Network based analysis to identify master regulators in prostate carcinogenesis

Mario Cangiano

Submitted in fulfilment of the requirements for the Degree of Doctor of Philosophy in Cancer Studies

Institute of Cancer Sciences

College of Medical, Veterinary and Life Sciences

University of Glasgow

March 2022

Table of contents

List of tables

List of figures

Acknowledgements

First of all, I would like to express my gratitude to the European Commission and the Marie Skłodowska-Curie program for Early Stage Researchers, for giving me the opportunity to pursue a PhD research project in prostate cancer. My experience in conducting a timely and impactful project was made possible by the excellent organisational infrastructure, provision of relevant expertise and availability of computational resources and datasets at GenomeScan B.V., my Dutch host institution, and the Beatson Institute for Cancer Research in Glasgow.

Crucial for my personal growth as a researcher was the supervision offered by Professor Hing Leung. With immense patience and understanding, he taught me the importance of rigor in both the research practice and the interpretation of the findings. I also very much appreciated the ability of, and input from my supervisory team at GenomeScan: Zoraide Granchi, Bart Janssen, Inès Beumer and Magda Grudniewska-Lawton who, collectively and constantly, motivated me and provided my work a solid structure through carefully arranged portfolio of the research activities.

Additional acknowledgements are due to the many colleagues I encountered along the journey, including the other Early Stage Researchers within the TransPot network across Europe, with whom I shared quality time during training and dissemination events; to all the principal investigators within the European consortium for their willingness to provide valuable advice and data for my research; GenomeScan's employees who made me feel at home in the Netherlands from day one; the laboratory staff at the Beatson institute, in particular Linda Rushworth and Rafael Sanchez Martinez, for the help with the analysis of the data and the bench validation of my *in-silico* results.

Moreover, my analysis would have not been feasible without the high quality transcriptomics and proteomics data provided by Dr Mark Salji (Beatson Institute and University of Glasgow), as a result of his work on prostate cancer orthografts

(see Chapters 3 and 4). I also want to express my gratitude to Prof. Ian Mills (University of Oxford and Queen's University Belfast) for generously sharing RNAseq data generated from intervention experiments in a preclinical prostate cancer subcutaneous xenograft model (see Chapter 5).

Finally, special thanks to Janneke, whose love and support pushed me safely to the end of the journey. Similarly, I cannot refrain mentioning my parents and family who sustained me every day while working from abroad.

Abstract

Prostate cancer (PCa) is the second most common tumor diagnosed in man, for which robust prognostic markers and novel targets for therapy are lacking. Major challenges in PCa therapeutical management arise from the marked intra- and inter-tumors heterogeneity, hampering the discernment of molecular subtypes that can be used to guide treatment decisions. For this reason, virtually all patients undergoing standard of care androgen deprivation therapy for locally advanced or metastatic cancer, will eventually progress into the more aggressive and currently incurable form of PCa, referred to as castration resistant prostate cancer (CRPC).

By exploiting the richness of information stored in gene-gene interactions, I tested the hypothesis that a gene regulatory network derived from transcriptomic profiles of PCa orthografts can reveal transcriptional regulators to be subsequently adopted as robust biomarkers or as target for novel therapies. Among the 1308 regulons identified from the preclinical models, Cox regression analysis coherently associated *JMJD6* regulon activity with disease-free survival in three clinical cohorts, outperforming three published prognostic gene signatures (TMCC11, BROMO-10 and HYPOXIA-28). Given its potential role in a number of cancers, in-depth investigations of JMJD6 mediated function in PCa is warranted to test if it has a driver role in tumor progression.

Encouraged by the predictive abilities of the gene regulatory network inferred from transcriptomics data, I explored the possibility of integrating the regulons structure with data from the proteomes of the same preclinical orthografts studied by RNA sequencing. This approach leverages the complementarity between gene and protein expression, to increase the robustness of the statistical analysis. Similar to gene-gene co-expression profiles, protein-protein co-expression data can provide a distinct representation of the molecular alterations underlying a biological phenotype. By implementing a pipeline to

integrate modules derived from transcriptomic based regulons and protein-protein interactions respectively from matched RNA-seq and quantitative proteomic data, I obtained 516 joint modules entailing a median of four protein complexes (range 1-41) per individual transcription factor regulon, providing new insight into its regulatory mechanisms. In the final step of the analysis, a permutation-based enrichment of the genes/proteins integrative modules implicated *MID1* (an E3 ubiquitin ligase belonging to the family of tripartite motif containing protein) to be a driver transcriptional regulator in CRPC. In fact, MID1 module was the only candidate for which gene-gene and protein-protein interactions were supported (p-value < 0.05) by both differentially expressed genes and proteins obtained from the CRPC vs PC contrast.

Finally, I wished to test the usefulness of a network based investigation as a tool to identify predictors of treatment response. To this end, I obtained transcriptomics data from an *in vivo* subcutaneous xenograft treatment experiment (namely mychophenolic acid or abiraterone/ARN-509 as stand alone treatment or in combination) and determined which regulons were inferred to be active in the tumours following treatment. The androgen receptor positive human LNCaP C4-2b prostate cancer cells were injected into mice. The effects of treatment were assessed by collecting serial tumor sizes and by performing RNAseq at the designed endpoint of the study

Noteworthy, the gene graph enrichment analysis provided novel hypothesis behind the anti-proliferative effect of mychophenolic acid (MPA), suggesting the *SET* proto-oncogene to be a target for MPA mediated suppression of proliferation. Of note, standard gene-set enrichment analysis, without input on specific gene-gene interactions, was not effective in prioritising the *SET* proto-oncogene, demonstrating the usefulness of the network based investigation.

Collectively, data presented in this thesis provides an alternative perspective for the analysis of multi-omics profiles from PCa and highlights the importance of gene-gene and protein-protein interactions in prostate cancer growth and progression.

Chapter 1 – Introduction

1.1 Prostate cancer

1.1.1 Epidemiology

Prostate cancer (PCa) is the second most frequent cancer diagnosed in males. Both its incidence and mortality rate are influenced by age (in particular if higher than 65 years) and ethnicity, with African-American men being the most affected by the disease[1]. Other relevant factors associated with PCa include family history, diet and inflammation.

Seven genomic regions were predicted as susceptibility loci for PCa: *HPC1* and *ELAC2* genes, involved in the innate immune defense and TGF-beta signaling pathway, respectively; the macrophage scavenger receptor 1 (*MSR1*) gene; *BRCA1*, *BRCA2* and *PALB2* genes associated with the aggressive form of the disease; small deletions in Xq26.3-q27.3 region noted in both sporadic and hereditary forms of the disease[1].

Concerning diet and metabolism, high intake of omega-6 fats, frequent consumption of red meat, high intake of calcium and low intake of vegetables were associated with a greater risk of PCa. Obesity is associated with more aggressive forms of the disease, probably due to the altered levels of metabolic and steroid hormones in circulation. On the contrary, the usage of metformin, a common hypoglycemic drug prescribed in the management of type II diabetes, was shown to reduce the risk of PCa diagnoses[2].

Lastly, inflammation, independently of its source, seems to be associated with increased risk of PCa, putatively by causing proliferative inflammatory atrophy. This can, in turn, develop into prostatic intraepithelial neoplasia, a known precursor of PCa[1].

1.1.2 Diagnosis

The main methodologies used for PCa diagnosis are digital rectal examination (DRE) and prostate-specific antigen (PSA) blood test, followed by transrectal ultrasound guided biopsy (TRUS).

In particular, total PSA (tPSA) is the most frequently used measurement for risk assessment in PCa, despite its specificity for the detection of prostate cancer being undermined by comparable levels of tPSA resulting from benign pathologies of the prostate including BPH (or benign prostatic hyperplasia) and prostatitis[3]. For this reason, the biomarker research is focusing on PSA test alternatives, by taking advantage, above all, of the advancements in genomics and proteomics technologies, with the aim to develop and implement personalised medicine strategies.

In the event of a positive result from the DRE or PSA test, 10 to 12 tissue samples are obtained by TRUS. Subsequently, a pathologist defines a primary and a secondary Gleason grade from the main patterns observed in the microscopical investigation of the cells. The Gleason scores is determined from the two most dominant morphologies, as ranging from three to five, according to the degree of abnormality of the tissue, and the sum of the two grades producing the overall Gleason sum score of the tumour. Combining Gleason scores, the clinical tumour stage and the PSA levels, individual tumours can be categorised according to a five-tiers (very low risk, low risk, intermediate risk, high risk, very high risk) based classification scheme[4].

Noteworthy, despite the accurate profiling of the tissue biopsies, PCa tumours are characterised by marked heterogeneity, namely distinct morphological and phenotypical profiles observable in different tumour cells, and multifocality, that is the presence of more than one independent cluster of tumour cells[5]. An additional limitation in the current diagnosis procedures relies in the inability to detect high-grade prostatic intraepithelial neoplasia (HGPIN), the most likely precursor of Pca. In fact, HGPIN does not induce a relevant increase in PSA concentration and leads to the development of adenocarcinomas, the most common type of Pca, within 10 years[6].

1.1.3 Primary treatment

Pca treatment strategies are typically guided by the extent of local and/or distant metastatic invasion within the body. Localised tumours, in other words tumours confined within the prostate gland and lacking an identifiable lymph node or distant metastasis on staging investigations (typically including isotopic bone scan, magnetic resonance imaging and/or computerised tomography), are suitable for radical local therapies. These mostly consist of radical surgery or radiation based therapies[4].

Radical prostatectomy (RP) entails the removal of both the prostate gland and the surrounding tissue for a thorough elimination of the tumour, with the intent of preserving continence and potency, if possible[7].

Radiotherapy (RT), either in the form of external radioactive beams (EBRT) or irradiation from within the patient's body (brachytherapy), hampers cancer cell division and offers similar cure rates to RP, although with reduced side effects. In the event comorbidity risk factors make patients less suitable for surgery, a combination of EBRT, brachytherapy and other non-surgical treatment options can be considered[8].

Expectant management consists either of watchful waiting, during which symptoms are only treated with palliation, or active surveillance, which involves PSA testing, MRIs, and biopsies to detect the earliest sign of cancer progression. These latter will trigger a change of management plan from active surveillance to curative intervention with radical surgery or radiation, in fact delaying intervention and thus avoiding potential side effects[4].

For metastatic prostate cancer, the first-line treatment is androgen deprivation therapy (ADT)[4], also known as hormonal therapy. Luteinising Hormone-Releasing Hormone (LHRH) agonists are currently the standard of care in hormonal therapy, replacing surgical castration after acknowledging better performances in terms of reversibility, physical and psychological discomfort.

Moreover, recent clinical trials suggest the adoption of the combined use of docetaxel, a chemotherapeutic inhibiting cell division, together with ADT as the new standard treatment for men showing metastasis at the time of the first investigation[9].

Lastly, several randomised controlled clinical trials have evaluated the application of both therapeutic methodologies for particular PCa cases. These studies led to new targeted recommendations, such as: ADT combined with radiotherapy for intermediate and high risk patients[8]; local therapy (without ADT) for lymph-node positive disease and for men with a limited number of metastasis[4]; EBRT supplemented with long-term ADT for local tumours with invasiveness abilities[10].

Despite the growing body of knowledge, it is still challenging to distinguish between patients who would benefit from active surveillance, definitive primary treatment or those who require treatment escalation[11].

1.1.4 Progression and second-line therapy

Despite the advancements in the management of PCa and initial control of the disease, none of the patients undergoing local therapy are cured[12]

, given that the tumour can progress either biochemically, when experiencing two consecutives rises of PSA within the castrate environment[9], or clinically, when a worsening of the patient's conditions occurs with or without any increase in PSA secretion[13].

The average time of PCa progression is five years after both RP and RT in clinically localised diseases[13], and two to three years in patients receiving ADT, leading to the development of the castration-resistant prostate cancer (CRPC)[14].

CRPC is an incurable and aggressive form of prostate cancer, as demonstrated by the average time to death of 22 months after the tumours develop resistance to chemical castration[12]. Several mechanisms of resistance development have been elucidated and, above all, the constitutive activation of the androgen receptor (AR) signaling, through either AR amplification/mutations/variants, activation via alternative pathways or intra-tumoral production of the androgens, remains a crucial driver of disease progression[15].

In order to counteract castration-induced compensatory mechanisms in CRPCs, second line therapies involving androgen-pathway targeting agents have been validated in clinical trials as either standalone or adjuvant treatment in combination with ADT in hormone-independent tumours[16]. Despite able to increase patient survival by months, these drugs are still not leading to a complete remission of the disease[17].

Among this drugs category we find nonsteroidal antiandrogens (NSAA) such as: enzalutamide and apalutamide, acting as competitive binders of androgens to the AR, prevent AR translocation into the nucleus and inhibit AR binding to chromosomal DNA[18,19]; abiraterone acetate, which hampers androgen synthesis by inhibiting the cytochrome P450 Family 17 Subfamily A (CYP17)[20].

1.1.5 Molecular characterisation of PCa

Multi-omics analyses have revealed both the key drivers of PCa development and the molecular subtypes guiding the selection of therapeutic interventions.

The *TMPRSS2-ERG* fusion is the most common genomic alteration in PCa, it is observed in 40% to 50% of tumour foci[21]. The fusion events were found via the cancer outlier profile analysis that detected the overexpression of ERG, located in the same chromosome as TMPRSS2[22]. ERG overexpression in prostate tumors activates C-MYC and inhibits prostate epithelial differentiation[23].

Further, aggressive primary and metastatic tumours showed high levels of copy number alterations across the whole genome. Lastly, despite not being characterised by high frequency mutational hotspots in the genomic DNA, the most mutated genes in PCa are *SPOP*, *TP53*, *FOXA1* and *PTEN*. As a result of the multi-omic analysis, major subtypes could be defined according to fusions involving members of the ETS family (*ERG*, *ETV*, *ETV4* and *FLI1*) and mutations of *SPOP*, *FOXA1* and *IDH1*[24].

For metastatic CRPC, a specific enrichment of alterations in *TP53* and the *AR*, together with *PIK3CA/B*, *R-Spondin*, *BRAF/RAF1*, *APC*, *β-catenin* and *ZBTB16/PLZF*, was identified. Moreover, within the mCRPC cohort, the full

inactivation of *BRCA2*, *BRCA1* or *ATM* genes was observed in 20% of the samples[25].

The mitochondrial DNA was also the object of multi-omics studies, revealing both recurrent mutational hotspots and a strong association with the nuclear mutation profiles, suggesting a role for mtSNVs in PCa development[26].

Subsequently, with the intent to obtain more information from genomics analysis, integrative studies have been performed. A tumour methylation quantitative trait loci (meQTL) analysis revealed 1178 SNPs associated with altered methylation in tumour tissue, including known driver genes such as *TCERG1L* and *AKT1*[27]. A comprehensive proteomic analysis of localised prostate cancers highlighted the convergence of the above mentioned genomic subtypes of PCa into five proteomic groups with distinct clinical manifestations[28]. Further, by integrating transcriptomics and proteomics data, changes in citric acid cycle (TCA) and MDH2 activities were observed during progression into CRPC, corroborating the importance of mitochondrial alterations in PCa[29].

1.2 Cancer biomarkers research

A biomarker is defined as "a biological molecule that is fairly evaluated as an indicator of normal physiological, pathological processes or pharmacological responses to a therapeutic intervention"[30].

The typical pipeline for biomarker research foresees five phases. Initially, a preclinical study is performed to identify potential biomarkers. Prioritised models are subsequently validated in phase 2 by performing a clinical assay. At this stage, either retrospective studies, in which participants present with the condition, or prospective studies, where participants have not yet developed the disease or outcome in question, are performed in phase 3 and 4, respectively. Lastly, control studies are performed in phase 5 through population screening[30].

Currently, cancer biomarkers research is placed in the context of personalised medicine, a patient management procedure driven by the accurate molecular and morphological classification of each tumour. The switch from the "one-size-fits-all" paradigm was made possible by the development of both high

throughput technologies for the study of different biological layers, such as genetic mutations, mRNA and protein expression levels, and suitable statistical analysis applied on the accumulated data[31].

In fact, as the technology advances, providing the researchers with higher accuracy in the measurements and larger lists of identifiable molecules from the experiment, newer data analysis strategies allow clearer recognitions of patterns in the data.

In particular, the integration of multiple layers of biological information in the same pipeline, leads to more robust results than single layer analysis, boosting our understanding of complex diseases such as cancer. In greater detail, the work illustrates the usefulness of multi-omics approach in improving functional annotation of genomic alterations, discovery of new therapeutic opportunities, uncovering interactions across layers of organization, extending tumour molecular profiling and assisting early cancer diagnosis[32].

1.2.1 Biomarker types

According to the aim for which they are developed, biomarkers fall into three categories: prognostic, if informative about the patient overall cancer outcome, regardless of the therapy; predictive, when providing information about the effect of a therapeutic intervention; diagnostic, if identifying the presence of a specific condition[33].

Biomarkers can be extracted from gas, i.e. volatile substances in the breath[34], liquid, such as PSA withdrawn from blood, or tissue samples, directly biopsied from the tumour[11]. Biomarker's sensitivity and specificity, together with the invasiveness and risk of the test, are crucial parameters in the selection of the best measurement for populational screening.

Lastly, biomarkers measurements can entail single or multiple features at the same time, in either an isolated or connected way. In case of isolated measurements, the status of one or more features is assessed independently and eventually combined into a univocal prediction. For example, the same protein could be tested in both the unmodified and phosphorylated forms to leverage

the ratio as a biomarker, while a set of genes can be assessed for the presence/absence of mutations to calculate a summary score that will function as the effective biomarker. Every summary statistic will follow a specific distribution. Similarly, each marker used will influence the output statistics in an independent way. For these reasons, it's important to validate both the members of the biomarker set and the summary score.

This dissertation is focused on network-driven biomarkers research, in which mechanistic interactions among molecules are considered to derive more realistic models. In fact, the robustness of the biomarker is provided by leveraging the causal relationships existing among the studied features and, at the same time, by lowering the chances to base a model on spurious associations.

For example, the analysis of concomitant inhibitory and activating relationships among genes can identify transcriptional regulators that act as biomarkers[35], while protein-protein co-expression patterns can reveal structures and stoichiometries that can be used for the development of a biomarker[36].

1.2.2 Prostate cancer approved biomarkers

PSA was introduced as a biomarker of prostate tissue in 1980s, and its usage for PCa detection was approved by the Food and Drug Administration (FDA) in 1994.

As mentioned before, the low specificity of the blood-based test, directly resulted in high risk of over-treatment for men with indolent disease. To solve the problem, derivatives of the tPSA measurements have been developed to incorporate different forms of both serum and serum-free PSA in the same scoring formula. The prostate health index (PHI) and the 4Kscore tests are examples of more complex liquid biopsies-based biomarkers. Moreover, a number of non-PSA liquid biomarkers have been approved by the FDA with the aim of increasing diagnostic accuracy for PCa, such as: a urine-based assay for the gene product of prostate cancer antigen 3 (*PCA3*), which encodes a long noncoding RNA suggested to be specific for the malignant prostate, being

overexpressed in 95% of PCa; the Select MDx formula combining both clinical factors and a urine-based assay for RNA levels of HOXC6 and DLX1; ExoDx Prostate Intelliscores (EPI), which is a pre-biopsy RNA-based assay leveraging the expression of PCA3 and ERG isolated from urinary exosomes[11].

Despite the advancements in PCa liquid biomarkers research, tissue biopsy based assay remains the gold standard to accurately represent the tumour environment[11]. These tests were mostly derived from the analysis of high-throughput data, for example: Confirm MDx is an epigenetic assay using a multiplex methylation specific PCR to detect the DNA methylation levels of three tumour suppressor genes (*GSTP1*, *APC* and *RASSF1*); Decipher is a 22-gene panel based assay, developed from RNA microarray technology for the prediction of metastasis; Oncotype Dx, based on the assessment of the expression levels of 12 tumour-specific and five reference genes from a needle biopsy; the automated immunofluorescence-based assay Promark, which measures the expression of eight proteins[11].

There is, however, an urgent need to identify biomarkers that would allow clinicians to define more accurate risk groups. In fact, at the moment, true personalised medicine workflows are absent in the clinical management of the disease[37].

The pre-analytical and analytical validations of the new biomarkers are heterogeneous. Only the PHI and the 4Kscore were shown to be able to discriminate aggressive and indolent prostate tumors, and ready for clinical use since they add value to the classical parameters[38].

Moreover, the clinical implementation of several newly discovered biomarkers has often been hindered by the erroneous design of preclinical trials, inappropriate statistical analyses etc[39]. Lastly, PSA remains an inexpensive and sensitive biomarker for disease detection and monitoring progression and recurrence following curative therapy of local disease[40].

1.2.3 Biomarker research in PCa – the state of the art

In addition to the FDA approved omics signatures for PCa risk stratification, a plethora of biomarkers have and are still being developed to define new predictive models with higher clinical utility.

The new signatures were discovered through the detailed study of specific alterations of the prostate tumours, showing promising preliminary results. For example, a transcriptomics study revealed that bromodomain-containing proteins (BRDs) are the mediators of the increased chromatin accessibility observed in CRPC. From this analysis, the BROMO-10 signature emerged, with validated prognostic abilities for biochemical recurrence[41].

Another prognostic signature derived from gene expression profiles is based on transcriptional targets of the tumour suppressor *TMEFF2* which, in turn, regulates the cell cycle. This signature was also validated using biochemical recurrence as the clinical outcome[42]. Lastly, starting from genes differentially regulated by hypoxia in PCa cell line, Yang et al. applied a network-based approach to derive a signature prognostic of both biochemical recurrence and metastatic outcome in PCa cohorts receiving primary treatment[43]. In particular, in Yang et al, the co-expression modules were not used to identify regulons but to shortlist hypoxia related genes. Their final signature doesn't exploit gene-gene interaction strength but only the expression of the selected genes.

Despite the predominant abundance of transcriptomic analysis, PCa biomarkers research from proteomics data has increased over time. Noteworthy, a first study of 4274 protein complexes activity in tissue samples revealed a distinctive pattern of enrichment, discriminating low-grade and high-grade tumours from normal prostate. In particular, 13 integrin complexes involved in cell adhesion were found downregulated in the tumours compared to the normal prostate. Moreover, four Prothymosin alpha complexes and four subcomplexes of the mictochondrial complex I were enriched in high-grade PCa while six protein complexes involved in RNA splicing were enriched in low-grade PCa[44]. All these studies highlighted the utility of protein complexes for PCa stratification.

Nevertheless, the necessity to integrate the information provided by multiple omics became evident after comparing the results of several studies. It was shown that different omics datasets, generated within the same biological context (or even sample), are characterised by a poor correlation of the measurements. For example, mRNAs expression levels are normally not reflected by the expression levels of the corresponding proteins due to processing mechanisms such as folding, posttranslational modifications and cellular localization[45]. Therefore, rather than risking biases associated with a single layer, it is logical to take advantage of the complementarity of the information provided by different omics datasets, to obtain an holistic perspective on the disease[46].

1.3 Network-based modelling

Biological mechanisms are the result of complex interactions among several molecules, such as protein-protein, protein-DNA or mRNA-mRNA causal networks. Graph theory is the mathematical discipline that studies complex networks. Therefore, its principles find application in biological systems as well, given the non-random structure of the interaction networks characterising them[47].

A biological graph [i.e. Figure 4.6 A] presents distinct topologies, namely arrangements of the relationships (referred as edges) among the features (referred as nodes), such as “scale-free” or “small-world” configurations. Moreover, the study of the interaction patterns can reveal additional characteristics such as: modules, which are portions of the network with higher than average connectivity; key nodes, whose importance can be due to the centrality of their position in the network or from the number of edges linked to them; motifs, in other words sub-topologies that occur significantly more frequently than expected by chance[48].

For these reasons, biomarker discovery can benefit from the network-based modelling of a disease through the accumulation of evidence derived from the

assessment of the graph topology. In particular, the identification of a putative biomarker can be performed by measuring the consistencies between the observed molecular phenotype and the set of relationships described by the network.

For example, gene sets enrichment analysis entails several statistical methods to assess the non-random over-representation of disease-specific features (i.e., differentially expressed genes or proteins) within pre-determined sets of functionally related molecules. Graph-based analysis instead, consider the mechanistic interactions among the members of these sets to filter out logically inconsistent relationships. Of note, the latter were shown to perform better than interactions-unaware strategies because more supporting evidences can be extracted from the data[49].

This thesis focusses on both the exploitation of network-driven modelling of genes and proteins in PCa for biomarker discovery as well as on the development of a novel graph-based enrichment analysis procedure.

1.3.1 Gene regulatory network

A gene regulatory network (GRN) is a graph representation of the interactions of transcriptional regulators and their target genes. The nodes of the network are the genes involved in the regulation mechanisms while the edges represent either inductive or inhibitory regulations which increase or reduce the expression of the targets, respectively.

GRNs can be constructed directly from high-throughput expression data through a procedure called reverse engineering, resulting in different class of models: logical models for qualitative networks (list of causal rules of regulation); continuous models, that relate the change in expression of each target with the expression level of the possible regulators; single molecule level models, to accurately describe the sequence and hierarchy of regulations among a small number of molecules; hybrid models, harboring both continuous and discrete aspects[50]. A classical GRN continuous model called ‘ARACNE foresees the

following steps: first, gene pairs that exhibit correlated transcriptional responses are identified by measuring the mutual information (MI) between their mRNA expression profiles. Indirect interactions are eliminated by applying a property of MI called the data processing inequality (DPI). Given a TF, application of the DPI, will generate predictions about which other genes might be its direct transcriptional targets or its upstream transcriptional regulators[51].

Although, it's worth noticing that despite the application of the DPI filter, the predicted TF-target relationships are mainly correlative but likely to give hints about the underlying biological mechanisms. In fact, a study comparing statistical methods to construct a GRN showed that GeneNet, WGCNA and ARACNE, perform well in constructing the global network structure[52].

For this analysis, data obtained from an RNAseq experiment, a next generation sequencing technique used to unbiasedly measure the mRNA expression[53] of individual genes or transcripts, was adopted to infer a continuous GRN.

1.3.2 Protein co-expression network

In parallel with the explosion of the number of large-scale transcriptomics analysis, the latest improvements in liquid chromatography and mass spectrometry enhanced the generation of high-throughput proteomics datasets as well. Furthermore, strengthened by extensive studies on transcriptomics datasets, some of the validated methods to reconstruct GRNs are being repurposed for the identification of protein co-expression modules from the more recently generated proteomics data. In fact, protein co-expression networks can be built via the same strategy as weighted gene co-expression network analysis[54].

It was shown that the analysis of both the topology and the modules of these graphs provide high sensitivity in the recognition of disease phenotypes[55]. In particular, their topology was found to be in partial agreement with the "scale-free" configuration and to contain non-random modules[56].

Noteworthy, protein co-expression networks, when compared to GRN, were shown to be even more predictive of the functional similarities among features since mRNA coexpression pattern was driven not only by cofunction but also by the colocalization of the genes, while protein coexpression was mainly driven by functional similarity, to the extent they can be used to predict protein complexes membership[32]. These latter are groups of functionally related and physically bound proteins that carry out major processes in a cell, such as gene transcription and splicing or protein synthesis and degradation.

The investigation of protein complexes can hence reveal particular characteristics of the tumour biology[57]. For example, in a breast cancer study, protein complexes reconstruction from proteomics data identified complexes that are consistently under- or over-expressed in specific tumour subtypes[36]. Similarly, in glioblastoma multiforme, using the weighted gene co-expression network analysis method, three main modules associated with three different membrane associated groups (mitochondrial, endoplasmic reticulum and vesicle fraction) were identified[58].

Inspired by these studies, I decided to exploit the complementary information provided by protein co-expression modules to improve the RNAseq derived GRN.

1.3.3 Integrated analysis of transcriptomics and proteomics data

To overcome the limitations inherent to each -omics layer, novel bioinformatics approaches have been developed to obtain a comprehensive perspective of biological systems.

In greater detail, single-level omics approaches lack the resolving power to establish causal relationships between molecular alterations and phenotypic manifestations. The joint analysis of different biological layers provides instead a clearer representation of the mechanisms underlying complex diseases as cancer[32].

As an example, mRNA profiling cannot capture post-transcriptional modifications influencing the amount of active protein, while proteomics profiling cannot detect low abundant proteins or novel proteoforms generated via alternative splicing[59]. Given the notable difficulties in directly correlating transcriptomics and proteomics profiles, several types of approaches have been proposed to jointly analyse the two sources of data, for example by: creating a reference set from the union of the two layers into a single framework; extracting common functional contexts, typically in terms of enriched signaling pathways; through topological and network analysis to find, for example, common regulators driving both expression profiles; by applying clustering approaches, aiming to find similarities between the groups identified in each individual dataset; dynamic modeling to describe temporal regulation of gene expression by leveraging the information provided by the proteomic profile[60].

Given the proven efficacy of the joint methods in cancer research, in Chapter 4 I propose a novel network-based joint enrichment analysis approach. The model is based on the construction of integrative modules generated by merging GRN regulons with associated protein complexes.

1.4 Summary of research hypothesis and aims of the PhD project

This PhD project aimed to evaluate and develop network-based methods for the analysis of PCa omics data. The final goal of the *in silico* approach was the identification of potential prognostic and diagnostic biomarkers that can then be mechanistically investigated through *in vitro* or *in vivo* experiments.

In Chapter 3, by exploiting the richness of information stored in gene-gene interactions, I tested the hypothesis that a gene regulatory network derived from transcriptomics profiles of PCa orthografts can reveal candidate transcriptional regulators with potential prognostic value.

In Chapter 4, I explored the possibility of integrating the regulon structure inferred from transcriptomics profiles in Chapter 3 with proteomic data generated from the same preclinical orthografts. The objective of the study was to assess the complementarity of the two omics layer and provide more robust insights into the regulatory mechanisms underlying CRPC.

In Chapter 5, I assessed the utility of the network-based investigation as a tool to identify changes associated with tumoral treatment responses. To this end, I leveraged transcriptomics data from *in vivo* subcutaneous xenografts, treated with either single or combination of drugs.

Chapter 2 – Materials and Methods

2.1 Preclinical models

2.1.1 Nude mice orthografts – experiment from Dr. Mark Salji

Experiments involving prostate orthografts derived from human prostate cancer cell lines were carried by Dr Mark Salji as part of his PhD project[12,61], including the subsequent transcriptomic and proteomic analyses. Hormone naïve human prostate cancer cell lines (CWR, LNCaP and VCaP) were implanted in androgen proficient nude mice to generate androgen dependent prostate orthografts (three biological triplicates for each cell line). Castration resistant (or androgen independent) prostate orthografts were generated from the respective isogenic derivatives (22Rv1, LNCaPAI and VCaP cells (with three biological triplicates for each cell line) by orthotopic implantation in castrated nude mice. All available samples (n=18) were used for the analyses.

