

OPEN
ARTICLE

A harmonized resource of integrated prostate cancer clinical, -omic, and signature features

Teemu D. Laajala^{1,2} , Varsha Sreekanth², Alex C. Soupir³ , Jordan H. Creed³, Anni S. Halkola¹, Federico C. F. Calboli^{1,4}, Kalaimathy Singaravelu¹, Michael V. Orman² , Christelle Colin-Leitzinger⁵ , Travis Gerke⁵, Brooke L. Fridley^{3,8} , Svitlana Tyekucheva^{6,8} & James C. Costello^{2,7,8}

Genomic and transcriptomic data have been generated across a wide range of prostate cancer (PCa) study cohorts. These data can be used to better characterize the molecular features associated with clinical outcomes and to test hypotheses across multiple, independent patient cohorts. In addition, derived features, such as estimates of cell composition, risk scores, and androgen receptor (AR) scores, can be used to develop novel hypotheses leveraging existing multi-omic datasets. The full potential of such data is yet to be realized as independent datasets exist in different repositories, have been processed using different pipelines, and derived and clinical features are often not provided or not standardized. Here, we present the *curatedPCaData* R package, a harmonized data resource representing >2900 primary tumor, >200 normal tissue, and >500 metastatic PCa samples across 19 datasets processed using standardized pipelines with updated gene annotations. We show that meta-analysis across harmonized studies has great potential for robust and clinically meaningful insights. *curatedPCaData* is an open and accessible community resource with code made available for reproducibility.

Introduction

Prostate cancer is the most common cancer type amongst men with an estimated incidence of 268,490 new cases per year in the United States, with an estimated 34,500 deaths per year¹. Molecular profiling of prostate cancer has led to insights into the relationship of genomic alterations and disease initiation, progression, and treatment response. However, no significant differences in disease free survival were found for patients that were stratified according to the 8-group prostate cancer (PCa) taxonomy defined by The Cancer Genome Atlas (TCGA) using single gene molecular alterations². Additionally, when primary tumors were compared to metastatic tumor samples, few changes in the frequency of these genomic alterations were observed²⁻⁴.

A reliable molecular biomarker that stratifies aggressive vs. indolent disease is increased frequency of Copy Number Alterations (CNAs)⁴⁻⁷; however, this finding provides little mechanistic or therapeutically actionable insight. Recent studies have shown that combinations of alterations, namely *TP53* & *RBI*⁸ and *CHD1* & *MAP3K7*⁹, drive aggressive disease, suggesting that molecular subtyping in PCa is complex. Many efforts have been put forward to develop predictive gene expression signatures with the goal of identifying which patients will progress to lethal disease¹⁰⁻¹⁶. Some of these signatures have been clinically successful^{11,17,18}; however, an overwhelming amount of gene expression profiling results lack replicability between studies resulting in inconsistent lists of candidate genes associated with PCa prognosis¹⁹. Additional challenges in reproducible

¹Department of Mathematics and Statistics, University of Turku, Turku, Finland. ²Department of Pharmacology, University of Colorado Anschutz Medical Campus, Aurora, CO, USA. ³Department of Biostatistics and Bioinformatics, Moffitt Cancer Center, Tampa, FL, USA. ⁴Natural Resources Institute Finland (Luke), F-31600, Jokioinen, Finland. ⁵Department of Cancer Epidemiology, Moffitt Cancer Center, Tampa, FL, USA. ⁶Department of Data Science, Dana-Farber Cancer Institute; Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. ⁷University of Colorado Cancer Center, University of Colorado Anschutz Medical Campus, Aurora, CO, USA. ⁸These authors jointly supervised this work: Brooke L. Fridley, Svitlana Tyekucheva, James C. Costello. e-mail: teelaa@utu.fi; svitlana@jimmy.harvard.edu; james.costello@cuanschutz.edu

PCa research remain. For example, the use of high-dimensional molecular data is dependent on a thorough validation of the statistical models in diverse datasets. Similar concerns apply to molecular subtyping. Many of these challenges can at least partially be addressed by harmonization of the ‘omic’ data preprocessing and annotations, matched with manual curation of the clinicopathologic features and outcomes for easy application of multi-study statistical learning²⁰, and cross-study validation²¹.

Data wrangling and data harmonization are critical for the consistent, reproducible, and benchmarked analysis of multi-omic cancer datasets. Efforts have been completed for ovarian cancer in the curatedOvarianData R package²², breast cancer in the curatedBreastData R package²³, and across cancer types in the curatedTCGAData R package²⁴. These packages have advanced the field in many ways. To this end, the R user community has put great effort into developing R class objects that help end-users to utilize data across different types - such as transcriptomics, copy number alterations, and somatic mutations - and between studies that vary in their specific study characteristics. The *MultiAssayExperiment*-class²⁵ (MAE) aggregates data of various types utilizing such R classes as *matrix*, *RaggedExperiment*, *SummarizedExperiment* across these data levels. This data class supports linking and simultaneous storage of sample- or patient-level clinical metadata fields that can be easily processed and stored together with their corresponding ‘omics’ data.

In addition to the primary ‘omic’ data types themselves, such as gene expression measurements by RNA sequencing or microarrays, there are now an array of innovative approaches to develop molecular signatures and deconvolution methods to estimate cell types present in bulk tissue. The *immunedecconv*-package²⁶ has proven to be a popular choice as a wrapper R package providing harmonized access to multiple popular cell type deconvolution methods such as EPIC²⁷, ESTIMATE²⁸, MCP-counter²⁹, quanTIseq³⁰, and xCell³¹. Estimating the prevalences of different cell types in the tumor specimen has allowed for investigating the relationship between immune cells and other cell frequencies in a tumor sample with clinical outcomes^{26–34}.

Given the value to the PCa research field in having a unified resource of molecular features across independent studies, we developed a curated, comprehensive, and harmonized PCa resource that contains multi-omic and clinical data from 19 PCa studies. The ‘omic’ data types were preprocessed and annotated, and clinical variables were mapped to a common data dictionary to ensure consistent annotation of the samples. Furthermore, we precomputed several prostate-specific genomic scores using the uniformly preprocessed and annotated gene expression data sets. Namely, we conveniently provide Decipher³⁵, Oncotype DX³⁶, and Prolaris³⁷ risk scores as well as Androgen Receptor (AR) scores². These pre-computed variables can be easily included in the downstream analyses as correlative or phenotypic variables. Leveraging the MAE class, we supply the data in the *curatedPCaData* R package (<https://github.com/Syksy/curatedPCaData>). The package provides open and accessible data and analysis pipelines with maximum flexibility for data analysts and prostate cancer researchers. We discuss the integrated datasets within the package and insights that have been gained by bringing together >3500 prostate tissue, primary PCa, and metastatic PCa tumor samples.

Results

A summary of the key study characteristics of the 19 datasets contained in the *curatedPCaData* package are in Table 1. The *curatedPCaData* package was developed using standardized workflows for raw data processing where available, mapping all clinical information for each dataset to a common data dictionary³⁸, and ensuring gene symbols are consistent and up-to-date using HUGO Gene Nomenclature Committee (HGNC) symbols across all datasets and data types (Figure S1). To harmonize, organize, and manage all datasets and data types, the *curatedPCaData* package was built using the data structures for multi-omic data integration as implemented in the *MultiAssayExperiment* R package²⁵.

For reproducibility and to provide users with example code, analyses and results presented in the following sections are made available as vignettes through the *curatedPCaData* package. Furthermore, the individual data components used to create the *MultiAssayExperiment* objects are made available via the *ExperimentHub* package’s storage service following current guidelines for data packages intended for the Bioconductor repository.

Molecular measurements are consistent across independent datasets. There is an expectation that multiple, independent datasets that report molecular features across cancer patient cohorts with similar clinical profiles will reveal similar biological findings. If results are inconsistent between patient cohorts, differences in data processing and annotations, major batch effects or potentially biological effects could be the explanation. To test the consistency of our processed molecular measurements across patient cohorts, we evaluated patterns of transcriptome, copy number alterations, and mutations.

Gene expression, as measured by microarrays or RNA sequencing, is the most common molecular measurement in the *curatedPCaData* package (Table 1). To evaluate the consistency of expression patterns, we first performed a pairwise correlation analysis of gene expression differences in Gleason grade ≥ 8 vs. Gleason grade ≤ 6 tumor samples using the genes that were in common between the datasets (Fig. 1a). Overall, we found that pairwise Pearson correlation between datasets was generally moderate to low and statistically significant. Compared to the TCGA dataset², the reported correlations were between 0.34 and 0.48 for Taylor *et al.*^{4,39}, Weiner *et al.*^{40,41}, Barwick *et al.*^{42,43}, and IGC^{44,45}. However, not all datasets were as correlated to TCGA. For example, the Friedrich *et al.*^{46,47} dataset only showed a correlation of 0.18, which could be attributed to the difference in the underlying platform as gene expression in TCGA was measured by RNA sequencing, and Friedrich *et al.* was measured using a custom Agilent microarray.

