Changing Selection And Epistasis On Drivers Over Tumorigenesis In Primary And Metastatic Prostate Cancer

Alexander Yang

Follow this and additional works at: https://elischolar.library.yale.edu/ymtdl

Part of the Medicine and Health Sciences Commons

Recommended Citation

This Open Access Thesis is brought to you for free and open access by the School of Medicine at EliScholar – A Digital Platform for Scholarly Publishing at Yale. It has been accepted for inclusion in Yale Medicine Thesis Digital Library by an authorized administrator of EliScholar – A Digital Platform for Scholarly Publishing at Yale. For more information, please contact elischolar@yale.edu.
Changing selection and epistasis on drivers over tumorigenesis in primary and metastatic prostate cancer

A Thesis Submitted to the Yale University School of Medicine in Partial Fulfillment of the Requirements for the Degree of Doctor of Medicine

by

Alexander Yang, 2022
Abstract

Background: Many attempts have been made to characterize and describe the driver genes and mutations responsible for prostate cancer tumorigenesis. We have quantified the cancer effect size—a direct measurement of the survival advantage a mutation confers—for 2699 primary and metastatic prostate tumor samples. Our measure of cancer effect treats tumorigenesis as an evolutionary process, subject to positive and negative selective pressures. We have applied this metric in a stage-specific manner to elucidate which mutations are selected for as prostate cancer develops.

Methods: We analyzed 2699 prostate cancer tumor exomes, genomes, and panel sequences (1648 primary tumors and 1051 metastatic samples). The Gleason grade groups were used to further divide the primary tumors into lower-risk (I/II) and higher-risk (III/IV/V) primary tumors. The deconstructSigs, dNdScv, and cancereffectsizeR packages were used for extraction of mutational signatures, calculation of gene mutation rate, and calculation of cancer effect sizes of somatic variants respectively. Furthermore, using a model of pairwise epistasis, we investigated pairs of genes, observing the effect that the presence or absence of a mutation in one gene has on the selection for mutations in the paired gene.

Results: The lower-risk, higher-risk, and metastatic tumors showed very similar mutational signatures, with the underlying gene mutation rates generally increasing from lower-risk tumors to higher-risk tumors to metastatic tumors. However, the genes and somatic variants that were most highly selected-for within each cohort were notably different. A distinct set of genes featured mutations that were selected for in the metastatic samples. Pairwise epistasis analysis suggested that there is an early role for SPOP in the development of prostate cancer: SPOP mutations increase the selection for mutations in several other tumor suppressors and oncogenes.

Conclusion: Application of cancer effect size analysis highlights which mutations and genes are selected for leading up to each stage of prostate cancer, emphasizing the genetic differences between lower-risk, higher-risk, metastatic prostate cancer. By incorporating pairwise epistasis analysis, we are able to support previous and hypothesize novel gene-gene interactions occurring during tumorigenesis.
Acknowledgements

I would like to thank Dr. Jeffrey Townsend for his mentorship throughout my years here at the Yale School of Medicine. He has had an enormous impact on my research journey, and I will carry the lessons I have learned under his guidance into my future research endeavors.

I would also like to thank other members of the Townsend lab, including but not limited to Dr. Christopher Cross, Dr. Vincent Cannatoro, Jeffrey Mandell, and J. Nick Fisk for their invaluable help on this project.

Lastly, I would like to show my appreciation for my family and friends, without whom both this thesis and my journey throughout medical school would not be possible.

Research reported in this publication was supported by the National Institute on Diabetes and Digestive and Kidney Diseases of the National Institutes of Health under Award Number T35DK104689. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.
# Table of Contents

Abstract ................................................................................................................................... i

Acknowledgements ................................................................................................................ ii

Table of Contents ................................................................................................................... 1

Introduction ........................................................................................................................... 2

Methods ............................................................................................................................... 13

Statement of Purpose ........................................................................................................... 17

Results .................................................................................................................................. 18

Discussion ............................................................................................................................ 34

Conclusion ........................................................................................................................... 40

References ............................................................................................................................ 41
Introduction

Background

Prostate cancer is the second most frequent cancer in men, with over 1.4 million new cases comprising 14.1% of all new cancer incidences among men in 2020. Furthermore, prostate cancer is the cancer with the fifth highest mortality, causing over 375,000 deaths world-wide in 2020\(^1\). While lung cancer is more frequent overall, prostate cancer is the cancer with the highest incidence in 112 countries. The overall 5-year survival of all prostate cancer is the highest of all cancers at 98%, which is partially attributable to widespread screening leading to the discovery of early, localized prostate cancer\(^1\). However, those who are diagnosed with distant metastases have only a 5-year survival of 30%\(^2\). Treatment of localized prostate cancer also decreases the risk of developing metastases, which greatly improves overall survival of these patients\(^3\).

Only three risk factors, all of which are non-modifiable, have been firmly associated with prostate cancer—age, race, and family history of prostate cancer\(^4\). The incidence of prostate cancer dramatically increases with age and over 85% of newly diagnosed patients are over the age of 60, and thus prostate cancer is usually highest in countries with a high human development index (HDI), particularly the United States and the United Kingdom\(^2\). However, Asian countries with a high HDI, such as Japan and South Korea, have about a ten-fold smaller incidence rate\(^5\). While Asian-Americans are less likely to be diagnosed with prostate cancer compared to Caucasians, African-Americans are 1.6 times as likely to be diagnosed and 2–3 times as likely to die from prostate cancer\(^6\). Furthermore, family history plays a large role in prostate cancer risk; men with first-degree relatives diagnosed with prostate cancer are twice as likely to develop the disease, and roughly 9% of individuals with prostate cancer have a family history of cancer\(^2\).
There are several additional risk factors that have been proposed or weakly associated with prostate cancer. Tobacco and alcohol are common carcinogens that have been strongly associated with many types of cancer, but neither have a strong association with the incidence of prostate cancer. A meta-analysis of 24 studies showed that in the pooled data, there was no change in prostate cancer incidence (RR: 1.04; 95% CI: 0.87–1.24), but when the data was stratified by amount the participants smoked, there was a statistically significant increase in prostate cancer incidence (RR: 1.11; 95% CI: 1.01–1.22), in addition to a higher mortality in current smokers (RR: 1.14; 95% CI 1.06–1.19). There have been inconsistent and heterogeneous results with regard to alcohol’s association with prostate cancer. A meta-analysis of 50 case-control studies and 22 cohort studies found that there was a borderline statistical association with alcohol consumption (RR: 1.06; 95% CI: 1.01–1.10); however, the authors concluded that this relationship had little clinical significance and did not show a material association between alcohol consumption and prostate cancer.

Dietary factors are another set of risks that have been explored with their regard to prostate cancer. The CLUE II study involved food frequency questionnaires and found that prostate cancer was associated with overall consumption of processed meat, but not total or red meat. It has been proposed that it is not the consumption of meat that increases prostate cancer risk, but actually the consumption of meat cooked at a high temperature and the aromatic hydrocarbons and mutagenic heterocyclic amines produced thereof. A large, prospective NIH-AARP study showed a positive correlation with meat cooked at high temperature and prostate cancer in African-American men (HR: 1.22, 95% CI: 1.03–1.44, for the highest uptake quartile). Because of the lower incidence of prostate cancer in high-HDI East Asian countries compared to their western counterparts, there have been studies on regional foods, such as soy products and green tea. A meta-analysis of 14 studies on soy consumption and prostate cancer risk showed a
26% reduction in risk of prostate cancer when the highest reported intake group was compared to the lowest\textsuperscript{10}. Another meta-analysis of 13 case-control studies showed a borderline significant inverse relationship between green tea consumption and prostate cancer risk\textsuperscript{11}. Interestingly, this meta-analysis did not find any relationship between black tea consumption and prostate cancer\textsuperscript{11}.