2.2 Transcriptomics analysis

2.2.1 Next generation sequencing based RNA analysis – experiment from Dr. Mark Salji

RNA was extracted from samples using RNeasy Mini Kit (Qiagen, 74104) after homogenisation with QIAshredder homogeniser columns (Qiagen, 79654). DNA was degraded with RNase-Free DNase Set (Qiagen, 79254). Quality of the purified RNA was then measured on a 2200 Tapestation (Agilent) using RNA screentape. Preparation of libraries for cluster generation and cDNA sequencing was done using Illumina TruSeq Stranded mRNA LT Kit (Illumina, 20020594). Quality and quantity of the DNA libraries was assessed using a 2200 Tapestation (D1000 screentape) and Qubit (Thermo Fisher Scientific, Q32851) respectively. The libraries were run on the Illumina Next Seq 500 using the High Output 75

cycles kit (2 x 36 cycles, 150bp paired end reads, single index). Fastq files were generated from the sequencer output using Illumina's bcl2fastq.

The data is published by Dr. Mark Salji[62].

2.2.2 Pre-processing

Raw fastq files underwent the following steps: adapter trimming using 'ILLUMINACLIP' step from 'Trimmomatic'[63] version 0.36; alignment to GRCh37.p13 human reference (and GRCm38.p4 mouse reference for orthografts) with TopHat[64] v2.0.14; ; human/mouse reads disambiguation using Disambiguate[65] v1.0.

2.2.3 Quality control

Both unaligned and mapped underwent quality control. FastQC[66] v0.11.9 was used to check sequence quality, base content, GC content, N content and duplication levels. Rseqc[67] v2.6.4 has been adopted to check mapped gene body coverage, inner distance, and read duplications.

2.2.4 Gene counts profiles

Gene level raw counts were calculated using 'HTSeq'[68] version 0.9.1 and GRCh37.p13 (and GRCm38.p4 for orthografts) annotation files (.gtf) for human and mouse reads respectively. Gene level FPKM quantification was performed using 'cufflinks'[69] v2.2.1 and GRCh37.p13 (and GRCm38.p4 for orthografts) gtfs for human and mouse reads respectively.

2.2.5 Group and single sample differential gene expression

The R package 'Deseq2'[70] v1.26 was used to calculate differentially expressed genes (DEG) among the treatments/phenotypical groups, using default parameters and gene-level raw counts as input.

For single sample analysis, differentially expressed genes were calculated by subtracting the Deseq2 normalised counts of each tumoral samples with the average of the panel made of all normal samples within each dataset under investigation. The significance of the difference was calculated through the interpolation on the standardized gaussian distribution, after dividing each difference for the standard deviation of the gene expression in the panel of normals.

2.2.6 Regulons enrichment

The full DEG lists for groups of samples were mapped onto the gene regulatory network identified from the preclinical orthografts. The inferred set of positive and negative gene-gene interactions, as well as each list of DEG, was given as input to the function 'nbea' from the package 'EnrichmentBrowser'[71] version v2.12.1, applying the 'GGEA'[49] (gene graph enrichment analysis) method with default parameters. A threshold of FDR ≤0.05 has been adopted to identify the enriched regulons (Appendix 1-3).

2.2.7 Pathway and statistical analysis

The R package 'clusterProfiler'[72] v2.1.0 was used to perform overrepresentation analysis of the lists of prioritized genes. In particular, the .gmt files obtained from the 'Molecular Signature Database (MSigDB)[73] v6.2 of C3 collection of transcription factors targets, H collection of experimentally validated Hallmarks of cancer and C5 collections of biological processes from gene ontologies, were

given as input to the 'enricher' function of 'clusterProfiler'. All statistical analysis of the preprocessed data were performed using R v4.03 (Appendix 3).

2.3 Gene regulatory network analysis

2.3.1 Datasets – obtained from Transpot consortium

Hormone naïve human prostate cancer cell lines (CWR, LNCaP and VCaP) were implanted into the prostates of androgen proficient (6 weeks old) nude male mice to generate androgen dependent prostate orthografts. Castration resistant (or androgen independent) prostate orthografts were generated from the 22Rv1, LNCaPAI and VCaP human PCa cell lines by orthotopic implantation into the prostates of castrated nude (6 weeks old) male mice. RNA-seq data were obtained from 18 orthografts derived from the six human PCa cell lines studied (n = 3 mice per cell line)[12], referred to as the UGLA dataset. All data were included for the inference of the gene regulatory network.

RNA-seq data from three clinical PCa cohorts were included in this study (Appendix 1): The University of Tampere (UTA-EGAD00001000609), the Erasmus Medical Center in Rotterdam (EMC-EGAD00001004215), and the International Cancer Genome Consortium (ICGC-EGAD00001004791). A summary of the clinicopathological characteristics of the cohorts is provided in Table 2-1.

The cohorts were chosen, and preferred to bigger dataset such as TCGA, for the standardized definition of biochemical recurrence.

Clinical cohorts	UTA		EMC		ICGC	
Number (n)	*n* = 27	%	*n* = 37	%	*n* = 85	%
age at diagnosis						
range	47–71		NA		32–52	
mean	60		NA		47	
median	61		NA		48	
na	0					
psa at diagnosis (ng/ml)						
range	3.5–48.1		0.3–36.2		3.1–743	
mean	10.4		11.8		30.48	
median	8.3		9.4		8.21	
na	0		0		0	
tumour stage						
t1	10	37.0	1	2.7	0	0.0
t2	16	59.3	15	40.5	61	71.8
t3	1	3.7	13	35.1	23	27.1
t4	0	0.0	8	21.6	1	1.2
na	0	0.0	0	0.0	0	0.0
gleason score						
<7	7	25.9	6	16.2	12	14.1
7	13	48.2	19	51.4	65	76.5
>7	7	25.9	0	0.0	8	9.4
na	0	0.0	12	32.4	0	0.0
therapy						
Radical prostatectomy	27	100	37	100	85	100

Table 2-2. Clinicopathological characteristics of patient cohorts (NA, data not available).

2.3.2 Regulons Identification and Filtering

The PCa gene-regulatory network was generated using the R package 'RTN'[74] version v2.4.6 (which reimplements the ARACNe/MRA pipelines), based on FPKM values (Fragments Per Kilobase of transcript per Million mapped reads) of the UGLA orthograft dataset and a list of 2065 transcription factors that were given as input (manually curated from MsigDb[73]).

Transcription factors significantly associated with one or more target genes were identified as regulators. The normalised counts matrix was then filtered by genes with FPKM equal or higher than one in at least one sample and standardised within the zero-to-one range. The function 'tna.shadow' from the R package 'Viper'[75] version 1.14.0 has been used to account for the 'shadow'

effect (the chance of obtaining false positive result) during the enrichment of a GRN, if a non-active regulator (in other words not causing the alterations in the expression levels) shares a significant proportion of its targets with a *bona fide* active transcription factor (Appendix 1). In greater details, the correlation data is permutated to generate a null model. The strength of the predicted intra regulons' relationships is then compared to the null model by taking into account regulons' sizes. Finally, a pvalue cutoff is used to discern more or less likely TF-target associations.

2.3.3 Gene Regulatory Network Metrics

The graph structure was analysed using the R package 'igraph' v1.2.5, exploiting the functions 'degree', 'betweenness', 'constraint' and 'closeness' to retrieve metrics at the 'nodes' level, providing complementary information about the importance of individual nodes within the network: (1) The 'degree' (or 'in-degree') of a node in a GRN is the number of transcriptional regulators involved in the control of the expression of a specific target gene. For different GRNs, the number of regulatory genes implicated for individual target genes varies, depending on complexity of the network; (2) 'Betweenness' is defined as the number of shortest paths passing through the node and can be interpreted as a measure of the influence of the node of interest over the global flow of information; (3) Burt's 'constraint' is a measure of the redundancy of the information received by the node and can be interpreted as its ability to converge different signals; (4) 'Closeness' quantifies the node's participation within a network. Finally, the Jaccard Index, a statistical measure defined as the ratio of the intersection and the union of two sets, was applied to highlight network nodes sharing a meaningful proportion of targets. The threshold of 0.1 was chosen to prioritise the nodes in this study. A threshold of Jaccard Index/Co-efficient set at 0.1 highlights pairs of regulons with intersection (sharing) of ≥10% of the target genes when considered across the full set of target genes for the respective regulons (Appendix 1).

2.3.4 Statistical analyses related to gene regulatory network analysis in clinical cohorts

Biochemical recurrence was defined as serum prostate-specific antigen (PSA) levels ≥ 2 ng/mL above nadir PSA (the lowest PSA level after treatment) and signifies clinical evidence of relapsed cancer. Relapse free survival (defined by absence of biochemical recurrence) was used to evaluate the prognostic utility of regulon signatures of interest in the UTA, EMC and ICGC clinical cohorts. The performance of our candidate *JMJD6* regulon signature as a prognostic marker was compared to three published signatures (using the formulas described in the original publications[41–43]: (1) For the TMCC11 signature, the per-sample average of the normalised counts of the genes belonging to the signature was used to stratify the patient cohort into two groups according to values above or below the 67th percentile. (2) For the HYPOXIA-28 signature, the normalised counts were multiplied by the coefficient associated to each gene of the signature and all the products were added together to generate a sample-specific overall score, and the patient cohort was stratified into two groups according to the median of its distribution. (3) For the BROMO-10 signature, the function 'gsva' from GSVA v1.38.2 was used to analyse data from the normalised counts to calculate a signature enrichment score per sample.

Patients were labelled according to the enrichment status of *JMJD6,* as predicted by GGEA, into active or inactive status groups. Hazard ratios (HR) for all the analysis were obtained by Cox proportional-hazard model regressions, using the 'coxph' function from the R package 'survival' version 3.1-8. Noteworthy, the small number of samples in the cohort was not enough to assess the assumptions for the CoxPH model.

Moreover, for multivariate analysis, Gleason score and the TNM (Tumour/Node/Metastasis) classification were added to the model formula in the form: 'Endpoint ~ JMJD6regulon_activity + second_variable'. Kaplan–Meier curves were obtained using the 'ggsurvplot' function from the R package 'survminer' v0.4.8 (Appendix 1).

2.3.5 Human prostate cell lines used in *in vitro* studies – experiment from Drs. Linda Rushworth

CWR22 (hormone naïve) cell lines were obtained from Case Western Reserve University, Cleveland, Ohio, and cultured in either RPMI medium (Gibco, Thermo Fisher Scientific, Waltham, MA, USA), supplemented with 10% foetal bovine serum (FBS, Gibco, Thermo Fisher Scientific, Waltham, MA, USA) and 2 mM glutamine ('CWR_FBS' label in Figures), or androgen-deprived medium consisting of phenol-free RPMI (Gibco, Thermo Fisher Scientific, Waltham, MA, USA) supplemented with 10% charcoal stripped serum (CSS, Gibco, Thermo Fisher Scientific, Waltham, MA, USA) and 2 mM glutamine ('CWR_CSS' label in Figures).

Their matching castration-resistant cell lines 22Rv1 were obtained from ATCC and cultured in androgen-deprived medium only, as described before ('22Rv1_CSS' label in Figures). All cells were kept in incubators set at 37°C and 5% CO_2.

2.3.6 Western blot analysis

Protein extraction were performed by Dr Linda Rushworth. In brief, media was removed and cells were washed twice with PBS. Llysis buffer (1% SDS with protease and phosphatase inhibitors) was added and the cells were scraped into tubes. The samples were sonicated 3 times x 10 seconds, spinned for 10 minutes at maximum speed to pellet cell debris. The supernatant was then moved into fresh tube. LDS (Lithium dodecyl sulfate) sample buffer was added 1:3 to the volume of lysate and boiled for ~5 minutes before loading the gel.

20 micrograms of proteins were loaded on an SDS-PAGE gel (Invitrogen, Thermo Fisher Scientific, Waltham, MA, USA) and transferred to a PVDF membrane (GE Healthcare, Chicago, IL, USA). Membranes were then blocked in 5% milk in TBS-Tween (TBST) and subsequently probed overnight, on a roller at 4 °C, with

primary JMJD6 antibodies (#60602, Cell Signaling Technology, Danvers, MA, USA) diluted 1:1000 in 5% milk TBST.

The following day the membranes were washed with TBST three times, 10 minutes each, and incubated with secondary antibody, diluted 1:10000 in 5% milk in TBST. Revelation was obtained by scanning the membrane with the LI-COR Odyssey CLx Imaging system (LI-COR Biosciences, Lincoln, NE, USA).

2.3.7 si-RNA mediated knock-down

For the knock-down experiment, CWR and 22Rv1 cell lines were transfected using the Lipofectamine RNAiMAX[76] Reagent (Invitrogen, Thermo Fisher Scientific, Waltham, MA, USA) and JMJD6 specific siRNA (ON-TARGETplus siRNA Reagents, Catalog ID:J-010363, Dharmacon, Horizon inspired cell solutions, Cambridge, UK).

Baseline cellular response to siRNAs was assessed using a control pool of four non-targeting siRNA (ON-TARGETplus Non-targeting Control Pool, Catalog ID:D-001810-10-05, Dharmacon, Horizon inspired cell solutions, Cambridge, UK). RNA or protein extraction was performed 72 hours after transfection.

2.3.8 *In vitro* growth assay

Evaluation of cell growth was performed using the Incucyte® Live-Cell Analysis system, following the 'Adherent cell line' protocol.

20000 CWR and 25000 22Rv1 cells were plated in a 96-well plate to evaluate the six conditions ('CWR_FBS_ControlsiRNAs', 'CWR_FBS_JMJD6siRNAs', 'CWR_CSS_ControlsiRNAs', 'CWR_CSS_JMJD6siRNAs', '22Rv1_CSS_ControlsiRNAs', '22Rv1_CSS_JMJD6siRNAs') with 10 technical replicates; phase-contrast images were captured every two hours and analysed using the integrated confluence algorithm, part of the Incucyte® Live-Cell Analysis suite.

2.4 Integrating transcriptomic and proteomic analysis

Transcriptomic analysis is based on methodology described in Section 2.2. Samples for proteomic analysis were carried out by Dr Mark Salji previously in the host laboratory.

2.4.1 SILAC based control using *in vitro* cultures of selected prostate cancer cells – experiment from Dr. Mark Salji

CWR, LNCAP, LNCAPAI and VCAP cell lines were grown and labelled with Arg-10 and Lys-8 using 100% dialysed FBS (dFBS) conditions to perform the SILAC experiment. For the heavy labelling, the RPMI SILAC media was used: Arg-10 and Lys-8 were introduced at the same concentration of the Standard RPMI 1640 conditions, namely 200mg/L and 40 mg/L, respectively. The SILAC standard was generated by seeding $1x10^6$ cells of each cell line in 6 cm dishes and by transferring them to their respective full or Charcoal Stripped serum Media to maintain approximately 50% confluence. Subsequently, the samples underwent the following steps: trypsinisation of the cells and centrifugation to remove excess trypsin; Centrifugation at 14,000 RCF for 15 minutes to remove DNA contamination; protein quantification via Bradford's assay with BSA standard curve; expansion of labeled cell lines in SILAC media to maintain incorporation; mixing at a 1:1:1:1 ratio to generate the super SILAC standard.

2.4.2 Processing of prostate orthografts for quantitative proteomic analysis – experiment from Dr. Mark Salji

Frozen tumours from the orthografts models reflecting hormone naïve (CWR, LNCAP, VCAP grown in intact mice) and correspondent CRPC tumours (22Rv1, LNCAPAI and VCAP grown in castrated mice) were split in four pieces and a quarter was processed for the proteomics experiment through grinding. Precellys homogenization at room temperature, involved ~20 mg of grinded tumour

together with 200 µl of 4% SDS lysis buffer. Subsequently, the lysate was boiled at 95 ^{o}C for 5 minutes and sonicated to fragments, to remove DNA by centrifugation. 20 mg of tumour lysates in 200 µl of 4% SDS buffer were quantified in triplicates using Bradford's assay. To reduce DDT concentration, samples were first diluted 1:5 in HPLC water. The filters applied for the selection of the samples were the following: standard deviation (SD) of <0.001 between triplicate values of 595 nm absorbance by spectrophotometry Bradford's assay and R^2 value with standard curve >0.99. Each of the 18-tumour sample (three biological replicates for each of the six models) was mixed at 1:1 ratio with the super SILAC standard for normalization purposes.

2.4.3 Protein quantification to determine differentially expressed proteins – experiment from Dr. Mark Salji

The raw data obtained from the mass spectrometer was processed with MaxQuant version 1.5.2.8 and searched with Andromeda search engine querying two different SwissProt databases: Homo sapiens (09/07/2016 92939 entries) and Mus musculus (20/06/2016; 57,258 entries). Protein hits coming from individual database were separated using "Split protein groups by taxonomy ID" option in MaxQuant. The "Re-quantify" and "Match Between Runs" options were also used. For quantification, multiplicity was set to 2 (doublets) and Arg0/Arg10, Lys0/Lys8 were used for ratio calculation of SILAC labelled peptides. Only unique peptides were used for protein group quantification. Digestion mode was set to "Specific" using the digestion enzyme trypsin and allowing for two miscleavages. Iodoacetamide derivative of cysteine was specified as a fixed modification, whereas: oxidation of methionine and acetylation of proteins N-terminus were specified as variable modifications. Peptides with less than seven amino acid residues were excluded from processing. Only protein groups identified with at least one unique peptide were used for quantification. The protein groups output file was then loaded into the Perseus platform version 1.5.2.4. Perseus was used to filter the data for confident identifications based on at least 1 unique peptide match and identified in at least 2 of 3 biological replicates in at least one group. A further median normalisation was performed on all samples

prior to Welch's t-test with permutation-based FDR set at 0.01 used to identify significantly changing proteins. Median normalisation of all samples was used to correct the data prior to applying FDR adjusted statistical testing. The pipeline generated four matrices for downstream analysis: LFQ of samples harbouring light isotopes (orthograft tumours), LFQ of heavy labelled samples (CWR, 22Rv1, LNCAP, LNCAPAI cell lines super SILAC standard), SILAC ratio of light versus heavy isotopes intensities, ratios of LFQ light and heavy intensities.

2.4.5 Protein network generation and modules splitting

The R package 'ProCoNa' v1.0.2 was used to build the protein co-expression network and the associated modules of co-expression from the light-labelled intensities matrix (orthograft data) and default parameters, except for the number of permutations increased from 100 to 1000 to achieve higher robustness.

Original modules obtained from the protein co-expression network were split using a two-steps procedure: the R package 'biclust' v2.0.2 was used to reveal influences from the orthografts' cell type of origin or CRPC status; the R package 'conclust' v1.1 was, subsequently, used to rearrange the predicted interactions by taking into account experimentally validated protein-protein interactions (Hippie and CORUM repositories).

The degree of overlap within the inferred protein modules and within the CORUM sets was calculated through the pairwise Jaccard index, namely the ratio of the intersection and the union set of the content of two lists. Only non-zero Jaccard indexes were retained for the generation of Figure 4-4 (Appendix 2).

2.4.6 Pathway analysis and integrative modules analysis (generation and enrichment)

The R package ‘ROTS’ v1.18 was used, with default parameters, to calculate differentially expressed proteins between PC and CRPC samples after log10 transforming the intensity values from the four normalized matrices generated by the SILAC quantification.

The R package ‘clusterProfiler’ v2.1.0 was used to perform overrepresentation analysis upon the lists of differentially expressed genes and proteins. In particular, the H collection (‘.gmt’ file) of experimentally validated Hallmarks of cancer was obtained from the MSigDB v6.2 and given as input to the ‘enricher’ function, applying default parameters.

RNA-seq derived regulons and CORUM protein complexes were integrated by means of at least one shared feature. The CORUM protein-protein interactions were weighted according to the adjacency values resulted from the protein co-expression analysis. To each regulon, one or more complexes were linked to and labelled either ‘posComplex’, when expected to be coherently expressed with the transcription factor positive target, and ‘negComplex’ when expected to be coherently repressed together with the transcription factor negative target. See Figure 4.6 for an example of integrative module.

The enrichment of the integrative modules consisted of the calculation of separate scores for regulons (TF score) and protein complexes (PC score) respectively. The R package ‘pracma’ v2.2.9 was used to calculate the cubic root among three weights for both TF and PC scores: edge weight obtained from the gene regulatory network and protein network; p-value associated to the CRPCvsPC comparison; log2fold change of the CRPCvsPC comparison. The scoring was repeated by reshuffling the differentially expressed genes and proteins tables using the R package ‘boot’ v1.3-24. 10,000 permutations allowed the calculation of a p-value for both the TF and PC scores for each integrative

modules. The threshold for significant results was defined at FDR= 0.05 (Appendix 2).

2.5 Transcriptomics analysis from *in-vivo* treated samples

2.5.1 In-vivo mouse models – Experiment from Prof. Ian Mills

In total, 87 mice were sacrificed for this experiment, split in the following groups: Treatment arm 1 (15 mice) – Mycophenolate mofetil (100 mg/kg) – delivered daily by oral gavage in a 0.9% saline solution; Treatment arm 2 (15 mice) – Abiraterone at 0.5 mmol/kg/d in vehicle delivered by intraperitoneal injection in 5% benzyl alcohol, 95% safflower oil together with prednisolone phosphate (water soluble) at 20 mg/kg delivered by intravenous injection – both with daily dosing; Treatment arm 3 (15 mice) – ARN-509 at 30 mg/kg administered daily by oral gavage in a formulation solution consisting of 15% alpha-tocopherol (Vitamin E)-TPGS and 65% of a 0.5% w/v Carboxymethylcellulose solution in 20 mM citrate buffer (pH 4.0); Treatment arm 4 (15 mice) – Mycophenolate mofetil (100 mg/kg) and ARN-509 (30mg/kg); Treatment arm 5 (15 mice) – Mycophenolate mofetil (100 mg/kg), Abiraterone (0.5 mmol/kg/d in vehicle) and prednisolone phosphate (20 mg/kg); Untreated (2 mice); Vehicle (10 mice) – 0.1 ml 5% benzyl alcohol and 95% safflower oil solution via intraperitoneal injection every day.

2.5.2 Growth assays – Experiment from Prof. Ian Mills

The LNCaP C4-2b derivative stably transduced with luciferase and expressing exogenous androgen receptor was selected and used as previously described [77,78]. This line was implanted subcutaneously at single flank sites at 106 cells (100 µL in 50% Matrigel (BD Biosciences) and 50% growth media) into the flanks of male SCID mice. Drug dosing, administration and formulations are outlined above for each treatment arm. Tumour size was measured twice weekly in

three dimensions (l x w x d) using calipers. In addition, in vivo luciferase imaging was used through the intraperitoneal injection of d-luciferin substrate (100 µl at a concentration of 15mg/ml, Xenogen). For this procedure the mice were anaesthetised 5 minutes post-injection using isfluorane (Baxter) and then imaged using a cooled charged-couple device IVIS camera. Downstream analysis of the imaging data was undertaken using Living Image 2.30 software. In all cases treatment commenced once tumours were established and had reached a volume of 300 mm3. Animals were euthanised once tumours reached a terminal volume of 1000 mm3 or at three weeks post-treatment, whichever occurred earlier.

2.5.3 RNA-seq – Experiment from Prof. Ian Mills

RNA was extracted from the cells and collected utilizing the column-based method through the RNeasy MinElute Clean-up Kit (QIAGEN, Hilden, Germany), at a concentration of 25 ng/µL in 20 µL, according to the manufacturer's instructions. The RNA library preparations were performed with the KAPA RNA HyperPrep Kit (KAPA Biosystems, Roche Holding AG) according to the manufacturer's instructions. Sequencing was performed on Illumina Next Seq 500, obtaining 25M 75 base pairs paired-end reads per sample. This was performed by the Genomic Core Technology Unit, CCRCB, Queen's University Belfast.

2.5.4 Fastq pre-processing

RNA-seq fastq files received from Queen's University Belfast were processed through the following pipeline to obtain gene-level normalized counts for each sample: adapter trimming using 'ILLUMINACLIP' step from 'Trimmomatic' version 0.36; alignment to GRCh37.p13 human reference and GRCm38.p4 mouse reference with TopHat v2.0.14; human/mouse reads disambiguation using Disambiguate 1.0; gene level raw counts calculation using 'HTSeq' version 0.9.1 and GRCh37.p13 and GRCm38.p4 annotation files (.gtf) for

human and mouse reads respectively; gene and transcripts level FPKM quantification using cufflinks-2.2.1 and GRCh37.p13 and GRCm38.p4 gtfs for human and mouse reads respectively.

2.5.5 Principal component analysis and heatmaps

Principal component analysis was performed using the R package ‘stats’ v4.0.3 and Deseq2 normalised counts, to visualise the overall effect of experimental covariates and batch effects within the data (Appendix 3).

The heatmpas were generated using the ‘heatmap3’ R package, with the following parameters: distance metric = as.dist(1 - cor(t(x), use = "pa")); hclustfun = hclust, method = "complete", scale = ‘none’.

2.5.6 Differentially expressed genes

The R package ‘Deseq2’ v1.26 was used to calculate DEG among the treatment's groups, using default parameters and gene-level raw counts as input.

Given the little number of untreated samples, both the no-treatment and the vehichles have been used as controls. The rationale was to have the most robust results possible with this experiment design.

2.5.7 Pathway analysis

The R package ‘clusterProfiler’ v2.1.0 was used to perform overrepresentation analysis of the lists of prioritized genes. In particular, the gmt files obtained from the MSigDB v6.2 of C3 collection of transcription factors targets, H collection of experimentally validated Hallmarks of cancer and C5 collections of biological processes from gene ontologies, were given as input to the

'enricher' function, together with the DEG lists. Only the over-representation analysis of the c5 collection provided statistically significant results at the q-value level (Appendix 3).

Chapter 3 – Gene Regulation Network Analysis

3.1 Introduction

Prostate cancer (PCa) is the second most common cancer among men and the fifth leading cause of death worldwide[79]. Tumour heterogeneity in PCa (between patients and among different tumour foci within individual patients) creates a major obstacle to the identification of clinically relevant molecular subtypes[80]. As a result, PCa treatment decisions are not based on tumour biology. Disease recurrence following treatment remains a significant problem, even following radical treatment such as radical prostatectomy or radical radiotherapy[9]. Despite the use of docetaxel chemotherapy or second generation androgen receptor pathway inhibitors along with androgen deprivation therapy (ADT), patients presenting with advanced and/or metastatic disease are at high risk of recurrent disease, which tend to be aggressive and incurable as either castration resistant (CRPC) or neuroendocrine PCa variants[4,5]. Therefore, there is an unmet need to improve our understanding of progressive PCa in order to identify new targets for therapy as well as prognostic biomarkers.

Inter-patient tumoral heterogeneity and intra-tumour heterogeneity among different tumour foci are well reported, making it unlikely that a single gene will be a representative biomarker of PCa progression[83]. Investigating a gene set-related network may leverage the correlations of the expression of multiple interacting genes[84]. Moreover, the study of gene-gene interactions can reveal commonalities that can be observed only at the functional level, when the alterations in different genes are associated with the same biological mechanism.

Several gene set-based panels are offered as prognostic tests for PCa patients. Commercial assays[85–87] including Decipher™, Oncotype DX® and Prolaris, together with scoring methods published in the literature, have been developed using microarray, Illumina or Nanostring transcriptome profiles[9–11] to apply mRNA expression data to predict the risk of cancer recurrence and/or progression.

While gene expression-based models have resulted in promising data for predicting cancer behaviour in vitro[88], significant improvements are required before a stratification/prognostic tool in PCa patients can be considered for routine clinical practice, including the prediction of the risk of cancer recurrence following treatment[4]. The limitations of existing commercial molecular PCa diagnostic tests may stem from potential biases introduced during the signature identification step (including factors related to patient ethnicity[89], immune[90] and stromal[91] components of the tumours) that may influence the gene expression profiles. Moreover, gene set-based methods typically focus on the expression of individual genes or gene sets, without the ability to incorporate biologically important information associated with gene-gene interactions[84].

Alterations in transcriptional programmes are frequently implicated in PCa progression[92]. Genes that co-operate within the same biological pathways are often under the regulatory control of shared (one or more) transcription factors. Conveniently, interacting genes tend to be associated at the expression levels[93], providing the chance to infer their relationships from transcriptomics data. Gene regulatory networks (GRNs) are graphs describing transcriptional regulators and their target genes as nodes, while the relationships (level of correlation) among the regulators and target genes are presented as the edges. Statistical and/or machine learning approaches have been applied to gene expression data[94] to predict the topology of GRNs, namely the arrangement of transcriptional regulators and their target genes as well as the direction of each transcription factor-target interaction (i.e., positive or negative regulation). Within GRNs, data on the agreement between the predicted regulations and differential gene expression analysis can be applied to explore the underlying biological mechanisms to explain specific phenotypes (such as cancers with lower or higher chances of recurrence/progression).

Preclinical models of human PCa cells grown as orthotopic xenografts in mice (orthografts) represent a useful tool to mimic progressive clinical disease. However, the use of preclinical PCa as a tool to identify potential GRNs involved

in progressive disease has not been tested.Despite the presence of intrinsic biases, our hypothesis is that a better overview of the underlying biology is provided by the regulons.

Here, to generate a robust scoring method, I derived GRNs from a collection of preclinical hormone naïve (dependent on androgens for growth) and castration resistant (growth despite androgen deprivation therapy) human PCa orthografts to capture the heterogeneous nature of clinical disease, leveraging the strength of correlations in the expression patterns of genes transcribed by tumour cells only. Filtering the GRNs for statistically significant associations led to the identification of putative regulons, signifying the network of target genes and shared transcription factor (or transcriptional regulator) involved. Integrating data from preclinical orthografts and clinical PCa cohorts, I modelled regulon signatures to identify patients at risk of cancer recurrence, and identified the *JMJD6* (Jumonji Domain Containing 6, arginine demethylase and lysine hydroxylase, a protein hydroxylase or histone demethylase) regulon as a prognostic marker in PCa (Figure 3-1). Lastly, preliminary knock-down experiments were performed on hormone naïve and hormone resistant prostate cancer cell lines to further test the involvement of JMJD6 in prostate carcinogenesis.

