Multi-omics evaluation of cell lines as models for metastatic prostate cancer

Xueying Liu[1], Weixing Yu[2], Xiuyuan Jin[1], Yugang Wang[2,*], Ke Liu[1,*]

Affiliations:

1. Department of Medical Dataology, School of Public Health, Cheeloo College of Medicine, Shandong University, Jinan, China.

2. Department of Biochemistry and Molecular Biology, School of Basic Medicine, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China.

* To whom correspondence should be addressed

Email addresses:

Yugang Wang: yugangw@hust.edu.cn

Ke Liu: keliu.iluke@email.sdu.edu.cn

Abstract

Cell lines are indispensable tools in prostate cancer research yet their suitability as models for distance metastasis remains incompletely characterized. Here, we conduct a systematic evaluation study using large-scale public multi-omics data. We reveal substantial genomic differences between cell lines and metastatic patient samples, and meanwhile pinpoint cell lines which more closely resemble metastatic prostate cancer. Notably, hypermutation significantly influences the tumor microenvironment, underscoring the importance of considering mutational burden in model selection. Surprisingly, the widely used PC3 cell line exhibits poor transcriptomic similarity to any prostate cancer subtype, revealing a previously unrecognized limitation. Furthermore, we find existing engineered stem-like cell lines fail to faithfully recapitulate the transcriptomic profiles of mesenchymal stem-like prostate cancer, whereas selected organoids exhibit higher fidelity. Our study provides guidance for cell line selection in metastatic prostate cancer research and highlights the urgent need to develop improved cell lines for the mesenchymal stem-like subtype.

Introduction

Cancer cell lines are indispensable preclinical models in oncology research. Their immortality, cost-effectiveness, and capacity for long-term cultivation and cloning make them vital for reproducible and scalable cancer research and drug development[1, 2, 3, 4, 5]. However, prolonged *in vitro* culture can lead to the accumulation of additional genetic and epigenetic alterations, resulting in substantial discrepancies compared to patient-derived tumor samples[6, 7, 8, 9]. Therefore, systematic evaluation of the molecular fidelity of cancer cell lines relative to patient-derived cancer cells is essential to ensure the translational relevance of cell line–based studies.

The widespread availability of large-scale open cancer genomics data has enabled researchers to comprehensively assess the suitability of cell lines as models of cancer research. For example, Domcke et al. compared genomic profiles between ovarian cancer cell lines and patient tumor samples and found several commonly used cell lines poorly recapitulated high-grade serous ovarian cancer[10]; Yu et al. performed a pan-cancer transcriptomic comparison between cell lines and patient tumor samples and proposed a new cell line panel TCGA-110-CL as a replacement of the canonical NCI-60 panel[11]; Jin et al. conducted similar pan-cancer analysis with their own computational approaches to select appropriate cell line models for different cancer types[12]; We previously performed multi-omics comparison between breast cancer cell lines and metastatic tumor samples and identified cell lines that faithfully represent distinct breast cancer subtypes[13]. These studies have deepened our understanding of cancer cell lines and highlighted the importance of selecting appropriate models based on molecular fidelity.

Prostate cancer is among the most common and lethal malignancies affecting men's health worldwide, posing a substantial public health challenge[14, 15]. Similar to other cancer types, cell lines have also been widely utilized to explore the underlying

biological mechanisms of this disease. Although several studies have evaluated the molecular representativeness of prostate cancer cell lines relative to tumor samples[11, 12, 16], three major limitations persist. First, metastasis is the leading cause of prostate cancer–related mortality[14, 17]; however, most of the evaluation studies mainly focus on primary tumors. Second, existed studies were performed in a pan-cancer manner, which applies uniform analytical criteria across various cancer types and overlooks the influence of prostate cancer specific characteristics in cell line selection[3]. Third, the majority of analyses rely exclusively on transcriptomic data, with limited integration of multi-omics information for a more comprehensive evaluation.

In this study, we leveraged the publicly available bulk and single-cell omics data to comprehensively evaluate the suitability of prostate cancer cell lines as models for metastatic prostate cancer. We proposed principles of cell line selection for different research scenarios and meanwhile point the strength and limitation of widely used cell line. Our study provides a framework for selecting prostate cancer cell lines tailored to specific biological contexts and highlight the need to develop more representative models for understudied prostate cancer subtypes.

Results

Systematic comparison of genomic profiles

We first compared the somatic mutation profiles between metastatic prostate cancer samples and prostate cancer cell lines, focusing on genes that were either highly mutated in metastatic prostate cancer or exhibited differential mutation rates between metastatic and primary prostate cancer. The somatic mutation data of primary prostate cancer, metastatic prostate cancer, and prostate cancer cell lines were downloaded from the Cancer Genome Atlas (TCGA), Stand Up to Cancer (SU2C), and the Cancer Cell Line Encyclopedia (CCLE), respectively[18, 19, 20]. A total of 68 highly mutated genes were identified using SU2C samples, along with 13 differentially mutated genes

between SU2C and TCGA samples (**Supplementary. Fig. 1a-c**). The integration of these two lists identified 73 unique genes (**Fig. 1a**). The median mutation rate of the 73 genes across the CCLE prostate cancer cell lines was 0.23, suggesting most of them could be recapitulated by only a few cell lines. *MUC16*, *SYNE2*, *SYNE1*, *OBSCN* have a mutation rate higher than 50%, with *MUC16* being the most frequently mutated gene (mutation rate = 0.62). *TP53*, a key tumor suppressor gene, displayed a mutation rate of 0.38. Surprisingly, there were four genes (*APC*, *SPOP*, *CDK12*, and *FAM186A*) which were entirely unmutated in all analyzed prostate cell lines, highlighting the limitation of using the analyzed cell lines in researching the role of them in metastasis-related study.