**Screening and Diagnosis**

Prostate cancer is primarily screened for using serum levels of prostate-specific antigen (PSA). Screening can be especially useful in prostate cancer, as early and localized prostate cancer is often asymptomatic but curable. The European Randomized Study for Screening for Prostate Cancer was initiated in the early 90s, featuring 162,243 men between the ages of 55 and 69, and it showed that PSA screening reduced mortality from prostate cancer by 20\%, but this was associated with a high rate of overdiagnosis\textsuperscript{12}. Overdiagnosis from PSA-based screening is due to PSA not being specific to prostate cancer; it can also be elevated in prostatitis or benign prostatic hyperplasia\textsuperscript{13}.

In 2008, the United States Preventative Services Taskforce (USPSTF) issued a Grade D recommendation (i.e. the harms outweigh the benefits) against PSA screening in men over the age of 75\textsuperscript{14}, and in 2012, the USPSTF extended this Grade D recommendation against PSA screening in all men regardless of age\textsuperscript{15}. This time period, 2008–2017, saw a decrease in PSA screening and prostate cancer incidence\textsuperscript{16}, as well as a decrease in prostate biopsy and radical prostatectomy\textsuperscript{17}. However, from 2008–2013, there was an increase in men above the age of 75 presenting with distant metastases at the time of prostate cancer diagnosis. Ultimately in 2017, the USPSTF rescinded its previous Grade D recommendation against PSA-testing in all men, continuing a Grade D recommendation against PSA-testing in men above the age of 70 and issuing a Grade C recommendation for men aged 55–69, essentially saying that PSA testing should be
offered in select patient populations (e.g. those with a family history) and each patient should discuss with his clinician about the benefits and harms\textsuperscript{18}.

There are parameters of PSA testing that are currently being studied to make PSA testing not only more specific to prostate cancer, but also a better indicator of risk and prognosis. For example, PSA velocity (PSAV) measures the change in PSA over time. In men without prostate cancer, PSAV increases by $\sim0.15$ ng/mL/year as opposed to 0.35-0.40 ng/mL/year\textsuperscript{19}. Furthermore, a PSAV greater than 2 ng/mL/year that is not caused by prostatitis may indicate that the disease is incurable\textsuperscript{20}. PSA density (PSAD) may also be used; it is simply PSA divided by prostate volume, factoring in that a larger prostate may produce more PSA\textsuperscript{19}. One additional variation of PSA is the free PSA-to-total PSA ratio, which is the (free PSA/total PSA). A free PSA-to-total PSA ratio less than 0.1 means there is a greater than 50% chance of a biopsy revealing prostate cancer, while a ratio greater than 0.25 is associated with a lower than 10% chance of a prostate cancer diagnosis\textsuperscript{21}. These additional PSA tests increase specificity and enable physicians to obtain a better picture of each individual patient’s clinical outlook.

In addition to PSA serum tests, digital rectal examinations (DRE) are also used to assess for any prostate irregularities or nodules. Irregularities in either or both PSA levels and DRE are indications for a transrectal ultrasound (TRUS) guided prostate biopsy\textsuperscript{22}. Currently, standard of care for diagnosis of prostate cancer involves taking 12 core samples under TRUS-guidance using an 18G needle and caudal block. While a transperineal approach can be taken and results in a similar rate of cancer detection, it requires spinal anesthesia, which can result in headaches as a side effect\textsuperscript{23}. Multiparametric magnetic resonance imaging (mpMRI) is used to guide prostate biopsy. It has a similar rate of overall prostate cancer detection as TRUS-guided biopsy. However, it has a higher rate of detection for significant cancer and a lower rate of detection for insignificant cancer\textsuperscript{24}. Systematic mpMRI prior to prostate biopsy has also
been suggested as a triage test to better determine who should get a TRUS-guided prostate biopsy\textsuperscript{25}.

Each of the prostate biopsy cores is then characterized histopathologically via its Gleason score. The Gleason scoring system is named after Dr. Donald Gleason, Chief of Pathology at the Veteran’s Hospital in Minnesota, who is credited with creating this system in the 1960s to assign a score to each prostate sample’s histology\textsuperscript{26}. The Gleason score is reported as the sum of two numbers, usually written $X + Y = Z$, where $X$ is the score primary pattern, $Y$ is the score of the secondary pattern, and $Z$ is the overall Gleason score\textsuperscript{27}. However, changes have been proposed and implemented to the original scoring system: two prostate tumors that have the same Gleason score but different primary and secondary patterns can have markedly different outcomes. Specifically, a tumor with the score of $3+4$ has a 4-year biochemical recurrence-free survival (BFS) of 82.7\%, while a $4+3$ tumor has a 4-year (BFS) of 65.1\%, even though both tumors have a Gleason score of 7. BFS may encompass several definitions, but in this study was defined by any post-operative serum PSA $> 0.2$ ng/mL. Because of the wide variety of outcomes based on different Gleason scores, Pierorazio et al proposed a modified system of prognostic Gleason grade groups in 2013\textsuperscript{28}. These changes were later discussed at the 2014 International Society of Urological Pathology Consensus Conference in an attempt to update prostate cancer grading, and the new Gleason grade groups were agreed upon in order to provide a more accurate stratification of tumors and simplify understanding of prognosis (i.e. Grade Groups 1–5)\textsuperscript{29}.

In addition to the Gleason score/grade groups, prostate cancer can be staged using the traditional TNM staging, in which $T$ represents the size of the main tumor, $N$ refers to the local lymph nodes affected, and $M$ is an indication of distant metastases. TNM staging is tailored uniquely to each cancer, and as such, the TNM staging for
prostate cancer also takes into account the Gleason grade group and serum PSA levels for the staging group30.

Treatment

Treatment for patients diagnosed with localized prostate cancer (no local lymph node involvement or distant metastases) are treated via (1) active surveillance, (2) surgery, in the form of a radical prostatectomy, or (3) radiotherapy2,31. Active surveillance consists of a schedule of follow-up PSA tests, physical examinations, and prostate biopsies to monitor for worsening prostate cancer, with the intent of curing those who develop worsening disease; it is used in low-risk patients with a greater than 10-year life expectancy to minimize treatment-related toxicity. Active surveillance is different from watchful waiting, which has palliative intent, no pre-defined follow-up schedule, and can be applied to any patient with estimated <10-year survival32. A prospective study of 1298 favorable risk patients undergoing active surveillance found that a 10-year cancer-specific survival rate and metastasis-free survival rate to be 99.9% and 99.4%, respectively33. Furthermore, the ProtecT trial looked at 2664 men with localized prostate who had been randomized into either active surveillance, surgery, or radiotherapy. The study found that after 10 years, there was no significant difference in prostate cancer-specific mortality, although surgery and radiotherapy was found to have lower incidence rates of disease progression and metastasis compared to active surveillance34.

Surgery and radiotherapy are both used to cure localized prostate cancer. These are usually offered to patients with intermediate risk prostate cancer and a 10-year expected survival of >10 years. While the ProtecT trial found the 10-year outcomes in both prostate cancer-specific survival as well as disease progression free survival to be comparable between surgery and radiotherapy, each was associated with different side effects; radiotherapy was associated with nocturia and bowel dysfunction, while surgery
was associated with worse urinary control and sexual function\textsuperscript{34}. Both these modalities are undergoing constant evolution and improvement. For example, the open radical prostatectomy is now being replaced by robotic radical prostatectomy, which has been found to have a better 1-year outcomes in both urinary continence and sexual function\textsuperscript{35,36}. Furthermore, intensity-modulated radiation therapy (IMRT) has been increasingly used over 3D-conformal radiation therapy. IMRT allows the radiation dose to better conform to the shape of the tumors, decreasing the dose of radiation given to the other organs at risk, such as the bladder and the rectum\textsuperscript{37,38}. Hypofractionation is a schedule for radiotherapy that gives increased doses of radiation in fewer treatment sessions, thus decreasing the length of treatment. Although hypofractionation schedules have been shown to have similar efficacies of traditional schedules\textsuperscript{39}, they are associated with more acute gastrointestinal and genitourinary toxicity\textsuperscript{40}.

