Methods

The MPCproject

The Metastatic Prostate Cancer Project (MPCproject) is a patient-driven genomics research initiative that enables metastatic or advanced prostate cancer patients to enroll through an online portal. Patients can elect to provide consent for acquisition of a saliva sample, a blood sample, medical records, and archived FFPE tissue samples. The data generated as a result are used to create a clinically-annotated genomic dataset that can be shared across multiple platforms. The following describes the methods used to generate the current dataset from The Metastatic Prostate Cancer Project (19 samples, 19 patients) in January of 2019. Please refer to MPCproject.org/data-release for links to data on public platforms.

Methods

Design

The Metastatic Prostate Cancer Project website (MPCproject.org) enables patients to enroll in the study remotely. The website and all associated messaging were developed through iterative feedback from metastatic or advanced prostate cancer patients. Patients registered with first name, last name, email address, and confirmation of a metastatic or advanced prostate cancer diagnosis. Registrants were asked to complete a 17-question survey, with all questions optional, about their experiences with metastatic or advanced prostate cancer (see link on MPCproject.org/data-release). Registrants acknowledged that responses would be stored in a secure database. Included in this acknowledgement is an understanding that patients may be re-contacted and that they can withdraw their information as indicated in the pre-submission statement:

> *"I understand that the information I entered here will be stored in a secure database and may be used to match me to one or more research studies conducted by the Metastatic Prostate Cancer Project. If the information that I entered matches a study being conducted by the Metastatic Prostate Cancer Project, either now or in the future, I agree to be contacted about possibly participating. I understand that if I would like my information deleted from the database, now or in the future, I can email info@MPCproject.org and my information will be removed from the database."*

All patients who at any time submitted a written request to withdraw from the study were immediately exited by study staff.

A link to an electronic informed consent document for formal enrollment in the study (see link on MPCproject.org/data-release) was sent to registrants who completed the survey. Registrants could direct any questions to study staff throughout the enrollment process. Email reminders were sent weekly for three weeks, and again at six weeks, to registrants who had not completed the consent process. Registrants who provided informed consent were then asked to complete a medical release form in order to provide their contact information, as well as a list of physicians and hospitals that provided their care for metastatic or advanced prostate cancer. Email reminders were sent weekly for three weeks, and again at six weeks, to registrants who had not completed the medical release form. Upon completion, signed copies of consent and medical release forms were sent electronically to the registrants. All data collected online were stored in a secure database.

Saliva kits were mailed to registrants who provided informed consent and who lived in The United States or

Canada. Saliva kits were labeled with a unique two-dimensional barcode and a pre-paid Business Reply Label addressed to the Broad Institute Genomics Platform prior to shipment. Each barcode identifier was assigned to a participant prior to shipment. Participants provided a saliva sample by following the included instructions (see link on MPCproject.org/data-release) and returned the kits free of charge. Saliva kits received at the Broad Institute were logged by their unique barcodes, and stored at room temperature until they were advanced to Whole Exome Sequencing (WES).

Blood kits were mailed to registrants who provided informed consent, opted in to the blood biopsy component of the study, and lived in The United States or Canada. Blood kits were labeled with a unique two-dimensional barcode and included a pre-paid FedEx ClinPak envelope addressed to the Broad Institute Genomics Platform prior to shipment. Each barcode identifier was assigned to a participant prior to shipment. Participants provided a blood sample by bringing the kit and instructions (see link on MPCproject.org/data-release**)** with them to their next regularly scheduled clinical appointment and requesting a courtesy draw. If a courtesy draw was not possible, patients were given the option to go to Quest Diagnostics with a voucher for their complementary blood draw. Blood kits were returned free of charge. Blood samples received at the Broad Institute were logged by their unique barcodes and fractionated. The resultant plasma and buffy coats were stored at -80C and plasma samples were advanced to Ultra-low Pass Whole Genome Sequencing (ULP-WGS) to ascertain the presence of circulating tumor DNA (ctDNA). Select plasma samples were advanced to WES. Buffy coats were advanced to WES where no saliva sample was available to generate matched germline data.