Cell lines were ranked based on the number of mutations present in a set of 73 genes associated with metastatic prostate cancer, and found the top four cell lines were LNCaP, DU145, 22RV1, and MDA-PCa-2b, respectively **(Fig 1b)**. Interestingly, we noticed these four cell lines were all inferred with microsatellite instability (MSI)[21] and had significantly higher mutation burden than other cell lines **(Fig. 1c)**. To adjust such bias, we re-ranked the cell lines according to a normalized ratio value (computed as number of mutated genes / mutation burden) and found the top three cell lines were PC3, VCaP, and NCI-H660 (**Fig. 1d**).

We next investigated whether those recurrent single-nucleotide alterations (which we call "mutation hotspot") identified in metastatic prostate cancer samples were also presented in prostate cancer cell lines. We defined mutation hotspots as non-synonymous mutations occurring in at least three SU2C samples. Based on this criterion, 58 mutation hotspots were identified (**Supplementary Table 1**). These mutation hotspots were distributed in 45 genes, with *TP53* and *AR* being the two genes harboring most mutation hotspots (**Fig. 1e and Supplementary. Fig. 1d)**. *TP53* mutation hotspots were predominantly located within the DNA-binding domain, a functionally critical

region required for the transcriptional regulation of tumor suppressor genes[22]. For *AR*, mutation hotspots were primarily clustered within the ligand-binding domain, which is pivotal for androgen receptor-targeted therapies[23].

Strikingly, only seven mutation hotspots were detected in at least one prostate cancer cell line (**Fig. 1f**), and all were presented exclusively in 22RV1, MDA-PCa-2b, and LNCaP. Of them, three were single nucleotide point mutations of *AR* and the other four were all frameshift mutations. Of the *AR* mutation hotspots, T878A and L702H mutations have been reported to be associated with resistance to AR inhibition therapy[24]. Notably, MDA-PCa-2b harbors both mutations (T878A and L702H), reflecting the clinical phenomenon of co-occurring multiple AR resistance mutations within individual tumors[24]. Among the frameshift mutation hotspots, K16fs mutation in *RPL22* was observed in all three cell lines, whereas the other three frameshift mutations were exclusive to 22RV1.

Besides somatic mutation profiles, we also compared copy number variation (CNV) profiles between metastatic and primary prostate cancer samples. Among the differential CNV genes between SU2C and TCGA samples, *AR* exhibited the highest level of amplification (**Fig. 1g**). A similar pattern was observed in site-specific analyses of differential CNV genes (**Supplementary. Fig. 1e-g**). This finding may not be surprising since *AR* is a critical driver of prostate cancer development and progression[25, 26, 27]. Notably, VCaP exhibited the highest level of *AR* amplification among all analyzed prostate cancer cell lines (**Fig. 1h**).

Hypermutated prostate cancer has increased cytotoxic CD8$^+$ T cell infiltration

Previous studies have demonstrated hypermutated colorectal tumor samples exhibiting increased level of tumor-infiltrating lymphocytes, elevated neoantigen burden, and potentially enhanced sensitivity to immunotherapy[28, 29]. Motivated by these findings,

we aim to assess whether hypermutation should be taken into consideration when selecting cell lines for studying the metastatic mechanisms of prostate cancer.

To identify hypermutated prostate cancer samples, we integrated samples from TCGA, SU2C, CCLE together and determined a mutation burden threshold above which samples were classified as hypermutated (**Fig. 2a**). Based on our criteria, we identified four hypermutated samples in TCGA, ten in SU2C, and four in CCLE. As expected, the aforementioned four hypermutated cell lines LNCaP, DU145, 22RV1, and MDA-PCa-2b were successfully identified. Notably, three of them harbor *RPL22* frameshift mutations, which is frequently observed in MSI tumors[30, 31]. To further validate our strategy, we computed the mutation frequencies of mismatch repair (MMR) genes[32] and found hypermutated samples exhibiting a significantly higher mutation frequency (**Fig. 2b and Supplementary Fig. 2a**). These results demonstrated the validity of the determined threshold value for identifying hypermutated prostate cancer samples.

We next performed differential gene expression analysis between hypermutated and non-hypermutated TCGA samples and identified 125 upregulated and 85 downregulated genes (**Supplementary Table 2**). Notably, genes encoding T cell markers (*CD3D* and *CD3E*), cytotoxic $CD8^+$ T cell markers (*CD8A*, *GZMA*, *GZMB*, *PRF1*), and immune checkpoint molecules (*PDCD1* and *CTLA4*) were consistently upregulated in the hypermutated group (**Fig. 2c**), suggesting increased cytotoxic $CD8^+$ T cell infiltration in the tumor microenvironment (TME) of hypermutated prostate cancer. To further validate this observation, we applied the Wilcoxon rank-sum test to these genes and obtained consistent results (**Supplementary Fig. 2b**). In addition, we employed two independent tools, TIDE[33] and xCell[34], to estimate cytotoxic $CD8^+$ T cell infiltration levels, both of which showed higher average infiltration in the hypermutated group, although the statistical significance was marginal—likely due to the limited sample size of hypermutated cases (**Supplementary Fig. 2c**). Finally, enrichment

analysis of the upregulated genes using the MSigDB Hallmark gene sets revealed six significantly enriched pathways, including "Interferon Gamma Response" (**Fig. 2d and Supplementary Table 3**), further supporting our findings.

