cell penile cancer appears to be a highly immunogenic malignancy. In both HPV positive and negative samples, methylation signatures of infiltrating immune cells was observed. In Chapter 5 the individual expression profiles of immune cells are assessed to give a clearer picture of the cancer immune microenvironment. The excitement and promise of the relatively new field of immunotherapy may provide an additional treatment modality for these patients who currently have poor and ineffective treatment options. Previously published work by Udager et al[212] demonstrated high levels of PDL-1 expression in up to 50% of patients. Those findings coupled with the findings of this chapter suggest that immunotherapy may prove to be effective in these patients, and further clinical work should be undertaken.

The methylation profiles of lymph node metastases are an amalgamation of the methylation profiles of cancer cells, lymphoid cells and infiltrating immune cells, all influenced by the surrounding microenvironment. As long as the tumour content of the lymph node metastatic tissue was sufficiently high (greater than 15%), then the lymph nodes tended to cluster with the matched primary tumour samples for that patient and not with the ‘normal’ lymph nodes or other lymph node metastases from other patients. The relatively lower tumour cell content of lymph node metastases was not unexpected and has been shown to be especially contaminated with infiltrating immune cells. This suggests that the main driver for some lymph node metastases clustering towards their matched normal sample was the dilution of their inherent tumour derived signature with ‘normal’ tissue and immune cells methylation signatures.

The majority of significant recurrent methylation changes were hyper methylation events of promoters in CpG islands. Many of these associated genes have been classified as definitive or candidate tumour suppressor genes. These include genes such as *RSPO2*, *CASP8* and *TERT.* The expression status of these genes will be assessed in Chapter 5. Pathway analysis using KEGG revealed clear large overrepresentation of methylation changes in previously recognised cancer pathways. The most significant of these pathways included RAP1, MAPK, RAS and PI3K-AKT – the latter particularly in HPV positive samples. A clear hypothesis is that there may be loss of expression of these tumour suppressor genes, potentially caused by hypermethylation of their gene promoters. If this proves to be the case, then demethylating agents such as azacitidine and decitabine could also be considered as a further treatment modality in penile cancer.

In conclusion, in the PenHet cohort of advanced penile cancer patients, there are extensive methylation changes, with both inter- and intra-tumour heterogeneity exhibited. Furthermore, there is evidence for significant involvement of infiltrating immune cells as well as distinct methylation profiles for HPV positive and negative samples. In addition, there appears to be less inter- and intra-tumour heterogeneity in HPV positive samples. This raises the possibility that penile cancer should be subclassified based on HPV status, with differing treatment modalities. In addition, immunotherapy may prove to be efficacious in these patients and should be further considered. Despite extensive noise at the CpG locus and the relatively small number of patients assessed in this PenHet cohort, there are a large number of recurrent methylation events (disproportionally hypermethylation events at promoter regions), which are validated in the PenOld cohort. This method of finding shared intra-tumour methylation events is therefore a way to refine the large numbers of methylation changes in order to find a more significant source of biomarkers and methylation oncogenic drivers. The significance of these findings will be discussed further in Chapter 5, where I will integrate the methylation changes with changes in expression.

5 Use of RNA sequencing to predict the key drivers in penile cancer by evaluating mRNA expression in conjunction with matched DNA mutations, copy number and methylation aberrations

5.1 Introduction

Changes in gene expression can be caused by genetic, epigenetic and micro-environmental changes. The previous two chapters assessed the genetic changes by means of whole exome sequencing (Chapter 3), and one type of epigenetic change by means of whole methylome analysis (Chapter 4). This chapter will assess the spectrum of gene expression changes by quantifying the changes in expression across the PenHet cohort. In addition, this chapter will integrate the gene expression results with those of the two previous chapters – creating associations between DNA mutations, copy number aberrations, methylation changes and gene expression. Furthermore, the inter- and intra-tumour heterogeneity will be modelled in the context of gene expression and will be compared with the models produced using the previous whole exome and methylome modalities. The main aim of this work is to discover the key early and late drivers of penile cancer to improve our understanding of the oncogenesis of the disease, and to hypothesise the genes that should most likely be targeted therapeutically.

The previous two results chapters demonstrated that significant intra-tumour heterogeneity exists within primary penile cancer samples. However, the effect of these mutation and methylation changes has yet to be determined.

This chapter will explore the changes in mRNA expression levels across the entire PenHet cohort. This will enable expression changes to be quantified between tumour and normal, along with modelling intra-tumour heterogeneity and the level of gene expression. As discussed in previous chapters, it is computationally challenging to distinguish the cancer driven methylation changes from random stochastic effects. One way to determine which changes at the level of DNA methylation are biologically significant is by determining which are associated with changes in the expression of the same gene. It is well recognised[17,213] that hypermethylation of promoter regions can suppress expression of that gene, whilst hypermethylation of gene bodies can enhance/activate gene expression. The converse has also been postulated.

5.2 Immune cell infiltration/contamination

The presence of infiltrating immune cells can influence the expression profiles derived from the bulk cancer tissues sampled by RNA sequencing. To determine the possible extent of immune cell contamination, xCell was used to estimate the relative proportions of immune and stroma content. The bar chart in Figure 71 displays the immune cell variability across all samples in the PenHet cohort. Unsurprisingly, as with the methylation profiles, the greatest immune contamination was seen in the lymph node metastases. As demonstrated there is also significant immune cell infiltration in the primary penile cancer samples (Figure 1). The top five ranked mean immune cell type scores were Th2 T cells (4.62), dendritic cells (3.45), Th1 T cells (3.16), B-cells (2.37) and basophils (2.02) (Figure 72). There is overrepresentation of Th1 and Th2 differentiated T cells as well as basophils cells when compared with the matched tissue adjacent normal samples. These scores were compared between primary, lymph node and HPV positive subsets. Th1 xCell score was significantly higher in the HPV positive samples (T-test, p = 0.0002) (Table 26). Th1 cytokine patterns in T helper cells have previously been described as the expected immune response to HPV infection in both cervical[214] and pharyngeal carcinomas[215]. Despite different modalities used (genomic, epigenomic or transcriptomic), a weak inverse relationship was found (R = –0.41, p = 0.0087) between xCell immune contamination scores and tumour cellularity (as calculated from whole exome sequencing data).

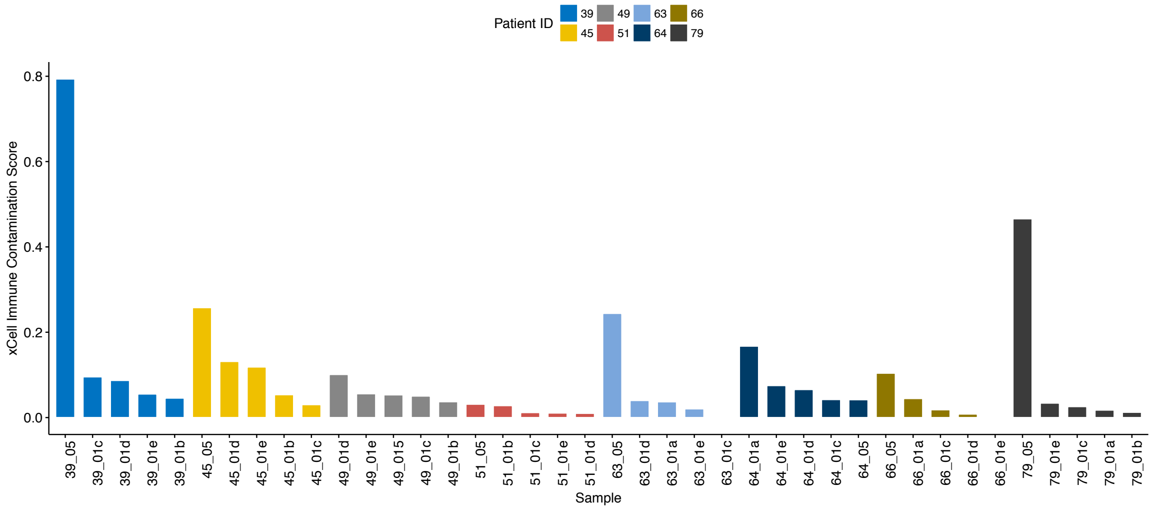


Figure 71: Bar chart depicting the relative immune cell contamination of all tumour samples. Immune cell contamination scores based on xCell (see Methods, Chapter 2). Tumour samples with the sample suffix _05 depict lymph node metastases. Samples with suffix _01a, _01b, _01c, _01d, _01e depict primary tumour samples. The two-digit sample prefix depicts the unique patient code from which the samples were taken. Further information on the samples can be found in Methods, Chapter 2.

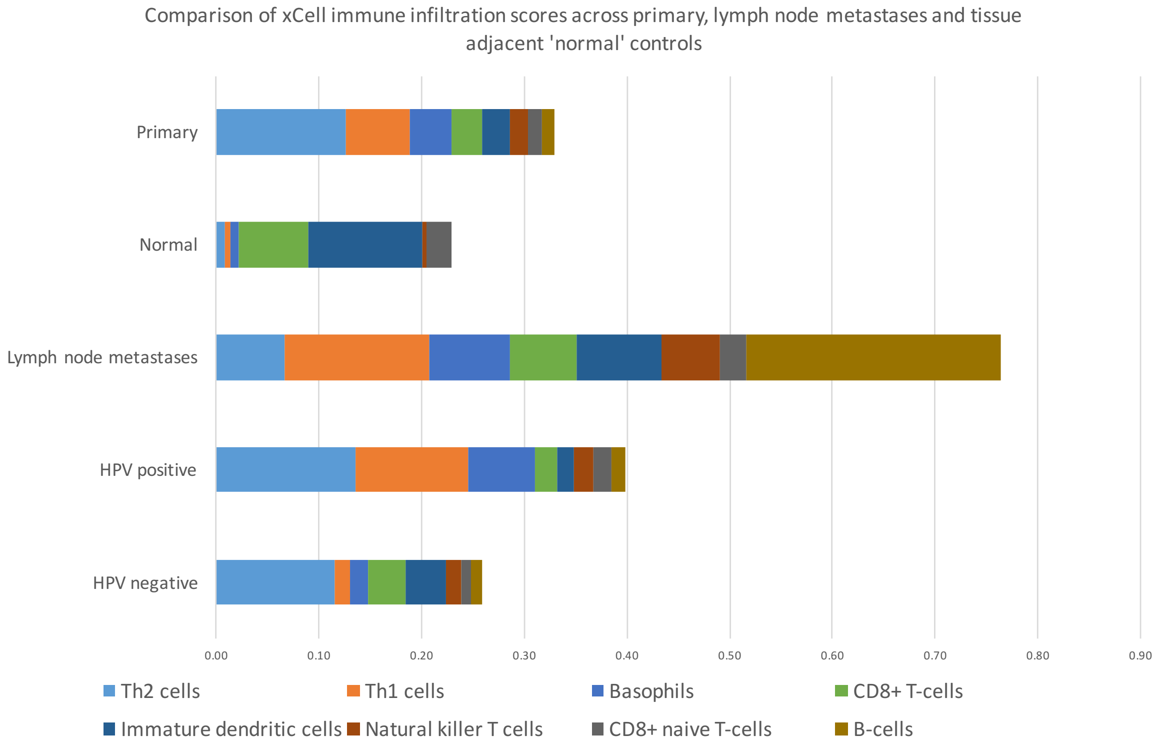


Figure 72: Comparison of relative xCell immune infiltration scores across groups of samples.

Table 26: Comparison of xCell immune infiltration scores comparing primary versus tissue adjacent as well as by HPV status. Cell types key: Th1 = Th1 cytokine response T helper cell, Th2 = Th2 cytokine response T helper cell, cDC = conventional dendritic cell / myeloid dendritic cell, aDC = activated dendritic cell, iDC = immature dendritic cell, NKT = natural killer cell.

Cell type	Primary	Normal	T-test (p-value)		HPV Positive	HPV Negative	T-test (p-value)
Th1 cells	0.06	0.01	0.0504		0.11	0.02	**0.0002**
Th2 cells	0.13	0.01	**0.0000**		0.14	0.01	0.3715
cDC	0.04	0.13	**0.0018**		0.01	0.07	**0.0021**
aDC	0.05	0.02	0.1507		0.05	0.06	0.7555
iDC	0.03	0.11	**0.0000**		0.02	0.04	0.0217
B-cells	0.01	0.00	0.2112		0.01	0.01	0.7812
Basophils	0.04	0.01	0.0278		0.07	0.02	**0.0001**
CD8+ T-cells	0.03	0.07	**0.0004**		0.02	0.04	0.0844
NKT	0.02	0.00	0.2369		0.02	0.01	0.5858
CD8+ naive T-cells	0.01	0.02	0.0133		0.02	0.01	**0.0099**

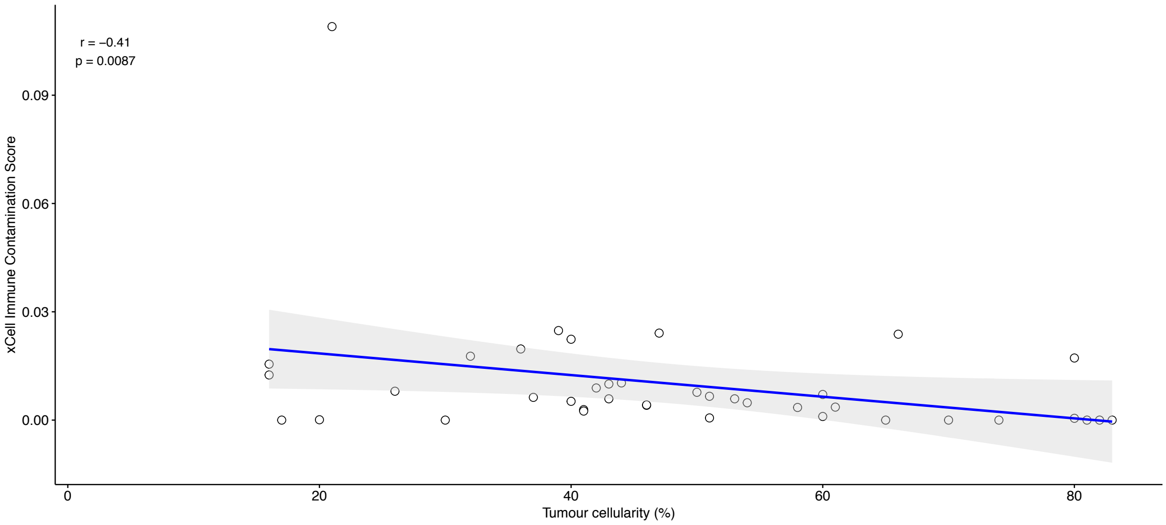


Figure 73: Comparison of xCell immune contamination scores with matched predicted tumour cellularity based on sequenza whole exome sequencing.

Although the presence of immune cells can be considered a contamination – reducing our ability to identify the cancer induced expression changes – their presence may also be an important indicator into the tumour microenvironment and potential susceptibility to immunotherapeutic agents. The differential expression of specific immune checkpoints is assessed below in Section 5.4.1.1.

5.3 Baseline expression characteristics of penile cancer

The relationship between samples within the PenHet cohort was assessed by plotting the 1,000 most variably expressed genes between primary and normal tumour samples (Figure 74). This plot clearly demonstrates that the normal samples cluster separately to tumour samples. Furthermore, the tumour samples appear to also cluster into two groups, reflecting the HPV status of each group (Figure 74). This is further demonstrated in the heatmap and clustering of all samples in Figure 75. Tumour samples, in general, clustered with the other tumour samples belonging to the same patient. The exception to this is in the case of the lymph node metastases, which appear to sometimes cluster closer to other lymph node metastases (Figure 75). One explanation for this is that the expression profile of lymph node metastases is influenced heavily by the presence of both immune cells – which are far more prominent in the lymph node metastases than primary tumour tissue – and the mixture of surrounding tissue cells in the bulk sequenced tissue sample.

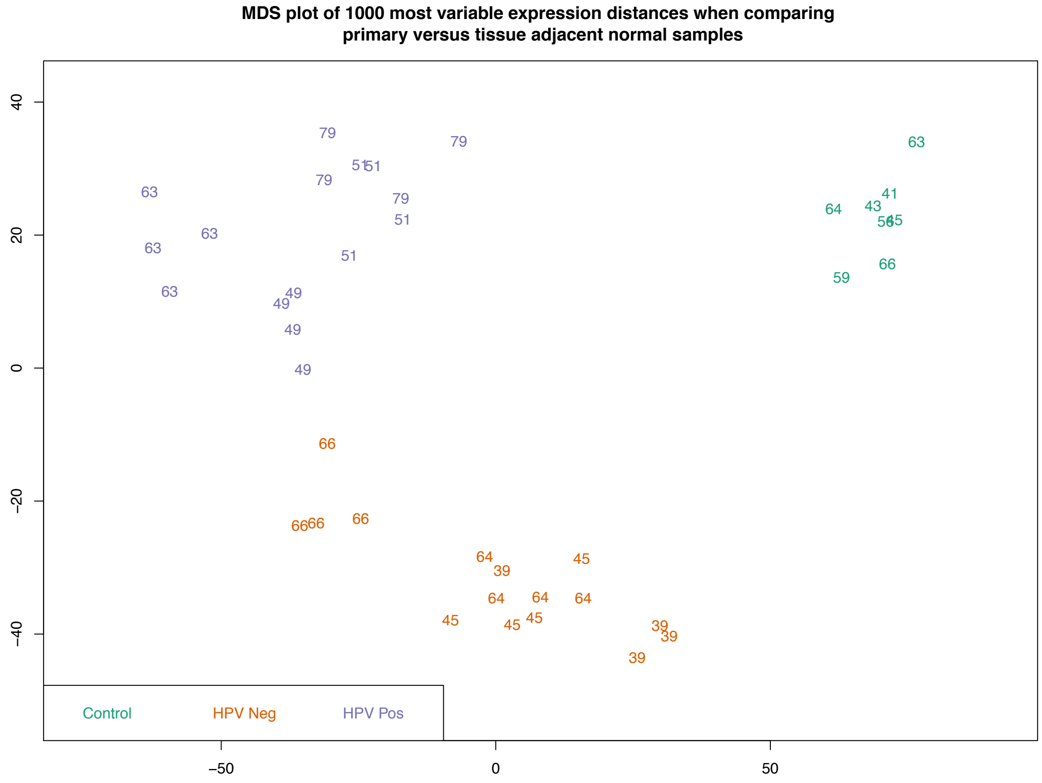


Figure 74: MDS plot of top 1,000 most variably expressed genes within primary tumour and tissue adjacent normal penile samples. Each number assigned to each data-point depicts the patient identifier. Green data points depict tissue adjacent 'normal' control samples. Orange data points belong to HPV negative samples and purple data points belong to HPV positive samples.

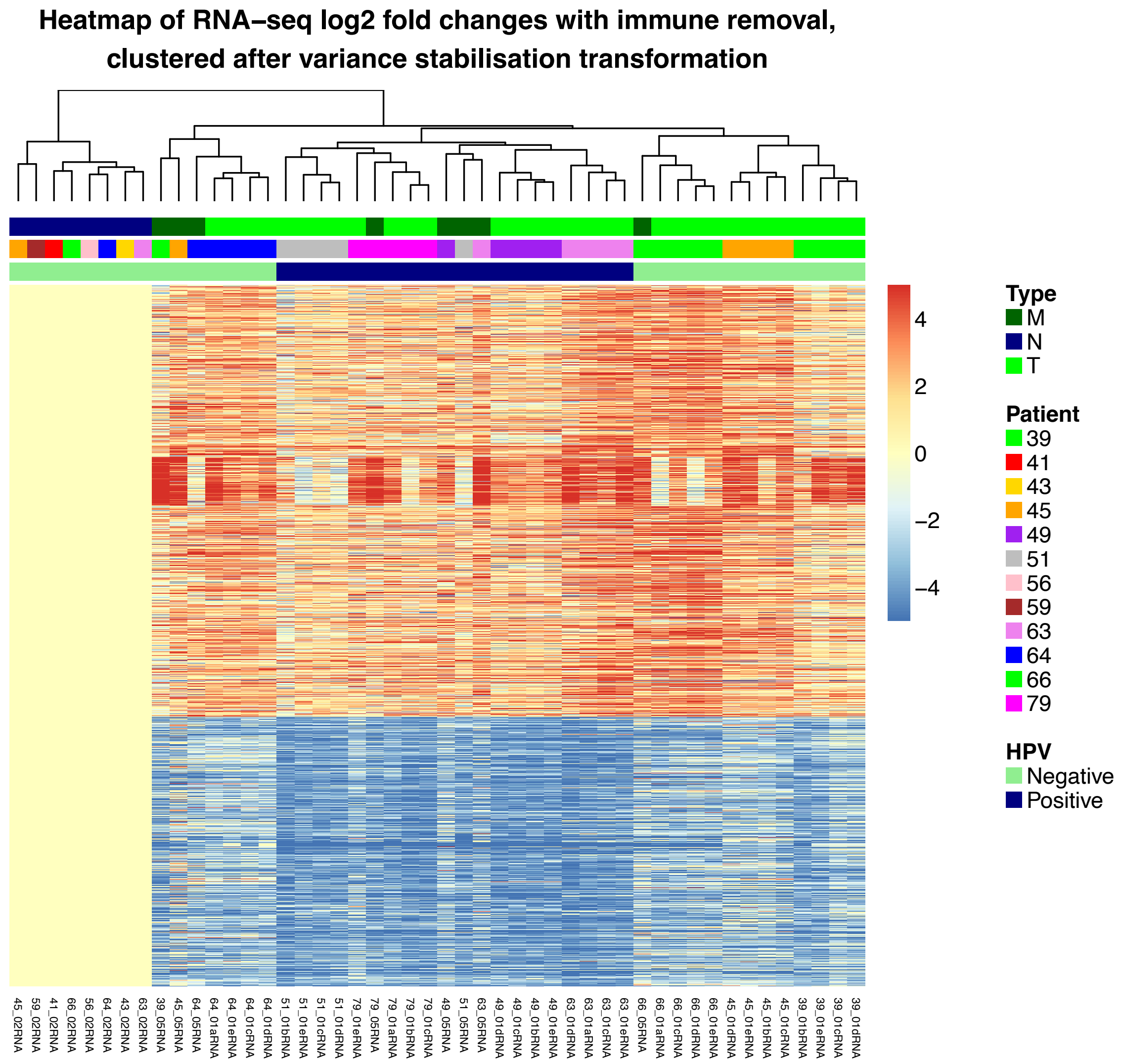


Figure 75: Heatmap of distances of differentially expressed genes when comparing all tumour and tissue adjacent 'normal' controls. Genes associated with the immune xCell signature have been removed from the analysis. Tumour samples with the sample suffix _05RNA depict lymph node metastases. Samples with suffix _01aRNA, _01bRNA, _01cRNA, _01dRNA, _01eRNA depict primary tumour samples. The two-digit sample prefix depicts the unique patient code the samples were taken from. Further information on the samples can be found in the Methods, Chapter 2.

5.4 Differential expression

Genes that are recurrently over- or under-expressed within samples are more likely to play a role in oncogenesis. Comparing differential gene expression between tumours and within tumours can be used to model the extent of inter- and intra-tumour heterogeneity. Differential gene expression was assessed using the R package DESeq2 as explained in the Methods (Chapter 2).

5.4.1 Primary versus normal

Primary tumour versus tissue adjacent normal samples were compared as described in the Methods (Chapter 2). The extent of differential expression was assessed by calculating log2 fold changes and adjusted p values. A macro view of the extent of differential expression between primary and tissue adjacent normal samples can be visualised in Figure 76. A comparison table of total number of over-expressed and under-expressed genes for a range of scenarios can be found in Table 27.

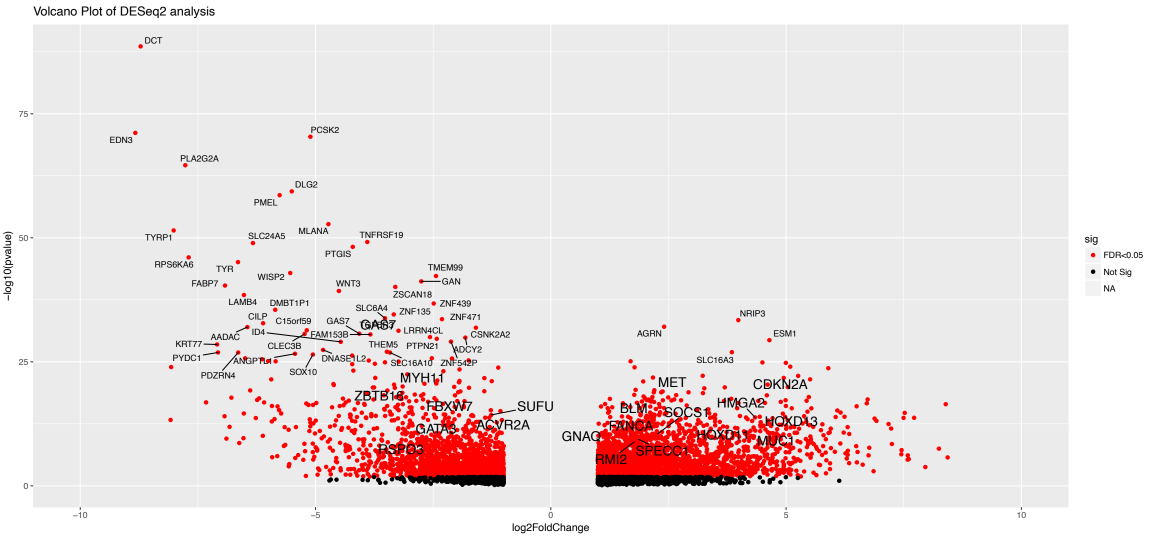


Figure 76: Volcano plot depicting the log2 fold change against level of significance for all genes. Annotated genes in the larger font size are significantly over-expressed or under-expressed that have previously been characterised in the COSMIC database.

Table 27: Comparison of total number of genes significantly differentially expressed with adjusted p value < 0.05 and an absolute log2 fold change of greater than 1 for primary versus normal, all tumour samples versus normal, HPV positive versus normal and HPV negative versus normal.

Comparison	Number of genes overexpressed	Number of genes underexpressed	Total number of genes differentially expressed	Percentage of genes overexpressed
Tumour versus tissue adjacent normal	3074	1673	4747	64.76%
Primary cancer versus tissue adjacent normal	2382	1836	4218	56.47%
HPV positive cancer samples versus normal	1377	1359	2736	50.33%
HPV negative cancer samples versus normal	1459	937	2396	60.89%

In order to assess the most likely drivers in more detail, the top 50 differentially expressed genes with the smallest p value were selected for downstream analysis. These genes had additionally passed the hard filter of an absolute minimum log2 fold change of 1. The number 50 was chosen to balance resolution with depth of analysis. The top 50 over- and under-expressed genes can be seen in Figure 77 to Figure 80.

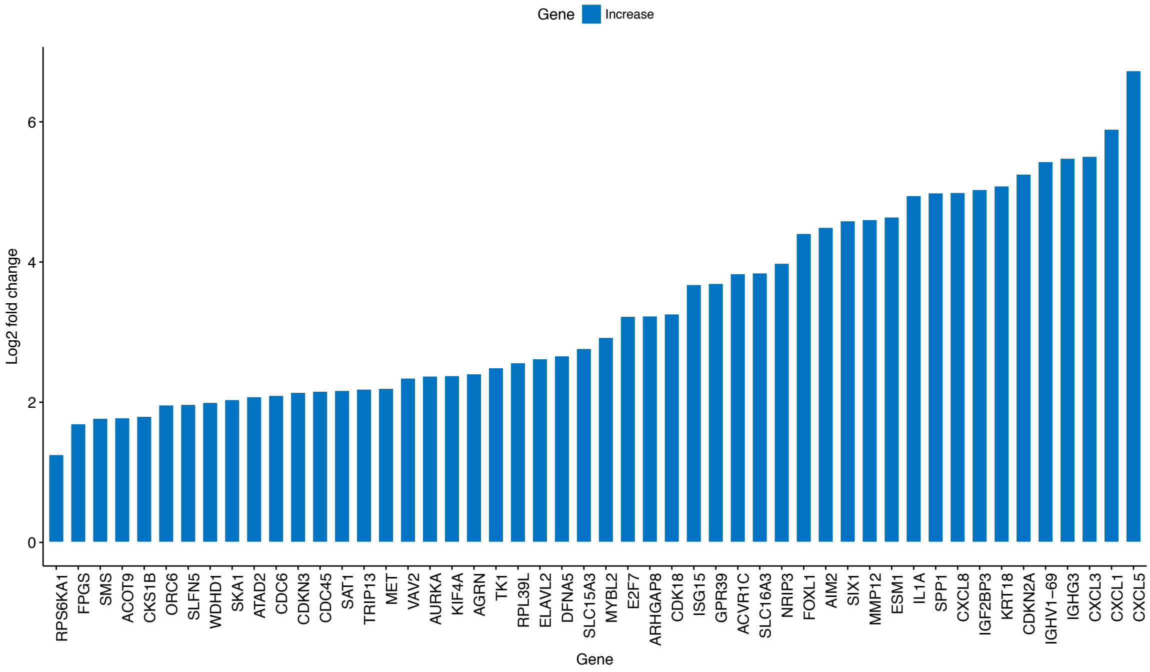


Figure 77: Top over-expressed genes as ranked by log2 fold change between primary and tissue adjacent normal samples.

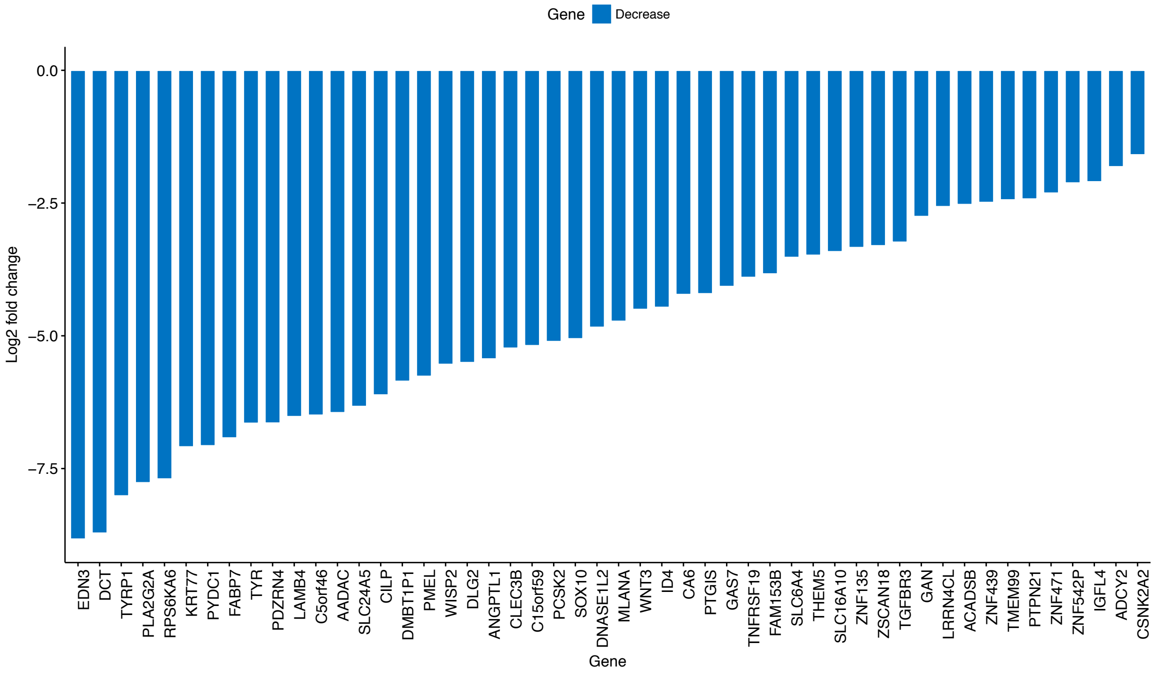


Figure 78: Top under-expressed genes as ranked by log2 fold change between primary and tissue adjacent normal samples.

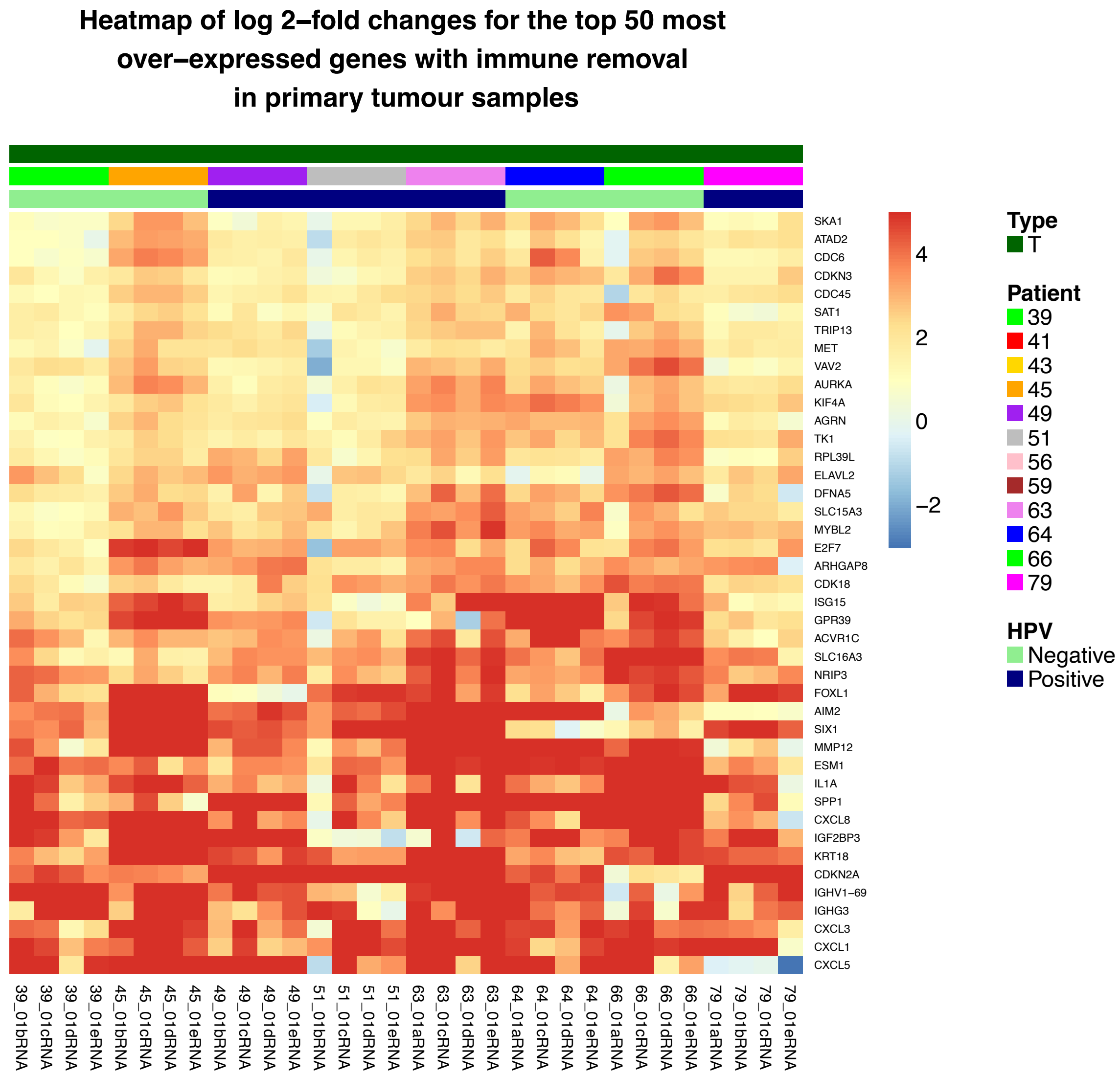


Figure 79: Heatmap of log2 fold changes for the 50 most over-expressed genes when comparing primary versus tissue adjacent tumour samples. Heatmap cell colours depict log2 fold changes each gene ranging from -5 (red) to +5 (blue).

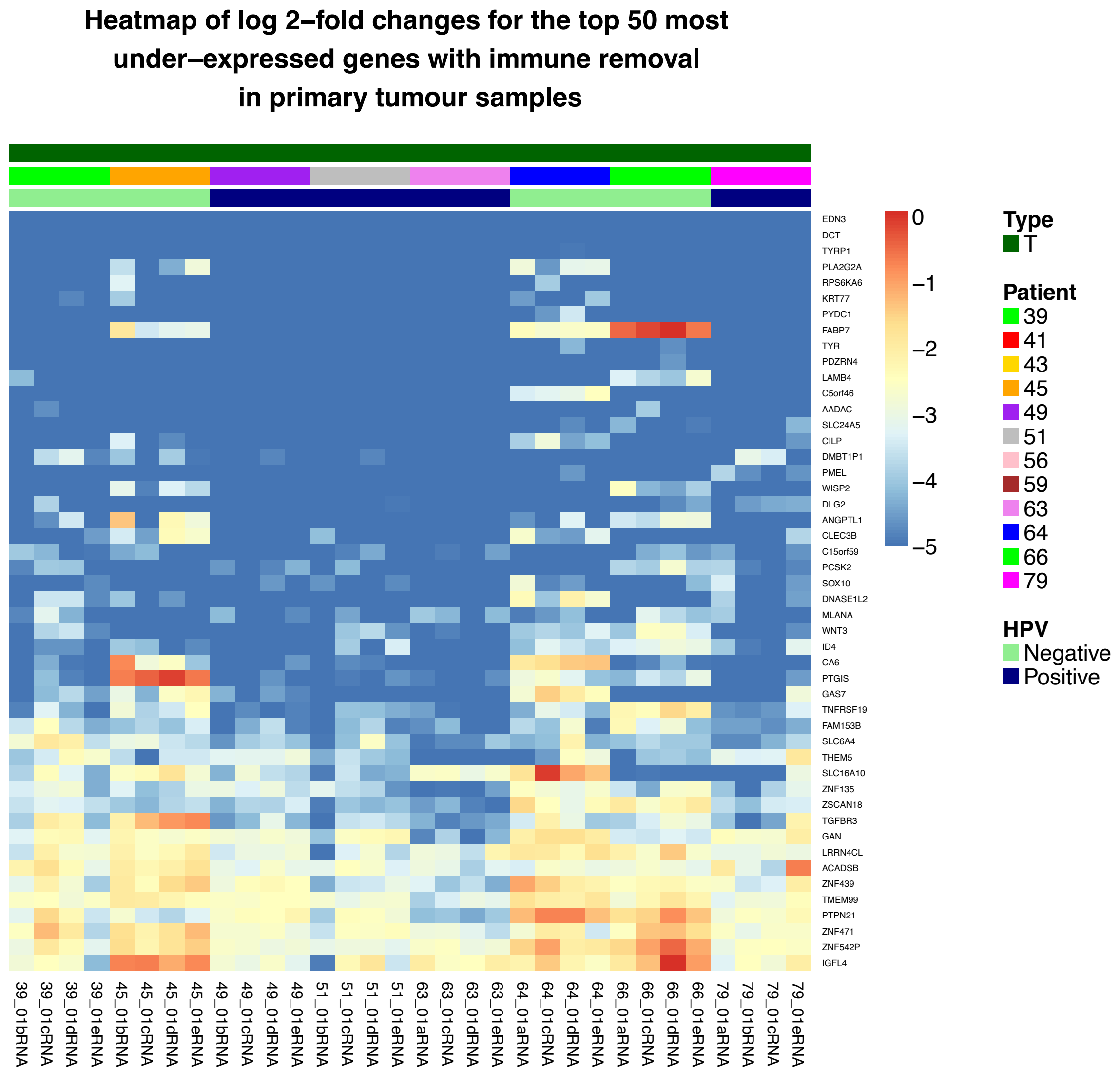


Figure 80: Heatmap of log2 fold changes for the 50 most under-expressed genes when comparing primary versus tissue adjacent tumour samples. Heatmap cell colours depict log2 fold changes each gene ranging from 0 (red) to -5 (blue).

To ensure that large log2 fold changes in potential gene drivers were not being filtered out, due to only occurring in a smaller number of samples, the adjusted p value was temporarily relaxed for the next four figures to 0.1. The genes were then further selected based on their presence in either the COSMIC database or the list of curated actionable genes discussed in the Methods (Chapter 2). These results are depicted in bar charts in Figure 81 and Figure 83, respectively. The heatmaps produced from this analysis demonstrate the frequency and amplitude of differential expression for these genes, in Figure 82 and Figure 84.

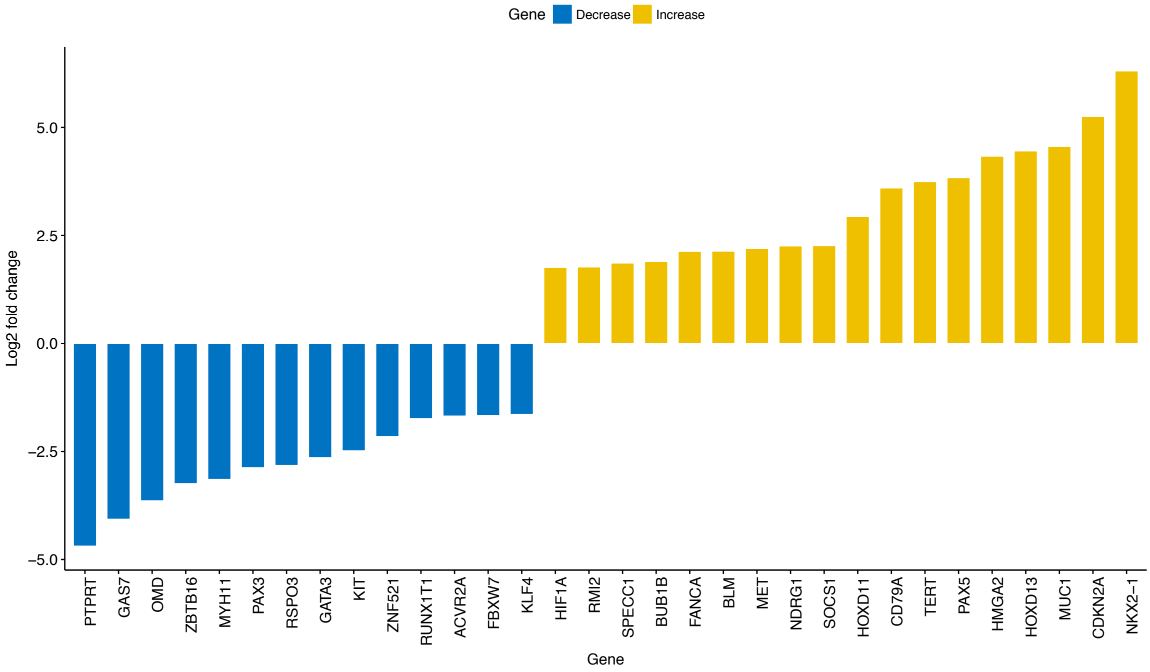


Figure 81: Bar chart with the height of each bar depicting the log2 fold change of the combined primary compared with tissue adjacent normal samples filtered by the presence in COSMIC gene database.

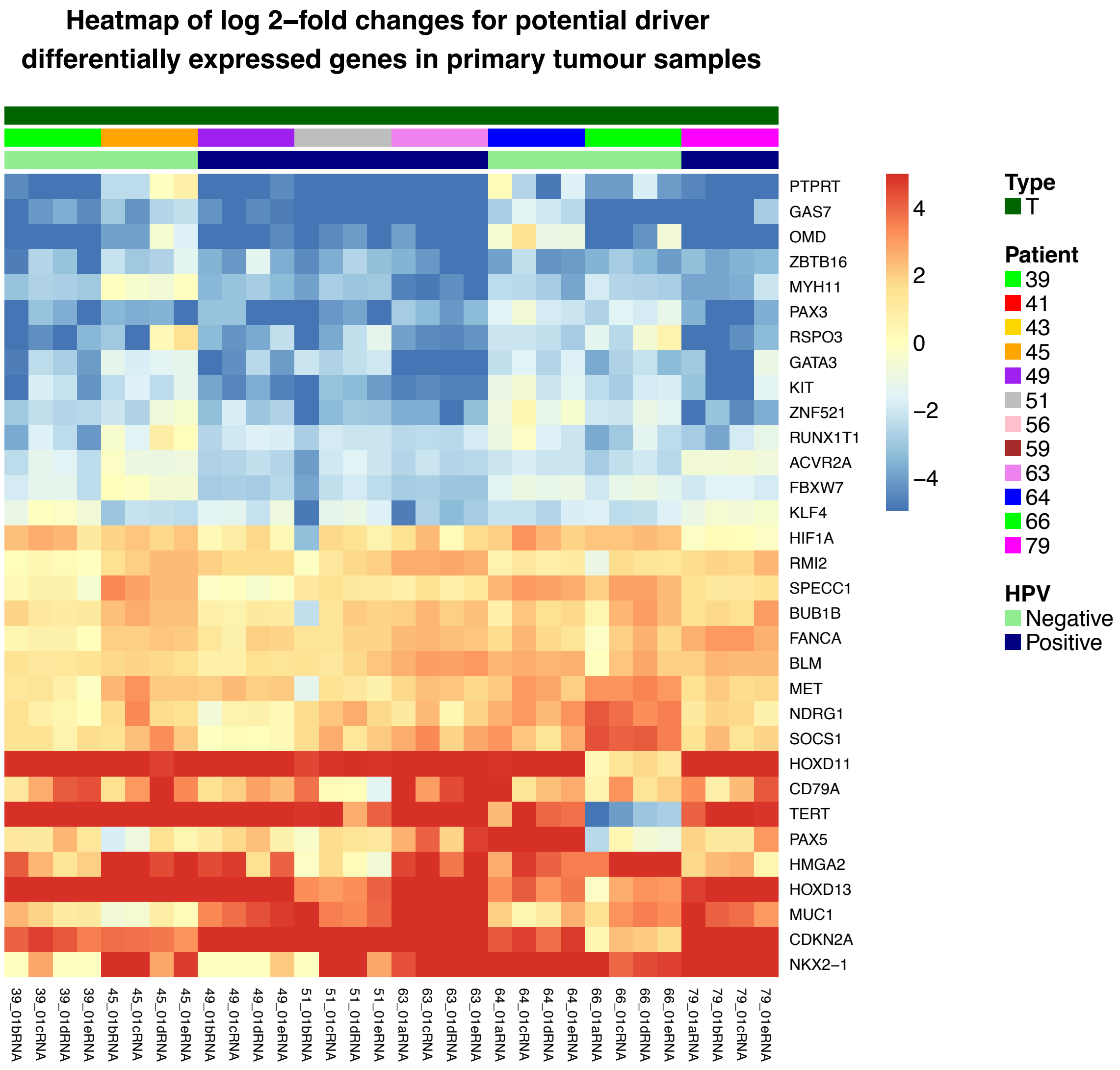


Figure 82: Heatmap of log2 fold changes for differentially expressed genes listed in the COSMIC database when comparing primary tumour samples with tissue adjacent controls. Heatmap cell colours depict log2 fold changes each gene ranging from -5 (red) to +5 (blue).

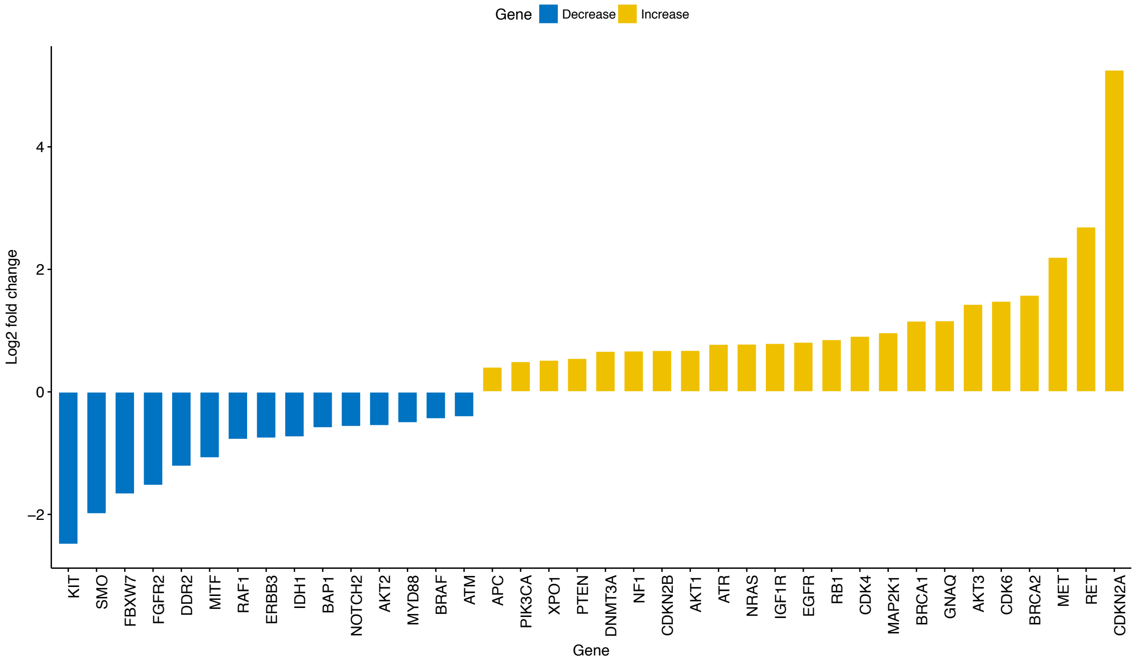


Figure 83: Bar chart with the height of each bar depicting the log2 fold change of the combined primary compared with tissue adjacent normal samples filtered whether the gene has previously been characterised as being therapeutically 'actionable'.

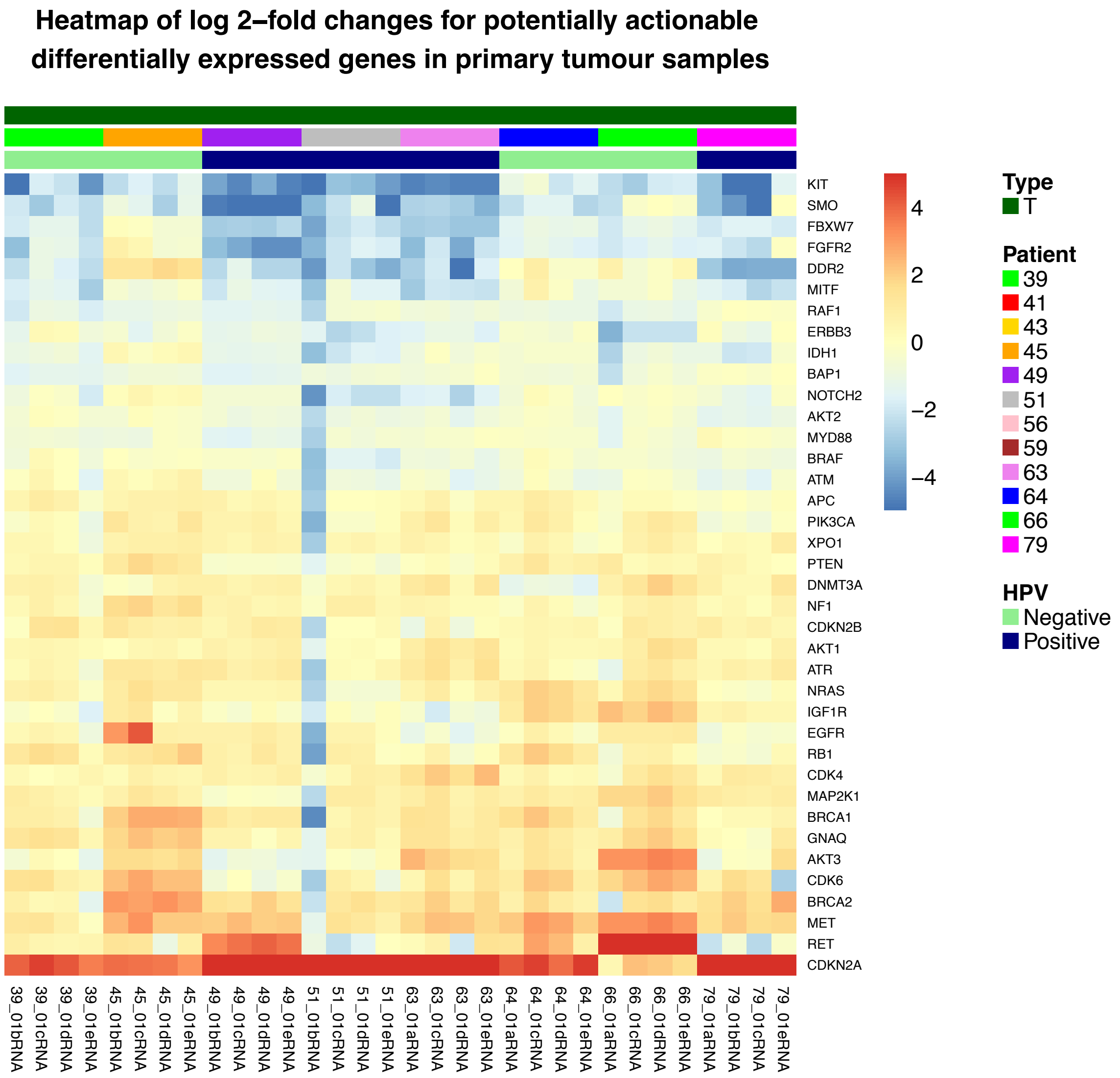


Figure 84: Heatmap of log2 fold changes for potentially actionable differentially expressed genes in primary tumour samples when comparing primary tumour samples with tissue adjacent controls. Heatmap cell colours depict log2 fold changes each gene ranging from -5 (red) to +5 (blue).

The differentially expressed genes also found in the COSMIC database were examined more closely by plotting the normalised transcript abundance for all primary versus adjacent tissue control samples, as shown in Figure 85 to Figure 88. These scatter plots indicate how many samples in the PenHet cohort also shared the same differential expression. In addition, clustering of certain samples can be visualised and examined more closely. An example of patient clustering can be seen in Figure 85 for *CDKN2A,* where the tumour samples appear to cluster into two distinct, well-defined groups. As seen below in Section 5.4.2, assessing differential expression in the context of HPV status these two groups can also be separated based on HPV status.

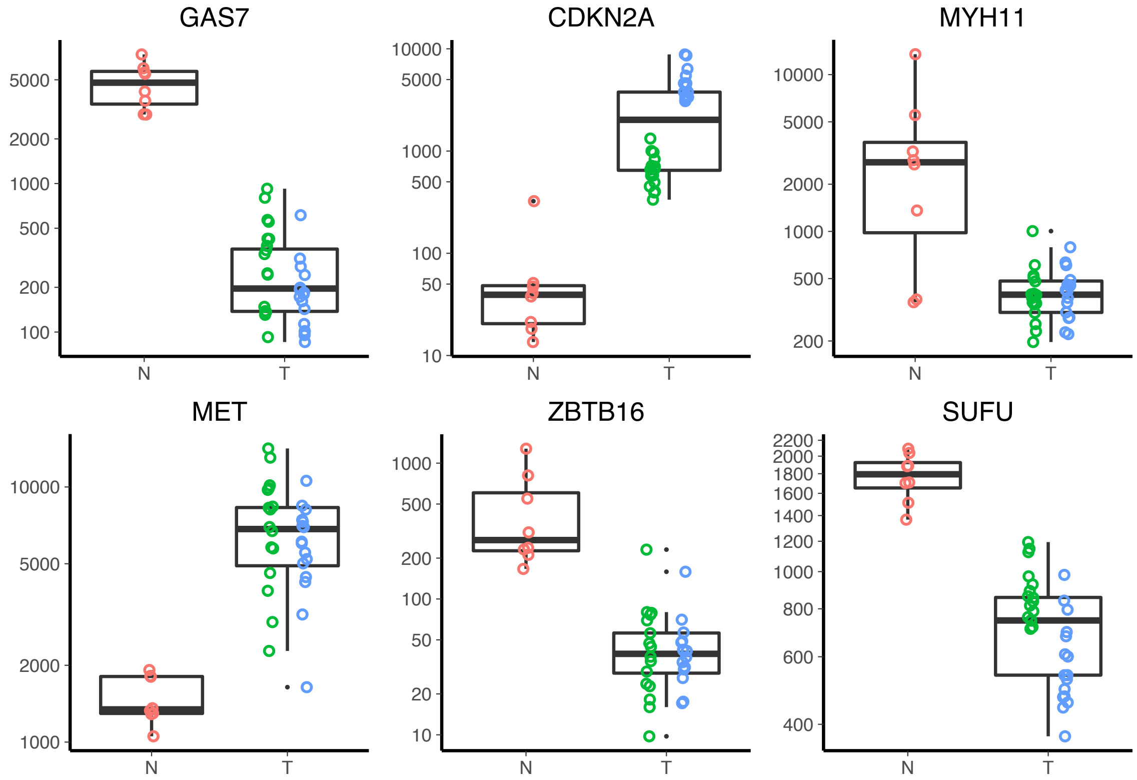


Figure 85: Combined scatter and box plot of most significant change in log2 fold changes in primary versus adjacent normal tissue samples found in the COSMIC gene database. N = tissue adjacent normal controls. T = primary tumour samples. HPV positive tumour samples scatter points are in blue, HPV negative scatter points are in green. Adjacent normal scatter points are in red. All of these have an adjusted p value < 0.0001 between tumour and normal samples.