Next, we identified the most commonly up- and down-regulated genes when comparing Gleason grade ≥ 8 vs. Gleason grade ≤ 6 tumor samples across multiple datasets (TCGA², IGC^{44,45}, Taylor *et al.*^{4,39}, Weiner *et al.*^{40,41}). We used the moderated t-test calculated through the *limma* R package to determine log fold changes and p-values for individual datasets. We then integrated the four datasets using Fisher’s method to combine p-values to identify genes that were consistently up- (n = 263) or down- (n = 501) regulated and significant

getPCa query name	Clinical end-points			Omics counts ^a	Sample counts ^b	Data source (Identifier)	Reference(s)
	Gleason or grade group	Recurrence	Survival				
abida	X		X	CNA: 444 GEX: 266	Metastatic: 444	cBioPortal (prad_su2c_2019)	Abida <i>et al.</i> ⁵⁶
backa	X			CNA: 56 MUT: 57	Metastatic: 2 Primary: 55	cBioPortal (prad_broad_2013)	Baca <i>et al.</i> ⁵⁸
barbieri	X			CNA: 109 GEX: 31 MUT: 112	Primary: 123	cBioPortal (prad_broad)	Barbieri <i>et al.</i> ⁵⁹
barwick	X	X		GEX: 146	Primary: 146	GEO (GSE18655) ⁴³	Barwick <i>et al.</i> ⁴²
chandran	X			GEX: 171	Metastatic: 25 Normal: 81 Primary: 65	GEO (GSE6919) ⁵⁵	Chandran <i>et al.</i> , Yu <i>et al.</i> ^{54,94}
friedrich	X		X	GEX: 255	BPH: 39 Normal: 52 Primary: 164	GEO (GSE134051) ⁴⁷	Friedrich <i>et al.</i> ⁴⁶
hieronymus	X		X	CNA: 104	Primary: 104	GEO (GSE54691) ⁷⁴	Hieronymus <i>et al.</i> ⁶
icgcca	X		X	GEX: 213	Primary: 213	ICGC Data portal (PRAD-CA)	Houlahan <i>et al.</i> ⁷⁵ Zhang <i>et al.</i> ⁹⁵
igc	X			GEX: 83	Primary: 83	GEO (GSE2109) ⁴⁵	IGC ⁴⁴
kim	X			GEX: 266	Primary: 266	GEO (GSE119616) ⁵⁰	Kim <i>et al.</i> ⁴⁹
kunderfranco	X			GEX: 67	Normal: 14 Primary: 53	GEO (GSE14206) ⁸⁹	Kunderfranco <i>et al.</i> ⁸⁸ Peraldo-Neia <i>et al.</i> ⁹⁶ Longoni <i>et al.</i> ⁹⁷
ren	X			GEX: 65 MUT: 65	Primary: 65	cBioPortal (prad_eururo_2017)	Ren <i>et al.</i> ⁵³
sun	X	X		GEX: 79	Primary: 79	GEO (GSE25136) ⁵²	Sun <i>et al.</i> ⁵¹
taylor	X	X		CNA: 194 GEX: 179 MUT: 43	Metastatic: 37 Normal: 29 Primary: 181	GEO (GSE21032) ³⁹ ; cBioPortal (prad_mskcc)	Taylor <i>et al.</i> ⁴
tcga	X	X	X	CNA: 492 GEX: 461 MUT: 495	Metastatic: 1 Normal: 52 Primary: 498	Xenabrowser	TCGA ²
true	X			GEX: 32	Primary: 32	GEO (GSE5132) ⁹⁸	True <i>et al.</i> ⁹⁹
wallace	X			GEX: 89	Normal: 20 Primary: 69	GEO (GSE6956) ⁸⁷	Wallace <i>et al.</i> ⁸⁶
wang	c			GEX: 148	BPH: 55 Atrophic: 21 Primary: 60	GEO (GSE8218) ⁷⁷	Wang <i>et al.</i> ⁷⁶ Jia <i>et al.</i> ¹⁰⁰
weiner	X			GEX: 838	Primary: 838	GEO (GSE157548) ⁴¹	Weiner <i>et al.</i> ⁴⁰

Table 1. Summary of studies in *curatedPCaData* and their corresponding *MultiAssayExperiment* (MAE) object contents when queried from *ExperimentHub* using the function *getPCa*. ^aCNA: Copy Number Alteration, GEX: Gene Expression, MUT: Mutations; ^bBPH: Benign Prostate Hyperplasia; ^cThe provided end-point was the proportions of cell types present in the sample determined by a pathologist (tumor, stroma, BPH, or atrophic gland).

(q -value < 0.01) across these datasets⁴⁸. Consistent with the biological processes associated with tumor growth and aggressiveness, the up-regulated genes are enriched for cell cycle-related processes, cell division, DNA replication, and DNA repair, while the down-regulated genes are enriched for positive regulation of apoptosis, negative regulation of ERK1 and ERK2 cascade, and cell-matrix adhesion. Using volcano plots for visualization and illustrative purposes, we highlighted the top 5 consistently up- (PRR16, RRM2, COMP, ASPN, PPFIA2) and top 5 consistently down-regulated genes (ANPEP, ACTG2, MYCBPC1, CD38, SLC2A3) (Fig. 1b).

Finally, for gene expression, we evaluated the consistency of correlation patterns in relation to prostate cancer-associated genes. For each dataset, we calculated the Pearson correlation of all genes within the dataset to Androgen Receptor (AR) and the ETS transcription factor, ERG. We then calculated the Spearman correlation of the correlation patterns to AR and ERG across datasets (Fig. 1c). For the majority of datasets measuring gene expression in primary prostate tumors, the correlation patterns for AR across datasets were consistent with some datasets being highly correlated, such as Kim *et al.*^{49,50} and Weiner *et al.*^{40,41}, or Taylor *et al.*^{4,39} and Sun *et al.*^{51,52}. Patterns for ERG expression were moderately to highly correlated, but there were some datasets with inverse correlation, such as Ren *et al.*⁵³ and Sun *et al.*^{51,52}, and Ren *et al.* and Barwick *et al.*^{42,43} While datasets with