The TME has been demonstrated to play pivotal roles in the distant metastasis of primary cancer[35, 36, 37] and our analysis suggests notable TME differences between hypermutated and non-hypermutated primary prostate cancer samples. Therefore, we propose that hypermutated cell lines (such as LNCaP) should be more appropriate for investigating the metastatic mechanisms of hypermutated prostate cancer. To the best of our knowledge, this consideration has not been addressed in previous pan-cancer cell line evaluation studies, highlighting the need for cancer-specific assessments.

Correlating cell lines with metastatic prostate cancer using bulk and single-cell RNA-seq data

Transcriptomic correlation analysis (TC analysis) has been proven to be an effective method for assessing the utility of cell lines[11, 13, 38]. Consequently, we ranked all 1,019 CCLE cell lines based on their transcriptomic similarity to MET500 prostate cancer samples and the top three cell lines were LNCaP, VCaP, and MDA-PCa-2b, respectively (**Fig. 3a and Supplementary Table 4**). We next examined whether the metastatic site influenced TC analysis results and found that cell line rankings were highly consistent among various metastatic sites (bone, liver, and lymph node) (**Fig. 3b and Supplementary Fig. 3a-c**).

To double confirm the analysis results derived from bulk RNA-seq data, we further ranked 1,019 CCLE cell lines according to their transcriptomic similarity with the malignant cells from a scRNA-seq dataset of metastatic prostate cancer[25]. Consistent with previous results, the top three cell lines with the highest similarity were VCaP, LNCaP, and MDA-PCa-2b (**Fig. 3c**). We further performed metastatic-site-specific

transcriptomic similarity analysis (bone and lymph node) and the results were highly correlated (**Fig. 3d and Supplementary Fig. 3d-e**).

We conducted a PubMed search of studies related to metastatic prostate cancer and found that the three most frequently cited cell lines (PC3, LNCaP, and DU145) accounted for over 90% of all citations (**Supplementary Fig. 3f and Supplementary Table 5**). Surprisingly, VCaP, which demonstrates both high genomic and transcriptomic similarity to metastatic prostate cancer samples in our analysis, was cited in only 1.4% of studies. This discrepancy highlights a notable mismatch between the actual suitability of cell lines and their prevalence in the literature, and underscores the importance of cell line evaluation for improving translational relevance in prostate cancer research.

Limitation of the PC3 cell line

According to PubMed search, PC3 is the most widely used cell line in studies investigating metastatic prostate cancer[39, 40]. However, our analysis suggests that this cell line shows limited similarity to metastatic prostate cancer samples (ranked 576 among 1,019 CCLE cell lines) (**Fig. 3a**). Prior to our research, several studies had revealed the heterogeneity of prostate cancer[41, 42, 43]. Considering that the metastatic prostate cancer samples used in the transcriptomic correlation analysis mainly exhibited adenocarcinoma histology (in which malignant cells primarily show luminal differentiation)[44, 45, 46], we further investigated whether PC3 could serve as a representative model for other subtypes of prostate cancer.

We performed principal component analysis (PCA) on gene expression profiles of CCLE prostate cancer cell lines and found they could be briefly classified into three clusters (**Fig. 4a**). Cluster 1 consisted of luminal lineage cell lines (such as VCaP[47]), as indicated by expression of luminal lineage markers such as *AR* and *KLK4*; cluster 2

only included NCI-H660[48], a small cell carcinoma cell line of neuroendocrine origin expressing *SYP*; cluster 3 comprised DU145 and PC3. Interestingly, our PCA results align with Han et al. 's research, which classified prostate cancer into three intrinsic subtypes: androgen receptor positive prostate cancer (ARPC), neuroendocrine prostate cancer (NEPC), and mesenchymal and stem-like prostate cancer (MSPC)[49]. Since PC3 and DU145 were separated with other cell lines on the first principal component (PC1), we performed enrichment analysis on the top 200 genes with the highest positive loadings along the PC1 axis using MSigDB Hallmark gene sets (**Fig. 4b and Supplementary Table 6**). The "Epithelial-Mesenchymal Transition" gene set exhibited the highest level of enrichment; in addition, the enrichment of "*IL-2/STAT5* Signaling" and "*IL-6/JAK/STAT3* Signaling" gene sets suggests activation of the *JAK/STAT* pathway, a known driver of cellular plasticity[46, 50]. Furthermore, PC3 exhibited relatively higher expression of *CD44* (a marker of stemness), along with lower expression of the luminal marker *AR* and the neuroendocrine marker *SYP* (**Fig. 4c-e**). We double confirmed the expression of these marker genes using a scRNA-seq dataset of PC3 (**Supplementary Fig. 4a**). Collectively, these results support the classification of PC3 as a cell line of the MSPC subtype.

