Unraveling the Complexity of Prostate Cancer: A Multi-omics study to delineate and exploit trajectories of disease progression

A doctoral dissertation presented by

Arianna Vallerga

Under the supervision of

Prof. Rolf Krause

Prof. Jean-Philippe Theurillat

Submitted to the

Faculty of Biomedical Sciences

Università della Svizzera Italiana

For the degree of

Ph.D. in Computational Biomedicine

June 2023

Table of Contents

It is important to acknowledge that most of the methodology used and the results presented in the following thesis are integral components of a recent publication:

Bolis, M.*, Bossi, D.*, **Vallerga, A***. *et al.* Dynamic prostate cancer transcriptome analysis delineates the trajectory of disease progression. *Nat Commun* **12**, 7033 (2021). https://doi.org/10.1038/s41467-021-26840-5 .

All the proteomics data shown are part of a dedicated manuscript currently in preparation.

* These authors contributed equally

A chi c'è stato,
c'è,
e ci sarà.

ABSTRACT

Prostate cancer (PCa) is a complex disease with a range of genetic and environmental factors contributing to its development and progression.

This PhD project aims to generate a harmonized Prostate Cancer Transcriptome Atlas with high-throughput transcriptional data sets from multiple studies, and to use this data to characterize and exploit disease progression. Trajectory inference analysis was applied to assign a pseudotime to each sample, describing the advancements along the path to disease evolution. Transcriptional changes in key signaling pathways throughout tumor progression along the trajectory were assessed, providing insights into the molecular mechanisms underlying prostate cancer progression. Functional validation of these findings was carried out using established human prostate cancer cell lines and patient-derived xenografts (PDX) models originating from surgically removed primary and advanced prostate cancers. The positioning of cell lines and PDX models along the trajectory was significantly associated with the originating disease stage and the dependence on androgens.

Furthermore, single-cell RNA sequencing (scRNA-seq) was performed on most PDX models *in vivo* to interrogate the individual cells' distribution along the trajectory of disease progression. The subpopulations were assessed to determine how they would evolve during the progression to androgen independence, with the identification of potential molecular targets for preventing the advancement of the disease.

Finally, to further enhance our understanding of prostate cancer progression, we added an additional level of complexity by integrating proteomics data into the analysis. We profiled the panel of PDX models at different stages of progression and found a strong correlation between mRNA and protein expression, providing further validation for our findings at the transcriptomic level. Additionally, the concordance of our results with data from patients further supports the relevance of our PDX models as preclinical tools for investigating disease progression.

Overall, this project provides a comprehensive understanding of the molecular mechanisms underlying prostate cancer progression, using transcriptomic and proteomic analysis and functional validation to generate new insights into the evolution of the disease. These findings have the potential to contribute to the development of novel diagnostic and therapeutic approaches for prostate cancer.

1 INTRODUCTION

1.1 Epidemiology of Prostate Cancer

1.1.1 Prevalence and incidence of prostate cancer

Prostate cancer is a malignant tumor that develops in the prostate gland of men. It is the most common non-skin malignancy diagnosed in men in the Western world and ranks as the fifth most common cancer overall [1]. Prostate cancer is generally asymptomatic in its early stages, making early detection and treatment difficult.

The prevalence of prostate cancer varies across different populations and geographic regions. In Europe, it is the most diagnosed cancer type in men. According to the latest statistics from the International Agency for Research on Cancer (IARC), there were approximately 450,000 new cases of prostate cancer and 95,000 deaths from the disease in Europe in 2020. Incidence and mortality rates vary across the continent, with the highest rates reported in the Northern and Western Europe [1]. The same is true for the United States, where as well it is the most diagnosed cancer among men, with an estimated 248,530 new cases and 34,130 deaths in 2021 [2]. Incidence rates are generally higher in North and South America compared to Central America and the Caribbean. In Asia, the incidence of prostate cancer is lower than in Western countries, yet it has been increasing in recent years [3].

1.1.2 Risk factors for prostate cancer

In general, the epidemiology of prostate cancer exhibits substantial regional variations worldwide, with incidence and mortality rates differing significantly. These differences may be attributable to a multitude of risk factors, which can be broadly categorized into non-modifiable and modifiable factors. Among these, age is considered the most significant, as the incidence of prostate cancer tends to rise with advancing age [4]. Other risk factors include a family history of prostate cancer, particularly in a father or brother, as well as

belonging to the African American race, with a higher incidence of prostate cancer than men of other races, and a diagnosis that is more likely to be at an advanced stage of the disease.

Modifiable risk factors include diet, obesity, and smoking. Specifically, a diet high in red meat and fat has been linked to an increased risk of developing prostate cancer [5]. Additionally, men who suffer from obesity have a higher risk of developing aggressive cancer and experiencing poorer outcomes. Overall, understanding these differences is important for developing effective prevention and treatment strategies that are tailored to the specific needs of different populations.

1.2 Physiology of the Prostate

1.2.1 Anatomy and histology of the prostate gland

The prostate is a small, walnut-sized gland found in the male reproductive system. It is located below the bladder, in front of the rectum, and surrounds the urethra, the tube that carries urine and semen out of the body. Its main function is to produce and secrete the fluid that provides nourishment and protection to sperm, which subsequently forms part of the semen. In addition to its primary role, the prostate gland also contributes to the regulation of urine flow, by secreting a fluid that can neutralize acidic urine in the urethra [6].

The prostatic gland contains three major glandular regions: the peripheral zone, which is the largest part and most prone to cancer; the central zone, which surrounds the ejaculatory ducts; and the transitional zone, which surrounds the urethra [7]. The gland is composed of both glandular and stromal tissue, with the glandular tissue being responsible for the main function of producing and secreting prostatic fluid.

The architecture of the glandular tissue is organized into acini and ducts. This organization plays a crucial role in its proper functioning, wherein the acini, small sac-like structures, are responsible for producing and storing prostate fluid, while the ducts, on the other hand, are small tubular structures that transport prostate fluid from the acini to the urethra. The histology of the prostate gland reveals that the glandular tissue comprises several types of cells, including luminal epithelial cells and basal cells. Luminal epithelial cells

are primarily responsible for producing and secreting prostate fluid, while basal cells provide support to the luminal cells and help regulate their proliferation.

The stromal tissue of the prostate is a complex network of cells and extracellular matrix components that play a critical role in maintaining the structure and function of the gland. The major components of the stromal tissue include fibroblasts, which are responsible for producing and organizing the extracellular matrix. The stromal tissue also contains smooth muscle cells, myofibroblasts, and endothelial cells that regulate blood flow to the glandular tissue and produce growth factors. Immune cells, such as macrophages, T cells, and B cells, are also present in the stromal tissue and play a role in the immune response to infection and inflammation, as well as influencing the growth and behavior of the glandular cells[8 , 9].

1.2.2 Hormonal regulation of the prostate

Prostate function and growth are regulated by male sex hormones, particularly testosterone and 4,5α-dihydrotestosterone (DHT). These hormones are produced by the testes and are responsible for the development and maintenance of the prostate gland throughout a man's life.

Testosterone is the primary male sex hormone, it is produced by the testicles and converted to DHT by an enzyme called 5-*alpha* reductase within the prostate gland [10]. DHT is a potent androgen, and it plays a crucial role in the development and maintenance of the prostate, as a key regulator of glandular growth and function, that is responsible for stimulating the growth and differentiation of prostate cells, as well as for regulating the secretory functions of the gland [11]. The hormonal regulation of the prostate gland is complex, and disruptions to this regulation can lead to a variety of disorders, including prostate cancer.

1.3 Diagnosis and treatment of prostate cancer

The diagnosis of prostate cancer is typically made by a combination of digital rectal examination, prostate-specific antigen (PSA) test, and prostate biopsy. A prostate cancer

diagnosis is typically confirmed based on a biopsy that involves the removal of small tissue samples from the prostate gland. These tissue samples are then analyzed by a pathologist to determine the histological characteristics of the tumor [12].

The treatment of prostate cancer depends on several factors, including the stage and grade of the cancer, the patient's overall health, and the patient's preferences. Treatment options include active surveillance, surgery, radiation therapy, and hormone therapy [13]. Notably, although the 5-year survival rate for patients with localized and regional tumors is almost 100%, this value decreases to 30% in patients affected by distant PCa [14].

Patients suffering from localized prostate cancer have three therapeutic alternatives: expectant management, surgery, and radiation (**Figure I.I**). Expectant management consists of a series of physical measurements and biochemical evaluations to strictly monitor any variation in disease risk which would require treatment intervention. It comprises watchful waiting, providing older patients with palliative care, and active surveillance, that focuses on younger and healthier patients, interested in deferring or avoiding the potential negative consequences of primary treatment with surgery or radiation [15].

Radical prostatectomy is defined nowadays as the partial or complete removal of the prostate through robotic-assisted laparoscopic surgery, allowing a significantly reduced invasiveness of the traditional open radical retropubic prostatectomy, while keeping oncological and functional outcomes [16].

Prostate Brachytherapy (PB) and External Beam Radiation Therapy (EBRT) are the two main types of radiotherapy used for PCa [17]. PB can be used either as monotherapy (low-intermediate-risk PCa) or in combination with EBRT (intermediate- high-risk PCa) and consists of placing radioactive sources inside or close to the neoplastic formation [18]. EBRT can be used alone or in combination with PB for intermediate-risk disease. Notably, the concurrent administration of EBRT and ADT represents a standard of care for patients with intermediate-high-risk PCa [19].

Clinically, patients suffering from advanced PCa are those affected by metastatic disease or characterized by a higher risk of disease progression and cancer-associated death

(**Figure I.I**). Metastatic disease is the leading cause of prostate cancer-associated deaths, linked to considerably diminished overall survival (OS). The standard treatment for metastatic prostate cancer has traditionally been androgen deprivation therapy (ADT) alone, which involves reducing the levels of male hormones (androgens) that stimulate the growth of prostate cancer cells. Luteinizing hormone-releasing hormone (LHRH) agonists (also called LHRH analogs) bind to the LHRH receptor in the pituitary gland, thus inhibiting testosterone synthesis and thereby blocking tumor progression and ameliorating symptoms. Initially, patients usually respond well to ADT, but over time, cancer can progress to a castration-resistant phase, where it continues to grow despite low levels of androgens (referred to as metastatic castration-resistant prostate cancer, mCRPC) [20].

In recent years, second-generation anti-androgens, such as abiraterone, apalutamide, darolutamide, and enzalutamide, have been shown to improve outcomes when combined with androgen deprivation therapy (ADT) in the treatment of metastatic prostate cancer. These drugs target the androgen receptor pathway and have demonstrated improved overall survival and delayed disease progression in clinical trials [21]. In particular, Abiraterone binds selectively and irreversibly cytochrome p450 17A1 (CYP17), thereby directly inhibiting androgens biosynthesis. On the other hand, enzalutamide binds and antagonizes the androgen receptor (AR), preventing it from translocating into the nucleus and activating AR-related genes.

Docetaxel has been shown to be an effective chemotherapy agent for the treatment of metastatic castration-resistant prostate cancer (mCRPC). As a microtubule-targeting agent, it works by disrupting the microtubule network in cancer cells, thereby blocking cell proliferation, and ultimately inducing cell death. Clinical trials have demonstrated that docetaxel improves overall survival in patients with mCRPC [22].

However, combining docetaxel with other chemotherapy agents or adding other drugs to docetaxel therapy has not consistently yielded incremental benefits. Several clinical trials have evaluated the addition of other drugs to docetaxel-based regimens, including

prednisone, bevacizumab, and cabazitaxel, but these studies have not consistently demonstrated improved outcomes [23].

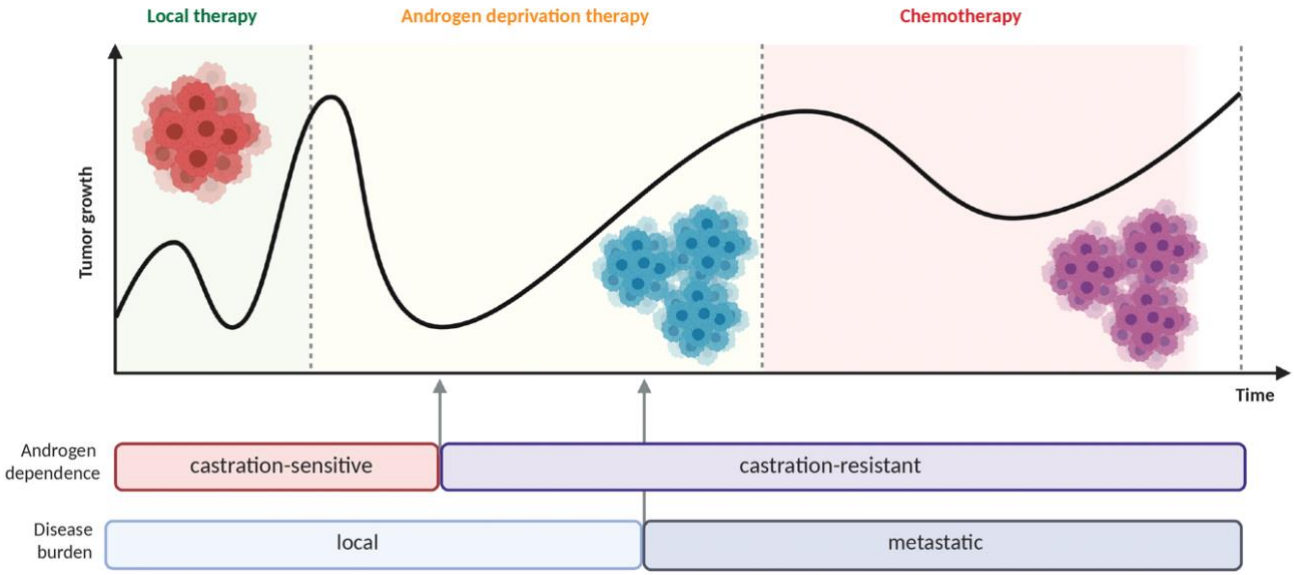


Figure I.I. The treatment options for PCa are heterogeneous. Localized disease is typically managed by active surveillance, surgery, and radiation. Following PSA rising, tumors are treated with ADT until the onset of local or metastatic CRPC. Advanced tumors are treated with second-generation antiandrogens and chemotherapy drugs. (Reprinted from "Natural History of Prostate Cancer", by BioRender.com (2023). Retrieved from hsps://app.biorender.com/biorender-templates).

1.4 Pathology of Prostate Cancer

1.4.1 Histopathology and grading of prostate cancer

One of the main challenges in prostate cancer research is the heterogeneity of the disease, which is characterized by the presence of different subtypes that have unique genetic, molecular, histological, and clinical features [24].

Histopathologically, the panel of prostate lesions can be categorized into several subtypes. Benign prostate hyperplasia (BPH) can be characterized by either an increase in the overall size of the prostate gland or abnormal growth in the number of epithelial and stromal cells without any alteration to the shape of the gland. To note, BPH is not a diagnostic feature considered specific for prostate adenocarcinoma *per se* [25].

Prostatic Intraepithelial Neoplasia (PIN) is histologically classified as intraductal dysplasia. It is characterized by the neoplastic proliferation of atypical epithelial cells within

pre-existing prostatic ducts and acini, surrounded by an intact basal cell layer [26]. Although low-grade PIN is typically associated with a benign course, high-grade PIN (HGPIN) lesions can be considered pre-cancerous events and early markers for adenocarcinoma [27]. Indeed, the onset of prostate carcinoma is driven by the expansion of malignant cells outside the basal membrane of HGPIN glands and their invasion of the surrounding stroma [28] (**Figure I.2**).

One of the most important features of prostate cancer (PCa) histology is the Gleason grading system. The Gleason grading system is used to grade prostate tumors based on the patterns of glandular architecture observed under the microscope. The grading system assigns a score ranging from 1 to 5 to two different patterns observed in the tissue sample. The two scores are then added together to produce a Gleason score that ranges from 2 to 10. Prostate tumors with a Gleason score of 6 or less are considered low-grade, while tumors with a score of 7 are intermediate-grade, and tumors with a score of 8 to 10 are high-grade. Higher Gleason scores are associated with a more aggressive disease course and poorer prognosis [29].

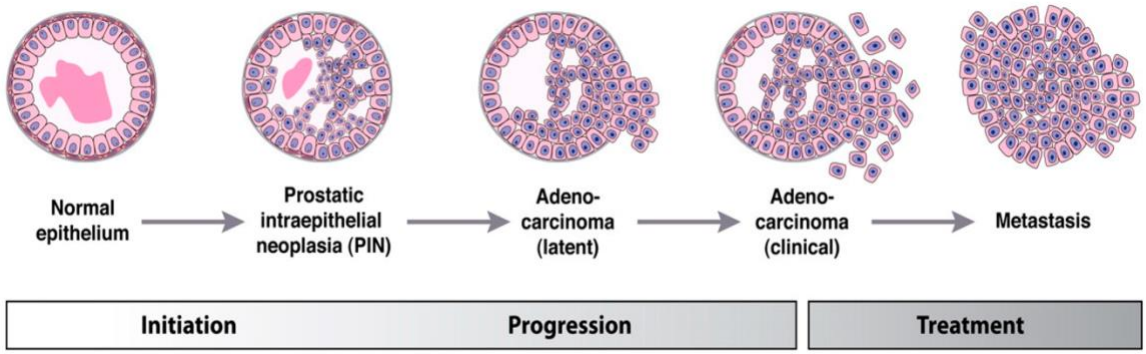


Figure I.2. **PCa depends on androgens and AR signaling**. PIN precedes the invasive adenocarcinoma, which then acquires metastatic potential and spreads outside the prostate, generally to the lymph nodes, bones, or brain. PCa growth highly depends on androgens, but advanced tumors can become androgen refractory (*adapted from [30]*).

Prostate malignant and invasive lesions are histologically classified as adenocarcinoma, neuroendocrine carcinoma, and sarcomatoid carcinomas [31]. Adenocarcinoma is the most prevalent form of prostate cancer and is characterized by an invasive cancerous tumor that originates from glandular cells and exhibits glandular characteristics. [32]. This tumor is androgen-dependent and classified as well, moderately, or poorly differentiated according to the glandular shape, number, and architecture [29]. On the

other hand, neuroendocrine and sarcomatoid tumors are far less common neoplasms and are typically characterized by androgen independence, low level of differentiation, and high metastatic potential [33] [34].

Around 15-20 % of patients with primary prostate cancer subsequently develop metastatic disease. Bone and local lymph nodes represent the dominant metastatic site for primary prostate tumors since 90% of patients with metastatic disease experience skeletal lesions [35]. On the other hand, neuroendocrine tumors also metastasize typically to visceral organs, such as bladder, liver and lungs, as well as the central nervous system (brain and spinal cord) [36]. The histopathological features of prostate cancer metastasis depend on the site of metastasis and can vary widely.

Lymph node metastases typically appear as small clusters or sheets of cancer cells within the lymph node parenchyma. The cells may be arranged in cords or trabeculae and may show a range of cellular atypia, including nuclear pleomorphism and hyperchromasia [37]. Bone metastases can manifest as osteoblastic, osteolytic, or mixed lesions, depending on the degree of bone formation or resorption induced by the cancer cells. The histopathology of bone metastases is characterized by the presence of cancer cells within the bone marrow space, which may disrupt the normal bone architecture and lead to bone pain and fractures [38]. Liver and lung metastases from prostate cancer typically resemble the primary tumor in terms of their histopathological features [39]. However, these metastases may exhibit more aggressive growth patterns, including vascular invasion and necrosis, as well as changes in cellular differentiation.

Lineage plasticity is a phenomenon where cancer cells can undergo a switch in their lineage or phenotype, leading to a change in their cellular characteristics and behavior. Metastatic castration-resistant prostate cancer (mCRPC) can undergo lineage plasticity and transform into the more aggressive neuroendocrine prostate cancer (NEPC) [40]. This process involves the loss of androgen receptor expression and the acquisition of neuroendocrine features, such as the expression of neuroendocrine markers and the production of neuropeptides [41].

NEPC is typically characterized by a loss of glandular differentiation seen in adenocarcinoma, with the tumor cells exhibiting a small, round, or ovoid morphology. The cells may be arranged in sheets or clusters, and often show a high degree of nuclear pleomorphism and mitotic activity. The tumor cells may also show neuroendocrine differentiation, as evidenced by the expression of markers such as chromogranin A and synaptophysin. Of note, NEPC can also coexist with adenocarcinoma, either as a distinct population of cells within the tumor or as a mixed phenotype. In these cases, the diagnosis of NEPC may require additional molecular testing to confirm the presence of neuroendocrine differentiation [42].

1.4.2 Molecular subtypes of prostate cancer

Molecular heterogeneity plays an important role in defining the complexity of the disease. Understanding this heterogeneity is essential for developing targeted therapies and improving treatment outcomes for patients with prostate cancer. Recent advances in molecular profiling of prostate cancer have revealed a high degree of heterogeneity within the disease, with distinct molecular subtypes exhibiting unique clinical and biological characteristics. Indeed, one key area of research has focused on identifying different molecular subtypes of prostate cancer.

The Cancer Genome Atlas (TCGA) Research Network conducted a large-scale analysis of 333 primary prostate cancer samples and identified seven molecular subtypes of prostate cancer, each with distinct genetic and molecular features[43]. These subtypes were characterized based on gene expression patterns, DNA mutations, and alterations in specific signaling pathways. The major molecular subtypes of prostate cancer identified have been defined by gene molecular classes based on distinct oncogenic drivers: fusions involving *ERG*, *ETV1*, *ETV4*, or *FLI1* (46%, 8%, 4%, and 1%, respectively); mutations in *SPOP* or *FOXA1*; or *IDH1* mutations (11%, 3%, and 1%, respectively). However, even within these subtypes, there exists significant diversity in DNA copy-number alterations, gene expression, and DNA methylation patterns. This heterogeneity in mutational profiles is reflective of the heterogeneous natural

history of primary prostate cancers. Nevertheless, the identification of these molecular subtypes has implications for the diagnosis, prognosis, and treatment of prostate cancer. It provides a framework for understanding the molecular heterogeneity of the disease and may help guide the development of targeted therapies tailored to specific subtypes of prostate cancer.

Other studies have specifically focused on identifying single genetic mutations and alterations that contribute to the development and progression of the disease. The *TMPRSS2-ERG* gene fusion is a genetic aberration that is commonly found in prostate cancer, present in approximately 40% of primary tumors. It is a result of a translocation event that fuses the androgen-regulated *TMPRSS2* gene with the *ERG* (Ets-related gene) oncogene, leading to the overexpression of *ERG* itself [44]. *SPOP* (Speckle-type POZ Protein) mutations are also relatively common events in prostate cancer, occurring in approximately 10-15% of the cases. Both ERG fusion and SPOP mutations are considered founder events that occur early in the development of the disease. Furthermore, mutations in the tumor suppressor genes *RB1*, *PTEN* and *TP53* are commonly observed drivers of prostate cancer, as well as alterations in androgen receptor signaling pathways [45]. These genetic mutations and alterations can impact the response to different treatments and may require tailored therapies as well.

In addition to genetic mutations, researchers have also identified different gene expression patterns in prostate cancer, that can be used to classify prostate cancer into different molecular subtypes, each with different clinical characteristics and treatment responses. As an example, in [46] *Robinson et al.* analyzed 150 metastatic prostate cancer samples using a variety of genomic and molecular profiling techniques, including whole-genome sequencing, whole-exome sequencing, RNA sequencing, and DNA copy number analysis. The goal was to identify the genomic and molecular characteristics of advanced prostate cancer and to develop new targeted treatments for the disease: they could identify several genetic alterations that were associated with the disease that converged into a few main signalings, including alterations in genes involved in the androgen receptor (AR)

signaling pathway, DNA damage repair, cell cycle regulation, PI3K/AKT/mTOR and WNT pathways.

Overall, the molecular heterogeneity of prostate cancer is a complex and evolving area of research. Understanding the unique genetic, molecular, and cellular features of each tumor is critical for developing personalized treatment approaches that can improve outcomes for patients with prostate cancer. Addressing these challenges will require a multi-disciplinary approach, including advances in molecular profiling, biomarker discovery, preclinical models, and treatment development.

1.4.3 Androgen receptor signaling in Prostate Cancer

Androgen receptor (AR) signaling plays a crucial role in the development and progression of prostate cancer. The androgen receptor is a transcription factor that regulates the expression of genes involved in cell growth, differentiation, and survival. Androgens, such as testosterone and dihydrotestosterone (DHT), bind to the androgen receptor and activate a transcriptional program related to cell differentiation and proliferation[47].