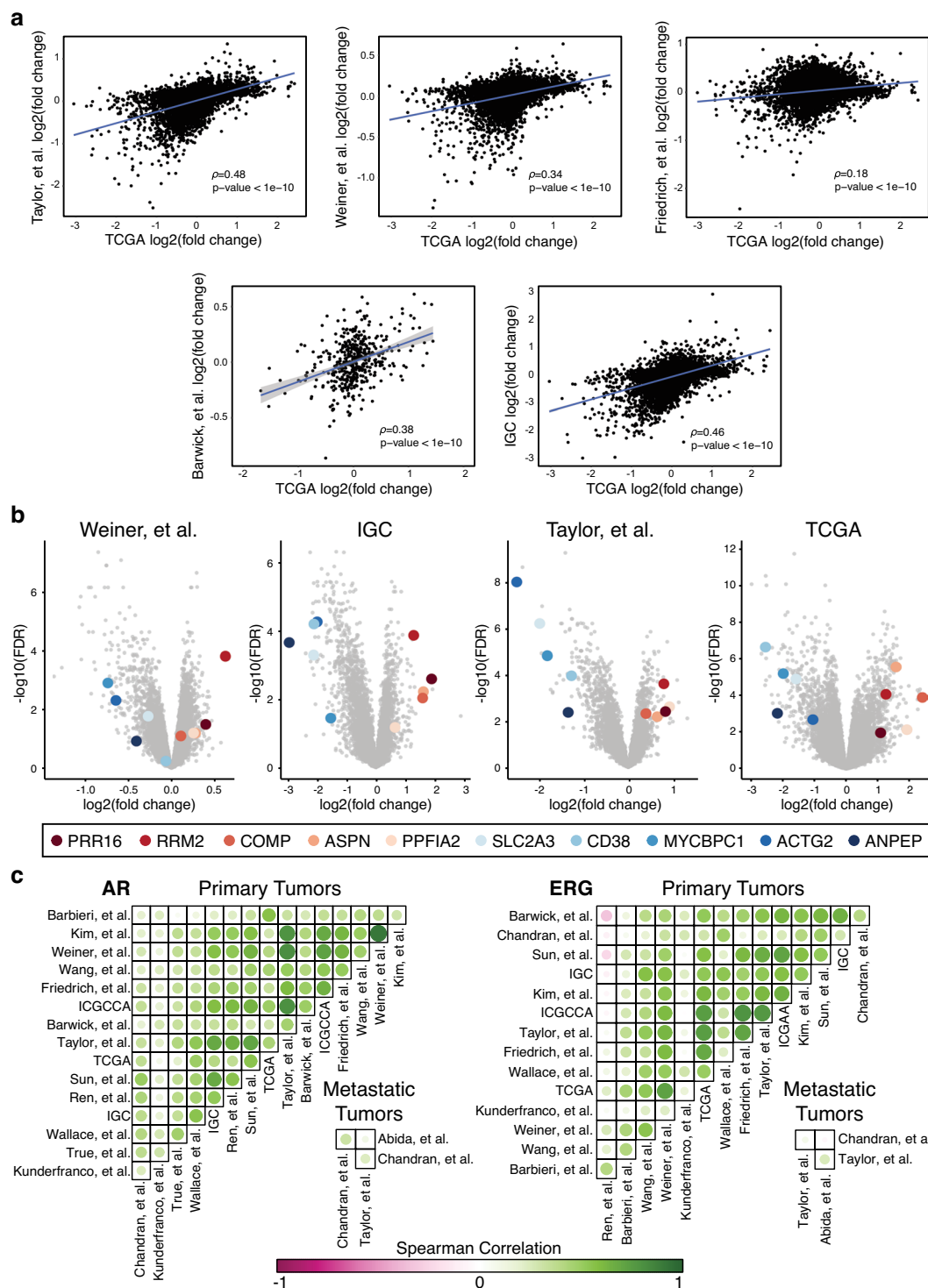


Fig. 1 Gene expression patterns across datasets. (a) Pearson correlation between datasets comparing differential expression of Gleason grade ≥ 8 vs. Gleason grade ≤ 6 samples for genes common between the datasets. (b) Volcano plots for differential gene expression comparing Gleason grade ≥ 8 vs. Gleason grade ≤ 6 samples. The highlighted genes are the top five up- and down-regulated genes identified across the four datasets using Fisher’s method to combine p-values. (c) Spearman’s rank correlations for all genes within the dataset were calculated compared to AR and the ETS transcription factor, ERG. The Spearman correlation was calculated for the correlation patterns between datasets and displayed for AR (left side) and ERG (right side) in both primary and metastatic tumors.

gene expression from metastatic tumors are few, the pattern of correlation between Chandran *et al.*^{54,55}, Abida *et al.*⁵⁶, and Taylor *et al.*^{4,39} were lower, likely due to the intrinsic heterogeneity of measuring gene expression from samples in the metastatic setting.

Prostate cancer is known to be heavily driven by copy number alterations, which will impact the molecular measurements of gene expression. For datasets with copy number alteration information, *curatedPCaData* provides discretized copy number calls according to GISTIC2 ($-2 =$ deep loss, $-1 =$ shallow loss, $0 =$ diploid, $1 =$ gain, $2 =$ amplification)⁵⁷. We evaluated the overall copy number landscape and found that independent datasets showed highly similar patterns of copy number gain and loss in primary tumors (Taylor *et al.*^{4,39}, TCGA², Baca *et al.*⁵⁸) (Fig. 2a), with samples from metastatic tumors (Abida *et al.*⁵⁶) showing an overall increase in copy number alterations as has been previously reported^{2,56}. We additionally evaluated the frequency of copy number alteration across several genes that have been shown to be associated with prostate cancer (PTEN, TP53, CHD1, MAP3K7, FOXA1, NKK3.1, USP10, SPOP^{2,4,9,58–64}), along with the TMPRSS2:ERG fusion^{2,65}. For these genes, we found the copy number alteration and mutation patterns to be consistent across datasets (Fig. 2b, note that not all datasets have all genes measured for mutations or copy number). We also tested for patterns of co-occurrence and mutual exclusivity between these genes. While general patterns of co-alteration were consistent between datasets, the statistical significance, as measured in the primary tumor setting (Taylor *et al.*^{4,39}, TCGA², Baca *et al.*⁵⁸), not surprisingly is highly dependent on the size of the dataset. In the metastatic setting (Abida *et al.*⁵⁶), the frequency of alteration is consistently much higher and many genes are statistically significantly co-altered (Fig. 2b).

Overall, these benchmarking analyses show that the molecular features in primary prostate cancer are generally reliably and consistently measured across datasets. Gene expression patterns are correlated across datasets. Copy number results were more robust across datasets, with mutational information limited to a few datasets. The consistent data processing and harmonization of gene names across datasets provide a ready to use resource for meta-analysis.

Derived features add value to published datasets. A value added in the *curatedPCaData* package, beyond data harmonization, is that features were systematically and consistently derived across datasets. Leveraging gene expression data, we inferred and evaluated estimates of risk (Oncotype DX⁶⁶, Decipher¹¹, and Prolaris¹⁰), AR scores, and microenvironment cell content leveraging the *ImmuneDeconv* R package³².

Prognostic risk scores are calculated from a select set of genes; thus, missing genes and assay platform differences can impact the reliability of the computed scores⁶⁷. To assess the impact of missing genes on risk score calculations, we benchmarked the risk scores included in *curatedPCaData* (Oncotype DX⁶⁶, Decipher¹¹, and Prolaris¹⁰) by removing different genes for calculating the risk scores, calculating the risk score with simulated missingness, followed by correlating the risk score derived from the incomplete gene set to the risk score calculated from the full gene list. Oncotype DX, a 12-gene signature, performed well overall when genes were missing from the gene list. As an example, with 5 genes missing over 100 random sampling iterations, the average correlation coefficient was 0.891 (median = 0.903) compared to the “ground truth” score using all genes (Figure S2a). Prolaris, a 34-gene signature, also proved to be highly robust whereby removing 10 random genes from the Prolaris gene list in the Kunderfranco *et al.* dataset had an average correlation with the original score of 0.973 (median = 0.974; Figure S2b). Decipher, a 17-gene signature, showed similar results to Oncotype DX where removing 5 genes resulted in an average correlation of 0.921 (median = 0.937; Figure S2c). Lastly, the AR score was calculated by taking the mean across scaled gene expression values and found to be robust to the removal of genes. There are 20 genes that are used to calculate the AR score and we found that by removing 10 at random still provides an average AR score with a correlation of 0.930 (median = 0.935; Figure S2d).

In addition to prognostic risk and AR score calculations, we performed cell type deconvolution, which infers immune cells and other stromal cells from bulk tissue gene expression profiling. For datasets with gene expression, we calculated immune and other cell estimates using EPIC²⁷, ESTIMATE²⁸, MCP-counter²⁹, quanTIseq³⁰, and xCell³¹ as implemented in the *immuneDeconv* R package³², and CIBERSORTx³⁴. While deconvolution methods vary in the types of cells that they estimate, the overall methodology has been shown to produce robust predictions and comparison between methods have been shown to be mostly consistent and robust, which is covered in depth by Sturm *et al.*³² and was a major motivation to develop the *immuneDeconv* R package. The following section highlights how the inferred cell content can be used to infer associations with clinical outcomes using *curatedPCaData*.

Endothelial cell content predicts patient outcomes. Leveraging the results from the immune and cell deconvolution methods from bulk transcriptome data, we evaluated the relationship between inferred cell types, patient outcomes, and disease progression. We found that the estimates of endothelial cell content as estimated by xCell³¹, MCP Counter²⁹, and EPIC²⁷ were predictive of biochemical recurrence. It was encouraging to also find that the results from the three independent methods were highly correlated (Fig. 3a), which provides support that the signal is reproducible and not an artifact of one deconvolution method. For illustrative purposes, we stratified patients in the TCGA² and Taylor *et al.*^{4,39} cohorts into the top 1/3 and bottom 2/3 by endothelial cell estimates, and estimated HRs using univariate Cox models for each method (EPIC, MCP-counter, and xCell). The univariate Cox models agreed on the Hazard Ratio (HR) estimates and statistical significance across the methods and datasets, with HR estimates ranging between 2.02 to 2.45 in TCGA and 1.96 to 3.54 in Taylor *et al.* (Fig. 3b). When Gleason grade group (≤ 6 , 7, ≥ 8) was modeled as a univariate Cox model predictor, its unit increase estimate for HR was of similar effect size as having the top tertile for endothelial cells with 2.15 and 3.52 for TCGA and Taylor *et al.*, respectively. Patient samples with a high endothelial score show significantly shorter times to biochemical relapse (Fig. 3c). Furthermore, we evaluated primary tumor datasets for the association between endothelial cell estimates and Gleason grade. Across the datasets that reported at least 10 patients per Gleason grade group and where we could infer endothelial cell content from gene expression data (TCGA², Taylor *et al.*^{4,39}, Friedrich *et al.*^{46,47}), we consistently found an increased estimated presence of endothelial cells in Gleason grade ≥ 8 compared to Gleason grade 7 or ≤ 6 (Fig. 3d).