We next evaluated whether PC3 could adequately recapitulate the transcriptomic profile of MSPC malignant cells. MSPC has been reported to be enriched in castration-resistant prostate cancer (CRPC) patients[49]. To this end, we analyzed a CRPC scRNA-seq dataset and found the malignant cells could be classified into ARPC, NEPC, and MSPC subtypes based on the expression of *AR*, *SYP*, and *CD44*, respectively (**Fig. 4f)**. We then ranked 1,019 CCLE cell lines according to their transcriptomic similarity with malignant cells in a subtype-specific manner. As expected, MDA-PCa-2b, VCaP, and LNCaP showed the highest similarity to ARPC, while NCI-H660 (a neuroendocrine prostate cancer cell line) exhibited a high similarity to NEPC (**Supplementary Fig. 4b-c**). Strikingly, for the MSPC subtype, PC3 only got a rank value of 775 although its

rank was the highest among prostate cancer cell lines (**Fig. 4g and Supplementary Table 7**). Moreover, the transcriptomic correlation between individual MSPC malignant cells and PC3 was significantly lower than that between ARPC malignant cells and VCaP, suggesting limited fidelity of PC3 as a model of MSPC (**Supplementary Fig. 4d**). Interestingly, we found MSPC malignant cells expressed basal lineage markers such as *TP63* and *CAV2* (**Fig. 4h and Supplementary Fig. 4e**), suggesting their basal differentiation. Consistent with this, lineage analysis suggested their transcriptomic profile was most similar to basal cell (**Fig. 4i and Supplementary. Fig. 4f**). However, PC3 lacked expression of key basal markers (such as *TP63*), which may explain its low transcriptomic similarity with MSPC malignant cells **(Supplementary Fig. 4g).**

Taken together, our findings highlight the limitation of PC3 cell line. Although it most closely resembles the MSPC subtype among CCLE prostate cancer cell lines, the overall transcriptomic similarity to primary MSPC malignant cells remains modest.

Patient-derived organoids better resemble the transcriptome of MSPC than PC3

Since PC3 does not fully recapitulate the transcriptomic features of MSPC, we next investigated other models that could be more representative. Previous studies have reported that AR inhibition by enzalutamide or *TP53/RB1* double-knockout can promote LNCaP (a hypermutated cell line of ARPC) into MSPC state[49, 50]. Therefore, we included these two engineered LNCaP-derived models in our analysis. In addition, we generated a VCaP-derived cell line by overexpressing *MET* (VCaP-MET-OE), as *MET* has been reported to promote EMT and sustain stemness in cancer cells[51, 52, 53]. As expected, bulk RNA-seq analysis suggested that *MET* overexpression induced upregulation of *CD44* ($\log_2$FC = 0.251, *P*-value = 0.028) and increased *JAK/STAT* pathway activity, while downregulating luminal markers *KLK2*, *KLK3*, and *KLK4*, suggesting the occurrence of de-differentiation (**Fig. 5a-b and Supplementary Table**

8).

We next computed the transcriptomic similarity with MSPC malignant cells for the above three engineered cell lines, respectively. Surprisingly, all of the engineered cells had significantly lower transcriptomic similarity with MSPC malignant cells than PC3 (**Fig. 5c**). In addition, lineage analysis revealed that these engineered cell lines retained their original luminal differentiation (**Fig. 5d**). Furthermore, when comparing the transcriptomic profiles of these engineered cell lines to the 1,019 CCLE cell lines, the highest correlations were consistently observed with their respective parental lines (**Supplementary Fig. 5a–g)**, suggesting the engineering did not trigger a luminal-basal lineage transition. MSPC malignant cells exhibit basal differentiation and this likely explains why the engineered cell lines failed to recapitulate the MSPC transcriptional landscape.

In recent years, patient-derived organoids (PDOs) have gained increasing attention for studying cancer biology. We next explored whether this type of model could better recapitulate the MSPC transcriptome. We computed the transcriptomic similarity between ten established PDOs (all of which were claimed as MSPC according to Tang et al.'s study)[54] and MSPC malignant cells. Notably, four PDOs had significantly higher transcriptomic similarity to MSPC malignant cells than PC3, with MSKPCa12 being the highest one (**Fig. 5e**). Further lineage analysis suggested that this organoid exhibited a basal differentiation phenotype (**Fig. 5f**). As expected, it displayed higher expression of basal-specific markers (such as *TP63* and *KRT5*) than PC3 **(Fig. 5g-h)**, which may explain its stronger transcriptomic resemblance of MSPC malignant cells.

Discussion

Cell lines are widely utilized in cancer research, particularly as models for metastatic cancers. Prior studies have demonstrated that cell lines derived from the same tumor

type exhibit considerable molecular heterogeneity, with certain cell lines more accurately recapitulating the characteristics of metastatic tumors than others[10, 38, 55, 56]. This underscores the critical importance of selecting cell line models that most faithfully represent the biology of metastatic cancers. Our study proposed useful guidance for cell line selection in the study of metastatic prostate cancer.

Adenocarcinoma is the predominant histological subtype of prostate cancer[57, 58]. Based on our multi-omics analysis results, VCaP appears to be the most suitable model for studying metastatic prostate cancer according to its relatively higher concordance with metastatic prostate adenocarcinoma patient samples in multiple aspects (for example, the high transcriptomic similarity with patient samples, and the highest level of *AR* amplification). However, according to PubMed search results, VCaP is rarely used in metastasis-related research, suggesting a gap between model suitability and actual utilization in the field.

Although VCaP represents a suitable model for metastatic prostate adenocarcinoma, it may not be appropriate for all contexts. One particular situation is hypermutated patient samples. Our analysis suggests that the microenvironment in hypermutated prostate cancer samples differs markedly from that in non-hypermutated tumors, particularly with respect to increased immune infiltration. This indicates that the hypermutation phenomenon should be considered when selecting cell line models. Given its high mutation burden and strong transcriptomic similarity to metastatic prostate adenocarcinoma samples, LNCaP was identified as a promising model for studying hypermutated prostate cancer.