In normal prostate tissue, androgen receptor signaling is tightly regulated and required for the maintenance of prostate homeostasis. In prostate cancer, however, androgen receptor signaling becomes dysregulated, leading to uncontrolled growth and proliferation of prostate cells. This dysregulation can occur through various mechanisms, including androgen receptor amplification and overexpression, mutations in the androgen receptor gene, and aberrant activation of downstream signaling pathways. The dysregulation of androgen receptor signaling is a key driver of prostate cancer progression and resistance to therapy [48].

1.5 Single-cell RNA-sequencing technology

Recent advances in sequencing technologies have revolutionized the field of genomics and enabled researchers to explore the complexity of biological systems at unprecedented resolutions. One of the most exciting developments in this field is single-cell RNA sequencing

(scRNA-seq) [49], which allows for the high-throughput sequencing of transcriptomes at the single-cell level. This technology has transformed our understanding of cellular heterogeneity and has the potential to uncover previously unknown cell types and states [50]. Traditionally, bulk RNA sequencing has been used to analyze gene expression in populations of cells, providing an average expression profile for a given tissue or cell type. However, this approach masks the inherent heterogeneity of individual cells within the population. By contrast, scRNA-seq provides a powerful tool for characterizing the transcriptional diversity of individual cells, revealing the existence of rare or previously unrecognized cell types.

Single-cell RNA sequencing technology involves the isolation of individual cells, the extraction and amplification of RNA from each cell, and the sequencing of the resulting cDNA libraries. The library preparation step is a critical component of the scRNA-seq workflow and involves converting the extracted RNA into a form that is compatible with sequencing. There are various library preparation methods available, each with its advantages and limitations. Some of the commonly used methods include Smart-seq [51], Smart-seq2 [52], CEL-seq [53], and Drop-seq [54]. These methods differ in their sensitivity, accuracy, and throughput, and researchers must choose the most appropriate method based on the specific research question and available resources.

Sequencing platforms for scRNA-seq have also evolved rapidly in recent years. In the present study, a 10x Genomics Chromium system has been used, mainly because of its high throughput capability, which enables tens of thousands of cells to be processed in a single run, which makes it a suitable platform for large-scale experiments. Briefly, the Chromium system works by partitioning individual cells into nanoliter-scale droplets, where each droplet contains a bead coated with a unique barcode sequence. Within each droplet, the cell's RNA is reverse transcribed, and the resulting cDNA is labeled with the unique barcode sequence, allowing for the identification of the cell of origin during downstream analysis. Following reverse transcription, the droplets are broken, and the barcoded cDNA is pooled for library preparation and sequencing. This approach allows for high-throughput sequencing of

thousands of individual cells, with each cell's transcriptome being represented by a unique barcode [55].

1.6 Mass spectronomy-based proteomics

Proteomics is the large-scale study of proteins and their functions within biological systems. Mass spectrometry-based proteomics has emerged as a powerful tool for the identification and quantification of proteins in complex biological samples. This technology has become an essential tool for investigating the proteome of organisms, tissues, and cells, and has led to a greater understanding of complex biological processes, including disease mechanisms and drug targets [56]. The field of mass spectrometry-based proteomics has undergone rapid advancements in recent years, with the development of new instrumentation, software, and sample preparation methods. These advancements have increased the sensitivity, accuracy, and throughput of proteomics experiments, enabling the detection and quantification of thousands of proteins in a single experiment.

One popular method for protein quantification is isobaric tagging, such as Tandem Mass Tag (TMT) labeling [57]. TMT-labeled mass spectrometry allows for the simultaneous quantification of multiple samples in a single experiment, making it a powerful tool for comparative proteomics. The TMT reagents covalently label peptides in each sample with a unique isotopic tag, allowing for multiplexed analysis of different samples. The resulting peptide mixture is analyzed by mass spectrometry, and the relative abundance of each peptide is determined by the intensity of its corresponding isotopic tag. TMT-labeled mass spectrometry has become a widely used method for the quantification of proteins in complex biological samples and has led to a greater understanding of complex biological processes, including disease mechanisms and drug targets.

2 AIMS OF THE STUDY

Comprehensive genomic studies have delineated key driver mutations linked to disease progression for most cancer types. However, the corresponding changes in the transcriptome remain largely elusive because of the scarcity of data related to the advanced disease for most tumor types and the significant bias associated with the cross-study analysis of data sets. Nevertheless, the assessment of gene expression may provide a complete and quantitative measure of the biological processes related to disease progression.

In the field of prostate cancer, genomic studies have been most notably conducted for primary or metastatic tumors that had progressed under androgen deprivation therapies (ADT) to castration-resistant disease (mCRPC) [58, 59]. Besides, more recent studies have also looked at genetic alterations in a smaller number of advanced prostate cancers that had lost androgen receptor expression, a process frequently associated with the neuroendocrine trans-differentiation [40, 60]. The comparison of mutation frequency across these studies revealed the particular importance of loss-of-function alterations in tumor suppressor genes for promoting disease progression and resistance to both ADT and second-generation androgen receptor signaling inhibitors [61]. Nevertheless, the binary nature of genomic data, the plethora of genetic alterations within the same gene (e.g. gene deletions, damaging point mutations of one or multiple alleles in tumor suppressor genes), and complex interactions across multiple driver genes largely complicate a quantitative assessment of the biologic behavior downstream (e.g. activation of cellular pathways, signaling cascades, patient survival).

In this scenario, the assessment of gene expression may provide a more complete and quantitative measure of the biological processes related to disease progression. Nevertheless, this approach requires the accurate integration of multiple data sets across studies to overcome the issue of introducing dataset-specific features. This has been until now the major problem due to the substantial amount of non-biological bias introduced during the generation and analysis of RNA sequencing data, the so-called “batch effect.” This

is the reason why there are no studies yet that attempt to integrate this data to nominate a trajectory of prostate cancer disease progression.

The project can be subdivided into several aims. The first aim of this project was to overcome the hurdles related to the integration of data from many different sources into a single cohort of high-throughput transcriptional data from 13 heterogeneous studies, constituting a comprehensive compendium of the disease. Indeed, we planned to obtain an extended cohort that may be representative not only of the intrinsic diversity of prostate cancer but also of its dynamic evolution in time.

The second aim of the project was the identification and quantification at the genomic and transcriptomic level of the roadmap to disease progression. To do this, we exploited the transcriptome atlas, to perform a trajectory analysis that allowed us to assign a value of a pseudo-time describing the advancement along the disease path.

We next set out to functionally validate our findings related to disease progression in *in vivo* animal models. In particular, we took advantage of a panel of patient-derived xenografts (PDX), which are models of cancer where the tissues from a patient's tumor are implanted into an immunodeficient mouse [62]. They are models frequently used to create an environment that allows for the natural growth of cancer, its monitoring, and corresponding treatment evaluations for the original patient, and therefore have been extensively used in cancer research. For our purpose, we used five models, originating either from primary prostate cancer (PNPCa), advanced castration-resistant AR-positive prostate tumors (LuCaP-35-78-147), and the most aggressive subtype of prostate cancer, neuroendocrine PCa (LuCap-145) [63, 64], to interrogate their positioning along the main trajectory of disease progression.

As a final phase of the project, we asked if transcriptional changes along the trajectory would also translate into differences in protein abundance levels, and therefore if proteomics data could also be integrated into the cancer progression line. To this aim, we took advantage of a panel composed of the five aforementioned models [64] and will quantify protein expression in such samples, using high coverage isobaric multiplexed-labeling and liquid-

chromatography-mass spectrometry [65, 66]. Given the fact that these models were already profiled for RNA sequencing, we finally aimed at adding a layer of complexity to our progression model, with the integration and correlation of protein data, starting from such *in vivo* models.

3 MATERIALS AND METHODS

3.1 Description of the study population, sample collection and processing

To build an integrated resource of transcriptional features representing all stages of prostate cancer progression, Marco Bolis collected raw sequencing data from a large panel of independent datasets. He gathered raw data for 1223 clinical samples (1104 excluding technical replicates, 1044 excluding multiple metastatic sites derived from the same individual). The resulting integrated cohort is representative of various stages of disease progression, namely, normal prostate specimens (n=174), primary tumors (n=714), castration-resistant prostate cancers (n=316), and castration-resistant prostate cancers showing features of neuroendocrine trans-differentiation (n=19).

Raw sequencing files were retrieved from following sources: 1) Gene Tissue Expression Database (GTEX); 2) The Cancer Genome Atlas (TCGA); 3) Atlas of RNA sequencing profiles of normal human tissues (GSE120795); 4) Integrative epigenetic taxonomy of primary prostate cancer (GSE120741); 5) Prognostic markers in locally advanced lymph node-negative prostate cancer (PRJNA477449); 6) The Long Noncoding RNA Landscape of Neuroendocrine Prostate Cancer and its Clinical Implications (PRJEB21092); 7) Integrative Clinical Sequencing Analysis of Metastatic Castration Resistant Prostate Cancer Reveals a High Frequency of Clinical Actionability (PRJNA283922; dbGaP: phs000915); 8) CSER - Exploring Precision Cancer Medicine for Sarcoma and Rare Cancers (PRJNA223419; dbGaP: phs000673); 9) Molecular Basis of Neuroendocrine Prostate Cancer (PRJNA282856; dbGaP: phs000909); 10) Heterogeneity of Androgen Receptor Splice Variant-7 (AR-V7) Protein Expression and Response to Therapy in Castration Resistant Prostate Cancer (CRPC) (GSE118435); 11) Molecular profiling stratifies diverse phenotypes of treatment-refractory metastatic castration-resistant prostate cancer (PRJNA520923; GEO: GSE126078). Depending on the specific dataset considered, *fastq* files were downloaded either by using GDC-client (TCGA) or sra-toolkit (SRA, dbGaP).

3.1.1 RNA-seq data processing of clinical samples

The overall quality of sequencing reads was evaluated using *FastQC (v.0.11.9)* [67]. No sample had been excluded due to low technical quality. Sequence alignments to the reference human genome (GRCh38) were performed using STAR (Spliced Transcripts Alignment to a Reference) aligner (v.2.6.1c), a very commonly used *'seed-and-extend'* algorithm. To significantly increase sensitivity to novel splice junction compared to the regular single-mapping, the *2-pass mode* was selected. Briefly, in the 2-pass mapping procedure, reads will be mapped twice: in the 1st pass, the novel junctions will be detected and inserted into the genome indices; in the 2nd pass, all reads will be re-mapped using annotated (from the GTF file) and novel (detected in the 1st pass) junctions. In particular, gene expression was quantified at the gene level in the 2nd pass by using the comprehensive annotations made available by *Gencode* (v29 GTF-File).

Strand-specific information was not maintained to avoid technical differences between stranded and unstranded libraries. Samples were adjusted for library size and normalized with the variance stabilizing transformation (vst) in the R statistical environment using DESeq2 (v1.28.1) pipeline [68].

3.1.2 Batch effects correction and Principal Component Analysis

In the process of integrating different datasets from a variety of sources, Marco Bolis and I verified that batch effects did not overwhelm the biological signal. Batch effects may derive not only from biological differences across datasets but also may be the consequence of a different sequencing technique (PolyA+; TotalRNA; Hybrid Capture Sequencing), different reference genomes, or originate from thus far other unknown sources. We aimed at specifically removing technical batches rather than real biological variation and tried to preserve biological differences which may be consequent of a different PSA level, age, tumor grade/stage, or other. Principal component analysis (PCA), by identifying the transcriptional features endowed with the highest variance across samples, is a very useful tool to detect relevant batch effects. When the latter are overwhelming, they are likely to appear among

the top principal components and cluster together samples sharing the same batch effect-related features. A PCA analysis performed on the complete set of 1223 samples (Figure S1B) showed that the largest source of batch effects was associated with the Hybrid Capture Sequencing technique (HCS), while no relevant differences could be associated with the dataset of origin. Only two of the CRPC datasets (phs000915, phs000673) contained samples sequenced using HCS, and for several of these, matched technical replicates sequenced using PolyA+ technology were also available. This allowed us to assess and remove technology-associated bias in gene expression (ComBat, PolyA+ samples set as reference batch).

I further reduced the possibility of confounding biological with technical variation by generating a training subset of our data, consisting of 883 PolyA+ samples (52 Normal prostates, 620 Primary tumors, 193 CRPCs, 19 NEPCs) and determined the top 2000 genes showing the highest amount of variation within the PolyA+ training set only. This way, for PCA representation we avoid the selection of genes that are possibly affected by the sequencing technique, despite the correction we had already performed on the data. Hence, the same 2000 genes were used to generate a PCA plot computed on the extended set of samples. The PCA is routinely generated using the most variable genes detected across the entire dataset. DESeq2's defaults are set to use the top 500 most variables genes only. This number is frequently applied when analyzing the transcription of protein-coding genes. Conversely, in our scenario we evaluated the expression of the comprehensive genomic annotations provided by Gencode, which also includes non-protein-coding genes, reaching a total amount of approx. 60000 genes. Thus, I increased the number of genes used for PCA analysis proportionally to the above-mentioned number (4 x 500 = 2000).

The results depicted in the PCA plot clearly show that the positioning of tumors at the same stages of cancer progression overlap with each other irrespective of the dataset of origin and the sequencing technology. This indicates that the different positioning of normal prostate, primary tumors, CRPCs, and NEPCs is due to a real biological signal and not consequent to an unwanted dataset-specific batch effect.

3.2 Trajectory analysis

Trajectory and pseudotime inference are frequently used in single-cell RNA sequencing data analysis to model developmental trajectories through smooth curves following dimensionality reduction and clustering. Here I applied one of these tools, slingshot [69] (v1.6.0), to infer progression-associated trajectory and pseudotime from our integrated set of bulk-RNA sequencing samples. Marco Bolis and I selected slingshot because of its capability to determine branches along the trajectory, if any. PCA positioning (principal component 1 and principal component 2; PC1-PC2) of the individual samples must be used as input for slingshot, along with the information that the computed trajectory had to start from the samples coming from the Normal tissues (i.e., non-malignant tissue).

Slingshot divides the problem of multiple lineage inference into several steps: first, the algorithm will order each sample into lineages by the construction of a minimum spanning tree, to stably identify the key elements of the global lineage structure, i.e., the number of lineages and where they branch. Paths through the tree are then smoothed by fitting simultaneous principal curves [70] and a sample's pseudo-time value is determined by its projection onto one or more of these curves.

The analysis was performed using 1106 samples, discarding all technical replicates, in order not to overweight some samples and influence the computation of the trajectory. Metastatic lesions from the same individual but localized in different organs were admitted for this analysis. Subsequently, we could associate a pseudotime for each sample, ranging from 0 to 250.

3.3 Differential gene expression analysis

The R package implemented *DESeq2 method [68]* was also used to detect differentially expressed genes. As the first step, the input for the *Deseq2* is defined as the un-normalized counts of sequencing reads obtained from the RNA-sequencing experiment, in the form of a matrix of integer values. The value in the i-th row and the j-th column of the matrix defines

the number of reads unambiguously mapped to gene i in sample j. Additionally, as an important step before starting the analysis, a pre-filter of low-count genes has been carried out, by removing rows in which all the read counts were equal to 0, to both reduce the memory size of the data objects and increase the speed of the transformation and testing functions.

The phases of the differential expression analysis are designed into a single function, *DESeq*. The analysis that it performs is based on the Negative Binomial (Gamma-Poisson) distribution and goes through three different steps: estimation of the size factors, estimation of the dispersion coefficients, and a Negative Binomial Generalized Linear Model (GLM) fitting.

The generalized linear model used in differential expression analysis is on the form:

$$K_{ij} \sim NB(\mu_{ij}, \alpha_i)$$

$$\mu_{ij} = s_j q_{ij}$$

$$\log_2(q_{ij}) = x_j \beta_i$$

where counts K_{ij} for gene i, sample j, are modeled using a Negative Binomial distribution with fitted mean μ_{ij} and a gene-specific dispersion parameter α_i. The fitted mean is composed of a sample-specific size factor s_j and a parameter q_{ij} proportional to the expected true concentration of fragments for sample j. The coefficients β_i give the log2 fold changes for gene i for each column x_j of the model matrix X [59].

The first step performed by the function is the estimation of size factor s_j (with the *median-of-ratios* method described in [71]), to obtain a normalized count matrix, followed by the estimation of dispersion coefficient α_i (e.g. equation (5) in [59]), that defines the relationship between the variance of the counts observed and its mean value, and the negative binomial GLM fitting for β_i, coupled with a Wald statistic.

Using the estimated size factors and dispersion estimates, the function *nbinomWaldTest* tests for significance of coefficients in a negative binomial generalized linear model (GLM). First, standard maximum likelihood estimates for the generalized linear model coefficients (β_i, or log2fold changes) are calculated. To obtain the Wald test p-values, the

coefficients are scaled by their standard errors $SE(\beta_i)$ and then compared to a standard normal distribution.

Notably, the *DESeq2* algorithm performs independent filtering, in order to filter out from the procedure those tests that have no, or little chance of showing significant evidence, before without even looking at their statistics. This allows to increase the detection power while controlling the experiment-wide type I error rate, using a two-stage approach that filters variables by a criterion independent of the statistics, and then tests only those variables that pass the filter. The independent filtering step is based on the mean of normalized counts for each gene. Genes with low mean expression levels across samples are more likely to have low statistical power to detect differential expression, so they are filtered out before hypothesis testing. By doing this, the number of tests performed is reduced, which can increase the power to detect truly differentially expressed genes without increasing the false discovery rate. Indeed, Wald test p-values of the subset of genes that have passed the filtering phase are adjusted using the Benjamini and Hochberg False Discovery Rate [72] procedure. In the end, the adjusted p-values for the genes which do not pass the filter threshold are set to NA. The filter threshold value and the number of rejections at each quantile of the filter statistic are available as metadata of the object returned as result.

3.4 Gene set enrichment analysis

Gene set enrichment analyses were performed using the *limma* package (*Camera*, use. ranks set to TRUE) [73]. Gene-Sets collections were retrieved either from the Molecular Signature Database (MSigDB) [74, 75], or from previous publications (AR/NE-Score) [76].

The *Camera* algorithm is a competitive gene set test that accounts for inter-gene correlation in gene expression data. It works by estimating the correlation between genes in a given gene set using a resampling-based method, and then uses this information to adjust the test statistic, that measures the enrichment of the set for differentially expressed genes relative to genes not in the set.

The inter-gene correlation within a gene set can inflate the statistical test and increase the false positive rate. To tackle this problem, the *Camera* function estimates the variance inflation factor (VIF, e.g. section *Variance inflation under correlation* in [62]) for each gene set, considering the inter-gene correlation present within the set. By adjusting the variance of the test statistic based on the VIF estimation, the accuracy of the final results is improved, effectively accounting for the correlation structure within the gene set. The estimate of the mean pair-wise correlation within each set of genes is implemented in the function *interGeneCorrelation*: the function calculates the mean pair-wise correlation between genes within each gene set based on the expression data. It uses a variance-stabilizing transformation to preprocess the data and estimate the correlation. The resulting estimate of the mean correlation within each gene set is then used to adjust the test statistics for inter-gene correlation. This gives a useful compromise between strict error rate control and interpretable gene set rankings.

For the analysis, I decided to use a slightly modified version of the presented algorithm, called *cameraPR*: this is a "pre-ranked" version of CAMERA where the genes are pre-ranked according to a pre-computed statistic. In this case, the statistical values given to the function arises the significance (p-value) of the statistical correlation parameter calculated through Pearson's procedure presented in the previous paragraph.

As a final result, CAMERA returns a matrix with a row for each gene set tested and a column for each of the following parameters: number of genes in the set, the direction of change ("up" or "down"), a two-tailed p-value and the adjusted p-value. All p-values were corrected for multiple testing using the Benjamini and Hochberg False Discovery Rate procedure, with the significance threshold set to 0.05. Additionally, gene set enrichment analysis significance was logarithmically transformed in the form of $-10\log_{10}$(p-adjusted), with a bold intercept (x=13.01) indicating the FDR threshold depicted in the corresponding plots.

3.5 Correlation of gene expression and pathways to pseudotime

Having defined a unique pseudotime value for each sample, I computed the correlation between pseudotime and mRNA expression for each gene. For this purpose, I used Pearson's correlation over Spearman's because I aimed at identifying the strength of the linear relationship between gene expression and pseudotime. However, to be more robust to outliers, I opted for 10 times repeated leave one-third out procedure. Precisely, I randomly selected 10 subsets composed of 66% of the samples and computed correlation coefficients between pseudotime and expression of each gene in all subsets. Finally, I averaged these values and ranked them according to their correlation coefficient to pseudotime. Subsequently, using this ranking I applied Camera to perform gene-set enrichment analysis procedure (use.ranks = *TRUE*) and determined which gene-set were mostly directly or inversely associated with pseudotime.

3.6 Correlation of mRNA expression and protein abundances

Proteomics data were retrieved from the Proteomics Identifier Database (PRIDE: projects PXD009868, PXD003430, PXD003452, PXD003515, PXD004132, PXD003615, PXD003636). The dataset includes 28 gland-confined prostate tumors and 8 adjacent non-malignant prostate tissue obtained from radical prostatectomy procedures, plus 22 bone metastatic prostate tumors obtained from patients operated to relieve spinal cord compression.

To compute the correlation between mRNA expression and protein abundance I first computed, for each gene, the average fold-change (log2) between CRPC and PRIMARY tumors based on mRNA expression. Then the same was applied to the proteomics data to obtain for each protein a log fold change representing differential abundance between CRPCs and primary tumors. For protein/mRNA correlation purposes, we discarded all genes that had not been evaluated in the proteomic data. Finally, I used Pearson's method to evaluate the strength of correlation and the associated statistical significance.

3.7 Retrieval of genetic information and correlation with progression

Matched genetic information respective to mutations and copy number status could be retrieved for 763 samples through cBioportal [77]. To determine associations between mutations and tumor progression, for each gene I compared the pseudotime of mutant vs wild-type samples, by performing statistical testing using the Wilcoxon-sum rank test. Mutations were ordered according to their False Discovery Rate adjusted p-values and analyses were performed separately in PRIMARY and CRPC+NEPC tumors, to determine the relative contribution of mutations at various stages of disease progression. I only screened for genes being mutated in more than 5 individuals (Supplementary Figure 1M). To determine associations between copy-number alterations and tumor progression, we associated for each gene a value of either -2 (homozygous deletion), -1 (heterozygous deletion), 0 (Wild-Type), 1 (Gain), 2(Amplification) and subsequently computed Pearson's correlation between these values and pseudotime. I restricted this last analysis to genes being frequently deleted or amplified in prostate tumors, namely, *MYC*, *AR*, *RB1*, *PTEN*, and *TP53* (Figure 1E). The above-described analyses were performed discarding technical replicates. Metastatic lesions from the same individual but localized in different organs were admitted for this analysis.

3.8 Quantification of immune infiltrates and correlation with progression

Quantification of immune infiltrates for all samples in our cohort was inferred from transcriptomic data using *CibersortX* [78] by using the default signature matrix "LM22" to deconvolve 22 immune cell subsets from bulk RNA-Seq (Absolute quantification mode). The abundance of inferred immune populations was correlated to pseudotime using the same strategy applied to correlate gene expression and pseudotime. I opted for 10 times repeated leave one-third out procedure. Precisely, I randomly selected 10 subsets composed of 66% of the samples and computed correlation coefficients between pseudotime and each immune population in all subsets. Finally, I averaged these values and ranked them according to their correlation coefficient to pseudotime. Pearson's correlation-associated p-values were corrected for multiple testing using the False Discovery Rate (FDR).

3.9 Cell lines and xenografts models

3.9.1 Cell lines

PC3, DU-145, 22rV1, MDA-PCa-2b, LAPC4, LNCaP, and VCaPcell lines were purchased from ATCC (American Tissue Culture Collection) (Manassas, USA). The LAPC4 cell line was a gift from Prof. Helmut Klocker, and the LNCaP-abl cell line was a gift from Prof. Myles Brown (DFCI, Boston).

3.9.2 Cell culture

PC3, DU-145, 22rV1, LAPC4, and LNCaP cell lines were cultured in RPMI-1640 (21875-034, Life Technologies) supplemented with 10% fetal bovine serum (FBS-11A, Capricorn Scientific), and 1% penicillin/streptomycin (15140-122, Life Technologies) with 5% CO_2 at 37 °C. LAPC4 was also supplemented with 1 nM dihydrotestosterone (DHT).

The LNCaP-abl cell line was cultured in phenol red-free RPMI-1640 (11835063, Life Technologies) containing 10% CSS (FBS, charcoal-stripped, A3382101, Life Technologies) and 1% penicillin/streptomycin with 5% CO_2 at 37 °C.