Study staff called the hospitals and physicians' offices listed in each participant's medical release form to confirm the fax number for the medical records department. A detailed request for medical records including clinic notes from treating providers, metastatic or advanced prostate cancer treatment data (including radiation and chemotherapy), pathology reports, operative reports, referrals, MD to MD exchange, and genetic testing reports from the date of diagnosis through the date the request was faxed to each facility. Medical records were received by fax, mail, or secure electronic message. All medical records were scanned and uploaded to a secure drive to facilitate abstraction. Each medical record was read to confirm a diagnosis of metastatic or advanced prostate cancer. The date of each procedure (e.g. biopsy, resection, prostatectomy, etc.), type of procedure, histology, and facility which performed the procedure were abstracted from all pathology reports in order to prioritize tissue samples for request for patients who opted in to the tissue component of the study. Metastatic or advanced prostate cancer samples were flagged for request. In collaboration with oncologists, study staff developed and followed a SOP for determining the minimum number of available tissue blocks required to request a portion of a patient's tissue without interfering with continued clinical care. If sufficient tissue were available, per pathology reports, tissue was requested with the explicit instruction to pathology departments not to exhaust any given sample.

Study staff called the pathology departments associated with each tissue sample to confirm the fax number for tissue requests. A form requesting a minimum of 5 and maximum of 23 5-micron unstained slides, or one tissue block, and one Hematoxylin and Eosin stain (H&E) slide was faxed to each pathology department. Requests explicitly stated that no sample should be exhausted in order to fulfill the request.

Tissue samples were received at the Broad Institute by mail. Tissue samples received as blocks were labeled with unique numerical identifiers and sent to the Dana- Farber/Harvard Cancer Center Specialized Histopathology Services (SHS) Core to be cut into three 30- micron scrolls per block. These scrolls were then labeled with unique barcode identifiers and submitted to the Broad Institute Genomics Platform for WES and Transcriptome Capture (RNA Seq). Tissue samples received as unstained slides were logged,

labeled with unique barcode identifiers, and submitted to the Broad Institute Genomics Platform for WES and RNA Seq.

Medical Record Abstraction

A data dictionary (see Appendix 1) comprising 13 fields based on expert pathologist review and 25 fields from medical records was developed. The date of primary diagnosis with metastatic or advanced prostate cancer was defined as the date of first definitive confirmation of disease by biopsy. Dates were abstracted to the greatest level of detail available in the record. Dates reported in the medical record only as a month and year were abstracted as the first of the month while dates reported only as a year were abstracted as the first of January of that year. For all other fields, only data explicitly reported in the medical record were abstracted; no data were inferred. In order to protect patient confidentiality, all dates were reported in shared data sets as elapsed time relative to the date of primary diagnosis and ages were grouped into 5-year increments.

Patient-Reported Data

Patient-reported data (PRD) comes from information provided by patients in the 17-question intake survey (see link on MPCproject.org/data-release). When applicable, data were cleaned to standardize format as well as protect patient confidentiality. In order to protect confidentiality, race categories that were reported fewer than five times in the dataset were reclassified as “other”.

Biological Sample Processing

DNA Isolation from Saliva

DNA was extracted via the Chemagic MSM I with the Chemagic DNA Blood Kit-96 from Perkin Elmer. This kit combines a chemical and mechanical lysis with magnetic bead-based purification.

Saliva samples were incubated at 50°C for 2 hours. The saliva was then transferred to a deep well plate placed on the Chemagic MSM I. The following steps were automated on the MSM I.

M-PVA Magnetic Beads were added to the saliva. Lysis buffer was added to the solution and mixed. The bead-bound DNA was then removed from solution via a 96-rod magnetic head and washed in three Ethanol-based wash buffers. The beads were then washed in a final water wash buffer. Finally, the beads were dipped in elution buffer to resuspend the DNA sample in solution. The beads were then removed from solution, leaving purified DNA eluate.

DNA samples were quantified using a fluorescence-based PicoGreen assay.

DNA Isolation from Whole Blood

DNA was extracted via the Chemagic MSM I with the Chemagic DNA Blood Kit-96 from Perkin Elmer. This kit combines a chemical and mechanical lysis with magnetic bead-based purification.

Whole Blood samples were incubated at 37°C for 5-10 minutes to thaw. The blood was then transferred to a deep well plate with protease and placed on the Chemagic MSM I. The following steps were automated on the MSM I.

M-PVA Magnetic Beads were added to the blood, protease solution. Lysis buffer was added to the

solution and vortexed to mix. The bead-bound DNA was then removed from solution via a 96-rod magnetic head and washed in three Ethanol-based wash buffers to eliminate cell debris and protein residue. The beads were then washed in a final water wash buffer. Finally, the beads were dipped in elution buffer to re-suspend the DNA. The beads were then removed from solution, leaving purified DNA eluate.

