nature portfolio

Corresponding author(s): Alexander W. Wyatt

Last updated by author(s): Jun 2, 2022

Reporting Summary

Nature Portfolio 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 Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the 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
☐	☒ A statement on 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 statistical parameters 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

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No custom software was used to collect the data. Raw sequencing data was generated in-house or obtained from collaborators. Sample meta-data was available in-house or provided by collaborators.
Data analysis	All statistical tests and data analyses were conducted in Python 3.7 (using pandas 0.25.0, numpy 1.16.4, scipy 1.6.2, and statsmodels 0.12.2), Julia 1.5.1 (using HypothesisTests 0.10.9, and Distributions 0.25.53), and R 3.6.1 (using dplyr 1.0.7) languages. Visualizations were generated using matplotlib version 3.3.4 (Python) and ggplot 3.1.1 (R). The following bioinformatics/genomic analysis software was used: cutadapt-1.11, seqkit-0.8, Bowtie-2.3.0, BWA (versions 0.7.15-r1140 and 0.7.17), samblaster-0.1.24, bedtools-2.25.0, samtools 1.12 (htslib 1.12), Picard 2.25.6, Mutato version 0.7, ANNOVAR (version 20191024), DeconstructSigs v1.47 (Python implementation available: https://github.com/vanallenlab/deconstruct_sigs_py), SigProfilerMatrixGenerator v1.2.1, SigProfileExtractor 1.1.4, Picard 2.25.6, GATK Mutect2, ASCAT 2.5.2, Battenberg 2.2.9, Breakfast software version 0.5, GRIDSS2 2.10.0, PyClone 0.13.1, and PhyloWGS version 2018-11-05. All custom code is available at https://github.com/annalam/cfdna-wgs-manuscript-code. Comprehensive methodology including all software (and versions) used is annotated within the manuscript and our Supplementary Information document.

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 and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The human reference genome hg38 was downloaded from UCSC. For cfDNA nucleosome depletion analyses, we used exon and TSS coordinates from the RefSeq Matched Annotation from NCBI and EMBL-EBI (MANE) database (genes not annotated in this database were omitted from analysis). Metastatic tissue RNA-Seq data (Cohort B) was obtained from previously published work (Quigley et al 2018; PMID: 30033370; dbGaP study accession: phs001648.v2.p1). Gene sets representing housekeeping genes and highly expressed genes in hematopoietic lineages were derived from previously published work (Eisenberg and Levanon 2013; PMID: 23810203; Ulz et al. 2016; PMID: 27571261) (Supplementary Tables 9-10). 3224 AR binding sites (ARBS) were from prior chromatin immunoprecipitation followed by ChIP-sequencing of 13 primary prostate cancer tissue samples (Pomerantz et al. 2015; PMID: 26457646; Morova et al. 2020; PMID: 32047165) (Supplementary Table 11).

All de-identified whole-genome sequencing data have been deposited in the European Genome-Phenome Archive (EGA) under the accession code EGAS00001005783, available for download by contacting the corresponding authors. All other data supporting the findings of this study (i.e. source data) are available within the article (including its Supplementary Information and Supplementary Tables).

Field-specific reporting

Please select the one below that is 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/documents/nr-reporting-summary-flat.pdf

Life sciences study design

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

Sample size	Sample size was not predetermined for this retrospective exploratory study (no power calculations were performed). Select blood samples from the prospective clinical trial NCT02125357 were included, in addition to blood and patient-matched metastatic tissue samples from collaborators (via NCT02432001), so long as inclusion criteria were met (as specified in Methods) and the patient provided consent. We analyzed all possible samples within the constraints of: 1) the denominator of available patient/samples collected through these two clinical studies (NCT02125357 and NCT02432001), 2) our sample/patient eligibility criteria (detailed in our Methods), and 3) overall project funding. The exploratory and descriptive nature of our study and lack of pre-specified analyses means that target sample sizes were not possible; all statistical analyses are descriptive and effect sizes / p-values are reported as appropriate.
Data exclusions	The only pre-specified inclusion criteria were those mandated as part of the original, previously-published, and publically-available clinical studies (NCT02125357 and NCT02432001) from which our samples were sourced. Any cohort-level exclusions in the context of this current study were based on our patient inclusion criteria (which was not pre-specified): only patients providing ≥2 serial cfDNA samples (NCT02125357) or matched metastatic tissue and ≥1 cfDNA samples (NCT02432001) with at least one sample with ≥30% cancer fraction (NCT02125357 and NCT02432001) were considered eligible for whole-genome sequencing (WGS). Patients with multiple high tumor fraction samples (≥30%) were prioritized for WGS. Samples that failed subsequent library generation or WGS were excluded from the final study cohort. When samples were excluded from sub-analyses, the rationale and denominators are clearly listed in the manuscript text and/or figure legend. Almost always this is due to categorical exclusion (e.g. an analysis quantifying the mutation rate in specifically in ctDNA samples would exclude tissue samples, etc.).
Replication	All analyses are descriptive in nature. No in vitro experiments requiring technical or biological replicates were performed. Repeat sequencing of identical plasma or tissue DNA samples was not performed. All in silico analyses/experiments/models were either deterministic (not requiring replication) or had high convergence (i.e. negligibly small probability of inter-replicate variance that was empirically tested). Random seeds were used for all relevant analyses for reproducibility (documented in our online code repository).
Randomization	This retrospective exploratory study did not involve any analyses requiring patient/sample randomization or controlling for participant covariates. All analyses were descriptive in nature, and all relevant samples were analyzed together (unless they were categorically irrelevant for the analysis at hand).
Blinding	Blinding was not performed for this retrospective, descriptive, and exploratory study (with one exception). We did incorporate investigator blinding into one in silico experiment to validate our bespoke subclonal reconstruction models. Specifically, investigator A was required to use our subclonal reconstruction procedure to infer the cancer fractions and subclone CCF fractions of several in silico simulated NGS samples (with ground truth clonal composition determined by investigator B; investigator A was blinded to this information until after subclonal reconstruction was executed)

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☐	☒ Clinical data
☒	☐ Dual use research of concern

Methods

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	The majority of participants were biologically male with metastatic castration-resistant prostate cancer of any histological variant. Two patients with metastatic bladder cancer were also included. Inclusion criteria (describing the population) are detailed in the methods.
Recruitment	We retrospectively selected patients who had initially enrolled on a prospective clinical trial (NCT02125357) and a prospective tissue-sequencing effort (NCT02432001) and had voluntarily consented to provide samples for future research purposes. Additional samples (n=9) serving as secondary controls were obtained in context of a provincial blood and tissue biobank (all patients voluntarily consented to provide samples for research purposes).
Ethics oversight	University of British Columbia Ethics Board; University of California San Francisco Institutional Review Board

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	No all-inclusive registration is available for this retrospective study. A subset of patients were enrolled in NCT02125357 and NCT02432001.
Study protocol	Patient inclusion criteria for this retrospective study are specified in the Methods. Patients were required to have metastatic cancer (mainly metastatic castration-resistant prostate cancer, but two patients with metastatic bladder cancer were also included).
Data collection	Patient samples were initially collected between 2013 and 2020 in the context of a prospective clinical trial (NCT02125357) or clinical study (NCT02432001). Additional samples (n=9) were collected as part of a provincial biobank program.
Outcomes	Clinical outcomes were not reported in this manuscript nor incorporated into any quantitative analyses.