VCaP cell lines were cultured in Dulbecco's modified Eagle's medium (DMEM) (61965059, Life Technologies) supplemented with 10% FBS and 1% penicillin/streptomycin with 5% CO_2 at 37 °C.

MDA-PCa-2b cell line was cultured in ATCC-formulated F-12K medium (30-2004) supplemented with 20% FBS, 25 ng/ml cholera toxin (C8252, Sigma), 10 ng/ml epidermal growth factor (EGF) (AF-100-15, PeproTech), 0.005 mM phosphoethanolamine (P1348, Sigma), 100 pg/ml hydrocortisone (H0135, Sigma), 45 nM selenium acid (211176, Sigma), 0.005 mg/ml human recombinant insulin (I1884, Sigma), and 1% penicillin/streptomycin with 5% CO_2 at 37 °C.

3.9.3 Animal experiments

All animal experiments were carried out accordingly to the protocol approved by the Swiss Veterinary Authority/Board (TI-42-2018 and TI-10-2010) and received approval from the ethical committee of the Institute of Oncology Research. All in vivo studies used 6–8-week-old male NRG (NOD-*Rag1*^{null} *IL2rg*^{null}, NOD rag gamma) mice.

For the subsequent analysis, we retrieved bulk RNA-seq data of several prostate cancer Xenografts models (i.e., PNPCa; LuCaP-78, LuCaP-23, LuCaP-35, LuCaP-145 [64], LNCAP, and LTL [80]) and their derived 3D cultures.

The LuCaP PDX series has been established by subcutaneous transplantation of tumor tissue of patients with metastatic prostate cancer tumors, from 1991 to 2005. PNPCa originated from a patient who presented with primary PCa (Gleason 9). Orchiectomy was performed directly after biopsy sampling; thus, the tumor was androgen-dependent at the time of collection. For the experiment in castration LNCaP or LuCaP-147 cells were suspended in PBS and 50% Matrigel and subcutaneously injected into the dorsal flanks of the mice (2×10^6 cells/mouse). Tumor growth was recorded using a digital caliper, and tumor volumes were calculated using the formula $(L \times W^2)/2$, where L is the length and W is the width of the tumor. Tumor volume was measured two times per week. When the tumor reached the dimension of 50–100 mm^2, mice were surgically castrated. For the GSK126 treatment, the mice were treated 1 week after castration by daily intraperitoneal injection at a dose of 100 mg/kg for 3 weeks. At the end of the experiment, mice were euthanized, and tumors were explanted and used for molecular assessment.

3.10 Integration of additional bulk RNA-Seq samples and pseudotime inference

Marco Bolis and I developed a method to include new prostate tumor samples in our current analysis by starting from raw counts, which allows the computation of pseudotime progression score and principal components without modifying the original data and plots. Ideally, RNA-Seq should be quantified using the sample genome (hg38) and references used for the current study (Gencode V29). Predictions can be performed sequentially, one sample

at a time. For each new sample, raw counts are merged with the ones composing the full set; the obtained numeric matrix undergoes the same normalization and processing steps up to the computation of the PCA. Here, coordinates may slightly differ from the original ones, due to the addition of a new sample which might exert a small effect on the global re-normalization of all samples. To address this issue, we developed a specific method to include new prostate tumor samples in our current analysis by starting from raw counts, which allows the computation of pseudotime and principal components without modifying the original data and plots. Hence, we apply a machine learning-based approach that generates at runtime three elastic net models, one for each of the top 3 principal components, and train them to predict the error between the original coordinates and ones that are recomputed following the addition of the extra sample of interest. Input features for training the model are the coordinates of the top 100 PCs of the original samples after the normalization with the addition of the new one.

The elastic net is a regularization method in regression analysis, that linearly combines the L^1 and L^2 penalty functions of the Lasso and Ridge regularizations during training [81].

Regression analysis is a statistical method used to explore the relationship between a dependent variable, also known as the outcome variable, and one or more independent variables, also known as predictor variables. The goal is to create a model that can predict the value of the dependent variable based on the values of the predictor variables. When dealing with a large number of predictors or features, as when dealing with gene expression data, there is a risk of overfitting the model to the training data, which can result in poor generalization to new data. Regularization is a technique used to prevent overfitting by adding a penalty term to the regression objective function that discourages the model from relying too heavily on any one predictor variable. In the case of the elastic net, the penalty term is a combination of both the L^1 (lasso) and L^2 (ridge) penalties.

The L^1 penalty (also known as the lasso penalty) encourages sparsity in the model, meaning that it sets some coefficients to exactly zero, effectively removing those predictors

from the model. The L^2 penalty (also known as the ridge penalty) encourages the coefficients to be small, but does not set any of them exactly to zero.

The Elastic net penalty is defined as:

$$\alpha * \lambda 1 * \|\beta\|_1 + (1-\alpha) * \lambda 2 * \|\beta\|_2{}^2$$

Where:

β is the vector of regression coefficients;

$\lambda 1$ and $\lambda 2$ are hyperparameters that control the strength of the L^1 and L^2 penalties, respectively.

$\|\beta\|_1$ and $\|\beta\|_2$ are the L^1 and L^2 norms of β. They are defined as:

$$\|\beta\|_1 = \sum |\beta_i|$$

$$\|\beta\|_2{}^2 = \sum \beta_i{}^2$$

The hyperparameter α controls the mixing between the L^1 and L^2 penalties. When **α = 0**, the penalty reduces to the ridge penalty, and when **α = 1**, the penalty reduces t the lasso penalty. The objective function for the elastic net regression model is then:

$$minimize \sum (y_i - \sum \beta_j * x_{ij})^2 + \alpha * \lambda 1 * \|\beta\|_1 + (1-\alpha) * \lambda 2 * \|\beta\|_2{}^2$$

Where:

y_i is the outcome variable for the i_{th} observation.

x_{ij} is the value of the $\mathbf{j}_{th}$ predictor variable for the $\mathbf{i}_{th}$ observation.

The elastic net penalty encourages the model to select a subset of predictor variables that are most important for predicting the outcome variable, while also shrinking the coefficients of those variables to reduce overfitting.

Hence, we applied the trained model to adjust the computed PC1, PC2 and PC3 coordinates of the extra sample, tuning the model using a Leave-half-out (LHO) cross-validation procedure, to optimize the hyperparameter $\lambda 1, \lambda 2$ and α, considering a resolution between 0.001 and 0.1.

The samples can at this point be added to the PCA plot and pseudotime can be determined.

3.11 Single-cell RNA-sequencing data processing

To perform scRNA-seq PDX tumor tissue, they were dissociated into single cells. After resuspension in PBS, single-cell suspensions were loaded into a 10x Chromium Controller (10x Genomics, Pleasanton, CA, USA), aiming for 10,000–5000 cells, with the Chromium Next GEM Single Cell 3' v3.1 Reagent Kit (PN-1000121, 10x Genomics), according to the manufacturer's instructions. Single-cell RNA sequencing (scRNA-seq) *in vitro* was conducted on the LNCaP cell line. LNCaP cells were maintained for 4 weeks in CSS alone or CSS supplemented with 1 nM DHT or with 1 µM GSK126. We conducted single-cell RNA sequencing (scRNA-seq) *in vivo* on the panel of xenograft models (PNPCa, LuCaP-23, LuCaP-35, LuCaP-78, LuCaP-145) at steady state, and on LuCaP147 xenograft model at different time points to investigate the impact of castration and GSK126 (*EZH2* inhibitor) treatment on gene expression profiles. The LuCaP-147 xenograft models were sampled before castration, 80 days after castration, and 120 days after castration, both with and without a 3-weeks treatment with GSK126.

3.11.1 Quantification of gene expression

Fastq files were generated by demultiplexing raw data using *cellranger mkfastq* (v3.1.0) To make single-cell gene-expression quantification more comparable to those of bulk RNA-Seq, I generated a custom genome with *cellranger mkref*, using the very same reference (GRCh38.p12) and annotations (*encode* v29) used for *STAR* when performing bulk RNA-Sequencing analysis. To discriminate between human and murine cells that may infiltrate the tumors in the in vivo setting, I created a Mouse-Human reference, by creating a hybrid genome (GRCh38.p12+GRCm38.p6) and hybrid gene annotations (*Gencode* v29 and M25, for human and mouse genes respectively). To avoid conflicts, mouse genomic coordinates were preceded by a prefix (i.e., mm_chr1, mm_chr2, etc.). Subsequently, *cellranger count* was used to quantify gene expression in the form of an h5-filtered matrix where Ensembl gene IDs are used as identifiers. To note, Cell Ranger uses *STAR* to perform a splicing-aware alignment of reads to the genome.

3.11.2 Data filtering and clustering

Expression quantification files were imported into R statistical environment using Seurat (v3.1.5) package. I discarded individual cells from our data matrix by using two filtering procedures: first, aiming at detecting transcriptional outliers, second, looking for putative doublets, which were also discarded. Briefly, I computed per-cell quality control metrics using *scatter* (v1.16.1). The total amount of mitochondrial and ribosomal gene expression was quantified for both human and mouse cells. The number of genes being detected per cell, the total amount of reads per cell, and the mitochondrial and ribosomal fraction of the transcriptome were used to determine the skewness-adjusted multivariate outlyingness for each cell (*robustbase* v0.93-6). Outliers were detected by median absolute deviation (MAD) and removed at both tails.

Counts were then normalized (*Seurat::NormalizeData*, method = LogNormalize, scale.factor = 1000) and the top 2000 most variable features were selected (*Seurat::FindVariableFeatures*, method = vst). Data were then scaled (*Seurat::ScaleData*) and principal component analysis was performed up to the top 50 components (*Seurat::RunPCA*). Subsequently, I identified and eliminated putative doublets using *DoubletFinder* (v2.0.3). Having identified outliers and doublets, I removed them from the original count data and went through the pre-processing step again (i.e., normalization, scaling, and pca-reduction). I proceeded to the determination of the k-nearest neighbors of each cell and the construction of a Shared Nearest Neighbor (SNN) Graph (*Seurat::FindNeighbors*), then we identified clusters using the shared nearest neighbor (SNN) modularity optimization-based clustering algorithm (*Seurat::FindClusters,* resolution = 0.5). Finally, I performed Umap dimensionality reduction on the first 10 Principal Components, annotated the previously identified clusters, and generated plots accordingly.

3.12 Identification of Cell-Cycle Phase and Cell-Type

I retrieved the list of cell cycle markers [82] and subdivided it into markers of the G2/M phase or S phase, according to Seurat's annotations. I then used this information to infer the cell cycle phase in our samples (*Seurat::CellCycleScoring*). Murine cells could be clearly distinguished from human cancer cells, because of the intrinsic differences that could be easily spotted thanks to the alignment and quantification performed using a hybrid human-mouse genome. Murine cell types were identified using SingleR (v1.2.4) [83], using ImmGen repository [84].

3.13 Dealing with Drop-out events

Drop-out events are very frequent in the single-cell experiment performed using chromium 10x technology. When drop-out occurs, the absence of data for a particular gene in a specific cell can introduce biases and distort the overall picture of gene expression patterns. This can lead to incorrect interpretations of the data, especially when trying to identify rare cell types or subtle differences between cells. To address these issues, I applied Markov Affinity-based Graph Imputation of Cells (*MAGIC* algorithm, RMagic v2.0.3) [85].

The method works by constructing a network of genes and interactions based on the diffusion of gene expression signals across the network. The diffusion process is modeled using a heat equation, where the gene expression values are treated as heat sources that spread across the network based on their similarity. Specifically, MAGIC utilizes data diffusion to recover an imputed count matrix from an observed count matrix, which represents the likely expression for each individual cell based on data diffusion between similar cells. The approach involves identifying the most similar cells and aggregating gene expression across these cells to impute gene expression that corrects for dropout and other sources of noise. However, since nearest neighbors in the raw data may not necessarily represent the most biologically similar cells due to data sparsity, the method constructs a weighted affinity matrix using data diffusion to represent a more accurate neighborhood of

similar cells. This matrix is then used to restore the data, increasing weights on cells that share similarity across a majority of biological processes with a sufficient number of cells.

3.14 Differential expression analysis and gene-set enrichment

Differential expression was performed between different cell clusters and between clusters subjected to different treatment conditions (Seurat::Findmarkers) using a hurdle model tailored to scRNA-seq data (MAST method [86]). Genes were subsequently ranked for $\log_2$ fold-change, and the CAMERA algorithm (*pre-ranked*) was used to determine gene-set enrichments for each comparison. Cell-specific gene-set enrichments were determined using single-sample GSEA (ssGSEA), computed using gene-expression values of each cell following RMagic imputation.

ssGSEA is a method widely used in gene expression analysis to determine the degree to which a predefined set of genes is coordinately upregulated or downregulated within an individual sample. Unlike traditional gene set enrichment analysis (GSEA), which compares gene expression between two groups or conditions, ssGSEA estimates the enrichment score for each sample individually, making it suitable for analyzing individual samples or small sample sizes. The algorithm calculates the enrichment score for a specific gene set for each sample by ranking the genes based on their expression levels and then comparing the cumulative distribution of the reference gene set with the rest of the genes in the sample.

3.15 Macrophage Reclustering

I could identify a sustained number of murine macrophages infiltrating all xenograft models, except for PNPCa cells. I isolated them and performed a cell-type-specific analysis by repeating all previously described processing steps (i.e., normalization, scaling, and pca-reduction). Dropout events were addressed using RMagic, and cell-specific enrichments were computed using a single sample GSEA.

3.16 Macrophage Polarization Index

The Macrophage Polarization Index, indicating polarization towards M1 or M2 phenotypes was computed for all single-cell RNA samples in our cohort using MacSpectrum [79], specifically designed to allow for the identification and characterization of different macrophage activation states within a heterogeneous mixture of cells.

The calculation of the Macrophage Polarization Index (MPI) takes advantage of a set of signature genes that were established to be differentially expressed between known M1 and M2 polarized macrophage populations. The MPI was calculated by dividing the sum of expression levels of M1 signature genes by the sum of expression levels of both M1 and M2 signature genes in every single cell of the sample. This ratio represents the degree to which the macrophage transcriptome resembles the M1 polarization state versus the M2 polarization state, with a higher MPI indicating a more M1-like polarization state and a lower MPI indicating a more M2-like polarization state.

3.17 Integration of scRNA-seq with bulk-RNA samples, PCA, and pseudotime inference

Single-cell experiments can be easily added together with bulk-RNA experiments into the same PCA analysis by simply summing up together gene counts for all individual cells into one meta-element. This has proven to be extremely comparable in terms of pseudotime inference and PCA positioning, as scRNA-seq and bulk RNA-Seq experiments performed on the same samples are superimposable to each other.

The same applies when dealing with single-cell derived clusters, with the one difference that the number of cells composing each cluster has not to be so critically low that the number of drop-out events results in a matrix composed of too many missing genes. If this is the case, or if just a single cell is to be integrated into the analysis, I suggest running RMagic to deal with the drop-out events, and then simply proceed as previously described.

3.18 Proteomic profiling on xenograft models

To integrate within the inferred progression line mass-spectrometry-based proteomic data, I took advantage of a workflow for multiplexed isobaric labeling and subsequent MS profiling that has been developed at the Proteomics facility of the Broad Institute to serialize proteome and post-translational modifications (PTM) profiling from a multiplexed tissue sample set. This new method integrates previously published methods in combination with recent improvements in MS instrumentation [66, 87, 88].

In a pilot experiment, I could measure in duplicates a total of 11,494 proteins across ten xenograft models ranging from hormone-sensitive primary PCa to late-stage CRPC and NEPC, mainly belonging to the *LuCaP* [64] PDX series (*LuCaP 23.1, LuCaP 35, LuCaP 78, LuCaP 147, LuCaP 145.2*), plus the one hormone naïve PNPCa [63], a previously characterized patient-derived organoid models (MSK-PCa1 [89]) and a small panel of *in house* CRPC xenografts models (*LNCaP, LNCaP-ABL, 22RV1*).

To extend the profiling with more PCa models, a second study was performed, with additional 18-plex isobaric proteome experiments (10,816 proteins quantified). For each 18-plex experiment, I will take advantage of a panel of PDX models made available at the *University of Washington and the Vancouver Prostate Cancer Center*. The new models added, which belong to the *LuCaP* [64]and the LTL [80] PDX series, will largely cover the trajectory of disease progression, deriving either from late-stage primary tumors (*LTL-330, LTL-471, LT-L-467, LTL-508*) or castration-resistant (*LuCaP 23.1CR, LuCaP 35CR, LuCaP96, LuCaP96CR, LuCaP 176*) to neuroendocrine models (*LuCaP 173.2, LuCaP 173.1, LuCaP 49, LuCaP96*).

To compute the correlation between mRNA expression and protein abundance in the panel of xenograft models, for each gene, Pearson's correlation coefficient between the expression across the human samples and the pseudotime. Then the same was applied to the proteomics data to obtain for each protein a correlation coefficient representing differential abundance across tumor progression. For protein/mRNA correlation purposes, I discarded all

genes that had not been evaluated in the proteomic data. Finally, I used Pearson's method to evaluate the strength of correlation and the associated statistical significance.

When dealing with human samples, I first set a threshold of pseudotime of 150, to correctly consider samples covering the same range of progression, both in human data and in the models. Subsequently, I proceed to compute the correlation as described.

4 RESULTS

4.1 Generation of the Prostate Cancer Transcriptome Atlas

To nominate gene expression changes related to disease progression, Marco Bolis re-processed and integrated high-throughput transcriptional data sets from 13 different studies, constituting thus far the most comprehensive compendium of the disease (Supplementary Fig1.A) [40, 46, 59, 61, 90]. The resulting principal component analysis (PCA) showed that samples' position at a given disease stage largely overlapped with another regardless of their origin. In contrast, samples from distinct disease stages differed in localization (Fig.1A). An appreciable "batch effect" related to the hybrid capture sequencing technique was detected and subsequently corrected (Supplementary Fig.1B).

Gene set enrichment analysis (GSEA) of the first two principal components (PC) revealed that PC1 correlated with enhanced proliferation while PC2 anti-correlated with canonical AR-signaling (Supplementary Fig.1C-D). Moreover, PC3 separated cancers harboring truncal mutations in SPOP and FOXA1 from the ones harboring gene fusions involving ETS family transcription factors (Supplementary Fig.1E)[43, 91-93]. Additional PCs accounting individually for less than 4% of the total variance did not reveal any association with tumor cell-specific features. Importantly, the stromal contribution was well represented by PC5 and to a much lesser extent associated with PC1-4 (Supplementary Fig.1F-H,). The latter indicates that the positioning of tissue samples in PC1-4 is only slightly influenced by the tumor purity.

4.2 Trajectory analysis quantifies the path to disease progression

Marco Bolis and I applied trajectory inference analysis to characterize disease progression. The approach identified the path to disease progression and assigned a pseudotime to each sample that describes the advancements along this specific path (Fig.1B).

Because PC3 was mainly influenced by truncal prostate cancer driver mutations, its addition to the trajectory inference analysis did not affect the assigned pseudotime (Supplementary Fig. 1I, J). Subsequently, we assessed corresponding gene expression changes to the initial two-dimensional trajectory (Fig.1C). Among the most up-regulated genes, we noticed key genes encoding for chromatin remodelers, which mediate gene silencing during development, such as DNA methyltransferases (*DNMTs*) and members of the polycomb-repressive-complex-2 (PRC2). Most importantly, the PRC2 member EZH2 emerged as the top up-regulated gene, corroborating its previously suggested role in disease progression (Fig.1C, Supplementary Fig.1K)[94-96]. Besides, among the most up-regulated genes, we noted *AR*-regulated genes that promote G2-M cell cycle progression, while AR-regulated differentiation genes were suppressed, as expected (Fig.1D)[97-99].

The progression path indicates that most prostate cancers evolve from normal tissue by continuously increasing AR-signaling (PC2). Then, under androgen deprivation therapy, the tumors progress to castration-resistant prostate cancer (CRPC) by increasing cell cycle genes and eventually de-differentiate to AR-negative disease with or without neuroendocrine features (NEPC) (Fig.1E). Notably, the transcriptional changes correlated well with the protein level changes in an independent set of primary and CRPC samples (Fig.1F)[100]. Because *EZH2* was not assessed in this dataset, we ascertained its upregulation with disease progression on a tissue microarray of 33 primaries and matched CRPC samples (Supplementary Fig.1L)[101].

Next, Marco Bolis and I evaluated whether genomic alterations in driver genes correlate with disease progression. We noted a significant correlation of point mutations in *PIK3CA*, *TP53*, *FOXA1*, *KMT2C*, and *PTEN* with progression in primary tumors and *FOXA1* in the metastatic counterpart (Fig. 1G). In primary tumors, we also noticed a positive correlation with *MYC* copy number and an inverse correlation with copy number of *RB1*, *PTEN*, and *TP53*, as expected. In contrast, in CRPC/NEPC samples, only RB1 loss seemed to correlate well with increased progression (Fig.1H, Supplementary Fig.1M, N). Next, we wondered if pseudotime would also predict survival in patients with metastatic disease. Indeed, increased pseudotime significantly correlated with overall survival (Fig. 1I). While loss-of-function mutations in RB1

and TP53 were also associated with poor survival, these alterations did not outcompete pseudotime in the multivariate analysis. Hence, pseudotime still reached significance when only RB1 wild-type tumors were considered (Supplementary Fig. 1O, P). The data suggest that pseudotime assessment may be useful to predict patient survival in an advanced disease setting.

Finally, I assessed transcriptional changes in key immune pathways throughout tumor progression along the trajectory. It has been widely appreciated during recent years that cancer growth is supported by changes in the tumor microenvironment, such as the polarization of macrophages from an M1- towards M2-like phenotype[102, 103]. Indeed, I noticed a potent downregulation of pro-inflammatory M1 markers and an increased and continuous shift towards M2-associated pro-tumorigenic effectors (Fig.1C,1J and Supplementary Figure 1Q-S). Interestingly, CD24 – a potent “don’t eat me” signal for M1 macrophages – was associated with progression as well[104].