Another situation is to study AR resistance. We found that the prostate cancer cell lines LNCaP and MDA-PCa-2b harbor *AR* mutation hotspots associated with clinical resistance, showing a certain degree of concordance with patient observations.

Specifically, MDA-PCa-2b carries both the T878A and L702H mutations, which are linked to resistance to flutamide and glucocorticoids, respectively. The T878A mutation was also detected in LNCaP, while 22RV1 harbors another recurrent *AR* mutation hotspot, H875Y. Notably, despite harboring these resistance-associated mutations, all three cell lines remain AR-dependent, suggesting that they are valuable *in vitro* models for studying AR-driven resistance mechanisms and evaluating novel antiandrogen therapies.

According to our PubMed search results, PC3 is the most widely used prostate cancer cell line; however, it exhibits low transcriptomic similarity to metastatic prostate adenocarcinoma. Although further analysis classified PC3 within the MSPC subtype, it does not faithfully model MSPC at the transcriptomic level, either. According to our lineage analysis, MSPC malignant cells exhibited strong basal differentiation, as indicated by the expression of basal lineage markers (such as *TP63*). Although PC3 also exhibits basal differentiation, several classical basal lineage markers were not expressed (data not shown). Therefore, we conclude that PC3 is in a "semi-basal differentiation" state and this may explain its transcriptomic discrepancy with MSPC malignant cells.

Besides CCLE cell lines, we also examined the suitability of three engineered cell lines (LNCaP-enza, LNCaP-TP53/RB1-KO, VCaP-MET-OE) and PDOs as models of MSPC. We found that the engineering applied on cell lines could only induce stem-like features while not triggering a full luminal-basal lineage transition. Not surprisingly, these engineered cell lines did not exhibit higher performance in resembling the transcriptome of MSPC than PC3. Notably, although several established prostate cancer organoids (such as MSKPCa12) appear to be better models of MSPC than PC3, cell lines are superior in their cost-effectiveness. Therefore, it is still valuable to further investigate how to establish a cell line that can adequately resemble the transcriptome of MSPC.

In summary, our comprehensive analysis identified suitable prostate cancer cell lines for different scenarios of metastatic prostate cancer research and meanwhile highlight the limitations of canonical cell lines. We anticipate that the adoption of our recommended cell lines will enhance the translational relevance of *in vitro* studies of metastatic prostate cancer and accelerate the development of clinically meaningful therapeutic strategies.

Materials and Methods

Selection of genes used for genomic profile comparison

The somatic mutation data of metastatic and primary prostate cancer were from SU2C and TCGA, respectively. In our analysis, non-adenocarcinoma samples were excluded. Given a gene, we performed Fisher's exact to evaluate whether its mutation frequency is significantly different between primary and metastatic prostate cancer samples. Since SU2C contains metastatic samples from three sites (liver, bone, and lymph node), the analysis was performed in site-specific manner. Only genes which are significant (adjusted P-value < 0.05) in at least two sites were retained.

Genes whose mutation frequency exceeded 5% in at least two metastatic sites (in the SU2C cohort) were defined as highly mutated. The non-synonymous somatic mutations occurring in at least three SU2C samples were defined as mutation hotspots.

Comparison of CNV profiles

We downloaded CNV segment data and then utilized CNTools[59] to generate gene-level copy number. For each gene, the Wilcoxon rank-sum test was applied to compare its CNV profile between SU2C and TCGA samples.

Transcriptomic correlation analysis with bulk RNA-seq and scRNA-seq data

In our previous research, we identified 1,000 genes which are highly varied across CCLE cell lines and proved their utility in transcriptomic correlation analysis[13]. Given a cell line and a patient sample (or single cell), we defined their transcriptomic similarity as the Spearman rank correlation across the 1,000 genes. The transcriptomic similarity between a cell line and several patient samples (or single cells) was computed as the median correlation value.

Differential gene expression analysis

DESeq2[60] was used to identify differentially expressed genes (adjusted P-value < 0.05 and abs($\log_2$FC) > 1) and the enrichment analysis was conducted using Enrichr[61]. Terms with a P-value < 0.05 were considered statistically significant.

PubMed search

Our PubMed search was performed on March 11th, 2025. For each prostate cancer cell line, we searched the database using the keyword "[cell line name] + metastasis" and counted the number of returned citations.

Analysis of scRNA-seq data

In the analysis of CRPC scRNA-seq dataset, we used inferCNV[62, 63] to infer copy number variation of epithelial cells (immune cells as reference). CNV values were extracted and subjected to k-means clustering to identify malignant cells. The marker genes used in our analysis were: *PTPRC* (immune cells) and *EPCAM* (epithelial cells).

Lineage analysis

We analyzed a normal prostate scRNA-seq dataset and identified three epithelial cell types: luminal cell (*KLK2*, *KLK4*, *ACPP*), basal cell (*KRT5*, *TP63*, *KRT17*, *KRT15*), and neuroendocrine cell (*ASCL1*, *FOXA2*, *MYCN*, *POU3F2*, *SIAH2*, *NCAM1*, *CHGA*, *CHGB*, *SYP*, *ENO2*)[64, 65]. We generated a pseudo-bulk profile for each epithelial cell

type and then identified the top 5,000 genes which are highly variable across the three cell types. Given a single malignant cell, we used these 5,000 genes to calculate its transcriptomic similarity with each of the three cell types and then assigned it to the lineage with the highest similarity value.

