AWARD NUMBER: W81XWH-17-1-0224

TITLE: Regulatory Networks of Immune Evasion in Metastatic Prostate Cancer

PRINCIPAL INVESTIGATOR: Marcin Cieślik, PhD

CONTRACTING ORGANIZATION: Regents of the University of Michigan
Ann Arbor, MI 48109

REPORT DATE: June 2019

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
June 2019	Annual	1 Jun 2018 - 31 May 2019

4. TITLE AND SUBTITLE

Regulatory Networks of Immune Evasion in Metastatic Prostate Cancer

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-17-1-0224

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Marcin Cieslik

E-Mail: mcieslik@med.umich.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

REGENTS OF THE UNIVERSITY OF
MICHIGAN
ANN ARBOR MI 48109

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT .

Metastatic castration resistant PCa (mCRPC) is a very heterogeneous disease at molecular level, which is manifested by the highly variable rates in overall survival, and differences in responses to drugs. Therefore, prognostic and predictive biomarkers are needed to guide treatment decisions. To develop the most effective immunotherapeutic strategies for mCRPC, it is essential to dissect tumor immunogenicity, immune infiltration, and immune escape in each tumor. However, to date very little is known about the role of adaptive and innate immunity in mCRPC and their association with survival. The central hypothesis of this research is that immune phenotypes, such as tumor infiltrating lymphocytes (TILs) and immune checkpoints, are important predictors of survival in mCRPC. Further, we hypothesize that mCRPC patients with better outcomes will be characterized by a stronger and more diverse immune phenotype that is in general associated with better responses to immunotherapy. Therefore, the overarching goal of this research is to dissect signatures of immune infiltration and escape in mCRPC and to identify novel mechanistic biomarkers of cancer immune evasion. The successful completion of this work will result in the first comprehensive immunogenomic landscape of mCRPC. Further, we will characterize mechanisms of immune evasion (e.g. immune-checkpoints) in mCRPC and their association with survival. Our work established critical dependencies between the genotypes of mCRPCs and their phenotypes. These genotype-phenotype relationships are then used to identify prognostic biomarkers, which is the foundation for accurate patient stratifications in the paradigm of personalized medicine. Hence, our work augments the precision oncology with an actionable perspective on the tumor immune microenvironment.

15. SUBJECT TERMS Prostate cancer, mCRPC, immunogenomics, genomics, cancer genetics, immune profiling, transcriptomics, immunotherapy, bioinformatics, immunoinformatics, CDK12 mutant prostate cancer.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
Unclassified	Unclassified	Unclassified	Unclassified	50	19b. TELEPHONE NUMBER *(include area code)*

Table of Contents

Page

INTRODUCTION

Metastatic castration resistant PCa (mCRPC) is a very heterogeneous disease at molecular level, which is manifested by the highly variable rates in overall survival, and differences in responses to drugs. Therefore, prognostic and predictive biomarkers are needed to guide treatment decisions. To develop the most effective immunotherapeutic strategies for mCRPC, it is essential to dissect tumor immunogenicity, immune infiltration, and immune escape in each tumor. However, to date very little is known about the role of adaptive and innate immunity in mCRPC and their association with survival. The central hypothesis of this research is that immune phenotypes, such as tumor infiltrating lymphocytes (TILs) and immune checkpoints, are important predictors of survival in mCRPC. Further, we hypothesize that mCRPC patients with better outcomes will be characterized by a stronger and more diverse immune phenotype that is in general associated with better responses to immunotherapy. Therefore, the overarching goal of this research is to dissect signatures of immune infiltration and escape in mCRPC and to identify novel mechanistic biomarkers of cancer immune evasion. The successful completion of this work will result in the first comprehensive immunogenomic landscape of mCRPC. Further, we will characterize mechanisms of immune evasion (e.g. immune-checkpoints) in mCRPC and their association with survival. Our work established critical dependencies between the genotypes of mCRPCs and their phenotypes. These genotype-phenotype relationships are then used to identify prognostic biomarkers, which is the foundation for accurate patient stratifications in the paradigm of personalized medicine. Hence, our work augments the precision oncology with an actionable perspective on the tumor immune microenvironment.

KEYWORDS

Prostate cancer, mCRPC, immunogenomics, genomics, cancer genetics, immune profiling, transcriptomics, immunotherapy, bioinformatics, immunoinformatics, CDK12 mutant prostate cancer.

ACCOMPLISHMENTS

What were the major goals of the project?

The proposed research listed the following major goals in the statement of work. These goals together aim to accomplish the objective of a better understanding of the immunogenomic landscape of mCRPC and its impact on patient outcomes.

1. Development of Computational Methods
2. Extended Immune Profiling Development
3. Validation of Computational Methods
4. Immunogenomic characterization of patient tissues
5. Statistical analyses of outcomes of patient data

Associated with those goals where the following milestones, due within the first and second reporting period (year 1 and 2). Milestones associated with the second reporting period are highlighted in (blue). Most notably, ahead of schedule we are transitioning towards the outcomes portion of the proposal, which has more immediate implications for patients (goal ahead of schedule are highlighted in green). This reflects the focus of this research on finding that directly benefit prostate cancer patients in addition to basic research and bioinformatics developments.

1. Finished of neoantigen pipeline (due Nov 2017, completed)
2. Neonatigen analyses of discovery cohort (due Dec 2018, completed)
3. Finished MImmScore workflow (due Nov 2017, completed)
4. Implementation of TIL-profiling pipeline (due March 2018, completed)
5. Execution of computational pipelines (due Dec 2018, completed)
6. Implementation of Immunotyper (due June 2018, 75% completed)
7. Integrative clustering using Immunotyper (due Dec 2018, 50% completed)
8. Integrative clustering using alternative tools (due Dec 2018, completed)
9. IHC optimized (due Sept 2017, completed)
10. TCRB-seq optimized (due Sept 2017, completed)
11. Perform TCRB-seq analysis (due Dec 2018, completed)
12. All IHC done for validation (due March 2018, completed)
13. Digital and computational quantification of immunophenotypes (due June 2019, 80% completed)
14. All TCRB-seq for validation (due Nov 2017, completed)
15. TIL-profiling validated (due June 2018, 90% completion)
16. MImmScore validated (due June 2018, completed)
17. Complete Immunogenomic Characterization (due June 2019, 75% completed)
18. Survival analysis of discovery cohort (due Dec 2019, 50% completed)
19. Discovery of genotype ~ phenotype ~ outcomes association (due June 2020, 25% completed)
20. Manuscripts published (due June 2020, strong publication record to date)

What was accomplished under these goals?

This is a multidisciplinary project which involves the close collaboration between bioinformaticians, statisticians, pathologists, and cancer geneticists. The major activities involve diverse work that spans multiple areas of expertise and approaches, which are unified through the proposed major goals as stated above. A key aspect of the second reporting period were stronger collaborations with clinicians (Dr. Alva, Dr. Reimers) which allowed the project to progress quicker towards clinical translation and outcomes research. The PI (Dr. Cieślik) has also fostered multiple national and international collaborations which allowed him to contribute the expertise and insights gained within the scope of this research proposal towards multiple collaborative research projects listed below.

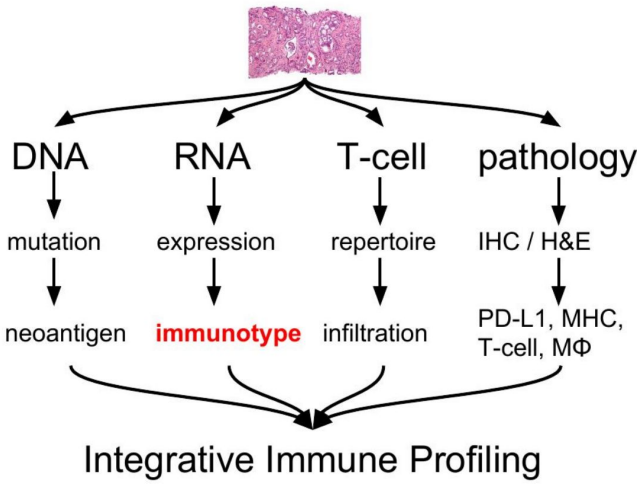


Figure 1: Integrative Immune Profiling

The above diagram explains how the different aspects, described in detail below, fit together to accomplish the major tasks of the proposal (above), and the overarching goal of this research, i.e. as stated in the proposal: “we introduce a novel Integrative Immune Profiling (IIP) strategy that leverages existing patient genomics and clinical data, but also incorporates dedicated immune profiling for T-cell infiltration using deep sequencing [...] to dissect signatures of immune infiltration and escape in mCRPC and to identify novel mechanistic biomarkers of cancer immune evasion.” We note that as computational methods advance there is a natural shift the technologies employed to achieve integrative immune profiling. In the context of this proposal, the recent availability of large volumes of whole-genome sequencing (WGS) data, with its rich insights into genomic instability and mutational processes, necessitate further developments of computational methods for the (DNA) portion of IIP. Other established methods, such as pathology require fewer additional tweaks.

Major activities:

(in all of the below descriptions, sections which have received significant effort during the second reporting period are highlighted in blue).

1. Implementation of bioinformatics pipelines (Major goal #1, above):
    - ➢ Specific Objective #1: to configure a secure and high-performance computer environment for the fast and controlled analysis of patient genomic data using bioinformatics pipelines and statistical algorithms. The following tasks were accomplished in order to make the swift processing of the large number of samples within this proposal (n=1,000). Specific Outcomes:
        - ○ Environments created and configured on local compute hardware.

- Scalable deployments implement on virtualized and secure cloud resources

Figure 2: Example run show highly efficient use of compute resources in immunogenomic analyses.

- All IT administration tasks have been automated and scripted, for example as shown below we have streamlined the query-submit-analyze-archive process on cloud resources:

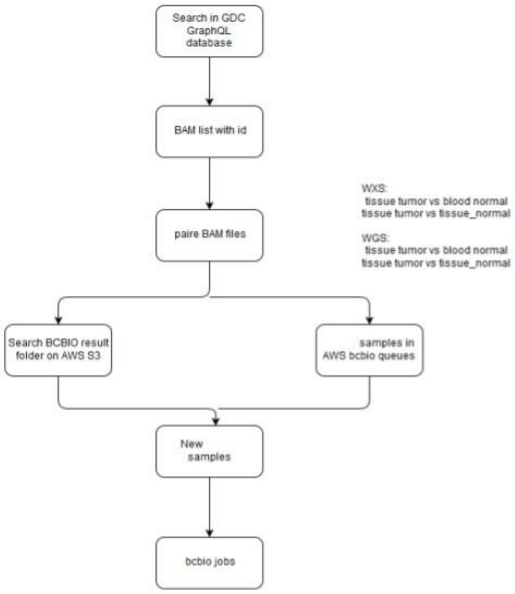


Figure 3: Implemented workflow for automated execution of bioinformatics and genomics pipelines (see below for description of pipelines and algorithms).

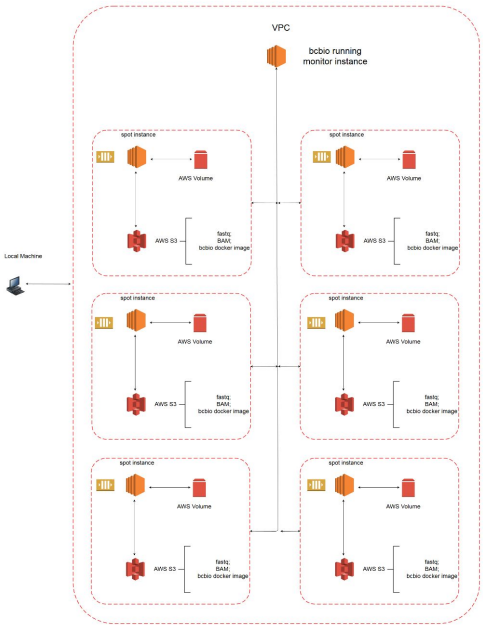


Figure 4: Automated work-scheduling on virtual private computing (VPC) clusters. Data-routing and details of the instance nodes are shown. A single monitor-instance manages the work-load of multiple worker-instances. The user manages the whole cluster by issuing programmatic calls to the monitor instance. Of note, our analysis workflows have been subsequently adopted to work in multiple compute environments which lessened or dependency on specific IT providers and lenders. This allows us to use the computational resources which offer the lowest cost and greatest ease-of-use.

To better understand and illustrate the reduction of the complexity of the in the genomic and bioinformatics pipelines and it's impact on the computational efficiency and maintenance burden we have carried out a systematic review of data flows. These analyses revealed that significant improvements are possible by streamlining data-flows, and introducing a principled mechanism for software maintenance / development. And the use of well thought-out software architecture.

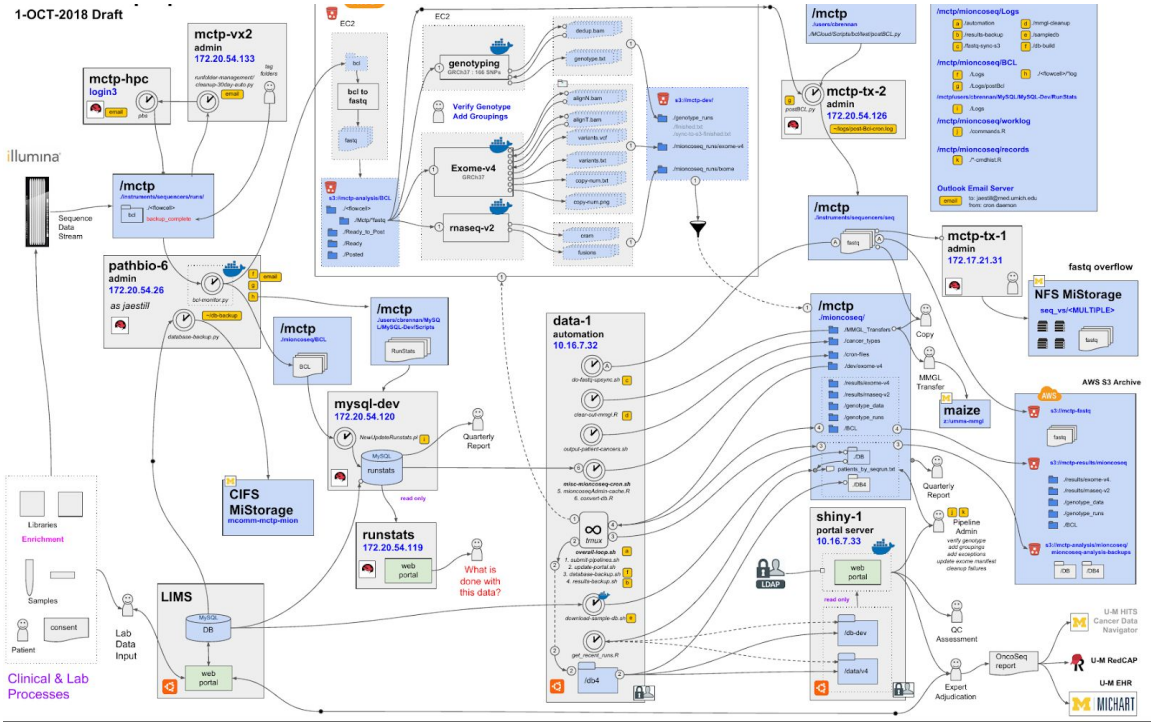


Figure 5: Illustration of extant genomic pipelines at the Michigan Center for Translational Pathology. Please not a complex flow of data and dependencies. While the architecture is reliable and provides accurate results. It is difficult to maintain slowing down the pace of progress and the delivery of improved genomics results to patients.

Therefore we have undertaken and green-field development strategy to improve our ability to translate genomic advances from research algorithms to patient diagnostics

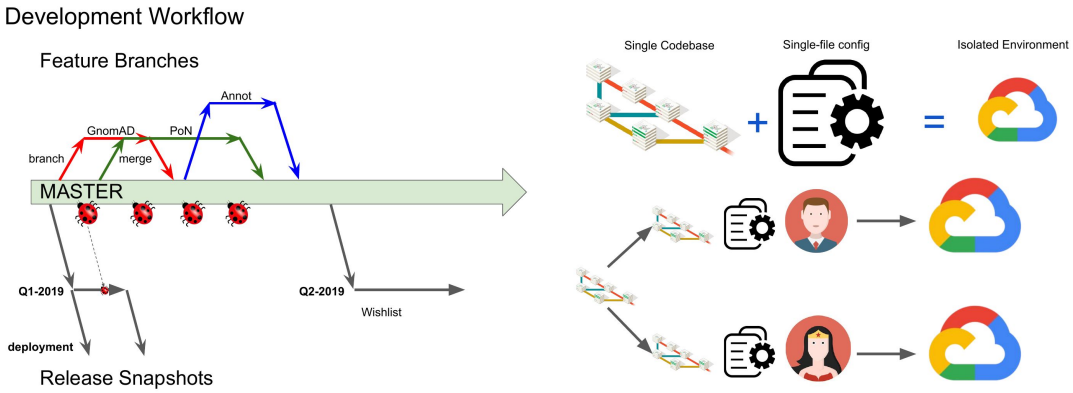


Figure 6: Implementation of rigorous software development practices (here shown is the feature-branch model of distributed version controlling) speeds up pace of development allowing us to meet or exceed the projects software-development timelines.

Next, we architected the operational organization of our compute pipelines both in terms of the genomics analyses:

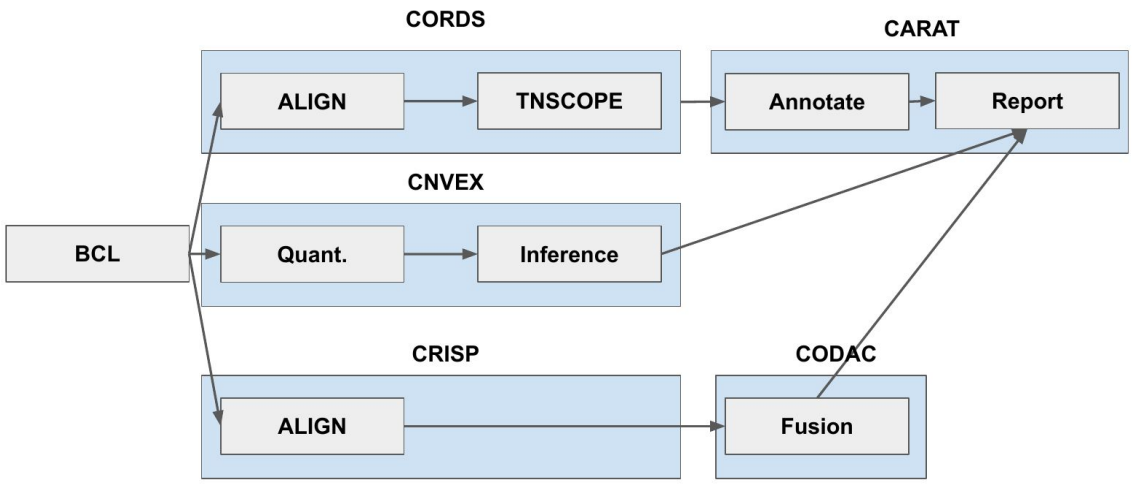


Figure 7: Computational pipelines at the MCTP are now organized in a logical fashion: CORDS - COmmon Runtime for DNA Sequencing, CNVEX - Copy Number Variant pipeline, CRISP - clinical RNA-seq pipeline, CODAC - comprehensive detection ana analysis of chimeras, CARAT - Clinically Accurate Reporting and Annotation of Tumors.

As well as in terms of the use of IT-solution to orchestrate the above analyses functionally using the available computational resources (e.g. clusters, compute nodes, development workstations).

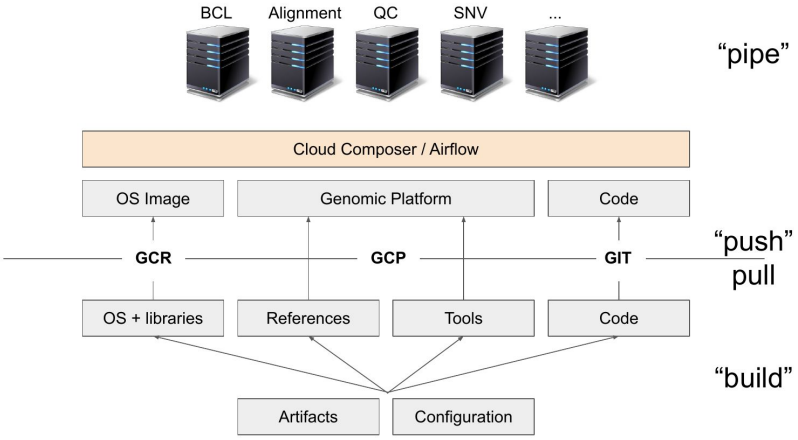


Figure 8: Use of vendor-provided IT-solution (Google Cloud Platform, Google Container Registry, GitHub) allows the ease of deployment and instantiation of computational analyses and pipelines. Black compute boxes reflect the different analyses while, gray diagram boxes represent data-flow and/or command-flow. From a set of Artifacts and Configurations a run on the Genomic Platform is instatiationed. Both data and commands are transferred using a push/pull mechanism.

➢ Specific Objective #2: to disseminate computational advances at the MCTP to the broader genomics research community.

Our efforts to make computational advances more broadly accessible resulted in the development of a software-project termed Turn-key Precision Oncology, which subsumes the immunogenomic IIP analyses proposed within this proposal. TPO has the overarching goal of making genomic analyses accessible to all researchers at all institutions and provide democratic access to advances in genomics. A turn-key solution to bioinformatics analyses of cancer genomes will allow other researchers to lounch research projects similar to ours and will improve the reach of the immunogenomic analyses proposed herein.

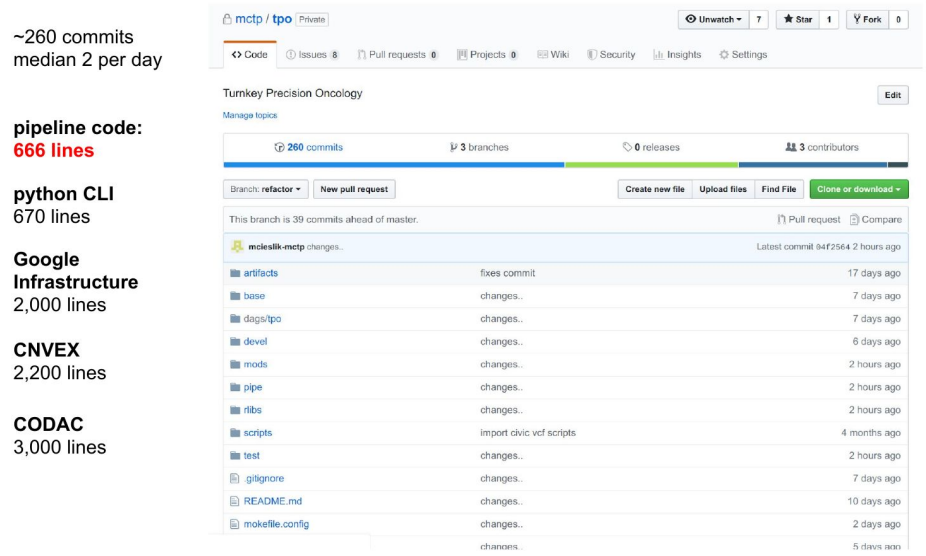


Figure 9: The TPO software project on the MCTP GitHub page. Basic summaries of commits and lines-of-code provided on the left.

➢ Specific Objective #3: To implement bioinformatics pipelines for the comprehensive analysis of genomic data. The pipelines represent the foundation for all the analyses and exploration proposed in the next stages of the proposal.

We start with a general description of the key pipelines CRISP and CODAC, which together implement the proposed workflow. CRISP, which stands for Clinical RNA-seq pipeline is responsible for the basic analysis of RNA-seq data (QC, alignment, immunogenomics, quantification), and takes raw sequencing data as input (FASTQ, FQ), the results from CRISP (quant. files, aligned data BAM) are fed into CODAC which carries out fusion detection and novel immunogenomic analyses, such as neoantigen

discovery. The results are saved as text (CSV) and binary (RDS) files and turned into user-friendly report html pages. Manuscripts detailed the operation of both components are in preparation.

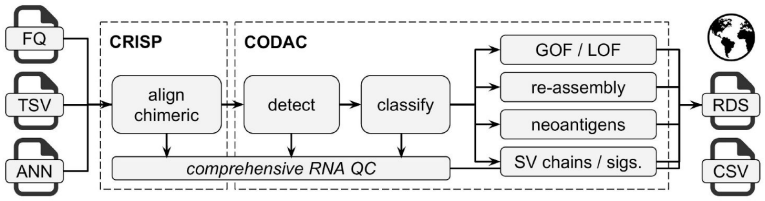


Figure 10: General overview of two of the implemented pipelines CRISP and CODAC and information flow in the immunogenomic analysis of RNA-seq data.
What follows are technical overviews of the two components.

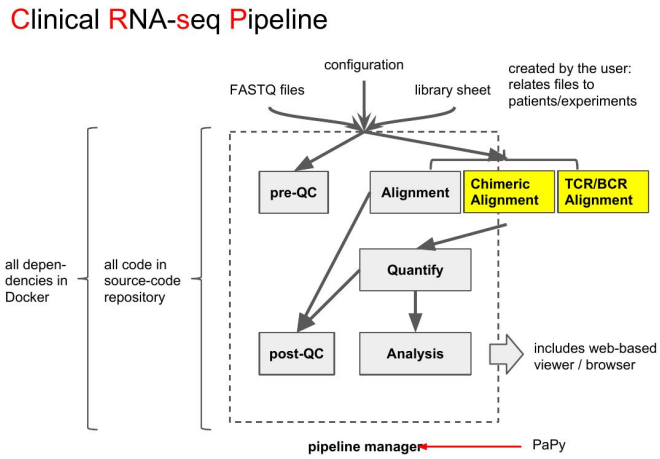


Figure 11: Details of the implemented bioinformatics pipeline for the analysis of RNA-seq data i.e. Clinical RNA-seq Pipeline (CIRSP)

As proposed, the pipeline includes special components for the immunogenomic i.e. T-cell receptor (TCR) and B-cell receptor assembly - for T-cell / B-cell repertoire analysis, and chimeric alignment for the better quantification of neoantigens (see below)

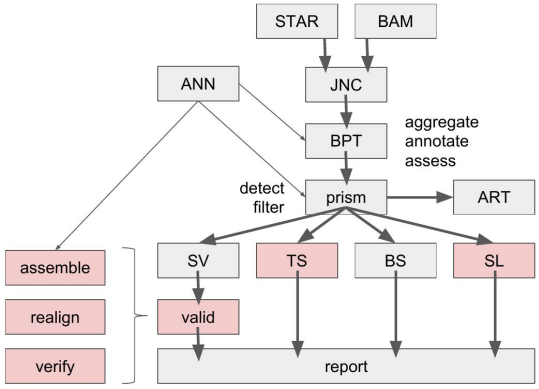


Figure 12: Details of the implemented bioinformatics pipelines for the Comprehensive Detection and analysis of Chimeras (CODAC)

Briefly, taking aligned BAM files as input, CODAC discovers chimeric junctions assembles them into breakpoints and classifies them using the PRISM algorithm (see below) which results in the classification of chimeric RNAs as structural variants (SV), trans-splices (TS), back-splices (BS), and stem-loops (SL). As proposed, the pipeline includes special components for the discovery and validation of neoantigens (valid. highlighted in red).

To prove CODACs improved utility and applicability in medical context, we demonstrate how it's functionality surpasses many existing fusion-calling algorithms:

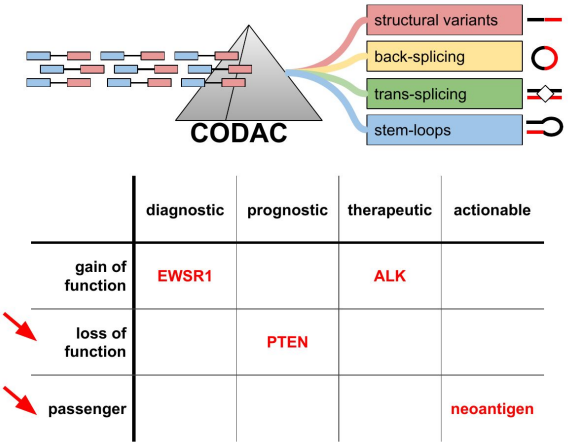


Figure 13: The CODAC "PRISM" algorithm distinguishes different classes of chimeric RNAs. This facilitates the discovery of neoantigens, which are central to this research proposal and can be classified as actionable passenger mutations. The red arrows highlights types of fusion which are typically neglected by other alternative implementations.

In the course of the second year of this project, we realized the need for accurate analysis of copy-number variation (CNV) data in the context of immunogenomics. This was spurred by recent publications, among many others, by (Davoli et. al, Science 2017), which reported that aneuploidies are associated with immune evasion, and (Stopsack et. al, PNAS, 2019), which reported a role of aneuploidies in prostate cancer progression.

This implied that accurate quantification of CNVS is necessary to accomplish the comprehensive immunogenomic profiling of mCRPC (as proposed). Therefore, we initiated the implementation of a state of the art software tool to quantify CNVs from a variatey of DNA-based assays including: targeted panel data (panel), whole exome

sequencing (WXS), whole genome sequencing (WGS). The tool, referred as CNVEX implements an innovative strategy which can combine WGS and targeted sequencing data to improve the resolution of CNVs.

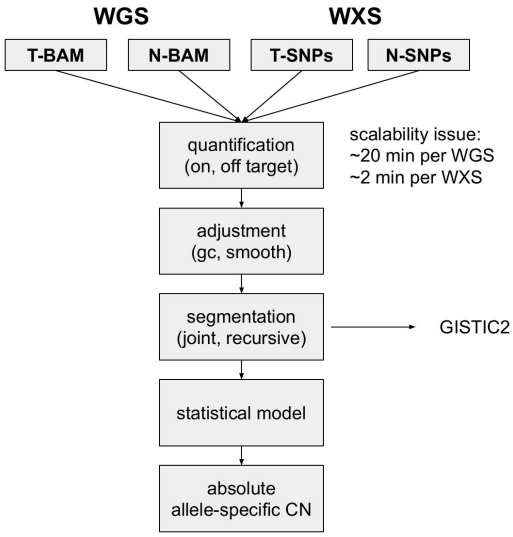


Figure 14: High-level overview of the CNVEX algorithm. Inputs to the algorithm constitute any combination of WGS and targeted sequencing data, followed by quantification, adjustment, smoothing, segmentations, and inference of absolute copy-numbers.

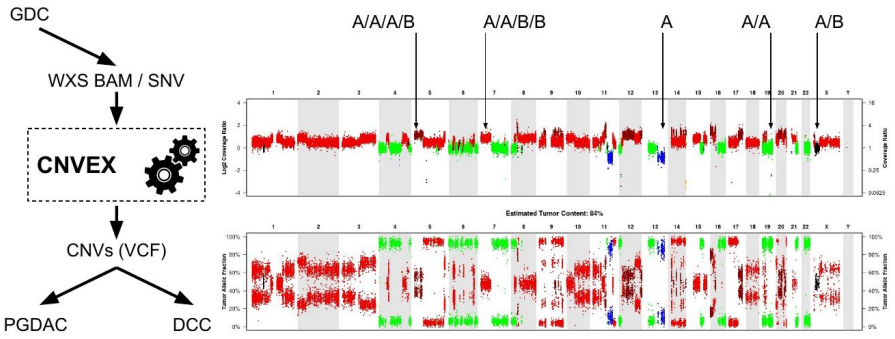


Figure 15: Desired outcome of the CNVEX algorithm - absolute inference of purity and ploidy levels. Only a purity/ploidy aware analysis can reliably identify events of whole-genome duplication, which are known to impact genomic instability and contribute to immune evasion.

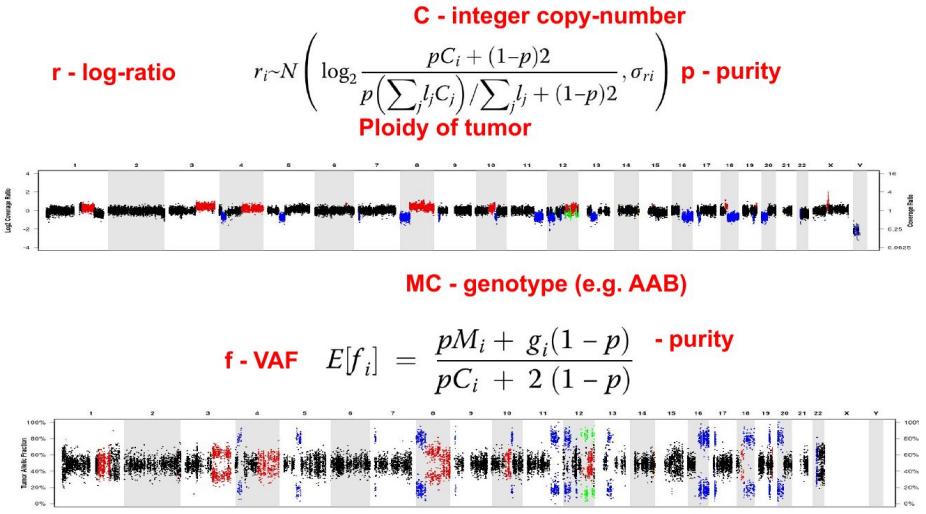


Figure 16: Mathematica model underlying the CNVEX algorithm. Top-panel log-ratios are modelled using a Gausian distribution and depend on absolute copy-numbers, tumor ploidy and purity. Bottom-panel Variant Allele Frequencies are distributed according to the binomial distributions and depend on purity and absolute copy number.

Overall, the CNVEX (while still in development) algorithm has been used to characterize CNVs across hundreds of patient samples. And provided important insights about the role of focal-tandem duplications (FTDs, see below) associated with the loss of CDK12 mutations.

➢ Specific Objective #3: Validation and benchmarking of bioinformatics pipelines. Clinical application of bioinformatics algorithms requires their validation, both using simulated data and by comparison orthogonal assays or other state-of-art methods. We have followed this general strategy to validate core components of the CRISP and CODAC pipelines.

The overall strategy is shown below. Briefly, performance is evaluated in terms of sensitivity and specificity. A training set is used to tune parameters of CODAC to (increase specificity) while relaxing the parameters of CRISP (increase sensitivity). The iterative application of this process results in improved calls and overall higher accuracy.

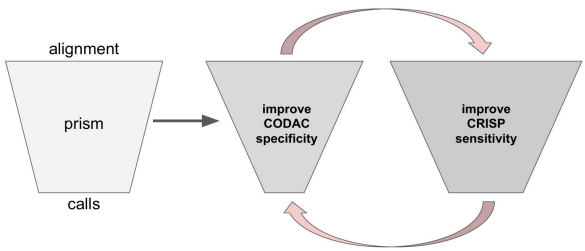


Figure 17: Iterative refinement of algorithm accuracy (fusion detection, neoantigen discovery, chimeric RNA classification)

After applying this strategy we tested the resulting algorithm against two competing approaches (GSNAP, TopHat-Fusion) and sanger sequencing (a wet-lab gold-standard technique. We achieved the following validation rates:

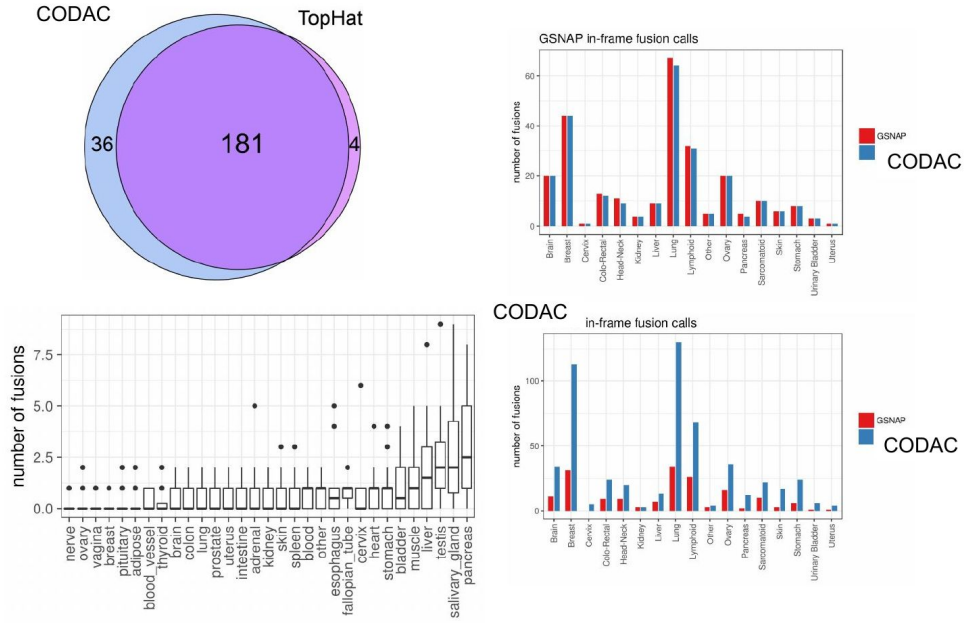


Figure 18: Validation and benchmarking of the CODAC algorithms. top-left improved sensitivity of fusion calling relative to Tophat, bottom-left very low number of false-positive calls in normal tissues, top-right excellent sensitivity in recovering GSNAP in-frame fusion calls, bottom-right overall improved sensitivity compared to GSNAP.

Accordingly, the most recent version of codac shows excellent accuracy approx. 95% sensitivity and 98% specificity as ascertained by comparisons with leading algorithms and sanger sequencing. The low number of false-positives are typically associated with extremely highly expressed genes and are easily “spotted” by experienced users. Further improvements will be achieved in future versions according to our development roadmap.

The extensive validation of the CNVEX algorithm is currently ongoing. The particular challenge lies in the accurate disambiguation of multiple highly-likely purity/ploidy combinations, as illustrated below.

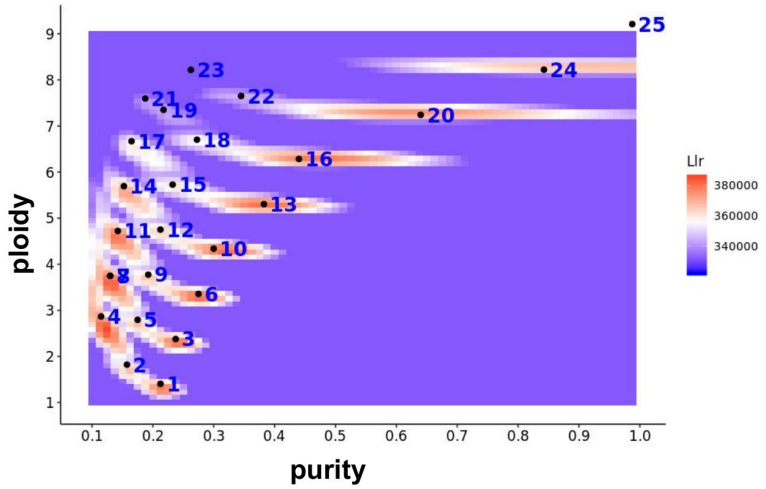


Figure 19: Purity-ploidy combinations have often very similar likelihoods (Llr), which requires detailed inference algorithms. Please not that the density of high-likelihood solutions is denser at low purity levels.

Therfore, validation of CNVEX is being done across a range of tumors with a particular emphasis on low-purity samples. These analyses are necessary to achieve good accuracy across the range of purities observed for mCRPC patients.

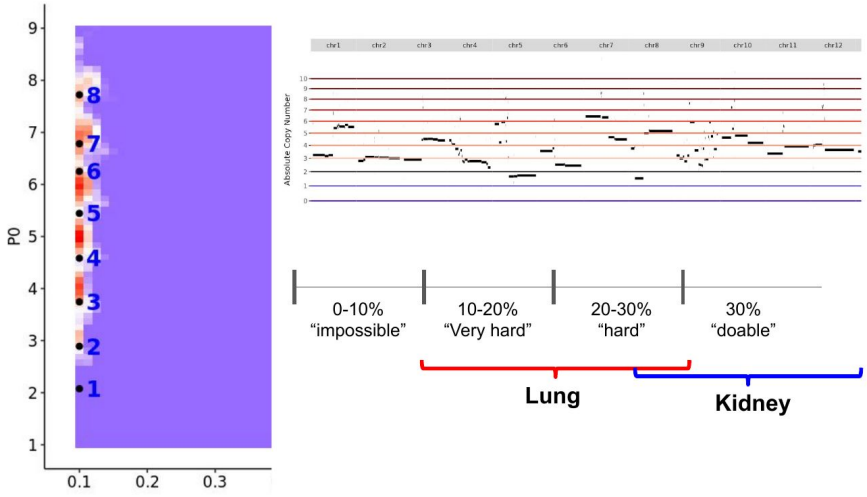


Figure 20: Illustration of CNVEX validations on a low-purity sample.

We also developed an interactive portal to examine the various purity-ploidy combinations, and evaluate the accuracy of probabilistic CNVEX choices. The portal is implemented as a shiny web-application.

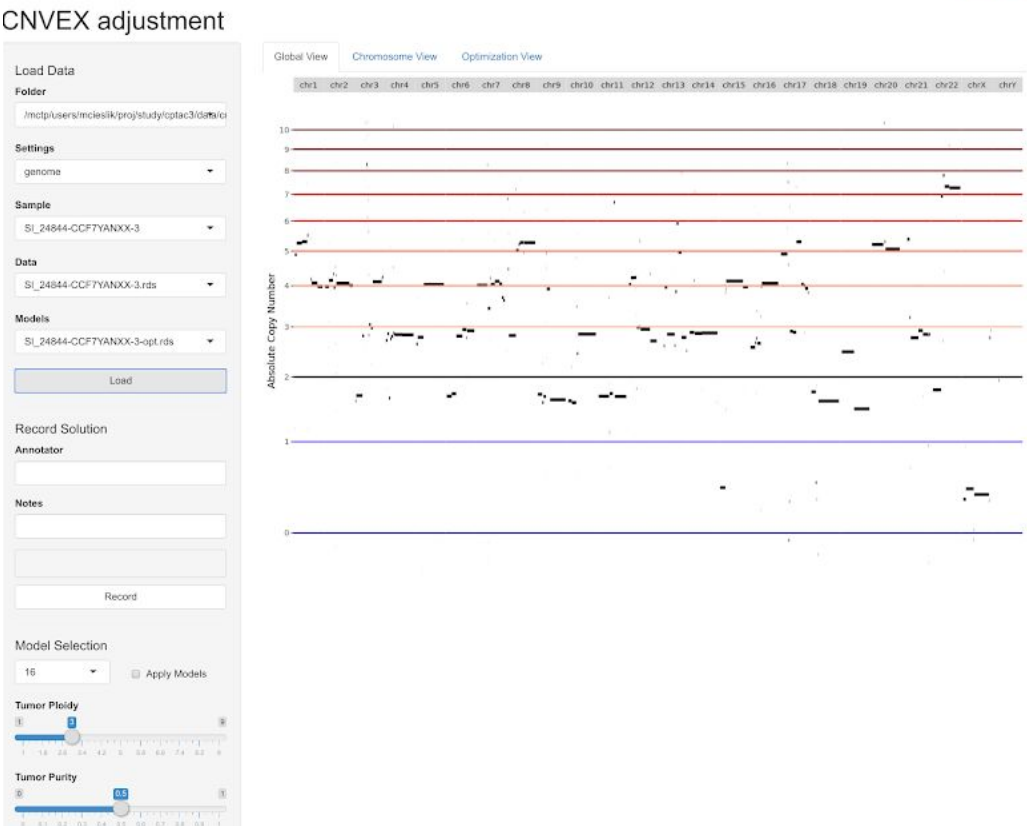


Figure 21: Browser-snapshot of a statistical copy-number analysis by CNVEX. Please note the ability to adjust purity and ploidy in real time (left-bottom corner), and the ability to explore multiple model selections.

➢ Specific Objective #4: Implementation of neoantigen pipeline. An critical aspect of modern immunogenomics is the identification and quantification of neoantigens. Towards that end, we have proposed the development of state-of-art pipelines to call neoantigens from next-gen sequencing data. The pipelines were implemented over the last year, by our bioinformatics team in collaboration with cancer immunologists, geneticists and pathologists. This collaborative effort results in algorithms of increased sophistication and breadth compared to alternatives, in terms of the types of genetic events generating neoantigens that are being ascertained.

Overall, our neoantigen pipelines utilize genetic calls produced by our genomic pipelines two implemented earlier MI-OncoSeq (internal), BCBIO (external, optimized and adapted to our needs), and CRISP/CODAC (internal, described in detail above). The results are being fed into a complex bioinformatics workflow, to detect mutant proteins, delineate altered peptides, genotype the patients HLA alleles, predict binding of the mutant peptides to the patient-specific HLA/MHC-I genotype and rank peptides based on their immunogenicity. The results of the neoantigen pipelines can then be used to desic neoantigen vaccines - or as is proposed herein to quantify the immunogenicity of a patient's tumor to aid in its classification from the immunological perspective.

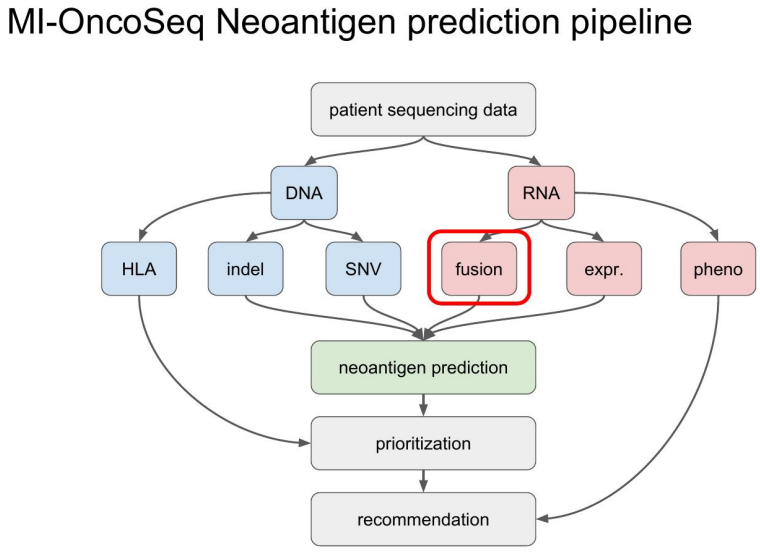


Figure 22: Flow-diagram of the implemented bioinformatics pipelines for neoantigen detection from integrative genomic sequencing data.

The above diagram shows how patient sequencing data in the form of DNA and RNA sequencing into appropriate genetic calls: HLA, indel, SNV, fusion, expression, and immune phenotypes is integrated in the process on neoantigen prediction, followed by prioritization based on parameters such as HLA binding and expression.

Next, we introduce the fusion neoantigen detection algorithm, which is a novelty compared to other state-of-art tools. Following discovery of chimeric junctions, and their aggregation into “breakpoints” overlapping reads are assembled into a contig (the quality of this assembly is judged by the number of reads covering contig), this contig is aligned back to the reference genome (as a form of additional support). If this alignment is valid we attempt to associate the contig with the the 5’ and 3’ transcripts in a process we refer to as stitching. If the contig only aligns to the 5’ transcript we call this a truncating fusion. If both 5’ and 3’ transcripts are stitched the result becomes an inframe or out-of-frame fusion we test this using in silico translation. The gist of the algorithm is presented in the two figures below.

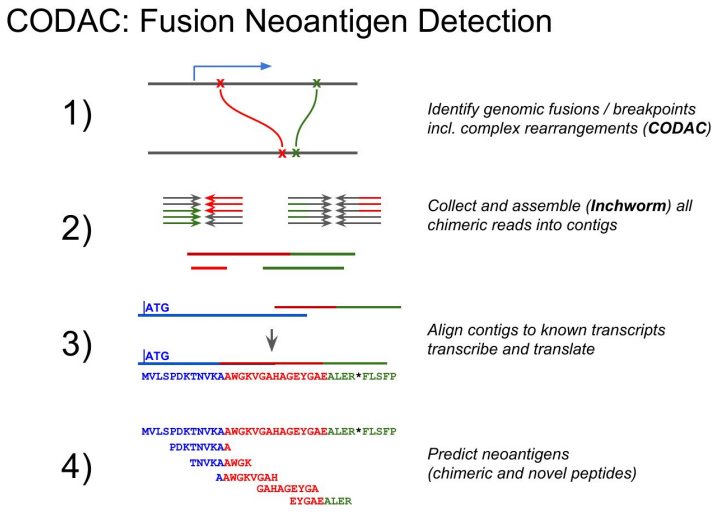


Figure 23: Assembly approach to fusion neoantigen through open-reading frame discovery.

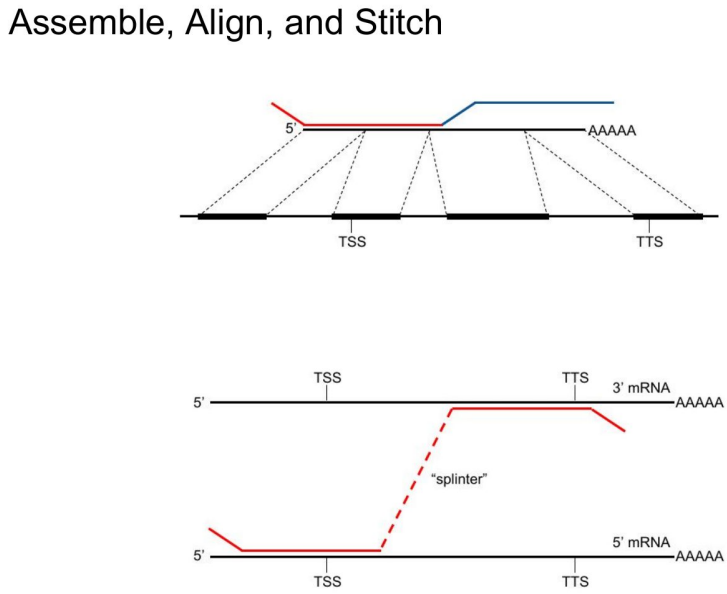


Figure 24: Stitching and Alignment step of the fusion-detection algorithm. “Splinter” refers to non-templated nucleotides that may be incorporated as part of the structural variant. Splinters have the potential to alter the frame.

The next step in the pipeline is to take the generated fusion proteins (as well as mutant peptides resulting from mutations detected from DNA sequencing) and score for their binding to the predicted MHC-I alleles. This is illustrated with the two following graphs:

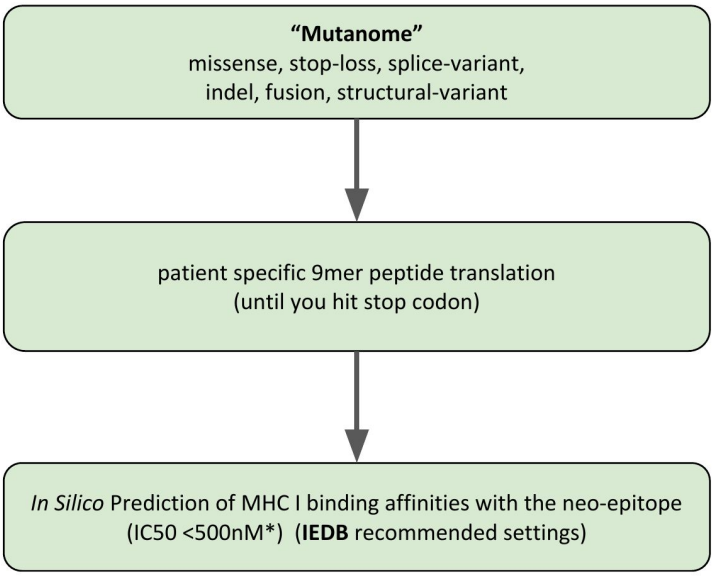


Figure 25: Overview of the peptide generation and scoring sub-module. The mutanome refers to different types of protein altering mutations. All mutation-overlapping 9mers are generated and used for binding scoring, which is done using IEDB.

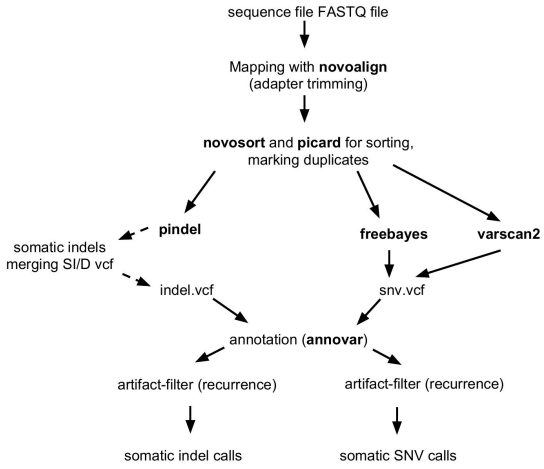


Figure 26: Delineation of protein altering mutations from DNA sequencing data. Both FreeBayes and VarScan2 are used to identify SNVs, and indels, respectively. Pindel is additionally used to increase the sensitivity of variant detection. All calls are subject to stringent filtering for false-positives e.g. artifacts.

A critical part of this procedure is HLA typing, which we carry-out using both HLAreporter and PHLAT (both of these algorithms were incorporated into the proposed pipelines), HLA-typing is necessary to achieve personalized binding prediction, which are needed for accurate immunogenomic profiling of individual patients:

Initial SI_11416			Population Freq				HLA quality profile: good quality			
MHC-I	Patient Allele(4digit)	[Allele]	[EUR]%	[CHN]%	[JPN]%	[AFR]%	10xcov%	20xcov%	30xcov%	50xcov%
HLA-A	A*02:06:01G	A*02:06	0%	6%	9%	0%	100	94	85	47
	A*31:01:02G	A*31:01	3%	3%	8%	0%	100	94	85	47
HLA-B	B*07:02:01G	B*07:02	13%	2%	6%	5%	100	95	86	53
	B*40:02:01G	B*40:02	1%	2%	8%	0%	100	95	86	53
HLA-C	C*03:04:01G	C*03:04	7%	12%	12%	5%	100	97	90	69
	C*07:02:01G	C*07:02	14%	18%	13%	6%	100	97	90	69

Initial SI_11416			Population Freq				HLA quality profile: low quality			
MHC-II	Patient Allele(4digit)	[Allele]	[EUR]%	[CHN]%	[JPN]%	[AFR]%	10xcov%	20xcov%	30xcov%	50xcov%
HLA-DPB1	DPB1*04:02:01G	DPB1*04:02	17%	3%	7%	11%	NA	NA	NA	NA
	DPB1*03:01:01G	DPB1*03:01	11%	7%	1%	14%	NA	NA	NA	NA
HLA-DQA1	DQA1*04:01:01G	DQA1*04:01	2%	1%	3%	2%	NA	NA	NA	NA
	DQA1*05:01:01G	DQA1*05:01	41%	18%	9%	17%	NA	NA	NA	NA
HLA-DQB1	DQB1*04:02:01	DQB1*04:02	3%	1%	4%	4%	NA	NA	NA	NA
	DQB1*02:01:01G	DQB1*02:01	23%	12%	1%	9%	NA	NA	NA	NA
HLA-DRB1	DRB1*03:01:01G	DRB1*03:01	12%	5%	0%	8%	NA	NA	NA	NA
	DRB1*08:02:01	DRB1*08:02	0%	1%	4%	0%	NA	NA	NA	NA
HLA-DRB3	DRB3*01:01:02G	DRB3*01:01	NA	NA	NA	NA	100	100	88	62

Figure 27: Example result of our HLA-typing (for both MHC class1 and MHC class 2, molecules). Perfect (100%) agreement is typically observed between PHLAT and HLAReporter.

To evaluate the relative contributions of mutations vs fusions to the overall neoantigen landscape (as proposed) we carried out extensive analysis of an index case with multiple available tissue specimens. In the results presented below we see that fusions contribute significantly to the overall neoantigen burden:

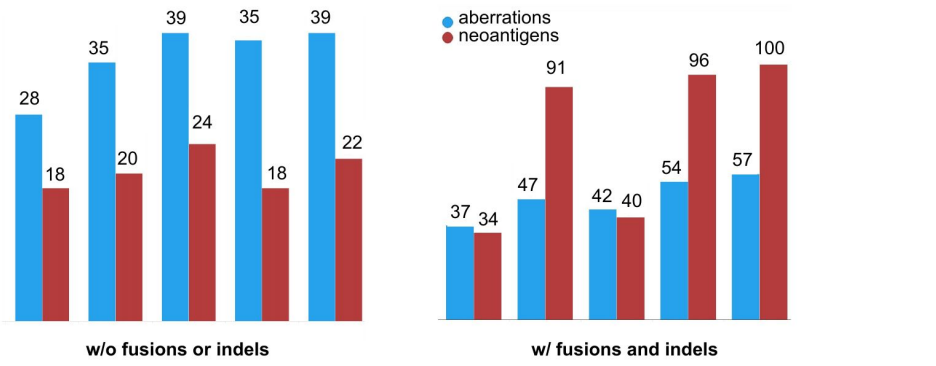


Figure 28: Relative contribution of fusion-derived neoantigens to the total neoantigen landscape in 5 selected index samples.

To test the inter-sample heterogeneity we further contrasted the overlap between the individual neoantigen landscapes derived from independent sequencing data sets (DNA and RNA, subject to the profiling as described above. We note extensive neoantigen heterogeneity between the different sets, suggesting the need to deeply profile each individual patient.