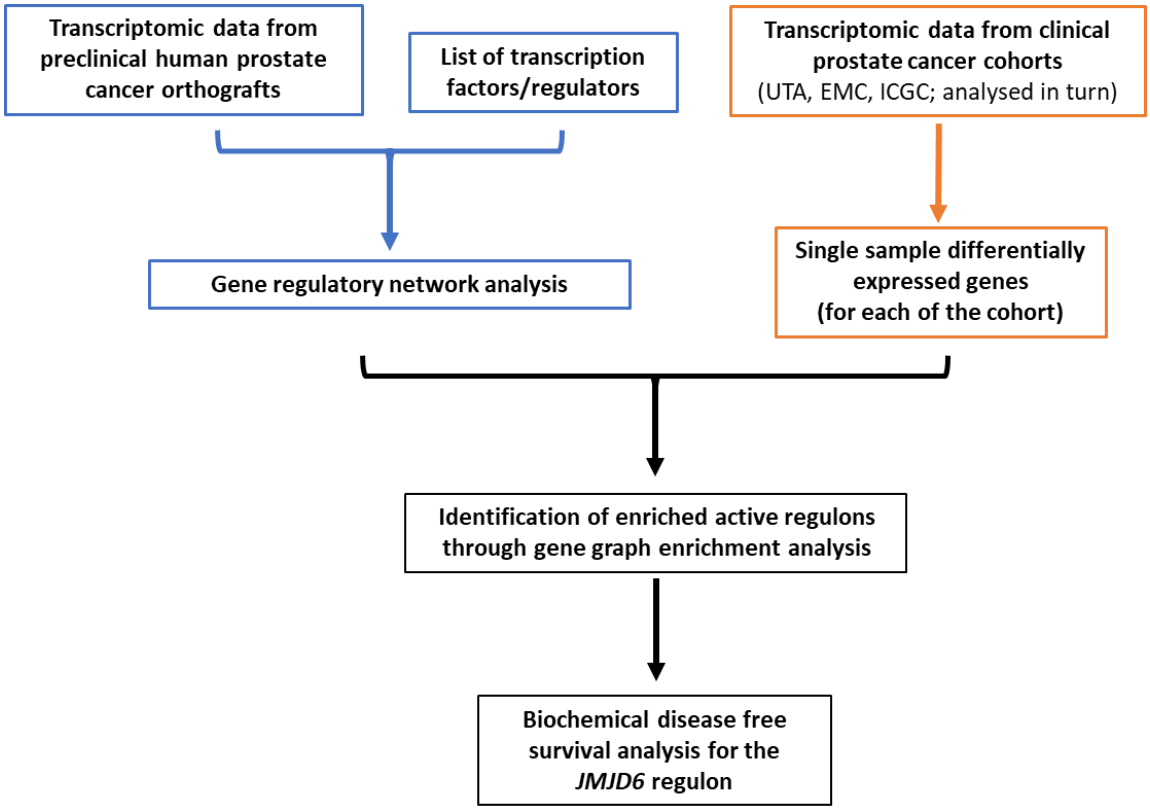


Figure 3-2. Workflow of the gene regulatory network analysis.

3.2 Results

3.2.1 Clinical cohorts used to integrate with transcriptomic data from preclinical orthografts

Summary of the three clinical PCa cohorts (namely UTA, EMC and ICGC) included in this study can be found in Section 2.3. The focus of the study was the identification of biomarkers able to predict the progression of PCa into CRPC. The CRPC orthograft data (9 samples) was used, together with the HN orthograts data (9 samples), only to infer the gene regulatory network.

Dataset from the UTA cohort[95] were obtained from 46 prostate tumour samples, including 28 untreated PCa samples from radical prostatectomy and 12 benign prostate hyperplasia control samples (obtained by radical prostatectomy, cystoprostatectomy or transurethral resection). RNA-seq data from treatment

naive PCa samples that passed mapping quality control, provided with information on progression free time (*n* = 27), were used in this study, along with the 12 benign samples.

The EMC dataset was obtained from 92 radical prostatectomy specimens (51 PCa with 41 adjacent benign prostate tissue)[21,22]. The tumour content was confirmed histologically. Only prostate tumour samples with the information on progression free time (*n* = 37) and all the benign control samples were used in the present study.

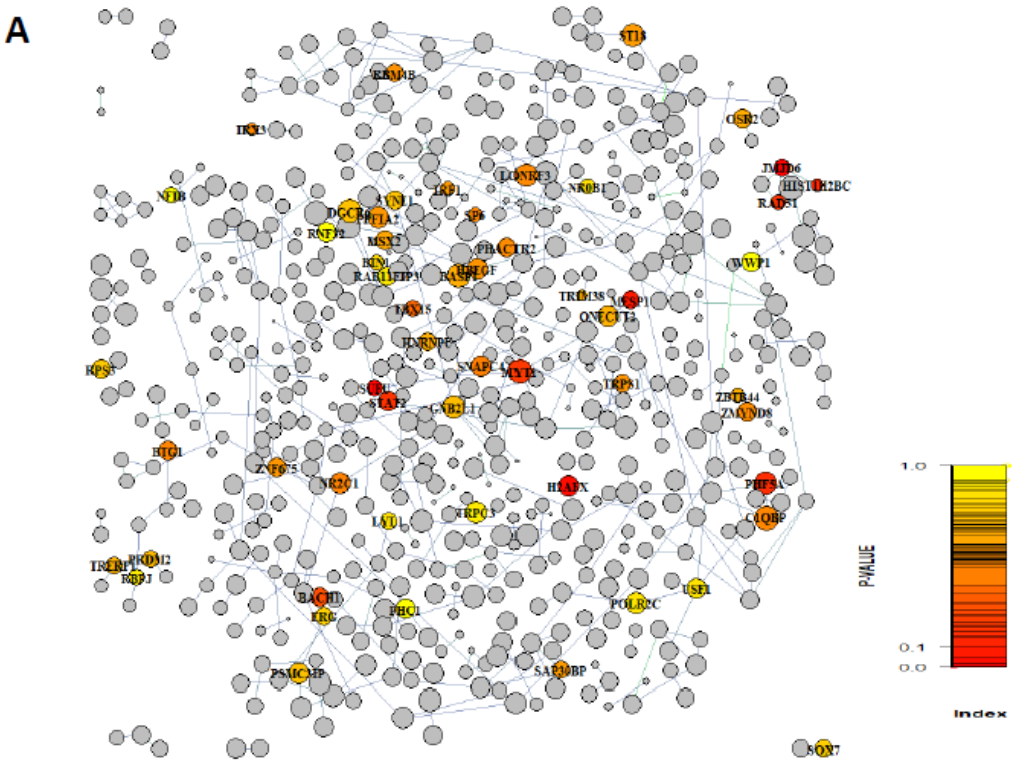
The ICGC dataset consists of 125 PCa specimens (and 8 matched benign control tissue) from 100 radical prostatectomy specimens[98]. Six tumour samples from the same prostatectomy specimens were sampled multiple times (from 3 to 6 biological replicates per patient) and were averaged at gene count level per patient, given the similarity in expression profiles. Samples from patients that did not receive neo-adjuvant therapy (*n* = 85) and all the benign samples (*n* = 8) were used in the present study.

3.2.2 Regulons Identification and Gene Regulatory Network from Preclinical Prostate Orthograft Models

The PCa gene-regulatory network was generated as described in Section 2.3.2. The expression profiles of 2065 manually curated transcription factors and co-factors[73] (Appendix 4 - Table S1) were correlated with the differentially expressed genes in 18 prostate orthografts derived from human PCa cells, namely CWR22Res, 22Rv1, LNCaP, LNCaP-AI and VCaP (*n* = 3, except for VCaP). VCaP derived orthografts were grown in both hormone proficient and castrated mice (*n* = 3 each). Out of the 2065 transcription factors, statistically significant associations with one or more target genes were found for 1643 regulators. Further removal of the 'shadow' effect (the chance of obtaining false positive result) during the enrichment of a GRN, if a non-active regulator shares a significant proportion of its targets with a *bona fide* active transcription factor, produced a final set of 1308 regulons, with a median of 20 genes per regulon (range 2–121) identified (Appendix 4 - Table S2). Interestingly, a large fraction of

transcription factors (*n* = 607; 46.4%), shared at least one target gene (Figure 3-2A).

Based on gene regulatory network metrics described in Section 2.3.3, the Jaccard Index, a statistical measure defined as the ratio of the intersection and the union of two sets, was applied to highlight network nodes sharing a meaningful proportion of targets. The threshold of 0.1 was chosen to prioritise the nodes to be shown in Figure 3-2A. A threshold of Jaccard Index/Co-efficient set at 0.1 highlights pairs of regulons with intersection (sharing) of ≥10% of the target genes when considered across the full set of target genes for the respective regulons. By decreasing the threshold, a higher number of genes would be present in the figure.



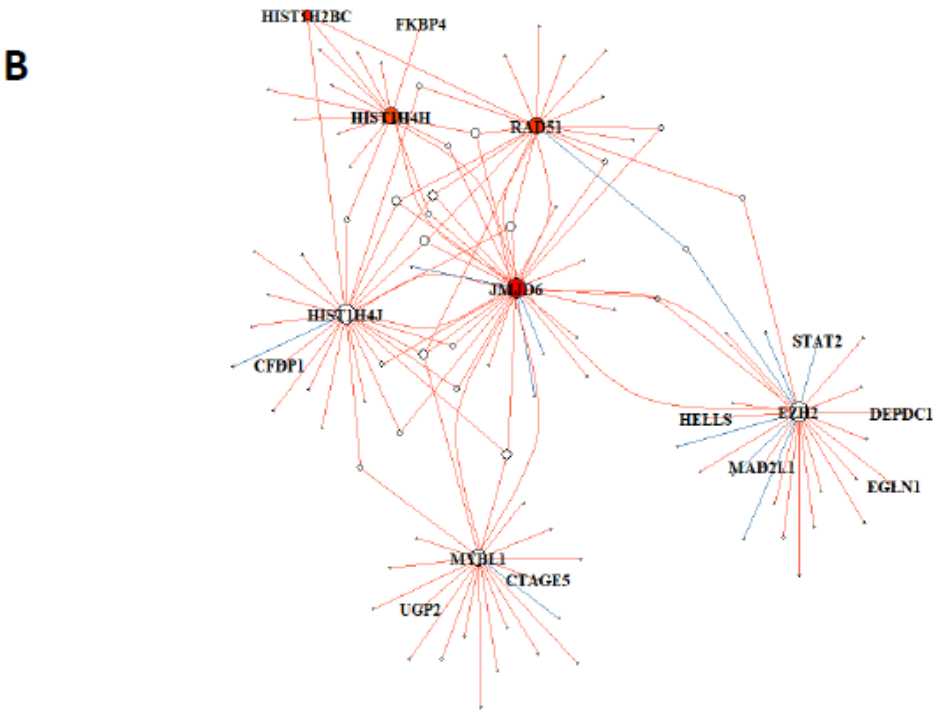


Figure 3-2. (A) Gene regulatory networks identified in preclinical human prostate cancer orthografts. The regulatory network of regulons (nodes of all colours) is presented with the edges linking pair of regulons sharing part of their targets. The commonality between pairs of regulons was calculated through the Jaccard index. Pairs with Jaccard Index ≥0.1 are shown. The colour of the nodes refers to the colour scale (range 1-0) represents the p-value of the enriched regulons associated with relapse free survival in the clinical (UTA and EMC) cohorts by cox regression analysis. Regulons in grey represent insignificant networks and therefore not included in further analysis. (B) The gene regulatory network topology cantered on the JMJD6 regulon. Red edges represent positive regulations while blue edges inhibitory relationships. (A,B) The names of regulators are annotated with HUGO gene symbol in black. The colour scale (range 1-0) represents the p-value of the enriched regulons associated with disease free survival.

Genes controlled by multiple transcription factors at the transcriptional level may suggest a functional requirement in controlling the expression of these target genes, thus signifying the likelihood of their biological importance. We searched for genes (as part of individual GRNs) predicted to be regulated by the highest number of transcriptional factors (Appendix 4 - Table S3). Up to 10 transcription factors per target gene were observed within the networks identified. Four target genes were associated with the highest number of transcription factors (n = 10), and interestingly all of these four genes have previously been implicated in PCa: *BUD31* encodes for a bona-fide AR-coactivator that enhances AR transactivation in prostate cells[99]; *PLOD3* is involved in tissue remodelling and plays a role in multiple tumour types including PCa[100]; *SDR42E1* is implicated in early prostate organogenesis as well as

carcinogenesis[101] and *XAGE1A* belongs to the cancer testis antigens family and its expression profile is linked to the aggressiveness of PCa[102]. Hence, a GRN-based analysis of prostate orthografts generated a network of candidate transcriptional regulators and their target genes that can be evaluated in clinical tumours.

3.2.3 Analysis of Differentially Expressed Genes (DEG) in Clinical PCa Patient Cohorts

Through comparison of each clinical tumour with the combined benign controls within the respective clinical cohorts, lists of differentially expressed genes (on a per sample basis) were generated on a per-sample basis initially in the UTA clinical cohort as part of a discovery analysis. The list of PCa associated genes (log2 fold changes and *p*-values) was then be used to identify the GRNs of interest, highlighting potential active regulons in individual tumours. In the UTA cohort (*n* = 27 PCa), we found a median of 2406 upregulated (range 1098–6419) and 282 downregulated (range 44–1173) genes per sample. In the EMC cohort as a validation dataset (*n* = 37 PCa), we observed a median of 2439 upregulated (range 827–7395) and 126 downregulated (range 1–925) genes for individual tumour samples.

We ranked the differentially expressed genes by the average frequency of alteration (up- or down- regulation) within the respective patient cohorts (Appendix 4 - Figure S1). Of note, many of the frequently altered genes (altered in > than 60% of the patients in the UTA and EMC cohorts) have been implicated in PCa, including *HPN*[103], *CLDN8* (an androgen regulated gene that promotes PCa cell proliferation and migration)[104], and *ONECUT2* (a known master regulator in PCa that suppresses the androgen axis)[105]. Hence, analysis of differentially expressed genes in the UTA and EMC cohorts highlighted candidate genes associated with PCa.

3.3.4 Gene Graph Enrichment Analysis

Data from transcription factor associated GRN identified in the preclinical prostate orthografts and individual gene sets from differentially expressed gene analysis on a per sample basis were integrated in a gene graph enrichment analysis (GGEA) to determine the activity status of the regulons (transcriptional regulators and their respective target genes) in the clinical tumours. The concordance of the positive and negative 'transcription factor-target gene' relationships was calculated for each sample within the UTA and EMC patient cohorts. GGEA[49] applies an enrichment approach to study the interactome surrounding the coregulators of interest to find supporting evidence of transcription factor activity.

In greater details, GGEA first maps the individual regulon under the investigation onto the full GRN to extract a subnetwork. Second, each edge of the subnetwork is scored for consistency with the expression data. Finally, the edge consistencies are summed up, normalized and statistically assessed via a permutation analysis.

Differentially expressed genes in individual tumours within the two cohorts were mapped onto the candidate GRNs highlighted in the orthograft models.

To corroborate enriched gene networks shared among independent cohorts, we ranked the regulons by the respective frequency of activation in the UTA and EMC patient datasets (Appendix 4 - Figure S2). Consistently, among the ten most frequently active transcription factors (regulators) in these two datasets, we found three known genes implicated in PCa progression: *BACH1* promotes invasion and migration of PCa cells by altering metastasis related genes[106]; *CITED2* (Cbp/P300 Interacting Transactivator With Glu/Asp Rich Carboxy-Terminal Domain 2) has recently been proposed as a therapeutic target to tackle PCa metastasis[107]; and *DNMT1* promotes PCa metastasis through the regulation of epithelial-mesenchymal transition and cancer stem cells[108]. Collectively, regulatory patterns identified in our preclinical orthograft PCa models successfully highlighted genes of potential clinical relevance.

3.2.5 Prognostic Utility of Regulon Activity Status in Radical Prostatectomy Clinical Cohorts

To evaluate the prognostic utility of the inferred regulons, we investigated the potential association between the enriched/not enriched status of regulons and the time to cancer relapse (signified by biochemical recurrence) following radical prostatectomy. We performed univariate CoxPH regression analysis in the UTA dataset in the first instance to identify enriched regulons associated with cancer recurrence (Appendix 4 - Table S4). Eleven statistically significant candidate regulons highlighted, with *JMJD6* as the top-ranking enriched regulon (p = 0.002; Table 3-1A, Figure 3-2B, Appendix 4 - Table S5). Analysing the EMC cohort as a validation dataset, fourteen enriched regulons were identified. Consistent with findings from the UTA cohort, *JMJD6* was also identified as the top-ranking enriched regulon (p = 0.003; Table 3-1B, Figure 3-2B, Appendix 4 - Table S5). Besides *JMJD6*, the *SUFU* regulon was enriched in both UTA and EMC cohorts. Analysing all available prostate cancer datasets in the cBio-portal (n = 22), altered *JMJD6* gene was detected in multiple cohorts, with the highest incidence of genetic abnormalities (up to 8%) detected in metastatic tumours (Appendix 4 - Figure S3). We reasoned that analysis of the *JMJD6* regulon as a network, rather than at a single gene level, would provide additional insight into its functional impact. Univariate regression analysis further revealed that the active *JMJD6* regulon was associated with early biochemical recurrence also in the EMC cohort (Table 3-2A). We further examined the status of the *JMJD6* regulon as a prognostic signature in the ICGC cohort for additional independent validation. Enrichment of the *JMJD6* regulon significantly correlated with time to biochemical recurrence in the ICGC cohort in univariate analysis (p = 0.00648). Kaplan-Meier analysis for biochemical free survival further confirmed reduced biochemical free survival in the presence of active status for the *JMJD6* regulon in patients within the UTA, EMC and ICGC cohorts (Figure 3-3).

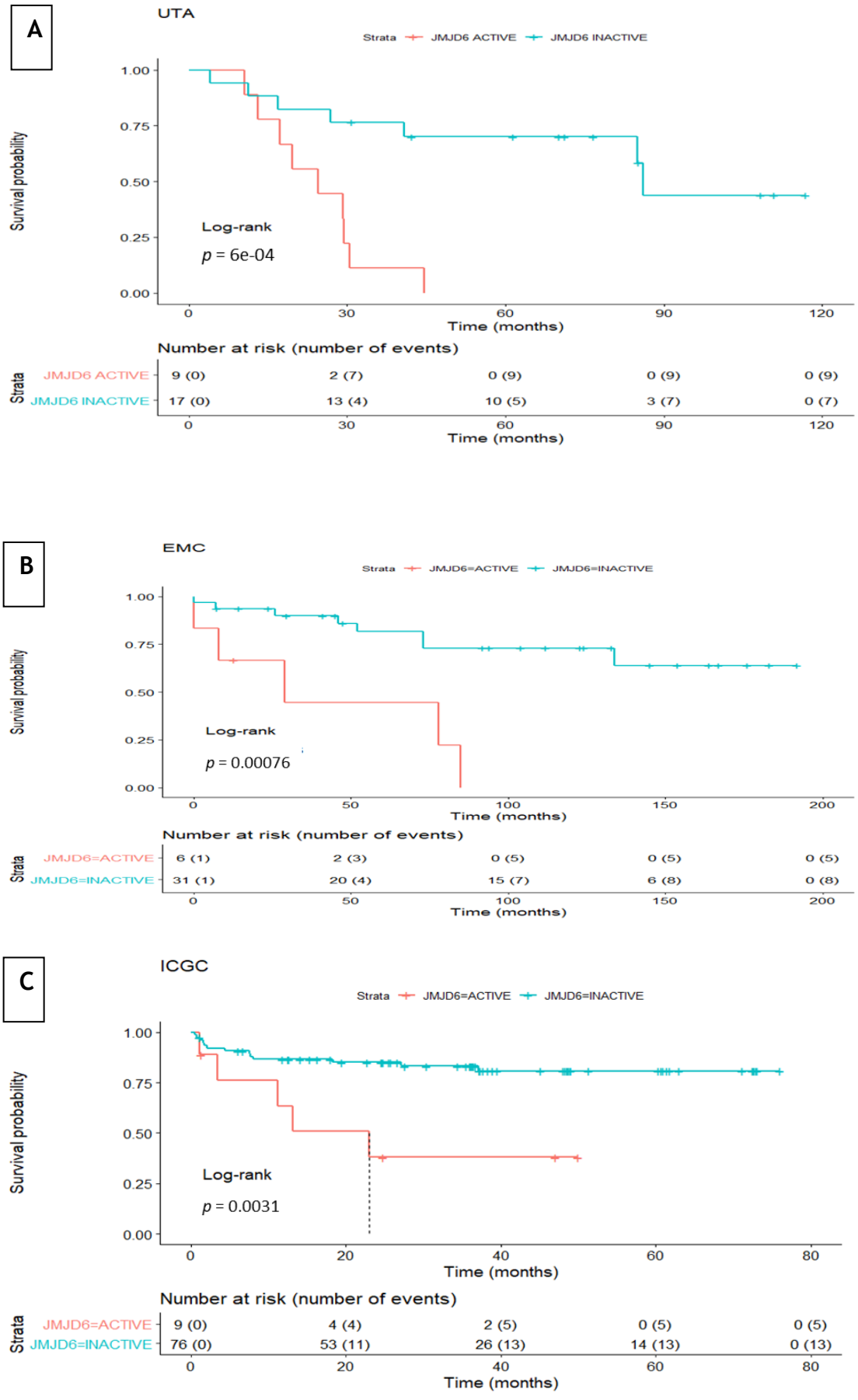


Figure 3-3. Disease free survival analysis of the JMJD6 regulon signature in clinical prostatectomy patient cohorts. The survival probability curves for patients in the UTA (A),

EMC (B) and IGCG (C) cohorts were prepared with patients stratified according to the presence or absence of the enriched JMJD6 regulon in red and turquoise, respectively.

(**A**)		
ENSEMBL ID	**HUGO SYMBOL**	***p* VALUE**
ENSG00000070495	*JMJD6*	0.002
ENSG00000196132	*MYT1*	0.006
ENSG00000100410	*PHF5A*	0.02
ENSG00000065057	*NTHL1*	0.02
ENSG00000159210	*SNF8*	0.02
ENSG00000171222	*SCAND1*	0.02
ENSG00000123091	*RNF11*	0.02
ENSG00000120798	*NR2C1*	0.02
ENSG00000107882	*SUFU*	0.03
ENSG00000146083	*RNF44*	0.04
(**B**)		
ENSEMBL ID	**HUGO SYMBOL**	***p* VALUE**
ENSG00000070495	*JMJD6*	0.003
ENSG00000095002	*MSH2*	0.006
ENSG00000107882	*SUFU*	0.007
ENSG00000136826	*KLF4*	0.01
ENSG00000119969	*HELLS*	0.01
ENSG00000151929	*BAG3*	0.02
ENSG00000105607	*GCDH*	0.02
ENSG00000092607	*TBX15*	0.02
ENSG00000188486	*H2AFX*	0.02
ENSG00000180596	*HIST1H2BC*	0.03

Table 3-1. Univariate cox regression analysis for regulons enrichment. Top ten genes are listed for the (A) UTA and (B) EMC cohorts.

To benchmark the *JMJD6* regulon as a prognostic marker for progressive/recurrent PCa, three reported independent signatures were selected for comparison: two androgen receptor related signatures (namely *TMEFF2* regulated cell cycle related gene signature[11] and the bromodomain related 10-genes signature[12]) as well as a 28-gene hypoxia signature[13]. The three signatures are referred to as TMCC11, BROMO-10 and HYPOXIA-28 respectively hereafter. Compared to *JMJD6* being prognostic in all three cohorts, TMCC11 was prognostic in the UTA and ICGC cohorts but not the EMC cohort, while BROMO-10 and HYPOXIA-28 significantly predicted recurrence in only one of the three cohorts, UTA and ICGC, respectively (Table 3-2A). Multivariate analyses of the three signatures (and of the *JMJD6* regulon status) were performed if the

respective univariate analysis were significant. In multivariate analysis, the *JMJD6* regulon status significantly predicted disease recurrence in UTA and EMC, but not ICGC (Table 3-2B). Among the three published signatures, none significantly prognosticate for cancer recurrence in multivariate analysis.

Collectively, our analysis highlights the feasibility of integrating data from preclinical human orthograft models of PCa with multiple clinical cohorts to generate information on the regulon landscape in identifying potential prognostic signatures. For the first time, our data identified the active status of the *JMJD6* regulon in patients at risk of PCa recurrence.

(A) Univariate analysis									
Clinical cohorts		**UTA**			**EMC**			**ICGC**	
Statistics	HR	95% CI	*p*	HR	95% CI	*p*	HR	95% CI	*p*
Clinicopathological variables									
Gleason score	2.7	1.6–4.7	**0.0004**	1.9	0.2–16	0.5	2	1.4–3	**0.0004**
Tumor stage	1.7	1–2.96	0.05	1.3	1–1.6	**0.02**	2.5	1.7–3.8	**<0.0001**
Signatures									
active *JMJD6* regulon	6	1.9–18	**0.002**	5.8	1.8–18.6	**0.003**	4.2	1.5–12	**0.006**
TMCC11	4.5	1.1–17.8	**0.03**	1	0.3–3.7	1.0	4	1.6–10.5	**0.004**
BROMO-10	0.06	0.0069–0.52	**0.01**	1.2	0.3–4.2	0.8	2.6	0.7–9.3	0.2
HYPOXIA-28	2.1	0.7–6.24	0.2	1.1	0.4–3.5	0.8	3.4	1.3–9.2	**0.01**
(B) Multivariate analysis									
		UTA			**EMC**			**ICGC**	
	HR	95% CI	P	HR	95% CI	P	HR	95% CI	P
JMJD6 regulon	6.5	1.3–32	**0.02**	4.4	1.3–14.6	**0.01**	1.2	0.3–4.8	0.7
Gleason score	1.6	0.8–3.1	0.2				1.2	0.6–2.4	0.6
Tumor stage	2.3	1.1–4.9	**0.03**	1.2	1–1.5	0.05	2.6	1.3–4.9	**0.004**
TMCC11	3.4	0.8–14.4	0.1				1.8	0.6–5.6	0.3
Gleason score	2.5	1.4–4.4	**0.002**				1.3	0.7–2.4	0.5
Tumor stage	1.58	0.8–3.2	0.2				2.2	1.1–4.4	**0.02**
BROMO-10	0.3	0.03–4.2	0.4						
Gleason score	2.15	1.08–4.27	**0.03**						
Tumor stage	1.5	0.8–2.8	0.2						
HYPOXIA-28							2.1	0.7–6.1	0.2
Gleason score							1.3	0.7–2.4	0.4
Tumor stage							2.2	1.13–4.24	**0.02**

Table 3-2. Cox regression univariate (A) and multivariate (B) survival analysis. p-values (P), Hazard ratios (HR) and 95% Confidence intervals (CI) are showed for each univariate regression. Multivariate analysis results, using Gleason score and/or tumour stage as covariates, are shown only for the variables whose association with biochemical recurrence was significant (p < 0.05) at univariate level. All significant p-values are highlighted in bold.

3.2.6 *In vitro* validation of JMJD6

Three experiments were performed to further assess the functional contribution of JMJD6 in prostate cancer.

I compared the level of JMJD6 expression in a panel of human prostate cancer cell lines along with the RWPE benign prostate epithelial cell line: LNCAP, LNCAPAI, C4-2, CWR22, 22Rv1, VCAP, DU145, PC3, PC3M. JMJD6 protein expression was higher in the prostate cancer cell lines when compared to the normal RWPE cell line (Figure 3-4), suggesting a potential role for JMJD6 in prostate cancer carcinogenesi.

To formally investigate JMJD6 mediated functions, I employed small interfering RNA (siRNA)-mediated knockdown (silencing) of *JMJD6* mRNA expression. Experiments were performed with cells cultured in either androgen proficient or androgen depleted conditions, with supplement using either full bovine serum or charcoal stripped serum (CSS, signifying steroid depletion) respectively. JMJD6 protein expression was strongly reduced following transfection with JMJD6 targeting siRNA (Figure 3-5), demonstrating the efficacy and specificity of the methodology.

Once JMJD6 expression was convincingly suppressed following siRNA transfection, I carried out *in vitro* proliferation assay using the Incucyte systems for live-cell Imaging and analysis platform. The growth of both hormone naïve CWR22 and its derived isogenic 22Rv1 cells was studied, culturing CWR22 cells in both androgen proficient and deprived conditions and 22Rv1 in androgen deprived condition. In this way, three sets of comparisons were possible to assess the impact of silencing JMJD6 in cellular proliferation, namely CWR22 cells cultured with and without androgens and 22Rv1 cells in 'castrated' condition. I observed marked suppression of growth in each control-treatment pair in the five days of incubation. Noteworthy, in the pair of cell lines tested, JMJD6 loss affected both hormone naïve and castration resistant cells (Figure 3-6). Interestingly, the difference observed in growth due to JMJD6 loss appears to be higher for cells cultured in androgen deprived condition. This is particularly evident at t=80 hours (Figure 3-7), when the CWR cell lines grown in the androgen proficient serum showed the largest difference in growth following JMJD6 knock down: 2.94 times the number of cells at time 0 for CWR_FBS_C

versus 1.87 times for CWR_FBS_J, with a relative reduction of viability of 36%; 31% reduction for CWR cultured in CSS; 25% reduction for the 22Rv1 pairs.

In summary, these preliminary data is consistent with the idea that the *JMJD6* gene and its associated regulon may play a role in prostate carcinogenesis as suggested by my *in-silico* gene regulatory network analysis.