Figure 1

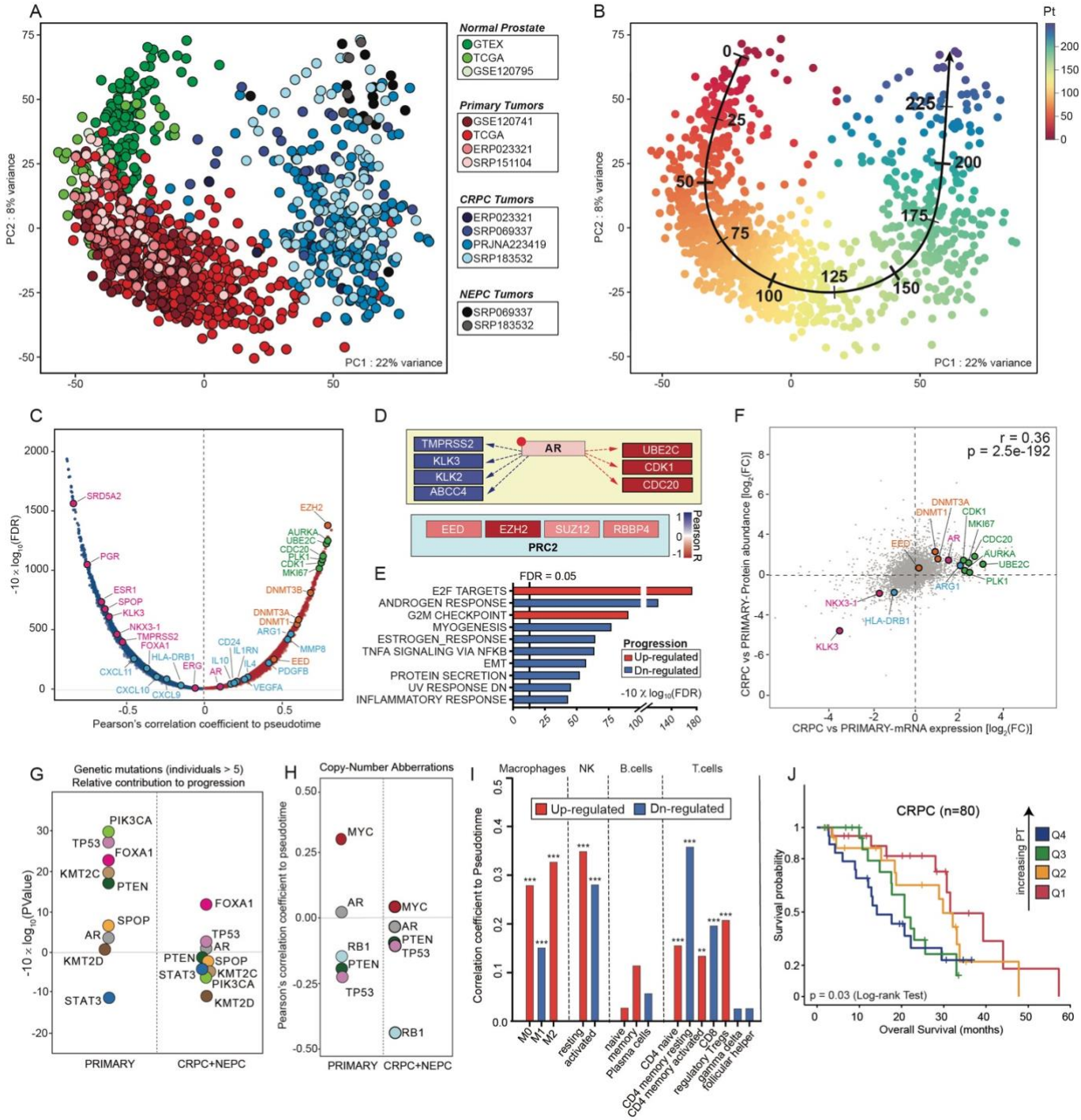


Figure 1 - Trajectory to Prostate Cancer Progression

(A) Principal component analysis (PCA) of pan-prostate cancer transcriptomes obtained from the indicated studies of normal, primary, castration-resistant (CRPC), and neuroendocrine prostate cancer (NEPC). (B) Unbiased trajectory analysis identifies the path to disease progression. Quantification of the path is indicated by inferred pseudotime. (C) Plot representing the correlation between mRNAs and pseudotime inferred along the trajectory. Polycomb repressive complex related genes are highlighted in orange, cell cycle-related genes in green, immune response in blue, and AR signaling in red. X-axis: Pearson's correlation coefficient between mRNAs and

pseudotime; Y-axis: The associated significance adjusted for False Discovery Rate (FDR) and expressed in the form of -10*x*log10(FDR). (D) Schematic representation of gene expression changes in AR-regulated target genes related to cell differentiation and proliferation and PRC2 components along the trajectory. Correlation coefficients between mRNA expression and pseudotime are depicted in a three-color scale (blue: -1; white: 0; red: +1). (E) Gene set enrichment analysis was performed on genes ranked for their Pearson's coefficient as determined by the correlation between mRNA expression and pseudotime inferred from the trajectory. Increasing pseudotime results in an increase of cell cycle-related genes and concomitant down-regulation of androgen-responsive genes. (F) Scatterplot revealing correlation between mRNAs and protein abundances, expressed in the form of fold-change (log-scale) between CRPCs and Primary tumors. (G) Pearson's coefficients, as determined from the correlation between somatic mutations (0:wild-type; 1:non-synonymous mutation) and inferred pseudotime along the trajectory. To dissect the relative impact on disease progression at different stages, coefficients were computed separately in primary and CRPC/NEPC samples. The analysis was performed only for genes mutated at least in 6 individuals. Significance level (p-values): * < 0.05, ** < 0.01, *** < 0.001. (H) Computed Pearson's correlation between samples' numeric copy number status (-2: homozygous deletion; -1: heterozygous deletion; 0: wild-type; 1:gain; 2:amplification) and inferred pseudotime, stratified for primary and metastatic tumors (CRPC, NEPC). (I) Histograms depicting the correlation between the inferred abundance of the indicated immune cell populations (as determined by Cibersortx) and pseudotime. P-values associated with Pearson's correlation coefficients were adjusted for multiple testing using the false discovery rate (FDR). (J) Kaplan-Meier curve for disease-free survival related to pseudotime using a 4-tiered scoring system (quartiles) reveals a significant association of higher pseudotime with impaired survival. Significance level (p-values): * < 0.05, ** < 0.01, *** < 0.001.

4.3 Integration of prostate cancer models in the transcriptome analysis

I next set out to further functionally validate our findings related to disease progression in eight established human prostate cancer cell lines and six patient-derived xenografts (PDX) models originating either from a surgically removed primary prostate cancer (PNPCa)[63] or CRPC (LuCaP-23.1, -35, -78, -145, -147)[64]. To this end, the transcriptional fingerprint of all models clustered towards the outer layer of the progression trajectory (Fig. 2A and Supplementary Figure 2A, B).

As expected, the PCA positioning of cell lines and the PDX models along the trajectory was highly significantly associated with the originating disease stage and the dependence on androgens (Supplementary Figure 2C). The hormone-naive PNPCa model was placed first, followed by the CRPC-derived models, positioned progressively according to their decreasing levels of AR dependency. Finally, I observed that the AR-negative (PC3, DU145) and neuroendocrine models (NCI-H660, LuCaP-145.2), were located at the end of the route (Fig. 2A and Supplementary Figure 2A, B). As expected, I also noted a corresponding upregulation of key proteins related to polycomb complexes (*EZH2*, *SUZ12*, *EED*), DNA methylation (*DNMT1*, *DNMT3A*/*B*), and G2-M cell cycle progression (Fig. 2B).

Multiple castration-resistant sublines of cell lines and PDX models have been generated over the last decades, enabling us to further functionally validate the disease progression trajectory in an isogenic system [105]. Indeed, I found that all sublines progressed on the trajectory (Fig. 2C and Supplementary Figure 2D-F). Most notably, the LTL-331 PDX model displayed a gradual transcriptional progression from late-stage primary prostate cancer to AR-negative, neuroendocrine disease within a timeframe of 32 weeks (Fig. 2C and Supplementary Figure 2D)[106]. At the molecular level, I also noted an increase in key proteins linked to the trajectory in LNCaP xenograft tumors upon tumor recurrence after castration (Fig. 2D). Altogether, the data suggest that progression along the trajectory can be recapitulated in human cell line and PDX models.

The *ex vivo* culture of prostate cancer cells has been traditionally a major challenge. That said, the adjustment of the 3D organoid culture system for prostate cancer has enabled

the *ex vivo* culture of PDX-derived cells and the generation of new prostate cancer organoid lines[89, 107]. I wondered if the transcriptional output of *ex vivo* cultures would mirror the corresponding PDX models *in vivo*. In general, I found that *ex vivo* organoid cultures displayed a more progressed transcriptional output compared to the corresponding *in vivo* models (Fig. 2E). In agreement, the AR-dependency was also largely diminished (Fig. 2F and Supplementary Figure 2C). This observation could be further validated when androgen-dependent LNCaP cells in standard 2D were cultured in the 3D organoid condition (Fig. 2F). Of note, the standard 2D culture matched better the corresponding xenograft model concerning the position on the progression trajectory (Fig. 2E). In aggregate, the data may suggest that the advances in culturing prostate cancer cells using the organoid system may come at the expense of transformation towards a more progressed and aggressive androgen-independent state.

Figure 2

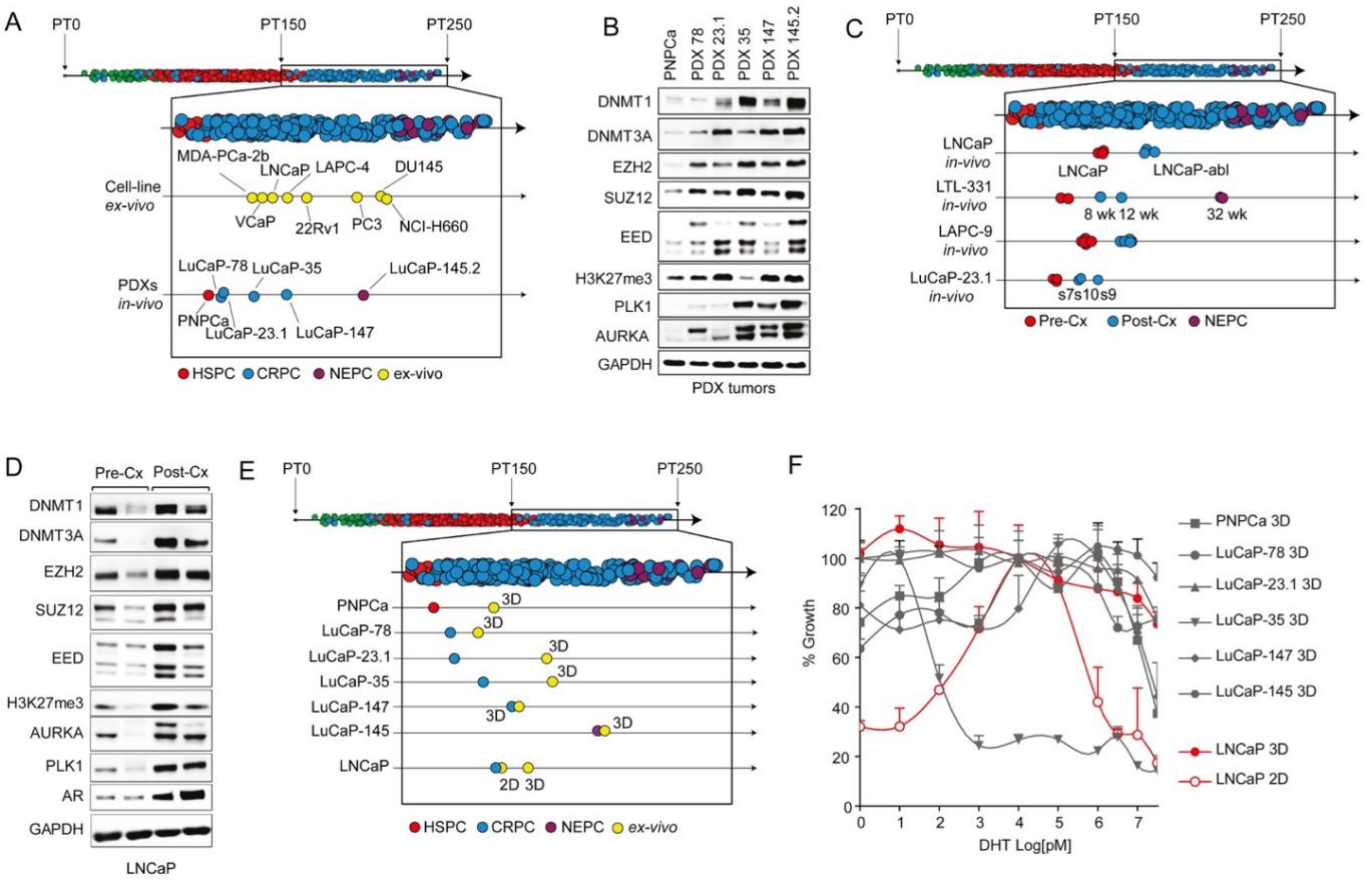


Figure 2 - Mapping of Human Prostate Cancer Models to the Trajectory

(A) Projection of the indicated human cell lines ex vivo and patient-derived xenograft (PDX) models in vivo to the trajectory (PT = pseudotime). (B) Immunoblotting analysis of the indicated proteins across PDX models indicates an upregulation of polycomb repressive complex 2 members and G2M cell cycle checkpoint genes. (C) Androgen-dependent xenograft models progress along the trajectory when developing castration resistance. (D) Corresponding immunoblot analysis of LNCaP xenograft models shows upregulation of AR, polycomb repressive complex 2 members, and G2M cell cycle checkpoint genes upon recurrence after castration (Cx). (E) 3D ex vivo cultures in Matrigel of the indicated PDX and the xenografted LNCaP cells show higher PT than their in vivo counterparts. (F) Corresponding dihydrotestosterone (DHT) dose-response curves of the indicated models in 3D using Matrigel versus standard 2D culture. In 3D conditions, DHT-dependency is largely abolished. (n= 3 independent experiments per condition) See also Supplementary Figure 2.

4.4 Single-cell resolution to the trajectory

Single-cell RNA sequencing (scRNA-seq) was performed on most aforementioned PDX models *in vivo* to interrogate the individual cells' distribution along the trajectory of disease progression. In each case, normal mouse stromal cells were identified and separated from human tumor cells (Fig. 3A, Supplementary Figure 3A-D). When comparing the merged single-cell data with the previously generated bulk RNA sequencing data, I noticed in each case an excellent concordance between the position of both data points on the PCA plot, suggesting that our single-cell data is sufficiently similar to allow the integration into the pan-prostate cancer transcriptome cohort (Fig. 3B, Supplementary Figure 3E-H).

Subsequently, I interrogated each PDX for the existence of separate subpopulations using the *Seurat* workflow [108] (Fig. 3A, Supplementary Figure 3A-D) and integrated the data into the PCA plot (see Method section). Overall, single cells of the various subpopulations within a given PDX model did not greatly differ in their position to the trajectory and displayed relatively little overlap across PDX models (Fig. 3B, Supplementary Figure 3E-L). As expected, subpopulations in cell cycle progression (i.e., S and G2M phase) are positioned higher on the trajectory (Fig. 3B, Supplementary Figure 3E-P). That said, the PDX model LuCaP-35 showed a wider distribution of subpopulations along the trajectory with distinct features linked to the S and G2M phase (H1-3 versus H4, 6), respectively, raising the possibility of being composed of two major, biologically diverse tumor clones (Supplementary Figure 3G, K, O).

Subsequently, I assessed if and how these subpopulations would evolve during the progression to androgen independence. For this purpose, I took advantage of the LuCaP-147 PDX tumor model that quickly develops castration resistance and compared the single-cell transcriptional profiles before and after castration (Fig. 3C). Upon regrowth, there was no major difference in the position and abundance of previously identified subpopulations (Fig. 3D). Instead, I noticed a concordant shift along the trajectory for each of the clusters h1-7, which was characterized by a shutdown of canonical AR signaling and upregulation of pro-proliferative MYC target genes, among others (Fig. 3E, F). Altogether, the data suggest that

resistance to castration in this setting occurs likely through reprogramming of the entire tumor cell population instead of a clonal selection of a particular cluster.

Subsequently, I wondered if the induction of resistance may be paralleled by changes in the tumor microenvironment. Indeed, after castration, an increase was observed in the abundance of tumor-associated macrophages that displayed a change in polarization from M1- to M2-like features (Fig. 3G, H). In line with this, I also observed a gradual reduction of TNF alpha signaling and inflammatory signatures – key features of M1 macrophages – in PDX models with increasing pseudotime along the trajectory (Fig. 3I). The results agree with the expression changes of M1- and M2-related transcripts along the trajectory of disease progression described earlier in Figure 1. Taken together, the data illustrates how bulk transcriptional changes related to disease progression can help to shed light on the emergence of androgen-independent prostate cancer at the single-cell level.

Figure 3

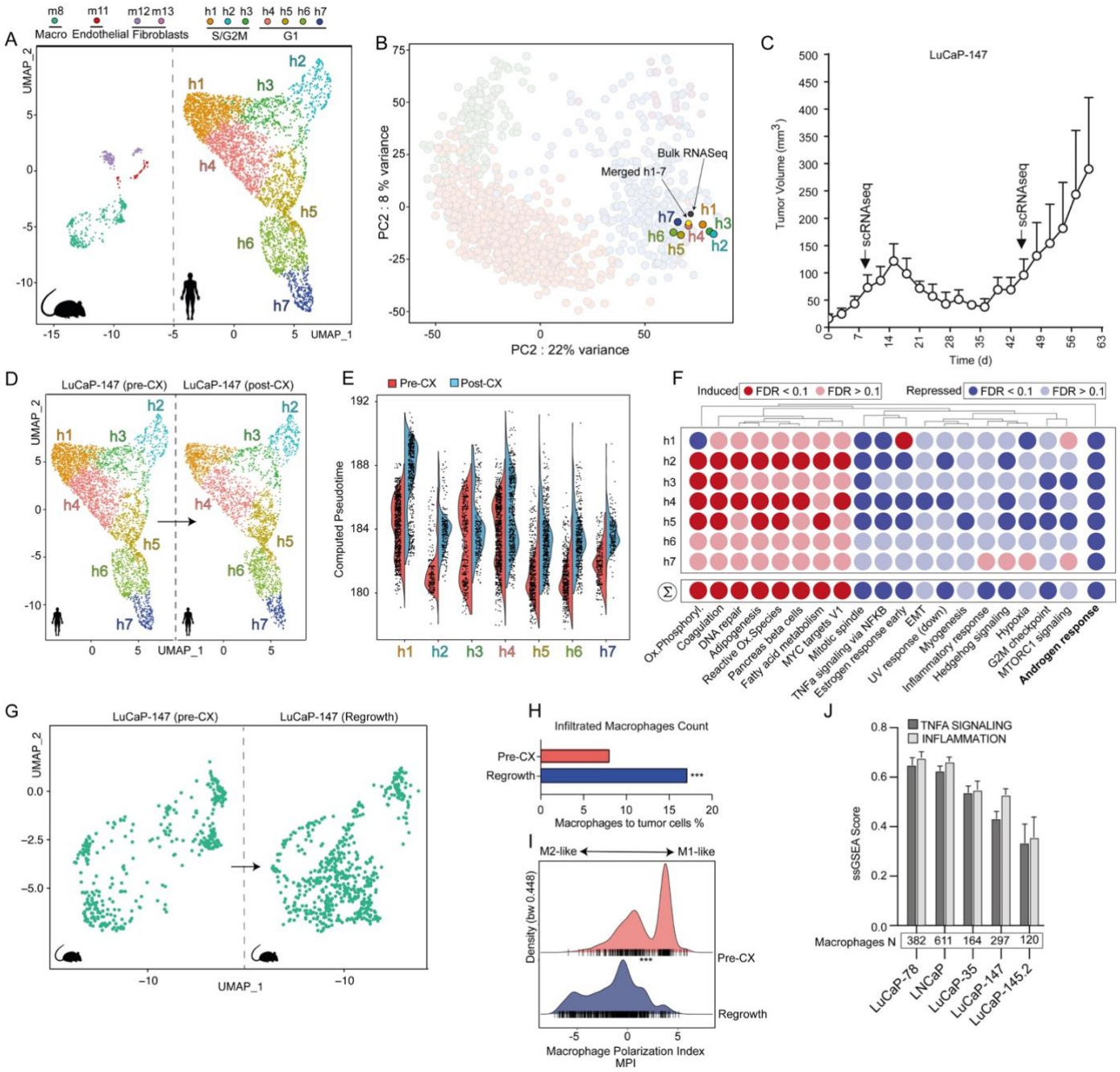


Figure 3 - Single-Cell Resolution to the Trajectory

(A) Dimensionality reduction of single-cell distribution of LuCaP-147 PDX model in vivo using Uniform Manifold Approximation and Projection (UMAP) and subsequent identification of cell clusters performed using Seurat [108] workflow. Human (right) and mouse cells (left) are separated from each other. A total amount of 7 and 4 clusters could be identified for human and mouse cells, respectively. For the latter, I indicated the cells of origin corresponding to the various clusters on top. Inference of cell types was performed with SingleR [83] through the exploitation of the ImmGen repository [84]. For human cell clusters, I indicated the inferred cell-cycle phase as

predicted using Seurat. (B) Projection of single-cell clusters on the PCA plot. The position of merged single-cell data corresponds to the one from bulk RNA sequencing data. Please refer to Star Methods for detailed information on scRNA-seq data integration with bulk RNA-Seq. (C) LuCaP-147 xenografts regress and regrow within 4 weeks after castration. (D) Comparison of tumor single-cell clusters before (left) and after castration (right). (E) The violin plot shows an increase in the pseudotime of individual cells within the cell clusters after castration. The pseudotime inference was performed for each cell following the imputation of missing genes using RMagic Field [85] to deal with drop-out events. (F) Gene-sets perturbed in LuCaP-147 xenografts' single-cell clusters at regrowth (post-castration) compared to pre-castration. Most hallmark gene sets are up-or down-regulated similarly. A marked downregulation of AR-responsive genes is noted. Differential expression for each cluster denoting the transcriptional changes occurring after castration was determined using the MAST algorithm [86]. Subsequently, I determined the gene set enrichments using Camera (pre-ranked) [73]. (G) Dimensionality reduction (UMAP) of murine macrophages before (left) and post-castration (right) highlights a notable increase in macrophage count at regrowth. (H-I) After castration, the percentage of infiltrated macrophage to tumor cell ratio not only increases (H) but also displays more M2-like transcriptional features according to the Macrophage Polarization Index (I), as determined by using MacSpectrum [79]. Significance levels (p-values) were determined using Wilcoxon rank-sum text: * < 0.05, ** < 0.01, *** < 0.001. (J) Single samples gene-set enrichment analysis of inflammation-related pathways performed following reclustering of murine macrophages extracted from the corresponding single-cell RNA-Seq experiments. Missing gene-expression values (dropout events) for each cell were imputed using RMagic. With increasing pseudotime along the trajectory, macrophages of xenograft models display less active TNFA and inflammatory signaling. PNPCa xenografts were excluded from the analysis because of the limited number of infiltrated macrophages. See also Supplementary Figure 3.

4.5 Co-targeting AR and EZH2 delays tumor progression

Because EZH2 emerged as a top-upregulated transcript within the trajectory of disease progression and had been shown to promote androgen independence [94, 96, 109, 110], I set out to investigate if co-targeting *AR* and *EZH2* may prevent or substantially delay disease progression. Indeed, I noted a dramatic change in the transcriptional output program of LNCaP cells when treated with the *EZH2* inhibitor GSK126 under androgen-deprived culture conditions in charcoal-stripped serum (CSS) (Fig. 4A). Previously detected LNCaP subpopulations (h1-6, h8) formed a new subpopulation (h7), suggesting a nearly complete rewiring of transcription, up-regulation of AR signaling, reduction of E2F-related cell cycle genes, and reversion of progression on the trajectory (Fig. 4B and Supplementary Figure 4A-D). In line with this, I noticed a strong reduction in colony formation when androgen-dependent LNCaP, VCaP, and LAPC4 cells were subjected to CCS and treated with GSK126, while forced expression of *EZH2* was sufficient to promote colony formation in the same setting (Supplementary Figure 4E).

Next, I tested whether my observations would translate into an *in vivo* setting. For this purpose, LNCaP cells were injected into the flank of immune-compromised mice and treated the emerging xenograft tumors with castration alone or in combination with three weeks of GSK126. In both cases, the tumors fully regressed. While the tumors of castrated mice regrew with a latency of around four weeks, GSK126 co-treated tumors took more than twice as much time to re-initiate tumor growth (Fig. 4C).

Subsequently, scRNA-sequencing was performed on the tumors pre- and post-castration to investigate transcriptional changes in tumor and stromal cell subpopulations. As noted previously for LuCaP-147, I found no major change in the tumor cell subpopulations (i.e., h1-6) that adapted to castration (Fig. 4D). Because GSK126 treatment *in vivo* had been stopped for three months before harvesting the tumors, the transcriptional changes in the tumor cells appeared less striking than in the aforementioned cell culture setting (Fig. 4A, D). That said, I observed after GSK126 treatment a continuous relative increase in tumor cell numbers of cluster h6 – the least progressed cluster on the trajectory that also displayed the

highest AR mRNA levels (Fig. 4D and Supplementary Figure 4F-I). This cluster showed a further increase in AR signaling and a reversion of disease progression after GSK126 treatment (Fig. 4E and Supplementary Figure 4I). Importantly, xenograft-associated macrophages also continuously increased in numbers and displayed a shift towards M2-like polarization in tumors adapted to castration as previously observed (Fig. 4D, F, G). Strikingly, I found a pronounced relative reduction of preferentially M2-like macrophages in GSK126-pretreated tumors, suggesting that GSK126-mediated changes on the tumor microenvironment may have contributed as well to the delayed regrowth of LNCaP xenografts (Fig. 4H).

In aggregate, the data suggest a rationale for joint targeting of AR and EZH2 in prostate cancer because the latter reverts tumor cell progression towards a more androgen-dependent state and at the same time counteracts adaptive changes in macrophages and fibroblasts that are intimately linked to disease progression.