Software tools and statistical methods

All of the analysis was conducted in R. The ggplot2[66], maftools[67] and ComplexHeatmap[68] packages were used for data visualization and the Seurat[69] package was used for processing scRNA-seq data. If not specified, the two-sided Wilcoxon rank-sum test was used to compute *P*-value and Benjamini-Hochberg procedure was applied to compute adjusted *P*-value.

Datasets

The genomic and transcriptomic data of CCLE cell lines were obtained from the DepMap data portal (https://depmap.org/portal/). The somatic mutation and copy number variation data of TCGA and SU2C samples were downloaded from cBioPortal (https://www.cbioportal.org/)[70, 71]. The bulk RNA-seq data of TCGA samples was downloaded from UCSC Xena data portal (https://xena.ucsc.edu/)[72]. The bulk RNA-seq data of metastatic prostate adenocarcinoma patients were from MET500 cohort[73]. The prostate adenocarcinoma scRNA-seq data was downloaded from Single Cell Portal (https://singlecell.broadinstitute.org/single_cell/study/SCP1244). The scRNA-seq data of normal human prostate, cell lines and organoids were downloaded from GEO (accession number: GSE157220, GSE117403, GSE162225, GSE175975 and GSE199190)[49, 50, 54, 65, 74]. The scRNA-seq data of CRPC was downloaded from SRA (accession number: PRJNA699369)[75].

Overexpression of *MET* in VCaP cell line

VCaP cells and HEK293T cells were obtained from Pricella Biotechnology Co., Ltd and American Type Culture Collection. VCaP and HEK293T cells maintained in

DMEM (catalog no. SH30285.01, Cytiva Life Sciences). All medium supplemented with 10% FBS (catalog no. 164210-50, Pricella Biotechnology). HEK293T cells (2 × 106) were seeded in a 100-mm cell culture dish and transfected with plasmids using PolyJet *in vitro* DNA transfection reagent PEI (40815ES03, Yeasen Biotechnology). The transfection experiments were operated according to the manufacturer's instructions. For overexpression of MET proteins, we co-transfected the HEK293T cells with the plasmids of pMD2.G, psPAX2 and pHAGE-Flag-MET to generate lentivirus. For lentivirus infection, 500 μl of virus-containing medium was added to the 5 × 104 VCaP cells with 1% polybrene.

Bulk RNA-seq library construction and sequencing

Total RNA was extracted from cells using the RNAsimple Total RNA Kit (no. DP419, TIANGEN). Illumina compatible libraries were prepared using the NEBNext Ultra RNA Library Prep Kit for Illumina (catalog no. E7530L, NEB). In brief, using fragmented mRNA as a template and random oligonucleotides as primers, the first strand of cDNA was synthesized in the M-MuLV reverse transcriptase system. Second strand cDNA synthesis was subsequently performed using DNA polymerase I and RNase H. The purified double-stranded cDNA undergoes end-repair, A-tailing and ligation of sequencing adapters. The 250–300-base pair cDNA is screened with AMPure XP beads, PCR amplification is performed and the PCR products are purified again with AMPure XP beads to obtain the library. After the library is constructed, use Qubit2.0 Fluorometer for preliminary quantification, dilute the library to 1.5 ng μl−1 and then use Agilent 5400 system to assess the quality of the library. After the insert size meets the expectation, real-time RT-PCR measures the effective concentration of the library. The qualified libraries were pooled and sequenced on Illumina platform with PE150 strategy in Novogene Bioinformatics Technology, according to effective library concentration and data amount required.

Code Availability

The code is available at github (https://github.com/bioklab/MetPCCellline).

Competing Interests

The researchers declare no competing interests.

Acknowledgements

The research is supported by National Natural Science Foundation of China (Fund 32370715), Science Fund for Distinguished Oversea Young Scholars of Shandong Province (2023HWYQ-015), Taishan Young Scholar Program of Shandong Province (tsqn202312020), and Cheeloo Young Scholar Program of Shandong University.

Authors' contribution

X.Y.L. and K.L. conceived the study. X.Y.L. performed the majority of computational analysis, Y.G.W. and W.X.Y. performed all the experimental assays, K.L. and Y.G.W. supervised the study. All authors contributed to writing, reviewing, and editing the manuscript and approved the manuscript.

Reference

1. Kaur G, Dufour JM. Cell lines: Valuable tools or useless artifacts. *Spermatogenesis* **2**, 1-5 (2012).

2. Goodspeed A, Heiser LM, Gray JW, Costello JC. Tumor-Derived Cell Lines as Molecular Models of Cancer Pharmacogenomics. *Molecular cancer research : MCR* **14**, 3-13 (2016).

3. Bose B, Bozdag S. Identifying cell lines across pan-cancer to be used in preclinical research as a proxy for patient tumor samples. *Commun Biol* **7**, 1101 (2024).

4. Dutil J, Chen Z, Monteiro AN, Teer JK, Eschrich SA. An Interactive Resource to Probe Genetic Diversity and Estimated Ancestry in Cancer Cell Lines. *Cancer research* **79**, 1263-1273 (2019).

5. Mouradov D*, et al.* Colorectal cancer cell lines are representative models of the main molecular subtypes of primary cancer. *Cancer research* **74**, 3238-3247 (2014).