For more advanced and metastatic cancer, androgen deprivation therapy (ADT) has been the mainstay treatment. Dating back to the 18\textsuperscript{th} century, there have been animal studies done that have shown a link between the pituitary, the testes, and the prostate\textsuperscript{41}. In 1941, Charles Huggins and Clarence Hodges at the University of Chicago were the first to start treating prostate cancer patients through ADT in the form of either surgical castration or medication castration via estrogen injection\textsuperscript{42}. The discovery that prostate cancer could be treated via hormonal control led to larger studies, one of which was conducted by the Veterans Administration Cooperative Urologic Research Group (VACURG) in the 1960s and concluded that treating prostate cancer with oral diethylstilbestrol (DES) was as effective as orchiectomy\textsuperscript{43}. However, this VACURG study found significant cardiovascular and thromboembolic side effects associated with increased estrogen. Working on a different part of the hormonal axis, Andrew Schally and his colleagues at Tulane university discovered the structure of and synthesized luteinizing hormone-releasing hormone (LHRH)\textsuperscript{44} and later showed that patients with
advanced prostate cancer who received daily LHRH agonists showed a 75% reduction in their testosterone levels by the third week of treatment, in addition to a significant reduction in metastases-related bone pain, showing that LHRH agonists were a viable alternative to orchiectomy or estrogen-derived medical castration\textsuperscript{45}.

Today, ADT, in the form of LHRH agonists, are the gold standard for metastatic prostate cancer, until there is biochemical and/or radiographic evidence of disease progression\textsuperscript{46}. Because ADT has several adverse effects, such as metabolic changes, sexual dysfunction, anemia, fatigue, decreased bone mineral density, and decreased muscle mass\textsuperscript{47}, there have been studies looking into an intermittent ADT schedule. A meta-analysis demonstrated non-inferiority of intermittent ADT compared to continuous ADT in overall survival, cancer-specific survival, and progression-free survival; and an improvement of physical and sexual function was reported with intermittent ADT\textsuperscript{48}. Disease that responds to ADT is defined as metastatic castration-sensitive prostate cancer (mCSPC).

Unfortunately, ADT is usually only palliative, as the disease often develops resistance to ADT within 2–3 years\textsuperscript{49}. Once metastatic prostate cancer stops responding to ADT, it is termed metastatic castration-resistant prostate cancer (mCRPC). While docetaxel was traditionally started after patients did not respond to ADT, there have been recent trials that assessed whether starting it earlier with ADT would improve outcomes. The CHAARTED and STAMPEDE trials demonstrated that the addition of docetaxel at the start of ADT for men with advanced or metastatic prostate cancer resulted in a significant increase in overall survival, in addition to an increase in adverse events\textsuperscript{50,51}. In addition to docetaxel, there are several additional drugs such as cabazitaxel\textsuperscript{52}, abiraterone\textsuperscript{53,54}, enzalutamide\textsuperscript{55,56}, and radium-\textsuperscript{223}\textsuperscript{57} that are currently being studied and used for metastatic prostate cancer.
There are also immunotherapy-based treatments, essentially “cancer vaccines”, are being researched for the treatment of prostate cancer, such as Sipuleucel-T, which consists of peripheral blood mononuclear cells that have been stimulated *ex vivo* with PA2020, a recombinant protein of prostate antigen and granulocyte-macrophage colony stimulating factor. The IMPACT trial was a phase III trial that showed that Sipuleucel-T increased median survival by 4 months in patients with metastatic castration-resistant prostate cancer, but did not increase time to disease progression. PROSTVAC is another immunotherapy which consists of two different poxviruses into which transgenes for PSA and three different costimulatory molecules have been inserted; the end goal of this therapy then is to increase the immunogenicity of PSA and trigger a targeted immune response. While this treatment showed an improvement in overall survival in Phase II trials, it failed to show the same benefit in a Phase III trial.

**Genetics**

With the advent of high-throughput sequencing over the last decade, many genomics-driven studies have explored the molecular landscape of PC using various techniques and approaches that highlight its complex heterogeneity. Moreover, a set of common PC-specific oncogenes and their associated genomic aberration have emerged, including point mutations in *SPOP*, *TP53*, and *FOXA1*; amplifications and copy number variants of the androgen receptor, *AR* (notably in castration-resistance prostate cancer); mutations in DNA–repair genes and cell–signaling genes such as *BRCA*, *PI3K*, *PTEN*, and *MYC*; gene fusions involving the ETS gene family; and germline variants in susceptibility loci associated with predisposition to prostate cancer development and metastatic progression including *ATM*. Furthermore, the WNT signaling pathway is often mutated in mCRPC, but not localized disease, involving alterations in genes such as *CTNNB1*, *SMAD4*, and *APC*.
The androgen receptor, AR, is of enormous importance throughout the pathogenesis of metastatic prostate cancer. Prostate cells (including cancer cells) all express the androgen receptor and are induced to proliferate by activated AR and its interaction with downstream signaling molecules. There are several mechanisms by which AR is altered following the induction of ADT, leading to castration resistance. mCRPC is characterized by both gain-of-function mutations and copy number amplifications of AR or regulators of AR (such as FOXA1), in addition to inactivating mutations and deletions of genes that repress AR signaling. In addition, changes to the post-translational modification of AR can also lead to constitutive activity in the absence of testosterone. Androgen receptors can also be alternatively spliced, leading to AR variants that lack the ligand-binding moiety and have weak constitutive activation. Over 70% of patients with mCRPC have alterations in AR, compared to only 2–6% patients with mCSPC having AR alterations.

Another pathway that is often affected in the development of prostate cancer is the PI3K pathway, including mutations in PIK3CA and PIK3CB, as well as AKT1, which is phosphorylated and activated by the directly upstream PI3 kinase. PTEN, a tumor suppressor that controls and suppresses activity of the PI3K pathway, is mutated and/or deleted in 30–60% of high-grade prostate cancers. Unregulated activation of the PI3K pathway, either by gain-of-function mutations in activators (PI3 kinases, AKT1) or mutation/deletions in suppressors (such as PTEN), results in uncontrolled tumor cell growth, survival, proliferation, and migration that is strongly implicated in progression to mCRPC.

Cancer Effect Size

Genomic studies in cancer have previously used P value, prevalence, and mutation rate to determine the oncogenes that have been used to define stages and subtypes of prostate cancer. However, these metrics are related to, but do not capture the effect that
somatic variants have on each other or on the processes of cancer development and progression\textsuperscript{88}. For example, the prevalence of mutations in a certain gene is subject to many factors unrelated to its effect on cancer cell lineage phenotype, such as the underlying mutation rate, size of the gene, and access to DNA repair. \( P \) values are a threshold for significance, rather than a quantification of effect\textsuperscript{89–91}.

In this thesis, I convey the evolutionary trajectory of prostate cancer, identifying the time-ordered and interdependent importance of mutations occupying key roles in the development of prostate cancer, all within a somatic evolutionary genetic approach using the cancereffectsize\texttt{R} software package, developed by the Townsend Lab at the Yale School of Public Health\textsuperscript{88}. This approach quantifies the cancer effect size (CES), a direct measurement of the survival advantage a mutation confers within an evolutionary process, subject to selective pressures\textsuperscript{88}. This basic approach has been applied to esophageal cancer\textsuperscript{92}, bladder cancer\textsuperscript{93,94}, and even prostate cancer\textsuperscript{95}. However, this CES method has not yet been performed using in a time-ordered manner, separating the tumors into different cohorts which can be presumed to be from different timepoints (i.e., localized vs metastatic). CES has also not been applied to perform epistatic analyses, investigating the interactions that the presence or absence of a mutation in a certain gene has on the selection for a mutation in another gene. Both these novel features of CES analysis are used to give a deeper insight into how interactions between genes lead to development of mutations that are responsible for the progression and development of metastatic prostate cancer.
Methods

Dataset and analysis

2,699 whole exome sequences (WES), whole genome sequences (WGS), and panel sequences for prostate tumors were obtained and assembled from four primary sources. In total, there were 1,648 unique primary tumors and 1,051 metastatic tumors. We categorized tumors with a Gleason score of 6 or 7 (3+4) as lower-risk (n = 479), and tumors with a Gleason score of 7 (4+3), 8, 9, or 10 into higher-risk (n = 406). This categorization is consistent with findings that tumors with a Gleason score of 7 have been shown to have very different prognoses depending if the score is 3+4 or 4+3. Correspondingly, our primary tumors are categorized into ISUP Gleason Grade Groups I/II and III/IV/V. There were 763 primary tumors for which no Gleason score had been attributed. These tumors were therefore omitted from risk classification and all analyses dependent on risk classification.