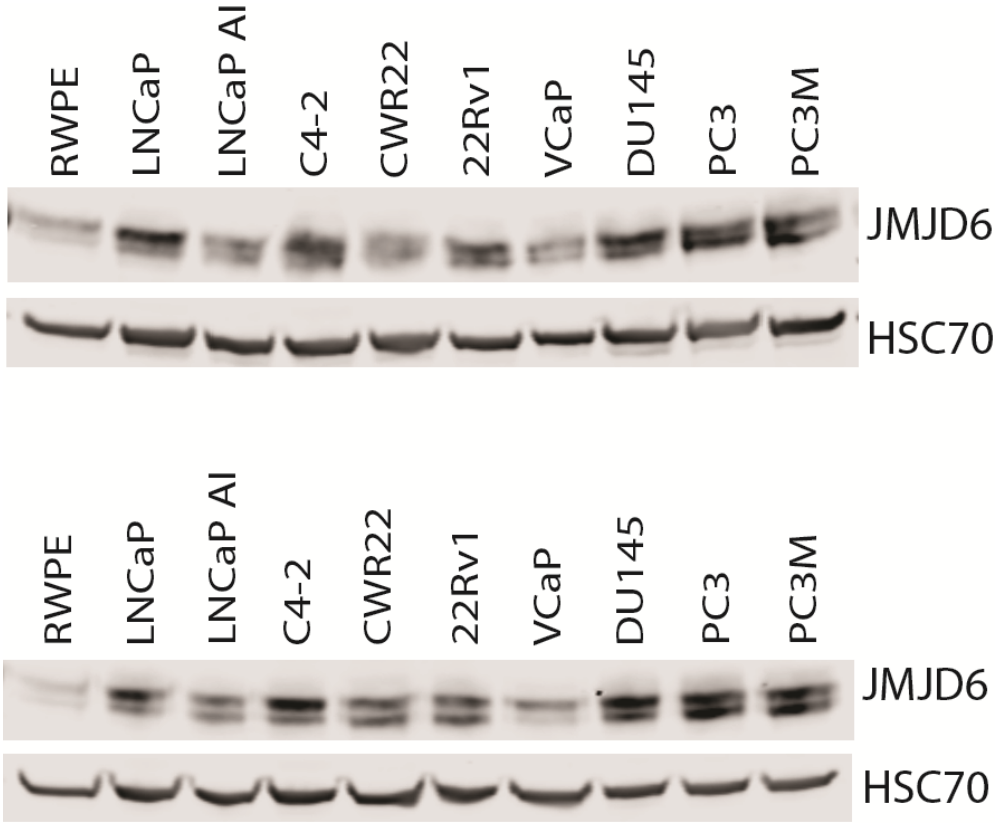


Figure 3-4. Western blot of prostate cell lines lysates (n= 2 biological replicates). RWPE cells are represent a benign prostate epithelial cell line while all other cell lines are models of prostate cancer (including hormone naive, hormone resistant, metastatic status).

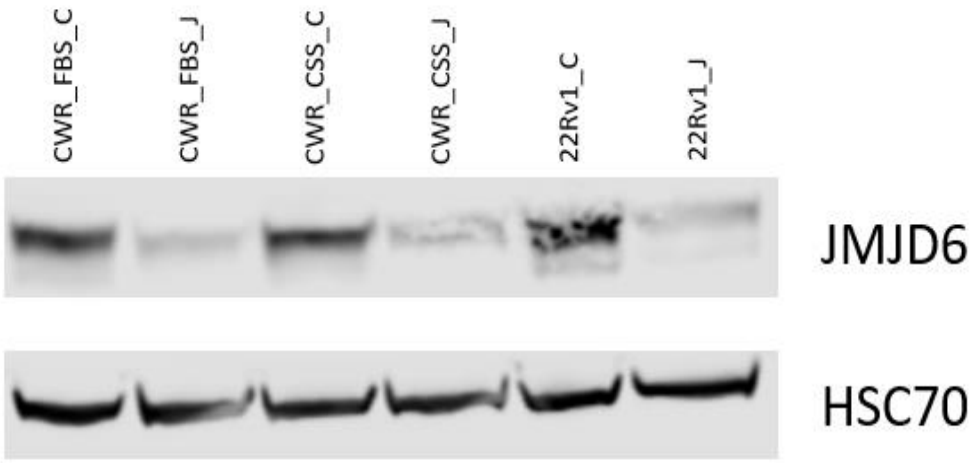


Figure 3-5. Western blot of siRNA mediated knockdown of *JMJD6* expression in CWR and 22Rv1 cell lines, cultured in FBS or CSS serum (n=1) respectively. Control samples, namely treated with untargeted siRNAs, are labeled with 'C', while JMJD6 targeted samples are labeled with 'J'.

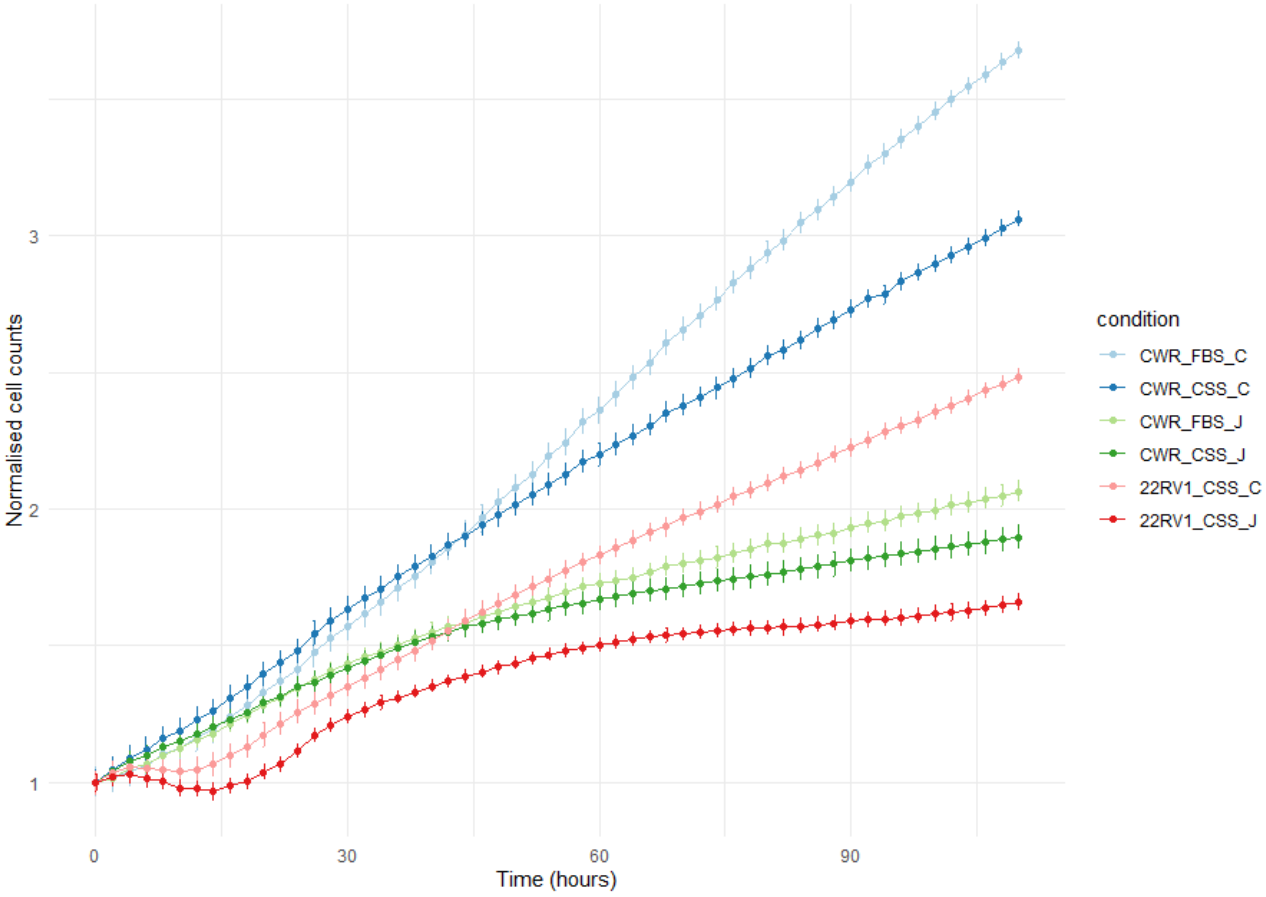


Figure 3-6. Normalised cell counts at different timepoints of incubation following siRNA mediated JMJD6 knock down with non-silencing control siRNA (n=10 technical replicates for each timepoint, n=1 experiment). X-axis refers to hours of incubation and Y-axis to the ratio of cell conflunce at time t versus t0. The norm value of the cell counts was obtained by

dividing the mean of the 10 replicates, for each condition and each timepoint, with the values mean at time 0, respectively to the each condtion. The error bars represent the standard error associated to the 10 replicatesfor each condition..

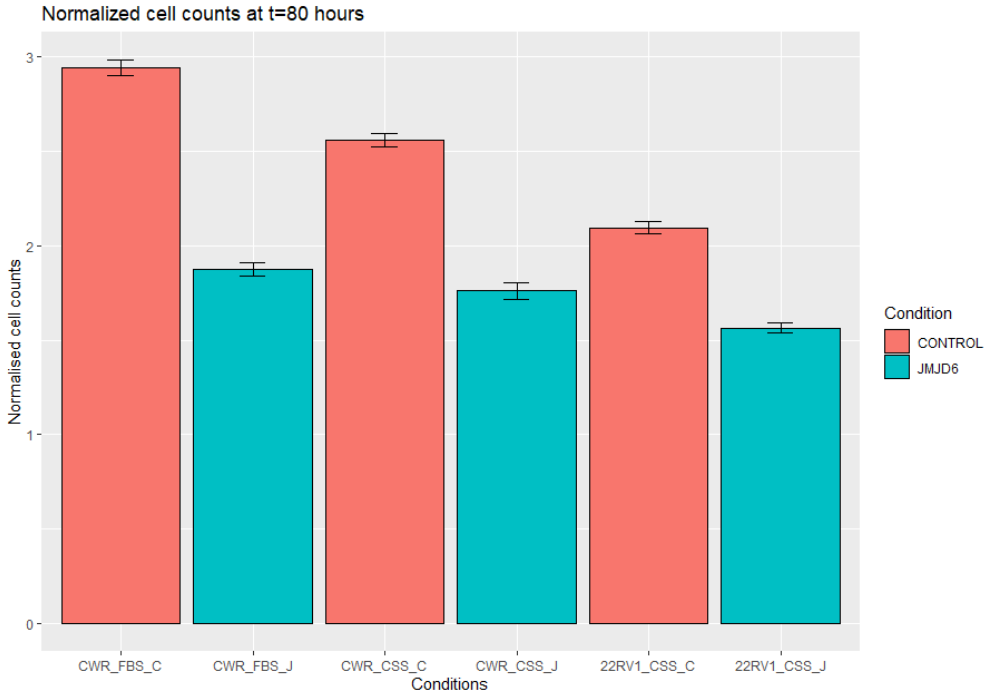


Figure 3-7. Boxplot of normalised cell counts at 80 hours of incubation (n=10 technical replicates for each timepoint, n=1 experiment). Cell counts were normalised by dividing the mean of the 10 replicates at t=80 with their mean at t=0, respectively to the each condtion. The error bars were obtained by diving the standard deviation of the 10 replicates at t=80 with the mean cell count at the same t, respectively to each condition.

3.3 Discussion

We hypothesised that the study of genes positively and negatively regulated by one or more transcription factors (collectively referred to as regulons) is a suitable approach to capture the general mechanisms driving tumour progression in PCa[109]. For the first time, we integrated datasets from preclinical human prostate orthografts and clinical cohorts to investigate if specific regulons were associated with the outcome of patients with PCa. By mapping transcriptomic gene graph enrichment-based signatures on to a network of interacting gene regulators, we identify the *JMJD6* regulon as a candidate prognostic signature for biochemical recurrent PCa. Our analysis is consistent with a recent report on GRN-based investigation in breast cancer[35] while our data on JMJD6 in PCa is

consistent with involvement of JMJD6 in oral[110], breast[111], neuroblastomas[112], melanoma[113] and ovarian[114] cancers.

The *JMJD6* regulon consists of 27 positive and 3 negative putative target genes (Appendix 4 - Table S5), including *RAD51, EZH2* and *SORL1*. *RAD51* is predicted to be upregulated by *JMJD6* (Figure 3-2B). *RAD51*, a critical gene for the DNA repair process, is upregulated in aggressive PCa[115], and is included as part of the panel in the U.S. Food and Drug Administration approved Prolaris gene expression assay[116]. Similarly, *EZH2* (Enhancer of zeste homolog 2) is associated with PCa progression[117], and predicted to be upregulated by JMJD6 (Appendix 4 - Table S5). Lastly, the expression of *SORL1*, a known hypoxia regulated gene[43], negatively correlates with JMJD6 expression.

We successfully identified regulons of interest from preclinical prostate orthografts and then investigated the prognostic value of our top candidate *JMJD6* regulon. Given the small number of preclinical samples available as a starting point to infer the GRNs in PCa, we were not able to robustly compare between hormone naïve and castration resistant orthografts. Instead, we combined the available orthografts to model tumour heterogeneity of clinical PCa. Importantly, some transcription factor-target genes relationships may not be revealed because of the limited sample number, thus creating potential biases with a subset of regulons appearing transcriptionally more important. Nonetheless, even with this limitation, the *JMJD6* regulon was identified as a key regulon enriched in two independent clinical cohorts, namely UTA and EMC, as well as the published independent ICGC clinical cohort. The ICGC cohort consists of relatively young patients (mean: 47, range: 35–52 years), compared to UTA (mean: 60, range: 47–71 years); such case selection bias may create confounding factors that contribute to the negative multivariate analysis for the *JMJD6* regulon in the ICGC cohort.

Although androgen receptor (AR) is essential for both prostate organogenesis and carcinogenesis, to our surprise, AR was not identified as an enriched regulon in

our analysis. AR may be functionally important in both benign and malignant prostatic epithelium, with distinct transcriptional profiles arising from functional re-programming. Even in CRPC, AR remains activated through by-pass mechanisms despite suppressed canonical (classical) androgen receptor pathway activities[118]. In addition, changes due to reprogramming of the AR as a transcription factor may not be fully highlighted by analysis of regulons as fixed transcription factor-target genes 'units'. Furthermore, AR splice variants (including AR-V7) are strongly implicated in CRPC. During the preparation of this chapter, a highly relevant publication highlighted the relationship between catalytic function of JMJD6 and the generation of AR-V7 mRNA in advanced prostate cancer[119]. Silencing of JMJD6 expression suppressed growth of LNCaP95 and 22Rv1 human CRPC cells, while combined *JMJD6* knockdown and anti-androgen treatment with enzalutamide produced substantially more anti-proliferative effects than each of the two treatments alone. Collectively, their data implicates JMJD6 to be important in PCa cell viability and proliferation, thus further supporting our GRN-based findings. Noteworthy, our single (due to time limitations and Covid traveling restrictions) gene knock-down experiment on CWR and 22Rv1 cell lines supported the idea that JMJD6 may contribute to prostate cancer growth, suggesting further detailed evaluations of JMJD6 function in prostate cancer including CRPC will provide more mechanistic information.

The strategy of standardising the analysis, by adopting a panel of benign controls within each dataset (benign prostatic hyperplasia for the UTA and ICGC cohorts; benign tissue adjacent to the tumour for EMC cohort), allowed the reduction of biases arising from different protocols for sample handling, sequencing and data processing. Indeed, by leveraging a panel of control samples within each cohort, it was possible to show commonalities among the independent data sets without resorting to batch correction.

JMJD6 belongs to the Jumonji C (JMJC) domain-containing family of proteins, thought to function mainly as a lysyl 5-hydroxylase and not as a demethylase[120], although enzymatically it has been shown to possess both catalytic activities. Its

ability to regulate the transcriptional activity of p53 through hydroxylation of a lysine in the p53 C-terminus is highly relevant in cancer biology. Upregulated JMJD6 expression is related to tumour growth, tumour metastasis and high tumour pathological classification[121-123]. To build on our findings, the classical Waddington epigenetic landscape[124] model can be applied to describe in more detail the mechanism of regulation for the target genes within the *JMJD6* regulon. Given its potential role in a number of tumour types, a novel JMJD6 specific inhibitor SKLB325 has recently been developed[114]. Should future research confirm *JMJD6* as a driver gene for progressive PCa, formal evaluation of JMJD6 targeted therapy will be warranted.

Chapter 4 - Integrative proteomics and transcriptomics analysis to identify functional modules in castration resistant prostate cancer

4.1 Background

4.1.1 Unbiased proteomic analysis

Unbiased proteomics profiling refers to the quantification of all detectable proteins in a biological sample without resorting to *a priori* information about the molecules. The opposite protein quantification paradigm is referred to as targeted proteomics analysis, testing specific hypothesis for a pre-selected subset of amino acid chains (or peptides)[125]. In both cases, mass spectrometry[126] (MS) is the most common analytical technique used to measure proteins abundance. A mass spectrometer determines the mass-to-charge (m/z) ratio of gas-phase ions derived from peptides generated by proteolytic digestion. The peaks observed in the resulting spectra are then searched within a database of predicted peptides mass values, generated by an *in-silico* digestion of each protein. Statistical analysis and significance thresholds are then applied to detect proteins from high scoring peptides matches. Relative or absolute protein quantification is finally obtained by either comparing peaks abundances of differentially labeled peptides or by using a known amount of labeled peptides as a reference, respectively[127].

The first unbiased method applied in proteomics was the two-dimensional gel electrophoresis in which proteins are electrically separated according to molecular weight and charge. Despite the development of newer techniques, the 2D gel arrays offers a convenient method for the recognition of protein isoforms of interest. Subsequently, metabolic labeling introduced the practice of incorporating known markers into biological samples to provide relative quantification of peptides based on the ratios of mass spectrometry generated peaks for respective isotopes pairs. A typical example of such techniques is the stable isotope labeling by amino acids in cell culture (SILAC). Other methodologies supporting unbiased proteomic quantification include:

(1) Proteolytic labeling of proteins labelled with Oxygen-18 (O-18) within the carboxylic acid group allows quantitation based on the ratio of the mass spectrometry derived peaks which signify the respective isotopes for peptides containing O-18 and O-16.
(2) Isobaric tags for relative and absolute quantification (iTRAQ) employs covalent labeling of tags of varying mass at the N-terminus and side chain amines of peptides generated from protein digestions. iTRAQ is particularly useful to compare the abundance of proteins from different sources in a single experiment.
(3) Addition of synthetic internal standards to the protein sample being analysed, to provide absolute quantification of multiple peptides, and then proteins, that share chemico-physical properties with the standard[128].

Building on data from transcriptomic analysis presented in Chapter 3 of this thesis, I wish to incorporate data from the proteomes of the same preclinical prostate tumour samples studied by RNA sequencing. The analysis discussed in this chapter is based on proteomics quantification of hormone naïve and hormone resistant prostate cancer orthografts using a SILAC based approach, in which arginine and lysine heavy isotopes were incorporated into the respective *in vitro* cultured human prostate cancer cells and combined to provide a SILAC standard for adding to the lysates (light isotopes, or without labelling) from different prostate cancer orthografts being examined. In this way, data from orthografts derived from different cancer cell lines and different mice (with and without castration, mimicking clinical androgen deprivation therapy) can be confidently compared, minimising risk of biases and maintaining reliability and robustness of the quantitation[129].

Similar to the regulon-based analysis on transcriptomic data, the proteomes of 18 orthografts from three sets of hormone naïve and castration resistant prostate orthografts were available for analysis, with cell line represented as triplicate orthorgrafts[130].

4.1.2 Consideration of a SILAC based approach to support quantitative proteomic analysis of prostate orthografts

Heavy labeled Arginine (Arg-10) and Lysine (Lys-8) isotopes were incorporated into four PCa cell lines (CWR, LNCAP, LNCAPAI and VCAP) lysates that, in turn, were mixed at 1:1:1:1 ratio to generate a super SILAC standard. The pooled cell lines standard can be leveraged to control for batch effects or sample specific biases. Each tumor sample, from a set of 18 mice PCa orthografts obtained from biological triplicates of three hormone naïve (CWR, LNCAP and VCAP) and matched CRPC (22Rv1, LNCAPAI and VCAPR) cell lines, were mixed at 1:1 ratio to the SILAC standard before undergoing digestion into peptides and liquid chromatography–mass spectrometry (LC-MS) online analysis.

The output of the mass spectrometer consists of the spectra containing peptide mass and intensity information. After processing of the raw data through software toolkits such as MaxQuant[131], the identity of the peptides can be deduced by searching for their characteristic spectra in a protein database such as the ones hosted in Swissprot[132]. The analysis can be performed adopting different proteomics search engines such as Andromeda[133] or Mascot[134]. Moreover, orthografts derived spectra can be used to query both human and murine databases to use the resulting hits as input for different downstream analysis.

Equal mixing (1:1 ratio) of light (from *in vivo* orthografts) and heavy (from *in vitro* cultured cell controls) isotopes-labelled lysates ensures robust comparison of the two quantifications and allows the ‘heavy’ intensities to function as a sample-specific normalisation factor. In addition to the use of SILAC intensities derived ratios, label-free quantification of individual proteins (peptides) can be obtained directly from each mass spectrometry based proteomic profiles. The Label Free Quantification (LFQ) algorithm introduced the concept of ‘delayed normalisation’, which exploits a different normalisation factor for each peptide fractionation step to increase quantification accuracy. This label-free procedure brings the additional advantage of directly evaluating the differences in protein abundance between different conditions[135].

In SILAC based analysis, data on a specific protein are considered informative if the respective SILAC ratios of heavy to light isotopes of relevant peptides in the experimental conditions being tested are different, i.e., ratio not equal to 1. In contrast, when label free quantitative analysis of the actual spectrometric peaks is studied, individual proteins are considered informative if their absolute values are greater than zero, i.e., detected with confidence. Of note, within a breast cancer study, SILAC intensities-based analysis showed a slightly lower coefficient of variation for protein quantification when compared to label free based analysis (median coefficient of variation of 13.7% against 16.3%, respectively). Nevertheless, the label free approach provided ~60% more informative proteins (1624 over the 1036 from SILAC) and higher reproducibility (~20% more proteins were quantified in all replicate samples)[136]. Proteomic data from both SILAC and LFQ based quantifications can then be used as inputs to carry out statistical analysis for multiple purposes, including the identification of differentially expressed proteins, the determination of protein production (as translational output) and their turnover rates, as well as inference of protein-protein interactions.

4.1.3 Relationship of gene expression at the RNA and protein levels

As the biology central dogma would suggest, protein expression should show a direct correspondence with mRNA expression levels, as the first are translated from the latter. Nevertheless, different factors regulate mRNAs and proteins life cycle[60], resulting in poorly correlated mRNA and proteins quantifications profiles from the same sample when high-throughput or omics methodologies are applied for correlative analysis, such as next generation based RNA sequencing and SILAC based protein analysis.

Therefore, an integrated transcriptomics and proteomics analysis may provide broader overall picture of molecular alterations that may explain the observed biological phenotypes. Different approaches can be implemented to take advantage of both data sources within the same analysis workflow. For example, the matched data on mRNA and protein expression levels can be combined into

single values for each sample, and subsequent merging of data from biological replicates for each sample type may more reliably represent different phenotypes, in other words hormone naïve versus castration resistant prostate cancer. Data at mRNA and protein levels can further support global analysis to consider the extent of enrichment for common functional pathways or biological processes ontologies. Similarly, topological network can be inferred and investigated by focusing on known or predicted mechanistic interactions among measured genes and proteins. Lastly, clustering approaches can be implemented to explore the possibility of a consensus subgroups within the studied samples, based on data common to both sources of biological information (namely mRNA and proteins)[60,59].

To my knowledge, despite previous attempts in correlative analysis between mRNA and protein quantifications[32], no methods have been developed to formally exploit the complementary nature of the transcriptome and the proteome of prostate cancer by merging data on co-expression modules. As discussed in Chapter 3 of this thesis, gene-gene co-expression can reveal relationships among transcriptional regulators and their targets[137]. In addition, coherent expression of proteins may directly implicate functional protein complexes[138–140], also illustrated in the CORUM repository[141], a comprehensive resource of mammalian protein complexes. Integrating transcriptomic and proteomic datasets can potentially help nominate master regulator(s) that play a key role in the underlying biology, which may otherwise not be identified. In this way, data obtained from an integrated mRNA-protein analysis will generate new hypothesis to explain the intersection between transcriptional and translational regulatory mechanism in driving castration resistant prostate cancer.

In this chapter, I explore the possibility of developing a novel integrative network analysis pipeline by coupling the regulons from the gene regulatory network generated from RNAseq data with established protein complexes listed in the CORUM repository[141]. The integration was achieved leveraging shared features between each regulons' target gene set and the protein complexes that contain at least one of the transcriptional targets from the regulons of interest. The generated integrative graph structure represents both the regulatory relationships and functional protein groups associated to a singular transcriptional regulator (Figure 4-1). In greater detail, networks of protein-

protein interactions were derived by mapping the data from SILAC analysis onto known protein complexes. In particular, the strength of known protein-protein relationships was calculated directly from the prostate cancer orthografts proteomics profiles. I was then able to perform a combined enrichment analysis, coupling differentially expressed genes and proteins to highlight differences between hormone naïve and castration resistant prostate orthografts.

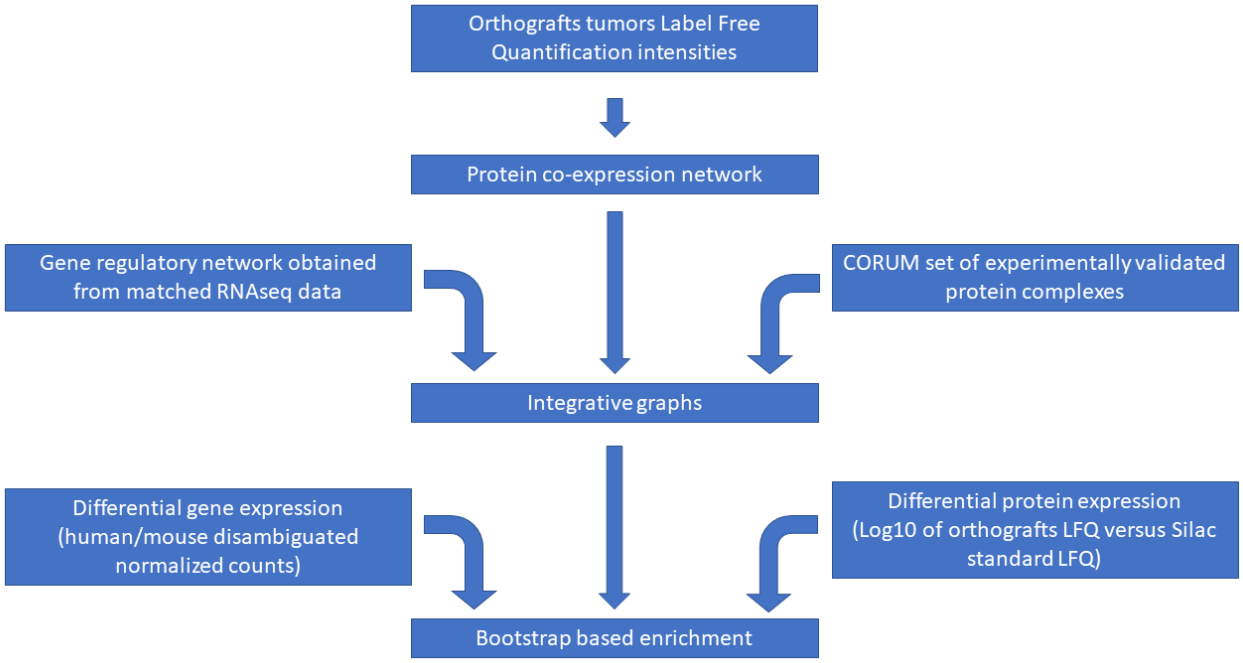


Figure 4-1. Analysis workflow HN vs CR PC.

4.2 Results

4.2.1 Evaluating the proteomics profiles of hormone naïve and castration resistant prostate orthografts

I wish to develop a robust methodology to carry out proteomic based analysis on data from the hormone naïve and castration resistant prostate orthografts. Different approaches to analyse data on the proteome were considered with a main objective to maximise the number of proteins (or the proportion of the proteome) yielding informative data. As described earlier in section 4.1.2, in SILAC based analysis, data on a specific protein are considered informative if the respective SILAC ratios of heavy to light isotopes on individual peptides between the experimental and control conditions (namely castration resistant and hormone naïve prostate orthografts) being tested are not equal to 1, signifying that the protein of interest is differentially expressed. In contrast, for label free quantitative analysis, individual proteins (represented by specific peptides) are considered informative if their absolute (and normalised) values are greater than zero, i.e. detected with confidence.

Mass spectrometry derived proteomic data can be presented in three ways based on different normalisation approaches:

(1) The abundance of individual proteins is measured by the spectrometric peaks generated by specific lytic peptides. Data from tumor tissues will contain light isotopes (Arg0 and Lys0) while the SILAC standard containing heavy isotopes (Arg10 and Lys8) were generated from selected human prostate cancer cell cultures with *in vitro* supplement of heavy isotopes in the medium. Normalisation of each protein detected was performed by generating the ratio for protein expression (peptides with light isotope) in the tumour to the respective peptide peaks for the heavy isotopes within the SILAC standard generated from cultured cells.

(2) Label free quantifications (LFQ) of the tumours can be analysed to provide quantitation of the respective peptides generated from individual proteins. (Data from murine proteins are included in the mass spectrometry data and can be included or excluded from analysis as appropriate.)

(3) Data from label free quantification of proteins in the orthografts can also be normalised to the respective label free data from *in vitro* cultured cell standard. (SILAC based data as a result of heavy isotope labelling are not utilised here.)

In summary, to compare the amount of data provided by the above analytical approaches, different criteria are considered. For data presented as a ratio, proteins with a ratio > or < not equal to 1 are differentially expressed between the *in vivo* and the *in vitro* experiments. In contrast, for label free quantitative analysis, peptides with absolute values >0 are detected with confidence and hence utilised in downstream analysis. Figure 4-2 presents the profile of data based on different approaches of data analysis as discussed above. Density plot from label free data generated from orthografts are referred to as LFQ light label, represented as Figure 4-2A. Mass spectrometry generated data from the *in vitro* cell lines with SILAC labelling are referred to as LFG heavy label, represented as Figure 4-2B. Panels C and D of Figure 4-2 represent analysis of the data as ratios, based on light isotope (orthografts) compared to heavy isotope data (standard generated from cell lines) for panel C, and label free value for peptides observed in orthografts compared to label free value for the same peptides observed in the standard from cultured cells for panel D. All panels are represented after log10 transforming the original data as described in the materials and methods section. No outliers among the 18 orthografts were observed in for all four profiles (Figure 4-2), confirming the robustness of the overall experiment. Label free quantification derived peptide profiles revealed a significant proportion of non-zero values (Figure 4-2A-B), signifying informative proteins. In contrast, peptide distributions based on ratios derived from SILAC and LFQ analysis revealed majority of the peptide with a value of zero in the logarithmic scale, signifying that the peptides are at the same abundance (Figure 4-2C-D).

I then calculated the pairwise Pearson's correlation coefficients within each dataset (as described in Figure 4-3) to assess if the different analytical approaches may show association with origin of individual cell lines and whether

they are hormone naïve and castration resistant (Figure 4-3). I reason that, given the known phenotypical differences of the cell lines used in this study, the patterns of clustering represent a proxy for the reliability of the overall proteomics profiles.