DNA samples were quantified using a fluorescence-based PicoGreen assay.

cfDNA Extraction from Whole Blood

Whole blood was collected in EDTA, CellSave, or Streck tubes and processed for plasma fractionation within 3 hours of blood draw. Blood tubes were centrifuged at 1900 g for 10 minutes and plasma was transferred to second tube before further centrifugation at 15000 g for 10 minutes. Supernatant plasma was stored at -80C until cfDNA extraction. cfDNA was extracted using the QIAsymphony DSP Circulating DNA Kit according to the manufacturer's instructions, with 6.3 mL of plasma as input and with a 60 uL DNA elution (Qiagen, 2017).

Library Construction

Library construction was performed using the KAPA Hyper Prep Kit according to the manufacturer's instructions (Kapa, 2016), with the following modifications: initial DNA input was normalized to 20 ng in 50 uL of TE buffer (10mM Tris HCl 1mM EDTA, pH 8.0) according to picogreen quantification. For adapter ligation, Illumina paired end adapters were replaced with palindromic forked adapters, purchased from Integrated DNA Technologies, with unique dual-indexed molecular barcode sequences to facilitate downstream pooling. Adapters were diluted to 3.75 uM before addition to the samples. During PCR, 10 cycles were used. During the post-enrichment SPRI cleanup, elution volume was reduced to 30µL to maximize library concentration.

Post Library Construction Quantification and Normalization

Library quantification was performed using the Invitrogen Quant-It broad range dsDNA quantification assay kit (Thermo Scientific Catalog: Q33130) with a 1:200 PicoGreen dilution. Following quantification, each library is normalized to a concentration of 25 ng/µL, using a 1X Low TE pH 7.0 solution.

Library Pool Creation for Ultra-low Pass Sequencing

In preparation for the sequencing of the ultra-low pass libraries (ULP), approximately, 2 µL of the normalized library is transferred into a new receptacle and further normalized to a concentration of 3ng/µL, again using a 1X Low TE pH 7.0 solution. Following normalization, up to 96 individual ultra-low pass WGS pool is created via equivolume pooling.

In-solution hybrid selection for exome or custom panels

After library construction, hybridization and capture were performed using the relevant components of Illumina's Nextera Rapid Capture Exome Kit and following the manufacturer's suggested protocol, with the following exceptions: first, all libraries within a library construction plate were pooled prior to hybridization. Second, the Midi plate from Illumina's Nextera Rapid Capture Exome Kit was replaced with a skirted PCR plate to facilitate automation. All hybridization and capture steps were automated on the Agilent Bravo liquid handling system.

Preparation of libraries for cluster amplification and sequencing

After post-capture enrichment, library pools were quantified using qPCR (automated assay on the Agilent Bravo), using a kit purchased from KAPA Biosystems with probes specific to the ends of the adapters. Based on qPCR quantification, pools were normalized to 1.5nM for exome libraries and 1.5nM for genome libraries.

Cluster amplification and sequencing

Cluster amplification of library pools was performed according to the manufacturer's protocol (Illumina) using Exclusion Amplification cluster chemistry and HiSeq X flowcells. Flowcells were sequenced on v2 Sequencing-by-Synthesis chemistry for HiSeq X flowcells. The flowcells are then analyzed using RTA v.2.7.3 or later. Each pool of whole genome libraries was run on paired 151bp runs, reading the dual-indexed sequences to identify molecular indices and sequenced across the number of lanes needed to meet coverage for all libraries in the pool.

Cluster amplification of library pools was performed according to the manufacturer's protocol (Illumina) using Exclusion Amplification cluster chemistry and HiSeq 4000 flowcells. Flowcells were sequenced on v1 Sequencing-by-Synthesis chemistry for HiSeq 4000 flowcells. The flowcells are then analyzed using RTA v.2.7.3 or later. Each pool of whole exome libraries was run on paired 76bp runs, reading the dual-indexed sequences to identify molecular indices and sequenced across the number of lanes needed to meet coverage for all libraries in the pool.

REFERENCES

1. Kapa Biosystems, "KAPA Hyper Prep Kit Technical Data Sheet," KR0961 – v5.16, June 2016.
2. Qiagen, "QIAsymphony® DSP Circulating DNA Kit Instructions for Use (Handbook)," Version 1, March 2017.