Mutation rates and signatures

We followed cancereffectisizeR's standard workflow, which leverages the deconstructSigs and dNdScv packages for mutational signature extraction and gene mutation rate calculation, respectively. Gene mutation rate calculation was informed by PRAD tissue-specific covariates included in the reference data available with cancereffectisizeR. Rates were calculated separately for low-risk and high-risk tumor groups. For mutational signature analysis, we attributed COSMIC v3.1 signatures to tumors using cancereffectisizeR’s trinuc_mutation_rates function; signatures that have been previously determined to be absent in PRAD were excluded from signature extraction.
Cancer effect sizes

The effect size, or scaled selection coefficient, of a point mutation is defined as the observed frequency of substitutions divided by their expected frequency, or $\gamma = \frac{\lambda}{\mu}$, under the assumption that observed mutations are fixed in the cancer cell population\textsuperscript{88}. We calculated effect sizes with cancereffectsize\textsuperscript{R}, version 2.2.0, using genomic reference data from the accompanying data package ces.refset.hg19, version 1.0.0. Effect sizes of SNVs were calculated using the model of stage-specific selection provided by cancereffectsize\textsuperscript{R}, which parameterized selection on newly arising variants as serially dynamic, from normal to lower-risk to higher-risk to metastatic tissue. To maximize power to observe differences in selection pressure on genes across risk groups, some effect sizes were estimated at the level of the gene, oncogenic gene region, or oncogenic domain instead of at the level of the SNV, using the cancereffectsize\textsuperscript{R} compound_variants feature. When possible, recurrent variants were included based on whether they matched a known oncogenic domain within the gene, including SPOP\textsuperscript{105}, AR,\textsuperscript{106} PIK3CA\textsuperscript{107}, CTNNB1\textsuperscript{108}, and TP53\textsuperscript{109}. The cancereffectsize\textsuperscript{R} R software package is available at https://github.com/Townsend-Lab-Yale/cancereffectsize\textsuperscript{R}.

Protein structural modeling

The structural model for the SPOP protein was sourced from the Research Collaboratory for Structural Bioinformatics Protein Data Bank (UID: 6I41)\textsuperscript{110}. The structural model for the ligand-binding domain of the androgen receptor was sourced from the RCSB PDB as well (UID:3L3X)\textsuperscript{111}. The protein structure was then visualized and important amino-acid residues were highlighted using the UCSF Chimera software\textsuperscript{112}. 
Epistasis analysis

Under a model of pairwise epistasis, we allow the scaled selection coefficients of a pair of variants to vary depending on the presence or absence of the other variant. For variants \(A\) and \(B\), there are four effect size parameters: \(\gamma_{A0}\), the effect of \(A\) in the absence of \(B\); \(\gamma_{AB}\), the effect of \(A\) in the presence \(B\); \(\gamma_{B0}\), the effect of \(B\) in the absence of \(A\); and \(\gamma_{BA}\), the effect of \(B\) in the presence of \(A\).

To derive the probabilities of observing each genotype in a sample, we simplify notation by defining rate \(\lambda_{A0} = \mu_A \gamma_{A0}\), etc. At time \(t\), \(P_{ab}\), the probability of lacking variants \(A\) and \(B\), is the product of the probabilities of zero fixation events in each:

\[
P_{ab}(t) = P_a(t)P_b(t) = e^{-\left(\lambda_{A0} + \lambda_{B0}\right)t}.
\]

To derive the probability of observing variant \(A\) but not \(B\), we begin with a differential equation justified by our assumption that fixation events are irreversible:

\[
P'_{Ab}(t) = P_{ab}(t)\lambda_{A0} - P_{ab}(t)\lambda_{BA}.
\]

We solve for \(P_{ab}(t)\) by rearranging, substituting for \(P_{ab}(t)\), and integrating using an integrating factor of \(e^{\lambda_{BA}t}\):

\[
\int e^{\lambda_{BA}t}(P'_{Ab}(t) + \lambda_{BA}P_{Ab}(t))dt = \lambda_{A0}\int e^{(\lambda_{BA} - \lambda_{A0} - \lambda_{B0})t}dt
\]

\[
P_{Ab}(t)e^{\lambda_{BA}t} = \frac{\lambda_{A0}e^{(\lambda_{BA} - \lambda_{A0} - \lambda_{B0})t}}{\lambda_{BA} - \lambda_{A0} - \lambda_{B0}} + c
\]

From our initial assumption that observed mutations arise from discrete fixation events, \(P_{ab}(0) = 0\), which leads to the unique solution

\[
P_{Ab}(t) = \frac{\lambda_{A0}(e^{-\left(\lambda_{A0} + \lambda_{B0}\right)t} - e^{-\lambda_{BA}t})}{\lambda_{BA} - \lambda_{A0} - \lambda_{B0}}.
\]

A symmetric derivation finds that \(P_{ab}(t)\), the probability of observing \(B\) but not \(A\), is
\[ P_{AB}(t) = \frac{\lambda_{B0}(e^{-(\lambda_{A0} + \lambda_{B0})t} - e^{-\lambda_{AB}t})}{\lambda_{AB} - \lambda_{A0} - \lambda_{B0}}. \]

Finally,
\[ P_{AB}(t) = 1 - P_{ab}(t) - P_{Ab}(t) - P_{ab}(t). \]

For any pair of variants, the likelihood of sample genotypes is a product of probabilities calculated with the above equations. Numerical optimization methods are used to infer the selection parameters by maximizing likelihood. A univariate confidence interval for each parameter is estimated by fixing the other parameters at their maximum likelihood estimates and calculating a Wald confidence interval. A full mathematical derivation for epistasis has been recently submitted to bioRxiv as a preprint.

To test for epistatic effects at the gene level, we assume that variants of each gene share equivalent pairwise selection dynamics with variants in the other. (To reduce the influence of passenger variants and calling error, we only consider variants that appear in multiple samples.) Within each gene, the rate of neutral variant fixation is the binomial probability of observing one or more of the variants under their independent neutral fixation rates. By defining variants \( A \) and \( B \) as the observation of one or more recurrent variants in the pair of genes, we construct genotype likelihoods as described above and infer gene-level epistatic effects.
Statement of Purpose

The purpose of this thesis project was to utilize the cancerEffectSizeR software package to calculate the selection intensity of mutations within the tumors of lower-risk and higher-risk primary and metastatic prostate cancer in a time-ordered manner. Analyses of epistasis were additionally done to investigate the dependence of one gene’s propensity to be mutated in relation to the presence or absence of a mutation in another gene.

The main objectives of this thesis are:

1. Quantify the selection intensity of point mutations in prostate cancer to illuminate differences in the mutational landscape between lower-risk, higher-risk, and metastatic prostate cancer
2. Correlate analyses of effect size with protein structural and biochemical analyses
3. Support previous research in the genetics and genomics of prostate cancer using analyses of cancer effect size
4. Hypothesize new gene-gene interactions within prostate cancer tumorigenesis using epistatic analyses
Results

While the mutational rate increases as prostate cancer progresses, the mutational signatures remain the same

A comparison of the lower-risk, higher-risk, and metastatic prostate tumors (Fig. 1) shows that the trinucleotide mutation profiles between these three cohorts exhibit very similar patterns. These trinucleotide mutation profiles were then used to generate the COSMIC signatures\textsuperscript{104}. As expected from the similarity in trinucleotide mutation profiles, the COSMIC signature profiles were extremely similar when comparing the three groups of tumors (Fig. 2). Certain signatures have been excluded from this analysis, as they are implausible in prostate cancer\textsuperscript{104}. Clock-like COSMIC signatures SBS1, SBS5, and SBS40 are heavily featured throughout, as have been previously noted\textsuperscript{113}. There is a marked enrichment in C→T mutations throughout all tumors, especially in the ACG/CCG/GCG trinucleotide contexts, which lines up directly with the COSMIC signature SBS1, which has been associated with spontaneous or enzymatic deamination of 5-methylcytosine to thymine. The enrichment of T→C mutations in the ATA/ATG/ATT contexts correlates with the COSMIC signature SBS5. Unlike other cancers such as melanoma or hepatocellular carcinoma, prostate cancer does not have any defining mutation patterns or causative factors. Rather, it is primarily a disease of age and genetics, and as such, the most heavily featured COSMIC signatures are clock-like signatures, which are correlated with the patient’s age\textsuperscript{114}.