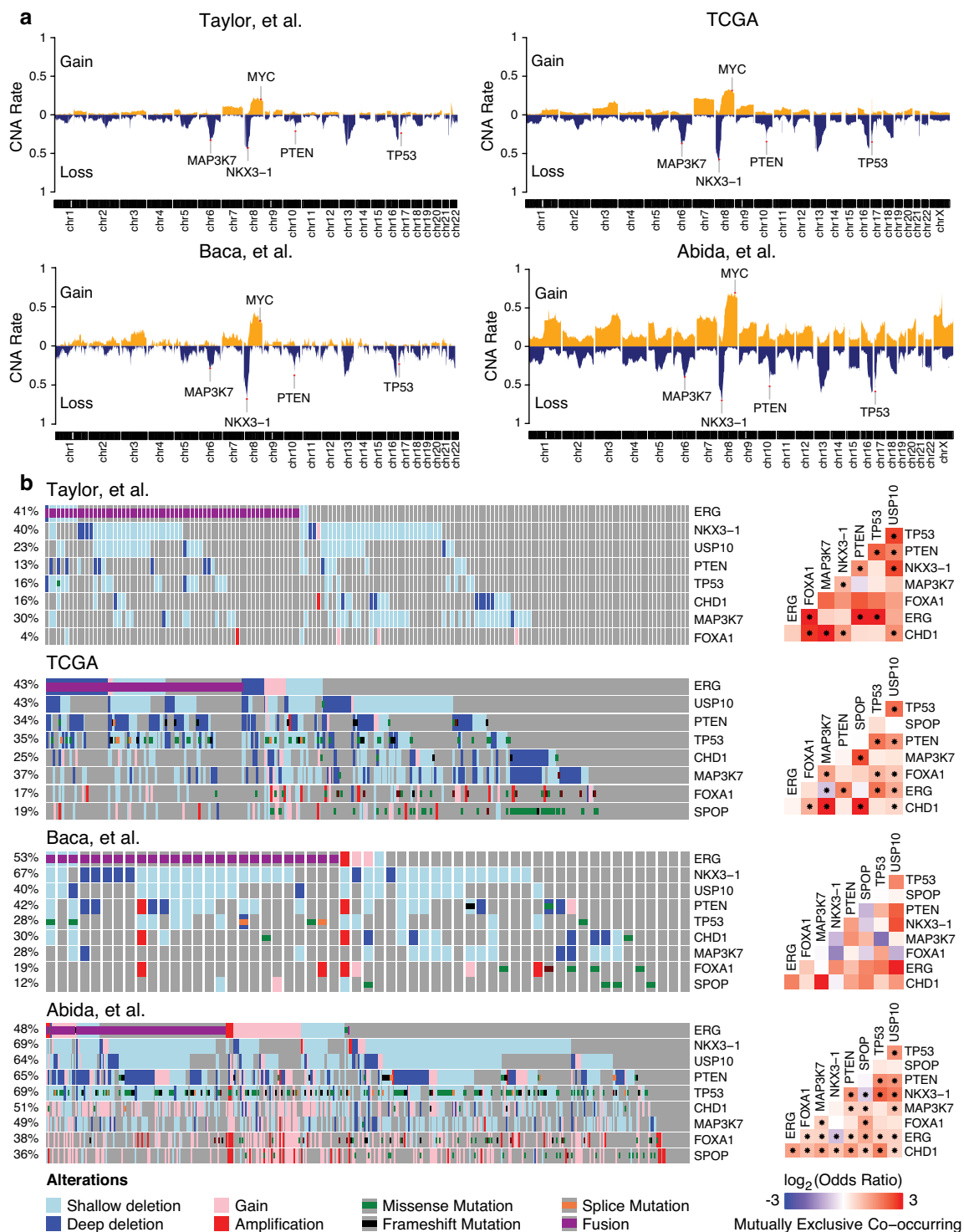


Fig. 2 Copy number and mutational landscapes across datasets. (a) Multiple known prostate cancer-associated genes (MAP3K7, MYC, NKX3-1, PTEN, TP53) displayed consistent copy number loss/deletion or gain/amplification across datasets. (b) Oncoprints (left side) for select prostate cancer-associated genes are displayed across datasets. Mutual exclusivity (right side) was calculated using Fisher’s exact test (* $p < 0.05$). Note that due to a lack of overlap in ‘omics’, some alteration percentages combining CNA and mutations are under-estimated; for example Taylor *et al.*^{4,39} used a targeted sequencing panel, and thus not all genes were measured for somatic mutations.

It has been established that the cellular content of the tumor microenvironment can be predictive of tumor progression and response to treatment, mostly in the context of immune cells³³. Similarly, angiogenesis and the vascularization of the tumor microenvironment have been associated with tumor progression and outcomes^{68–71},

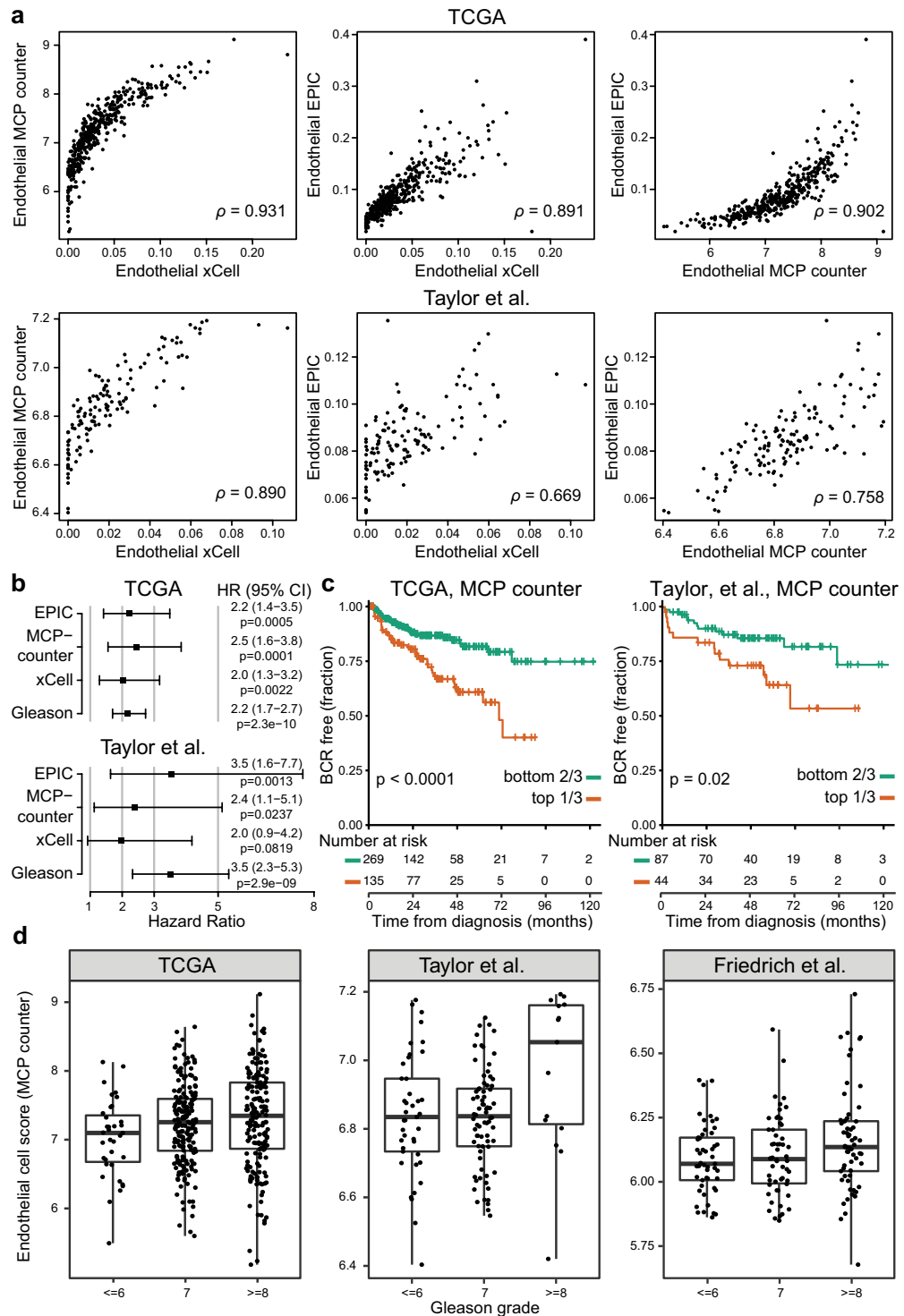


Fig. 3 Estimates of endothelial cell content are associated with clinical outcomes. **(a)** The endothelial cell scores calculated from gene expression across TCGA² and Taylor *et al.*^{4,39} were highly correlated (Spearman correlation) across the three estimation methods, EPIC²⁷, MCP-counter²⁹, and xCell³¹. **(b)** Forest plots for univariate Cox proportional hazard models illustrate that all three methods and Gleason grade were predictive of biochemical recurrence. **(c)** Endothelial cell score top tertiles, as illustrated using MCP-counter’s estimates, showed a statistically significant stratification for worse outcome in TCGA and Taylor *et al.* datasets. **(d)** In addition to being associated with biochemical recurrence, the estimates from MCP-counter are associated with tumor Gleason grade groups.

with specific studies linking endothelial cell content to prostate cancer aggressiveness^{72,73}. Our findings are consistent with previous results and demonstrate the strength of leveraging the inferred features across multiple, independent datasets through *curatedPCaData*.