The heatmap generated from orthografts derived label free data (LFQ light, L) showed good correlation based on the origin of the cell lines, with minimum correlation value of 0.49 (Figure 4-3A), which contrasts to the heatmap from LFQ data from heavy isotope labeled cells showing a strong overall similarity with a min coefficient of 0.9 (Figure 4-3B), which was not surprising given the nature of pooling to generate the *in vitro* standard from cultured cells. Samples from SILAC and LFQ derived ratios (Figure 4-3C, D respectively) did not show strong similarity to each, with large range of the correlation value from 0.2 to 1.0.

Dendrograms in panels A and C of Figure 4-3 highlight the evidence of clustering among different cell lines, which was not observed in panels B and D. Of note, LFQ ratios-based analysis of the orthografts (Figure 4-3D) revealed the presence of two macro-clusters, more likely reflecting the separation of individual cell lines within different sets of hormone naïve and castration resistant prostate cancer models. This creates the opportunity for meaningful analysis to compare between hormone naïve and castration resistant prostate cancer.

Based on these observations, I reason that analysis of normalised LFQ values from orthografts (light labeled data, with higher proportion of non-zero values and moderate overall correlation within the set) can be applied to generate a protein-protein interaction network.

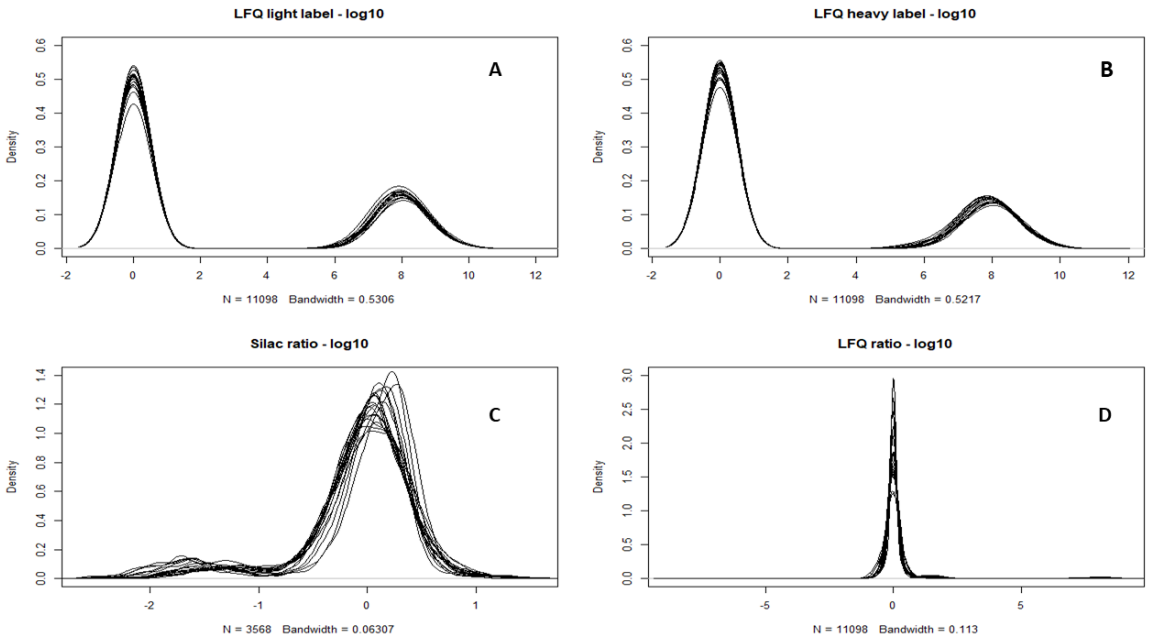


Figure 4-2. Density plot of the four available prostate cancer models proteomics profiles (namely LFQ light, LFQ heavy, SILAC ratio and LFQ ratio, see text for detail description). The x axis shows the logarithmic value of the normalised intensity (added to 1) (A and B) or ratio of intensities between light and heavy labeled samples (C and D). N refers to the number of data points used. Bandwith refers to the automatically calculated smoothing parameter from the R 'density()' function.

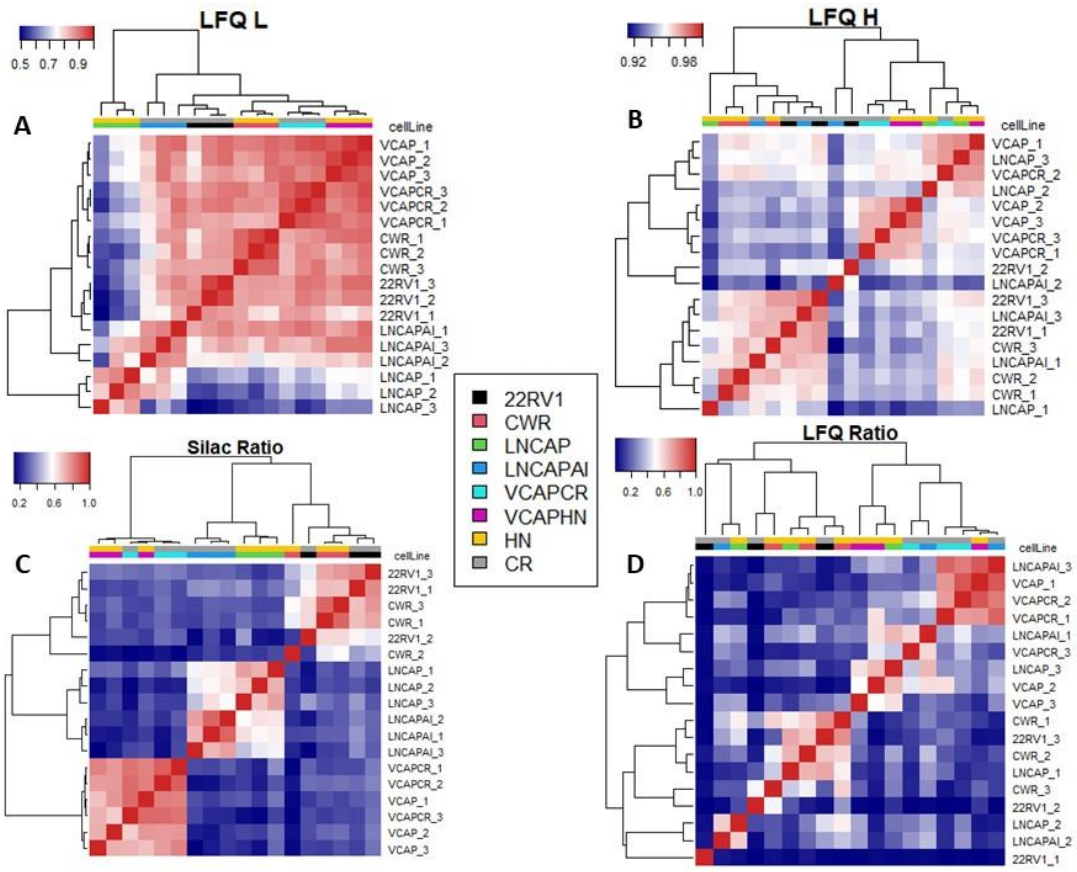


Figure 4-3. Heatmaps of raw Pearson's correlations values among the samples of each independent proteomics quantification set: LFQ L (light), LFQ H (heavy), SILAC Ratio and LFQ Ratio as explained in text. Colour bars below dendrograms highlight the cell line identity and their status as hormone naïve (HN) or castration resistant (CR) (bottom and top respectively of the bars) Blue and Red shades reflect low and high levels of correlations, respectively, according to the colour scale at the top left of each heatmap.

Besides clustering among different samples, differentially expressed proteins were also identified using the different analytic approaches. As described earlier that heatmap on LFQ ratio revealed the largest range of variation, i.e., data from different samples appear to be highly different. Hence, evaluating the intra-dataset variability in terms of detected differentially expressed proteins between hormone naïve and hormone resistant orthografts, I observed highest number of differentially expressed proteins following analysis based on LFQ ratios (Appendix 5): LFQ Light, 12 proteins; LFQ heavy, 9 proteins; SILAC ratio, 5 proteins and LFQ ratio, 227 proteins. Hence, data from LFQ ratios were applied for subsequent downstream enrichment analysis.

4.2.2 Protein co-expression analysis

The proteins quantifications of all the 18 orthografts (LFQ values normalised data from tumours) was given as input to the R package 'ProCoNa'[54] to calculate a co-expression adjacency matrix in the prostate orthografts. 'ProConNa' based analysis generated an initial list of 18 broad co-expression modules (or sets of proteins with correlated expression levels) containing a median of 401 proteins (range 73 – 2066). In contrast, within the the CORUM repository, the characterised modules have a median of three members (proteins) per set (range 1 -143), suggested that to meaningfully apply clustering methodologies it is necessary to identify smaller sets of protein groups.

I therefore considered a two-steps approach entailing a biclustering algorithm to investigate the influence of the three cell lines of origin of the orthografts samples, followed by a constrained clustering algorithm to resize each protein group by leveraging literature information. Biclustering (or two mode clustering)[142] consists in the simultaneous clustering of rows (peptides/proteints) and columns (18 samples analysed) of a matrix, in which data on prostate cancer, hormone naïve versus castration resistance, or cell-line specific data are incorporated. Constrained clustering[143] (a form of semi-supervised learning algorithms) takes advantage of known validated and/or non-viable protein-protein interactions to logically constrain the modules. The first step of the pipeline with biclustering, separated the original 18 co-expression modules from ProCoNa analysis into 189 subgroups, with a median of 9 members/proteins per group (range 2 – 792). At least one subgroup was found within each original module with a median of 9 subgroups identified per module.

For constrained clustering, we applied data from two repositories to define a list of confident interacting proteins preserved within individual clusters, namely the CORUM[141] and the HIPPIE[144] (Human Integrated Protein-Protein Interaction rEference) repositories, incorporating data on validated interacting protein-protein complexes as well as information on unfeasible relationships. Following constrained clustering, a final set of 579 submodules of proteins was identified, with a median of 4 proteins per complex (range 2 – 682). In greater detail, 93 out of the original 189 modules obtained from the biclustering analysis were further split, with a median of 7 subgroups identified per module.

The clustering procedure allowed the generation of submodules of comparable size to that observed in the CORUM set, which allows me to carry out a comparison between clustering informed analysis of the ProCoNa data with the gold standard dataset on protein-protein interactions presented in the CORUM set. The distribution of pairwise non-zero Jaccard indexes among the inferred submodules from the CORUM and ProCoNa based pipeline respectively are presented in Figure 4-4. For the CORUM data, there was a good distribution of the Jaccard indexes which signifies the ability of the dataset on protein-protein interaction to provide useful information for downstream analysis. In contrast, data from the orthografts following ProCoNa based analysis did not demonstrate the range of Jaccard indexes expected from the experimentally validated set of protein complexes, likely because the clustering algorithms used were promoting the formation of subgroups independent from each other (i.e. not sharing any member with between each others). These results suggest that the protein complexes are highly dissimilar from each other, and unlikely to yield meaningful information.

Since the complexes generated by Procona didn't reflect the characteristics of the gold standard, I choose to adopt the CORUM sets as the 'backbone' for the rest of the pipeline, while retaining the adjacency matrix obtained from LFQ L data to 're-weight' protein-protein interactions within each protein complex. The underlying rationale is that the nature of protein-protein interactions can change in prostate carcinogenesis such as the transition from hormone naïve to castration resistant disease[145], thus data generated may provide insight into molecular classification of castration resistant prostate cancer[36].

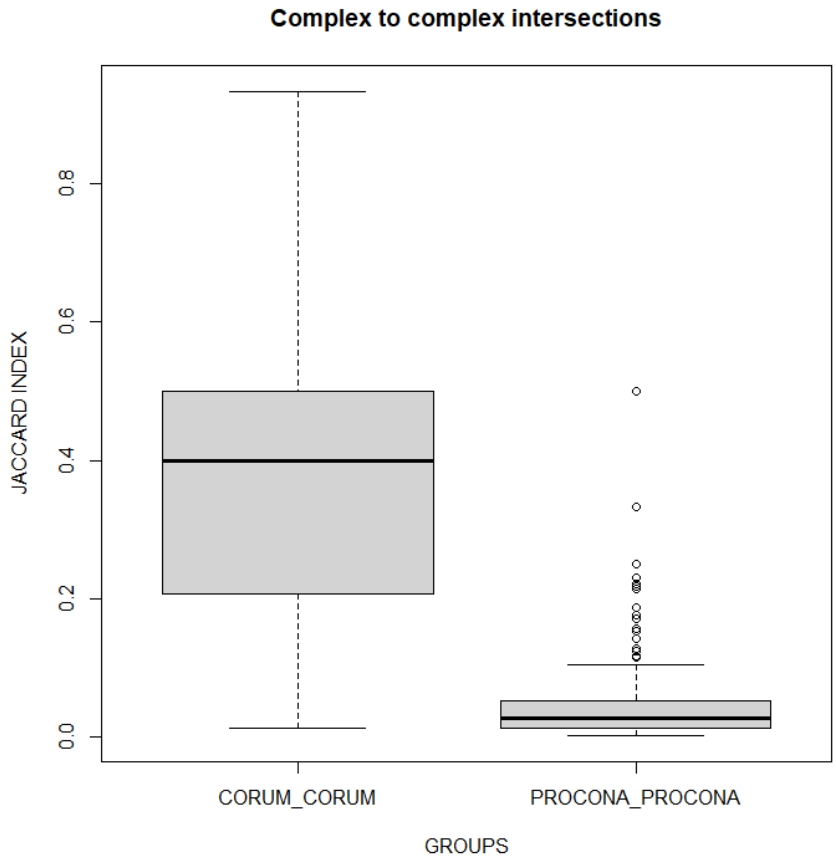


Figure 4-4. Boxplots of pairwise, non-zero Jaccard indexes distributions of CORUM sets (CORUM_CORUM) and modules inferred from orthografts data (PROCONA_PROCONA). The number of non-zero Jaccard indexes was ~7 times higher for the intra-CORUM calculations than ProCoNa based analysis (n= 1761 versus 260 among our modules respectively).

4.2.3 Integrative modules

By integrating data from the gene regulatory network in the previous chapter (Chapter 3) with protein complexes represented in the CORUM data, 516 integrative modules (i.e., regulon linking to protein complex(es), Appendix 6) from the 1308 regulons based on RNAseq data that have identifiable protein linkage as a potential transcriptional target. Within each highlighted module, at least one member of the protein complex is required to be a target gene (or genes) within the regulon. I observed a median of 4 protein complexes (range 1-41) per module, thus highlighting the complementary relationship between co-expressed genes (within regulons) and functional protein groups (within protein-protein interacting complexes).

To assess the usefulness of the assembled modules, I searched for known connections between protein complexes and transcriptional regulators previously implicated in prostate cancer. Germline mutations of *BRCA2* have been associated with poor prognosis for patients with prostate cancer[146,147]. In agreement with a previous observation associating *BRCA2* mutations with the dysregulation of the MED12L/MED12 axis[147], we found the BRCA2-module to contain the 'HOMER3-IP3R-TRPC1_complex', which is constituted by mediator complex proteins and cyclin dependent kinases. Furthermore, *TP53* is a tumour suppressor gene involved in virtually every cancer type. Our analysis (Appendix 6) predicted the association of its regulon with the 'TFIIH transcription factor' complex, which is known to physically interact[148] with TP53, supporting the validity of the inference pipeline I have developed. Lastly, *GATA2*, a key transcription factor involved in prostate cancer adaptation to the castrate environment[149], has been integrated with the URI complex. Consistent with my findings, a study based on the integration of genome-wide Chip-Seq data[150] also support *URI1* as a direct target of GATA2. Collectively, these data indicate the integration method to be a valid approach to leverage functional interactions between co-expressed genes and proteins.

4.2.4 Differentially expressed genes and proteins

To refine the list of highlighted integrative modules, I compared data at mRNA and protein levels from hormone naïve (HN) and castration resistant (CR) prostate orthografts (n=9 for each group), generating a list of differentially expressed genes (DEGs) and differentially expressed proteins (DEPs).

A **B**

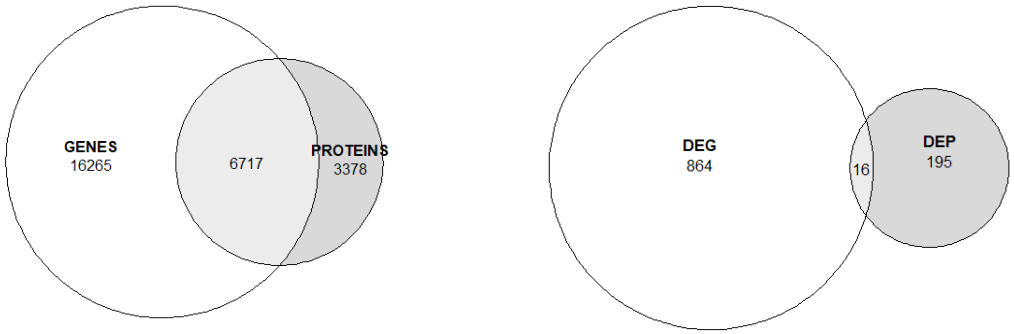


Figure 4-5. A: Euler plot of the sets of measured genes and proteins from the matched transcriptomics/proteomics analysis on 18 orthografts. B: Euler plot of the sets of identified significantly differentially expressed genes (DEG) and proteins (DEP) when hormone naïve and castration resistant orthografts were compared.

Deseq2 pipeline, applied to RNAseq data, provided an higher number of differential features when compared to the ROTS (An R package for reproducibility-optimized statistical testing) algorithm, applied to proteomic quantification based on the LFQ ratio (Figure 4-5), which is in agreement with other studies[151]. Among the 880 mRNA and 211 proteins found to be statistically differentially expressed between hormone naïve and castration resistant orthografts, I observed 16 gene products differentially expressed at both transcript and protein levels (Figure 4-5B, Table 4-1). Interestingly, all the 16 features showed coherent directionality in the fold change. Moreover, among these genes, both known and novel genes in prostate cancer are highlighted. For instance, consistent with my findings, *IGFBP2* and *SIGIRR* were overexpressed in progressive prostate cancer[152,153], while downregulated *SEPP1* expression was associated with shorter patient survival[154].

Enrichment analysis was then performed using 50 genesets based on the 'Hallmarks of cancer' curated genesets[155], a stringent set of experimentally validated and cancer-specific biological mechanisms. In this way, I was able to consider the proteomic and transcriptomic data from a functional perspective and to facilitate hypothesis generating in the context of the underlying the biology in CRPC (Table 4-2 and 4-3). Of note, the androgen receptor pathway was most significantly upregulated based on analysis of transcriptomic analysis. Interestingly, the pathway containing MYC target genes is found most enriched in CRPC following differentially expressed protein-based analysis.

Gene/Protein Hugo symbol	**Transcript**		**Peptides**	
	log2FoldChange	qvalue	log2FoldChange	qvalue
TRAM1	0.496	0.009	0.405	0.000
DSP	1.071	0.038	0.313	0.018
IMPAD1	0.234	0.031	0.176	0.023
IGFBP2	1.528	0.001	0.527	0.000
SDC1	2.042	0.000	0.614	0.018

UAP1	-1.490	0.001	-0.439	0.023
ACSL3	-2.496	0.000	-0.761	0.000
STEAP4	-7.450	0.000	-4.838	0.018
HRSP12	0.899	0.009	0.480	0.000
EPHX1	2.234	0.012	0.508	0.044
ALAD	0.908	0.000	0.256	0.018
ACSL1	1.450	0.038	0.360	0.026
MINA	0.336	0.050	0.272	0.013
SIGIRR	1.115	0.006	0.246	0.018
CD47	1.240	0.002	0.383	0.026
SEPP1	-1.381	0.012	-4.129	0.033

Table 4-1. Common differentially expressed genes and proteins shared between CR vs HN orthografts.

Geneset name	Gene Ratio	pvalue	qvalue
HALLMARK_ANDROGEN_RESPONSE	18/213	9.880E-07	4.370E-05
HALLMARK_UV_RESPONSE_DN	17/213	5.240E-04	1.157E-02
HALLMARK_ESTROGEN_RESPONSE_EARLY	18/213	7.762E-03	1.144E-01
HALLMARK_HYPOXIA	17/213	1.634E-02	1.416E-01
HALLMARK_NOTCH_SIGNALING	5/213	1.781E-02	1.416E-01
HALLMARK_BILE_ACID_METABOLISM	11/213	1.921E-02	1.416E-01
HALLMARK_FATTY_ACID_METABOLISM	13/213	4.218E-02	2.664E-01

Table 4-2. List of significantly over-representated pathways based on differentially expressed genes comparing CR vs HN prostate orthografts. Pathways with p values <0.05 are shown.

Geneset name	GeneRatio	pvalue	Qvalue
HALLMARK_MYC_TARGETS_V1	9/61	0.002	0.031
HALLMARK_OXIDATIVE_PHOSPHORYLATION	9/61	0.002	0.031
HALLMARK_FATTY_ACID_METABOLISM	6/61	0.022	0.280
HALLMARK_MYC_TARGETS_V2	3/61	0.046	0.438

Table 4-3.Over-representation analysis using differentially expressed proteins from the CR vs HN contrast. Pathways with p values <0.05 are shown.

From the DEG, among the seven 'Hallmarks of cancer' pathways with significant p-values, two pathways, (namely Androgen Response and Ultraviolet Response) were found to have significant q values at 0.00004 and 0.01 respectively. On the other hand, pathway analysis of DEP revealed two pathways to be significantly enriched, signifying upregulation of the MYC and oxidative phosphorylation pathways, both with q values at 0.03.

Despite the relatively small number of common features found, these results support the feasibility of a joint enrichment analysis of transcriptomics/ proteomics integrative modules, leveraging at the same time both DEG and DEP.

4.2.5 Modules enrichment

To obtain a broader understanding of the information stored in the transcriptomics/proteomics integrative modules, I ranked the integrative modules based on the percentage of differentially expressed genes (DEG) and proteins (DEP) represented within individual enriched pathways, in relationship to either the regulon's target genes or the proteins belonging to the complexes linked to the regulon, respectively (Tables 4-4 and 4-5). With the large variation of size of modules and percentage of genes/proteins affected, there were, respectively, nine and five modules presenting at least 50 percent of the target genes or proteins differentially expressed. Despite that, no common modules observed between Tables 4-4 and 4-5.

Integrative module	number of genes	number of degs	number of degs/number of genes
ENSG00000118217_ATF6_1:161736083-161933860	2	2	1
ENSG00000005810_MYCBP2_13:77618791-77901185	25	20	0.8
ENSG00000174576_NPAS4_11:66188474-66194178	8	6	0.75
ENSG00000077092_RARB_3:25215822-25639423	39	28	0.718
ENSG00000176165_FOXG1_14:29235049-29238870	17	11	0.647
ENSG00000163848_ZNF148_3:124944404-125094198	8	5	0.625
ENSG00000056972_TRAF3IP2_6:111877656-111927481	12	7	0.583
ENSG00000164061_BSN_3:49591921-49708978	28	15	0.536
ENSG00000110171_TRIM3_11:6469842-6495689	6	3	0.5

Table 4-4. Integrative modules ranking according to the percentage of differentially expressed genes (degs) within individual modules. Modules with 0.5 or more of the network genes showing differential expression are included.

Integrative module	number of proteins	number of deps	number of deps/number of proteins
ENSG00000007866_TEAD3_6:35441373-35464853	1	1	1
ENSG00000125107_CNOT1_16:58553854-58663790	2	1	0.5
ENSG00000048052_HDAC9_7:18126571-19042039	2	1	0.5
ENSG00000162775_RBM15_1:110881127-110889299	2	1	0.5
ENSG00000173258_ZNF483_9:114287438-114340124	2	1	0.5
ENSG00000104907_TRMT1_19:13215715-13228381	75	27	0.36
ENSG00000075975_MKRN2_3:12598512-12625212	75	27	0.36
ENSG00000129535_NRL_14:24549315-24584223	76	27	0.355
ENSG00000072310_SREBF1_17:17713712-17740325	78	27	0.346
ENSG00000159461_AMFR_16:56395363-56459450	79	27	0.342
ENSG00000153879_CEBPG_19:33864235-33873592	80	27	0.338
ENSG00000066135_KDM4A_1:44115828-44171186	3	1	0.333
ENSG00000171148_TADA3_3:9821543-9834695	3	1	0.333
ENSG00000186153_WWOX_16:78133309-79246564	81	27	0.333
ENSG00000162227_TAF6L_11:62538774-62554814	83	27	0.325
ENSG00000130382_MLLT1_19:6212965-6279959	84	27	0.321
ENSG00000166200_COPS2_15:49398267-49447858	85	27	0.318
ENSG00000132773_TOE1_1:45805341-45809647	92	28	0.304
ENSG00000084072_PPIE_1:40157853-40229586	89	27	0.303
ENSG00000168495_POLR3D_8:22102616-22112113	93	28	0.301
ENSG00000163812_ZDHHC3_3:44956748-45017677	90	27	0.3
ENSG00000093010_COMT_22:19929129-19957498	91	27	0.297
ENSG00000112130_RNF8_6:37321747-37362514	98	28	0.286
ENSG00000166197_NOLC1_10:103911932-103923627	97	27	0.278
ENSG00000169131_ZNF354A_5:178138592-178157703	101	28	0.277
ENSG00000089234_BRAP_12:112079949-112123790	102	27	0.265
ENSG00000140694_PARN_16:14529557-14726585	106	27	0.255
ENSG00000104976_SNAPC2_19:7985200-7988135	108	27	0.25
ENSG00000174405_LIG4_13:108859786-108870716	4	1	0.25
ENSG00000204977_TRIM13_13:50570023-50594617	4	1	0.25

Table 4-5. Integrative modules ranking according to the percentage of differentially expressed proteins (deps) within individual modules. Modules with 0.25 or more of the network proteins showing differential expression are included.

To support statistical analysis of the modules, I developed a bespoke analysis by randomly reshuffling the fold changes and the respective p-values of

differentially expressed genes and protein to unbiasedly assess if any of the modules were enriched in CRPC. Using this approach, I wish to study the concordance between transcription factors and mRNA levels of gene targets on one hand, and protein-protein interaction and DEP on the other hand.

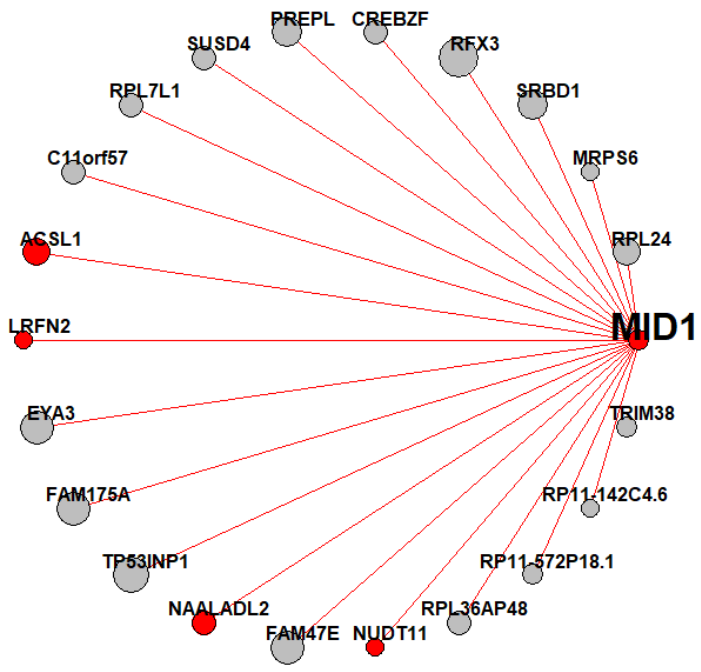
My permutation-based enrichment analysis provided separated results for regulons and protein complexes concordances (Appendix 7). For each integrative graph, the enrichment p-values were calculated as the fraction of times the scores generated by randomly reshuffling the inputs were higher (if positive) or lower (when negative) than the real score.

The analysis highlighted the only integrative module (with regulons and protein complexes bootstrapping p-values of 0) significant at both regulons and protein complexes level, namely the MID1 (Midline 1, Midline 1 RING Finger Protein) regulatory module (Figure 4-6A) to be implicated in CRPC. The *MID1* integrative module is composed of a regulon of 21 putative target genes and 177 proteins from five CORUM complexes, namely Ribosome, _cytoplasmic306, 60S_ribosomal_subunit_cytoplasmic308, 28S_ribosomal_subunit_mitochondrial315, 55S_ribosome_mitochondrial320 and Nop56p-associated_pre-rRNA_complex3055. Within the *MID1* regulon, I found non-random co-upregulated mRNA expression for *ACSL1*, *LRFN2*, *NAALADL2* and *NUDT11* genes as transcriptional targets within the *MID1* regulon) (Figure 4-6A).

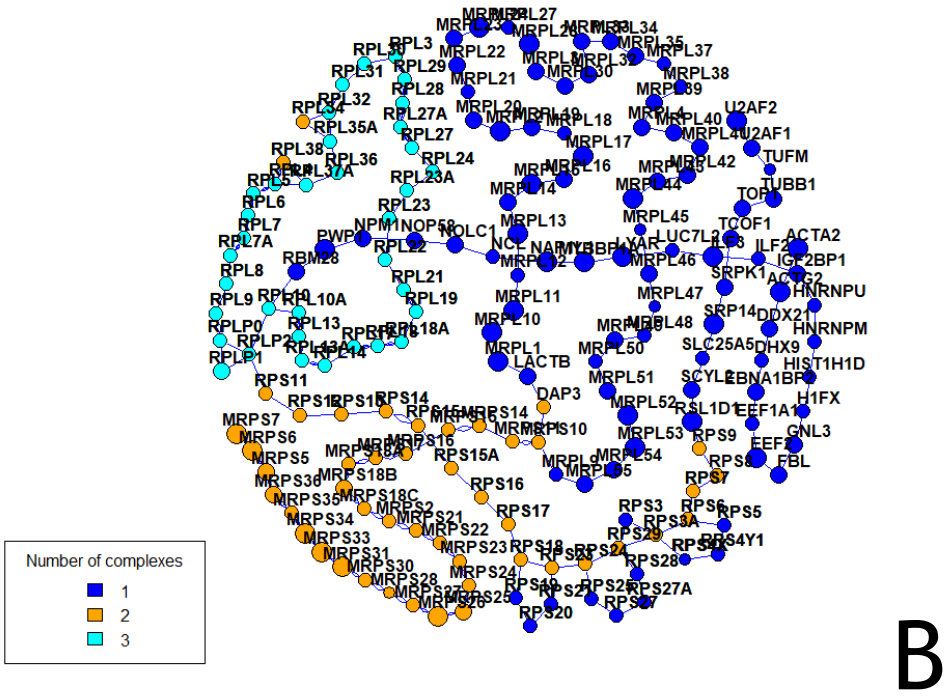
Given the overlapping nature of the CORUM sets, 82 proteins from the integrative module belong to multiple complexes (Figure 4-6B). I then investigated the protein complexes associated with the *MID1* regulon (Figure 4-6B). I observed upregulated expression of both large and small mitochondrial ribosomal subunits complexes (including MRPS14, DAP3, MRPS22, MRPS25, MRPS5, MRPS21, MRPS34, MRPS6, MRPS26, MRPS18A, MRPS33, MRPS17, MRPS7, MRPS2. MRPS18B, MRPL3, MRPL19, MRPL49, MRPL14, MRPL50, MRPL32, MRPL20, MRPL9, MRPL4, MRPL18, MRPL15, MRPL27, HNRNPU). Of note, within the

prioritised regulon, *ACSL1* is the only gene upregulated at both mRNA and protein levels.