The average mutation rate of each gene in all tumors, without any stratification by stage is shown in Fig. 3A. The samples were subdivided into lower-risk (n = 479), higher-risk (n = 406), and metastatic tissue (n = 1051). Higher-risk tumors tend to have a higher gene-level mutation rate than the lower-risk tumors, with the five selected genes all possessing a higher mutation rate in the higher-risk primary tumors (Fig. 3B). Furthermore, the gene-level mutation rate of the metastatic tissue was generally greater
than those of the primary tumors (Fig. 3C). Again, the five selected genes had an increased mutation rate within the metastatic tumors. Of note, the mutation rates of the androgen receptor gene \(AR\) and tumor suppressor \(TP53\) were drastically higher in the metastatic cohort. This overall increase in mutation rate is consistent with the current literature\(^{115–118}\).

**Figure 1.** Percent of single-nucleotide mutations within each trinucleotide context in lower-risk primary, higher-risk primary, and metastatic tissue.
Figure 2. Distributions of the weights assigned to COSMIC signatures in (A) lower-risk (inside, darker color) and higher-risk (outside, lighter color) primary tumors and (B) all primary (inside, darker color) and metastatic (outside, lighter color) tissue. Signatures not shown because they contributed no weight were signatures 4, 7a, 7b, 7c, 7d, 9, 10a, 10b, 11, 14, 15, 16, 17a, 17b, 19, 20, 21, 22, 23, 24, 25, 26, 28, 29, 30, 31, 32, 34, 35, 36, 38, 42, 44, 84, 85, 88, 89, 90, 91, 92, 93, and 94.
Figure 3. The gene-level mutation rates—with known prostate cancer drivers \textit{SPOP} (pink), \textit{PI3KCA} (dark blue), \textit{TP53} (light green), \textit{FOXA1} (blue), and \textit{AR} (green) indicated—for (A) all tumors without any staging, (B) lower-risk versus higher-risk tumors (equality: dashed red line), and (C) primary vs. metastatic tissue (equality: dashed red line). The five reference genes indicated by larger, colored points are recurrently mutated in the tumors and are well-known drivers of prostate cancer.
Lower-risk, higher risk, and metastatic prostate cancer tumors are subject to markedly different selective regimes on genes

The lower-risk, higher-risk, and metastatic subsets of prostate tumors have a remarkably different set of genes that have been strongly selected (Fig. 4). The 16 genes shown are genes that have been well established within the literature as drivers of tumorigenesis, with most of them being prostate-cancer specific49,63,64,78,84,119–124. The seven genes that appear to be associated with lower-risk prostate cancer are CUL3, SPOP, PIK3CA, CTNNB1, AKT1, TP53, and FOXA1. The genes that were associated with higher-risk prostate cancer were ROCK1, RHOA, PIK3CB, ATM, and MUC16. MUC16 was selected-for in both higher-risk primary tumors and metastases as well. Lastly, in the metastatic tissue, mutations in AR, APC, and KMT2D were highly selected for. Given the androgen receptor’s role in the transition from mCSPC to mCRPC, it is understandable that there is such a high selection for mutations in AR within the metastatic prostate tissue. This categorization method was used to characterize the genes as “lower-risk”, “higher-risk”, and “metastatic” for the subsequent analyses.
Figure 4. Across selected oncogenic sites/regions/domains within 16 genes, the gene-level scaled selection coefficients of lower-risk primary tumors (red), higher-risk primary tumors (green), and metastatic tissues (blue) in 16 known prostate-cancer associated genes.
**SPOP variants within prostate cancer and their scaled selection coefficients**

Mutations in SPOP are very well-known drivers of prostate cancer; *SPOP* is the most-commonly mutated tumor suppressor gene in prostate cancer\(^{73}\) and is responsible for mediating the ubiquitination and subsequent degradations of several oncoproteins—including AR\(^{125}\). Many residues of SPOP are mutated throughout all the prostate cancer samples (**Fig. 5**). Nine SPOP residues are recurrently mutated, out of which six residues (Y87, F102, F125, D130, F133, W131) have multiple distinct substitutions, while three residues only have one substitution (S119, K129E, K134N). By looking at these SPOP mutations with respect to the residues’ location on the protein structure and taking into account the amino acid change, we can obtain a functional understanding of why certain somatic variants may have a drastically higher scaled selection coefficient. All mutated residues occur in (Y87, F102, K129, D130, F131, W133) or around (S119, F125, K134) the binding groove of SPOP\(^{105}\). SPOP F133V and SPOP W131G are the two variants that have the highest scaled selection coefficient in the entire group of primary tumors. SPOP F133V has a scaled selection coefficient significantly greater than all but two other mutations (non-overlapping 95% CIs with all but W131G and F102C). Interestingly, the amino-acid changes of those two mutations are drastic: W→G and F→V mutations constitute a change from two of the bulkiest R-groups to two of the least bulky, with these changes occurring at two residues right in the middle of the binding pocket. Such significant changes within the binding pocket likely abrogate the ability of SPOP to mediate degradation of its substrates, facilitating tumorigenesis. Analysis of somatic variants in SPOP might demonstrate a direct connection between differences in biochemistry directly affecting differences in tumorigenicity.
Figure 5. Scaled selection coefficients of recurrent single-nucleotide variant amino-acid substitutions in SPOP in all tumors, ranging from strongly selected (red) to weaker-selected (yellow), as well as 95% confidence intervals for the scaled strength of selection and raw prevalences (out of 2699 tumors with sequence data). Inset: SPOP protein structure (grey), with the binding moiety of the SPOP ligand, BRD3 (green); all substituted sites (inset residues—stronger selection:red to weaker selection:yellow) are counterfactually depicted as simultaneously featuring the strongest-selected recurrent amino-acid substitutions.

AR variants within prostate cancer and their scaled selection coefficients

Overexpression of the AR signaling is a sufficient and principal driver of mCRPC, involving point mutations, gene amplifications, and alternative splice variants\textsuperscript{126,127}. The cancer\textit{effect sizes} R package allows us to investigate the effect sizes of the AR point mutations within our tumors. Within the samples, there are 9 distinct recurrent mutations at 7 different sites in AR (Fig. 6). All these recurrent mutations occur within the ligand-binding domain (LBD) of the androgen receptor\textsuperscript{106}. L702H was the mutation with the highest scaled selection coefficient, followed by T878A and H875Y. The next two mutations were both at the same residue: W742L and W742C. Distinct AR point
mutations have been found to exhibit distinct effects, even though they all appear to occlude the ligand binding site. The T878A, H875Y, and W742C point mutations are all activating mutations that respond to anti-androgens, such as flutamide and bicalutamide as agonists, rather than antagonists\textsuperscript{128–130}. However, L702H, the mutation with the highest scaled selection coefficient in our analysis, is not activated by anti-androgens, but rather by glucocorticoids such as cortisol and cortisone\textsuperscript{131,132}. These four aforementioned mutations comprise up to 20% of all CRPC cases\textsuperscript{84}. While they all occur within the LBD of \textit{AR}, they result in two different pathways to escape androgen dependence and circumvent ADT. It is notable that the variant with the highest scaled selection coefficient, L702H, has a different effect on the androgen receptor from the other variants.