Exome Express ICE Methods

Library Construction

Library construction was performed as described in Fisher et al., with the following modifications: initial genomic DNA input into shearing was reduced from 3µg to 10-100ng in 50µL of solution. For adapter ligation, Illumina paired end adapters were replaced with palindromic forked adapters, purchased from Integrated DNA Technologies, with unique dual-indexed molecular barcode sequences to facilitate downstream pooling. With the exception of the palindromic forked adapters, the reagents used for end repair, A-base addition, adapter ligation, and library enrichment PCR were purchased from KAPA Biosciences in 96-reaction kits. In addition, during the post-enrichment SPRI cleanup, elution volume was reduced to 30µL to maximize library concentration, and a vortexing step was added to maximize the amount of template eluted.

In-solution hybrid selection

After library construction, hybridization and capture were performed using the relevant components of Illumina's Nextera Rapid Capture Exome Kit and following the manufacturer's suggested protocol, with the following exceptions: first, all libraries within a library construction plate were pooled prior to hybridization. Second, the Midi plate from Illumina's Nextera Rapid Capture Exome Kit was replaced with a skirted PCR plate to facilitate automation. All hybridization and capture steps were automated on the Agilent Bravo liquid handling system.

Preparation of libraries for cluster amplification and sequencing

After post-capture enrichment, library pools were quantified using qPCR (automated assay on the Agilent Bravo), using a kit purchased from KAPA Biosystems with probes specific to the ends of the adapters. Based on qPCR quantification, libraries were normalized to 2nM, then denatured using 0.1 N NaOH on the Hamilton Starlet. After denaturation, libraries were diluted to 20pM using hybridization buffer purchased from Illumina.

Cluster amplification and sequencing

Cluster amplification of denatured templates was performed according to the manufacturer's protocol

(Illumina) using HiSeq 4000 cluster chemistry and HiSeq 4000 flowcells. Flowcells were sequenced on v1 Sequencing-by-Synthesis chemistry for HiSeq 4000 flowcells. The flowcells are then analyzed using RTA v.1.18.64 or later. Each pool of whole exome libraries was run on paired 76bp runs, reading the dual- indexed sequences to identify molecular indices and sequenced across the number of lanes needed to meet coverage for all libraries in the pool.

REFERENCES

1. Fisher S, Barry A, Abreu J, Minie B, Nolan J, Delorey TM, Young G, Berlin AM, Blumenstiel B, Cibulskis K, Friedrich D, Johnson R, Juhn F, Reilly B, Shammas R, Stalker J, Sykes SM, Thompson J, Walsh J, Zimmer A,
2. Zwirk Z, Gabriel S, Nicol R, Nusbaum C. A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. Genome Biology 2011, 12:R1.

Sequencing Data Analysis

Whole exome sequences were captured using Illumina technology and the sequence data processing and analysis was performed using the Picard and FireCloud pipelines at the Broad Institute. The Picard pipeline (http://picard.sourceforge.net) was used to produce a BAM file with aligned reads. This includes alignment to the GRCh37 human reference sequence using the BWA aligner[1] and estimation and recalibration of base quality score with the Genome Analysis Toolkit (GATK)[2]. All sample pairs passed through the FireCloud pipeline were subjected to QC testing to test for any tumor/normal and inter-individual contamination as previously described[3,4]. The MuTect algorithm was used to identify somatic mutations[4]. Furthermore, we filtered for false-positive somatic mutation calls using a panel of normals (PoN), oxoG filter[5], and an FFPE filter[6] to remove artifacts introduced during the sequencing or formalin fixation process. Small somatic insertions and deletions were detected using the Strelka algorithm[7].

Somatic mutations including single-nucleotide variants, insertions, and deletions were annotated using Oncotator[8]. The germline somatic variants were analyzed using the HaplotypeCaller module of GATK[2]. Significance of identified somatic mutations was analyzed using MutSig2CV[9], which uses patient and gene-specific mutation rates to estimate a background model of predicted mutation incidence across the genome. MutSig2CV then factors in biological co-variates such as replication timing and gene-expression level on a gene-by-gene basis to account for the increased mutational rate of certain classes of genes.
To analyze somatic copy number alterations (SCNA) from whole exome data, we used GATK CNV, which assesses homolog-specific copy ratios from segmental estimates of multipoint allelic copy ratios at heterozygous loci incorporating the statistical phasing software (BEAGLE) and population haplotype panels (HAPMAP3)[10,11].

For copy number alteration significance analysis, segmented copy number data was analyzed by GISTIC 2.0, to identify significantly recurring focal and arm-level amplification/deletion peaks.[12]Allele-specific SCNAs and tumor ploidy/purity status were assessed using FACETS.[13]

Prostate cancer tumor samples which passed QC testing and had a purity of 20% or more were submitted to cBioPortal (http://www.cbioportal.org/index.do).