Discussion

The *curatedPCaData* R package provides a harmonized and centralized resource for prostate cancer studies with multi-omic and clinical data that can be leveraged easily for cancer research. The cross-study analyses presented herein demonstrate the strength of leveraging multiple studies in PCA; however, it is important to understand and incorporate relative differences between studies, their aims, design, and the underlying composition in such data analysis. For example, Abida *et al.*⁵⁶ focused on the progressed metastatic form of the disease and reported a significant number of disease-related deaths suitable for overall survival modeling. On the other hand, Friedrich *et al.*^{46,47}, Hieronymus *et al.*^{6,74}, ICGC-CA⁷⁵, and TCGA² also reported overall survival, but they present a more indolent form of the disease with a lower count of deaths, making survival modeling more challenging. Furthermore, biochemical recurrence is often used as a surrogate for progression-free survival and is reported in Barwick *et al.*^{42,43}, Sun *et al.*^{51,52}, Taylor *et al.*^{4,39} and TCGA²; of these four datasets, we focused our Cox models for recurrence on Taylor *et al.* and TCGA, as Barwick *et al.* used a very targeted custom DASL gene panel (<1,000 genes) making cell composition estimation unreliable for most methods. Sun *et al.* only report recurrence as a binary outcome without follow-up times, rendering it unsuitable for Cox proportional hazards models or survival estimation using the Kaplan-Meier method. Despite the differences in reported variables, a considerable amount of clinical information is made available across independent datasets to draw associations with molecular features.

Researchers should also consider the original study aims, as these will be reflected in which metadata fields and 'omics' that will be available. For example, Weiner *et al.*^{40,41} studied ethnicity-related PCA-trends, thus the patients had accurate demographics-related metadata commonly available, while samples were just described as being primary tumors. In contrast, Wang *et al.*^{76,77} studied how sample composition (tumor cells, stroma, atrophic gland, or benign prostate hyperplasia) could be differentiated based on gene expression, thus providing metadata suitable for tumor purity estimation, but provided no clinical end-points or patient characteristics. While we have gone through great effort to minimize technical and reporting variability, some fundamental study characteristics will inevitably not be comparable. Thus, combining studies should be planned with care to avoid introducing confounding effects. To this end, *curatedPCaData* offers assistance in bringing together studies suitable for efficiently tackling specific prostate cancer related research questions.

Additional consideration should be given to how studies reported the common end-point of Gleason grade. In *curatedPCaData*, we provided summarized results across studies as Gleason grade groups (≤ 6 , 7 , ≥ 8), though studies might have additional information to report. For example, Weiner *et al.*^{40,41} reported an International Society of Urologic Pathologists (ISUP) disease stage ranging from 1–5, for which the suggested mapping to the traditional Gleason grade was done⁷⁸. Multiple studies reported Gleason as the sum of major + minor Gleason grades or a grade group (≤ 6 , 7 , ≥ 8), thus groupings were offered as an endpoint with an equal level of granularity, while a finer level of detail was offered in alternate clinical metadata columns when available. In ambiguous cases, the primary publications and the supplementary material were mined, along with contacting the primary authors in many cases, in an effort to offer accurate and up-to-date information on both the clinical metadata and the primary data. For this purpose, a great deal of manual labor was required to curate the *curatedPCaData* datasets. The resulting datasets were thus standardized to be as comparable as possible, while retaining details essential to the studies. To this end, we offer a great variety of R package vignettes alongside *curatedPCaData* with numerous examples and extra data characteristics, which assist the end-user in planning their analyses.

One benefit of *curatedPCaData* is that it greatly lowers the barrier for accessing data to rapidly test hypotheses and generate novel hypotheses supported by multiple, independent datasets. The code used to generate the MAE objects is offered within the R package and GitHub repository. The processed MAE objects exported from the package are the main focus of the package; however, from a developer point of view, they also offer natural potential for future extensions such as: a) adding new studies and exporting them as new MAE objects using the pipelines developed in *curatedPCaData*; b) supplementing the existing MAE slots with newly derived variables or even adding other primary 'omics' data; or c) extending the existing clinical metadata fields to include new fields.

Currently, *curatedPCaData* offers a base R Shiny⁷⁹ interface to the package as well, with plans to extend the visual browser-based access to the data. While ongoing efforts such as the NCI Genomic Data Commons⁸⁰, cBioPortal⁸¹, or the International Cancer Genome Consortium⁸² already aim to provide a standardized approach to tackling complex 'omics' traits in cancer, *curatedPCaData* is the first harmonized, multi-study, hands-on data resource intended for analysts with a strong focus on PCA and allowing for maximum flexibility of the analyses, using the R statistical software⁸³. As such, the presented proof-of-concept analyses provide merely a staging platform for more efficient exploration of multi-omics signatures coupled with clinical metadata for the wider research community for prostate cancer.

Methods

Data acquisition. Gene expression, copy number alterations, and mutation data were downloaded from Gene Expression Omnibus (GEO)⁸⁴ using *GEOquery* (R package version 2.64.2) and from cBioPortal⁸¹ using *cBioPortalData* (R package version v2.8.2) and *cgdsr* (R package version v1.3.0) (Figure S1a). In addition to downloading raw data from GEO, *GEOquery* was used for downloading the latest array-specific annotations and all three R packages were further utilized to download clinical metadata accompanying the raw data. Raw CEL-file files for Affymetrix-arrays were RMA-normalized in *oligo* (R package version v1.62.1) with functions *read.celfiles*, *rma*, *getNetAffx*, and *exprs*. Agilent arrays were processed using *limma* (R package version v3.52.2) with the functions *read.maimages*, *backgroundCorrect*, *normalizeBetweenArrays*, and *avereps*. For custom arrays such as the DASL array in Barwick *et al.*^{42,43}, quantile normalization was used together with log-transformation. No additional normalization was done on the gene expression data from cBioPortal, since cBioPortal offers pre-normalized data. For data with raw copy number alteration available, these were processed using *rCGH* (R package

version v1.26.0) with functions *readAgilent*, *adjustSignal*, *segmentCGH*, and *EMnormalize*. This yielded log-ratios, which were input to GISTIC2⁵⁷ when available. Copy number alteration matrices from cBioPortal with pre-existing GISTIC2 calls were stored with the discretized calls consistently across all the datasets. A summary of the acquired datasets and their sources is presented in Table 1.

The TCGA Prostate Cancer (PRAD) dataset was downloaded from Xena Browser⁸⁵, due to better data quality and providing tumor samples and normal samples separately, instead of providing relative tumor to normal gene expression found in cBioPortal processed data. We also removed low-quality samples which were excluded from the TCGA publication due to RNA degradation from the gene expression matrix to provide users with the most reliable information. We followed uniform naming conventions for all the metadata fields and leveraged data in the original publications to obtain maximum information in case information wasn't readily available in these public repositories³⁸.

All layers of data, namely the gene expression, copy number alterations, and mutations, underwent a harmonization process to ensure uniform gene naming conventions. Note that some datasets have matched normal samples to call somatic mutations and some datasets do not have matched normal samples and are thus tumor-only variants. The mutation calling status is noted in the "Mutation_status" field. The latest hg38 gene symbols, aliases, and locations were downloaded using *biomaRt* (R package version v2.52.0). We then mapped all the gene names to the up-to-date dictionary to ensure consistency in HGNC symbols across all datasets. A liftover from hg19 to hg38 was done as part of the harmonization using the *liftOver* function from *rtracklayer* (R package version v1.56.1), for mutations called with an older genome assembly to ensure uniformity.

Clinicopathological features were processed using R scripts customized to each dataset. Features were collected from supplementary annotation files and processed to map features to the data dictionary. The data dictionary ensured common terminology and some additional features, such as Gleason grade group (where not supplied by the primary publication), were inferred using a predefined set of rules. The scripts for each dataset are made available in *curatedPCaData*.

Derived features. A number of derived features were computed for the final MAE-objects (Figure S1b). Using gene expression data, we calculated cell proportions, genomic risk scores, and AR scores. The *immune-deconv*³² (R package version v2.1.0) wrapper package was used to estimate cell proportions from EPIC²⁷, ESTIMATE²⁸, MCP-counter²⁹, quanTIseq³⁰, and xCell³¹. As the implementation of CIBERSORTx³⁴ required external access using the free academic license, it was run with default parameters on their web interface and quantile normalization disabled with the normalized gene expression data as input and LM22 signature matrix used to infer cell types. The output CIBERSORTx matrices were then downloaded and integrated into the MAEs.