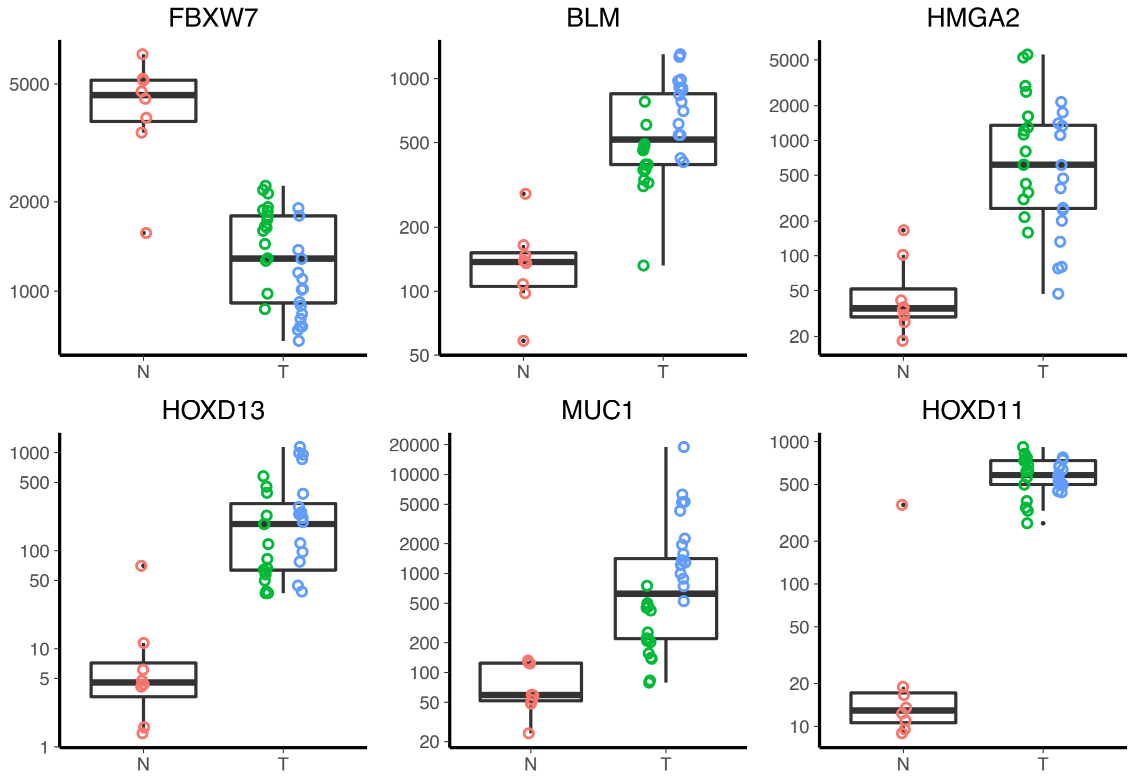


Figure 86: Combined scatter and box plot of most significant change in log2 fold changes in primary versus adjacent normal tissue samples found in the COSMIC gene database. N = tissue adjacent normal controls. T = primary tumour samples. HPV positive tumour samples scatter points are in blue, HPV negative scatter points are in green. Adjacent normal scatter points are in red. All of these have an adjusted p value < 0.0001 between tumour and normal samples.

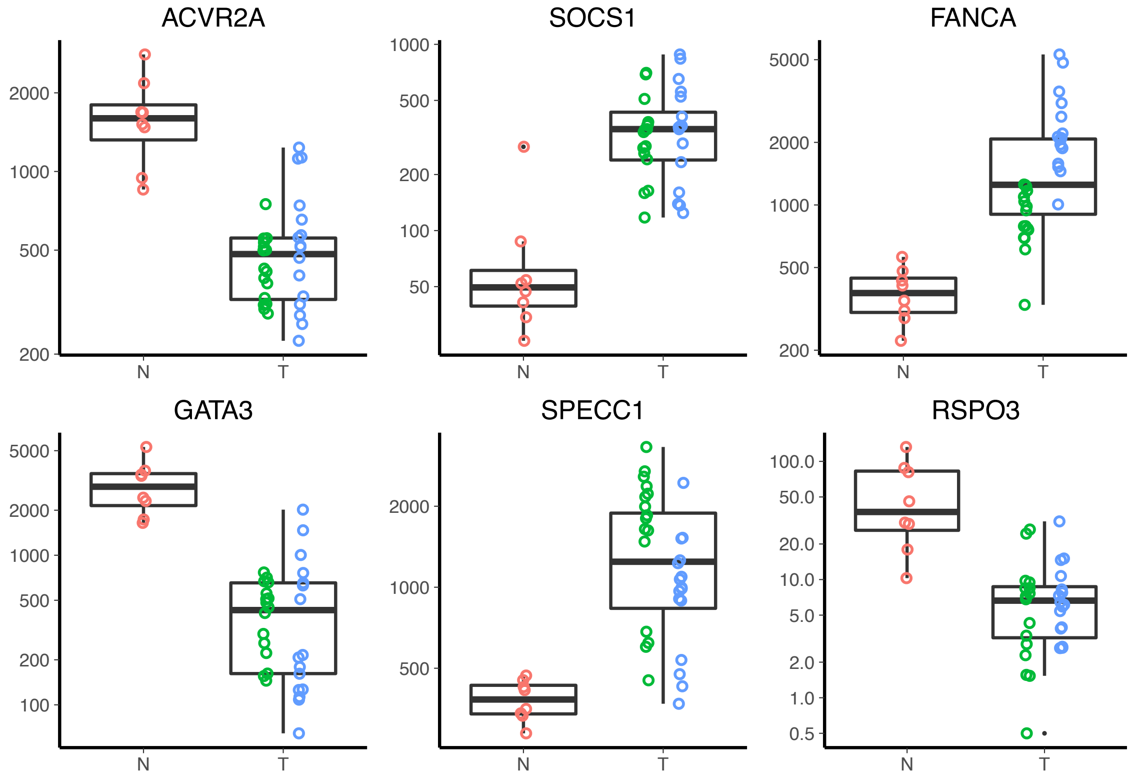


Figure 87: Combined scatter and box plot of most significant change in log2 fold changes in primary versus adjacent normal tissue samples found in the COSMIC gene database. N = tissue adjacent normal controls. T = primary tumour samples. HPV positive tumour samples scatter points are in blue, HPV negative scatter points are in green. Adjacent normal scatter points are in red. All of these have an adjusted p value < 0.0001 between tumour and normal samples.

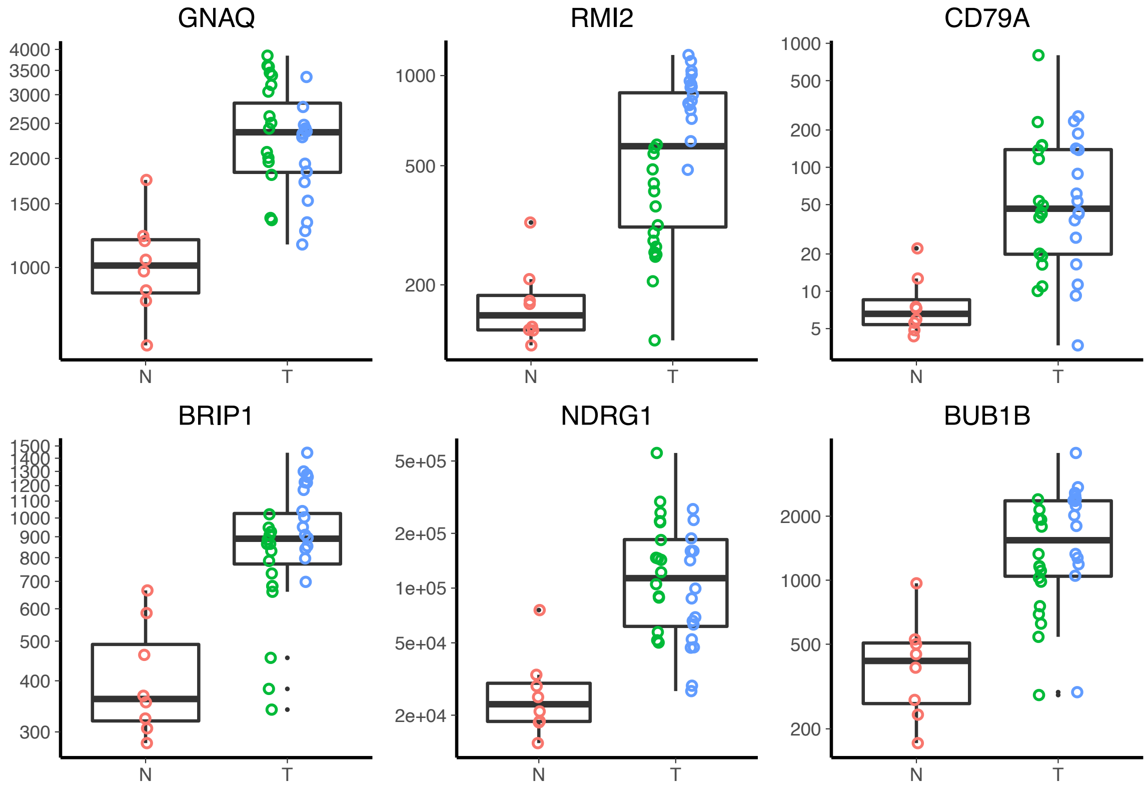


Figure 88: Combined scatter and box plot of most significant change in log2 fold changes in primary versus adjacent normal tissue samples found in the COSMIC gene database. N = tissue adjacent normal controls. T = primary tumour samples. HPV positive tumour samples scatter points are in blue, HPV negative scatter points are in green. Adjacent normal scatter points are in red. All of these have an adjusted p value < 0.0001 between tumour and normal samples.

In order to gain a greater understanding into the pathways and processes that may be disturbed by the differential expression, the differentially expressed genes were assessed by KEGG pathway analysis. Table 28 displays perturbed pathways when comparing primary versus tissue adjacent normal samples. After performing p value correction for multiple testing, only two pathways remained statistically overrepresented: the generalised Cell-Cycle pathway and the JAK-STAT pathway (both p = 0.003, adjusted p = 0.0998). The JAK-STAT pathway can be visualised in Figure 89. Although not statistically significant, the remaining overrepresented KEGG pathways were very similar to those found in the methylation analysis in Chapter 4. KEGG diagrams of these other most overrepresented pathways (mTOR, PI3K-AKT and TP53) can be seen in the Appendix.

Table 28: KEGG pathway analysis representing differentially expressed pathways when comparing primary tumour samples with tissue adjacent normal samples.

Original p-value	Adjusted p-value	Number of genes	Pathway identified	Pathway name
0.0014	0.0999	47	hsa04110	Cell cycle
0.0019	0.0999	51	hsa04630	Jak-STAT signaling pathway
0.0049	0.1026	21	hsa00140	Steroid hormone biosynthesis
0.0060	0.1042	17	hsa00480	Glutathione metabolism
0.0114	0.1710	28	hsa00830	Retinol metabolism
0.0178	0.2120	28	hsa04916	Melanogenesis
0.0182	0.2120	13	hsa00071	Fatty acid metabolism
0.0244	0.2317	36	hsa04270	Vascular smooth muscle contraction
0.0262	0.2317	13	hsa02010	ABC transporters
0.0265	0.2317	21	hsa00350	Tyrosine metabolism
0.0351	0.2835	22	hsa04340	Hedgehog signaling pathway
0.0416	0.3122	12	hsa00565	Ether lipid metabolism
0.0462	0.3232	34	hsa04972	Pancreatic secretion
0.0512	0.3347	29	hsa03320	PPAR signaling pathway
0.0542	0.3347	40	hsa04360	Axon guidance
0.0097	0.3408	33	hsa04620	Toll-like receptor signaling pathway
0.0606	0.3534	17	hsa04975	Fat digestion and absorption
0.0170	0.4043	36	hsa04650	Natural killer cell mediated cytotoxicity
0.0194	0.4043	68	hsa04510	Focal adhesion
0.0231	0.4043	53	hsa04062	Chemokine signaling pathway
0.0270	0.4048	21	hsa04622	RIG-I-like receptor signaling pathway
0.0826	0.4242	25	hsa04912	GnRH signaling pathway
0.0845	0.4242	11	hsa00053	Ascorbate and aldarate metabolism
0.0848	0.4242	11	hsa00280	Valine, leucine and isoleucine degradation
0.0917	0.4375	13	hsa04960	Aldosterone-regulated sodium reabsorption
0.0971	0.4431	15	hsa00564	Glycerophospholipid metabolism
0.1059	0.4633	18	hsa00600	Sphingolipid metabolism
0.1336	0.5611	39	hsa04310	Wnt signaling pathway
0.0459	0.5866	21	hsa03030	DNA replication
0.0558	0.5866	14	hsa04623	Cytosolic DNA-sensing pathway
0.0607	0.5866	21	hsa04210	Apoptosis
0.0626	0.5866	10	hsa00520	Amino sugar and nucleotide sugar metabolism
0.0670	0.5866	50	hsa04380	Osteoclast differentiation
0.0801	0.6236	10	hsa03440	Homologous recombination
0.0832	0.6236	19	hsa04115	p53 signaling pathway
0.0946	0.6624	17	hsa00240	Pyrimidine metabolism

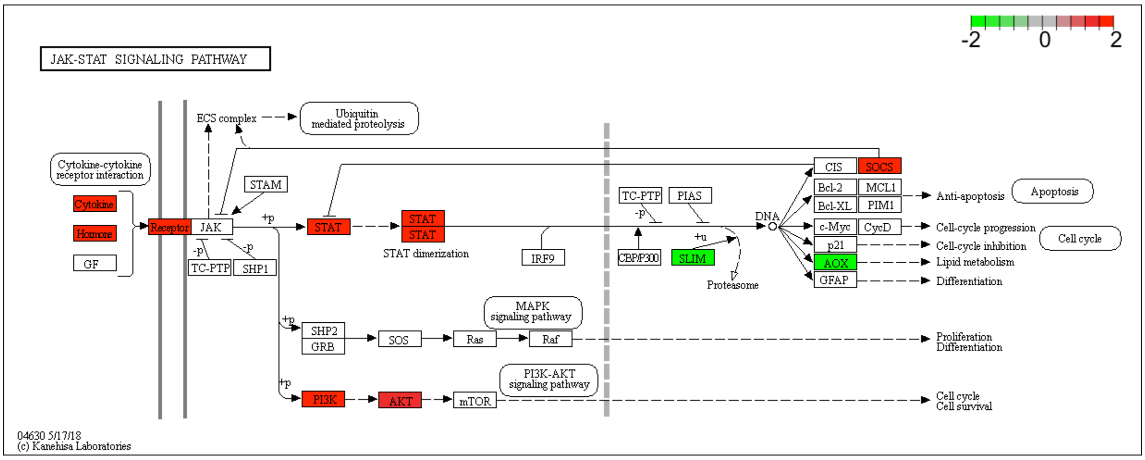


Figure 89: Diagram of differentially expressed genes when comparing primary tumour samples with tissue adjacent normal samples in the JAK-STAT pathway. Green = loss of expression. Red = over-expressed.

5.4.1.1 Expression of immune checkpoints

Immunotherapy, involving immune checkpoint inhibitors, has gained deserved attention as it has been proven to have enduring therapeutic activity in many cancers including melanoma[78], lung[79] and bladder carcinomas[80]. One factor these cancers have in common is a high mutational load. Further biomarkers have been proposed, including over-expression of *PDL1*, *LAG3*, *IDO1* and the presence of CD8 +ve T cells, and there is at least some evidence for predicting immunotherapy response [216]. The specific expression of these immune checkpoints and associated biomarkers when comparing primary versus matched tissue adjacent normals can be found in Section 5.4.1.1 and Figure 90 below.

Given that immunotherapy is an increasingly used treatment modality, the expression of eight genes previously found to be immune suppressors or activators were tested[217]. Statistically significant increased expression was found in the tumour samples for *CD274 (PDL-1)*, *CTLA4*, *LAG3*, *IDO1, TIM-3* and *KIR*, as demonstrated in the plots in Figure 90. Increased expression in *PCCD1 (PD1)* was found in only a few individual samples. Increased expression of checkpoint inhibitors CTLA-4 and PDL-1 was also demonstrated in all samples bar two, displaying increased expression of CTLA-4 outside the interquartile range of the tissue adjacent histopathologically normal samples. Research into uncovering further inhibitory immune receptors has led to the discovery of lymphocyte activation gene-3 (*LAG3*)[218] and Indoleamine 2,3-dioxigenase 1 (*IDO1*)[219] both of which can promote immune tolerance to tumour antigens. Both of these genes were differentially expressed, particularly *IDO1* ($p < 0.0001$). Further work is also underway in development of immune modulators for *TIM3*[220], *CD137*[221] and *KIR*[222].

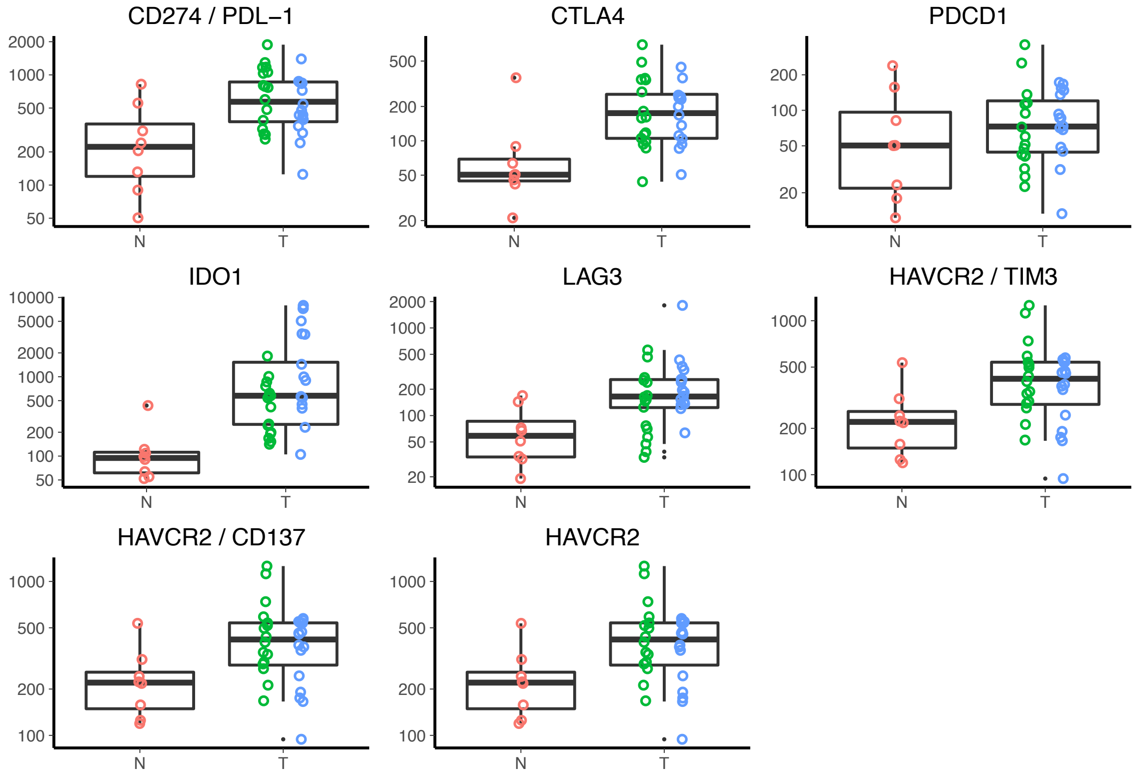


Figure 90: Combined scatter and box plot of log2 fold expression changes of immune checkpoint molecules. Comparison between tumour and tissue adjacent normal samples. N = tissue adjacent normal controls. T = primary tumour samples. HPV positive tumour samples scatter points are in blue, HPV negative scatter points are in green. Adjacent normal scatter points are in red.

5.4.2 HPV positive versus normal in comparison with HPV negative versus normal

As with the previous methylation and mutation analysis, the expression profiles of all primary tumour samples clustered by HPV, as shown in Figure 74. This stratification of samples and patients can result in a reduced sensitivity to discover driver genes. The differentially expressed genes found above were significant despite this stratification of patients. They therefore represent genes which are important in penile cancer despite HPV status. The proportion of these changes that are recurrent throughout the PenHet cohort will be assessed in Section 5.7.

In order to determine whether HPV status is associated to a particular set of differentially expressed genes, DESeq2 was used to compare HPV positive samples and HPV negative samples with adjacent normal tissue. The results of these two differential expression experiments were then compared to produce unique signatures of genes whose expression is associated with HPV status. The statistical power for detecting differentially expressed genes is dramatically reduced when performing this subgroup analysis. Therefore, an even greater log2 fold change is required for a particular gene to be classified as differentially expressed in the HPV positive or negative cohort.

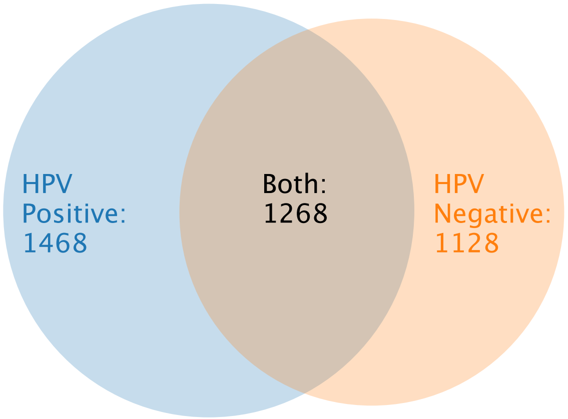


Figure 91: Venn diagram depicting the number of differentially expressed genes found to be unique to and shared between HPV positive and HPV negative primary tumour samples. The circle on the left refers to HPV positive samples, and the circle on the right refers to HPV negative samples. The 1,268 in the middle section are differentially expressed genes which are shared between both HPV positive and HPV negative samples.

Figure 91 is a Venn diagram depicting the overlap of expression changes in HPV positive compared with tissue adjacent normal, in comparison to HPV negatives compared with tissue adjacent normal. 1,468 and 1,128 genes were found to be unique to HPV positive and HPV negative samples respectively. 1,268 genes were found to be common to both. Of the total number of genes differentially expressed in HPV positive samples, 1,377 (50.3%) were over-expressed, compared with 1,459 (60.9%) over-expressed in HPV negative samples (Table 27). These shared and unique genes were assessed in greater detail by focusing on the previously described important drivers. A filtered dataset of genes also present in the COSMIC database was used for the results in Figure 92 and Figure 93.

A closer look at some of these genes are assessed in scatterplots 1-4 in Figure 94 to Figure 97. These are selected on the basis of having the greatest differences in log2 fold changes between the HPV positive and negative samples. Genes that stand out include *CDKN2A* and *TP53*. *TP53*, which is mutated in all HPV negative samples, appears to have a direct loss of expression solely in those HPV negative samples. In the case of *CDKN2A*, there appears to be a step-fold increase in expression of *CDKN2A* when comparing normal controls with HPV negative and HPV positive samples.

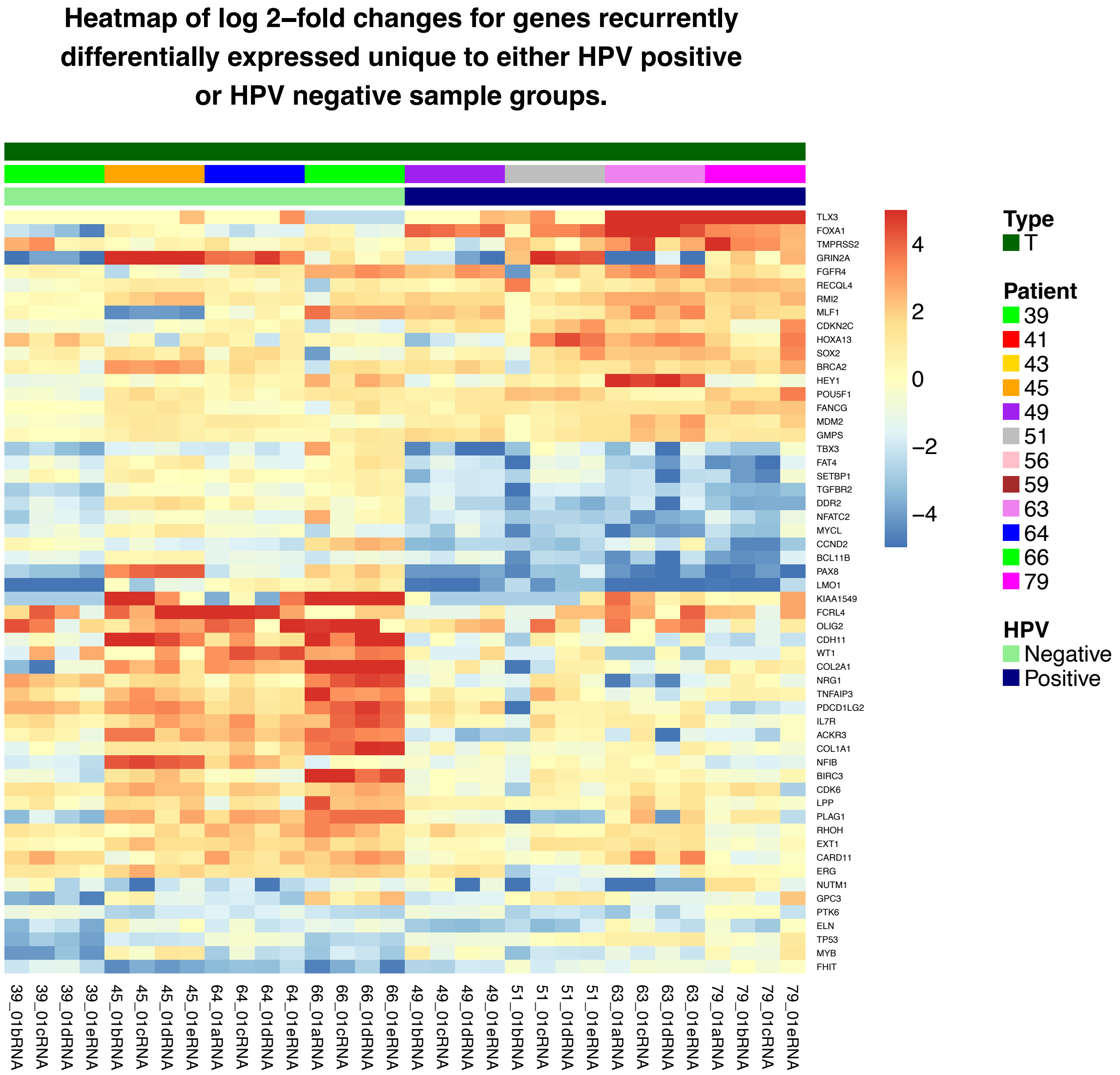


Figure 92: Genes previously described in the COSMIC database that are differentially expressed solely in the HPV positive or negative samples of the PenHet cohort. Heatmap cell colours depict log2 fold changes each gene ranging from -5 (red) to +5 (blue).

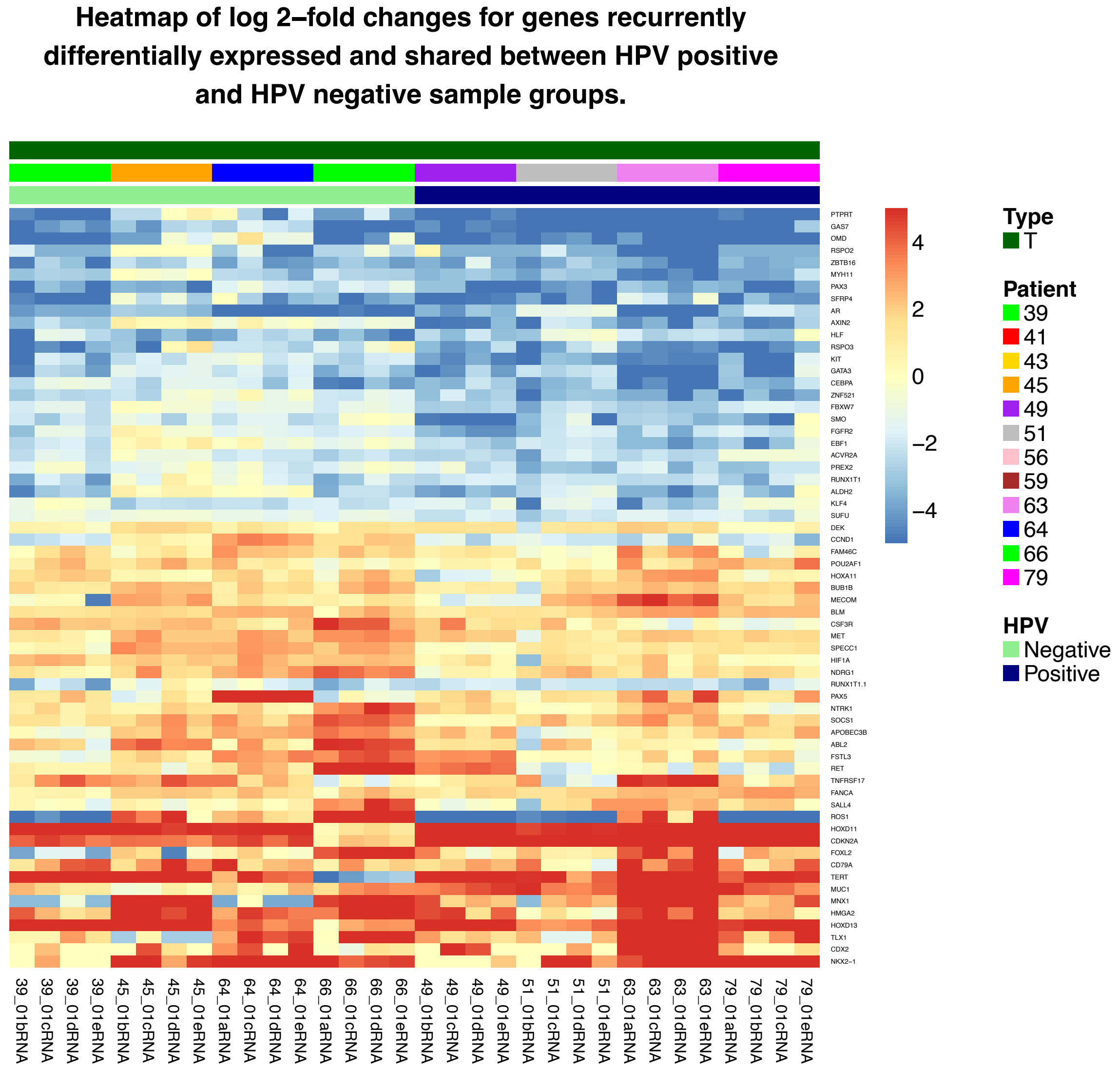


Figure 93: Genes previously described in the COSMIC database that are differentially expressed and shared throughout both HPV positive or negative samples in the PenHet cohort. Heatmap cell colours depict log2 fold changes each gene ranging from -5 (red) to +5 (blue).

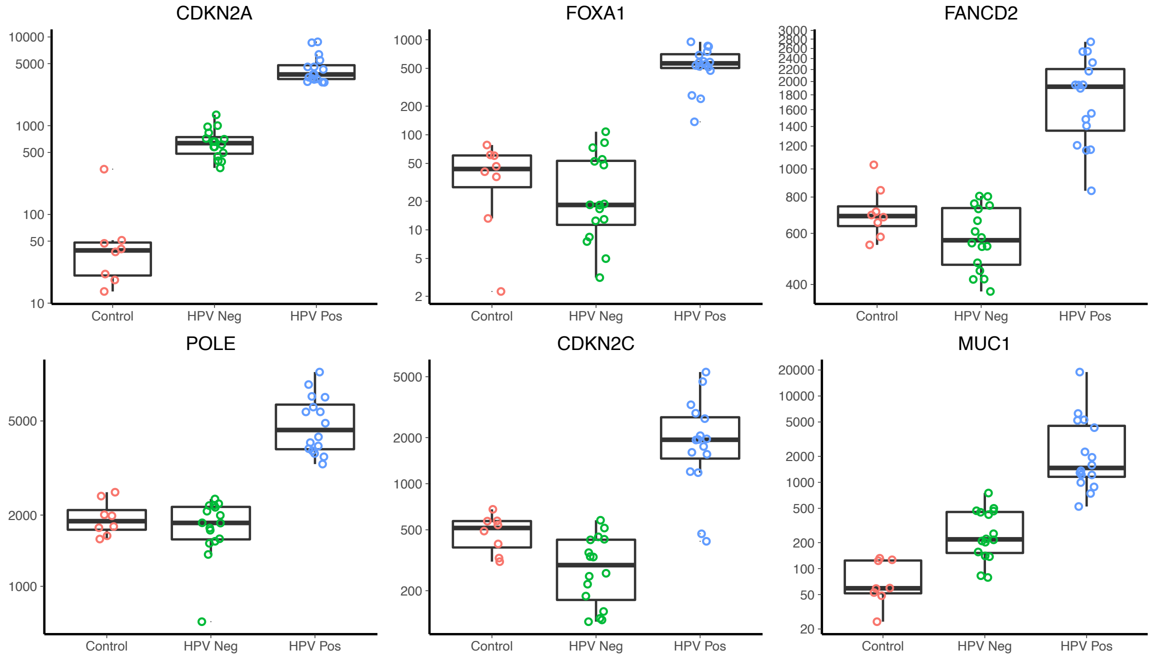


Figure 94: Scatter plot (1) of most significant log2 fold gene expression changes HPV positive in comparison to HPV negative primary tumour samples versus adjacent normal tissue samples for genes found in the COSMIC gene database. Control = tissue adjacent normal controls. HPV Neg = HPV negative primary tumour samples. HPV Pos = HPV positive primary tumour samples.

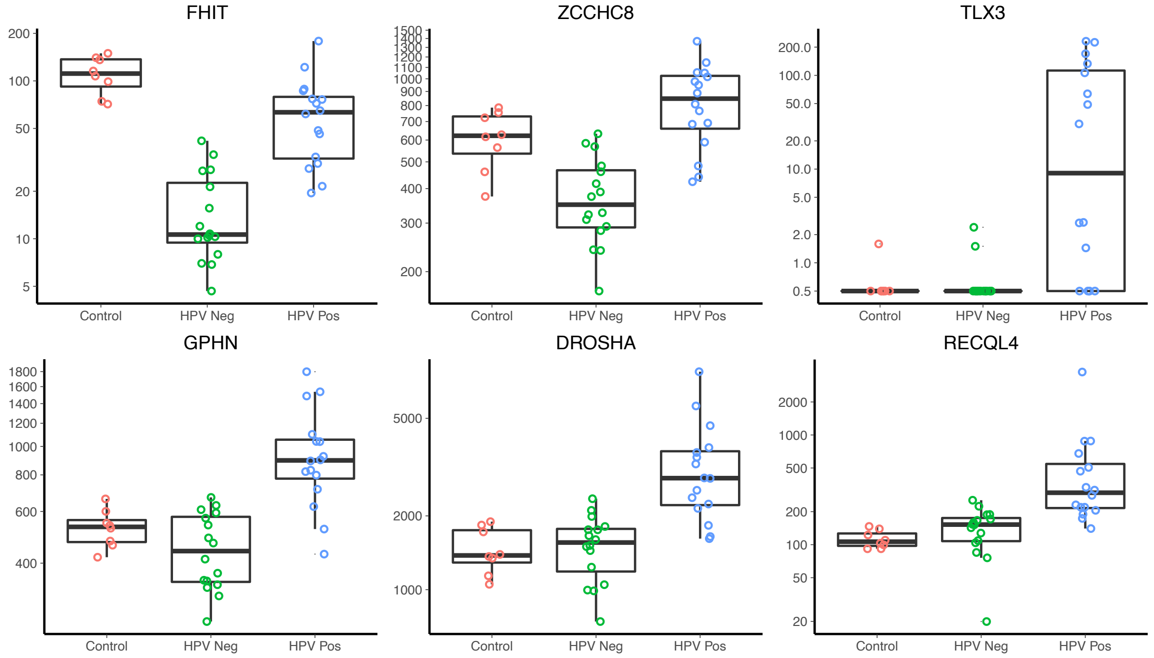


Figure 95: Scatter plot (2) of most significant log2 fold gene expression changes HPV positive in comparison to HPV negative primary tumour samples versus adjacent normal tissue samples for genes found in the COSMIC gene database. Control = tissue adjacent normal controls. HPV Neg = HPV negative primary tumour samples. HPV Pos = HPV positive primary tumour samples.

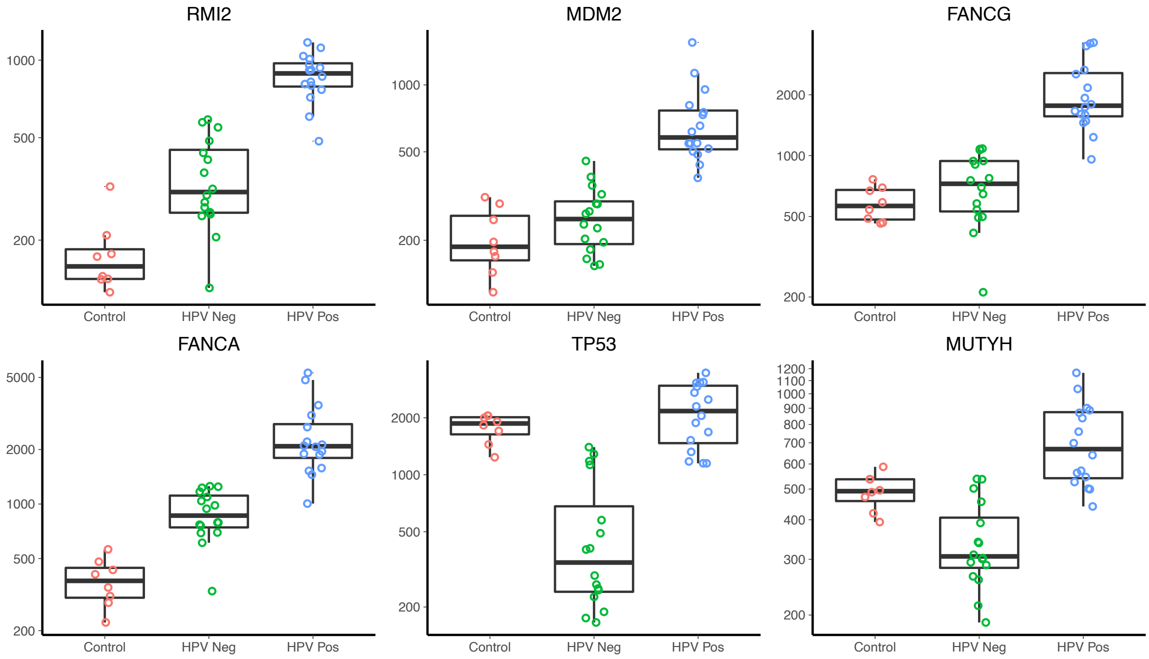


Figure 96: Scatter plot (3) of most significant log2 fold gene expression changes HPV positive in comparison to HPV negative primary tumour samples versus adjacent normal tissue samples for genes found in the COSMIC gene database. Control = tissue adjacent normal controls. HPV Neg = HPV negative primary tumour samples. HPV Pos = HPV positive primary tumour samples.

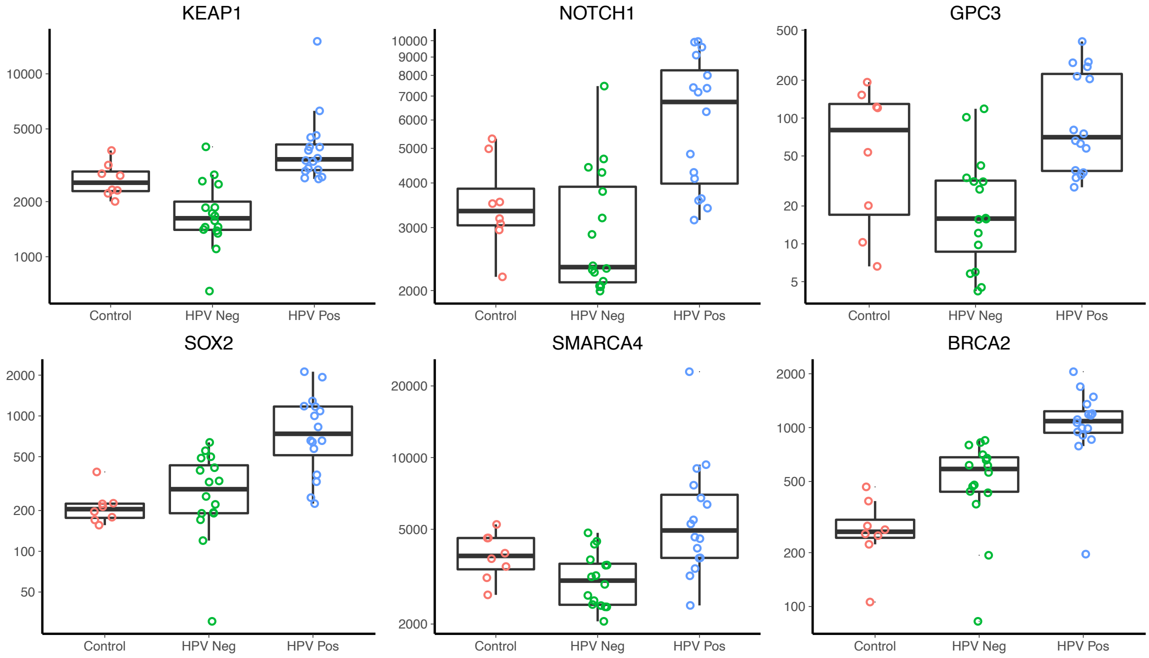
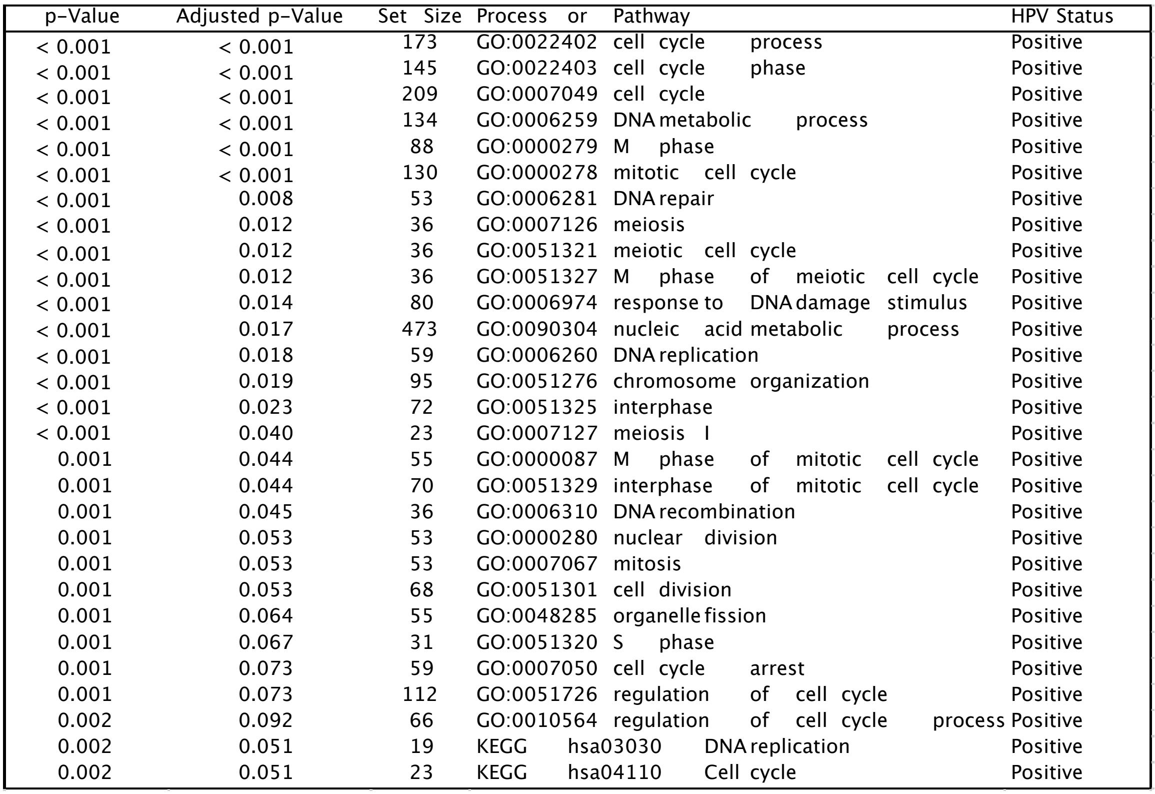


Figure 97: Scatter plot (4) of most significant log2 fold gene expression changes HPV positive in comparison to HPV negative primary tumour samples versus adjacent normal tissue samples for genes found in the COSMIC gene database. Control = tissue adjacent normal controls. HPV Neg = HPV negative primary tumour samples. HPV Pos = HPV positive primary tumour samples.

Gene set enrichment analysis was undertaken to ascertain if there were particular processes enriched for either HPV positive or HPV negative samples. This was undertaken as explained in the Methods (Chapter 2), using the R package ‘gage’. The statistical power to demonstrate aberrant expression in only HPV positive or HPV negative samples was limited due to the low numbers in the individual cohorts. Genes that were differentially expressed in both HPV positive and negative samples were excluded to elucidate what processes may uniquely pertain to either group of patients.

The pathways that reached or approached statistical significance are shown in Table 29 and Table 30. KEGG pathway analysis revealed disturbed cell adhesion processes in the HPV negative samples, whereas pathways in HPV positive samples focused on cell cycle and DNA replication. GEO biological processes assessment revealed only statistically significantly derangement in HPV positive samples with greater than expected involvement of cell cycle processes, cell cycle regulation and DNA repair. Analysis of these individual pathways can be visualised in Figure 98 and Figure 99 which show DNA damage and cell cycle pathways respectively in HPV positive samples, and in Figure 100 which shows cell adhesion for HPV negative samples. As demonstrated in these figures, there is large-scale disturbed expression of cell cycle control genes as well as cell adhesion molecules. In particular, there is loss of expression of tumour suppressor *cyclin D1(CCND1)* and over-expression of oncogene *MDM2.*

Table 29: Gene set enrichment analysis of HPV positive samples when compared with tissue adjacent controls. KEGG and GO terms representing differential expression across a pathway or biological process.



p-Value	Adjusted p-Value	Set Size	Process or Pathway	HPV Status
< 0.001	< 0.001	173	GO:0022402 cell cycle process	Positive
< 0.001	< 0.001	145	GO:0022403 cell cycle phase	Positive
< 0.001	< 0.001	209	GO:0007049 cell cycle	Positive
< 0.001	< 0.001	134	GO:0006259 DNA metabolic process	Positive
< 0.001	< 0.001	88	GO:0000279 M phase	Positive
< 0.001	< 0.001	130	GO:0000278 mitotic cell cycle	Positive
< 0.001	0.008	53	GO:0006281 DNA repair	Positive
< 0.001	0.012	36	GO:0007126 meiosis	Positive
< 0.001	0.012	36	GO:0051321 meiotic cell cycle	Positive
< 0.001	0.012	36	GO:0051327 M phase of meiotic cell cycle	Positive
< 0.001	0.014	80	GO:0006974 response to DNA damage stimulus	Positive
< 0.001	0.017	473	GO:0090304 nucleic acid metabolic process	Positive
< 0.001	0.018	59	GO:0006260 DNA replication	Positive
< 0.001	0.019	95	GO:0051276 chromosome organization	Positive
< 0.001	0.023	72	GO:0051325 interphase	Positive
< 0.001	0.040	23	GO:0007127 meiosis I	Positive
0.001	0.044	55	GO:0000087 M phase of mitotic cell cycle	Positive
0.001	0.044	70	GO:0051329 interphase of mitotic cell cycle	Positive
0.001	0.045	36	GO:0006310 DNA recombination	Positive
0.001	0.053	53	GO:0000280 nuclear division	Positive
0.001	0.053	53	GO:0007067 mitosis	Positive
0.001	0.053	68	GO:0051301 cell division	Positive
0.001	0.064	55	GO:0048285 organelle fission	Positive
0.001	0.067	31	GO:0051320 S phase	Positive
0.001	0.073	59	GO:0007050 cell cycle arrest	Positive
0.001	0.073	112	GO:0051726 regulation of cell cycle	Positive
0.002	0.092	66	GO:0010564 regulation of cell cycle process	Positive
0.002	0.051	19	KEGG hsa03030 DNA replication	Positive
0.002	0.051	23	KEGG hsa04110 Cell cycle	Positive

Table 30: Table of results from gene set enrichment analysis of HPV negative samples when compared with tissue adjacent controls. KEGG and GO terms representing differential expression across a pathway or biological process.

p-Value	Adjusted p-Value	Set Size	Process or Pathway	HPV Status
0.002	0.075	19	KEGG hsa04512 ECM-receptor interaction	Negative

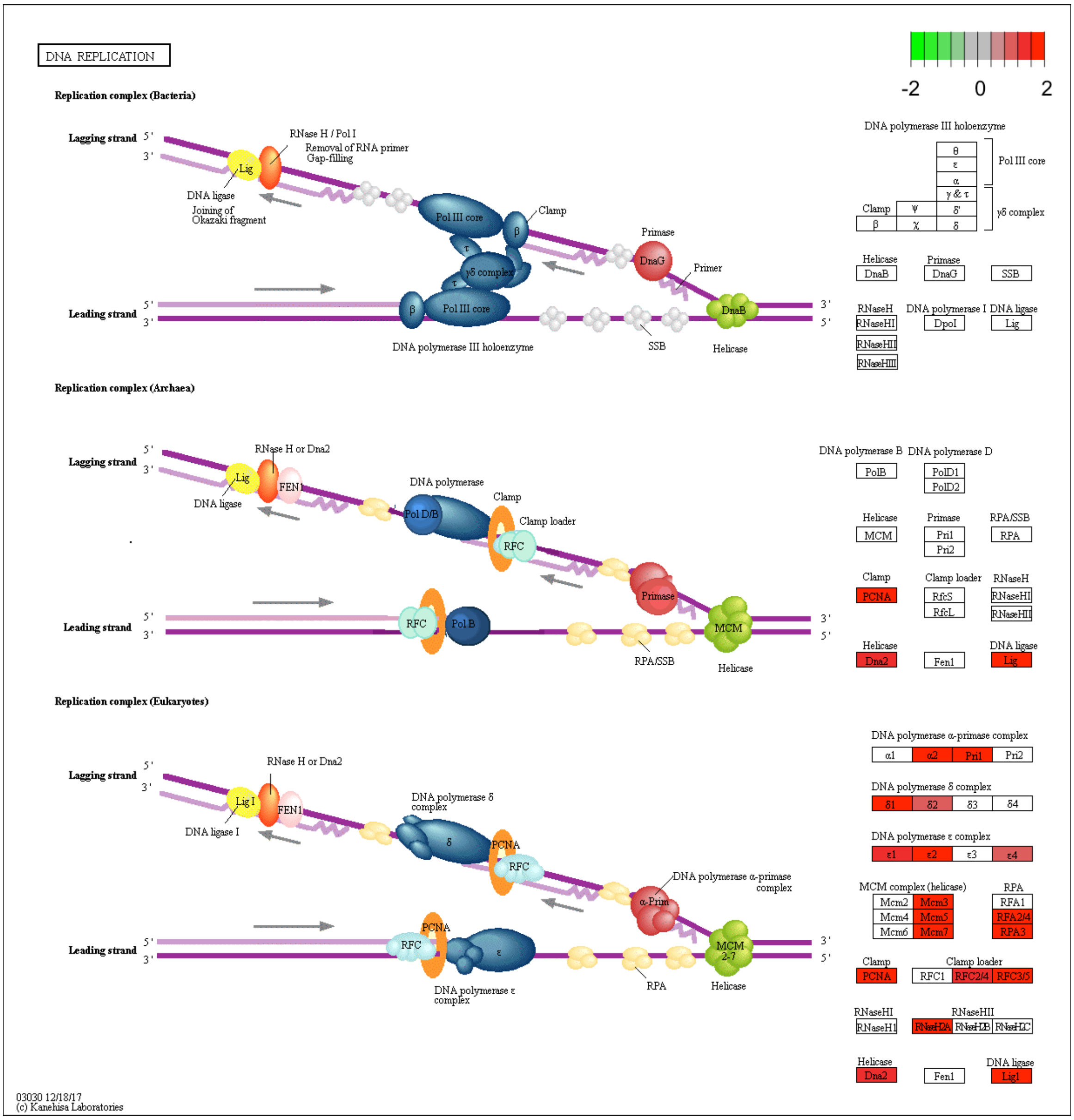


Figure 98: KEGG pathway associated with differential expression in HPV positive samples. Genes in red boxes are over-expressed, whilst those in green have loss of expression.

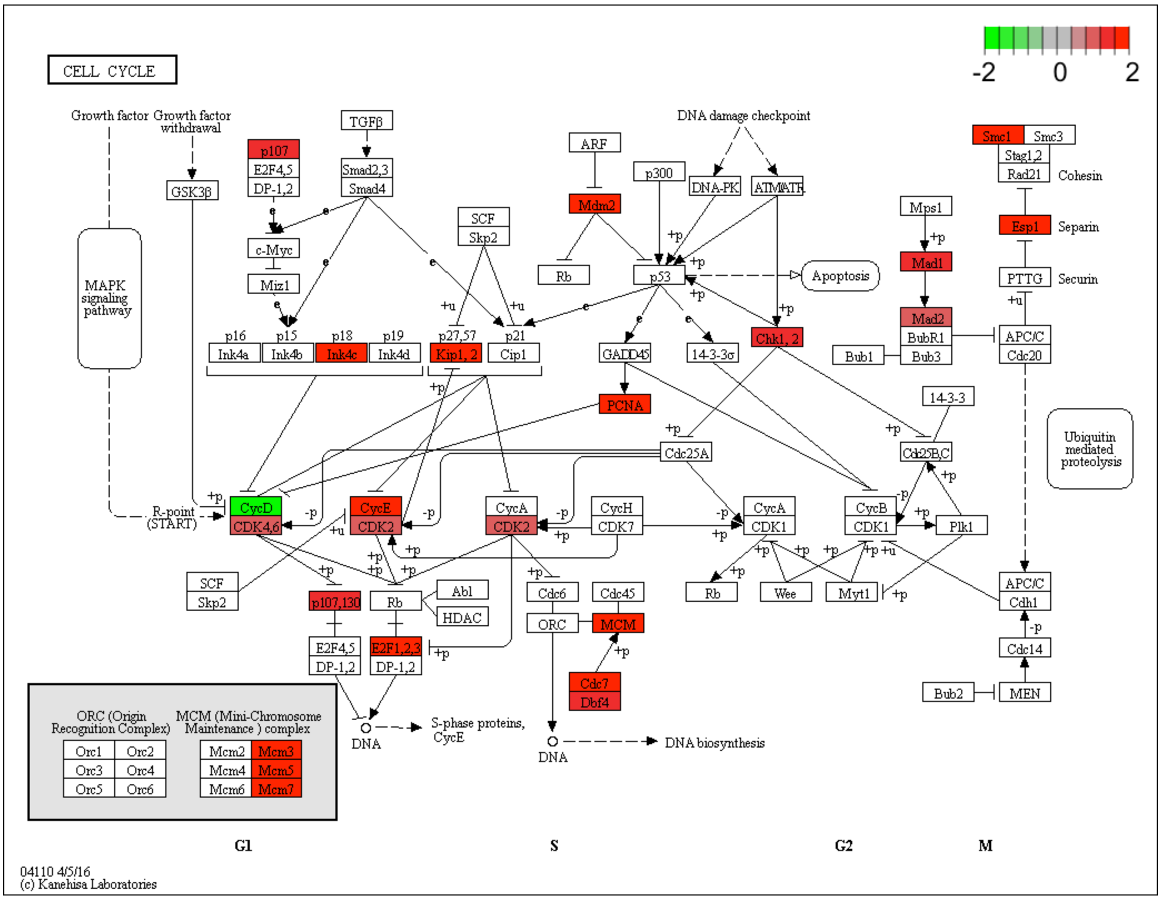


Figure 99: KEGG pathway associated with differential expression in HPV positive samples. Genes in red boxes are over-expressed, whilst those in green have loss of expression.

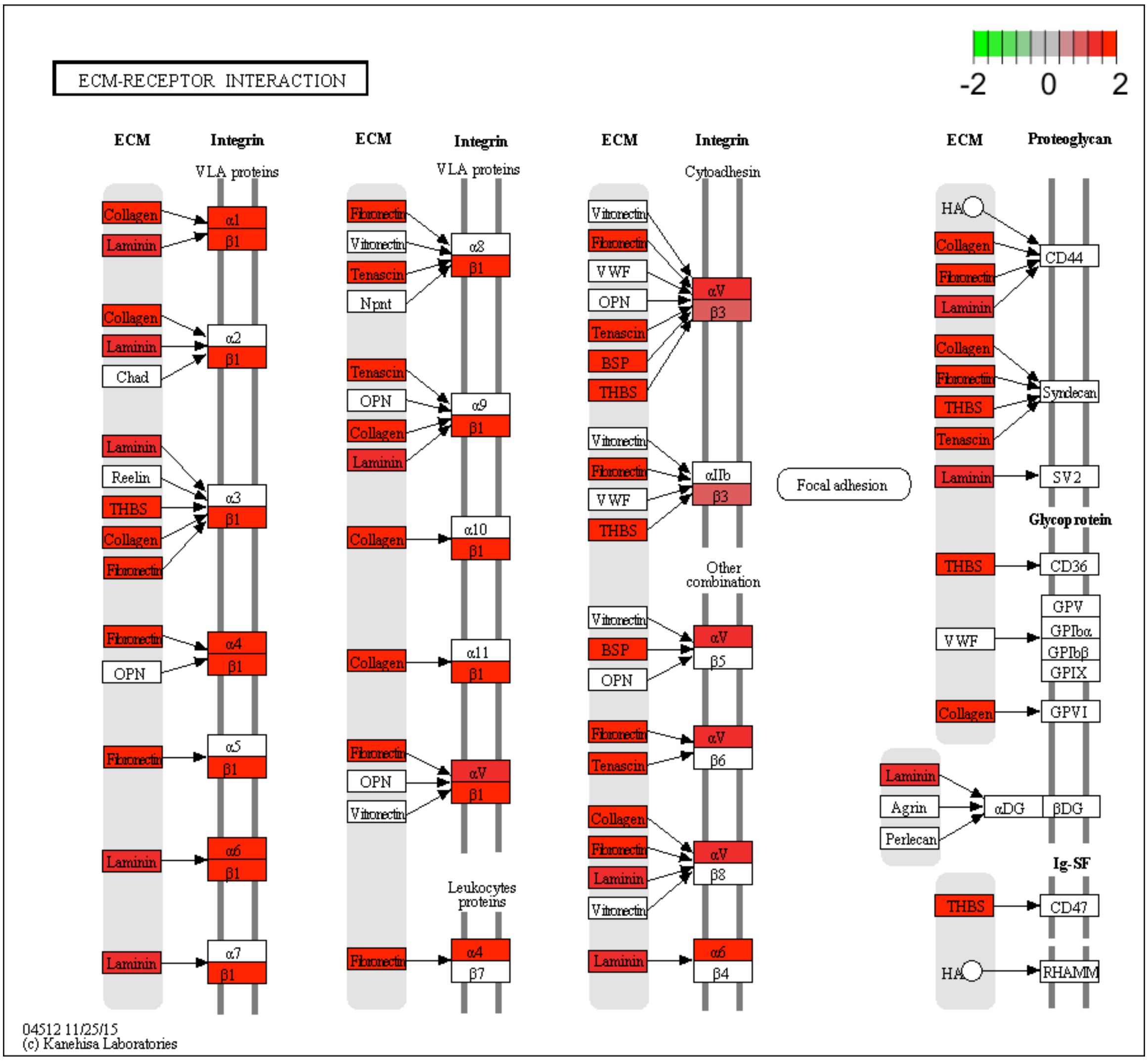


Figure 100: KEGG pathway associated with differential expression in HPV negative samples. Genes in red boxes are over-expressed, whilst those in green have loss of expression.