6. Gillet JP, Varma S, Gottesman MM. The clinical relevance of cancer cell lines. *Journal of the National Cancer Institute* **105**, 452-458 (2013).

7. Weinstein JN. Drug discovery: Cell lines battle cancer. *Nature* **483**, 544-545 (2012).

8. Ertel A, Verghese A, Byers SW, Ochs M, Tozeren A. Pathway-specific differences between tumor cell lines and normal and tumor tissue cells. *Molecular cancer* **5**, 55 (2006).

9. Salvadores M, Fuster-Tormo F, Supek F. Matching cell lines with cancer type and subtype of origin via mutational, epigenomic, and transcriptomic patterns. *Science advances* **6**, (2020).

10. Domcke S, Sinha R, Levine DA, Sander C, Schultz N. Evaluating cell lines as tumour models by comparison of genomic profiles. *Nat Commun* **4**, 2126 (2013).

11. Yu K*, et al.* Comprehensive transcriptomic analysis of cell lines as models of primary tumors across 22 tumor types. *Nat Commun* **10**, 3574 (2019).

12. Jin H*, et al.* Systematic transcriptional analysis of human cell lines for gene expression landscape and tumor representation. *Nat Commun* **14**, 5417 (2023).

13. Liu K*, et al.* Evaluating cell lines as models for metastatic breast cancer through integrative analysis of genomic data. *Nat Commun* **10**, 2138 (2019).

14. Xin S*, et al.* ScRNA-seq revealed an immunosuppression state and tumor microenvironment heterogeneity related to lymph node metastasis in prostate cancer. *Exp Hematol Oncol* **12**, 49 (2023).

15. Rebello RJ*, et al.* Prostate cancer. *Nature reviews Disease primers* **7**, 9 (2021).

16. Peng D*, et al.* Evaluating the transcriptional fidelity of cancer models. *Genome medicine* **13**, 73 (2021).

17. Hao M*, et al.* HIC1 loss promotes prostate cancer metastasis by triggering epithelial-mesenchymal transition. *The Journal of pathology* **242**, 409-420 (2017).

18. Weinstein JN, *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nature genetics* **45**, 1113-1120 (2013).

19. Abida W, *et al.* Genomic correlates of clinical outcome in advanced prostate cancer. *Proc Natl Acad Sci U S A* **116**, 11428-11436 (2019).

20. Barretina J, *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603-607 (2012).

21. Ghandi M, *et al.* Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503-508 (2019).

22. Hernández Borrero LJ, El-Deiry WS. Tumor suppressor p53: Biology, signaling pathways, and therapeutic targeting. *Biochimica et biophysica acta Reviews on cancer* **1876**, 188556 (2021).

23. Hunter I, Jamieson C, McEwan IJ. The androgen receptor amino-terminal domain: structure, function and therapeutic potential. *Endocrine oncology (Bristol, England)* **5**, e240061 (2025).

24. Robinson D, *et al.* Integrative clinical genomics of advanced prostate cancer. *Cell* **161**, 1215-1228 (2015).

25. He MX, *et al.* Transcriptional mediators of treatment resistance in lethal prostate cancer. *Nat Med* **27**, 426-433 (2021).

26. Augello MA, Den RB, Knudsen KE. AR function in promoting metastatic prostate cancer. *Cancer metastasis reviews* **33**, 399-411 (2014).

27. Heinlein CA, Chang C. Androgen receptor in prostate cancer. *Endocrine reviews* **25**, 276-308 (2004).

28. Lin A, Zhang J, Luo P. Crosstalk Between the MSI Status and Tumor Microenvironment in Colorectal Cancer. *Frontiers in immunology* **11**, 2039 (2020).

29. Seo MK, Kang H, Kim S. Tumor microenvironment-aware, single-transcriptome prediction of microsatellite instability in colorectal cancer using meta-analysis. *Scientific reports* **12**, 6283 (2022).

30. Ferreira AM*, et al.* High frequency of RPL22 mutations in microsatellite-unstable colorectal and endometrial tumors. *Human mutation* **35**, 1442-1445 (2014).

31. Weinstein HNW*, et al.* RPL22 is a tumor suppressor in MSI-high cancers and a splicing regulator of MDM4. *Cell Rep* **43**, 114622 (2024).

32. Cortes-Ciriano I, Lee S, Park WY, Kim TM, Park PJ. A molecular portrait of microsatellite instability across multiple cancers. *Nat Commun* **8**, 15180 (2017).

33. Jiang P*, et al.* Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. *Nat Med* **24**, 1550-1558 (2018).

34. Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol* **18**, 220 (2017).

35. Lambert AW, Zhang Y, Weinberg RA. Cell-intrinsic and microenvironmental determinants of metastatic colonization. *Nature cell biology* **26**, 687-697 (2024).

36. Biray Avci C, Goker Bagca B, Nikanfar M, Takanlou LS, Takanlou MS, Nourazarian A. Tumor microenvironment and cancer metastasis: molecular mechanisms and therapeutic implications. *Frontiers in pharmacology* **15**, 1442888 (2024).

37. Joyce JA, Pollard JW. Microenvironmental regulation of metastasis. *Nature reviews Cancer* **9**, 239-252 (2009).

38. Chen B, Sirota M, Fan-Minogue H, Hadley D, Butte AJ. Relating hepatocellular carcinoma tumor samples and cell lines using gene expression data in translational research. *BMC medical genomics* **8 Suppl 2**, S5 (2015).