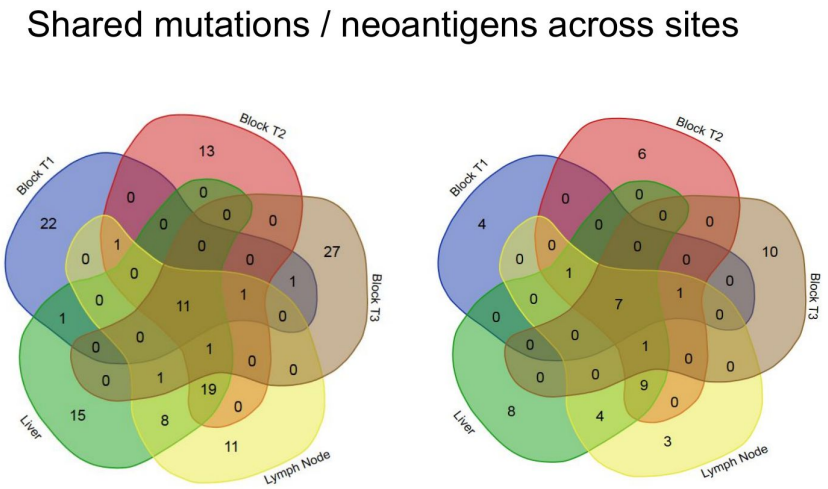


Figure 29: Overlap of DNA-based and RNA-based neoantigen calls between different biopsy sites (tissue blocks) from an index case patient.

The whole validated pipeline allows us to arrive at ranked lists of neoantigens individually for each patient:

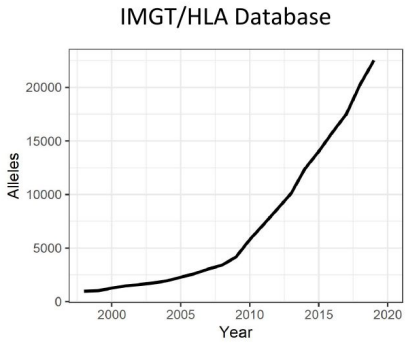
Gene	Location	Mutation Type	Protein Change	COSMIC Count	HLA Allele	pep length	MT peptide	IEDB_Recomended (method)	Binding
DAGLA	chr11:61507051	SNV	p.I591V	0	HLA-C*03:04	9	VALSASTPL	netmhcpan	SB
DAGLA	chr11:61507051	SNV	p.I591V	0	HLA-B*07:02	9	HPSDLTVAL	Consensus (ann/smm)	SB
CC2D1A	chr19:14034599	SNV	p.E639Q	0	HLA-A*02:06	9	FQQRTFSVI	Consensus (ann/smm)	SB
XPO5	chr6:43540245	SNV	p.N100D	0	HLA-A*02:06	9	LIADGTLNI	Consensus (ann/smm)	SB
TNRC6A	chr16:24804952	SNV	p.Q1112L	0	HLA-C*03:04	9	SSSVGPLAL	netmhcpan	WB
PSMG2	chr18:12720514	SNV	p.P138L	0	HLA-A*31:01	9	LQLRSTLFR	Consensus (ann/smm)	WB
DAGLA	chr11:61507051	SNV	p.I591V	0	HLA-A*02:06	9	VALSASTPL	Consensus (ann/smm)	WB
XPO5	chr6:43540245	SNV	p.N100D	0	HLA-C*03:04	9	IADGTLNIL	netmhcpan	WB
XPO5	chr6:43540245	SNV	p.N100D	0	HLA-B*40:02	9	MELIADGTL	Consensus (ann/smm)	WB
DAGLA	chr11:61507051	SNV	p.I591V	0	HLA-C*03:04	9	WTHPSDLTV	netmhcpan	WB
PIK3CA	chr3:178952085	SNV	p.H1047R	3302	HLA-A*31:01	9	FMKQMNDAR	Consensus (ann/smm)	WB
TNRC6A	chr16:24804952	SNV	p.Q1112L	0	HLA-C*03:04	9	WGSSSVGPL	netmhcpan	WB
DAGLA	chr11:61507051	SNV	p.I591V	0	HLA-A*02:06	9	WTHPSDLTV	Consensus (ann/smm)	WB
PSMG2	chr18:12720514	SNV	p.P138L	0	HLA-A*31:01	9	RSTLFRYLL	Consensus (ann/smm)	WB
TRIM56	chr7:100731115	INDEL	p.E175A	0	HLA-A*02:06	9	AQCPQHPGA	Consensus (ann/smm)	WB
DAGLA	chr11:61507051	SNV	p.I591V	0	HLA-C*03:04	9	HPSDLTVAL	netmhcpan	WB
XPO5	chr6:43540245	SNV	p.N100D	0	HLA-C*03:04	9	LIADGTLNI	netmhcpan	WB

Figure 30: Ranked list of predicted neoantigens based on their assigned HLA alleles and binding prediction: SB - strong binder, WB - weak binder. The IEDB recommended scoring algorithm is listed (ann, smm, netmhcpan).

Over the last years multiple mechanisms have been discovered which explain how tumors evade the immune system. Those mechanisms can be either reversible or irreversible and occur at the level of the genotype or phenotype. Recently, there is a stronger recognition that mutations in HLA-genes are a powerful mechanism to achieve

immunological invisibility. However, it is unknown whether mutations in HLA occur in metastatic prostate cancer. Detection of HLA mutations is challenging as they can occur through two mechanisms - single nucleotide / indel variants and loss of heterozygosity. Towards comprehensive immunogenomic characterization of mCRPC we are developing innovative algorithms to detect both classes of HLA mutations. Clinically, these mutations are of paramount importance because, based on existing data and research it is suggested that patients with a loss of HLA do not respond to immunotherapy. Unfortunately, no reliable tools exist to measure HLA mutations and HLA copy-loss and/or loss-of-heterozygosity (LOH) exist to conduct these analyses. The detection of HLA mutations is particularly challenging due to the extreme polymorphism of those loci.

- MHC class 1 – HLA-A, HLA-B, HLA-C
- Extreme polymorphism



HLA Class I			
Gene	***A***	***B***	***C***
Alleles	5,018	6,096	4,852
Proteins	3,398	4,248	3,193
Nulls	272	204	222

Figure 31: Extreme polymorphism of HLA loci in the human population. The number of known HLA alleles increases without a sign of saturations. In the Human population for each HLA gene at least 3,000 different proteins are known to exist.

Therefore we initiated two research projects to develop methods to detect HLA mutations from tumor sequencing data. The first project is to develop a novel algorithm to call mutations in HLA based on dynamic reference selection, and utilizing the benefits of alt-aware mapping against the GRCh38 reference. The development of which was initiated in 2019. The second project centered around detection of LOH will be initiated later in 2019/2020.

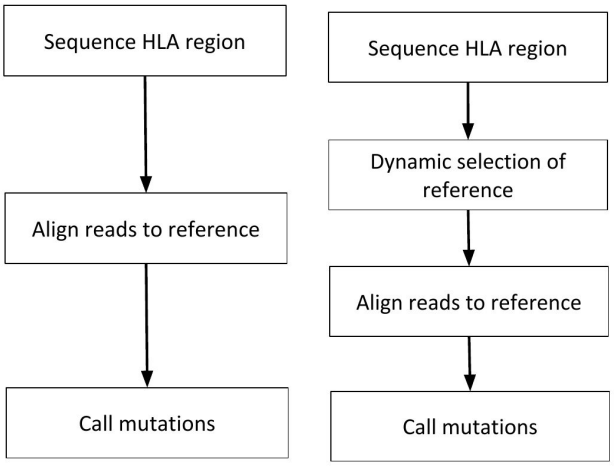


Figure 32: Illustration of our approach of Dynamic Reference Selection to Call mutations in HLA genes. A software tool - HLARS - which implements this overall algorithm is currently in development.

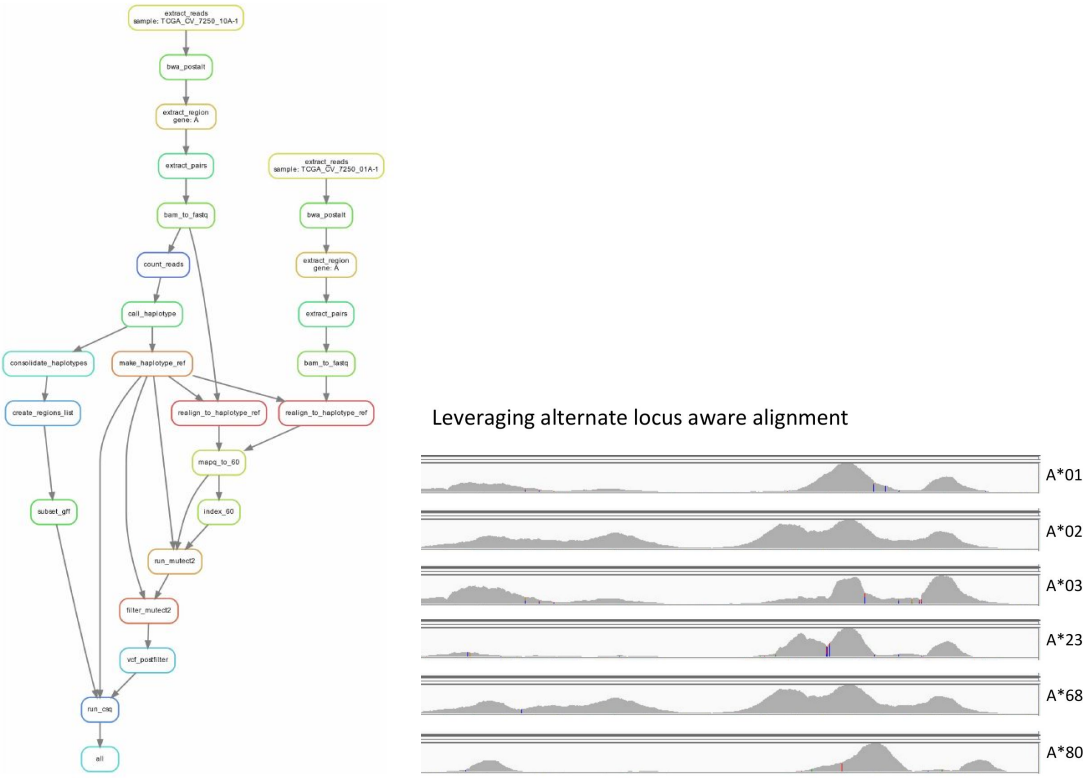


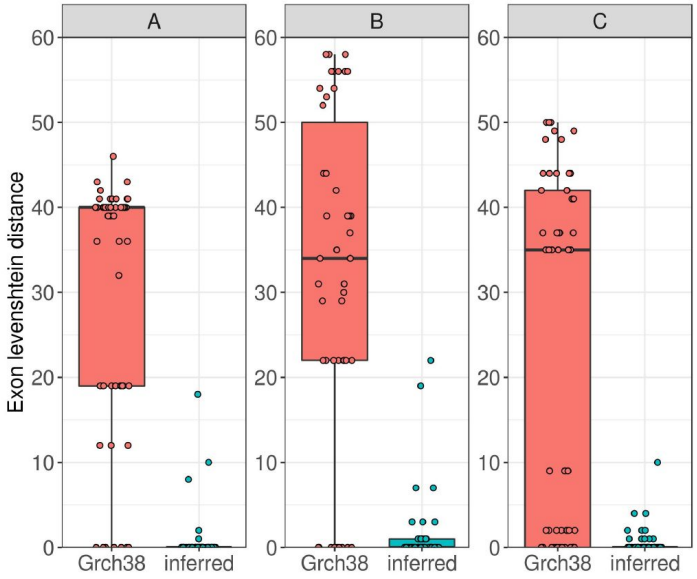
Figure 33: HLARS - HLA Reference Selection is a complex algorithms which starts with aligned sequencing data and proceeds through steps of HLA-typing, re-alignment and mutation calling. It leverages alt-aware alignment against all HLA loci at the same time.

HLARS implements a simple read de-convolution algorithm based on algebraic

transformations read alignment biases

$$\underset{A}{\begin{bmatrix} .90 & 0 & .44 & .01 & 0 & .02 \\ 0 & .97 & .01 & 0 & .32 & .01 \\ .32 & 0 & .90 & .02 & .01 & .06 \\ .01 & 0 & .02 & .99 & .01 & .02 \\ 0 & .22 & .03 & .01 & .99 & .01 \\ 0 & 0 & .04 & 0 & & 0.98 \end{bmatrix}} \underset{x}{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix}} = \underset{b}{\begin{bmatrix} 1371 \\ 2868 \\ 2834 \\ 49 \\ 715 \\ 117 \end{bmatrix}}$$

Figure 34: Mathematical example behind HLARS, here be refers to the observed read-counts for different HLA alleles while A is a read-bias alignment matrix inferred from simulations of reads followed by alignment using an alt-aware algorithm.



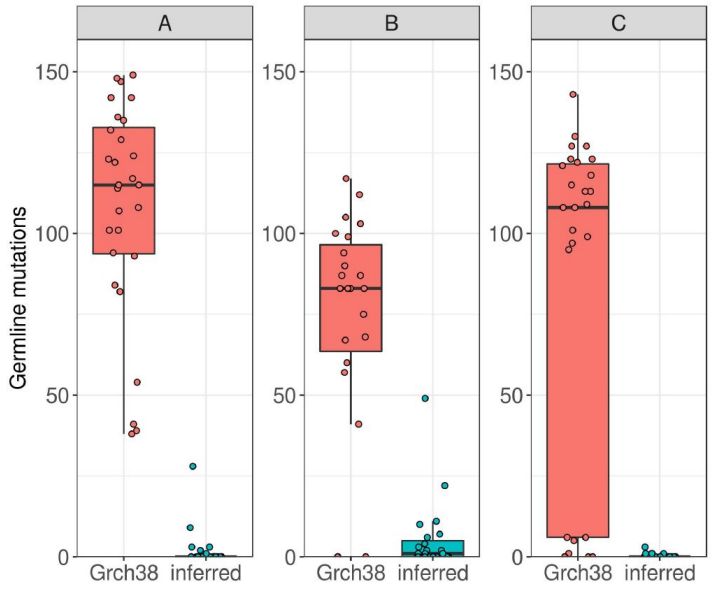
Mean

	Grch38	Dyn.
HLA-A	30	0.7
HLA-B	32	1.6
HLA-C	23	0.5

Median

	Grch38	Dyn.
HLA-A	40	0
HLA-B	34	0
HLA-C	35	0

Figure 35: HLARS validation using reference genomic sequences data regarding the expected difference in the chosen reference relative to the reference included in the canonical HLA locus in GRCh38.



Mean

	Grch38	Dyn.
HLA-A	108	1.6
HLA-B	74	5.2
HLA-C	82	0.3

Median

	Grch38	Dyn.
HLA-A	115	0
HLA-B	83	1
HLA-C	108	0

Figure 35: Same as **Figure 34** with the exception that now the number of germline mutations is counted.

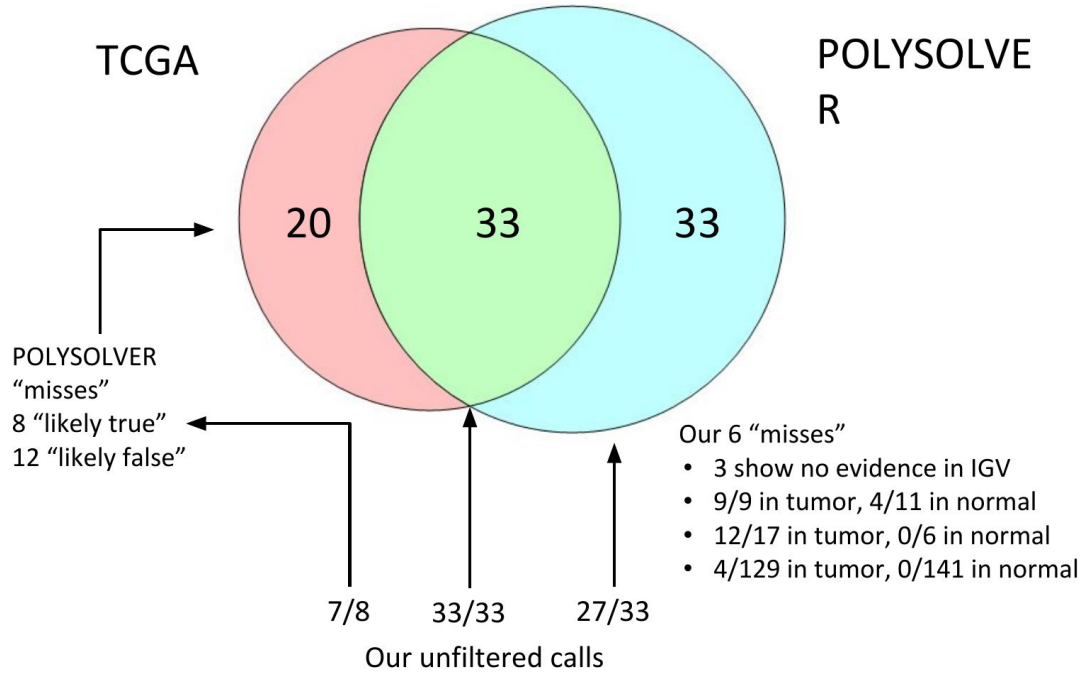


Figure 36: Validation of HLARS against Polysolver (a competing algorithm) and the standard approach (TCGA), followed by manual evaluation of conflicting calls. These preliminary results suggest that HLARS has improved sensitivity relative to existing approaches.

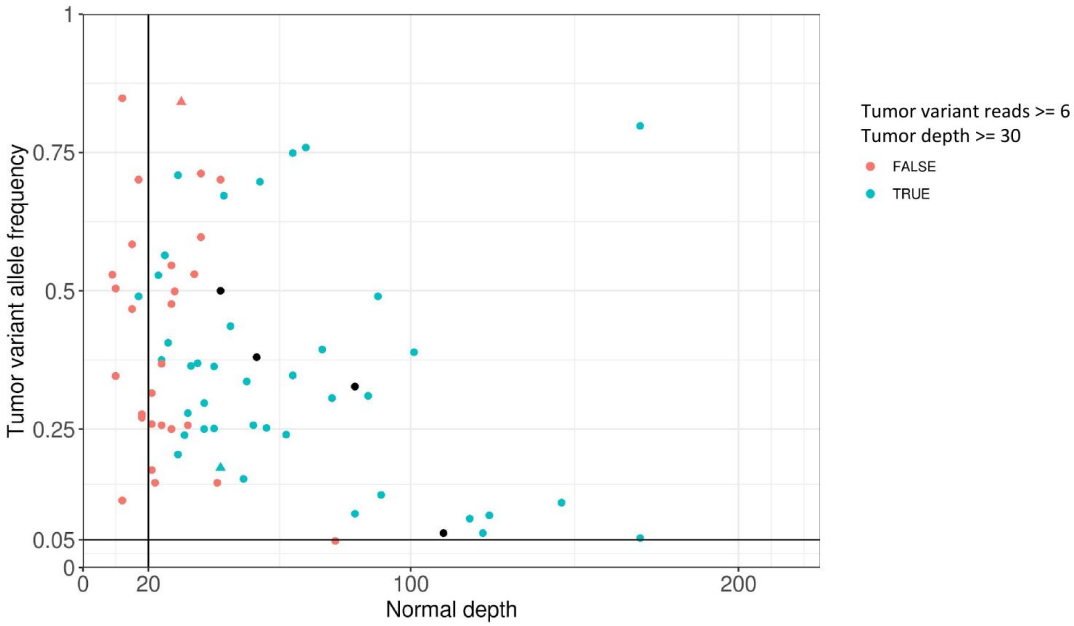


Figure 37: HLARS calls are contrasted with Polysolver calls on the plane of normal coverage and tumor variant allele frequency. HLARS can detect all mutations previously discovered by Polysolver and additionally detect 4 mutations out of ~80 cases which suggest significantly higher sensitivity.

Overall, the above validation results suggest that HLARS has state-of-art sensitivity and specificity. Further it suggests that the overall rate of mutations in HLA have been under-estimated by a factor of 50% in primary tumors. The development of HLARS may therefore result in the implementation of novel predictive biomarkers for immunotherapy, as patients with HLA mutations (now estimated at ~8% in primary cancer), are unlikely to respond to immunotherapy.

➢ Specific Objective #4: To finalize the MImmScore workflow.

An important aspect of our work was to finalize the MImmScore workflow which has been presented as preliminary data in our proposal. Briefly, the MI-ImmScore is capable of quantifying immune infiltration across cancer types. As shown below the MImmScore correctly predicts that prostate cancer (PRAD) has low immune infiltration, while melanoma (SKCM) has high immune infiltration

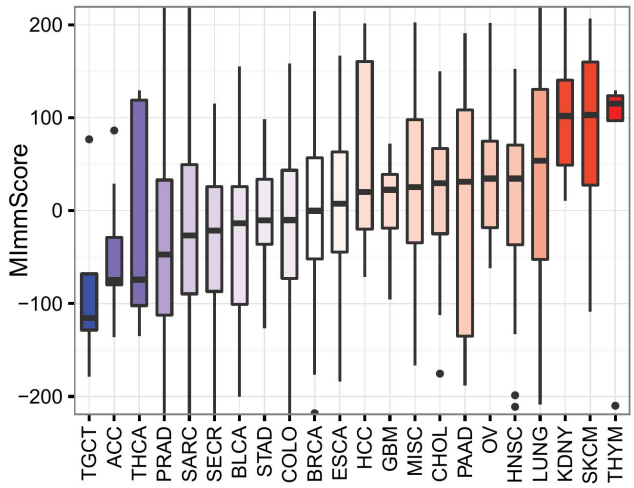


Figure 38: MImmScore across different cohorts of metastatic tumors. Higher MImmScores correspond to a larger magnitude of immune infiltration, important cohorts: BRCA - breast cancer, PRAD - prostate cancer, KDNY - kidney cancer, SKCM- melanoma.

Next, we set forth to validate that the MImmScore is reproducible and stable across different biopsy sites. We found that MImscore levels are robust to technical variation and accurately reflect differences in infiltration levels across different biopsy sites.

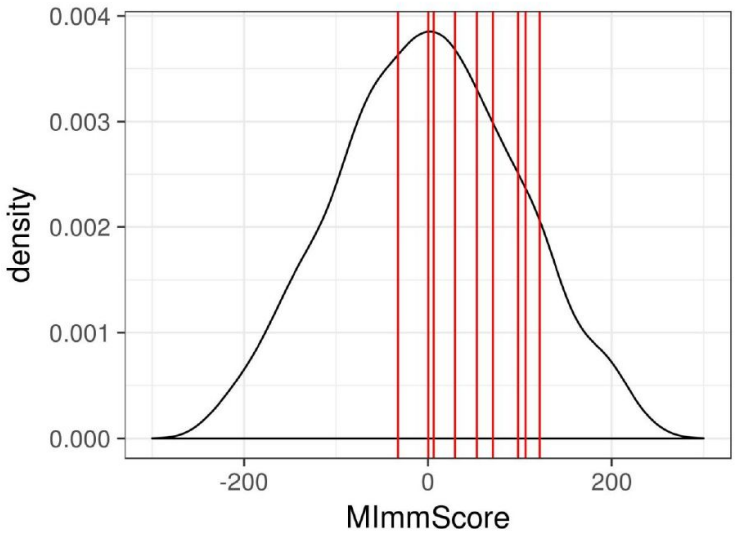


Figure 39: Distribution of MImmScores across a large cohort (n=497) of metastatic tumors. In the following image the black line indicates the distribution of MImmScore levels across a large cohort of metastatic tumors.
The red lines indicate individual samples from one index patients (same as above), subject to RNA-sequencing and MImmScore analysis (as described in the referenced publications).

Next we decided to generalize the MImmScore to whole-cohort analyses. As proposed we focused on prostate cancer. We stratified prostate cancer patients by mutation status (ETS - ETS fusion positive, ATM - ATM deficient, HRD - homologous recombination deficient not ATM, CDK12 - biallelic loss of ATM and MMRD - mismatch repair deficient).

We found that a mathematical summary of individual-level MImmScores results in a highly stable cohort-level MImmScore, which can be used to sub-classify metastatic tumors based on immune status (this is the overarching goal of the proposal):

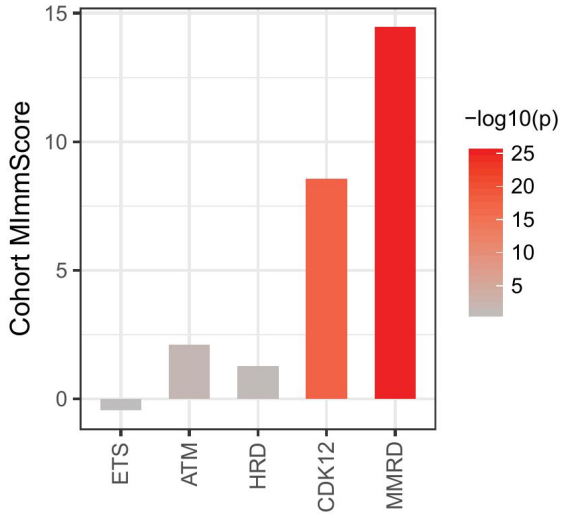


Figure 40: Cohort level MImmScore applied to a cohort of 300+ metastatic prostate cancers. Higher Cohort MImmScores signify that immune infiltration is higher in a given sub-cohort as opposed to individual patients.

We also proceeded to validate the MImmScore against independent experimental methods to determine the level of T-cell infiltration. We used T-cell repertoire sequencing using capture RNA-seq based data, to measure the numbers and clonal expansion of tumor infiltrating T-cells. This assay is also RNA-based and hence subject to different biases as the RNA-based MImmScore.

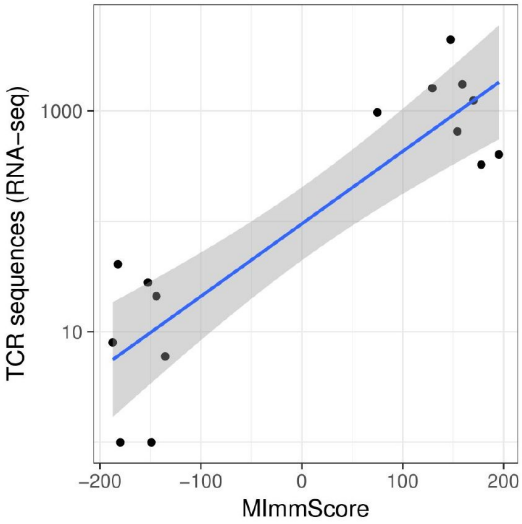


Figure 41: Correlation of MImmScores and total numbers of infiltrating T-cells as estimated from the numbers of reads overlapping their unique CDR3 sequences.

2. Extended Immune Profiling Development (Major Goal #2) and Validation of Computational Methods (Major Goal #3).

   Validation of computational methods, particularly those which have an inferrential component (such as methods to deconvolute infiltrating immune subtypes), are critical for their firm establishment and routine use. As proposed we followed three main strategies to validated our experimental methods:

   1. Validation against non-inferential algorithms.
   2. Validation against DNA-based assays.
   3. Validation against immunohistochemistry-based assays.

   We followed all these approaches to independently validate the individual modules of the proposed bioinformatics pipelines (i.e. the ones presented above)

   ➢ Specific Objective #5: Experimental validation of Fusion calling

   Our goal was to validate the sensitivity and specificity of fusion calling in addition to the algorithmic methods described above. RNA-seq fusions can be validated by FISH or by DNA-sequencing. Given the higher accuracy of DNA-based methods we opted for this approach. An additional method to validate fusions is to carry out Sanger sequencing of the cDNA.

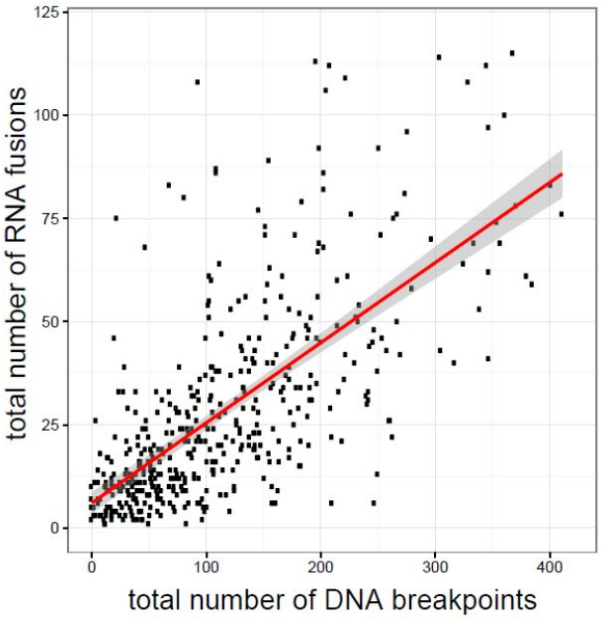


Figure 42: Validation of overall fusion burden calls against DNA-based prediction of breakpoints based on Whole Exome Sequencing (WES) Copy Number Variant Analysis (CNV). The red line show the high correlation between both approaches.

As shown above we achieve high correlation in the total number of breakpoints detected through fusion detection against standard WES CNV analysis. This demonstrated that our algorithms are asymptotically accurate. We also noted a striking increase in fusions

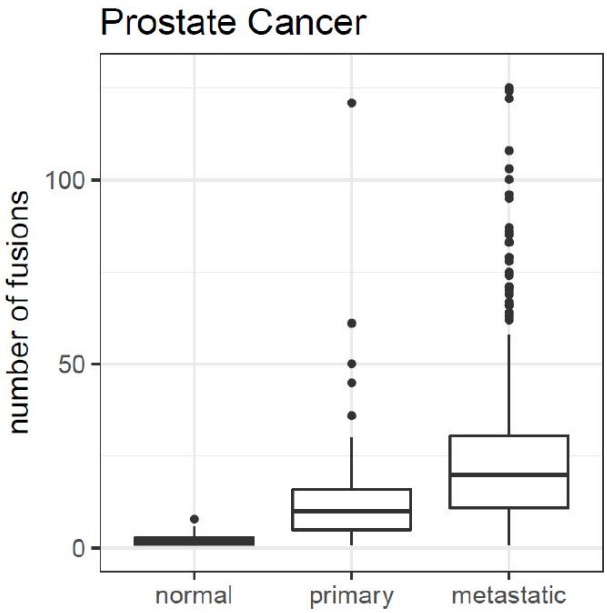


Figure 43: Systematic increase in the number of detected gene fusions between normal, primary, and metastatic prostate tissues.

To further prove that the resulting calls are individually, as opposed to in bulk, accurate. We carried out Sanger sequencing of candidate breakpoints.

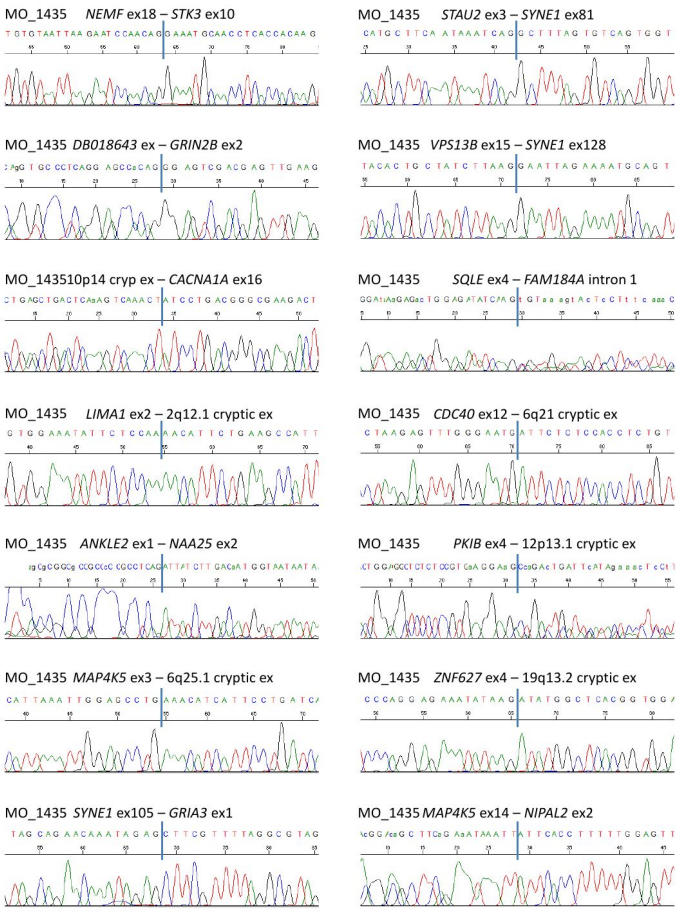


Figure 44: Example sequencing traces of candidate fusions. The breakpoint (chimeric) is indicated in the middle of the trace with a blue line.

Shown above are example traces of the hundreds of fusions we attempted to validate. Overall we estimate our validation (i.e. specificity) at 98%.

➢ Specific Objective #6: Experimental validation of T-cell repertoire (TIL) profiling methods.

T-cell repertoire profiling is a powerful technique to profile TIL infiltration into tumor tissues. It relies on the accurate delineation of CDR3 overlapping sequences from next generation sequencing (NGS) data. This means that either DNA or RNA can be sequenced in order to confirm TIL presence. Multiple algorithms have been proposed for the task of CDR3 sequence assembly and alignment. And we used MIXCR in our implemented bioinformatics pipelines.

To validated TCRB-seq we first assessed it's agreement with expected biological genotypes and phenotypes. Specifically, we expect that TIL infiltration is correlated with the immunogenicity of a tumor, and that it results in the increased clonality of the T-cells. We test this on a set of index cases:

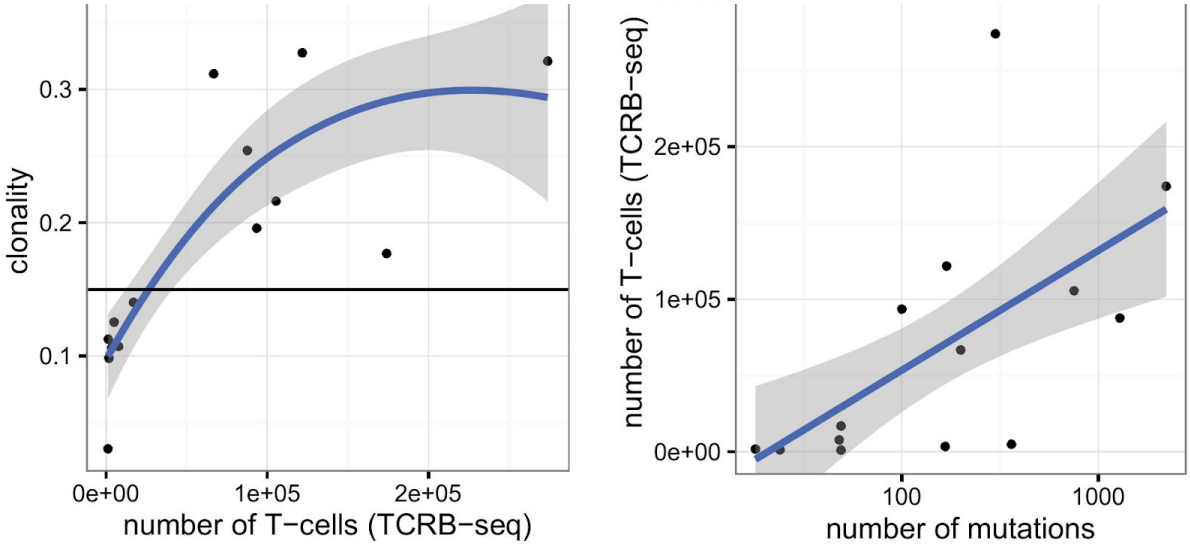


Figure 45: Association of TCRB-seq (i.e. the total number of T-cells estimated from the total number of CDR3 reads), with clonality (i.e. the Gini index of T-cell clonal expansion), and the total number of mutations per patient.

We also used a similar analysis to see if increased T-cell infiltration correlated with a increase in the total number of T-cells, and decreased in the Treg/Tcell CD8+ ratio:

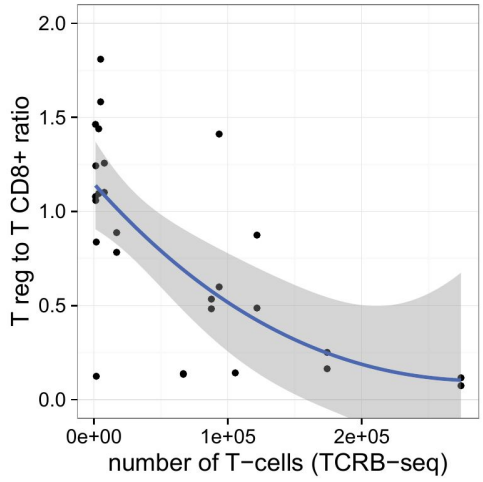


Figure 46: Association of TCRB-seq (i.e. the total number of T-cells estimated from the total number of CDR3 reads), with the ration of suppressive to activating T-cells. A significant trend was observed.

The above results show that the observed results (increased presence and clonal expansion of T-cells) are in agreement with expectation, and thus rationally validate the quantification approach.

Next, we comprehensively validated RNA-seq based TIL estimates to DNA-based TIL-estimates using a large number of paired samples for which both DNA and RNA was available:

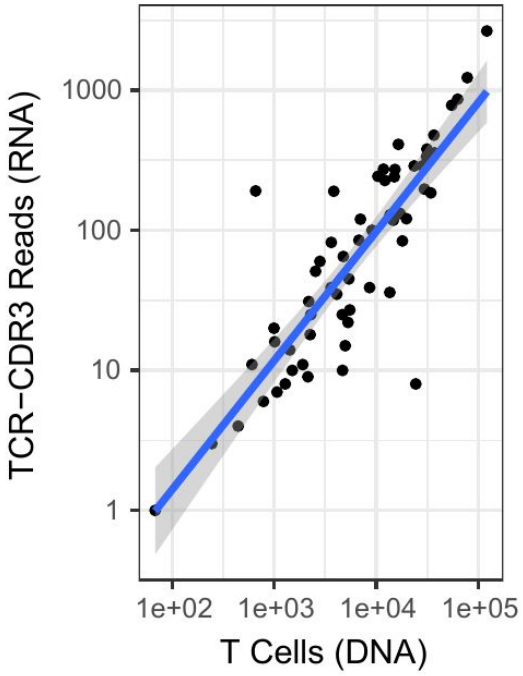


Figure 47: Comprehensive validation of DNA and RNA based T-cell repertoire sequencing. High correlation is observed, although DNA-based methods show a higher overall sensitivity.

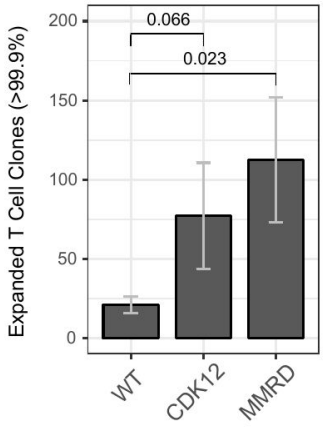


Figure 48: clonal expansion of T-cells in prostate cancer tumors with differing types of driver mutations

As shown above we also demonstrated that clonal expansion of T-cells is concordant with increased mutations (neoantigen burdens) associated with different types of metastatic prostate cancer

A similar concordant conclusion can be reached by inspecting the results from RNA-seq based TCR as shown below:

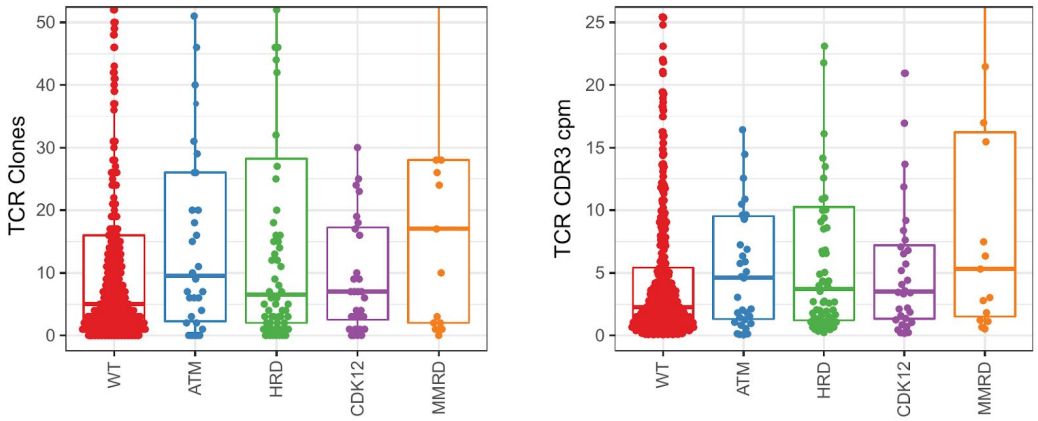


Figure 49: Different levels of T-cell clones (clonal expansion) and total numver of CDR3 reads in different genetic classes of prostate cancer.

Next we validated that a similar increase in clonal expansion can be detected through DNA-based sequencing of the T-cell receptor. For each genotype we used index samples from 10 different prostate cancer patients and aggregated the T-cell clones. The number of templates is an estimate of the total number of T-cells in a given clone.

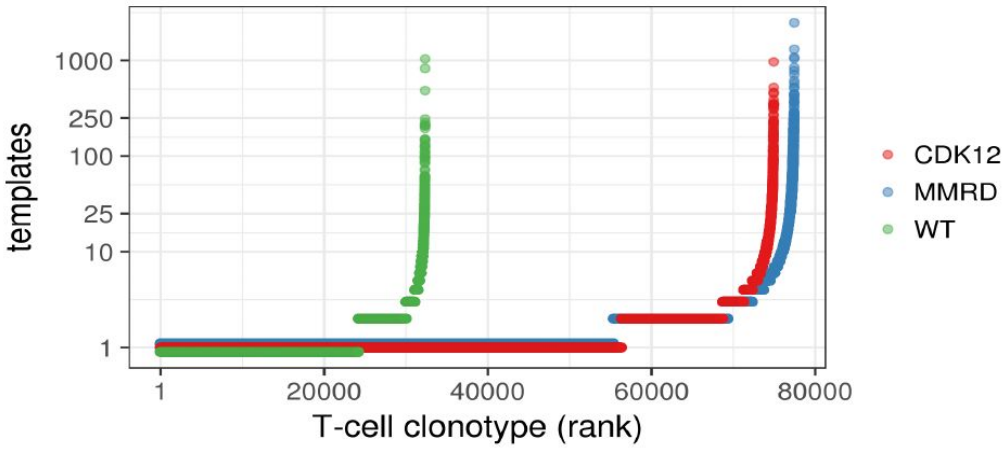


Figure 50: Hockey-stick plot of T-cell clones present in aggregate samples of a given genotype. The results are concordant with RNA-seq.

Finally, we proceed to validate immune infiltration, using IHC. We selected a large number of samples for which both RNA, DNA, and FFPE tissue was available - we proceeded to stain for infiltrating T-cells in the tissue blocks. Sections were quantified for their immune infiltration levels and contrasted with samples predicted to have not infiltration.

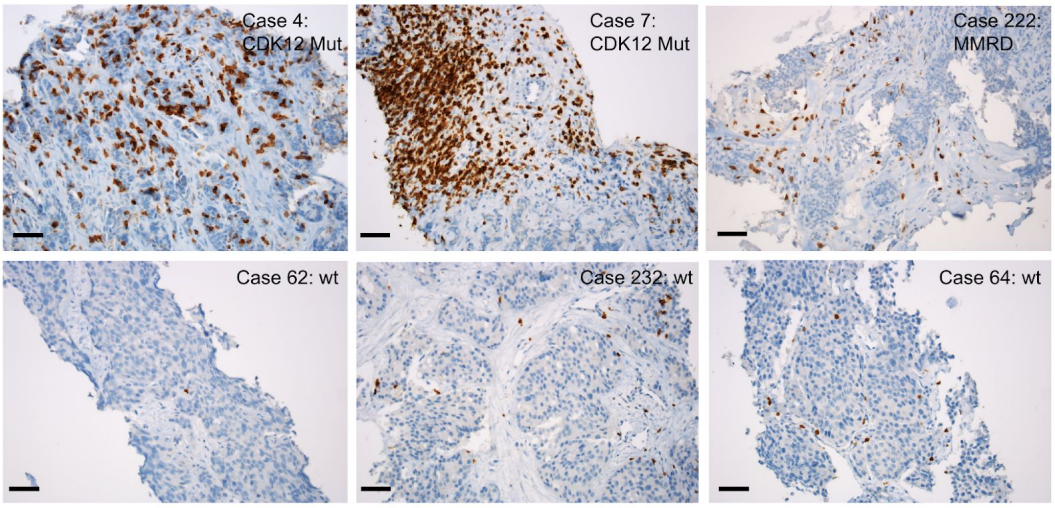


Figure 51: Representative examples of differences in T-cell infiltration levels between

➢ Specific Objective #6: Extended Immune Profiling Development

We also proposed the development of additional (extended) methods for immune profiling. Accordingly, We have implemented a number of bioinformatics algorithms algorithms and statistical methods, based on the premise that the metastatic transcriptomes contains sufficient information for the quantification of different types of immune cells. We focused on two areas typically neglected in similar attempts: pathway activity of the infiltrating immune cells and chemokine signaling. Overall these approaches represent the state-of-art in terms of characterizing the immune landscape of metastatic prostate cancer.

To develop the methods we identified CDK12 mutant prostate cancer as a molecular subtypes of prostate cancer with extensive immune infiltration. We then proceeded to develop differential expression analysis and the summary of the resulting logFC at the level of individual chemokines/cytokines or pathways.

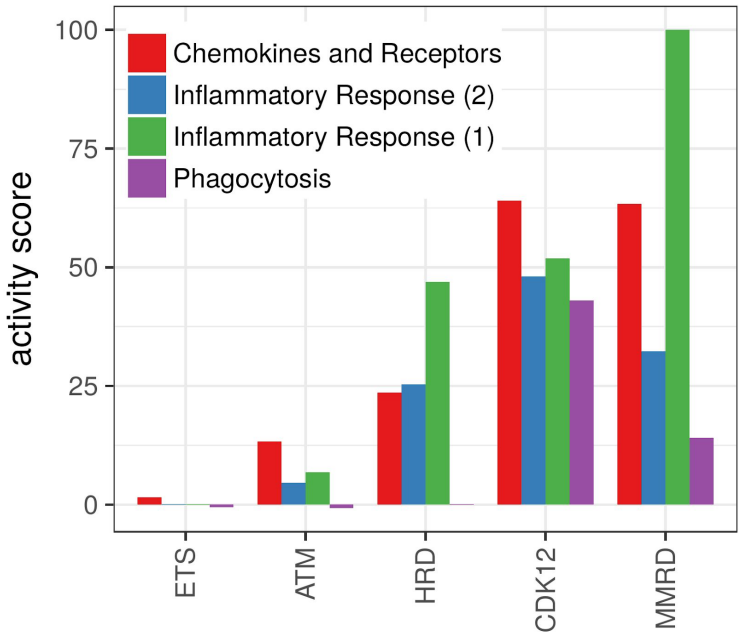


Figure 52: Increasing levels of immune infiltration in metastatic prostate cancers, according to their genetic status. CDK12 mutant tumors, and mismatch repair deficient tumors (MMRD) are the most immunogenic.

The above, graph shows that as expected - activities of immune pathways are increased in CDK12 mutant tumors, and mismatch repair deficient tumors.

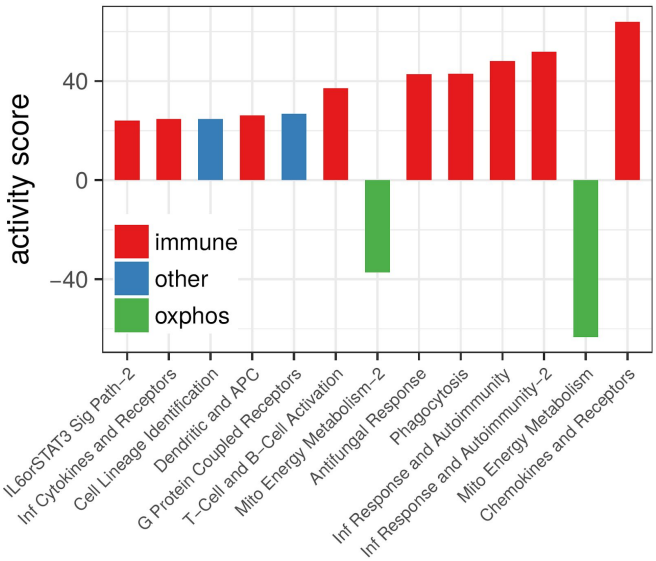


Figure 53: Differential pathway activities in CDK12 mutant tumors show, that immune pathways show a high level of activation in this molecular subtypes of prostate cancer.

As shown above, our approaches enable us to drill-down with high resolution on the

individual pathways up or downregulated in a given molecular subtypes of prostate cancer. Hence we are able to better understand the immunological characteristics of prostate cancers.

This approach can be even further refined to drill down on individual genes, here we focused on the specific chemokines that are being up-regulated or down-regulated in metastatic prostate cancer:

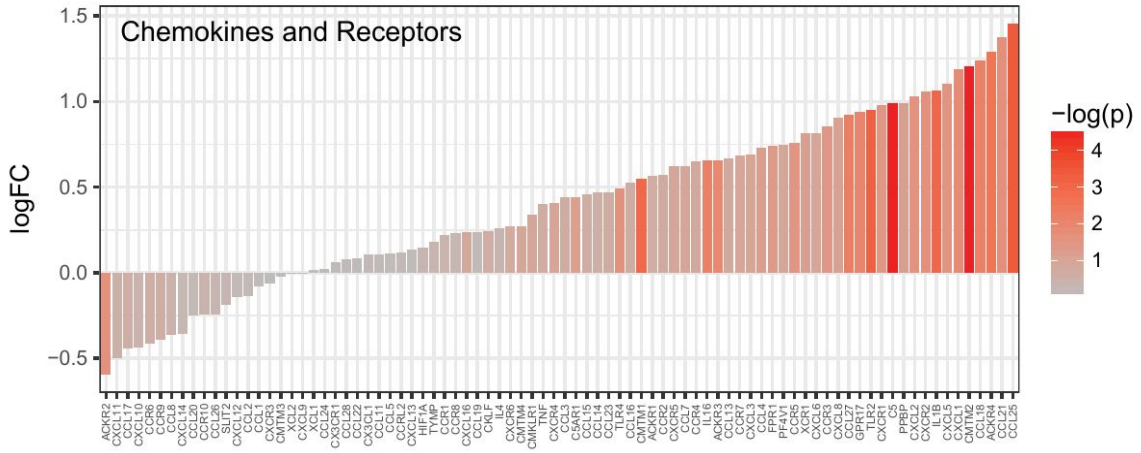


Figure 54: Differential expression of Chemokines (including cytokines) and their receptors in CKD12 mutant prostate cancer as compared to other types of tumors.

Formal statistical description of those methods are ongoing. Briefly those statistical approaches levarache advanced techniques such as moderated negative-binomial linear models, mixed effects models, independent component analyses. Further the methods leverage extensive databases of annotated molecular pathways with a focus of adaptive and innate immunity in human and model organisms.

3. Association of immunogenomic phenotypes with outcomes in metastatic castration-resistant prostate cancer.

As summarized above we have rapidly transitioned from the descriptive phase of the project towards clinical and outcomes research. Our discovery that CDK12 mutant tumors are associated with a distinct immunogenic phenotype, resulted in the initiation of a clinical trial to test the efficacy of immune-checkpoint inhibitors in patients with CDK12 mutations in prostate cancer and other indications.

Journal of Clinical Oncology
An American Society of Clinical Oncology Journal

Enter words / phrases / DOI / ISBN / authors / keywords / etc.

Newest Articles | Issues | Browse By Topic | Special Content | Authors | Subscribe

GENITOURINARY (PROSTATE) CANCER

IMPACT: Immunotherapy in patients with metastatic cancers and CDK12 mutations.

Melissa Andrea Reimers, Wassim Abida, Jonathan Chou, Daniel J. George, Elisabeth I. Heath, Rana R. McKay, ...

Show More

Abstract Disclosures

Abstract

TPS5091

Background: Tumors with biallelic CDK12 loss have been identified as a distinct subtype in metastatic castration resistant prostate cancer (mCRPC) and other cancer types. The CDK12 biallelic loss mCRPC genomic signature, distinct from homologous recombination deficient (HRD) and ETS fusion signatures, is characterized by excessive tandem duplications, genomic instability, gene fusion-caused putative neoantigens, and increased tumor T cell infiltration. Early clinical experience with anti-PD-1 immunotherapy in CDK12 loss mCRPC patients (pts) is notable for deep and sustained PSA as well as radiographic responses. We hypothesize that CDK12 biallelic loss is a potential biomarker of immune checkpoint immunotherapy (CI) efficacy in mCRPC and other cancers. **Methods:** IMPACT (NCT03570619) is a multi-center, open label, phase 2 study of pts with metastatic cancers that harbor CDK12 biallelic loss. mCRPC pts will be enrolled in cohort A (n = 25) in a Mini-Max Simon Two-Stage design, and all other pts in single-stage cohort B (n = 15). All pts will receive induction therapy with nivolumab 3 mg/kg IV and ipilimumab 1 mg/kg IV q3 weeks for up to 4 cycles, followed by maintenance nivolumab at 480 mg IV q4 weeks (up to 52 weeks in total). Eligible pts must have identified biallelic CDK12 loss on any CLIA/CAP approved next generation sequencing assay and a histologic diagnosis of metastatic prostate adenocarcinoma or other metastatic carcinoma. No prior ICI is allowed. The primary endpoint is the overall response rate (ORR) in cohort A per PCWG3 criteria. An ORR of 30% is targeted in cohort A. Secondary endpoints include safety, secondary efficacy measures, quality of life, and survival measures. Exploratory objectives include tumor whole exome analysis and changes in immune profiles with therapy. Comprehensive and serial monitoring of peripheral blood immune cell populations will be performed via T cell clonal diversity assessment and multi-parametric flow cytometry. Changes in myeloid and lymphoid populations will be assessed from whole blood. Polarization and effector function of T cells and activation of antigen presenting cells will be further characterized from isolated peripheral blood mononuclear cells. Study accrual is ongoing. Clinical trial information: NCT03570619.

Figure 55: The IMPACT trial for CDK12 mutant prostate cancer has been reported at ASCO 2019.

In parallel we have contributed to a number of studies which assess outcomes in prostate cancers, and genetic drivers thereof. In a publication in Cell 2018 we have contributed to the understanding of the prevalence of and role of structural variation in metastatic prostate cancer genomes, and further confirmed our discovery that CDK12 mutations result in a distinct pattern of focal tandem duplications and genomic instability. This project also fostered deeper collaboration between the University of Michigan and academic institutions on the west coast, which resulted in a quicker dissemination of our findings and insights related to immunogenomic phenotypes.

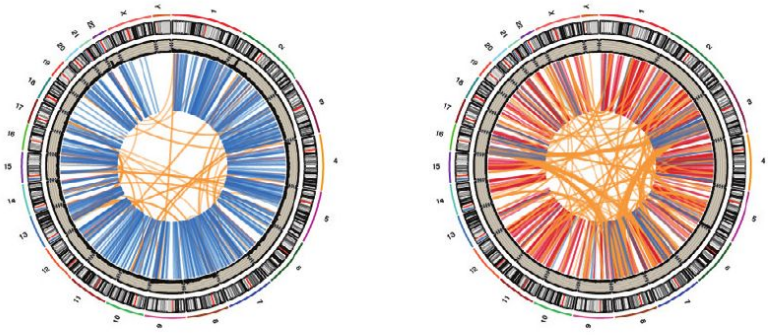


Figure 56: Independent confirmation of our finding that CDK12 loss is associated with a tandem-duplicator phenotype (right). Deletions associated with BRCA2 homozygous deletion are shown as control.

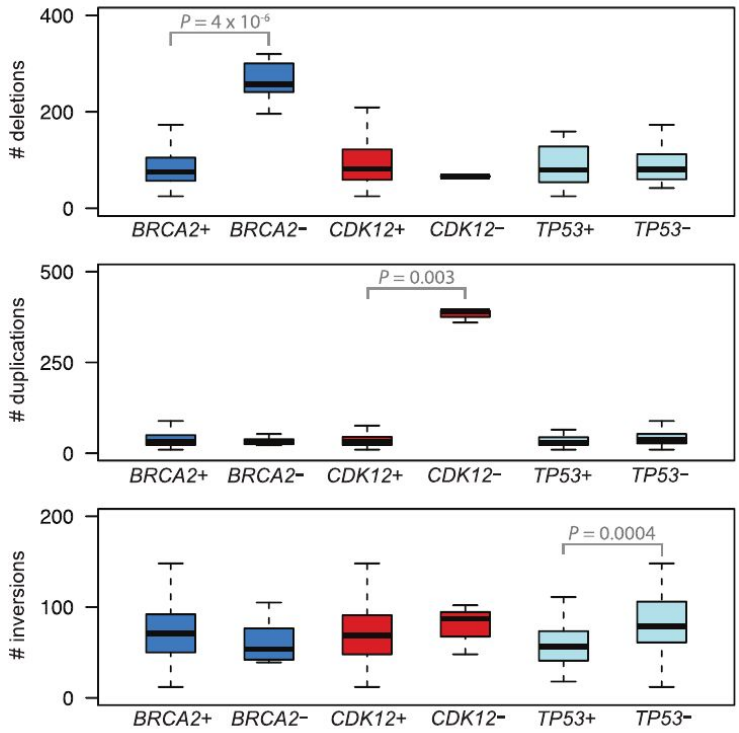


Figure 57: Further independent quantifications which confirm our assessments of genomic instability in mCRPC.

We have found that TP53 mutant tumors have higher overall levels of genomic instability particularly inversions. The implications of this finding for patients are unknown, particularly on how these aberrations impact the tumor immune phenotypes. Hence, towards comprehensive immunogenomic characterization of mCRPCs we will assess the role of TP53 independently and in conjunction with other types of aberrations.