5.4.3 Lymph node metastases versus normal

Expression in lymph node metastases was assessed to determine whether there was a unique signature of differentially expressed genes, which may be driving the metastatic phenotype. Unfortunately, no matched histologically normal lymph node tissue was available from the patients in the PenHet cohort. Adjacent histologically normal skin was used as a sub-optimal control, as with the primary tumour samples. Therefore, care must be taken when analysing results of the lymph node metastases. Some of the difference in expression may be attributable to the differences in the genes expressed within lymphoid tissue. In addition, as demonstrated above, in Figure 71, there is a significant enrichment for immune expression signatures in the lymph node metastases when compared with the primary tumour samples. Furthermore, with only eight lymph node metastases in the PenHet cohort, there is reduced power to detect tumour specific expressed genes.

Nevertheless, when a differential expression analysis was undertaken, similar findings were found to the primary versus tissue adjacent normal samples. To reduce the bias attributable to the increased immune cell mix in the lymph node metastases, all genes found in the xCell immune signature were removed from the analysis. As demonstrated in the Venn diagram in Figure 101, 64% (3,699) of differentially expressed genes found were also found to be differentially expressed between primary tumour tissue and adjacent tissue. Therefore, only 36% of genes (2,078) were found to be uniquely expressed in lymph node metastases compared with normal lymphoid tissue. A bar chart displaying a filtered list of genes also present in the COSMIC database of potential gene drivers together with their respective log2 fold changes can be found in Figure 102.

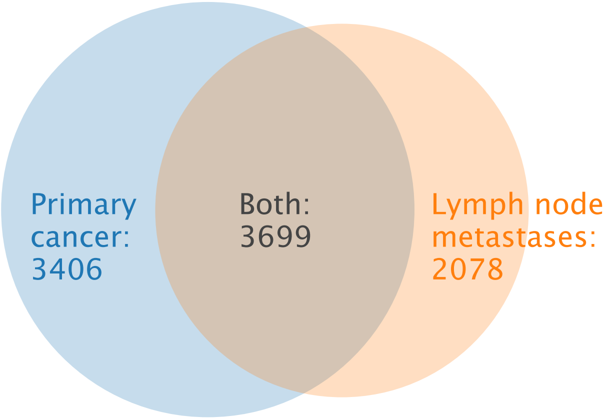


Figure 101: Venn diagram displaying the overlap between differentially expressed genes in lymph node metastatic tissue compared to primary penile cancer samples.

Gene set enrichment analysis was performed on all the differentially expressed genes, using an identical method to that detailed in Section 5.4.1. No statistically significant pathways were found to be enriched when assessing pathways in the KEGG database. However, despite removal of xCell immune signature genes, immune cell mediated processes were enriched in lymph node metastases from the GO database. All enriched GO processes with an adjusted p value of < 0.01 are displayed in Table 31. It is unclear what proportion of these changes relate to infiltrating immune cells or sequencing of normal lymphoid tissue. Excluding these immune processes, there was also a trend for enrichment of cell migration and motility biological processes but not at sufficient level to provide statistical significance. Cell migration enrichment had a p value of

0.0077 but an adjusted p value of 0.148. Cell migration enrichment had a p value of 0.012 but an adjusted p value of 0.186. One would expect these two processes to be enriched in metastatic tissue, but further samples would be required to provide enough power to demonstrate this.

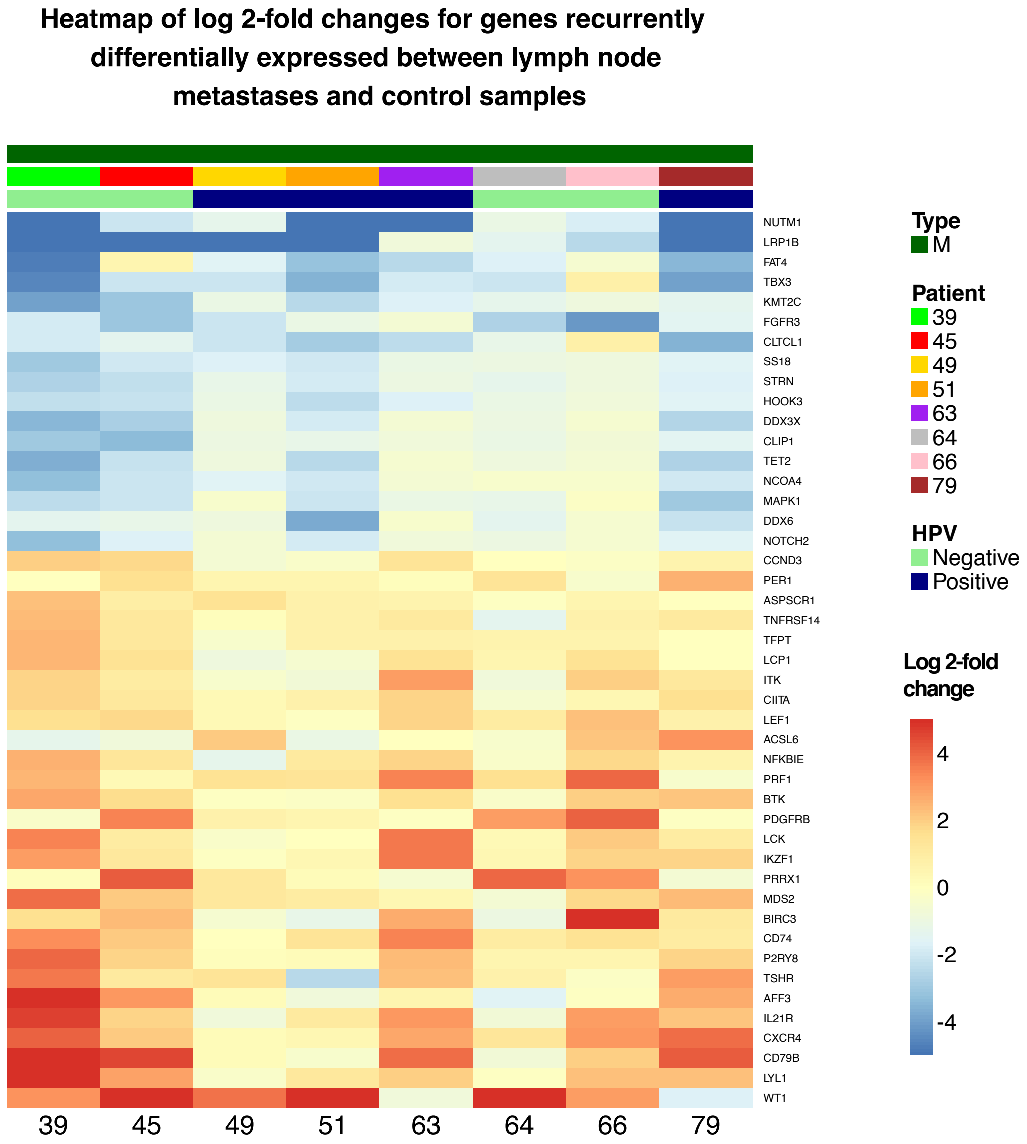
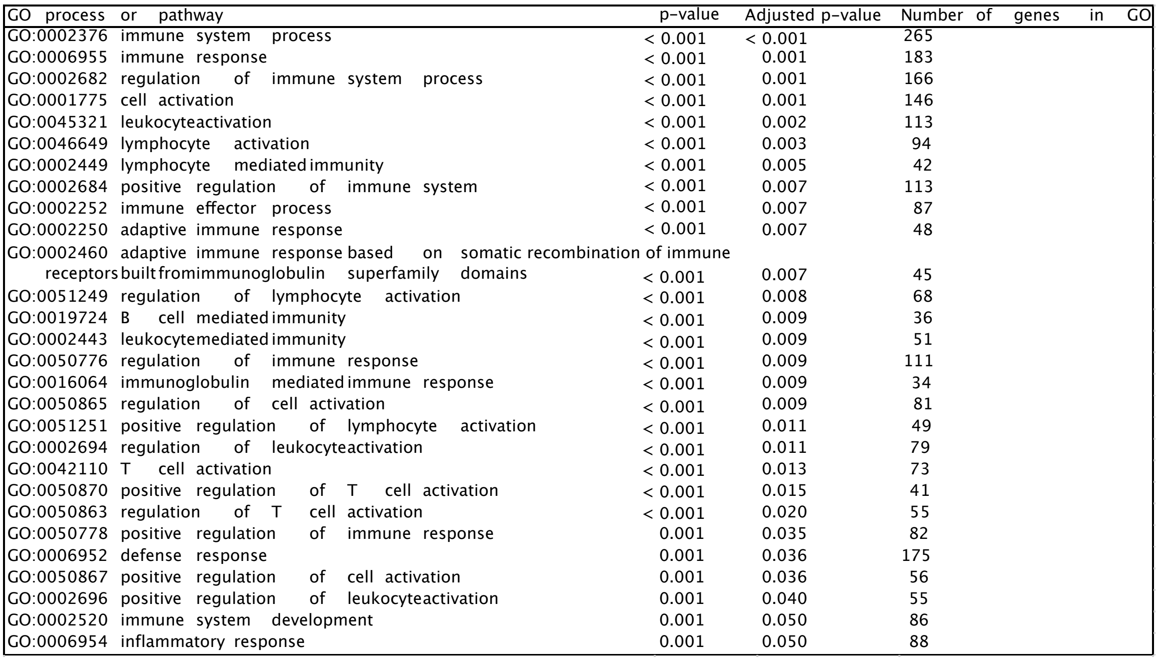


Figure 102: Differentially expressed genes unique to lymph node metastases. Heatmap cell colours depict log2 fold changes each gene ranging from -5 (red) to +5 (blue).

Table 31: Gene set enrichment analysis for GO processes when comparing differentially methylated genes between lymph node metastases and normal control samples.



GO process or pathway	p-value	Adjusted p-value	Number of genes in GO
GO:0002376 immune system process	< 0.001	< 0.001	265
GO:0006955 immune response	< 0.001	0.001	183
GO:0002682 regulation of immune system process	< 0.001	0.001	166
GO:0001775 cell activation	< 0.001	0.001	146
GO:0045321 leukocyteactivation	< 0.001	0.002	113
GO:0046649 lymphocyte activation	< 0.001	0.003	94
GO:0002449 lymphocyte mediated immunity	< 0.001	0.005	42
GO:0002684 positive regulation of immune system	< 0.001	0.007	113
GO:0002252 immune effector process	< 0.001	0.007	87
GO:0002250 adaptive immune response	< 0.001	0.007	48
GO:0002460 adaptive immune response based on somatic recombination of immune receptors builtfromimmunoglobulin superfamily domains	< 0.001	0.007	45
GO:0051249 regulation of lymphocyte activation	< 0.001	0.008	68
GO:0019724 B cell mediated immunity	< 0.001	0.009	36
GO:0002443 leukocytemediated immunity	< 0.001	0.009	51
GO:0050776 regulation of immune response	< 0.001	0.009	111
GO:0016064 immunoglobulin mediated immune response	< 0.001	0.009	34
GO:0050865 regulation of cell activation	< 0.001	0.009	81
GO:0051251 positive regulation of lymphocyte activation	< 0.001	0.011	49
GO:0002694 regulation of leukocyteactivation	< 0.001	0.011	79
GO:0042110 T cell activation	< 0.001	0.013	73
GO:0050870 positive regulation of T cell activation	< 0.001	0.015	41
GO:0050863 regulation of T cell activation	< 0.001	0.020	55
GO:0050778 positive regulation of immune response	0.001	0.035	82
GO:0006952 defense response	0.001	0.036	175
GO:0050867 positive regulation of cell activation	0.001	0.036	56
GO:0002696 positive regulation of leukocyteactivation	0.001	0.040	55
GO:0002520 immune system development	0.001	0.050	86
GO:0006954 inflammatory response	0.001	0.050	88

5.5 External validation

The results displayed above are based on RNA sequencing from only eight patients with squamous cell penile carcinoma. Whilst undertaking the analysis described above, it is important to consider that the results may not be representative of larger independent cohorts of invasive penile cancer. In addition to using stringent adjusted p values for downstream processing, two additional methods were therefore utilised in this chapter to increase the confidence of detecting truly positive differentially expressed genes. The first method is deployed in the integration sections below where the expression data is integrated into the whole exome sequencing and methylation data to find associative drivers acting across these datasets (Sections 5.8-5.10). The second method is to evaluate whether the selected genes can be corroborated in an independent cohort of penile squamous cell carcinoma patients.

An external cohort of penile cancer patients was sought to increase the statistical significance of the differentially expressed genes discovered in the PenHet cohort. Only a single study has previously performed gene expression analysis of penile cancer samples[89]. Marchi et al used Agilent 4x44k Human genome gene expression microarrays in penile squamous cell carcinomas to determine a set of differentially expressed genes in a cohort of 39 patients[89]. The external array data (GEO: GSE57955) was processed as per the original study (Methods, Chapter 2). Expression levels of 14,393 identical genes were able to be detected by both the gene expression arrays of the external cohort and the RNA-seq method utilised in the PenHet cohort. Differentially expressed genes were evaluated as described in the Methods in Chapter 2. The baseline patient characteristics of the Marchi cohort are also described in the Methods. However, any attempt to use the Marchi cohort to corroborate the findings in this chapter has to be made with caution as the Marchi patients have very different baseline characteristics. They are more heterogeneous: only 27% have lymph node metastasis and 20% have low grade 1 disease, compared with 100% of patients in the PenHet cohort having aggressive grade 2-3 disease and lymph node metastases. Nevertheless, genes which are differentially expressed in both cohorts potentially represent important genes, both biologically and in terms of biomarkers, for all patients with penile squamous cell carcinoma.

To evaluate what proportion of genes in the PenHet cohort were also differentially expressed in the Marchi cohort, the specific genes assessed on the Marchi panel had to be ascertained, as well as their levels of differential expression. 1,385 genes were found to be differentially expressed in the PenHet cohort, and present on the Marchi gene panel. Of these 1,385, 630 (45%) were also found to be significantly differentially expressed in the external Marchi cohort. Considering that the patient cohorts are different in terms of grade and stage of disease, this was a surprisingly large number of genes, which could form the basis for further work on these candidates for genes involved in both low- and high-grade disease.

This external Marchi cohort is used below in Sections 5.6.1 and 5.7 to externally validate differentially expressed genes in both the trunk and branches of phylogenetic trees created for each patient.

5.6 Intra-tumour heterogeneity (ITH)

Sampling different regions from the same tumour can reveal how heterogeneous or homogeneous a tumour is. By determining what proportion of aberrations are shared throughout the tumour, an evolutionary model of the tumour can be created that determines which aberrations occurred early – and are therefore shared throughout the tumour – and which aberrations occurred late – therefore only shared by a proportion of the tumour. This is important clinically, as discussed in the Introduction (Chapter 1), as theoretically if targeted therapies are utilised, the highest efficacy would be achieved from targeting molecular aberrations that are present throughout all cells of the tumour.

In the previous two chapters assessing DNA mutations, copy number aberration and methylation aberrations, extensive intra-tumour changes were found. The same phenomenon of heterogeneity was found when assessing gene expression. When clustering all samples, all tumour samples tended to cluster by patient first, as shown in Figure 74 and Figure 75. The tissue adjacent normal samples clustered together, indicating that they are more similar to each other than they are to their matched cancer samples. The exceptions to this included the lymph node metastases from patients 39, 45, 49, 51 and 63. The lymph node metastases from these five patients clustered together into two groups by HPV status. One possible theory that was explored was whether the clustering by patient or by sample type was being driven by immune cell content. However, Figure 71 shows that the lymph node samples from patients 39 and 79 had the highest xCell immune contamination scores and yet the lymph node metastasis from patient 39 clustered with the other lymph node metastases, while the lymph node metastasis from patient 79 clustered with the remaining primary tumour samples. Therefore, this theory could not explain the lymph node clustering. The differences between lymph node metastases and the primary tumour will be explored in terms of intra-tumour heterogeneity in the phylogenetic tree, in Section 5.6.1.

This section of this chapter will assess the extent of RNA expression intra-tumour heterogeneity and model phylogenetic relationships between regions of each patient's cancer. Following this assessment within RNA expression, a comparison will be made with integrating the results from Chapters 3 and 4 on SNV and methylation aberrations.

The mRNA expression profiles of each sample within the primary tumour as well as the lymph node metastasis were compared to assess for intra-tumour heterogeneity. These expression changes were then grouped into categories depending on what proportion of samples within a patient contained the specific change. Using the same methods as for the regional sample analysis for mutations and methylation aberrations in Chapters 3 and 4, each gene was classified as truncal if present in all cancer regions, shared if present in more than one region, and private if only present in one region.

Regional phylogenetic trees can be produced by assessing the genes that fall into each of these three categories of events. Differentially expressed genes that are conserved across all regions make up the trunk of each tree, differentially expressed genes that are unique to each region make up the terminal branches, and all other differentially expressed genes make up the shared branches. Unlike in Chapter 3 on DNA mutations, the exact clonal structure of the tumour samples has not been elucidated. This challenge is currently beyond the scope of this thesis, as it would involve accurately deconvoluting the immune cell and stromal fraction together with compensating for the inherent biases in transcript counting to produce a cancer cell fraction of expression events.

As a result, early clonal changes can only be approximated. One method employed by other researchers[193] and utilised here is to find differentially expressed genes that are recurrent throughout all the primary tumour samples. Strictly speaking these are truncal shared expression events that are likely to be early events and possibly clonal in origin. However, as demonstrated in Chapter 3, it is possible for recurrent genomic events to appear as clonal when they are not. Therefore, although the terms 'truncal', 'early' and 'clonal' are utilised interchangeably in many publications, caution should be exercised when interpreting these results. Fortunately for the purposes of this analysis, the exact subclonal structure of the primary cancer is not required to be calculated, as most of the insights can be gleaned from splitting the differential expression events into truncal versus non-truncal.

Several methods are used throughout this thesis to produce this topological configuration and assess the relatedness of each sample to another. The two most commonly used methods involve 'binarising' the data, and either assessing the Euclidian distance between each sample[223,224] or using a maximum parsimony ratchet method, as detailed in the Methods (Chapter 2). Extensive heterogeneity was found in all patient samples in the PenHet cohort, as demonstrated in Figure 75 and the phylogenetic trees plotted in the next section, 5.6.1. Differential expression tended to cluster by patient, but remarkable intra-tumour expression heterogeneity was still observed.

Due to the inherent noise already discussed in this section, intra-tumour expression heterogeneity was conservatively called. Expression information from all regions within each tumour was utilised to set a dynamic level for classifying a change as truncal/early. This was accomplished by reducing the minimum log2 fold change required from 1 to 0.58 for calling a gene differentially expressed if the log2 fold change was above the threshold of 1 for the remaining three regions sampled.

5.6.1 Phylogenetic trees

Regional phylogenetic trees, of relative differences in gene expression, were constructed based on the Euclidean distances between samples (see Methods, Chapter 2), as shown in Figure 103 to Figure 110.

Genes that are either recurrently shared/truncal or recurrently on the branches of these regional phylogenetic trees, are discussed in further detail below in Sections 5.6.1.1 and 5.6.1.2. In general, the lymph node metastases (depicted as a region with a suffix 05 on the phylogenetic trees), tend to branch off earlier than the other primary tumour samples. This pattern is irrespective of HPV status. After this branch event there appears to be continued differential expression both shared amongst the primary tumour samples and unique to each tumour sample. In addition, the lymph node metastasis also appears to accumulate further changes in gene expression.

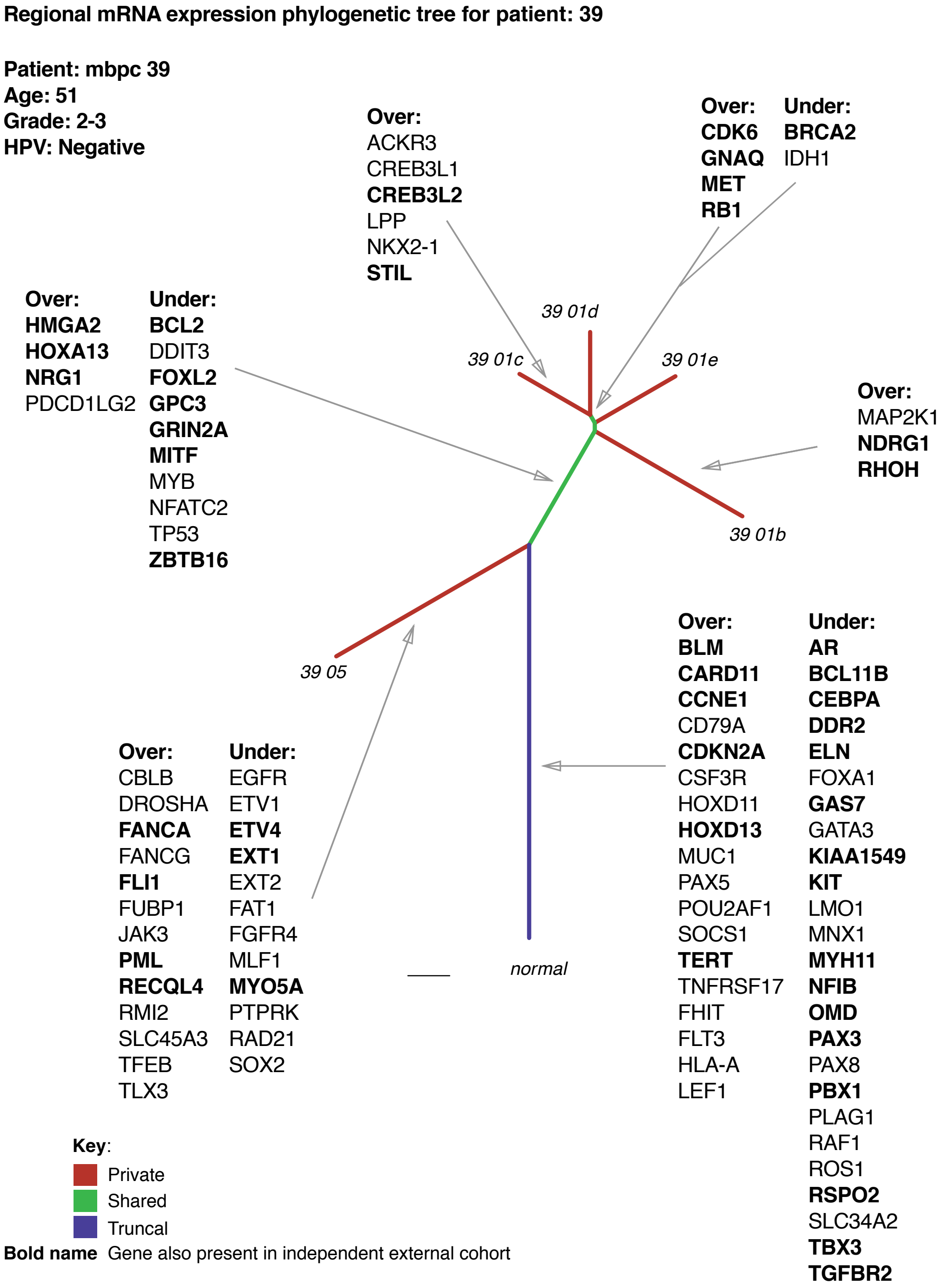


Figure 103: Regional mRNA differential expression phylogenetic tree based constructed using the 'ape' package based on Euclidean distances between tissue samples for patient 39. Genes in bold are also differentially expressed in the independent external cohort (Section 5.5). Genes filtered as potential 'drivers' based on the presence in previous studies of other cancers in the COSMIC database. Private changes are genes that are differentially expressed only in one region of the tumour. Shared changes are present in multiple but not all regions, whilst truncal changes are shared throughout the primary and metastasis. Regions with the suffix 01 are primary tumour regions, whereas regions with the suffix 05 are from lymph node metastatic tissue.

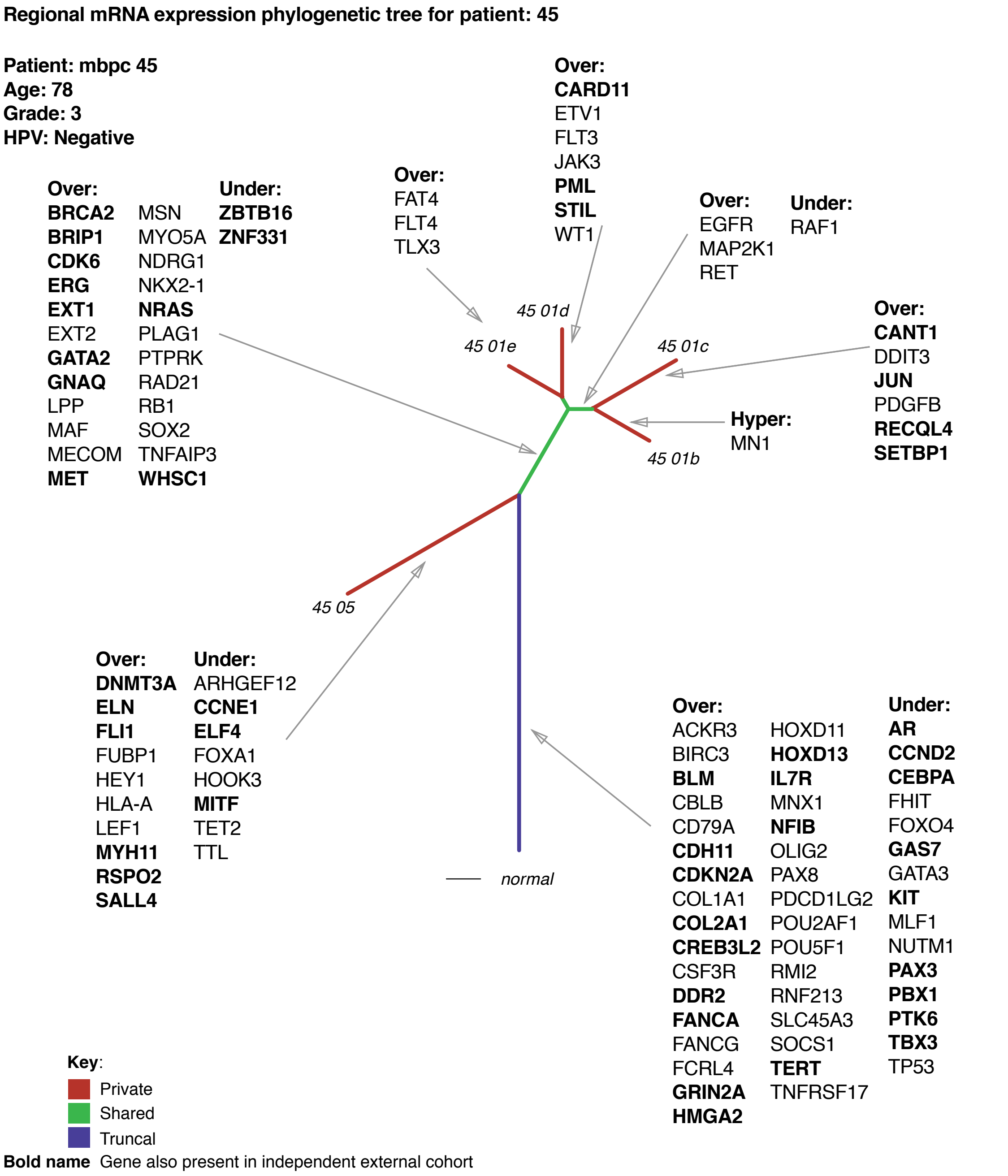


Figure 104: Regional mRNA differential expression phylogenetic tree based constructed using the 'ape' package based on Euclidean distances between tissue samples for patient 45 . Genes in bold are also differentially expressed in the independent external cohort (Section 5.5). Genes filtered as potential 'drivers' based on the presence in previous studies of other cancers in the COSMIC database. Private changes are genes that are differentially expressed only in one region of the tumour. Shared changes are present in multiple but not all regions, whilst truncal changes are shared throughout the primary and metastasis. Regions with the suffix 01 are primary tumour regions, whereas regions with the suffix 05 are from lymph node metastatic tissue.

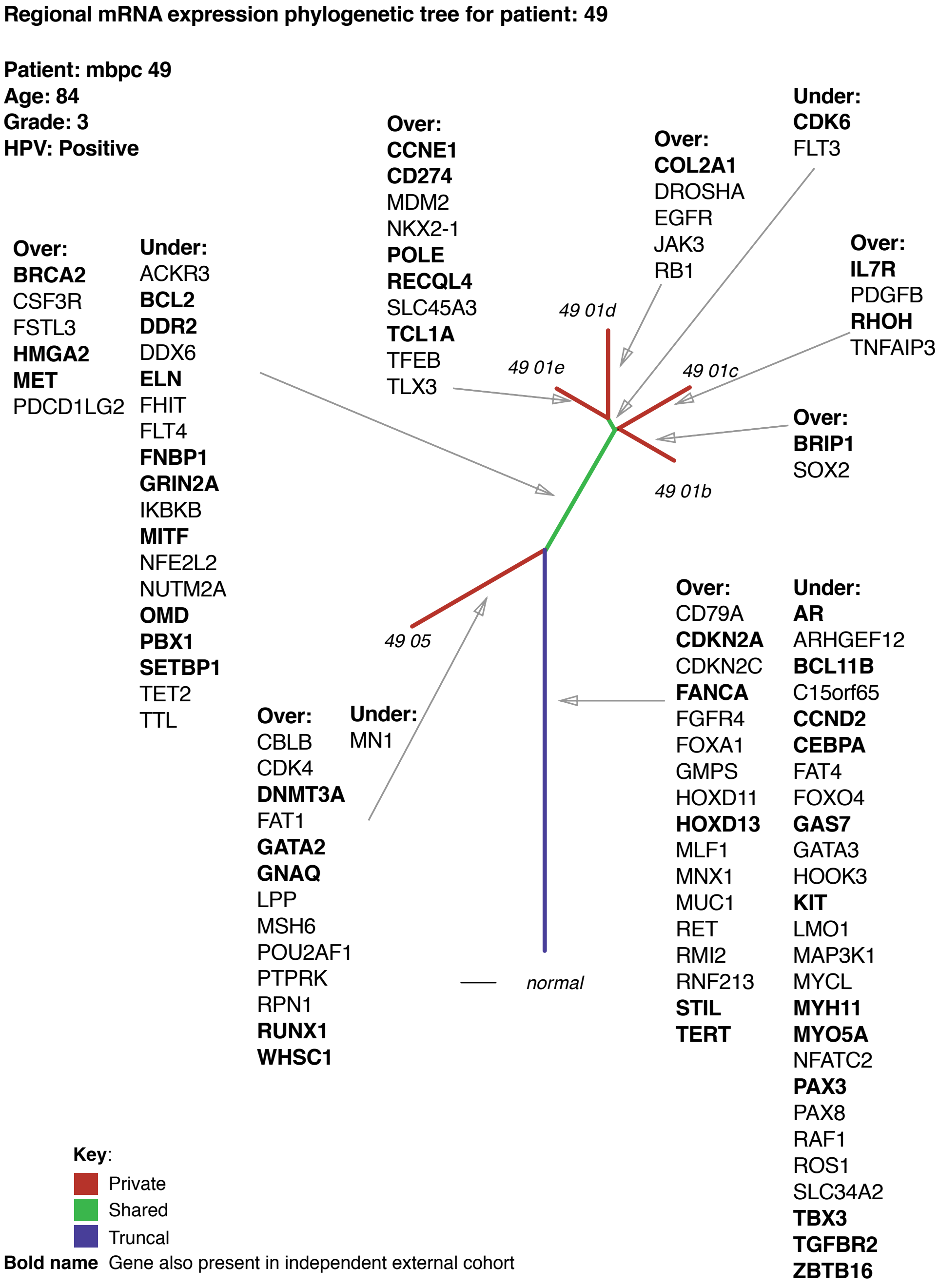


Figure 105: Regional mRNA differential expression phylogenetic tree based constructed using the 'ape' package based on Euclidean distances between tissue samples for patient 49. Genes in bold are also differentially expressed in the independent external cohort (Section 5.5). Genes filtered as potential 'drivers' based on the presence in previous studies of other cancers in the COSMIC database. Private changes are genes that are differentially expressed only in one region of the tumour. Shared changes are present in multiple but not all regions, whilst truncal changes are shared throughout the primary and metastasis. Regions with the suffix 01 are primary tumour regions, whereas regions with the suffix 05 are from lymph node metastatic tissue.

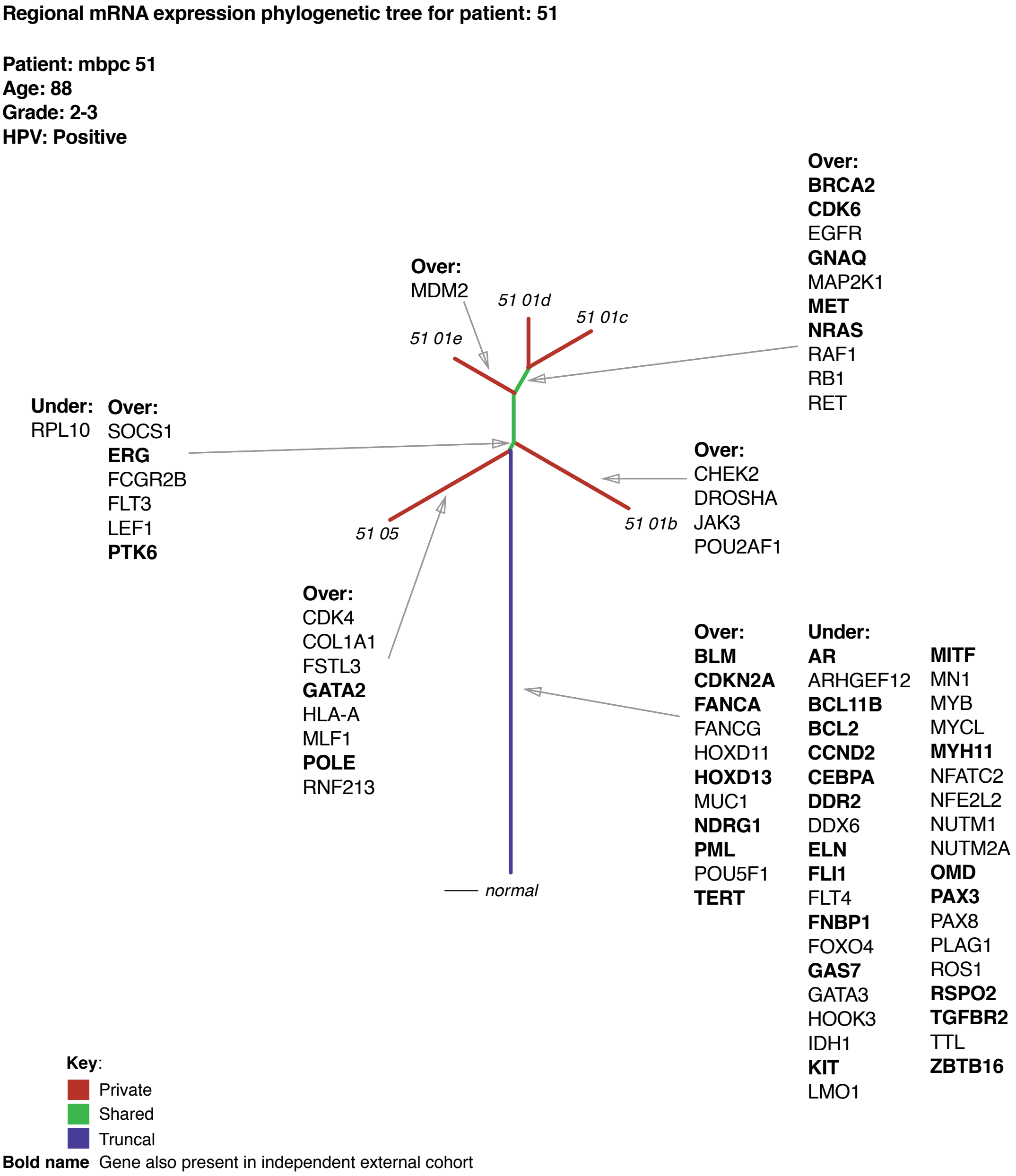


Figure 106: Regional mRNA differential expression phylogenetic tree based constructed using the 'ape' package based on Euclidean distances between tissue samples for patient 51. Genes in bold are also differentially expressed in the independent external cohort (Section 5.5). Genes filtered as potential 'drivers' based on the presence in previous studies of other cancers in the COSMIC database. Private changes are genes that are differentially expressed only in one region of the tumour. Shared changes are present in multiple but not all regions, whilst truncal changes are shared throughout the primary and metastasis. Regions with the suffix 01 are primary tumour regions, whereas regions with the suffix 05 are from lymph node metastatic tissue.

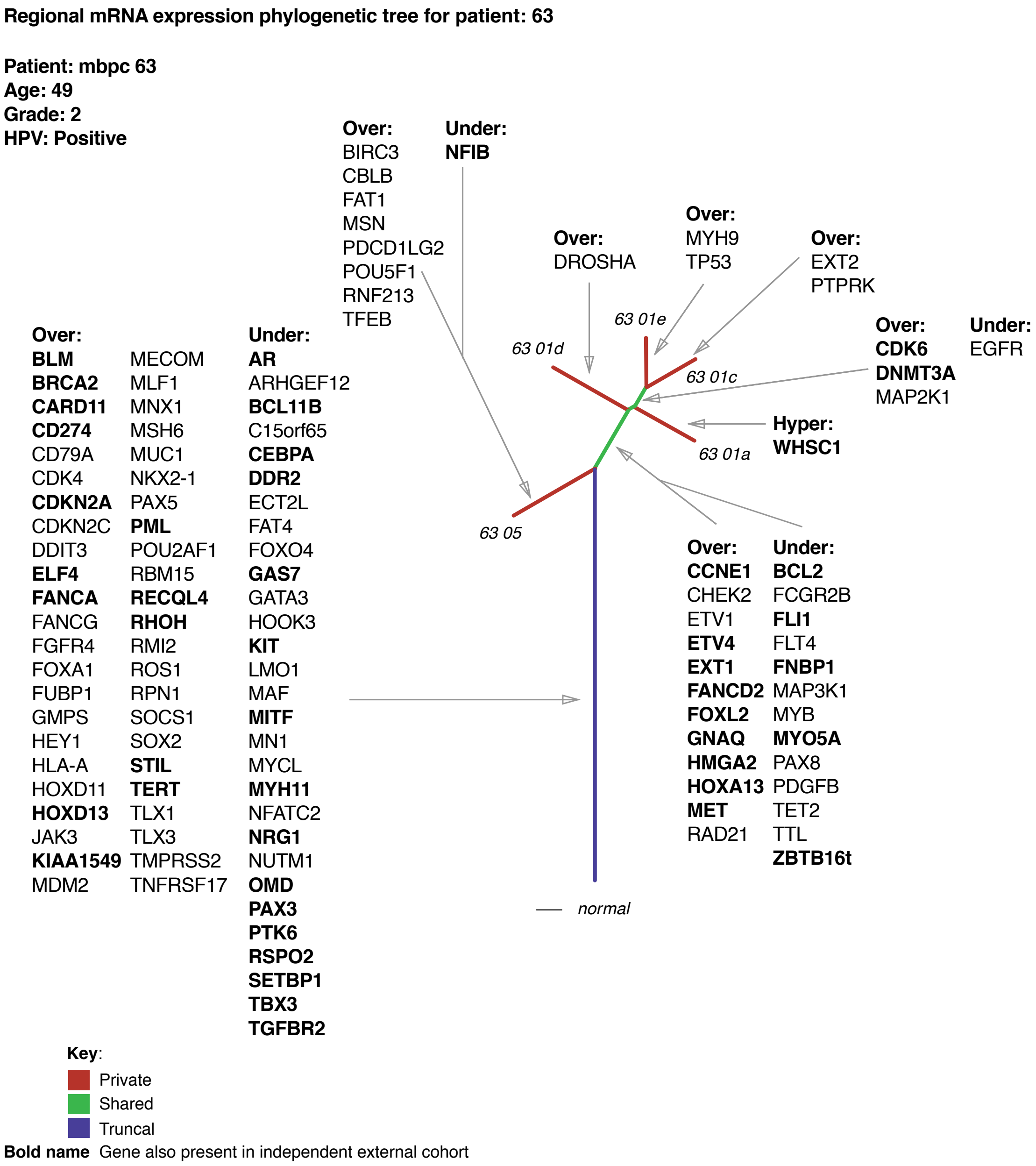


Figure 107: Regional mRNA differential expression phylogenetic tree based constructed using the 'ape' package based on Euclidean distances between tissue samples for patient 63. Genes in bold are also differentially expressed in the independent external cohort (Section 5.5). Genes filtered as potential 'drivers' based on the presence in previous studies of other cancers in the COSMIC database. Private changes are genes that are differentially expressed only in one region of the tumour. Shared changes are present in multiple but not all regions, whilst truncal changes are shared throughout the primary and metastasis. Regions with the suffix 01 are primary tumour regions, whereas regions with the suffix 05 are from lymph node metastatic tissue.

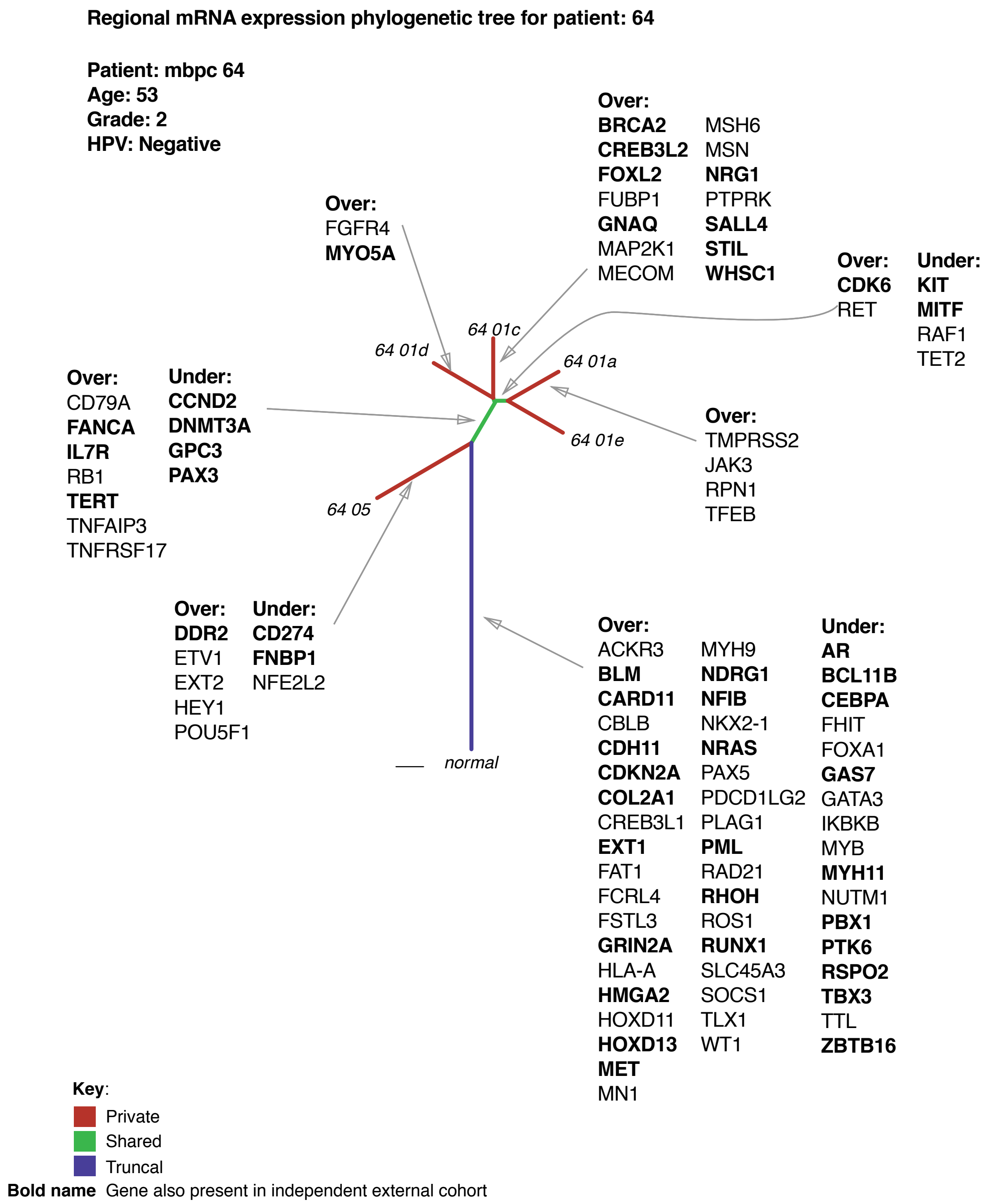


Figure 108: Regional mRNA differential expression phylogenetic tree based constructed using the 'ape' package based on Euclidean distances between tissue samples for patient 64. Genes in bold are also differentially expressed in the independent external cohort (Section 5.5). Genes filtered as potential 'drivers' based on the presence in previous studies of other cancers in the COSMIC database. Private changes are genes that are differentially expressed only in one region of the tumour. Shared changes are present in multiple but not all regions, whilst truncal changes are shared throughout the primary and metastasis. Regions with the suffix 01 are primary tumour regions, whereas regions with the suffix 05 are from lymph node metastatic tissue.

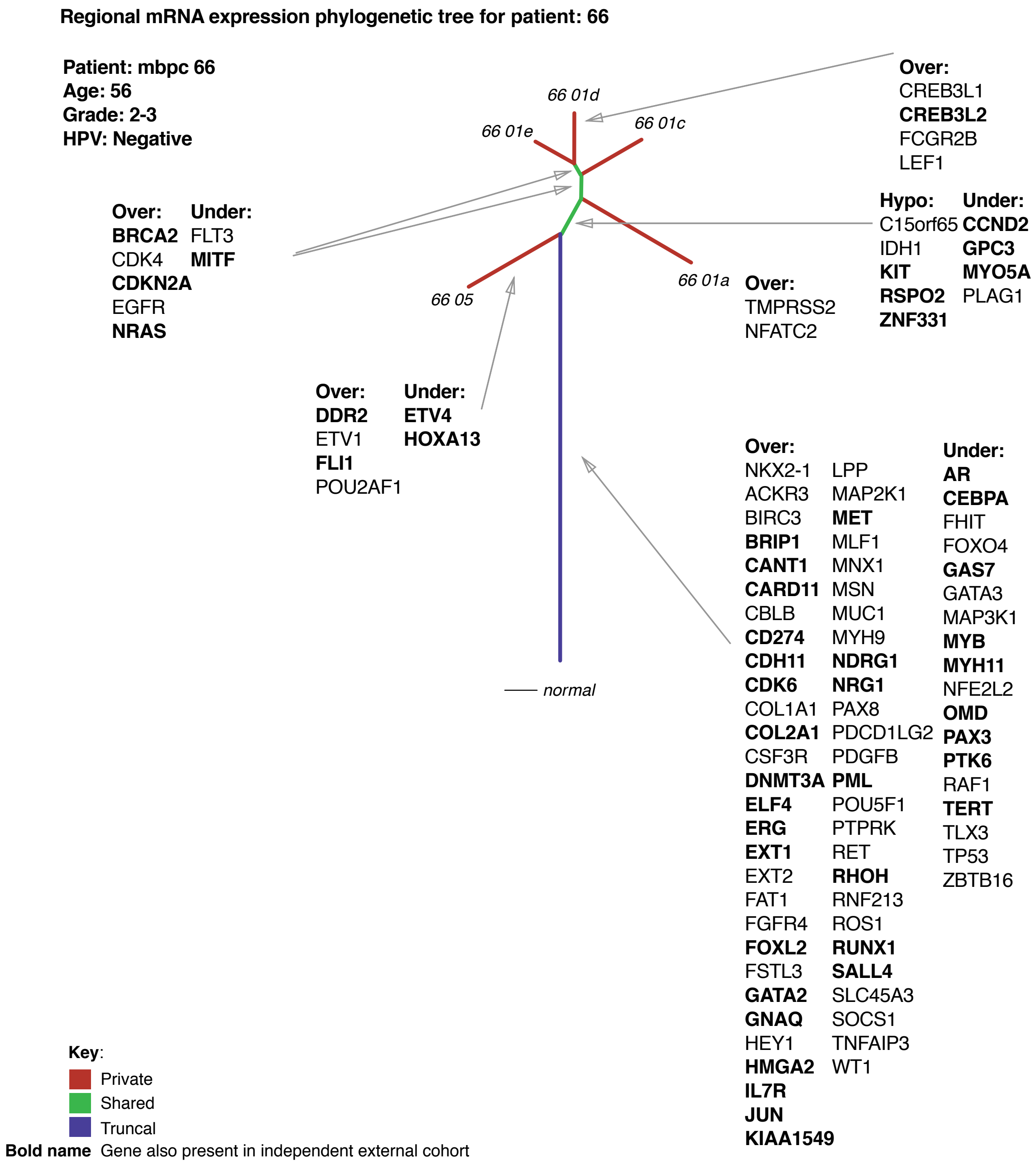


Figure 109: Regional mRNA differential expression phylogenetic tree based constructed using the 'ape' package based on Euclidean distances between tissue samples for patient 66. Genes in bold are also differentially expressed in the independent external cohort (Section 5.5). Genes filtered as potential 'drivers' based on the presence in previous studies of other cancers in the COSMIC database. Private changes are genes that are differentially expressed only in one region of the tumour. Shared changes are present in multiple but not all regions, whilst truncal changes are shared throughout the primary and metastasis. Regions with the suffix 01 are primary tumour regions, whereas regions with the suffix 05 are from lymph node metastatic tissue.

Regional mRNA expression phylogenetic tree for patient: 79

Patient: mbpc 79
Age: 59
Grade: 3
HPV: Positive

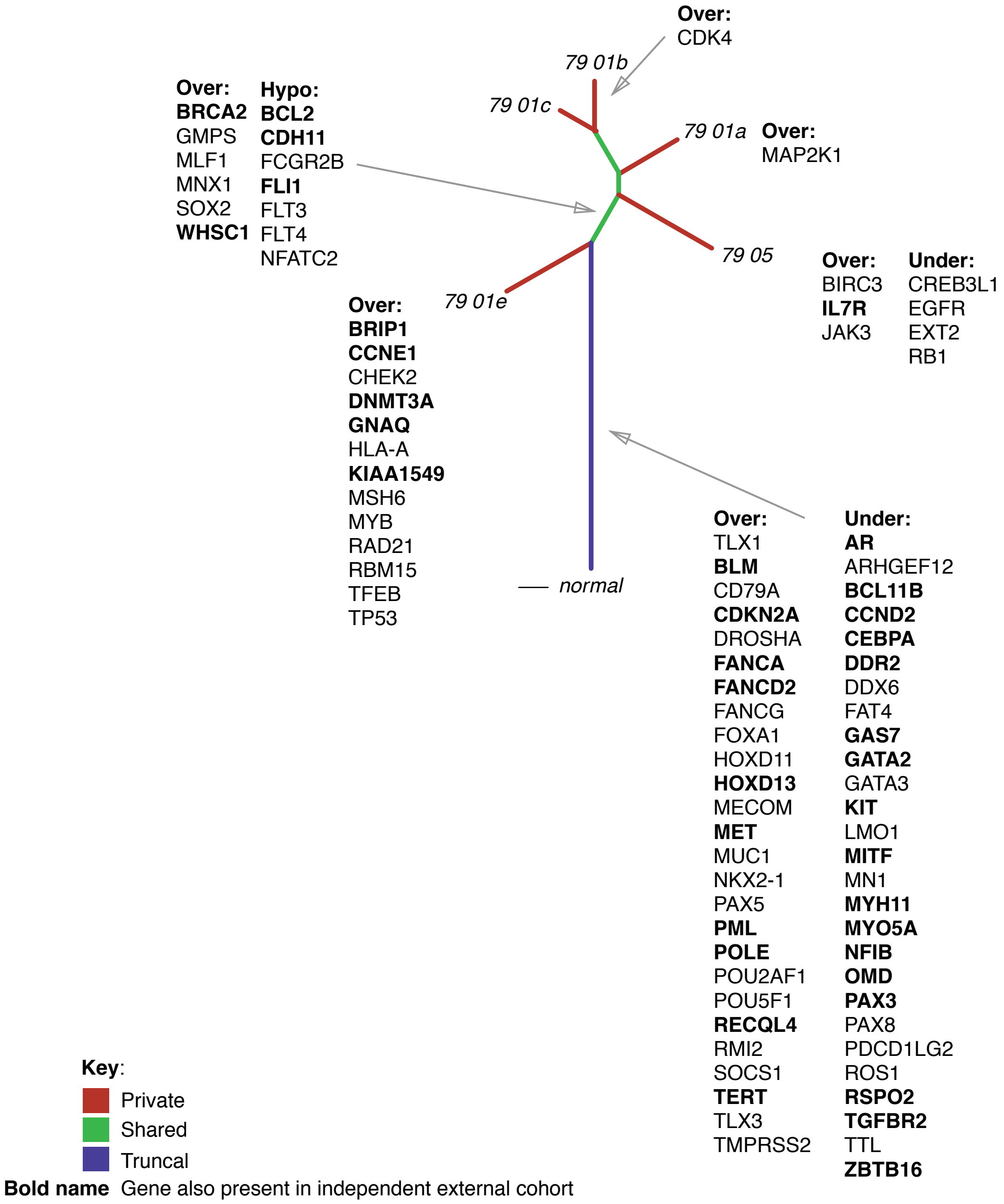


Figure 110: Regional mRNA differential expression phylogenetic tree based constructed using the 'ape' package based on Euclidean distances between tissue samples for patient 79. Genes in bold are also differentially expressed in the independent external cohort (Section 5.5). Genes filtered as potential 'drivers' based on the presence in previous studies of other cancers in the COSMIC database. Private changes are genes that are differentially expressed only in one region of the tumour. Shared changes are present in multiple but not all regions, whilst truncal changes are shared throughout the primary and metastasis. Regions with the suffix 01 are primary tumour regions, whereas regions with the suffix 05 are from lymph node metastatic tissue.

5.6.1.1 Early/truncal differentially expressed genes

Due to the large numbers of differentially expressed genes, only those genes also found in the COSMIC database were annotated on the phylogenetic trees above.

These genes were assessed to see whether they only ever appear in the trunk of the regional phylogenetic trees. These represent likely early changes in expression as they have been conserved across the primary tumour. Seven genes were found in the trunks of the phylogenetic trees. These can be visualised in the differential gene expression heatmap displayed in Figure 111. These were:

- Reduced expression of *DDR2, GAS7*, *GATA3, MYH11* and *ZBTB16*
- Increased expression of *MLF1* and *TERT*

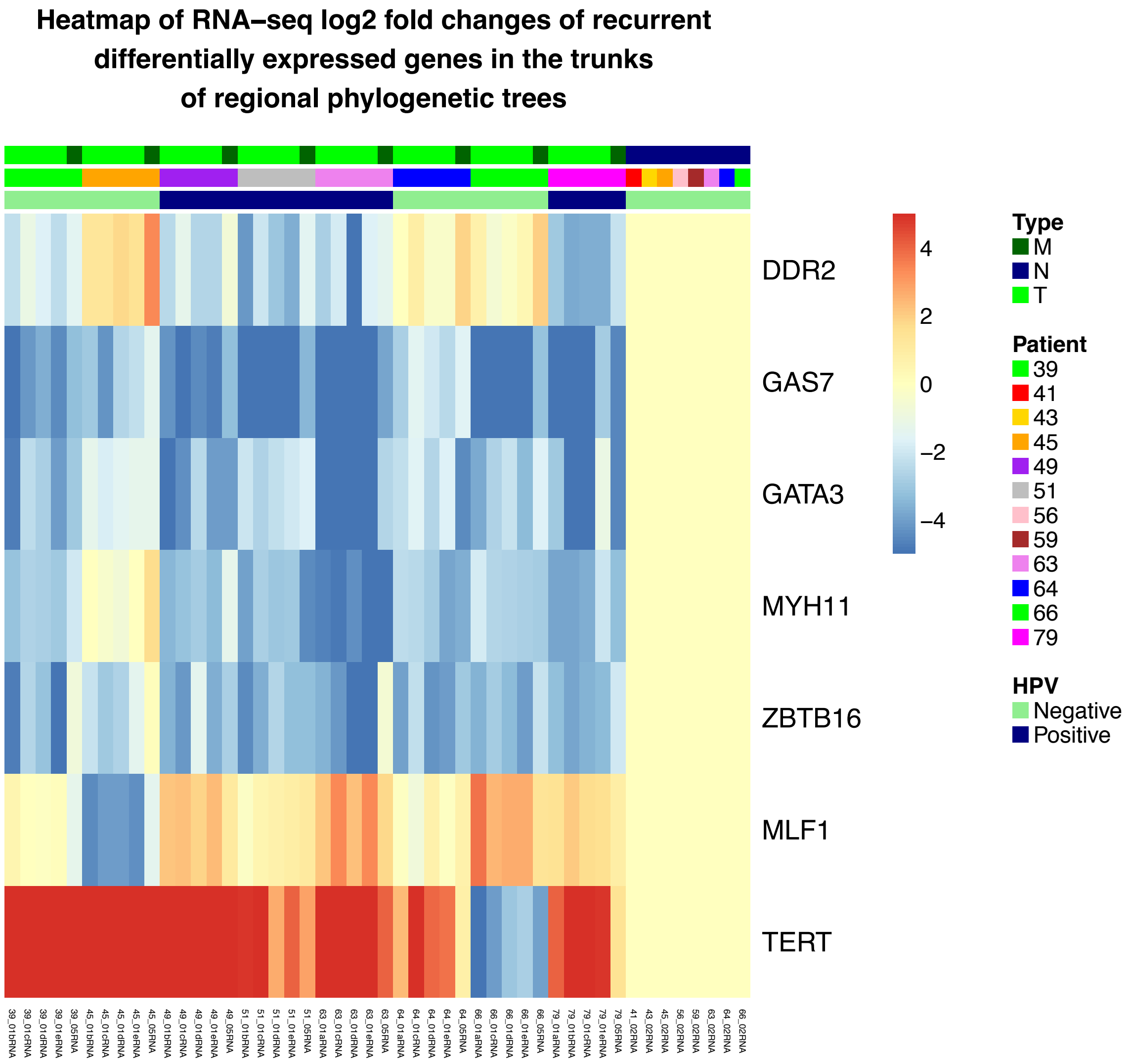


Figure 111: Heatmap of differential gene expression of the seven genes recurrently found in the trunks of the regional phylogenetic trees. Heatmap cell colours depict log2 fold changes each gene ranging from -5 (red) to +5 (blue).