MID1 is a regulatory protein within a microtubule-associated complex that binds mRNA to promote translation. It is interesting to note that one of its validated targets is androgen receptor, a well-established driver in prostate carcinogenesis including treatment resistance[156,157]. Similarly, mitochondria activity is known to play a key role in cancer progression[158] and, in particular, mitochondrial ribosomal subunits have been associated with tumor progression in prostate cancer[159] and proposed as therapeutical target in many other tumor types[160]. Hence, it will be interesting to test if there was a mechanistic link between the functional status of the *MID1* regulon and increased tumoral mitochondrial activity in CRPC.



A



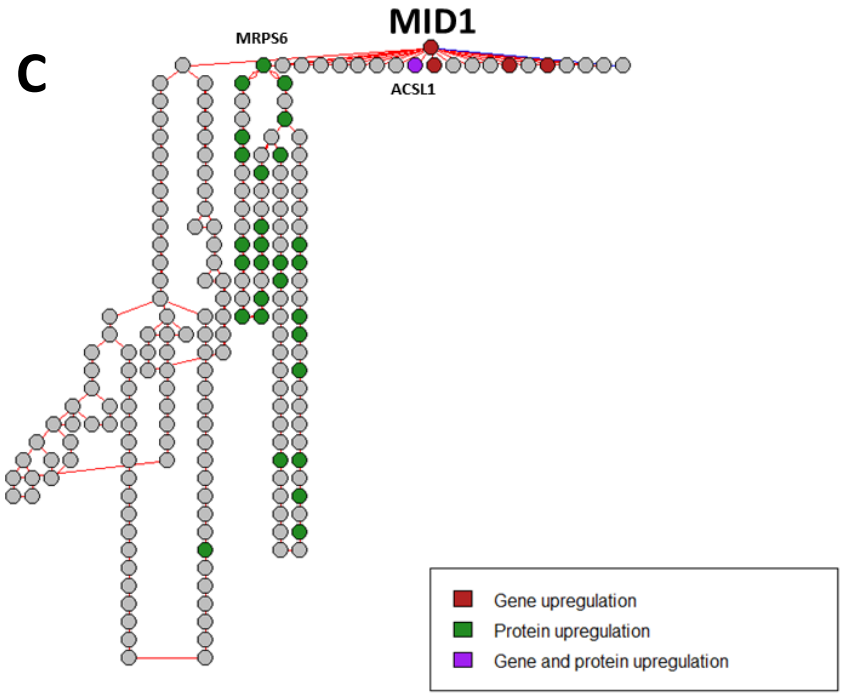


Figure 4-6. MID1 integrative module enrichment.
(A) A MID1 regulon portion of the integrative module. Genes are showed as nodes and type of regulation is represented by edges color (red= upregulation). Hugo symbol of each node is reported. The size of each node is proportional to the strength of the correlations between MID1 and the target. Node colors reflect CRPC vs PC differential expression (grey = not significant, red = significant over-expression).

(B) Associated protein complexes of the MID1 regulon. Nodes represent proteins and edges represents relationships with one or more protein complex. Node colour reflect the number of CORUM complexes each protein belongs to.
(C) View of the full MID1 integrative module. Nodes represent both target genes and proteins belonging to the CORUM complexes. Node color reflects differential expression information (grey= not significant, red = gene upregulation, green = protein upregulation, purple= both gene and protein upregulation). Relevant nodes have been highlighted by showing the correspondent hugo names.

4.3 Discussion

4.3.1 MID1 function in tumorigenic pathways

MID1 (Midline 1) is an E3 ubiquitin ligase belonging to the family of tripartite motif (TRIM) containing proteins. It was primarily discovered as the causative gene for the Optiz BBB/G syndrome through its function as a scaffold for the assembly of a large microtubule-associated ribonucleoprotein complex aimed to the regulation of the protein phosphatase 2A (PP2A)[161]. This interaction leads to the upregulation of the mTORC1 signaling, which has a marked effect on cell proliferation and tumorigenesis[162]. Translational Regulatory functions of MID1 were subsequently reported, involving its ability to bind mRNAs at a purine-rich sequence motif called MIDAS (MID1 association sequence), leading to an increase of production of the proteins encoded by the target RNAs up to 20-fold[157]. Importantly, in cancer cells, MID1 controls both the subcellular localisation and transcriptional activity of *GLI3* (GLI Family Zinc Finger 3), a key transcription factor of the tumorigenic sonic hedgehog signaling pathway. In addition, highly relevant in prostate cancer, MID1 also binds to transcripts of androgen receptor to induce translation of AR protein[163]. Interestingly, there appears to be a functional feedback-loop between AR and MID1 function. Withdrawal of androgens resulted in upregulated MID1 expression which in turn increased the expression of AR protein[164]. In addition, treatment with metformin, an approved medicine for type II diabetes, on both AR positive (LNCaP, VCaP, DuCaP, LNCaP-abl) and AR negative (PC-3 and Du-145) human prostate cancer cells, resulted in suppressed cell growth via the disruption of the association between AR mRNA and MID1 complex[156].

This data supports our findings and suggest the investigation of the physical binding of MID1 and the mRNAs of the identified DEGs, as well as the putative

protein-protein interactions with the enriched CORUM complexes. Functional validations experiments are recommended, to evaluate the effect of overexpressing versus knocking-down MID1, on the growth in androgen dependent and independent conditions.

4.3.2 Differentially expressed genes implication in CRPC

The differentially expressed genes matching the predicted MID1 transcriptional regulations are known players in PCa.

The long-chain fatty acyl-CoA synthetases 1 (ACSL1) is part of the MID1 module (Figure 4-6A,C). ACSL1 is a key element in lipid metabolism and has been implicated to promote prostate cancer progression through increased fatty acid beta-oxidation, mitochondrial respiration, and ATP production. The NUDT11 phospho-hydrolases, also part of the MID1 module (Figure 4-6A), may have a signaling role in prostate cancer, with two single nucleotide variants identified in a genome wide expression quantitative trait loci analysis[165]. LRFN2 has been implemented in a prognostic gene expression signature for recurrence and metastatic-lethal progression of prostate cancer, with upregulated expression in high grade cancer[166]. Lastly, N-acetyl-L-aspartyl-L-glutamate peptidase-like 2 (NAALADL2) overexpression was associated with poor patient survival outcome following radical prostatectomy, with putative effects in promoting cancer motility and metastasis[167]. Moreover, a recent integrated transcriptomic, proteomic and metabolomics analysis of CRPC from our group highlighted potential impact of NAALADL2 expression in upregulating sphingolipid metabolism and nucleotide synthesis required for CRPC tumors growth[61].

4.3.3 Enriched complexes involvement in CRPC

The identified differentially expressed proteins highlighting the enrichment of the MID1 module is associated with co-upregulation of proteins mainly belonging

to the CORUM complexes ‘28S_ribosomal_subunit,_mitochondrial’ and ‘55S_ribosome,_mitochondrial’.

Many of the associated protein complexes of the MID1 regulon are related to translation (Figure 4-6B). This can be functionally linked to the activity of ACSL1 (also found overexpressed at the gene level), and is in agreement with previously published results from a proteomic study, similarly relying on the CORUM sets[141]. Consistent with my findings here, mitochondrial ribosomes were found to be the most over-expressed complexes in high grade clinical prostate cancer tumours, along with sets of proteins involved In ribosome biogenesis, RNA splicing and cytoplasmic ribosomes[57]. Additional investigations supported the key role of mitochondria in prostate cancer progression with upregulated MTCO2 expression, a marker for mitochondrial content[158] and the frequent mutations in the mitochondrial genome[26]. Noteworthy, the protein levels of MRPS18-B correlates with disease progression due to the promotion of epithelial to mesenchymal cell transition[159]. Moreover, it is known to be part of a cytoplasmic complex together with the Ring finger protein 2 (RNF2) (another E3 ubiquitine ligase) that maintains cell stemness[168]. It is therefore important for future research to characterise cytoplasmatic interactions between MRPS18-B and other mitochondrial ribosome subunits within the MID1 module.

Lastly, among the subset of differentially expressed proteins relevant to MID1 integrative module enrichment, we found Heterogeneous Nuclear Ribonucleoprotein U (HNRNPU) (Figure 4-6B). We believe this detection provides additional evidence to a previous identification of HNRNPU mRNA upregulation in CRPC tissues, correlated to an increase of AR-v7 expression[169]. Moreover, given its involvement in the formation of ribonucleoprotein complexes, HNRNPU may play a crucial role in the formation of the cytoplasmic MID1 complex.

My approach introduced a novel hypothesis about the involvement of MID1 in the lipid metabolism alterations observed in PCa, as well as the association with the upregulation of mitochondrial ribosomes activity.

Chapter 5 – Analysis of implicated regulon following treatment in a preclinical *in vivo* model

After studying the usefulness of regulons for predictive (Chapter 3) and diagnostic (Chapter 4) biomarker discovery in PCa, I evaluated the ability of the regulons to explain the molecular effects of different drug treatments.

5.1 Introduction

As highlighted previously in this thesis, the androgen receptor (AR) remains an important therapeutic target in locally advanced and metastatic prostate cancer[170]. Novel androgen receptor pathway inhibitors include highly potent and specific antagonists of the AR such as enzalutamide[171] and apalutamide[172] (ARN-509) as well as abiraterone which functions as an inhibitor of adrenal and testicular bio-synthesis of androgens.

Recent research in the Mills' laboratory, a collaborating partner within the TransPot consortium, showed that *de novo* purine biosynthesis and the conversion of inosine monophasphasate to xanthosine monophosphate (a key intermediate in purine metabolism) is tightly regulated by the *MYC* oncogene. Their findings highlighted that *IMPDH2* (Inosine Monophosphate Dehydrogenase 2) gene product functions as an enzyme to support nucleotide synthesis required for carcinogenesis. IMPDH2 expression was found to be upregulated following androgen deprivation therapy *in vitro* as well as upregulated in clinical prostate tumours. Of interest, combining anti-androgen treatment with mycophenolic acid (MPA, a clinically approved IMPDH inhibitor) resulted in enhanced treatment efficacy[173]. Consistent with their findings, suppression of guanine nucleotide synthesis was found to sensitise prostate cancer cells to AR inhibitors such as Abiraterone and Enzalutamide[174].

To further explore the functional effects of combining suppression of AR function and *de novo* nucleotide synthesis in prostate carcinogenesis, members of the Mills' laboratory applied the androgen receptor positive human LNCaP C4-2b prostate cancer cells in an *in vivo* subcutaneous xenograft experiment and tested for potential synergistic interaction between suppression of androgen receptor function and the use of mycophenolic acid. Serial tumour sizes were obtained to assess the effects of treatments. Transcriptomic analysis was performed to support comparative analysis in different treatment groups. After receiving the raw RNAseq data,in this chapter, I seek to leverage regulon-based analysis to identify candidate master regulators associated with different treatment combinations.

5.2 Results

5.2.1 Assessment of tumour size following treatment

In total 23 xenograft samples were randomly obtained from the *in vivo* efficacy experiment and included for transcriptomic analysis (Table 5-1): n=4 for mycophenolate mofetil (clinical formulation of mycophenolic acid, MPA, Arm 1) treatment alone, n=4 for abiraterone treatment alone (Arm 2), n=3 for apalutamide (ARN-509) alone (Arm 3), n=3 for combined mycophenolate mofetil and apalutamide treatment (Arm 4), n=4 for combined mycophenolate mofetil and abiraterone treatment (Arm 5), and control samples including no treatment (n=2) or vehicle control treatment (benzyl alcohol and safflower oil, n=3).

All treatment arms (Arms 1-5) achieved a significant reduction in tumour growth rate when compared to the controls, with no tumours following treatment achieving terminal volume within the 3-week time-course. Of note, terminal volume was achieved at around 16 days post-treatment in the control arm (Figure 5-1, orange line). Whilst there were no statistically significant differences in the growth kinetics among various treatment arms, there was a qualitatively lower growth rate in the MPA/abiraterone combination (Arm 5, Figure 5-1, light blue line) relative to abiraterone alone (Figure 5-1, light blue

line (Arm 5) versus red line (Arm 2)), with mean tumour weights at 0.5 g for MPA/Abiraterone (Arm 5) compared to 1 gm for abiraterone-only (Arm 2) and 0.72 g for MPA alone (Arm 1).

GROUP	TREATMENT
ARM1_A	Mycophenolate_mofetil
ARM1_B	Mycophenolate_mofetil
ARM1_C	Mycophenolate_mofetil
ARM1_D	Mycophenolate_mofetil
ARM2_A	Abiraterone
ARM2_B	Abiraterone
ARM2_C	Abiraterone
ARM2_D	Abiraterone
ARM3_A	ARN-509
ARM3_B	ARN-509
ARM3_C	ARN-509
ARM4_A	Mycophenolate_mofetil_and_ARN_509
ARM4_C	Mycophenolate_mofetil_and_ARN_509
ARM4_D	Mycophenolate_mofetil_and_ARN_509
ARM5_A	Mycophenolate_mofetil_and_Abiraterone
ARM5_B	Mycophenolate_mofetil_and_Abiraterone
ARM5_C	Mycophenolate_mofetil_and_Abiraterone
ARM5_D	Mycophenolate_mofetil_and_Abiraterone
Untr_A	Untreated
Untr_B	Untreated
Vehicle_A	benzyl_alcohol_and_safflower_oil
Vehicle_B	benzyl_alcohol_and_safflower_oil
Vehicle_C	benzyl_alcohol_and_safflower_oil

Table 5-1. Samples sheet

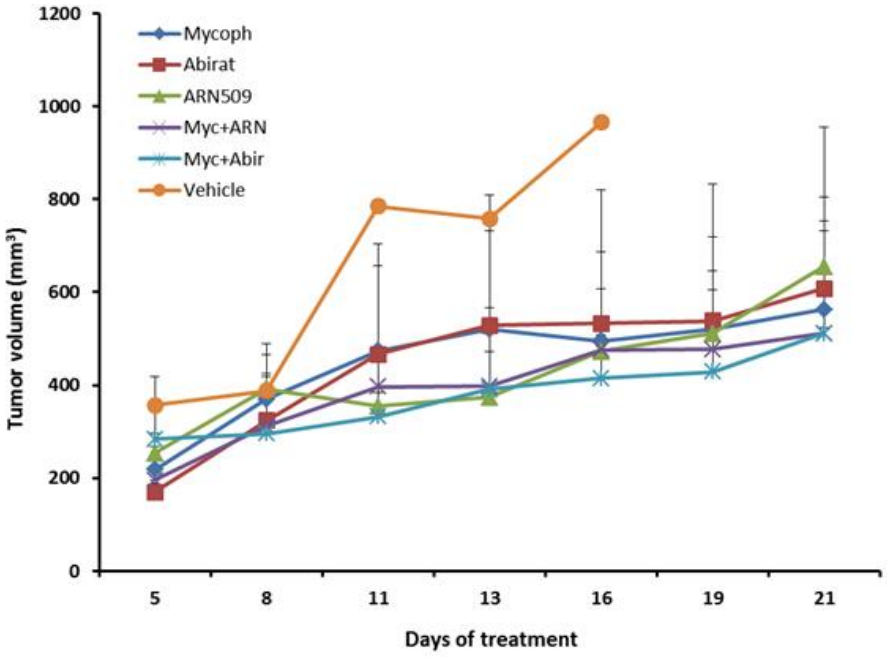


Figure 5-1. Graph showing tumour volumes following different treatments over a 21 day study period. (Myc, Mycoph = Mycophenolate_mofetil, Abir = Abiraterone, ARN = ARN-509) (data prepared and provided by Ian Mills).

5.2.2 Reads disambiguation

Reads disambiguation analysis was performed to evaluate biases in the number of human reads from the xenografts as a quality control measure to 'qualify' data for further downstream analysis. The range of human reads obtained from the sequencing experiment was 6.7 million (M) to 16.1 M (Figure 5-2), with varying reads for human (and murine) gene sequences observed in samples within individual treatment groups (Arms 1-5 and control samples), probably reflecting different composition between tumour (from LNCaP cells) and stromal (from murine host) compartments. Moreover, the generally low amount of sequencing reads can reflect a lower Initial amount of RNA or quality.

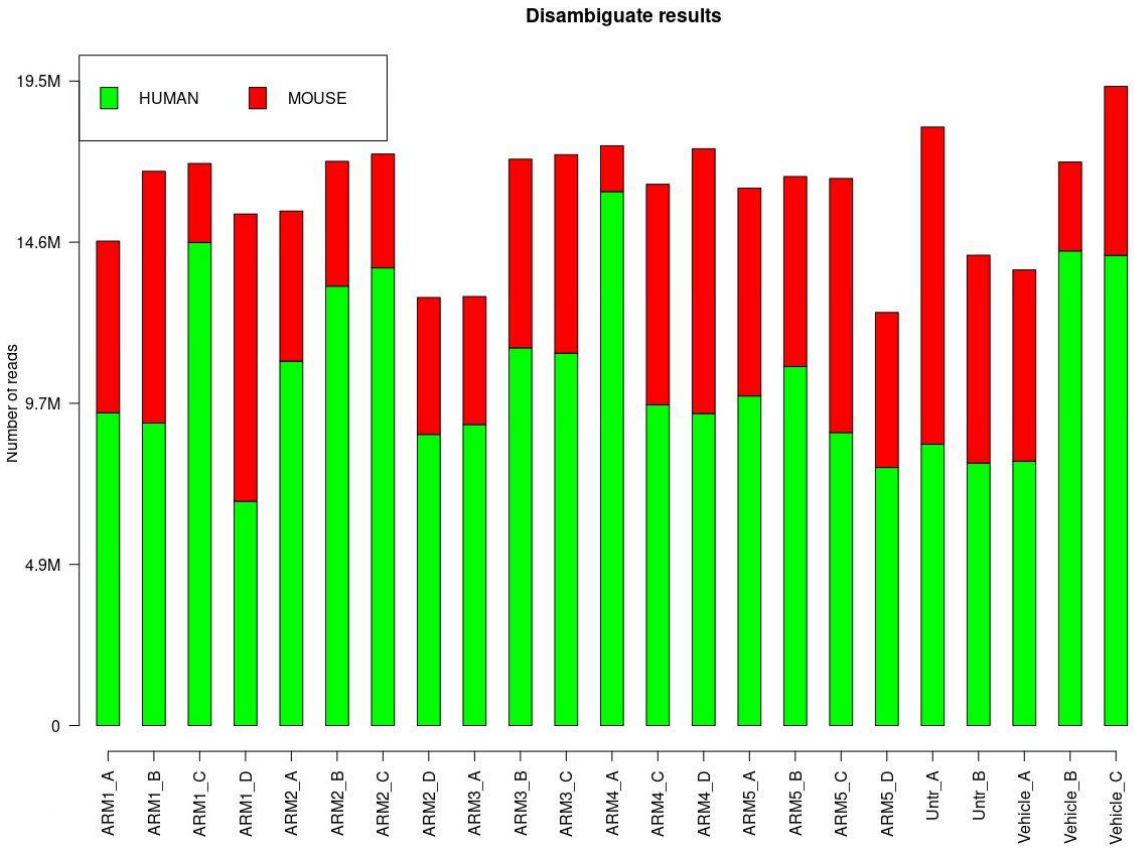


Figure 5-2. Barplot of sequenced human (green) and mouse (red) unambiguous reads for each sample. Treatment groups are identified by name of Arms (1-5), and within each treatment arm samples are labelled as alphabets.

5.2.3 Visual examination of expression profiles

To evaluate the level of heterogeneity of the transcriptome among the samples at the molecular level, Pearson's linear correlations were calculated among fragments per kilobase per millions (FPKM), gene-level normalised counts profiles (Figure 5-3). Regardless of the efficacy of treatment under investigation, samples from the same treatment grouping should be consistently similar to support robust analysis between different treatment regimes. However, samples within the same treatment groups did not cluster together. The median Pearson's correlation among all the samples was disappointingly found to be 0.8, suggesting no evidence of 'clusters' at the molecular level and that all samples were transcriptomically similar. Given the substantial overlap of the samples profiles, irrespective of the experimental groups, the overall quality of the data received was compromised. Hence, we could not comment on the pathology.

These results could be in part explained by a large variation within the xenograft tumour micro-environment resulting in overlapping transcriptomes.

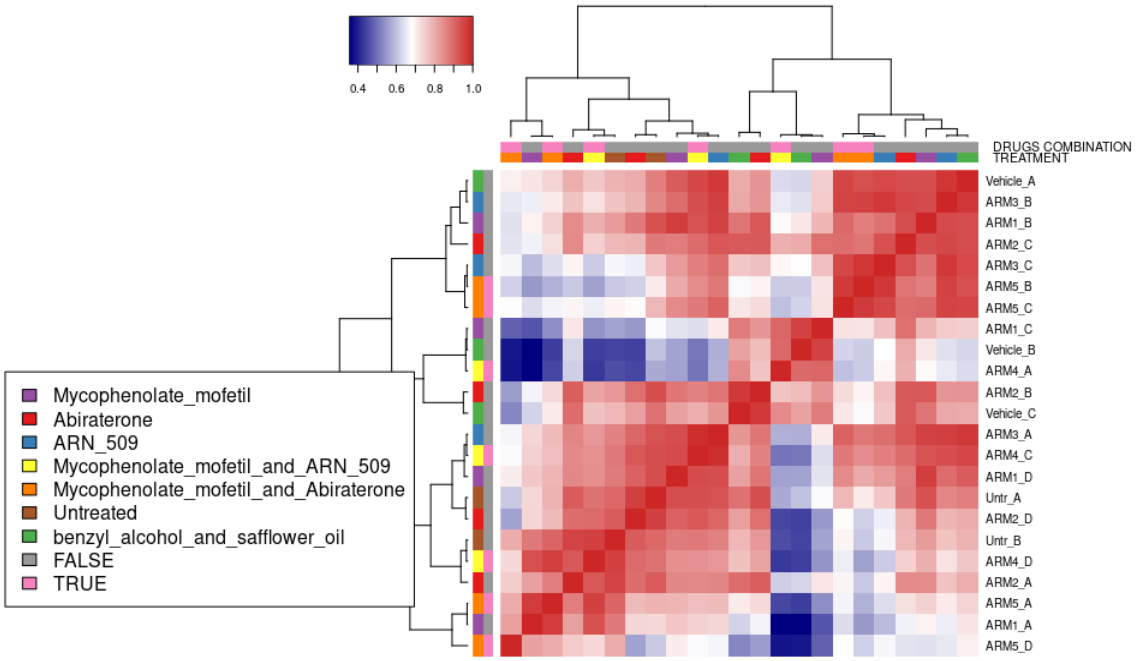


Figure 5-3. Symmetric heatmap of analysis of sample-to-sample gene expression profiles by Pearson's linear correlation. Colour red signifies higher correlations while colour blue low correlation values. The annotation contains labeling for both treatment group and single/combinatorial treatment type.

The heatmap generated from Pearson analysis suggested substantial overlap of samples irrespective of the treatment grouping. I reasoned that the two anti-androgen treated groups are biologically similar and can therefore be combined to increase the size of the treatment group to improve the statistical power of the analysis. Applying Principal Component Analysis of the combined groups (Figure 5-4), once again, I did not observe any evidence of clustering based on the treatment types (control, single agent, or combined treatment).

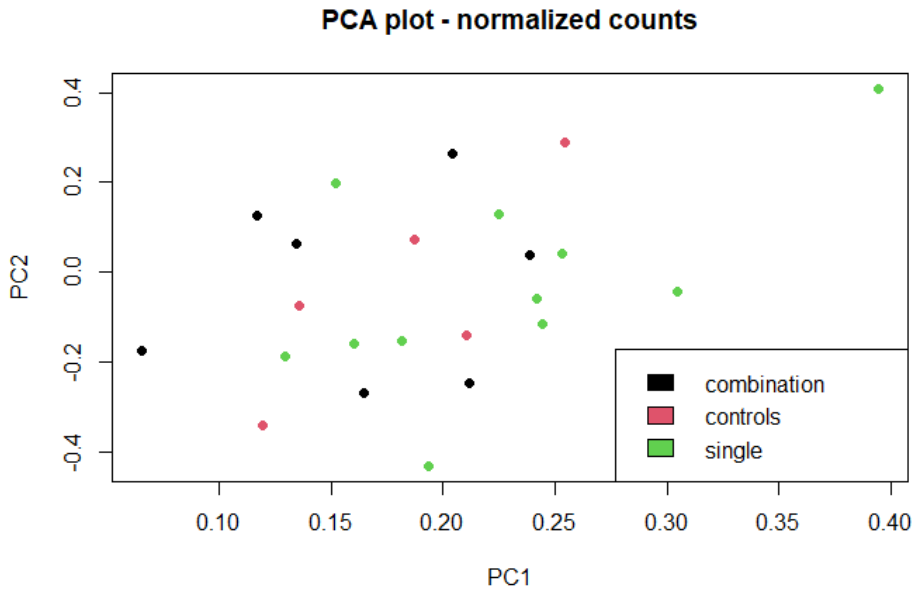


Figure 5-4. Principal component analysis (PCA) for different treatment groups categorised as controls (untreated/vehicle treated), single treatment (MPA, Abiraterone, ARN-509) alone, or combined treatment (MPA + either Abiraterone or ARN-509).

5.2.4 Differentially expressed genes

Accepting the low correlation among samples within individual treatment groups, I assessed the number of differentially expressed genes following different treatments when compared to control treatment (untreated or vehicle treatment, Table 5-1). Abiraterone is increasingly prescribed as the choice of second line androgen deprivation therapy following clinical evidence of CRPC. In addition, recent clinical trials have confirmed its dramatic efficacy when combined with standard of care hormonal therapy to achieve combined androgen blockage, typically with combining LHRH antagonist and abiraterone as first line treatment for metastatic prostate cancer. I was therefore interested to investigate the effects of adding mycophenolate mofetil (clinical formulation for mycophenolic acid, MPA) to abiraterone.

A three-way comparison of differentially expressed genes following treatment with abiraterone or mycophenolic acid alone as well as combined treatment was carried out and presented as a Venn diagram in Figure 5-5. As a results of the variation between samples within individual treatment groups, the number

of differentially expressed genes was relatively low. Of note, treatment with MPA alone (Arm 1) showed the least number of differentially expressed genes when compared to an androgen receptor pathway inhibitor such as abiraterone alone (Arm 2). This may reflect the mode of action of the treatment, with androgen receptor being a transcription factor and thus abiraterone treatment affecting more genes that MPA.

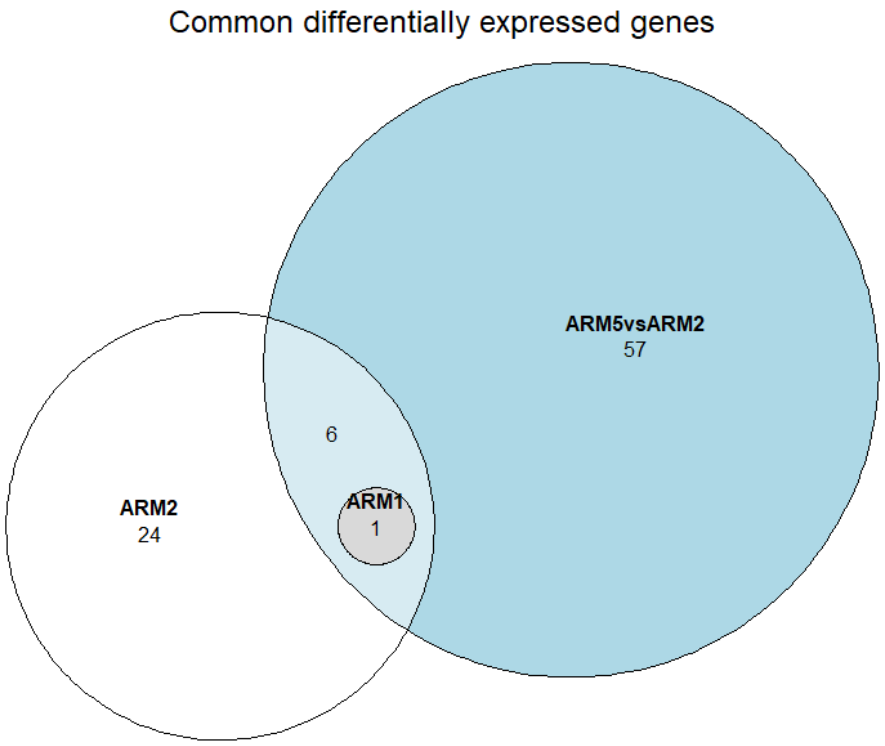


Figure 5-5. Venn diagram illustrating number of differentially expressed genes among treatment groups with MPA (Arm 1), abiraterone (Arm 2) and MPA + abiraterone (Arm 5) when compared to control samples (untreated or vehicle treated). For this figure, only significant (adjusted p-value < 0.05) differentially expressed genes were used.

I observe that while MPA alone (Arm 1) did not have a strong impact on gene expression, the addition of MPA to abiraterone treatment (Arm 5) altered the profile of significantly differentially expressed genes when compared to abiraterone treatment alone (Arm 2). I therefore probed the molecular consequence of combined MPA/abiraterone treatment (Arm 5) when compared to abiraterone alone (Arm 2). Relative to Arm 2, 64 genes were significantly (adjusted p-value < 0.05) differentially expressed in Arm 5 (Table 5-2).