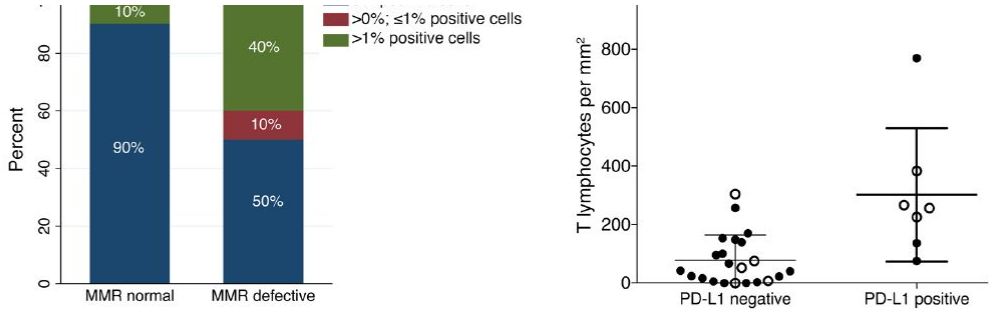


Figure 58: Independent validation of our finding that mismatch repair deficient tumors are associated with a higher level of immune infiltration and potentially prognostically significant (see above), and predictive for immunotherapy response.

Detailed genomic analyses carried out as part of this proposal continue being shared with collaborators within the SU2C consortium which results in many discoveries including prognostication of mCRPCs.

As proposed, we have carried-out a detailed analysis clinical and outcomes parameters in our cohort of mCRPCs, as well as multi-institutional validations.

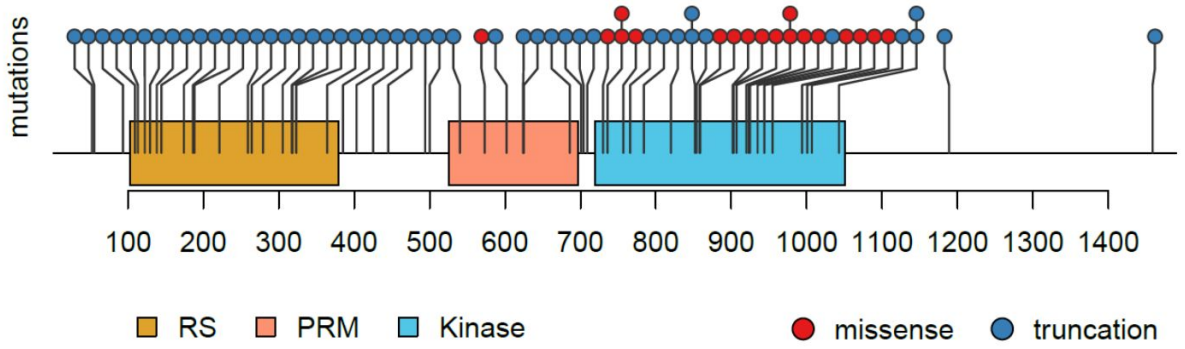


Figure 59: Multi-institutional analysis of CDK12 mutations in tumor specimens as well as detected from cell-free DNA / blood samples. The overall distribution of mutations in CDK12 has been confirmed, in particular the presence of missense mutations in the kinase domain. As expected very few mutations are found to truncate the protein after the kinase domain.

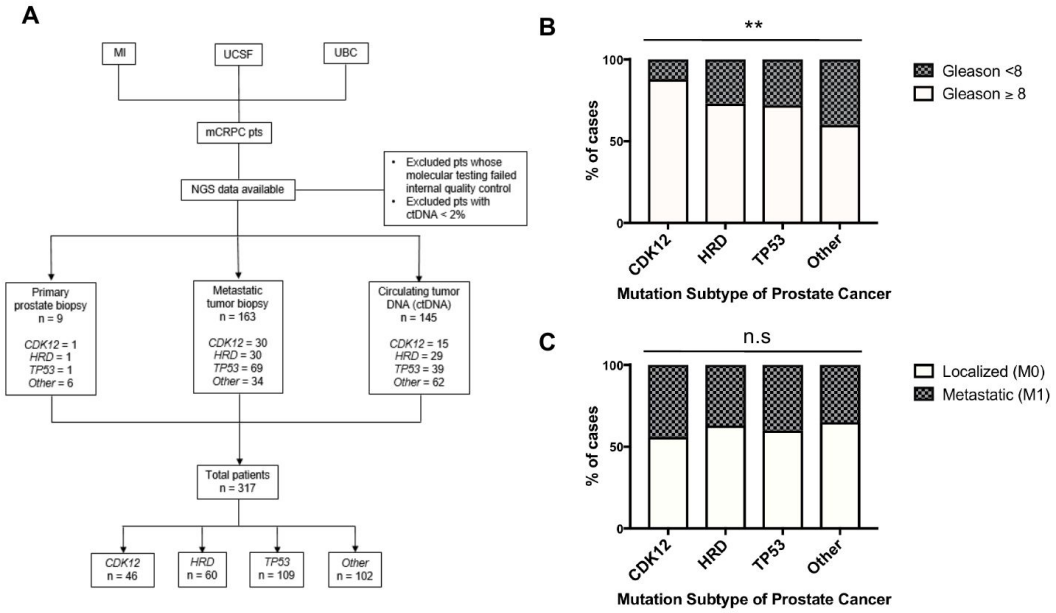


Figure 60: (A) CONSORT diagram depicting the patients included in this cohort based on NGS source and frequency of each mutation group. MI = University of Michigan, UCSF = University of California San Francisco, UBC = Vancouver Prostate Centre and BC Cancer. (B, C) Patients were stratified by genomic mutation types. Graph depicts the proportion of patients with

Gleason score ≥ 8 or < 8 at diagnosis (B) and the proportion of patients with localized (M0) or metastatic disease (M1) at diagnosis (C) for each subgroup. ** denotes p-value = 0.009 by Chi-square testing; n.s. denotes non-significant.

Therapy	CDK12		HRD		TP53		Other		p-value			
	n = 46 (%)	Median, months (Range)	n = 60 (%)	Median, months (Range)	n = 109 (%)	Median, months (Range)	n = 102 (%)	Median, months (Range)	Overall	CDK12 vs. HDR	CDK12 vs. TP53	CDK12 vs. Other
Abiraterone	33/46 (71.7%)	5.95 (0.3 - 11.64)	42/60 (70%)	6.17 (0 - 53.39)	68/109 (62.4%)	5.28 (0.92 - 26.58)	59/102 (57.8%)	8.06 (0.89 - 49.77)	0.201	0.361	0.834	0.047
Enzalutamide	29/46 (63%)	6.18 (1.38 - 31.02)	33/60 (55%)	4.28 (0.95 - 37.14)	56/109 (51.4%)	3.67 (0.46 - 40.2)	48/102 (47.1%)	6.7 (0.03 - 31.28)	0.133	0.821	0.228	0.412
Docetaxel	29/46 (63%)	3.49 (0.69 - 9.93)	30/60 (50%)	3.6 (1.38 - 16.58)	67/109 (61.5%)	3.45 (0.03 - 24.41)	35/102 (34.3%)	3.5 (0.03 - 21.05)	0.634	0.379	0.752	0.612
Carboplatin	8/46 (17.4%)	2.24 (1.15 - 4.11)	10/60 (16.7%)	3.44 (1.41 - 16.58)	15/109 (13.8%)	3.91 (1.55 - 9.67)	5/102 (4.90%)	3.85 (0.69 - 7.6)	0.335	0.286	0.057	0.608
Anti-PD-1/PDL-1	5/46 (10.9%)	2.07 (0.69 - 2.76)	2/60 (3.33%)	1.05 (0 - 2.11)	1/109 (1%)	0.69 (0.69 - 0.69)	3/102 (2.94%)	1.61 (0.69 - 2.11)	0.614	N/A	N/A	0.549
Olaparib	3/46 (6.5%)	4.14 (1.51 - 4.38)	11/60 (18.3%)	4.61 (0 - 15.82)	0	N/A	1/102 (1%)	4.11 (4.11 - 4.11)	0.873	0.586	N/A	N/A
Cabazitaxel	21/46 (45.6%)	2.76 (0.69 - 15.16)	10/60 (16.7%)	3.45 (0.69 - 7.57)	27/109 (24.8%)	1.45 (0.63 - 7.99)	12/102 (11.8%)	7.53 (0.2 - 25.07)	0.008	0.526	0.16	0.022
Abiraterone and Enzalutamide	42/46 (91.3%)	8.19 (1.32 - 32.01)	50/60 (83.3%)	8.59 (0.3 - 53.39)	57/109 (52.3%)	6.78 (1.12 - 40.2)	32/102 (31.4%)	10.08 (1.97 - 54.77)	0.473	0.363	0.986	0.192

Figure 61: Detailed examination of clinical parameters and drug treatments in our mCRPC cohorts.

We have also conducted a number of survival analyses to demonstrate whether CDK12 mutations are associated with specific clinical presentations and outcomes. These analyses considered CDK12 mutations in addition to other genetic aberrations such as TP53 mutations and homologous recombination deficiency (HRD) and evaluated whether CDK12 loss is an independent prognostic factor in prostate cancer progression.

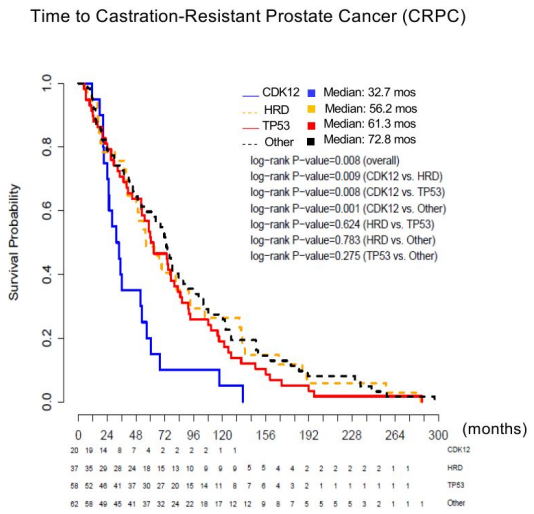


Figure 62: Kaplan-Meier Curves for (A) time to development of metastasis and (B) time to development of CRPC in patients presenting with localized disease at presentation, stratified by mutation type. For (A), the overall p-value = 0.0141, and the greatest difference was in the CDK12 compared with Other cohort (34.9 months versus 64.7 months, p-value = 0.0023). For (B), the overall p-value = 0.008, and the greatest difference was in the CDK12 compared with Other cohort (32.7 months versus 72.8 months, p-value = 0.001).

Patient Number	Institution	CDK12 Mutations Hit 1	Hit 2	Comments	Mutation Status
1	UCSF	W1043*	H113fs	truncation after kinase domain	Biallelic
2	UCSF	R221fs	T500fs		Biallelic
3	UCSF	L55fs	Y709fs		Biallelic
4	UCSF	S138fs	S174fs		Biallelic
5	UCSF	L907Q	I730del	kinase domain mutation	Biallelic
6	UCSF	S385fs	Q602*		Biallelic
7	UCSF	K144fs	R364fs		Biallelic
8	UCSF	P540fs	K852*		Biallelic
9	UCSF	splice site 2866+1G>A			Monoallelic
10	UCSF	R701fs	R858G	kinase domain mutation	Biallelic
11	UCSF	R784*	LOH by copy loss		Biallelic
12	UCSF	F573I			Monoallelic
13	UCSF	Q994L	LOH by copy loss	kinase domain mutation	Biallelic
14	UCSF	G736E	W920C	kinase domain mutation	Biallelic
15	MI	R188fs	P686fs		Biallelic
16	MI	R93fs	K186fs		Biallelic
17	MI	D52fs	S323*		Biallelic
18	MI	Y279fs	LOH in tumor		Biallelic
19	MI	deletion of exons 6-11			Monoallelic
20	MI	R902G	K757del	kinase domain mutation	Biallelic
21	MI	Homozygous deletion	Homozygous deletion		Biallelic
22	MI	T624fs	S625fs		Biallelic
23	MI	L1189fs	LOH by copy gain		Biallelic
24	MI	W1043*	P955R	truncation after kinase domain	Biallelic
25	MI	I704fs	copy neutral LOH		Biallelic
26	MI	G766fs	N854fs		Biallelic
27	MI	Detailed information not available			Missing data
28	MI	Detailed information not available			Missing data
29	MI	Detailed information not available			Missing data
30	MI	S318fs			Monoallelic
31	MI	G923V	copy neutral LOH		Biallelic
32	UBC	L852fs	I925S		Biallelic
33	UBC	D494fs	K1007fs		Biallelic
34	UBC	L122fs			Monoallelic
35	UBC	V425fs			Monoallelic
36	UBC	W1459fs	copy neutral LOH	truncation after kinase domain	Biallelic
37	UBC	L820*	copy neutral LOH		Biallelic
38	UBC	H109fs	LOH by copy loss		Biallelic
39	UBC	E129*	L1001R		Biallelic
40	UBC	G317fs			Monoallelic
41	UBC	K403*	copy neutral LOH		Biallelic
42	UBC	K264fs	copy neutral LOH	splice site mutation	Biallelic
43	UBC	P903R		kinase domain mutation	Monoallelic
44	UBC	Y305*	I935fs		Biallelic
45	UBC	D859A		kinase domain mutation	Monoallelic
46	UBC	Y259*	K445fs		Biallelic

Figure 63: CDK12 mutation breakdown with biallelic hits identified stratified by institution, in 34 of 43 patients (79%). 12 of 14 UCSF patients had documented biallelic loss. 12 of 17 MI patients had documented biallelic loss. Detailed mutation data were unavailable for 3 patients in the MI cohort. Note that although a second somatic alteration was not detected in these samples, there are several mechanisms of somatic alteration that can confound detection, particularly at lower levels of ctDNA. By our best estimate, at least 10 of 15 UBC patients had documented biallelic.

We also continue to look for novel associations between genotypes and immune phenotypes in a systematic fashion across the whole spectrum of mCRPC.

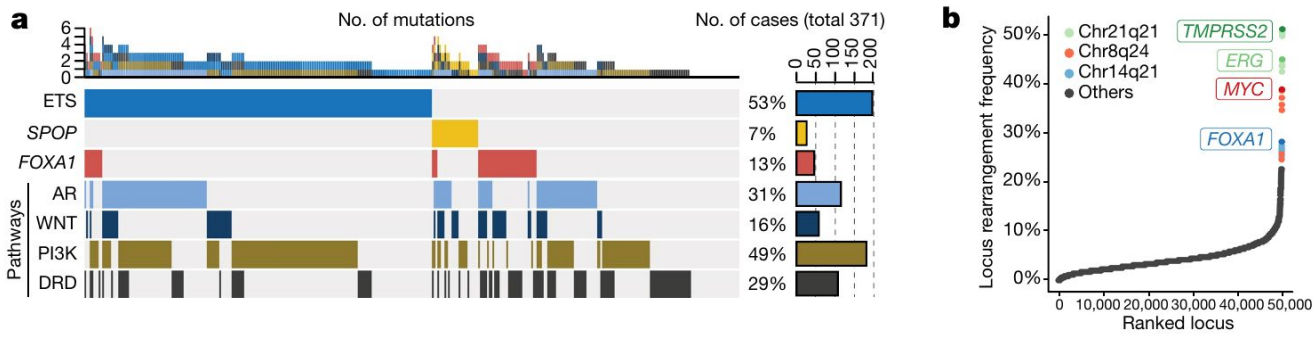


Figure 64: Characterization of the most recurrent genomic aberrations in mCRPC at the pathway level. We are currently evaluating the impact of FOXA1, AR, WNT, PI3K aberration on the immune landscape and immune phenotype of metastatic tumors.

Structural aberrations are highly recurrent in metastatic tumors (e.g. right panel Figure 64), but it is unknown whether these are associated with immune-related phenotypes. We are developing computational (see above CNVEX, HLARS) and statistical methods to associate the recurrence of structural variants and their impact on cell-intrinsic (pathway activity) and cell-extrinsic (e.g. immunity) mechanisms.

Notably, we found that mutations in mCRPC are not independent i.e. the presence of one mutations impacts the prevalence of other types of mutations. For example FOXA1 mutations are enriched for mutations in DNA-repair genes, but depleted of ETS and SPOP mutations. These types of associations have very significant implications in statistical analyses as one type of mutation tents to confound other finding. CDK12 mutations were an independent class largely independent of other mutations (with the exception of ETS), which allowed for relatively simple statistical analyses. For other types of aberrations (such as the ongoing FOXA1 mutations) these analyses will be more difficult, and require careful calibration, robust annotation of samples and possibly larger cohort sizes.

Overall, our immunogenomic studies gain statistical power as more cohorts are made available and are included in the analysis. This in turn creates demand for better computational tools (see above) which drives further discoveries thanks to the more careful genetic annotation of samples. Re-use of genomic data together with the development of innovative algorithms (such as MImmScore, CNVEX, CODAC, etc.) is therefore a viable strategy to make additional discoveries based on already existing data sets.

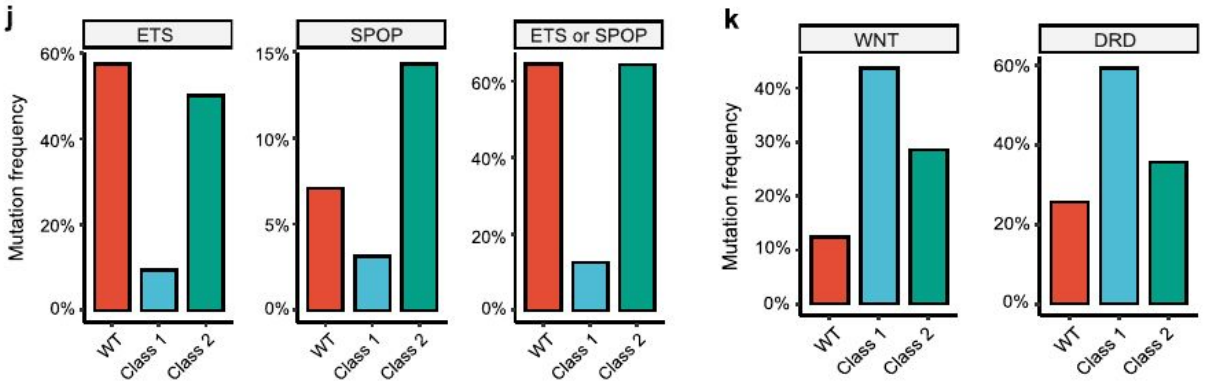


Figure 65: Genetic associations between FOXA1 mutations and other types of genetic aberration in mCRPC tumors (n=444).

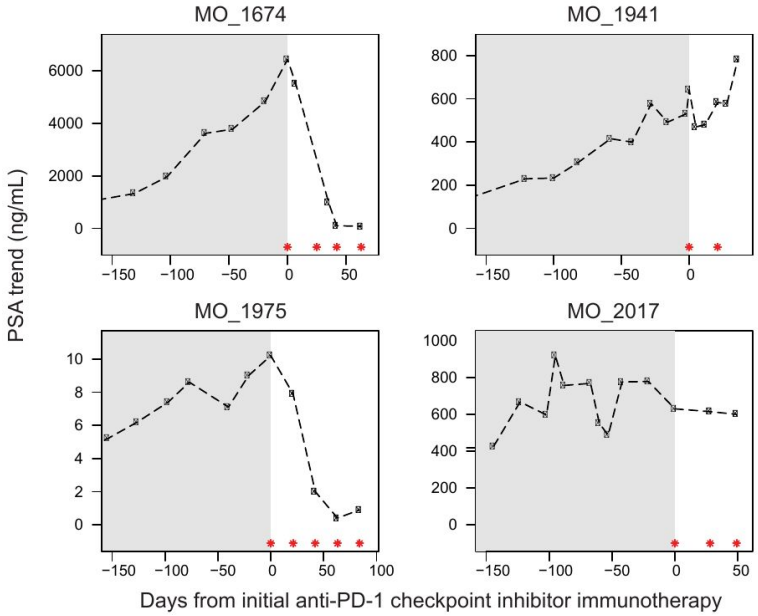


Figure 66: Significant responses to immunotherapy are observed for prostate cancer patients with high levels of immune infiltration

Conclusions:

1. We have shown significant progress in the realization of all major goals of the proposal spanning, computational developments, analyses of data, outcomes research and the identification of novel immune phenotypes in prostate cancer .Specifically, most of the validations outlined above have been achieved by analyzing the large numbers of samples proposed within this grant.
2. We have reached all but one (75% completion) milestones within the first two reporting periods, and have already made significant progress on the final outcomes-research based aspects of our study.

3. Our results have demonstrated the feasibility of immune phenotyping of metastatic tumors, and have shown that these findings have direct impact on patients (i.e. identification of CDK12 as a novel class of immunogenomic prostate cancer.
4. We have continued to validate our experimental and computational approaches, and expanded these assessments to other algorithms CNVEX, and HLARS.
5. We have developed bioinformatics pipelines and the underlying computational infrastructure, over the last year these infrastructures have been modernized to deal with the expected availability of WGS data for mCRPC.
6. We have already made good progress towards the milestones within the third reporting period.
7. Several manuscripts resulting from this work have been published or are currently under review.

Stated Goals Not Met

None so far. The development of Immunotyper is well underway and only insignificantly delayed.

What opportunities for training and professional development has the project provided?

This project was not intended to provide training, however I would like to note the the Michigan Center for Translational Pathology is dedicated to provide an exceptional training and learning environment. All of the researchers on staff benefit from the rich interactions and interdisciplinary nature of the Center. All our researchers routinely participate in professional development activities, write manuscripts, present at conferences, and engage in discussion during internal scientific meetings. Personally, I supervise all my reports and work with them one-on-one on the challenging aspects of their work, including this proposal. The University of Michigan provides organizes multiple workshops per year on topics ranging from developing analytical skills, using high-performance computing, to financial advice or even work-life balance. All staff routinely involved in research attend and present at scientific conferences, including AACR. In September 2018 I have started an independent research laboratory, where I train the next generation of bioinformaticians. Two of my graduate students work on aspects of this project. In particular Brodie Mumphrey is developing the HLARS algorithm. Other students working with me have received prestigious DoD and NIH awards and/or are on competitive training programs.

How were the results disseminated to communities of interest?

Our finding of CDK12 mutant prostate cancer has been extensively features at non-scientific venues.

What do you plan to do during the next reporting period to accomplish the goals?

So far the project is executed as planned. None of the milestones have been significantly delayed and all of our foundational experiments generated positive results and validations suggesting that they will be ready for the next phase of this proposal. All of our bioinformatics pipelines were implemented on-time, are highly scalable, and automated, hence we do not foresee any additional challenges. Therefore the original plan as presented in the research proposal, and further specified in the statement

of work remains valid. We propose no changes to the overall aims and are on-track to meet the milestones within the second reporting period. The last reporting period, as proposed originally, will focus on outcomes research, it will also involve a more in-depth characterization of copy-number variation and structural variation and their role in cancer immunity in prostate cancer. We will also carry out a study on the prognostic role of HLA mutations in metastatic PCa, as well as delineate the immunological subtypes of mCRPC as originally proposed using Immunotyper, which is currently expanded to included the covariates based on the genetic status of HLA.

IMPACT:

What was the impact on the development of the principal discipline(s) of the project?

The project while still in an early phase (before reaching the 3rd reporting phase where the majority of clinical findings are expected), however we already made an important contribution to our understanding of the immunogenomic foundations of prostate cancer through the delineation of CDK12 mutant prostate cancer. We have shown that integrative immunogenomic profiling is a viable and effective strategy to characterize a tumor immune phenotype, and a very efficient and cost-effective discovery platform. This will allow prostate cancer patients to be stratified based on their immune-status in an orthogonal way as compared to a stratification based on, for example, gleason grade. This in turn will result in a more granular understanding of prostate cancer as a disease. Such a refinement is beneficial, because it allows clinicians and researchers to better understand the heterogeneity of the disease, and better match patients with the right therapy. In the example provided above, albeit preliminary we show that based on immunological stratification we are potentially able to assign patients to the right treatment. Our data shows that CDK12 mutant tumors may respond to immune checkpoint therapy, but also do not respond to AR-inhibition. This has very important implication for the therapeutic choice in late-stage prostate cancer. Overall, the response of prostate cancer patients to immunotherapy is low at around 10-15%, however in our preliminary data we have observed responses around 50%, which is unprecedented in solid tumors outside melanoma and lung cancer. In the long term this study, and the approaches herein will inform the design of clinical trials, as evaluated and validated experimental and bioinformatics approaches may serve as effective biomarkers to guide the enrollment of patients into the right treatment baskets. This has the potential to improve response rates by giving the right drug to the right patient. Finally, we have recently identified mutations in FOXA1, and conclusively shown that these mutations are oncogenic.

What was the impact on other disciplines?

All the approaches presented herein can be generalized to other types of metastatic tumors, as such it affects other cancer-related disciplines beyond the prostate cancer research field. For example we applied several of the TCR-seq based methods (as explained above) to triple negative breast cancer with excellent results. In that type of cancer we were able to very quickly identify prognostic biomarkers. Since we were able to corroborate our result in breast cancer, it stands to reason that the proposed approaches may quantify and characterize fundamental principles of anti-tumor immunity and as such be much more broadly useful. Specifically, the bioinformatics pipelines, computational algorithms, and

statistical methods as developed and validated above can be readily extended to other types of tumors. In fact several of these methods are already being adapted to patients with other tumor types enrolled as part of the MI-OncoSeq program, which includes all indications. More generally, the computational methods presented herein will have an important impact on fields such as Bioinformatics, Computational Biology, and the emerging field of immunogenomics. To facilitate the rapid dissemination of our computational advances we provide access to most of our pipelines (after they are validated) as open source (see below). Algorithms developed for the characterization of CNVs and HLAs will also have broad utility outside of prostate cancer in particular they will be valuable in high-grade ovarian cancer (where CDK12 is also mutated), which is highly aneuploidy, and in head and neck cancer where mutations in HLA are particularly prevalent. The TPO pipeline will be broadly useful outside of prostate cancer research (it is cancer-type agnostic), and in fact technical solutions to workflow automation and computational genomics may be applicable outside the field of cancer.

What was the impact on technology transfer?

We are in the early stages of evaluating the patentability and technology transfer of several approaches developed. A methodological development, termed TAGTILE, has been submitted to the University of Michigan for evaluation and licensing to commercial entities. This technology revolves around transcriptomic profiling of clinical tissues.

What was the impact on society beyond science and technology?

Approaches developed as part of this project are being rapidly transferred to the MI-OncoSeq project. The MI-OncoSeq precision oncology projects goal is to match patients with the best possible treatments based on their tumors molecular profile. Hence research done as part of this proposal has a potential impact on all MI-OncoSeq patients in that it is carefully and diligently being used to gain insights of patient tumors and eventually may affect their treatment. Our goal is to make the approaches developed as part of this project available widely in the spirit of open-science, therefore all algorithms, technical solutions, and data are deposited and made widely available on the appropriate dissemination platforms.

CHANGES/PROBLEMS:

Changes in approach and reasons for change

Nothing to report.

Actual or anticipated problems or delays and actions or plans to resolve them

Nothing to report.

Changes that had a significant impact on expenditures

Nothing to report.

Significant changes in use or care of human subjects, vertebrate animals, biohazards, and/or select agents

Nothing to report.

PRODUCTS:

Journal Publications:

1. Published, Federal support acknowledged:

   Wu, Y.-M., Cieślik, M., Lonigro, R.J., Vats, P., Reimers, M.A., Cao, X., Ning, Y., Wang, L., Kunju, L.P., de Sarkar, N., Heath, E.I., Chou, J., Feng, F.Y., Nelson, P.S., de Bono, J.S., Zou, W., Montgomery, B., Alva, A., PCF/SU2C International Prostate Cancer Dream Team, Robinson, D.R., Chinnaiyan, A.M., 2018. Inactivation of CDK12 Delineates a Distinct Immunogenic Class of Advanced Prostate Cancer. Cell 173, 1770–1782.e14.

   Parolia, A., Cieslik, M., Chu, S.-C., Xiao, L., Ouchi, T., Zhang, Y., Wang, X., Vats, P., Cao, X., Pitchiaya, S., Su, F., Wang, R., Feng, F.Y., Wu, Y.-M., Lonigro, R.J., Robinson, D.R., Chinnaiyan, A.M., 2019. Distinct structural classes of activating FOXA1 alterations in advanced prostate cancer. Nature.

A number of additional publications based on this grant are currently in review in European Urology, Nature Medicine.

Software:
The following software packages implement most of the results described in the accomplishment sections. Independent packages will be released together with publications.

1. CRISP (see above):
   https://github.com/mcieslik-mctp/crisp-build
2. CODAC (see above):
   https://github.com/mcieslik-mctp/codac
3. TPO (see above):
   No public repository so-far (pre-release)
4. CNVEX (see above)
   https://github.com/mcieslik-mctp/codac

PARTICIPANTS & OTHER COLLABORATING ORGANIZATIONS

What individuals have worked on the project?

Name:	Marcin Cieslik
Project Role:	Principal Investigator
Researcher Identifier (e.g. ORCID ID):	NA
Nearest person month worked:	3
Contribution to Project:	Project leader, design and supervise analyses
Funding Support:	NIH, Prostate Cancer Foundation

Name:	Yuping Zhang
Project Role:	Research Staff
Researcher Identifier (e.g. ORCID ID):	NA
Nearest person month worked:	2
Contribution to Project:	Bioinformatician, implementation of most bioinformatics pipelines
Funding Support:	NA

Name:	Shasha Li
Project Role:	Postdoctoral Fellow
Researcher Identifier (e.g. ORCID ID):	NA
Nearest person month worked:	2
Contribution to Project:	Bioinformatician
Funding Support:	NA

Name:	Ying Liu

Project Role:	Sequencing Technician
Researcher Identifier (e.g. ORCID ID):	NA
Nearest person month worked:	3
Contribution to Project:	Library preparation, Next Generation Sequencing
Funding Support:	NA

Has there been a change in the active other support of the PD/PI(s) or senior/key personnel since the last reporting period?

Nothing to Report.

What other organizations were involved as partners?

Nothing to Report.

SPECIAL REPORTING REQUIREMENTS

COLLABORATIVE AWARDS: Not applicable
QUAD CHARTS: Not applicable

APPENDICES:

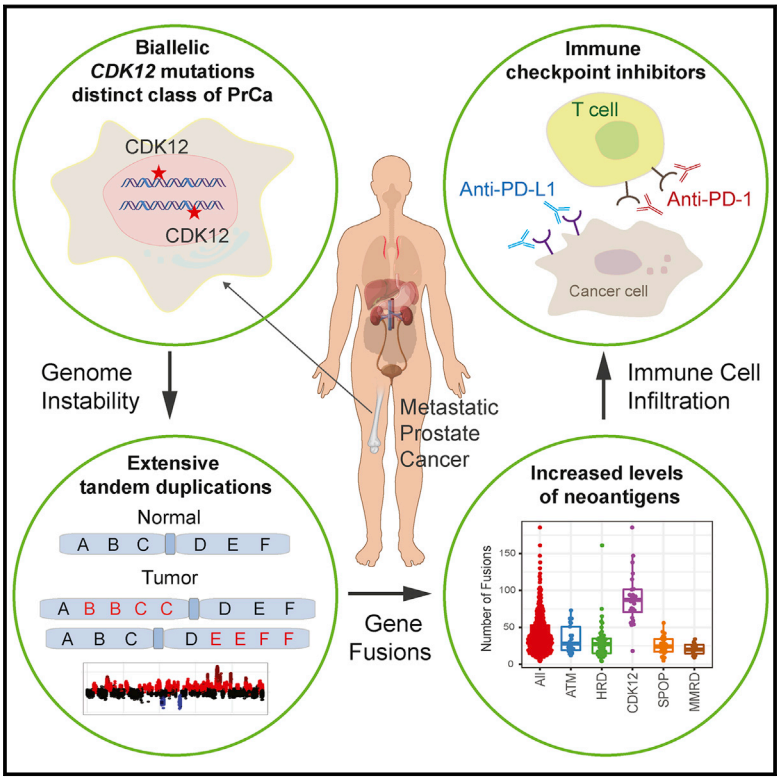
- Appendix #1: Cell Manuscript on CDK12 mutation in mCRPC
- Appendix #2: Nature Manuscript on FOXA1 mutations in mCRCP

Article

Cell

Inactivation of *CDK12* Delineates a Distinct Immunogenic Class of Advanced Prostate Cancer

Graphical Abstract



Highlights

- *CDK12* biallelic inactivating mutations define a distinct subtype of prostate cancer
- *CDK12* loss is associated with genomic instability and focal tandem duplications
- *CDK12* loss leads to increased gene fusions, neoantigen burden, and T cell infiltration
- Patients with *CDK12* mutant tumors may benefit from immune checkpoint inhibition

Authors

Yi-Mi Wu, Marcin Cieślik, Robert J. Lonigro, ..., PCF/SU2C International Prostate Cancer Dream Team, Dan R. Robinson, Arul M. Chinnaiyan

Correspondence

danrobi@umich.edu (D.R.R.), arul@umich.edu (A.M.C.)

In Brief

Loss of both alleles of the *CDK12* gene defines a molecular subtype of metastatic castration-resistant prostate cancer that is potentially targetable with immune checkpoint inhibitors.

Wu et al., 2018, Cell *173*, 1770–1782
June 14, 2018
https://doi.org/10.1016/j.cell.2018.04.034

CellPress

Cell

Article

Inactivation of *CDK12* Delineates a Distinct Immunogenic Class of Advanced Prostate Cancer

Yi-Mi Wu,[1,2,20] Marcin Cieślik,[1,2,20] Robert J. Lonigro,[1] Pankaj Vats,[1] Melissa A. Reimers,[3] Xuhong Cao,[1] Yu Ning,[1] Lisha Wang,[1] Lakshmi P. Kunju,[1,2,4] Navonil de Sarkar,[5] Elisabeth I. Heath,[6,7] Jonathan Chou,[8] Felix Y. Feng,[8,9,10,11] Peter S. Nelson,[5,12,13] Johann S. de Bono,[14,15] Weiping Zou,[1,2,16] Bruce Montgomery,[12,17] Ajjai Alva,[1,3] PCF/SU2C International Prostate Cancer Dream Team, Dan R. Robinson,[1,2,*] and Arul M. Chinnaiyan[1,2,4,18,19,21,*]

[1]Michigan Center for Translational Pathology, University of Michigan, Ann Arbor, MI 48109, USA
[2]Department of Pathology, University of Michigan, Ann Arbor, MI 48109, USA
[3]Department of Internal Medicine, University of Michigan, Ann Arbor, MI 48109, USA
[4]Rogel Cancer Center, University of Michigan, Ann Arbor, MI 48109, USA
[5]Division of Human Biology, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA
[6]Department of Oncology, Wayne State University School of Medicine, Detroit, MI 48201, USA
[7]Molecular Therapeutics Program, Barbara Ann Karmanos Cancer Institute, Detroit, MI 48201, USA
[8]Department of Medicine, University of California at San Francisco, San Francisco, CA 94143, USA
[9]Department of Radiation Oncology, University of California at San Francisco, San Francisco, CA 94143, USA
[10]Department of Urology, University of California at San Francisco, San Francisco, CA 94143, USA
[11]Helen Diller Family Comprehensive Cancer Center, University of California at San Francisco, San Francisco, CA 94143, USA
[12]Department of Medicine, University of Washington, Seattle, WA 98109, USA
[13]Division of Clinical Research, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA
[14]Cancer Biomarkers Team, Division of Clinical Studies, The Institute of Cancer Research, London SM2 5NG, UK
[15]Prostate Cancer Targeted Therapy Group and Drug Development Unit, The Royal Marsden NHS Foundation Trust, London SM2 5NG, UK
[16]Department of Surgery, University of Michigan, Ann Arbor, MI 48109, USA
[17]Veterans Affairs Puget Sound Health Care System, University of Washington, Seattle, WA 98109, USA
[18]Department of Urology, University of Michigan, Ann Arbor, MI 48109, USA
[19]Howard Hughes Medical Institute, University of Michigan, Ann Arbor, MI 48109, USA
[20]These authors contributed equally
[21]Lead Contact
*Correspondence: danrobi@umich.edu (D.R.R.), arul@umich.edu (A.M.C.)
https://doi.org/10.1016/j.cell.2018.04.034

SUMMARY

Using integrative genomic analysis of 360 metastatic castration-resistant prostate cancer (mCRPC) samples, we identified a novel subtype of prostate cancer typified by biallelic loss of *CDK12* that is mutually exclusive with tumors driven by DNA repair deficiency, *ETS* fusions, and *SPOP* mutations. *CDK12* loss is enriched in mCRPC relative to clinically localized disease and characterized by focal tandem duplications (FTDs) that lead to increased gene fusions and marked differential gene expression. FTDs associated with *CDK12* loss result in highly recurrent gains at loci of genes involved in the cell cycle and DNA replication. *CDK12* mutant cases are baseline diploid and do not exhibit DNA mutational signatures linked to defects in homologous recombination. *CDK12* mutant cases are associated with elevated neoantigen burden ensuing from fusion-induced chimeric open reading frames and increased tumor T cell infiltration/clonal expansion. *CDK12* inactivation thereby defines a distinct class of mCRPC that may benefit from immune checkpoint immunotherapy.

INTRODUCTION

Comprehensive genomic analyses have substantially furthered our understanding of primary prostate cancer (PCa) and metastatic castration-resistant prostate cancer (mCRPC) (Barbieri et al., 2012; Beltran et al., 2016; Fraser et al., 2017; Grasso et al., 2012; Robinson et al., 2015; Cancer Genome Atlas Research Network, 2015). These studies have discovered common genetic drivers of prostate cancer, such as fusions of ETS genes (Tomlins et al., 2005), amplification of *AR*, and loss of *CDKN2A*, *PTEN*, *RB1*, and *TP53* (Robinson et al., 2015). Integrative genomic studies have further delineated distinct molecular subtypes in primary and metastatic prostate cancer and specific molecular pathways that contribute to prostate cancer onset and progression, including *AR*, *WNT*, *SPOP*, and *PI3K/AKT/MTOR* signaling (Barbieri et al., 2012; Beltran et al., 2016; Robinson et al., 2017; Cancer Genome Atlas Research Network, 2015).

This knowledge is being actively translated into promising drug targets. Recently, recurrent germline and somatic mutations in genes involved in DNA repair provided a rationale for the use of poly ADP ribose polymerase (PARP) and immune checkpoint inhibitors in homologous recombination-deficient (HRD) and mismatch repair-deficient (MMRD) metastatic prostate cancer, respectively (Le et al., 2015; Mateo et al., 2015; Robinson et al., 2015). Intriguingly, in both cases, the genomic

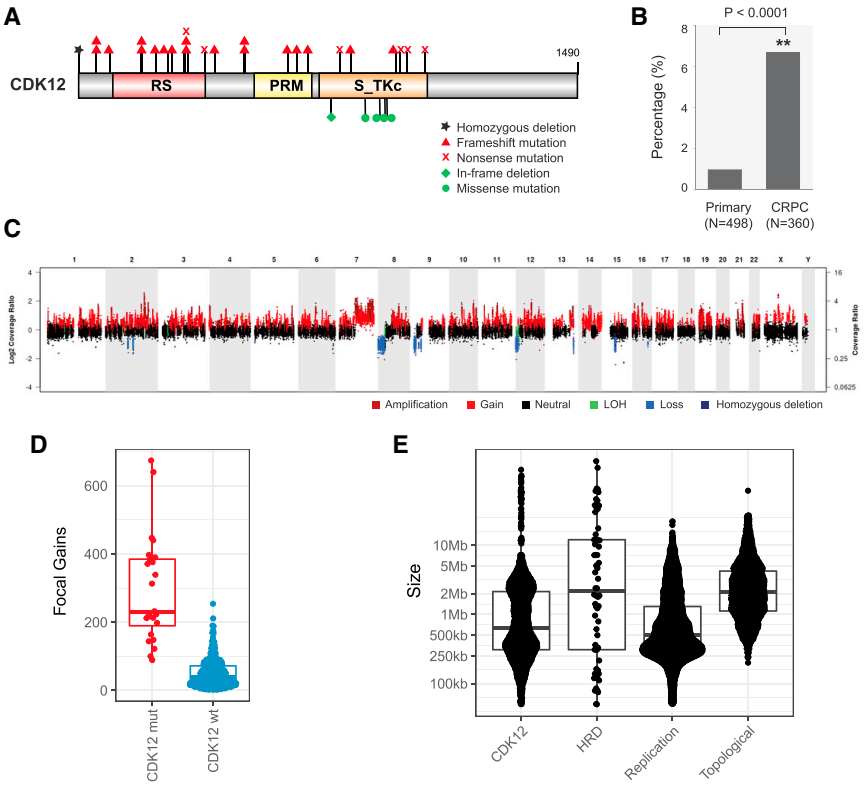


Figure 1. Biallelic Loss of *CDK12* Is Enriched in mCRPC and Results in Focal Tandem Duplications

(A) Schematic of mutations in CDK12.

(B) Increased frequency of *CDK12* loss in metastatic castration-resistant prostate cancer (CRPC) compared to primary disease.

(C) Characteristic pattern of genomic instability found in all cases with *CDK12* loss. Copy gains are indicated in shades of red. LOH, loss of heterozygosity.

(D) Number of focal copy gains (<8 Mb) by *CDK12* mutational status, as determined by whole-exome analysis.

(E) Size of copy gains (tandem duplications), as ascertained by whole-genome sequencing of index cases with *CDK12* mutations (*CDK12*) and homologous recombination deficiency (HRD). Sizes of replication domains and topological domains in normal tissues are shown for comparison. See also Figures S1–S3 and Tables S1, S2, S3, S4, and S5.

instability engendered by the deficiency becomes a "double-edged sword." On one hand, it is the mechanism by which the tumor generates secondary oncogenic drivers, while on the other, it makes the tumor susceptible to a specific therapy. For example, cancer cells with MMRD have a high mutation burden that generates tumor neoantigens, thereby making the patients favorable candidates for intervention with immunotherapies (Le et al., 2015).

CDK12 is a cyclin-dependent kinase that associates with its activating partner, cyclin K, to form a heterodimeric complex that regulates several critical cellular processes (Blazek et al., 2011; Cheng et al., 2012). CDK12 consists of different functional domains: a centrally located kinase domain, several RS (arginine/serine) motifs near the N terminus, and a proline-rich motif (PRM) that can function as a binding site for additional proteins (Ko et al., 2001). CDK12 directly regulates transcription by phosphorylating serine residues of the hepta-peptide repeats (YSPTSPS) within the C-terminal domain of RNA polymerase II essential for transcriptional elongation (Bartkowiak et al., 2010; Blazek et al., 2011; Cheng et al., 2012). Multiple studies have also suggested a role for *CDK12* in controlling genomic stability through regulation of genes involved in the DNA damage response (*ATR*, *BRCA1*, *FANCD2*, *FANCI*, etc.) (Blazek et al., 2011; Juan et al., 2016). Depletion or loss of function of *CDK12* have further been observed to sensitize ovarian cancer cells to PARP inhibitors through defects in HR (Bajrami et al., 2014; Ekumi et al., 2015; Joshi et al., 2014).

In previous studies, we found recurrent *CDK12* mutations in metastatic prostate cancer (Robinson et al., 2015), while similar observations were later made in serous ovarian tumors (Popova et al., 2016). Herein, we delineate a novel genetically unstable subtype of mCRPC associated with biallelic inactivation of *CDK12*. We show that *CDK12* mutants are genetically, transcriptionally, and phenotypically distinct from HRD and MMRD tumors. Further, we identify that *CDK12* mutant tumors have synthetic genetic dependencies and a characteristic immunophenotype, which provide candidate targets for precision therapy.

RESULTS

CDK12 Mutations Are Enriched in Cases of mCRPC

We previously reported that 4.7% of mCRPC patients harbored biallelic aberrations of *CDK12*. To confirm this observation, we have compiled an extended multi-site metastatic prostate cancer cohort of 360 patients (CRPC360), comprising SU2C (Robinson et al., 2015), MI-OncoSeq (Robinson et al., 2017), and UMich rapid autopsy cases (Grasso et al., 2012) (Table S1), a majority of which have matched whole-exome and transcriptome data (Table S2). The combined datasets were reanalyzed using the MI-Oncoseq workflow (Robinson et al., 2017), producing harmonized call sets of somatic, germline, and structural variants. We also analyzed, using the MI-Oncoseq workflow, sequence data from 498 cases of primary prostate cancer in the TCGA (The Cancer Genome Atlas) dataset. We detected aberrations of *CDK12* in 25/360 of mCRPC patients (6.9%), 95% confidence interval (CI) [4.6%, 10.2%] (Figure 1A). This is significantly higher than in primary PCa, 6/498 patients (1.2%) (Figure 1B; Table S3) ($p < 0.0001$ Fisher's exact test). Examination of data across additional primary and metastatic prostate cancer datasets revealed a similar difference in the frequency of biallelic *CDK12* mutations between primary and metastatic cancer (Table S4) (Abida et al., 2017; Beltran et al., 2016; Fraser et al., 2017;

Kumar et al., 2016). CRPC genomes are more highly mutated than those of localized tumors; however, the magnitude of the increased mutation rate is not sufficient to explain the increased frequency of biallelic loss of *CDK12*. The majority of *CDK12* mutations (83%) were truncating and resulted in the loss of the kinase domain. Missense mutations were clustered around conserved residues in the kinase domain (Figure S1). All patients showed biallelic inactivation of *CDK12*. *CDK12* has been shown to have a very low tolerability for germline loss-of-function variants (Juan et al., 2016), and, consistently, no germline aberrations were detected in our cohort (Table S5).

CDK12 Mutant Tumors Are Baseline Diploid with an Excess of Focal Tandem Duplications

A significant increase in genomic instability is a hallmark of metastatic tumors (Negrini et al., 2010). While primary prostate cancers are largely diploid, metastatic tumors often show extensive LOH, aneuploidies, and a significant increase in mutational burden (Robinson et al., 2017). We examined the landscape of *CDK12*-mutated mCRPC cases and observed a distinctive genomic landscape (Figures 1C and S2), similar to that identified in a subset of ovarian cancers (Popova et al., 2016). The prototypical *CDK12* mutant tumor was baseline diploid and had few arm-level copy-number aberrations except gain of 8q, but notably, hundreds of focal copy-number gains were dispersed across the genome. While focal gains were present on all chromosomes within a sample, other focal events, such as high-level amplifications or deletions, were rare or absent. *CDK12* biallelic inactivation was strongly associated with this form of genomic instability ($p < 0.00001$, Fisher's exact test). All cases with *CDK12* inactivation, and only cases with *CDK12* mutation, exhibited this form of genome instability in both the metastatic and primary cohorts (Figure S2; Table S3). No other genes were positively associated with this genome instability. ETS fusions and *PTEN* mutations were depleted in cases with *CDK12* mutations ($p < 0.00001$ for both, Fisher's exact test). None of the *CDK12*-mutated tumors exhibited a neuroendocrine phenotype.

The genomic phenotype of *CDK12* mutant tumors was compared to other cases in the CRPC360 cohort, particularly those associated with frequent primary genetic drivers (PGDs) of prostate cancer: *ATM* mutations, HRD, *SPOP* mutations, and MMRD. Like *CDK12* mutant cases, *SPOP*- and MMRD-driven tumors were mostly diploid, while a large subset of *ATM*- and HRD-driven tumors showed large-scale aneuploidy (Figure S3A). The high number of focal gains was consistently observed in *CDK12* mutant cases compared to those in the cohort with wild-type *CDK12* (Figure 1D). Detection of genomic structural variants (SV) from whole-genome sequencing (WGS) data confirmed that the gains were focal tandem duplications (FTDs) (Figure S3B) and enriched in gene-dense regions (Figure S3C). Strikingly, comparison of *CDK12* mutant and HRD index cases revealed a bimodal distribution of FTD sizes in *CDK12* mutant, but not HRD, tumors (Figure S3D). The modes of this distribution were consistent with the sizes of replication domains (RD), but not topological domains (TD) (Figure 1E). Specifically, the ~2.4 Mb peak was close to the mode of the early/late RDs, while the ~0.4 Mb peak matched the size of transitional RDs (Hiratani et al., 2008) (Figure S3E). Breakpoint sequence assembly revealed that FTDs were enclosed by error-prone junctions indicative of a non-homologous end joining (NHEJ)-mediated repair process (Figure S3F). We refer to these events as *CDK12*-associated FTDs (*CDK12*-FTDs) to distinguish them from *BRCA*-dependent events and focal amplifications.

CDK12 Mutants Represent a Specific Class of Prostate Cancer with a Distinct Transcriptional Phenotype

We next tested for genetic associations between *CDK12* loss and the most frequent PGDs of prostate cancer to determine whether *CDK12* mutant cases were a unique class of mCRPC. Strikingly, *CDK12* aberrations were mutually exclusive with all of the PGDs analyzed (ETS fusions, *SPOP* mutations, HRD, *ATM* mutations, and MMRD) (Figure 2A).

Several of the established prostate cancer PGDs have been associated with characteristic gene expression profiles (Herschkowitz et al., 2008; Parikh et al., 2014; Saal et al., 2007). We hypothesized that *CDK12* loss may similarly constrain a specific transcriptional phenotype. To test this, we compared the expression profiles of mCRPC cases with aberrations in specific PGDs or *CDK12* to a reference set of cases ($n = 92$) that were wild-type for all the PGDs, including *CDK12* (PGD-WT). Interestingly, we found that *CDK12* aberrations were associated with the highest number of differentially expressed genes (DEGs) (Figure S4A), independent of differences in the number of cases for each PGD, and across a wide range of effect-size (Figure 2B), and p value cutoffs. The most up- (e.g., *AIFM2*, *ARID3C*, *TBX4*) or downregulated (e.g., *TSACC*, *CDNF*, *ABCC12*) genes have not been previously studied in the context of prostate cancer (Figure S4B). To establish a causal link between this transcriptional phenotype and loss of *CDK12*, we performed a small interfering RNA (siRNA)-mediated knockdown experiment in LNCaP cells. Depletion of *CDK12* at the RNA and protein levels resulted in growth arrest (Figures S4C–S4E) and profound transcriptional changes. In addition, DEGs associated with *CDK12* mutations in patients were almost perfectly recapitulated *in vitro* (Figure 2C), which allowed us to define a transcriptional signature of *CDK12*-loss in mCRPC (Table S6).

While most *CDK12* mutants retained active androgen receptor (AR) signaling (Figure S4F) (Beltran et al., 2016), their expression signature was distinct from the equivalent signatures for the other PGDs (Figures 2D and S4G). Gene set enrichment analysis (GSEA) (Subramanian et al., 2005) across the MSigDB (Liberzon et al., 2015) revealed significantly perturbed curated gene sets (Figure S4H). The most prominently altered were those related to oxidative phosphorylation (down), inflammatory response (up), hormone receptor signaling (down), and epithelial dedifferentiation (down). To understand this further, we delineated a core set of 28 genes downregulated in both metaplastic and stem-like breast cancer (i.e., two of the most significant gene sets). Strikingly, the majority of those genes were significantly downregulated in *CDK12* mutant mCRPC (Figure S4I). Although the shift from oxidative to glycolytic metabolism (Warburg effect) is a hallmark of many cancer types (Vander Heiden et al., 2009), it is not a characteristic of most prostate cancers (Cutruzzolà et al., 2017).

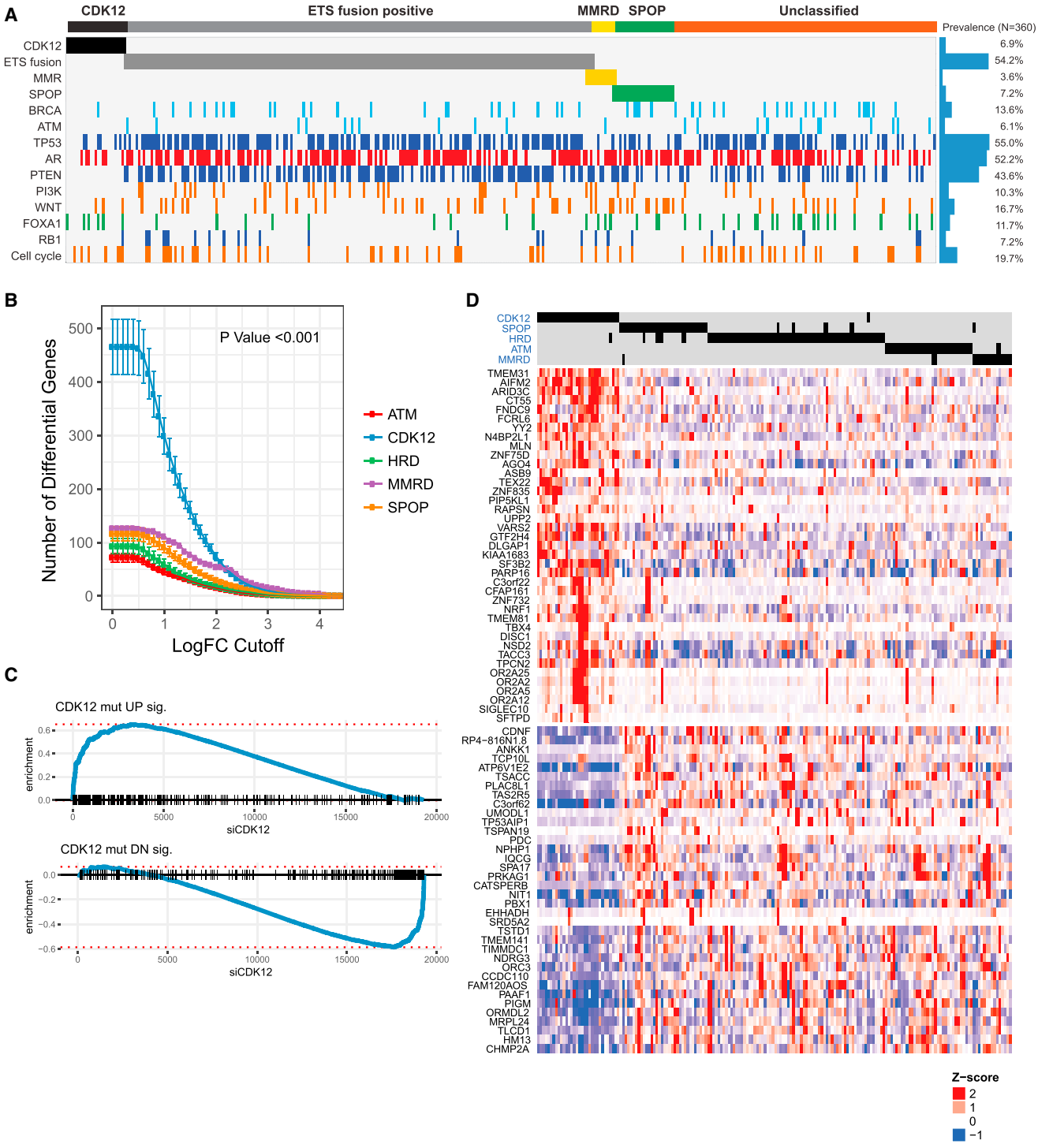


Figure 2. *CDK12* Mutant Prostate Cancer Is a Novel Molecular Subtype of mCRPC

(A) Mutual exclusivity of *CDK12* loss, ETS fusions, mismatch repair deficiency (MMRD), *SPOP* mutations, and homologous recombination deficiency (HRD).

(B) Number of significantly differentially expressed genes (DEGs) for the prostate tumors with different primary genetic drivers.

(C) Enrichment plot for signatures of up- (top) and downregulated (bottom) genes in *CDK12* mutant tumors. Genes are ranked by their fold change following si*CDK12* knockdown in LNCaP cells, with *CDK12*-loss signature genes indicated as black dashes. The increased relative frequency (enrichment score) of genes at either end of this spectrum is shown as a blue line.

(D) Heatmap of the top DEGs in *CDK12*-mutant prostate cancer. Differential expression for all samples (columns) in this heatmap is relative to tumors that are wild-type for primary genetic drivers of prostate cancer (as in B).

See also Figure S4 and Tables S1 and S6.

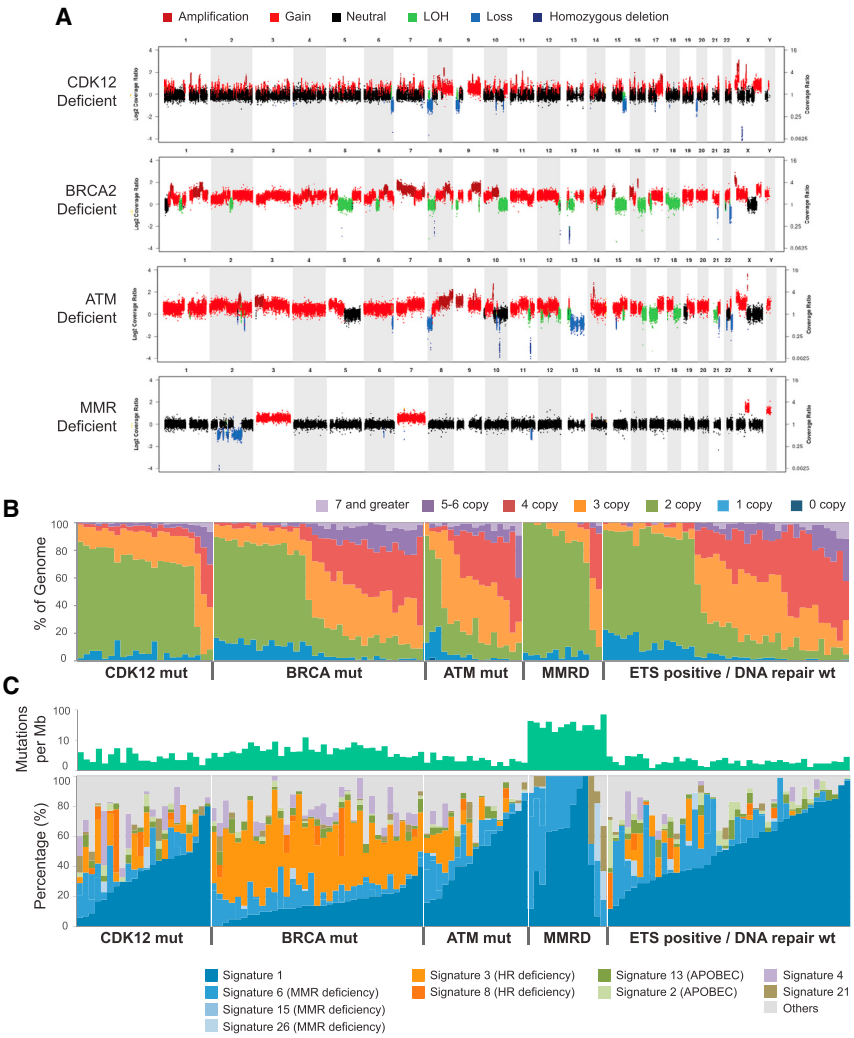


Figure 3. *CDK12* Loss Results in a Distinct Pattern of Genomic Instability

(A) Representative copy-number plots for prostate tumors with deficiencies in key DNA damage response or repair pathways.