Five out of seven of these were also found differentially expressed in the independent external cohort (Section 5.5) – namely *DDR2, GATA3, MYH11*, *ZBTB16 and TERT.*

DDR2 is a tyrosine kinase that has been previously shown in squamous cell lung carcinoma[225] and ovarian cancer[226] to be associated with invasion, metastasis and poor prognosis. It is also potentially therapeutically targetable[227] by molecules such as dasatinib[228], and so may represent a future therapeutic target in penile squamous cell carcinomas.

GATA3 is a transcription factor and therefore master regulator of many cell processes[229]. Within the context of cancer, it appears to both regulate the cell and interface with the immune system[230]. Loss of *GATA3* in bladder cancer has previously been described as promoting cell invasion and migration[231].

GAS7 was found in Chapter 4 to be hypermethylated at its promoter transcription start site, with a clear significant differentially methylated region present over this CpG island, as shown in Figure 112. In this chapter, a direct relationship is found with reciprocal loss of expression of *GAS7* across all tumour samples, as seen in Figure 113. *GAS7* has previously been described as a biomarker for oral squamous cell carcinoma[232] and demonstrated to be important in preventing metastases[233]. In addition, loss of expression of *GAS7* been demonstrated as one pathway to gefitinib resistance[234]. This is potentially clinically important for penile cancer as gefitinib is one of the tyrosine kinase inhibitors currently being trailed as a targeted therapy for penile cancer.

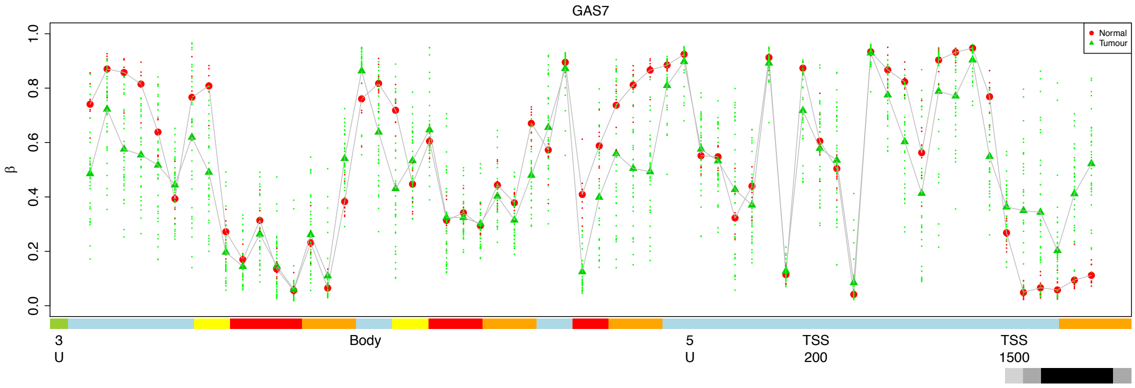


Figure 112: Gene methylation plots demonstrating the beta methylation value for each sample across GAS7. GAS7 is annotated with CpG island location (black horizontal bar at the bottom of the plot) as well as transcription start site (TSS), 5'UTR and gene body. Each green triangle indicates a primary tumour sample beta methylation value at an individual locus. Each red dot represents a normal control sample value. A hypermethylated differentially methylated region can clearly be seen across the CpG island (marked in black). Adjacent to the CpG island are two dark grey sections representing CpG shores and to the outside of those light grey sections representing CpG shelves.

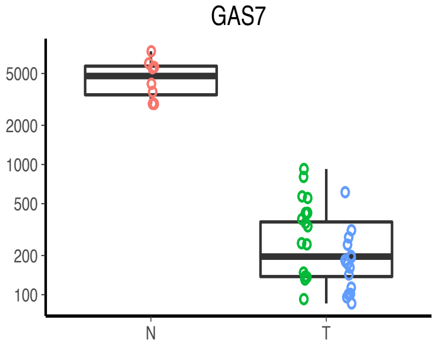


Figure 113: Normalised counts of transcripts collapsed for the gene GAS7 comparing tumour (T) versus normal (N) samples. Combined scatter and box plots. Red scatter points are for the normal control samples, green scatter points for the HPV negative samples and blue scatter points for the HPV positive samples.

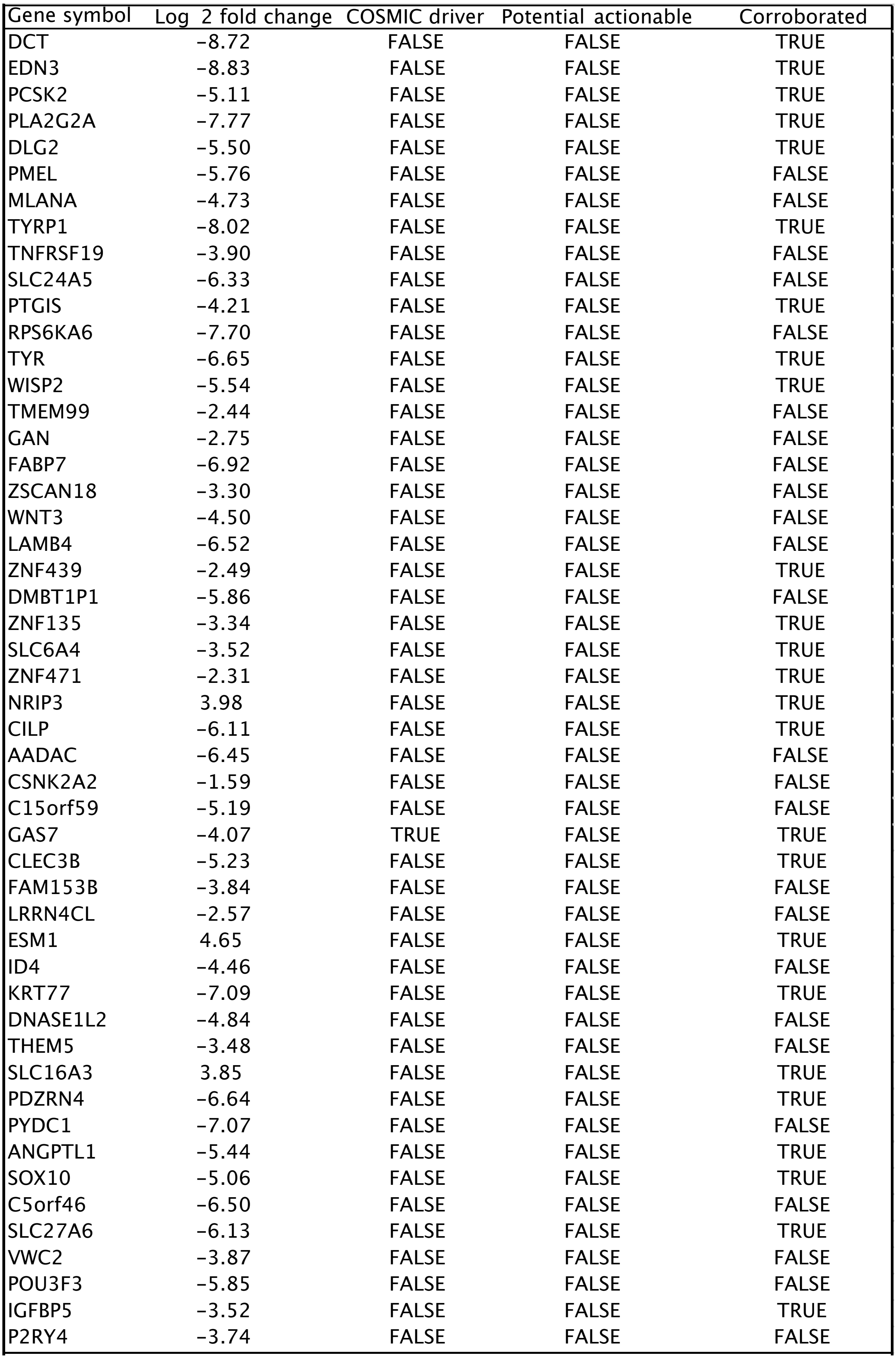
MYH11 has been hypothesised to play an important role in cell migration and adhesion. Loss of expression of *MYH11* was previously found to be associated with a poor prognosis in colorectal adenocarcinomas[235].

ZBTB16, also known as *PLZF*, is a tumour suppressor inhibiting proliferation and metastases[236]. *PLZF* has also been shown to display loss of expression in castration resistance prostate cancer[237].

TERT has been well characterised as a gene potentially subverted in cancer to express telomerase as a method of immortalising cells and promoting cell proliferation[198]. Despite over-expression of *TERT* in almost all samples, no recurrent mutations, either SNVs or CNAs, were discovered on whole exome sequencing. Furthermore, there was no recurrent promoter hypomethylation or gene body hypermethylation, which could have been a further cause of the aberrant over expression. Further work, would need to be undertaken to determine the pathogenicity of *TERT* over expression and if alternative epigenetic causes may be driving its over expression.

A table of all early/truncal expressed genes can be seen in the Appendix. The top 50 with the lowest adjusted p values are displayed in Table 32. Only one out of these 50 genes, *GAS7*, has previously been described as a candidate cancer driver. Of these 50 genes, 42 were assessed on the Marchi external gene expression array. Of these 42, 25 (59%) were validated. This is in comparison to only 45% of all the significant differentially expressed genes discovered in the PenHet cohort. This demonstrates that despite the differences in disease grade and stage at least 50% of the most significant early/truncal changes found in the invasive penile cancer cohort can be found and validated in the more heterogenous external cohort.

Table 32: Table of the top 50 differentially expressed genes, all with adjusted p values < 0.001, always found shared amongst primary tumour samples in the PenHet cohort.



Gene symbol	Log 2 fold change	COSMIC driver	Potential actionable	Corroborated
DCT	-8.72	FALSE	FALSE	TRUE
EDN3	-8.83	FALSE	FALSE	TRUE
PCSK2	-5.11	FALSE	FALSE	TRUE
PLA2G2A	-7.77	FALSE	FALSE	TRUE
DLG2	-5.50	FALSE	FALSE	TRUE
PMEL	-5.76	FALSE	FALSE	FALSE
MLANA	-4.73	FALSE	FALSE	FALSE
TYRP1	-8.02	FALSE	FALSE	TRUE
TNFRSF19	-3.90	FALSE	FALSE	FALSE
SLC24A5	-6.33	FALSE	FALSE	FALSE
PTGIS	-4.21	FALSE	FALSE	TRUE
RPS6KA6	-7.70	FALSE	FALSE	FALSE
TYR	-6.65	FALSE	FALSE	TRUE
WISP2	-5.54	FALSE	FALSE	TRUE
TMEM99	-2.44	FALSE	FALSE	FALSE
GAN	-2.75	FALSE	FALSE	FALSE
FABP7	-6.92	FALSE	FALSE	FALSE
ZSCAN18	-3.30	FALSE	FALSE	FALSE
WNT3	-4.50	FALSE	FALSE	FALSE
LAMB4	-6.52	FALSE	FALSE	FALSE
ZNF439	-2.49	FALSE	FALSE	TRUE
DMBT1P1	-5.86	FALSE	FALSE	FALSE
ZNF135	-3.34	FALSE	FALSE	TRUE
SLC6A4	-3.52	FALSE	FALSE	TRUE
ZNF471	-2.31	FALSE	FALSE	TRUE
NRIP3	3.98	FALSE	FALSE	TRUE
CILP	-6.11	FALSE	FALSE	TRUE
AADAC	-6.45	FALSE	FALSE	FALSE
CSNK2A2	-1.59	FALSE	FALSE	FALSE
C15orf59	-5.19	FALSE	FALSE	FALSE
GAS7	-4.07	TRUE	FALSE	TRUE
CLEC3B	-5.23	FALSE	FALSE	TRUE
FAM153B	-3.84	FALSE	FALSE	FALSE
LRRN4CL	-2.57	FALSE	FALSE	FALSE
ESM1	4.65	FALSE	FALSE	TRUE
ID4	-4.46	FALSE	FALSE	FALSE
KRT77	-7.09	FALSE	FALSE	TRUE
DNASE1L2	-4.84	FALSE	FALSE	FALSE
THEM5	-3.48	FALSE	FALSE	FALSE
SLC16A3	3.85	FALSE	FALSE	TRUE
PDZRN4	-6.64	FALSE	FALSE	TRUE
PYDC1	-7.07	FALSE	FALSE	FALSE
ANGPTL1	-5.44	FALSE	FALSE	TRUE
SOX10	-5.06	FALSE	FALSE	TRUE
C5orf46	-6.50	FALSE	FALSE	FALSE
SLC27A6	-6.13	FALSE	FALSE	TRUE
VWC2	-3.87	FALSE	FALSE	FALSE
POU3F3	-5.85	FALSE	FALSE	FALSE
IGFBP5	-3.52	FALSE	FALSE	TRUE
P2RY4	-3.74	FALSE	FALSE	FALSE

5.6.1.2 Late/branch differentially expressed genes

The same method applied in Section 5.6.1.1 to discover genes that are only found in the trunk of the regional phylogenetic trees was applied to assess which genes are solely found in the branches. These likely represent changes in expression that took place after the last common ancestor, as not all regions of the tumour harbour these changes. Three genes previously described in the COSMIC database were found only in the branches of these trees. These were:

- *EGFR (over-expression)*
- *TCL1A (over-expression)*
- *TFEB (over-expression)*

Over-expression of *EGFR* (Epidermal growth factor receptor) is an important finding in the context of the recently launched clinical trials of EGFR inhibitors in penile cancer. *EGFR* was found over-expressed solely in the branches of two trees. Changes in *EGFR* expression therefore appear – based on this small dataset – to be a relatively later occurrence not shared by multiple tumour regions. This raises the likelihood that treatment with EGFR inhibitors may fail due to the targeted therapy only targeting part of the tumour. This is in contrast to the findings by Jamal et al when assessing the clonal status of *EGFR* in non-small cell lung carcinoma[172].

TCL is an oncogene and coactivator of *AKT*. It is found hyper-expressed in T cell lymphomas and has been proposed as a potential immunogenic antigen in B cell lymphomas[238]. In the PenHet cohort *TCL1A* is always found on the branches of the regional phylogenetic trees representing a likely later aberration during the oncogenesis of penile cancer.

TFEB is a transcription factor and has been previously described as regulating lysosomal function and autophagy[239]. Within cancer it has been described as having a key role in driving metastases in lung[240], pancreatic[241] and breast cancer[242].

5.6.1.3 Comparison of phylogenetic trees from DNA mutation, copy number and methylation studies

The matched regional phylogenetic trees constructed for SNV, copy number and methylation data can be seen alongside the trees from RNA-seq in Figure 114 to Figure 121. There is a clear similarity of the structure of the matched trees. This is not surprising as each region sampled contains a different relative dominance of clones and subclones. These clones are physical entities with unique genetic, epigenetic and expression signatures. However, as the DNA was extracted from almost identical populations of cells as the RNA was extracted from, one would expect the same clones to be present irrespective of whether genetics, epigenetics or levels of expression were used to assess them. Concordance between the regional phylogenetic trees within each patient was assessed using the Mantel test. Figure 114 displays the phylogenetic trees for patient 39 where there is almost complete concordance between the phylogenetic trees representing the SNV, CNV, DMP and expression data. For all other patients there were at least two trees demonstrating concordance ($p < 0.05$). As suggested, one explanation for this concordance is that each region will contain the same ratio of cells of each subclone, irrespective of molecular aberration investigated. However, if one subclone has a disproportionately large number of aberrations of one particular molecular type – for example, methylation as in number of DMPs – then that molecular aberration type will exert a disproportionately large effect on the distance between the region containing most of that subclone and the region containing the least number of cells of that subclone.

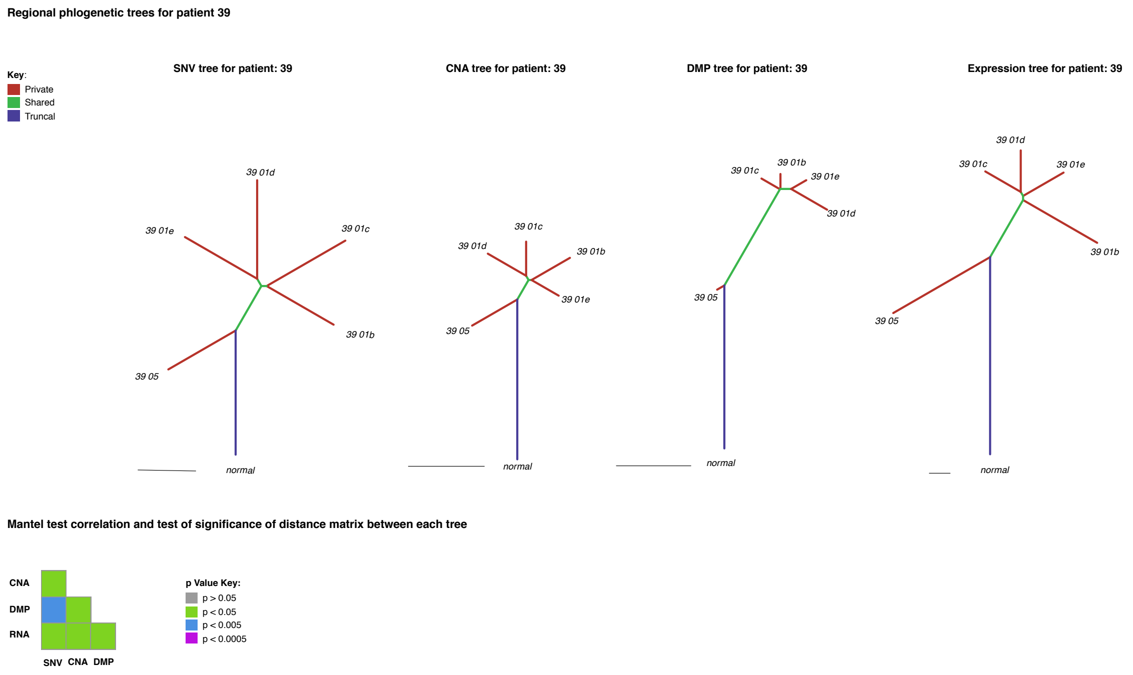


Figure 114: Top: Comparison of regional phylogenetic trees from patient 39 (HPV negative) generated from single nucleotide variants, copy number aberrations, differentially methylated positions and mRNA expression. Regional trees generated using the 'ape' R package. Horizontal line at the bottom of each tree represents a normalised scale of Euclidean distance. Bottom: Mantel test of significance of correlation between distance matrices generated between each molecular aberration type. Regions ending in 05 represent LN metastases, regions ending in a letter represent primary tumour samples.

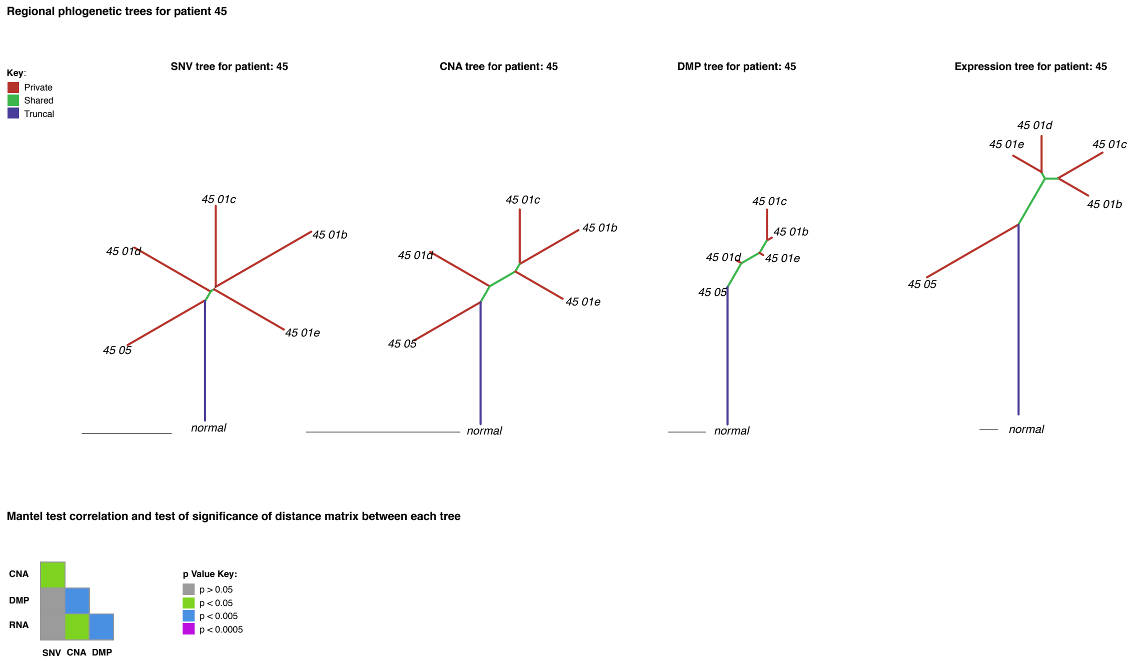


Figure 115: Top: Comparison of regional phylogenetic trees from patient 45 (HPV negative) generated from single nucleotide variants, copy number aberrations, differentially methylated positions and mRNA expression. Regional trees generated using the 'ape' R package. Horizontal line at the bottom of each tree represents a normalised scale of Euclidean distance. Bottom: Mantel test of significance of correlation between distance matrices generated between each molecular aberration type. Regions ending in 05 represent LN metastases, regions ending in a letter represent primary tumour samples.

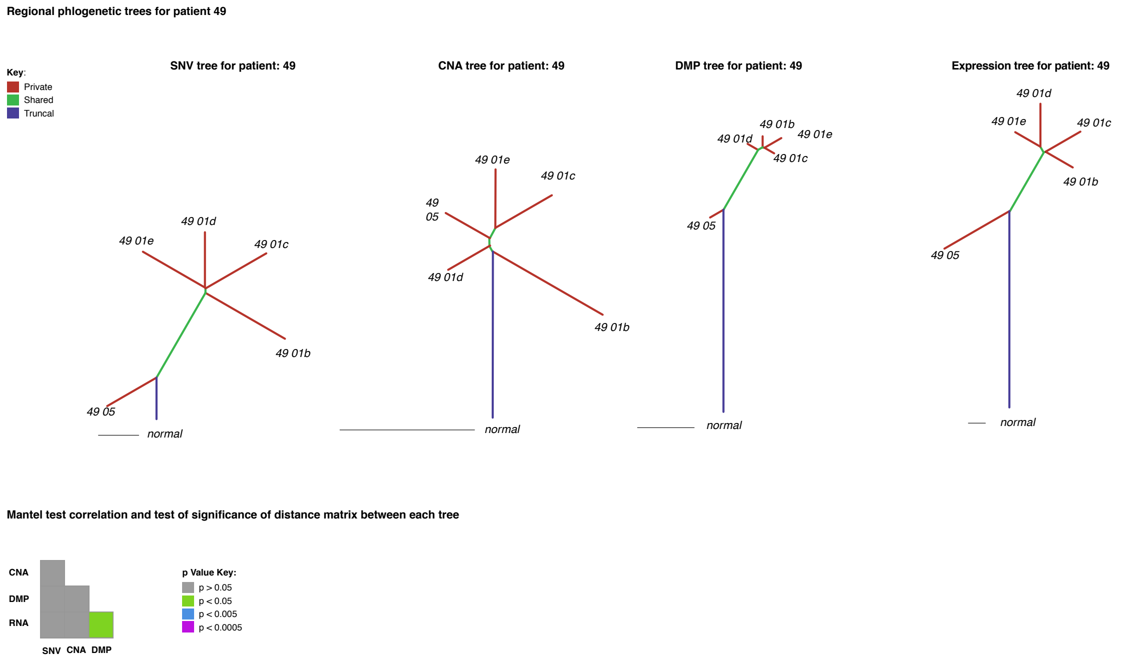


Figure 116: Top: Comparison of regional phylogenetic trees from patient 49 (HPV positive) generated from single nucleotide variants, copy number aberrations, differentially methylated positions and mRNA expression. Regional trees

generated using the 'ape' R package. Horizontal line at the bottom of each tree represents a normalised scale of Euclidean distance. Bottom: Mantel test of significance of correlation between distance matrices generated between each molecular aberration type. Regions ending in 05 represent LN metastases, regions ending in a letter represent primary tumour samples.

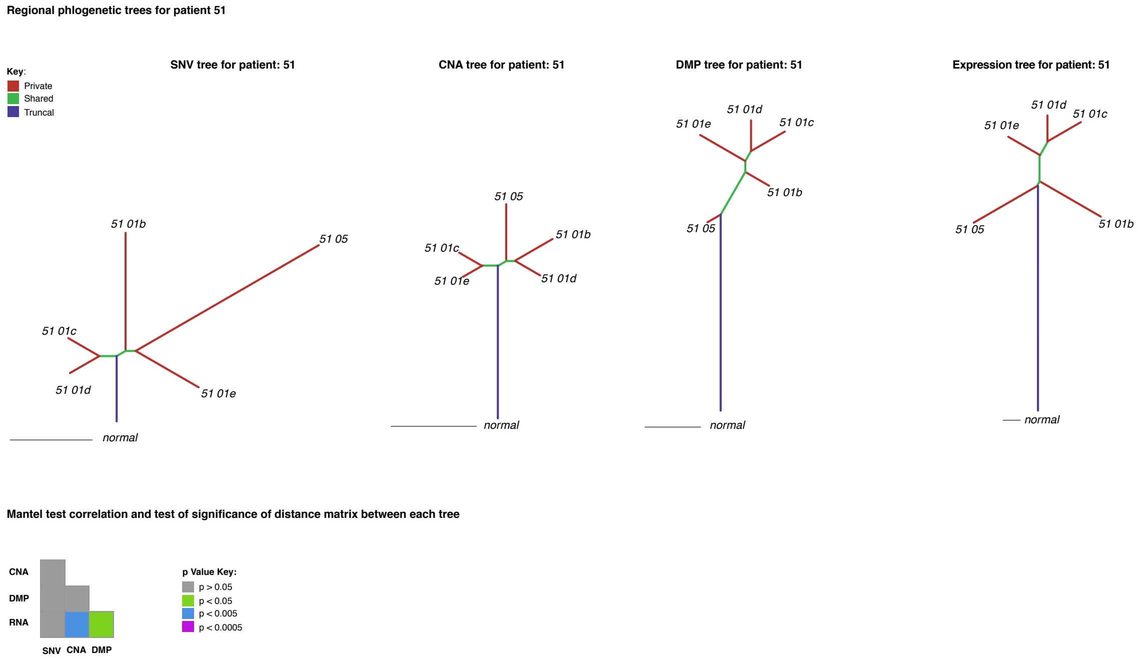


Figure 117: Top: Comparison of regional phylogenetic trees from patient 51 (HPV positive) generated from single nucleotide variants, copy number aberrations, differentially methylated positions and mRNA expression. Regional trees generated using the 'ape' R package. Horizontal line at the bottom of each tree represents a normalised scale of Euclidean distance. Bottom: Mantel test of significance of correlation between distance matrices generated between each molecular aberration type. Regions ending in 05 represent LN metastases, regions ending in a letter represent primary tumour samples.

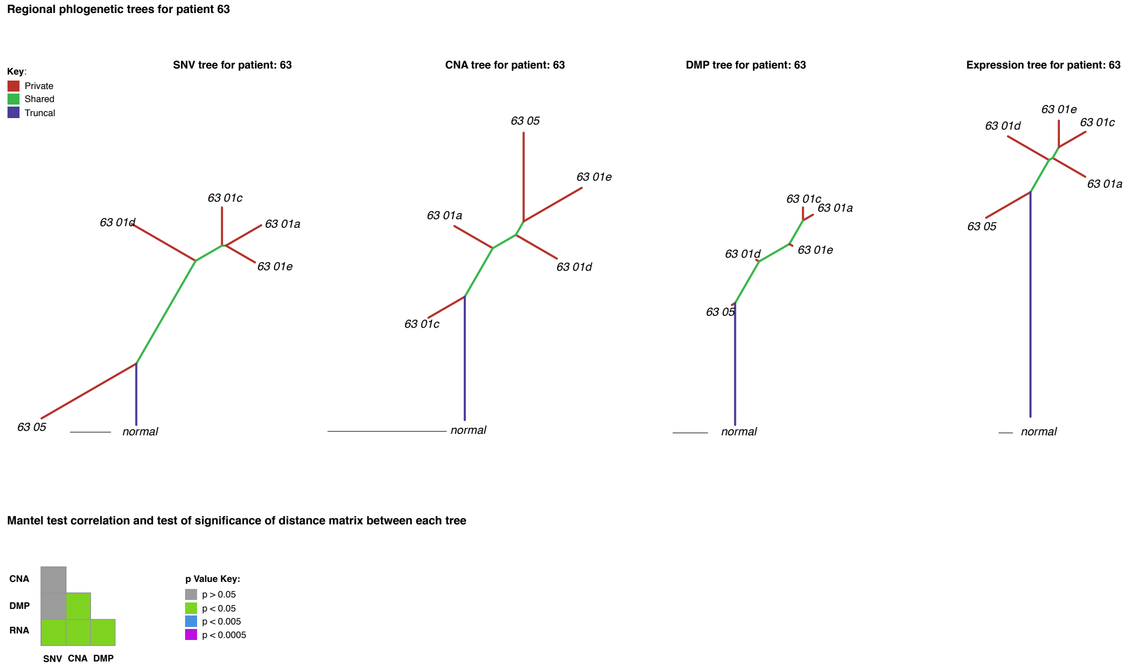


Figure 118: Top: Comparison of regional phylogenetic trees from patient 63 (HPV positive) generated from single nucleotide variants, copy number aberrations, differentially methylated positions and mRNA expression. Regional trees generated using the 'ape' R package. Horizontal line at the bottom of each tree represents a normalised scale of Euclidean distance. Bottom: Mantel test of significance of correlation between distance matrices generated between each molecular aberration type. Regions ending in 05 represent LN metastases, regions ending in a letter represent primary tumour samples.

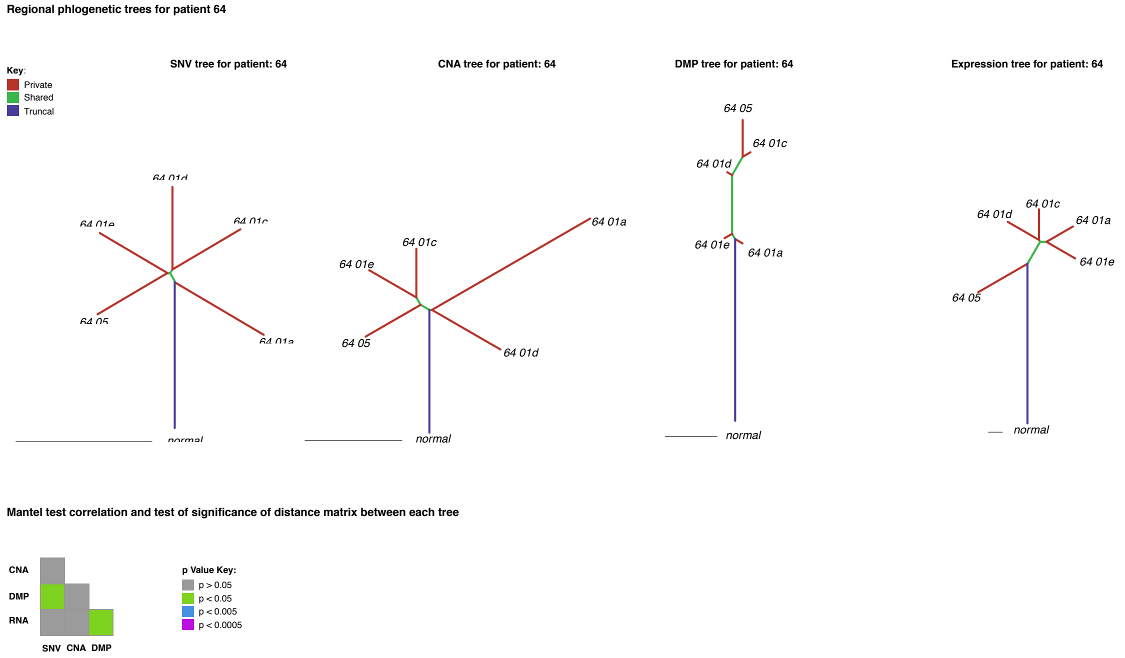


Figure 119: Top: Comparison of regional phylogenetic trees from patient 64 (HPV negative) generated from single nucleotide variants, copy number aberrations, differentially methylated positions and mRNA expression. Regional trees generated using the 'ape' R package. Horizontal line at the bottom of each tree represents a normalised scale of

Euclidean distance. Bottom: Mantel test of significance of correlation between distance matrices generated between each molecular aberration type. Regions ending in 05 represent LN metastases, regions ending in a letter represent primary tumour samples.

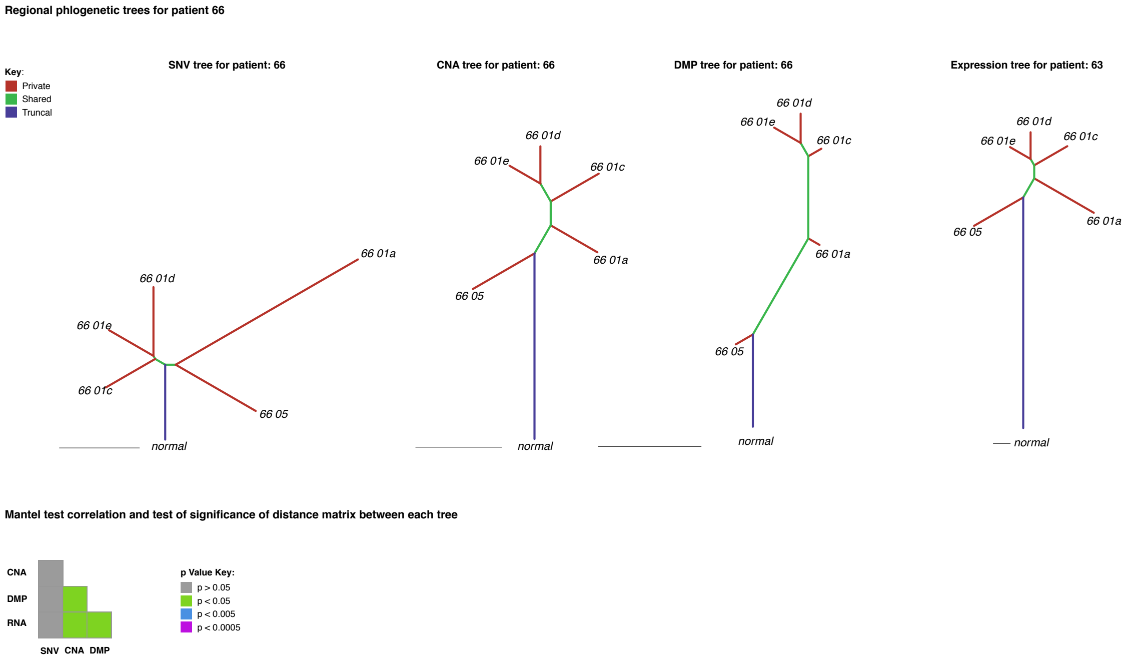


Figure 120: Top: Comparison of regional phylogenetic trees from patient 66 (HPV negative) generated from single nucleotide variants, copy number aberrations, differentially methylated positions and mRNA expression. Regional trees generated using the 'ape' R package. Horizontal line at the bottom of each tree represents a normalised scale of Euclidean distance. Bottom: Mantel test of significance of correlation between distance matrices generated between each molecular aberration type. Regions ending in 05 represent LN metastases, regions ending in a letter represent primary tumour samples.

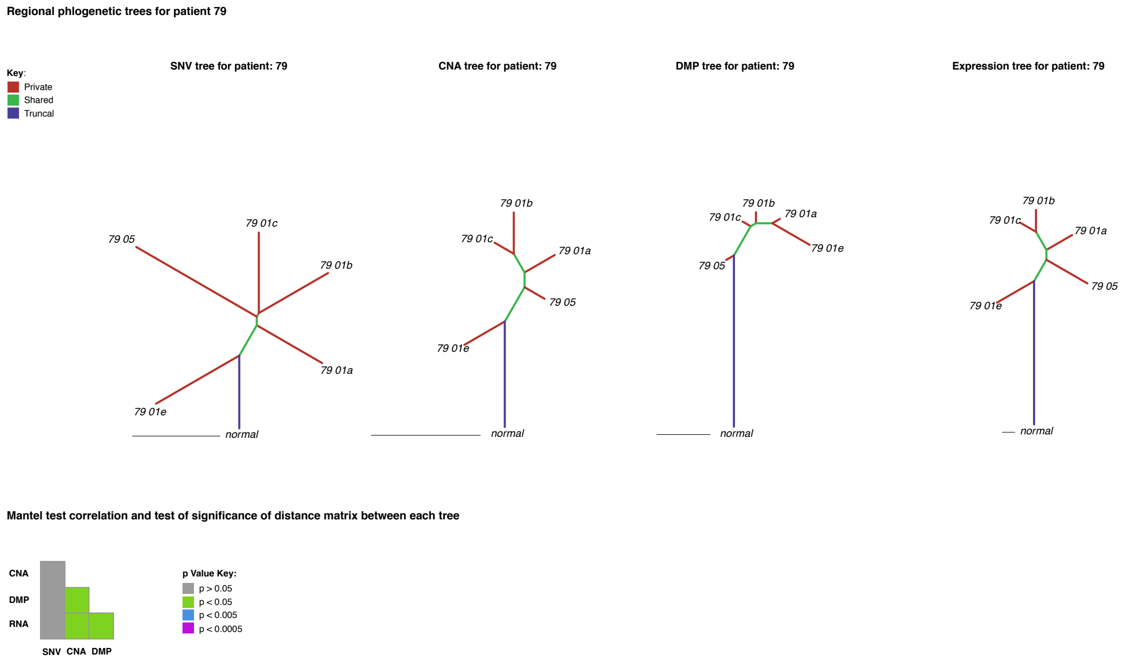


Figure 121: Top: Comparison of regional phylogenetic trees from patient 79 (HPV positive) generated from single nucleotide variants, copy number aberrations, differentially methylated positions and mRNA expression. Regional trees generated using the 'ape' R package. Horizontal line at the bottom of each tree represents a normalised scale of Euclidean distance. Bottom: Mantel test of significance of correlation between distance matrices generated between each molecular aberration type. Regions ending in 05 represent LN metastases, regions ending in a letter represent primary tumour samples.

5.6.2 Intra-tumour heterogeneity (ITH) scoring

As with genomic and epigenomic heterogeneity, we sought to quantify the level of intra-tumour heterogeneity. The extent of intra-tumour heterogeneity can be scored by several methods. Most simply the following can be calculated for each patient:

$$ITH = 1 - n$$

Where n is the proportion of differentially expressed genes that are truncal in origin.

Here, truncal refers to genes that are differentially expressed in all primary tumour regions. As in the previous chapter, assessing methylation changes (Chapter 4), the lymph node metastasis is excluded from this calculation. This is because the sensitivity of detecting a differentially expressed gene previously found in the primary tumour is lower in the lymph node metastasis. The lowered sensitivity is caused by the signal from the metastatic lymph node tumour cells being diluted, as well as influenced, by the presence of infiltrating immune cells. Furthermore, the signal is additionally influenced by the tissue specific expression of the surrounding 'histopathologically normal' lymph node itself. This is compounded by the unavailability of a 'normal' lymph node control in this sample set.

The total number of differentially expressed genes for each patient is displayed in Table 33. Table 34 displays scores for intra-tumour heterogeneity as calculated using the formula for ITH above. Table 35 compares the relative ITH scores across patients and biological aberrations. As demonstrated there is a significant increase in the ITH score for HPV negative samples when compared with HPV positive samples ($p < 0.0001$). This is demonstrated throughout assessment of SNVs, DMPs and differentially expressed genes. This data suggests that HPV positive samples are less heterogeneous than HPV negative samples. Table 35 also suggests that gene expression changes are the least heterogeneous with the lowest ITH scores when compared with SNVs or DMPs. One may hypothesise that invasive cancer encompasses a genetic and epigenetic instability phenotype where cells accumulate many aberrations – particularly methylation DMPs – due to this instability and stochastic effects. This results in large numbers of likely passenger genetic and epigenetic events creating ever increasing branches signifying further heterogeneity on the phylogenetic trees. It would seem possible that cancer cells can tolerate large numbers of these passenger aberrations. On the other hand, with gene expression there may be less scope for a viable cell surviving with large numbers of genes differentially expressed, therefore there is less scope for viable subclones and hence ITH.

Table 33: Table displaying the total number of differentially expressed genes per patient.

Patient identifier	**HPV status**	**Hyper-expressed genes**	**Hypo-expressed genes**	**Hyper and Hypo expressed genes**	**Total number of differentially expressed genes**
39	Negative	826	1007	397	2230
45	Negative	1196	761	262	2219
49	Positive	1004	956	292	2252
51	Positive	742	1043	448	2233
63	Positive	1155	919	176	2250
64	Negative	1133	903	191	2227
66	Negative	1202	838	200	2240
79	Positive	933	1015	285	2233
	HPV positive (mean)	958.5	983.25	300.25	2242
	HPV negative (mean)	1089.25	877.25	262.5	2229
	Total (mean)	1023.875	930.25	281.375	2235.5

Table 34: Table displaying the proportion of differentially expressed genes in the trunk as well as ITH scores for the patient expression level regional phylogenetic trees.

Patient identifier	HPV status	Number of differentially expressed genes in trunk	Differentially expressed genes in trunk as percentage of total	ITH score
39	Negative	1024	45.92%	54.08%
45	Negative	1192	53.72%	46.28%
49	Positive	1356	60.21%	39.79%
51	Positive	1196	53.56%	46.44%
63	Positive	1520	67.56%	32.44%
64	Negative	1204	54.06%	45.94%
66	Negative	1346	60.09%	39.91%
79	Positive	1270	56.87%	43.13%
	HPV positive (mean)	1335.5	59.55%	40.45%
	HPV negative (mean)	1191.5	53.45%	46.55%
			p < 0.0001	p < 0.0001
	Total (mean)	1263.5	56.50%	43.50%

Table 35: Table comparing the ITH scores generated by assessing the aberrations in single nucleotide variants (SNVs), differentially methylated positions (DMPs) and differential expression.

Patient identifier	HPV status	SNV ITH	DMP ITH	RNA ITH
39	Negative	35%	70%	54%
45	Negative	70%	65%	46%
49	Positive	54%	44%	40%
51	Positive	80%	72%	46%
63	Positive	23%	66%	32%
64	Negative	65%	79%	46%
66	Negative	85%	83%	40%
79	Positive	67%	70%	43%
	HPV positive (mean)	56%	64%	40%
	HPV negative (mean)	64%	73%	47%

5.7 Recurrent intra-tumour and inter-patient differential expression (recurrent truncal events)

One method of determining the likelihood of driver gene status is to determine which genes are recurrently over- or under-expressed throughout the cohort. This was completed above in Section 5.4. Furthermore, genes that are additionally present in all primary tumour sections of all patients are likely to be important in the oncogenesis of penile cancer. The presence of these genes in every patient primary sample also indicates that these genes are early truncal changes and therefore represent a special case of truncal events that are present in every patient of the PenHet cohort.

All the differentially expressed genes in Section 5.4.1 were therefore assessed to determine what proportion of them are present in all patient samples. 179 genes were found to be differentially expressed in all primary tumour regions from all patients. This gene list can be found in the Appendix. A bar chart of the genes with the 50 largest log2 fold changes can be seen in Figure 122. Of all these changes, only three genes have previously been described in the COSMIC database: *GAS7*, *GATA3* and *TERT*. One can hypothesise that these 179 genes are likely to be more important and more likely to be replicated in independent studies than genes that are only differentially expressed in a portion of primary tumour cancer samples.

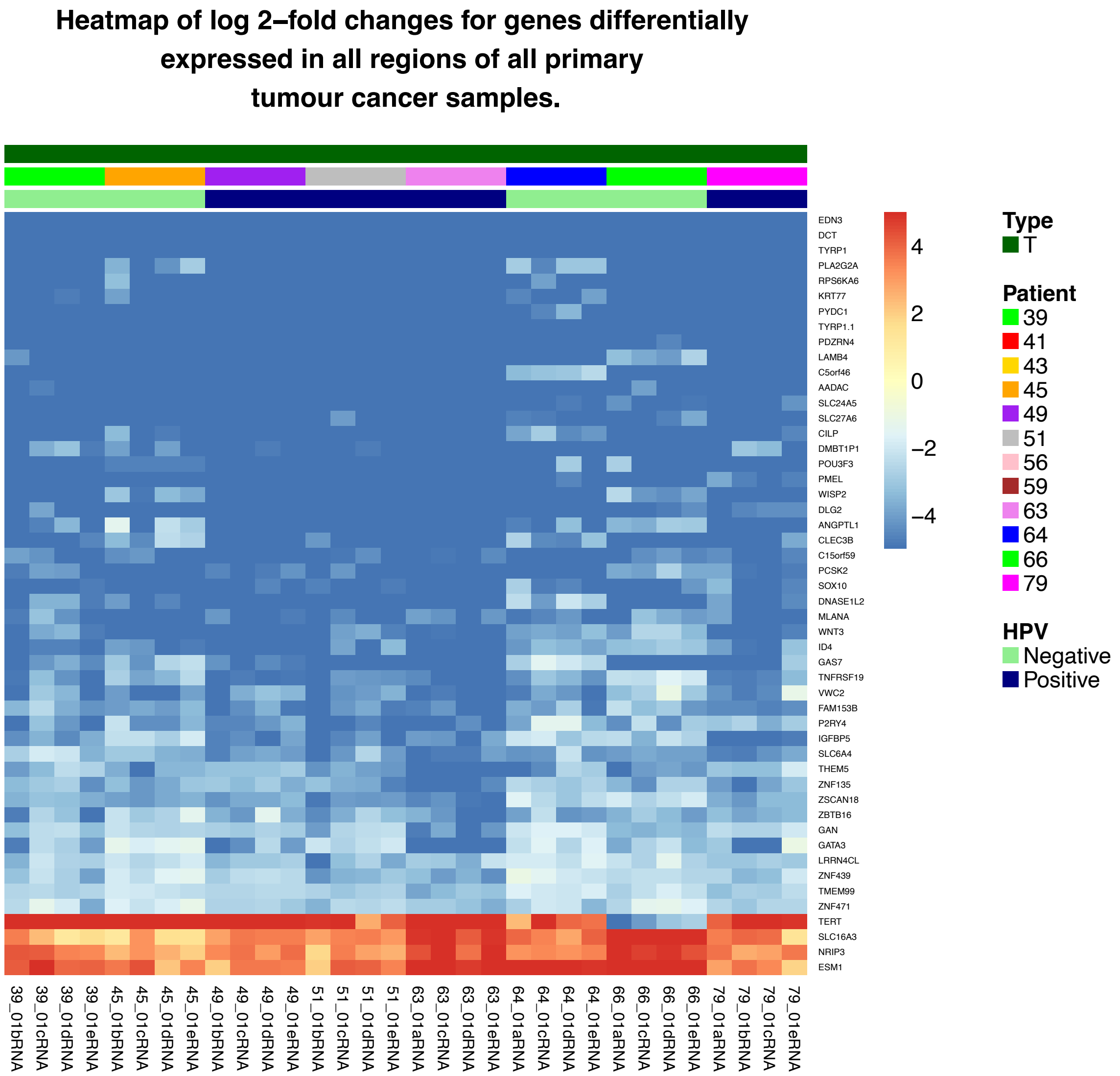


Figure 122: Bar chart displaying the 50 genes with the smallest adjusted p values that are differentially expressed in all primary tumour regions of all patients. Regions ending in 05 represent LN metastases, regions ending in a letter represent primary tumour samples. Heatmap cell colours depict log2 fold changes for each gene ranging from -5 (blue) to +5 (red).

All the differentially expressed genes were assessed for their presence in the Marchi external independent cohort of patients as described above. By reducing the dataset to include just the recurrent differentially expressed genes, there was a significant enrichment in the proportion of genes (37.3% in the recurrent group versus 15.2% in the non-recurrent group, $p < 0.00001$) that could be corroborated in this external cohort. If this list of genes is filtered further to only the top 50 with the largest adjusted p values, then the proportion of genes validated in the external dataset rises to 50%. One hypothesis that arises from this result is that truncal genes appear to be more likely to validate than non-truncal genes. However, a caveat here is that truncal genes are more detectable by their very nature – in other words, due to them being recurrent in a larger proportion of tumour cells. Therefore, there may be an element of bias whereby late branch events are more difficult to detect, as they are in a lower proportion of any region of bulk tumour cells, but feasibly could be just as likely to occur within a tumour as a truncal event. Currently, there appears to be a higher validation in an external cohort, and functional analysis would be needed before one could determine whether a particular gene is vital in the oncogenesis of penile squamous cell carcinoma.

5.8 Predicting recurrent SNV mutation drivers

Determining which DNA mutations are pathogenic and drive oncogenesis in penile cancer would be useful clinically to predict success of targeted biological therapies and academically to improve our understanding of the disease. One method of predicting which mutations are influencing a change in expression is to determine whether any genes were consistently associated with a change in gene expression across the PenHet cohort.

The whole exome sequencing data generated in Chapter 3 was integrated into the RNA-sequencing data by the use of R package 'xseq'[127] as described in the Methods (Chapter 2). Xseq utilises a hierarchical Bayes statistical model to systematically quantify the impact of somatic mutations on expression profiles. Xseq was deployed in cis mode to produce a list of drivers. Potential genetic drivers were ranked by probabilities. Only one gene had a probability of greater than 95% of being a significant driver. This was due to the limited statistical power owing to the lack of mutations that were recurrent and the relatively low number of patients in the PenHet cohort.

TP53 was the only gene found to be recurrently mutated and differentially expressed in the same samples with a probability of 99%. It is of particular interest as it was demonstrated in Chapter 3 to be a clonal mutation in all the HPV negative samples. This integrated analysis now shows that there is an association between early truncal mutations in *TP53* of the HPV negative samples and corresponding loss of *TP53* expression in these samples. Interestingly, although there was a clonal early SNV mutation of *TP53* for all the HPV negative samples, there was no loss of expression in the tumour samples from patient 64. Unlike the nonsense SNV mutation discovered in the other HPV negative patients – where there is loss of gene expression – all the tumour samples from patient 64 had a had a missense mutation R235H SNV in *TP53* with no loss of gene expression. R235H has previously been characterised as a pathogenic SNV and although gene expression may be maintained, the resulting protein may no longer be functional.

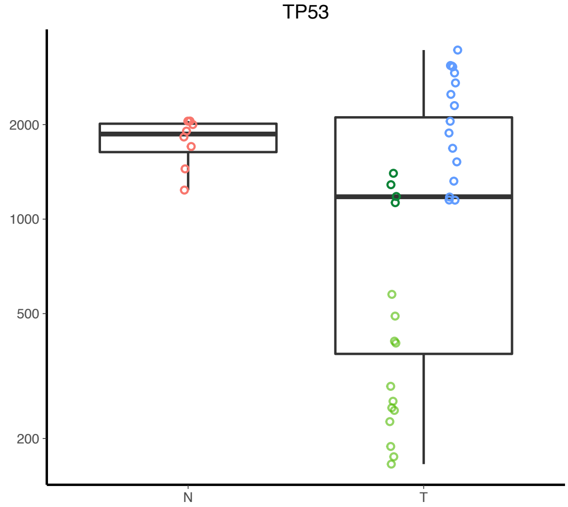


Figure 123: Normalised transcript counts reduced to gene level for gene TP53. Results displayed for both HPV positive samples and HPV negative samples. Control samples were obtained from tissue adjacent 'normal' controls as described above. Dark green scatter points represent HPV negative RNA samples with concurrent missense mutations, whilst light green scatter plots represent HPV negative RNA samples with concurrent nonsense mutations. Blue scatter points represent HPV positive samples.

Due to the low mutation rate in the PenHet cohort it was difficult to detect recurrent genetic drivers. In the case of *TP53*, there was an almost universal association between mutation and change in expression. No other genes displayed such a universal association. Just three genes displayed a probability of more than 40%, reflecting a weaker association between mutations and associative changes in gene expression than in *TP53*. These genes were:

- *CDKN2A* (63%), where there is a shared splice site mutation in all primary tumour regions of patient 66
- *USP8* (60%), where there is a shared nonsense mutation (Q368X) in all primary tumour regions of patient 49
- *NOTCH2* (52%), where there is a shared missense mutation (P863A) in all primary tumour regions of patient 63

Box and scatterplots for these three genes, shown in Figure 124, help to visualise the association between mutation and change in expression. Unlike in the case of *TP53,* differential expression was still found in many of the samples despite no genetic mutation being found. In these cases, other molecular aberrations such as epigenetic events may be driving the differential expression in the samples not harbouring the specific mutations.

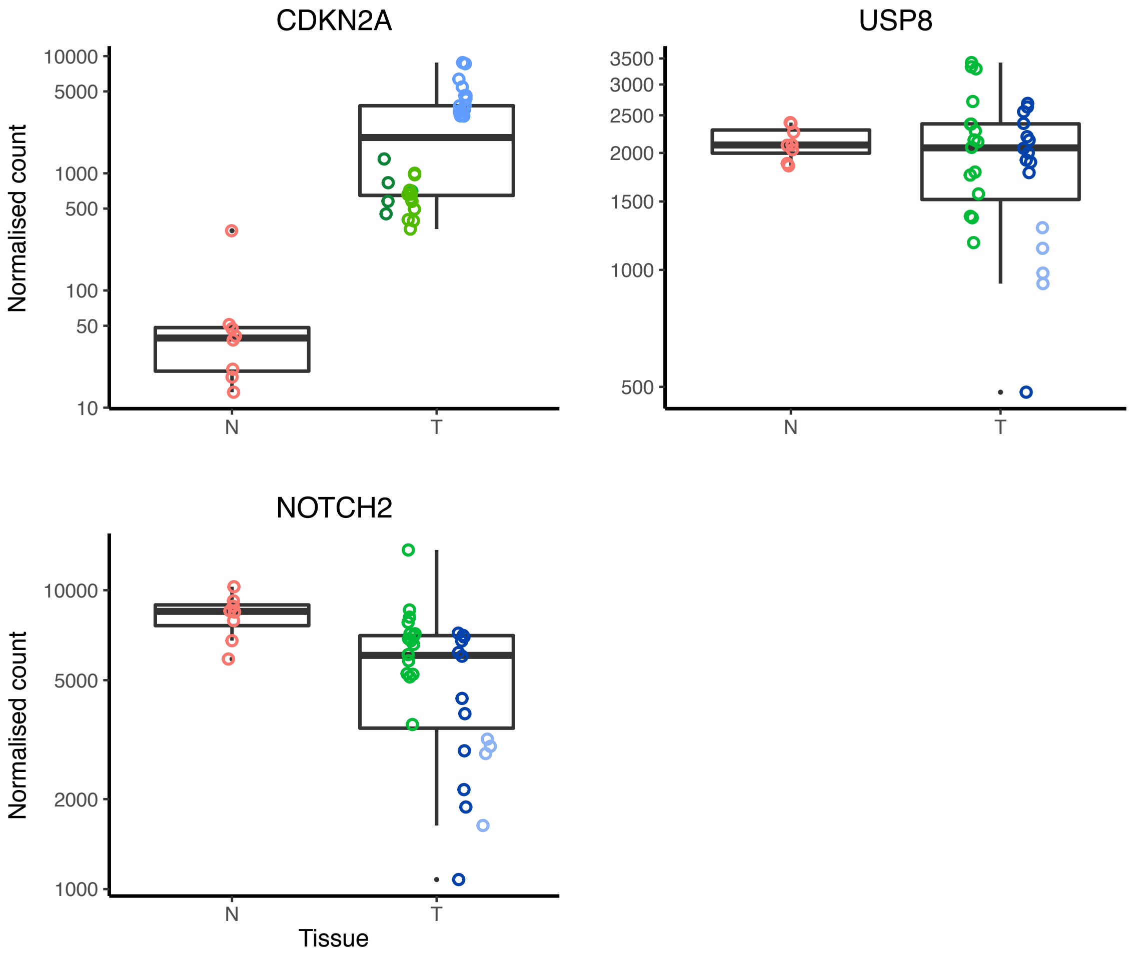


Figure 124: Additional genes identified by xseq where a mutation is associated with change in gene expression. Red scatter points represent HPV negative samples. Green scatter points represent HPV negative samples. Blue scatter points represent HPV positive samples.