Figure 4

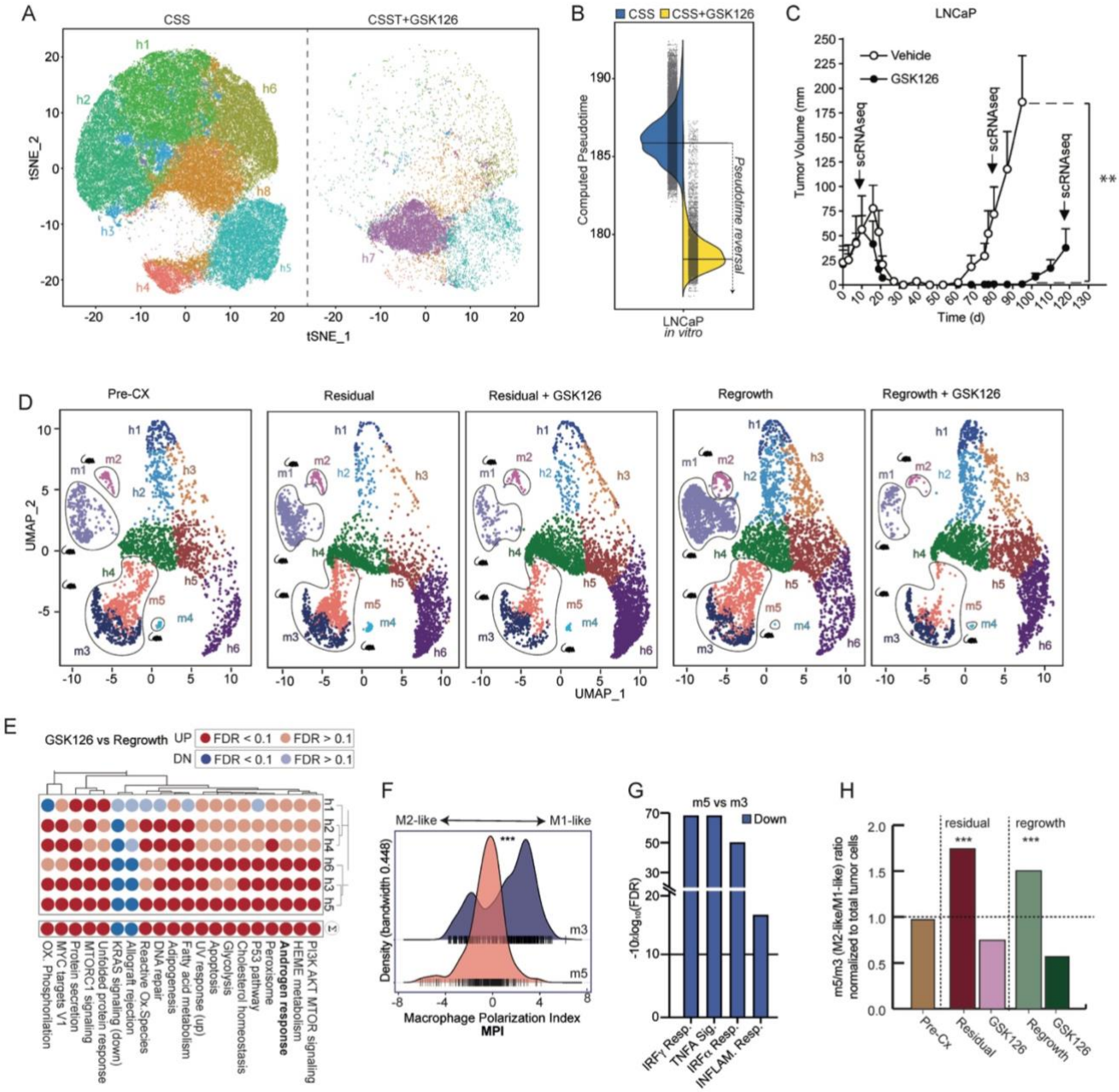


Figure 4 - EZH2 Inhibition Cooperates with Castration

(A) Dimensionality reduction (TSNE) of single-cell RNASeq performed on LNCaP cells cultured in vitro with charcoal-stripped serum (CSS) in the presence (right) or absence (left) of the EZH2 inhibitor GSK126. Identification of cell clusters (h1-h8) was performed using the Seurat workflow. EZH2 inhibition has a dramatic impact on LNCaP cells, as most of the clusters disappear, while the remaining cells undergo such deep transcriptional modifications that give rise to a novel cluster (h7). (B) Pseudotime of individual LNCaP cultured in charcoal-stripped serum (CSS) is significantly reduced upon GSK126 treatment. Pseudotime was computed for each cell, following the imputation of missing genes (drop-outs) using RMagic. (C) GSK126 treatment for 3 weeks upon castration significantly delays the regrowth of LNCaP xenografts after castration. (D) Dimensionality Reduction (UMAP) of LNCaP xenografts

performed on scRNA-seq experiments derived from mice before castration (left), 80 days after castration (left-center), 80 days after concomitant castration and EZH2 inhibition with GSK126 (center), 120 days after castration (right-center), and 120 days after concomitant castration and EZH2 inhibition with GSK126 (right). Murine cells can be subdivided into 5 clusters corresponding to different cell populations according to SingleR (m1:fibroblasts; m2:endothelial cells; m3,m5:macrophages; m4:monocytes). Human malignant cells can be separated into 6 clusters. An increase in the relative number of cells in cluster h6 and a concomitant reduction of murine macrophages following EZH2 inhibition is observed. (E) Upon GSK126 pre-treatment, for each cluster, I determined differentially expressed genes (MAST algorithm) and performed gene set enrichment using Camera (pre-ranked). Results highlight a global increase in androgen-responsive genes. (F) The density plot of macrophage polarization index (MPI) reveals that macrophage cluster m5 (which decreases upon GSK126 administration) shows M2-like transcriptional features, while cluster m3 corresponds to an increased M1-like polarization as determined by MacSpectrum. (G) Differential expression (MAST algorithm) shows that M1-like inflammatory signaling pathways are downregulated in m5 compared to the m3 cluster. (H) Histogram representing the relative proportion between m5- and m3-cluster before castration, 80 days after castration (Residual) and 120 days after castration (Regrowth). Significance level (p-values): * < 0.05, ** < 0.01, *** < 0.001. See also Supplementary Figure 4.

4.6 Proteomic profiling of the panel of xenografts models

This study employed a panel of xenograft models that comprehensively represented the spectrum of prostate cancer progression, ranging from hormone-sensitive to castration-resistant and neuroendocrine prostate cancer, that had been transcriptomic profiled and integrated into the progression line (Fig. 2A and Supplementary Figure B), with the procedure described in the corresponding section (*3.17 Methods*). In a *pilot study*, those xenografts models had been profiled using a unique workflow established at the Proteomic Platform of the Broad Institute (Cambridge, USA) that enables the protein and PTM profiling from the same sample set by isobaric multiplexed-labeling and liquid-chromatography-mass spectrometry (LC-MS/MS). Ten models and a common reference have been profiled using TMT-10/11 plexes, in duplicates. In addition to those models, in a subsequent experiment, additional xenografts have been profiled, using a unique platform that allows to similarly profile 18 samples at once (TMT-18 plex, in duplicates). I visually depicted the samples belonging to each experiment in both 2D and mono-dimensional plots, integrated into the Transcriptomic Atlas (Fig. 5A, B and Fig. 5C).

By using principal component analysis based on protein abundance, I successfully clustered the models based on their level of progression, indicating that this level of quantification can effectively distinguish between the various stages of prostate cancer development, in both the 11- and 18-plex experiment (Fig. 5D,E). Additionally, In the 11-plex study, I discovered a robust correlation (Pearson's correlation coefficient) between PC1 and the pseudotime computed for each model, providing further evidence for the effectiveness of this approach (Fig. 5F).

Before proceeding with the subsequent analysis, I investigate the overlap in proteins quantified between the 18-plex and 11-plex experiments, and notably the two were found to be substantially superimposable, with more the 75% of concordance (Supplementary Figure 5A). Indeed, PCA analysis based on protein abundance in the 18-plex experiment effectively clustered the models based on their pseudotime as well (Supplementary Figure 1B).

4.7 Correlation between Protein abundance and Pseudotime

To examine the relationship between protein abundance and pseudotime, I conducted correlation analysis for each replicate of the mass spectrometry data. My findings revealed that protein expression changes along the disease trajectory correlated well with the corresponding gene expression changes. As expected, PRC2 complex member EZH2, SUZ12, and EED were among the proteins most associated with pseudotime, together with genes belonging to cell proliferation and mitotic spindle formation (AURKA, PLK1, MKI67) and DNA methyltransferases (DNMT1 and DNMT3A), consistently with the previously shown transcriptomic data in human samples (Fig. 5G, Supplementary Figure 5C).

In contrast, I observed a negative modulation of genes related to androgen receptor (AR) signaling within tumor progression. Indeed, my observation of a negative correlation between genes related to androgen receptor (AR) signaling and pseudotime is in line with the well-established phenomenon of progressive acquisition of androgen independence observed in advanced prostate cancers: as tumors progress and become more advanced, they often become less dependent on androgen. These findings were further supported by a gene set enrichment analysis (Supplementary figure 5D).

4.8 Correlation between mRNA expression and protein abundance

All the xenograft models used in this study had been previously profiled for RNA-sequencing [61, 111] (Supplementary Figure 5E). Therefore, I could compare the correlation of mRNA expression and pseudotime to that of protein abundance and pseudotime, discovering a strong correlation between the two measures ($R = 0.58$, p-val < 0.001) (Supplementary Figure 5F). Subsequent GSEA analysis highlighted that pathways positively correlated between mRNA expression and protein abundance were primarily related to cell proliferation (such as E2F targets and G2M checkpoint), while negatively correlated pathways were primarily related to AR signaling and fatty acid metabolism (Fig. 5G). This is in line with the evidence that modulation of lipid metabolism during cancer development and

progression is one of the hallmarks of cancer in solid tumors and is directly linked to AR signaling modulation [112].

Subsequently, I aimed at finding outliers in the correlation, specifically focusing on proteins that showed a positive correlation at the protein level but did not show any corresponding increase in mRNA expression, and *vice versa* (Supplementary Figure 5G, *left*). Unfortunately, the analysis did not identify any significant outliers meeting our predetermined threshold of a correlation, when coming to GSEA (Supplementary Figure 5G, *right*). This suggests that beyond the level of expression of individual genes or proteins, there is a substantial coherence in the directional flow of pathways and signaling.

4.9 Relationship between human and xenografts data

Next, I set to compare the transcriptomic profile of our xenograft model from the 18-plex experiment, with data coming from patients. To achieve this, I set a threshold of 150 in pseudotime value for patient samples, as the real size of the path described with our longitudinal models (Supplementary Figure 5H, *left*).

This analysis revealed a strong correlation between the transcriptomic profiles of our xenograft models and patient samples. Specifically, I observed a Pearson's correlation coefficient of 0.6 between the two datasets (P-value < 0.001), indicating a robust association between the molecular mechanisms underlying prostate cancer progression in our xenograft models and those observed in patients (Supplementary Figure 5H, *right*). Gene set enrichment analysis revealed that a lot of signaling pathways were significantly modulated throughout the progression from late-stage primary to NEPC, both in human and xenograft data: among the most significantly up-regulated we can find gene set related to epithelial to mesenchymal transition, interferon signaling, and inflammation, whereas among the top down-regulated signaling AR response, MTORC signaling and cholesterol homeostasis are the most regulated (Fig. 5I).

These findings suggest that our xenograft models accurately recapitulate key aspects of prostate cancer progression, providing a valuable platform for studying the disease, with accurate recapitulating of the key aspects of prostate cancer progression.

Figure 5

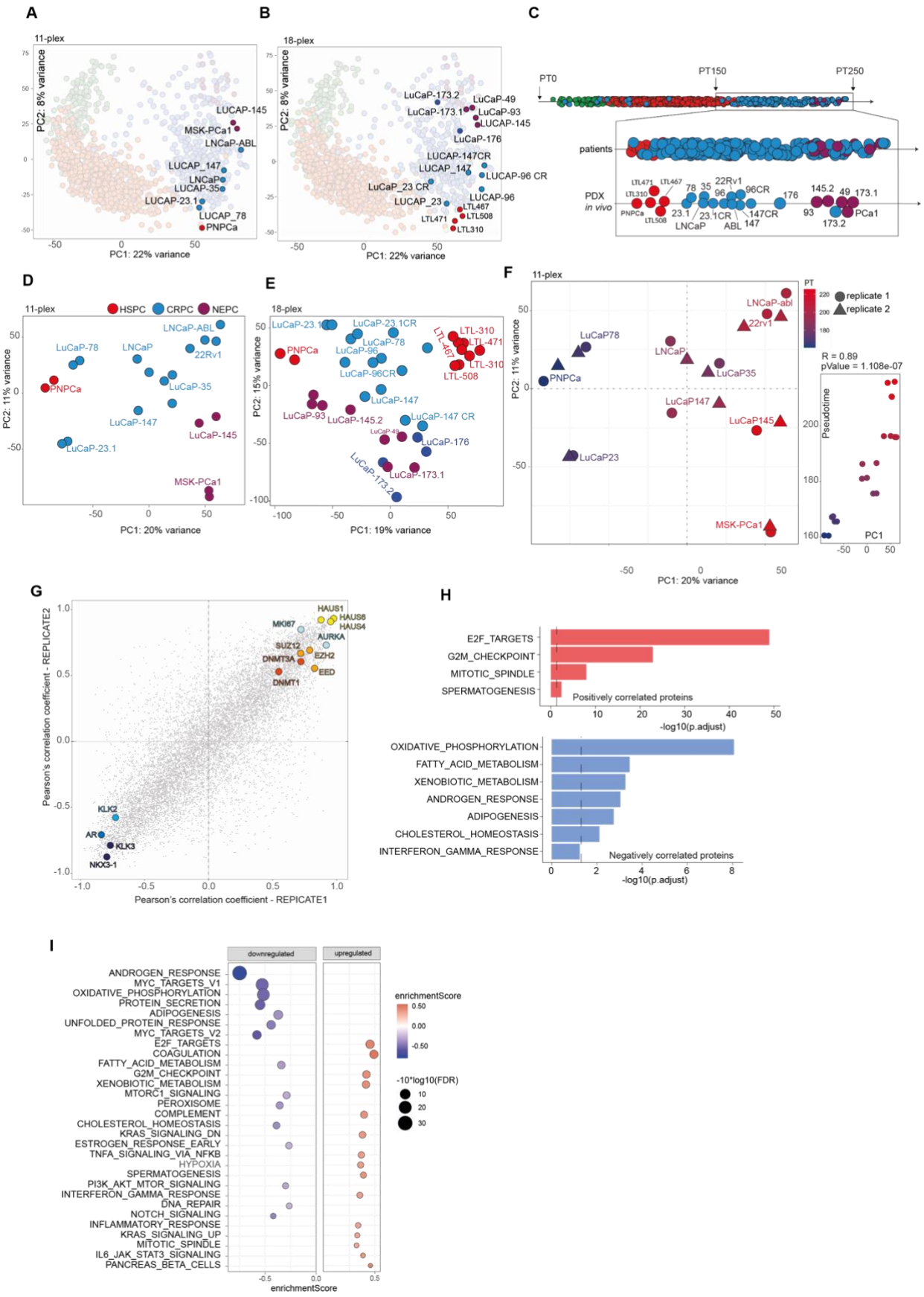

A
11-plex
PC2: 8% variance
PC1: 22% variance
LUCAP-145
MSK-PCa1
LNCaP-ABL
LUCAP_147
LNCaP
LUCAP-35
LUCAP-23.1
LUCAP_78
PNPCa
B
18-plex
LuCaP-173.2
LuCaP-49
LuCaP-173.1
LuCaP-93
LUCAP-145
LuCaP-176
LUCAP-147CR
LUCAP_147
LUCAP-96 CR
LuCaP_23 CR
LUCAP_23
LUCAP-96
LTL467
LTL508
LTL471
LTL310
C
PT0
PT150
PT250
patients
PDX
in vivo
D
11-plex
HSPC
CRPC
NEPC
PC2: 11% variance
PC1: 20% variance
E
18-plex
PC2: 15% variance
PC1: 19% variance
F
11-plex
PC2: 11% variance
PC1: 20% variance
replicate 1
replicate 2
R = 0.89
pValue = 1.108e-07
Pseudotime
PC1
G
Pearson's correlation coefficient - REPLICATE2
Pearson's correlation coefficient - REPICATE1
HAUS1
HAUS6
HAUS4
MKI67
AURKA
SUZ12
EZH2
DNMT3A
EED
DNMT1
KLK2
AR
KLK3
NKX3-1
H
E2F_TARGETS
G2M_CHECKPOINT
MITOTIC_SPINDLE
SPERMATOGENESIS
Positively correlated proteins
-log10(p.adjust)
OXIDATIVE_PHOSPHORYLATION
FATTY_ACID_METABOLISM
XENOBIOTIC_METABOLISM
ANDROGEN_RESPONSE
ADIPOGENESIS
CHOLESTEROL_HOMEOSTASIS
INTERFERON_GAMMA_RESPONSE
Negatively correlated proteins
-log10(p.adjust)
I
downregulated
upregulated
ANDROGEN_RESPONSE
MYC_TARGETS_V1
OXIDATIVE_PHOSPHORYLATION
PROTEIN_SECRETION
ADIPOGENESIS
UNFOLDED_PROTEIN_RESPONSE
MYC_TARGETS_V2
E2F_TARGETS
COAGULATION
FATTY_ACID_METABOLISM
G2M_CHECKPOINT
XENOBIOTIC_METABOLISM
MTORC1_SIGNALING
PEROXISOME
COMPLEMENT
CHOLESTEROL_HOMEOSTASIS
KRAS_SIGNALING_DN
ESTROGEN_RESPONSE_EARLY
TNFA_SIGNALING_VIA_NFKB
HYPOXIA
SPERMATOGENESIS
PI3K_AKT_MTOR_SIGNALING
INTERFERON_GAMMA_RESPONSE
DNA_REPAIR
NOTCH_SIGNALING
INFLAMMATORY_RESPONSE
KRAS_SIGNALING_UP
MITOTIC_SPINDLE
IL6_JAK_STAT3_SIGNALING
PANCREAS_BETA_CELLS
enrichmentScore
-10*log10(FDR)

Figure 5 - Correlation of protein abundance with pseudotime in the panel of xenograft models.

(A) Two-dimensional plot depicting the progression of the xenograft models from hormone-sensitive prostate cancer (HSPC) to castration-resistant prostate cancer (CRPC) and neuroendocrine prostate cancer (NEPC) for the 11-plex. (B) Same representation for the 18-plex. (C) Mono-dimensional plot showing the progression of the xenograft models from HSPC to CRPC and NEPC, with the color gradient representing the phenotype as in (A, B). (D) Principal component analysis based on protein abundance clearly cluster the samples based on the progression stage, both in the 11-plex. (E) The same is true for samples in the 18-plex. (F) Principal component analysis based on protein abundance for the 11-plex clearly shows a correlation between PC1 and the pseudotime value. The color gradient represents the pseudotime computed for each model, with blue indicating the earliest timepoint and red indicating the latest. (G) Scatter plot depicting the correlation between protein abundance between the duplicates for each xenograft model (11-plex). Pearson's correlation coefficient for replicate1 (REPLICATE1) and replicate 2 (REPLICATE2) were computed and depicted in the figure. Each point represents a protein. Among the positively correlated proteins, we can find members of the Polycomb repressor complex 2 (PRC2) EZH2, EED, SUZ12, together with proteins involved in the mitotic spindle formation (HAUS1, HAUS6, HAUS4) and methyltransferases DNMT1 and DNMT3A, known to play a critical role in maintaining the methylation pattern during DNA replication, and to establish de novo DNA methylation patterns, respectively. Genes belonging to androgen receptor signaling are found to be negatively correlated. (H) GSEA based on the most positively (threshold of correlation coefficient = 0.5) and negatively (threshold of correlation coefficient = -0.5) correlated proteins in the comparison between mRNA expression and protein abundance, shows enrichment in gene set related to proliferation and downregulation of AR signaling and cholesterol homeostasis. (I) Dot plot representing GSEA analysis based on the most correlating genes between patient' samples (threshold pf PT=150) and the xenograft models. Gene set linked to TGFβ signaling, interferon and inflammation in general are found to be positively correlating, whereas gen sets related to AR and PI3K/MTORC signaling are found to be negatively correlating.

4.10 Longitudinal samples to exploit stages of tumor progression

To further explore the intricate relationship between mRNA and proteome dynamics throughout different stages of disease progression, I took advantage of a unique patient-derived xenograft model of NEPC trans-differentiation: a typical hormone-naive AR/PSA-positive adenocarcinoma (LTL-331)[80] that, upon host castration, initially regresses but rapidly relapses as terminally differentiated NEPC (LTL-331R), within a framework of 32 weeks (Fig. 6A and Supplementary Figure 2D). The PDX has been profiled for proteomic at the time of the ingraft, after 8 and 12 weeks, when the tumor decreases in size after the castration of the mice, and 32 weeks, after the final relapse as NEPC. By comparing protein expression patterns across different stages of the disease, we aimed to identify key proteins or protein signatures that may be associated with disease progression or treatment response.

Principal component analysis based on the 2000 most variable proteins demonstrated a clear clustering pattern, suggesting that the proteomic landscape undergoes significant changes over time (Fig. 6B). Particularly, the intermediate time points at 8 and 12 weeks exhibited a similar proteomic profile, indicating a potential transitional state during disease progression. Direct correlation analysis on protein expression at 8 and 12 weeks (Fig. 6C) further confirmed a robust and significant association between the observed time points.

In order to examine the differential expression patterns between consecutive phases of progression within the LTL-331 model, I performed differential expression analysis at mRNA and protein levels across each time point, and Fold Changes representing the relative changes in gene/protein expression levels were calculated for each comparison. The 8-week and 12-week time points were merged for consistency. Notably, a strong correlation was observed when comparing the 32-week time point with the initial 0-week time point (R = 0.56, Fig. 6D. However, as the comparison moved towards the middle time point, the correlation gradually diminished (Fig. 6E). This loss of correlation suggests that the proteomic landscape undergoes intricate changes during the intermediate stages of disease progression, possibly reflecting complex molecular events and signaling pathways involved in prostate cancer development, that are not directly transcriptionally regulated. Moreover, these findings suggest that the

molecular changes occurring during the intermediate time points may differ from those observed at the extremes: this could reflect dynamic biological processes and signaling pathways specific to the transitional stages of disease progression.

To better investigate the pattern of protein modulation at various stages of progression, I employed the k-means clustering method on the most variable proteins (Supplementary Figure 6A, *top*). K-means clustering is a popular unsupervised learning algorithm that partitions data into k clusters based on their similarity. By selecting the most variable proteins, I focused only on the proteins that exhibit significant changes in expression levels across different stages of the progression, reducing noise and minimizing the influence of proteins that do not significantly change throughout cancer evolution. A specific cluster exhibited a notable pattern: there was a rapid increase in abundance at 8 weeks, followed by a consistent level of activation up to 12 weeks, and then a sharp decline upon reaching 32 weeks (Cluster 4, Supplementary Figure 6A, *bottom*). I applied a predetermined threshold to filter and refine the cluster, specifically targeting proteins with statistically significant positive regulation during the time of 0 to 8 weeks, while encountering significant downregulation between 12 and 32 weeks (Fig. 6F).

By performing an over-representation analysis (ORA) using the list of proteins exhibiting the pattern observed in cluster 4, it would be possible to determine that gene sets associated with this behavior are predominantly linked to inflammation (TNFα signaling, interferon α/γ signaling, complement ad coagulation) and epithelial-mesenchymal transition (EMT). Specifical activation of the immune response signaling pathways is of particular interest in this context, as often links to the activation of immune surveillance, which refers to the continuous monitoring and detection of abnormal cells, including cancer cells, by the immune system. In the context of cancer, innate immune cells, including macrophages, natural killer (NK) cells, dendritic cells (DCs), and neutrophils, contribute to immune surveillance [113].

Although the experimental setup involved NRG immunocompromised mice with restricted activity limited to macrophages and dendritic cells, it remains intriguing to explore

whether there is a comparable induction of inflammation pathways in the mouse stromal counterpart. Principal component analysis (PCA) conducted on the mouse stroma, focusing on the 500 most variable genes, revealed a distinct separation between different time points, indicating temporal changes in protein expression (Supplementary Figure 6B). However, one replicate of the 12-week time point was unexpectedly close to the week 0-time point, suggesting an anomaly or potential similarity. Nevertheless, I performed differential expression analysis across stages, that revealed the activation of a clear inflammatory pattern (Fig. 6H and Supplementary Figure 6C) at 8 and 12 weeks, that is completely reverted at 32 weeks-relapse (Fig. 6I and Supplementary Figure 6D).