Due to the different platforms (sequencing, different brands, and versions of microarrays) used to assess gene expression, not all datasets have the same set of genes. To determine the impact that gene missingness on the precomputed scores would have on those studies without all genes, we benchmarked the Oncotype DX⁶⁶, Decipher¹¹, and Prolaris¹⁰ risk scores and the AR score. This was performed by identifying the study in *curatedPCaData* that contained the most genes belonging to the scoring method. By using this study, we were able to get as close to what the true score would be. Assessing the impact of missing genes was performed by randomly removing genes to simulate missing between 1 and 10 genes for Prolaris¹⁰ risk score (34 genes in the complete signature) and AR score (20 genes), and removing between 1 and 5 for Oncotype DX⁶⁶ and Decipher¹¹ risk scores (12 and 20 genes, respectively). Since the number of gene combinations that can be made by simulating 10 missing genes for a risk score such as Prolaris¹⁰ is large, the combinations were sampled to cut down on vignette and package build time. The number of combinations used for assessing impact of missingness in Decipher¹¹, Oncotype DX⁶⁶, and AR scores was 100 while Prolaris risk score used 50 combinations.

We implemented the Oncotype DX⁶⁶, Decipher¹¹, and Prolaris¹⁰ risk scores based on the instructions in their original publications supported by the implementation outlined in Creed *et al.*⁶⁷ The gene list (n = 12 matching genes) for Oncotype DX matched perfectly with several studies: Abida *et al.*⁵⁶, Kim *et al.*^{49,50}, Ren *et al.*⁵³, Sun *et al.*^{51,52}, Taylor *et al.*^{4,39}, TCGA², Wallace *et al.*^{86,87}, and Weiner *et al.*^{40,41} We considered TCGA to be the most complete dataset as well as most widely used, thus we used the gene expression from TCGA for testing the variability of the Oncotype DX score due to missing genes (Table 2). The gene list (n = 17 matching genes in TCGA) for Decipher did not have a 1-to-1 match with any study in *curatedPCaData*, but did have the highest number of matching genes in Ren *et al.*⁵³ (18 genes were a 1-to-1 match with two genes from Decipher missing) while Abida *et al.*⁵⁶, Friedrich *et al.*^{46,47}, and TCGA² had slightly fewer number of matching genes (17 genes were a 1-to-1 with 3 genes missing). We used TCGA gene expression for benchmarking inferred risk scores from Decipher. Prolaris required the largest number of genes (n = 34 matching genes) to calculate risk. Kunderfranco *et al.*^{88,89} had the highest number of matching genes with 32 1-to-1 matches and only 2 genes missing. The next highest 1-to-1 match was ICGC-CA⁷⁵ where 29 genes were 1-to-1 matches. Because of the high number of matching genes, we selected Kunderfranco *et al.* as the benchmarking study for Prolaris (Table 2).

AR-scores were calculated for the 20 genes identified originally in Hieronymus *et al.*⁹⁰ and then calculated as the sum of z-scores of AR signaling genes as described by TCGA². There were 8 studies that matched all 20 genes used to calculate the AR score; we leveraged TCGA gene expression for benchmarking.

Statistical analysis. While the primary focus is on providing readily processed MAE-objects with *MultiAssayExperiment* (R package version v1.21.6), *curatedPCaData* delivers several application examples as R vignettes and documentation, with relevant statistical methodology applied therein. Cox proportional hazard models and Kaplan-Meier (KM) curves were fitted with *survival* (R package version v3.3-1) and plotted using *survminer* (R package version v0.4.9), and the corresponding p-values were calculated using log-rank tests.

MAE Object	Score with Perfect Match	Prolaris			Oncotype DX			Decipher			Androgen Receptor (AR)		
		Genes with Single Match	Genes with Multiple Matches	Genes Missing	Genes with Single Match	Genes with Multiple Matches	Genes Missing	Genes with Single Match	Genes with Multiple Matches	Genes Missing	AR: one match	AR: multiple matched	AR: no match
abida ⁵⁶	Oncotype DX	28	2	4	12	—	—	17	—	3	19	1	—
backa ⁵⁸		—	—	34	—	—	12	—	—	20	—	—	20
barbieri ⁵⁹		27	2	5	11	—	1	15	—	5	19	1	—
barwick ^{42,43}		6	—	28	1	—	11	2	—	18	1	—	19
chandran ^{54,55,94}		14	2	18	11	—	1	10	—	10	15	—	5
friedrich ^{46,47}		28	2	4	11	—	1	17	—	3	19	1	—
hieronymus ^{6,74}		—	—	34	—	—	12	—	—	20	—	—	20
icgcca ^{75,95}		29	1	4	11	—	1	15	—	5	17	1	2
igc ^{44,45}		27	2	5	11	—	1	13	—	7	17	1	2
kim ^{49,50}	Oncotype DX	28	3	3	12	—	—	16	—	4	19	1	—
kunderfranco ^{88,89,96,97}		32	2	—	11	—	1	16	—	4	18	1	1
ren ⁵³	Oncotype DX	28	2	4	12	—	—	18	—	2	19	1	—
sun ^{51,52}	Oncotype DX	27	2	5	12	—	—	13	—	7	18	1	1
taylor ^{4,39}	Oncotype DX	28	3	3	12	—	—	16	—	4	19	1	—
tcga ²	Oncotype DX	28	2	4	12	—	—	17	—	3	19	1	—
true ^{98,99}		9	—	25	6	—	6	5	—	15	12	1	7
wallace ^{86,87}	Oncotype DX	27	2	5	12	—	—	13	—	7	18	1	1
wang ^{76,77,100}		27	2	5	11	—	1	13	—	7	17	1	2
weiner ^{40,41}	Oncotype DX	28	3	3	12	—	—	16	—	4	19	1	—

Table 2. The intersection between Prolaris, Oncotype DX, Decipher, and Androgen Receptor (AR) score' genes and genes that are found in studies within curatedPCaData R Package. A gene from the score or its aliases matched either with a single gene in the dataset (1-to-1 match), gene from the score matched or its aliases had multiple matches in the dataset (1-to-many), or the gene from the score calculation was missing from the dataset altogether.

Differential gene expression was calculated as the average log-transformed expression of Gleason grade ≥ 8 samples minus the average log-transformed expression of Gleason grade ≤ 6 samples. Statistical significance was determined by comparing the log-transformed gene expression of Gleason grade ≥ 8 compared to Gleason grade ≤ 6 samples using the moderated t-test as implemented in *limma* (R package version v3.52.2). The final p-values were adjusted for multiple testing using Benjamini-Hochberg correction. Pearson correlation was used to compare differential expression in Fig. 1a. The genes reported in Fig. 1b were identified using Fisher's method to combine p-values for statistical significance. The log fold change was then tested to ensure consistent up- and down-regulation of the associated gene, meaning a gene needed to have $\log_{2}FC > 0$ or $\log_{2}FC < 0$ across all four datasets tested. The top up- and down-regulated gene sets were tested for pathway and biological process enrichment using the DAVID web server⁹¹. The correlations reported in Fig. 1c were calculated using Spearman's rank correlation.

Genes were defined to be co-occurring or mutually exclusive based on the odds ratio (OR) which is calculated as: $OR = (Both * Neither) / (B \text{ Not } A * A \text{ Not } B)$ where A and B stand for alterations in genes A and B respectively. We define any alteration in copy number or mutations that are not silent as an alteration. The significance of mutual exclusivity/co-occurrence was computed using the Fisher's Exact Test and the Benjamini-Hochberg correction was applied to determine the adjusted p-values. Mutual exclusivity plots for different data sets shown in Fig. 2b (right side) provide information on whether or not a set of important genes in PCa are significantly altered together.

Statistical modeling used to identify interesting derived features predictive of biochemical recurrence were based on 10-fold cross-validation (CV) of Cox models regularized using LASSO from *glmnet* (R package version v4.1-4)⁹². There were three methods that calculated endothelial cell abundance scores (EPIC²⁷, MCP-counter²⁹, and xCell³¹). Among these methods, endothelial cell abundance scores were predictive in at least one of these datasets, when predictive features were chosen according to the optimal regularization coefficient λ in the CV-curve.

Spearman's rank correlation was used to assess the non-linear association between endothelial cell scores in Fig. 3a. Cox proportional hazards models were fit as univariate models with biochemical recurrence as an endpoint, by introducing one of the endothelial scores at a time to a separate model compared with using Gleason score sum as a univariate predictor; these were then plotted together as a forest plot in Fig. 3b.

Data availability

All the data presented herein are available as *MultiAssayExperiments*²⁵ via the *curatedPCaData* R package (<https://github.com/Syksy/curatedPCaData>) along with code that can be used to reproduce these objects. The original raw data repositories along with unique identifiers are listed, such as GEO accession IDs or cBioPortal identifiers listed in Table 1.

Code availability

All the code used to generate the processed datasets, as well as the resulting R package are available openly on GitHub (<https://github.com/Syksy/curatedPCaData>). The DOI-linked copy of the package's GitHub repository is available via Zenodo⁹³.