CDKN2A: A shared splice site mutation was found in the primary tumour samples from patient 66 represented by the dark green scatter points. The light green scatter points represent the remaining HPV negative samples.

USP8: A shared nonsense mutation (Q368X) was found in the primary tumour samples from patient 49 represented by the light blue scatter points. The dark blue scatter points represent the remaining HPV positive samples.

NOTCH2: A shared missense mutation (P863A) was found in the primary tumour samples from patient 63 represented by the light blue scatter points. The dark blue scatter points represent the remaining HPV positive samples.

5.9 Predicting copy number aberration (CNA) drivers

Copy number aberrations have the potential to cause over-expression of oncogenes and under or loss of expression of tumour suppressor genes. To determine, which CNAs were potentially driving a change in gene expression the CNAs discovered in Chapter 3 were integrated with the RNA sequencing data by the use of R package 'iGC'[128]. Each gene overlapping a region of CNA was paired with cis gene expression data. Student's t-test with unequal variance was used to identify differentially expressed genes ($p < 0.001$) that were significantly associated with CNA.

348 genes were found to have concurrence of a copy number gain or loss, and simultaneous gain or loss of expression respectively. The complete list of genes can be found in the Appendix. When filtering these genes for the presence in the COSMIC database, 10 genes were found and in all cases were early truncal CNA events:

- CCND1 (gain)
- NFIB (gain)
- FANCD2 (loss)
- FHIT (loss)
- MLH1 (loss)
- FGFR1 (loss)
- PCM1 (loss)
- PCSK7 (loss)
- TGFBR2 (loss)
- IKBKB (loss)

The details of the associated CNA and change in expression are displayed in Table 36.

Table 36: Results from integration of copy number data with differential gene expression analysed by iGC to determine candidate copy number aberration drivers. Genes filtered for presence in the COSMIC gene database. Full results available in the Appendix.

Gene symbol	Adjusted p value	Copy gain in proportion	Copy neutral in proportion	Copy loss in proportion	Copy gain mean expression	Copy neutral expression	Copy loss mean expression	Change in expression	In COSMIC database	Expression change
CCND1	0.000	0.25	0.75	0.00	2.45	-1.51	NA	3.96	TRUE	gain
NFIB	0.000	0.25	0.75	0.00	3.17	-0.76	NA	3.93	TRUE	gain
FANCD2	0.007	0.00	0.63	0.38	NA	0.74	-0.34	-1.08	TRUE	loss
FHIT	0.013	0.00	0.63	0.38	NA	-1.61	-3.12	-1.51	TRUE	loss
MLH1	0.022	0.00	0.63	0.38	NA	-0.17	-0.91	-0.73	TRUE	loss
FGFR1	0.037	0.00	0.78	0.22	NA	-0.87	0.60	1.46	TRUE	loss
PCM1	0.038	0.00	0.78	0.22	NA	-0.21	-0.63	-0.42	TRUE	loss
PCSK7	0.045	0.00	0.59	0.41	NA	-0.91	-1.32	-0.41	TRUE	loss
TGFBR2	0.049	0.00	0.63	0.38	NA	-2.22	-0.82	1.40	TRUE	loss
IKBKB	0.050	0.00	0.78	0.22	NA	-0.86	-1.61	-0.75	TRUE	loss

These 10 genes should be investigated further to determine the associated functional affects and association with disease outcomes. It is interesting to note that the CNAs of these ten genes were shared and therefore likely represent early events in the PenHet cohort. In terms of external validation, *CCND1* has previously been found in penile cancer[243] to be amplified and associated with a shorter time to progression and death. *CCND1* regulates G1 cell cycle progression by controlling the phosphorylation of RB1. Deregulation of cyclin D1 is therefore a clear pathway to genomic instability and oncogenesis. Copy number gain and increased expression has previously been found in other squamous cell carcinomas including oral[129] and oesophageal carcinomas[244].

5.10 Integrating methylation and RNA expression data

Changes in CpG methylation within cancer cells have the potential to cause a change in gene expression. Therefore, within penile cancer there may exist a group of genes that are aberrantly methylated, causing over-expression of oncogenes or loss/under-expression of tumour suppressor genes. To determine whether any methylation events were associated with gene expression, all statistically significant differentially methylated positions were integrated with the RNA sequencing data. For this analysis, data from primary versus tissue adjacent normal samples were utilised.

A comparison of log2 fold changes at different differentially methylated positions (DMPs) was undertaken. All DMPs were grouped via the CpG genomic features into: promoters, 5'UTR, Body, 1st exon and 3'UTR. This was then compared to the log2 fold change in expression of genes associated with that genomic feature. A comparison of gene expression (log2 fold change) by DMP genomic feature and whether a DMP was hyper- or hypo-methylated can be visualised in Figure 125 and Figure 126. The overall trend is for hypermethylation at gene promoters to be associated with a loss of gene expression. The opposite effect is observed at the gene body.

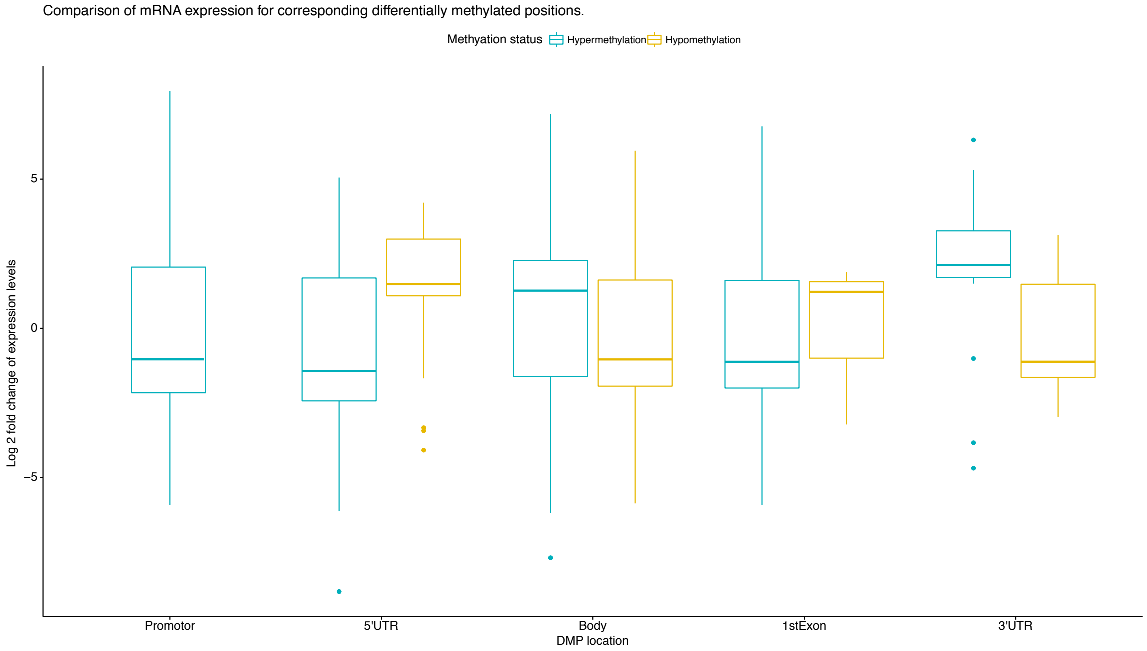


Figure 125: Comparison of differential expression of a gene with concurrent differential methylation based on CpG location.

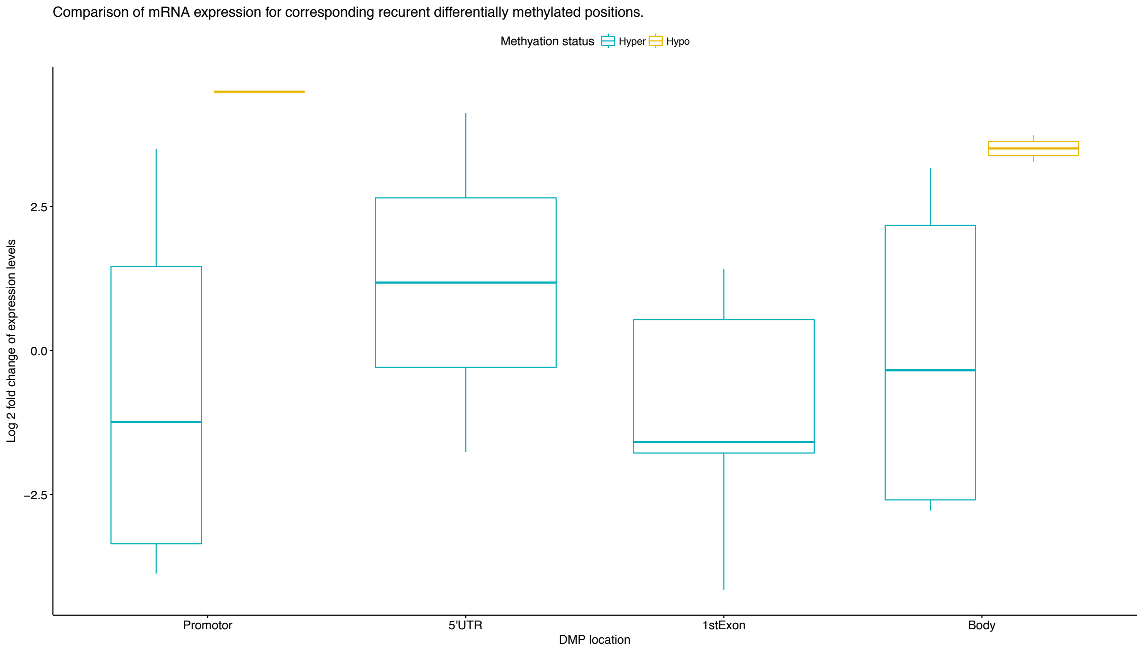


Figure 126: Comparison of differential expression of a gene with concurrent differential methylation based on CpG location. DMPs were filtered for only those which were recurrent throughout all patients in the PenHet cohort.

The previous chapter (Chapter 4) determined a list of recurrently differentially methylated genes. This list was then integrated into the gene expression data to determine whether there was any associated change in expression. The results of this can be seen in Table 37, below. As clearly demonstrated, in 27 out of 93 cases there is an associative change in expression. This is significantly more than one would expect by chance (p = 0.00008).

Table 37: Recurrent differentially methylated positions across all eight patients in the PenHet cohort. Where differential expression was found, this is noted in the final two columns.

Probe ID	Gene	In COSMIC	Gene location	CpG location	Hyper or hypomethylated	Truncal	Validated in external dataset	Log 2 fold expression change	Differential expression
cg16845394	RSPO2	TRUE	TSS200	S_Shore	Hyper	Trunk	TRUE	-3.36743547	Under
cg26799474	CASP8	TRUE	5'UTR		Hypo	Trunk	TRUE		None
cg13823136	ST6GALNAC5	FALSE	TSS200	Island	Hyper	Trunk	TRUE		None
cg14794428	ASCL1	FALSE	TSS200	N_Shore	Hyper	Trunk	TRUE		None
cg07601320	RP11-96H19.1	FALSE	TSS200		Hyper	Trunk	TRUE		None
cg26394244	NID2	FALSE	TSS200	Island	Hyper	Trunk	TRUE		None
cg03278146	C18orf42	FALSE	TSS200	Island	Hyper	Trunk	TRUE		None
cg06433694	CTC-512J12.4	FALSE	TSS200		Hyper	Trunk	TRUE		None
cg15241920	TTYH1	FALSE	TSS200	N_Shore	Hyper	Trunk	TRUE		None
cg27477373	AC006116.21	FALSE	TSS200	Island	Hyper	Trunk	TRUE		None
cg08701621	ZNF135	FALSE	TSS200	Island	Hyper	Trunk	TRUE	-3.338122443	Under
cg12919006	AC079154.1	FALSE	TSS200	Island	Hyper	Trunk	TRUE		None
cg13356896	BOLL	FALSE	TSS200	Island	Hyper	Trunk	TRUE	2.946951303	Over
cg26492446	BHLHE23	FALSE	TSS200	Island	Hyper	Trunk	TRUE		None
cg14859460	GRM6	FALSE	TSS200	Island	Hyper	Trunk	TRUE		None
cg02467990	VWC2	FALSE	TSS200	Island	Hyper	Trunk	TRUE	-3.869141936	Under
cg24928391	SOX17	FALSE	TSS200	Island	Hyper	Trunk	TRUE	-1.239784517	Under
cg14653281	GRIN3A	FALSE	TSS200	Island	Hyper	Trunk	TRUE		None
cg10636246	AIM2	FALSE	TSS1500		Hypo	Trunk	TRUE	4.497122628	Over
cg21675115	EDNRB	FALSE	TSS1500	S_Shore	Hyper	Trunk	TRUE	-3.548728665	Under
cg08217024	MDGA2	FALSE	TSS1500	Island	Hyper	Trunk	TRUE		None
cg09624466	OTX2-AS1	FALSE	TSS1500	Island	Hyper	Trunk	TRUE		None
cg04037038	FRZB	FALSE	TSS1500	Island	Hyper	Trunk	TRUE		None
cg00970325	PAQR9	FALSE	TSS1500	Island	Hyper	Trunk	TRUE	3.500672994	Over
cg03304610	GALNTL6	FALSE	TSS1500	Island	Hyper	Trunk	TRUE	-2.993617852	Under
cg18325622	MARCH11	FALSE	TSS1500	Island	Hyper	Trunk	TRUE		None
cg09591286	ZNF804B	FALSE	TSS1500	N_Shore	Hyper	Trunk	TRUE		None
cg07792478	MIR124-2	FALSE	TSS1500	Island	Hyper	Trunk	TRUE		None
cg21578219	IGSF21	FALSE	5'UTR	Island	Hyper	Trunk	TRUE		None
cg10224098	RNF220	FALSE	5'UTR	Island	Hyper	Trunk	TRUE		None
cg14780632	GAL3ST3	FALSE	5'UTR	Island	Hyper	Trunk	TRUE		None
cg14699728	NPAS4	FALSE	5'UTR	Island	Hyper	Trunk	TRUE		None
cg23989821	C14orf39	FALSE	5'UTR	Island	Hyper	Trunk	TRUE	4.123056099	Over
cg02401399	AC002116.7	FALSE	5'UTR	Island	Hyper	Trunk	TRUE		None
cg26246807	ZIK1	FALSE	5'UTR	Island	Hyper	Trunk	TRUE		None
cg06714480	CERKL	FALSE	5'UTR	N_Shelf	Hyper	Trunk	TRUE	0.79291019	Over
cg06818532	BBX	FALSE	5'UTR		Hyper	Trunk	TRUE		None
cg05663573	CLDN11	FALSE	5'UTR	N_Shore	Hyper	Trunk	TRUE	-0.950792218	Under
cg15031661	FMN2	FALSE	1stExon	Island	Hyper	Trunk	TRUE	-1.559293268	Under
cg20168230	GRIK3	FALSE	1stExon	Island	Hyper	Trunk	TRUE		None
cg19839798	FAM155A	FALSE	1stExon	Island	Hyper	Trunk	TRUE	-1.832861363	Under
cg04945331	SOX14	FALSE	1stExon	Island	Hyper	Trunk	TRUE		None
cg09221867	PCDH10	FALSE	1stExon	Island	Hyper	Trunk	TRUE		None
cg05716166	RALYL	FALSE	1stExon	N_Shore	Hyper	Trunk	TRUE		None
cg08738570	TMEM240	FALSE	Body	N_Shore	Hyper	Trunk	TRUE		None
cg09671258	LHX4	FALSE	Body	Island	Hyper	Trunk	TRUE		None
cg07046369	PAX2	FALSE	Body	Island	Hyper	Trunk	TRUE		None
cg18419977	SLC22A18	FALSE	Body	N_Shelf	Hypo	Trunk	TRUE	2.368988287	Over
cg17203063	KCNC2	FALSE	Body	Island	Hyper	Trunk	TRUE		None
cg25317585	FGF14	FALSE	Body	Island	Hyper	Trunk	TRUE	-2.528437996	Under
cg13012916	RP11-896J10.3	FALSE	Body	Island	Hyper	Trunk	TRUE		None
cg18716164	VSTM2B	FALSE	Body	Island	Hyper	Trunk	TRUE		None
cg06428620	PCDHGA1	FALSE	Body	N_Shore	Hyper	Trunk	TRUE		None
cg18789958	HCN1	FALSE	Body	N_Shore	Hyper	Trunk	TRUE		None
cg20348196	HLA-DRA	FALSE	Body		Hypo	Trunk	TRUE		None
cg21179088	VSTM2A	FALSE	Body	Island	Hyper	Trunk	TRUE		None
cg11395386	MECP2	FALSE	Body	N_Shelf	Hyper	Trunk	TRUE	0.52654875	Over
cg09755589	RP11-276E17.2	FALSE	Intergenic		Hypo	Trunk	TRUE		None
cg26986871	KLRC4-KLRK1	FALSE	Intergenic		Hypo	Trunk	TRUE		None
cg26132774	RNF219-AS1	FALSE	Intergenic	Island	Hyper	Trunk	TRUE		None
cg01630690	NKX2-1-AS1	FALSE	Intergenic	Island	Hyper	Trunk	TRUE		None
cg00001747	LINC01158	FALSE	Intergenic	Island	Hyper	Trunk	TRUE		None
cg25946059	LINC01237	FALSE	Intergenic		Hypo	Trunk	TRUE		None
cg02421985	FOXO1B	FALSE	Intergenic	Island	Hyper	Trunk	TRUE		None
cg04349084	RP11-175E9.1	FALSE	Intergenic		Hypo	Trunk	TRUE		None
cg18249580	RP11-32K4.1	FALSE	Intergenic	N_Shore	Hyper	Trunk	TRUE		None
cg22001496	C8orf34	FALSE	Intergenic	Island	Hyper	Trunk	TRUE	-0.828190154	Under
cg22770135	TRPC7	FALSE	TSS200		Hyper	Trunk	FALSE		None
cg27136241	RP5-1186N24.3	FALSE	TSS200	S_Shore	Hyper	Trunk	FALSE		None
cg08063125	ZNF667	FALSE	TSS1500	Island	Hyper	Trunk	FALSE	-0.755183835	Under
cg26597242	GABRA1	FALSE	TSS1500		Hyper	Trunk	FALSE		None
cg15423872	FAM110B	FALSE	TSS1500	N_Shore	Hyper	Trunk	FALSE		None
cg10109500	GHSR	FALSE	1stExon	Island	Hyper	Trunk	FALSE		None
cg12880658	CDO1	FALSE	1stExon	Island	Hyper	Trunk	FALSE	-4.15868335	Under
cg17627654	SHANK2	FALSE	Body	Island	Hyper	Trunk	FALSE	3.170789382	Over
cg08567279	LECT1	FALSE	Body	Island	Hyper	Trunk	FALSE		None
cg27555582	ABCC9	FALSE	Intergenic	Island	Hyper	Trunk	FALSE		None
cg11014373	RP11-13J10.1	FALSE	Intergenic	Island	Hyper	Trunk	FALSE		None
cg19971388	GATA4	FALSE	Intergenic	Island	Hyper	Trunk	FALSE	4.922503087	Over
cg24931138	TERT	TRUE	Body	S_Shore	Hypo	Branch	FALSE	3.748486902	Over
cg09842118	RNASE3	FALSE	5'UTR		Hypo	Branch	TRUE		None
cg22459052	SLC6A7	FALSE	Body		Hypo	Branch	TRUE	3.27393201	Over
cg12930338	MS4A6E	FALSE	TSS200		Hypo	Branch	FALSE		None
cg19828416	OR4D1	FALSE	TSS1500	N_Shelf	Hypo	Branch	FALSE		None
cg27109600	SATB2	FALSE	TSS1500	N_Shore	Hyper	Branch	FALSE	0.795993288	Over
cg06580419	AC107218.3	FALSE	TSS1500		Hypo	Branch	FALSE		None
cg14627175	DACT2	FALSE	TSS1500	S_Shore	Hypo	Branch	FALSE		None
cg19825483	RYR2	FALSE	Body		Hypo	Branch	FALSE		None
cg06375949	MSX1	FALSE	Body	N_Shore	Hyper	Branch	FALSE		None
cg11234281	ZNF32-AS3	FALSE	Intergenic		Hypo	Branch	FALSE		None
cg23253961	CTD-2277K2.1	FALSE	Intergenic		Hypo	Branch	FALSE		None
cg03116035	RP11-205M3.3	FALSE	Intergenic		Hypo	Branch	FALSE		None

Potential epigenetic drivers were assessed using the 'MethylMix'[129,130] R package, as explained in the Methods (Chapter 2). Methylation gene drivers are genes that when aberrantly methylated cause an over-expression of oncogenes or silencing/loss of expression of tumour suppressor genes. Unlike in Table 37 where the total log2 fold change is calculated for each recurrent DMP, MethylMix identifies differential and functional DNA methylation by using a beta mixture model to identify subpopulations of samples with different DNA methylation compared with normal tissue. Functional DNA methylation is predicted by MethylMix based on correlations with matched gene expression data. Eighteen genes were found to be recurrently differentially methylated, as well as highly associated with changes in gene expression. These 18 genes were: *NPTX2*, *CBLN1*, *ZNF682*, *DUT*, *GBGT1*, *ZNF85*, *PAPLN*, *ZNF577*, *CXCL14*, *ADAP1*, *ZNF727, CDX2, HOXD13, OLIG2, WT1, ZNF135, AIM2* and *ZNF429*. Graphs 1-5 in Figure 127 to Figure 131 demonstrate the association between change in beta methylation score and log2 fold expression change between tumour and normal.

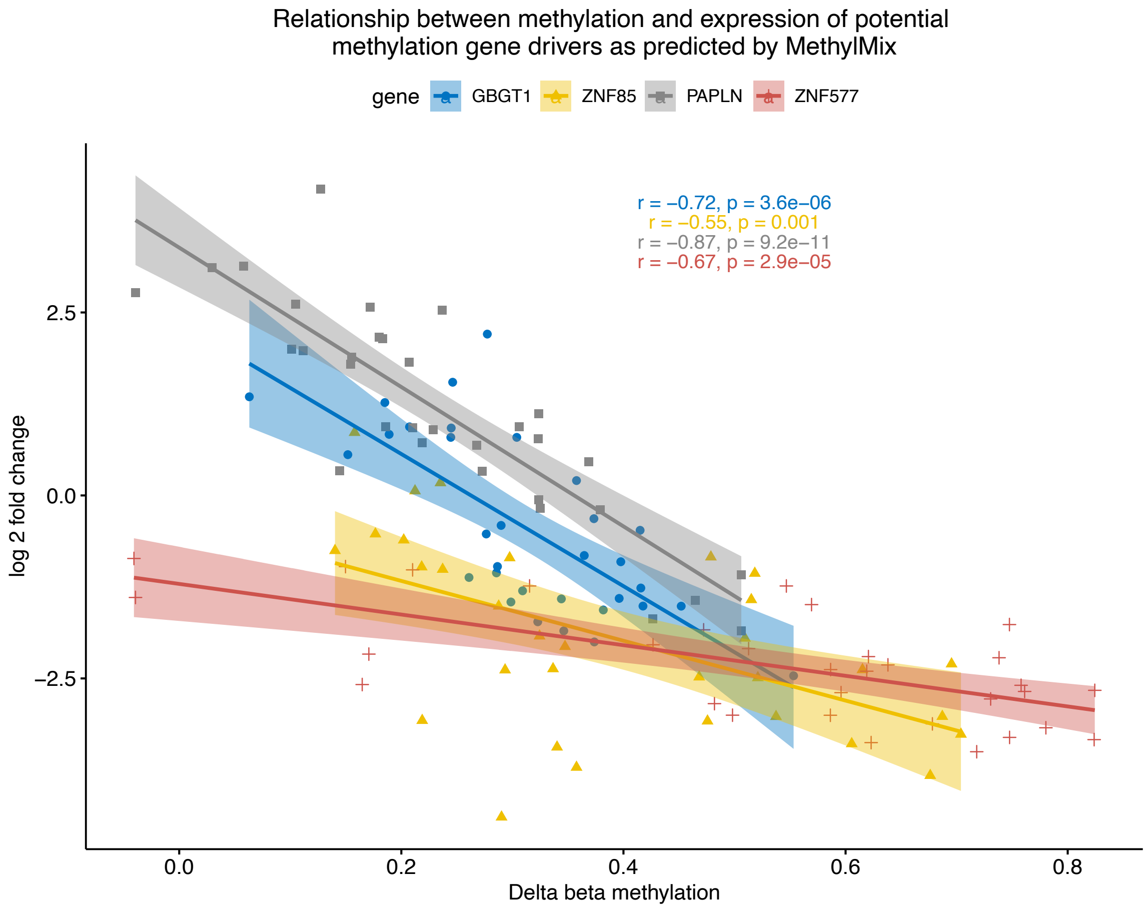


Figure 127: Graph (1) depicting the relationship between log2 fold change of gene expression against delta promoter beta methylation value for tumour versus adjacent normal tissue for the genes GBGT1, ZNF85, PAPLN and ZNF577. Genes selected based on output from MethylMix integration analysis.

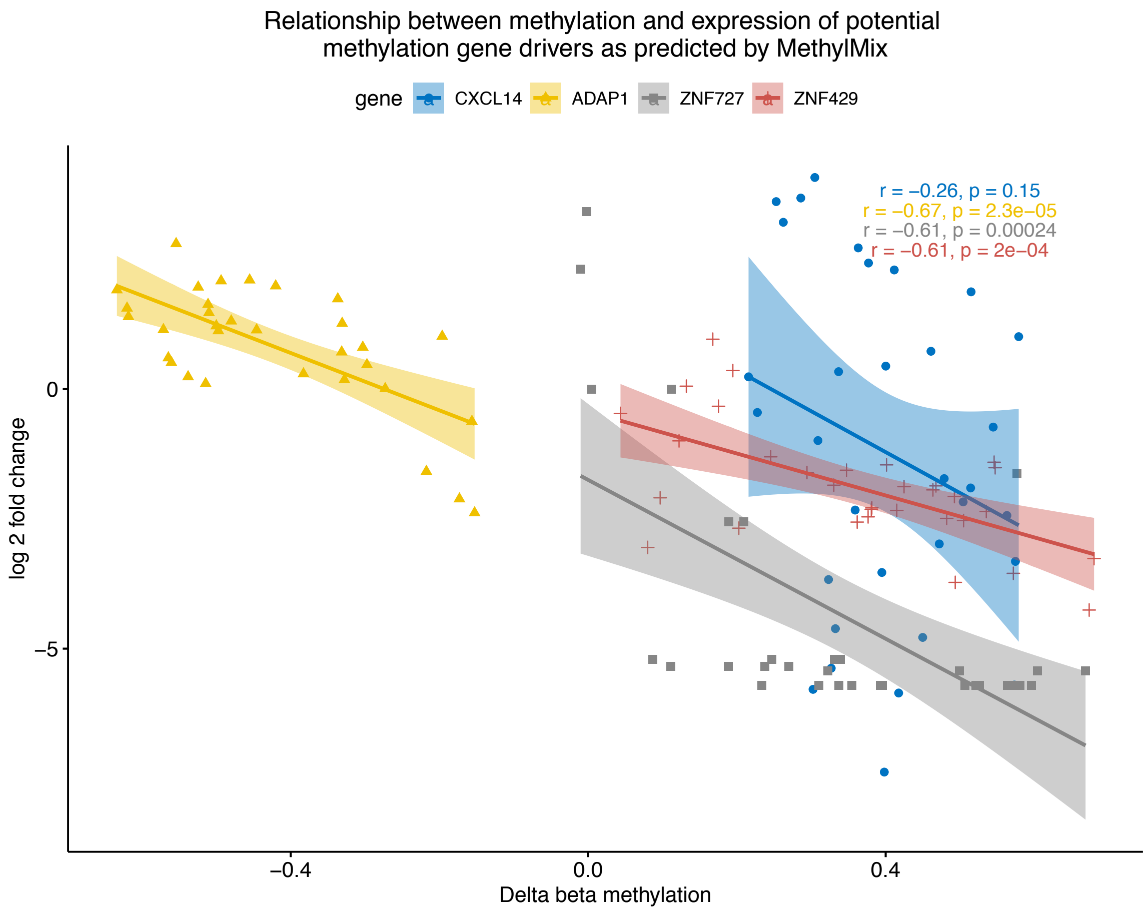


Figure 128: Graph (2) depicting the relationship between log2 fold change of gene expression against delta promoter beta methylation value for tumour versus adjacent normal tissue for the genes CXCL14, ASAP1, ZNF727 and ZNF429. Genes selected based on output from MethylMix integration analysis.

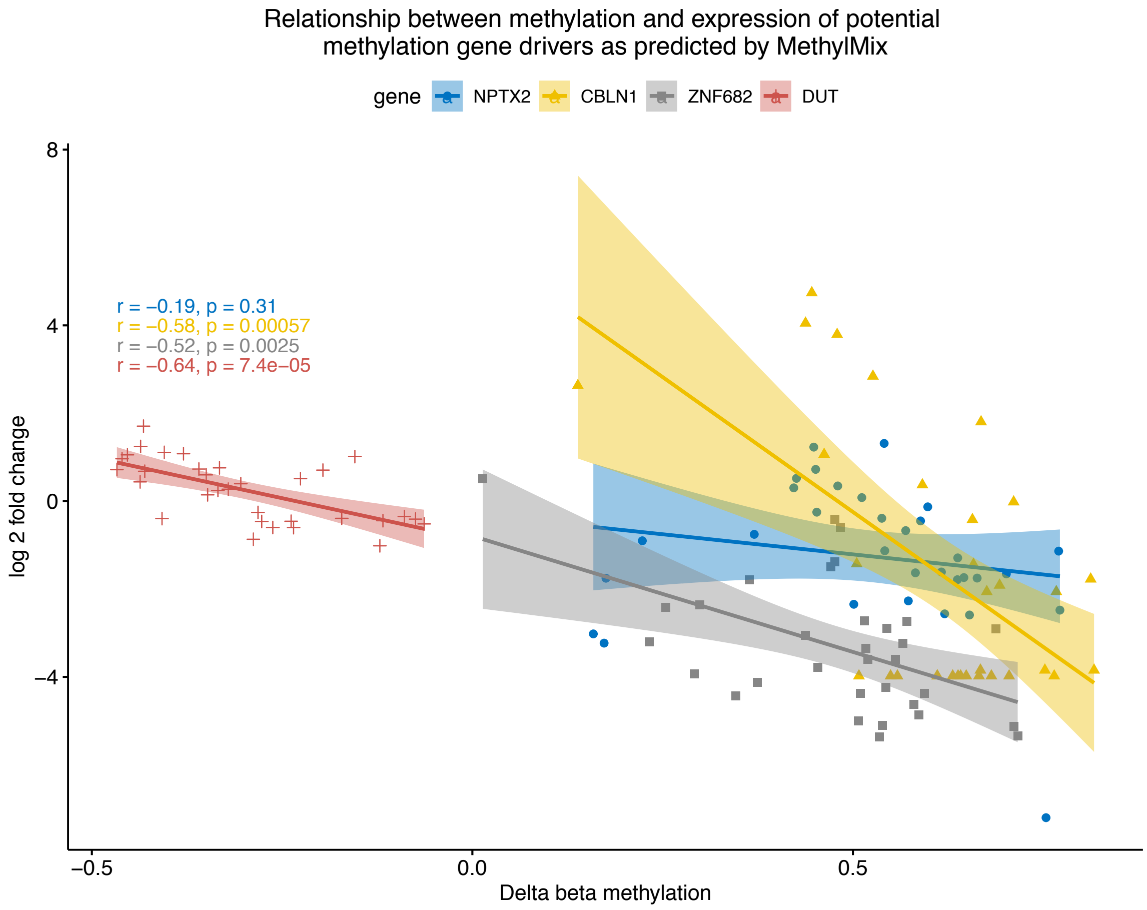


Figure 129: Graph (3) depicting the relationship between log2 fold change of gene expression against delta promoter beta methylation value for tumour versus adjacent normal tissue for the genes NPTX2, CBLN1, ZNF682, and DUT. Genes selected based on output from MethylMix integration analysis.

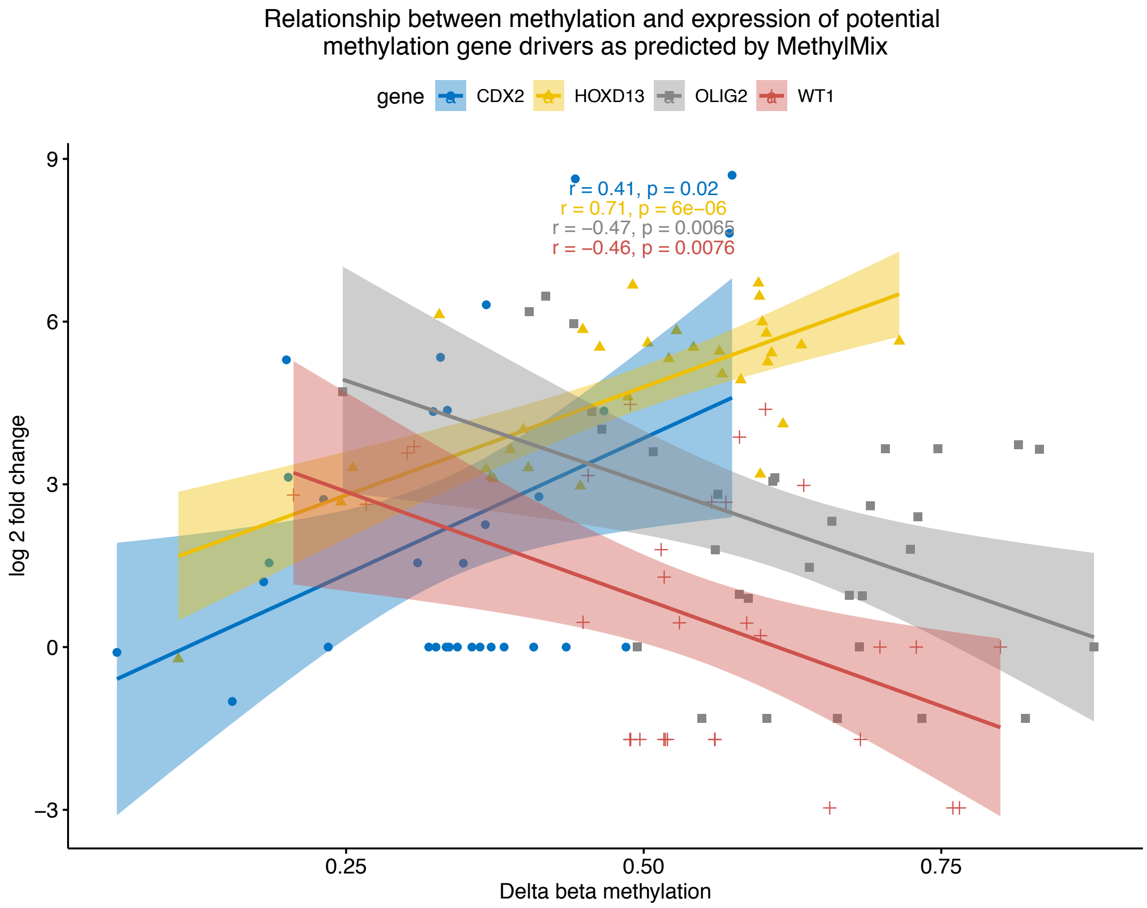


Figure 130: Graph (4) depicting the relationship between log2 fold change of gene expression against delta promoter beta methylation value for tumour versus adjacent normal tissue for the genes CDX2, HOXD13, OLIG2 and WT1. Genes selected based on output from MethylMix integration analysis.

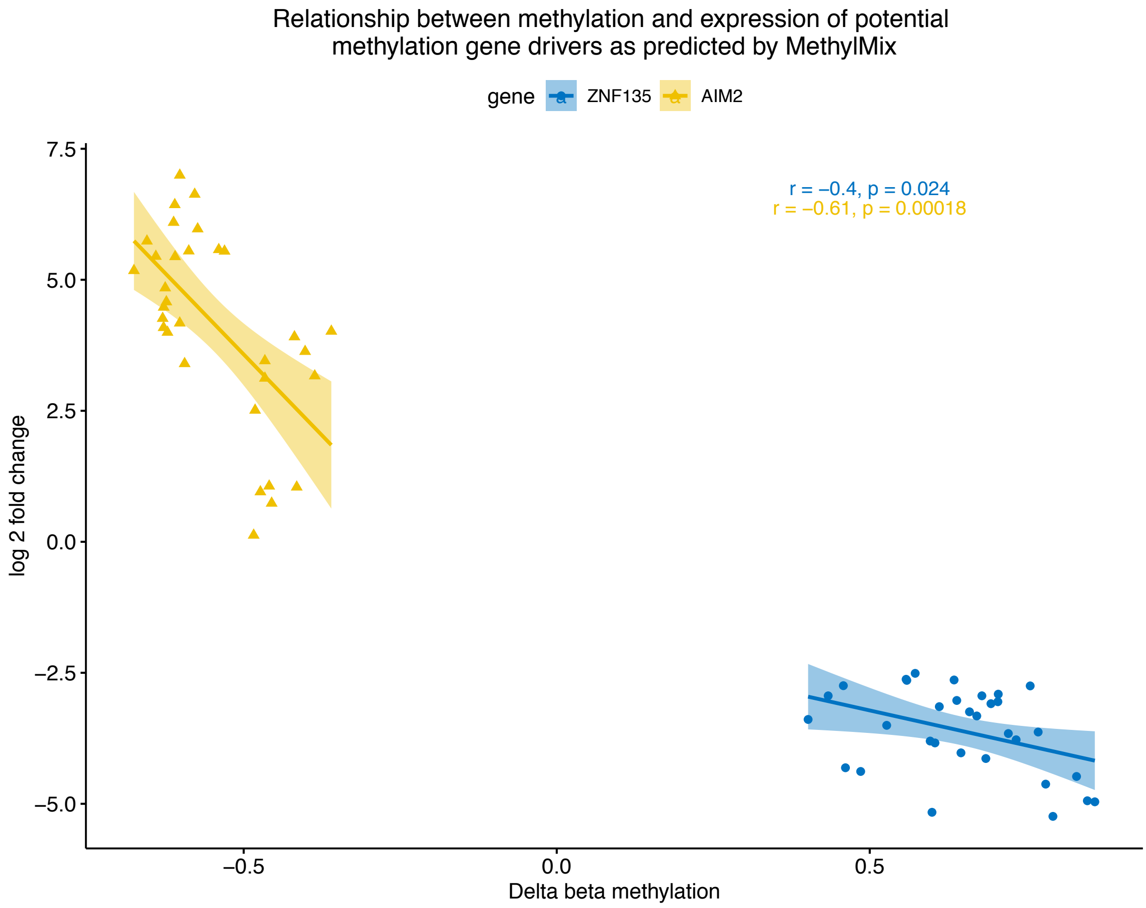


Figure 131: Graph (5) depicting the relationship between log2 fold change of gene expression against delta promoter beta methylation value for tumour versus adjacent normal tissue for the genes ZNF135 and AIM2. Genes selected based on output from MethylMix integration analysis.

5.10.1 Expression of methyltransferases

Methyltransferases are responsible for the transfer of methyl groups and therefore play an important role in regulating methylation[245].

Previous studies have demonstrated that infection with oncogenic HPV can increase expression of methyltransferases[246]. The expression of *DNMT1*, *DNMT3A*, *DNMT3B*, *TET1* and *TET2* was therefore assessed to determine whether there was any relationship between expression of methyltransferases and perturbed methylation.

Figure 132 reveals that HPV infection is associated with *DNMT1* and *DNMT3A* hyper-expression. Although all HPV positive samples displayed higher levels of *DNMT3A* expression, three samples, had particularly high levels appearing to cluster together with normalised counts of approximately 1,150. These three samples belong to regions of the primary tumour of patient 63 (01a, 01c, 01e). On further investigation, these three samples shared the same missense mutation (S770L) in *DNMT3A*.

One could hypothesise that the increased expression of these methyltransferases could lead to increased aberrant methylation[245]. Indeed, assessing the number of differentially methylated positions in the tumours of HPV patients compared with those of HPV negative patients reveals a mean increase of 38% per patient. Furthermore, the patient with the largest number of DMPs was patient 63, in whom the shared mutation in *DNMT3A* with the greatest over-expression was found. The scatter points belonging to these three *DNMT3A* mutations with the greatest over-expression are circled in purple on the scatter plot in Figure 132.

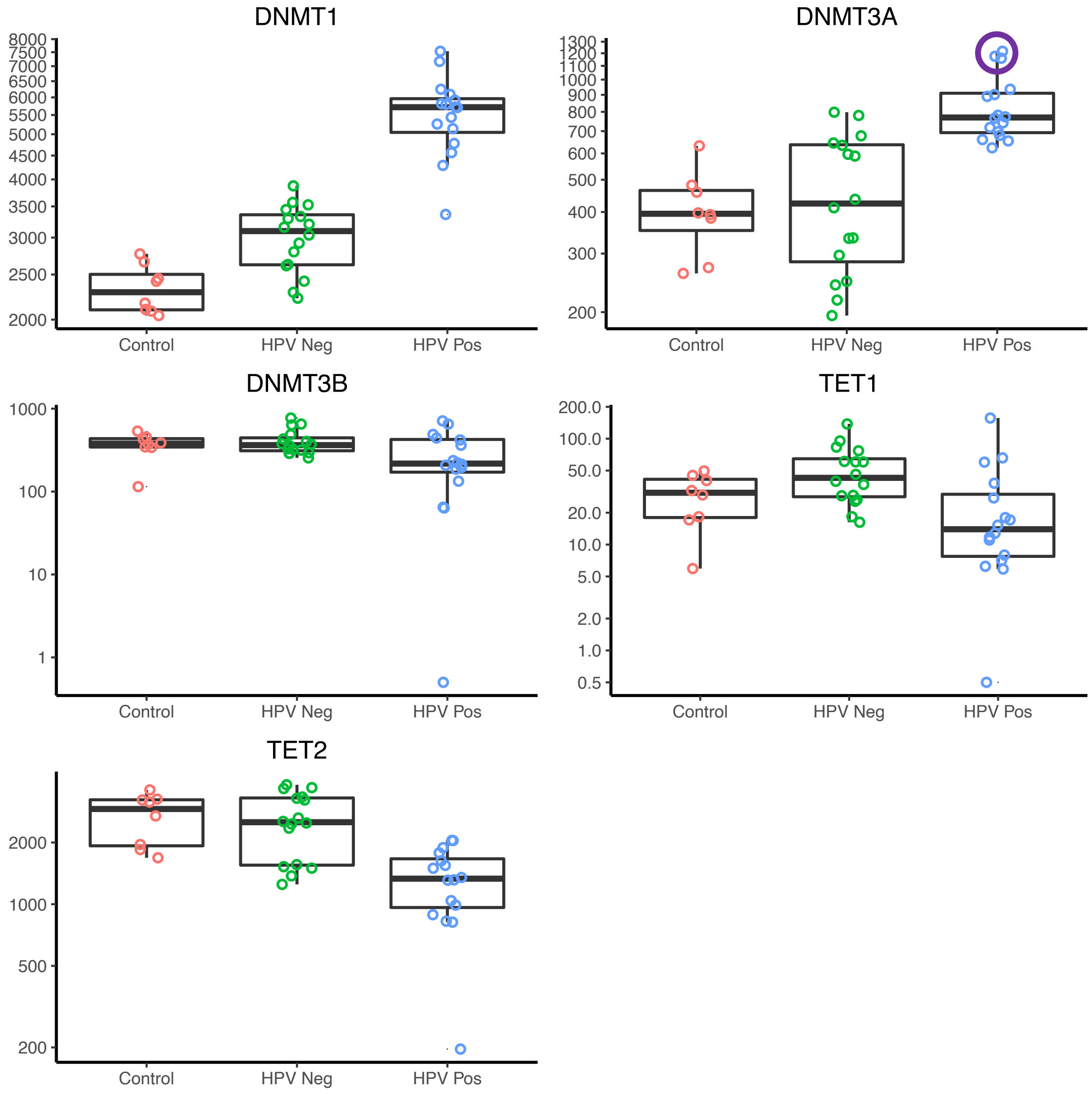


Figure 132: Comparison of differential expression of methyltransferases between HPV positive and HPV negative tumour samples.

5.11 Discussion

5.11.1 Immune system involvement and immunotherapy

There is an intimate relationship between a growing cancer and the host’s immune system. The host’s immune system plays an important role in cancer prevention. This is clearly demonstrated in the increased rates of malignancy in immunosuppressed patients. Previous studies have demonstrated the large extent of infiltrating immune cells within solid malignancies. This relationship is further highlighted by the remarkable success in treating initially melanoma and now many other solid cancers by using immunotherapy agents such as immune checkpoint inhibitors. In Chapter 3 (DNA mutation analysis) the presence of immune cells only has an indirect effect on the detection of mutations originating within the cancer cells. As the tissue is bulk sequenced, the mutated cancer cells are sequenced mixed with adjacent normal tissue and any infiltrating immune cells. In this situation, the presence of immune cell content only acts to reduce the cancer cell fraction of the bulk tissue being sequenced and so may reduce the sensitivity of calling any particular mutation. In both Chapter 4 (methylation analysis) and this chapter on RNA expression, the infiltrating tumour cells have a distinct methylation or expression profile when compared with the tumour cells or tissue adjacent controls. In these circumstances, there is a danger of attributing a transcript as being over- or under-expressed in comparison to the tumour cells, when in fact the differential expression may be caused by the presence of infiltrating immune cells. It is for this reason that a conservative approach was used in the analysis of methylation data to exclude any methylated CpG that could be seen as differentially methylated in a range of immune cell populations compared with the tissue adjacent normal samples. Due to the dangers of combining analyses from different RNA expression studies, there is no equivalent expression dataset used to reduce this potential impact.

Tumour infiltrating lymphocytes, particularly CD8+ cells, have been demonstrated in melanoma to be associated with checkpoint inhibitor response[247,248]. The presence of expression signatures from infiltrating immune cells within the PenHet samples raises the possibility that there may be high levels of infiltrating immune cells. This could be explored further by undertaking cell sorting and immunohistochemistry. If increased levels of CD8+ T-cells are found, this would suggest that immunotherapy may be efficacious in penile cancer. A further important insight is that the variability in infiltrating immune cells does not seem to be affected by the variability in HPV status.

Immune modulators, which were previously found to play a role in immune suppression in the context of sustained tumour immune antigen presentation, were found to be differentially expressed in the PenHet cohort. Significantly increased expression was found across a range of immune inhibitors, including *CTLA4*, *PDL1*, *IDO1*, *LAG3*, *TIM3*, and *KIR*. Recurrent increased expression across almost all samples was found for *CTLA4* and *IDO1*. RNA-seq methods utilised in this chapter are not currently approved in the clinical setting for assessment of immune modulator expression, such as PDL-1. When using RNA-seq transcript abundance data, there is not currently a cut-off score for designating a tumour as PDL-1 positive, in contrast to when immunohistochemistry (IHC) is used. Conroy et al have demonstrated that RNA-seq normalised transcript counts are proportional to IHC scoring[249]. For this to be used clinically, a set of comparator controls will be required. However, despite this limitation the large differential expression found in the PenHet cohort should prompt further work to ascertain whether targeting these molecules could prove to be more efficacious than the current platinum-based chemotherapy regimens.

5.11.2 Human papillomavirus (HPV)

The global expression of genes across all samples clustered by HPV status. The same phenomenon was demonstrated in the previous two chapters, Chapter 3 on DNA mutations and Chapter 4 on methylation aberration analysis. This gives further evidence that the largest component of inter-patient heterogeneity can likely be explained by the presence of activating oncogenic HPV 16. One can therefore hypothesise that different oncogenic pathways exist in these separate patient groups; they will likely respond best to differing targeted treatment modalities. An example of these different pathways is the sole presence of clonal *TP53* mutations and associated loss of expression in the HPV negative samples. A further example is the over-expression of the methyltransferases *DNMT1* and *DNMT3A* in HPV positive samples. Over-expression of these methyltransferases is associated with increasing number of DMPs found within these samples. Development of HPV positive and negative penile cancer cell lines would help to elucidate the most important pathways that could be targeted in these two subtypes of penile cancer.

Currently, patients are not routinely screened for HPV status, so there is a lack high-quality studies to determine the different outcomes for these patient groups. In 2018, the NHS has now agreed to vaccinate young boys as well as young girls. It will be important to assess over the coming decades how the rates of chronic HPV infection and HPV-related malignancies fall. Due to the long lead time from initial infection with HPV and resulting malignancy many decades later, the prevalence of HPV-related penile cancer may not fall for many decades.

5.11.3 Intra-tumour heterogeneity (ITH)

ITH was found throughout all three experimental modalities in this thesis (whole exome sequencing, methylation arrays and RNA sequencing). The greatest amount of ITH was found in CpG methylation, whilst the least was found in gene expression. This suggests that a proportion of branch differential methylation events do not result in any change in gene expression. Despite the extensive heterogeneity, recurrent changes in gene expression were found for every patient. These recurrent changes can be defined as truncal early events, in comparison to the changes found solely in a subset of regions sampled.

Modelling of each tumour's underlying clonal and subclonal structure was undertaken in Chapter 3, revealing subclones in the primary tumour of every patient in the PenHet cohort. The comparison regional phylogenetic trees, where SNV, copy number, methylation and gene expression were compared for each patient, revealed statistically similar structures as determined by the Mantel test. This is because each region investigated is made up of a proportion of cells belonging to the initial clone as well as subclones, irrespective of which type of molecular aberration is interrogated. However, specific subclones may exert greater differences in one modality – for example, methylation – than another – for example, genetic mutations. This imbalance can cause differences to arise between the regional phylogenetic trees.

ITH of genes that can be targeted may be important indicators of clinical efficacy. These are discussed in further detail in Section 5.11.3.1, below.

5.11.3.1 Targeted therapies

The recurrent DNA mutations and aberrant methylation changes discovered in the previous two chapters, once validated in larger cohorts, may become useful biomarkers to be used in the diagnosis of metastatic disease, to measure treatment response and for long-term surveillance. But more importantly, when combined with expression data, they can be used to hypothesise on the important driving cancer pathways present in advanced squamous cell penile cancer. This data can then be used to hypothesise on future therapeutics.

Due to the rarity of penile cancer, very little progress has been made in undertaking clinical trials to take advantage of the developments in targeted treatment modalities. Only in 2018, following significant efforts, there are a few small-cohort trials into which patients with penile cancer can be recruited. The major targeted therapies that are being tested include EGFR inhibitors and direct tyrosine kinase inhibitors.

The results from both the mutation analysis and expression analysis suggest that *EGFR* is only expressed in the later stages of cancer development at the subclonal level. If these results are generalisable to the general population of metastatic penile cancer patients, this may therefore represent a major challenge in successfully treating these patients.

Unlike *EGFR*, *DDR2* was found to be hyper-expressed universally in the trunk of patients in the PenHet cohort. This likely represents an early event in the oncogenesis of penile cancer. As a tyrosine kinase, it is also potentially targetable with an inhibitor such as dasatinib, which is currently in clinical testing and development.

GAS7 was found to exhibit promoter hypermethylation and loss of expression also universally in the trunk of patients in the PenHet cohort. Loss of expression of *GAS7* has previously been found to represent one pathway to gefitinib resistance. Gefitinib is one of the EGFR inhibitors currently under investigation in a clinical trial for penile cancer. This finding of loss of *GAS7* may represent further evidence for loss of efficacy of EGFR inhibitors in penile cancer.

Inhibition of the oncogene *c-MET*, for example by crizotinib, has been demonstrated as efficacious in a subgroup of patients with non-small cell lung cancer. In the PenHet cohort *c-MET* was found to be almost universally hyper-expressed within penile cancer. It has been proposed that activated *c-MET* can mobilise neutrophils and suppress induced T cell effector functions, thereby reducing the efficacy of immunotherapies. Combined *c-MET* inhibition and immunotherapy has been proposed as a synergistic therapy, and these results suggested in this PenHet cohort would support preliminary work in this area for penile cancer.

5.11.4 Conclusion

In conclusion, integrated analysis of DNA mutation, methylation and expression reveals extensive intra-tumour heterogeneity within invasive penile cancer. This heterogeneity has been modelled into complex tumour structures involving initial clones and subsequent subclones. Early events in HPV positive samples appear to involve integration of HPV viral genomes into the host human genomes. This is accompanied by further clonal disruption of *PIK3CA* and *MTOR* pathways. APOBEC mutagenesis causes characteristic mutational signatures detected in these HPV positive samples. Furthermore, over-expression of methyltransferases may result in further epigenetic mutagenesis. HPV negative samples are characterised by early clonal mutations in *TP53*, frequently causing loss of *TP53* expression. Loss of *TP53* is likely to be a key driving event in the oncogenesis of these samples. *CCND1*, *c-MET* and *CDKN2A* have previously been described as cancer drivers in many other malignancies and appear to be early truncal events within penile cancer.

In addition, over-expression of immune checkpoints *CTLA4* and *IDO1* provides preliminary evidence that a new class of therapeutics, namely immunotherapy check point inhibitors, may prove to be efficacious in penile cancer. This thesis provides evidence that these checkpoint inhibitors should be further investigated to determine efficacy. It is hoped that such investigations could impact those current and future patients in whom a diagnosis of invasive penile cancer currently carries a high chance of morbidity and mortality.

6 Discussion

6.1 Introduction

Penile cancer is a rare but mutilating disease. This rarity has resulted in a lack of investment in molecular research to help understand its development and progression, as well as limited new therapeutic development. The current standard of care for patients with expected lymph node metastatic disease has remained unchanged for decades: potential lymph node dissection and systemic treatment with platinum-based therapies. Despite these treatments, penile cancer patients continue to suffer from high morbidity and mortality.

Prior to the publication of this thesis, there was a paucity of molecular analyses to better understand the disease, uncover molecular drivers, and inform the development of new therapies. Given the limited number of patients available for clinical trials, it is paramount that preliminary evidence is generated quickly to ascertain which new generation of therapeutics would be most likely to succeed – be it immunotherapy, targeted molecular therapy or methylation inhibitors. Intra-tumour heterogeneity could cause targeted therapies to fail; on the other hand, it could provide antigenic substrate for immunotherapies.

This thesis represents the first comprehensive analysis of combined whole exome, methylome and mRNA expression of advanced penile squamous cell carcinoma. The overarching aim of this body of work was to characterise the likely genetic, epigenetic (DNA methylation) and changes in expression that contribute to this aggressive disease. Crucially, these changes are described below in the context of modelling the extent of intra-tumour heterogeneity, to determine which changes may contribute to early development and which to later development of penile squamous cell carcinoma.

6.2 Oncogenic drivers in penile cancer

A range of genetic, epigenetic (DNA methylation) and expression-based candidate drivers were found to be recurrently disrupted in the PenHet cohort. Among these were well known oncogenes such as *TERT* and *cMET*, in which increased expression was observed across all patients. In addition, clonal mutations were found in tumour suppressors *TP53* and *PIK3CA* in HPV negative and HPV positive samples, respectively.

Many of these differentially expressed drivers were also found in the external validation cohort published by Marchi et al[89], including *TERT*, *CDKN2A*, *NFIB*, *RSPO2*, *GRIN2A, MET* and *GAS7* in the trunks of the regional phylogenetic trees and *BRACA2*, *AR*, *KIT*, *GATA2*, *CCNE1* and *CDK6* in the branches.

When expression data was integrated into the genetic and epigenetic datasets from Chapters 3 and 4 respectively, the following drivers were found to potentially influence expression: *TP53*, *CDKN2A*, *USP8*, *NOTCH2*, *CCND1*, *NFIB*, *FHIT*, *RSPO2*, *ZNF135*, *VWC2*, *EDNRB*, *GALNTL6*, *NPTX2*, *CBLN1*, *ZNF682*, *DUT*, *GBGT1*, *ZNF85*, *PAPLN*, *ZNF577*, *CXCL14*, *ADAP1*, *ZNF727* and *ZNF429.* Each of these candidate genetic and epigenetic drivers are discussed in more detail in Chapters 3 and 4. The next steps in determining the validity of these candidate drivers is to replicate in larger studies and functionally validate them in tumour models such as penile cancer cell lines or potentially with organoids.

For the purpose of this discussion, I will focus on the intra-tumour heterogeneity, as it enables the construction of a model of penile cancer oncogenesis. Such a model provides insights into the key early molecular aberrations, which could form the basis of targetable therapies and new biomarkers.

6.3 Human papillomavirus (HPV)

Infection with high-risk HPV is a key driver of penile cancer development. HPV driven tumours show distinct epigenetic and genetic events compared with non-HPV driven tumours. Four out of eight of the patients in the PenHet cohort were found to be infected with HPV type 16. No other HPV viral subtypes were found. HPV viral reads were found both as intact complete reads, and as concatemers which indicate viral integration into host genomic sequences. It was interesting to find that for patient 79, in whom there was a very high captured viral load (30 whole HPV 16 genome equivalents captured), there was sufficient evidence to suggest that as well as integrating into other locations in a non-clonal manner, HPV had integrated into the same locus on chromosome 18 in all tumour samples. This signifies that the integration of HPV

16 was likely to have been a very early clonal event, possibly preceding many of the other genomic and epigenomic aberrations found. Clonal concatemer reads were not identified in the other three HPV positive patients, however this does not mean that HPV had not integrated in a clonal manner. Rather, it means that HPV may have integrated into a location not captured as part of the exome capture process (which only captures < 2% of the genome).

As expected when defining the mutational signatures represented in each tumour, the presence of mutational signature 2 was completely correlated with the presence of HPV viral reads, indicating enhanced APOBEC mutational activity in the HPV positive samples. This indicates that despite not being able to identify the presence of a clonal HPV integration site for all HPV positive patients, these tumours are undoubtedly driven by oncogenic HPV. APOBEC mutational signatures have previously been associated with cancers driven by oncogenic HPV subtypes[170].