(B) Spectrum of copy-number aberrations in tumors with distinct genetic drivers.

(C) Spectrum of inferred mutational signatures in tumors with distinct genetic drivers.

See also Figure S4.

CDK12 Mutant Tumors Display Characteristic Copy-Number and Mutational Signatures Distinct from DNA Repair-Deficient Prostate Cancer

Previous studies suggested that *CDK12* is involved in controlling genomic stability through regulation of HR or other DNA damage response effectors (Blazek et al., 2011; Ekumi et al., 2015; Joshi et al., 2014; Juan et al., 2016). Our CRPC360 transcriptional data also showed a unique signature for *CDK12* mutant tumors (Figure 2D). Large-scale copy-number gains were evident in the *BRCA2*- and *ATM*-deficient cases, as compared to *CDK12* mutant or MMRD cases (Figure 3A). To quantitate and contrast the *CDK12* mutant pattern with the other PGDs on a larger scale, we tallied absolute copy-number levels from whole-exome sequencing (WES) data across the entire CRPC360 cohort (Figure 3B). These analyses showed that *BRCA* and *ATM* mutated, as well as ETS fusion-positive, tumors had the highest percentage of copy-number gains, while the majority of *CDK12* mutant and MMRD tumors did not exhibit changes in ploidy (Figure 3B).

Genomic signatures are a powerful approach to study the mutagenic imprints of environmental and genetic factors. To determine if loss of *CDK12* activity is associated with a distinct signature, we computed mutational burden as well as mutational signature across various genetic drivers (Figure 3C). As expected, MMRD cases had the highest mutational burden and a signature consistent with microsatellite instability (signature 6) (Alexandrov et al., 2013). HRD tumors had the next highest mutational burden, and *BRCA*-loss was associated with an evident signature 3 (Polak et al., 2017). The remaining PGDs, including *CDK12*, had a baseline level of SNVs and were dominated by age-related 5-methylcytosine deamination (signature 1). Combined, these data support that the *CDK12* mutant subtype is distinct from either the HRD or MMRD type of prostate cancer. In particular, *CDK12*-mutants are different from tumors with HRD, which was previously presumed to be the pathway through which *CDK12* regulated genomic stability. Notably, the expression of *BRCA1* or *BRCA2* was not affected by *CDK12* mutational status (Figure S4J) and neither was the expression of other genes encoding long transcripts and cognate proteins (Figure S4K), a class previously suggested to be regulated by *CDK12* (Blazek et al., 2011).

CDK12-FTDs Result in Highly Recurrent Gains of Genes Involved in the Cell Cycle and DNA Replication

The large number of FTDs present in all *CDK12* mutant tumors introduces the possibility of detecting synthetic genetic dependencies or epistasis. One approach is to look for loci with recurrent *CDK12*-FTDs at the cohort level. To identify such genomic regions, we developed a Monte Carlo null model to simulate the expected distribution of FTD recurrences, given their number and size. We applied both stringent (2 Mb, "narrow") and relaxed (8 Mb, "wide") definitions (Figures S5A–S5B). Using both models, we detected a total of 27 loci with recurrent focal gains at false-discovery rates of 3.5% and 5%, respectively (Figures 4A and S5C). Indicative of strong

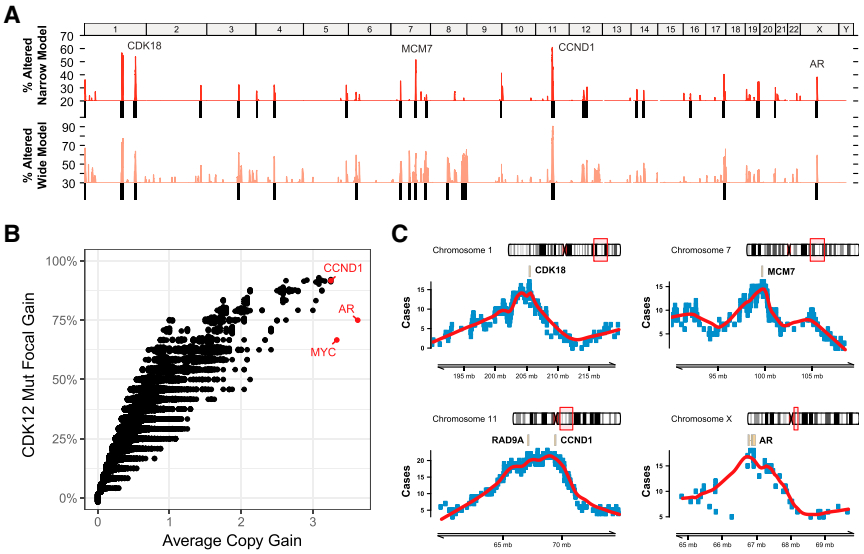


Figure 4. Recurrence of Focal Tandem Duplications Associated with *CDK12* Loss

(A) Genome-wide frequency (percentage of *CDK12* mutant patients) of focal tandem duplications (FTDs) based on a narrow (<2 Mb) and wide (<8 Mb) definition of focality.

(B) FTD recurrence and average copy-number gain of FTDs at the individual gene level. Genes with the highest average copy-number are highlighted in red.

(C) Delineation of minimal common regions (MCR) for loci with the most recurrent gains specific to *CDK12*-loss tumors. Genes related to the cell cycle are highlighted in each MCR. The *AR* locus is presented as a positive control.

See also Figures S5 and S6.

positive selection, several of these loci showed copy-number gains in almost all *CDK12* mutant cases (Figure S5C). Strikingly, their recurrence was significantly lower in *CDK12* wild-type tumors, which suggests a synthetic dependency (Figure S5D). As a prominent exception, the *MYC* and *AR* loci (Figure S5E) were recurrently amplified, regardless of *CDK12* status, which underscores their fundamental role in prostate cancer. Although most of the *CDK12*-FTDs result in the gain of one additional copy (Figure 3B), we observed that the most recurrent genes also had the highest copy-number gains (*MYC*, *AR*, *CCND1*), suggestive of gene dosage selective pressure (Figure 4B).

The delineation of minimal common regions (MCR) is an established strategy to identify genetic targets that are subject to positive selection and, hence, responsible for the recurrent copy-number aberrations (Mermel et al., 2011). In order to nominate such candidate genes in *CDK12* mutant mCRPC, we summarized FTDs into MCRs at the most recurrent loci (Figure 4C). The *AR* locus, whose MCR was centered on the *AR* gene as expected, represents a positive control for this approach. Of the recurrent loci, two were consolidated into a narrow MCR harboring a single candidate gene (*MCM7* and *CDK18*), while one required further prioritization. The chr11_q13.2 locus is characterized by high gene density and the presence of *RAD9A* and *CCND1*, all of which could contribute to the FTD recurrence of this region. *CCND1* was also associated with the highest copy-number gains, comparable in magnitude with amplifications at the *MYC* and *AR* loci (Figure 4B). Strikingly, candidate genes under positive selection, *MCM7*, *RAD9A*, *CDK18*, and *CCND1*, have crucial roles in DNA replication and genome stability. Amplifications of *MYC* and *AR* are among the most recurrent genetic events in mCRPC and not specific to *CDK12* mutants. Correspondingly, their molecular functions are pleiotropic; both regulate the cell cycle (Bretones et al., 2015; Yuan et al., 2006) and contribute independently to proliferation of prostate cancer cells (Bernard et al., 2003).

CDK12-FTDs Induce Expression in a Dosage-Dependent and Independent Manner

In order to better understand some of the functional consequences of *CDK12*-FTDs, we interrogated both global and gene-specific associations between copy-number and expression levels. To assess global effects of *CDK12*-FTDs, we probed changes in average expression levels associated with the focal increases in copy-number (Figures S5F and S5G). We observed a significant increase in the number of DEGs at each absolute copy-number level (Figure S5F). To demonstrate the feasibility of identifying gene-specific effects given our sample size, we interrogated the expression of three genes associated with the highest average copy-number gains and high recurrence: *CCND1*, *MYC*, and *AR* (Figure 4B). A significant dose-dependent relationship for *CCND1* and *AR*, but not *MYC*, was observed (Figure S6A). We expanded this analysis to other cancer-related genes and identified similar trends for key oncogenes in the MAPK, AKT, and MTOR pathways (Figure S6B). Strikingly, dosage dependence was much less robust for receptor tyrosine kinases (RTK), which were dominated by singleton expression outliers (Figure S6C). A global analysis was performed to determine the contribution of *CDK12*-FTDs to the prevalence of expression outliers. Overall, outliers were more frequent in *CDK12*-FTDs, and their frequency increased with copy-number gains (0.5% to 4%) (Figure S5G).

Mutant *CDK12* Prostate Cancers Exhibit a Unique Structural Signature Characterized by Increased Gene Fusions

Transcriptome sequencing data were used to delineate signatures of structural genomic instability across the different classes of PGDs. Interestingly, as shown in Figure 5A, *CDK12* mutant tumors had the highest fusion burden, consistent with the large number of focal copy-number events (Figure 1D) and their enrichment in gene-rich regions (Figure S3C). The prototypical *CDK12* mutant case exhibited a large number of fusions (Figure 5A) generated by tandem duplications and relatively fewer by translocations, inversions, or deletions (Figure 5B). This contrasts with HRD and MMRD tumors, which have a significantly lower fusion burden dominated by translocations. Next, we

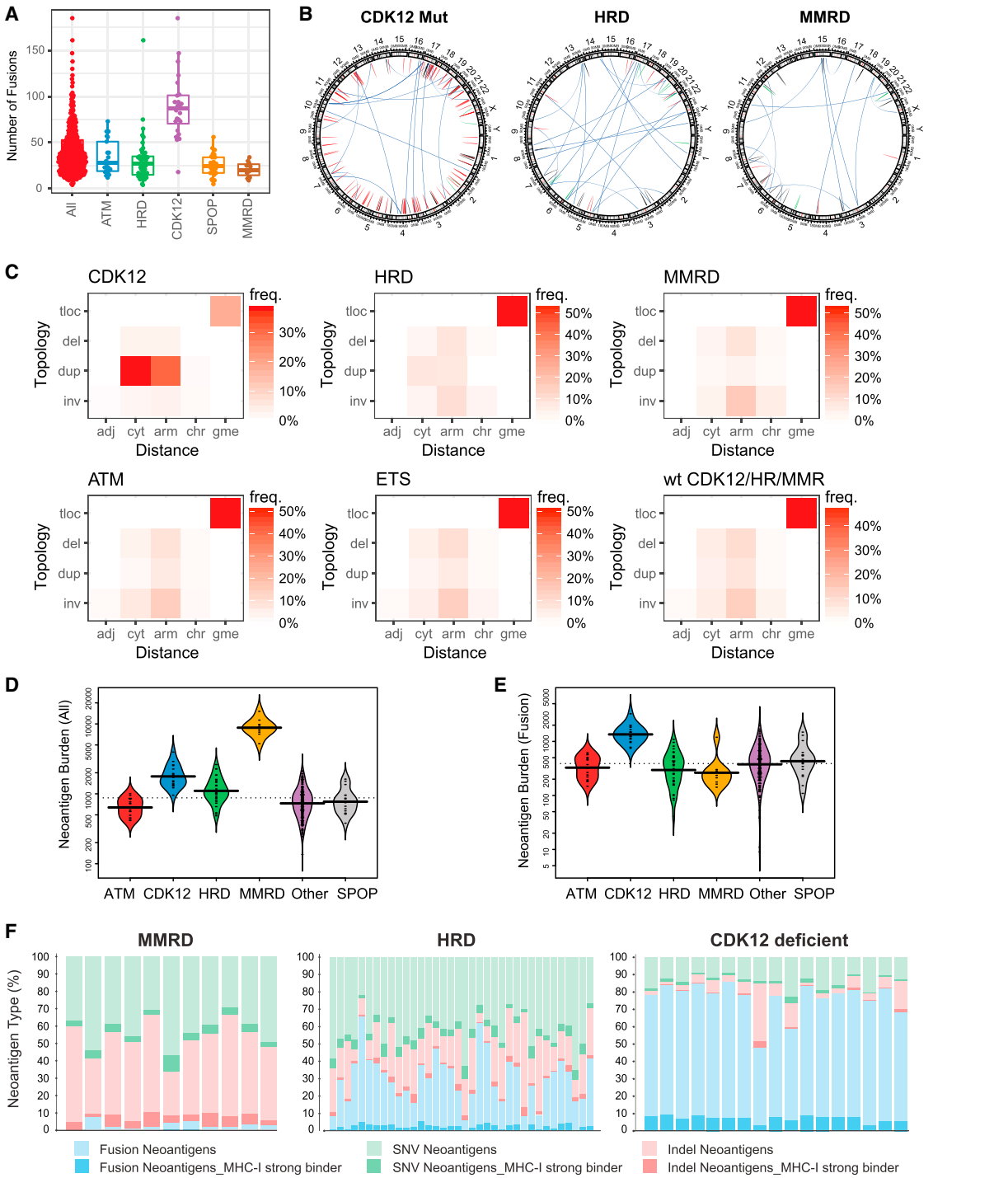


Figure 5. Signatures of Structural Variation and Neoantigen Presentation in *CDK12* Mutant Tumors

(A) Total number of detected gene fusions for prostate tumors with different genetic drivers.

(B) Representative examples of Circos plots showing the pattern structural variation in tumors with major types of genomic instability. Structural variants (SVs) detected from RNA-seq are classified into translocations, deletions, duplications, and inversions based on the topology of the breakpoints. Color code: blue, translocations; red, duplication; green, inversion; black, deletion.

(C) Classification of SVs based on the topology and distance between the breakpoints. adj, breakpoints in adjacent loci; cyt, in same cytoband; arm, on same chromosome arm; gme, genomic translocation; inv, inversion; dup, duplication; del, deletion; tloc, translocation. Heatmap color indicates frequency of a SV class across all index cases. (numbers of patients: CDK12 = 24, HRD = 47, MMRD = 11, ATM = 21, ETS = 190, WT = 31).

(legend continued on next page)

devised "fusion-grams" to quantitatively compare signatures of structural variants between the varying prostate cancer classifications (Figure 5C). In a fusion-gram, structural variants are classified according to the observed distance and topology of their breakpoints (i.e., deletion, duplication, inversion, translocation). For *CDK12* mutant tumors, the majority of fusions (70%) were classified as duplications within a cytoband or chromosome arm. All other PGDs had signatures dominated by translocations (~49%) and fewer overall duplications (11%) than deletions (18%) or inversions (22%), further supporting the uniqueness of *CDK12* mutant PCa.

Because *CDK12*-FTDs are associated with expression outliers (Figure S5G) and a large number of gene fusions (Figure 5A), we hypothesized that some of those events are potential secondary genetic cancer drivers. We searched for candidate driver events where a chromosomal aberration resulted in either outlier expression of an oncogene or formation of a likely oncogenic gene fusion. In addition to the singleton outlier RTKs (Figure S6C), we found two cases of *BRAF* fusions (*KIAA1549-BRAF* and *HIPK2-BRAF*) generated as a result of a *CDK12*-FTD (Figures S6D and S6E). While we, and others, have previously reported *BRAF* fusions in prostate cancer (Palanisamy et al., 2010), duplications involving the *KIAA1549-HIPK2-BRAF* locus have thus far been noted as hallmarks of pilocytic astrocytoma (Yu et al., 2009). Surprisingly, we also found a promoter hijacking event leading to outlier expression of *ETV1* (Figure S6F). However, not all secondary events could be inferred as direct consequences of *CDK12*-FTDs. For example, we found a translocation leading to extremely high expression of full-length *FGFR2* (Figure S6G). Importantly, *FGFR* fusions can be found in many solid tumors and are compelling targets for precision therapy (Wu et al., 2013).

CDK12 Mutant Tumors Are Characterized by Increased Gene Fusion-Induced Neoantigen Open Reading Frames

Tumor immunogenicity is associated with mutational burden and neoantigen load (Le et al., 2015). We reasoned that gene fusions and their chimeric protein products yield significant numbers of neoantigens in *CDK12* mutant tumors. We carried out comprehensive prediction of novel peptides from mutation and fusion calls (STAR Methods) and found that MMRD, HRD, and *CDK12* mutant tumors had a significantly higher neoantigen burden compared to other mCRPC molecular subtypes (Figure 5D). Strikingly, the mutational mechanism by which the neoantigens were generated was specific to each subtype. While neoantigens in MMRD and HRD tumors were formed by indels and SNVs, fusions contributed most of the novel epitopes in *CDK12* mutant mCRPC (Figures 5E and 5F). The calculated neoantigen burden from fusions was the highest in *CDK12* mutant tumors among the other PGDs (Figure 5E). Importantly, these analyses also identified neoantigens with strong MHC class-I binding affinities that are predicted to be candidate epitopes for immunotherapy (Figure 5F).

CDK12 Mutant Tumors Show High Immune Infiltration and Imprints of Immune Evasion

We found significant activation of the cancer inflammatory hallmark gene-set in *CDK12* mutant tumors and LNCaP cells transfected with siRNA to *CDK12* (Figure 6A). Compared to wild-type tumors (PGD-WT, see above), *CDK12* mutant cases showed increased expression of chemokines and their receptors (Figure S7A). Overall, we observed reduced or low expression of chemokines that can recruit regulatory T cells (*CCL17*, *CCL20*, *CCL22*) (Curiel et al., 2004; Zou, 2006) and an increase in chemokines that support dendritic cell migration into the tumor microenvironment (*CCL21*, *CCL25*). Interestingly, certain direct pro-tumor chemokines, including *CCL18* and *CXCL8* (Nagarsheth et al., 2017), were enriched in patients with *CDK12* mutations. To determine whether this immune phenotype was specific to *CDK12* mutant tumors, we contrasted the activation of the top signatures across genetically unstable mCRPC subtypes. Strikingly, only MMRD and *CDK12* mutant tumors showed robust activation of chemokine signaling/inflammatory response and high immune infiltration as estimated by the cohort MImmScore (Robinson et al., 2017) (Figures 6B and S7B). Taken together, these data indicate that *CDK12* mutant tumors are immunogenic and infiltrated by leukocytes but evolve chemokine-mediated mechanisms of immune evasion.

Antigen recognition by T cells leads to their clonal expansion. To detect whether increased neoantigen burden was mirrored by an increase in T cell clonality (McGranahan et al., 2016), we performed T cell repertoire analysis using TCRb sequencing on a set of 60 tumors across all molecular subtypes (n = 10 per group). We found that, compared to genomically stable tumors, *CDK12* mutant tumors showed higher overall levels of T cell infiltration (Figure 6C) and larger numbers of expanded T cell clones (Figure 6D), regardless of the template cutoff used (Figure S7C). To confirm these trends, we performed T cell repertoire profiling of RNA sequencing (RNA-seq) data (Bolotin et al., 2015). First, we established that RNA and DNA-based estimates of T cell infiltration were in agreement (Figure S7D). We found that relative to wild-type cases (PGD-WT), MMRD, HRD, and *CDK12* mutant tumors all had a significant increase in both the number of detected T cell clones (Figure S7E) and the total number of CDR3 sequences (Figure S7F). Importantly, immunohistochemical (IHC) staining of CD3 on representative index cases further confirmed the presence of tumor-infiltrating T cells in a subset of *CDK12* mutant tumors (Figure 6E).

Pilot Clinical Study to Determine *CDK12* Mutant Prostate Cancer Response to Checkpoint Inhibitor Immunotherapy

Of eleven *CDK12* mutant patients identified in the MI-Oncoseq program, a total of five late stage, pre-treated mCRPC patients had some exposure to immunotherapy in the form of the immune checkpoint inhibitor anti-PD1. One patient received one dose of

(D and E) Antigen burden in tumors with distinct types of genetic instability. Overall burden based on single nucleotide variants, insertions/deletions, and fusions is shown in (D). Fusion-specific burden is shown in (E).

(F) Distribution of neoantigens based on genetic variant type and predicted MHC class-I (MHC-I) binding affinity.

See also Figure S6.

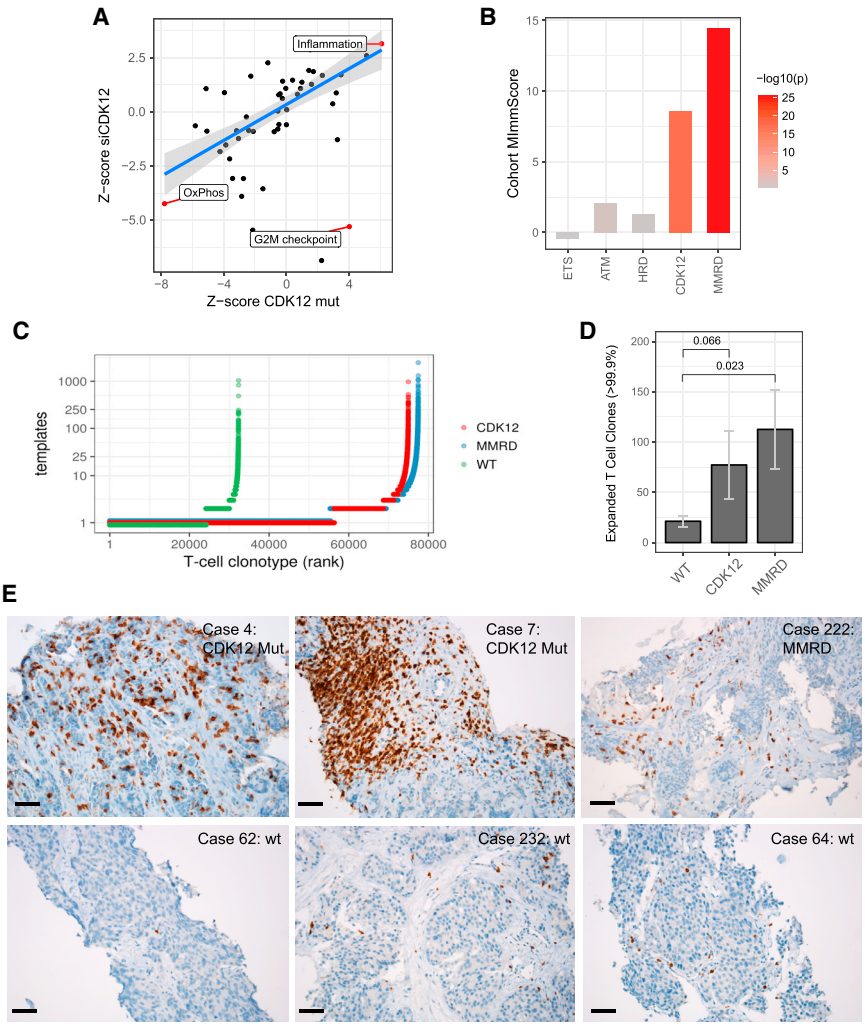


Figure 6. Immunogenomic Properties of *CDK12* Mutant Tumors

(A) Differential expression of MSigDB cancer hallmark gene-sets in *CDK12* mutant patients and in LNCaP cells depleted with *CDK12* by siRNA. Highlighted hallmarks are significant (false discovery rate [FDR] <0.05, limma moderated t test).

(B) Levels of global immune infiltration across prostate tumors with distinct genetic drivers compared to genetically stable (PGD wild-type) tumors. The "Cohort MImmScore" is defined as the gene-set enrichment *Z* score and p value based on random-set test and moderated cohort DE log2 fold-changes.

(C) Overview of T cell clonotypes across *CDK12* mutant (n = 10), MMRD (n = 10), and WT (n = 10) tumors. T cell clonotypes (i.e., identical CDR3 sequences) are ranked by their frequency (number of templates). *CDK12* mutant and MMRD tumors show, overall, an increase in the total number of T cells (x axis), and higher levels of clonal expansion (y axis).

(D) Comparison of clonal expansion between immunogenic (MMRD, *CDK12*) and wild-type mCRPC tumors (t test). Expanded clones are defined as those with the highest number of clonal expansion (estimated number of templates >99.9 percentile across all cohorts; n > 12).

(E) Immunohistochemistry (IHC) performed on formalin-fixed paraffin-embedded tumor sections using anti-CD3 antibody. Six representative cases are shown, including two *CDK12* mutant tumors, one MMRD tumor, and three tumors which are wild-type for *CDK12*, MMR genes, and HR genes. Black bar, 50 μm.

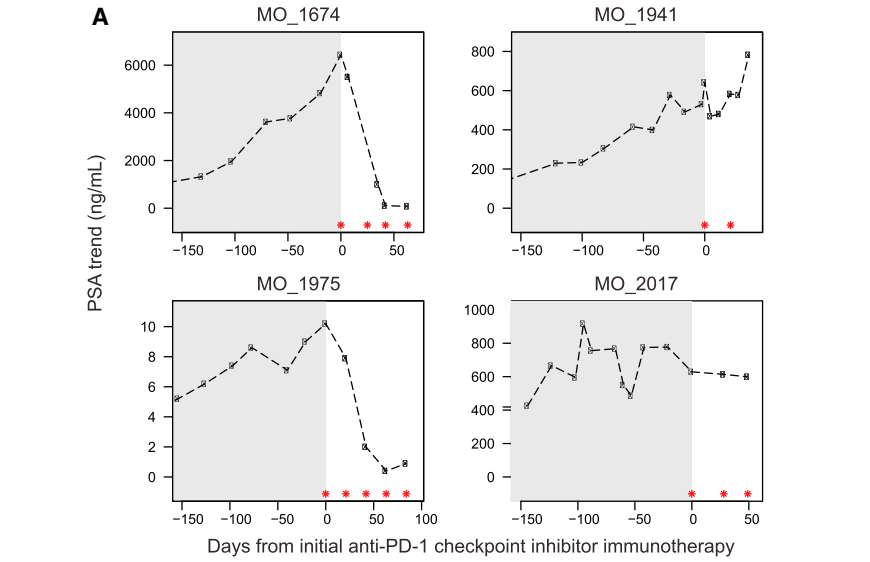
See also Figure S7.

anti-PD1 as part of combination therapy on a clinical trial and was excluded, as he did not receive anti-PD1 monotherapy and could, therefore, not be compared to the other treated patients. Detailed prostate-specific antigen (PSA) response data are presented on the four patients treated with anti-PD1 monotherapy for whom we also have associated clinical data and detailed sequencing information (Figure 7A). Strikingly, two of the four patients had an exceptional response in terms of PSA decline. This was surprising as checkpoint inhibitor immunotherapy has typically not been efficacious in prostate cancer, with the exception of patients with mismatch repair defects (Le et al., 2015).

One patient (MO_1674) was treated with anti-PD1 checkpoint inhibitor immunotherapy and displayed a marked PSA response after four doses of therapy, but eventually succumbed to multisystem organ failure, possibly due to anti-PD1 induced pneumonitis (Nishino et al., 2015). Patient MO_1941 received only two doses of anti-PD1 with a subsequently rising PSA and is deceased. Two patients are still alive on active therapy (MO_2017, MO_1975). Patient MO_2017 had heavily pre-treated disease, with prior disease progression on abiraterone, enzalutamide, docetaxel, and cabazitaxel. Pre-treatment PSA prior to initiation of immunotherapy was 628.8 ng/mL with a modest improvement in PSA after three doses of anti-PD-1 and subsequent PSA decline to 599.2 ng/mL. Patient MO_1975 had a Gleason 9 metastatic prostatic adenocarcinoma and prior lymph node progression on abiraterone and enzalutamide. Evaluation of a metastatic lymph node biopsy demonstrated robust CD3 staining by IHC (Figure 7B). To date, the patient has received five doses of anti-PD1 with a significant PSA decrement (Figure 7A), as well as marked decline in pelvic lymph node disease burden (Figure 7C). These early clinical results support the hypothesis that metastatic prostate cancer patients who harbor biallelic *CDK12* loss may have a higher likelihood of response to immunotherapy than an unselected metastatic prostate cancer population. Further study in the context of a clinical trial is warranted.

DISCUSSION

In this report, we comprehensively characterized biallelic loss of *CDK12* as a novel PGD of prostate cancer. Importantly, through an integrative genomic approach, we demonstrate that *CDK12* mutations are mutually exclusive with other PGDs, such as *SPOP* mutations and ETS fusions. *CDK12* mutant tumors present unique characteristics at the genetic, transcriptomic, and



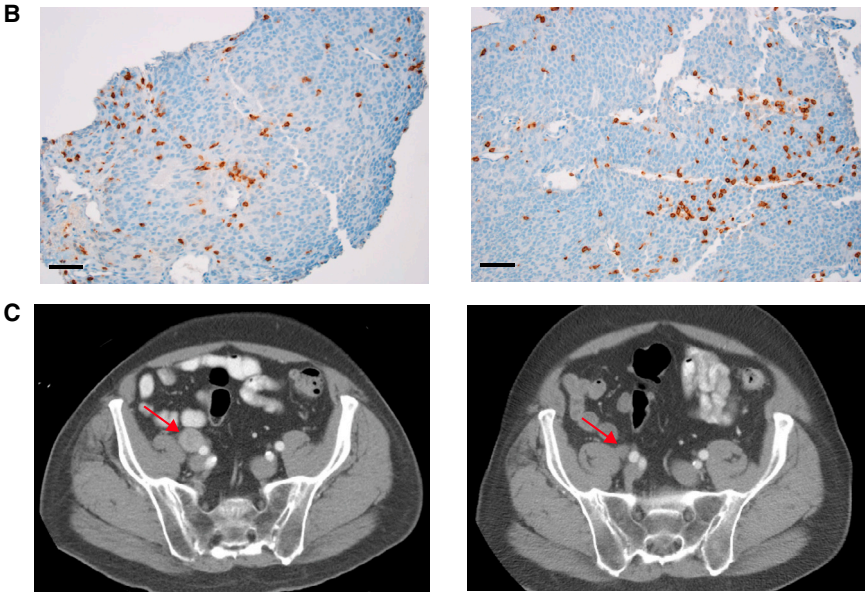


Figure 7. Response of *CDK12* Mutant Patients to Anti-PD1 Checkpoint Inhibitor Immunotherapy

(A) PSA levels of four *CDK12* mutant prostate cancer patients treated with anti-PD1 monotherapy. Gray shading represents PSA levels prior to anti-PD-1 therapy. Asterisks indicate anti-PD1 doses of 200 mg intravenously (i.v.).

(B) Representative CD3 IHC images of metastatic lymph node biopsies of patient MO_1975 prior to anti-PD1 treatment. Cells exhibited membranous and cytoplasmic staining of CD3, highlighting the presence of T lymphocytes. Black bar, 50 μm.

(C) Computed tomography (CT) imaging of patient MO_1975 pre- and post-immunotherapy treatment. Arrows indicate metastatic lymph node.

immunophenotypic levels and have the potential to be therapeutically targeted.

At the genetic level, *CDK12* mutant tumors show a characteristic pattern of genomic instability. Previous findings, primarily from cell-based assays, suggested that *CDK12* impacts genome stability through defects in HR (Bajrami et al., 2014; Blazek et al., 2011; Ekumi et al., 2015; Joshi et al., 2014; Juan et al., 2016). However, our data, and the observations made previously in *CDK12* mutant ovarian tumors (Popova et al., 2016), are inconsistent with that model. In contrast to HRD mCRPC tumors, which are characterized by translocations and aneuploidies, *CDK12* mutant tumors are diploid with a large number of focal tandem duplications (*CDK12*-FTD) and few translocations. *CDK12* mutant cases also lack mutational signatures of HRD (Figure 3C) and maintain the expression levels of *BRCA1* and *BRCA2* (Figure S4J). Overall, the genomic phenotypes of HR and *CDK12* deficiency are clearly distinct.

Several lines of evidence indicate that *CDK12*-FTDs are a result of aberrant DNA re-replication during S-phase: (1) *CDK12*-FTDs have a characteristic bimodal size distribution that matches the length of replication domains but not topologically associated domains or HR defects; (2) *CDK12* mutant cases have a synthetic dependency on aberrations in genes involved in DNA replication: *MCM7*, *RAD9A*, *CCND1*, and *CDK18*; (3) *CDK12*-FTDs result most frequently in the gain of one additional copy, consistent with the firing of an additional origin of replication; and (4) knockdown of *CDK12* results in growth arrest (Figure S4E). It remains unknown whether FTDs are generated through a one-time catastrophic event, a slow ongoing mutational process, or rescue of the phenotype by one of the recurrent gains.

At the transcriptional level, *CDK12* mutant tumors are associated with over 300 DEGs, which makes them the most prominent molecular subtype in our analysis (Figures 2B and S4A). Perhaps most importantly for translational purposes, *CDK12* mutant cases exhibit a characteristic immunophenotype. *CDK12* mutant tumors show high overall immune infiltration (Figure 6B), increased levels of tumor-infiltrating lymphocytes (Figures 6C–6E and S7E), and altered chemokine signaling.

This immunological phenotype may be influenced by the elevated neoantigen burden in *CDK12* mutant tumors. While single-nucleotide variants (SNVs) and indels are the main source of neoantigens in MMRD and HRD tumors, neoantigens in *CDK12* mutant tumors are mostly from FTD-induced fusions. Although the detection of neoantigens from fusions is still at an early

stage, the present study is, to our knowledge, the first to demonstrate the analytical value of neoantigen prediction from RNA-seq data. Fusions are analogous to indels in that they can generate neoantigens through in-frame and frameshift mechanisms. The latter, often referred to as neo-ORFs (Hacohen et al., 2013), are particularly interesting because they generate completely novel epitopes that are potentially highly immunogenic. In line with this possibility, high levels of CCL21 and CCL25 may mediate dendritic cell tumor trafficking and neoantigen-specific T cell clonal expansion (Chan et al., 1999; Gosling et al., 2000; Vicari et al., 1997).

A large number of studies have established the complex and important roles of immunity in the development and progression of prostate cancer (Strasner and Karin, 2015). These findings gave rise to a number of clinical trials for several classes of immunotherapeutics, which have been met with mixed results. For example, a phase 3 trial comparing ipilimumab (anti-CTLA4 immune checkpoint inhibitor) with placebo failed in patients with mCRPC (Kwon et al., 2014). A plausible explanation is in the genetics of prostate cancer. Compared to other tumors, prostate cancer has a low mutation rate, few neoantigens, and, consequently, is less visible to the adaptive immune system. In spite of that, exceptional responses to anti-CTLA4 (Cabel et al., 2017) and anti-PD1 (Graff et al., 2016) treatment have been observed clinically. These findings clearly show that strategies are needed to identify those patients that will benefit from immunotherapy. Taken together, our data suggest that *CDK12* mutant prostate cancer is intrinsically immunogenic (Sharma et al., 2017), and *CDK12* mutations may identify a subset of patients where immunotherapy would be efficacious. Indeed, we observed an exceptional response (PSA decline) with anti-PD1 monotherapy in two out of four mCRPC patients in this study (Figure 7A). Furthermore, identification of *CDK12* mutation-associated neoantigens may help in the design of personalized tumor vaccines. The immune phenotype of *CDK12*-mutated tumors may also broadly suggest a combinational strategy for prostate cancer treatment involving inhibition of *CDK12* and immune checkpoint blockade.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Cell lines
  - Human subjects and patient inclusion
- METHOD DETAILS
  - Kinase domain alignment
  - siRNA-mediated knockdown of CDK12
  - Immunostaining of T lymphocytes
  - Integrative clinical sequencing
  - T cell receptor β repertoire deep sequencing
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Whole-genome sequencing data analysis
  - Exome data analysis
  - Assignment of pathway status
  - RNA-seq data analysis
  - Differential expression analysis
  - Pathway and gene set enrichment analyses
  - Mutation signature analysis
  - CDK12 mutation frequency analysis
  - CDK12-FTD recurrence analysis
  - Copy-number expression aggregation
  - Structural variant and fusion-gram analysis
  - HLA-typing analysis
  - Integrative *in silico* neoantigen translation
  - IEDB peptide binding prediction
  - T cell repertoire analysis from RNA-seq data
  - Statistical Analysis
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures and six tables and can be found with this article online at https://doi.org/10.1016/j.cell.2018.04.034.

ACKNOWLEDGMENTS

We gratefully acknowledge all patients who participated. We thank Stephanie Ellison, Ph.D., for assistance in preparing this manuscript and Fuzon Chung for *CDK12* knockdown in cell lines. We also acknowledge the efforts of the MI-Oncoseq team. This work was supported by the Prostate Cancer Foundation (PCF), StandUp2 Cancer (SU2C)-Prostate Cancer Foundation Prostate Dream Team (SU2C-AACR-DT0712), Department of Defense (DOD) (W81XWH-15-1-0562), Early Detection Research Network (U01 CA214170), and Prostate SPORE (P50 CA186786, P50 CA097186). M.C. is supported by a PCF Young Investigator Grant and a DOD Prostate Cancer Research Program Idea Development Award (PC160429). A.M.C. is a Howard Hughes Medical Institute Investigator, Taubman Scholar, and American Cancer Society Professor.

AUTHOR CONTRIBUTIONS

Y.-M.W., M.C., R.J.L., P.V., and D.R.R. prepared figures and tables. Y.-M.W., X.C., Y.N., and D.R.R. carried out sequencing and coordinated patient cohorts. M.C. coordinated transcriptomic, fusion, and immunogenomic analyses. R.J.L., Y.-M.W., and D.R.R. carried out whole-exome analyses for ploidy and copy-number. Neoantigen analyses were performed by P.V. and M.C. CD3 IHC staining was performed by L.P.K. and L.W. Performance of patients on anti-PD1 immunotherapy was evaluated by M.A.R. and A.A. M.A.R., F.Y.F., J.C., B.M., and N.d.S. compiled clinical data. Patients were enrolled by E.I.H., P.S.N., J.S.d.B, B.M., A.A., and the International SU2C/PCF Prostate Cancer Dream Team. W.Z. provided advice and analysis of the immunology components of this study. M.C., Y.-M.W., R.J.L., P.V., D.R.R., and A.M.C. wrote the paper. D.R.R. and A.M.C. supervised the study.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 23, 2017
Revised: February 23, 2018
Accepted: April 24, 2018
Published: June 14, 2018

REFERENCES

Abida, W., Armenia, J., Gopalan, A., Brennan, R., Walsh, M., Barron, D., Danila, D., Rathkopf, D., Morris, M., Slovin, S., et al. (2017). Prospective genomic profiling of prostate cancer across disease states reveals germline and

somatic alterations that may affect clinical decision making. JCO Precis. Oncol. Published online May 31, 2017. https://doi.org/10.1200/PO.17.00029.

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.L., et al.; Australian Pancreatic Cancer Genome Initiative; ICGC Breast Cancer Consortium; ICGC MMML-Seq Consortium; ICGC PedBrain (2013). Signatures of mutational processes in human cancer. Nature *500*, 415–421.

Bajrami, I., Frankum, J.R., Konde, A., Miller, R.E., Rehman, F.L., Brough, R., Campbell, J., Sims, D., Rafiq, R., Hooper, S., et al. (2014). Genome-wide profiling of genetic synthetic lethality identifies CDK12 as a novel determinant of PARP1/2 inhibitor sensitivity. Cancer Res. *74*, 287–297.

Barbieri, C.E., Baca, S.C., Lawrence, M.S., Demichelis, F., Blattner, M., Theurillat, J.-P., White, T.A., Stojanov, P., Van Allen, E., Stransky, N., et al. (2012). Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat. Genet. *44*, 685–689.

Bartkowiak, B., Liu, P., Phatnani, H.P., Fuda, N.J., Cooper, J.J., Price, D.H., Adelman, K., Lis, J.T., and Greenleaf, A.L. (2010). CDK12 is a transcription elongation-associated CTD kinase, the metazoan ortholog of yeast Ctk1. Genes Dev. *24*, 2303–2316.

Beltran, H., Prandi, D., Mosquera, J.M., Benelli, M., Puca, L., Cyrta, J., Marotz, C., Giannopoulou, E., Chakravarthi, B.V.S.K., Varambally, S., et al. (2016). Divergent clonal evolution of castration-resistant neuroendocrine prostate cancer. Nat. Med. *22*, 298–305.

Bernard, D., Pourtier-Manzanedo, A., Gil, J., and Beach, D.H. (2003). Myc confers androgen-independent prostate cancer cell growth. J. Clin. Invest. *112*, 1724–1731.

Blazek, D., Kohoutek, J., Bartholomeeusen, K., Johansen, E., Hulinkova, P., Luo, Z., Cimermancic, P., Ule, J., and Peterlin, B.M. (2011). The Cyclin K/Cdk12 complex maintains genomic stability via regulation of expression of DNA damage response genes. Genes Dev. *25*, 2158–2172.

Bolotin, D.A., Poslavsky, S., Mitrophanov, I., Shugay, M., Mamedov, I.Z., Putintseva, E.V., and Chudakov, D.M. (2015). MiXCR: software for comprehensive adaptive immunity profiling. Nat. Methods *12*, 380–381.

Bretones, G., Delgado, M.D., and León, J. (2015). Myc and cell cycle control. Biochim. Biophys. Acta *1849*, 506–516.

Cabel, L., Loir, E., Gravis, G., Lavaud, P., Massard, C., Albiges, L., Baciarello, G., Loriot, Y., and Fizazi, K. (2017). Long-term complete remission with Ipilimumab in metastatic castrate-resistant prostate cancer: case report of two patients. J. Immunother. Cancer *5*, 31.

Cancer Genome Atlas Research Network (2015). The molecular taxonomy of primary prostate cancer. Cell *163*, 1011–1025.

Chan, V.W.F., Kothakota, S., Rohan, M.C., Panganiban-Lustan, L., Gardner, J.P., Wachowicz, M.S., Winter, J.A., and Williams, L.T. (1999). Secondary lymphoid-tissue chemokine (SLC) is chemotactic for mature dendritic cells. Blood *93*, 3610–3616.

Cheng, S.W., Kuzyk, M.A., Moradian, A., Ichu, T.A., Chang, V.C., Tien, J.F., Vollett, S.E., Griffith, M., Marra, M.A., and Morin, G.B. (2012). Interaction of cyclin-dependent kinase 12/CrkRS with cyclin K1 is required for the phosphorylation of the C-terminal domain of RNA polymerase II. Mol. Cell. Biol. *32*, 4691–4704.

Cieslik, M., Chugh, R., Wu, Y.M., Wu, M., Brennan, C., Lonigro, R., Su, F., Wang, R., Siddiqui, J., Mehra, R., et al. (2015). The use of exome capture RNA-seq for highly degraded RNA with application to clinical cancer sequencing. Genome Res. *25*, 1372–1381.

Curiel, T.J., Coukos, G., Zou, L., Alvarez, X., Cheng, P., Mottram, P., Evdemon-Hogan, M., Conejo-Garcia, J.R., Zhang, L., Burow, M., et al. (2004). Specific recruitment of regulatory T cells in ovarian carcinoma fosters immune privilege and predicts reduced survival. Nat. Med. *10*, 942–949.

Cutruzzolà, F., Giardina, G., Marani, M., Macone, A., Paiardini, A., Rinaldo, S., and Paone, A. (2017). Glucose metabolism in the progression of prostate cancer. Front. Physiol. *8*, 97.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., and Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics *21*, 3439–3440.

Ekumi, K.M., Paculova, H., Lenasi, T., Pospichalova, V., Bösken, C.A., Rybarikova, J., Bryja, V., Geyer, M., Blazek, D., and Barboric, M. (2015). Ovarian carcinoma CDK12 mutations misregulate expression of DNA repair genes via deficient formation and function of the Cdk12/CycK complex. Nucleic Acids Res. *43*, 2575–2589.

Fraser, M., Sabelnykova, V.Y., Yamaguchi, T.N., Heisler, L.E., Livingstone, J., Huang, V., Shiah, Y.-J., Yousif, F., Lin, X., Masella, A.P., et al. (2017). Genomic hallmarks of localized, non-indolent prostate cancer. Nature *541*, 359–364.

Gehring, J.S., Fischer, B., Lawrence, M., and Huber, W. (2015). SomaticSignatures: inferring mutational signatures from single-nucleotide variants. Bioinformatics *31*, 3673–3675.

Gosling, J., Dairaghi, D.J., Wang, Y., Hanley, M., Talbot, D., Miao, Z., and Schall, T.J. (2000). Cutting edge: identification of a novel chemokine receptor that binds dendritic cell- and T cell-active chemokines including ELC, SLC, and TECK. J. Immunol. *164*, 2851–2856.

Graff, J.N., Alumkal, J.J., Drake, C.G., Thomas, G.V., Redmond, W.L., Farhad, M., Cetnar, J.P., Ey, F.S., Bergan, R.C., Slottke, R., and Beer, T.M. (2016). Early evidence of anti-PD-1 activity in enzalutamide-resistant prostate cancer. Oncotarget *7*, 52810–52817.

Grasso, C.S., Wu, Y.-M., Robinson, D.R., Cao, X., Dhanasekaran, S.M., Khan, A.P., Quist, M.J., Jing, X., Lonigro, R.J., Brenner, J.C., et al. (2012). The mutational landscape of lethal castration-resistant prostate cancer. Nature *487*, 239–243.

Hacohen, N., Fritsch, E.F., Carter, T.A., Lander, E.S., and Wu, C.J. (2013). Getting personal with neoantigen-based therapeutic cancer vaccines. Cancer Immunol. Res. *1*, 11–15.

Herschkowitz, J.I., He, X., Fan, C., and Perou, C.M. (2008). The functional loss of the retinoblastoma tumour suppressor is a common event in basal-like and luminal B breast carcinomas. Breast Cancer Res. *10*, R75.

Hiratani, I., Ryba, T., Itoh, M., Yokochi, T., Schwaiger, M., Chang, C.W., Lyou, Y., Townes, T.M., Schübeler, D., and Gilbert, D.M. (2008). Global reorganization of replication domains during embryonic stem cell differentiation. PLoS Biol. *6*, e245.

Joshi, P.M., Sutor, S.L., Huntoon, C.J., and Karnitz, L.M. (2014). Ovarian cancer-associated mutations disable catalytic activity of CDK12, a kinase that promotes homologous recombination repair and resistance to cisplatin and poly(ADP-ribose) polymerase inhibitors. J. Biol. Chem. *289*, 9247–9253.

Juan, H.C., Lin, Y., Chen, H.R., and Fann, M.J. (2016). Cdk12 is essential for embryonic development and the maintenance of genomic stability. Cell Death Differ. *23*, 1038–1048.

Ko, T.K., Kelly, E., and Pines, J. (2001). CrkRS: a novel conserved Cdc2-related protein kinase that colocalises with SC35 speckles. J. Cell Sci. *114*, 2591–2603.

Kumar, A., Coleman, I., Morrissey, C., Zhang, X., True, L.D., Gulati, R., Etzioni, R., Bolouri, H., Montgomery, B., White, T., et al. (2016). Substantial interindividual and limited intraindividual genomic diversity among tumors from men with metastatic prostate cancer. Nat. Med. *22*, 369–378.

Kwon, E.D., Drake, C.G., Scher, H.I., Fizazi, K., Bossi, A., van den Eertwegh, A.J., Krainer, M., Houede, N., Santos, R., Mahammedi, H., et al.; CA184-043 Investigators (2014). Ipilimumab versus placebo after radiotherapy in patients with metastatic castration-resistant prostate cancer that had progressed after docetaxel chemotherapy (CA184-043): a multicentre, randomised, double-blind, phase 3 trial. Lancet Oncol. *15*, 700–712.

Law, C.W., Chen, Y., Shi, W., and Smyth, G.K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. *15*, R29.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. (2013). Software for computing and annotating genomic ranges. PLoS Comput. Biol. *9*, e1003118.

Layer, R.M., Chiang, C., Quinlan, A.R., and Hall, I.M. (2014). LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. *15*, R84.

Le, D.T., Uram, J.N., Wang, H., Bartlett, B.R., Kemberling, H., Eyring, A.D., Skora, A.D., Luber, B.S., Azad, N.S., Laheru, D., et al. (2015). PD-1 blockade in tumors with mismatch-repair deficiency. N. Engl. J. Med. *372*, 2509–2520.

Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. *1*, 417–425.

Mateo, J., Carreira, S., Sandhu, S., Miranda, S., Mossop, H., Perez-Lopez, R., Nava Rodrigues, D., Robinson, D., Omlin, A., Tunariu, N., et al. (2015). DNA-repair defects and olaparib in metastatic prostate cancer. N. Engl. J. Med. *373*, 1697–1708.

McGranahan, N., Furness, A.J., Rosenthal, R., Ramskov, S., Lyngaa, R., Saini, S.K., Jamal-Hanjani, M., Wilson, G.A., Birkbak, N.J., Hiley, C.T., et al. (2016). Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. Science *351*, 1463–1469.

Mermel, C.H., Schumacher, S.E., Hill, B., Meyerson, M.L., Beroukhim, R., and Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. *12*, R41.

Nagarsheth, N., Wicha, M.S., and Zou, W. (2017). Chemokines in the cancer microenvironment and their relevance in cancer immunotherapy. Nat. Rev. Immunol. *17*, 559–572.

Negrini, S., Gorgoulis, V.G., and Halazonetis, T.D. (2010). Genomic instability–an evolving hallmark of cancer. Nat. Rev. Mol. Cell Biol. *11*, 220–228.

Newton, M.A., Quintana, F.A., Den Boon, J.A., Sengupta, S., and Ahlquist, P. (2007). Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. Ann. Appl. Stat. *1*, 85–106.

Nishino, M., Sholl, L.M., Hodi, F.S., Hatabu, H., and Ramaiya, N.H. (2015). Anti-PD-1-related pneumonitis during cancer immunotherapy. N. Engl. J. Med. *373*, 288–290.

Olshen, A.B., Venkatraman, E.S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics *5*, 557–572.

Palanisamy, N., Ateeq, B., Kalyana-Sundaram, S., Pflueger, D., Ramnarayanan, K., Shankar, S., Han, B., Cao, Q., Cao, X., Suleman, K., et al. (2010). Rearrangements of the RAF kinase pathway in prostate cancer, gastric cancer and melanoma. Nat. Med. *16*, 793–798.

Parikh, N., Hilsenbeck, S., Creighton, C.J., Dayaram, T., Shuck, R., Shinbrot, E., Xi, L., Gibbs, R.A., Wheeler, D.A., and Donehower, L.A. (2014). Effects of TP53 mutational status on gene expression patterns across 10 human cancer types. J. Pathol. *232*, 522–533.

Polak, P., Kim, J., Braunstein, L.Z., Karlic, R., Haradhavala, N.J., Tiao, G., Rosebrock, D., Livitz, D., Kübler, K., Mouw, K.W., et al. (2017). A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. Nat. Genet. *49*, 1476–1486.

Popova, T., Manié, E., Boeva, V., Battistella, A., Goundiam, O., Smith, N.K., Mueller, C.R., Raynal, V., Mariani, O., Sastre-Garau, X., and Stern, M.H. (2016). Ovarian cancers harboring inactivating mutations in CDK12 display a distinct genomic instability pattern characterized by large tandem duplications. Cancer Res. *76*, 1882–1891.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

Robinson, D., Van Allen, E.M., Wu, Y.M., Schultz, N., Lonigro, R.J., Mosquera, J.M., Montgomery, B., Taplin, M.E., Pritchard, C.C., Attard, G., et al. (2015). Integrative clinical genomics of advanced prostate cancer. Cell *161*, 1215–1228.

Robinson, D.R., Wu, Y.M., Lonigro, R.J., Vats, P., Cobain, E., Everett, J., Cao, X., Rabban, E., Kumar-Sinha, C., Raymond, V., et al. (2017). Integrative clinical genomics of metastatic cancer. Nature *548*, 297–303.

Saal, L.H., Johansson, P., Holm, K., Gruvberger-Saal, S.K., She, Q.B., Maurer, M., Koujak, S., Ferrando, A.A., Malmström, P., Memeo, L., et al. (2007). Poor prognosis in carcinoma is associated with a gene expression signature of aberrant PTEN tumor suppressor pathway activity. Proc. Natl. Acad. Sci. USA *104*, 7564–7569.

Sergushichev, A. (2016). An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. bioRxiv. https://doi.org/10.1101/060012.

Sharma, P., Hu-Lieskovan, S., Wargo, J.A., and Ribas, A. (2017). Primary, adaptive, and acquired resistance to cancer immunotherapy. Cell *168*, 707–723.

Smyth, G.K. (2005). Limma: linear models for microarray data. In Bioinformatics and Computational Biology Solutions Using R and Bioconductor, R. Gentleman, V.J. Carey, W. Huber, R.A. Irizarry, and S. Dudoit, eds. (New York, NY: Springer New York), pp. 397–420.

Strasner, A., and Karin, M. (2015). Immune infiltration and prostate cancer. Front. Oncol. *5*, 128.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550.

Tomlins, S.A., Rhodes, D.R., Perner, S., Dhanasekaran, S.M., Mehra, R., Sun, X.W., Varambally, S., Cao, X., Tchinda, J., Kuefer, R., et al. (2005). Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science *310*, 644–648.

Turlach, B., and Weingessel, A. (2013). Quadprog: Functions to solve quadratic programming problems.

Vander Heiden, M.G., Cantley, L.C., and Thompson, C.B. (2009). Understanding the Warburg effect: the metabolic requirements of cell proliferation. Science *324*, 1029–1033.

Vicari, A.P., Figueroa, D.J., Hedrick, J.A., Foster, J.S., Singh, K.P., Menon, S., Copeland, N.G., Gilbert, D.J., Jenkins, N.A., Bacon, K.B., and Zlotnik, A. (1997). TECK: a novel CC chemokine specifically expressed by thymic dendritic cells and potentially involved in T cell development. Immunity *7*, 291–301.

Waldron, L., and Riester, M. (2017). HGNChelper-package: Handy functions for working with HGNC gene symbols and Affymetrix probeset identifiers. https://bitbucket.org/lwaldron/hgnchelper.

Wu, Y.M., Su, F., Kalyana-Sundaram, S., Khazanov, N., Ateeq, B., Cao, X., Lonigro, R.J., Vats, P., Wang, R., Lin, S.F., et al. (2013). Identification of targetable FGFR gene fusions in diverse cancers. Cancer Discov. *3*, 636–647.

Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., Treviño, V., Shen, H., Laird, P.W., Levine, D.A., et al. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. Nat. Commun. *4*, 2612.

Yu, J., Deshmukh, H., Gutmann, R.J., Emnett, R.J., Rodriguez, F.J., Watson, M.A., Nagarajan, R., and Gutmann, D.H. (2009). Alterations of BRAF and HIPK2 loci predominate in sporadic pilocytic astrocytoma. Neurology *73*, 1526–1531.

Yuan, X., Li, T., Wang, H., Zhang, T., Barua, M., Borgesi, R.A., Bubley, G.J., Lu, M.L., and Balk, S.P. (2006). Androgen receptor remains critical for cell-cycle progression in androgen-independent CWR22 prostate cancer cells. Am. J. Pathol. *169*, 682–696.