To sum up, the analysis of the correlation between mRNA and proteome reveals an intriguing pattern: a robust correlation is observed between these molecular layers when examining the extreme time points. However, as the investigation transitions towards intermediate stages, a gradual loss of correlation between mRNA and proteome becomes evident. This finding raises fundamental questions regarding the dynamics of gene expression and protein synthesis during these transitional phases. Interestingly, during this intermediate time frame, a pronounced upregulation of inflammatory pathways is observed, indicating their active involvement in the underlying biological processes, potentially linked to immune surveillance. The presence of inflammation suggests the existence of a stromal counterpart that plays a pivotal role during this phase: the preliminary analysis suggests a complex interplay between the mRNA and proteome dynamics, highlighting the importance of considering temporal factors and the activation of inflammatory pathways when investigating the regulatory mechanisms that govern cellular processes involved in prostate cancer progression.

FIGURE 6

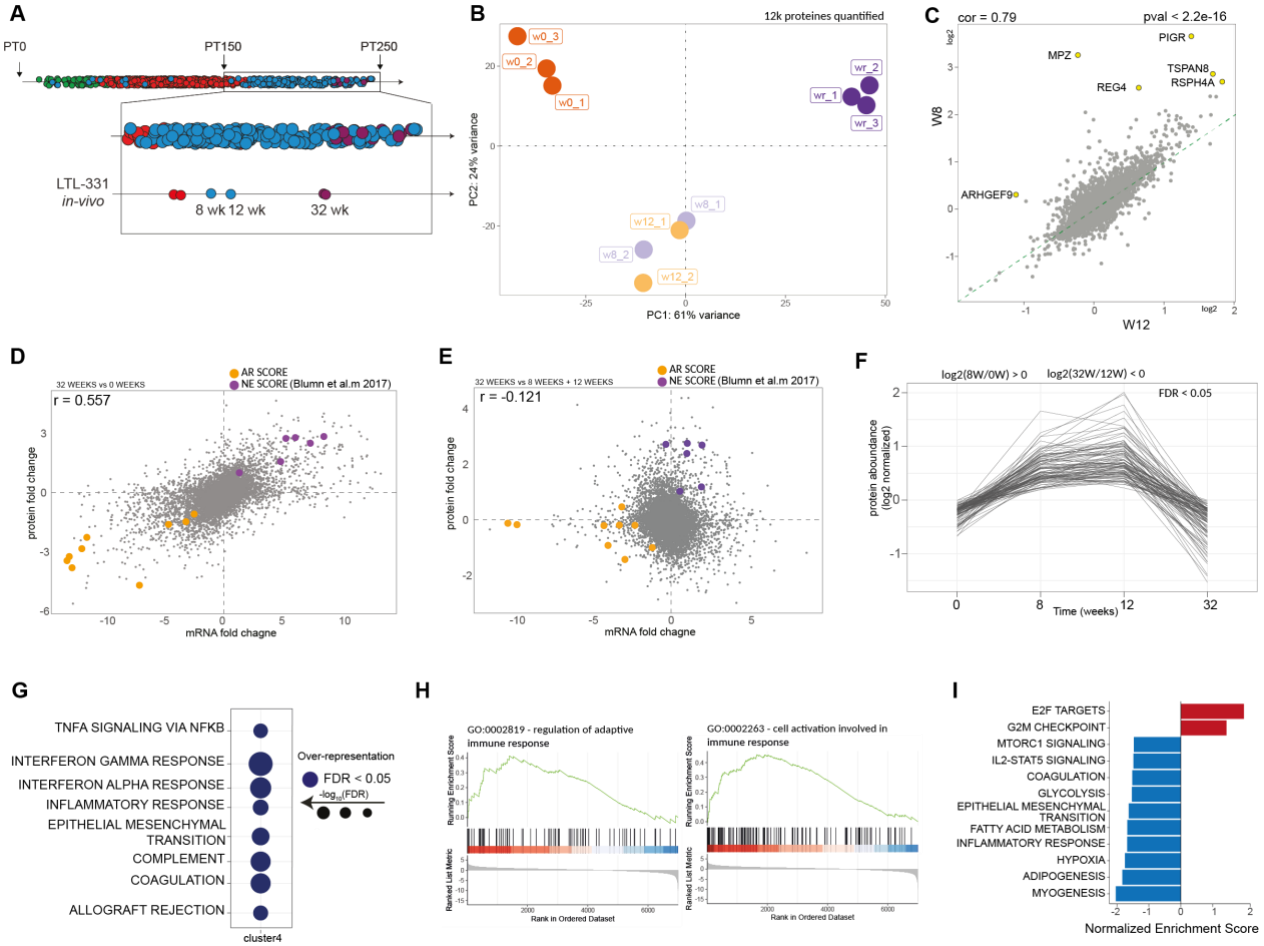
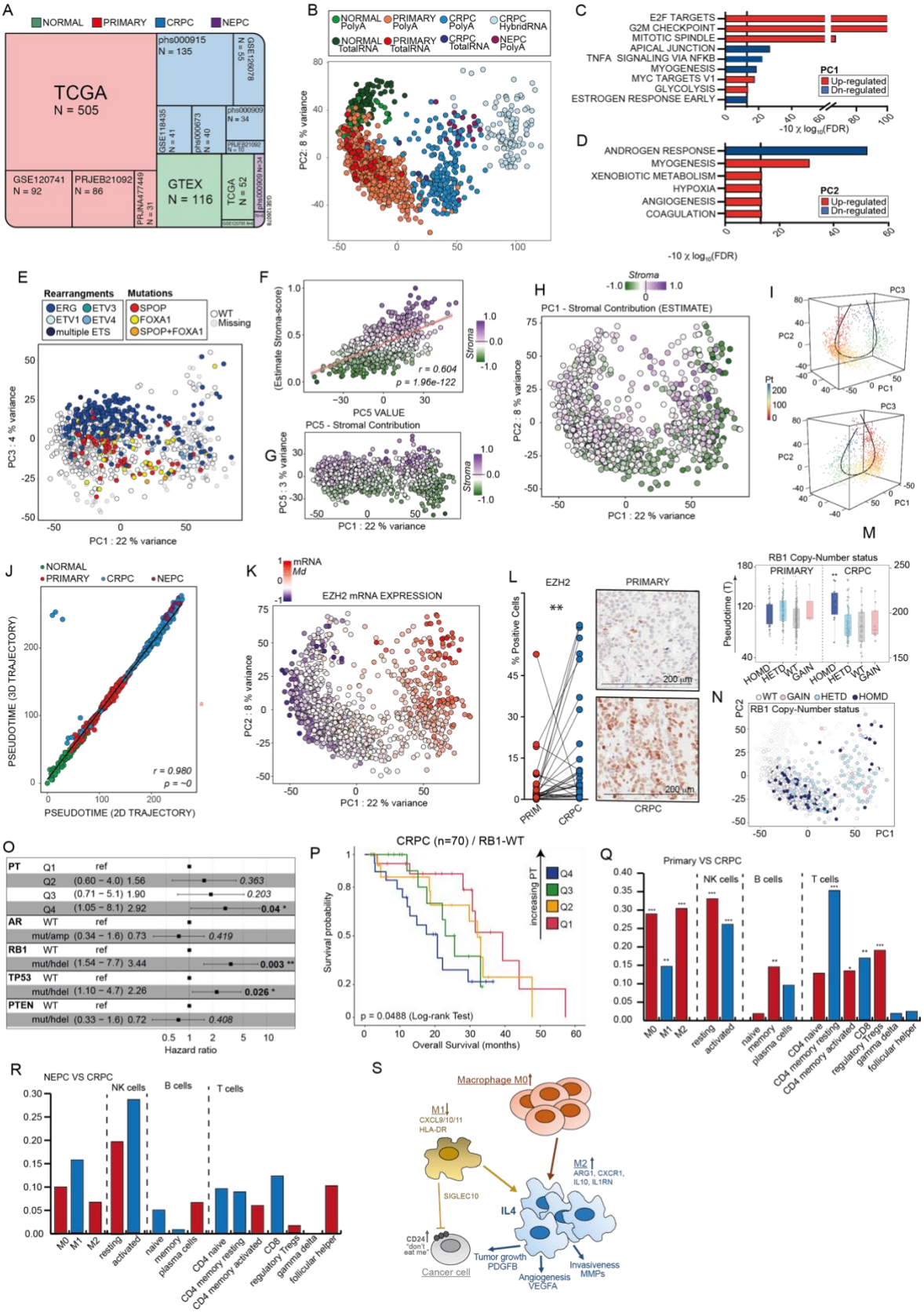


Figure 6 – Proteomic profiling of longitudinal xenograft models

(A) Mono-dimensional plot showing the progression of the LTL-331 xenograft models from HSPC to CRPC and NEPC, with the color gradient representing the phenotype. (B) Principal component analysis (PCA) based on the 2000 most variable proteins clearly clusters the samples based on the progression stage, with the highest similarity found between intermediate time points. (C) Scatterplot representing. (D-E) Scatterplot revealing a correlation between mRNAs and protein abundances, expressed in the form of fold-change (log-scale) between 32- and 0-week time points or 8-plus 12-week time points, respectively. The rationale behind the choice of unifying the intermediate stages stems from their high level of similarity in terms of protein (and gene) expression. (F) The graph represents the trend of expression of protein abundance, for those belonging to cluster 4 and further selected based on the threshold reported in the figure. (G) Dotplot depicting Over-representation analysis on the selected proteins belonging to cluster 4. Proteins were selected based on the direction and statistical significance of their regulation (See also Supplementary Figure 6A). (H) GSEA plot of two GO terms enriched in the differential analysis between the 8-and 0-week time points. (I) Barplot depicting GSEA analysis on the most significantly differentially expressed proteins between the 32- and 12-week time points. The length of the bars represents the Normalized Enrichment Score. Red=up-regulated gene signature; blue=down-regulated.

5 SUPPLEMENTARY FIGURES

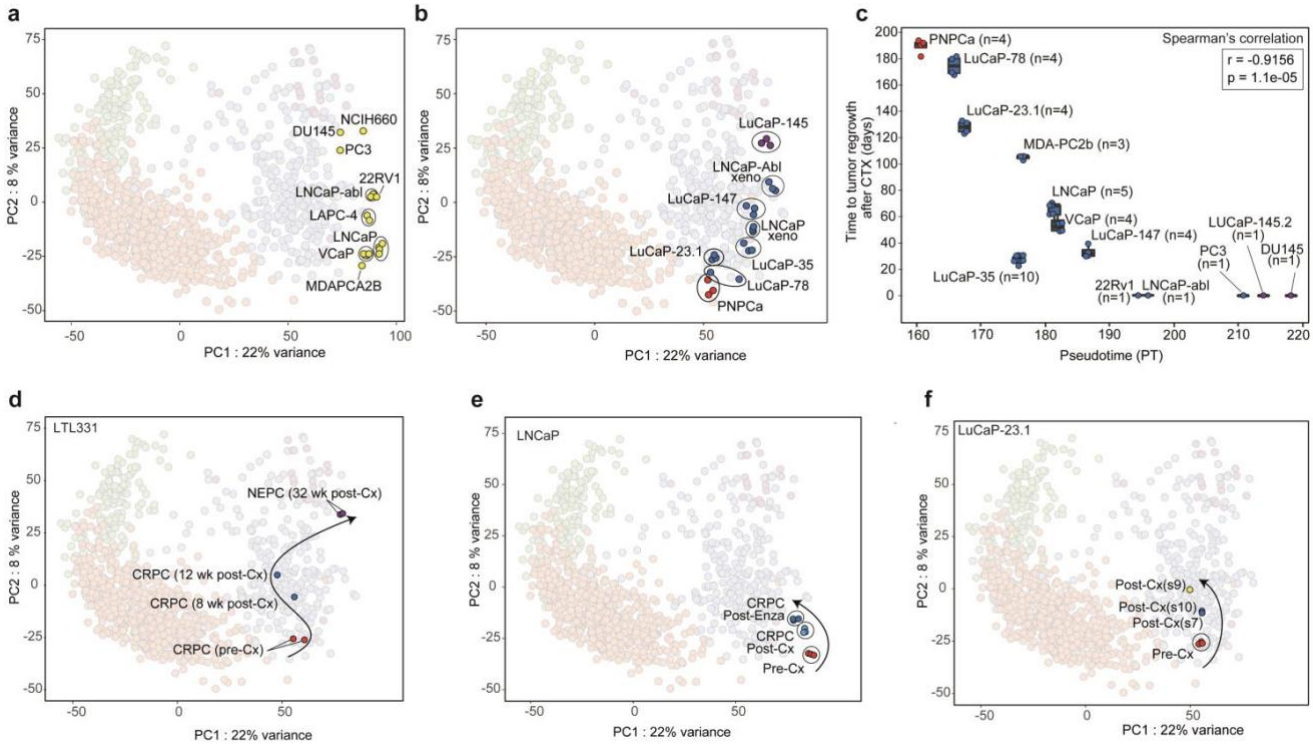
Supplementary Figure 1



Supplementary Figure 1.

(A) Graphical representation of the RNA sequencing cohorts, their accession numbers, the total number of samples in each dataset, and tumor stages as indicated. (B) Position of individual tumors in the PCA after re-processing of the raw data by selecting the top 2000 most variable genes. Hybrid capture-based RNA sequencing samples derived fromCPRC highlighted in light blue show a marked but consistent shift in the PC1 and PC2. No significant differences are observed in the first two principal components for TotalRNA when compared to PolyA+ samples. (C) Gene-sets enrichments performed using *Camera* algorithm on genes ranked according to their relative contribution (coefficient) to the positioning of samples along the PC1 axis. The analysis performed on Hallmark gene sets reveals an increase of cell cycle-related gene sets along PC1. (D) Corresponding analysis performed on genes ranked according to their contribution to PC2 shows a decrease in androgen-responsive genes along this axis. (E) PCA plot representing the PC1/PC3 pane can be used to discern SPOP/FOXA1 mutant prostate cancers from those harboring gene fusions involving ETS transcription factors. (F) Stromal score correlation to the PC5 component. Pearson's correlation coefficient and associated p-values are reported. The stromal score was computed using ESTIMATE (Estimation of Stromal and Immune cells in Malignant Tumor tissues using Expression data) across the samples. (G) Correlation between PC1 and PC5 values. The score is scaled between -1 and 1 and is represented in a three-color scale (green: lowest value; white: median value; violet: highest value). X-axis: the value of principal component 5 (PC5); Y-axis: The associated Estimate stroma-score, scaled between 0 and 1. (H) Corresponding analysis does not show a major influence of the stromal component to the positioning of samples in the PC1-2 plot. (I) Two distinct views of 3- dimensional PCA plot with the corresponding 3D trajectory. (J) Plot representing the correlation between the 2D- and 3D inferred pseudotime in the entire cohort (r = Pearson's correlation coefficient). (K) EZH2 mRNA expression increases gradually along the main trajectory. Expression levels of each sample are reported within the PCA plot representing the PC1/PC2 pane. Gene expressionlevels are scaled between -1 and 1 and are represented in a three-color scale (blue: lowest value; white: median value; red: highest value). (L) IHC analysis reveals upregulation of EZH2 in CRPC tumors compared to the matched primarytumors. Left: Quantification of EZH2 positive cells. Right: IHC images of a primary and its corresponding CRPC counterpart. (M) Boxplots representing different pseudo-time distributions for RB1-specific copy number alterations (homozygous, heterozygous, wild-type, gains). (N) Corresponding PCA plot highlighting RB1 copy-number status acrosssamples. (O) Multivariate analysis using the Cox Proportional Hazard model shows the relative contribution of pseudotime, AR status, RB1 status, TP53 status and PTEN status to survival. (P) Kaplan-Meier curve for overall survival related to pseudotime using a 4-tiered scoring system (quartiles) in the subset of CRPCs patients with wild-type RB1. (Q) Histograms depicting the correlation between the inferred abundance of the indicated immune cell populations (as determined by Cibersortx) and pseudo-time in a subset composed of primary and CRPCs tumors. P-values associated with Pearson's correlation coefficients were adjusted for multiple testing using the false discovery rate (FDR). (R) Histograms depicting the correlation between the inferred abundance of the indicated immune cell populations (as determined by Cibersortx) and pseudo-time in a subset composed of CRPCs and neuroendocrine tumors. P-values associated with Pearson's correlation coefficients were adjusted for multiple testing using the false discovery rate (FDR). S) Schematic representation of expression changes in genes linked to the tumor environment. Transcripts specific to M1-macrophages decrease along the trajectory while those of the M2 counterpart increase.

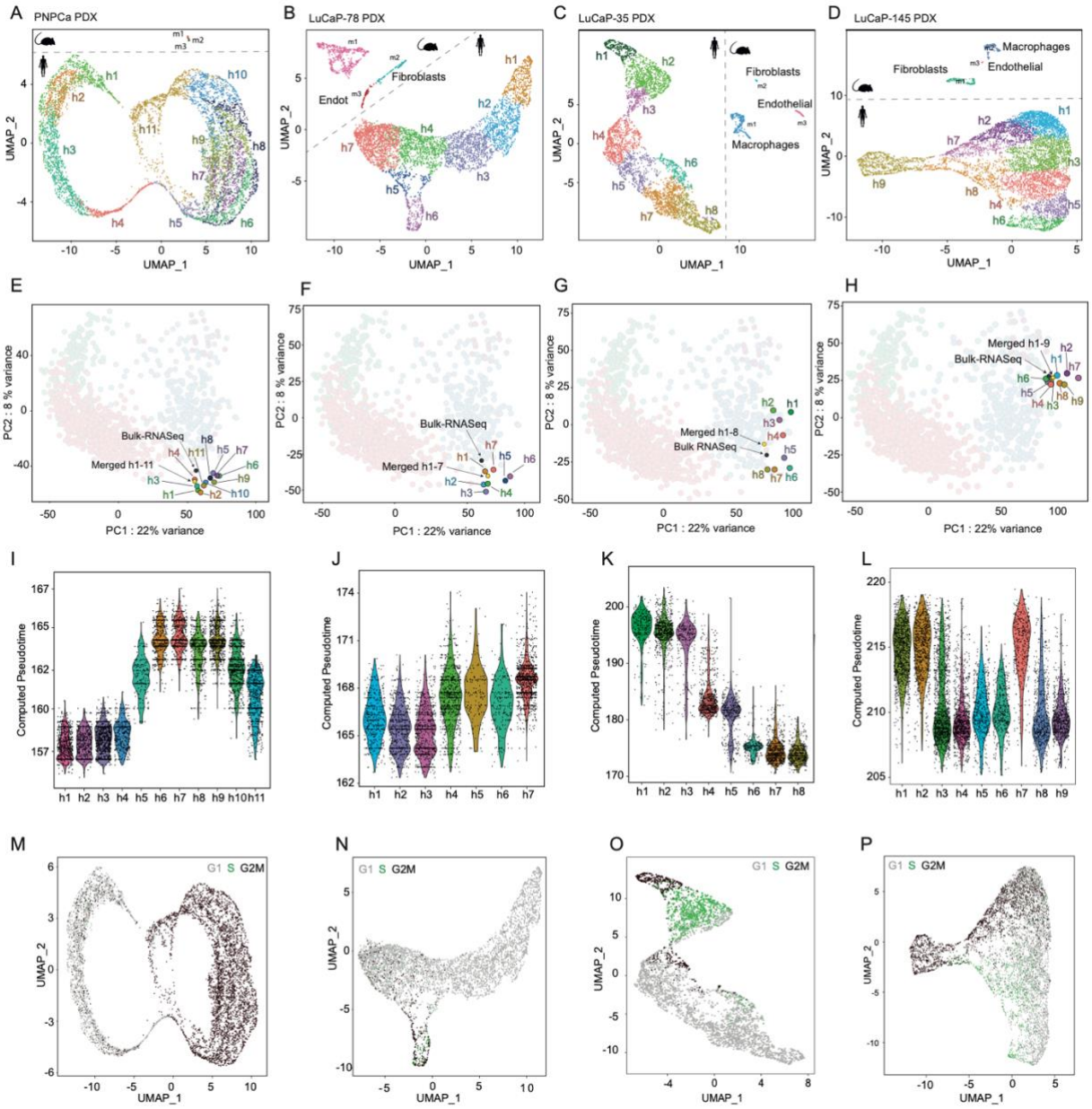
Supplementary figure 2



Supplementary Figure 2

(A) Integration of the indicated ex-vivo cultured prostate cancer cell lines within the PCA plot. (B) Corresponding analysis for Xenografts. For the PNPCa model, the normal and primary tumor tissue's PCA position is reported and dramatically differs from the one found in immune-compromised mice. Red = Primary Tumor, Blue = CRPC, Violet = NEPC. (C) Plot showing the correlation (Spearman, r = -0.9156/ p = 1.1e-05) between the pseudotime values inferred for each xenograft model and the corresponding latency of tumor regrowth after castration. X-axis: pseudotime value for the PDX and xenograft models; Y-axis: time to tumor regrowth after castration. Correlation with pseudotime was computed using the average latency for each xenograft: PNPCa=187.75; LuCaP-78=173; LuCaP23.1=127; LuCaP35=26.8; MDAPCA2B=104; LNCaP=64.20; VCaP=53.75; LuCaP147=33; 22Rv1, LNCaP-abl, PC3, LuCaP145.2 and DU145 = 0. Red = Primary Tumor, Blue = CRPC, Violet = NEPC. 22Rv1, LNCaP-abl, PC3, LuCaP-145.2 and DU145 are DHT insensitive and don't stop growing after castration, thus 1 single experiment for each of these Xenograft models is reported, as their Time to Tumor regrowth equals 0. (D-F) After castration, the indicated PDX models and LNCaP xenograft progress along the main trajectory. Red = Primary Tumor, Blue = CRPC, Violet = NEPC.

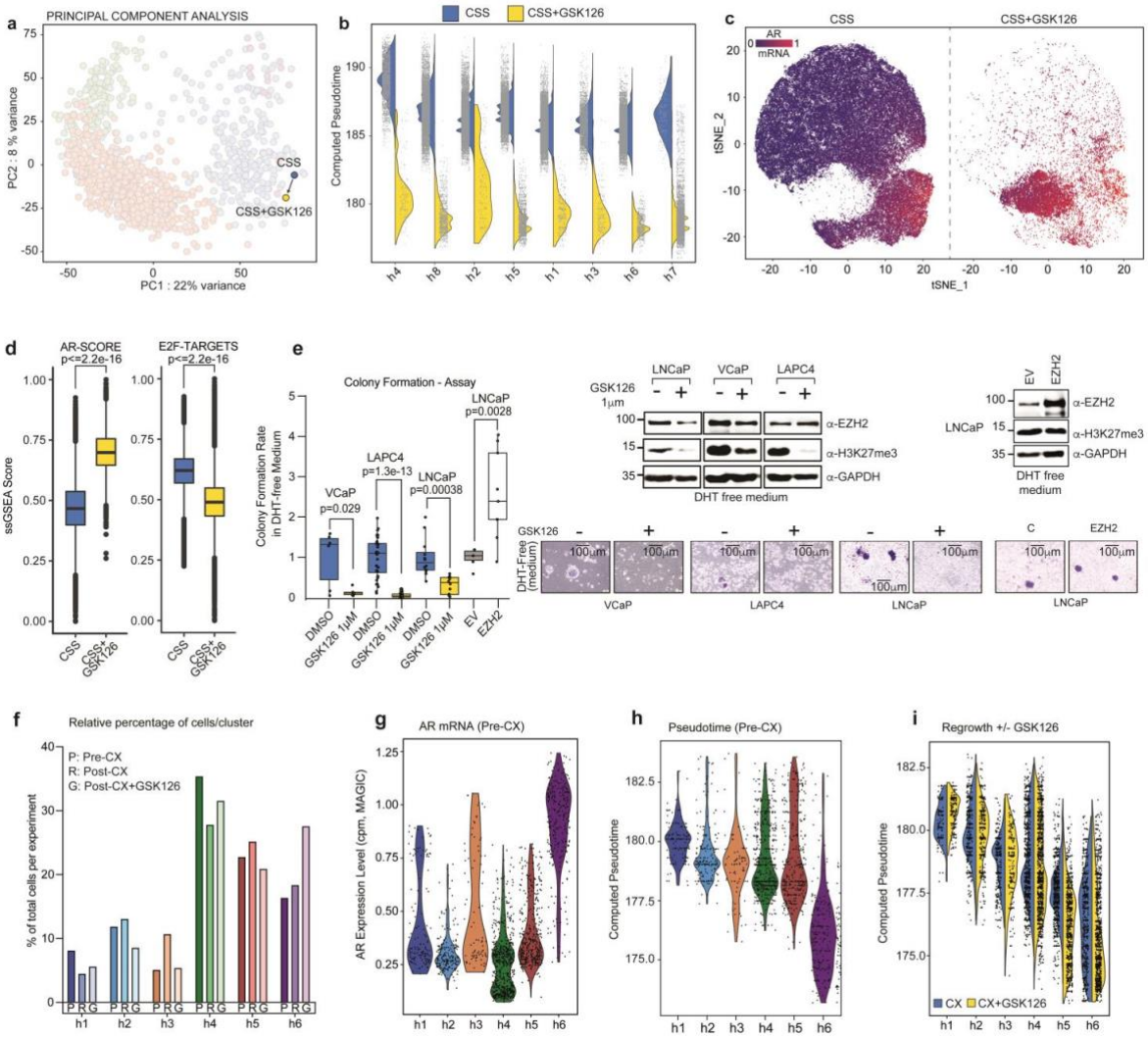
Supplementary figure 3



Supplementary figure 3

(A-D) Single-cell representation of the indicated PDX models *in vivo* using dimensionality reduction by Uniform Manifold Approximation and Projection (UMAP) and subsequent identification of tumor single-cell clusters using Seurat's workflow. (E-H) The Integration of merged single-cell data of the indicated PDX model on the PCA plot shows a comparable position to the corresponding bulk RNA sequencing data. Individual single-cell clusters are also integrated into the PC1/PC2 pane. The highest dispersion of clusters along the main trajectory is seen for LuCaP-35. (I-L) Violin plots indicate the pseudo-time of individual cells within the different cell clusters. The pseudo-time inference was performed following the imputation of missing genes (dropout events) by using *RMagic*. (M-P) Attribution of cells to either the G1, S or G2M cellcycle phase for each PDX model. Cell Cycle Phase was determined using Seurat's workflow. Also see Figure 3.