The distinct effect of HPV infection was also observed when all samples from all patients were subjected to hierarchical clustering – all the samples appeared to cluster by HPV infection status. This was the case irrespective of genetic, epigenetic or expression modalities. These data further highlight that penile squamous cell carcinoma is driven by a range of genomic and epigenetic processes, which are highly dependent on the HPV status. Examples of drivers found to be dependent on HPV status included clonal mutations and loss of expression in *TP53*, found only in the HPV negative samples. In addition, clonal mutations in *mTOR* or *PIK3CA* were only found in the HPV positive samples. In addition HPV positive patients had increased expression of methyltransferases including *DNMT1* and *DNMT3A*. Infection with HPV 16 and associated increases in expression of *DNMT1* and *DNMT3A* has been previously reported in cervical cancer[250,251]. HPV 16 infection therefore may induce increased expression of *DNMT1* and *DNMT3A*, which may result in the increased aberrant methylation found in HPV positive samples when compared with HPV negative samples.

All patients irrespective of HPV status appeared to display enhanced expression of immune checkpoint proteins *CTLA4*, *IDO1* and *LAG3*. These proteins are discussed in the following Section, 6.4.

6.4 Penile cancer and the immune system

The immune system plays in intricate part in surveillance and protection against malignancies. It is not surprising that genetic, epigenetic and expression signatures associated with a substantial infiltration of immune cells, particularly Th1 and Th2 T cells, CD8+ T cells and B cells, were found within the tumour samples. This was especially the case for the lymph node metastases. Interestingly, Th1 cells were found enriched in the HPV positive samples (p = 0.0002), which may reflect the anti-viral activity of Th1 positive T cells[214].

It has previously been suggested that a higher tumour mutational burden increases the likelihood of response to immunotherapy agents [11], such as PDL1 inhibitors. Mutational load was therefore investigated for penile cancer to determine whether the cancer has a similar mutational burden to other cancers that have shown oncological efficacy with immunotherapy agents. Tumour mutational load was found to be 1.4/megabase(Mb) with a range of 0.625-7.36 mutations/Mb (Chapter 3). Although this mutational load is higher than that seen for many cancers, it is significantly lower than for melanoma, lung and bladder cancer where high mutational loads are often seen and correlated with response to immunotherapies[252,253].

As well as assessing for the presence of infiltrating immune cells, the expression levels of key immune checkpoints were also assessed. Over-expression of immune checkpoint inhibitor proteins, including *CTLA4*, *IDO1* and *LAG3,* was especially pronounced across all patients. CTLA4 inhibitors are already available clinically, and clinical trials involving IDO1 and LAG3 are currently underway[218,254]. Previous published work demonstrates recurrent PDL1 expression based on immunohistochemistry in penile cancer[212]. Given these data, it is conceivable that these patients would make good candidates for immune checkpoint blockade as part of the management in metastatic penile squamous cell carcinoma.

6.5 Intra-tumour heterogeneity (ITH)

The assessment of ITH can provide valuable information regarding the absolute amount of ITH in a particular tumour type compared with others; the clonal and subclonal structures of the cancer; the cancer cell fraction of a particular mutation; and the relative timings of aberrations. The cancer cell fraction and the derived clonal/subclonal structure is important clinically, as – all things being equal – targetable subclonal mutations within a cancer would be less efficacious than a targetable clonal mutation.

Extensive ITH has been found when examining genetic, epigenetic and expression changes throughout the PenHet cohort. When the extent of ITH was scored, the greatest amount was found in the methylome analysis (68.5%), a lower amount was found in genetic changes (60%), and the lowest amount of heterogeneity was observed at the RNA expression level (43.5%). Furthermore, there was more ITH in the HPV negative samples than the HPV positive samples, throughout all three types of molecular aberration. One possible explanation for this is that HPV driven disease occurs through a more uniform set of driving alterations (set into place by the integration of viral oncogenes E6 and E7), whereas in HPV negative disease there are a large number of non-recurrent drivers leading to alternative pathways of genomic instability.

DNA methylation is controlled by a highly regulated set of genes, including *DNMT1*, *DNMT3A*, *DNMT3B*, *TET1* and *TET2*. *DNMT1* and *DNMT3A* were found to be significantly over-expressed (compared with normal) in the trunk of the phylogenetic trees of HPV positive patients. These genes are methyltransferases, which normally regulate the methylation of CpGs. However, when over-expressed, they cause aberrant DNA methylation and disruption of normal epigenetic machinery. Previously published work demonstrates that both viral oncoproteins E6 and E7 can induce the expression of the *DNMTs* (DNA methyl transferases)[246]. In addition, E7 may directly bind to the methyltransferases, further inducing its activity. Indeed, the HPV positive patients demonstrate over-expression of *DNMT1* and *DNMT3A*, perhaps causing the increased number of DMPs found within these samples. Unregulated disruption of the methylome leads to large-scale changes in gene expression affecting genes involved in apoptosis, cell cycle control and cell adhesion.

A simplification of the oncogenesis of penile cancer, taking into account the early versus late status of drivers discovered in the PenHet cohort, can be predicted as follows:

In HPV positive patients, failure to clear oncogenic HPV 16 results in integration of its genome, leading to continuous expression of viral oncogenes E6 and E7. This integration may be particularly prominent at the fragile sites previously described in squamous cell HPV driven cervical cancer[147]. Expression of HPV oncoproteins drives genetic instability by disrupting cell cycle control mechanisms such as pRB, inhibiting apoptosis by targeting p53 ubiquitination, and increasing the expression and activity of methyltransferases *DNMT1* and *DNMT3A*[246]. This leads to further changes in gene expression of key drivers, resulting in further disruption of cell cycle control and DNA repair processes. The induced changes in *DNMT1* and *DNMT3A* expression may be responsible for the spectrum of methylation heterogeneity seen in HPV positive patients. Early clonal mutations in *PIK3CA/mTOR* pathways were also found in all four patients with HPV positive disease. Henderson et al discovered that when mutations in *PIK3CA* occur in HPV positive samples, APOBEC mutagenesis may be responsible for TCW mutations in the *PIK3CA* mutational hotspots E542K and E545K in the helical domain[170]. When searched for in the PenHet cohort, the clonal *PIK3CA* mutations that occurred in the two HPV positive patients (51 and 63) were both TCW mutations of E545K, as they were in the head and neck cancers analysed by Henderson et al[170]. One could hypothesise, therefore, that infection of HPV is an early – or even pre-malignant – event. This causes activation of APOBEC mutagenesis, likely as a form of viral defence. If the virus is not cleared, the viral oncogenes can induce genomic instability, as discussed in the Introduction (Chapter 1). Further uncontrolled APOBEC mutagenesis may then cause characteristic mutations in genes such as *PIK3CA*, a proto-oncogene. All these events may take place very early in tumorigenesis before the formation of the last common ancestor.

In HPV negative patients, the early clonal mutation in *TP53* appears to be the only recurrent genetic clonal driver. Mutation of *TP53* is associated with matched loss of expression of TP53 in these same samples. *TP53* is a vitally important tumour suppressor lost in many other cancers including HPV negative head and neck squamous cell carcinoma. It has roles in DNA repair, cell cycle control, initiation of apoptosis and senescence in response to telomere shortening. Loss of *TP53* in this cohort of patients, therefore, leads to a vicious cycle of further genomic instability and accumulated mutations.

Some oncogenic pathways are activated in both HPV positive and negative patients at an early stage, including: cell immortalisation by expression of *TERT*, inducing telomerase activity; activation of *cMET*, a receptor tyrosine kinase leading to cell growth; proliferation and motility; hyper-expression of *DDR2*, a tyrosine kinase that can be potentially targetable with dasatinib[228]; and promoter hypermethylation with loss of expression of *GAS7*, a tumour suppressor[255] whose loss has been found to lead to gefitinib resistance[234]. In addition, early truncal promoter hypermethylation and associated loss of expression was found for *RSPO2*, *ZNF135*, *EDNRB* and *GALNTL6*.

Expression of targetable oncogenic proteins such as *EGFR*, were only found in later subclones of both HPV positive and negative patients. Increased expression of *CTLA4*, *PDL1*, *IDO1*, *LAG3*, *TIM3* and *KIR* all took place recurrently across both HPV positive and negative patients. These proteins can cause a localised immune suppression against tumour antigens and are likely expressed after prolonged immune exposure to the growing tumour.

No copy number drivers were found to be recurrent in every patient throughout the cohort. However, copy gain of *cyclin D1* and *NFIB* along with the loss of *FHIT* were the most recurrent significant drivers found to directly affect expression of the *CIS* gene.

6.6 Clinical implications

The fact that HPV positive and negative patients have such different mutation, methylation and expression profiles suggests that HPV status will affect response to new therapeutics that are developed. It is therefore of paramount importance that the HPV status of all patients is determined, particularly for all clinical trials performed. An example of the potential clinical implication can be seen in head and neck squamous cell carcinomas, where HPV status confers a three-year overall survival advantage with systemic platinum-based chemotherapy (82.4% versus 57.1%, $p < 0.001$)[256].

Within penile cancer, as of 2018 there were only two types of therapeutics being considered across all trials in development or active recruitment on clinicaltrials.gov. One type is therapies that target tyrosine kinases with direct EGFR inhibitors such as cetuximab or tyrosine kinase inhibitors such as gefitinib. The other is immunotherapies that target immune checkpoints[252].

Recently presented work on the use of dacomitinib, an irreversible *EGFR* inhibitor, in a small 28 patient phase 2 trial, confirms an objective partial response rate of 32% with a 12 month median survival of 54.9%[257]. Based solely on this small cohort of patients, dacomitinib does not appear to dramatically improve the survival of these patients. One potential explanation for this poor response rate, as demonstrated for the first time in this thesis, is that *EGFR* is only subclonally mutated and expressed. Therefore, *EGFR* inhibition that only targets a portion of the patient's cancer burden, results in reduced efficacy and high relapse rates. However, it may have a role in a subset of patients who are *EGFR* positive or in those who are unable to receive standard platinum/taxane-based chemotherapy[258]. This thesis therefore demonstrates that targeting *EGFR* may not provide an enduring response for patients.

Inhibition of *DDR2* by dasatinib may prove a more effect therapeutic possibility than *EGFR* inhibition, as *DDR2* was found to be over-expressed in the trunk of the regional expression phylogenetic trees in the PenHet cohort. This indicates that it is a shared change in expression throughout all regions examined. It is therefore likely to be a relatively early change in the development of penile cancer in these patients, and as such may prove a more attractive therapeutic target.

Cyclin D was demonstrated to be amplified and over-expressed in an early truncal manner, with an average log 2-fold change of 3.96. Several CDK4/6 inhibitors have already been licensed in oestrogen receptor positive metastatic breast cancer – such as Palbociclib, Abemaciclib and Ribociclib[259]. This raises the possibility that the Cyclin D-CDK4/6 pathway could be targeted successfully in penile squamous cell carcinoma.

In earlier stage patients, there may be a role for HPV therapeutic vaccination where the oncogenic proteins E6 and E7 are targeted before they have a chance to cause irreversible genomic instability[260]. However, for late stage patients, such those in the PenHet cohort, the downstream and knock-on effects of long-term integrated HPV are likely to be irreversible even if the activity of E6 and E7 can be reduced.

There may also be a role for DNA methyltransferase inhibitors, such as azacytidine and decitabine[261], particularly in the HPV positive patients where *DNMT1* and *DNMT3A* are almost universally over-expressed.

6.7 Next steps and giving hope to metastatic penile cancer patients

The body of this work is based on the PenHet cohort, which consists of 48 samples from eight patients. To further progress this work, these findings need to be validated both in larger cohorts of patients and functionally to determine which targets can be used clinically.

To validate the findings uncovered in this thesis, larger scale powered experiments should be undertaken. If there were no financial constraints, whole genome sequencing and whole genome bisulfite sequencing should be undertaken serially for patients with both pre-malignant and malignant penile lesions. The background noise in the sequencing experiments could be reduced by increasing the purity of tumour samples. This could be achieved by carrying out single cell sequencing or at least laser dissecting tumour samples to improve purity.

The findings should be functionally validated. A new HPV negative penile cancer cell line has recently been developed, which could potentially be used to validate some of these findings[262]. Further work to create an HPV positive penile cell line would also be very useful. The cell lines should be interrogated for molecular aberrations found in the above experiments. Targeted agents could be tested against the cell line to demonstrate efficacy before human trials are commenced. Development of organoids may also prove to be valuable in the testing of new therapeutics against targets discovered.

The ideal next step would be to initiate a dedicated phase 3 trial comparing traditional chemotherapy with targeted therapies and immunotherapies. Based on the findings in this thesis, the targeted therapies most likely to be successful would include CDK 4/6 inhibitors and MET inhibition. However, for a rare disease such as penile cancer this would take considerable collaboration between many large centres and may not be desirable for pharmaceutical companies.

An alternative to dedicated phase 3 trials would be to encourage recruitment of patients with squamous penile cancer into ‘basket trials’ – trials where patients are selected based on the molecular biology of their cancer rather than the cancer type. This would enable penile cancer patients to join a study of multiple cancer types, thereby gaining access to new treatments. Use of biomarkers for increased expression of CDK4/6, MET or PDL1 may be beneficial in selecting patients who are most likely to benefit from a particular trial.

For patients who are not eligible for any trials, targeted therapies or immunotherapies may be granted on a compassionate use basis. Retrospective analysis of patients who receive therapies through compassionate use may still provide valuable evidence of efficacy.

6.8 Conclusions

Advanced penile squamous cell carcinoma is a heterogeneous disease at both the population and tumour levels. At the population level the greatest cause for the molecular differences is the presence of oncogenic HPV 16. Advanced penile cancer appears to be driven by both genetic and epigenetic drivers, causing a large number of aberrations. The main drivers common to all patients are *TERT*, *CDKN2A* and *cMET*. In addition, all tumours show an increased expression of immune checkpoint genes, including *CTLA4*, *PDL1*, *IDO1*, *LAG3*, *TIM3* and *KIR*. Irrespective of the presence or absence of HPV, penile cancer appears to exhibit a large tumour mutational load, which has been demonstrated to be a biomarker for immunotherapy success[11]. This combined with the high expression of immune checkpoint genes may suggest that these tumours will be amenable to treatment with novel immunotherapies.

I speculate that the main driver of ITH within HPV positive patients is over-expression of methyltransferases and APOBEC mutagenesis, caused by the integration of HPV at an early stage. In HPV negative patients, however, there does not appear to be one common pathway leading to ITH – besides the mutation and loss of expression of *TP53*. Instead, there appear to be many non-recurrent drivers in HPV negative patients. It will be interesting to observe over the coming decades how the rates of chronic HPV infection and HPV-related malignancies may fall following the introduction of HPV vaccination in young men. Due to the long lead time from initial infection with HPV and resulting malignancy many decades later, the prevalence of HPV-related penile cancer may not fall for many decades.

The data presented in the thesis has the potential to inform the direction of ongoing and new clinical trials for the treatment of penile cancer. This thesis shows that there is limited evidence that the currently investigated EGFR inhibitors would lead to enduring response in patients, as *EGFR* appears to be a subclonal aberration. Areas for further therapeutic investigation based on this body of work could include: immunotherapies, including T cell checkpoint inhibitors; targeted therapies of *cMET*; and methylation inhibitors. The next stage of research in this area should be to validate these findings in a functional setting using newly created penile cancer cell lines[263] or potentially penile organoids. The successful validation of these findings may provide hope and a new generation of pharmaco-therapeutics for patients with metastatic penile squamous cell carcinoma.

7 References

1. Rodney S, Feber A, Muneer A, Kelly JD. Textbook of Penile Cancer. In: Muneer A, Horenblas S, eds. *Molecular Biology of Penile Cancer*. 2nd ed. Springer; 2016:334.

2. Rodney S, Feber A, Arya M, Muneer A. Molecular markers in penile cancer. *Curr Probl Cancer*. 2015;39(3):137-145. doi:10.1016/j.currproblcancer.2015.03.005.

3. Rodney S, Muneer A. HPV and Penile Cancer: Perspectives on the Future Management of HPV-Positive Disease. *Oncology (Williston Park, NY)*. 2016;30(3):250-252.

4. Arya M, Kalsi J, Kelly J, Muneer A. Malignant and premalignant lesions of the penis. *BMJ*. 2013;346(mar06 1):f1149-f1149. doi:10.1136/bmj.f1149.

5. Moses KA, Winer A, Sfakianos JP, et al. Contemporary management of penile cancer: greater than 15 year MSKCC experience. *The Canadian journal of urology*. 2014;21(2):7201-7206.

6. Ornellas AA, Chin EWK, Nóbrega BLB, Wisnescky A, Koifman N, Quirino R. Surgical treatment of invasive squamous cell carcinoma of the penis: Brazilian National Cancer Institute long-term experience. - PubMed - NCBI. *J Surg Oncol*. 2008;97(6):487-495. doi:10.1002/jso.20980.

7. Rautava J, Syrjänen S. Biology of human papillomavirus infections in head and neck carcinogenesis. - PubMed - NCBI. *Head and Neck Pathology*. 2012;6(S1):3-15. doi:10.1007/s12105-012-0367-2.

8. Muñoz N. Human papillomavirus and cancer: the epidemiological evidence. *J Clin Virol*. 2000;19(1-2):1-5.

9. Ghittoni R, Accardi R, Hasan U, Gheit T, Sylla B, Tommasino M. The biological properties of E6 and E7 oncoproteins from human papillomaviruses. - PubMed - NCBI. *Virus Genes*. 2009;40(1):1-13. doi:10.1007/s11262-009-0412-8.

10. Münger K, Baldwin A, Edwards KM, et al. Mechanisms of human papillomavirus-induced oncogenesis. *J Virol*. 2004;78(21):11451-11460. doi:10.1128/JVI.78.21.11451-11460.2004.

11. Rizvi NA, Hellmann MD, Snyder A, et al. Mutational landscape determines sensitivity to PD-1 blockade in non–small cell lung cancer. *Science (New York, NY)*. 2015;348(6230):124-128. doi:10.1126/science.aaa1348.

12. Miralles-Guri C, Bruni L, Cubilla AL, Castellsague X, Bosch FX, de

Sanjose S. Human papillomavirus prevalence and type distribution in penile carcinoma. *J Clin Pathol*. 2009;62(10):870-878. doi:10.1136/jcp.2008.063149.

13. Moreno V, Bosch FX, Muñoz N, et al. Effect of oral contraceptives on risk of cervical cancer in women with human papillomavirus infection: the IARC multicentric case-control study. *The Lancet*. 2002;359(9312):1085-1092. doi:10.1016/S0140-6736(02)08150-3.

14. Stewart CL, Soria AM, Hamel PA. Integration of the pRB and p53 cell cycle control pathways. *J Neurooncol*. 2001;51(3):183-204.

15. Poetsch M, Schuart B-J, Schwesinger G, Kleist B, Protzel C. Screening of microsatellite markers in penile cancer reveals differences between metastatic and nonmetastatic carcinomas. *Mod Pathol*. 2007;20(10):1069-1077. doi:10.1038/modpathol.3800931.

16. Merlo A, Herman JG, Mao L, et al. 5′ CpG island methylation is associated with transcriptional silencing of the tumour suppressor p16/CDKN2/MTS1 in human cancers. *Nature Medicine*. 1995;1(7):686-692. doi:10.1038/nm0795-686.

17. Razin A, Cedar H. DNA methylation and gene expression. *Microbiological Reviews*. 1991;55(3):451-458.

18. Poetsch M, Hemmerich M, Kakies C, et al. Alterations in the tumor suppressor gene p16(INK4A) are associated with aggressive behavior of penile carcinomas. - PubMed - NCBI. *Virchows Arch*. 2010;458(2):221-229. doi:10.1007/s00428-010-1007-4.

19. Zhu Y, Zhou XY, Yao XD, Dai B, Ye DW. The prognostic significance of p53, Ki-67, epithelial cadherin and matrix metalloproteinase-9 in penile squamous cell carcinoma treated with surgery. - PubMed - NCBI. *BJU Int*. 2007;100(1):204-208. doi:10.1111/j.1464-410X.2007.06908.x.

20. Gunia S, Kakies C, Erbersdobler A, Hakenberg OW, Koch S, May M. Expression of p53, p21 and cyclin D1 in penile cancer: p53 predicts poor prognosis. *J Clin Pathol*. 2012;65(3):232-236. doi:10.1136/jclinpath-2011-200429.

21. MARTINS ACP, FARIA SM, COLOGNA AJ, SUAID HJ, TUCCI S. IMMUNOEXPRESSION OF p53 PROTEIN AND PROLIFERATING CELL NUCLEAR ANTIGEN IN PENILE CARCINOMA. *The Journal of Urology*. 2002;167(1):89-93. doi:10.1016/S0022-5347(05)65389-X.

22. Cubilla AL, Lloveras B, Alejo M, et al. Value of p16(INK)^4(a) in the pathology of invasive penile squamous cell carcinomas: A report of 202 cases. - PubMed - NCBI. *The American Journal of Surgical Pathology*. 2011;35(2):253-261. doi:10.1097/PAS.0b013e318203cdba.

23. Lopes A, Bezerra ALR, Pinto CAL, Serrano SV, de MellO CA, Villa LL. p53 as a new prognostic factor for lymph node metastasis in penile carcinoma: analysis of 82 patients treated with amputation and bilateral lymphadenectomy. *The Journal of Urology*. 2002;168(1):81-86.

24. Gunia S, Erbersdobler A, Hakenberg OW, Koch S, May M. p16INK4a is a Marker of Good Prognosis for Primary Invasive Penile Squamous Cell Carcinoma: A Multi-Institutional Study. *The Journal of Urology*. 2012;187(3):899-907. doi:10.1016/j.juro.2011.10.149.

25. Licitra L, Perrone F, Bossi P, et al. High-risk human papillomavirus affects prognosis in patients with surgically treated oropharyngeal squamous cell carcinoma. - PubMed - NCBI. *JCO*. 2006;24(36):5630-5636. doi:10.1200/JCO.2005.04.6136.

26. Kayes OJ, Loddo M, Patel N, et al. DNA Replication Licensing Factors and Aneuploidy Are Linked to Tumor Cell Cycle State and Clinical Outcome in Penile Carcinoma. *Clin Cancer Res*. 2009;15(23):7335-7344. doi:10.1158/1078-0432.CCR-09-0882.

27. May M, Burger M, Otto W, et al. Ki-67, mini-chromosome maintenance 2 protein (MCM2) and geminin have no independent prognostic relevance for cancer-specific survival in surgically treated squamous cell carcinoma of the penis. *BJU Int*. 2013;112(4):E383-E390. doi:10.1111/j.1464-410X.2012.11735.x.

28. Stankiewicz E, Ng M, Cuzick J, et al. The prognostic value of Ki-67 expression in penile squamous cell carcinoma. - PubMed - NCBI. *J Clin Pathol*. 2012;65(6):534-537. doi:10.1136/jclinpath-2011-200638.

29. Berdjis N, Meye A, Nippgen J, et al. Expression of Ki-67 in squamous cell carcinoma of the penis. - PubMed - NCBI. *BJU Int*. 2005;96(1):146-148. doi:10.1111/j.1464-410X.2005.05584.x.

30. PAPADOPOULOS O, BETSI E, TSAKISTOU G, et al. Expression of cyclin D1 and Ki-67 in squamous cell carcinoma of the penis. *Anticancer Res*. 2007;27(4B):2167-2174.

31. Guimarães GC, de Oliveira Leal ML, Sousa Madeira Campos R, et al. Do proliferating cell nuclear antigen and MIB-1/Ki-67 have prognostic value in penile squamous cell carcinoma? - PubMed - NCBI. *Urology*. 2007;70(1):137-142. doi:10.1016/j.urology.2007.03.003.

32. Andersson P, Kolaric A, Windahl T, Kirrander P, Söderkvist P, Karlsson MG. PIK3CA, HRAS and KRAS Gene Mutations in Human Penile Cancer. *The Journal of Urology*. 2008;179(5):2030-2034. doi:10.1016/j.juro.2007.12.040.

33. Gou H-F, Li X, Qiu M, et al. Epidermal Growth Factor Receptor (EGFR)-

RAS Signaling Pathway in Penile Squamous Cell Carcinoma. Viglietto G, ed. *PLoS ONE*. 2013;8(4):e62175. doi:10.1371/journal.pone.0062175.

34. Protzel C, Kakies C, Kleist B, Poetsch M, Giebel J. Down-regulation of the metastasis suppressor protein KAI1/CD82 correlates with occurrence of metastasis, prognosis and presence of HPV DNA in human penile squamous cell carcinoma. *Virchows Arch*. 2008;452(4):369-375. doi:10.1007/s00428-008-0590-0.

35. Breiteneder-Geleff S, Soleiman A, Kowalski H, et al. Angiosarcomas express mixed endothelial phenotypes of blood and lymphatic capillaries: podoplanin as a specific marker for lymphatic endothelium. - PubMed - NCBI. *The American Journal of Pathology*. 1999;154(2):385-394. doi:10.1016/S0002-9440(10)65285-6.

36. Li Y-Y, Zhou C-X, Gao Y. Podoplanin promotes the invasion of oral squamous cell carcinoma in coordination with MT1-MMP and Rho GTPases. *Am J Cancer Res*. 2015;5(2):514-529.

37. Minardi D, d'Anzeo G, Lucarini G, et al. D2-40 immunoreactivity in penile squamous cell carcinoma: a marker of aggressiveness. - PubMed - NCBI. *Human Pathology*. 2011;42(11):1596-1602. doi:10.1016/j.humpath.2010.12.020.

38. Thompson EW, Price JT. Mechanisms of tumour invasion and metastasis: emerging targets for therapy. *Expert Opinion on Therapeutic Targets*. 2005;6(2):217-233. doi:10.1517/14728222.6.2.217.

39. Sousa Madeira Campos R, Lopes A, Cardoso Guimarães G, Lopes Carvalho A, Augusto Soares F. E-cadherin, MMP-2, and MMP-9 as prognostic markers in penile cancer: analysis of 125 patients. - PubMed - NCBI. *Urology*. 2006;67(4):797-802. doi:10.1016/j.urology.2005.10.026.

40. Kudo Y, Siriwardena BSMS, Hatano H, Ogawa I, Takata T. Periostin: novel diagnostic and therapeutic target for cancer. *HISTOLOGY AND HISTOPATHOLOGY*. 2007;22(10):1167-1174. doi:10.14670/HH-22.1167.

41. Gunia S, Jain A, Koch S, et al. Periostin expression correlates with pT-stage, grading and tumour size, and independently predicts cancer-specific survival in surgically treated penile squamous cell carcinomas. *J Clin Pathol*. 2013;66(4):297-301. doi:10.1136/jclinpath-2012-201262.

42. Han Z-D, He H-C, Bi X-C, et al. Expression and Clinical Significance of CD147 in Genitourinary Carcinomas. *Journal of Surgical Research*. 2010;160(2):260-267. doi:10.1016/j.jss.2008.11.838.

43. Soares FA, da Cunha IW, Guimarães GC, Nonogaki S, Campos RSM, Lopes A. The expression of metaloproteinases-2 and -9 is different according to the patterns of growth and invasion in squamous cell

carcinoma of the penis. - PubMed - NCBI. *Virchows Arch*. 2006;449(6):637-646. doi:10.1007/s00428-006-0299-x.

44. Feber A, Arya M, de Winter P, et al. Epigenetics markers of metastasis and HPV-induced tumorigenesis in penile cancer. *Clin Cancer Res*. 2015;21(5):1196-1206. doi:10.1158/1078-0432.CCR-14-1656.

45. Kuasne H, Cólus IM de S, Busso AF, et al. Genome-wide methylation and transcriptome analysis in penile carcinoma: uncovering new molecular markers. *Clin Epigenetics*. 2015;7(1):30. doi:10.1186/s13148-015-0082-4.

46. Ferreux E, Lont AP, Horenblas S, et al. Evidence for at least three alternative mechanisms targeting the p16INK4A/cyclin D/Rb pathway in penile carcinoma, one of which is mediated by high-risk human papillomavirus. *J Pathol*. 2003;201(1):109-118. doi:10.1002/path.1394.

47. Busso-Lopes AF, Marchi FA, Kuasne H, et al. Genomic Profiling of Human Penile Carcinoma Predicts Worse Prognosis and Survival. *Cancer Prevention Research*. 2015;8(2):149-156. doi:10.1158/1940-6207.CAPR-14-0284.

48. Dang CV. MYC on the Path to Cancer. *Cell*. 2012;149(1):22-35. doi:10.1016/j.cell.2012.03.003.

49. Ashkenazi A. Targeting death and decoy receptors of the tumour-necrosis factor superfamily. - PubMed - NCBI. *Nat Rev Cancer*. 2002;2(6):420-430. doi:10.1038/nrc821.

50. Kato H, Torigoe T. Radioimmunoassay for tumor antigen of human cervical squamous cell carcinoma. *Cancer*. 1977;40(4):1621-1628.

51. Touloupidis S, Zisimopoulos A, Giannakopoulos S, Papatsoris AG, Kalaitzis C, Thanos A. Clinical usage of the squamous cell carcinoma antigen in patients with penile cancer. *International Journal of Urology*. 2007;14(2):174-176. doi:10.1111/j.1442-2042.2007.01694.x.

52. Minardi D, Lucarini G, Filosa A, et al. Prognostic value of CD44 expression in penile squamous cell carcinoma: a pilot study. *Cell Oncol*. 2012;35(5):377-384. doi:10.1007/s13402-012-0098-0.

53. Jakobsen JK, Krarup KP, Sommer P, et al. DaPeCa-1: diagnostic accuracy of sentinel lymph node biopsy in 222 patients with penile cancer at four tertiary referral centres – a national study from Denmark. *BJU Int*. 2016;117(2):235-243. doi:10.1111/bju.13127.

54. Hakenberg OW, Compérat EM, Minhas S, Necchi A, Protzel C, Watkin N. EAU Guidelines on Penile Cancer: 2014 Update. *European Urology*. 2015;67(1):142-150. doi:10.1016/j.eururo.2014.10.017.

55. Van Poppel H, Watkin NA, Osanto S, Moonen L, Horwich A, Kataja V.

Penile cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol*. 2013;24(suppl_6):vi115-vi124. doi:10.1093/annonc/mdt286.

56. Pizzocaro G, Piva L, Nicolai N. [Treatment of lymphatic metastasis of squamous cell carcinoma of the penis: experience at the National Tumor Institute of Milan]. *Arch Ital Urol Androl*. 1996;68(3):169-172.

57. Pizzocaro G, Piva L. Adjuvant and Neoadjuvant Vincristine, Bleomycin, And Methotrexate for Inguinal Metastases from Squamous Cell Carcinoma of the Penis. *Acta Oncologica*. 2009;27(6):823-824. doi:10.3109/02841868809094366.

58. Nowell PC. The clonal evolution of tumor cell populations. *Science*. 1976;194(4260):23-28.

59. Williams MJ, Werner B, Barnes CP, Graham TA, Sottoriva A. Identification of neutral tumor evolution across cancer types. *Nat Genet*. 2016;48(3):238-244. doi:10.1038/ng.3489.

60. Inda MDM, Bonavia R, Mukasa A, et al. Tumor heterogeneity is an active process maintained by a mutant EGFR-induced cytokine circuit in glioblastoma. *Genes & Development*. 2010;24(16):1731-1745. doi:10.1101/gad.1890510.

61. Notta F, Chan-Seng-Yue M, Lemire M, et al. A renewed model of pancreatic cancer evolution based on genomic rearrangement patterns. *Nature*. 2016;538(7625):378-382. doi:10.1038/nature19823.

62. Baca SC, Prandi D, Lawrence MS, et al. Punctuated Evolution of Prostate Cancer Genomes. *Cell*. 2013;153(3):666-677. doi:10.1016/j.cell.2013.03.021.

63. Marusyk A, Polyak K. Tumor heterogeneity: Causes and consequences. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*. 2010;1805(1):105-117. doi:10.1016/j.bbcan.2009.11.002.

64. Davis A, Gao R, Navin N. Tumor Evolution: Linear, Branching, Neutral or Punctuated? *BBA - Reviews on Cancer*. January 2017:1-58. doi:10.1016/j.bbcan.2017.01.003.

65. Yap KL, Kiyotani K, Tamura K, et al. Whole-Exome Sequencing of Muscle-Invasive Bladder Cancer Identifies Recurrent Mutations of UNC5C and Prognostic Importance of DNA Repair Gene Mutations on Survival. *Clin Cancer Res*. 2014;20(24):6605-6617. doi:10.1158/1078-0432.CCR-14-0257.

66. McGranahan N, Swanton C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell*. 2017;168(4):613-628.

doi:10.1016/j.cell.2017.01.018.

67. Mroz EA, Tward AD, Pickering CR, Myers JN, Ferris RL, Rocco JW. High intratumor genetic heterogeneity is related to worse outcome in patients with head and neck squamous cell carcinoma. *Cancer*. 2013;119(16):3034-3042. doi:10.1002/cncr.28150.

68. MIYAUCHI T, YAGUCHI T, KAWAKAMI Y. Inter-patient and Intra-tumor Heterogeneity in the Sensitivity to Tumor-targeted Immunity in Colorectal Cancer. *JpnJclinImmun*. 2017;40(1):54-59. doi:10.2177/jsci.40.54.

69. Fisher R, Pusztai L, Swanton C. Cancer heterogeneity: implications for targeted therapeutics. *British Journal of Cancer*. 2013;108(3):479-485. doi:10.1038/bjc.2012.581.

70. Hirsch FR, Ottesen G, Pødenphant J, Olsen J. Tumor heterogeneity in lung cancer based on light microscopic features. A retrospective study of a consecutive series of 200 patients, treated surgically. *Virchows Arch A Pathol Anat Histopathol*. 1983;402(2):147-153.

71. Inukai M, Toyooka S, Ito S, et al. Presence of Epidermal Growth Factor ReceptorGene T790M Mutation as a Minor Clone in Non–Small Cell Lung Cancer. *Cancer Res*. 2006;66(16):7854-7858. doi:10.1158/0008-5472.CAN-06-1951.

72. Campbell PJ, Pleasance ED, Stephens PJ, et al. Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *PNAS*. 2008;105(35):13081-13086. doi:10.1073/pnas.0801523105.

73. Carter P, Presta L, Gorman CM, et al. Humanization of an anti-p185HER2 antibody for human cancer therapy. *PNAS*. 1992;89(10):4285-4289.

74. Schumacher TN, Schreiber RD. Neoantigens in cancer immunotherapy. *Science*. 2015;348(6230):69-74. doi:10.1126/science.aaa4971.

75. Jamal-Hanjani M, Quezada SA, Larkin J, Swanton C. Translational Implications of Tumor Heterogeneity. *Clin Cancer Res*. 2015;21(6):1258-1266. doi:10.1158/1078-0432.CCR-14-1429.

76. McGranahan N, Favero F, de Bruin EC, Birkbak NJ, Szallasi Z, Swanton C. Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Science Translational Medicine*. 2015;7(283):283ra54-283ra54. doi:10.1126/scitranslmed.aaa1408.

77. McGranahan N, Furness AJS, Rosenthal R, et al. Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science*. 2016;351(6280):1463-1469. doi:10.1126/science.aaf1490.

78. Ott PA, Hodi FS, Robert C. CTLA-4 and PD-1/PD-L1 blockade: new

immunotherapeutic modalities with durable clinical benefit in melanoma patients. - PubMed - NCBI. *Clin Cancer Res*. 2013;19(19):5300-5309. doi:10.1158/1078-0432.CCR-13-0143.

79. Langer CJ. Emerging Immunotherapies in the Treatment of Non–small Cell Lung Cancer (NSCLC): The Role of Immune Checkpoint Inhibitors. *American Journal of Clinical Oncology*. 2015;38(4):422-430. doi:10.1097/COC.0000000000000059.

80. Powles T, Eder JP, Fine GD, et al. MPDL3280A (anti-PD-L1) treatment leads to clinical activity in metastatic bladder cancer. - PubMed - NCBI. *Nature*. 2014;515(7528):558-562. doi:10.1038/nature13904.

81. Mellman I, Coukos G, Dranoff G. Cancer immunotherapy comes of age. *Nature*. 2011;480(7378):480-489. doi:10.1038/nature10673.

82. Alves JM, Prieto T, Posada D. Multiregional Tumor Trees Are Not Phylogenies. *Trends in Cancer*. 2017;3(8):546-550. doi:10.1016/j.trecan.2017.06.004.

83. Jiang Y, Qiu Y, Minn AJ, Zhang NR. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. - PubMed - NCBI. *PNAS*. 2016;113(37):E5528-E5537.

84. Roth A, Khattra J, Yap D, et al. PyClone: statistical inference of clonal population structure in cancer. *Nat Methods*. 2014;11(4):396-398. doi:10.1038/nmeth.2883.

85. Nik-Zainal S, Van Loo P, Wedge DC, et al. The life history of 21 breast cancers. - PubMed - NCBI. *Cell*. 2012;149(5):994-1007. doi:10.1016/j.cell.2012.04.023.

86. Jiao W, Vembu S, Deshwar AG, Stein L, Morris Q. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinformatics*. 2014;15(1):35. doi:10.1186/1471-2105-15-35.

87. Qi Y, Pradhan D, El-Kebir M. Implications of non-uniqueness in phylogenetic deconvolution of bulk DNA samples of tumors. *Algorithms Mol Biol*. 2019;14(1):19. doi:10.1186/s13015-019-0155-6.

88. Solsona E, Algaba F, Horenblas S, Pizzocaro G, Windahl T. EAU Guidelines on Penile Cancer. *European Urology*. 2004;46(1):1-8. doi:10.1016/j.eururo.2004.03.007.

89. Marchi FA, Martins DC, Barros-Filho MC, et al. Multidimensional integrative analysis uncovers driver candidates and biomarkers in penile carcinoma. *Scientific Reports*. 2017;7(1):9379. doi:10.1038/s41598-017-06659-1.

90. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. bioinformatics.babraham.ac.uk. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed November 8, 2018.

91. Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*. October 2015:btv566. doi:10.1093/bioinformatics/btv566.

92. Zheng-Bradley X, Streeter I, Fairley S, Richardson D, Clarke L, Flicek P. Alignment of 1000 Genomes Project reads to reference assembly GRCh38. *Gigascience*. 2017;6(7):68. doi:10.1093/gigascience/gix038.

93. Chandrani P, Kulkarni V, Iyer P, et al. NGS-based approach to determine the presence of HPV and their sites of integration in human cancer genome. *British Journal of Cancer*. 2015;112(12):1958-1965. doi:10.1038/bjc.2015.121.

94. Niu B, Ye K, Zhang Q, et al. MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. *Bioinformatics*. 2013;30(7):1015-1016. doi:10.1093/bioinformatics/btt755.

95. Commissioner OOT. Press Announcements - FDA approves first cancer treatment for any solid tumor with a specific genetic feature.

96. Boyiadzis MM, Kirkwood JM, Marshall JL, Pritchard CC, Azad NS, Gulley JL. Significance and implications of FDA approval of pembrolizumab for biomarker-defined disease. *j immunotherapy cancer*. 2018;6(1):1974. doi:10.1186/s40425-018-0342-x.

97. Picard Tools - By Broad Institute. broadinstitute.github.io. http://broadinstitute.github.io/picard/. Accessed September 4, 2018.

98. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics*. 2015;31(12):2032-2034. doi:10.1093/bioinformatics/btv098.

99. SeqCap EZ Exome Probes v3.0 Target Enrichment Probes - Roche Sequencing Solutions. sequencing.roche.com. /content/rochesequence/en/products-solutions/by-category/target-enrichment/hybridization/seqcap-ez-exome-v3-kit.html. Accessed September 4, 2018.

100. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):e164-e164. doi:10.1093/nar/gkq603.

101. Kim H, Zheng S, Amini SS, et al. Whole-genome and multisector exome sequencing of primary and post-treatment glioblastoma reveals patterns

of tumor evolution. *Genome Res*. 2015;25(3):gr.180612.114-gr.180612.327. doi:10.1101/gr.180612.114.

102. Forbes SA, Bhamra G, Bamford S, et al. The Catalogue of Somatic Mutations in Cancer (COSMIC). *Current protocols in human genetics / editorial board, Jonathan L Haines [et al]*. 2008;CHAPTER:Unit. doi:10.1002/0471142905.hg1011s57.

103. Favero F, Joshi T, Marquard AM, et al. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann Oncol*. 2014;26(1):64-70. doi:10.1093/annonc/mdu479.

104. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. *Nature*. 2013;500(7463):415-421. doi:10.1038/nature12477.

105. Rosenthal R, McGranahan N, Herrero J, Taylor BS, Swanton C. deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol*. 2016;17(1):31. doi:10.1186/s13059-016-0893-4.

106. Wallace NA, Münger K. The curious case of APOBEC3 activation by cancer-associated human papillomaviruses. Coyne CB, ed. *PLOS Pathogens*. 2018;14(1):e1006717. doi:10.1371/journal.ppat.1006717.

107. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics*. 2010;11(1):367. doi:10.1186/1471-2105-11-367.

108. Landau DA, Carter SL, Stojanov P, et al. Evolution and Impact of Subclonal Mutations in Chronic Lymphocytic Leukemia. *Cell*. 2013;152(4):714-726. doi:10.1016/j.cell.2013.01.019.

109. Hellmann MD, Nathanson T, Rizvi H, et al. Genomic Features of Response to Combination Immunotherapy in Patients with Advanced Non-Small-Cell Lung Cancer. *Cancer cell*. 2018;33(5):843-852.e844. doi:10.1016/j.ccell.2018.03.018.

110. Revell LJ, Chamberlain SA. Rphylip: an R interface for PHYLIP. Freckleton R, ed. *Methods in Ecology and Evolution*. 2014;5(9):976-981. doi:10.1111/2041-210X.12233.

111. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*. 2004;20(2):289-290. doi:10.1093/bioinformatics/btg412.

112. Flensburg C, Sargeant T, Oshlack A, Majewski I. SuperFreq: Integrated mutation detection and clonal tracking in cancer. *bioRxiv*. July 2018:380097. doi:10.1101/380097.

113. Miller CA, McMichael J, Dang HX, et al. Visualizing tumor evolution with the fishplot package for R. *BMC Genomics 2016 17:1*. 2016;17(1):880. doi:10.1186/s12864-016-3195-z.

114. Van Loo P, Nordgard SH, Lingjaerde OC, et al. Allele-specific copy number analysis of tumors. *PNAS*. 2010;107(39):16910-16915. doi:10.1073/pnas.1009843107.

115. Aryee MJ, Jaffe AE, Corrada-Bravo H, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*. 2014;30(10):1363-1369. doi:10.1093/bioinformatics/btu049.

116. Zhou W, Laird PW, Shen H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res*. 2017;45(4):e22-e22. doi:10.1093/nar/gkw967.

117. Reinius LE, Acevedo N, Joerink M, et al. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. - PubMed - NCBI. Ting AH, ed. *PLoS ONE*. 2012;7(7):e41361. doi:10.1371/journal.pone.0041361.

118. Buja A, Swayne DF, Littman ML, Dean N, Hofmann H, Chen L. Data Visualization With Multidimensional Scaling. *Journal of Computational and Graphical Statistics*. 2012;17(2):444-472. doi:10.1198/106186008X318440.

119. Du P, Zhang X, Huang C-C, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*. 2010;11(1):587. doi:10.1186/1471-2105-11-587.

120. Coolen MW, Stirzaker C, Song JZ, et al. Consolidation of the cancer genome into domains of repressive chromatin by long-range epigenetic silencing (LRES) reduces transcriptional plasticity. - PubMed - NCBI. *Nat Cell Biol*. 2010;128(3):683–246. doi:10.1038/ncb2023.

121. Peters TJ, Buckley MJ, Statham AL, et al. De novo identification of differentially methylated regions in the human genome. *Epigenetics & Chromatin 2015 8:1*. 2015;8(1):6. doi:10.1186/1756-8935-8-6.

122. Fortin J-P, Triche TJ Jr, Hansen KD. Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. - PubMed - NCBI. *Bioinformatics*. 2016;11(4):btw691–560. doi:10.1093/bioinformatics/btw691.

123. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*. 2016;34(5):525-527. doi:10.1038/nbt.3519.

124. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. - PubMed - NCBI. *F1000Res*. 2016;4:1521. doi:10.12688/f1000research.7563.2.

125. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. - PubMed - NCBI. *Genome Biol*. 2014;15(12):31. doi:10.1186/s13059-014-0550-8.

126. Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol*. 2017;18(1):1960. doi:10.1186/s13059-017-1349-1.

127. Ding J, McConechy MK, Horlings HM, et al. Systematic analysis of somatic mutations impacting gene expression in 12 tumour types. *Nat Comms*. 2015;6(1):1546. doi:10.1038/ncomms9554.

128. Lai Y-P, Wang L-B, Wang W-A, et al. iGC—an integrated analysis package of gene expression and copy number alteration. *BMC Bioinformatics*. 2017;18(1):457. doi:10.1186/s12859-016-1438-2.

129. Gevaert O. MethylMix: an R package for identifying DNA methylation-driven genes. - PubMed - NCBI. *Bioinformatics*. 2015;31(11):1839-1841. doi:10.1093/bioinformatics/btv020.

130. Cedoz P-L, Prunello M, Brennan K, Gevaert O. MethylMix 2.0: an R package for identifying DNA methylation genes. - PubMed - NCBI. *Bioinformatics*. 2018;30:1363. doi:10.1093/bioinformatics/bty156.

131. Dietz EJ. Permutation Tests for Association Between Two Distance Matrices. *Syst Biol*. 1983;32(1):21-26. doi:10.1093/sysbio/32.1.21.

132. Phipson B, Maksimovic J, Oshlack A. missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. - PubMed - NCBI. *Bioinformatics*. 2015;32(2):btv560–288.

133. Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. - PubMed - NCBI. *Genome Biol*. 2010;11(2):R14. doi:10.1186/gb-2010-11-2-r14.

134. Luo W, Friedman MS, Shedden K, Hankenson KD, Woolf PJ. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics*. 2009;10(1):161. doi:10.1186/1471-2105-10-161.

135. Luo W, Brouwer C. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics*. 2013;29(14):1830-1831. doi:10.1093/bioinformatics/btt285.

136. Gerlinger M, Rowan AJ, Horswell S, et al. Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *N Engl J Med*. 2012;366(10):883-892. doi:10.1056/NEJMoa1113205.

137. Feber A, Worth DC, Chakravarthy A, et al. CSN1 Somatic Mutations in Penile Squamous Cell Carcinoma. *Cancer Res*. 2016;76(16):4720-4727. doi:10.1158/0008-5472.CAN-15-3134.

138. Carter SL, Cibulskis K, Helman E, et al. Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol*. 2012;30(5):413-421. doi:10.1038/nbt.2203.

139. Byrne EH, Fisher DE. Immune and molecular correlates in melanoma treated with immune checkpoint blockade. - PubMed - NCBI. *Cancer*. 2017;123(S11):2143-2153. doi:10.1002/cncr.30444.

140. Brown SD, Warren RL, Gibb EA, et al. Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival. *Genome Res*. 2014;24(5):743-750. doi:10.1101/gr.165985.113.

141. Chalmers ZR, Connelly CF, Fabrizio D, et al. Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Medicine 2017 9:1*. 2017;9(1):34. doi:10.1186/s13073-017-0424-2.

142. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. *Nat Comms*. 2015;6(1):239. doi:10.1038/ncomms9971.

143. Le DT, Uram JN, Wang H, et al. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *http://dxdoiorg/101056/NEJMoa1500596*. 2015;372(26):2509-2520. doi:10.1056/NEJMoa1500596.

144. Sudenga SL, Ingles DJ, Pierce Campbell CM, et al. Genital Human Papillomavirus Infection Progression to External Genital Lesions: The HIM Study. - PubMed - NCBI. *European Urology*. 2016;69(1):166-173. doi:10.1016/j.eururo.2015.05.032.

145. McBride AA, Warburton A. The role of integration in oncogenic progression of HPV-associated cancers. Spindler KR, ed. *PLOS Pathogens*. 2017;13(4):e1006211. doi:10.1371/journal.ppat.1006211.

146. Doolittle-Hall J, Cunningham Glasspoole D, Seaman W, Webster-Cyriaque J. Meta-Analysis of DNA Tumor-Viral Integration Site Selection Indicates a Role for Repeats, Gene Expression and Epigenetics. *Cancers*. 2015;7(4):2217-2235. doi:10.3390/cancers7040887.

147. Liu Y, Lu Z, Xu R, Ke Y. Comprehensive mapping of the human papillomavirus (HPV) DNA integration sites in cervical carcinomas by HPV capture technology. *Oncotarget*. 2016;7(5):5852-5864. doi:10.18632/oncotarget.6809.

148. Peter M, Rosty C, Couturier J, Radvanyi F, Teshima H, Sastre-Garau X. MYC activation associated with the integration of HPV DNA at the MYC

locus in genital tumors. *Oncogene*. 2006;25(44):5985-5993. doi:10.1038/sj.onc.1209625.

149. Mao Y, Liu J, Zhang D, Li B. MiR-1290 promotes cancer progression by targeting nuclear factor I/X(NFIX) in esophageal squamous cell carcinoma (ESCC). *Biomedicine & Pharmacotherapy*. 2015;76:82-93. doi:10.1016/j.biopha.2015.10.005.

150. Rahman NIA, Abdul Murad NA, Mollah MM, Jamal R, Harun R. NFIX as a Master Regulator for Lung Cancer Progression. *Front Pharmacol*. 2017;8:517. doi:10.3389/fphar.2017.00540.

151. Hu Y, Guo X, Wang J, et al. A novel microRNA identified in hepatocellular carcinomas is responsive to LEF1 and facilitates proliferation and epithelial-mesenchymal transition ... - PubMed - NCBI. *Oncogenesis*. 2018;7(2):1245. doi:10.1038/s41389-017-0010-x.

152. Zack TI, Schumacher SE, Carter SL, et al. Pan-cancer patterns of somatic copy number alteration. *Nat Genet*. 2013;45(10):1134-1140. doi:10.1038/ng.2760.

153. Xue W, Kitzing T, Roessler S, et al. A cluster of cooperating tumor-suppressor gene candidates in chromosomal deletions. *PNAS*. 2012;109(21):8212-8217. doi:10.1073/pnas.1206062109.

154. Akervall JA, Michalides RJ, Mineta H, et al. Amplification of cyclin D1 in squamous cell carcinoma of the head and neck and the prognostic value of chromosomal abnormalities and cyclin D1 overexpression. *Cancer*. 1997;79(2):380-389.

155. Miyamoto R, Uzawa N, Nagaoka S, Hirata Y, Amagasa T. Prognostic significance of cyclin D1 amplification and overexpression in oral squamous cell carcinomas. *Oral Oncology*. 2003;39(6):610-618.

156. Borg A, Baldetorp B, Fernö M, Killander D, Olsson H, Sigurdsson H. ERBB2 amplification in breast cancer with a high rate of proliferation. *Oncogene*. 1991;6(1):137-143.

157. Carthon BC, Ng CS, Pettaway CA, Pagliaro LC. Epidermal growth factor receptor-targeted therapy in locally advanced or metastatic squamous cell carcinoma of the penis. *BJU Int*. 2014;113(6):871-877. doi:10.1111/bju.12450.

158. Momand J, Jung D, Wilczynski S, Niland J. The MDM2 gene amplification database. *Nucleic Acids Res*. 1998;26(15):3453-3459. doi:10.1093/nar/26.15.3453.

159. Mariusz L Hartman MC. MITF in melanoma: mechanisms behind its expression and activity. *Cellular and Molecular Life Sciences*.

2015;72(7):1249-1260. doi:10.1007/s00018-014-1791-0.

160. Koon HB MD, Ippolito GC PhD, Banham AH PhD, Tucker PW PhD. FOXP1: a potential therapeutic target in cancer. *Expert Opinion on Therapeutic Targets*. 2007;11(7):955-965. doi:10.1517/14728222.11.7.955.

161. Semenova EA, Kwon M-C, Monkhorst K, et al. Transcription Factor NFIB Is a Driver of Small Cell Lung Cancer Progression in Mice and Marks Metastatic Disease in Patients. *CellReports*. 2016;16(3):631-643. doi:10.1016/j.celrep.2016.06.020.

162. Yang ZQ, Imoto I, Pimkhaokham A, et al. A Novel Amplicon at 9p23-24 in Squamous Cell Carcinoma of the Esophagus That Lies Proximal to GASC1 and Harbors NFIB. *Japanese Journal of Cancer Research : Gann*. 2001;92(4):423-428. doi:10.1111/j.1349-7006.2001.tb01112.x.

163. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. - PubMed - NCBI. *Nat Genet*. 2006;38(2):209-213. doi:10.1038/ng1706.

164. Warren C, Westrich J, Doorslaer K, Pyeon D. Roles of APOBEC3A and APOBEC3B in Human Papillomavirus Infection and Disease Progression. *Viruses*. 2017;9(8):233. doi:10.3390/v9080233.

165. Neddermann P, Gallinari P, Lettieri T, et al. Cloning and expression of human G/T mismatch-specific thymine-DNA glycosylase. *J Biol Chem*. 1996;271(22):12767-12774. doi:10.1074/jbc.271.22.12767.

166. Nichols AC. High frequency of activating PIK3CA mutations in human papillomavirus-positive oropharyngeal cancer. - PubMed - NCBI. *JAMA Otolaryngol Head Neck Surg*. 2013;139(6):617-622. doi:10.1001/jamaoto.2013.3210.

167. Rocco JW. Mutant Allele Tumor Heterogeneity (MATH) and Head and Neck Squamous Cell Carcinoma. *Head and Neck Pathology*. 2015;9(1):1-5. doi:10.1007/s12105-015-0617-1.

168. de Bruin EC, McGranahan N, Mitter R, et al. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science*. 2014;346(6206):251-256. doi:10.1126/science.1253462.

169. Forbes SA, Beare D, Boutselakis H, et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res*. 2017;45(D1):D777-D783. doi:10.1093/nar/gkw1121.

170. Henderson S, Chakravarthy A, Su X, Boshoff C, Fenton TR. APOBEC-Mediated Cytosine Deamination Links PIK3CA Helical Domain Mutations

to Human Papillomavirus-Driven Tumor Development. *CellReports*. 2014;7(6):1833-1841. doi:10.1016/j.celrep.2014.05.012.

171. Dacomitinib (PF-00299804) in Advanced/Metastatic Squamous Cell Carcinoma of the Penis - Full Text View - ClinicalTrials.gov.

172. Jamal-Hanjani M, Wilson GA, McGranahan N, et al. Tracking the Evolution of Non–Small-Cell Lung Cancer. *http://dxdoiorg/101056/NEJMoa1616288*. 2017;376(22):2109-2121. doi:10.1056/NEJMoa1616288.

173. Van Allen EM, Miao D, Schilling B, et al. Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science (New York, NY)*. 2015;350(6257):207-211. doi:10.1126/science.aad0095.

174. Daud AI, Loo K, Pauli ML, et al. Tumor immune profiling predicts response to anti–PD-1 therapy in human melanoma. *Journal of Clinical Investigation*. 2016;126(9):3447-3452. doi:10.1172/JCI87324.

175. Morris LGT, Riaz N, Desrichard A, et al. Pan-cancer analysis of intratumor heterogeneity as a prognostic determinant of survival. *Oncotarget*. 2016;7(9):10051-10063. doi:10.18632/oncotarget.7067.

176. Feinberg AP, Irizarry RA. Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *PNAS*. 2010;107(suppl_1):1757-1764. doi:10.1073/pnas.0906183107.

177. Jones PA. The role of DNA methylation in mammalian epigenetics. - PubMed - NCBI. *Science*. 2001;293(5532):1068-1070. doi:10.1126/science.1063852.

178. Egger G, Liang G, Aparicio A, Jones PA. Epigenetics in human disease and prospects for epigenetic therapy. - PubMed - NCBI. *Nature*. 2004;429(6990):457-463. doi:10.1038/nature02625.

179. Houseman E, Accomando WP, Koestler DC, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. - PubMed - NCBI. *BMC Bioinformatics*. 2012;13(1):86. doi:10.1186/1471-2105-13-86.

180. Onuchic V, Hartmaier RJ, Boone DN, et al. Epigenomic Deconvolution of Breast Tumors Reveals Metabolic Coupling between Constituent Cell Types. - PubMed - NCBI. *Cell Reports*. 2016;17(8):2075-2086.

181. Garcia-Gomez A, Rodríguez-Ubreva J, Ballestar E. Epigenetic interplay between immune, stromal and cancer cells in the tumor microenvironment. *Clinical Immunology*. March 2018. doi:10.1016/j.clim.2018.02.013.

182. Baylin SB. DNA methylation and gene silencing in cancer. - PubMed - NCBI. *Nat Rev Clin Oncol*. 2005;2(S1):S4-S11. doi:10.1038/ncponc0354.

183. Rechache NS, Wang Y, Stevenson HS, et al. DNA Methylation Profiling Identifies Global Methylation Differences and Markers of Adrenocortical Tumors. *The Journal of Clinical Endocrinology and Metabolism*. 2012;97(6):E1004-E1013. doi:10.1210/jc.2011-3298.

184. Bibikova M, Barnes B, Tsan C, et al. High density DNA methylation array with single CpG site resolution. *Genomics*. 2011;98(4):288-295. doi:10.1016/j.ygeno.2011.07.007.

185. Huang R-L, Chang C-C, Su P-H, et al. Methylomic Analysis Identifies Frequent DNA Methylation of Zinc Finger Protein 582 (ZNF582) in Cervical Neoplasms. Peng D, ed. *PLoS ONE*. 2012;7(7):e41060. doi:10.1371/journal.pone.0041060.

186. Shen-Gunther J, Wang C-M, Poage GM, et al. Molecular Pap smear: HPV genotype and DNA methylation of ADCY8 , CDH8 , and ZNF582 as an integrated biomarker for high-grade cervical cytology. *Clin Epigenetics*. 2016;8(1):96. doi:10.1186/s13148-016-0263-9.