Received: 18 January 2023; Accepted: 27 June 2023;

Published online: 05 July 2023

References

- Siegel, R. L., Miller, K. D., Fuchs, H. E. & Jemal, A. Cancer statistics, 2022. *CA Cancer J. Clin.* **72**, 7–33 (2022).
- Cancer Genome Atlas Research Network. The Molecular Taxonomy of Primary Prostate Cancer. *Cell* **163**, 1011–1025 (2015).
- Robinson, D. *et al.* Integrative clinical genomics of advanced prostate cancer. *Cell* **161**, 1215–1228 (2015).
- Taylor, B. S. *et al.* Integrative genomic profiling of human prostate cancer. *Cancer Cell* **18**, 11–22 (2010).
- Grasso, C. S. *et al.* The mutational landscape of lethal castration-resistant prostate cancer. *Nature* **487**, 239–243 (2012).
- Hieronimus, H. *et al.* Copy number alteration burden predicts prostate cancer relapse. *Proc. Natl. Acad. Sci. USA* **111**, 11139–11144 (2014).
- Hieronimus, H. *et al.* Tumor copy number alteration burden is a pan-cancer prognostic factor associated with recurrence and death. *Elife* **7**, (2018).
- Ku, S. Y. *et al.* Rb1 and Trp53 cooperate to suppress prostate cancer lineage plasticity, metastasis, and antiandrogen resistance. *Science* **355**, 78–83 (2017).
- Rodrigues, L. U. *et al.* Coordinate loss of MAP3K7 and CHD1 promotes aggressive prostate cancer. *Cancer Res.* **75**, 1021–1034 (2015).
- Cuzick, J. *et al.* Prognostic value of an RNA expression signature derived from cell cycle proliferation genes in patients with prostate cancer: a retrospective study. *Lancet Oncol.* **12**, 245–255 (2011).
- Erho, N. *et al.* Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy. *PLoS One* **8**, e66855 (2013).
- Na, R., Wu, Y., Ding, Q. & Xu, J. Clinically available RNA profiling tests of prostate tumors: utility and comparison. *Asian J. Androl.* **18**, 575–579 (2016).
- Spratt, D. E. *et al.* Individual Patient-Level Meta-Analysis of the Performance of the Decipher Genomic Classifier in High-Risk Men After Prostatectomy to Predict Development of Metastatic Disease. *J. Clin. Oncol.* **35**, 1991–1998 (2017).
- Klein, E. A. *et al.* A 17-gene assay to predict prostate cancer aggressiveness in the context of Gleason grade heterogeneity, tumor multifocality, and biopsy undersampling. *Eur. Urol.* **66**, 550–560 (2014).
- Penney, K. L. *et al.* mRNA expression signature of Gleason grade predicts lethal prostate cancer. *J. Clin. Oncol.* **29**, 2391–2396 (2011).
- Sinnott, J. A. *et al.* Prognostic Utility of a New mRNA Expression Signature of Gleason Score. *Clin. Cancer Res.* **23**, 81–87 (2017).
- Yamoah, K. *et al.* Novel Biomarker Signature That May Predict Aggressive Disease in African American Men With Prostate Cancer. *J. Clin. Oncol.* **33**, 2789–2796 (2015).
- Tomlins, S. A. *et al.* Characterization of 1577 primary prostate cancers reveals novel biological and clinicopathologic insights into molecular subtypes. *Eur. Urol.* **68**, 555–567 (2015).
- Chen, Z., Gerke, T., Bird, V. & Prosperi, M. Trends in Gene Expression Profiling for Prostate Cancer Risk Assessment: A Systematic Review. *Biomed Hub* **2**, 1–15 (2017).
- Patil, P. & Parmigiani, G. Training replicable predictors in multiple studies. *Proc. Natl. Acad. Sci. USA* **115**, 2578–2583 (2018).
- Bernau, C. *et al.* Cross-study validation for the assessment of prediction algorithms. *Bioinformatics* **30**, i105–112 (2014).
- Ganzfried, B. F. *et al.* curatedOvarianData: clinically annotated data for the ovarian cancer transcriptome. *Database* **2013**, bat013 (2013).
- Planey, K. *curatedBreastData: Curated breast cancer gene expression data with survival and treatment information.* (R package).
- Ramos, M. *et al.* Multiomic Integration of Public Oncology Databases in Bioconductor. *JCO Clin Cancer Inform* **4**, 958–971 (2020).
- Ramos, M. *et al.* Software for the Integration of Multiomics Experiments in Bioconductor. *Cancer Res.* **77**, e39–e42 (2017).
- Sturm, G. *et al.* Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics* **35**, i436–i445 (2019).
- Racle, J., de Jonge, K., Baumgaertner, P., Speiser, D. E. & Gfeller, D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *Elife* **6**, (2017).
- Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4**, 2612 (2013).
- Becht, E. *et al.* Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol.* **17**, 218 (2016).
- Finotello, F. *et al.* Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. *Genome Med.* **11**, 34 (2019).
- Aran, D., Hu, Z. & Butte, A. J. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* **18**, 220 (2017).
- Sturm, G., Finotello, F. & List, M. Immunedeconv: An R Package for Unified Access to Computational Methods for Estimating Immune Cell Fractions from Bulk RNA-Sequencing Data. *Methods Mol. Biol.* **2120**, 223–232 (2020).
- Gentles, A. J. *et al.* The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat. Med.* **21**, 938–945 (2015).
- Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
- Herlemann, A. *et al.* Decipher identifies men with otherwise clinically favorable-intermediate risk disease who may not be good candidates for active surveillance. *Prostate Cancer Prostatic Dis.* **23**, 136–143 (2020).
- Knezevic, D. *et al.* Analytical validation of the Oncotype DX prostate cancer assay - a clinical RT-PCR assay optimized for prostate needle biopsies. *BMC Genomics* **14**, 690 (2013).
- NICE Advice - Prolaris gene expression assay for assessing long-term risk of prostate cancer progression: © NICE (2016). Prolaris gene expression assay for assessing long-term risk of prostate cancer progression. *BJU Int.* **122**, 173–180 (2018).
- Laajala, T. D. *et al.* curatedPCaData: metadata template. *Zenodo* <https://doi.org/10.5281/zenodo.7995819> (2023).
- Taylor, B.S., Schultz, N., Hieronimus, H. & Sawyers, C.L. *GEO*, <https://identifiers.org/geo/GSE21032> (2010).
- Weiner, A. B. *et al.* Plasma cells are enriched in localized prostate cancer in Black men and are associated with improved outcomes. *Nat. Commun.* **12**, 935 (2021).
- Davicioni, E. *GEO* <https://identifiers.org/geo/GSE157548> (2020).
- Barwick, B. G. *et al.* Prostate cancer genes associated with TMPRSS2-ERG gene fusion and prognostic of biochemical recurrence in multiple cohorts. *Br. J. Cancer* **102**, 570–576 (2010).
- Barwick, B.G., Seth, A., Leyland-Jones, B.R. & Abramovitz, M. *GEO*, <https://identifiers.org/geo/GSE18655> (2009).
- The International Genomics Consortium. *IGC* <https://intgen.org/> (2009).