\textbf{Figure 6.} Scaled selection coefficients of recurrent single-nucleotide variant amino-acid substitutions in SPOP in primary tumors, ranging from strongly selected (red) to weaker-selected (yellow), as well as 95% confidence intervals for the scaled strength of selection and raw prevalences (out of 2699 tumors with sequence data). Inset: AR protein structure (grey), with the binding moiety of the AR ligand, dihydrotestosterone (teal); all substituted sites (inset residues—stronger selection:red to weaker selection:yellow) are counterfactually depicted as simultaneously featuring the strongest-selected recurrent amino-acid substitutions. The residues from S760 to M781 were removed from the structure to provide a better viewing window into the ligand binding pocket.
**Epistasis in prostate cancer**

Paired gene-wise epistasis analysis was applied to the previously selected 16 genes, resulting in 240 different permutations. Each gene exhibited 15 different epistatic effects upon the other genes, and each gene received 15 different epistatic effects from the other genes. The majority of the other genes exhibited negative interactions with SPOP (Fig. 7A). That is, selection for a SPOP mutation decreases when there are mutations in these genes, with the exception of RHOA and CTNNB1. In addition, under the epistatic model, selection for mutations in these six genes increased with a background SPOP mutation, including RHOA, AKT1, AR, and PIK3CA. These findings suggest that SPOP is an early gene in the chain of gene mutations that are mutated along the genetic trajectory associated with the progression of prostate cancer.

The next dominant gene in prostate cancer whose epistatic gene interactions were investigated was TP53 (Fig. 7B). AR showed a modest increase in selection, and MUC16 showed a significant increase in selection after a TP53 mutation. Mutations in KMT2C, APC, and PIK3CB, among a few other genes, increase the selection for a mutation in TP53, although these interactions appeared modest. The presence of a mutation in SPOP does not seem to affect the selection for a TP53 mutation. However, it appears that the presence of a mutation in TP53 greatly reduces the selection for a SPOP mutation.

Lastly, the epistatic interactions between AR and the other genes were explored (Fig. 7C). The selection for a mutation in AR is massively increased after a mutation in KMT2D. This epistatic change in selection for AR is almost tenfold greater than the next positive interaction in AR. There are modest increases in selection for a mutation in AR after a mutation in ATM, SPOP, PIK3CA, CTNNB1, and FOXA1 as well. A mutation in AR only increases the selection for mutations in two genes: ATM and MUC16. A deeper look at these epistatic interactions can result in a hypothesized time-ordered trajectory of somatic mutations during prostate cancer development. For example, a mutation in
*MUC16* results in a negative change in selection for a mutation in *AR*. However, a mutation in *AR* results in a much higher increase in the selection for a mutation in *MUC16*. This points to mutations in *AR* occurring chronologically before mutations in *MUC16*.

Of the 240 possible interactions between the 16 genes, 36 showed a positive epistatic interaction. Typically, positive epistatic interactions aligned either with progression within the same group or progression from lower-risk stages to higher-risk stages to metastasis (*Table 1*). Interestingly, 19 of the 36 positive interactions involved two “lower-risk” genes, indicating that positive selection for specific genes is an important factor in the early tumorigenesis of prostate cancer. These findings show the evolutionary trajectory of prostate cancer (*Fig. 8*), identifying the time-ordered and interdependent importance of mutations occupying key roles in the development of prostate cancer. While this is a simplified diagram of only a fraction of the interactions that were analyzed, it shows how epistatic analysis of cancer effect sizes can be used to construct a sequentiality map of mutational progression in cancer.
Figure 7. Pairwise epistatic effects of 16 selected genes (A) Change in scaled selection coefficient for mutated SPOP due to epistatic effects of mutations in 15 genes (aquamarine), and change in scaled selection coefficient for mutations in 15 genes due to epistatic effects of mutated SPOP (orange). (B) Change in scaled selection coefficient for mutated TP53 due to epistatic effects of mutations in 15 genes (aquamarine), and change in scaled selection coefficient for mutations in 15 genes due to epistatic effects of mutated SPOP (orange). (C) Change in scaled selection coefficient for mutated AR due to epistatic effects of mutations in 15 genes (aquamarine), and change in scaled selection coefficient for mutations in 15 genes due to epistatic effects of mutated AR (orange).
Table 1. Estimated positive epistatic interactions for prostate cancer growth