187. Cheng S-J, Chang C-F, Lee J-J, et al. Hypermethylated ZNF582 and PAX1 are effective biomarkers for detection of oral dysplasia and oral cancer. - PubMed - NCBI. *Oral Oncology*. 2016;62:34-43. doi:10.1016/j.oraloncology.2016.09.007.

188. van Kempen PM, Noorlag R, Braunius WW, Stegeman I, Willems SM, Grolman W. Differences in methylation profiles between HPV-positive and HPV-negative oropharynx squamous cell carcinoma: A systematic review. *Epigenetics*. 2014;9(2):194-203. doi:10.4161/epi.26881.

189. Colacino JA, Dolinoy DC, Duffy SA, et al. Comprehensive Analysis of DNA Methylation in Head and Neck Squamous Cell Carcinoma Indicates Differences by Survival and Clinicopathologic Characteristics. Ramqvist T, ed. *PLoS ONE*. 2013;8(1):e54742. doi:10.1371/journal.pone.0054742.

190. Yang H-J. Aberrant DNA methylation in cervical Carcinogenesis. *Chinese Journal of Cancer*. 2013;32(1):42-48. doi:10.5732/cjc.012.10033.

191. van Kempen PM, Noorlag R, Braunius WW, Stegeman I, Willems SM, Grolman W. Differences in methylation profiles between HPV-positive and HPV-negative oropharynx squamous cell carcinoma. *Epigenetics*. October 2013. doi:10.4161/epi.26881.

192. Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. *Nat Rev Genet*. 2011;12(8):529-541. doi:10.1038/nrg3000.

193. Brocks D, Assenov Y, Minner S, et al. Intratumor DNA Methylation Heterogeneity Reflects Clonal Evolution in Aggressive Prostate Cancer. *CellReports*. 2014;8(3):798-806. doi:10.1016/j.celrep.2014.06.053.

194. Gerlinger M, Horswell S, Larkin J, et al. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nature Publishing Group*. 2014;46(3):225-233. doi:10.1038/ng.2891.

195. Hao J-J, Lin D-C, Dinh HQ, et al. Spatial intratumoral heterogeneity and temporal clonal evolution in esophageal squamous cell carcinoma. *Nat Genet*. 2016;48(12):1500-1507. doi:10.1038/ng.3683.

196. Wilhelm F, Simon E, Böger C, Behrens H-M, Krüger S, Röcken C. Novel Insights into Gastric Cancer: Methylation of R-spondins and Regulation of LGR5 by SP1. *Mol Cancer Res*. 2017;15(6):776-785. doi:10.1158/1541-7786.MCR-16-0472.

197. Wu C, Qiu S, Lu L, et al. RSPO2–LGR5 signaling has tumour-suppressive activity in colorectal cancer. *Nat Comms*. 2014;5:43. doi:10.1038/ncomms4149.

198. Pestana A, Vinagre J, Sobrinho-Simões M, Soares P. TERT biology and function in cancer: beyond immortalisation. - PubMed - NCBI. *Journal of Molecular Endocrinology*. 2017;58(2):R129-R146. doi:10.1530/JME-16-0195.

199. Gevaert O, Tibshirani R, Plevritis SK. Pancancer analysis of DNA methylation-driven genes using MethylMix. *Genome Biol*. 2015;16(1):17. doi:10.1186/s13059-014-0579-8.

200. Huang R-L, Su P-H, Liao Y-P, et al. Integrated Epigenomics Analysis Reveals a DNA Methylation Panel for Endometrial Cancer Detection Using Cervical Scrapings. *Clin Cancer Res*. 2017;23(1):263-272. doi:10.1158/1078-0432.CCR-16-0863.

201. Schlensog M, Magnus L, Heide T, et al. Epigenetic loss of putative tumor suppressor SFRP3 correlates with poor prognosis of lung adenocarcinoma patients. *Epigenetics*. 2016;14(3):1-14. doi:10.1080/15592294.2016.1229730.

202. Salas LA, Johnson KC, Koestler DC, O'Sullivan DE, Christensen BC. Integrative epigenetic and genetic pan-cancer somatic alteration portraits. *Epigenetics*. 2017;12(7):561-574. doi:10.1080/15592294.2017.1319043.

203. Wang K, Liang Q, Li X, et al. MDGA2 is a novel tumour suppressor cooperating with DMAP1 in gastric cancer and is associated with disease outcome. - PubMed - NCBI. *Gut*. 2016;65(10):1619-1631. doi:10.1136/gutjnl-2015-309276.

204. Hsu L-S, Lee H-C, Chau G-Y, Yin P-H, Chi C-W, Lui WY. Aberrant methylation of EDNRB and p16 genes in hepatocellular carcinoma (HCC) in Taiwan. *Oncol Rep*. 2006;15(2):507-511.

205. Lo K-W, Tsang Y-S, Kwong J, To K-F, Teo PML, Huang DP. Promoter hypermethylation of the EDNRB gene in nasopharyngeal carcinoma. *Int J Cancer*. 2002;98(5):651-655.

206. Chen C, Wang L, Liao Q, et al. Hypermethylation of EDNRB promoter contributes to the risk of colorectal cancer. - PubMed - NCBI. *Diagn Pathol*. 2013;8(1):792362. doi:10.1186/1746-1596-8-199.

207. Yin D, Jia Y, Yu Y, et al. SOX17 methylation inhibits its antagonism of Wnt signaling pathway in lung cancer. *Discovery medicine*. 2012;14(74):33-40.

208. Fu D-Y, Wang Z-M, Li-Chen, et al. Sox17, the canonical Wnt antagonist, is epigenetically inactivated by promoter methylation in human breast cancer. *Breast Cancer Res Treat*. 2009;119(3):601-612. doi:10.1007/s10549-009-0339-8.

209. Jia Y, Yang Y, Zhan Q, et al. Inhibition of SOX17 by microRNA 141 and methylation activates the WNT signaling pathway in esophageal cancer. - PubMed - NCBI. *The Journal of Molecular Diagnostics*. 2012;14(6):577-585. doi:10.1016/j.jmoldx.2012.06.004.

210. Chai AWY, Cheung AKL, Dai W, et al. Metastasis-suppressing NID2, an epigenetically-silenced gene, in the pathogenesis of nasopharyngeal carcinoma and esophageal squamous cell carcinoma. *Oncotarget*. 2016;7(48):78859-78871. doi:10.18632/oncotarget.12889.

211. Meissner A. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. - PubMed - NCBI. *Nucleic Acids Res*. 2005;33(18):5868-5877. doi:10.1093/nar/gki901.

212. Udager AM, Liu TY, Skala SL, et al. Frequent PD-L1 expression in primary and metastatic penile squamous cell carcinoma: potential opportunities for immunotherapeutic approaches. *Ann Oncol*. 2016;27(9):1706-1712. doi:10.1093/annonc/mdw216.

213. Herman JG, Baylin SB. Gene Silencing in Cancer in Association with Promoter Hypermethylation. *N Engl J Med*. 2003;349(21):2042-2054. doi:10.1056/NEJMra023075.

214. Scott M, Stites DP, Moscicki AB. Th1 cytokine patterns in cervical human papillomavirus infection. *Clin Diagn Lab Immunol*. 1999;6(5):751-755.

215. Bleotu C, Chifiriuc MC, Grigore R, et al. Investigation of Th1/Th2 cytokine profiles in patients with laryngo-pharyngeal, HPV-positive cancers. *Eur Arch Otorhinolaryngol*. 2012;270(2):711-718. doi:10.1007/s00405-012-2067-7.

216. Gibney GT, Weiner LM, Atkins MB. Predictive biomarkers for checkpoint

inhibitor-based immunotherapy. *The Lancet Oncology*. 2016;17(12):e542-e551. doi:10.1016/S1470-2045(16)30406-5.

217. Dempke WCM, Fenchel K, Uciechowski P, Dale SP. Second- and third-generation drugs for immuno-oncology treatment—The more the better? *European Journal of Cancer*. 2017;74:55-72. doi:10.1016/j.ejca.2017.01.001.

218. Andrews LP, Marciscano AE, Drake CG, Vignali DAA. LAG3 (CD223) as a cancer immunotherapy target. *Immunol Rev*. 2017;276(1):80-96. doi:10.1111/imr.12519.

219. Li F, Zhang R, Li S, Liu J. IDO1: An important immunotherapy target in cancer treatment. - PubMed - NCBI. *International Immunopharmacology*. 2017;47:70-77. doi:10.1016/j.intimp.2017.03.024.

220. Anderson AC. Tim-3: an emerging target in the cancer immunotherapy landscape. - PubMed - NCBI. *Cancer Immunology Research*. 2014;2(5):393-398. doi:10.1158/2326-6066.CIR-14-0039.

221. Yonezawa A, Dutt S, Chester C, Kim J, Kohrt HE. Boosting Cancer Immunotherapy with Anti-CD137 Antibody Therapy. - PubMed - NCBI. *Clin Cancer Res*. 2015;21(14):3113-3120.

222. Muntasell A, Ochoa MC, Cordeiro L, et al. Targeting NK-cell checkpoints for cancer immunotherapy. - PubMed - NCBI. *Current Opinion in Immunology*. 2017;45:73-81. doi:10.1016/j.coi.2017.01.003.

223. Quek K, Li J, Estecio M, et al. DNA methylation intratumor heterogeneity in localized lung adenocarcinomas. - PubMed - NCBI. *Oncotarget*. 2017;8(13):21994-22002.

224. Hlady RA, Zhou D, Puszyk W, Roberts LR, Liu C, Robertson KD. Initiation of aberrant DNA methylation patterns and heterogeneity in precancerous lesions of human hepatocellular cancer. - PubMed - NCBI. *Epigenetics*. 2017;12(3):215-225. doi:10.1080/15592294.2016.1277297.

225. Leo S Payne PHH. Discoidin domain receptor 2 signaling networks and therapy in lung cancer. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer*. 2014;9(6):900-904. doi:10.1097/JTO.0000000000000164.

226. Grither WR, Divine LM, Meller EH, et al. TWIST1 induces expression of discoidin domain receptor 2 to promote ovarian cancer metastasis. *Oncogene*. 2018;37(13):1714-1729. doi:10.1038/s41388-017-0043-9.

227. Grither WR, Longmore GD. Inhibition of tumor–microenvironment interaction and tumor invasion by small-molecule allosteric inhibitor of DDR2 extracellular domain. *PNAS*. 2018;115(33):E7786-E7794.

doi:10.1073/pnas.1805020115.

228. Mässenhausen von A, Sanders C, Brägelmann J, et al. Targeting DDR2 in head and neck squamous cell carcinoma with dasatinib. - PubMed - NCBI. *Int J Cancer*. 2016;139(10):2359-2369. doi:10.1002/ijc.30279.

229. Chou J, Provot S, Werb Z. GATA3 in development and cancer differentiation: Cells GATA have it! *J Cell Physiol*. 2010;222(1):42-49. doi:10.1002/jcp.21943.

230. Pei X-H, Bai F, Smith MD, et al. CDK Inhibitor p18INK4c Is a Downstream Target of GATA3 and Restrains Mammary Luminal Progenitor Cell Proliferation and Tumorigenesis. *Cancer cell*. 2009;15(5):389-401. doi:10.1016/j.ccr.2009.03.004.

231. Li Y, Ishiguro H, Kawahara T, Kashiwagi E, Izumi K, Miyamoto H. Loss of GATA3 in bladder cancer promotes cell migration and invasion. *Cancer Biology & Therapy*. 2014;15(4):428-435. doi:10.4161/cbt.27631.

232. Li Y-F, Hsiao Y-H, Lai Y-H, et al. DNA methylation profiles and biomarkers of oral squamous cell carcinoma. *Epigenetics*. 2015;10(3):229-236. doi:10.1080/15592294.2015.1006506.

233. Chang J-W, Kuo W-H, Lin C-M, et al. Wild-type p53 upregulates an early onset breast cancer-associated gene GAS7 to suppress metastasis via GAS7–CYFIP1-mediated signaling pathway. *Oncogene*. 2018;37(30):4137-4150. doi:10.1038/s41388-018-0253-9.

234. Ping W, Gao Y, Fan X, Li W, Deng Y, Fu X. MiR-181a contributes gefitinib resistance in non-small cell lung cancer cells by targeting GAS7. *Biochemical and Biophysical Research Communications*. 2018;495(4):2482-2489. doi:10.1016/j.bbrc.2017.12.096.

235. Wang R-J, Wu P, Cai G-X, et al. Down-regulated MYH11 expression correlates with poor prognosis in stage II and III colorectal cancer. *Asian Pac J Cancer Prev*. 2014;15(17):7223-7228.

236. Shen H, Zhan M, Zhang Y, et al. PLZF inhibits proliferation and metastasis of gallbladder cancer by regulating IFIT2. *Cell Death & Disease 2018 9:2*. 2018;9(2):71. doi:10.1038/s41419-017-0107-3.

237. Hsieh CL, Botta G, Gao S, et al. PLZF, a Tumor Suppressor Genetically Lost in Metastatic Castration-Resistant Prostate Cancer, Is a Mediator of Resistance to Androgen Deprivation Therapy. *Cancer Res*. 2015;75(10):1944-1948. doi:10.1158/0008-5472.CAN-14-3602.

238. Weng J, Rawal S, Chu F, et al. TCL1: a shared tumor-associated antigen for immunotherapy against B-cell lymphomas. *Blood*. 2012;120(8):1613-1623. doi:10.1182/blood-2011-09-382838.

239. Gennaro Napolitano AB. TFEB at a glance. *Journal of Cell Science*. 2016;129(13):2475-2481. doi:10.1242/jcs.146365.

240. Kundu ST, Grzeskowiak CL, Fradette JJ, et al. TMEM106B drives lung cancer metastasis by inducing TFEB -dependent lysosome synthesis and secretion of cathepsins. *Nat Comms*. 2018;9(1):2731. doi:10.1038/s41467-018-05013-x.

241. KLEIN K, WERNER K, TESKE C, et al. Role of TFEB-driven autophagy regulation in pancreatic cancer treatment. - PubMed - NCBI. *International Journal of Oncology*. 2016;49(1):164-172. doi:10.3892/ijo.2016.3505.

242. Giatromanolaki A, Sivridis E, Kalamida D, Koukourakis MI. Transcription Factor EB Expression in Early Breast Cancer Relates to Lysosomal/Autophagosomal Markers and Prognosis. *Clinical Breast Cancer*. 2017;17(3):e119-e125. doi:10.1016/j.clbc.2016.11.006.

243. McDaniel AS, Hovelson DH, Cani AK, et al. Genomic Profiling of Penile Squamous Cell Carcinoma Reveals New Opportunities for Targeted Therapy. *Cancer Res*. 2015;75(24):5219-5227. doi:10.1158/0008-5472.CAN-15-1004.

244. Hu X, Moon JW, Li S, et al. Amplification and overexpression of CTTN and CCND1 at chromosome 11q13 in Esophagus squamous cell carcinoma (ESCC) of North Eastern Chinese Population. *International Journal of Medical Sciences*. 2016;13(11):868-874. doi:10.7150/ijms.16845.

245. Zhang Z-M, Lu R, Wang P, et al. Structural basis for DNMT3A-mediated de novo DNA methylation. *Nature*. 2018;554(7692):387-391. doi:10.1038/nature25477.

246. Sen P, Ganguly P, Ganguly N. Modulation of DNA methylation by human papillomavirus E6 and E7 oncoproteins in cervical cancer. *Oncology Letters*. 2018;15(1):11-22. doi:10.3892/ol.2017.7292.

247. Hamid O, Schmidt H, Nissan A, et al. A prospective phase II trial exploring the association between tumor microenvironment biomarkers and clinical activity of ipilimumab in advanced melanoma. *J Transl Med*. 2011;9(1):1-16. doi:10.1186/1479-5876-9-204.

248. Balatoni T, Mohos A, Papp E, et al. Tumor-infiltrating immune cells as potential biomarkers predicting response to treatment and survival in patients with metastatic melanoma receiving ipilimumab therapy. *Cancer Immunol Immunother*. 2017;67(1):141-151. doi:10.1007/s00262-017-2072-1.

249. Conroy JM, Pabla S, Nesline MK, et al. Next generation sequencing of PD-L1 for predicting response to immune checkpoint inhibitors. *j*

immunotherapy cancer. 2019;7(1):1-11. doi:10.1186/s40425-018-0489-5.

250. Sen S, Mandal P, Bhattacharya A, et al. Impact of viral and host DNA methylations on HPV16-related cervical cancer pathogenesis:. *Tumor Biology*. 2017;39(5):101042831769979. doi:10.1177/1010428317699799.

251. JIMÉNEZ-WENCES H, PERALTA-ZARAGOZA O, FERNÁNDEZ-TILAPA G. Human papilloma virus, DNA methylation and microRNA expression in cervical cancer (Review). *Oncol Rep*. 2014;31(6):2467-2476. doi:10.3892/or.2014.3142.

252. Lyu G-Y, Yeh Y-H, Yeh Y-C, Wang Y-C. Mutation load estimation model as a predictor of the response to cancer immunotherapy. *npj Genomic Medicine 2018 3:1*. 2018;3(1):12. doi:10.1038/s41525-018-0051-x.

253. Goodman AM, Kato S, Bazhenova L, et al. Tumor Mutational Burden as an Independent Predictor of Response to Immunotherapy in Diverse Cancers. *Mol Cancer Ther*. 2017;16(11):2598-2608. doi:10.1158/1535-7163.MCT-17-0386.

254. Seeber A, Klinglmair G, Fritz J, et al. High IDO-1 expression in tumor endothelial cells is associated with response to immunotherapy in metastatic renal cell carcinoma. *Cancer Sci*. 2018;109(5):1583-1591. doi:10.1111/cas.13560.

255. Tseng R-C, Hsieh F-J, Hsu H-S, Wang Y-C. Minimal deletion regions in lung squamous cell carcinoma: association with abnormality of the DNA double-strand break repair genes and their applic... - PubMed - NCBI. *Lung Cancer*. 2008;59(3):332-339. doi:10.1016/j.lungcan.2007.08.038.

256. Benson E, Li R, Eisele D, Fakhry C. The clinical impact of HPV tumor status upon head and neck squamous cell carcinomas. *Oral Oncology*. 2014;50(6):565-574. doi:10.1016/j.oraloncology.2013.09.008.

257. Necchi A, Vullo Lo S, Perrone F, et al. First-line therapy with dacomitinib, an orally available pan-HER tyrosine kinase inhibitor, for locally advanced or metastatic penile squamous cell... - PubMed - NCBI. *BJU Int*. 2017;121(3):348-356. doi:10.1111/bju.14013.

258. Kloth DD, Iacovelli L, Arbuckle R, McIntosh AC. The escalating role of epidermal growth factor receptor inhibitors in cancer management: clinical considerations for the health system pharmacist. *P T*. 2010;35(4):219-229.

259. Eggersmann TK, Degenhardt T, Gluz O, Wuerstlein R, Harbeck N. CDK4/6 Inhibitors Expand the Therapeutic Options in Breast Cancer: Palbociclib, Ribociclib and Abemaciclib. *BioDrugs*. 2019;33(2):125-135. doi:10.1007/s40259-019-00337-6.

260. Chabeda A, Yanez RJR, Lamprecht R, Meyers AE, Rybicki EP, Hitzeroth II. Therapeutic vaccines for high-risk HPV-associated diseases. - PubMed - NCBI. *Papillomavirus Research*. 2018;5:46-58. doi:10.1016/j.pvr.2017.12.006.

261. Juergens RA, Wrangle J, Vendetti FP, et al. Combination Epigenetic Therapy Has Efficacy in Patients with Refractory Advanced Non-Small Cell Lung Cancer. *Cancer Discovery*. 2011;1(7):598-607. doi:10.1158/2159-8290.CD-11-0214.

262. Chen J, Yao K, Li Z, et al. Establishment and characterization of a penile cancer cell line, penl1, with a deleterious TP53 mutation as a paradigm of HPV-negative penile carcinogenesis. *Oncotarget*. 2016;7(32):51687-51698. doi:10.18632/oncotarget.10098.

263. Zhou Q-H, Deng C-Z, Li Z-S, et al. Molecular characterization and integrative genomic analysis of a panel of newly established penile cancer cell lines. *Cell Death & Disease 2018 9:2*. 2018;9(6):S2. doi:10.1038/s41419-018-0736-1.

8 Appendix

8.1 Whole exome sequencing output

Table 38: Sequencing output statistics for whole exome sequencing. Passing filter and mapping statistics provided by Oxford kindly provided by Oxford Genomics Centre, University of Oxford. Number of reads, duplication rate and coverage obtained from the output of quality control software Qualimap. Depth of coverage was calculated over captured regions. Whole exome sequencing captured regions provided in the design specification files from Nimblegen.

Sample	Patient	Tissue type	Number of reads	Passing filter (%)	Duplications (%)	Mapped (%)	Coverage (X)
39_01b	39	Primary cancer	102,349,058	100	24%	99.1	86
39_01c	39	Primary cancer	120,477,058	100	24%	98.2	101
39_01d	39	Primary cancer	176,901,437	100	27%	99.3	151
39_01e	39	Primary cancer	114,721,548	100	22%	98.4	89
39_03	39	'Normal'	129,112,456	100	25%	99.7	106
39_05	39	LN Metastasis	117,930,530	100	22%	99.0	95
45_01b	45	Primary cancer	164,718,194	100	26%	98.0	138
45_01c	45	Primary cancer	123,509,992	100	23%	99.5	100
45_01d	45	Primary cancer	124,416,050	100	23%	99.3	98
45_01e	45	Primary cancer	130,938,438	100	25%	99.2	107
45_03	45	'Normal'	109,906,475	100	26%	99.4	94
45_05	45	LN Metastasis	119,822,121	100	24%	99.4	97
49_015	49	LN Metastasis	107,449,861	100	22%	98.5	90
49_01b	49	Primary cancer	138,253,668	100	25%	99.4	118
49_01c	49	Primary cancer	116,488,044	100	24%	98.9	101
49_01d	49	Primary cancer	146,327,309	100	25%	99.3	130
49_01e	49	Primary cancer	121,483,002	100	23%	99.4	103
49_03	49	'Normal'	105,765,728	100	23%	99.2	90
51_01b	51	Primary cancer	168,614,521	100	25%	99.7	146
51_01c	51	Primary cancer	82,341,210	100	22%	98.6	70
51_01d	51	Primary cancer	147,898,818	100	25%	98.7	129
51_01e	51	Primary cancer	117,410,485	100	22%	99.7	97
51_03	51	'Normal'	146,774,390	100	27%	96.4	133
51_05	51	LN Metastasis	149,260,266	100	22%	99.5	120
63_01a	63	Primary cancer	126,444,117	100	24%	99.0	105
63_01c	63	Primary cancer	149,933,535	100	25%	99.7	125
63_01d	63	Primary cancer	141,294,315	100	26%	99.6	123
63_01e	63	Primary cancer	123,771,626	100	24%	99.3	102
63_03	63	'Normal'	151,739,818	100	25%	99.4	128
63_05	63	LN Metastasis	128,722,227	100	23%	99.0	106
64_01a	64	Primary cancer	124,040,556	100	24%	99.1	103
64_01c	64	Primary cancer	116,090,567	100	24%	99.0	101
64_01d	64	Primary cancer	111,669,086	100	24%	99.0	98
64_01e	64	Primary cancer	139,847,143	100	24%	97.7	122
64_03	64	'Normal'	133,881,517	100	26%	99.3	118
64_05	64	LN Metastasis	96,751,327	100	24%	97.4	84
66_01a	66	Primary cancer	138,327,376	100	24%	99.7	113
66_01c	66	Primary cancer	108,034,232	100	21%	99.0	84
66_01d	66	Primary cancer	133,908,402	100	25%	97.4	111
66_01e	66	Primary cancer	78,459,281	100	21%	98.9	62
66_03	66	'Normal'	112,907,325	100	23%	98.9	91
66_05	66	LN Metastasis	132,962,266	100	22%	99.8	101
79_01a	79	Primary cancer	117,534,574	100	24%	99.5	96
79_01b	79	Primary cancer	140,083,118	100	23%	99.4	113
79_01c	79	Primary cancer	156,963,984	100	23%	99.5	124
79_01e	79	Primary cancer	125,847,451	100	22%	99.0	99
79_03	79	'Normal'	125,945,300	100	21%	99.4	95
79_05	79	LN Metastasis	159,125,023	100	22%	99.6	122

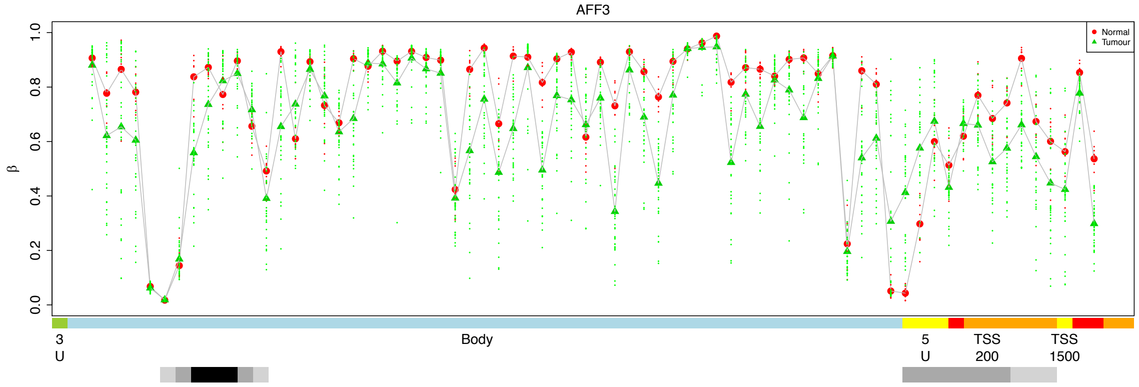
8.2 Immune associated DMP GO terms (top 100)

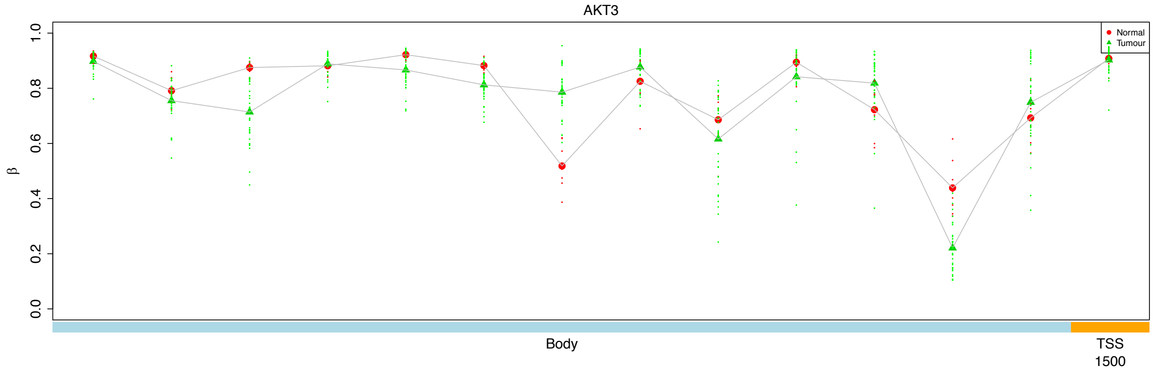
Table 39: Top 100 GO terms for immune associated DMPs.

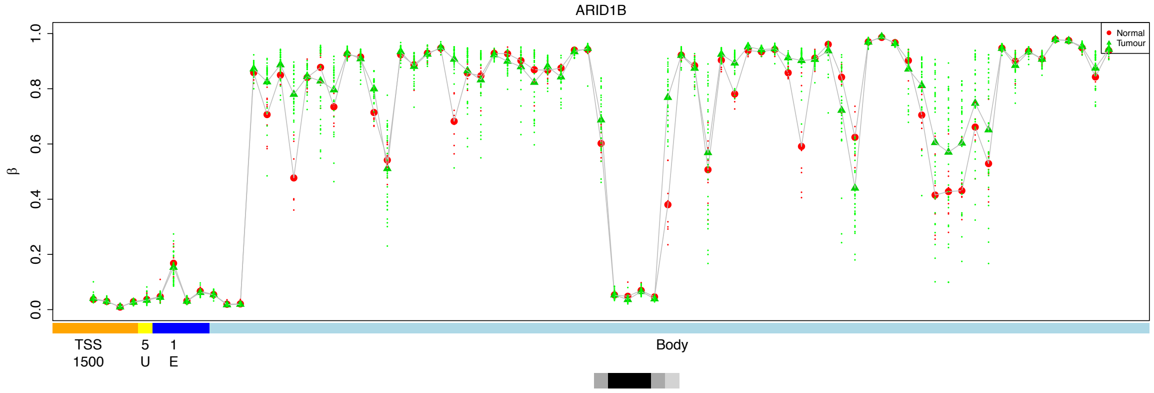
GO term	FDR
immune system process	7.01E-28
immune response	1.15E-24
cell activation	9.60E-20
leukocyte activation	9.60E-20
regulation of response to stimulus	9.60E-20
positive regulation of response to stimulus	3.69E-18
regulation of signaling	1.62E-17
single-organism cellular process	1.71E-17
regulation of cell communication	6.08E-17
single-organism localization	2.77E-16
immune effector process	5.47E-16
cell surface receptor signaling pathway	6.05E-16
positive regulation of biological process	7.09E-16
regulation of signal transduction	1.04E-15
regulation of immune system process	1.42E-15
defense response	3.40E-15
localization	3.53E-15
cellular response to chemical stimulus	4.51E-15
vesicle	5.97E-15
intracellular signal transduction	6.55E-15
response to stimulus	8.93E-15
single-organism transport	1.35E-14
response to oxygen-containing compound	1.35E-14
biological adhesion	4.47E-14
plasma membrane part	5.18E-14
cell adhesion	5.64E-14
secretion	9.28E-14
positive regulation of signaling	1.16E-13
positive regulation of signal transduction	1.17E-13
vesicle-mediated transport	1.38E-13
protein binding	2.28E-13
positive regulation of cell communication	2.34E-13
inflammatory response	2.86E-13
cellular response to oxygen-containing compound	2.94E-13
response to external stimulus	3.16E-13
regulation of localization	3.16E-13
regulation of intracellular signal transduction	5.16E-13
cellular response to organic substance	1.51E-12
positive regulation of cellular process	1.82E-12
lymphocyte activation	2.41E-12
response to organic substance	2.48E-12
secretion by cell	2.69E-12
leukocyte mediated immunity	2.73E-12
phosphorus metabolic process	3.18E-12
myeloid leukocyte activation	7.73E-12
phosphate-containing compound metabolic process	1.11E-11
positive regulation of intracellular signal transduction	1.55E-11
positive regulation of immune system process	1.81E-11
transport	2.62E-11
single organism cell adhesion	2.83E-11
establishment of localization	3.42E-11
cell communication	3.42E-11
regulation of multicellular organismal process	4.07E-11
single organism signaling	4.81E-11
signaling	5.31E-11
regulation of immune response	5.41E-11
cytoplasm	5.41E-11
extracellular organelle	6.09E-11
extracellular vesicle	6.46E-11
movement of cell or subcellular component	7.18E-11
leukocyte activation involved in immune response	7.83E-11
extracellular exosome	7.83E-11
extracellular region part	7.92E-11
hemopoiesis	1.00E-10
leukocyte differentiation	1.06E-10
cell activation involved in immune response	1.07E-10
myeloid leukocyte mediated immunity	1.29E-10
locomotion	1.29E-10
leukocyte migration	1.49E-10
regulation of molecular function	2.01E-10
regulation of response to external stimulus	2.16E-10
cytoplasmic vesicle	2.20E-10
whole membrane	2.28E-10
intracellular vesicle	2.35E-10
regulation of cell differentiation	3.11E-10
hematopoietic or lymphoid organ development	3.22E-10
regulation of transport	4.50E-10
cell migration	5.30E-10
T cell activation	5.30E-10
endocytosis	5.71E-10
cytokine production	5.82E-10
leukocyte degranulation	6.20E-10
regulation of MAPK cascade	7.23E-10
immune system development	7.31E-10
regulation of cell proliferation	8.28E-10
neutrophil mediated immunity	8.73E-10
receptor binding	9.83E-10
cell proliferation	1.08E-09

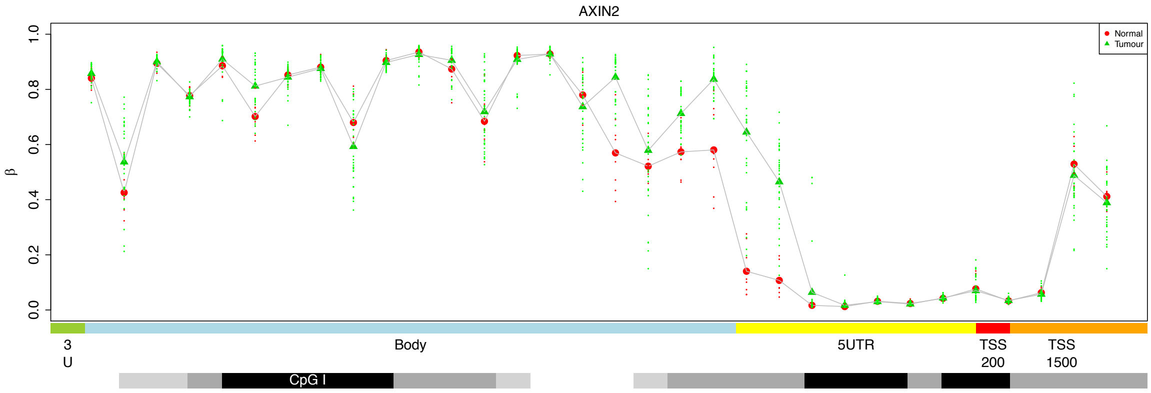
8.3 Canonical gene plots of DMPs associated with potential oncogenic driver genes

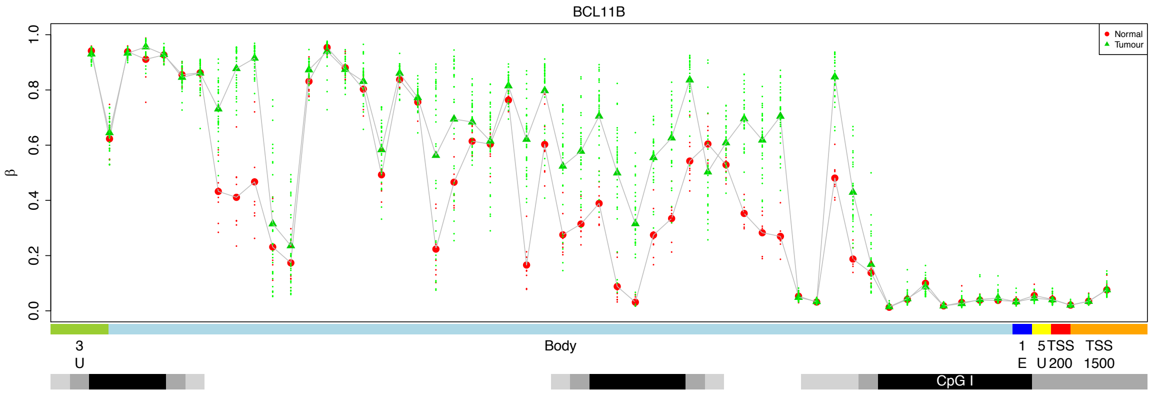
Gene methylation plots demonstrating the beta methylation value for each sample across a gene. Genes are annotated with CpG island location (black horizontal bar at the bottom of the figure) as well as transcription start site (TSS), 5'UTR and gene body. Each red point indicates a primary tumour sample beta methylation value at an individual locus. Each green point represents a normal control sample value.

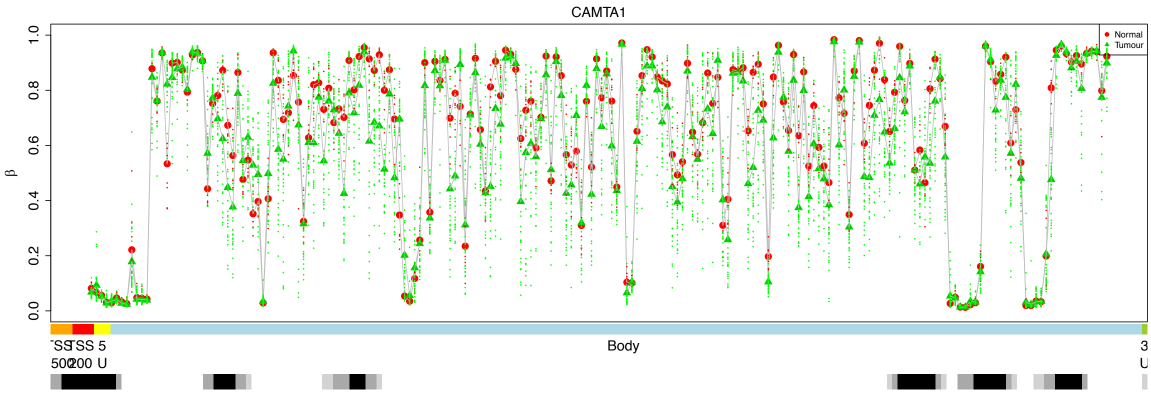


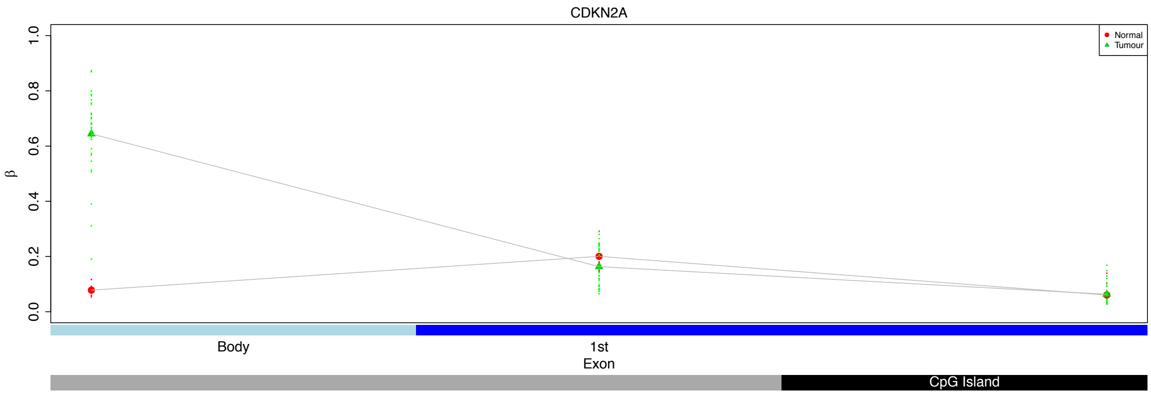


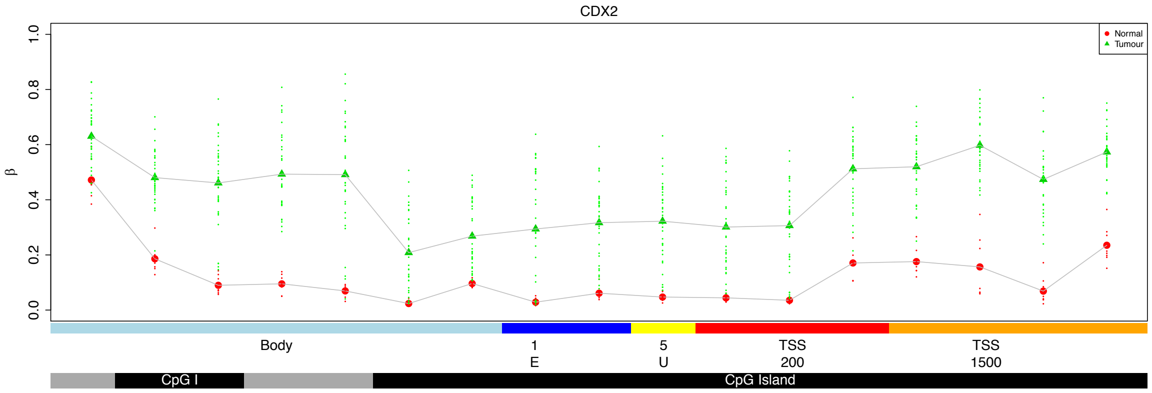


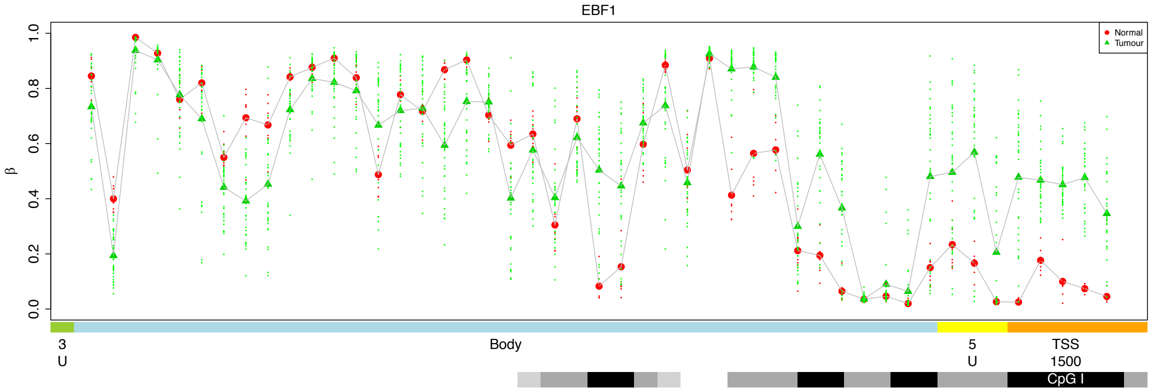

AXIN2
β
0.0
0.2
0.4
0.6
0.8
1.0
Normal
Tumour
3
U
Body
5UTR
TSS
200
TSS
1500
CpG I

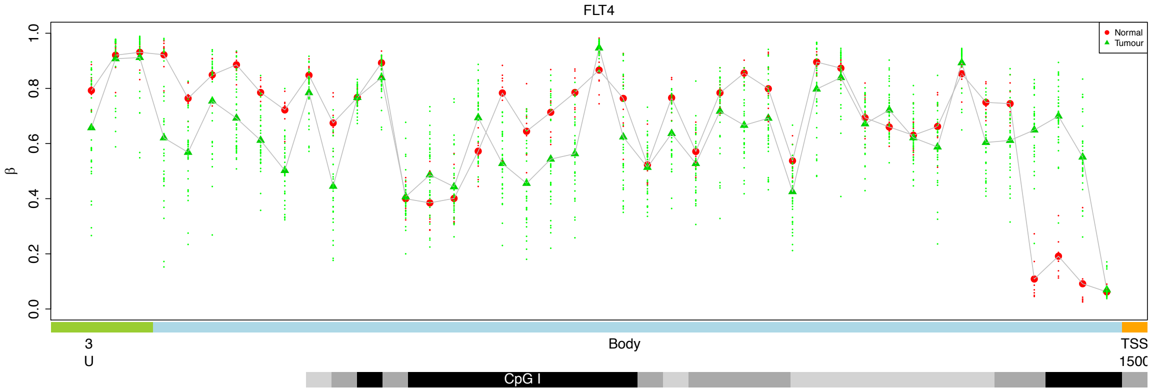

BCL11B
β
0.0
0.2
0.4
0.6
0.8
1.0
Normal
Tumour
3
U
Body
1
E
5
U
TSS
200
TSS
1500
CpG I

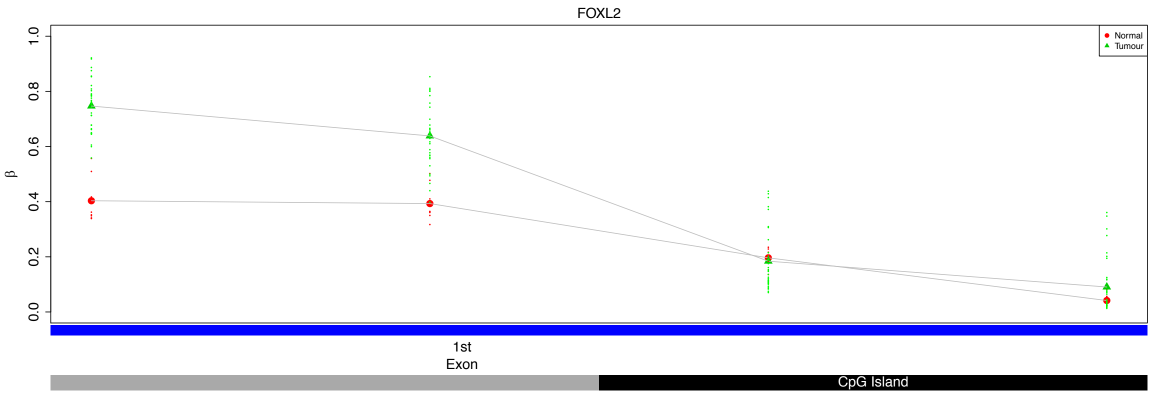

CAMTA1
β
0.0
0.2
0.4
0.6
0.8
1.0
Normal
Tumour
TSS
1500
TSS
200
5
U
Body
3
U

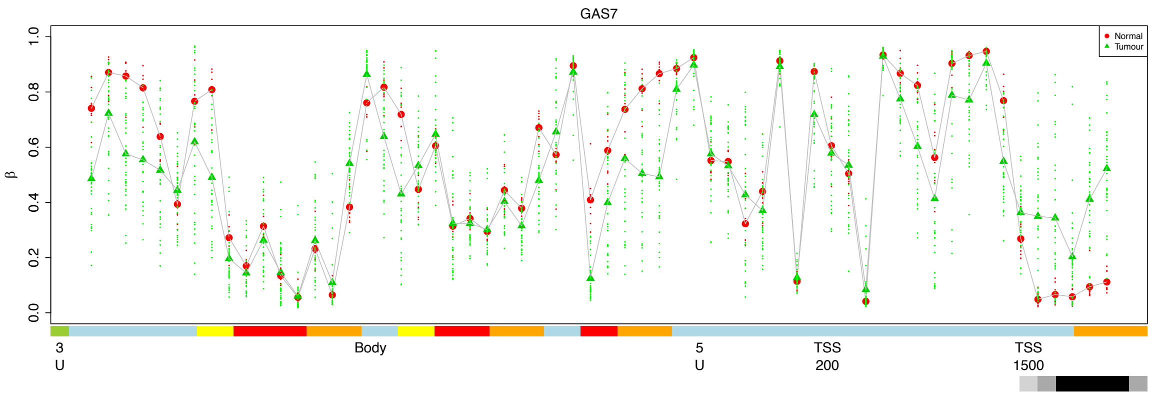

CDKN2A
β
0.0
0.2
0.4
0.6
0.8
1.0
Normal
Tumour
Body
1st
Exon
CpG Island

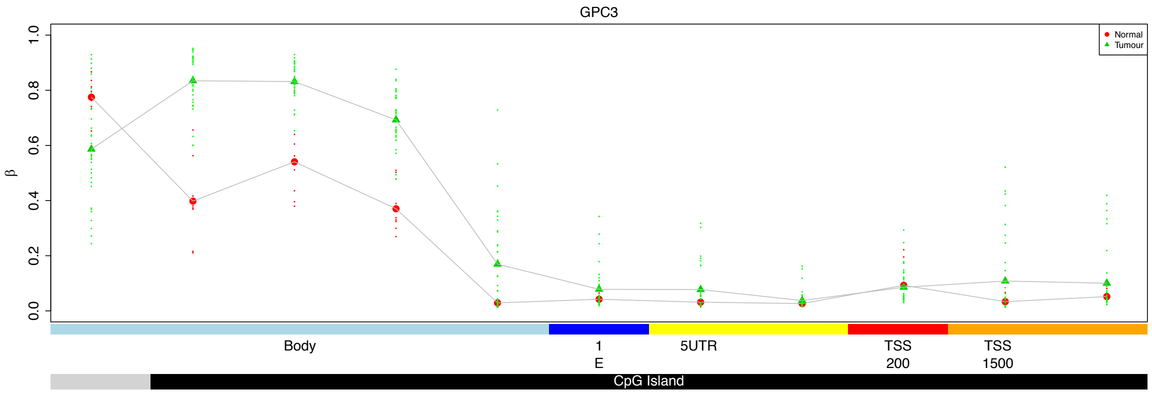

CDX2
Normal
Tumour
β
0.0
0.2
0.4
0.6
0.8
1.0
Body
1
E
5
U
TSS
200
TSS
1500
CpG I
CpG Island

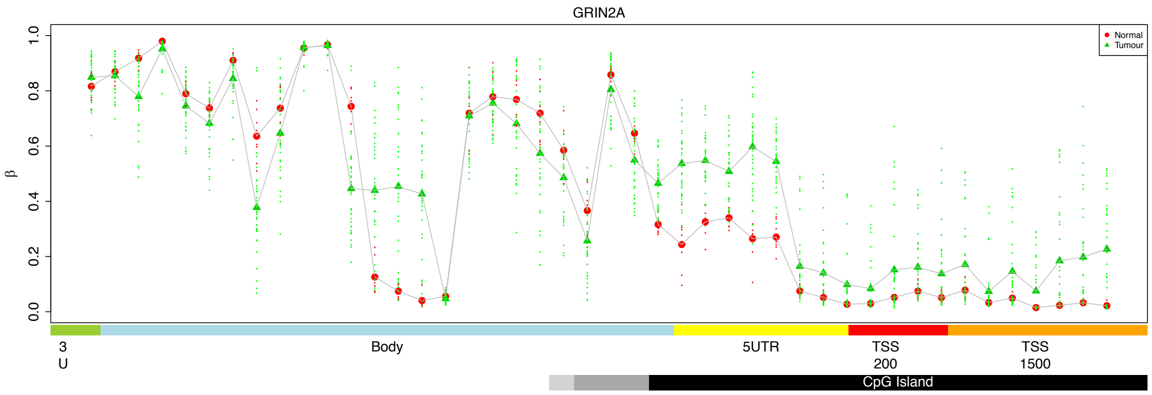

EBF1
Normal
Tumour
β
0.0
0.2
0.4
0.6
0.8
1.0
3
U
Body
5
U
TSS
1500
CpG I

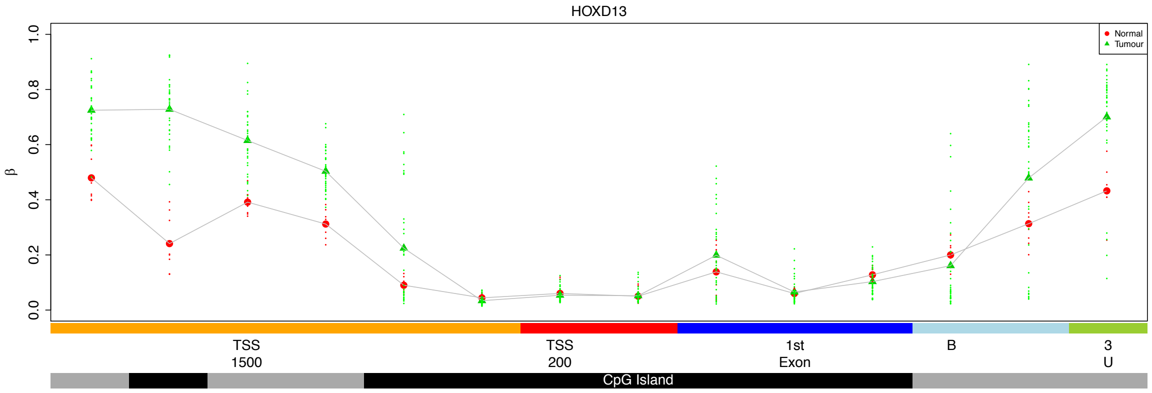

FLT4
Normal
Tumour
β
0.0
0.2
0.4
0.6
0.8
1.0
3
U
Body
TSS
150(
CpG I

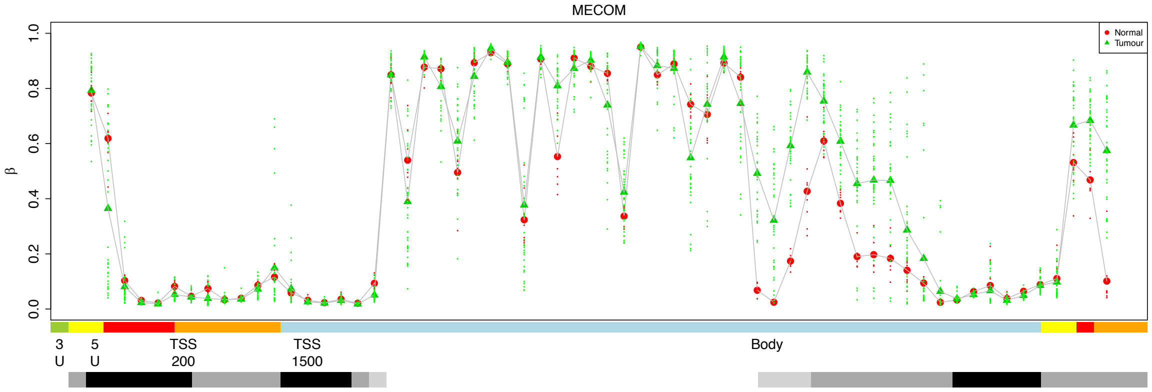

FOXL2
β
0.0
0.2
0.4
0.6
0.8
1.0
Normal
Tumour
1st
Exon
CpG Island

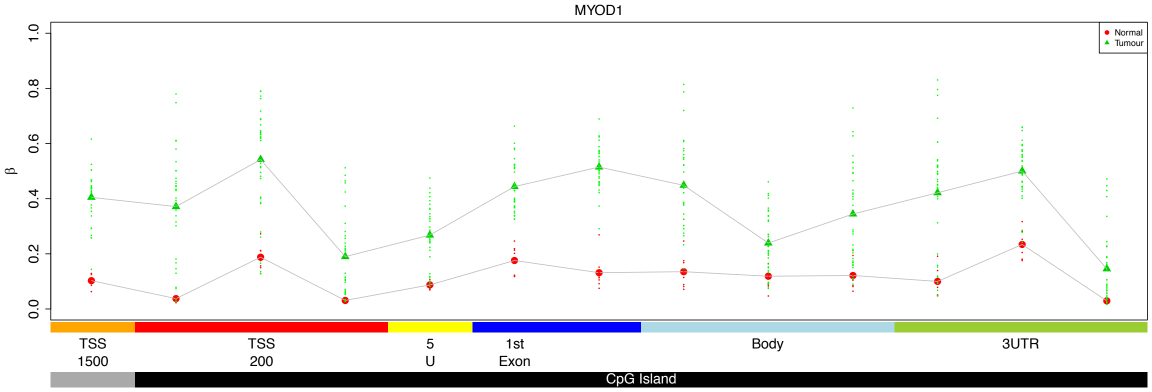

GAS7
β
0.0
0.2
0.4
0.6
0.8
1.0
Normal
Tumour
3
U
Body
5
U
TSS
200
TSS
1500

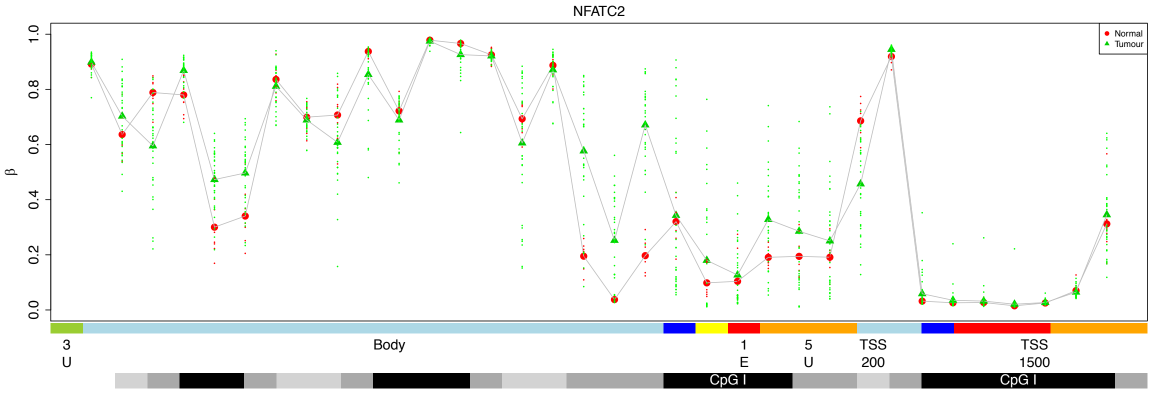

GPC3
β
0.0
0.2
0.4
0.6
0.8
1.0
Normal
Tumour
Body
1
E
5UTR
TSS
200
TSS
1500
CpG Island

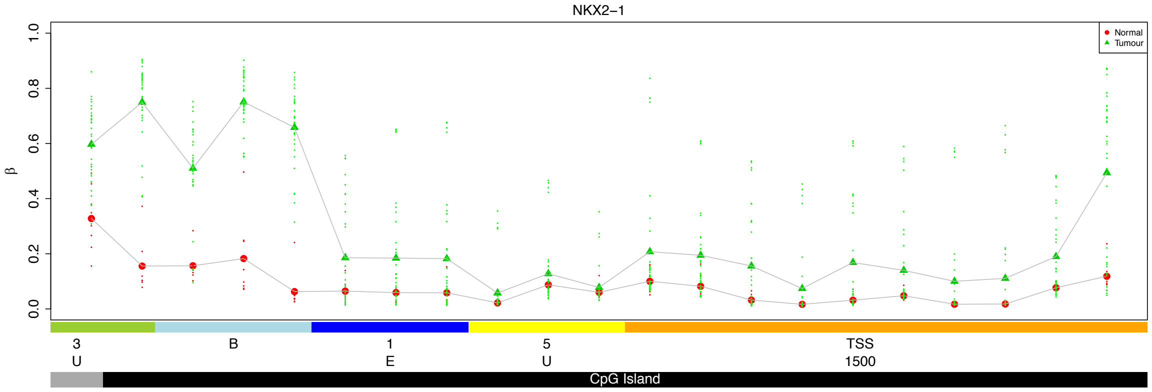

GRIN2A
β
0.0
0.2
0.4
0.6
0.8
1.0
Normal
Tumour
3
U
Body
5UTR
TSS
200
TSS
1500
CpG Island