Ensembl gene ID	Log2 fold change	Adjusted p-value	Hugo symbol
ENSG00000003056	0.892	0.023	M6PR
ENSG00000014257	2.845	0.015	ACPP
ENSG00000050426	1.020	0.009	LETMD1
ENSG00000060971	-1.057	0.018	ACAA1
ENSG00000063854	-0.882	0.005	HAGH
ENSG00000076351	1.503	0.019	SLC46A1
ENSG00000103064	1.414	0.005	SLC7A6
ENSG00000104267	7.170	0.008	CA2
ENSG00000108924	2.072	0.009	HLF
ENSG00000120875	3.943	0.005	DUSP4
ENSG00000121039	2.204	0.030	RDH10
ENSG00000122176	7.794	0.008	FMOD
ENSG00000123643	1.457	0.047	SLC36A1
ENSG00000124788	3.519	0.047	ATXN1
ENSG00000130589	-2.063	0.011	HELZ2
ENSG00000131711	-2.406	0.004	MAP1B
ENSG00000132003	-1.480	0.039	ZSWIM4
ENSG00000136848	2.661	0.035	DAB2IP
ENSG00000139514	1.229	0.004	SLC7A1
ENSG00000141959	-1.360	0.011	PFKL
ENSG00000142515	2.681	0.005	KLK3
ENSG00000144339	5.275	0.043	TMEFF2
ENSG00000151640	-2.670	0.045	DPYSL4
ENSG00000154124	1.093	0.031	FAM105B
ENSG00000155850	2.452	0.047	SLC26A2
ENSG00000156269	6.661	0.005	NAA11
ENSG00000162032	-1.646	0.043	SPSB3
ENSG00000162545	4.742	0.043	CAMK2N1
ENSG00000164181	1.956	0.037	ELOVL7
ENSG00000164300	1.441	0.023	SERINC5
ENSG00000165731	5.060	0.018	RET
ENSG00000166831	3.178	0.049	RBPMS2
ENSG00000167393	-2.541	0.007	PPP2R3B
ENSG00000167657	-1.536	0.039	DAPK3
ENSG00000167751	1.499	0.048	KLK2
ENSG00000169016	1.317	0.047	E2F6
ENSG00000169567	2.865	0.033	HINT1
ENSG00000169884	2.608	0.015	WNT10B
ENSG00000171448	1.361	0.015	ZBTB26
ENSG00000171885	5.490	0.005	AQP4
ENSG00000172987	8.021	0.004	HPSE2
ENSG00000173214	1.399	0.032	KIAA1919
ENSG00000174684	1.381	0.031	B3GNT1
ENSG00000178385	2.170	0.039	PLEKHM3
ENSG00000181800	6.059	0.005	CELF2-AS1

ENSG00000185236	-0.656	0.032	RAB11B
ENSG00000185567	-3.920	0.005	AHNAK2
ENSG00000187210	2.007	0.031	GCNT1
ENSG00000187792	1.572	0.033	ZNF70
ENSG00000188257	4.590	0.030	PLA2G2A
ENSG00000197635	1.281	0.039	DPP4
ENSG00000197961	1.000	0.047	Than
ENSG00000204314	2.297	0.049	PRRT1
ENSG00000207389	-28.527	0.000	RNU1-4
ENSG00000214717	-1.084	0.005	ZBED1
ENSG00000215568	8.394	0.004	GAB4
ENSG00000242284	5.751	0.005	CT45A5
ENSG00000244567	3.315	0.015	AC096772.6
ENSG00000248118	5.009	0.015	LINC01019
ENSG00000255310	2.640	0.032	AF131215.2
ENSG00000260778	-3.283	0.008	MIR940
ENSG00000262885	5.870	0.032	CTD-2144E22.11
ENSG00000265369	5.741	0.035	U3
ENSG00000269640	4.776	0.048	CTD-2521M24.9

Table 5-2. Significant differentially expressed genes from the ARM5vsARM2 comparison

5.2.5 Regulons enrichment

Next, I interrogated the list of differentially expressed genes for transcription factors as potential regulons as inferred from a previous analysis (see Chapter 3) in the Arm 5 vs Arm 2 comparison.

The Gene Graph Enrichment Analysis (GGEA) was inputted with the full list of genes from the differential analysis between Arm 5 vs Arm 2, revealing the *SET* proto-oncogene as the only significantly enriched regulon (adjusted p-value 0.000999, Table 5-3).

Ensembl gene id	Hugo symbol	Genomic coordinates	Regulation
ENSG00000002330	BAD	11:64037301-64052176	Positive
ENSG00000062370	ZNF112	19:44830707-44871377	Positive
ENSG00000063046	EIF4B	12:53399941-53435993	Positive
ENSG00000070061	IKBKAP	9:111629796-111696396	Positive
ENSG00000100129	EIF3L	22:38244874-38285414	Positive
ENSG00000105373	GLTSCR2	19:48248778-48260315	Positive

ENSG00000108848	LUC7L3	17:48796904-48833574	Positive
ENSG00000110066	SUV420H1	11:67922329-67981295	Positive
ENSG00000110583	NAA40	11:63706430-63724800	Positive
ENSG00000116251	RPL22	1:6241328-6269449	Positive
ENSG00000119396	RAB14	9:123940414-123985292	Positive
ENSG00000131115	ZNF227	19:44711699-44741421	Positive
ENSG00000138439	FAM117B	2:203499900-203634480	Positive
ENSG00000142534	RPS11	19:49999621-50002946	Positive
ENSG00000142676	RPL11	1:24018268-24022915	Positive
ENSG00000147403	RPL10	X:153618314-153637504	Positive
ENSG00000148154	UGCG	9:114659045-114697649	Positive
ENSG00000159128	IFNGR2	21:34775201-34851655	Positive
ENSG00000159131	GART	21:34876237-34915797	Positive
ENSG00000160208	RRP1B	21:45079428-45115958	Positive
ENSG00000167770	OTUB1	11:63753324-63769283	Positive
ENSG00000172113	NME6	3:48334753-48343175	Positive
ENSG00000175061	FAM211A-AS1	17:16342135-16381992	Positive
ENSG00000175893	ZDHHC21	9:14611068-14693469	Positive
ENSG00000177410	ZFAS1	20:47894714-47905797	Positive
ENSG00000177733	HNRNPA0	5:137087074-137090039	Positive
ENSG00000178464	CTD-2192J16.15	19:12754088-12754733	Positive
ENSG00000179698	KIAA1875	8:145162628-145173218	Positive
ENSG00000185658	BRWD1	21:40556101-40693485	Positive
ENSG00000197258	EIF4BP6	7:104308195-104310023	Positive
ENSG00000197756	RPL37A	2:217362911-217443903	Positive
ENSG00000199753	SNORD104	17:62223442-62223512	Positive
ENSG00000200237	SNORA70	19:9930629-9930770	Positive
ENSG00000204178	TMEM57	1:25757387-25826700	Positive
ENSG00000204713	TRIM27	6:28870778-28891766	Positive
ENSG00000205581	HMGN1	21:40714240-40721573	Positive
ENSG00000207165	SNORA70	X:153628621-153628756	Positive
ENSG00000207166	SNORA68	19:17973396-17973529	Positive
ENSG00000213280	RP11-212P7.1	7:128210294-128210742	Positive
ENSG00000224078	SNHG14	15:25295778-25492435	Positive
ENSG00000224546	EIF4BP3	9:98908288-98910138	Positive
ENSG00000225031	EIF4BP7	X:110862904-110864717	Positive
ENSG00000227034	RP11-234N17.1	1:99473773-99474030	Positive
ENSG00000228223	HCG11	6:26522075-26526807	Positive
ENSG00000233913	CTC-575D19.1	5:168043316-168044059	Positive
ENSG00000235720	RP11-340I6.6	7:63353663-63355019	Positive
ENSG00000241399	CD302	2:160625363-160654753	Positive
ENSG00000260521	CTD-2576F9.1	15:95398702-95400143	Positive
ENSG00000272888	AC013394.2	15:93425936-93441975	Positive

Table 5-3. SET regulon composition.

Interestingly, SET, also known as inhibitor 2 of protein phosphatase 2A (I2PP2A), is implicated to upregulate c-Myc and downregulate histone acetylation in prostate cancer cells[175]. Moreover, several genes previously linked to PCa were upregulated within the SET regulon, including BAD, a pro-apoptotic protein, with BAD upregulated expression promoting prostate cancer proliferation[176]; EIF4B, a translational initiator protein implicated in prostate carcinogenesis[177]; OTUB1 mediates prostate cancer invasion *in vivo* through RhoA[178]; and the LUC7L3 splicing factor is associated with both relapse free and overall survival in TCGA data[179].

To better understand the function of the enriched regulon, I performed an over-representation analysis of the Gene Ontologies using predicted SET targets. The analysis disclosed significantly enriched ontologies (q value < 0.05) (Table 5-4 and Figure 5-6) including ribonuclear protein complex, ribosome biogenesis and translational initiation.

Gene ontology id	p-value	q-value
go_ribonucleoprotein_complex_biogenesis	1.6e-09	5.8e-07
go_translational_initiation	4.2e-09	7.5e-07
go_multi_organism_metabolic_process	1.1e-07	1.3e-05
go_establishment_of_protein_localization_to_endoplasmic_reticulum	8.6e-07	6.9e-05
go_ribonucleoprotein_complex_subunit_organization	9.5e-07	6.9e-05
go_nuclear_transcribed_mrna_catabolic_process_nonsense_mediated_decay	1.6e-06	9.2e-05
go_protein_localization_to_endoplasmic_reticulum	2.0e-06	9.2e-05
go_rna_catabolic_process	2.1e-06	9.2e-05
go_rrna_metabolic_process	4.0e-06	1.6e-04
go_establishment_of_protein_localization_to_membrane	4.9e-06	1.8e-04
go_protein_targeting_to_membrane	6.6e-06	2.2e-04
go_viral_life_cycle	8.4e-06	2.5e-04
go_ribosome_biogenesis	1.2e-05	3.3e-04

go_establishment_of_protein_localization_to_organelle	2.9e-05	7.5e-04
go_protein_localization_to_membrane	3.7e-05	8.8e-04
go_ncrna_processing	4.2e-05	9.5e-04
go_organic_cyclic_compound_catabolic_process	7.4e-05	1.6e-03
go_protein_targeting	5.9e-04	1.2e-02
go_formation_of_translation_preinitiation_complex	6.5e-04	1.2e-02
go_ribosomal_large_subunit_assembly	7.2e-04	1.3e-02
go_cytoplasmic_translation	2.3E-03	3.9E-02

Table 5-4. Significantly enriched Gene Ontologies utilising SET predicted targets.

The identification of the SET regulon from comparing the transcriptome of tumours following MPA + abiraterone versus abiraterone alone treatment raises the possibility of SET as a master regulator following combined MPA/Abiraterone treatment, supporting the consideration of new hypothesis to link genes and potential biological processes in driving prostate cancer progression.

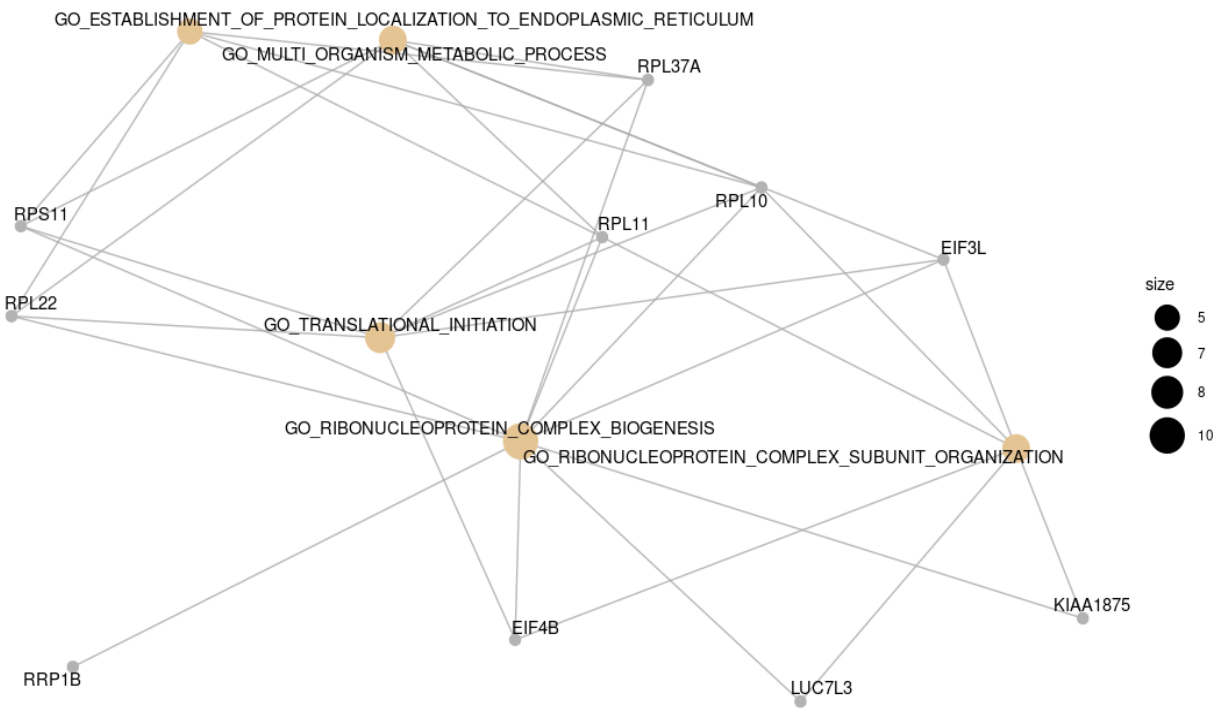


Figure 5-6. Cnet plot of the results of the overrepresentation analysis, mapping SET regulons genes to gene ontologies biological processes. Yellow nodes represent gene ontologies while

grey nodes single genes. The size of the yellow dots reflects the number of genes within each gene ontology.

5.3 Discussion

In this chapter, I explored the value of regulon analysis in a preclinical treatment model and tested the usefulness of the regulon analysis as described in previous chapters. There was substantial overlap for the transcriptome from tumours following different treatment arms. These variations may arise from different sources including varying compositions of malignant epithelium and host stromal tissue, variation between individual mice, bias introduced due to random selection of representative tumours from each of the treatment arms for next generation sequencing, and varying bioavailability of treatment agents due to differences in how the animals metabolised the drugs of interest. Despite these limitations, analysis of differentially expressed genes revealed interesting insights into the effects of adding MPA to an androgen receptor pathway inhibitor. Data from differentially expressed genes and pathway analysis presented in this chapter provide new and complementary information to be correlated to active biological processes.

Firstly, our differential gene expression analysis confirmed the previous observation of a reactivation of the androgen receptor by MPA, highlighted by the overexpression of AR-target genes[180] like KLK2 and KLK3. More importantly, the regulons-based analysis pointed to a novel transcriptional regulation induced by MPA, unrelated to the androgen receptor signalling. Of note, The AR regulon was not highlighted by the regulons enrichment. This is likely due to the fact that the hormone naive and hormone resistant ortografts used for the inference of the gene regulatory network may have shown different TF-target relationships, which can hamper the identification of consensus putative target genes.

Further, a previous work from Ian Mill's group revealed that MPA treatment could upregulate TP53 expression and promote the inhibition of c-Myc expression

and activity by perturbing nucleolar biogenesis. The root cause of these effects was attributed to the inhibition of *de novo* guanine nucleotide biosynthesis[174]. Guanine nucleotides are required to sustain the assembly of nucleoli, an organelle that supports RNA processing (ribosomal RNAs in particular) as well as the interaction between TP53 and MDM2 that results in concomitant ubiquitination and degradation of *TP53* as an important tumour suppressor[181,182].

Based on the finding that the ribonucleoprotein complex and ribosome biogenesis ontologies are significantly overrepresented among the SET positive target genes, I hypothesise that inhibition of the *SET* proto-oncogene may be mechanistically involved in producing the anti-proliferative effect of MPA on the PCa xenografts. For future research, it will be interesting to investigate the impact of C-6 ceramide, a bioactive tumour suppressor lipid, in MPA/Abiraterone mediated effects on tumour growth as C-6 ceramide was reported to target and inhibit SET function[175]. This analysis could lead to the development of a more efficient drug combination for further preclinical studies and ultimately clinical trials. Moreover, functional validation experiments are still warranted, namely overexpression and knock-out of SET proto-oncogene in both in-vivo and in-vitro CRPC models.

Once functionally validated, SET could be both studied as a biomarker to validate the cellular response to MPA, as well as an alterantive therapeutic target.

Chapter 6 – General discussion

6.1 Verification of the research hypothesis

The aim of this dissertation was the evaluation of network-based methods for the analysis of PCa omics data. Three different applications were tested: in Chapter 3, the regulons of a gene regulatory network, derived from transcriptomics profiles of pre-clinical models, were used as statistical units for a survival analysis of patients samples; in Chapter 4, both regulons and protein complexes information was combined to generate multi-layer integrative modules for the investigation of differences between CRPC and PC status; in Chapter 5, the same gene regulatory network was appraised by the ability to detect regulational changes induced by different treatments combinations.

Noteworthy, the GNR proved to be useful for the prioritisation of transcriptional regulators with prognostic ability for biochemical recurrence of PCa. In fact, to the authors knowledge, the analytical method applied in Chapter 3 was the first to reveal the JMJD6 regulon as a candidate biomarker for PCa progression. The relevance of the JMJD6 protein in PCa biology was furtherly supported by our pilot knock-down experiment as well as studies performed by other groups[119,183].

Nevertheless, two major issues have still to be addressed to gain the most out of network-based analysis.

First, a consolidated approach to validate the results of any network modelling is currently lacking. *Ad-hoc* experiments to indirectly confirm the predicted gene-gene relationships by artificially rewiring the network are recommended[184], since the isolated analysis of a single gene may not provide an exhaustive explanation of the cause-effect mechanisms underlying the phenotype of interest.

Secondly, the activity of the transcriptional regulators may change over time according to the chromatin accessibility of the target genes[185]. This finding suggests considering multiple omics layers to infer gene-gene relationships.

For example, chip-seq data and methylation measurements can be used to assess the availability of the promoter or enhancer region of the target gene to further support the predicted relationships.

The joint analysis of regulons and protein complexes was effective in highlighting and exploiting the complementarity of the information provided by differentially expressed genes and proteins, identified from the CRPC vs PC contrast. Interestingly, the over-representation analysis performed using the individual sets of differentially expressed features revealed two different aspects of the same tumour biology, and our joint enrichment analysis highlighted a transcriptional regulator that had not been prioritised by neither of the isolated lists.

However, our workflow showed some limitations: the known difference in the number of identifiable differentially expressed genes and proteins from the same contrast[151] may have biased the enrichment by giving higher importance to the regulons than to the protein complexes. Moreover, the conjunction of regulons and protein complexes by means of shared features only, did not consider the full spectrum of possible gene-proteins interactions, such as those involving proteins with RNA-binding capacity. Lastly, given the limited number of significant results obtained applying the permutations-based method developed here, a rigorous benchmarking is required to find a suitable balance between the number of true and false positive results.

The assessment of treatment response of *in-vivo* PCa xenografts to single or combination of drugs proved the usefulness of the inferred GRN in a different experimental context. In fact, the prioritised transcriptional regulations of SET proto-oncogene represent a plausible link with the known effects of the drug under investigation (MPA).

Despite these results, the interpretation of the enrichment of regulons containing a large number of genes is complicated by the fact that most transcription factors regulate genes in multiple biological processes[186]. In order to facilitate the generation of cause-effect hypothesis and eventually increase the robustness of the analysis, gene ontologies memberships could be taken into consideration within the enrichment step of the analysis. Finally, functional validations are required to assess the robustness of the single findings and, one-to-one comparisons with PSA and other standards of care, are needed to fully demonstrate the clinical utility of the newly putative biomarkers.

6.2 GRN application in cancer research

Regulatory networks are systems biology approaches effective in capturing cooperative or mutual interactions among genes, especially in contexts in which the phenotype is due to a rewiring of the gene expression[35].

GRNs play an important role in cancer research and have been applied to study several tumour types, for example: master regulators were identified from a GRN derived from microarray gene expression experiment to understand the upstream events leading to the development of breast cancer[187]; co-activated transcription factors were similarly studied from an oesophageal squamous cell carcinoma GRN to gain insights into the carcinogenesis mechanisms[188]; a network comprising transcription factors, mRNA and lncRNAs expression, revealed 15 core modules associated with lung adenocarcinoma as putative targets for therapeutic strategies[189].

Similar examples can be found in PCa literature as well, given the key role that transcription factors, like the androgen receptor, play in the disease. Several studies of PCa biology have been based on GRNs developed from either microarray or RNAseq expression profiles.

Initially, sub-networks of the genes SREBFI, STAT6 and PBXI have been associated with the development of prostate cancer, while SLC22A3 regulon explained the differentiation between high and low grade cancers[190].

In a subsequent study, gene expression, miRNAs expression and clinical data have been jointly used to generate a GRN from PCa patients samples to find modules linked to clinical parameters such as the aggressiveness of the disease[191].

Similarly, an independent study exploited gene and miRNA expression profiles to reveal HOXD10, BCL2 and PGR regulons as the most influent factors of primary PCa, as well as STAT3, JUN and JUNB as key players in the metastatic disease. These results were validated through a survival analysis based on the high/low expression levels of the prioritized transcription factors[192].

The studies cited above, while sharing part of the methodologies for the inference and analysis of the GRN with the work described in this thesis, lacked the utilisation of the regulon enrichment as a predictive biomarker.

In greater detail, our network-based analysis was motivated by the consideration of the concordance among the full set of regulatory relationships, rather than the expression levels of the individual transcriptional regulators, as the biomarker for the observed phenotypes.

With this view in mind, rather than looking for isolated biomarkers, the author suggests shifting the research focus on the rewiring of molecular interactions occurring during the development of complex diseases such as cancer.

6.3 Future studies

With the experience obtained from my research activity, I envision future network-based studies of cancer omics data, to improve the analysis workflow by finding solutions to the encountered issues.

To increase the accuracy of the inferred relationships and hence the interpretability of the regulational modules, it would be recommended to introduce additional omics layers in the analysis, starting from epigenetics data. The information obtained from histone modifications and chromatin accessibility

experiments can reveal a more comprehensive map of the phenotype specific regulations, including those involving distal enhancer regions and transcription factors cooperation[193].

Similarly, single-cell derived data can reveal the extent of heterogeneity confounding the omics profile, so to discern stochastic associations from mechanistic regulatory relationships[194]. Such strategy could lead to more robust and interpretable regulons to be used as prognostic/diagnostic markers.

An additional aspect worth investigating is the influence of proteins post-translational modifications (PTMs) over the network of protein-protein interactions. In fact, human proteins undergoing PTMs were found associated with specific protein interaction network properties[195]. By taking into account the modification status of the proteins, it is possible to increase accuracy in the reconstruction of phenotype specific interaction networks, in order to replace less suitable external databases.

Lastly, in order to ensure the robustness, reproducibility, and comparability across -omics studies, it is needed to further investigate both wet and dry lab protocols to account for the lack of gold-standard unified sample processing workflows, and post-processing data analysis methods such as normalization, transformation, and scaling[32]. This, together with standardization in the annotation of clinical data, would minimize the barrier for the translation omics derived biomarkers into clinical practice.

References

1. Rawla, P. Epidemiology of Prostate Cancer. *World J. Oncol.* **10**, 63–89 (2019).

2. Ahn, H. K., Lee, Y. H. & Koo, K. C. Current status and application of metformin for prostate cancer: A comprehensive review. *Int. J. Mol. Sci.* **21**, 1–18 (2020).

3. Descotes, J. L. Diagnosis of prostate cancer. *Asian J. Urol.* **6**, 129–136 (2019).

4. Litwin, M. S. & Tan, H. J. The diagnosis and treatment of prostate cancer: A review. *JAMA - J. Am. Med. Assoc.* **317**, 2532–2542 (2017).

5. Salami, S. S. *et al.* Transcriptomic heterogeneity in multifocal prostate cancer. *JCI Insight* **3**, (2018).

6. Klink, J. C., Miocinovic, R., Galluzzi, C. M. & Klein, E. A. High-Grade prostatic intraepithelial neoplasia. *Korean J. Urol.* **53**, 297–303 (2012).

7. Mottet, N. *et al.* EAU-ESTRO-SIOG Guidelines on Prostate Cancer. Part 1: Screening, Diagnosis, and Local Treatment with Curative Intent. *Eur. Urol.* **71**, 618–629 (2017).

8. Gay, H. A. & Michalski, J. M. Radiation Therapy for Prostate Cancer. *Mo. Med.* **115**, 146–150 (2018).

9. Cornford, P. *et al.* EAU-ESTRO-SIOG Guidelines on Prostate Cancer. Part II: Treatment of Relapsing, Metastatic, and Castration-Resistant Prostate Cancer. *Eur. Urol.* **71**, 630–642 (2017).

10. Moris, L. *et al.* Benefits and Risks of Primary Treatments for High-risk Localized and Locally Advanced Prostate Cancer: An International Multidisciplinary Systematic Review[Formula presented]. *Eur. Urol.* **77**, 614–627 (2020).

11. Kornberg, Z., Cooperberg, M. R., Spratt, D. E. & Feng, F. Y. Genomic biomarkers in prostate cancer. *Transl. Androl. Urol.* **7**, 459–471 (2018).

12. Salji, M. J., Ma, M. & Mrcs, C. Quantitative Proteomics and Metabolomics of Castration Resistant Prostate Cancer. (2018).

13. Soloway, M., Roach, M. & Iii, M. R. Prostate cancer progression after therapy of primary curative intent: A review of data from the prostate-specific antigen era. *Cancer* **104**, 2310–2322 (2005).

14. Karantanos, T., Corn, P. G. & Thompson, T. C. Prostate cancer progression after androgen deprivation therapy: Mechanisms of castrate resistance and novel therapeutic approaches. *Oncogene* **32**, 5501–5511 (2013).

15. Tilki, D., Schaeffer, E. M. & Evans, C. P. Understanding Mechanisms of Resistance in Metastatic Castration-resistant Prostate Cancer: The Role of the Androgen Receptor. *Eur. Urol. Focus* **2**, 499–505 (2016).

16. Shah, H. & Vaishampayan, U. Therapy of Advanced Prostate Cancer: Targeting the Androgen Receptor Axis in Earlier Lines of Treatment. *Target. Oncol.* **13**, 679–689 (2018).

17. Rafael Sánchez Martínez. Characterisation and Role of SLFN5 in Castration Resistant Prostate Cancer. (University of Glasgow, 2020).

18. Saad, F. Evidence for the efficacy of enzalutamide in postchemotherapy metastatic castrate-resistant prostate cancer. *Ther. Adv. Urol.* **5**, 201–210 (2013).

19. Chong, J. T., Oh, W. K. & Liaw, B. C. Profile of apalutamide in the treatment of metastatic castration-resistant prostate cancer: evidence to date. *Onco. Targets. Ther.* **11**, 2141–2147 (2018).

20. Hoy, S. M. Abiraterone acetate: A review of its use in patients with metastatic castration-resistant prostate cancer. *Drugs* **73**, 2077–2091 (2013).

21. Hessels, D. & Schalken, J. A. Recurrent gene fusions in prostate cancer: Their clinical implications and uses. *Curr. Urol. Rep.* **14**, 214–222 (2013).

22. Wang, Z. *et al.* Significance of the TMPRSS2:ERG gene fusion in prostate

cancer. *Mol. Med. Rep.* **16**, 5450–5458 (2017).

23. Sun, C. *et al.* TMPRSS2-ERG fusion, a common genomic alteration in prostate cancer activates C-MYC and abrogates prostate epithelial differentiation. *Oncogene* **27**, 5348–5353 (2008).

24. Abeshouse, A. *et al.* The Molecular Taxonomy of Primary Prostate Cancer. *Cell* **163**, 1011–1025 (2015).

25. Robinson, D. *et al.* Integrative clinical genomics of advanced prostate cancer. *Cell* **161**, 1215–1228 (2015).

26. Hopkins, J. F. *et al.* Mitochondrial mutations drive prostate cancer aggression. *Nat. Commun.* **8**, 656 (2017).

27. Houlahan, K. E. *et al.* Genome-wide germline correlates of the epigenetic landscape of prostate cancer. *Nat. Med.* **25**, 1615–1626 (2019).

28. Sinha, A. *et al.* The Proteogenomic Landscape of Curable Prostate Cancer. *Cancer Cell* **35**, 414-427.e6 (2019).

29. Latonen, L. *et al.* Integrative proteomics in prostate cancer uncovers robustness against genomic and transcriptomic aberrations during disease progression. *Nat. Commun.* **9**, (2018).

30. Kamel, H. F. M. & Al-Amodi, H. S. A. B. Exploitation of Gene Expression and Cancer Biomarkers in Paving the Path to Era of Personalized Medicine. *Genomics, Proteomics Bioinforma.* **15**, 220–235 (2017).

31. Wang*, J. Z. and E. Cancer Biomarker Discovery for Precision Medicine: New Progress. *Current Medicinal Chemistry* vol. 26 7655–7671 (2019).

32. Menyhárt, O. & Győrffy, B. Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis. *Comput. Struct. Biotechnol. J.* **19**, 949–960 (2021).

33. Oldenhuis, C. N. A. M., Oosting, S. F., Gietema, J. A. & de Vries, E. G. E. Prognostic versus predictive value of biomarkers in oncology. *Eur. J. Cancer* **44**, 946–953 (2008).

34. Fuchs, P., Loeseken, C., Schubert, J. K. & Miekisch, W. Breath gas aldehydes as biomarkers of lung cancer. *Int. J. cancer* **126**, 2663–2670

(2010).

35. Castro, M. A. A. *et al*. Regulators of genetic risk of breast cancer identified by integrative network analysis. *Nat. Genet.* **48**, 12–21 (2015).

36. Ryan, C. J., Kennedy, S., Bajrami, I., Matallanas, D. & Lord, C. J. A Compendium of Co-regulated Protein Complexes in Breast Cancer Reveals Collateral Loss Events. *Cell Syst.* **5**, 399-409.e5 (2017).

37. Couñago, F. *et al*. Clinical applications of molecular biomarkers in prostate cancer. *Cancers (Basel).* **12**, 1–25 (2020).

38. Lamy, P. J. *et al*. Prognostic Biomarkers Used for Localised Prostate Cancer Management: A Systematic Review. *Eur. Urol. Focus* **4**, 790–803 (2018).

39. Saini, S. PSA and beyond: alternative prostate cancer biomarkers. *Cell. Oncol.* **39**, 97–106 (2016).

40. Prensner, J. R., Rubin, M. A., Wei, J. T. & Chinnaiyan, A. M. Beyond PSA: The next generation of prostate cancer biomarkers. *Sci. Transl. Med.* **4**, (2012).

41. Urbanucci, A. *et al*. Androgen Receptor Deregulation Drives Bromodomain-Mediated Chromatin Alterations in Prostate Cancer. *Cell Rep.* **19**, 2045–2059 (2017).