Supplementary figure 4

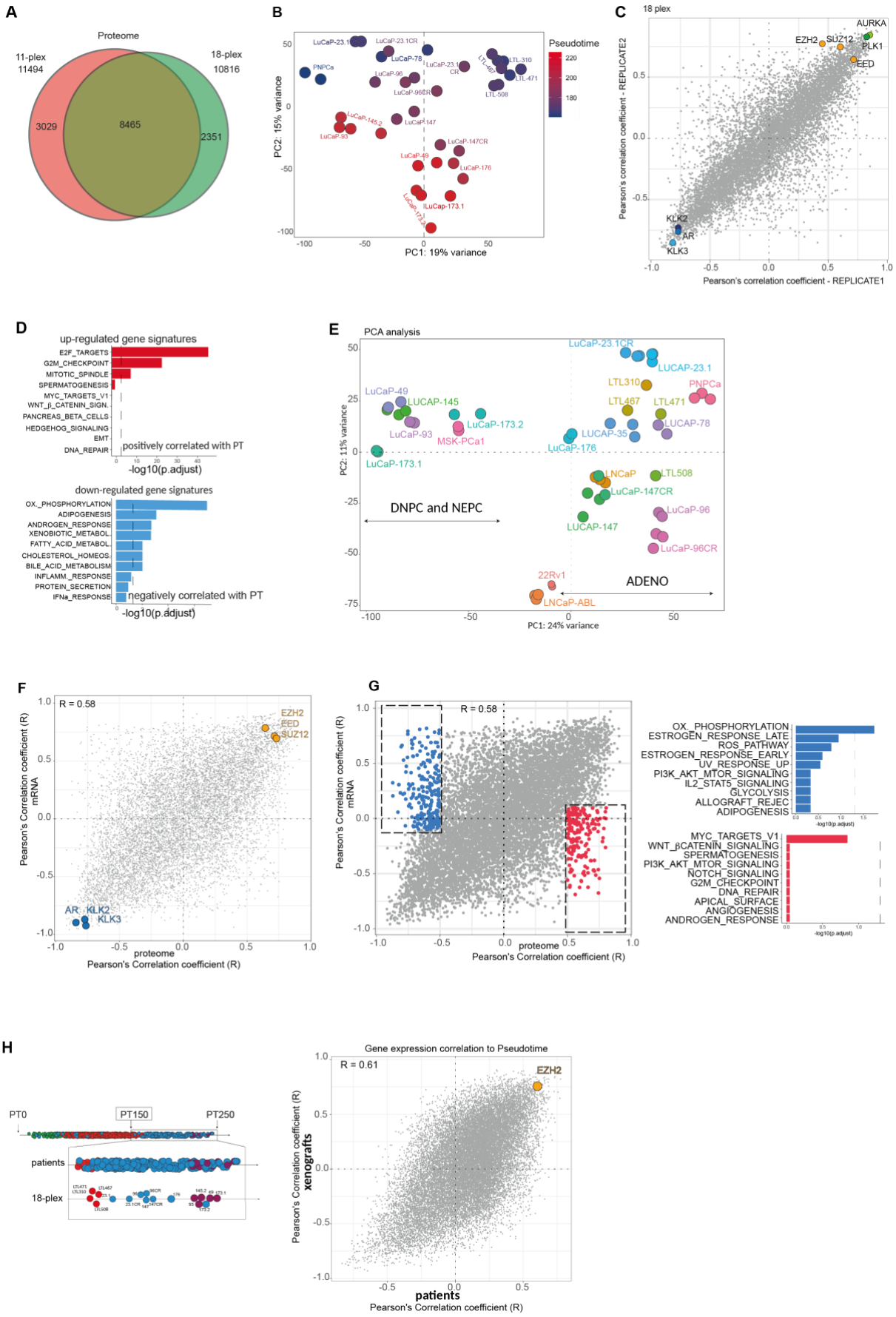


Supplementary figure 4

(A) Single-cell RNA-Seq data of LNCaP cells cultured in charcoal-stripped serum (CSS) was merged and integrated within the PCA plot (PC1/PC2). PCA positioning shows a decrease in pseudo-time upon EZH2 inhibition by GSK126. (B) Pseudotime of individual LNCaP cultured in charcoal-stripped serum (CSS) is significantly reduced upon GSK126 treatment in each cell cluster (h1-h7). Pseudo-time was computed for each cell, following the imputation of missing genes (drop-outs) using RMagic. (C) Dimensionality reduction (TSNE) of single-cell RNA-Seq performed on LNCaP cells cultured in vitro with charcoal-strippedserum (CSS) in the presence (right) or absence (left) of the EZH2 inhibitor GSK126. Upon GSK126 treatment, as most cell clusters disappear, there is an increase in AR mRNA expression in the transcriptionally rewired LNCaP cells that give rise to a novel cluster characterized by higher AR expression levels. (D) Corresponding quantification of AR-SCOREand E2F target genes (Hallmark gene set) computed for each cell before (left) and following (right) EZH2 inhibition by GSK126. Missing gene-expression values (dropout events) for each cell were imputed using *RMagic*. (E) GSK126 inhibits colony formation of LNCaP cells when cultured in CSS, while EZH2 over-expression increases the number of colonies under the same

condition. (n=3 independent experiments per condition) (F) Following EZH2 inhibition by GSK126, there is a common trend towards decreasing of cells in most clusters, except for cluster h6, which shows an opposite behavior. (G) Violin plots depicting AR expression levels show that the h6 cell cluster is characterized by higher levels of the latter. Missing gene-expression values (dropout events) for each cell were imputed using *RMagic*. (H) The pseudo-time inference was performed for each cell, and cluster h6 resulted to be associated with a less progressed phenotype. Clusters (h1-h6) are depicted using different colors as indicated in the figure panel. (I) Violin plots comparing pseudotime before (blue) or following (yellow) EZH2 inhibition by GSK126. Cluster h6 displays the highest AR expression levels, the least progression on the main trajectory, and a reduction in pseudo-time after GSK126 treatment. Significance was assessed using the Wilcoxon sum-rank test and p-values were adjusted for multiple comparisons using false discovery rate (FDR): * < 0.05, ** < 0.01, *** < 0.001.

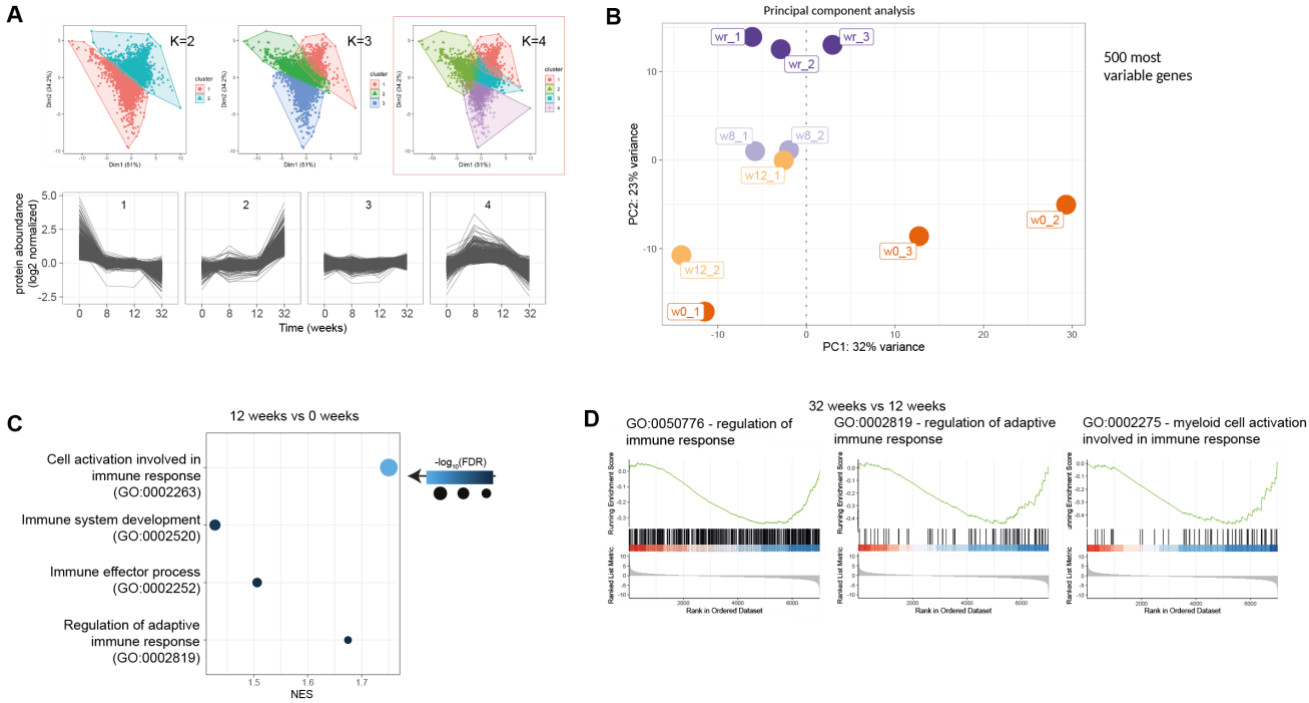
Supplementary figure 5



Supplementary figure 5

(A)Venn diagram representing the overlap in protein quantified between the 11-and 18-plexes. As expected, I found a good overlap (around 75%) between the two experiments, opening the possibility to merge them into one to maximize the statistical power of the analysis. (B) Principal component analysis based on protein abundance for the 18-plex clearly shows a correlation between PC2 and the pseudotime value. The color gradient represents the pseudotime computed for each model, with blue indicating the earliest timepoint and red indicating the latest. (C) (Scatterplot depicting the correlation between protein abundance and pseudotime across replicates in the xenograft models (18-plex). Each point represents a protein; Pearson's correlation coefficient for each protein is depicted in the x-axis and y-axis for replicate 1 and 2, respectively. Among the positively correlated proteins there are members of the Polycomb repressor complex 2 (PRC2) EZH2, EED, SUZ12. On the other hand, genes belonging to androgen receptor signaling are negatively correlated. (D) Barplot representing GSEA analysis on the most significantly correlating proteins, ranked by correlation coefficient. (E) Principal component analysis on gene expression data for the entire panel of the xenograft models. PC1 separates models based on AR dependency, being double negative prostate cancers (DNPC) and NEPC clearly divided from AR-positive adenocarcinomas. (F) The scatterplot shows the correlation between mRNA expression and protein abundance within the xenograft models (18-plex). (G) *left*. Selection of outliers that have a correlation at protein levels greater than 0.5 or less than -0.5, and that have no or poor correlation at the mRNA level. *Right*. Barplot from GSEA analysis shows no significant enrichments for gene sets based on the selected outliers. (H) (left) Mono-dimensional plot highlighting patient samples selected for correlation analysis based on a threshold of pseudotime (150), to exclude from the investigation the portion of the trajectory that is not covered by the PDX models. (right) Scatterplot representing gene expression correlation to pseudotime for the xenograft models (y-axis) and for the patient' samples (x-axis). The correlation coefficient is 0.61 (p-value <0.001). EZH2 gene is highlighted in orange.

Supplementary Figure 6



Supplementary figure 6

(A) *Top*: Results of the k-means clustering analysis on proteomic data, progressively increasing the number of clusters. The final decision about the k-value (4) has been made using the Elbow method. *Bottom*: trend abundances across the progression line, for proteins belonging to each one of the clusters identified through the k-means algorithm. (B) Principal component analysis (PCA) conducted on the mouse stroma using the 500 most variable genes demonstrated a clear distinction between different time samples, indicating temporal variations in protein expression, clearly associated with the different stages of tumor progression. (C) Gene Set Enrichment Analysis (GSEA) on Gene Oncology Collection (Biological Processes) shows up-regulation of gene set related to immune response when considering the comparison between the 12-week and the 0-time points in mouse stromal tissue. (D) GSEA plot representing Gene Ontology enrichment on selected gene sets, highlighting that upon tumor relapse, inflammation-related pathways are clearly down-modulated.

6 DISCUSSION

In the present study, transcriptional profiles of prostate cancers at various disease stages were combined into a comprehensive prostate cancer transcriptome atlas with negligible study-related interference (i.e., “batch effects”). Mining the atlas reveals a rather uniform trajectory towards disease progression from normal prostate, primary, and metastatic castration-resistant prostate cancer. The trajectory is characterized by a gradual upregulation of genes related to EZH2-mediated polycomb signaling and cell cycle progression, most namely G2M checkpoints and mitotic spindle genes. The latter may provide an explanation why taxanes (i.e., docetaxel, cabazitaxel) which disrupt microtubule function during cell division, remain a cornerstone of prostate cancer treatment in the hormone-sensitive and castration-resistant metastatic setting [23, 114-120].

EZH2 has been previously described to be critically involved in prostate cancer as an activator of AR signaling[94]. It is also a key component of polycomb repressor complex 2-mediated gene silencing – a developmental pathway implicated in de-differentiation and prostate cancer progression [96, 121, 122]. In agreement with the latter, we find EZH2 the top-upregulated gene in the progression trajectory along with other PRC2 members. In line with a function in driving disease progression and de-differentiation towards the loss of AR expression, I demonstrate how EZH2 inhibition reverts the transcriptional output of prostate cancer cells along the progression trajectory. The findings may have important implications for the treatment of prostate cancer patients in a hormone-naïve or early CRPC because it may prevent the de-differentiation of cancer cells as an escape-mechanisms to AR-directed therapeutic interventions. In line with previous reports, I noticed along the trajectory a change of macrophage polarization from inflammatory M1 to pro-tumorigenic M2[102, 103]. Our findings further underscore the anti-tumor potential of pharmacologically re-educating macrophages towards M1. Castration was sufficient in our PDX models to induce a change toward M2-polarization after a relatively short period in line with previous reports [123], suggesting that therapeutic interventions per se may be at least in part the underlying cause. Importantly, in the same setting the inhibition of EZH2 substantially blocked the castration-induced polarization change towards M2, uncovering a thus far underappreciated role for EZH2 in macrophage polarization and another rationale towards co-targeting AR and EZH2 in prostate cancer.

It is mostly unknown how disease progression in prostate cancer emerges at the single-cell level. Using a series of PDX models reflecting different progression stages from hormone-naïve to AR-negative late-stage disease enabled the addition of single-cell resolution to the progression trajectory. Our results suggest that resistance to androgen deprivation may occur through transcriptional adaptation of tumor cells towards a more progressed state. In line with this, a recent study has proposed that prostate regeneration (a process that shares many molecular features with prostate cancer progression) is driven by nearly all persisting luminal cells, not just by rare stem cells[124]. That said, in the study, we have used a relatively uniform xenograft tumor model that has been already derived from CRPC and thus adapt swiftly to castration in mice. Conceivably, resistance to androgen receptor inhibition over a longer period may also involve the selection of stem-cell-like subpopulations irrespective of the presence of genetic drivers of CRPC (e.g., AR amplification or point mutations)[125, 126, 127].

In conclusion, me and Marco Bolis successfully merged and exploited the RNA sequencing data from several prostate cancer studies, covering different disease stages. Based on that, we delineate the roadmap to prostate cancer progression in an unprecedented, qualitative, and quantitative manner.

Furthermore, I was able to effectively show how individual tumor cells can be tracked along the progression trajectory in response to pharmacological perturbations. Because transcriptome data of advanced metastatic disease will become more readily available for other tumor types, the current study may serve as a blueprint for their analysis and exploitation.

Subsequently, I aimed at introducing a second layer of complexity, through the integration within the inferred and exploited progression line of mass-spectrometry-based proteomics data on a panel of models of advanced disease. Proteomics has recently emerged as a powerful tool for investigating the molecular mechanisms underlying prostate cancer complexity [128, 129]. However, most studies had been conducted thus far on localized disease [130, 131], with the limitation of not being representative of advanced disease and lacking the view of the intricacy of events driving tumor evolution.

By leveraging our range of models that cover a spectrum from late-stage primary tumors to advanced disease, I have been able to identify molecular patterns and signaling

pathways that become more active as prostate cancer advances, specifically at the protein level. In particular, the use of PDX models has allowed me to accurately differentiate the signals coming from human tumor cells from those originating from mouse stroma. Specifically, I have observed a consistent upregulation of proteins involved in chromatin modification, cell cycle, and cell proliferation, as previously documented in the few studies that were carried out on advanced disease [132]. Conversely, I found downregulation of proteins associated with androgen receptor (AR) signaling.

There is growing evidence that the correlation between mRNA and protein expression in prostate cancer is complex and variable [133]. This relationship can be influenced by many factors, including post-transcriptional regulation, protein degradation, and protein modification [134, 135]. Furthermore, different proteins may exhibit varying degrees of correlation with their corresponding mRNA levels, with some proteins showing a strong correlation while others exhibit a weak or no correlation.

In our study, when comparing the correlation of mRNA expression and pseudotime to that of protein abundance and pseudotime, a strong concordance between the two measures was observed, way stronger than comparing mRNAs and proteins directly. This is in line with a previous study [136], where it has been reported that differentially expressed mRNAs correlate significantly better with their protein product than non-differentially expressed mRNAs, across a time series of ovarian cancer xenograft models.

This may suggest that mRNA expression levels can be used as a proxy for protein abundance in some cases and highlight the importance of considering both mRNA and protein levels in understanding the molecular changes associated with tumor progression.

However, these findings may appear to contradict previous reports that stated a poor correlation between mRNA and protein expression in prostate cancer. There may be several reasons to justify those observations. Conceivably, the higher accuracy of our state-of-the-art quantitative proteomic approach using TMT-labeling in aggregate with the better preservation of PDX tissues as opposed to previous proteomics studies based on clinical tissue samples with significant and variable stroma contaminations, has contributed to a better

correlation of mRNA and protein expression. Anyhow, a large spectrum of post-transcriptional and post-translational modifications can lead to discordance between mRNA and protein levels. As an example, miRNA regulation, and protein degradation can contribute to differences in mRNA and protein expression levels, as well as the effect of protein-protein interactions, which may not be directly reflected in mRNA expression levels [137, 138].

In order to better investigate the evolving interplay between mRNA and protein, I conducted an analysis using a longitudinal model [80]. The correlation analysis revealed as expected a strong association between mRNA and proteome at the extreme time points. However, as I examined the intermediate stages, a progressive loss of correlation between the two became evident. Intriguingly, during this transitional phase, there was a notable upregulation of inflammatory pathways, indicating their active engagement. These findings found partial confirmation by the analysis of the stromal counterpart, where a similar trend in inflammatory pathways has been found. Clearly, this approach is far from being exhaustive: in order to gain a more comprehensive understanding of this aspect, it may be necessary to incorporate additional approaches such as syngeneic models or biopsies from minimal residual disease in patients. The latter can provide valuable insights into the dynamics of mRNA and protein interactions in the context of disease progression and treatment response, in the transitional phase that is usually missed in clinical samples. Additionally, the complex dynamics observed between mRNA and protein emphasize the importance of considering temporal factors and the involvement of inflammatory processes in comprehending the underlying cellular mechanisms driving tumor progression.

Despite the correlation demonstrated between the transcriptomic profiles of our xenograft models and patient samples revealed by our analysis, further evaluation and analysis may be necessary to fully understand the complexity of the molecular mechanisms underlying prostate cancer progression. Indeed, to obtain a truly comprehensive multi-omics model, it would be better to consider post-translational modifications (PTM) in addition to the abundance of proteins. Post-translational modifications, such as phosphorylation, ubiquitination, and acetylation, can significantly affect protein function and stability, and play

important roles in many cellular processes including cancer progression. Therefore, incorporating information on post-translational modifications in addition to protein abundance and mRNA expression data could provide a more comprehensive understanding of the molecular mechanisms underlying prostate cancer progression.

Additionally, including more samples from primary tumors (and potentially more samples from patients profiled for proteomics and PTMs) in the multi-omics profiling would certainly enhance the comprehensiveness and accuracy of the model. By including a larger and more diverse sample set, it would be possible to capture the heterogeneity of prostate cancer progression not only in the regulation of transcription, as done with the Transcriptome Atlas, but also in the regulation that occurs beyond, directly linked to post-translational modification of proteins. In this way, it would be possible to obtain a more complete understanding of the molecular changes that occur during prostate cancer progression and potentially identify subgroups with unique molecular profiles that can be targeted in the perspective of an increasingly accurate personalized medicine.

References

1 Ferlay J, Colombet M, Soerjomataram I, Parkin DM, Pineros M, Znaor A *et al*. Cancer statistics for the year 2020: An overview. *Int J Cancer* 2021.

2 Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. *CA Cancer J Clin* 2022; 72: 7-33.

3 Center MM, Jemal A, Lortet-Tieulent J, Ward E, Ferlay J, Brawley O *et al*. International variation in prostate cancer incidence and mortality rates. *Eur Urol* 2012; 61: 1079-1092.

4 Parker C, Castro E, Fizazi K, Heidenreich A, Ost P, Procopio G *et al*. Prostate cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol* 2020; 31: 1119-1134.

5 Nuhn P, De Bono JS, Fizazi K, Freedland SJ, Grilli M, Kantoff PW *et al*. Update on Systemic Prostate Cancer Therapies: Management of Metastatic Castration-resistant Prostate Cancer in the Era of Precision Oncology. *Eur Urol* 2019; 75: 88-99.

6 Singh O, Bolla SR. Anatomy, Abdomen and Pelvis, Prostate. *StatPearls*: Treasure Island (FL), 2023.

7 McNeal JE. Normal histology of the prostate. *Am J Surg Pathol* 1988; 12: 619-633.

8 Tuxhorn JA, Ayala GE, Smith MJ, Smith VC, Dang TD, Rowley DR. Reactive stroma in human prostate cancer: induction of myofibroblast phenotype and extracellular matrix remodeling. *Clin Cancer Res* 2002; 8: 2912-2923.

9 Tuong ZK, Loudon KW, Berry B, Richoz N, Jones J, Tan X *et al*. Resolving the immune landscape of human prostate at a single-cell level in health and cancer. *Cell Rep* 2021; 37: 110132.

10 Kosaka T, Miyajima A, Oya M. Is DHT Production by 5alpha-Reductase Friend or Foe in Prostate Cancer? *Front Oncol* 2014; 4: 247.

11 Gelmann EP. Molecular biology of the androgen receptor. *J Clin Oncol* 2002; 20: 3001-3015.

12 Mottet N, Bellmunt J, Bolla M, Briers E, Cumberbatch MG, De Santis M *et al*. EAU-ESTRO-SIOG Guidelines on Prostate Cancer. Part 1: Screening, Diagnosis, and Local Treatment with Curative Intent. *Eur Urol* 2017; 71: 618-629.

13 Descotes JL. Diagnosis of prostate cancer. *Asian J Urol* 2019; 6: 129-136.

14 Society AC. Cancer facts & figures 2021. In: Society AC (ed), 2021.

15 Filson CP, Marks LS, Litwin MS. Expectant management for men with early stage prostate cancer. *CA Cancer J Clin* 2015; 65: 265-282.