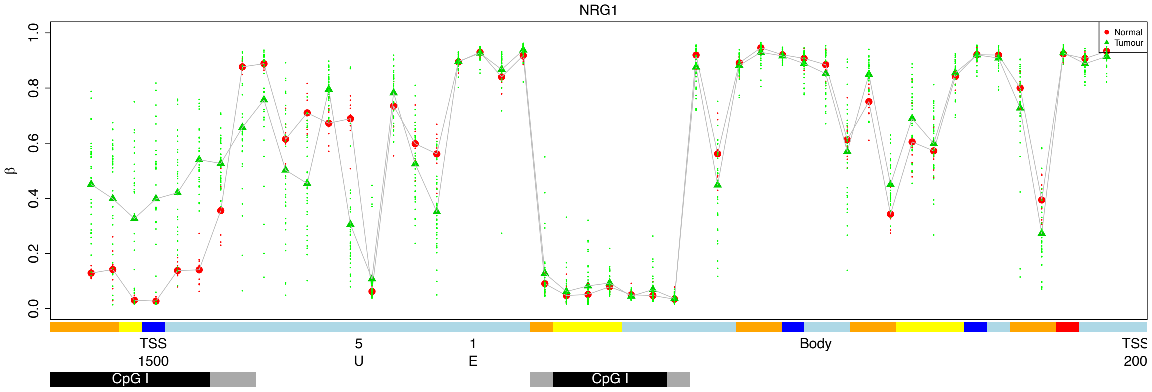

HOXD13
Normal
Tumour
β
0.0
0.2
0.4
0.6
0.8
1.0
TSS
1500
TSS
200
1st
Exon
B
3
U
CpG Island

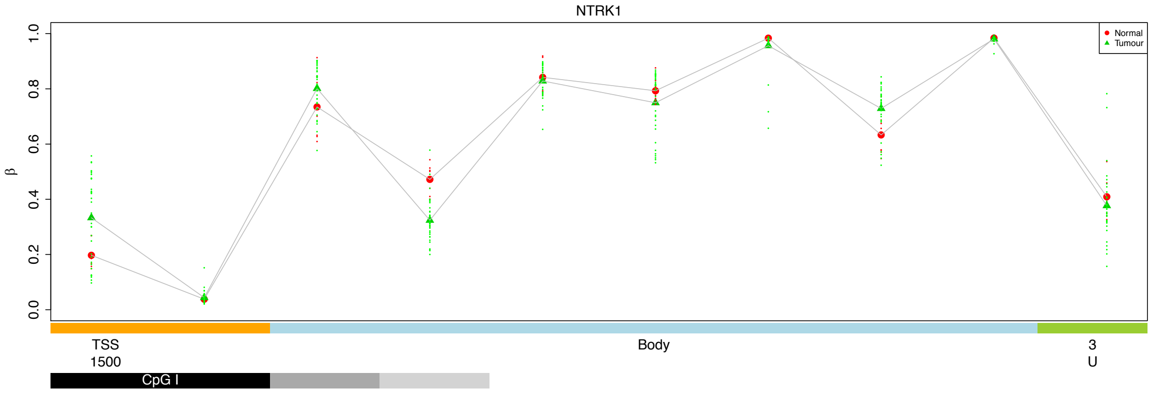

MECOM
Normal
Tumour
β
0.0
0.2
0.4
0.6
0.8
1.0
3
U
5
U
TSS
200
TSS
1500
Body

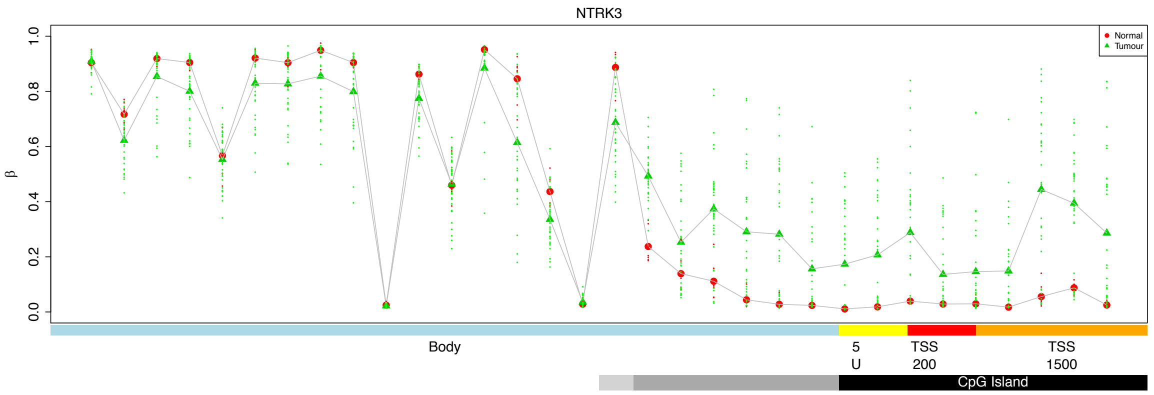

MYOD1
Normal
Tumour
β
0.0
0.2
0.4
0.6
0.8
1.0
TSS
1500
TSS
200
5
U
1st
Exon
Body
3UTR
CpG Island

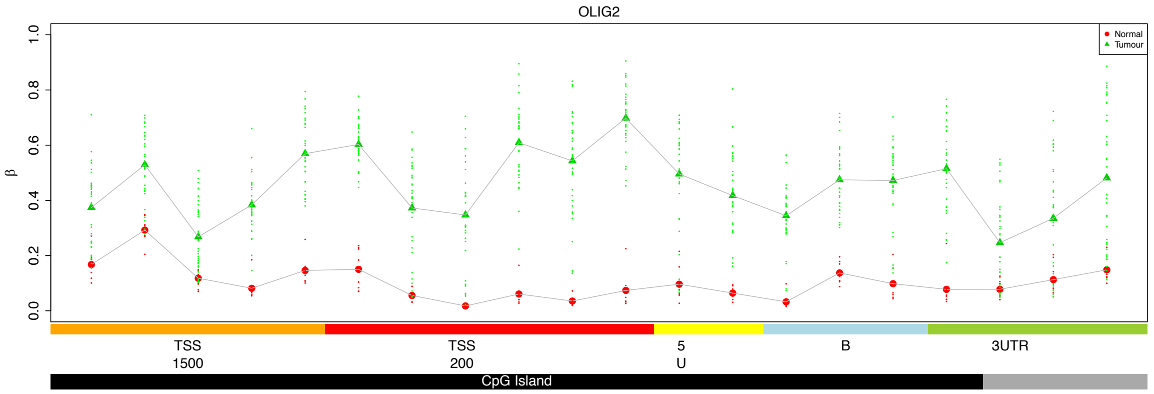

NFATC2
β
0.0
0.2
0.4
0.6
0.8
1.0
Normal
Tumour
3
U
Body
1
E
5
U
TSS
200
TSS
1500
CpG I
CpG I

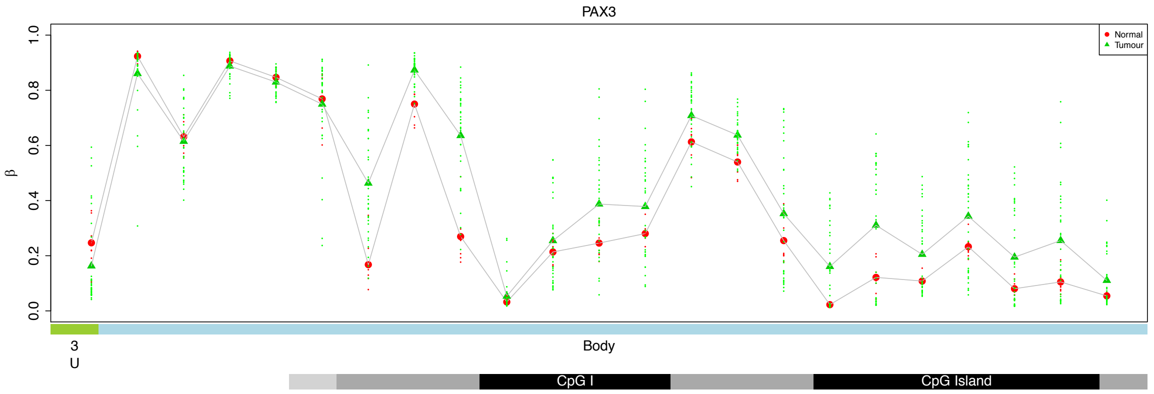

NKX2−1
β
0.0
0.2
0.4
0.6
0.8
1.0
Normal
Tumour
3
U
B
1
E
5
U
TSS
1500
CpG Island

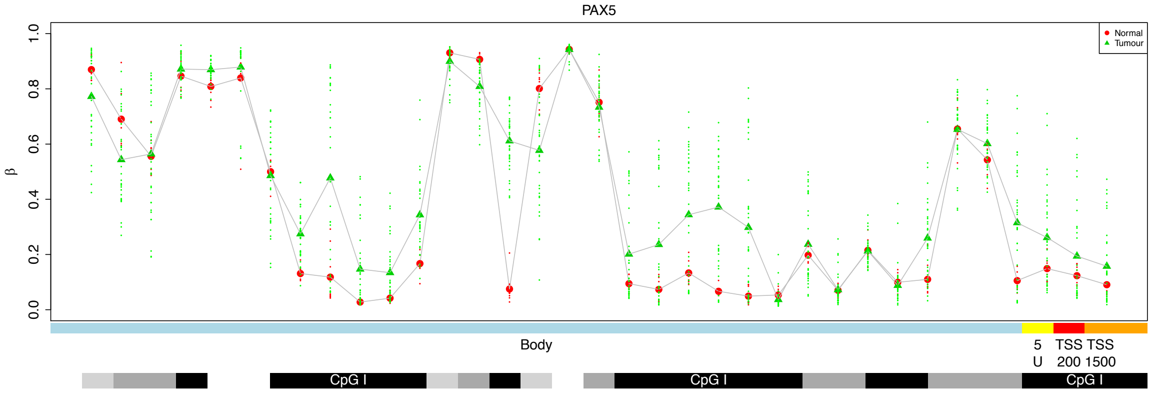

NRG1
β
0.0
0.2
0.4
0.6
0.8
1.0
Normal
Tumour
TSS
1500
5
U
1
E
Body
TSS
200
CpG I
CpG I

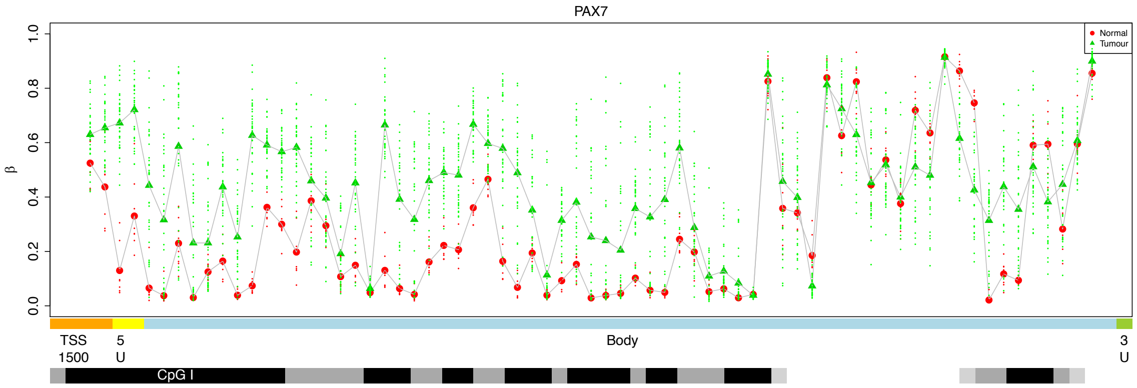

NTRK1
β
0.0
0.2
0.4
0.6
0.8
1.0
Normal
Tumour
TSS
1500
Body
3
U
CpG I

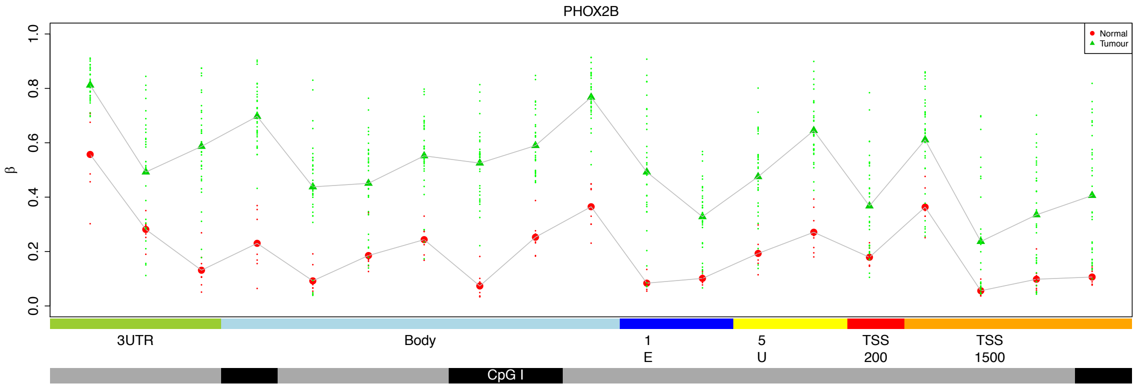

NTRK3
β
0.0
0.2
0.4
0.6
0.8
1.0
Normal
Tumour
Body
5
U
TSS
200
TSS
1500
CpG Island

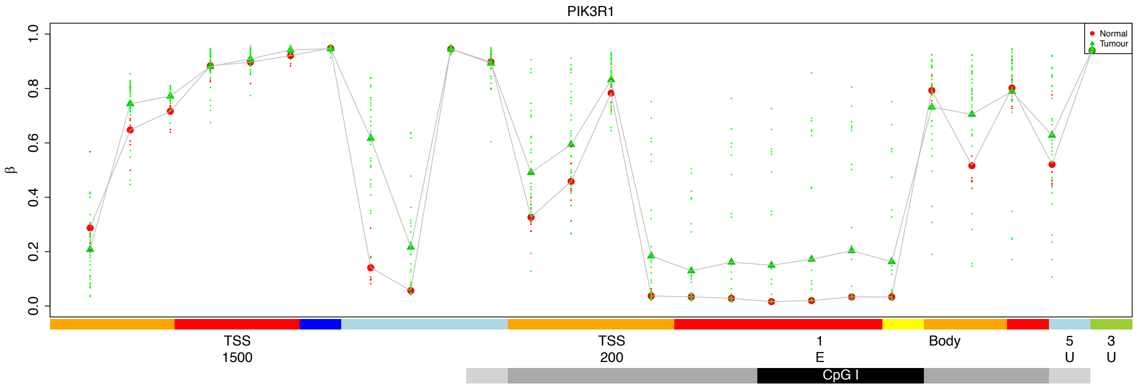

OLIG2
β
0.0
0.2
0.4
0.6
0.8
1.0
Normal
Tumour
TSS
1500
TSS
200
5
U
B
3UTR
CpG Island

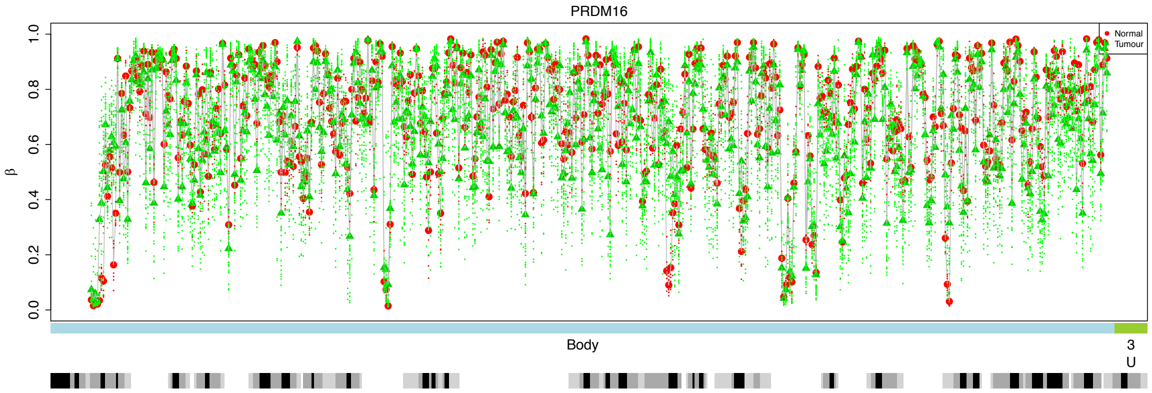

PAX3
β
0.0
0.2
0.4
0.6
0.8
1.0
Normal
Tumour
3
U
Body
CpG I
CpG Island

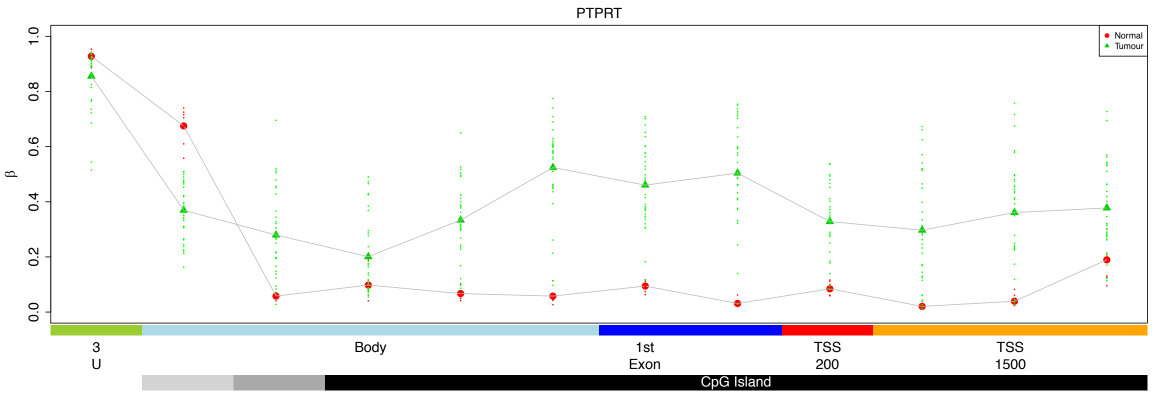

PAX5
β
0.0
0.2
0.4
0.6
0.8
1.0
Normal
Tumour
Body
5
U
TSS
200
TSS
1500
CpG I
CpG I
CpG I

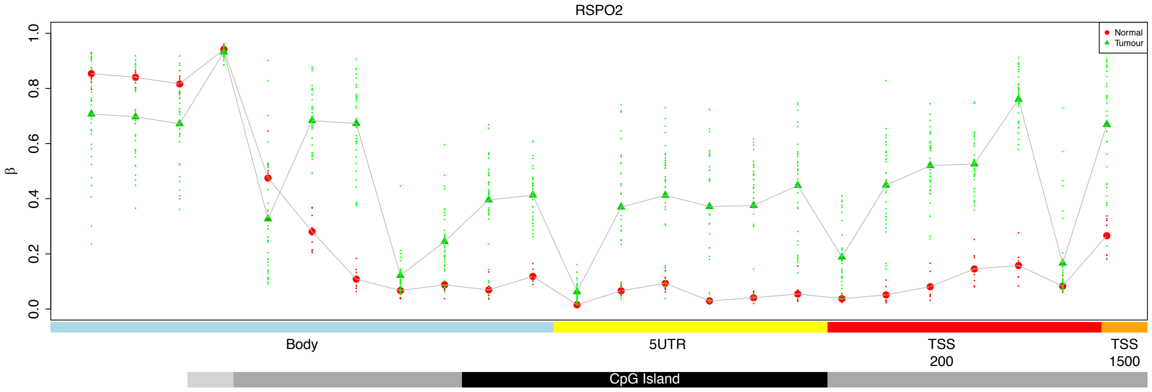

PAX7
β
0.0
0.2
0.4
0.6
0.8
1.0
Normal
Tumour
TSS 1500
5 U
Body
3 U
CpG I

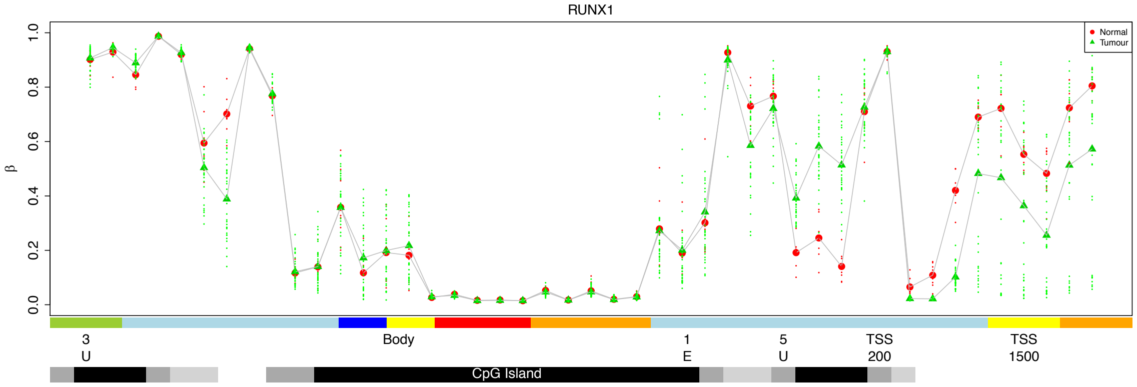

PHOX2B
β
0.0
0.2
0.4
0.6
0.8
1.0
Normal
Tumour
3UTR
Body
1 E
5 U
TSS 200
TSS 1500
CpG I

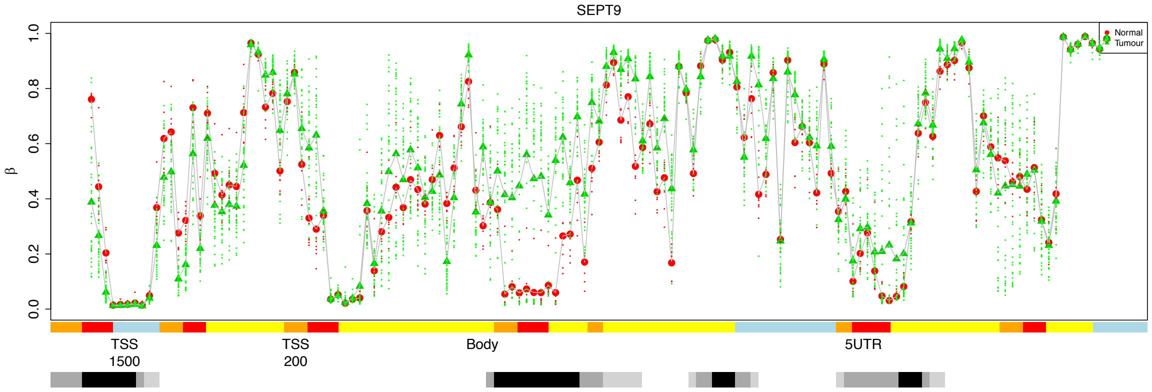

PIK3R1
β
0.0
0.2
0.4
0.6
0.8
1.0
Normal
Tumour
TSS 1500
TSS 200
1 E
Body
5 U
3 U
CpG I

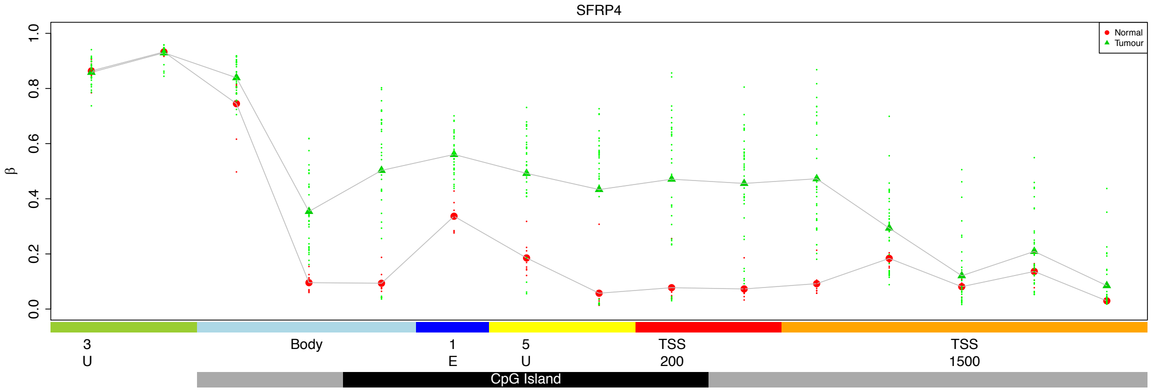

PRDM16
β
0.0
0.2
0.4
0.6
0.8
1.0
Normal
Tumour
Body
3 U

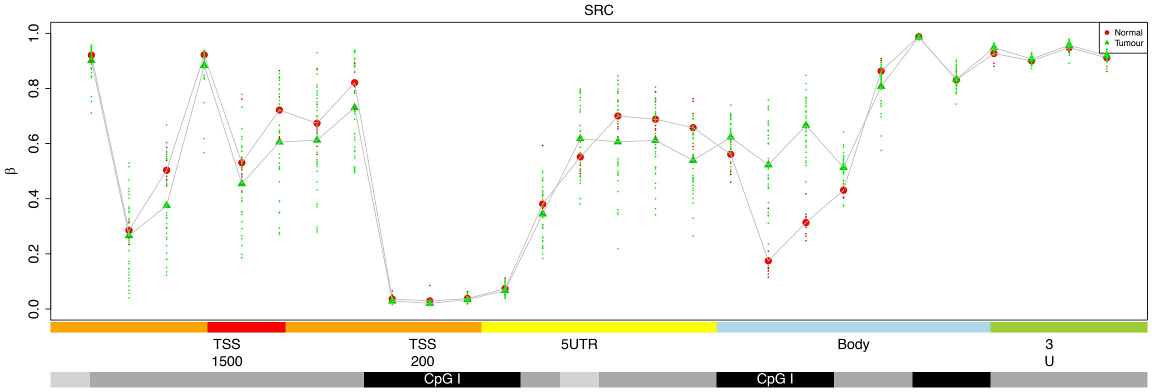

PTPRT
β
0.0
0.2
0.4
0.6
0.8
1.0
Normal
Tumour
3
U
Body
1st
Exon
TSS
200
TSS
1500
CpG Island

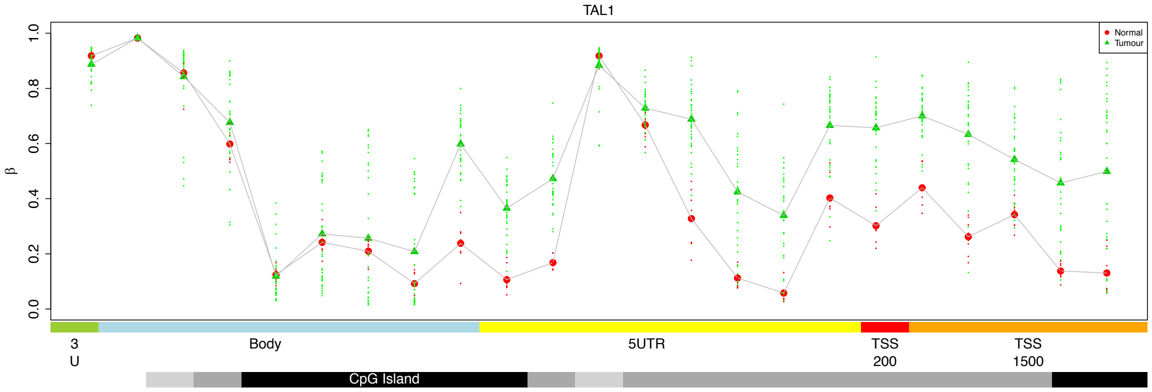

RSPO2
β
0.0
0.2
0.4
0.6
0.8
1.0
Normal
Tumour
Body
5UTR
TSS
200
TSS
1500
CpG Island

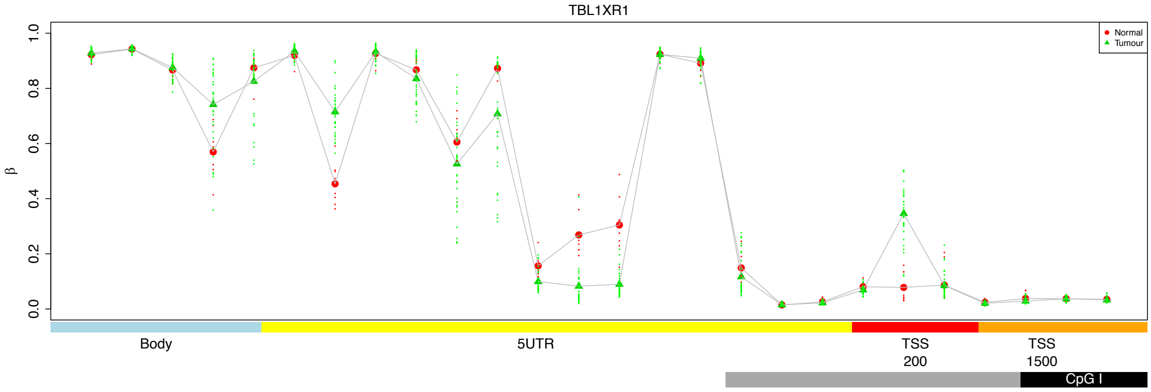

RUNX1
β
0.0
0.2
0.4
0.6
0.8
1.0
Normal
Tumour
3
U
Body
1
E
5
U
TSS
200
TSS
1500
CpG Island

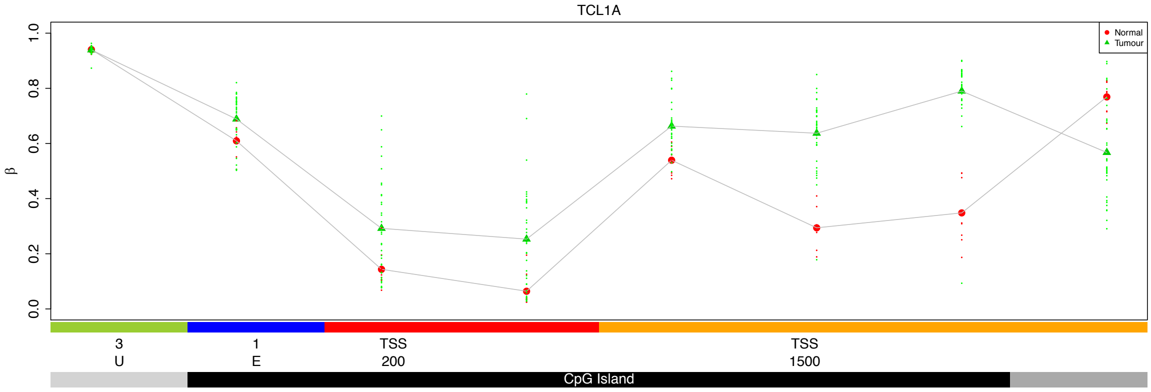

SEPT9
β
0.0
0.2
0.4
0.6
0.8
1.0
Normal
Tumour
TSS
1500
TSS
200
Body
5UTR

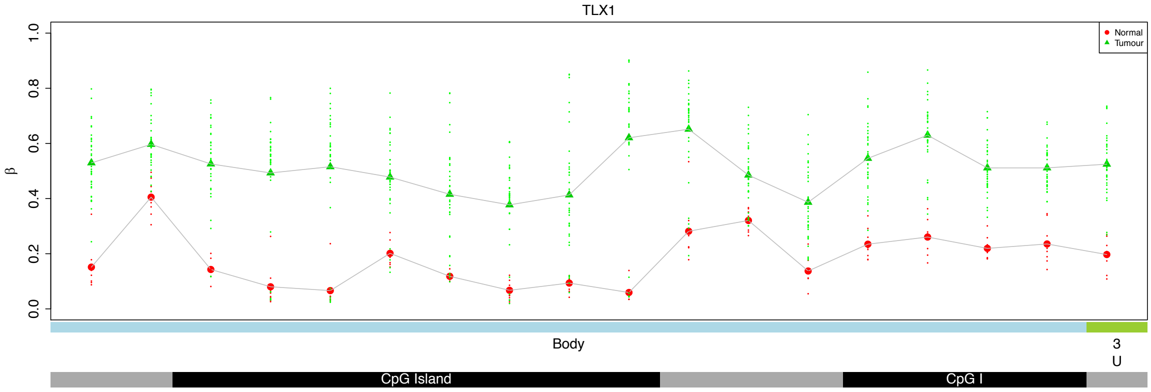

SFRP4
Normal
Tumour
β
1.0
0.8
0.6
0.4
0.2
0.0
3
U
Body
1
E
5
U
TSS
200
TSS
1500
CpG Island

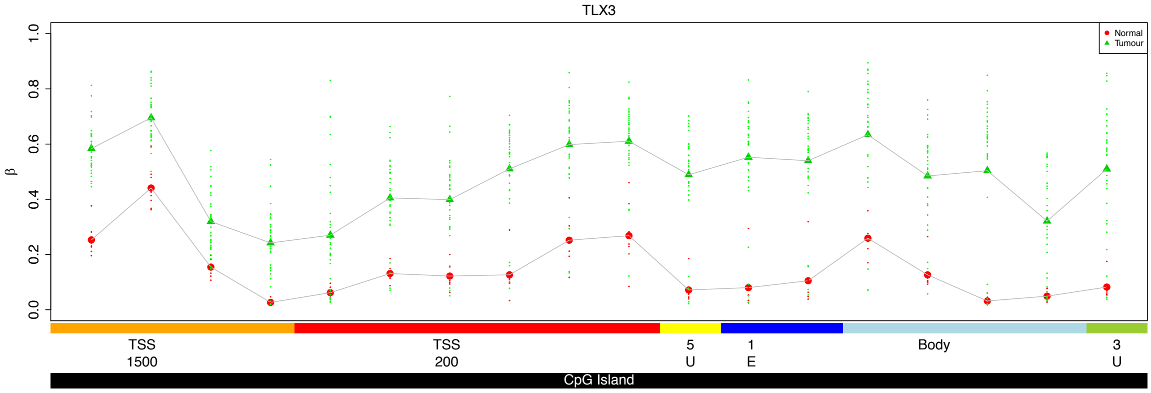

SRC
Normal
Tumour
β
1.0
0.8
0.6
0.4
0.2
0.0
TSS
1500
TSS
200
5UTR
Body
3
U
CpG I
CpG I

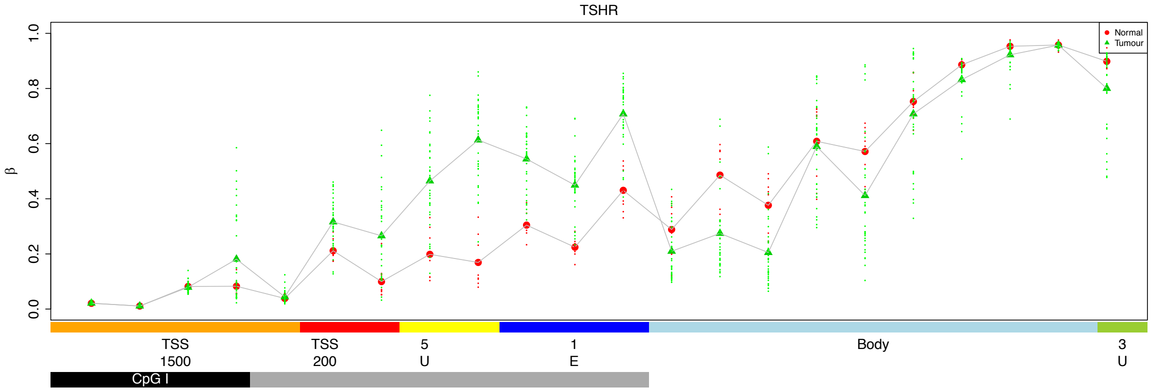

TAL1
Normal
Tumour
β
1.0
0.8
0.6
0.4
0.2
0.0
3
U
Body
5UTR
TSS
200
TSS
1500
CpG Island

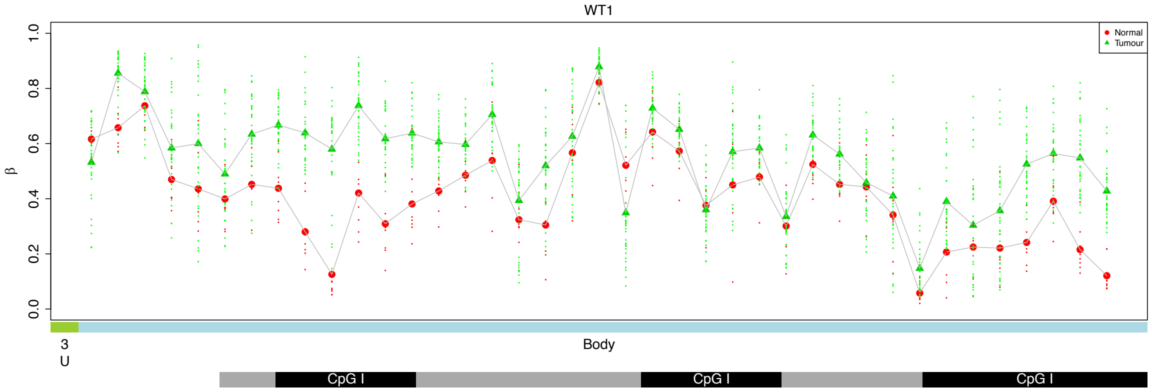

TBL1XR1
β
0.0
0.2
0.4
0.6
0.8
1.0
Normal
Tumour
Body
5UTR
TSS 200
TSS 1500
CpG I

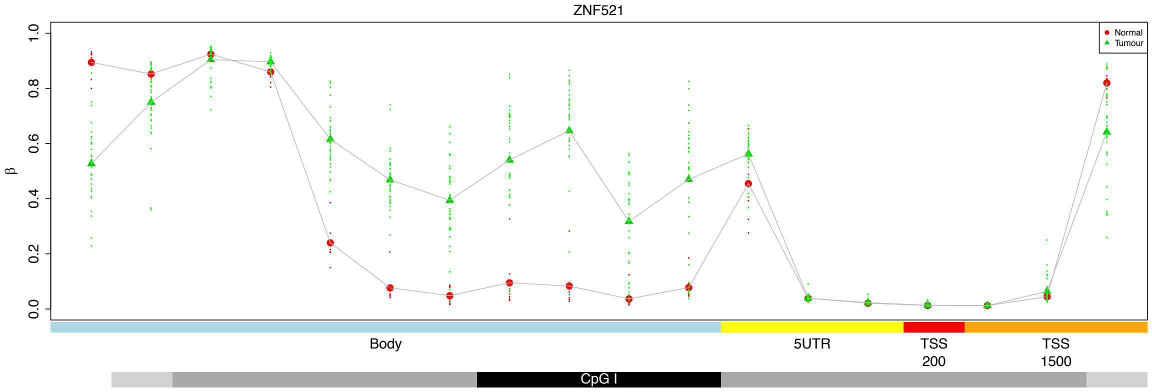

TCL1A
β
0.0
0.2
0.4
0.6
0.8
1.0
Normal
Tumour
3 U
1 E
TSS 200
TSS 1500
CpG Island


TLX1
β
0.0
0.2
0.4
0.6
0.8
1.0
Normal
Tumour
Body
3 U
CpG Island
CpG I


TLX3
β
0.0
0.2
0.4
0.6
0.8
1.0
Normal
Tumour
TSS 1500
TSS 200
5 U
1 E
Body
3 U
CpG Island


TSHR
β
0.0
0.2
0.4
0.6
0.8
1.0
Normal
Tumour
TSS
1500
TSS
200
5
U
1
E
Body
3
U
CpG I


WT1
β
0.0
0.2
0.4
0.6
0.8
1.0
Normal
Tumour
3
U
Body
CpG I
CpG I
CpG I


ZNF521
β
0.0
0.2
0.4
0.6
0.8
1.0
Normal
Tumour
Body
5UTR
TSS
200
TSS
1500
CpG I

8.4 DMRs found in driver DMPs

Gene name	DMR coordinates	Number of CpGs	Adjusted p-value	DMR direction	Within a promotor?
SEPT9	chr17:75385086-75385432	3	< 0.00001	Hypermethylation	FALSE
SEPT9	chr17:75405842-75406074	2	< 0.00001	Hypermethylation	FALSE
SEPT9	chr17:75315081-75315244	3	< 0.00001	Hypermethylation	FALSE
AXIN2	chr17:63553581-63556315	6	< 0.00001	Hypermethylation	TRUE
CBFA2T3	chr16:89006877-89008134	5	< 0.00001	Hypermethylation	TRUE
CDX2	chr13:28540622-28543520	13	< 0.00001	Hypermethylation	TRUE
EBF1	chr5:158526263-158528040	11	< 0.00001	Hypermethylation	TRUE
EBF1	chr5:158522427-158522899	3	< 0.00001	Hypermethylation	FALSE
EBF1	chr5:158524270-158524649	3	< 0.00001	Hypermethylation	FALSE
ERBB4	chr2:213400697-213402433	10	< 0.00001	Hypermethylation	FALSE
FOXL2	chr3:138654993-138659021	21	< 0.00001	Hypermethylation	FALSE
GAS7	chr17:10101010-10101195	2	< 0.00001	Hypermethylation	FALSE
GPC3	chrX:133118088-133119308	5	< 0.00001	Hypermethylation	FALSE
GRIN2A	chr16:10133433-10133501	3	< 0.00001	Hypermethylation	FALSE
MECOM	chr3:169376298-169376618	3	< 0.00001	Hypermethylation	FALSE
MECOM	chr3:169377725-169379010	8	< 0.00001	Hypermethylation	FALSE
NCOR2	chr12:124873347-124874007	3	< 0.00001	Hypermethylation	FALSE
NCOR2	chr12:124990897-124991139	4	< 0.00001	Hypermethylation	FALSE
NFATC2	chr20:50157455-50158996	4	< 0.00001	Hypermethylation	FALSE
NR4A3	chr9:102590743-102591302	2	< 0.00001	Hypermethylation	FALSE
NTRK3	chr15:88798331-88801474	18	< 0.00001	Hypermethylation	TRUE
OLIG2	chr21:34395093-34402565	30	< 0.00001	Hypermethylation	TRUE
PAX3	chr2:223176167-223177785	14	< 0.00001	Hypermethylation	FALSE
PAX3	chr2:223154140-223154201	3	< 0.00001	Hypermethylation	FALSE
PAX8	chr2:114033360-114034595	6	< 0.00001	Hypermethylation	FALSE
PHOX2B	chr4:41746614-41754857	26	< 0.00001	Hypermethylation	TRUE
PRDM16	chr1:2990490-2990678	2	< 0.00001	Hypermethylation	FALSE
PRDM16	chr1:3309864-3310911	4	< 0.00001	Hypermethylation	FALSE
PRDM16	chr1:2987332-2987961	5	< 0.00001	Hypermethylation	FALSE
PTPRT	chr20:41817011-41819125	10	< 0.00001	Hypermethylation	TRUE
RSPO2	chr8:109092720-109096151	20	< 0.00001	Hypermethylation	TRUE
RUNX1	chr21:36399146-36399540	4	< 0.00001	Hypermethylation	FALSE
SEPT9	chr17:75368902-75371764	14	< 0.00001	Hypermethylation	TRUE
SFRP4	chr7:37955508-37957021	13	< 0.00001	Hypermethylation	TRUE
TAL1	chr1:47694517-47697496	14	< 0.00001	Hypermethylation	TRUE
TLX1	chr10:102893925-102896869	16	< 0.00001	Hypermethylation	FALSE
TLX1	chr10:102898409-102900491	6	< 0.00001	Hypermethylation	FALSE
TLX3	chr5:170734312-170740937	29	< 0.00001	Hypermethylation	TRUE
TLX3	chr5:170742118-170744407	9	< 0.00001	Hypermethylation	FALSE
WT1	chr11:32454718-32461240	40	< 0.00001	Hypermethylation	TRUE
ZNF521	chr18:22927454-22931003	11	< 0.00001	Hypermethylation	FALSE
Sep-09	chr17:75229687-75229837	2	< 0.00001	Hypomethylation	FALSE
AFF3	chr2:100624627-100625353	2	< 0.00001	Hypomethylation	FALSE
CAMTA1	chr1:7122541-7123346	5	< 0.00001	Hypomethylation	FALSE
CBFA2T3	chr16:89029055-89030042	2	< 0.00001	Hypomethylation	FALSE
CBFA2T3	chr16:89098327-89098782	2	< 0.00001	Hypomethylation	FALSE
EBF1	chr5:158456264-158456317	2	< 0.00001	Hypomethylation	FALSE
GAS7	chr17:9940121-9941095	5	< 0.00001	Hypomethylation	FALSE
GAS7	chr17:9821299-9821552	2	< 0.00001	Hypomethylation	FALSE
NCOR2	chr12:124876101-124876650	3	< 0.00001	Hypomethylation	FALSE
PRDM16	chr1:3272662-3273407	3	< 0.00001	Hypomethylation	FALSE
PRDM16	chr1:3155161-3155965	3	< 0.00001	Hypomethylation	FALSE
PRDM16	chr1:2994051-2994372	2	< 0.00001	Hypomethylation	FALSE
RUNX1	chr21:36168008-36168288	2	< 0.00001	Hypomethylation	FALSE
RUNX1	chr21:36421467-36421955	6	< 0.00001	Hypomethylation	TRUE
SFRP4	chr7:37991679-37991986	2	< 0.00001	Hypomethylation	FALSE

Table 40: DMRs associated with potential driver genes when comparing the methylation of primary tumour with tissue adjacent normal samples.

Table 41: Table of the all differentially expressed genes, always found shared amongst primary tumour samples in the PenHet cohort.

Gene symbol	Log 2 fold change	Adjusted p value	COSMIC gene driver
DDR2	-1.22	1.83E-02	TRUE*
GAS7	-4.07	1.45E-28	TRUE*
MYH11	-3.15	2.24E-18	TRUE*
TERT	3.75	1.20E-07	TRUE*
ZBTB16	-3.25	5.46E-15	TRUE*
ACVR2A	-1.68	4.55E-10	TRUE*
GATA3	-2.65	5.55E-09	TRUE*
MLF1	1.82	8.28E-03	TRUE*
AADACL2	-4.98	4.87E-07	FALSE
ANGPTL1	-5.44	1.11E-24	FALSE
BCHE	-4.27	7.64E-07	FALSE
C14orf132	-1.97	5.56E-06	FALSE
C8orf48	-3.42	1.04E-13	FALSE
CAT	-1.12	8.33E-03	FALSE
CFD	-2.43	2.15E-07	FALSE
CIDEA	-5.81	1.11E-14	FALSE
CILP	-6.11	1.41E-30	FALSE
CLEC3B	-5.23	1.62E-28	FALSE
CYB5A	-2.06	1.29E-05	FALSE
CYP3A5	-2.86	1.07E-15	FALSE
DCT	-8.72	5.78E-85	FALSE
DLG2	-5.50	1.97E-56	FALSE
DOCK3	-2.61	1.48E-06	FALSE
DPP6	-4.42	4.27E-16	FALSE
ECM2	-1.98	5.27E-06	FALSE
EDN3	-8.83	8.01E-68	FALSE
EDNRB	-3.55	1.03E-16	FALSE
EFHC2	-3.73	5.01E-20	FALSE
EPHX2	-1.92	1.91E-17	FALSE
ESM1	4.65	2.51E-27	FALSE
FAM107A	-3.10	2.96E-10	FALSE
FGF7	-2.71	1.32E-12	FALSE
FLG	-5.42	7.12E-11	FALSE
GFRA1	-4.51	6.47E-13	FALSE
GSTA4	-3.15	1.34E-11	FALSE
GSTM2	-2.17	3.31E-08	FALSE
GULP1	-1.60	6.60E-03	FALSE
HPSE	-1.42	5.20E-03	FALSE
IGFBP5	-3.52	4.60E-23	FALSE
IL18	-2.94	3.13E-13	FALSE
KRT10	-5.12	2.66E-12	FALSE
KRT18	5.09	3.18E-22	FALSE
KRT77	-7.09	1.72E-26	FALSE
LOR	-4.97	8.24E-07	FALSE
MFSD7	-3.33	9.84E-19	FALSE
MYOCD	-2.86	8.45E-06	FALSE
NPY1R	-4.28	2.83E-07	FALSE
NRIP3	3.98	3.43E-31	FALSE
NUDT10	-5.11	1.01E-11	FALSE
OGN	-4.36	2.08E-15	FALSE
ONECUT2	4.15	1.24E-12	FALSE
OSM	4.89	1.63E-15	FALSE
PAQR4	1.76	5.86E-09	FALSE
PARD3B	-2.21	2.06E-06	FALSE
PCP4L1	-2.71	4.81E-04	FALSE
PCSK2	-5.11	3.17E-67	FALSE
PDCD4	-2.25	1.87E-12	FALSE
PDE8B	-3.23	5.57E-14	FALSE
PDZRN4	-6.64	6.34E-25	FALSE
PLA2G2A	-7.77	1.32E-61	FALSE
PLP1	-4.59	3.52E-11	FALSE
PTGER3	-3.72	6.26E-13	FALSE
PTGIS	-4.21	1.38E-45	FALSE
PTPRH	3.75	6.71E-11	FALSE
QPCT	-3.38	3.67E-13	FALSE
RFC2	0.88	9.85E-04	FALSE
SGCG	-5.06	2.64E-06	FALSE
SLC16A3	3.85	5.69E-25	FALSE
SLC27A6	-6.13	1.28E-23	FALSE
SLC5A1	-2.53	8.37E-05	FALSE
SLC6A4	-3.52	1.57E-31	FALSE
SMOC2	-3.16	3.46E-17	FALSE
SOSTDC1	-5.93	6.67E-12	FALSE
SOX10	-5.06	1.55E-24	FALSE
STX12	-0.79	6.01E-04	FALSE
STXBP6	-4.56	4.17E-12	FALSE
TAC1	-5.92	2.19E-05	FALSE
TCEAL5	-4.36	2.12E-02	FALSE
TCP11	2.96	5.83E-03	FALSE
TFAP2B	-5.67	1.99E-07	FALSE
TYR	-6.65	1.38E-42	FALSE
TYRP1	-8.02	9.21E-49	FALSE
UPK1A	-5.05	5.95E-15	FALSE
WISP2	-5.54	2.01E-40	FALSE
ZNF135	-3.34	2.79E-32	FALSE
ZNF439	-2.49	1.89E-34	FALSE
ZNF471	-2.31	2.26E-31	FALSE
AADAC	-6.45	7.81E-30	FALSE
ACER1	-4.82	3.67E-09	FALSE
ACKR1	-3.86	7.50E-10	FALSE
ACSM3	-2.44	1.29E-12	FALSE
ACY3	3.99	2.54E-15	FALSE
ADGRD1	-1.84	4.87E-07	FALSE
AMTN	7.50	1.44E-12	FALSE
ANKRD30BP3	-3.98	4.19E-08	FALSE
ANO10	-1.12	4.74E-22	FALSE
APOB	-4.23	3.52E-10	FALSE
ARSF	-5.00	4.52E-12	FALSE
ARSFP1	-5.15	5.95E-15	FALSE
ASAP3	-2.41	1.57E-19	FALSE
BAALC-AS2	-5.20	3.21E-15	FALSE
BLMH	-2.89	7.59E-17	FALSE
BNIPL	-2.95	8.03E-07	FALSE
C11orf24	0.92	8.20E-03	FALSE
C15orf59	-5.19	3.16E-29	FALSE
C5orf46	-6.50	8.40E-24	FALSE
CABLES1	-1.63	6.44E-04	FALSE
CBX3P7	-5.94	9.41E-20	FALSE
CD151	1.12	3.42E-06	FALSE
CDX4	-6.20	3.86E-16	FALSE
CES4A	-1.84	6.39E-04	FALSE
CHAD	-2.06	2.89E-07	FALSE
CLDN8	-6.83	6.45E-11	FALSE
CSNK2A2	-1.59	1.01E-29	FALSE
CYP2J2	-2.66	3.10E-13	FALSE
CYP4F22	-3.26	3.53E-04	FALSE
CYP4Z1	-2.98	7.77E-06	FALSE
CYP4Z2P	-6.27	1.56E-12	FALSE
DAPL1	-4.74	3.34E-08	FALSE
DEGS2	-2.84	5.22E-06	FALSE
DMBT1P1	-5.86	3.35E-33	FALSE
DNAH8	-6.93	7.51E-13	FALSE
DNASE1L2	-4.84	2.09E-25	FALSE
DUSP13	-3.30	2.70E-07	FALSE
EFCC1	-2.45	3.47E-05	FALSE
ELOVL1	-1.29	9.86E-05	FALSE
ENDOU	-4.26	1.18E-12	FALSE
EPHA10	1.44	8.28E-03	FALSE
EPHX3	-2.40	1.77E-03	FALSE
FAAHP1	-3.33	3.95E-12	FALSE
FABP7	-6.92	5.63E-38	FALSE
FAM153B	-3.84	1.96E-28	FALSE
FRG2HP	-6.89	9.20E-09	FALSE
GALNTL6	-2.99	4.47E-12	FALSE
GAN	-2.75	8.56E-39	FALSE
GDF7	-3.81	7.42E-13	FALSE
GPR149	-5.93	9.36E-15	FALSE
GRHL1	-1.16	2.02E-02	FALSE
GULOP	-3.09	1.49E-09	FALSE
HMCN2	-3.23	1.40E-16	FALSE
ID4	-4.46	5.19E-27	FALSE
IFNAR2	1.22	1.72E-13	FALSE
IGSF9B	-4.30	9.35E-11	FALSE
JAKMIP3	-1.74	1.42E-02	FALSE
KCNA1	-2.76	2.33E-05	FALSE
KCNJ13	-3.15	1.53E-06	FALSE
KIAA0319	-3.02	1.82E-13	FALSE
KRT1	-4.20	5.32E-06	FALSE
KRT2	-5.57	2.22E-10	FALSE
KRTDAP	-3.23	6.67E-04	FALSE
LAMB4	-6.52	3.85E-36	FALSE
LCE1B	-5.73	2.36E-13	FALSE
LCE6A	-5.06	7.55E-08	FALSE
LINC00346	3.06	9.21E-11	FALSE
LRMP	-3.47	1.31E-17	FALSE
LRRN4CL	-2.57	6.44E-28	FALSE
MLANA	-4.73	5.60E-50	FALSE
MSMB	-4.87	8.03E-09	FALSE
NFE2	-1.71	6.83E-03	FALSE
NMB	2.68	2.51E-09	FALSE
NPAS1	-2.73	2.11E-14	FALSE
NTS	-3.78	3.42E-18	FALSE
ODF4	-3.71	1.79E-06	FALSE
OTC	-3.86	1.21E-04	FALSE
OVCH2	-3.93	1.24E-12	FALSE
P2RY4	-3.74	8.49E-23	FALSE
PACRG	-2.08	4.34E-09	FALSE
PALMD	-1.49	1.04E-03	FALSE
PANK1	-2.43	6.04E-19	FALSE
PARK2	-2.29	3.85E-15	FALSE
PCDHB3	-3.81	3.10E-13	FALSE
PCDHGA12	-2.84	1.11E-08	FALSE
PCDHGB7	-2.12	1.13E-07	FALSE
PCNX1	0.81	1.00E-02	FALSE
PHF2P2	-8.07	3.82E-22	FALSE
PHYHIP	-4.17	3.79E-14	FALSE
PLPPR1	-6.12	2.88E-09	FALSE
PMEL	-5.76	9.51E-56	FALSE
PON3	-2.75	6.61E-05	FALSE
POU2F3	-3.24	1.14E-10	FALSE
POU3F3	-5.85	2.98E-23	FALSE
PREP	-0.58	3.42E-02	FALSE
PSAPL1	-6.09	2.07E-15	FALSE
PSORS1C2	-4.18	8.90E-10	FALSE
PYDC1	-7.07	6.34E-25	FALSE
RBFOX1	-4.26	2.25E-02	FALSE
RPL7AP64	-3.76	1.49E-10	FALSE
RPS6KA6	-7.70	1.67E-43	FALSE
RRM2P3	-3.92	1.53E-12	FALSE
SCN11A	-4.20	1.84E-21	FALSE
SDSL	2.30	2.18E-07	FALSE
SERPINA12	-4.97	1.40E-06	FALSE
SERPINB12	-6.24	7.64E-16	FALSE
SLC24A5	-6.33	2.60E-46	FALSE
SLC25A15P1	-6.52	2.36E-13	FALSE
SLC45A2	-5.77	5.40E-14	FALSE
SLC46A2	-5.01	1.05E-09	FALSE
SLC5A8	-3.76	1.30E-08	FALSE
SLITRK2	-5.66	7.15E-11	FALSE
SPINT3	-4.92	3.67E-03	FALSE
SPTSSB	-3.67	4.28E-06	FALSE
SUGCT	3.39	1.37E-12	FALSE
TACR1	-3.58	1.73E-09	FALSE
THEM5	-3.48	4.84E-25	FALSE
THRB	-2.06	4.07E-08	FALSE
TMEM200B	2.34	3.77E-09	FALSE
TMEM45A	-2.83	2.01E-06	FALSE
TMEM45B	-2.00	1.37E-02	FALSE
TMEM99	-2.44	7.61E-40	FALSE
TNFRSF19	-3.90	1.73E-46	FALSE
TPRG1	-2.28	7.26E-06	FALSE
TRIM46	2.25	4.08E-10	FALSE
TRPM1	-6.39	1.25E-17	FALSE
UBA52P6	3.35	2.90E-09	FALSE
USP24P1	-5.57	8.52E-04	FALSE
USP31	1.11	1.85E-05	FALSE
VWA3A	-4.12	1.05E-12	FALSE
VWC2	-3.87	2.19E-23	FALSE
WNT3	-4.50	6.20E-37	FALSE
WNT9B	-3.05	1.02E-20	FALSE
XCR1	-3.10	2.25E-10	FALSE
XG	-1.48	1.72E-02	FALSE
ZNF582	-3.14	1.45E-10	FALSE
ZNF626	-2.75	1.21E-16	FALSE
ZNF677	-2.28	2.39E-21	FALSE
ZSCAN18	-3.30	9.86E-38	FALSE