<table>
<thead>
<tr>
<th>Gene pair</th>
<th>Scaled selection coefficient</th>
<th>Effect on progression</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR</td>
<td>2,277</td>
<td></td>
</tr>
<tr>
<td>KMT2D → AR</td>
<td>92,682</td>
<td>Metastatic → Metastatic</td>
</tr>
<tr>
<td>AKT1</td>
<td>4,013</td>
<td></td>
</tr>
<tr>
<td>CTNNB1 → AKT1</td>
<td>80,630</td>
<td>Lower-risk → Lower-risk</td>
</tr>
<tr>
<td>RHOA</td>
<td>3,406</td>
<td></td>
</tr>
<tr>
<td>SPOP → RHOA</td>
<td>52,143</td>
<td>Lower-risk → Lower-risk</td>
</tr>
<tr>
<td>AKT1</td>
<td>4,522</td>
<td></td>
</tr>
<tr>
<td>FOXA1 → AKT1</td>
<td>35,671</td>
<td>Lower-risk → Lower-risk</td>
</tr>
<tr>
<td>CTNNB1</td>
<td>3,710</td>
<td></td>
</tr>
<tr>
<td>FOXA1 → CTNNB1</td>
<td>33,921</td>
<td>Lower-risk → Lower-risk</td>
</tr>
<tr>
<td>ATM</td>
<td>1,299</td>
<td></td>
</tr>
<tr>
<td>AR → ATM</td>
<td>24,074</td>
<td>Metastatic → Higher-risk</td>
</tr>
<tr>
<td>SPOP</td>
<td>8,6481</td>
<td></td>
</tr>
<tr>
<td>CTNNB1 → SPOP</td>
<td>29,378</td>
<td>Lower-risk → Lower-risk</td>
</tr>
<tr>
<td>MUC16</td>
<td>1,337</td>
<td></td>
</tr>
<tr>
<td>AR → MUC16</td>
<td>29,477</td>
<td>Metastatic → Higher-risk/Metastatic</td>
</tr>
<tr>
<td>FOXA1</td>
<td>756</td>
<td></td>
</tr>
<tr>
<td>ATM → FOXA1</td>
<td>17,809</td>
<td>Higher-risk → Lower-risk</td>
</tr>
<tr>
<td>FOXA1</td>
<td>685</td>
<td></td>
</tr>
<tr>
<td>MUC16 → FOXA1</td>
<td>13,866</td>
<td>Higher-risk/Metastatic → Lower-risk</td>
</tr>
<tr>
<td>PIK3CA</td>
<td>1,968</td>
<td></td>
</tr>
<tr>
<td>KMT2C → PIK3CA</td>
<td>13,930</td>
<td>Higher-risk → Lower-risk</td>
</tr>
<tr>
<td>AR</td>
<td>2,335</td>
<td></td>
</tr>
<tr>
<td>ATM → AR</td>
<td>13,655</td>
<td>Higher-risk → Metastatic</td>
</tr>
<tr>
<td>PIK3CA</td>
<td>1,899</td>
<td></td>
</tr>
<tr>
<td>FOXA1 → PIK3CA</td>
<td>10,809</td>
<td>Lower-risk → Lower-risk</td>
</tr>
<tr>
<td>SPOP</td>
<td>8,912</td>
<td></td>
</tr>
<tr>
<td>RHOA → SPOP</td>
<td>17,634</td>
<td>Higher-risk → Lower-risk</td>
</tr>
<tr>
<td>Gene 1</td>
<td>Gene 2</td>
<td>Interaction</td>
</tr>
<tr>
<td>-----------</td>
<td>------------</td>
<td>-------------</td>
</tr>
<tr>
<td>FOXA1</td>
<td>ATM</td>
<td>ATM → FOXA1</td>
</tr>
<tr>
<td>MUC16</td>
<td>TP53</td>
<td>TP53 → MUC16</td>
</tr>
<tr>
<td>PIK3CA</td>
<td>CTNNB1</td>
<td>CTNNB1 → PIK3CA</td>
</tr>
<tr>
<td>KMT2D</td>
<td>CTNNB1</td>
<td>CTNNB1 → KMT2D</td>
</tr>
<tr>
<td>AKT1</td>
<td>SPOP</td>
<td>SPOP → AKT1</td>
</tr>
<tr>
<td>MUC16</td>
<td>FOXA1</td>
<td>FOXA1 → MUC16</td>
</tr>
<tr>
<td>AR</td>
<td>SPOP</td>
<td>SPOP → AR</td>
</tr>
<tr>
<td>TP53</td>
<td>KMT2C</td>
<td>KMT2C → TP53</td>
</tr>
<tr>
<td>AR</td>
<td>PIK3CA</td>
<td>PIK3CA → AR</td>
</tr>
<tr>
<td>TP53</td>
<td>APC</td>
<td>APC → TP53</td>
</tr>
<tr>
<td>TP53</td>
<td>PIK3CB</td>
<td>PIK3CB → TP53</td>
</tr>
<tr>
<td>AR</td>
<td>CTNNB1</td>
<td>CTNNB1 → AR</td>
</tr>
<tr>
<td>PIK3CA</td>
<td>SPOP</td>
<td>SPOP → PIK3CA</td>
</tr>
<tr>
<td>TP53</td>
<td>CUL3</td>
<td>CUL3 → TP53</td>
</tr>
<tr>
<td>AR</td>
<td>FOXA1</td>
<td>FOXA1 → AR</td>
</tr>
<tr>
<td>APC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene Relationship</td>
<td>Frequency</td>
<td>Phenotype Change</td>
</tr>
<tr>
<td>-------------------</td>
<td>-----------</td>
<td>------------------</td>
</tr>
<tr>
<td>SPOP → APC</td>
<td>2,526</td>
<td>Lower-risk → Metastatic</td>
</tr>
<tr>
<td>TP53</td>
<td>1,027</td>
<td></td>
</tr>
<tr>
<td>PIK3CA → TP53</td>
<td>2,417</td>
<td>Lower-risk → Lower-risk</td>
</tr>
<tr>
<td>CTNNB1</td>
<td>4,112</td>
<td></td>
</tr>
<tr>
<td>PIK3CA → CTNNB1</td>
<td>4,929</td>
<td>Lower-risk → Lower-risk</td>
</tr>
<tr>
<td>AR</td>
<td>2,289</td>
<td></td>
</tr>
<tr>
<td>TP53 → AR</td>
<td>3,079</td>
<td>Lower-risk → Lower-risk</td>
</tr>
<tr>
<td>TP53</td>
<td>1,032</td>
<td></td>
</tr>
<tr>
<td>CTNNB1 → TP53</td>
<td>1,815</td>
<td>Lower-risk → Lower-risk</td>
</tr>
<tr>
<td>TP53</td>
<td>1,041</td>
<td></td>
</tr>
<tr>
<td>ATM → TP53</td>
<td>1,750</td>
<td>Higher-risk → Lower-risk</td>
</tr>
<tr>
<td>FOXA1</td>
<td>742</td>
<td></td>
</tr>
<tr>
<td>SPOP → FOXA1</td>
<td>1,310</td>
<td>Lower-risk → Lower-risk</td>
</tr>
</tbody>
</table>
Figure 8. Conceptual diagram of the somatic evolutionary genetic trajectory of progression, based on scaled selection coefficients and epistatic effects of highly selected driver mutations. Mutations in SPOP and CTNNB1 are strongly selected along the trajectory from normal tissue to lower-risk, and when present increase selection for mutations in PIK3CA leading to lower-risk progression. Mutations in SPOP and RHOA appear to have mutually positive epistatic effects upon each other. Mutations in KMT2C are strongly selected along the trajectory from normal tissue to higher-risk tumor, and when present, increase selection for mutations of PIK3CA and TP53. Mutations in PIK3CB also increase the selection for mutations in TP53. Mutations in TP53 when present increase selection for mutations of AR and MUC16 leading to progression from lower-risk to higher-risk/metastatic progression. Mutations in AR when present increase selection for mutations of ATM and MUC16 and further higher-risk/metastatic progression. Lastly, mutations in CTNNB1 increase selection for mutations in KMT2D, which increase selection for mutations in AR, thus progressing lower-risk tumors towards metastasis.
**Discussion**

Here, we have shown that the trinucleotide mutation profiles, and thus COSMIC mutational signatures, underlying primary prostate tumors and metastatic tissue are highly similar. We have divided the primary tumors into “lower-risk” and “higher-risk” tumors according to the Gleason grade. The gene-level mutation rates increased as the tumors progressed from primary to metastasis, with higher-risk primary tumors having an overall higher mutation rate compared to lower-risk primary tumors. Using analyses of cancer effect size, we elucidated which genes and mutations are selected in the prognosis and progression of prostate cancer from normal tissue to primary tumors to metastases, respectively. Protein structural and biochemical analysis can be combined with this effect size analysis to get a full picture of the processes that contribute to the genesis of certain mutations. In the case of SPOP, all the sites of mutation are within the binding groove and the variants with the largest effect size tend to be the ones with a biggest change (i.e. from a bulky R-group to small R-group). Furthermore, by applying analyses of epistasis and calculating the effects of mutations on gene-level scaled selection coefficients, we are able to support previous and hypothesize novel gene-gene interactions, as well as postulate the position of genes along the trajectory of progression in prostate cancer.

Within the tumor evolution from normal tissue to lower-risk tumors, mutations in two genes that have important roles in prostate cancer, SPOP and CUL3, were strongly selected. Speckle-type POZ Protein (SPOP) is particularly well-known as a driver that is found to be frequently mutated in prostate cancer tumors73. SPOP has two domains—a MATH domain that binds to substrates and mediates ubiquitination and subsequent degradation, and a BTB domain that assembles with CUL3, forming part of a E3 ubiquitin ligase. Mutations in CUL3 and SPOP have been shown to be mutually exclusive96, due to the negative epistatic interaction between mutations in these two
genes shown in our analysis, demonstrating the importance of both these genes to the ubiquitination process.

Based on our epistasis analysis, mutations in SPOP result in an epistatic increase in selection for mutations in both PI3KCA and AR. Two drivers of prostate tumorigenesis are overactivation of PI3K/mTOR signaling and dysregulation of androgen-receptor (AR) signal pathway. These two pathways have been shown to negatively regulate each other through reciprocal feedback86. Mutations in SPOP disrupt this feedback, leading to an upregulation of both PI3K and AR pathways133. In addition, tumors in the TCGA database with SPOP mutations show an association with higher AR activity119,134. The androgen receptor has been shown to be targeted for degradation by wild-type SPOP, but this process is blunted with mutant SPOP variants135.

SPOP mutations have long been thought to be early events in prostate cancer tumorigenesis73-136. By dividing the primary prostate tumors into “lower-risk” and “high-risk” tumors—based on their Gleason grade—we showed that different genes and variants are selected for and responsible for the advancing tumorigenesis at different stages of tumor evolution. The variety of SPOP mutations seem to be highly selected-for in mainly the primary tumors that were a lower Gleason grade, and therefore lower-risk. SPOP mutant-positive prostate tumors were analyzed and the SPOP mutations were found to be highly clonal and early events that preceded any other SPOP-associated mutations or deletions136. Previous studies have shown that TP53 mutations tend to present later and be associated with more advanced prostate tumors that remained localized137,138. SPOP mutations and TP53 alterations have been found to be strongly mutually exclusive119, indicating that they represent different molecular subtypes of prostate cancer. In our analysis, the presence of mutations in SPOP had little effect on the selection of, while mutations in TP53 resulted in negative epistatic change in selection for SPOP mutations. Furthermore, our epistasis analysis demonstrated that a
mutation in \textit{KMT2C} increases the selection for a mutation in \textit{TP53}, with little change in selection when the order of mutations is reversed. This may explain the previously demonstrated co-occurring relationship between \textit{TP53} and \textit{KMT2C} in prostate tumors\textsuperscript{139}.