Zou, W. (2006). Regulatory T cells, tumour immunity and immunotherapy. Nat. Rev. Immunol. *6*, 295–307.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Rabbit polyclonal anti-CDK12	Cell Signaling	11973; RRID: AB_2715688
Rabbit monoclonal anti-CD3 (2GV6)	Roche	790-4341
Biological Samples		
Tumor/ Normal tissues from prostate cancer patients	University of Michigan MI-ONCOSEQ collection	See STAR Methods and Table S1
Tumor/ Normal tissues from prostate cancer patients	University of Michigan Rapid autopsy program	See STAR Methods and Table S1
Tumor/ Normal tissues from prostate cancer patients	SU2C-PCF, Multiple tissue source sites	See STAR Methods and Table S1
Chemicals, Peptides, and Recombinant Proteins		
Actinomycin D	Sigma-Aldrich	A1410-10MG
RQ1 RNase-Free DNase	Promega	M6101
Superscript II Reverse Transcriptase	Invitrogen	18064-071
RNase H	Invitrogen	18021-071
DNA Polymerase I	New England Biolabs	M0209L
USER Enzyme	New England Biolabs	M5505L
Phusion High-Fidelity DNA Polymerase	New England Biolabs	M0530L
Critical Commercial Assays		
AllPrep DNA/RNA/miRNA Universal Kit	QIAGEN	80224
KAPA Hyper Prep Kit for Illumina	Kapa Biosystems	KK8504
SureSelect XT Human All Exon V4 library	Agilent Technologies	5190-4632
SureSelectXT Reagent kit	Agilent Technologies	G9611B
RNA 6000 Nano kit	Agilent Technologies	5067-1511
DNA 1000 kit	Agilent Technologies	5067-1504
QIAGEN Multiplex PCR Kit	QIAGEN	206143
immunoSEQ hsTCRB Kit	Adaptive Biotechnologies	ISK10101
Deposited Data		
BAM files of mCRPC in Mi-Oncoseq program, University of Michigan Clinical Sequencing Exploratory Research (CSER)	This study and Robinson et al., 2017	dbGaP (phs000673.v2.p1)
BAM files of the SU2C-PCF CRPC150 cohort	Robinson et al., 2015	dbGaP (phs000915.v1.p1)
Mutation calls and clinical annotation of the SU2C-PCF CRPC150 and extended cohort	Robinson et al., 2015	cBio portal, http://www.cbioportal.org/study?id=prad_p1000
BAM files of mCRPC in Rapid autopsy cohort at the University of Michigan	Grasso et al., 2012	dbGAP (phs000554.v1.p1)
Experimental Models: Cell Lines		
LNCaP	ATCC	CRL-1740
HeLaS3	ATCC	CCL-2.2
Oligonucleotides		
NEBNext Multiplex Oligos for Illumina	New England Biolabs	E7535L
NEBNext Multiplex Oligos for Illumina Index Set 2	New England Biolabs	E7500L

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Random Primers	Invitrogen	48190-011
ON-TARGETplus CDK12 siRNA	GE Healthcare	L-004031-00-0005
Software and Algorithms		
NCBI Multiple Sequence Alignment Viewer	NCBI	https://www.ncbi.nlm.nih.gov/projects/msaviewer/#
CRISPR Design	Zhang Lab, MIT 2017	http://crispr.mit.edu
MiXCR	Bolotin et al., 2015	https://github.com/milaboratory/mixcr
GenomicRanges	Lawrence et al., 2013	https://bioconductor.org/packages/release/bioc/html/GenomicRanges.html
Clinical RNA-seq Pipeline (CRISP)	This paper and Robinson et al., 2017	https://github.com/mcieslik-mctp/bootstrap-rnascape
Comprehensive Detection and Analysis of Chimeras (CODAC)	This paper and Robinson et al., 2017	https://github.com/mcieslik-mctp/codac
Ggplot2	http://ggplot2.org/book/	https://cran.r-project.org/web/packages/ggplot2/index.html
DNACopy	Olshen et al., 2004	http://bioconductor.org/packages/release/bioc/html/DNAcopy.html
biomaRt	Durinck et al., 2005	https://bioconductor.org/packages/release/bioc/html/biomaRt.html
HGNChelper	Waldron and Riester, 2017	https://www.rdocumentation.org/packages/HGNChelper/versions/0.3.4
fgsea	Sergushichev, 2016	https://github.com/ctlab/fgsea
edgeR	Robinson et al., 2010	http://bioconductor.org/packages/release/bioc/html/edgeR.html
limma	Ritchie et al., 2015	http://bioconductor.org/packages/release/bioc/html/limma.html
Novoalign	Novocraft	http://www.novocraft.com/products/novoalign
Picard	Broad Institute	https://github.com/broadinstitute/picard
Freebayes	https://github.com/ekg/freebayes	https://github.com/ekg/freebayes
Pindel	https://github.com/genome/pindel	https://github.com/genome/pindel
SnpEff	http://snpeff.sourceforge.net	http://snpeff.sourceforge.net
SnpSift	http://snpeff.sourceforge.net/SnpSift.html	http://snpeff.sourceforge.net/SnpSift.html
Other		
SeqCap EZ HE-Oligo Kit A	Roche	06777287001
SeqCap EZ HE-Oligo Kit B	Roche	06777317001
Agencourt RNAClean XP	Beckman Coulter	A63987
AMPURE XP beads	Beckman Coulter	A63882
Dynabeads MyOne Streptavidin T1	Invitrogen	65602

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Arul M. Chinnaiyan (arul@med.umich.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell lines

LNCaP (male, prostate carcinoma) and HeLaS3 (female, cervical adenocarcinoma) cell lines were obtained from the American Type Culture Collection. LNCaP cells were cultured in RPMI1640 medium, and HeLaS3 cells were cultured in Ham's F-12K medium, both

supplemented with 10% fetal bovine serum (FBS; Invitrogen) and 1% penicillin/streptomycin (Invitrogen). Cell lines were maintained at 37°C in a 5% CO_2 cell culture incubator. Cell lines were genotyped to confirm their identity at the University of Michigan Sequencing Core and tested routinely for Mycoplasma contamination.

Human subjects and patient inclusion

Sequencing of clinical samples was approved by the Institutional Review Board of the University of Michigan (Michigan Oncology Sequencing Protocol, MI-ONCOSEQ, IRB # HUM00046018, HUM00067928, HUM00056496). Patients with clinical evidence of metastatic castration-resistant prostate cancer (mCRPC) that could be feasibly accessed by image-guided biopsy were eligible for inclusion. Consecutive cases from SU2C, mCRPC enrolled in Mi-Oncoseq, and the University of Michigan rapid autopsy series, with at least 25% tumor content as determined by post-sequencing analysis of zygosity shift and copy-number adjusted variant allele fraction using the Mi-Oncoseq clinical analysis pipeline, were included in this study (see Table S1 for source cohort). All patients provided written informed consent to obtain fresh tumor biopsies and to perform comprehensive molecular profiling of tumor and germline exomes and tumor transcriptomes.

METHOD DETAILS

Kinase domain alignment

Alignment of the kinase domains of 30 members of the human CDK and MAPK families of protein kinases were performed using BLASTp followed by visualizations using NCBI Multiple Sequence Alignment Viewer 1.6.0 with no master sequence set. Amino acid residues were shaded by conservation using NCBI Multiple Sequence Alignment Viewer 1.6.0 using frequency based differences, with highly conserved residues shaded red, moderately conserved residues shaded blue, and nonconserved residues shaded gray (https://www.ncbi.nlm.nih.gov/projects/msaviewer/#).

siRNA-mediated knockdown of CDK12

For the *CDK12* knockdown experiment, a pooled ON-TARGETplus siRNA targeting *CDK12* (Dharmacon/ GE Healthcare) was transfected into LNCaP cells using Oligofectamine (Life Sciences). To ensure an efficient knockdown of *CDK12*, cells were transfected again with the same siRNA 48 hours later (48-hr time point), and incubated for another 24 hours (72-hr time point). Scrambled siRNA was used as a negative control (ON-TARGETplus Non-targeting Pool, Dharmacon/ GE Healthcare). For CDK12 protein detection, cells were lysed in RIPA buffer containing protease inhibitor cocktail (Pierce). Expression of CDK12 protein was measured by western blotting using anti-CDK12 antibody (Cell Signaling). For the cell proliferation assay, LNCaP cells were trypsinized 72 hours post-transfection, and plated in triplicate in 24-well plates. The cells were incubated at 37°C and 5% CO_2 atmosphere using the IncuCyte live-cell imaging system (Essen Biosciences). Cell proliferation was assessed by kinetic imaging confluence measurements at 3-hour time intervals.

Immunostaining of T lymphocytes

Immunohistochemistry (IHC) was performed on formalin-fixed paraffin-embedded tumor tissue sections using CONFIRM anti-CD3 (2GV6) rabbit monoclonal antibody (Ventana Medical Systems). IHC was carried out using an automated protocol developed for the Benchmark XT automated slide staining system and detected using the UltraView Universal DAB detection kit (Ventana Medical Systems). Hematoxylin II (Ventana-Roche) was used as counterstain. Human tonsil sections were used as the positive control. CD3-positive T lymphocytes exhibited membranous and cytoplasmic staining.

Integrative clinical sequencing

Integrative clinical sequencing was performed using standard protocols in our Clinical Laboratory Improvement Amendments (CLIA) compliant sequencing lab (Robinson et al., 2015, 2017). In brief, tumor genomic DNA and total RNA were purified from the same sample using the AllPrep DNA/RNA/miRNA kit (QIAGEN). Matched normal genomic DNA from blood, buccal swab, or saliva was isolated using the DNeasy Blood & Tissue Kit (QIAGEN). RNA sequencing was performed by exome-capture transcriptome platform (Cieslik et al., 2015). Exome libraries of matched pairs of tumor/normal DNAs were prepared as described before (Robinson et al., 2015, 2017), using the Agilent SureSelect Human All Exon v4 platform (Agilent). All the samples were sequenced on the Illumina HiSeq 2000 or HiSeq 2500 (Illumina Inc) in paired-end mode. The primary base call files were converted into FASTQ sequence files using the bcl2fastq converter tool bcl2fastq-1.8.4 in the CASAVA 1.8 pipeline.

T cell receptor β repertoire deep sequencing

Amplification and sequencing of [TCRB / IGH / IGKL / TCRAD / TCRG] CDR3 was performed using the immunoSEQ Platform (Adaptive Biotechnologies). Same DNA aliquot obtained from frozen tumor tissues was used as for the exome sequencing. The immunoSEQ Platform combines multiplex PCR with high throughput sequencing and a sophisticated bioinformatics pipeline for [TCRB / IGH / IGKL / TCRAD / TCRG] CDR3 analysis that includes internal PCR amplification controls. PCR reactions were performed

on 60 mCRPC tumor samples with 2 μg of DNA, and PCR fragments were sequenced on the Illumina MiSeq. Computational analysis of sequencing data, including the estimation of the total number of templates, identification, and clonotypes was performed using the vendor-supplied analysis portal.

QUANTIFICATION AND STATISTICAL ANALYSIS

Whole-genome sequencing data analysis

The bcbio-nextgen pipeline version 1.0.3 was used for the initial steps of tumor whole-genome data analysis. Paired-end reads were aligned to the GRCh38 reference using BWA (bcbio default settings), and structural variant calling was done using LUMPY (Layer et al., 2014) (bcbio default settings), with the following post-filtering criteria: "(SR>=1 & PE>=1 & SU>=7) & (abs(SVLEN)>5e4) & DP<1000 & FILTER=="PASS"." The following settings were chosen to minimize the number of expected germline variants: (FDR < 0.05 for germline status for both deletions and duplications). Replication domain sizes for normal tissues were obtained from GSE53984, and transactivation domain sizes for prostate cancer cell lines were obtained from GSE73782.

Exome data analysis

The FASTQ sequence files from whole exome libraries were processed through an in-house pipeline constructed for analysis of paired tumor/normal data. The sequencing reads were aligned to the GRCh37 reference genome using Novoalign (version 3.02.08) (Novocraft) and converted into BAM files using SAMtools (version 0.1.19). Sorting, indexing, and duplicate marking of BAM files used Novosort (version 1.03.02). Mutation analysis was performed using freebayes (version 1.0.1) and pindel (version 0.2.5b9). Variants were annotated to RefSeq (via the UCSC genome browser, retrieved on 8/22/2016), as well as COSMIC v79, dbSNP v146, ExAC v0.3, and 1000 Genomes phase 3 databases using snpEff and snpSift (version 4.1g). SNVs and indels were called as somatic if they were present with at least 6 variant reads and 5% allelic fraction in the tumor sample, and present at no more than 2% allelic fraction in the normal sample with at least 20X coverage; additionally, the ratio of variant allelic fractions between tumor and normal samples was required to be at least six in order to avoid sequencing and alignment artifacts at low allelic fractions. Minimum thresholds were increased for indels observed to be recurrent across a pool of hundreds of platform- and protocol-matched normal samples. Specifically, for each such indel, a logistic regression model was used to model variant and total read counts across the normal pool using PCR duplication rate as a covariate, and the results of this model were used to estimate a predicted number of variant reads (and therefore allelic fraction) for this indel in the sample of interest, treating the total observed coverage at this genomic position as fixed. The variant read count and allelic fraction thresholds were increased by these respective predicted values. This filter eliminates most recurrent indel artifacts without affecting our ability to detect variants in homopolymer regions from tumors exhibiting microsatellite instability. Germline variants were called using ten variant reads and 20% allelic fraction as minimum thresholds, and were classified as rare if they had less than 1% observed population frequency in both the 1000 Genomes and ExAC databases.

Exome data was analyzed for copy number aberrations and loss of heterozygosity by jointly segmenting B-allele frequencies and log2-transformed tumor/normal coverage ratios across targeted regions using the DNAcopy (version 1.48.0) implementation of the Circular Binary Segmentation algorithm. The Expectation-Maximization Algorithm was used to jointly estimate tumor purity and classify regions by copy number status. Additive adjustments were made to the log2-transformed coverage ratios to allow for the possibility of non-diploid tumor genomes; the adjustment resulting in the best fit to the data using minimum mean-squared error was chosen automatically and manually overridden if necessary.

Assignment of pathway status

For pathway status depicted in Figure 2A, the following criteria were applied: (1) *TP53*, *RB1*, *PTEN*, and *ATM* cases with biallelic inactivation by mutation, copy loss, copy neutral LOH, gene fusion or known pathogenic germline allele were scored as mutant for that pathway; (2) For BRCA pathway, biallelic inactivations of *BRCA2*, *BRCA1*, *PALB2*, or *RAD51B/C* were scored as mutant; (3) For PI3K pathway activation, activating mutations or amplifications of *PIK3CA*, *PIK3CB*, truncating or iSH2 mutations in *PIK3R1*, or known activating mutations in *AKT1* were included; (4) For WNT pathway activation, biallelic inactivation of *APC*, *ZNRF3*, or *RNF43*, recurrent activating mutations of *CTNNB1*, or fusions and overexpression of *RSPO* family ligands were included; (5) For cell cycle aberrations, amplifications of *CCND1*, *CCND2*, *CCND3*, *CCNE1*, and *CDK4*, or biallelic inactivations of *RB1*, *CDKN2A*, *CDKN1B*, and *CDKN2C* were included. For all genes, amplification was defined as an absolute copy-number of seven or more.

RNA-seq data analysis

RNA-seq data processing, including quality control, read trimming, alignment, and expression quantification by read counting, was carried out as described previously (Robinson et al., 2017), using our standard clinical RNA-seq pipeline "CRISP" (available at https://github.com/mcieslik-mctp/bootstrap-rnascape) and our toolkit for the comprehensive detection of chimeric RNAs "CODAC" (available at https://github.com/mctp/codac). Both pipelines were run with default settings for paired-end RNA-seq data of at least 75bp. The only changes were made for unstranded transcriptome libraries sequenced at the Broad Institute, for which quantification using "featureCounts" (Liao et al., 2014) was used in unstranded mode "-s0." Briefly, three separate alignment passes (STAR 2.4.0g1) against the GRCh38 (hg38) reference with known splice-junctions provided by the (Gencode 27) are made for the purposes of expression quantification and fusion discovery. The first pass is a standard paired-end alignment followed by gene expression

quantification. The second and third pass are for the purpose of gene fusion discovery and enable STAR's chimeric alignment mode (chimSegmentMin: 10, chimJunctionOverhangMin: 1, alignIntronMax: 150000, chimScoreMin: 1). Fusion detection was also carried out using CODAC with default parameters to balance sensitivity and specificity (annotation preset:balanced). CODAC uses MOTR v2 a custom reference transcriptome based on a subset of Gencode 27. Fusion-Grams were prepared using CODAC (v 3.2.2) based on its standard prediction of topology (inversion, duplication, deletion, translocation), and distance (adjacent – breakpoints in two directly adjacent loci, cytoband – breakpoints within the same cytoband based on UCSC genome browser, arm – breakpoints within the same chromosome arm).

Differential expression analysis

All differential expression analyses were done using limma R-package (Smyth, 2005), with the default settings for the ''voom'' (Law et al., 2014), ''lmFit,'' ''eBayes,'' and ''topTable'' functions. The contrasts were designed as follows. First, a set of ''all wild-type'' samples were identified. These samples were wild-type (WT) for mutations in all primary genetic drivers (PGDs) of prostate cancer, i.e., *ETS* fusions, homologous recombination deficiency (*BRCA1/2*, *PALB2*, etc.), *ATM* mutations, mismatch repair deficiency, *SPOP* mutations, and *CDK12* mutations. These samples were formed a baseline to which all other groups were compared. Next, we constructed separate design matrices with coefficients for each of the primary genetic drivers, in addition to coefficients for *TP53* status, different biopsy sites (bone marrow, lymph node, soft tissue), and type of RNA-seq library (capture RNA-seq versus polyA RNA-seq). For example, *CDK12* mutant samples were contrasted with the wild-type samples, with separate coefficients for *TP53* status, library type, etc. This allowed us to estimate the log fold-changes and adjusted p values associated with each of the genetic drivers and some of the confounding variables (technical i.e., library type, and biological e.g., biopsy site, TP53 mutation status). Liver biopsies were excluded from this analysis because of the large variability in the expression of liver-specific genes in these biopsies. These estimated moderated log fold-changes and adjusted (FDR) p values were used in all of the other downstream analyses.

To estimate the number of differentially expressed genes (DEGs) associated with each PGD (Figure 2B), we had to correct for the fact that we had different statistical power to detect those differences for different groups, since the number of samples are much higher among certain groups, e.g., for *ETS*-positive prostate cancer than for *SPOP* mutant prostate cancer. Hence, we followed a sampling approach where we selected a random set of 13 samples (equal to the size of the smallest category, mismatch repair-deficient), and carried out the differential expression analysis as described before. We repeated this analysis 32 times to generate estimates of the average number of DEGs. We plot the number of DEGs, given a fixed p value, as a function of absolute logFC cutoff.

Pathway and gene set enrichment analyses

All enrichment analyses have been carried out using the Random-Set approach (Newton et al., 2007) using the shrunken log fold changes estimated above. Gene signatures were obtained from the MSigDB (Liberzon et al., 2015), and the collection of pathway gene sets curated by SABiosciences (a QIAGEN company). Identifiers (entrez gene ids, gene symbols) were mapped onto Ensemble gene_id's using Bioconductor and biomaRt (Durinck et al., 2005). If necessary, outdated gene symbols were corrected using HGNChelper (Waldron and Riester, 2017). The AR signaling score (Figure S4F) was computed using the signature by Beltran et al. (Beltran et al., 2016). Briefly, gene expression levels were converted into percentiles across the whole cohort. These percentiles were transformed using the quantile function for the normal distribution ''qnorm'' in R. For each sample these ''inverse-normal'' scores were summed to obtain the raw AR signaling score. Given that expression of AR targets strongly depends on tumor content, we constructed a linear model (R: lm), with tumor contents as a covariate and the raw score as a dependent variable. The final ''AR signaling scores'' were computed as the residual i.e., ''raw score – predicted.'' The cohort MImmScore is the cohort-level generalization of the MImmScore, as described previously (Robinson et al., 2017). It is based on the same set of immune system-related genes, but rather than scoring the immunological activity of one sample versus all other samples (MImmScore), it scores the immunological activity in one cohort versus the WT cohort (as described above). The moderated fold-changes (see section: Differential expression analysis) and the Yoshihara et al. gene set are used as input to the Random-Set method (Yoshihara et al., 2013). The resulting Z-scores and adjusted p values are shown in Figure 6B. Hallmark (Figure 6A) and immune pathway analyses (Figure S7B) were based on the Hallmark sets from MsigDB and the SABiosciences gene sets. For Figure S7B, activity scores were computed as ''Z-score * -log10(p-value)'' based on the Z-scores and p values from the Random-Set method. The intersection of genes in the LIEN_BREAST_CARCINOMA_METAPLASTIC_VS_DUCTAL_DN and LIM_MAMMARY_STEM_CELL_DN signatures was designated as ''Stem and Metaplastic dn'' in Figure S4I.

Mutation signature analysis

Mutation signature analysis was performed by interpreting the set of somatic mutations in the context of 30 known mutational signatures from the COSMIC database (https://cancer.sanger.ac.uk/cosmic/signatures). The empirical distribution of the set of trinucleotide changes around somatic single nucleotide variants was extracted for each sample using the Bioconductor SomaticSignatures package, version 2.10.0 (Gehring et al., 2015). The R package quadprog, version 1.5-5 (Turlach and Weingessel, 2013), was then used to estimate a set of 30 non-negative weights each representing the contribution of a known COSMIC signature to the observed set of trinucleotide changes. Results were visualized using the plotMutationSpectrum function from the SomaticSignatures package.

CDK12 mutation frequency analysis

Using estimates of 2.3 Mutations/Mb in CRPC and 0.95 Mutations/Mb in localized tumors (determined from cohorts sequenced and analyzed uniformly here), we expect the rate of *CDK12* mutations to increase by about 2.5-fold in CRPC. Using an empirical distribution of mutation rates for 277 localized prostate tumors, scaled to a median of 2.3 Mutations/Mb to reflect this increase, we sampled with replacement from this distribution 498 times (the size of the localized cohort), and simulated a number of mutations from the *CDK12* locus (0.0045 Mb) using a Poisson distribution and computed the number of samples with one or two simulated *CDK12* mutations. Across 1000 such iterations, we found a mean of 6 samples with single mutations in *CDK12* and 0.06 samples with two or more mutations in *CDK12*; the maximum number of samples with two or more mutations in *CDK12* across the 1000 simulations was 1. Therefore, even if the mutation rate in the localized cohort was inflated to reflect the observed mutation rate in CRPC, we would expect at most 1/498 (0.2%) extra double hits, far less than the difference observed between localized and CRPC samples.

CDK12-FTD recurrence analysis

To identify regions recurrently amplified in *CDK12* mutant cases, we first developed a random model to estimate number of peaks at any genomic region controlling for differences in gene density (since our copy-number calls are based on whole-exome sequencing data). First, we determined the sizes of all copy gains relative to the baseline copy-number; these events included all regions with three+ copies and regions with two copies on X. Next, we filtered all CNVs for focal tandem duplications (FTDs) using a narrow (< 2Mb) and wide (< 10Mb) cutoff, resulting in two separate sets of FTDs in each sample. We developed two independent null models (background models) based on the two sets of FTDs (i.e., the narrow and the wider FTDs) in both the *CDK12*-mutant and *CDK12*-wild-type sets of cases. The overall statistical procedure was to: 1) sample random peaks (i.e., generate the same number of peaks as in any of the four input sets (narrow *CDK12*-wt, narrow *CDK12*-mut, wide *CDK12*-wt, wide *CDK12*-mut)- if a peak overlapped a region that is not covered by our capture kit, it was randomized again; 2) compute coverage at all loci in the genome; 3) compute how many loci are covered by more than a given number of random peaks. This procedure was repeated 800 times for each of the four sets of peaks. This allowed us to determine what the average (across all 800 randomization) number of loci was which were covered by a least given number of peaks, i.e., the expected number of false-positive calls. Based on these models, we chose cutoffs (i.e., the minimum number of peaks) that define a region as significant based on a pre-defined empirical false-discovery rate (i.e., the expected proportion of false-positive calls among all calls). Finally, regions exceeding the predefined threshold were merged into a contiguous peak based on a distance threshold of 1Mb. Regions significant in the *CDK12* mutant cases (i.e., narrow *CDK12*-mut, wide *CDK12*-mut) were also subsequently merged to define a final set of loci with recurrent (narrow or wide) gains in *CDK12* mutant cases.

Copy-number expression aggregation

When aggregating copy-number and expression at the gene level, we defined 100kb windows centered around the canonical promoter for each gene. We overlapped those promoter regions with the copy-number segments and assigned each gene to exactly one segment. If a promoter region overlapped multiple segments, we chose the one with the higher copy-number. To analyze expression differences in each sample, we followed a strategy very similar to the one above (Differential expression analysis section). We contrasted each individual *CDK12* mutant sample with the all-wild-type group; therefore, for each gene in each sample, we computed a shrunken log fold change (relative to the all-wild-type group) and p value (based on the variance estimate in the all-wild-type group). The following thresholds were used to compute the number of genes meeting differential expression criteria: Differentially Expressed Gene: Nominal p value < 0.1. Outlier Expressed Gene: p value < 1e-3 and log fold change > 3.322 and RPKM > 4 and percentile > 0.95.

Structural variant and fusion-gram analysis

Fusion-grams were plotted using data directly from the CODAC chimeric RNA discovery pipeline (see above), which includes gene-gene fusions as well as a number of types of truncating gene fusions. All of these events were categorized into broad classes of likely duplications, deletions, inversions, and translocations, based on the topology of their breakpoints, and also based on the distance between the breakpoints from GRCh38 cytobands and loci adjacency. To compute a fusion-gram, the frequency of events within a given class combination (distance x topology) was determined relative to the total number of events across all samples of a genetic subtype (e.g., *CDK12* mutant cases). Similarly, to create fusion circos plots, we have color coded the CODAC variants based on the inferred topology of the breakpoints. To create circos plots that are representative both in terms of the number of structural variants and their topology within each genetic class, we first combined all of the structural variants across all cases within a group, and then sampled a random set of structural variants proportional to their average number.

HLA-typing analysis

PHLAT (Version 1.0) was used to determine the HLA haplotype of individuals for MHCI (HLA-A, HLA-B, HLA-C) at four-digit resolution using exome sequencing data from the patient's matched normal sample.

Integrative *in silico* neoantigen translation

Mutation analysis from exome sequencing of patient's matched tumor and normal pair along with fusion analysis from patient's transcriptome sequencing was carried out. Somatic mutations from single/dinucleotide variants as well as small insertion/deletions from

the cohort were used to identify the specific amino acid coding change. Missense mutations with > 1 RPKM expression were selected and processed using Annovar (Version 07.16.17) and in-house perl script to get 17-mer amino acid neopeptides. Mutations with start-loss, stop-gain, and splice sites were excluded from the analysis. Indels and fusions with > 1RPKM expression were selected. Inframe, indel, and fusion neopeptides of 17-mer length were created in the similar way as missense mutations. Frameshifts, indels, and fusions create novel open reading frames producing several neoantigenic peptides that are highly distinct from self. These frameshift peptides were generated until a stop codon was hit, or we reached the read evidence. Neopeptides created from indels and fusions with length less than 9-mer or with an immediate stop codon were excluded from further analysis.

IEDB peptide binding prediction

All of the neopeptides from single mutations, dinucleotides, small insertion/deletions, and fusions were than used to assess MHCI binding using the IEDB_recommended parameter from Immune Epitope Database (IEDB) (Version IEDB_MHC-2.17) and predicted high affinity MHCI binding neopeptide against patient autologous haplotypes. All neopeptides with an IEDB percentile rank < 2 were considered as high affinity binding epitopes.

T cell repertoire analysis from RNA-seq data

Repertoire analysis was carried out using MiXCR (Bolotin et al., 2015) using the recommended workflow and setting for RNA-seq data, i.e., ''-g -s hsa -p rna-seq -OallowPartialAlignments=true,'' and two rounds of ''assemblePartial'' followed by ''extendAlignments'' and ''assemble.'' MiXCR was run on all unmapped reads, paired-end reads mapped to the T cell receptor loci. The number of reads mapped to the T cell receptor loci and normalized to the number of aligned reads and the number of different CDR3 sequences were used as the TCR CDR3 cpms and TCR clones. To verify the accuracy of this approach, we compared the RNA-based estimates to TCRb DNA-based sequencing and found them in excellent agreement (Figure S7D).

Statistical Analysis

Fisher exact tests were performed for *CDK12* mutation incidence in CRPC versus primary prostate cancer in the Results section and Figure 1B; n = 360 for CRPC and n = 498 for primary tumors. Fisher exact tests were performed for *CDK12* mutation status versus *PTEN* mutation status and *CDK12* incidence versus ETS fusion status in the Results section; n = 360. A t test was used to evaluate expanded T cell clone values in differing subclasses of CRPC in Figure 6D; n = 10 for each subclass.

DATA AND SOFTWARE AVAILABILITY

The accession number for the new sequencing data (19 mCRPC cases) reported in this paper is Database of Genotypes and Phenotypes (dbGaP): phs000673.v3.p1. These data are deposited under the same dbGaP ID as Robinson et al. (2017) since they belong to a continuous sequencing program with the same IRB-approved protocol (MI-Oncoseq program, University of Michigan Clinical Sequencing Exploratory Research).

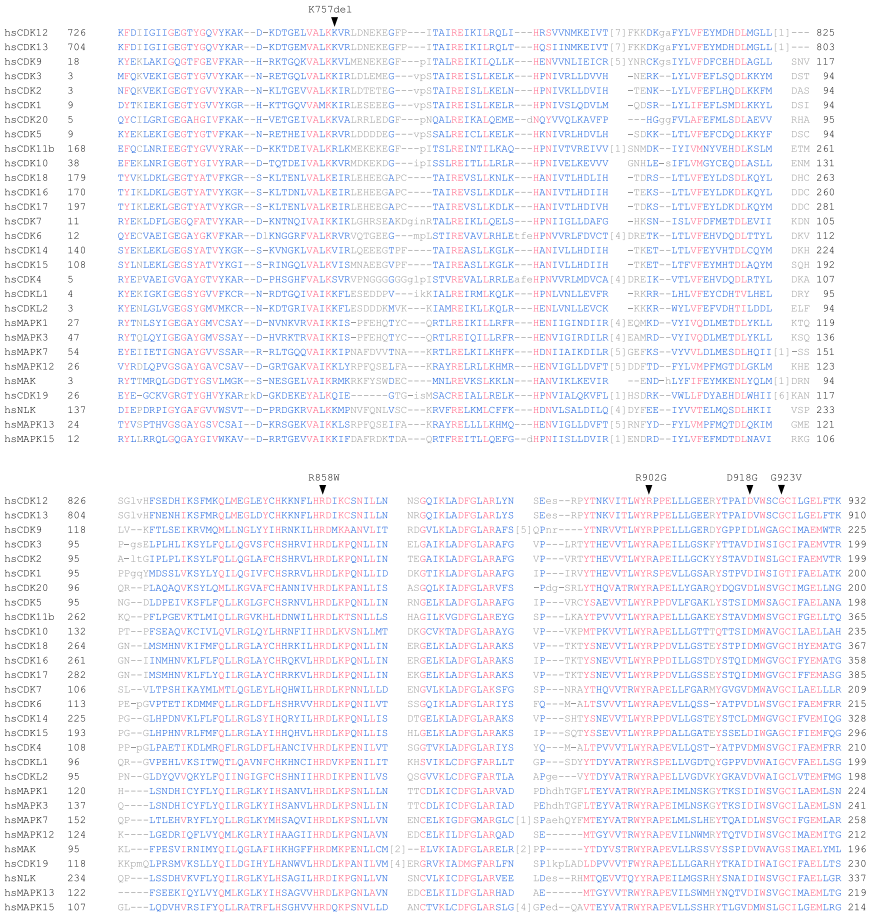


Figure S1. Alignment of the Kinase Domains of CDK12 and CDK Subfamily Kinases, Related to Figure 1A

Highly conserved residues are in red, semi-conserved residues are in blue, and divergent residues are in gray. Missense mutations identified in *CDK12* are indicated by arrowheads.

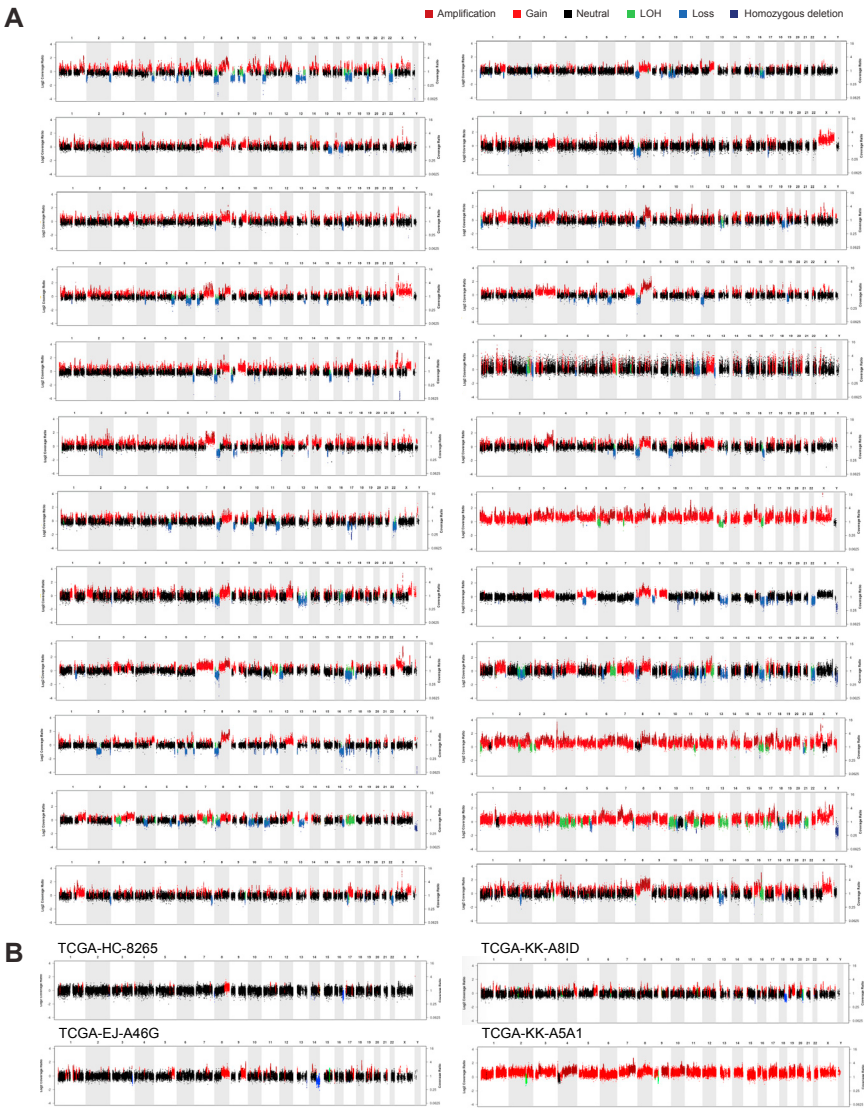


Figure S2. Copy-Number Plots of *CDK12* Mutant Tumors, Related to Figure 1

Gene copy-number landscape was assessed by whole-exome sequencing matched to germline. Chromosomes are numbered above each plot. Copy-number changes are indicated by different colors. LOH, loss of heterozygosity. Representative mCRPC cases are shown in (A), and primary prostate cancer cases are shown in (B).

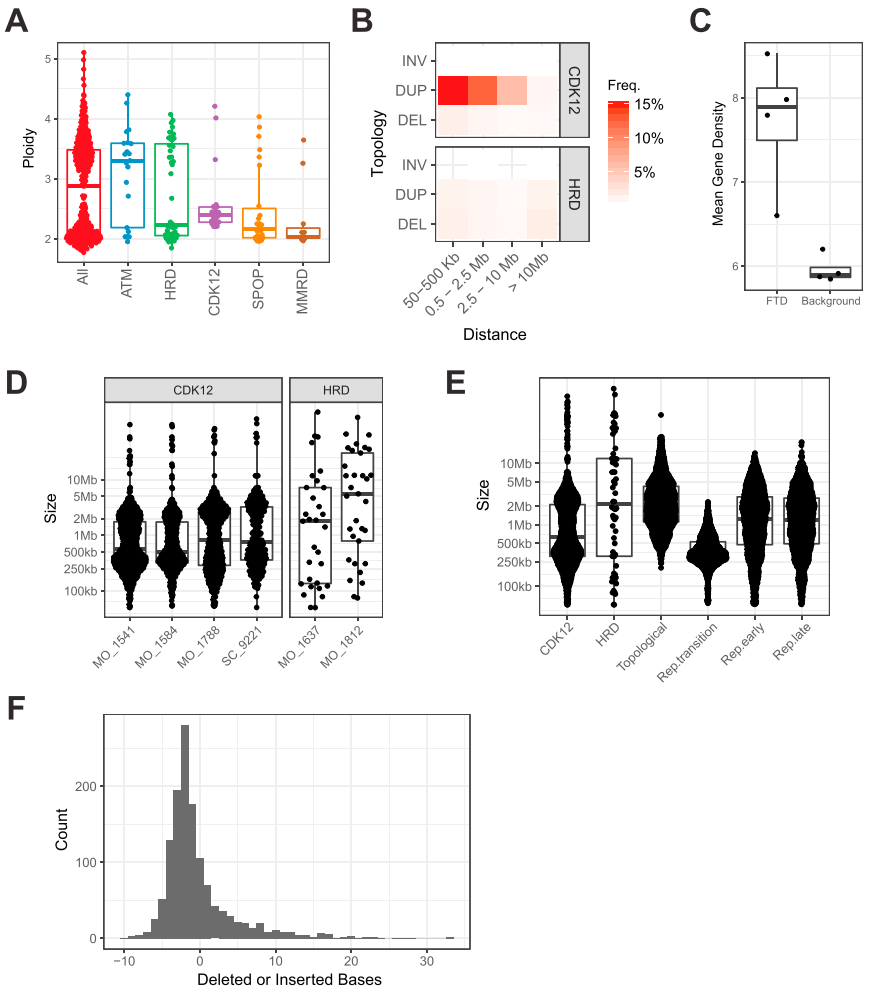


Figure S3. Genetic Instability of *CDK12* Mutant Tumors, Related to Figure 1

(A) Ploidy of tumors associated with distinct primary genetic drivers of prostate cancer.

(B) Fusion-gram inferred from structural variants detected by whole-genome sequencing.

(C) Density of genes within and outside focal tandem duplications (FTDs).

(D) Size of FTDs of example cases of tumors with aberrations in *CDK12* and homologous recombination deficiency (HRD).

(E) Size of FTDs of tumors with mutant *CDK12* or HRD compared with the size of topological domains or replication domains (transitional, early, or late).

(F) Distribution of the number of inserted or deleted based at tandem duplication breakpoints.

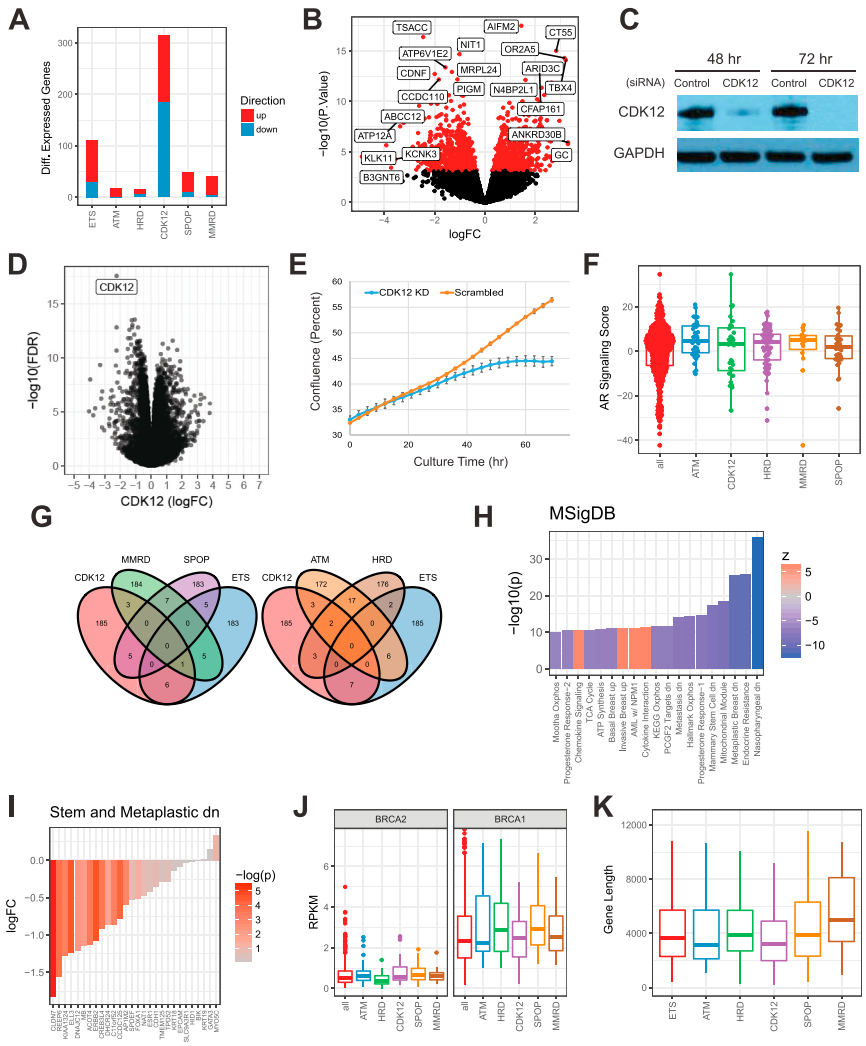


Figure S4. Transcriptional Characteristics of *CDK12* Mutant Tumors, Related to Figures 2 and 3 and Table S6

(A) Number of differentially expressed genes (DEGs) in prostate tumors with common primary genetic drivers relative to tumors with no aberrations in any of those genes.

(B) Volcano plot of DEGs in *CDK12* mutant tumors. The most significant and differential genes are highlighted.

(C) Depletion of CDK12 protein expression in LNCaP-CDK12 KD cells. CDK12 was knocked down by siRNA in LNCaP cells.

(D) Volcano plot of DEGs in LNCaP-CDK12 KD cells, demonstrating the magnitude and significance of the CDK12 knockdown.

(E) Effect of CDK12 knockdown on cell proliferation in LNCaP cells.

(F) AR signaling in prostate tumors with common primary genetic drivers. The cumulative score is based on the expression of known AR targets.

(G) Overlap between top 200 most DEGs for each of the genetic molecular subtypes of prostate cancer.

(H) Most significant pathways and signatures from the MSigDB associated with *CDK12* loss.

(I) Differential expression of genes common to the "Metaplastic Breast dn" and "Mammary Stem Cell dn" signatures from (H).

(J) Expression of *BRCA1* and *BRCA2* across genetic subtypes of prostate cancer is shown.

(K) Role of CDK12 in the transcription of long transcripts. Lengths of differentially expressed genes across genetic subtypes of prostate cancer are shown.

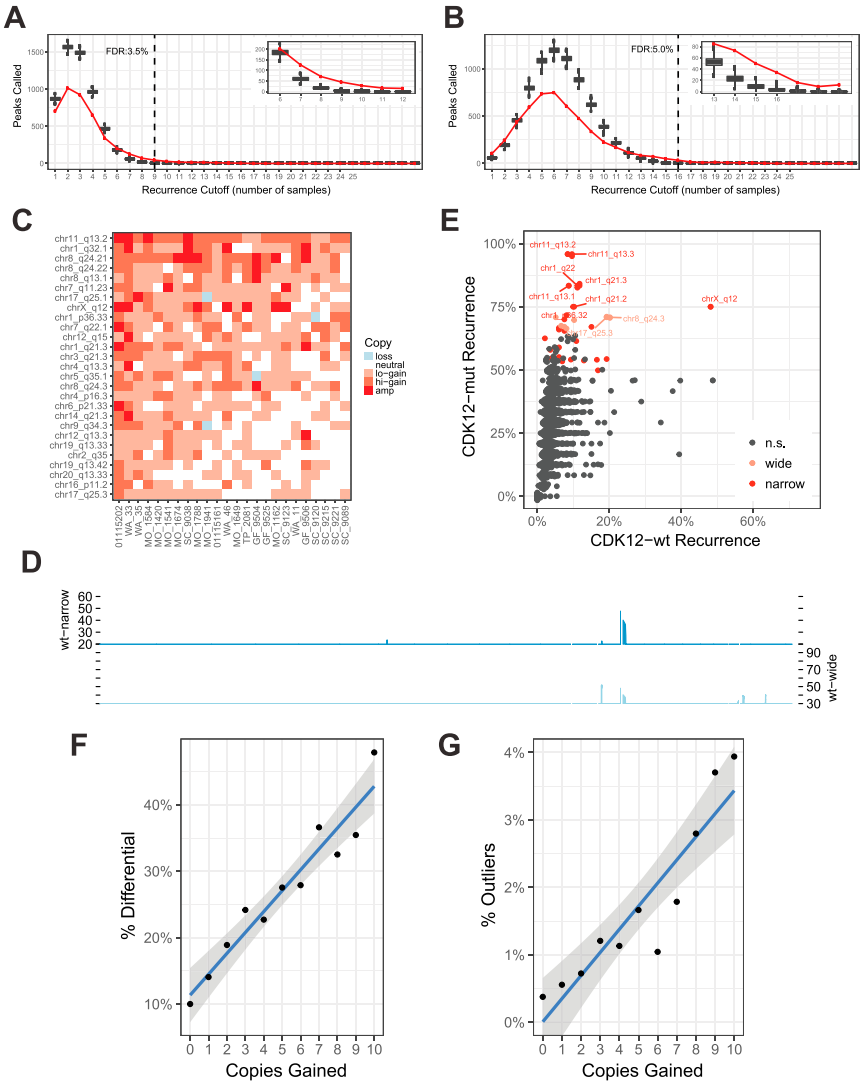


Figure S5. Recurrence of *CDK12*-Associated FTDs and Effect on Expression/Upregulation of Genes within *CDK12*-Associated FTDs, Related to Figure 4

(A and B) Empirical model to call genomic regions with recurrent focal tandem duplications. Number of loci (putative peaks, y axis) called at a given recurrence threshold (x axis) are shown. Red line indicates the observed (empirical) distribution. Black boxplots indicate the observed number of sites at a given cutoff generated by placing the peaks randomly across the genome. Dotted line indicates a cutoff which achieves the indicated false-discovery rate i.e., number of expected false positives. (A) narrow model (peaks < 2Mb). (B) wide model (peaks < 8Mb).

(C) Copy-number aberrations across loci with the most recurrent *CDK12*-associated FTDs (*CDK12*-FTDs) and all *CDK12* mutant mCRPC cases.

(D) Genome-wide frequency (percentage of *CDK12* wild-type patients) of FTDs based on a narrow (< 2Mb) and wide (< 8Mb) definition of focality.

(E) Frequency of *CDK12*-FTDs at the most recurrent loci in *CDK12* mutant and wild-type tumors.

(F) Effect of *CDK12*-FTDs on the frequency of differential expression.

(G) Dose-independent effect of *CDK12*-FTDs on the frequency of gene expression outliers.

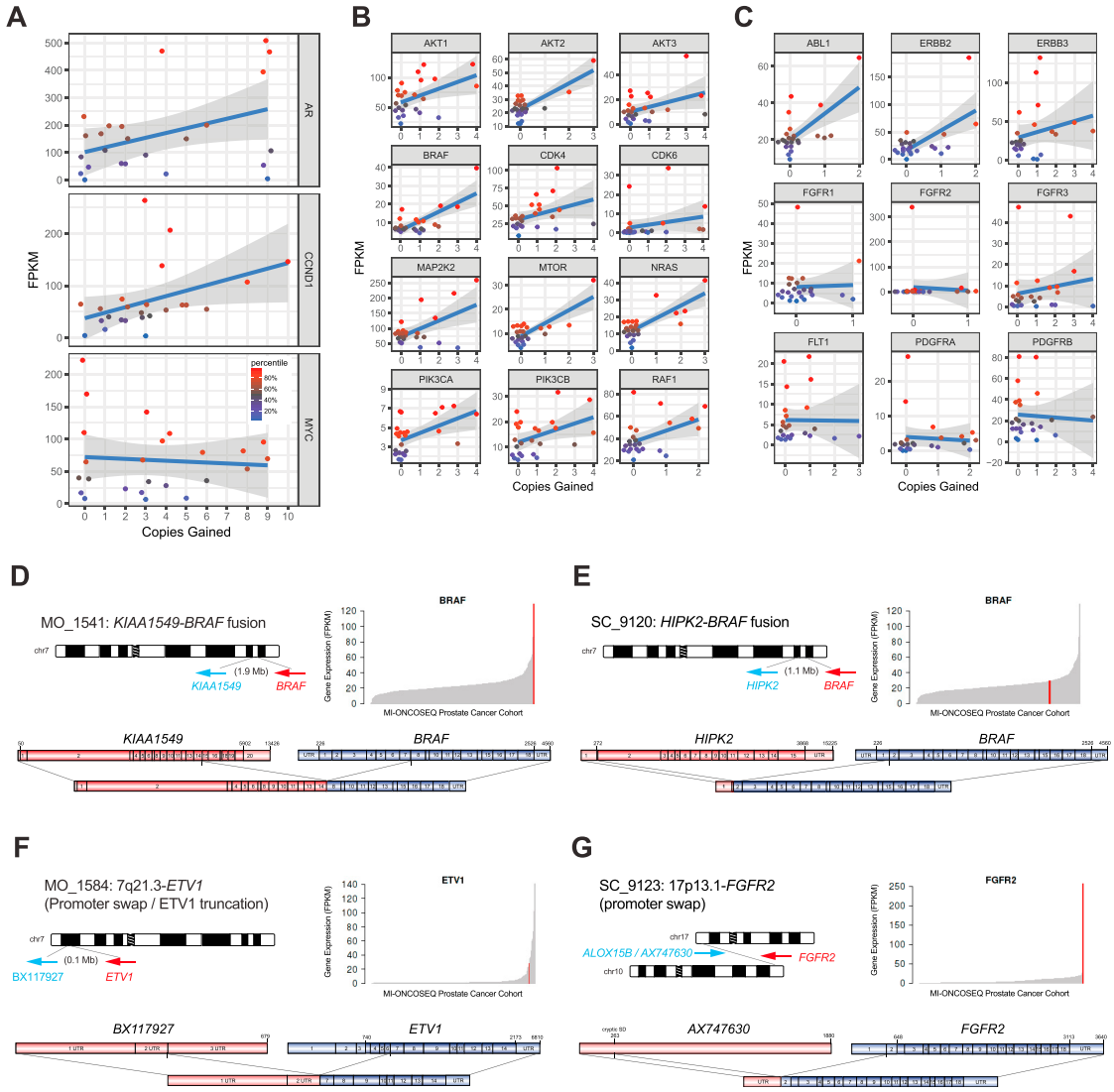


Figure S6. Effect of *CDK12*-FTDS on the Expression of Select Genes, Related to Figures 4 and 5

(A) Genes with the highest average copy-number gains in *CDK12* mutant tumors.

(B) Genes associated with oncogenic signaling pathways (e.g., *MAPK*, *AKT*, *MTOR*).

(C) Oncogenic tyrosine kinases.

(D–G) Schematic diagram of driver gene fusions identified in *CDK12*-deficient cases. *KIAA1549-BRAF* fusion is shown in D, *HIPK2-BRAF* fusion is shown in E, *BX117927-ETV1* fusion is shown in F, and AX747630-FGFR2 fusion is shown in G.

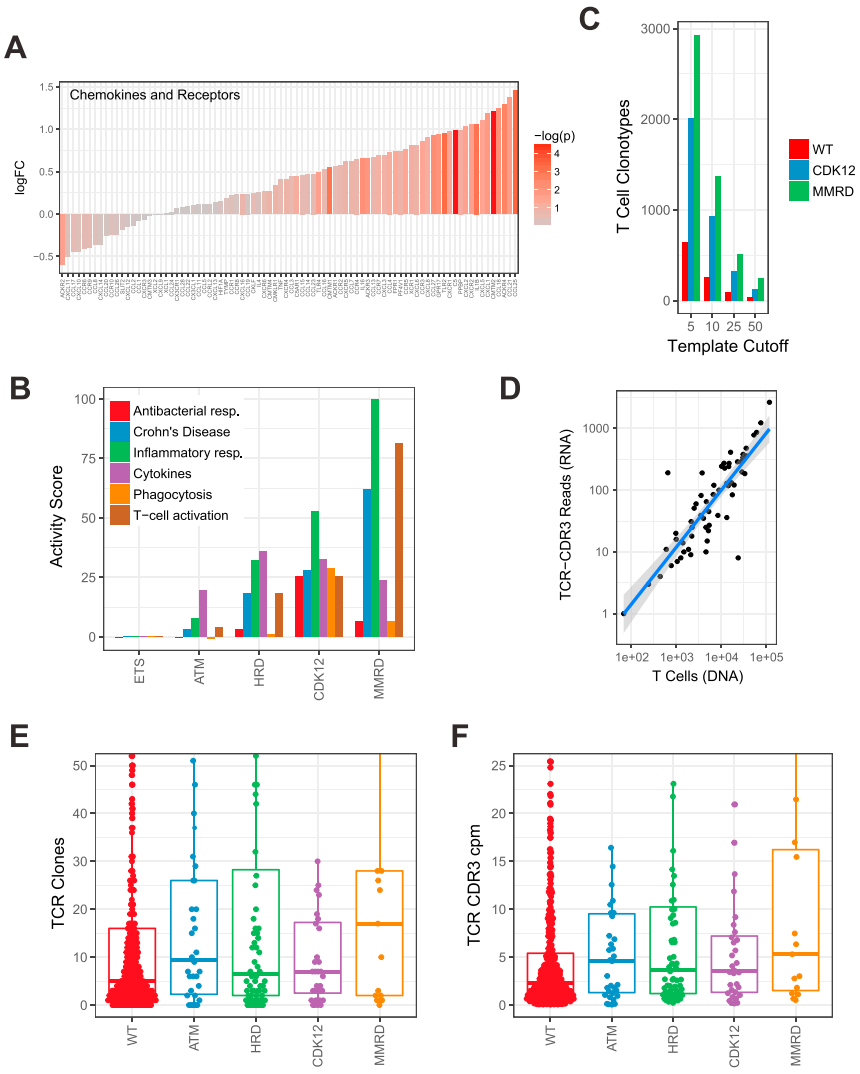


Figure S7. Immunophenotypic Characteristics of *CDK12* Mutant Tumors, Related to Figure 6

(A) Differential expression of chemokines and receptors in *CDK12* mutant tumors.

(B) Activity score for the most significant immune-related pathways across genetically unstable types of prostate cancer.

(C) Measurement of expanded T cell clones using different template cutoffs.

(D) RNA-seq and DNA-based (Adaptive) estimation of T cell infiltration in tumors. Total number of reads (RNA-seq) and estimated templates (Adaptive) is plotted for T cell CDR3 sequences.

(E) Number of distinct T cell clones (based on unique CDR3 sequences) from RNA-seq data.

(F) Number of T cell receptor CDR3 sequences (counts per million of aligned reads) from RNA-seq data.

LETTER

https://doi.org/10.1038/s41586-019-1347-4

Distinct structural classes of activating FOXA1 alterations in advanced prostate cancer

Abhijit Parolia[1,2,3,12], Marcin Cieslik[1,2,4,12], Shih-Chun Chu[1,2], Lanbo Xiao[1,2], Takahiro Ouchi[1,2], Yuping Zhang[1,2], Xiaoju Wang[1,2], Pankaj Vats[1,2], Xuhong Cao[1,2,5], Sethuramasundaram Pitchiaya[1,2], Fengyun Su[1,2], Rui Wang[1,2], Felix Y. Feng[6,7,8,9], Yi-Mi Wu[1,2], Robert J. Lonigro[1,2], Dan R. Robinson[1,2] & Arul M. Chinnaiyan[1,2,5,10,11]*

Forkhead box A1 (FOXA1) is a pioneer transcription factor that is essential for the normal development of several endoderm-derived organs, including the prostate gland[1,2]. FOXA1 is frequently mutated in hormone-receptor-driven prostate, breast, bladder and salivary-gland tumours[3–8]. However, it is unclear how FOXA1 alterations affect the development of cancer, and FOXA1 has previously been ascribed both tumour-suppressive[9–11] and oncogenic[12–14] roles. Here we assemble an aggregate cohort of 1,546 prostate cancers and show that FOXA1 alterations fall into three structural classes that diverge in clinical incidence and genetic co-alteration profiles, with a collective prevalence of 35%. Class-1 activating mutations originate in early prostate cancer without alterations in ETS or SPOP, selectively recur within the wing-2 region of the DNA-binding forkhead domain, enable enhanced chromatin mobility and binding frequency, and strongly transactivate a luminal androgen-receptor program of prostate oncogenesis. By contrast, class-2 activating mutations are acquired in metastatic prostate cancers, truncate the C-terminal domain of FOXA1, enable dominant chromatin binding by increasing DNA affinity and—through TLE3 inactivation—promote metastasis driven by the WNT pathway. Finally, class-3 genomic rearrangements are enriched in metastatic prostate cancers, consist of duplications and translocations within the *FOXA1* locus, and structurally reposition a conserved regulatory element—herein denoted FOXA1 mastermind (FOXMIND)—to drive overexpression of *FOXA1* or other oncogenes. Our study reaffirms the central role of FOXA1 in mediating oncogenesis driven by the androgen receptor, and provides mechanistic insights into how the classes of FOXA1 alteration promote the initiation and/or metastatic progression of prostate cancer. These results have direct implications for understanding the pathobiology of other hormone-receptor-driven cancers and rationalize the co-targeting of FOXA1 activity in therapeutic strategies.

FOXA1 independently binds to and de-compacts condensed chromatin to reveal the binding sites of partnering nuclear hormone receptors[15,16]. In prostate luminal epithelial cells, FOXA1 delimits tissue-specific enhancers[17], and reprograms androgen receptor (AR) activity in prostate cancer[14]. Accordingly, FOXA1 and AR are co-expressed in prostate cancer cells, in which FOXA1 activity is indispensable for cell survival and proliferation[14] (Extended Data Fig. 1a–i). It is notable that, in AR-dependent prostate cancer, *FOXA1* is the third most-highly mutated gene[4,5] and—as shown here—is located at one of the most-highly rearranged genomic loci. Counterintuitively, recent studies have suggested these alterations are inactivating[18,19] and have described FOXA1 as a tumour suppressor in AR-driven metastatic prostate cancer[9–11]. However, FOXA1 alterations have not yet been fully characterized or experimentally investigated in cancer.

To study these alterations, we first curated an aggregate cohort of prostate cancer that comprised 888 localized and 658 metastatic samples[4,5,8,20], of which 498 and 357, respectively, had matched RNA-sequencing (RNA-seq) data. Here, *FOXA1* mutations recurred at a frequency of 8–9% in primary disease, which increased to 12–13% in metastatic castration-resistant prostate cancer (mCRPC) (Fig. 1a, Extended Data Fig. 1j). RNA-seq calls of structural variants revealed a high prevalence (Fig. 1b, Supplementary Table 1) and density (Extended Data Fig. 1k) of rearrangements within the *FOXA1* locus. The presence of structural variants was confirmed by whole-exome and whole-genome sequencing (Extended Data Fig. 1l, m, Supplementary Tables 2, 3). Overall, we estimated the recurrence of *FOXA1* locus rearrangements to be 20–30% in mCRPC (Extended Data Fig. 1n). All *FOXA1* mutations were heterozygous and *FOXA1* itself was copy-amplified in over 50% of cases with no biallelic deletions (Extended Data Fig. 2a, b). We also found a stagewise increase in FOXA1 expression in prostate cancer (Extended Data Fig. 2c, Supplementary Discussion).