Assessment of Tumor Mutation Burden (TMB)

TMB (mutation per megabase) was calculated as the total number of mutations (non-synonymous + synonymous) detected for a given sample divided by the length of the total genomic target region captured with the whole exome sequencing[13]. Samples with TMB ≥ 10 mutations per megabase were classified as hypermutated.

REFERENCES

1. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–60 (2009).
2. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–303 (2010).
3. Berger, M. F. *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470,** 214–20 (2011).
4. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31,** 213–9 (2013).
5. Costello, M. *et al*. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res.* **41**, e67– (2013).
6. Van Allen, E. M. *et al.* Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nat. Med.* **20**, 682–688 (2014).
7. Saunders, C. T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28,** 1811–7 (2012).
8. Ramos, A. H. *et al.* Oncotator: cancer variant annotation tool. *Hum. Mutat.* **36,** E2423-9 (2015).
9. Lawrence, M. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature **499**, 214-218 (2013)
10. Browning, B. L. & Yu, Z. Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. *Am. J. Hum. Genet.* 85, 847–861 (2009).
11. International HapMap, C. et al. Integrating common and rare genetic variation in diverse human populations. *Nature* 467, 52–58 (2010).
12. Beroukhim, R. *et al.* Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc. Natl Acad. Sci. USA* **104**, 20007–20012(2007)
13. Shen, R. & Seshan, V. E. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res.* **44**, e131 (2016).
14. Chalmers *et al*. Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden *Genome Medicine* 9:34 (2017)

Appendix 1: Data Dictionary

Medical Record Data	
Sex	Male Female
DOB	Date of birth (used to calculate age at diagnosis)
Date of Diagnosis of Metastatic Disease	Date of confirmed metastatic disease
Date of Diagnosis of Biochemical Recurrence	Date of confirmed biochemical recurrence
Date of PSA at diagnosis of primary prostate cancer	Date of PSA closest to diagnosis with primary prostate cancer
PSA at diagnosis of primary prostate cancer	Numeric value
Diagnostic histology	Prostatic Adenocarcinoma Other Unknown
Diagnostic primary Gleason score	3, 4, 5
Diagnostic secondary Gleason score	3, 4, 5
Diagnostic combined Gleason score	6, 7, 8, 9, 10
Cores taken for diagnostic biopsy	Numeric value
Highest percent core in diagnostic biopsy	Numeric value
Diagnostic T Stage	T1 T2 T3 Unknown
Diagnostic N Stage	N0 N1 Nx Unknown
Diagnostic M Stage	M0 M1 Unknown
Diagnostic Alkaline Phosphatase Score	Numeric Value
Date of diagnostic Alkaline Phosphatase score	Date of closest Alkaline phosphatase to diagnosis with primary prostate cancer
Metastatic sites at initial diagnosis	List – location(s) of metastatic disease at diagnosis of prostate cancer. These are considered de-novo metastatic sites.

Metastatic sites at diagnosis of metastatic disease	List – location(s) of metastatic disease at diagnosis of metastatic disease. List will include sites of metastatic disease at initial diagnosis if patient was diagnosed with de-novo metastatic disease
PSA at diagnosis of metastatic disease	Numeric value
Date of PSA test closest to metastatic diagnosis	Date of PSA test closest to metastatic diagnosis
PSA at diagnosis of biochemical recurrence	Numeric value
Date of PSA test closest to biochemical recurrence	Date of PSA test closest to biochemical recurrence diagnosis
Radiation received	Yes No
Treatment	Drug name (generic)
Treatment start date	MM/DD/YYYY
Treatment stop date	MM/DD/YYY
Mode of treatment	Biochemical recurrence Metastatic (Met) Pre-Metastatic (Pre-Met)
Location of treatment	List – location(s) where treatment was targeted

Pathology Data	
Sample Naive	Calculated variable that is set to YES if a sample was taken prior to the start of any medical therapy
Sample Collection Time	Days sample was taken from primary diagnosis
Biopsy location	Location in the body where procedure was performed
Biopsy procedure type	Core Biopsy Prostatectomy Dissection Excision Other
Gleason Primary Pattern	3, 4, 5
Gleason Secondary Pattern	3, 4, 5
Total Gleason Score	6, 7, 8, 9, 10
Seminal Vesicle Invasion	Present Absent Unknown N/A

Lymphovascular Invasion	Present Absent Unknown N/A
Extraprostatic Invasion	Present Absent Unknown N/A
Perineural Invasion	Present Absent Unknown N/A
Margin Status	Positive Negative Unknown N/A