42. Georgescu, C. *et al*. A TMEFF2-regulated cell cycle derived gene signature is prognostic of recurrence risk in prostate cancer. *BMC Cancer* **19**, 423 (2019).

43. Yang, L. *et al*. Development and Validation of a 28-gene Hypoxia-related Prognostic Signature for Localized Prostate Cancer. *EBioMedicine* **31**, 182–189 (2018).

44. Zhou, B. *et al*. Quantitative proteomic analysis of prostate tissue specimens identifies deregulated protein complexes in primary prostate cancer. *Clin. Proteomics* **16**, 1–18 (2019).

45. Olivier, M., Asmis, R., Hawkins, G. A., Howard, T. D. & Cox, L. A. The need for multi-omics biomarker signatures in precision medicine. *Int. J.*

Mol. Sci. **20**, (2019).

46. Lu, M. & Zhan, X. The crucial role of multiomic approach in cancer research and clinically relevant outcomes. *EPMA J.* **9**, 77–102 (2018).

47. Pavlopoulos, G. A. *et al.* Using graph theory to analyze biological networks. *BioData Min.* **4**, 1–27 (2011).

48. Aittokallio, T. & Schwikowski, B. Graph-based methods for analysing networks in cell biology. *Brief. Bioinform.* **7**, 243–255 (2006).

49. Geistlinger, L., Csaba, G., Küffner, R., Mulder, N. & Zimmer, R. From sets to graphs: Towards a realistic enrichment analysis of transcriptomic systems. *Bioinformatics* (2011) doi:10.1093/bioinformatics/btr228.

50. Yu, H. T., Wu, C., Liu, W. W., Fu, X. P. & He, J. Modeling of gene regulatory networks. *Acad. J. Second Mil. Med. Univ.* **27**, 737–740 (2006).

51. Margolin, A. A. *et al.* Reverse engineering cellular networks. *Nat. Protoc.* **1**, 662–671 (2006).

52. Allen, J. D., Xie, Y., Chen, M., Girard, L. & Xiao, G. Comparing statistical methods for constructing large scale gene networks. *PLoS One* **7**, 17–19 (2012).

53. Sims, D., Sudbery, I., Ilott, N. E., Heger, A. & Ponting, C. P. Sequencing depth and coverage: Key considerations in genomic analyses. *Nat. Rev. Genet.* **15**, 121–132 (2014).

54. Gibbs, D. L. ProCoNA : Protein Co-expression Network Analysis. 1–9 (2014).

55. Vella, D., Zoppis, I., Mauri, G., Mauri, P. & Di Silvestre, D. From protein-protein interactions to protein co-expression networks: a new perspective to evaluate large-scale proteomic data. *Eurasip J. Bioinforma. Syst. Biol.* **2017**, (2017).

56. Gibbs, D. L. *et al.* Protein co-expression network analysis (ProCoNA). 1–10 (2013).

57. Sciences, B., Comprehensive, S. O. & Angeles, L. Complexes in Primary Prostate Cancer. **1**, (2018).

58. Kanonidis, E. I., Roy, M. M., Deighton, R. F. & Le Bihan, T. Protein co-

expression analysis as a strategy to complement a standard quantitative proteomics approach: Case of a glioblastoma multiforme study. *PLoS One* **11**, 1–22 (2016).

59. Kumar, D. *et al*. Integrating transcriptome and proteome profiling: Strategies and applications. *Proteomics* **16**, 2533–2544 (2016).

60. Haider, S. & Pal, R. Integrated Analysis of Transcriptomic and Proteomic Data. *Curr. Genomics* **14**, 91–110 (2013).

61. Salji, M. J. *et al*. Multi-Omics Analysis Identifies Local N-Acetyl Aspartate Accumulation as a Feature of Castration Resistant Prostate Cancer. *iScience* (2021) doi:10.2139/ssrn.3762111.

62. Salji, M.J., Blomme, A., Däbritz, J.H.M., Repiscak, P., Lilla, S., Patel, R. & Sumpton, D., van den Broek, N.J.F., Daly, R., Zanivan, S., Leung, H. . Multi-omics and Pathway Analysis Identify Potential Roles for Tumour N-Acetyl Aspartate Accumulation in Murine Models of Castration Resistant Prostate Cancer. *iScience* (2022) doi:https://doi.org/10.1016/j.isci.2022.104056.

63. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

64. Kim, D. *et al*. TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, 1–13 (2013).

65. Ahdesmäki, M. J., Gray, S. R., Johnson, J. H. & Lai, Z. Disambiguate: An open-source application for disambiguating two species in next generation sequencing data from grafted samples. *F1000Research* **5**, 1–11 (2017).

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

67. Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185 (2012).

68. Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).

69. Trapnell, C. *et al*. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).

70. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 1–21 (2014).

71. Geistlinger, L. Seamless navigation through combined re- sults of set- & network-based enrichment analysis. 1–23 (2017).

72. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).

73. Liberzon, A. *et al*. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).

74. Fletcher, M. N. C. *et al*. Master regulators of FGFR2 signalling and breast cancer risk. *Nat. Commun.* **4**, 1–12 (2013).

75. Alvarez, M. J. *et al*. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat. Genet.* **48**, 838–847 (2016).

76. Zhao, M. *et al*. Lipofectamine RNAiMAX: An Efficient siRNA Transfection Reagent in Human Embryonic Stem Cells. *Mol. Biotechnol.* **40**, 19–26 (2008).

77. Clegg, N. J. *et al*. ARN-509: a novel antiandrogen for prostate cancer treatment. *Cancer Res.* **72**, 1494–1503 (2012).

78. Tran, C. *et al*. Development of a Second-Generation Antiandrogen for Treatment of Advanced Prostate Cancer. *Science (80-.).* **324**, 787 LP – 790 (2009).

79. Siegel, R. L., Miller, K. D., Fuchs, H. E. & Jemal, A. Cancer Statistics, 2021. *CA. Cancer J. Clin.* **71**, 7–33 (2021).

80. Ciccarese, C. *et al*. Prostate cancer heterogeneity: Discovering novel molecular targets for therapy. *Cancer Treat. Rev.* **54**, 68–73 (2017).

81. Amaral, T. M. S., Macedo, D., Fernandes, I. & Costa, L. Castration-Resistant Prostate Cancer: Mechanisms, Targets, and Treatment. *Prostate Cancer* **2012**, 1–11 (2012).

82. Parimi, V., Goyal, R., Poropatich, K. & Yang, X. J. Neuroendocrine differentiation of prostate cancer: a review. *Am. J. Clin. Exp. Urol.* **2**, 273–85 (2014).

83. Myers, J. S., von Lersner, A. K., Robbins, C. J. & Sang, Q.-X. A. Differentially Expressed Genes and Signature Pathways of Human Prostate Cancer. *PLoS One* **10**, e0145322 (2015).

84. Zeng, T., Sun, S. Y., Wang, Y., Zhu, H. & Chen, L. Network biomarkers reveal dysfunctional gene regulations during disease progression. *FEBS J.* **280**, 5682–5695 (2013).

85. Prolaris Cell Cycle Progression Test for Localized Prostate Cancer: A Health Technology Assessment. *Ont. Health Technol. Assess. Ser.* **17**, 1–75 (2017).

86. Dalela, D., Löppenberg, B., Sood, A., Sammon, J. & Abdollah, F. Contemporary Role of the Decipher® Test in Prostate Cancer Management: Current Practice and Future Perspectives. *Rev. Urol.* **18**, 1–9 (2016).

87. Van Den Eeden, S. K. *et al.* A Biopsy-based 17-gene Genomic Prostate Score as a Predictor of Metastases and Prostate Cancer Death in Surgically Treated Men with Clinically Localized Disease. *Eur. Urol.* **73**, 129–138 (2018).

88. Dempster, J. *et al.* Gene expression has more power for predicting in vitro cancer cell vulnerabilities than genomics. (2020) doi:10.1101/2020.02.21.959627.

89. Creed, J. H. *et al.* Commercial Gene Expression Tests for Prostate Cancer Prognosis Provide Paradoxical Estimates of Race-Specific Risk. *Cancer Epidemiol. Biomarkers & Prev.* **29**, 246 LP – 253 (2020).

90. Wu, S.-Q., Su, H., Wang, Y.-H. & Zhao, X.-K. Role of tumor-associated immune cells in prostate cancer: angel or devil? *Asian J. Androl.* **21**, 433–437 (2019).

91. Krušlin, B., Ulamec, M. & Tomas, D. Prostate cancer stroma: an important

factor in cancer growth and progression. *Bosn. J. basic Med. Sci.* **15**, 1–8 (2015).

92. Fitzgerald, K. A. *et al.* The role of transcription factors in prostate cancer and potential for future RNA interference therapy. *Expert Opin. Ther. Targets* **18**, 633–649 (2014).

93. Grimes, T., Potter, S. S. & Datta, S. Integrating gene regulatory pathways into differential network analysis of gene expression data. *Sci. Rep.* **9**, 1–12 (2019).

94. Mochida, K., Koda, S., Inoue, K. & Nishii, R. Statistical and machine learning approaches to predict gene regulatory networks from transcriptome datasets. *Front. Plant Sci.* **871**, 1–7 (2018).

95. Ylipää, A. *et al.* Transcriptome sequencing reveals PCAT5 as a Novel ERG-Regulated long Noncoding RNA in prostate cancer. *Cancer Res.* **75**, 4026–4031 (2015).

96. Chen, S. *et al.* Widespread and Functional RNA Circularization in Localized Prostate Cancer. *Cell* **176**, 831-843.e22 (2019).

97. Hendriksen, P. J. M. *et al.* Evolution of the androgen receptor pathway during progression of prostate cancer. *Cancer Res.* **66**, 5012–5020 (2006).

98. Gerhauser, C. *et al.* Molecular Evolution of Early-Onset Prostate Cancer Identifies Molecular Risk Markers and Clinical Trajectories. *Cancer Cell* **34**, 996-1011.e8 (2018).

99. Hsu, C. L. *et al.* Identification of a new androgen receptor (AR) co-regulator BUD31 and related peptides to suppress wild-type and mutated AR-mediated prostate cancer growth via peptide screening and X-ray structure analysis. *Mol. Oncol.* **8**, 1575–1587 (2014).

100. Baek, J. H. *et al.* PLOD3 suppression exerts an anti-tumor effect on human lung cancer cells by modulating the PKC-delta signaling pathway. *Cell Death Dis.* (2019) doi:10.1038/s41419-019-1405-8.

101. MSigDb. SCHAEFFER_PROSTATE_DEVELOPMENT_48HR_UP. https://www.gsea-msigdb.org/gsea/msigdb/cards/SCHAEFFER_PROSTATE_DEVELOPMENT_48H

R_UP.

102. Suyama, T. *et al*. Expression of cancer/testis antigens in prostate cancer is associated with disease progression. *Prostate* **70**, 1778–1787 (2010).

103. Goel, M. M., Agrawal, D., Natu, S. M. & Goel, A. Hepsin immunohistochemical expression in prostate cancer in relation to Gleason's grade and serum prostate specific antigen. *Indian J. Pathol. Microbiol.* **54**, 476–481 (2011).

104. Ashikari, D., Takayama, K. I., Obinata, D., Takahashi, S. & Inoue, S. CLDN8, an androgen-regulated gene, promotes prostate cancer cell proliferation and migration. *Cancer Sci.* **108**, 1386–1393 (2017).

105. Rotinen, M. *et al*. ONECUT2 is a targetable master regulator of lethal prostate cancer that suppresses the androgen axis. *Nat. Med.* **24**, 1887–1898 (2018).

106. Shajari, N. *et al*. Silencing of BACH1 inhibits invasion and migration of prostate cancer cells by altering metastasis-related gene expression. *Artif. Cells, Nanomedicine Biotechnol.* **46**, 1495–1504 (2018).

107. Shin, S. H. *et al*. Aberrant expression of CITED2 promotes prostate cancer metastasis by activating the nucleolin-AKT pathway. *Nat. Commun.* **9**, (2018).

108. Lee, E. *et al*. DNMT1 Regulates Epithelial-Mesenchymal Transition and Cancer Stem Cells, Which Promotes Prostate Cancer Metastasis. *Neoplasia (United States)* **18**, 553–566 (2016).

109. Garcia-Alonso, L. *et al*. Transcription factor activities enhance markers of drug sensitivity in cancer. *Cancer Res.* **78**, 769–780 (2018).

110. Lee, C. R. *et al*. Elevated expression of JMJD6 is associated with oral carcinogenesis and maintains cancer stemness properties. *Carcinogenesis* **37**, 119–128 (2015).

111. Gao, W. wei *et al*. JMJD6 Licenses ERα-Dependent Enhancer and Coding Gene Activation by Modulating the Recruitment of the CARM1/MED12 Co-activator Complex. *Mol. Cell* **70**, 340-357.e8 (2018).

112. Wong, M. *et al*. JMJD6 is a tumorigenic factor and therapeutic target in neuroblastoma. *Nat. Commun.* **10**, 1–15 (2019).

113. Liu, X. *et al*. JMJD6 promotes melanoma carcinogenesis through regulation of the alternative splicing of PAK1, a key MAPK signaling component. *Mol. Cancer* **16**, 1–18 (2017).

114. Zheng, H. *et al*. Jumonji domain-containing 6 (JMJD6) identified as a potential therapeutic target in ovarian cancer. *Signal Transduct. Target. Ther.* **4**, (2019).

115. Mitra, A. *et al*. Overexpression of RAD51 occurs in aggressive prostatic cancer. *Histopathology* **55**, 696–704 (2009).

116. Na, R., Wu, Y., Ding, Q. & Xu, J. Clinically available RNA profiling tests of prostate tumors: Utility and comparison. *Asian J. Androl.* **18**, 575–579 (2016).

117. Testa, U., Castelli, G. & Pelosi, E. Cellular and Molecular Mechanisms Underlying Prostate Cancer Development: Therapeutic Implications. *Medicines* **6**, 82 (2019).

118. Mills, I. G. Maintaining and reprogramming genomic androgen receptor activity in prostate cancer. *Nat. Rev. Cancer* **14**, 187–198 (2014).

119. Paschalis, A. *et al*. JMJD6 Is a Druggable Oxygenase That Regulates AR-V7 Expression in Prostate Cancer. *Cancer Res.* **81**, 1087–1100 (2021).

120. Unoki, M. *et al*. Lysyl 5-hydroxylation, a novel histone modification, by Jumonji domain containing 6 (JMJD6). *J. Biol. Chem.* **288**, 6053–6062 (2013).

121. Poulard, C. *et al*. Role of JMJD6 in Breast Tumourigenesis. *PLoS One* **10**, e0126181 (2015).

122. Zhang, J., Ni, S.-S., Zhao, W.-L., Dong, X.-C. & Wang, J.-L. High expression of JMJD6 predicts unfavorable survival in lung adenocarcinoma. *Tumour Biol. J. Int. Soc. Oncodevelopmental Biol. Med.* **34**, 2397–2401 (2013).

123. Aprelikova, O. *et al*. The epigenetic modifier JMJD6 is amplified in

mammary tumors and cooperates with c-Myc to enhance cellular transformation, tumor progression, and metastasis. *Clin. Epigenetics* **8**, 38 (2016).

124. Fard, A. T. & Ragan, M. A. Modeling the attractor landscape of disease progression: A network-based approach. *Front. Genet.* **8**, 1–11 (2017).

125. Borràs, E. & Sabidó, E. What is targeted proteomics? A concise revision of targeted acquisition and targeted data analysis in mass spectrometry. *Proteomics* **17**, (2017).

126. Pan, S. *et al.* Mass Spectrometry Based Targeted Protein Quantification: Methods and Applications. 787–797 (2009).

127. Baldwin, M. A. Protein identification by mass spectrometry: Issues to be considered. *Mol. Cell. Proteomics* **3**, 1–9 (2004).

128. Fenselau, C. A review of quantitative methods for proteomic studies. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* **855**, 14–20 (2007).

129. Chen, X., Wei, S., Ji, Y., Guo, X. & Yang, F. Quantitative proteomics using SILAC: Principles, applications, and developments. *Proteomics* **15**, 3175–3192 (2015).

130. Martinez, R. S. *et al.* SLFN5 Regulates LAT1-Mediated mTOR Activation in Castration-Resistant Prostate Cancer. *Cancer Res.* **81**, 3664–3678 (2021).

131. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).

132. Bairoch, A. & Apweiler, R. The SWISS-PROT Protein Sequence Data Bank and Its New Supplement TREMBL. *Nucleic Acids Res.* **24**, 21–25 (1996).

133. Cox, J. *et al.* Andromeda: A peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).

134. Brosch, M., Yu, L., Hubbard, T. & Choudhary, J. Accurate and sensitive peptide identification with Mascot Percolator. *J. Proteome Res.* **8**, 3176–3181 (2009).

135. Cox, J. *et al.* Accurate proteome-wide label-free quantification by delayed

normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* **13**, 2513–2526 (2014).

136. Liu, N. Q., Dekker, L. J. M. & Umar, A. Quantitative Proteomic Analysis of Microdissected Breast Cancer Tissues: Comparison of Label-Free and SILAC-based Quantification with Shotgun, Directed, and Targeted MS Approaches. (2014) doi:10.1007/978-1-4939-0685-7.

137. Serin, E. A. R., Nijveen, H., Hilhorst, H. W. M. & Ligterink, W. Learning from co-expression networks: Possibilities and challenges. *Front. Plant Sci.* **7**, 1–18 (2016).

138. Wang, J. *et al.* Proteome profiling outperforms transcriptome profiling for coexpression based gene function prediction. *Mol. Cell. Proteomics* **16**, 121–134 (2017).

139. Kerrigan, J. J., Xie, Q., Ames, R. S. & Lu, Q. Production of protein complexes via co-expression. *Protein Expr. Purif.* **75**, 1–14 (2011).

140. Stefan, A., Ceccarelli, A., Conte, E., Silva, A. M. & Hochkoeppler, A. The multifaceted benefits of protein co-expression in Escherichia coli. *J. Vis. Exp.* 1–9 (2015) doi:10.3791/52431.

141. Giurgiu, M. *et al.* CORUM: The comprehensive resource of mammalian protein complexes - 2019. *Nucleic Acids Res.* **47**, D559–D563 (2019).

142. Cheng, Y. & Church, G. M. Biclustering of expression data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8**, 93–103 (2000).

143. Hiep, T. K., Duc, N. M. & Trung, B. Q. Local search approach for the pairwise constrained clustering problem. *ACM Int. Conf. Proceeding Ser.* **08-09-Dece**, 115–122 (2016).

144. Alanis-Lobato, G., Andrade-Navarro, M. A. & Schaefer, M. H. HIPPIE v2.0: Enhancing meaningfulness and reliability of protein-protein interaction networks. *Nucleic Acids Res.* **45**, D408–D414 (2017).

145. Singh, A. *et al.* Stoichiometric analysis of protein complexes by cell fusion and single molecule imaging. *Sci. Rep.* **10**, 1–12 (2020).

146. Nombela, P. *et al.* BRCA2 and other DDR genes in prostate cancer. *Cancers*

(Basel). **11**, 1–15 (2019).

147. Taylor, R. A. *et al*. Germline BRCA2 mutations drive prostate cancers with distinct evolutionary trajectories. *Nat. Commun.* **8**, (2017).

148. Schaeffer, L., Moncollin, V., Weeda, G., Forrester, K. & Harris, C. C. p53 modulation of TFIIH. **10**, (1995).

149. Rosa-Ribeiro, R. *et al*. Transcription factors involved in prostate gland adaptation to androgen deprivation. *PLoS One* **9**, (2014).

150. Lachmann, A. *et al*. ChEA: Transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics* **26**, 2438–2444 (2010).

151. Wang, J. *et al*. In-depth method assessments of differentially expressed protein detection for shotgun proteomics data with missing values. *Sci. Rep.* **7**, 1–8 (2017).

152. DeGraff, D. J., Aguiar, A. A. & Sikes, R. A. Disease evidence for IGFBP-2 as a key player in prostate cancer progression and development of osteosclerotic lesions. *Am. J. Transl. Res.* **1**, 115–130 (2009).

153. Quispe-Tintaya, W. 乳鼠心肌提取 HHS Public Access. *Physiol. Behav.* **176**, 139–148 (2017).

154. Penney, K. L. *et al*. Selenoprotein P genetic variants and mrna expression, circulating selenium, and prostate cancer risk and survival. *Prostate* **73**, 700–705 (2013).

155. Liberzon, A. *et al*. The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst.* **1**, 417–425 (2015).

156. Demir, U., Koehler, A., Schneider, R., Schweiger, S. & Klocker, H. Metformin anti-tumor effect via disruption of the MID1 translational regulator complex and AR downregulation in prostate cancer cells. *BMC Cancer* **14**, 1–9 (2014).

157. Aranda-Orgille, B. *et al*. Protein Phosphatase 2A (PP2A)-specific ubiquitin ligase MID1 is a sequence-dependent regulator of translation efficiency controlling 3-Phosphoinositide-dependent Protein Kinase-1 (PDPK-1). *J.*

Biol. Chem. **286**, 39945–39957 (2011).

158. Grupp, K. *et al.* High mitochondria content is associated with prostate cancer disease progression. *Mol. Cancer* **12**, 1–11 (2013).

159. Mushtaq, M. *et al.* The MRPS18-2 protein levels correlate with prostate tumor progression and it induces CXCR4-dependent migration of cancer cells. *Sci. Rep.* **8**, 1–14 (2018).

160. Chiu, H. Y., Tay, E. X. Y., Ong, D. S. T. & Taneja, R. Mitochondrial Dysfunction at the Center of Cancer Therapy. *Antioxidants Redox Signal.* **32**, 309–330 (2020).

161. Aranda-Orgillés, B. *et al.* The Opitz syndrome gene product MID1 assembles a microtubule-associated ribonucleoprotein complex. *Hum. Genet.* **123**, 163–176 (2008).

162. Liu, E., Knutzen, C. A., Krauss, S., Schweiger, S. & Chiang, G. G. Control of mTORC1 signaling by the Opitz syndrome protein MID1. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 8680–8685 (2011).

163. Winter, J., Basilicata, M. F., Stemmler, M. P. & Krauss, S. The MID1 protein is a central player during development and in disease. *Front. Biosci. - Landmark* **21**, 664–682 (2016).

164. Köhler, A. *et al.* A hormone-dependent feedback-loop controls androgen receptor levels by limiting MID1, a novel translation enhancer and promoter of oncogenic signaling. *Mol. Cancer* **13**, 1–16 (2014).

165. Penney, K. L. *et al.* Association of prostate cancer risk variants with gene expression in normal and tumor tissue. *Cancer Epidemiol. biomarkers Prev. a Publ. Am. Assoc. Cancer Res. cosponsored by Am. Soc. Prev. Oncol.* **24**, 255–260 (2015).

166. Jhun, M. A. *et al.* Abstract 4957: Gene expression signature of Gleason score is associated with prostate cancer outcomes in a radical prostatectomy cohort. **8**, 4957–4957 (2017).

167. Whitaker, H. C. *et al.* N-acetyl-L-aspartyl-L-glutamate peptidase-like 2 is overexpressed in cancer and promotes a pro-migratory and pro-metastatic phenotype. *Oncogene* **33**, 5274–5287 (2014).

168. Muhammad Mushtaq. Role of mitochondrial ribosomal protein S18-2 in cancerogenesis and in regulation of stemness and differentiation. (Karolinska Institutet, Stockholm, Sweden, 2017).

169. Takayama, K. ichi *et al*. Dysregulation of spliceosome gene expression in advanced prostate cancer by RNA-binding protein PSF. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 10461–10466 (2017).

170. Ferraldeschi, R., Welti, J., Luo, J., Attard, G. & De Bono, J. S. Targeting the androgen receptor pathway in castration-resistant prostate cancer: Progresses and prospects. *Oncogene* **34**, 1745–1757 (2014).

171. Mateo, J. *et al*. Managing Nonmetastatic Castration-resistant Prostate Cancer. *Eur. Urol.* **75**, 285–293 (2019).

172. Rathkopf, D. E. & Scher, H. I. Apalutamide for the treatment of prostate cancer. *Expert Rev. Anticancer Ther.* **18**, 823–836 (2018).

173. Majd, N. *et al*. A Review of the Potential Utility of Mycophenolate Mofetil as a Cancer Therapeutic. *J. Cancer Res.* **2014**, 1–12 (2014).

174. Barfeld, S. J. *et al*. Myc-dependent purine biosynthesis affects nucleolar stress and therapy response in prostate cancer. *Oncotarget* **6**, 12587–12602 (2015).

175. Mukhopadhyay, A., Tabanor, K., Chaguturu, R. & Aldrich, J. V. Targeting inhibitor 2 of protein phosphatase 2A as a therapeutic strategy for prostate cancer treatment. *Cancer Biol. Ther.* **14**, 962–972 (2013).

176. Smith, A. J., Karpova, Y., D'Agostino, R., Willingham, M. & Kulik, G. Expression of the Bcl-2 protein BAD promotes prostate cancer growth. *PLoS One* **4**, (2009).

177. Hernández, G., Ramírez, J. L., Pedroza-Torres, A., Herrera, L. A. & Jiménez-Ríos, M. A. The secret life of translation initiation in prostate cancer. *Front. Genet.* **10**, 1–10 (2019).

178. Iglesias-Gato, D. *et al*. OTUB1 de-ubiquitinating enzyme promotes prostate cancer cell invasion in vitro and tumorigenesis in vivo. *Mol. Cancer* **14**, 1–14 (2015).

179. Cao, Z. X. *et al*. Comprehensive investigation of alternative splicing and development of a prognostic risk score for prostate cancer based on six-gene signatures. *J. Cancer* **10**, 5585–5596 (2019).

180. Zenata, O., Dvorak, Z. & Vrzal, R. Mycophenolate Mofetil induces c-Jun-N-terminal kinase expression in 22Rv1 cells: An impact on androgen receptor signaling. *J. Cancer* **9**, 1915–1924 (2018).

181. Tsai, R. Y. L. & McKay, R. D. G. A nucleolar mechanism controlling cell proliferation in stem cells and cancer cells. *Genes Dev.* **16**, 2991–3003 (2002).

182. Meng, L., Lin, T. & Tsai, R. Y. L. Nucleoplasmic mobilization of nucleostemin stabilizes MDM2 and promotes G2-M progression and cell survival. *J. Cell Sci.* **121**, 4037–4046 (2008).

183. Tong, D. The role of JMJD6/U2AF65/AR-V7 axis in castration-resistant prostate cancer progression. *Cancer Cell Int.* **21**, 4–9 (2021).

184. Damle, S. S. & Davidson, E. H. Synthetic in vivo validation of gene network circuitry. *Proc. Natl. Acad. Sci.* **109**, 1548 LP – 1553 (2012).

185. Duren, Z., Chen, X., Jiang, R., Wang, Y. & Wong, W. H. Modeling gene regulation from paired expression and chromatin accessibility data. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E4914–E4923 (2017).

186. Ledezma-Tejeida, D., Altamirano-Pacheco, L., Fajardo, V. & Collado-Vides, J. Limits to a classic paradigm: Most transcription factors regulate genes in multiple biological processes. *bioRxiv* 479857 (2018) doi:10.1101/479857.

187. Tovar, H., García-Herrera, R., Espinal-Enríquez, J. & Hernández-Lemus, E. Transcriptional master regulator analysis in breast cancer genetic networks. *Comput. Biol. Chem.* **59**, 67–77 (2015).

188. Zhao, Y. *et al*. Construction of disease-specific transcriptional regulatory networks identifies co-activation of four gene in esophageal squamous cell carcinoma. *Oncol. Rep.* **38**, 411–417 (2017).

189. Li, D. *et al*. Transcription factor and lncRNA regulatory networks identify key elements in lung adenocarcinoma. *Genes (Basel).* **9**, 1–14 (2018).

190. Yeh, H. Y. *et al*. Identifying significant genetic regulatory networks in the prostate cancer from microarray data based on transcription factor analysis and conditional independency. *BMC Med. Genomics* **2**, 1–19 (2009).

191. Bonnet, E., Michoel, T. & Van De Peer, Y. Prediction of a gene regulatory network linked to prostate cancer from gene expression, microRNA and clinical data. *Bioinformatics* **27**, i638–i644 (2011).

192. Sadeghi, M. *et al*. MicroRNA and transcription factor gene regulatory network analysis reveals key regulatory elements associated with prostate cancer progression. *PLoS One* **11**, 1–19 (2016).

193. Jung, S. & del Sol, A. Multiomics data integration unveils core transcriptional regulatory networks governing cell-type identity. *npj Syst. Biol. Appl.* **6**, 26 (2020).

194. Fiers, M. W. E. J. *et al*. Mapping gene regulatory networks from single-cell omics data. *Brief. Funct. Genomics* **17**, 246–254 (2018).

195. Duan, G. & Walther, D. The Roles of Post-translational Modifications in the Context of Protein Interaction Networks. *PLOS Comput. Biol.* **11**, e1004049 (2015).

Appendices and supplementary material

1. Code - Chapter 3 – functions_and_libraries.R, network_analysis.R, datasets_preparation.R, single_sample_GGEA.R, survival_analysis_from_regulons.R
2. Code - Chapter 4 – functions.R, analysis.R
3. Code - Chapter 5 – analysis.R
4. Article – Supplementary Information
5. Supplementary Table 1 - Differentially expressed proteins tables
6. Supplementary Table 2 - Integrative modules composition
7. Supplementary Table 3 - Integrative modules enrichment results

Publications

1. Cangiano M, Grudniewska M, Salji MJ, Nykter M, Jenster G, Urbanucci A, Granchi Z, Janssen B, Hamilton G, Leung HY, Beumer IJ. Gene Regulation Network Analysis on Human Prostate Orthografts Highlights a Potential Role for the *JMJD6* Regulon in Clinical Prostate Cancer. Cancers (Basel). 2021 Apr 26;13(9):2094. doi: 10.3390/cancers13092094. PMID: 33925994; PMCID: PMC8123677.https://pubmed.ncbi.nlm.nih.gov/33925994/.
2. Lan Yu, Mervi Toriseva, Syeda Afshan, Mario Cangiano, Vidal Fey, Andrew Erickson, Heikki Seikkula, Kalle Alanen, Pekka Taimen, Otto Ettala, Martti Nurmi, Peter Bostrom, Tuomas Mirtti, Markku Kallajoki, Johanna Tuomela, Inès J Beumer, Matthias Nees, Pirkko Härkönen. Increased expression and altered cellular localization of fibroblast growth 'factor' receptor like 1 (FGFRL1) are associated with prostate cancer progression (Under review).