When we mapped mutations onto the protein domains of FOXA1, we found two structural patterns: (1) missense and in-frame insertion and deletion (indel) mutations were clustered at the C-terminal end of the forkhead domain (FKHD); and (2) truncating frameshift mutations were restricted to the C-terminal half of the protein (Fig. 1c). *FOXA1* structural variants predominantly consisted of tandem duplications and translocations, which clustered in close proximity to the *FOXA1* gene without disrupting its coding sequence (Fig. 1d). Thus, we categorized FOXA1 alterations into three structural classes: class 1, which comprises all the mutations within the FKHD; class 2, which comprises mutations in the C-terminal end after the FKHD; and class 3, which comprises structural variants within the *FOXA1* locus (Fig. 1c, d, Extended Data Fig. 2d). We also found similar classes of FOXA1 alterations in breast cancer (Extended Data Fig. 2e, f).

We found that the majority of *FOXA1* mutations in primary prostate cancer belonged to class 1, which showed no enrichment in the metastatic disease (Fig. 1e). Conversely, class-2 mutations were significantly enriched in metastatic prostate cancer; in the rare primary cases with class-2 mutations, the mutant allele was detected at sub-clonal frequencies (Fig. 1e, f, Extended Data Fig. 2g, h). We found no cases that possessed both class-1 and class-2 mutations. Class-3 structural variants were also significantly enriched in mCRPC (odds ratio = 3.46) (Fig. 1g). Overall, we found the cumulative frequency of FOXA1 alterations to be over 34% in mCRPC (Fig. 1h). Assessment of concurrent alterations revealed that class-1 mutations are mutually exclusive with other primary events (for example, ETS fusions) (odds ratio = 0.078), whereas class-2-mutant mCRPC are enriched for *RB1* deletions (odds ratio = 4.17) (Extended Data Fig. 2i, j). Both mutational classes were further enriched for alterations in DNA repair, mismatch repair and

[1]Michigan Center for Translational Pathology, University of Michigan, Ann Arbor, MI, USA. [2]Department of Pathology, University of Michigan, Ann Arbor, MI, USA. [3]Molecular and Cellular Pathology Program, University of Michigan, Ann Arbor, MI, USA. [4]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA. [5]Howard Hughes Medical Institute, University of Michigan, Ann Arbor, MI, USA. [6]Helen Diller Family Comprehensive Cancer Center, University of California at San Francisco, San Francisco, CA, USA. [7]Department of Radiation Oncology, University of California at San Francisco, San Francisco, CA, USA. [8]Department of Urology, University of California at San Francisco, San Francisco, CA, USA. [9]Department of Medicine, University of California at San Francisco, San Francisco, CA, USA. [10]Department of Urology, University of Michigan, Ann Arbor, MI, USA. [11]Rogel Cancer Center, University of Michigan, Ann Arbor, MI, USA. [12]These authors contributed equally: Abhijit Parolia, Marcin Cieslik. *e-mail: arul@umich.edu

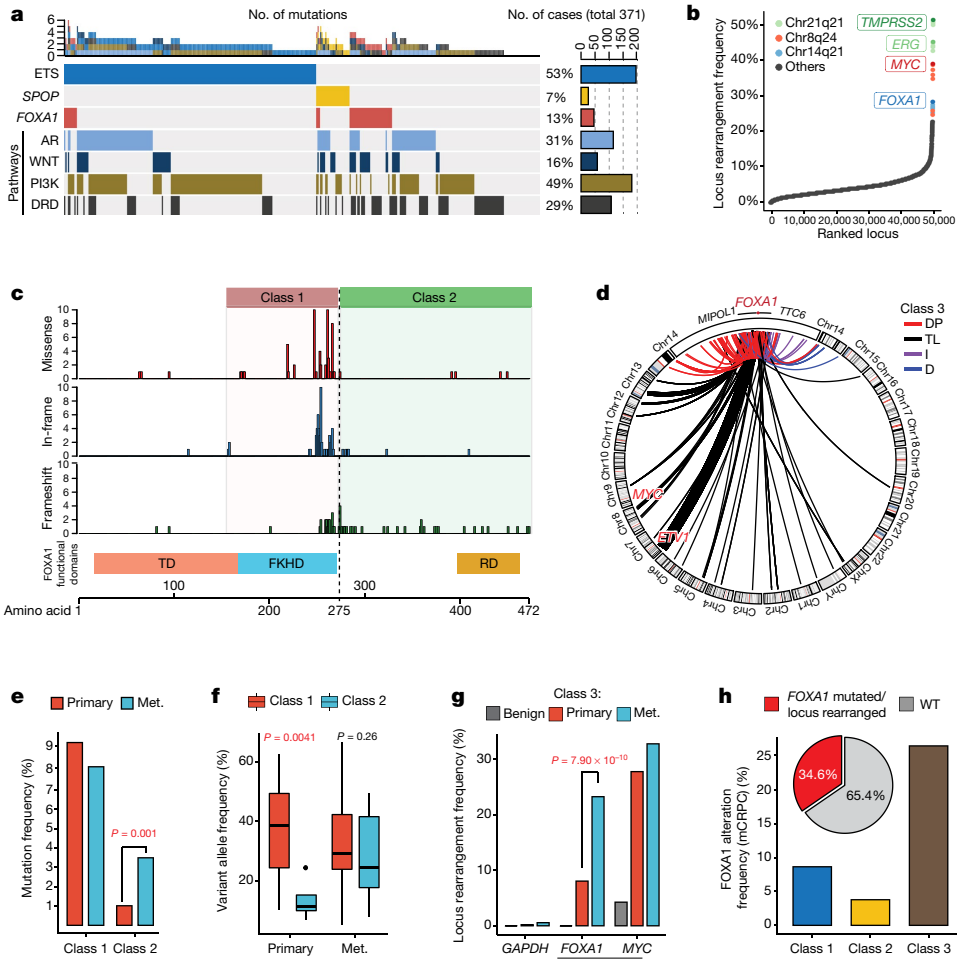


Fig. 1 | Structural classes of FOXA1 alterations. a, *FOXA1* mutations and key alterations in mCRPC. Alterations in ETS, AR, WNT, PI3K and DNA repair (DRD) were aggregated at the pathway or group level. **b**, Locus-level recurrence of RNA-seq structural variations. **c**, Structural classification of *FOXA1* mutations. TD, transactivation domain; RD, regulatory domain. **d**, Structural classification of *FOXA1* locus rearrangements. DP, tandem duplications; TL, translocations; I, inversions; D, deletions. **e**, Frequency of *FOXA1* mutational classes by prostate cancer stage ($n = 888$ primary, 658 metastatic (met.)) (two-sided Fisher's exact test). **f**, Variant allele frequency by stage and class (two-sided *t*-test). Box plot centre, median; box, quartiles 1–3, whiskers, quartiles 1–3 ± 1.5 × interquartile range (IQR). **g**, Locus-level recurrence of structural variants based on RNA-seq by prostate cancer stage (two-sided Fisher's exact test). **h**, Integrated (RNA-seq and whole-exome sequencing) recurrence of FOXA1-alteration classes in mCRPC (Stand Up 2 Cancer and Michigan Center for Translational Pathology (MCTP) cohort, $n = 370$).

WNT signalling pathways (Extended Data Fig. 2i, k), and had higher levels of expression of *FOXA1* mRNA relative to the wild-type cases (Extended Data Fig. 2l). Together, these data suggest that class-1 mutations emerge in localized prostate cancer, whereas class-2 and class-3 mutations are acquired or enriched, respectively, in the course of disease progression.

Class-1 mutations consist of missense and in-frame indels that cluster at the C-terminal edge of the winged-helix DNA-binding FKHD. The majority of the class-1 mutations were located either within the wing-2 region (residues 247–269) or a 3D hotspot that spatially protrudes towards wing 2[21] (Fig. 2a, b, Extended Data Fig. 3a, b). Notably, these mutations did not alter FKHD residues that make base-specific interactions with the DNA[22,23] (Fig. 2a, Extended Data Fig. 3c). In FOXA proteins, wing-2 residues make base-independent (that is, non-specific) contacts with the DNA backbone[23,24] that reportedly impede its nuclear movement[24]. Thus, we hypothesized that class-1 mutants with altered wing-2 regions would display faster nuclear mobility.

We cloned representative class-1 mutants of FOXA1: I176M (mutation of the 3D hotspot), R261G (missense) and R265–Q271del (in-frame deletion), all of which retained nuclear localization (Extended Data Fig. 3d). In fluorescence recovery after photobleaching (FRAP) assays, we found class-1 mutants had 5–6× faster nuclear mobility irrespective of the mutation type (Fig. 2c, d, Extended Data Fig. 3e, g). By contrast, class-2 mutants with intact wing-2 regions were sluggish in their nuclear movement (Fig. 2d, Extended Data Fig. 3f, g). Using single particle tracking, we verified that class-1 mutants have a higher overall rate of nuclear diffusion, with 3–4-fold fewer slow particles and shorter chromatin dwell times (Extended Data Fig. 3h, i). In chromatin immunoprecipitation with parallel DNA sequencing (ChIP–seq) assays, we found that ectopically expressed class-1 mutants in HEK293 cells bind DNA at the consensus FOXA1 motif (Extended Data Fig. 3j, k). In prostate cancer cells, the class-1 cistrome entirely overlapped with wild-type binding sites, with similar enrichment for FOXA1 and AR cofactor motifs, AR-binding sites and genomic distribution (Extended Data Fig. 3l–s). Furthermore, in growth rescue experiments using untranslated-region-specific small interfering (si) RNAs that targeted the endogenous *FOXA1* transcript, we found that exogenous class-1 mutants fully compensated for the wild-type protein (Extended Data Fig. 4a).

Next, we asked how class-1 mutations affect AR signalling. Similar to wild-type FOXA1, both class-1 and class-2 mutants interacted with the AR signalling complex (Extended Data Fig. 4b–d). In reporter assays, class-1 mutants induced 3–6-fold higher activation of AR signalling (Fig. 2e), which was evident even under stimulation with castrate levels of androgen or treatment with enzalutamide (Extended Data Fig. 4e, f). In parallel assays, class-2 mutants showed no differences relative to wild-type FOXA1 (Fig. 2e). Transcriptomic analyses of class-1 tumours

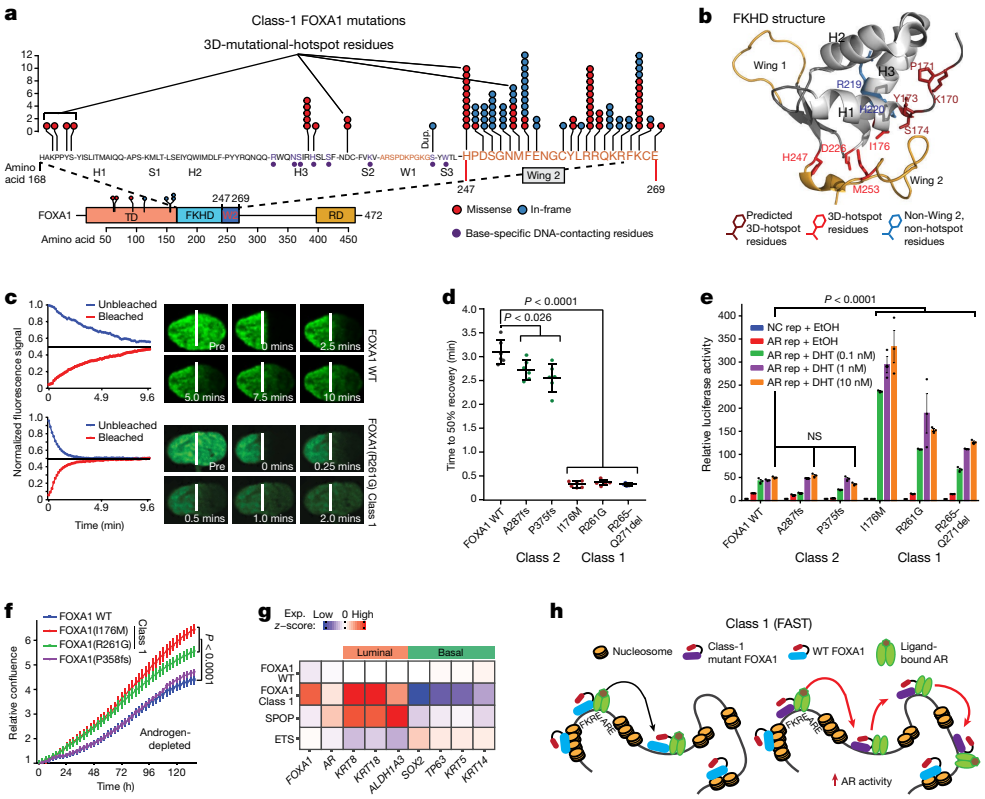


Fig. 2 | Functional characterization of class-1 mutations of *FOXA1*. **a**, Distribution of class-1 mutations on the protein map of FOXA1 functional domains and FKHD secondary structures. Dup., duplication. **b**, Crystal structure of the FKHD with visualization of non-wing-2 (that is, outside of amino acids 247–269) mutations. Mutations in the 3D hotspot are in red. **c**, FRAP kinetic plots (left) and representative time-lapse images from pre-bleaching to the equilibrated state (right; $n = 6$ biological replicates). Images are uniformly brightened for signal visualization. WT, wild type. **d**, FRAP durations until 50% recovery ($n = 6$ nuclei per variant). **e**, Negative control (NC) or AR reporter (rep) activity with overexpression of FOXA1 variants and dihydrotestosterone (DHT) stimulation ($n = 3$ biological replicates). **f**, Growth (IncuCyte) of 22RV1 cells that overexpress FOXA1 variants in androgen-depleted medium ($n = 5$ biological replicates). In **d**–**f**, mean ± s.e.m. is shown, and *P* values are from two-way analysis of variance (ANOVA) and Tukey's test. **g**, Relative expression of luminal and basal markers in class-1 ($n = 38$) tumours compared with wild-type ($n = 457$), SPOP ($n = 48$) and ETS ($n = 243$) primary prostate cancer tumours. **h**, Class-1 model. Wing-2-disrupted FOXA1 shows increased chromatin mobility and chromatin sampling frequency, which results in stronger transcriptional activation of oncogenic AR signalling. FKRE, forkhead-responsive element; ARE, androgen-responsive element.

from patients revealed the activation of hyperproliferative and pro-tumorigenesis pathways, and further enrichment of primary prostate cancer genes (Extended Data Fig. 4g–i). Notably, AR was predicted[25] to be the driver transcription factor for class-1 upregulated genes, which we experimentally confirmed for several targets (Extended Data Fig. 4j–l). Concordantly, overexpression of class-1 mutants in 22RV1 cells increased growth in androgen-depleted medium (Fig. 2f) but not in androgen-supplemented medium, and rescued proliferation upon treatment with enzalutamide (Extended Data Fig. 4m, n). For class-1 downregulated genes, the basal transcription factors TP63 and SOX2 were predicted to be transcriptional drivers (Extended Data Fig. 4j). Consistently, in class-1 specimens from patients, both of these transcription factors were significantly downregulated, with a concomitant downregulation of basal, and upregulation of luminal, markers (Fig. 2g, Extended Data Fig. 4o, p). In addition, class-1 tumours had a higher AR transcriptional signature, and a lower neuroendocrine transcriptional signature (Extended Data Fig. 4q). Together, these data suggest that class-1 mutations that alter the wing-2 region increase the nuclear speed and genome-scanning efficiency of FOXA1 without affecting its DNA sequence specificity (Supplementary Discussion), and drive a luminal AR program of prostate oncogenesis (Fig. 2h).

Class-2 mutations consist of frameshifting alterations that truncate the C-terminal regulatory domain of FOXA1 (Fig. 3a). Thus, we characterized the class-2 cistrome by using N-terminal and C-terminal antibodies, with the C-terminal antibody binding exclusively to wild-type FOXA1 (Extended Data Fig. 5a, b). Notably, mCRPC-derived LAPC4 cells endogenously contained a *FOXA1* class-2 mutation (that is, a frameshift at amino acid P358 (P358fs)), and both wild-type and mutant variants interacted with the AR complex (Extended Data Fig. 5c–f). However, in ChIP–seq assays, only the N-terminal antibody detected FOXA1 binding to the DNA. By contrast, N-terminal and C-terminal FOXA1 cistromes substantially overlapped in wild-type prostate cancer cells (Fig. 3b, Extended Data Fig. 5g–i). Even with 13-fold overexpression of wild-type FOXA1 in LAPC4 cells, the endogenous class-2 mutant retained its binding dominance (Fig. 3b, Extended Data Fig. 5j, k). Conversely, overexpression of the FOXA1(P358fs) mutant in LNCaP cells markedly diminished the endogenous wild-type cistrome (Fig. 3b). In in vitro assays, class-2 mutants showed markedly stronger binding to the *KLK3* enhancer element (Fig. 3c, Extended Data Fig. 6a–d), and biolayer interferometry confirmed that the FOXA1(P358fs) mutant has an approximately fivefold-higher DNA-binding affinity (Extended Data Fig. 6e). In CRISPR-engineered class-2-mutant 22RV1 clones (Extended Data Fig. 6f, g), FOXA1 ChIP–seq data reaffirmed the cistromic dominance of class-2 mutants (Fig. 3d). Knockdown of either mutant FOXA1 or AR in 22RV1 or LNCaP class-2 CRISPR clones significantly attenuated proliferation (Fig. 3e, Extended Data Fig. 6h, i). Consistently, in rescue experiments, the FOXA1(P358fs) mutant fully compensated for the loss of wild-type FOXA1 (Extended Data Fig. 4a).

The class-2 cistrome was considerably larger than the wild-type cistrome (Extended Data Fig. 6j–l), and the acquired sites were enriched for the CTCF motif and distal regulatory regions (Extended Data Fig. 7a–e, Supplementary Discussion). In transcriptomic and motif analyses of the class-2 clones, LEF and TCF were predicted as

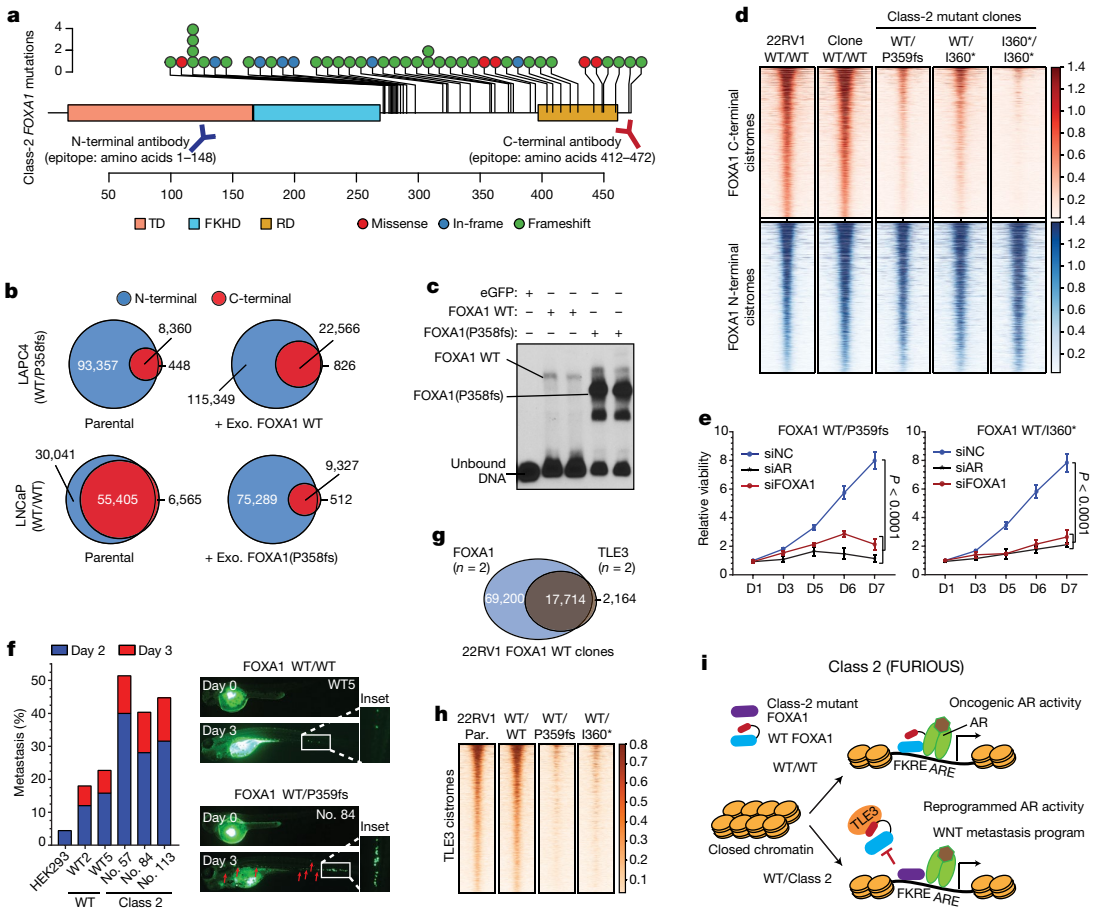


Fig. 3 | Functional characterization of class-2 mutations of *FOXA1*. **a**, Class-2 mutations and antibody epitopes on the protein map of FOXA1. **b**, N-terminal and C-terminal FOXA1 cistromes in FOXA1 wild-type (*FOXA1*$^{WT/WT}$ (WT/WT)) or mutant (*FOXA1*$^{WT/P358fs}$ (WT/P358fs)) prostate cancer cells that are untreated (left) or have exogenous (exo.) overexpression of FOXA1 variants (right). **c**, Electromobility shift of FOXA1 variants bound to the *KLK3* enhancer ($n = 3$ biological replicates). For gel source data, see Supplementary Fig. 1. **d**, FOXA1 ChIP–seq read-density heat maps in independent class-2-mutant 22RV1 CRISPR clones (*FOXA1*$^{WT/P359fs}$ (WT/P359fs), *FOXA1*$^{WT/I360*}$ (WT/I360*) and *FOXA1*^{I360*I360*} (I360*/I360*)). **e**, Growth of class-2-mutant 22RV1 clones treated with non-targeting (siNC), *AR*- or *FOXA1*-targeting siRNAs ($n = 5$ biological replicates; two-way ANOVA and Tukey's test). Mean ± s.e.m. is shown. D, day. **f**, Left, metastasis frequency in zebrafish embryos injected with HEK293 (negative control), wild-type 22RV1 clones or class-2-mutant 22RV1 clones ($n \geq 30$ embryos per group). Right, representative images of embryos, showing the disseminated prostate cancer cells. **g**, Overlap of wild-type FOXA1- and TLE3-binding sites in 22RV1 CRISPR clones ($n = 2$ biological replicates each). **h**, TLE3 ChIP–seq read-density heat maps in 22RV1 parental (par.) cells and distinct FOXA1 wild-type and class-2-mutant 22RV1 CRISPR clones. **i**, Class-2 model. Truncated FOXA1 shows dominant chromatin binding and displaces wild-type FOXA1 and TLE3 from the chromatin, which results in increased WNT signalling.

the top regulatory transcription factors for the upregulated genes (Extended Data Fig. 7g, h). The LEF–TCF complex is the primary nuclear effector of WNT signalling and remains inactive until it is bound by β-catenin[26]. Consistently, we found a marked accumulation of transcriptionally active β-catenin—that is, non-phosphorylated at S31, S37 and T41—in distinct mutant clones, as well as a concomitant increase in the expression of the WNT targets LEF1 and AXIN2 (Extended Data Fig. 7i, j). Class-2 clones showed 2–3-fold higher invasiveness in Boyden chamber assays (Extended Data Fig. 7k, l), and a higher rate and extent of metastatic dissemination in zebrafish embryos (Fig. 3f, Extended Data Fig. 7m). In these assays, class-1 mutant cells showed no differences relative to wild-type cells (Extended Data Fig. 7n). Furthermore, treatment with the WNT inhibitor XAV939 completely abrogated the class-2 invasive phenotype (Extended Data Fig. 7o). Investigating the mechanism that underlies this invasiveness, we found that FOXA1 transcriptionally activates and—through its C-terminal domain—recruits TLE3 (a bona fide WNT co-repressor[27]) to the chromatin (Extended Data Fig. 8a–e). Class-2 mutants had lost this interaction, which led to the untethering of TLE3 from chromatin and downstream activation of WNT signalling (Fig. 3g, h, Extended Data Fig. 8e–k, Supplementary Discussion). Together, these data suggest that class-2 mutations confer cistromic dominance and abolish TLE3-mediated repression of the WNT program of metastasis (Fig. 3i).

Class-3 rearrangements occur within the *PAX9* and *FOXA1* locus that is linearly conserved across the deuterostome superphylum[28] (Fig. 4a). Notably, almost all break ends were clustered within the *FOXA1* topologically associating domain (Extended Data Fig. 9a). We found that the genes located within the *FOXA1* topologically associating domain had the highest expression in the normal prostate, and the non-coding RP11-356O9.1 transcript had a prostate-specific expression (Extended Data Fig. 9b). Furthermore, in patient tumours, expression of RP11-356O9.1 was strongly correlated with *FOXA1* and *TTC6* expression (Extended Data Fig. 9c). Thus, to identify prostate-specific enhancers of the *FOXA1* topologically associating domain, we performed the assay for transposase-accessible chromatin using sequencing (ATAC-seq) and interrogated chromatin features in AR^{+} and AR^{-} prostate cells. Notably, a CTCF-bound intronic site in RP11-356O9.1 (hereafter denoted as *FOXMIND*) and a site within the 3′ untranslated region of *MIPOL1* were accessible and marked with active enhancer modifications only in AR^{+}FOXA1^{+} prostate cancer cells (Fig. 4b, Extended Data Fig. 9d). This strongly suggested that these conserved sites function as enhancer elements. Consistently, CRISPR knockout of these loci in VCaP cells led to a significant decrease in the expression of *FOXA1*

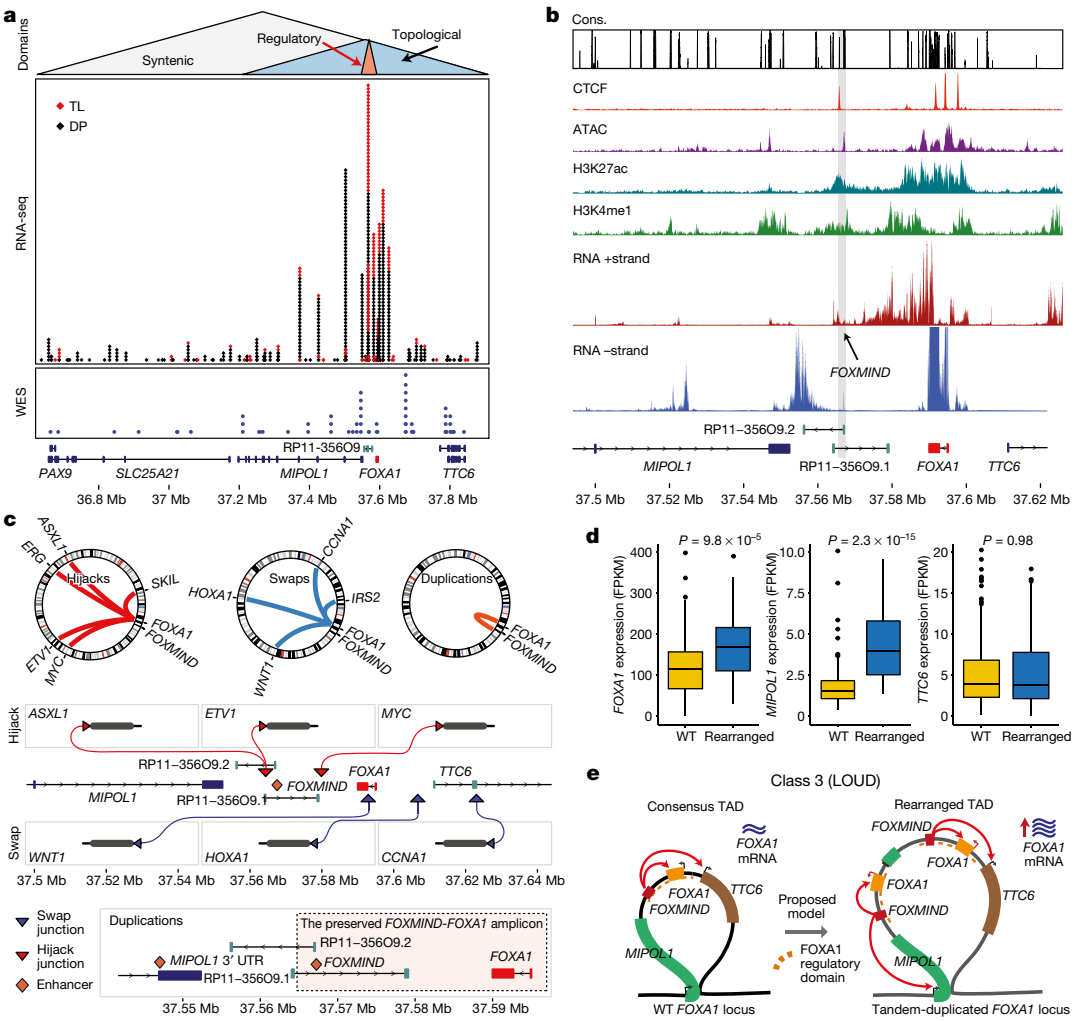


Fig. 4 | Genomic characterization of class-3 rearrangements of the *FOXA1* locus. a, Break ends in relation to the *FOXA1* syntenic, topological and regulatory domains. WES, whole-exome sequencing. **b**, Representative functional genomic tracks at the *FOXA1* locus. Base-level conservation (cons.), DNA accessibility (ATAC), enhancer-associated histone modifications (H3K27me1 and H3K27Ac), CTCF chromatin binding and stranded RNA-seq read densities are visualized. The *FOXMIND* enhancer is highlighted. **c**, Structural patterns of translocations and duplications. Hijacks occur between *FOXMIND* and *FOXA1*; swaps occur upstream of *FOXA1*. Duplications amplify the highlighted *FOXMIND-FOXA1* regulatory domain. **d**, Transcriptional changes in the *FOXA1*, *MIPOL1* and *TTC6* genes in wild-type ($n = 320$) and rearranged ($n = 50$) cases (two-sided *t*-test). Box plot centre, median; box, quartiles 1–3; whiskers, quartiles 1–3 ± 1.5 × IQR. FPKM, fragments per kilobase of transcript per million mapped reads. **e**, Class-3 model. Tandem duplications within the *FOXA1* topologically associating domain (TAD) amplify *FOXMIND* to drive overexpression of *FOXA1*.

and *TTC6*—but not of *MIPOL1*, which has its promoter outside of the *FOXA1* topologically associating domain (Extended Data Fig. 9d, e).

We found that translocations were largely within a 50-kb region between *FOXA1* and the 3′ untranslated region of *MIPOL1*, whereas break-end junctions from duplications mostly flanked the *FOXMIND-FOXA1* region (Fig. 4a, Extended Data Fig. 9f). For translocations, we delineated two patterns: (1) the hijacking of the *FOXMIND* enhancer; and (2) insertions upstream of the *FOXA1* promoter (Fig. 4c). The first pattern subsumes previously reported in-frame fusion genes that involve RP11-356O9.1, *ETV1*[29] and *SKIL*[30], as well as a newly reported *ASXL1* fusion (Supplementary Table 4). The second pattern inserts an oncogene (such as *CCNA1*) upstream of *FOXA1* (Fig. 4c). Notably, both mechanisms resulted in outlier expression of the translocated gene (Extended Data Fig. 9g). For duplications, which constitute 70% of all rearranged cases, we found *FOXMIND* and *FOXA1* to be co-amplified in 89% of the rearranged cases and never separated (Fig. 4c, bottom, Extended Data Fig. 9h), thus preserving the *FOXMIND-FOXA1* regulatory domain.

Next, while assessing the transcriptional effect of duplications, we found that levels of *FOXA1* mRNA were poorly correlated with copy number (Extended Data Fig. 10a), but highly sensitive to focal structural variants. Tandem duplications (ascertained at the RNA and DNA levels) significantly increased expression of *FOXA1* and *MIPOL1*, but not of *TTC6* (Fig. 4d). Translocations resulted in a modest decrease in expression levels of *FOXA1* (Extended Data Fig. 10b), despite a significant co-occurrence with tandem duplications (odds ratio = 3.89, Extended Data Fig. 10c). To investigate this further, we carried out haplotype-resolved, linked-read sequencing of MDA-PCA-2b cells, which contain a translocation of *FOXMIND* and *ETV1*. Here, *ETV1* translocation was accompanied by a focal tandem duplication in the non-translocated *FOXA1* allele (Extended Data Fig. 10d). The translocated *FOXA1* allele was inactivated, which resulted in monoallelic transcription (Extended Data Fig. 10e) without a net loss in *FOXA1* expression (266 fragments per kilobase of transcript per million mapped reads, 95th percentile in mCRPC). By contrast, RP11-356O9.1 retained biallelic expression (Extended Data Fig. 10f). In LNCaP cells, which also contain an *ETV1* translocation into the *FOXA1* locus, deletion of *FOXMIND* caused a significant reduction in *ETV1* expression (Extended Data Fig. 10g). Thus, translocations result in the loss of *FOXA1* expression from the allele *in cis*, which is rescued by tandem duplications of the allele *in trans*. Altogether, we propose a coalescent model in which class-3 structural variants duplicate or reposition

FOXMIND to drive overexpression of *FOXA1* or other oncogenes (Fig. 4e).

In summary, we identify three structural classes of FOXA1 alterations that differ in genetic associations and oncogenic mechanisms. We establish *FOXA1* as a principal oncogene in AR-dependent prostate cancer that is altered in 34.6% of mCRPC. Given the unique pathogenic features of the three classes, we have named them the 'FAST' (class-1), 'FURIOUS' (class-2) and 'LOUD' (class-3) alterations of FOXA1 (Figs. 2h, 3i, 4e, Supplementary Table 5, Supplementary Discussion). Structurally equivalent FOXA1 alterations are also found in other hormone-receptor-driven cancers, thus positioning FOXA1 as a promising target for therapeutic strategies in these malignancies.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1347-4.

Received: 22 May 2018; Accepted: 3 June 2019;
Published online: 26 June 2019

1. Gao, N. et al. Forkhead box A1 regulates prostate ductal morphogenesis and promotes epithelial cell maturation. *Development* **132**, 3431–3443 (2005).
2. Friedman, J. R. & Kaestner, K. H. The Foxa family of transcription factors in development and metabolism. *Cell. Mol. Life Sci.* **63**, 2317–2328 (2006).
3. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* **507**, 315–322 (2014).
4. Robinson, D. et al. Integrative clinical genomics of advanced prostate cancer. *Cell* **161**, 1215–1228 (2015).
5. Cancer Genome Atlas Research Network. The molecular taxonomy of primary prostate cancer. *Cell* **163**, 1011–1025 (2015).
6. Ciriello, G. et al. Comprehensive molecular portraits of invasive lobular breast cancer. *Cell* **163**, 506–519 (2015).
7. Dalin, M. G. et al. Comprehensive molecular characterization of salivary duct carcinoma reveals actionable targets and similarity to apocrine breast cancer. *Clin. Cancer Res.* **22**, 4623–4633 (2016).
8. Zehir, A. et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat. Med.* **23**, 703–713 (2017).
9. Jin, H.-J., Zhao, J. C., Ogden, I., Bergan, R. C. & Yu, J. Androgen receptor-independent function of FoxA1 in prostate cancer metastasis. *Cancer Res.* **73**, 3725–3736 (2013).
10. Jin, H.-J., Zhao, J. C., Wu, L., Kim, J. & Yu, J. Cooperativity and equilibrium with FOXA1 define the androgen receptor transcriptional program. *Nat. Commun.* **5**, 3972 (2014).
11. Song, B. et al. Targeting FOXA1-mediated repression of TGF-β signaling suppresses castration-resistant prostate cancer progression. *J. Clin. Invest.* **129**, 156–582 (2019).
12. Robinson, J. L. L. et al. Androgen receptor driven transcription in molecular apocrine breast cancer is mediated by FoxA1. *EMBO J.* **30**, 3019–3027 (2011).
13. Robinson, J. L. L. et al. Elevated levels of FOXA1 facilitate androgen receptor chromatin binding resulting in a CRPC-like phenotype. *Oncogene* **33**, 5666–5674 (2014).
14. Pomerantz, M. M. et al. The androgen receptor cistrome is extensively reprogrammed in human prostate tumorigenesis. *Nat. Genet.* **47**, 1346–1351 (2015).
15. Cirillo, L. A. et al. Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. *Mol. Cell* **9**, 279–289 (2002).
16. Iwafuchi-Doi, M. et al. The pioneer transcription factor FoxA maintains an accessible nucleosome configuration at enhancers for tissue-specific gene activation. *Mol. Cell* **62**, 79–91 (2016).
17. Lupien, M. et al. FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell* **132**, 958–970 (2008).
18. Barbieri, C. E. et al. Exome sequencing identifies recurrent *SPOP*, *FOXA1* and *MED12* mutations in prostate cancer. *Nat. Genet.* **44**, 685–689 (2012).
19. Yang, Y. A. & Yu, J. Current perspectives on FOXA1 regulation of androgen receptor signaling and prostate cancer. *Genes Dis.* **2**, 144–151 (2015).
20. Grasso, C. S. et al. The mutational landscape of lethal castration-resistant prostate cancer. *Nature* **487**, 239–243 (2012).
21. Gao, J. et al. 3D clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets. *Genome Med.* **9**, 4 (2017).
22. Clark, K. L., Halay, E. D., Lai, E. & Burley, S. K. Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. *Nature* **364**, 412–420 (1993).
23. Li, J. et al. Structure of the forkhead domain of FOXA2 bound to a complete DNA consensus site. *Biochemistry* **56**, 3745–3753 (2017).
24. Sekiya, T., Muthurajan, U. M., Luger, K., Tulin, A. V. & Zaret, K. S. Nucleosome-binding affinity as a primary determinant of the nuclear mobility of the pioneer transcription factor FoxA. *Genes Dev.* **23**, 804–809 (2009).
25. Wang, Z. et al. BART: a transcription factor prediction tool with query gene sets or epigenomic profiles. *Bioinformatics* **34**, 2867–2869 (2018).
26. Behrens, J. et al. Functional interaction of β-catenin with the transcription factor LEF-1. *Nature* **382**, 638–642 (1996).
27. Daniels, D. L. & Weis, W. I. β-catenin directly displaces Groucho/TLE repressors from Tcf/Lef in Wnt-mediated transcription activation. *Nat. Struct. Mol. Biol.* **12**, 364–371 (2005).
28. Wang, W., Zhong, J., Su, B., Zhou, Y. & Wang, Y.-Q. Comparison of *Pax1/9* locus reveals 500-Myr-old syntenic block and evolutionary conserved noncoding regions. *Mol. Biol. Evol.* **24**, 784–791 (2007).
29. Tomlins, S. A. et al. Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. *Nature* **448**, 595–599 (2007).
30. Annala, M. et al. Recurrent SKIL-activating rearrangements in ETS-negative prostate cancer. *Oncotarget* **6**, 6235–6250 (2015).

METHODS

Cell culture. Most cell lines were originally purchased from the American Type Culture Collection (ATCC) and were cultured as per standard ATCC protocols. LNCaP-AR and LAPC4 cells were gifts from the laboratory of C. Sawyers (Memorial Sloan Kettering Cancer Center). Unless otherwise stated, for all the experiments LNCaP, PNT2, LNCaP-AR, C42B, 22RV1, DU145 and PC3 cells were grown in the RPMI 1640 medium (Gibco) and VCaP cells in the DMEM with Glutamax (Gibco) medium supplemented with 10% full bovine serum (FBS; Invitrogen). LAPC4 cells were grown in IMEM (Gibco) supplemented with 15% FBS and 1 nM of R1881. For the immortalized normal prostate cells: RWPE1 cells were grown in keratinocyte medium with regular supplements (Lonza); PNT2 cells were grown in RPMI medium with 10% FBS. HEK293 cells were grown in DMEM (Gibco) medium with 10% FBS. All cells were grown in a humidified 5% CO_2 incubator at 37 °C. All cell lines were tested once a fortnight to be free of mycoplasma contamination and genotyped every month at the University of Michigan Sequencing Core using Profiler Plus (Applied Biosystems) and compared with corresponding short tandem repeat profiles in the ATCC database to authenticate their identity in culture between passages and experiments.

Antibodies. For immunoblotting, the following antibodies were used: FOXA1 N-terminal (Cell Signaling Technologies: 58613S; Sigma-Aldrich: SAB2100835); FOXA1 C-terminal (Thermo Fisher Scientific: PA5-27157; Abcam: ab23738); AR (Millipore: 06-680); LSD1 (Cell Signaling Technologies: 2139S); vinculin (Sigma Aldrich: V9131); H3 (Cell Signaling Technologies: 3638S); GAPDH (Cell Signaling Technologies: 3683); β-actin (Sigma Aldrich: A5316); β-catenin (Cell Signaling Technologies: 8480S); vimentin (Cell Signaling Technologies: 5741S); phospho(S33/S37/T41)-β-catenin (Cell Signaling Technologies: 8814S); LEF1 (Cell Signaling Technologies: 2230S); AXIN2 (Abcam: ab32197); and TLE3 (Proteintech: 11372-1-AP).

For co-immunoprecipitation and ChIP–seq experiments, the following antibodies were used: FOXA1 N-terminal (Cell Signaling Technologies: 58613S); FOXA1 C-terminal (Thermo Fisher Scientific: PA5-27157); AR (Millipore: 06-680); V5 tag (R960-25); and TLE3 (Proteintech: 11372-1-AP).

Immunoblotting and nuclear co-immunoprecipitation. Cell lysates were prepared using the RIPA lysis buffer (Thermo Fisher Scientific; cat. no. 89900) and denatured in the complete NuPage 1× LDS/reducing agent buffer (Invitrogen) with 10 min heating at 70 °C. Between 10 and 25 μg of total protein was loaded per well, separated on 4–12% SDS polyacrylamide gels (Novex) and transferred onto 0.45-μm nitrocellulose membrane (Thermo Fisher Scientific; cat. no. 88018) using a semi-dry transfer system (Trans-blot Turbo System; BioRad) at 25 V for 1 h. The membrane was incubated for 1 h in blocking buffer (Tris-buffered saline, 0.1% Tween (TBS-T), 5% non-fat dry milk) and incubated overnight at 4 °C with primary antibodies. When samples were run on multiple gels for an experiment, multiple loading control proteins (GAPDH, β-actin, total H3 and vinculin) were probed on each membrane separately. Host-species-matched secondary antibodies conjugated to horseradish peroxidase (HRP; BioRad) were used at 1/20,000 dilution to detect primary antibodies and blots were developed using enhanced chemiluminescence (ECL Prime, Thermo Fisher Scientific) following the manufacturer's protocol.

For nuclear co-immunoprecipitation assays, 8–10 million cells ectopically overexpressing different V5-tagged FOXA1 variants and wild-type AR (or TLE3) were fractionated to isolate intact nuclei using the NE-PER kit reagents (Thermo Fisher Scientific; cat. no. 78835) and lysed in the complete IP lysis buffer (Thermo Fisher Scientific; cat. no. 87788). Nuclear lysates were incubated for 2 h at 4 °C with 30 μl of magnetic protein-G Dynabeads (Thermo Fisher Scientific; cat. no. 10004D) for pre-clearing. A fraction of the pre-cleared lysate was saved as input and the remainder was incubated overnight (12–16 h) with 10 μg of target protein antibody at 4 °C with gentle mixing. Next day, 50 μl of Dynabeads protein-G beads were added to the lysate–antibody mixture and incubated for 2 h at 4 °C. Beads were washed three times with IP buffer (150 nM NaCl; Thermo Fisher Scientific) and directly boiled in 1× NuPage LDS/reducing agent buffer (ThermoFisher Scientific; cat. no. NP0007 and NP0009) to elute and denature the precipitated proteins. These samples were then immunoblotted as described above with the exception of using protein A-HRP secondary (GE HealthCare; cat. no. NA9120-1ML) antibody for detection.

RNA extraction and quantitative polymerase chain reaction. Total RNA was extracted using the the miRNeasy Mini Kit (Qaigen), with the inclusion of the on-column genomic DNA digestion step using the RNase-free DNase Kit (Qaigen), following the standard protocols. RNA was quantified using the NanoDrop 2000 Spectrophotometer (ThermoFisher Scientific) and 1 μg of total RNA was used for complementary DNA (cDNA) synthesis using the SuperScript III Reverse Transcriptase enzyme (Thermo Fisher Scientific) following the manufacturer's instructions. Twenty nanograms of cDNA was input per polymerase chain reaction (PCR) using the FAST SYBR Green Universal Master Mix (Thermo Fisher Scientific) and every sample was quantified in triplicate. Gene expression was calculated relative to *GAPDH* and *HPRT1* (loading control) using the $\Delta\Delta C_t$ method and normalized to the control group for graphing. Quantitative PCR (qPCR) primers were designed using the Primer3Plus tool (http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi) and synthesized by Integrated DNA Technologies.

Primer used in this study are listed below: *GAPDH*: forward (F), TGCACCACCA ACTGCTTAGC and reverse (R), GGCATGGACTGTGGTCATGAG; *HPRT1*: F, AGGCGAACCTCTCGGCTTTC and R, CTAATCACGACGCCAGGGCT; *ACTB*: F, AGGATGCAGAAGGAGATCACTG and R, AGTACTTGCGCTCAGGAGGAG; *AR*: F, CAGTGGATGGGCTGAAAAAT and R, GGAGCTTGGTGAGCTGGTAG; *FOXA1*-3′: F, GAAGACTCCAGCCTCCTCAACTG and R, TGCCTTGAAGTCCA GCTTATGC; *FOXA1*-5′: F, CTACTACGCAGACACGCAGG and R, CCGCTCGTAGTCATGGTGTT; *TLE3*: F, AAGGACAGCTTGAGCCGATA and R, TTTGGTCTTGGAGGAAGGTG; *TTC6*: F, CGAACAGAGCCAGGAGGT AG and R, GTTCTCCCTGGGCTCCTAAC; *MIPOL1*: F, GCAAACGGTTAGAGC AGGAG and R, GGGTCTGGATTTCCTCTTCC; *ETV1*: F, TACCCCATGGACC ACAGATT and R, CACTGGGTCGTGGTACTCCT; *TUBB*: F,CTGGACCGCATC TCTGTGTACT and R,GCCAAAAGGACCTGAGCGAACA.

siRNA-mediated gene knockdown. Cells were seeded in a 6-well plate at the density of 100,000–250,000 cells per well. After 12 h, cells were transfected with 25 nM of gene-targeting ON-TARGETplus SMARTpool siRNAs or non-targeting pool siRNAs as negative control (Dharmacon) using the RNAiMAX reagent (Life Technologies; cat. no. 13778075) on two consecutive days, following the manufacturer's instructions. Both total RNA and protein were extracted on day 3 (total 72 h) to confirm efficient (>80%) knockdown of the target genes. For crystal-violet staining, at day 9 growth medium was aspirated and cells were first fixed with 4% formaldehyde solution, followed by a 30-min incubation in 0.5% crystal-violet solution in 20% methanol, and then scanned. Catalogue numbers and guide sequences (5′ to 3′) of siRNA SMARTpools (Dharmacon) used are: non-targeting control (cat. no. D-001810-10-05; UGGUUUACAUGUCGACUAA, UGGUUUACAUGUUGUGUGA, UGGUUUACAUGUUUUCUGA, UGGUUUA CAUGUUUUCCUA); *AR* (cat. no. L-003400-00-0005; GAGCGUGGACUUUCCG GAA, UCAAGGAACUCGAUCGUAU, CGAGAGAGCUGCAUCAGUU, CAGAAAUGAUUGCACUAUU); *FOXA1* (cat. no. L-010319-00-0005; GCACUGCAAUACUCGCCUU, CCUCGGAGCAGCAGCAUAA, GAACAGCU ACUACGCAGAC, CCUAAACACUUCCUAGCUC); *TLE3* (cat. no. L-019929-00-0005; GCCAUUAUGUGAUGUACUA, GCAUGGACCCGAUAGGUAU, GAACCACCAUGAACUCGAU, UCAGGUCGAUGCCGGGUAA).

The *FOXA1* SMARTpool consists of siRNAs targeting 5′ as well as 3′ ends of the *FOXA1* transcript. Thus, both wild-type and class-2 mutant transcripts are degraded using the SMARTpool siRNAs. This was experimentally confirmed in LAPC4 cells that endogenously contain a *FOXA1* class-2 mutation (Extended Data Fig. 1d, e).

CRISPR–Cas9-mediated gene or enhancer knockout. Cells were seeded in a 6-well plate at the density of 200,000–300,000 cells per well and infected with viral particles with lentiCRISPR-V2 plasmids coding either non-targeting (sgNC) or single guide RNAs (sgRNAs) targeting the exon 1 or the FKHD of *FOXA1* (both resulting in FOXA1 inactivation). This was followed by three days of puromycin selection, after which proliferation assays were carried out as described below. The lentiCRISPR-V2 vector was a gift from the laboratory of F. Zhang (Addgene plasmid no. 52961).

sgRNA sequences used are as follows: sgNC no. 1: 5′-GTAGCGAACGTGTCC GGCGT-3′; sgNC no. 2: 5′-GACCGGAACGATCTCGCGTA-3′; sg*FOXA1* exon 1: 5′-GTAGTAGCTGTTCCAGTCGC-3′; sg*FOXA1* FKHD: 5′-GCCGTTCTCGAACATGTTGC-3′.

Alternatively, for functional interrogation of the *FOXA1* topologically associating domain (TAD) enhancer elements, VCaP or LNCaP cells were transfected with pairs of sgRNAs targeting the *MIPOL1* untranslated region (UTR) or *FOXMIND* or a control locus within the *FOXA1* TAD. Transfected cells were then selected with puromycin (1.0 μg/ml) for 48 h, followed by incubation for an additional 72 h. Total RNA was extracted and qPCR was performed as described above.

Pairwise sgRNA sequences are as follows (5′ to 3′): control sgRNA (sgCtrl): CA CCGATTAGCCTCAACTATACCA and CACCGTGCAATATCTGAATCACACG; sgMIPOL1 UTR: CACCGTGAAAAAAAACGACAGTCTG and CACCGAACTC AAGTCAGCAGCAAAG; sgFOXMIND 1: CACCGCTTTAATAAAGCTATTTGC and CACCGATAGAGTGACTAATGCCCTG; sgFOXMIND 2: CACCGTAACAGT TGACCTACTAAC and CACCGATTTAGATAAGGGGATAGAA; sgFOX-MIND 3: CACCGCTTTAATAAAGCTATTTGC and CACCGATTTAG ATAAGGGGATAGAA.

CRISPR knockout screen. For the genome-wide CRISPR knockout screen, a two-vector system was used. First, LNCaP cells were engineered to stably overexpress the enzymatically active Cas9 protein. These cells were then treated with the human GeCKO knockout sgRNA library (GecKO V2) that was a gift from the Zhang laboratory (Addgene; cat. no. 1000000049). This was followed by puromycin

selection for 48 h, after which a fraction of these cells was processed to isolated genomic DNA as the input sample. The remaining cells were then cultured for 30 days, and genomic DNA was extracted at this time point. sgRNA sequences were amplified using common adaptor primers and sequenced on the Illumina HiSeq 2500 (125-nucleotide read length). Sequencing data were analysed as described[31] and depletion or enrichment of individual sgRNAs at 30 days was calculated relative to the input sample. Note that only a subset of genes—including essential controls, epigenetic regulators and transcription factors from the GeCKO-V2 screen—was plotted in Extended Data Fig. 1i.

Proliferation assays. For siRNA growth assays, cells were directly plated in a 96-well plate at the density of 2,500–8,000 cells per well and transfected with gene-specific or non-targeting siRNAs, as described above, on day 0 and day 1. Every treatment was carried out in six independent replicate wells. CellTiter-Glo reagent (Promega) was used to assess cell viability at multiple time points after tranfection, following the manufacturer's protocol. Data were normalized to readings from siNC treatment on day 1, and plotted as relative cell viability to generate growth curves.

Alternatively, for CRISPR sgRNA growth assays, cells were treated as described above for target-gene inactivation and seeded into a 24-well plate at 20,000 cells per well, with 2 replicates per group. After 12 h, plates were placed into the IncuCyte live-cell imaging machine (IncuCyte) set at the phase-contrast option to record cell confluence every 3 h for between 7 and 9 days. Similarly, for class-1 growth assays (Fig. 2f), stable doxycycline-inducible 22RV1 cells were grown in 10% charcoal-stripped-serum (CSS)-supplemented medium for 48 h. Androgen-starved cells were then seeded into a 96-well plate at 5,000 cells per well in 10% CSS medium with or without addition of doxycycline (1 μg/ml) to induce control or mutant protein expression (6 replicates per group). Once adherent, treated cells were placed in the IncuCyte live-cell imaging machine set at phase contrast to record cell confluence every 3 h for between 7 and 9 days. In all IncuCyte assays, confluence measurements from all time points were normalized to the matched measurement at 0 h and plotted as relative confluence to generate growth curves.

Cloning of representative FOXA1 mutants. Wild-type *FOXA1* coding sequence was purchased from Origene (cat. no. SC108256) and cloned into the pLenti6/V5 lentiviral vector (Thermo Fisher Scientific; cat. no. K4955-10) using the standard TOPO cloning protocol. Class-1 missense mutations (I176M, H247Q and R261G) were engineered from the wild-type FOXA1 vector using the QuikChange II XL Site-Directed Mutagenesis Kit (Agilent Tech) as per the manufacturer's instructions. All point mutations were confirmed using Sanger sequencing through the University of Michigan Sequencing Core Facility. Engineered mutant plasmids were further transfected in HEK293 cells to confirm expression of the mutant protein. For truncated class-2 variants, the wild-type coding sequence up to the amino acid before the intended mutation was cloned. All FOXA1 variants had the V5 tag fused on the C terminus. Selected mutants were cloned into a doxycycline-inducible vector (Addgene: pCW57.1; cat. no. 41393) to generate stable lines. For FRAP and single particle tracking assays, the pCW57.1 vector was edited to incorporate an in-frame GFP or Halo coding sequences at the C-terminal end, respectively.

FRAP assay and data quantification. PNT2 cells were seeded in a 6-well plate at 200,000 cells per well, and transfected with 2 μg of doxycycline-inducible vectors that encoded different variants of FOXA1 fused to GFP on the C-terminal end. After 24 h, cells were plated in glass-bottom microwell dishes (MatTek; #P35G-1.5-14-C) in phenol-free growth medium supplemented with doxycycline (1 μg/ml). Cells were then incubated for 48 h to allow for robust expression of the exogenous GFP-tagged protein and strong adherence to the glass surface. Microwell dishes were placed in humidity-controlled chamber set at 37 °C (Tokai-Hit) and mounted on the SP5 Inverted 2-Photon FLIM Confocal microscope (Leica). FRAP Wizard from the Leica Microsystems software suite was used to conduct and analyse FRAP experiments. Fluorescence signals were automatically computed in regions of interest using in-built tools in the FRAP Wizard. Roughly half of the nucleus was photobleached using the argon laser at 488 nm and 100% intensity for 20–30 iterative frames at 1.2-s intervals. Laser intensity was reduced to 1% for imaging post bleaching. Immediately after photobleaching, 2 consecutive images were collected at 1.2-s intervals followed by images taken at 10-s intervals for 60 frames (that is, 10 min).

For data analyses, recovery of signal in the bleached half and loss of signal in the unbleached half were measured as average fluorescence intensities in at least 80% of the respective areas, excluding the immediate regions flanking the separating border. All intensity curves were generated from background-subtracted images. The fluorescence signal measured in a region of interest was normalized to the signal before bleaching using the following formula[32]: $R = (I_t - I_{bg})/(I_o - I_{bg})$, in which I_o is the average intensity in the region of interest before bleaching, I_t is the average intensity in the region of interest at any time-point after bleaching and I_{bg} is the background fluorescence signal in a region outside of the cell nucleus. Raw recovery kinetic data from above were fitted with best hyperbolic curves using the GraphPad Prism software and the time until 50% recovery was calculated from the resulting best-fit equations. For representative time-lapse nuclei images shown in the FRAP figures, the fluorescence signal was uniformly brightened for ease of visualization.

Single particle tracking experiment and data quantification. PNT2 cells were transiently transfected with doxycycline-inducible vectors encoding C-terminal Halo-tagged wild-type or class-1 mutant variants of FOXA1. Transfected cells were seeded in glass-bottomed DeltaT culture dishes (Bioptechs; cat. no. 04200417C) and incubated for 24 h with 0.01 μg/ml of doxycycline. Cells were then treated with phenol-red-free medium containing 2% FBS and 5 nM cell permeable JF549 Halo ligand dye[33] for 30 min at 37 °C. Cells were subsequently washed twice, 10 min per wash at 37 °C, with phenol-red-free medium containing 2% FBS. Before imaging, cells were washed once with the 1× HBSS buffer and were imaged in the buffer.

Single particle tracking was performed on an Olympus IX81 microscope via HILO illumination, as previously described[34], at a spatial accuracy of 30 nm and temporal resolution of 33 ms. Image analysis was performed as previously described[35]. In brief, tracking was done in Imaris (bitplane) and particles that were at least visible for four continuous frames were used for further analysis. Diffusion constants were calculated as previously described[36], assuming a Brownian diffusion model under steady-state conditions. Dwell time histograms were fit to a double-exponential function to extract fast and slow dwell times of 'bound' particles that displayed a frame-to-frame displacement of <300 nm. All particles that were visible for less than 4 consecutive frames, or those that moved >300 nm between frames, were counted as 'unbound' particles. At least 5 cells were imaged for each transcription factor variant and >500 particles were tracked to extract diffusion constants and dwell time.