45. Curley, E. *GEO*, <https://identifiers.org/geo:GSE2109> (2005).
46. Friedrich, M. *et al.* The Role of lncRNAs TAPIR-1 and -2 as Diagnostic Markers and Potential Therapeutic Targets in Prostate Cancer. *Cancers* **12** (2020).
47. Baretton, GB. *et al.* *GEO*, <https://identifiers.org/geo:GSE134051> (2020).
48. Laajala, T. D. *et al.* *curatedPCaData: differential gene expression analysis*. *Zenodo* <https://doi.org/10.5281/zenodo.7988148> (2023).
49. Kim, H. L. *et al.* Validation of the Decipher Test for predicting adverse pathology in candidates for prostate cancer active surveillance. *Prostate Cancer Prostatic Dis.* **22**, 399–405 (2019).
50. duPlessis, M. *et al.* *GEO* <https://identifiers.org/geo:GSE119616> (2018).
51. Sun, Y. & Goodison, S. Optimizing molecular signatures for predicting prostate cancer recurrence. *Prostate* **69**, 1119–1127 (2009).
52. Goodison, S. & Sun, Y. *GEO* <https://identifiers.org/geo:GSE25136> (2010).
53. Ren, S. *et al.* Whole-genome and Transcriptome Sequencing of Prostate Cancer Identify New Genetic Alterations Driving Disease Progression. *Eur. Urol.* <https://doi.org/10.1016/j.eururo.2017.08.027> (2017).
54. Chandran, U. R. *et al.* Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process. *BMC Cancer* **7**, 64 (2007).
55. Monzon, FA. *GEO*, <https://identifiers.org/geo:GSE6919> (2007).
56. Abida, W. *et al.* Prospective Genomic Profiling of Prostate Cancer Across Disease States Reveals Germline and Somatic Alterations That May Affect Clinical Decision Making. *JCO Precis Oncol* **2017** (2017).
57. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
58. Baca, S. C. *et al.* Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666–677 (2013).
59. Barbieri, C. E. *et al.* Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nat. Genet.* **44**, 685–689 (2012).
60. Kaffenberger, S. D. & Barbieri, C. E. Molecular Subtyping of Prostate Cancer. *Curr. Opin. Urol.* **26**, 213–218 (2016-5).
61. Song, M. S., Salmena, L. & Pandolfi, P. P. The functions and regulation of the PTEN tumour suppressor. *Nat. Rev. Mol. Cell Biol.* **13**, 283–296 (2012).
62. Liu, W. *et al.* Genetic markers associated with early cancer-specific mortality following prostatectomy. *Cancer* **119**, 2405–2412 (2013).
63. Liu, W. *et al.* Deletion of a small consensus region at 6q15, including the MAP3K7 gene, is significantly associated with high-grade prostate cancers. *Clin. Cancer Res.* **13**, 5028–5033 (2007).
64. Wu, M. *et al.* Suppression of Tak1 promotes prostate tumorigenesis. *Cancer Res.* **72**, 2833–2843 (2012).
65. Tomlins, S. A. *et al.* Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310**, 644–648 (2005).
66. Cullen, J. *et al.* A Biopsy-based 17-gene Genomic Prostate Score Predicts Recurrence After Radical Prostatectomy and Adverse Surgical Pathology in a Racially Diverse Population of Men with Clinically Low- and Intermediate-risk Prostate Cancer. *Eur. Urol.* **68**, 123–131 (2015).
67. Creed, J. H. *et al.* Commercial Gene Expression Tests for Prostate Cancer Prognosis Provide Paradoxical Estimates of Race-Specific Risk. *Cancer Epidemiol. Biomarkers Prev.* **29**, 246–253 (2020).
68. Rak, J. W., St Croix, B. D. & Kerbel, R. S. Consequences of angiogenesis for tumor progression, metastasis and cancer therapy. *Anticancer Drugs* **6**, 3–18 (1995).
69. Zuazo-Gatzelu, I. & Casanovas, O. Unraveling the Role of Angiogenesis in Cancer Ecosystems. *Front. Oncol.* **8**, 248 (2018).
70. Choi, H. & Moon, A. Crosstalk between cancer cells and endothelial cells: implications for tumor progression and intervention. *Arch. Pharm. Res.* **41**, 711–724 (2018).
71. Oshi, M. *et al.* Abundance of Microvascular Endothelial Cells Is Associated with Response to Chemotherapy and Prognosis in Colorectal Cancer. *Cancers* **13** (2021).
72. Bahmad, H. F. *et al.* Tumor Microenvironment in Prostate Cancer: Toward Identification of Novel Molecular Biomarkers for Diagnosis, Prognosis, and Therapy Development. *Front. Genet.* **12**, 652747 (2021).
73. Quinn, D. I., Henshall, S. M. & Sutherland, R. L. Molecular markers of prostate cancer outcome. *Eur. J. Cancer* **41**, 858–887 (2005).
74. Hieronymus, H., Schultz, N., Taylor, B. S. & Sawyers, C. L. *GEO* <https://identifiers.org/geo:GSE54691> (2014).
75. Houlihan, K. E. *et al.* Genome-wide germline correlates of the epigenetic landscape of prostate cancer. *Nat. Med.* **25**, 1615–1626 (2019).
76. Wang, Y. *et al.* In silico estimates of tissue components in surgical samples based on expression profiling data. *Cancer Res.* **70**, 6448–6455 (2010).
77. Wang, Y. *GEO*, <https://identifiers.org/geo:GSE8218> (2007).
78. Egevad, L., Delahunt, B., Strigley, J. R. & Samarutunga, H. International Society of Urological Pathology (ISUP) grading of prostate cancer - An ISUP consensus on contemporary grading. *APMIS* **124**, 433–435 (2016).
79. Chang, W. *et al.* shiny: Web Application Framework for R. <https://shiny.rstudio.com/> (2022).
80. Grossman, R. L. *et al.* Toward a Shared Vision for Cancer Genomic Data. *N. Engl. J. Med.* **375**, 1109–1112 (2016).
81. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**, 11 (2013).
82. Zhang, J. *et al.* The International Cancer Genome Consortium Data Portal. *Nat. Biotechnol.* **37**, 367–369 (2019).
83. R Core Team (2018) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.r-project.org/>.
84. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets-update. *Nucleic Acids Res.* **41**, D991–995 (2013).
85. Goldman, M. J. *et al.* Visualizing and interpreting cancer genomics data via the Xena platform. *Nat. Biotechnol.* **38**, 675–678 (2020).
86. Wallace, T. A. *et al.* Tumor immunobiological differences in prostate cancer between African-American and European-American men. *Cancer Res.* **68**, 927–936 (2008).
87. Ambs, S., Hudson, R. & Yi, M. *GEO*, <https://identifiers.org/geo:GSE6956> (2008).
88. Kunderfranco, P. *et al.* ETS transcription factors control transcription of EZH2 and epigenetic silencing of the tumor suppressor gene Nkx3.1 in prostate cancer. *PLoS One* **5**, e10547 (2010).
89. Kunderfranco, P. *et al.* *GEO*, <https://identifiers.org/geo:GSE14206> (2010).
90. Hieronymus, H. *et al.* Gene expression signature-based chemical genomic prediction identifies a novel class of HSP90 pathway modulators. *Cancer Cell* **10**, 321–330 (2006).
91. Sherman, B. T. *et al.* DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res.* **50**, W216–21 (2022).
92. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
93. Laajala, T. D. *et al.* *curatedPCaData 0.99.1*. *Zenodo* <https://doi.org/10.5281/zenodo.7996377> (2023).
94. Yu, Y. P. *et al.* Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy. *J. Clin. Oncol.* **22**, 2790–2799 (2004).
95. Zhang, Y. *et al.* Promoting cell proliferation, cell cycle progression, and glycolysis: Glycometabolism-related genes act as prognostic signatures for prostate cancer. *Prostate* **81**, 157–169 (2021).

96. Peraldo-Neia, C. *et al.* Epidermal Growth Factor Receptor (EGFR) mutation analysis, gene expression profiling and EGFR protein expression in primary prostate cancer. *BMC Cancer* **11**, 31 (2011).
97. Longoni, N. *et al.* Aberrant expression of the neuronal-specific protein DCDC2 promotes malignant phenotypes and is associated with prostate cancer progression. *Oncogene* **32**, 2315–2324, 2324.e1–4 (2013).
98. True, L. *et al.* *GEO*, <https://identifiers.org/geo:GSE5132> (2006).
99. True, L. *et al.* A molecular correlate to the Gleason grading system for prostate adenocarcinoma. *Proc. Natl. Acad. Sci. USA* **103**, 10991–10996 (2006).
100. Jia, Z. *et al.* Diagnosis of prostate cancer using differentially expressed genes in stroma. *Cancer Res.* **71**, 2476–2487 (2011).

Acknowledgements

This work is supported by grants CA241647 to J.C.C., S.T., and B.F., CA231978 to J.C.C., the Finnish Cultural Foundation and the Finnish Cancer Institute as FICAN Cancer Researcher to T.D.L., in part by the Biostatistics and Bioinformatics Shared Resource at the H. Lee Moffitt Cancer Center & Research Institute, an NCI designated Comprehensive Cancer Center (P30CA076292), and in part by the Biostatistics and Bioinformatics Shared Resource at the University of Colorado Cancer Center, an NCI designated Comprehensive Cancer Center (P30CA046934). The authors would like to extend gratitude to the curated datasets' original authors, who provided irreplaceable advice and additional information for their studies.

Author contributions

T.D.L., V.S., A.C.S., J.H.C., A.S.H., F.C.F.C., K.S., C.C.L. developed and wrote the R package, documentation and constructed the exported data objects; T.D.L., V.S., J.H.C., F.C.F.C., C.C.L., T.G., S.T., J.C.C. designed the harmonized data processing pipeline; T.D.L., V.S., A.C.S., M.V.O., B.L.F., S.T., J.C.C. contributed R vignettes; T.D.L., V.S., A.C.S., J.H.C., F.C.F.C., K.S., T.G., B.L.F., S.T., J.C.C. contributed original analyses; T.D.L., V.S., A.C.S., M.V.O. visualized data and analyses; T.G., B.L.F., S.T., J.C.C. supervised the project and obtained funding; T.D.L., V.S., A.C.S., S.T., J.C.C. drafted the manuscript; All authors read and approved the final manuscript.

Competing interests

J.C.C. is co-founder of PrecisionProfile and OncoRX Insights. All other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41597-023-02335-4>.

Correspondence and requests for materials should be addressed to T.D.L., S.T. or J.C.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023