\textit{MUC16}, which was selected for in higher-risk primary and metastatic tumors, produces a protein product named CA-125, which is almost solely associated with ovarian cancer\textsuperscript{139,140}. It is the most effective screening assay for ovarian cancer\textsuperscript{141} and is essentially the analog to what PSA is to prostate cancer. Induced expression of \textit{MUC16} in cell lines has been shown to increase growth and tissue invasiveness, as well as increase activation of the AKT and ERK pathways of 3T3 cell lines\textsuperscript{124}. In addition, \textit{MUC16}/CA125 has been showed to bind to the mesothelial epithelium of the peritoneum, and this is a purported contribution of CA125 towards ovarian cancer metastasis\textsuperscript{142}. While CA125 was used as a biomarker for detection of ovarian cancer and monitoring of treatment response, it appears that it may also have an important role in ovarian cancer pathogenesis and development of metastatic disease.

It is interesting then that we see selection for mutations in \textit{MUC16} as well in prostate cancer, and it is especially surprising that we see increased epistatic selection for mutations in \textit{MUC16} when there is a mutation in either \textit{AR} or \textit{TP53}. Because of its strong association with ovarian cancer, there are few studies done with other cancers in mind. However, 10\% of non-gynecological tumors showed CA125 expression, and 28\% of patients with non-gynecological tumors had an increase in blood level of CA125\textsuperscript{143}. A case series of patients with advanced prostate cancer with increased serum CA125 levels showed that out of the 11 patients, eight had low or undetectable serum levels of PSA\textsuperscript{144}. The authors concluded that CA125 may be another useful biomarker for detecting prostate cancer or monitoring treatment.

Currently, there are several preliminary studies investigating \textit{MUC16} as a therapeutic target in ovarian cancer\textsuperscript{145}. DMUC5754A is a humanized anti-MUC16
antibody, conjugated to the microtubule-disrupting agent monomethyl auristatin E, that was given to patients with advanced ovarian or pancreatic cancer and showed promising anti-tumor activity in patients with MUC16-high tumors. REGN4018 is a human bispecific antibody to MUC16 and CD3, which showed extremely promising results in mice models and is currently in a phase I trial (NCT03564340).

Our findings show that MUC16 is not only selected-for in higher-risk primary prostate tumors as well as metastases, but also that mutations in the androgen receptor, AR, greatly increase selection for such mutations in MUC16. Given that AR has a well-established role in metastatic prostate cancer and that MUC16 has been implicated in the development of metastatic ovarian, lung, and pancreatic cancer, MUC16 may have an important role in metastatic prostate cancer as well. Furthermore, our analysis also showed increased epistatic selection from mutations in MUC16 after mutations in TP53.

The connection between TP53 and MUC16 has been noted in ovarian cancer—(1) patients with ovarian tumors that contained TP53 mutations had a significantly higher level of TP53 auto-antibodies ($P = 0.0019$), and (2) TP53 auto-antibodies have been shown to have a significant lead time over increase in CA125 levels. In the future, considering expression of and mutations in MUC16, as well as serum CA125 levels may be useful in not only detecting and monitoring prostate cancer, but also treating metastatic prostate cancer.

The several other epistatic interactions involving AR are a testament to its prominent role in metastatic prostate cancer. There is an extremely high selection for mutations in AR after a mutation in histone-lysine N-methyltransferase 2D (KMT2D, sometimes known as MLL2). Epigenetic involvement in prostate cancer tumorigenesis and metastasis is a relatively new aspect of prostate cancer research, but KMT2D has an established positive correlation with AR. Lv et al demonstrate that KMT2D expression is correlated with not only AR, but also the AR-downstream target CAMKK2. KMT2D
and WD5R complex together to increase AR mRNA transcription in the presence of ACK-induced histone phosphorylation\textsuperscript{151}. Elucidating epigenetic control of AR may play an important role in future therapeutic options for mCRPC.

A major limitation of our analysis is that we have only been able to analyze single-nucleotide variants using the \texttt{cancereffectsize} software package. There are additional genomic processes involved with prostate cancer tumorigenesis, including (1) multiple-nucleotide variants; (2) copy number variations (both amplifications and deletions); (3) fusion genes, such as \textit{TMPRSS2-ERG}; and (4) epigenetic changes including histone methylation and acetylation. Future research may expand our capability to analyze CNVs, which should enhance our understanding of important genes in not only prostate cancer.

The classification of genes within our prostate cancer samples as “lower-risk”, “higher-risk”, and “metastatic” was solely based off of our CES analysis of gene-level selection. This methodology is not especially rigorous, but was done out of practicality, as an internal method of organizing the genes, to give a framework for the subsequent analyses of epistatic interactions.

Lastly, the cancer effect size values produced by the CES analysis give us a metric for a gene or a somatic variant’s contribution towards tumorigenesis, but does not tell why that specific gene or variant has been selected. Cancer effect sizes should be paired with other analyses to fully explore the mechanism of contribution. For example, we performed structural analyses with \textit{SPOP} and \textit{AR} to find that all of the mutations occurred within the ligand binding domain of the genes, with some of the highest selected-for mutations in \textit{SPOP} corresponding to a significant change in amino acid. Furthermore, function of the highly selected mutations should be investigated with traditional methods in molecular biology to establish their role in pathogenesis. The \textit{SPOP} mutations associated with prostate cancer impair BET degradation, while those
associated with endometrial cancer enhance degradation; this leads to prostate cancer with \( SPOP \) mutations to have resistance against BET inhibitors, while \( SPOP \) mutation-associated endometrial cancer responds well to such treatment\(^{152}\).

In the future, cancer effect sizes can be used in conjunction with translational and clinical research to guide treatment via precision medicine, not only to treat existing mutations within the tumor, but also predict which mutations are likely to arise as a result of this treatment. For example, using CES analysis, \( KRAS \ G_{12C} \) targeted therapy in lung tumors was determined to most likely give rise to \( BRAF \ V_{600E} \) mutations\(^{153}\). In addition, knowing the evolutionary trajectory of mutations within a disease process, such as prostate cancer, will provide suggestions as to which genes to target first.
Conclusion

Application of cancer effect size analysis highlighted which mutations and genes are selected for leading up to each stage of prostate cancer, clarifying the genetic differences between lower-risk, higher-risk, metastatic prostate cancer. Protein structural analyses of SPOP and AR provided insight as to why those mutations occurred at those specific residues. We showed that SPOP is a gene that plays a key role in early tumorigenesis in prostate cancer—not only through the prominent selection coefficients of SPOP mutations in lower-risk prostate tumors, but also through its epistatic interactions that increase the selection for mutations in many other genes known to drive prostate cancer. The epistatic interactions between the many genes involved with prostate cancer tumorigenesis were then synthetically integrated into a hypothesis for the evolutionary trajectory of prostate cancer tumorigenesis.
References


44. Schally AV, Kastin AJ, Arimura A. Hypothalamic Follicle-Stimulating Hormone (FSH) and Luteinizing Hormone (LH)-Regulating Hormone: Structure, Physiology, and Clinical Studies**The work described here has been supported by research grants from the Veterans Administration, United States Public Health Service Grants AM-07467 and AM-09094, and the Population Council, New York, N.Y.††This paper was presented at the Twenty-Seventh Annual Meeting of the American Fertility Society, New Orleans, Louisiana, March 25–27, 1971. Fertil Steril 1971;22(11):703–21.


70. Kumar A, Coleman I, Morrissey C, et al. Substantial interindividual and limited intraindividual genomic diversity among tumors from men with metastatic prostate


91. Hojat M, Xu G. A visitor’s guide to effect sizes - statistical significance versus practical (clinical) importance of research findings.


114. Cannataro VL, Mandell JD, Townsend JP. Attribution of cancer origins to endogenous, exogenous, and actionable mutational processes [Internet]. bioRxiv. 2020; Available from: http://dx.doi.org/10.1101/2020.10.24.352989