Dual luciferase AR reporter assay. HEK293 cells stably overexpressing the wild-type AR protein (that is, HEK293-AR) were used for the AR reporter assays. HEK293-AR cells were seeded in a 12-well plate at 300,000 cells per well and transfected with 2 μg of the pLenti6/V5 vector encoding different variants of FOXA1, or GFP (control). After 8 h, medium was replaced with 10% CSS-supplemented phenol-free medium (androgen-depleted) and cells were transfected with the AR reporter Firefly luciferase or negative-control constructs from the Cignal AR-Reporter(luc) kit (Qiagen; cat. no. CCS-1019L) as per the manufacturer's instructions. Both constructs were premixed with constitutive *Renilla* luciferase vector as control. After 12 h, cells were treated with different dosages of DHT or enzalutamide (at 10 μM dosage); and additional 24 h later dual luciferase activity was recorded for every sample using the Dual-Glo Luciferase assay (Promega; E2980) and luminescence plate reader (Promega-GLOMAX-Multi Detection System). Each treatment condition had four independent replicates. Firefly luciferase signals were normalized with the matched *Renilla* luciferase signals to control for variable cell number and/or transfection efficiencies, and normalized signals were plotted relative to the negative control reporter constructs.

Electrophoretic mobility shift assay. HEK293 cells were plated in 10-cm dishes at 1 million per plate and transfected with 10 μg of the pLenti6/V5 vector coding GFP (control) or different variants of FOXA1. After 48 h, cells were trypsinized and nuclear lysates were prepared using the NE-PER kit reagents (Thermo Fisher Scientific). Immunoblots were run to confirm comparable expression of recombinant FOXA1 variants in 2 μl (that is, equal volume) of final nuclear lysates. Next, FOXA1 and AR ChIP–seq data were used to identify the *KLK3* enhancer element. Sixty base pairs of the *KLK3* enhancer, centred at the *FOXA1* consensus motif 5′-GTAAACAA-3′, were synthesized as single-stranded oligonucleotides (IDT) and biotin-labelled using the Biotin 3′-End DNA labelling kit (Thermo Fisher Scientific), and then annealed to generate a labelled double-stranded DNA duplex.

Binding reactions were carried out in 20-μl volumes containing 2 μl of the nuclear lysates, 50 ng/μl poly(dI.dC), 1.25% glycerol, 0.025% Nonidet P-40 and 5 mM $MgCl_2$. Biotin-labelled *KLK3* enhancer probe (10 fmol) was added at the very end with gentle mixing. Reactions were incubated for 1 h at room temperature, size-separated on a 6% DNA retardation gel (100 V for 1 h; Invitrogen) in 0.5× TBE buffer, and transferred on the Biodyne Nylon membrane (0.45 μm; Thermo Fisher Scientific) using a semi-dry system (BioRad). Transferred DNA was crosslinked to the membrane using the UV light at 120 mJ/cm^2 for 1 min. Biotin-labelled free and protein-bound DNA was detected using HRP-conjugated streptavidin (Thermo Fisher Scientific) and developed using chemiluminescence according to the manufacturer's protocol.

Protein synthesis and purification. First, wild-type FOXA1 and FOXA1(P358fs) proteins were purified using the *Escherichia coli* bacterial expression system and nickel-affinity chromatography. In brief, wild-type FOXA1 or FOXA1(P358fs) coding sequences were cloned into the pFC7A (HQ) Flexi vector (Promega; cat. no. C8531) with a C-terminal HQ tag, following the manufacturer's protocol. These expression constructs were used to transform Single Step (KRX) Competent *E. coli* cells (Promega; cat. no. L3002), which have been modified for synthesis of mammalian proteins. A starter broth of 2 ml was inoculated with a single colony of transformed bacterial cells and incubated at 37 °C with constant shaking at

250 rpm until an optical density at 600 nm (OD_{600}) of 0.4–0.5 was reached. The starter brother was then used to inoculate 1,000 ml of LB broth containing ampicillin, and protein synthesis was induced using 0.1% v/v of rhamanose. Induced culture was incubated at 20 °C for 16 h with constant shaking at 250 rpm. Bacterial cells were then pelleted by centrifugation at 4,000 rpm for 30 min and mechanically lysed through sonication in 50 mM Tris (pH 7.4), 150 mM NaCl, 1 mM $MgCl_2$, 0.5 mM EDTA, 1 mM DTT and 1% glycerol in the presence of protease inhibitors (Roche). HisLink Purification Resin (Promega; cat. no. V8821) was used to purify untagged recombinant proteins from the crude bacterial lysates as per the manufacturer's protocol (this also includes removal of the His tag). Purified protein fractions were then tested for purity by Coomaisse staining relative to the crude input lysates, and purified protein concentrations were estimated using protein standards of known concentrations (Thermo Fisher Scientific; cat. no. 23208). The identities of purified proteins were confirmed via immunoblotting using an N-terminal FOXA1 antibody (Cell Signaling Technology; cat. no. 58613S).

Biolayer interferometry assay. Biolayer interferometry (BLI) assays were carried out using the Octet-RED96 system (PALL ForteBio) and in-built analysis software. In brief, a biotin-labelled, 60-bp *KLK3* enhancer element centred at the FOXA1 consensus motif was immobilized on the Super Streptavidin Biosensors (PALL ForteBio, part no. 18-5057) with the loading step carried out for 1,000 s with shaking at 500 rpm. This was followed by baseline measurements for 120 s and association for 900 s using varying concentrations of purified FOXA1 proteins (3.125–100 nM; two replicate biosensors per concentration). A control DNA element with no FOXA1 motif was used in the negative-control reaction with 100 nM of the protein. The association step was followed by the dissociation step for 3,000 s. Signal from all the biosensors was adjusted for the background signal from the control sensors and normalized data of DNA binding kinetics were analysed using the Octet-RED96 (PALL ForteBio) analysis software, as previously described[37].

Generation of CRISPR clones and stable lines. 22RV1 or LNCaP cells were seeded in a 6-well plate at 200,000 cells per well and transiently transfected with 2.5 μg of lentiCRISPR-V2 (Addgene; 52961) vector using the Lipofectamine 3000 reagent (cat. no. L3000008), encoding the Cas9 protein and sgRNA that cuts either at amino acid 271 (5′-GTCAAGTGCGAGAAGCAGCCG-3′) or 359 (5′-GCCGGGCCCGGAGCTTATGGG-3′) in exon 2 of *FOXA1*. Cells were treated with non-targeting control sgRNA (5′-GACCGGAACGATCTCGCGTA-3′) vector to generate isogenic wild-type clones. Transfected cells were selected with puromycin (Gibco) for 3–4 days and sorted by fluorescence-activated cell sorting as single cells into 96-well plates. Cells were maintained in 96-well plates for 4–6 weeks, with replacement of the growth medium every 7 days to allow for the expansion of clonal lines. Clones that successfully seeded were further expanded and genotyped for *FOXA1* using Sanger sequencing, and immunoblotting with the N-terminal FOXA1 antibody. Sequence- and expression-validated 22RV1 and LNCaP clones with distinct class-2 mutations were used for growth, invasion and metastasis assays as described.

To generate stable cells, doxycycline-inducible vectors coding different variants of FOXA1 or GFP (control) were packaged into viral particles at the University of Michigan Vector Core. Prostate cancer cells were seeded in a 6-well plate at 100,000–250,000 cells per well and infected with 0.5 ml of 10× viral titres packaged at the University of Michigan Vector Core. This was followed by 3–4 days of puromycin (Gibco) selection to generate stable lines.

Rescue growth and functional compensation experiments. Stable 22RV1 cells with doxycycline-inducible expression of empty vector (control), FOXA1 wild type, or distinct FOXA1 mutants were seeded in a 6-well plate in the completed growth medium supplemented with 1.0 μg/ml of doxycycline. Notably, the exogenous genes only contain the coding sequence of *FOXA1* without its intron and UTRs. After 24 h, cells were transfected with 30 nM of either distinct 3′ UTR-specific *FOXA1*-targeting siRNAs or a non-targeting control siRNA using the RNAiMAX (Life Technologies; cat. no. 13778075) reagent. *FOXA1* UTR-specific siRNAs were purchased from Thermo Fisher Scientific (cat. no. siNC, 4390844 (sequence is proprietary); siRNA no. 3, s6687 (sense sequence: 5′-GCAAUACUCUUAACCAUAA-3′); siRNA no. 4, 5278 (sense sequence: 5′-AACACATAAAATTAGTTTC-3′); and siRNA no. 5 – 107428 (sense sequence: 5′-AAGTTATAGGGAGCTGGAT-3′)). On the following day, cells were counted and seeded in a 96-well plate at a density of 5,000 cells per well with 6 replicates for each treatment condition. Cell growth was then assessed using the IncuCyte assay, as described above.

Testing the GFP-tagged wild-type FOXA1 variant. 22RV1 cells were seeded in 10-cm dishes and transfected with 8 μg of mammalian expression plasmids encoding either FOXA1(WT) or FOXA1(WT)–GFP (the exact construct used in the FRAP assay) using the Lipofectamine 3000 (Life Technologies; cat. no. L3000008) reagent, as per the manufacturer's protocol. Transgene expression was induced using 1.0 μg/ml of doxycycline and cells were cultured for 96 h with doxycycline replenishment every 48 h. Total RNA was extracted and RNA-seq was performed as described. A portion of these cells was used for the rescue growth experiments using UTR-specific *FOXA1* siRNAs as described above.

Matrigel invasion assay. 22RV1 CRISPR clones were grown in 10% CSS-supplemented medium for 48 h for androgen starvation. A matrigel-coated invasion chamber was used, which was additionally coated with a light-tight polyethylene terephthalate membrane to allow for fluorescent quantification of the invaded cells (Biocoat: 24-well format, no. 354166). Fifty thousand starved cells were resuspended in serum-free medium and were added to each invasion chamber. Twenty per cent FBS-supplemented medium was added to the bottom wells to serve as a chemoattractant. After 12 h, medium from the bottom well was aspirated and replaced with 2 μg/ml Calcein-green AM dye (Thermo Fisher Scientific; C3100MP) in 1× HBSS (Gibco) and incubated for 30 min at 37 °C. Invasion chambers were then placed in a fluorescent plate reader (Tecan-Infinite M1000 PRO) and fluorescent signals from the invaded cells at the bottom were averaged across 16 distinct regions per chamber to determine the extent of invasion.

ChIP–seq. ChIP experiments were carried out using the HighCell# ChIP-Protein G kit (Diagenode) as per the manufacturer's protocol. Chromatin from five million cells was used per ChIP reaction with 6.5 μg of the target protein antibody. In brief, cells were trypsinized and washed twice with 1× PBS, followed by crosslinking for 8 min in 1% formaldehyde solution. Crosslinking was terminated by the addition of 1/10 volume 1.25 M glycine for 5 min at room temperature followed by cell lysis and sonication (Bioruptor, Diagenode), resulting in an average chromatin fragment size of 200 bp. Fragmented chromatin was then used for immunoprecipitation using various antibodies, with overnight incubation at 4 °C. ChIP DNA was de-crosslinked and purified using the iPure Kit V2 (Diagenode) using the standard protocol. Purified DNA was then prepared for sequencing as per the manufacturer's instructions (Illumina). ChIP samples (1–10 ng) were converted to blunt-ended fragments using T4 DNA polymerase, *E. coli* DNA polymerase I large fragment (Klenow polymerase) and T4 polynucleotide kinase (New England BioLabs (NEB)). A single A base was added to fragment ends by Klenow fragment (3′ to 5′ exo minus; NEB) followed by ligation of Illumina adaptors (Quick ligase, NEB). The adaptor-ligated DNA fragments were enriched by PCR using the Illumina Barcode primers and Phusion DNA polymerase (NEB). PCR products were size-selected using 3% NuSieve agarose gels (Lonza) followed by gel extraction using QIAEX II reagents (Qiagen). Libraries were quantified and quality checked using the Bioanalyzer 2100 (Agilent) and sequenced on the Illumina HiSeq 2500 Sequencer (125-nucleotide read length).

Zebrafish embryo metastasis experiment. Wild-type AB^{TL} zebrafish were maintained in aquaria according to standard protocols. Embryos were generated by natural pairwise mating and raised at 28.5 °C on a 14 h light/10 h dark cycle in a 100-mm Petri dish containing aquarium water with methylene blue to prevent fungal growth. All experiments were performed with 2–7-day-old embryos post-fertilization, and were done in approved University of Michigan fish facilities using protocols approved from the University of Michigan Institutional Animal Care and Use Committee (UM-IACUC). Cell injections were carried out as previously described[38]. In brief, GFP-expressing normal (control) or cancer cells were resuspended in PBS at the concentration of 1×10^7 cells/ml. Forty-eight hours after fertilization, wild-type embryos were dechorionated and anaesthetized with 0.04 mg/ml tricaine. Approximately 10 nl (approximately 100 cancer cells) were microinjected into the perivitelline space using a borosilliac micropipette tip with filament. Embryos were returned to aquarium water and washed twice to remove tricaine, then moved to a 96-well plate with one embryo per well and kept at 35 °C for the duration of the experiment. All embryos were imaged at 24-h intervals to follow metastatic dissemination of injection cells. Water was changed daily to fresh aquarium water. More than 30 fish were injected for each condition (wild-type no. 2, $n = 30$; wild-type no. 5, $n = 50$; no. 57, $n = 35$; no. 84, $n = 57$; no. 113, $n = 38$) and metastasis was visually assessed daily up to 5 days after injection (that is, for a total of 7 days post-fertilization) by counting the total number of distinct cellular foci in the body of the embryos. All of the metastasis studies were terminated at seven days post-fertilization in accordance with the approved embryo protocols. Embryos were either imaged directly in the 96-well plates or placed onto a concave glass slide to capture representative images using a fluorescent microscope (Olympus-IX71). For quantification, evidently distinct cell foci in the embryo body were counted 72 h after the injections.

For all these experiments, relevant ethical regulations were carefully followed. No statistical methods were used to predetermine sample size for any of the cohort analyses or experiments. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment unless otherwise stated.

ATAC-seq and data analysis. ATAC-seq was performed as previously described[39]. In brief, 25,000 normal prostate or prostate cancer cells were washed in cold PBS and resuspended in cytoplasmic lysis buffer (CER-I from the NE-PER kit, Invitrogen, cat. no. 78833). This single-cell suspension was incubated on ice for 10 min with gentle mixing by pipetting at every 2 min. The lysate was centrifuged at 1,300*g* for 5 min at 4 °C. Nuclei were resuspended in 2× TD buffer, then incubated with Tn5 enzyme for 30 min at 37 °C (Nextera DNA Library Preparation Kit;

cat. no. FC-121-1031). Samples were immediately purified by Qiagen minElute column and PCR-amplified with the NEBNext High-Fidelity 2X PCR Master Mix (NEB; cat. no. M0541L). qPCR was used to determine the optimal PCR cycles to prevent over-amplification. The amplified library was further purified by Qiagen minElute column and SPRI beads (Beckman Coulter; cat. no. A63881). ATAC-seq libraries were sequenced on the Illumina HiSeq 2500 (125-nucleotide read length).

Paired-end .fastq files were uniquely aligned to the hg38 human genome assembly using Novoalign (Novocraft) (with the parameters -r None -k -q 13 -k -t 60 -o sam –a CTGTCTCTTATACACATCT), and converted to .bam files using SAMtools (version 1.3.1). Reads mapped to mitochondrial or duplicated reads were removed by SAMtools and PICARD MarkDuplicates (version 2.9.0), respectively. Filtered .bam files from replicates were merged for downstream analysis. MACS2 (2.1.1.20160309) was used to call ATAC-seq peaks. The coverage tracks were generated using the program bam2wig (http://search.cpan.org/dist/Bio-ToolBox/) with the following parameters:–pe–rpm–span–bw. Bigwig files were then visualized using the IGV (Broad Institute) open source genome browser.

ChIP–seq data analysis. Paired-end 125-bp reads were trimmed and aligned to the GRCh38 human reference using the STAR (version 2.4.0g1) aligner with splicing disabled; the resulting reads were filtered using samtools 'samtools view -@ 8 -S -1 -F 384'. The resulting .bam file was sorted and duplicate-marked using Novosort, and converted into a bigwig file for visualization using 'bedtools genomecov -bg -split -ibam' and 'bedGraphToBigWig'. The coverage signal was normalized to total sequencing depth/1 $\times$ 10^6 reads. Peak calling was performed using MACS2 with the following settings: 'macs2 callpeak–call-summits–verbose 3 -g hs -f BAM -n OUT–qvalue 0.05'. ChIP peak profile plots and read-density heat maps were generated using deepTools2[40], and cistrome overlap analyses were carried out using the ChIPpeakAnno[41] package in R. It is important to note that, given the cistromic dominance of class-2 mutants, in heterozygous class-2 mutant clones part of the FOXA1 antibody binds to the wild-type protein that does not interact with, or immunoprecipitate, the DNA. This confounds all analyses involving peak-read density comparisons between the wild-type and class-2-mutant FOXA1 ChIP–seq data; we therefore largely avoided this strategy in our study. For the same reason, the read densities from only the heterozygous clones were factored by 1.5 for heat map generation in Fig. 3d.

De novo and known motif enrichment analysis. All de novo and known motif enrichment analyses were performed using the HOMER (v.4.10) suite of algorithms[42]. Peaks were called by the findPeaks function (-style factor -o auto) at 0.1% false discovery rate; de novo motif discovery and enrichment analysis of known motifs were performed with findMotifsGenome.pl (-size 200 -mask). For motif analysis of common wild-type- and mutant-specific chromatin binding sites, the top 5,000 peaks ranked by score were used as input. A common set of background sequences was generated by di-nucleotide shuffling of the input sequences using the fasta-shuffle-letters function from MEME[43]. Alternatively, we ranked peaks by the relative signal fold change between mutant and wild type, and selected the top and bottom 5,000 peaks (keeping the requirement that mutant-specific peaks are not called in the wild-type cistrome, and vice versa) for motif discovery. For class-2 mutants, only heterozygous 22RV1 clones were used, which more accurately recapitulate the clinical presentation of *FOXA1* mutations. Also, for both mutational classes, cistromes from biological replicates were merged to define a union cistrome that was compared to the union wild-type cistrome generated from matched FOXA1 wild-type cells. For the supervised motif analyses, we identified all instances of the FOXA canonical motif (5′-T[G/A]TT[T/G]AC-3′) within cistromes (ChIP–seq peaks) of class-1 and wild-type FOXA1 proteins using motifmatchR, and calculated nucleotide frequencies in the flanking positions.

Cohorts, datasets and resources. This study uses previously published public or restricted patient genetic data. Genetic calls for primary prostate cancer and breast cancer were obtained from the Genomic Data Commons (GDC)[44] for the prostate cancer PRAD[5] and breast cancer BRCA[6,45] cohorts, respectively. Raw RNA-seq data (paired-end reads from unstranded polyA libraries) for the samples were downloaded from the GDC and processed with our standard clinical RNA-seq pipeline CRISPR/CODAC (see below). For The Cancer Genome Atlas (TCGA) PRAD and BRCA cohorts, we downloaded mutational calls from multiple sources (GDC, cBio Portal and UCSC Xena) and additionally used the BAM-slicing tool to download sequence alignments from whole-exome sequencing libraries to the *FOXA1* locus. We then used our internal pipeline (see below) to call single-nucleotide variants and indels within *FOXA1*. We also used the downloaded aligned data for manual review of *FOXA1* mutation calls. Mutation calls for advanced primary and metastatic cases were obtained from the MSK-IMPACT cohort (downloaded from the cBio portal[46]). The main MCTP mCRPC cohort includes 360 previously reported cases (the location of all raw .bam files is provided in ref. [47]), the 10 additional mCRPC cases included here (but not in ref. [47]) will be included in the Database of Genotypes and Phenotypes (dbGaP) under accession code phs000673.v3.p1, and belong to a continuous sequencing program with the same IRB-approved protocol (MI-Oncoseq program, University of Michigan Clinical Sequencing Exploratory Research). The genetic sequencing data (WXS) for rapid autopsy cases are available from dbGaP with accession codes hs000554.v1.p1and phs000567.v1.p1. De-identified somatic mutation calls, RNA-seq fusion calls, processed and segmented copy-number data, and RNA-seq expression matrices across the full 370 cases of the MCTP mCRPC cohort are available on request from the authors.

Preparation of whole-exome sequencing and RNA-seq libraries. Integrative clinical sequencing (comprising exome sequencing and polyA and/or capture RNA-seq) was performed using standard protocols in our Clinical Laboratory Improvement Amendments-compliant sequencing laboratory. In brief, tumour genomic DNA and total RNA were purified from the same sample using the AllPrep DNA/RNA/miRNA kit (Qiagen). Matched normal genomic DNA from blood, buccal swab or saliva was isolated using the DNeasy Blood & Tissue Kit (Qiagen). RNA-seq was performed using the exome-capture transcriptome platform[48]. Exome libraries of matched pairs of tumour and normal DNA were prepared as previously described[49], using the Agilent SureSelect Human All Exon v4 platform (Agilent). All the samples were sequenced on an Illumina HiSeq 2000 or HiSeq 2500 (Illumina) in paired-end mode. The primary base call files were converted into FASTQ sequence files using the bcl2fastq converter tool bcl2fastq-1.8.4 in the CASAVA 1.8 pipeline.

Analysis of whole-exome sequencing data. The .fastq sequence files from whole-exome libraries were processed through an in-house pipeline constructed for analysis of paired tumour and normal data. The sequencing reads were aligned to the GRCh37 reference genome using Novoalign (version 3.02.08) (Novocraft) and converted into .bam files using SAMtools (version 0.1.19). Sorting, indexing, and duplicate marking of .bam files used Novosort (version 1.03.02). Mutation analysis was performed using freebayes (version 1.0.1) and pindel (version 0.2.5b9). Variants were annotated to RefSeq (via the UCSC genome browser, retrieved on 22 August 2016), as well as COSMIC v.79, dbSNP v.146, ExAC v.0.3 and 1000 Genomes phase 3 databases using snpEff and snpSift (v.4.1g). Single nucleotide variants and indels were called as somatic if they were present with at least 6 variant reads and 5% allelic fraction in the tumour sample, and present at no more than 2% allelic fraction in the normal sample with at least 20$\times$ coverage. Additionally, the ratio of variant allelic fractions between tumour and normal samples was required to be at least six to avoid sequencing and alignment artefacts at low allelic fractions. Minimum thresholds were increased for indels observed to be recurrent across a pool of hundreds of platform- and protocol-matched normal samples. Specifically, for each such indel, a logistic regression model was used to model variant and total read counts across the normal pool using PCR duplication rate as a covariate, and the results of this model were used to estimate a predicted number of variant reads (and therefore allelic fraction) for this indel in the sample of interest, treating the total observed coverage at this genomic position as fixed. The variant read count and allelic fraction thresholds were increased by these respective predicted values. This filter eliminates most recurrent indel artefacts without affecting our ability to detect variants in homopolymer regions from tumours exhibiting microsatellite instability. Germline variants were called using 10 variant reads and 20% allelic fraction as minimum thresholds, and were classified as rare if they had less than 1% observed population frequency in both the 1000 Genomes and ExAC databases. Exome data were analysed for copy-number aberrations and loss of heterozygosity by jointly segmenting B-allele frequencies and $\log_2$-transformed tumour/normal coverage ratios across targeted regions using the DNAcopy (version 1.48.0) implementation of the Circular Binary Segmentation algorithm. The expectation–maximization algorithm was used to jointly estimate tumour purity and classify regions by copy-number status. Additive adjustments were made to the $\log_2$-transformed coverage ratios to allow for the possibility of non-diploid tumour genomes; the adjustment resulting in the best fit to the data using minimum mean-squared error was chosen automatically and manually overridden if necessary.

Detection of copy-number break ends from whole-exome sequencing. The output of our clinical whole-exome sequencing pipeline includes segmented copy-number data, inferred absolute copy numbers and predicted parent-specific genotypes (for example, AAB), detection of loss of heterozygosity, and detection of copy-neutral loss of heterozygosity (uniparental disomy). Together, these data enable the detection of joint discontinuities in the copy-number profile (log-ratio and B-allele frequencies) at exon-level resolution. A subset of genomic rearrangements results in changes in copy number or allelic shifts, and the presence of such discontinuities in paired tumour-normal whole-exome sequencing data are therefore strongly indicative of a somatic breakpoint. For example, one copy gain will result in a segment with an increased log-ratio, and a corresponding zygosity deviation (see above). This segment will be discontinuous with adjacent segments, which will result in the call of a whole-exome sequencing break end (discontinuity) on either side of the copy gain. The size of the break end depends on the density of covered exons and in general the resolution is better in genic versus intergenic regions. We assessed the presence of such breakpoints within the gene-dense and exon-dense *FOXA1* locus; all copy-number break ends met statistical thresholds of the circular binary segmentation (CBS) algorithm (see above) at either the log-ratio or B-allele level.

Genetic characterization of mCRPC tumour samples at the pathway level. The co-occurrence or mutual exclusivity of *FOXA1* alterations with other previously described genetic events in prostate cancer has been carried out at the pathway level, but grouping putative functionally equivalent (and largely genetically mutually exclusive) events. All known types of ETS fusion (*ERG*, *ETV1*, *FLI1*, *ETV4* and *ETV5*) were considered as ETS-positive tumours, PI3K alterations included *PTEN* homozygous loss, *PIK3CA* activating mutations and *PIK3R1* inactivating mutations, AR pathway alterations included *AR*, *NCOR1*, *NCOR2* and *ZBTB16* mutations or deletions, but excluded *AR* amplifications and copy gains. The KMT category included mutations in all recurrently mutated lysine methyltransferases. The WNT category included inactivating alterations in *APC* and activating mutations in *CTNNB1*. DRD included cases with mutations in *BRCA1*, *BRCA2*, *PALB2* and *ATM* (all common mismatch repair genes), and *CDK12*.

Assessment of two-hit biallelic alterations. To assess the frequency of genetic inactivations of both alleles we integrated mutational, copy-number and RNA-seq (fusion) data. A gene was considered as having both alleles inactivated for any combination (pair) of the following events: copy loss, mutation, truncating fusion and copy-number breakpoint, in addition to homozygous deletion of both copies and two independent mutations. Ambiguous cases were manually reviewed to increase the accuracy and ascertain whether both events, for example, copy-number breakpoint and gene fusion, are probably independent events.

Unified mutation calling and variant classification of FOXA1. Mutation calls for *FOXA1* obtained or downloaded from the GDC and TCGA flagship manuscripts[5,6] as well as our internal pipelines were lifted over to GRCh38 (using the Bioconductor package rtracklayer) and annotated with respect to the canonical RefSeq *FOXA1* isoform. For TCGA samples or cases, multiple call sets were available and we manually reviewed all discrepancies in *FOXA1* mutation calls, resulting in a unified call set with improved sensitivity and specificity. Mutational effect (consequence) was simplified into three categories: missense, in-frame indel and frameshift (the last category included stop-gain, stop-loss and splice-site mutations). The resulting mutations were dichotomized into class 1 and class 2 based on their position relative to amino acid residue 275. Variant allele frequencies were only available for TCGA and the in-house mCRPC cohorts.

Analysis of whole-genome sequencing data. The bcbio-nextgen pipeline version 1.0.3 was used for the initial steps of tumour whole-genome data analysis. Paired-end reads were aligned to the GRCh38 reference using BWA (bcbio default settings), and structural variant calling was done using LUMPY[50] (bcbio default settings), with the following post-filtering criteria: "(SR> = 1 & PE> = 1 & SU> = 7) & (abs(SVLEN)>5e4) & DP <1000 & FILTER == "PASS". The following settings were chosen to minimize the number of expected germline variants: false discovery rate (FDR) < 0.05 for germline status for both deletions and duplications. Additionally, common structural germline variants were filtered.

Analysis of 10X genomics long-read sequencing data. High-molecular mass DNA from MDA-PCA-2b and LNCaP cell lines was isolated and processed into linked-read next-generation sequencing libraries per the manufacturer's instructions (10X WGS v2 kit). The resulting paired-end sequencing data were sequenced on an Illumina Hi-Seq 2500 instrument and analysed (demultiplexing, alignment, phasing and structural variant calls) using the longranger 2.2.1 pipeline with all default settings. The resulting libraries met all 10X-recommended quality control parameters including molecule size, average phasing length, and sequencing coverage (~50×). Here, we focused on structural variant calls within the *FOXA1* TAD and confirmed the presence of the previously reported *FOXMIND-ETV1* fusions; that is, translocation for MDA-PCA-2b, and balanced insertional translocation for LNCaP. Both cell lines were confirmed to contain three copies of *FOXA1* (that is, one translocated allele and two duplicated alleles).

RNA-seq data pre-processing and primary analysis. RNA-seq data processing—including quality control, read trimming, alignment, and expression quantification by read counting—was carried out as previously described[49], using our standard clinical RNA-seq pipeline CRISP (available at https://github.com/mcieslik-mctp/bootstrap-rnascape). The pipeline was run with default settings for paired-end RNA-seq data of at least 75 bp. The only changes were made for unstranded transcriptome libraries sequenced at the Broad Institute and the TCGA and CCLE cohorts, for which quantification using featureCounts[51] was used in unstranded mode '-s0'. The resulting counts were transformed into fragments per kilobase of transcript per million mapped reads using upper-quartile normalizations as implemented in EdgeR[52]. For mCRPC samples *FOXA1* expression estimates were adjusted by tumour content estimated from whole-exome sequencing (see above) given the highly prostate-specific *FOXA1* expression profile. For the quantification of *FOXMIND* expression levels, a custom approach was necessary given the poor annotation and unspliced nature of this transcript. First, we delineated regions of sense and antisense transcription from the *FOXMIND* ultra-conserved regulatory elements, chr14:37564150-37591250:+ and chr14:37547900-37567150:-, respectively. Next, to make the expression estimates reliable in unstranded libraries, we identified regions of substantial overlap between the sense and antisense RP11-356O9.1 transcripts, and *FOXA1* and *MIPOL1*. These overlaps have been excluded from quantification, resulting in the following trimmed target regions: chr14:37564150-37589500, and chr14:37553500-37567150. Within these regions, the average base-level coverage normalized to sequencing depth was computed as an expression estimate.

Differential expression analyses. All differential expression analyses were done using limma R-package[53], with the default settings for the voom[54], lmFit, eBayes and topTable functions. The contrasts were designed as follows to identify transcriptional signatures of class-1 mutants. Given the mutual exclusivity of the genotypes in primary and metastatic tumours, the overall MCTP mCRPC cohort of 371 cases was partitioned into 4 groups: (1) ETS-fused or *SPOP*-mutant tumours, (2) class-1 mutant tumours, (3) class-2 mutant tumours, and (4) tumours that were wild type for ETS, *SPOP* and *FOXA1*. To avoid confounding effects, the class-2 and ETS and SPOP groups were excluded from class-1 transcriptional analyses. Next, the class-1 samples were contrasted with the wild-type samples with additional independent regressors for assay type (capture vs polyA, as previously described[49], and mutational status (see above) for the following genes and pathways: PI3K, WNT, DRD, *RB1* and *TP53*. In other words, we constructed a design matrix with coefficients for class-1 mutational status, in addition to coefficients for confounding variables and recurrent genetic heterogeneity. This allowed us to estimate the fold changes (expressed logarithmically) and adjusted *P* values associated with *FOXA1* mutations and other genotypes (for example, PI3K status). An analogous procedure was carried out for the primary class-1 samples (TCGA) and for class-2 mutations in mCRPC (MCTP), but given the lack of mutual-exclusivity between class-2 mutations and ETS and SPOP group, only class-1 mutations were excluded.

Pathway and signature enrichment analyses. The Molecular Signatures Database (MSigDB)[55] was used as a source of gene sets comprising cancer hallmarks, molecular pathways, oncogenic signatures and transcription factor targets. The enrichment of signatures was assessed using the parametric random-set method[56], and visualized using the gene-set enrichment analysis (GSEA) enrichment statistic[57] and barcode plots. All *P* values have been adjusted for multiple-hypothesis testing using a false discovery rate correction. To identify putative transcription factors regulating differentially expressed genes, we used the transcription factor prediction tool BART[25]. BART was run with all default settings, and the provided transcription factor databases. We used voom- and limma-based gene-level fold-changes as input to the algorithm.

Detection of structural variants from RNA-seq. The detection of chimeric RNAs (gene fusions, structural variants, circular RNAs and read-through events) was carried out using our previously published[49] in-house toolkit for the comprehensive detection of chimeric RNAs, CODAC (available at https://github.com/mctp/codac). In brief, three separate alignment passes (STAR 2.4.0g1) against the GRCh38 (hg38) reference with known splice junctions provided by Gencode v.27 (ref. [58]) are made for the purposes of expression quantification and fusion discovery. The first pass is a standard paired-end alignment followed by gene-expression quantification. The second and third pass are for the purpose of gene fusion discovery and to enable the chimeric alignment mode of STAR (chimSegmentMin: 10, chimJunctionOverhangMin: 1, alignIntronMax: 150000, chimScoreMin: 1). Fusion detection was carried out using CODAC with default parameters to balance sensitivity and specificity (annotation preset:balanced). CODAC uses MOTR v.2, a custom reference transcriptome based on a subset of Gencode 27 (available with CODAC). Prediction of topology (inversion, duplication, deletion and translocation), and distance (adjacent, breakpoints in two directly adjacent loci; cytoband, breakpoints within the same cytoband based on UCSC genome browser; arm, breakpoints within the same chromosome arm). The high specificity of our pipeline has been assessed through Sanger sequencing[49]. To create fusion circos plots, we have colour-coded the CODAC variants on the basis of the inferred topology of the breakpoints. Unbiased discovery of recurrently rearranged loci has been carried out by breaking the genome into 1.5-Mb windows with a step of 0.5 Mb. For each window, the percentage of patients with at least one RNA break end has been calculated. The resulting genomic windows were ranked and clustered by proximity for visualization. CODAC has the ability to make fusion calls independent of known transcriptome references or annotations and is therefore capable of detecting fusions involving intergenic or poorly annotated regions.

Classification of *FOXA1* locus genomic rearrangements. Structural variants within the *FOXA1* locus have been partitioned into two broad topological patterns: (1) translocations (including inversions and deletions involving distal loci on the same chromosome) and (2) focal duplications. The translocations have been further subdivided into hijacking and swapping events on the basis of their position relative to *FOXMIND* (GRCh38: chr14:37564150-37591250) and *FOXA1*. Hijacking translocations position a translocation partner within the *FOXMIND-FOXA1* regulatory domain (defined as GRCh38: chr14:37547501-37592000, based on manual review of chromatin conformation Hi-C, CTCF, H3K4me1, H3K27ac, evolutionary conservation and synteny data). Swapping

translocations preserve the *FOXMIND-FOXA1* regulatory domain but insert the translocation partner upstream of the *FOXA1* promoter, frequently 'swapping-out' the *TTC6* gene. Notably, one isoform of *TTC6* gene can be transcribed from the bi-directional *FOXA1* promoter. Focal duplications within the *FOXA1* locus have been derived from the CODAC structural-variant output file. In brief, for each case independently, all RNA-seq fusion junctions annotated by CODAC as tandem duplications and overlapping the *FOXA1* topologically associating domain (GRCh38: chr14:37210001-37907919) have been collated and used to infer the minimal duplicated region. Because RNA-seq chimeric junctions generally coincide with splice junctions (limited resolution) and generally cannot be phased (ambiguous haplotype), the inference of minimal duplicated regions makes the necessary and parsimonious assumption that overlapping tandem duplications are due to a single somatic genetic event, and not multiple independent events.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All raw data for the graphs, immunoblot and gel electrophoresis figures are included in the Source Data or Supplementary Information. All materials are available from the authors upon reasonable request. All the raw next-generation sequencing, ChIP and RNA-seq data generated in this study have been deposited in the Gene Expression Omnibus (GEO) repository at NCBI (accession code GSE123625).

Code availability

All custom data analysis software and bioinformatics algorithms used in this study are publically available on Github: https://github.com/mcieslik-mctp/ and https://github.com/mctp/.

31. Shalem, O. et al. Genome-scale CRISPR–Cas9 knockout screening in human cells. *Science* **343**, 84–87 (2014).
32. Phair, R. D. et al. Global nature of dynamic protein–chromatin interactions in vivo: three-dimensional genome scanning and dynamic interaction networks of chromatin proteins. *Mol. Cell. Biol.* **24**, 6393–6402 (2004).
33. Grimm, J. B. et al. A general method to improve fluorophores for live-cell and single-molecule microscopy. *Nat. Methods* **12**, 244–250 (2015).
34. Pitchiaya, S. et al. Dynamic recruitment of single RNAs to processing bodies depends on RNA functionality. *Mol. Cell* **74**, 521–533 (2019).
35. Swinstead, E. E. et al. Steroid receptors reprogram FoxA1 occupancy through dynamic chromatin transitions. *Cell* **165**, 593–605 (2016).
36. Pitchiaya, S., Androsavich, J. R. & Walter, N. G. Intracellular single molecule microscopy reveals two kinetically distinct pathways for microRNA assembly. *EMBO Rep.* **13**, 709–715 (2012).
37. Shah, N. B. & Duncan, T. M. Bio-layer interferometry for measuring kinetics of protein–protein interactions and allosteric ligand effects. *J. Vis. Exp.* **84**, e51383 (2014).
38. Teng, Y. et al. Evaluating human cancer cell metastasis in zebrafish. *BMC Cancer* **13**, 453 (2013).
39. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
40. Ramírez, F. et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).
41. Zhu, L. J. et al. ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics* **11**, 237 (2010).
42. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
43. Bailey, T. L. et al. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–W208 (2009).
44. Wilson, S. et al. Developing cancer informatics applications and tools using the NCI genomic data commons API. *Cancer Res.* **77**, e15–e18 (2017).
45. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
46. Cerami, E. et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).
47. Wu, Y.-M. et al. Inactivation of *CDK12* delineates a distinct immunogenic class of advanced prostate cancer. *Cell* **173**, 1770–1782 (2018).
48. Cieslik, M. et al. The use of exome capture RNA-seq for highly degraded RNA with application to clinical cancer sequencing. *Genome Res.* **25**, 1372–1381 (2015).
49. Robinson, D. R. et al. Integrative clinical genomics of metastatic cancer. *Nature* **548**, 297–303 (2017).
50. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84 (2014).
51. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2013).
52. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
53. Smyth, G. K., McCarthy, D. J. & Smyth, G. K. in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (eds Dudoit, S. & Carey, V. J.) 397–420 (Springer, New York, 2005).
54. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
55. Liberzon, A. et al. The molecular signatures database hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
56. Newton, M. A., Quintana, F. A., Boon, J. A. D., Sengupta, S. & Ahlquist, P. Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *Ann. Appl. Stat.* **1**, 85–106 (2007).
57. Subramanian, A., Kuehn, H., Gould, J., Tamayo, P. & Mesirov, J. P. GSEA-P: a desktop application for Gene Set Enrichment Analysis. *Bioinformatics* **23**, 3251–3253 (2007).
58. Searle, S. et al. The GENCODE human gene set. *Genome Biol.* **11**, P36 (2010).

Acknowledgements We thank D. Macha, L. Wang, S. Zelenka-Wang, I. Apel, M. Tan, Y. Qiao, A. Delekta, K. Juckette and J. Tien for technical assistance, and S. Gao for assistance with the manuscript. This work was supported by the Prostate Cancer Foundation (PCF), Early Detection Research Network (UO1 CA214170), NCI Prostate SPORE (P50 CA186786) and Stand Up 2 Cancer-PCF Dream Team (SU2C-AACR-DT0712) grants to A.M.C. A.M.C. is an NCI Outstanding Investigator, Howard Hughes Medical Institute Investigator, A. Alfred Taubman Scholar and American Cancer Society Professor. A.P. is supported by a Predoctoral Department of Defense (DoD) - Early Investigator Research Award (W81XWH-17-1-0130). M.C. is supported by a DoD - Idea Development Award (W81XWH-17-1-0224) and a PCF Young Investigator Award.

Author contributions A.P., M.C. and A.M.C. conceived and designed the study; A.P. performed all the experiments with assistance from L.X., T.O., X.W. and S.P. M.C. carried out bioinformatics analyses with assistance from A.P., Y.Z., R.J.L. and P.V. S.-C.C. and A.P. performed zebrafish in vivo experiments. A.P. is responsible for the following experimental figures: Figs. 2b–f, h, 3b–i, 4e, as well as Extended Data Figs. 1a–i, 3b–n, 4a–f, k–n, 5a–k, 6a–l, 7i–o, 8a–h, j, 9a, d, e, 10g. M.C. is responsible for the following computational figures: Figs. 1a–h, 2a, g, 3a, 4a–d, as well as Extended Data Figs. 1j–n, 2a–l, 3a, p, q, 4g–j, o–q, 7a–c, g, h, 9b, c, f–h, 10a–f. Y.Z. is responsible for the following computational figures: Extended Data Figs. 3o, r, s, 7d–f, 8i, k. F.S. and R.W. generated ChIP–seq and RNA-seq libraries. X.C. performed sequencing. F.Y.F. provided genomic validation data. Y.-M.W. and D.R.R. coordinated clinical sequencing. A.P., M.C. and A.M.C. wrote the manuscript and organized the figures.

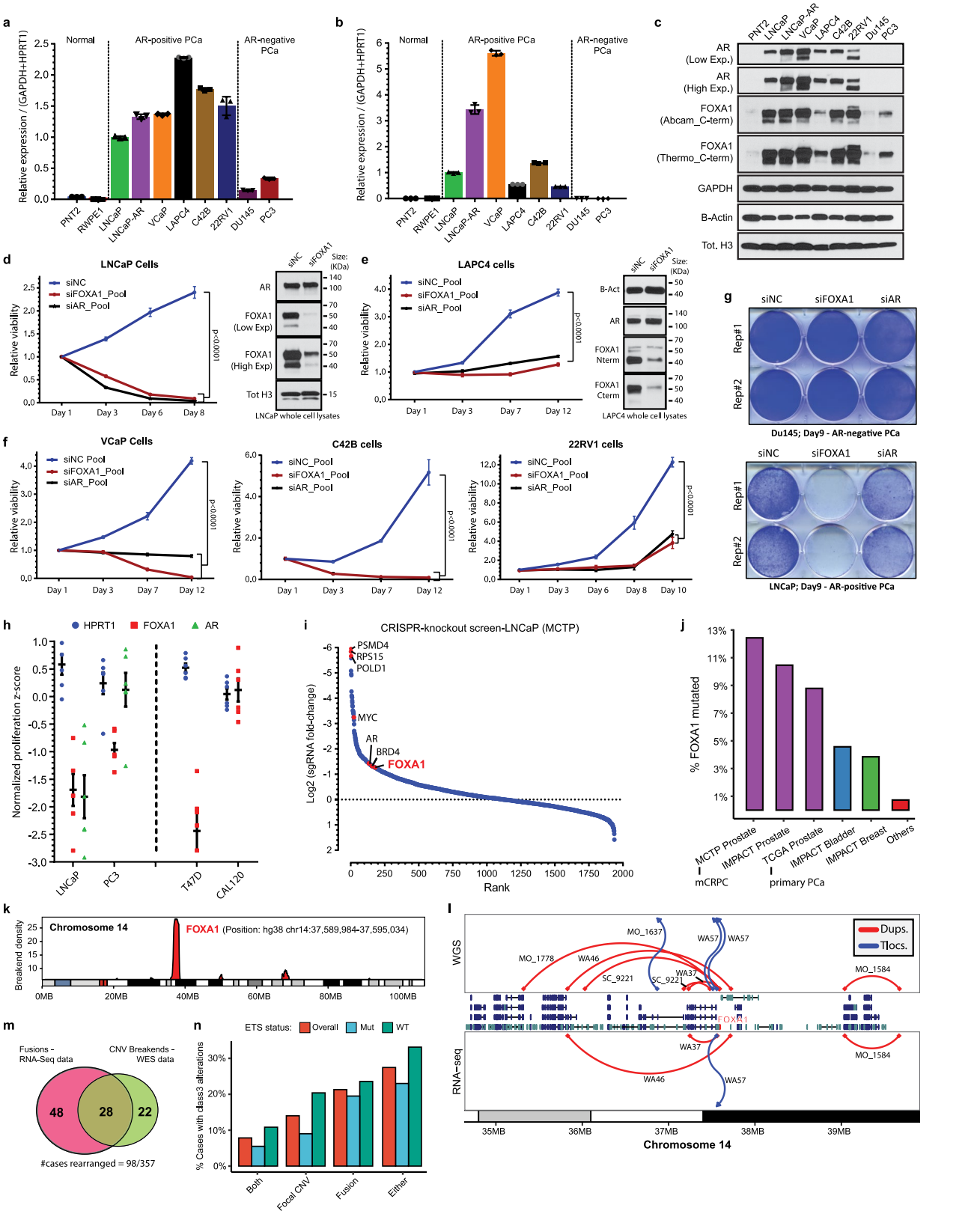
Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1347-4.

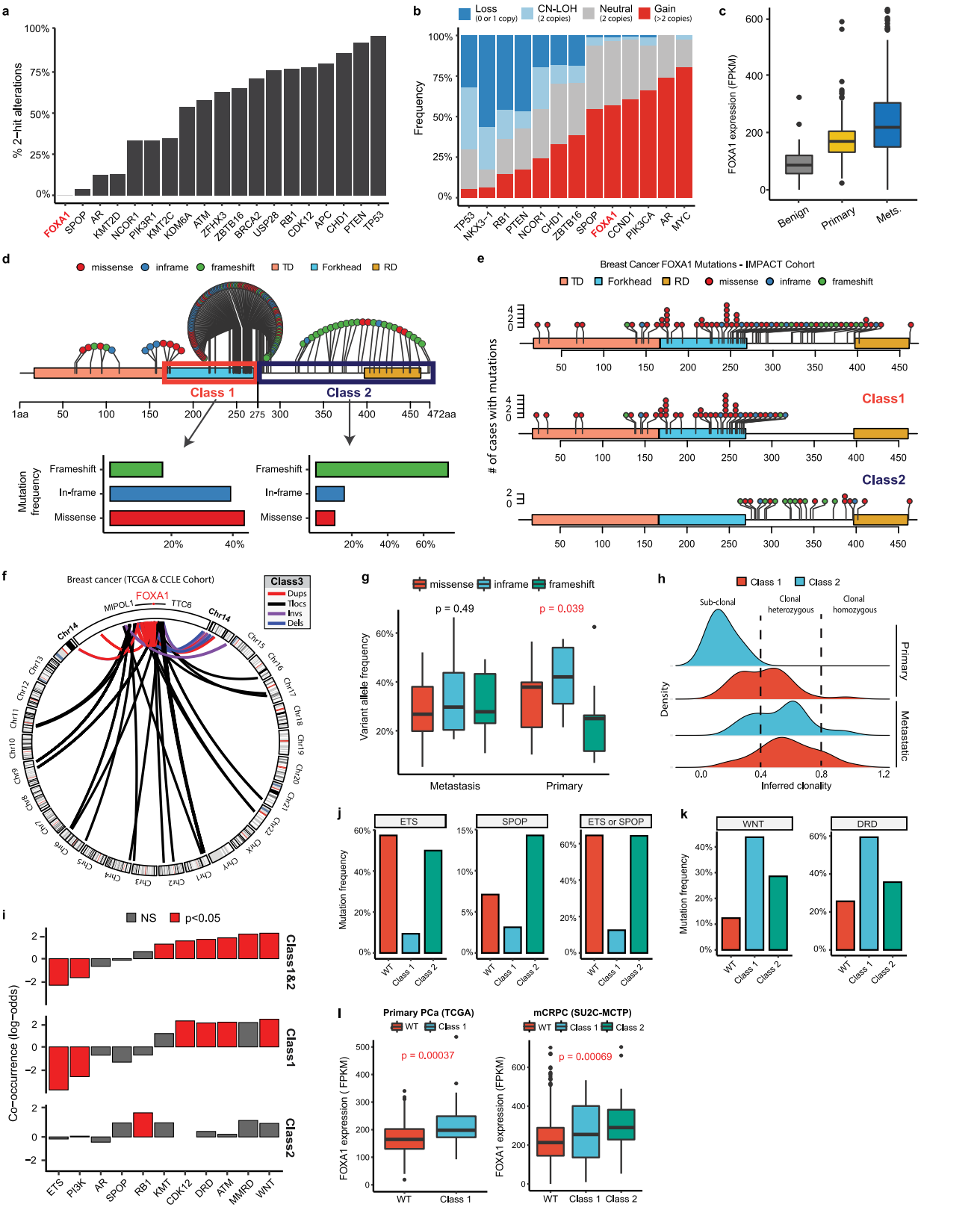
Correspondence and requests for materials should be addressed to A.M.C.

Peer review information *Nature* thanks Myles Brown, William Nelson, Mark A. Rubin and the other anonymous reviewer(s) for their contribution to the peer review of this work.



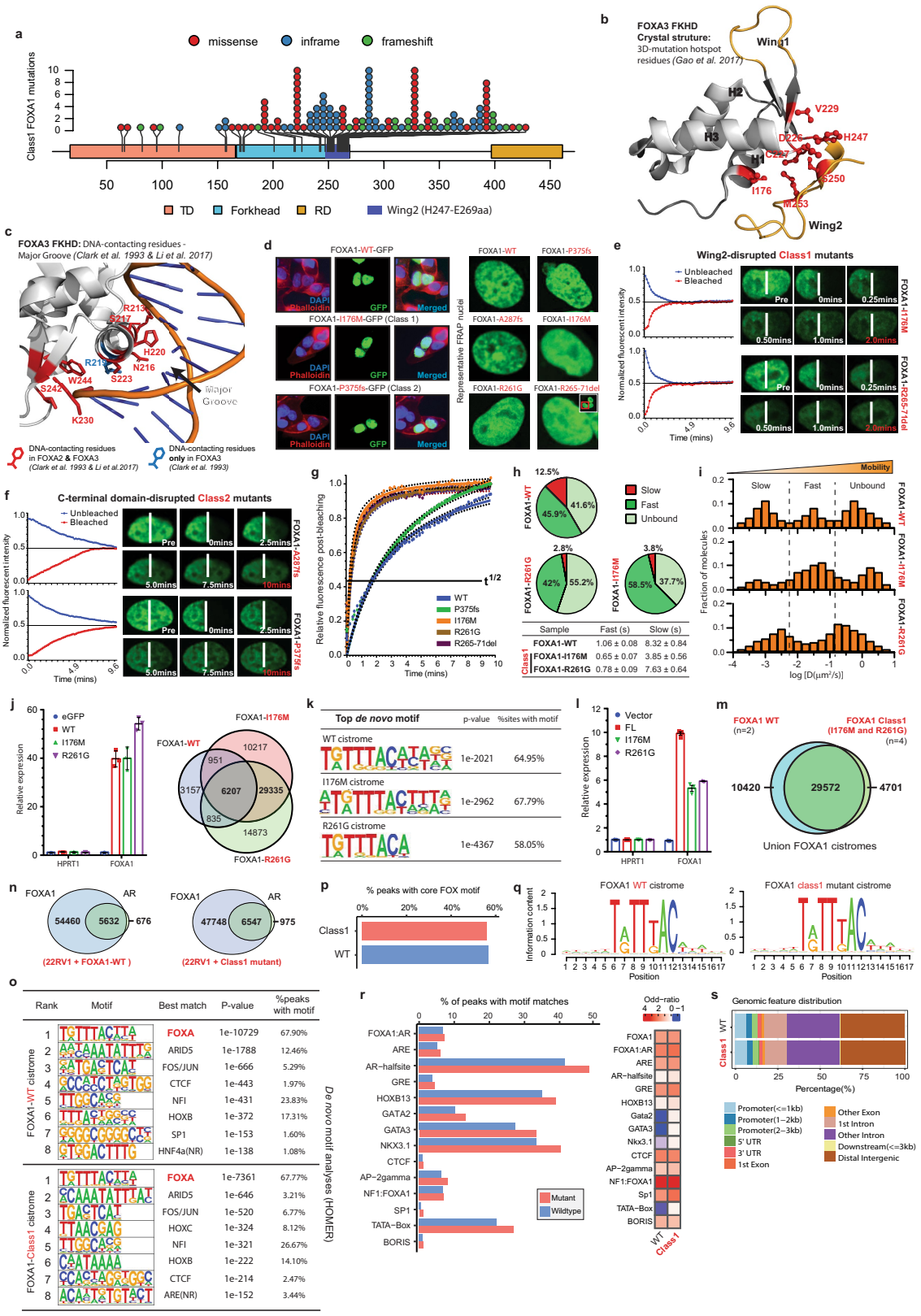
Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Functional essentiality and recurrent alterations of FOXA1 in AR$^+$ prostate cancer. a–**c**, AR (**a**) and FOXA1 (**b**) mRNA (qPCR) and (**c**) protein expression in a panel of prostate cancer cells ($n = 3$ technical replicates). Mean ± s.e.m. is shown and dots are individual data points. **d**–**f**, Growth curves of AR$^+$ prostate cancer cells treated with non-targeting control (siNC), *AR*- or *FOXA1*-targeting siRNAs (25 nM at day 0 and 1; $n = 6$ biological replicates). Immunoblots confirm knockdown of FOXA1 protein in LNCaP and LAPC4 72 h after siRNA treatment. For all gel source data, see Supplementary Fig. 1. **g**, Crystal-violet stain of AR$^-$ DU145 prostate cancer and LNCaP (control) cells treated with siNC, *AR*- or *FOXA1*-targeting siRNAs. Results represent 3 independent experiments ($n = 2$ biological replicates). **h**, Averaged proliferation *z*-scores for 6 independent *FOXA1*-targeting sgRNAs extracted from publically available CRISPR Project Achilles data (BROAD Institute) in prostate and breast cancer cells. *HPRT1* and *AR* data serve as negative and positive controls, respectively. Mean ± s.e.m. is shown; dots are proliferative *z*-scores for independent sgRNAs. **i**, Ranked depletion or enrichment of sgRNA read counts from GeCKO-V2 CRISPR knockout screen in LNCaP cells (at day 30) relative to the input sample. Only a subset of genes—including essential controls, chromatin modifiers and transcription factors—is visualized. **j**, Recurrence of *FOXA1* mutations across TCGA, MSK-IMPACT and SU2C cohorts. **k**, Density of break ends (RNA-seq chimeric junctions) within overlapping 1.5-Mb windows along chr14 in mCRPC tumours. **l**, Whole-genome sequencing (WGS) of seven mCRPC index cases with distinct patterns of *FOXA1* translocations (Tlocs) and duplications (Dups), nominated by RNA-seq (WA46, WA37, WA57 and MO_1584) or whole-exome sequencing (MO_1778, SC_9221 and MO_1637). **m**, Concordance of RNA-seq (chimeric junctions) and whole-exome-sequencing-based *FOXA1* locus rearrangements calls (mCRPC cohort). CNV, copy-number variation. **n**, Frequency of *FOXA1* locus rearrangements in mCRPC based on RNA-seq and whole-exome sequencing.



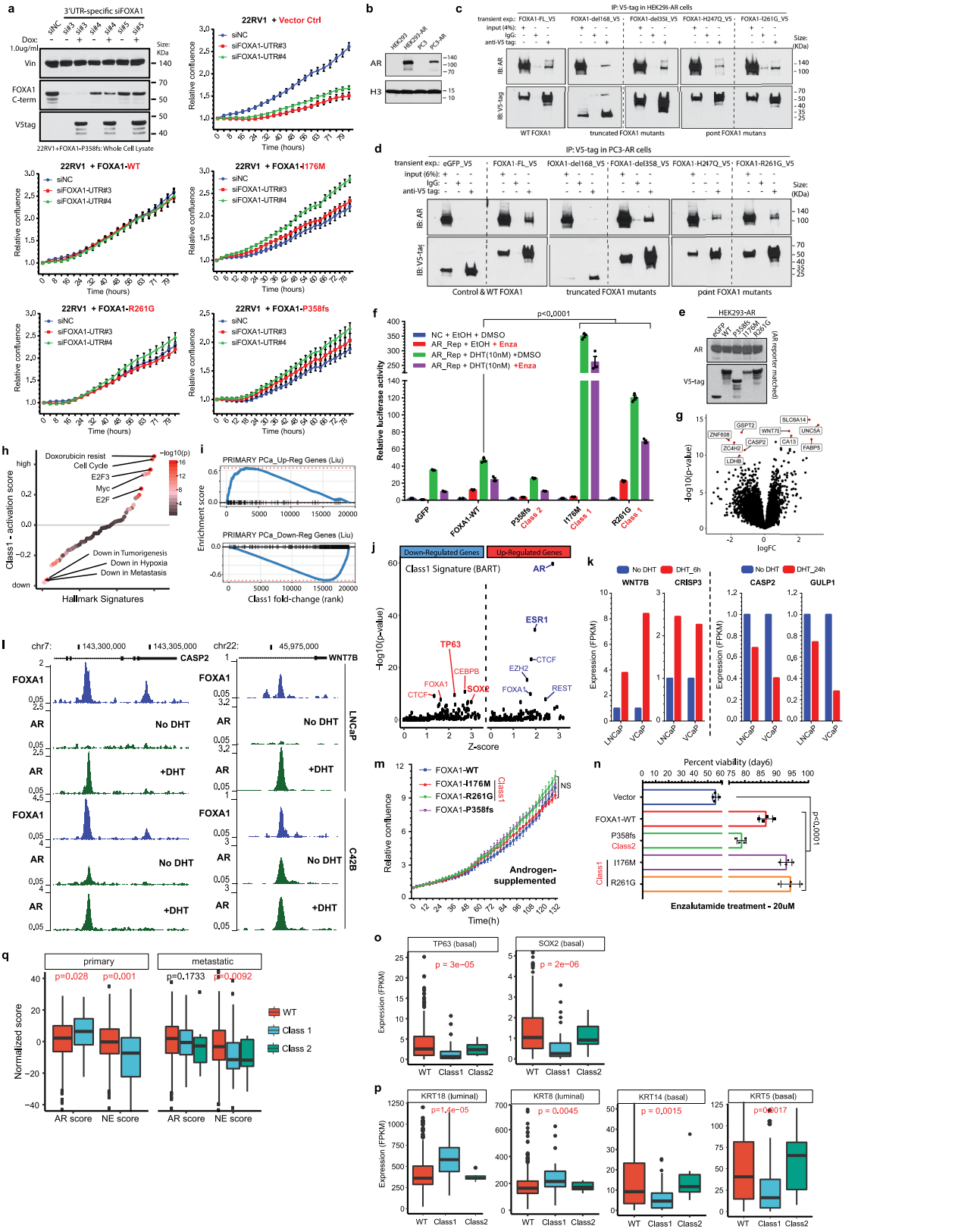
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Genomic characteristics of the three classes of FOXA1 alterations in prostate and breast cancer. a, **b**, Bi-allelic inactivation (**a**) and copy-number variations (**b**) of *FOXA1* across mCRPC ($n = 371$). CN-LOH, copy-neutral loss of heterozygosity. **c**, FOXA1 expression (RNA-seq) in benign ($n = 51$), primary ($n = 501$) and metastatic ($n = 535$) prostate cancer. **d**, Distribution and functional categorization of *FOXA1* mutations (all cases in the aggregate cohort) on the protein map of FOXA1. **e**, Aggregate and class-specific distribution of *FOXA1* mutations in advanced breast cancer (MSK-IMPACT cohort). **f**, Structural classification of *FOXA1* locus rearrangements in breast cancer (TCGA and CCLE cell lines). **g**, **h**, Variant allele frequency of *FOXA1* mutations by tumour stage (**g**) and clonality estimates of class-1 and class-2 mutations (**h**) in tumour-content-corrected primary prostate cancer ($n = 500$) and mCRPC ($n = 370$) specimens. **i**, Mutual exclusivity or co-occurrence of *FOXA1* mutations (two-sided Fisher's exact test). Mutations in AR, WNT, and PI3K were aggregated at the pathway level. ETS, ETS gene fusions; DRD, DNA repair defects and included alterations in *BRCA1*, *BRCA2*, *ATM* and *CDK12*; MMRD, mismatch repair deficiency (total $n = 371$). **j**, Mutual exclusivity of ETS and/or SPOP ($n = 26$) alterations with FOXA1 ($n = 46$) alterations distinguished by class in mCRPC ($n = 371$). **k**, Co-occurrence of WNT ($n = 58$) and DRD ($n = 107$) pathway alterations with FOXA1 alteration classes in mCRPC ($n = 371$). **l**, Stage- and class-specific increase in FOXA1 expression levels in primary ($n = 500$) and metastatic prostate cancer ($n = 357$). Left, two-sided *t*-test. Right, two-way ANOVA. For all box plots, centre shows median, box marks quartiles 1–3 and whiskers span quartiles 1–3 $\pm$ 1.5 $\times$ IQR.