16 Coughlin GD, Yaxley JW, Chambers SK, Occhipinti S, Samaratunga H, Zajdlewicz L *et al*. Robot-assisted laparoscopic prostatectomy versus open radical retropubic prostatectomy: 24-month outcomes from a randomised controlled study. *Lancet Oncol* 2018; 19: 1051-1060.

17 Vanneste BG, Van Limbergen EJ, van Lin EN, van Roermund JG, Lambin P. Prostate Cancer Radiation Therapy: What Do Clinicians Have to Know? *Biomed Res Int* 2016; 2016: 6829875.

18 Gay HA, Michalski JM. Radiation Therapy for Prostate Cancer. *Mo Med* 2018; 115: 146-150.

19 Nabid A, Carrier N, Vigneault E, Van Nguyen T, Vavassis P, Brassard MA *et al*. Androgen deprivation therapy and radiotherapy in intermediate-risk prostate cancer: A randomised phase III trial. *Eur J Cancer* 2021; 143: 64-74.

20 Achard V, Putora PM, Omlin A, Zilli T, Fischer S. Metastatic Prostate Cancer: Treatment Options. *Oncology* 2022; 100: 48-59.

21 Wang EC, Lee WR, Armstrong AJ. Second generation anti-androgens and androgen deprivation therapy with radiation therapy in the definitive management of high-risk prostate cancer. *Prostate Cancer Prostatic Dis* 2022.

22 Berthold DR, Pond GR, Soban F, de Wit R, Eisenberger M, Tannock IF. Docetaxel plus prednisone or mitoxantrone plus prednisone for advanced prostate cancer: updated survival in the TAX 327 study. *J Clin Oncol* 2008; 26: 242-245.

23 de Bono JS, Oudard S, Ozguroglu M, Hansen S, Machiels JP, Kocak I *et al*. Prednisone plus cabazitaxel or mitoxantrone for metastatic castration-resistant prostate cancer progressing after docetaxel treatment: a randomised open-label trial. *Lancet* 2010; 376: 1147-1154.

24 Haffner MC, Zwart W, Roudier MP, True LD, Nelson WG, Epstein JI *et al*. Genomic and phenotypic heterogeneity in prostate cancer. *Nat Rev Urol* 2021; 18: 79-92.

25 Chokkalingam AP, Nyren O, Johansson JE, Gridley G, McLaughlin JK, Adami HO *et al*. Prostate carcinoma risk subsequent to diagnosis of benign prostatic hyperplasia: a population-based cohort study in Sweden. *Cancer* 2003; 98: 1727-1734.

26 Ma X, Liu Y, Liu Y, Alexandrov LB, Edmonson MN, Gawad C *et al*. Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. *Nature* 2018; 555: 371-376.

27 Brawer MK. Prostatic intraepithelial neoplasia: a premalignant lesion. *Hum Pathol* 1992; 23: 242-248.

28 Bostwick DG, Liu L, Brawer MK, Qian J. High-grade prostatic intraepithelial neoplasia. *Rev Urol* 2004; 6: 171-179.

29 Epstein JI, Egevad L, Amin MB, Delahunt B, Srigley JR, Humphrey PA *et al*. The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma: Definition of Grading Patterns and Proposal for a New Grading System. *Am J Surg Pathol* 2016; 40: 244-252.

30 Sekhoacha M, Riet K, Motloung P, Gumenku L, Adegoke A, Mashele S. Prostate Cancer Review: Genetics, Diagnosis, Treatment Options, and Alternative Approaches. *Molecules* 2022; 27.

31 Humphrey PA. Histological variants of prostatic carcinoma and their significance. *Histopathology* 2012; 60: 59-74.

32 Alizadeh M, Alizadeh S. Survey of clinical and pathological characteristics and outcomes of patients with prostate cancer. *Glob J Health Sci* 2014; 6: 49-57.

33 Hansel DE, Epstein JI. Sarcomatoid carcinoma of the prostate: a study of 42 cases. *Am J Surg Pathol* 2006; 30: 1316-1321.

34 Yamada Y, Beltran H. Clinical and Biological Features of Neuroendocrine Prostate Cancer. *Curr Oncol Rep* 2021; 23: 15.

35 Klusa D, Lohaus F, Furesi G, Rauner M, Benesova M, Krause M *et al*. Metastatic Spread in Prostate Cancer Patients Influencing Radiotherapy Response. *Front Oncol* 2020; 10: 627379.

36 Conteduca V, Oromendia C, Eng KW, Bareja R, Sigouros M, Molina A *et al*. Clinical features of neuroendocrine prostate cancer. *Eur J Cancer* 2019; 121: 7-18.

37 Datta K, Muders M, Zhang H, Tindall DJ. Mechanism of lymph node metastasis in prostate cancer. *Future Oncol* 2010; 6: 823-836.

38 Msaouel P, Pissimissis N, Halapas A, Koutsilieris M. Mechanisms of bone metastasis in prostate cancer: clinical implications. *Best Pract Res Clin Endocrinol Metab* 2008; 22: 341-355.

39 Gandaglia G, Karakiewicz PI, Briganti A, Passoni NM, Schiffmann J, Trudeau V *et al*. Impact of the Site of Metastases on Survival in Patients with Metastatic Prostate Cancer. *Eur Urol* 2015; 68: 325-334.

40 Beltran H, Prandi D, Mosquera JM, Benelli M, Puca L, Cyrta J *et al*. Divergent clonal evolution of castration-resistant neuroendocrine prostate cancer. *Nat Med* 2016; 22: 298-305.

41 Akamatsu S, Inoue T, Ogawa O, Gleave ME. Clinical and molecular features of treatment-related neuroendocrine prostate cancer. *Int J Urol* 2018; 25: 345-351.

42 Beltran H, Eng K, Mosquera JM, Sigaras A, Romanel A, Rennert H *et al*. Whole-Exome Sequencing of Metastatic Cancer and Biomarkers of Treatment Response. *JAMA Oncol* 2015; 1: 466-474.

43 Cancer Genome Atlas Research N. The Molecular Taxonomy of Primary Prostate Cancer. *Cell* 2015; 163: 1011-1025.

44 Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW *et al*. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* 2005; 310: 644-648.

45 Spans L, Clinckemalie L, Helsen C, Vanderschueren D, Boonen S, Lerut E *et al*. The genomic landscape of prostate cancer. *Int J Mol Sci* 2013; 14: 10822-10851.

46 Robinson D, Van Allen EM, Wu YM, Schultz N, Lonigro RJ, Mosquera JM *et al*. Integrative clinical genomics of advanced prostate cancer. *Cell* 2015; 161: 1215-1228.

47 Cai C, Balk SP. Intratumoral androgen biosynthesis in prostate cancer pathogenesis and response to therapy. *Endocr Relat Cancer* 2011; 18: R175-182.

48 Chen CD, Welsbie DS, Tran C, Baek SH, Chen R, Vessella R *et al*. Molecular determinants of resistance to antiandrogen therapy. *Nat Med* 2004; 10: 33-39.

49 Chen G, Ning B, Shi T. Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. *Front Genet* 2019; 10: 317.

50 Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. *Nat Protoc* 2018; 13: 599-604.

51 Ramskold D, Luo S, Wang YC, Li R, Deng Q, Faridani OR *et al*. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol* 2012; 30: 777-782.

52 Picelli S, Faridani OR, Bjorklund AK, Winberg G, Sagasser S, Sandberg R. Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc* 2014; 9: 171-181.

53 Hashimshony T, Wagner F, Sher N, Yanai I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep* 2012; 2: 666-673.

54 Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M *et al*. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* 2015; 161: 1202-1214.

55 Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R *et al*. Massively parallel digital transcriptional profiling of single cells. *Nat Commun* 2017; 8: 14049.

56 Aebersold R, Mann M. Mass spectrometry-based proteomics. *Nature* 2003; 422: 198-207.

57 Thompson A, Schafer J, Kuhn K, Kienle S, Schwarz J, Schmidt G *et al*. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem* 2003; 75: 1895-1904.

58 Grasso CS, Wu YM, Robinson DR, Cao X, Dhanasekaran SM, Khan AP *et al*. The mutational landscape of lethal castration-resistant prostate cancer. *Nature* 2012; 487: 239-243.

59 Kumar A, Coleman I, Morrissey C, Zhang X, True LD, Gulati R *et al*. Substantial interindividual and limited intraindividual genomic diversity among tumors from men with metastatic prostate cancer. *Nat Med* 2016; 22: 369-378.

60 Aggarwal R, Huang J, Alumkal JJ, Zhang L, Feng FY, Thomas GV *et al*. Clinical and Genomic Characterization of Treatment-Emergent Small-Cell Neuroendocrine Prostate Cancer: A Multi-institutional Prospective Study. *J Clin Oncol* 2018; 36: 2492-2503.

61 Labrecque MP, Coleman IM, Brown LG, True LD, Kollath L, Lakely B *et al*. Molecular profiling stratifies diverse phenotypes of treatment-refractory metastatic castration-resistant prostate cancer. *J Clin Invest* 2019; 129: 4492-4505.

62 Lai Y, Wei X, Lin S, Qin L, Cheng L, Li P. Current status and perspectives of patient-derived xenograft models in cancer research. *J Hematol Oncol* 2017; 10: 106.

63 Karkampouna S, La Manna F, Benjak A, Kiener M, De Menna M, Zoni E *et al*. Patient-derived xenografts and organoids model therapy response in prostate cancer. *Nat Commun* 2021; 12: 1117.

64 Nguyen HM, Vessella RL, Morrissey C, Brown LG, Coleman IM, Higano CS *et al*. LuCaP Prostate Cancer Patient-Derived Xenografts Reflect the Molecular Heterogeneity of Advanced Disease an--d Serve as Models for Evaluating Cancer Therapeutics. *Prostate* 2017; 77: 654-671.

65 Udeshi ND, Mani DC, Satpathy S, Fereshetian S, Gasser JA, Svinkina T *et al*. Rapid and deep-scale ubiquitylation profiling for biology and translational research. *Nat Commun* 2020; 11: 359.

66 Mertins P, Qiao JW, Patel J, Udeshi ND, Clauser KR, Mani DR *et al*. Integrated proteomic analysis of post-translational modifications by serial enrichment. *Nat Methods* 2013; 10: 634-637.

67 Andrews S. FastQC: A Quality Control Tool for High Throughput Sequence Data, 2015.

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

69 Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N *et al*. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* 2018; 19: 477.

70 Stuetzle THW. *Principal Curves*. Journal of the American Statistical Association, 1989.

71 Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol* 2010; 11: R106.

72 Hochberg YBY. *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing*, 1995.

73 Wu D, Smyth GK. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Res* 2012; 40: e133.

74 Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 2011; 27: 1739-1740.

75 Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* 2015; 1: 417-425.

76 Bluemn EG, Coleman IM, Lucas JM, Coleman RT, Hernandez-Lopez S, Tharakan R *et al*. Androgen Receptor Pathway-Independent Prostate Cancer Is Sustained through FGF Signaling. *Cancer Cell* 2017; 32: 474-489 e476.

77 Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA *et al*. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* 2012; 2: 401-404.

78 Steen CB, Liu CL, Alizadeh AA, Newman AM. Profiling Cell Type Abundance and Expression in Bulk Tissues with CIBERSORTx. *Methods Mol Biol* 2020; 2117: 135-157.

79 Li C, Menoret A, Farragher C, Ouyang Z, Bonin C, Holvoet P *et al*. Single cell transcriptomics based-MacSpectrum reveals novel macrophage activation signatures in diseases. *JCI Insight* 2019; 5.

80 Lin D, Wyatt AW, Xue H, Wang Y, Dong X, Haegert A *et al*. High fidelity patient-derived xenografts for accelerating prostate cancer discovery and drug development. *Cancer Res* 2014; 74: 1272-1283.

81 Zou H, Hastie T. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society Series B (Statistical Methodology)* 2005; 67: 301-320.

82 Kowalczyk MS, Tirosh I, Heckl D, Rao TN, Dixit A, Haas BJ *et al*. Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Res* 2015; 25: 1860-1872.

83 Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A *et al*. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol* 2019; 20: 163-172.

84 Shay T, Kang J. Immunological Genome Project and systems immunology. *Trends Immunol* 2013; 34: 602-609.

85 van Dijk D, Sharma R, Nainys J, Yim K, Kathail P, Carr AJ *et al*. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* 2018; 174: 716-729 e727.

86 Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK *et al*. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol* 2015; 16: 278.

87 Gillette MA, Satpathy S, Cao S, Dhanasekaran SM, Vasaikar SV, Krug K *et al*. Proteogenomic Characterization Reveals Therapeutic Vulnerabilities in Lung Adenocarcinoma. *Cell* 2020; 182: 200-225 e235.

88 Krug K, Jaehnig EJ, Satpathy S, Blumenberg L, Karpova A, Anurag M *et al*. Proteogenomic Landscape of Breast Cancer Tumorigenesis and Targeted Therapy. *Cell* 2020; 183: 1436-1456 e1431.

89 Gao D, Vela I, Sboner A, Iaquinta PJ, Karthaus WR, Gopalan A *et al*. Organoid cultures derived from patients with advanced prostate cancer. *Cell* 2014; 159: 176-187.

90 Abida W, Cyrta J, Heller G, Prandi D, Armenia J, Coleman I *et al*. Genomic correlates of clinical outcome in advanced prostate cancer. *Proc Natl Acad Sci U S A* 2019; 116: 11428-11436.

91 Barbieri CE, Baca SC, Lawrence MS, Demichelis F, Blattner M, Theurillat JP *et al*. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nat Genet* 2012; 44: 685-689.

92 Shoag J, Liu D, Blattner M, Sboner A, Park K, Deonarine L *et al*. SPOP mutation drives prostate neoplasia without stabilizing oncogenic transcription factor ERG. *J Clin Invest* 2018; 128: 381-386.

93 Bernasocchi T, El Tekle G, Bolis M, Mutti A, Vallerga A, Brandt LP *et al*. Dual functions of SPOP and ERG dictate androgen therapy responses in prostate cancer. *Nat Commun* 2021; 12: 734.

94 Xu K, Wu ZJ, Groner AC, He HH, Cai C, Lis RT *et al*. EZH2 oncogenic activity in castration-resistant prostate cancer cells is Polycomb-independent. *Science* 2012; 338: 1465-1469.

95 Yu J, Yu J, Mani RS, Cao Q, Brenner CJ, Cao X *et al*. An integrated network of androgen receptor, polycomb, and TMPRSS2-ERG gene fusions in prostate cancer progression. *Cancer Cell* 2010; 17: 443-454.

96 Varambally S, Dhanasekaran SM, Zhou M, Barrette TR, Kumar-Sinha C, Sanda MG *et al*. The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature* 2002; 419: 624-629.

97 Wang Q, Li W, Zhang Y, Yuan X, Xu K, Yu J *et al*. Androgen receptor regulates a distinct transcription program in androgen-independent prostate cancer. *Cell* 2009; 138: 245-256.

98 Pomerantz MM, Li F, Takeda DY, Lenci R, Chonkar A, Chabot M *et al*. The androgen receptor cistrome is extensively reprogrammed in human prostate tumorigenesis. *Nat Genet* 2015; 47: 1346-1351.

99 Pomerantz MM, Qiu X, Zhu Y, Takeda DY, Pan W, Baca SC *et al*. Prostate cancer reactivates developmental epigenomic programs during metastatic progression. *Nat Genet* 2020; 52: 790-799.

100 Iglesias-Gato D, Thysell E, Tyanova S, Crnalic S, Santos A, Lima TS *et al*. The Proteome of Prostate Cancer Bone Metastasis Reveals Heterogeneity with Prognostic Implications. *Clin Cancer Res* 2018; 24: 5433-5444.

101 Federer-Gsponer JR, Muller DC, Zellweger T, Eggimann M, Marston K, Ruiz C *et al*. Patterns of stemness-associated markers in the development of castration-resistant prostate cancer. *Prostate* 2020; 80: 1108-1117.

102 Di Mitri D, Mirenda M, Vasilevska J, Calcinotto A, Delaleu N, Revandkar A *et al*. Re-education of Tumor-Associated Macrophages by CXCR2 Blockade Drives Senescence and Tumor Inhibition in Advanced Prostate Cancer. *Cell Rep* 2019; 28: 2156-2168 e2155.

103 Kowal J, Kornete M, Joyce JA. Re-education of macrophages as a therapeutic strategy in cancer. *Immunotherapy* 2019; 11: 677-689.

104 Barkal AA, Brewer RE, Markovic M, Kowarsky M, Barkal SA, Zaro BW *et al*. CD24 signalling through macrophage Siglec-10 is a target for cancer immunotherapy. *Nature* 2019; 572: 392-396.

105 Pauli C, Hopkins BD, Prandi D, Shaw R, Fedrizzi T, Sboner A *et al*. Personalized In Vitro and In Vivo Cancer Models to Guide Precision Medicine. *Cancer Discov* 2017; 7: 462-477.

106 Akamatsu S, Wyatt AW, Lin D, Lysakowski S, Zhang F, Kim S *et al*. The Placental Gene PEG10 Promotes Progression of Neuroendocrine Prostate Cancer. *Cell Rep* 2015; 12: 922-936.

107 Beshiri ML, Tice CM, Tran C, Nguyen HM, Sowalsky AG, Agarwal S *et al*. A PDX/Organoid Biobank of Advanced Prostate Cancers Captures Genomic and Phenotypic Heterogeneity for Disease Modeling and Therapeutic Screening. *Clin Cancer Res* 2018; 24: 4332-4345.

108 Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, 3rd *et al*. Comprehensive Integration of Single-Cell Data. *Cell* 2019; 177: 1888-1902 e1821.

109 Berger A, Brady NJ, Bareja R, Robinson B, Conteduca V, Augello MA *et al*. N-Myc-mediated epigenetic reprogramming drives lineage plasticity in advanced prostate cancer. *J Clin Invest* 2019; 129: 3924-3940.

110 Mu P, Zhang Z, Benelli M, Karthaus WR, Hoover E, Chen CC *et al*. SOX2 promotes lineage plasticity and antiandrogen resistance in TP53- and RB1-deficient prostate cancer. *Science* 2017; 355: 84-88.

111 Winters B, Brown L, Coleman I, Nguyen H, Minas TZ, Kollath L *et al*. Inhibition of ERG Activity in Patient-derived Prostate Cancer Xenografts by YK-4-279. *Anticancer Res* 2017; 37: 3385-3396.

112 Siltari A, Syvala H, Lou YR, Gao Y, Murtola TJ. Role of Lipids and Lipid Metabolism in Prostate Cancer Progression and the Tumor's Immune Environment. *Cancers (Basel)* 2022; 14.

113 Binnewies M, Roberts EW, Kersten K, Chan V, Fearon DF, Merad M *et al*. Understanding the tumor immune microenvironment (TIME) for effective therapy. *Nat Med* 2018; 24: 541-550.

114 Hall ME, Huelster HL, Luckenbaugh AN, Laviana AA, Keegan KA, Klaassen Z *et al*. Metastatic Hormone-sensitive Prostate Cancer: Current Perspective on the Evolving Therapeutic Landscape. *Onco Targets Ther* 2020; 13: 3571-3581.

115 Kyriakopoulos CE, Chen YH, Carducci MA, Liu G, Jarrard DF, Hahn NM *et al*. Chemohormonal Therapy in Metastatic Hormone-Sensitive Prostate Cancer: Long-Term Survival Analysis of the Randomized Phase III E3805 CHAARTED Trial. *J Clin Oncol* 2018; 36: 1080-1087.

116 Petrylak DP, Tangen CM, Hussain MH, Lara PN, Jr., Jones JA, Taplin ME *et al*. Docetaxel and estramustine compared with mitoxantrone and prednisone for advanced refractory prostate cancer. *N Engl J Med* 2004; 351: 1513-1520.

117 Gandaglia G, Fossati N, Suardi N, Montorsi F, Briganti A. STAMPEDE trial and patients with non-metastatic prostate cancer. *Lancet* 2016; 388: 234-235.

118 Clarke NW, Ali A, Ingleby FC, Hoyle A, Amos CL, Attard G *et al*. Addition of docetaxel to hormonal therapy in low- and high-burden metastatic hormone sensitive prostate cancer: long-term survival results from the STAMPEDE trial. *Ann Oncol* 2019; 30: 1992-2003.

119 Sweeney CJ, Chen YH, Carducci M, Liu G, Jarrard DF, Eisenberger M *et al*. Chemohormonal Therapy in Metastatic Hormone-Sensitive Prostate Cancer. *N Engl J Med* 2015; 373: 737-746.

120 Tannock IF, de Wit R, Berry WR, Horti J, Pluzanska A, Chi KN *et al*. Docetaxel plus prednisone or mitoxantrone plus prednisone for advanced prostate cancer. *N Engl J Med* 2004; 351: 1502-1512.

121 Yu J, Yu J, Rhodes DR, Tomlins SA, Cao X, Chen G *et al*. A polycomb repression signature in metastatic prostate cancer predicts cancer outcome. *Cancer Res* 2007; 67: 10657-10663.

122 Xiao L, Tien JC, Vo J, Tan M, Parolia A, Zhang Y *et al*. Epigenetic Reprogramming with Antisense Oligonucleotides Enhances the Effectiveness of Androgen Receptor Inhibition in Castration-Resistant Prostate Cancer. *Cancer Res* 2018; 78: 5731-5740.

123 Escamilla J, Schokrpur S, Liu C, Priceman SJ, Moughon D, Jiang Z *et al*. CSF1 receptor targeting in prostate cancer reverses macrophage-mediated resistance to androgen blockade therapy. *Cancer Res* 2015; 75: 950-962.

124 Karthaus WR, Hofree M, Choi D, Linton EL, Turkekul M, Bejnood A *et al*. Regenerative potential of prostate luminal cells revealed by single-cell analysis. *Science* 2020; 368: 497-505.

125 Laudato S, Aparicio A, Giancotti FG. Clonal Evolution and Epithelial Plasticity in the Emergence of AR-Independent Prostate Carcinoma. *Trends Cancer* 2019; 5: 440-455.

126 Linja MJ, Visakorpi T. Alterations of androgen receptor in prostate cancer. *J Steroid Biochem Mol Biol* 2004; 92: 255-264.

127 Antonarakis ES, Lu C, Wang H, Luber B, Nakazawa M, Roeser JC *et al*. AR-V7 and resistance to enzalutamide and abiraterone in prostate cancer. *N Engl J Med* 2014; 371: 1028-1038.

128 Sadeesh N, Scaravilli M, Latonen L. Proteomic Landscape of Prostate Cancer: The View Provided by Quantitative Proteomics, Integrative Analyses, and Protein Interactomes. *Cancers (Basel)* 2021; 13.

129 Intasqui P, Bertolla RP, Sadi MV. Prostate cancer proteomics: clinically useful protein biomarkers and future perspectives. *Expert Rev Proteomics* 2018; 15: 65-79.

130 Sinha A, Huang V, Livingstone J, Wang J, Fox NS, Kurganovs N *et al*. The Proteogenomic Landscape of Curable Prostate Cancer. *Cancer Cell* 2019; 35: 414-427 e416.

131 De Vargas Roditi L, Jacobs A, Rueschoff JH, Bankhead P, Chevrier S, Jackson HW *et al*. Single-cell proteomics defines the cellular heterogeneity of localized prostate cancer. *Cell Rep Med* 2022; 3: 100604.

132 Lygirou V, Fasoulakis K, Stroggilos R, Makridakis M, Latosinska A, Frantzi M *et al*. Proteomic Analysis of Prostate Cancer FFPE Samples Reveals Markers of Disease Progression and Aggressiveness. *Cancers (Basel)* 2022; 14.

133 Latonen L, Afyounian E, Jylha A, Nattinen J, Aapola U, Annala M *et al*. Integrative proteomics in prostate cancer uncovers robustness against genomic and transcriptomic aberrations during disease progression. *Nat Commun* 2018; 9: 1176.

134 Chen G, Gharib TG, Huang CC, Taylor JM, Misek DE, Kardia SL *et al*. Discordant protein and mRNA expression in lung adenocarcinomas. *Mol Cell Proteomics* 2002; 1: 304-313.

135 Liu Y, Beyer A, Aebersold R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* 2016; 165: 535-550.

136 Koussounadis A, Langdon SP, Um IH, Harrison DJ, Smith VA. Relationship between differentially expressed mRNA and mRNA-protein correlations in a xenograft model system. *Sci Rep* 2015; 5: 10775.

137 Mann M, Jensen ON. Proteomic analysis of post-translational modifications. *Nat Biotechnol* 2003; 21: 255-261.

138 Aebersold R, Mann M. Mass-spectrometric exploration of proteome structure and function. *Nature* 2016; 537: 347-355.