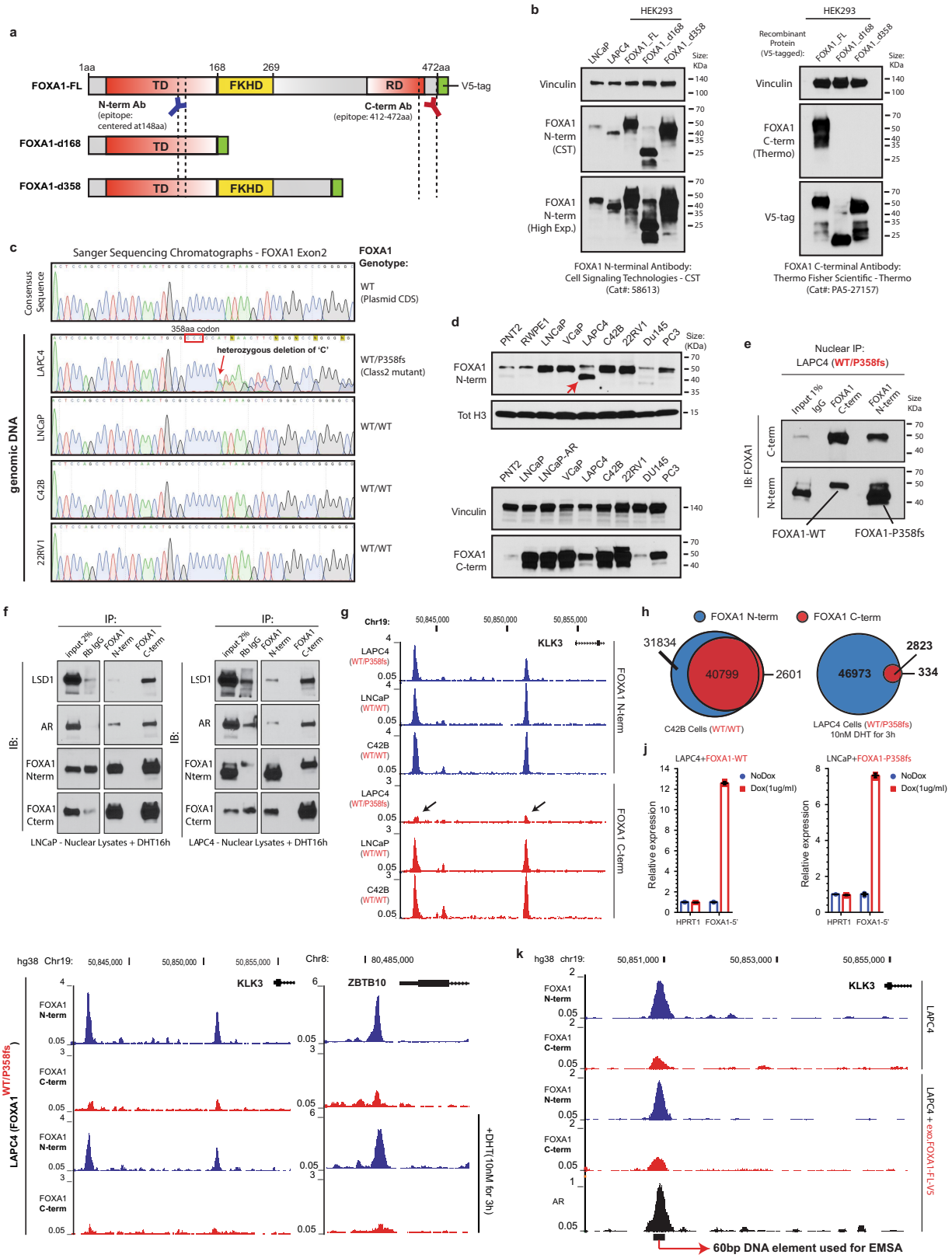
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Biophysical and cistromic characteristics of the class-1 FOXA1 mutants. a, Distribution of class-1 mutations on the protein map of FOXA1. **b**, Three-dimensional structure of FKHD (FOXA3) with visualization of all mutated residues collectively identified as the 3D-mutational hotspot in FOXA1 across cancers. **c**, DNA-bound 3D structure of FKHD with visualization of all residues shown through crystallography to make direct base-specific contacts with the DNA in FOXA2 and FOXA3 proteins. **d**, Representative fluorescent images of nuclei expressing different variants of FOXA1 fused to GFP at the C termini. **e**, **f**, FRAP kinetic plots (left) and representative time-lapse images (right) from pre-bleaching (pre) to 100% recovery (red timestamps) for wing-2-altered class-1 mutants (**e**) and truncated class-2 mutants (that is, A287fs and P375fs) (**f**) ($n = 6$ nuclei per variant; quantified in Fig. 2d). White lines indicate the border between bleached and unbleached areas. **g**, Representative FRAP kinetics in the bleached area for indicated FOXA1 variants. $t_{1/2}$ line indicates the time to 50% recovery. Coloured dots show raw data; superimposed solid curves show a hyperbolic fit with 95% confidence intervals. **h**, Single particle tracking quantification of chromatin-bound (slow and fast) and unbound (freely diffusing) particles of wild-type and class-1 FOXA1 variants, and average chromatin dwell times (mean ± s.d.) for the bound fractions ($n \geq 500$ particles per variant). **i**, Diffusion constant histograms of single particles of wild-type or distinct class-1 FOXA1 mutants. Particles were categorized into chromatin-bound (slow and fast) or unbound fractions using cut-offs marked by dashed lines ($n \geq 500$ particles per variant imaged in 3–5 distinct nuclei). **j**, Left, mRNA expression (qPCR) of labelled FOXA1 variants in stable, isogenic HEK293 cells ($n = 3$ technical replicates). Right, overlaps between FOXA1 wild-type and class-1 mutant cistromes from these cells ($n = 2$ biological replicates). **k**, Top de novo motifs identified from the three FOXA1 cistromes from HEK293 cells (HOMER, hypergeometric test). **l**, mRNA expression (qPCR) of labelled FOXA1 variants in stable, isogenic 22RV1 cells ($n = 3$ technical replicates). For **j** and **l**, centres show mean values and lines mark s.e.m. **m**, Overlap between wild-type ($n = 2$ biological replicates) and class-1 ($n = 4$ biological replicates) cistromes from stable 22RV1 overexpression models. **n**, Overlap between the FOXA1 wild-type and AR union cistromes generated from 22RV1 cells overexpressing wild-type ($n = 2$ biological replicates) or class-1 mutant (I176M or R216G; $n = 2$ biological replicates each) FOXA1 variants. **o**, De novo motif results for the wild-type or class-1 mutant FOXA1-binding sites from prostate cancer cells (HOMER, hypergeometric test). **p**, **q**, Per cent of wild-type or class-1 binding sites with perfect match to the core FOXA1 motif (5′-T[G/A]TT[T/G]AC-3′) (**p**) and the consensus FOXA1 motifs identified from these sites (**q**). **r**, Left, per cent of wild-type or class-1 binding sites containing known motifs of the labelled FOXA1 or AR cofactors. Right, enrichment of the cofactor motifs in the two cistromes relative to the background ($n =$ top 5,000 peaks by score for each variant, see Methods). **s**, Genomic distribution of wild-type and class-1 binding sites in prostate cancer cells.



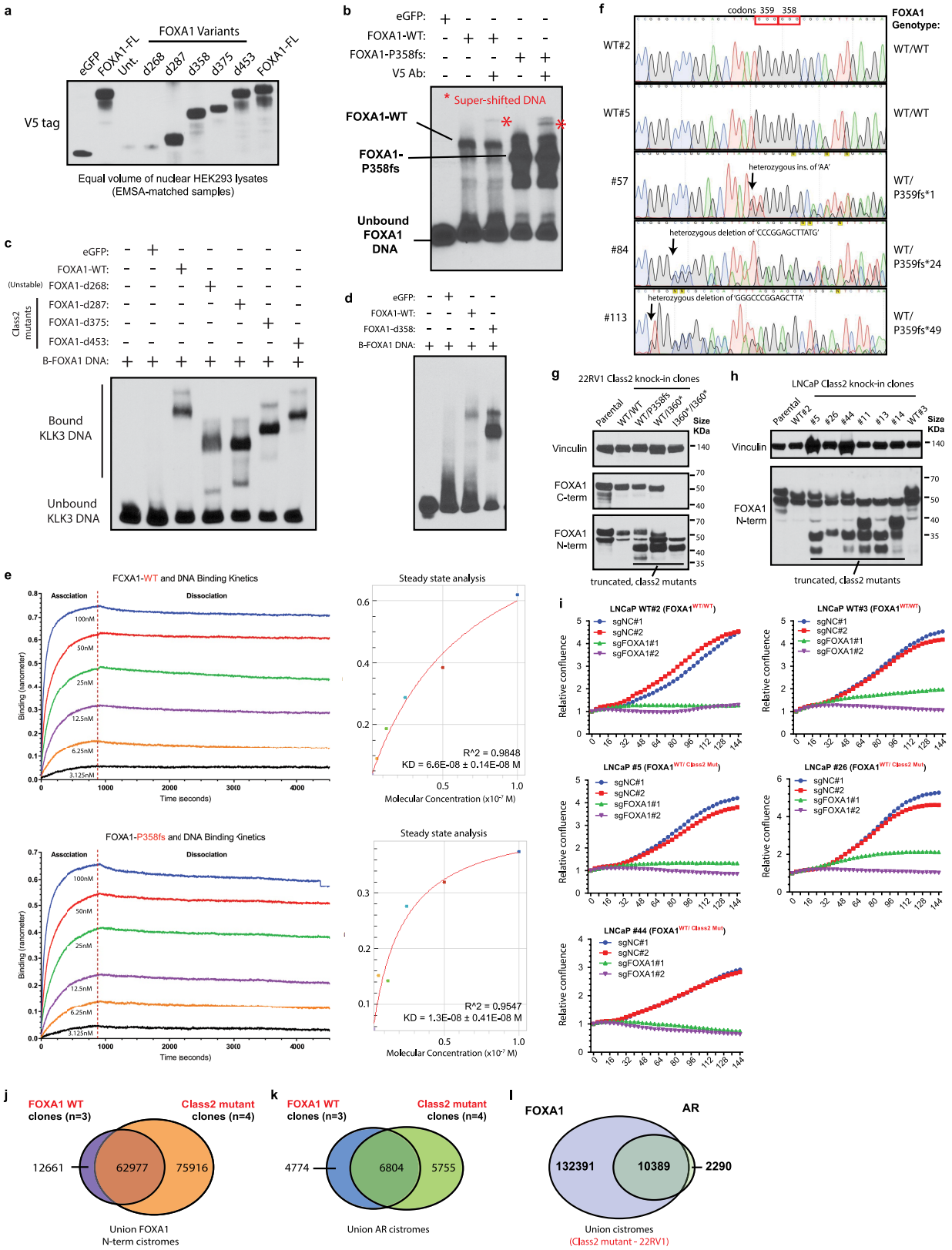
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Functional effect of *FOXA1* mutations on oncogenic AR signalling. a, Immunoblot showing expression of endogenous and V5-tagged exogenous FOXA1 proteins in doxycycline (dox)-inducible 22RV1 cells transfected with distinct UTR-specific *FOXA1*-targeting siRNAs (no. 3–5) or a non-targeting control siRNA (siNC). These results represent two independent experiments. IncuCyte growth curves of 22RV1 cells overexpressing empty vector (control), wild-type or mutant FOXA1 variants upon treatment with UTR-specific *FOXA1*-targeting siRNAs ($n = 5$ biological replicates). Mean ± s.e.m. is shown. **b**, Immunoblots confirming stable overexpression of the wild-type AR protein in HEK293 and PC3 cells. **c**, **d**, Co-immunoprecipitation assay of indicated recombinant FOXA1 variants using a V5-tag antibody in HEK293 (**c**) and PC3 (**d**) cells stably overexpressing the AR protein (referred to as HEK293-AR and PC3-AR cells). eGFP is a negative control. FOXA1-FL, full-length wild-type FOXA1. del168 and del358 are truncated FOXA1 variants with only the first 168 amino acids (that is, before the FKHD) or 358 amino acids of the FOXA1 protein. H247Q and R261G are missense class-1 mutant variants. **e**, Immunoblots confirming comparable expression of AR and recombinant FOXA1 variants in AR reporter assay-matched HEK293 lysates. Immunoblots show representative results from 2 or 3 independent experiments and class-1 and class-2 mutants serve as biological replicates. For all gel source data (**a**, **b**–**e**), see Supplementary Fig. 1. **f**, AR dual-luciferase reporter assays with transient overexpression of indicated FOXA1 variants in HEK293-AR cells with or without DHT stimulation and enzalutamide treatment ($n = 3$ biological replicates per group). Mean ± s.e.m. is shown (two-way ANOVA and Tukey's test). **g**, Genes differentially expressed in class-1 tumours from patients ($n = 38$) compared to FOXA1 wild-type tumours (see Methods). The most significant genes are shown in red and labelled (limma two-sided test). **h**, Differential expression of cancer-hallmark signature genes in class-1 mutant prostate-cancer tumours (GSEA statistical test). **i**, Localized, primary prostate cancer gene signature showing concordance between class-1 tumour and primary prostate cancer genes. **j**, BART prediction of specific transcription factors mediating observed transcriptional changes. The significant and strong (z-score) mediators of transcriptional responses in class-1 tumours are labelled (BART, Wilcoxon rank-sum test). **k**, mRNA expression (RNA-seq) of class-1 signature genes in LNCaP and VCaP cells either starved for androgen (no DHT) or stimulated with DHT (10 nM). RNA-seq from two distinct prostate cancer cell lines is shown. **l**, Representative FOXA1 and AR ChIP–seq normalized signal tracks at the *WNT7B* or *CASP2* gene loci in LNCaP and C42B cells. ChIP–seq assays were carried out in two distinct prostate cancer cell lines with similar results. **m**, Growth curves (IncuCyte) of 22RV1 cells overexpressing distinct FOXA1 variants in complete, androgen-supplemented growth medium ($n = 2$ biological replicates). Mean ± s.e.m. is shown. **n**, Per cent viable 22RV1 stable cells, overexpressing either empty vector, wild-type or mutant FOXA1 variants upon treatment with enzalutamide (20 μM for 6 days; $n = 4$ biological replicates). Mean ± s.e.m. is shown. P values in **m** and **n** were calculated using two-way ANOVA and Tukey's test. **o**, **p**, mRNA expression (RNA-seq) of labelled basal and luminal transcription factors or canonical markers in FOXA1 wild-type, class-1 or class-2 mutant tumours in primary prostate cancer (total $n = 500$; two-way ANOVA). **q**, Extent of AR and neuroendocrine (NE) pathway activation in FOXA1 wild-type, class-1 or class-2 mutant cases from both primary ($n = 500$) and metastatic ($n = 370$) prostate cancer. Both AR and NE scores were calculated using established gene signatures (see Methods). Left, two-sided t-test; right, two-way ANOVA. For all box plots, centre shows median, box marks quartiles 1–3 and whiskers span quartiles 1–3 ± 1.5 × IQR.



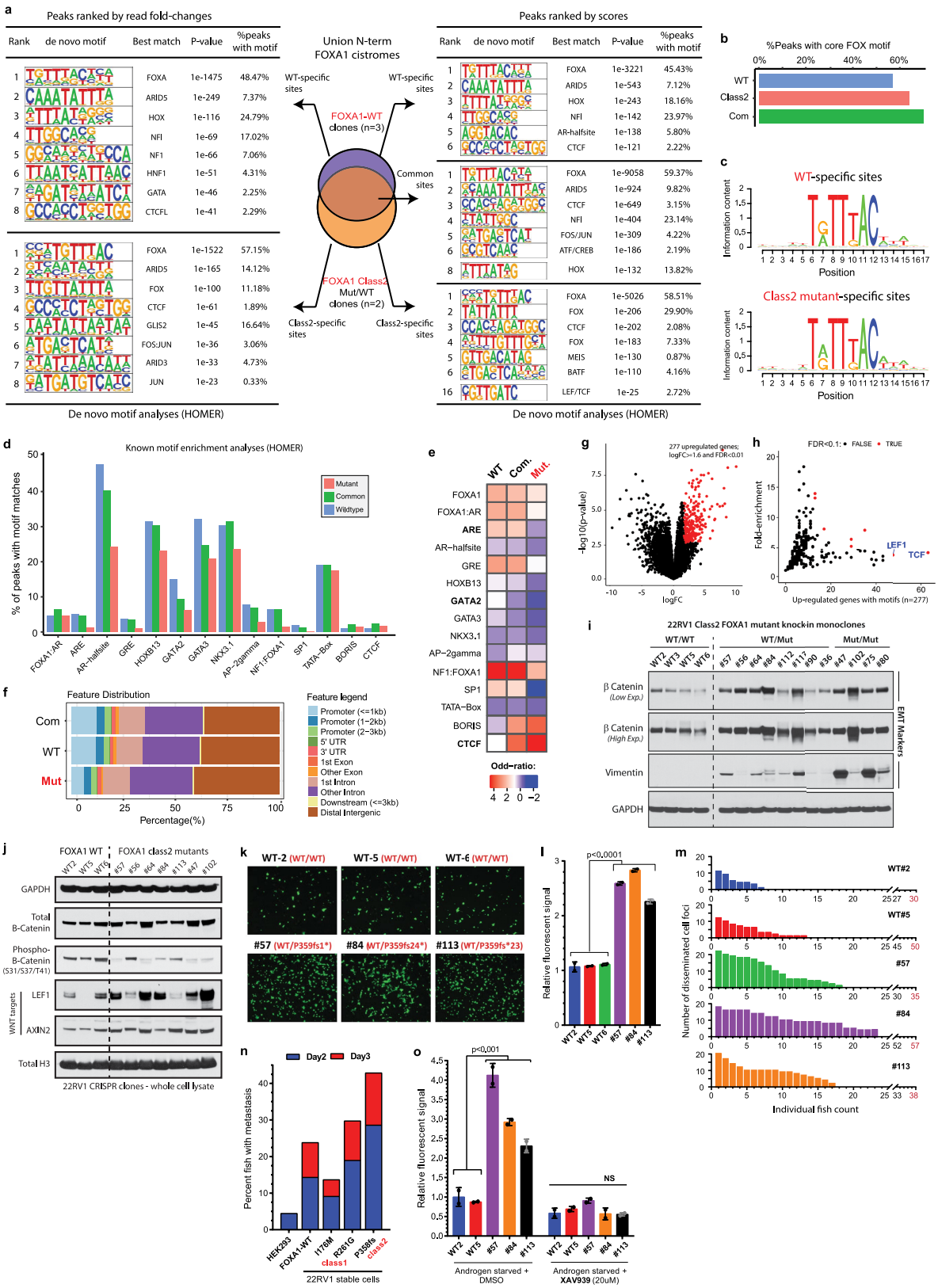
Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | DNA-binding dominance of the class-2 FOXA1 mutants. a, FOXA1 protein maps showing the recombinant proteins used to validate the N-terminal (N-term) and C-terminal (C-term) FOXA1 antibodies. **b**, Immunoblots depicting detection of all variants by the N-terminal antibody (left), and of only the full-length wild-type FOXA1 protein by the C-terminal antibody (right). These results were reproducible in two independent experiments. Antibody details are included in the Methods. **c**, Sanger sequencing chromatograms showing the heterozygous class-2 mutation in LAPC4 cells after the P358 codon in exon 2 ($n = 2$ technical replicates). All other tested prostate cancer cell lines were wild type for FOXA1. **d**, Immunoblots confirming the expression of the truncated FOXA1 variant in LAPC4 at the expected approximately 40-kDa size (top, red arrow). The short band is detectable only with the N-terminal (top) FOXA1 antibody and not the C-terminal (bottom) antibody. These results were reproducible in two independent experiments. **e**, Co-immunoprecipitation and immunoblotting of FOXA1 using N-terminal and C-terminal antibodies from LAPC4 nuclei with species-matched IgG used as control. **f**, Nuclear co-immunoprecipitation of FOXA1 from LAPC4 or LNCaP cells stimulated with DHT (10 nM for 16 h) using N-terminal and C-terminal antibodies. Species-matched IgG are controls. Immunoprecipitations and immunoblots in **d**–**f** were reproducible in two and three independent experiments, respectively. For gel source data (**b**, **d**, **e**, **f**), see Supplementary Fig. 1. **g**, FOXA1 N-terminal and C-terminal ChIP–seq normalized signal tracks from FOXA1 wild-type or class-2 mutant prostate cancer cells at canonical AR target *KLK3*. **h**, Left, overlap between global N-terminal and C-terminal FOXA1 cistromes in untreated C42B cells. Right, overlap between global N-terminal and C-terminal FOXA1 cistromes in LAPC4 cells treated with DHT (10 nM for 3 h). **i**, FOXA1 ChIP–seq normalized signal tracks from N-terminal and C-terminal antibodies in LAPC4 cells with or without DHT stimulation (10 nM for 3 h) at *KLK3* and *ZBTB10* loci. ChIP–seq assays in **g** and **i** were carried out in two distinct FOXA1 wild-type prostate cancer cells. For LAPC4 ChIP–seq experiments, results were reproducible in two independent experiments. **j**, mRNA (qPCR) expression of *FOXA1* in LAPC4 cells with exogenous overexpression of wild-type FOXA1 (left), and in LNCaP cells with exogenous overexpression of the P358fs mutant (right) ($n = 3$ technical replicates). Mean ± s.e.m. is shown and dots are individual data values. **k**, FOXA1 ChIP–seq normalized signal tracks from N-terminal and C-terminal antibodies in parental LAPC4 cells and LAPC4 cells overexpressing wild-type FOXA1 at the *KLK3* locus. This experiment was independently repeated twice with similar results. The 60-bp AR- and FOXA1-bound *KLK3* enhancer element used for electrophoretic mobility shift assay (EMSA) is shown.



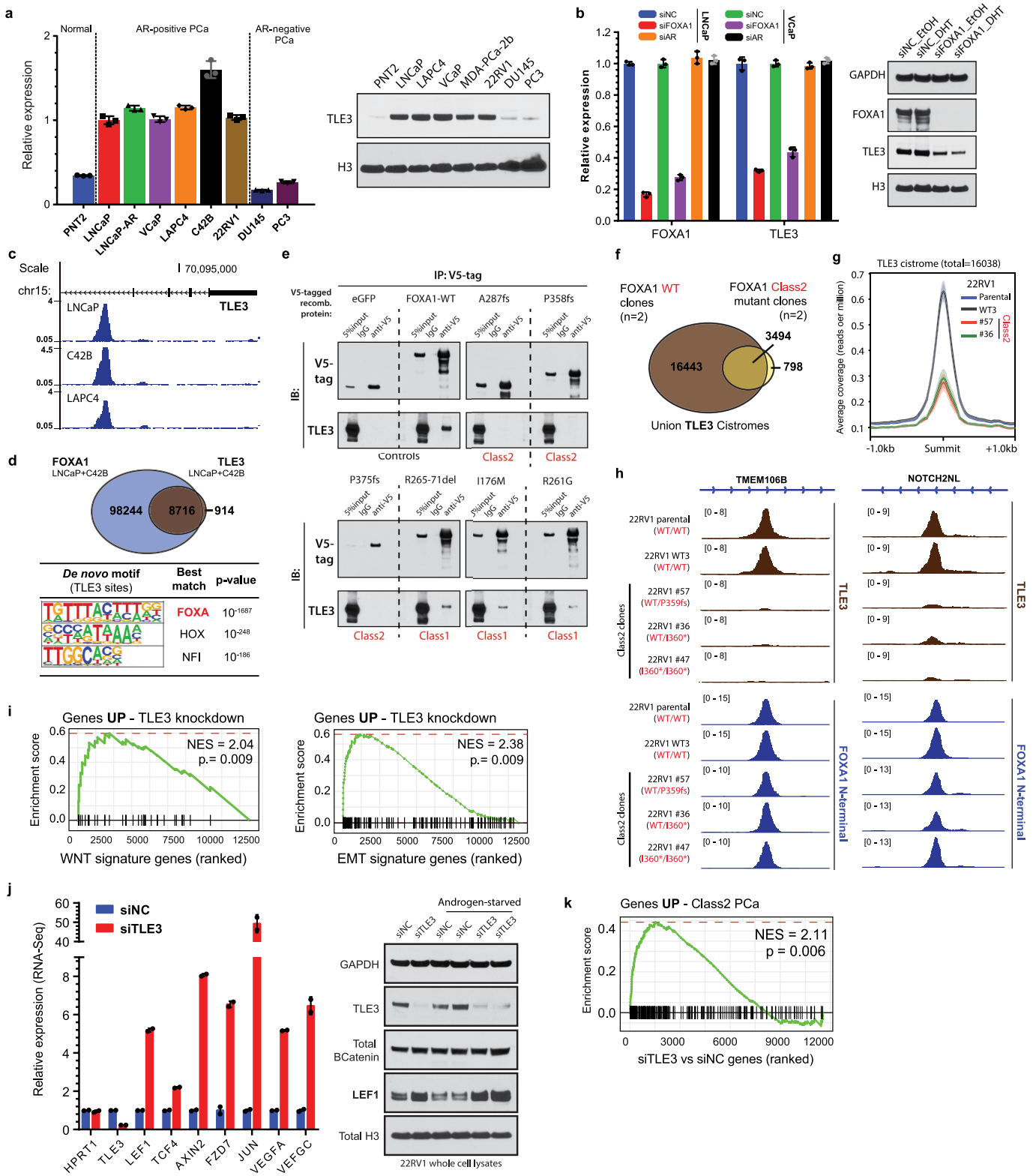
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | DNA-binding affinity and functional essentiality of the class-2 FOXA1 mutants. a, Immunoblot showing comparable expression of recombinant FOXA1 variants in equal volume of nuclear HEK293 lysates used to perform EMSAs. **b**, Higher exposure of EMSA with recombinant wild-type or P358fs mutant and *KLK3* enhancer element, showing the super-shifted band with addition of the V5 antibody (red asterisks; matched to Fig. 3f). **c**, **d**, EMSA with recombinant wild-type or different class-2 mutants (truncated at 268, 287, 358, 375 and 453 amino acids) and *KLK3* enhancer element. Class-2 mutants display higher affinity than wild-type FOXA1. Each class-2 mutant serves as a biological replicate and these results were reproducible in two independent experiments. **e**, DNA association and dissociation kinetics at varying concentrations of purified wild-type or P358fs class-2 FOXA1 mutants from the biolayer-interferometry assay performed using OctetRED system. Overall binding curves and equilibrium dissociation constants (mean ± s.d.) are shown. These results were reproducible in two independent experiments. **f**, Sanger sequencing chromatograms from a set of 22RV1 CRISPR clones confirming the introduction of distinct indels in the endogenous *FOXA1* allele, resulting in a premature stop codon ($n = 2$ technical replicates). Protein mutations are identified on the right. **g**, Immunoblots showing the expression of endogenous wild-type or class-2 mutant FOXA1 variants in parental and distinct CRISPR-engineered 22RV1 clones. **h**, Immunoblots showing expression of FOXA1 (N-terminal antibody) in parental and CRISPR-engineered LNCaP clones expressing distinct class-2 mutants with truncations closer to the FKHD domain. For gel source data (**a**–**d**, **g**, **h**), see Supplementary Fig. 1. **i**, Growth curves of wild-type or mutant clones upon treatment with the non-targeting or *FOXA1*-targeting sgRNAs and CRISPR–Cas9 protein (see Methods). For **i**, distinct class-2 clones and distinct sgRNAs serve as biological replicates. **j**, **k**, Overlap between union FOXA1 (**j**) and AR (**k**) cistromes from wild-type ($n = 3$ biological replicates) and class-2-mutant ($n = 4$ biological replicates) 22RV1 clones. **l**, Overlap between union FOXA1 and AR cistromes from class-2 mutant 22RV1 cells.



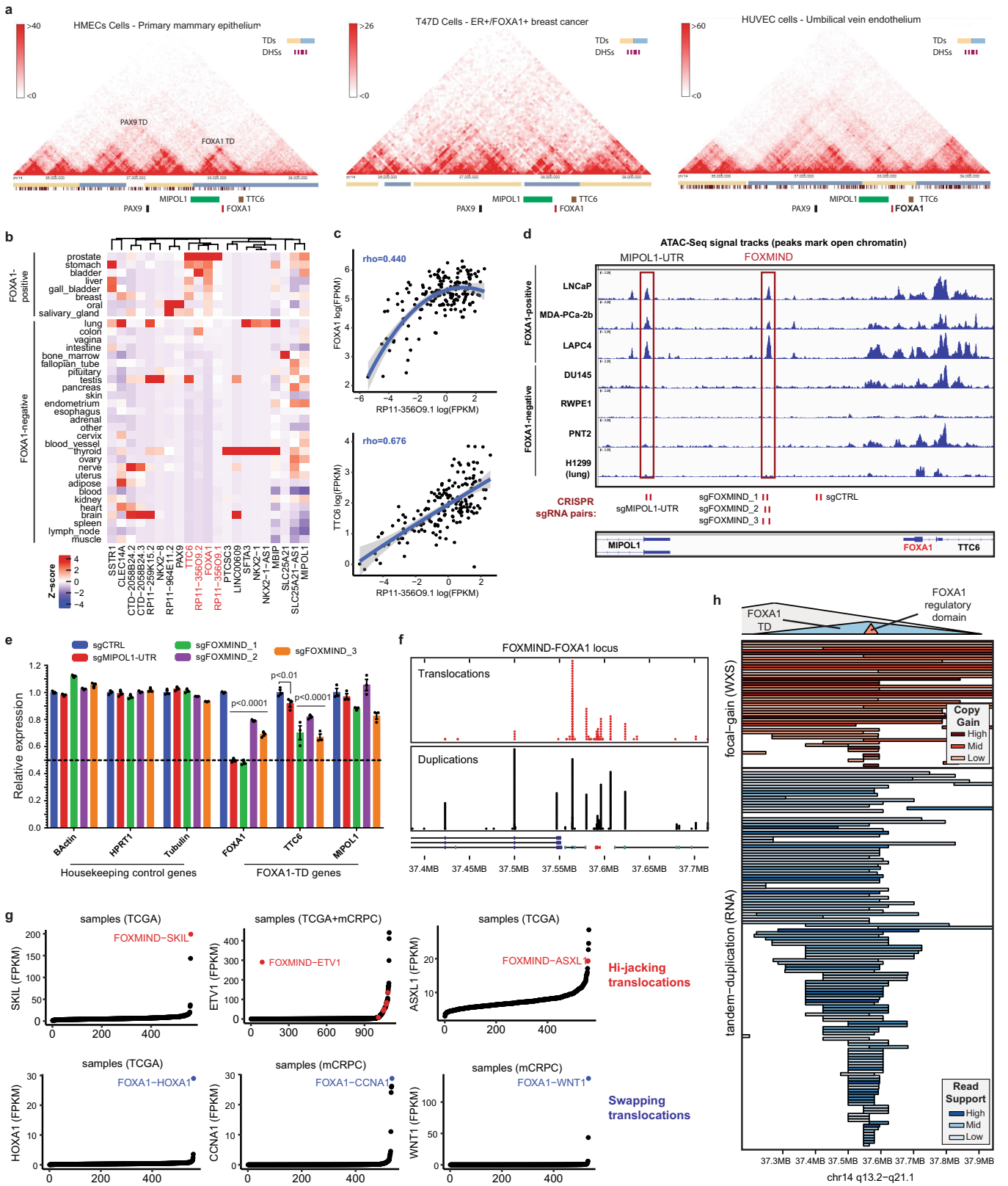
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Cistromic and WNT-driven phenotypic characteristics of the class-2 FOXA1 mutants. a, De novo motif analyses of the wild-type-specific, common and class-2-specific FOXA1-binding site subsets defined from either sequencing-read fold changes (left) or peak-calling scores (right) of ChIP–seq data. Wild-type and class-2 cistromes were generated from $n = 3$ and $n = 2$ independent biological replicates, respectively. Only the top 5,000 or 10,000 peaks from each subset were used as inputs for motif discovery (see Methods) (HOMER, hypergeometric test). **b**, **c**, Per cent of wild-type or class-2 binding sites with perfect match to the core FOXA1 motif (5′-T[G/A]TT[T/G]AC-3′) (**b**) and the consensus FOXA1 motifs identified from these sites (**c**). **d**, **e**, Per cent of binding sites in the three FOXA1-binding-site subsets containing known motifs of the labelled FOXA1 or AR cofactors (**d**), and enrichment of the cofactor motifs in the three binding site subsets relative to the background (**e**). **f**, Genomic distribution of wild-type-specific, common and class-2-specific binding sites in prostate cancer cells. **g**, Differential expression of genes in FOXA1 class-2 mutant CRISPR clones relative to FOXA1 wild-type clones ($n = 2$ biological replicates (limma two-sided test)). **h**, Distinct transcription factor motifs within the promoter (2-kb upstream) of differentially expressed genes. Transcription factors with the highest enrichment (fold change, per cent of upregulated genes with the motif and significance) are highlighted and labelled (two-tailed Fisher's exact test). **i**, Immunoblots showing the expression of β-catenin and vimentin in a panel of wild-type and heterozygous or homozygous class-2 mutant 22RV1 CRISPR clones. **j**, Immunoblots showing the phosphorylation status of β-catenin and expression of direct WNT target genes in select class-2 mutant 22RV1 clones. Immunoblots in **i** and **j** are representative of two independent experiments; every individual clone serves as a biological replicate. For gel source data, see Supplementary Fig. 1. **k**, Representative images of Boyden chambers showing invaded cells stained with calcein AM dye. **l**, Quantified fluorescence signal from invaded cells ($n = 2$ biological replicates per group; two-way ANOVA and Tukey's test). Mean ± s.e.m. is shown and dots are individual data points. **m**, Absolute counts of disseminated cell foci in individual zebrafish embryos as a measure of metastatic burden. **n**, Per cent metastasis at day 2 and day 3 in zebrafish embryos injected with either the normal HEK293 cells (negative controls) or 22RV1 prostate cancer cells virally overexpressing wild-type, class-1 or class-2 mutant FOXA1 variants ($n > 20$ for each group). **o**, Fluorescent signal from the invaded wild-type or class-2-mutant 22RV1 cells after androgen starvation (5% charcoal-stripped serum medium for 72 h) or treatment with the WNT inhibitor XAV939 (20 μM for 24 h; $n = 2$ biological replicates per group; two-way ANOVA and Tukey's test). Mean ± s.e.m. and individual data points are shown.



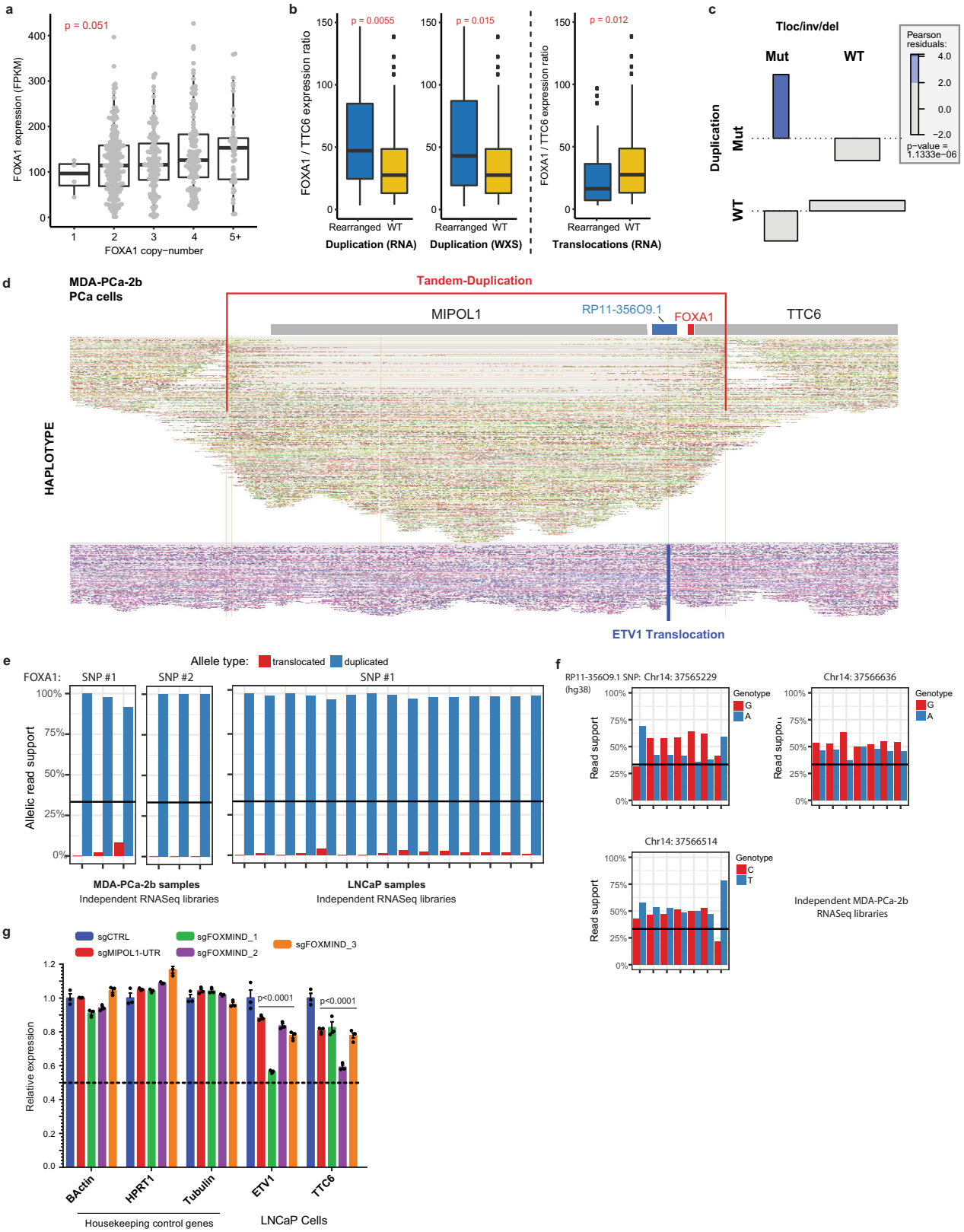
Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Functional association of FOXA1 and TLE3 in prostate cancer. a, mRNA (qPCR) and protein (immunoblot) expression of TLE3 in a panel of prostate cancer cells. Mean ± s.e.m. and individual data points are shown. **b**, Left, mRNA expression of *FOXA1* and *TLE3* in LNCaP and VCaP cells treated with siRNAs targeting either *FOXA1* or *AR* ($n = 3$ technical replicates). Two FOXA1 wild-type prostate cancer cells serve as biological replicates. Mean ± s.e.m. and individual data points are shown. Right, protein expression of FOXA1 and TLE3 in matched LNCaP lysates. **c**, FOXA1 N-terminal ChIP–seq normalized signal tracks from LNCaP, C42B and LAPC4 prostate cancer cells at the *TLE3* locus. Each cell line serves as a biological replicate. **d**, Overlap of the union wild-type FOXA1- and TLE3-binding sites from LNCaP and C42B prostate cancer cells ($n = 1$ for each), and top de novo motifs discovered (HOMER, hypergeometric test) in the TLE3 cistrome. **e**, Co-immunoprecipitation assays of labelled recombinant FOXA1 wild-type, class-1 or class-2 variants using a V5-tag antibody in HEK293 cells overexpressing the TLE3 protein. V5-tagged GFP protein was used as a negative control. These results were reproducible in two independent experiments and distinct class-1 and class-2 mutants serve as biological replicates. **f**, Overlap of union TLE3 cistromes from isogenic wild-type ($n = 2$ biological replicates) or heterozygous class-2-mutant ($n = 2$ biological replicates) 22RV1 CRISPR clones. **g**, ChIP peak profile plots from TLE3 ChIP–seq in isogenic FOXA1 wild-type or class-2-mutant 22RV1 clones ($n = 2$ biological replicates each). **h**, Representative TLE3 and FOXA1 ChIP–seq read signal tracks from independent 22RV1 CRISPR clones with or without endogenous *FOXA1* class-2 mutation ($n = 2$ biological replicates each). **i**, GSEA showing significant enrichment of WNT (left) and EMT (right) pathway genes in 22RV1 cells treated with *TLE3*-targeting siRNAs ($n = 2$ biological replicates for each treatment; GSEA enrichment test). **j**, Left, mRNA (RNA-seq) expression of direct WNT target genes in 22RV1 upon siRNA-mediated knockdown of *TLE3* ($n = 2$ biological replicates). Right, Immunoblot showing LEF1 upregulation upon TLE3 knockdown in 22RV1 prostate cancer cells with and without androgen starvation (representative of two independent experiments). For gel source data (**a**, **b**, **e**, **j**), see Supplementary Fig. 1. **k**, Gene enrichment plots showing significant enrichment of class-2 upregulated genes upon *TLE3* knockdown in 22RV1 cells ($n = 2$ biological replicates for each treatment; GSEA enrichment test).



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Topological, physical and transcriptional characteristics of the *FOXA1* locus in normal tissues and prostate cancer. a, HI-C data (from: http://promoter.bx.psu.edu/hi-c/view.php) depicting conserved topological domains within the *PAX9* and *FOXA1* syntenic block in normal and FOXA1$^+$ cancer cell lines. DHSs, DNase I hypersensitive sites. **b**, Highly tissue-specific patterns of gene expression within the *PAX9* and *FOXA1* syntenic block. Tissues were dichotomized into FOXA1$^+$ and FOXA1$^-$ on the basis of FOXA1 expression levels; genes were subject to unsupervised clustering. *z*-score normalization was performed for each gene across all tissues. **c**, Correlation of RP11-356O9.1 (Methods) and *FOXA1* or *TTC6* expression levels across metastatic tissues ($n = 370$; Spearman's rank correlation coefficient). The 95% confidence interval is shown. **d**, Representative ATAC-seq ($n = 1$) read signal tracks from normal basal epithelial prostate (RWPE1 and PNT2 cells) or prostate cancer cells. Cells are grouped on the basis of expression of FOXA1, and differentially pioneered loci are marked with red boxes. CRISPR sgRNA pairs used for genomic deletion of the labelled elements are shown at the bottom. Distinct FOXA1$^+$ and FOXA1$^-$ cell lines serve as biological replicates for ATAC-seq. **e**, mRNA (qPCR) expression of housekeeping control genes, genes located within the *FOXA1* topologically associated domain, and *MIPOL1* in VCaP cells treated with CRISPR sgRNA pairs targeting a control site (sgCTRL), *FOXMIND* or the *MIPOL1* UTR regulatory element (see Extended Data Fig. 2c for sgRNA binding sites). Distinct sgRNA pairs cutting at *FOXMIND* serve as biological replicates. Mean ± s.e.m. is shown ($n = 3$ technical replicates; two-way ANOVA and Tukey's test). **f**, Distribution of tandem duplication and translocation break ends (chimeric junctions or copy-number segment boundaries) focused at the *FOXMIND-FOXA1* regulatory domain. **g**, Outlier expression of genes involved in translocations with the *FOXA1* locus. Translocations positioning a gene between *FOXMIND* and *FOXA1* (hijacking) are shown on top (red). Translocations positioning a gene upstream of the *FOXA1* promoter (swapping) are shown on the bottom (blue). **h**, Inferred duplications within the *FOXA1* locus on the basis of RNA-seq (tandem break ends) and whole-exome sequencing (copy-gains), zoomed-in at the *FOXA1* topologically associating domain.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Transcriptional and genomic characteristics of class-3 FOXA1 rearrangements in prostate cancer. a, Dosage sensitivity of the *FOXA1* gene. Expression of *FOXA1* (RNA-seq) across mCRPC tumours ($n = 370$) as a function of gene ploidy (as determined by absolute copy number at the *FOXA1* locus (two-way ANOVA)). **b**, Relative expression of *FOXA1* (within the minimally amplified region) to *TTC6* (outside the amplified region) in rearranged ($n = 50$) (duplication or translocation) versus wild-type ($n = 320$) *FOXA1* loci (two-sided *t*-test). For all box plots, centre shows median, box marks quartiles 1–3 and whiskers span quartiles 1–3 ± 1.5 × IQR. **c**, Association plot visualizing the relative enrichment of cases with both translocation and duplications within the *FOXA1* locus ($n = 370$). Overabundance of cases with both events is quantified using Pearson residuals. Significance of this association is based on the χ^2 test without continuity correction. Inv, inversion; del, deletion. **d**, *FOXA1* locus visualization of linked-read (10X platform) whole-genome sequencing of the MDA-PCA-2b cell line. Alignments on the haplotype-resolved genome are shown in green and purple. Translocation and tandem-duplication calls are indicated in blue and red, respectively. **e**, Monoallelic expression of *FOXA1* cell lines with *FOXMIND-ETV1* translocations in MDA-PCA-2b ($n = 6$ biological replicates) and LNCaP ($n = 15$ biological replicates). Phasing of *FOXA1* SNPs to structural variants is based on linked-read sequencing (Methods). **f**, Biallelic expression of the RP11-356O9.1 transcript assessed using three distinct SNPs in MDA-PCA-2b cells that contain *ETV1* translocation into the *FOXA1* locus ($n = 7$ biological replicates). **g**, mRNA (qPCR) expression of *ETV1* and *TTC6* upon sgRNA-mediated disruption of the *FOXMIND* or the *MIPOL1* UTR enhancer in LNCaP cells, which also contain *ETV1* translocation into the *FOXA1* locus (see Extended Data Fig. 9d for sgRNA binding sites). Distinct sgRNA pairs cutting at *FOXMIND* serve as biological replicates. Mean ± s.e.m. are shown ($n = 3$ technical replicates; two-way ANOVA and Tukey's test).

nature research

Corresponding author(s): Arul M. Chinnaiyan

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	No software was usded for data collection.
Data analysis	All custom codes used for data analyses are freely available from the following public repositories: https://github.com/mcieslik-mctp/papy https://github.com/mcieslik-mctp/hpseq https://github.com/mcieslik-mctp/bootstrap-rnascape https://github.com/mcieslik-mctp/codac https://github.com/mcieslik-mctp/crisp https://github.com/mcieslik-mctp/ https://github.com/mctp/ GraphPad Prism 7 and in-built statistical tools version... Leica Microsystems – Leica Application Suite X (LAS X) SAMtools Version 1.3.1 PICARD Mark Duplicates Version 2.9.0

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

For all public data used in this study, accession codes are provided in the Methods. For sequencing data specifically collected in this study, we have deposited the raw ChIP and RNA sequencing files to the GEO repository; accession #: GSE123625.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine sample size. All samples size details for the analyses carried out in this study are reported in the Methods section. We curated an aggregate PCa cohort comprising of 888 localized and 658 metastatic samples, 498 and 357 with matched RNA-sequencing (RNA-seq) data, respectively. Other experimental sample sizes are included in the figure legends and the Methods section.
Data exclusions	No data was excluded from the published publically-available patient sequencing studies. For biologically experiments, no data exclusions were made.
Replication	For all experiments, there are at least two independent biological repeats and multiple technical repeats in each. In all instances, all attempts at replicating the experiments produced similar results.
Randomization	For zebrafish metastasis studies, embryos were randomly assigned to treatment groups with n>=30 for all.
Blinding	No experimental designs in this study required blinding. Additionally, no data quantification was manually performed that may require the blinding step to be incorporated into data analyses.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☐	☒ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	For immunoblotting, the following antibodies were used: FOXA1_N-terminal (Cell Signaling Technologies: 58613S; Sigma-Aldrich: SAB2100835) ; FOXA1_C-terminal (ThermoFisher Scientific: PA5-27157; Abcam: ab23738); AR (Millipore: 06-680); LSD1 (Cell Signaling Technologies: 2139S); Vinculin (Sigma Aldrich: V9131); H3 (Cell Signaling Technologies: 3638S); GAPDH (Cell Signaling Technologies: 3683); B-Actin (Sigma Aldrich: A5316); B-Catenin (Cell Signaling Technologies: 8480S); Vimentin (Cell Signaling Technologies: 5741S); Phospho(S33/S37/T41)-B-Catenin (Cell Signaling Technologies: 8814S); LEF1 (Cell Signaling Technologies: 2230S) ; AXIN2 (Abcam: ab32197), and TLE3 (Proteintech: 11372-1-AP).

The Vinculin and total H3 antibodies were used at 1:2000 dilution. All the remaining antibodies were used at 1:1000 dilution.

For co-immunoprecipitation and ChIP-Seq experiments, the following antibodies were used: FOXA1_N-terminal (Cell Signaling Technologies: 58613S); FOXA1_C-terminal (ThermoFisher Scientific: PA5-27157); AR (Millipore: 06-680); V5-tag (R960-25); TLE3 (Proteintech: 11372-1-AP).

For ChIPs, 10ug of all antibodies were used with 7.5-10M cells.

Validation

All antibodies used in this study are from reputed commercial vendors and have been validated by the vendors (see website).QC data is directly available from all the vendor listed above and these antibodies have been routinely used in other publications. Additionally, two key antibodies used in this study for FOXA1 ChIP_Seq were validated using recombinant proteins in this study. Data is included in Extended Data Figure 11. Also, the FOXA1 and TLE3 antibodies have been validated in this study using the siRNA targeting these proteins and concomitant disappearance of the specific protein bands upon immunoblotting.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)

Most cell lines were originally purchased from the American Type Culture Collection (ATCC) and were cultured as per the standard ATCC protocols. LNCaP-AR and LAPC4 cells were gifts from Dr. Charles Sawyers lab (Memorial Sloan-Kettering Cancer Center, New York, NY). Until otherwise stated, for all the experiments LNCaP, PNT2, LNCaP-AR, C42B, 22RV1, DU145, PC3 cells were grown in the RPMI 1640 medium (Gibco) and VCaP cells in the DMEM with Glutamax (Gibco) medium supplemented with 10% Full Bovine Serum (FBS; Invitrogen). LAPC4 cells were grown in IMEM (Gibco) medium supplemented with 15%FBS and 1nM of R1881. Immortalized normal prostate cells: RWPE1 were grown in keratinocyte media with regular supplements (Lonza); PNT2 were grown in RPMI medium with 10%FBS. HEK293 cells were grown in DMEM (Gibco) medium with 10% FBS. All cells were grown in a humidified 5% CO2 incubator at 37 celsius.

Authentication

All cell lines were biweekly tested to be free of mycoplasma contamination and genotyped every month at the University of Michigan Sequencing Core using Profiler Plus (Applied Biosystems) and compared with corresponding short tandem repeat (STR) profiles in the ATCC database to authenticate their identity in culture between passages and experiments.

Mycoplasma contamination

All cells were biweekly tested for mycoplasma contamination using the MycoAlert PLUS Mycoplasma Detection Kit (Lonza) and were found to be continually negative. More details are included in the Methods section.

Commonly misidentified lines
(See ICLAC register)

None

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals

Wild type ABTL zebrafish (male and female) were maintained in aquaria according to standard protocols. Embryos were generated by natural pairwise mating and raised at 28.5°C on a 14h light/10h dark cycle in a 100 mm petri dish containing aquarium water with methylene blue to prevent fungal growth. All experiments were performed on post-fertilization 2 to 7 days old embryos and were done in approved University of Michigan fish facilities under protocols approved from the University of Michigan Institution Animal Care and Use Committee.

Wild animals

NA.

Field-collected samples

NA.

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
May remain private before publication.

We have deposited the raw ChIP and RNA sequencing files to the GEO repository; accession #: GSE123625.

Files in database submission

ChIPSeq files
LAPC4 parental_FOXA1_CST
LAPC4 parental_FOXA1_TFS
LAPC4+DHT_FOXA1_TFS
LAPC4+DHT_FOXA1_CST
LNCaP parental_FOXA1_CST
LNCaP parental_FOXA1_TFS
C42B parental_FOXA1_CST
C42B parental_FOXA1_TFS
LAPC4+FOXA1-WT-V5_FOXA1-CST

LAPC4+FOXA1-WT-V5_FOXA1-TFS
LNCaP+FOXA1-P358fs-V5_FOXA1-CST
LNCaP+FOXA1-P358fs-V5_FOXA1-TFS
22RV1 parental_FOXA1-CST
22RV1 parental_FOXA1-TFS
22RV1 parental_ARMilli
22RV1 CRISPR_#WT3_FOXA1-CST
22RV1 CRISPR_#WT3_FOXA1-TFS
22RV1CRISPR_#WT3_ARMilli
22RV1 CRISPR_#36_FOXA1-CST
22RV1 CRISPR_#36_FOXA1-TFS
22RV1 CRISPR_#36_ARMilli
22RV1 CRISPR_#57_FOXA1-CST
22RV1 CRISPR_#57_FOXA1-TFS
22RV1 CRISPR_#57_ARMilli
22RV1 CRISPR_#70_FOXA1-CST
22RV1 CRISPR_#70_FOXA1-TFS
HEK293+eGFP-V5_FOXA1-TFS
HEK293+FOXA1-WT-V5_FOXA1-TFS
HEK293+FOXA1-I176M-V5_FOXA1-TFS
HEK293+FOXA1-R261G-V5_FOXA1-TFS
22RV1+FOXA1-WT-V5_FOXA1-CST_Rep1
22RV1+FOXA1-WT-V5_FOXA1-CST_Rep2
22RV1+FOXA1-I176M-V5_FOXA1-CST__Rep1
22RV1+FOXA1-I176M-V5_FOXA1-CST__Rep2
22RV1+FOXA1-R261G-V5_FOXA1-CST__Rep1
22RV1+FOXA1-R261G-V5_FOXA1-CST_Rep2
22RV1+FOXA1-WT-V5_AR-Milli_Rep1
22RV1+FOXA1-WT-V5_AR-Milli__Rep2
22RV1+FOXA1-I176M-V5_AR-Milli_Rep1
22RV1+FOXA1-I176M-V5_AR-Milli__Rep2
22RV1+FOXA1-R261G-V5_AR-Milli_Rep1
22RV1+FOXA1-R261G-V5_AR-Milli__Rep2
LNCaP parental_TLE3
C42B parental_TLE3
LAPC4 parental_TLE3
22RV1 parental_TLE3
22RV1 CRISPR_#WT3_TLE3
22RV1 CRISPR_#57_TLE3
22RV1 CRISPR_#36_TLE3
RNASeq files:
22RV1 siNC_72h_Rep1
22RV1 siNC_72h_Rep2
22RV1 siTLE3_72h_Rep1
22RV1 siTLE3_72h_Rep2
22RV1 parental_Rep1
22RV1 parental_Rep2
22RV1 CRISPR_#WT2_Rep1
22RV1 CRISPR_#WT2_Rep2
22RV1 CRISPR_#WT3_Rep1
22RV1 CRISPR_#WT3_Rep2
22RV1 CRISPR_#57_Rep1
22RV1 CRISPR_#57_Rep2

Genome browser session
(e.g. UCSC)

Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.

Methodology

Replicates

Multiple biological as well as technical replicates are included.

Sequencing depth

See Methods

Antibodies

Validated by the Vendors. Additionally, validation data for two key antibodies is included in the Extended Data Figure 11.

Peak calling parameters

See Methods

Data quality

See Methods

Software

See Methods