39. Moya L*, et al.* Characterisation of cell lines derived from prostate cancer patients with localised disease. *Prostate Cancer Prostatic Dis* **26**, 614-624 (2023).

40. Molter CW, Muszynski EF, Tao Y, Trivedi T, Clouvel A, Ehrlicher AJ. Prostate cancer cells of increasing metastatic potential exhibit diverse contractile forces, cell stiffness, and motility in a microenvironment stiffness-dependent manner.

Frontiers in cell and developmental biology **10**, 932510 (2022).

41. Erickson A, *et al.* A Systematic Review of Prostate Cancer Heterogeneity: Understanding the Clonal Ancestry of Multifocal Disease. *European urology oncology* **4**, 358-369 (2021).

42. Løvf M, *et al.* Multifocal Primary Prostate Cancer Exhibits High Degree of Genomic Heterogeneity. *European urology* **75**, 498-505 (2019).

43. Song H, *et al.* Single-cell analysis of human primary prostate cancer reveals the heterogeneity of tumor-associated epithelial cell states. *Nat Commun* **13**, 141 (2022).

44. Guo W, *et al.* Single-cell transcriptomics identifies a distinct luminal progenitor cell type in distal prostate invagination tips. *Nature genetics* **52**, 908-918 (2020).

45. Zhang D, Zhao S, Li X, Kirk JS, Tang DG. Prostate Luminal Progenitor Cells in Development and Cancer. *Trends in cancer* **4**, 769-783 (2018).

46. Chan JM, *et al.* Lineage plasticity in prostate cancer depends on JAK/STAT inflammatory signaling. *Science* **377**, 1180-1191 (2022).

47. Korenchuk S, *et al.* VCaP, a cell-based model system of human prostate cancer. *In vivo (Athens, Greece)* **15**, 163-168 (2001).

48. Carney DN, *et al.* Establishment and identification of small cell lung cancer cell

lines having classic and variant features. *Cancer research* **45**, 2913-2923 (1985).

49. Han H, *et al.* Mesenchymal and stem-like prostate cancer linked to therapy-induced lineage plasticity and metastasis. *Cell Rep* **39**, 110595 (2022).

50. Deng S, *et al.* Ectopic JAK-STAT activation enables the transition to a stem-like and multilineage state conferring AR-targeted therapy resistance. *Nat Cancer* **3**, 1071-1087 (2022).

51. Jeon HM, Lee J. MET: roles in epithelial-mesenchymal transition and cancer stemness. *Annals of translational medicine* **5**, 5 (2017).

52. van Leenders GJ, *et al.* Activation of c-MET induces a stem-like phenotype in human prostate cancer. *PloS one* **6**, e26753 (2011).

53. Dong J, Li J, Li Y, Ma Z, Yu Y, Wang CY. Transcriptional super-enhancers control cancer stemness and metastasis genes in squamous cell carcinoma. *Nat Commun* **12**, 3974 (2021).

54. Tang F, *et al.* Chromatin profiles classify castration-resistant prostate cancers suggesting therapeutic targets. *Science* **376**, eabe1505 (2022).

55. Virtanen C, *et al.* Integrated classification of lung tumors and cell lines by expression profiling. *Proc Natl Acad Sci U S A* **99**, 12357-12362 (2002).

56. Sinha R, *et al.* Analysis of renal cancer cell lines from two major resources

enables genomics-guided cell line selection. *Nat Commun* **8**, 15165 (2017).

57. Xing P, *et al.* Treatment strategies and drug resistance mechanisms in adenocarcinoma of different organs. *Drug resistance updates : reviews and commentaries in antimicrobial and anticancer chemotherapy* **71**, 101002 (2023).

58. Raychaudhuri R, Lin DW, Montgomery RB. Prostate Cancer: A Review. *Jama* **333**, 1433-1446 (2025).

59. Jianhua Zhang. The CNTools package.https://bioconductor.org/packages/release/bioc/html/CNTools.html.).

60. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).

61. Kuleshov MV, *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* **44**, W90-97 (2016).

62. Patel AP, *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396-1401 (2014).

63. Puram SV, *et al.* Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell* **171**, 1611-1624.e1624 (2017).

64. Dong B, *et al.* Single-cell analysis supports a luminal-neuroendocrine transdifferentiation in human prostate cancer. *Commun Biol* **3**, 778 (2020).

65. Henry GH, *et al.* A Cellular Anatomy of the Normal Adult Human Prostate and Prostatic Urethra. *Cell Rep* **25**, 3530-3542.e3535 (2018).

66. The ggplot2 package. https://cran.r-project.org/web/packages/ggplot2/index.html.).

67. The maftools package.https://www.bioconductor.org/packages/release/bioc/html/maftools.html.).

68. The ComplexHeatmap package. https://bioconductor.org/packages/release/bioc/html/ComplexHeatmap.html.).

69. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* **33**, 495-502 (2015).

70. Gao J, *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science signaling* **6**, pl1 (2013).

71. Cerami E, *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery* **2**, 401-404 (2012).

72. Vivian J, *et al.* Toil enables reproducible, open source, big biomedical data

analyses. *Nat Biotechnol* **35**, 314-316 (2017).

73. Robinson DR*, et al.* Integrative clinical genomics of metastatic cancer. *Nature* **548**, 297-303 (2017).

74. Kinker GS*, et al.* Pan-cancer single-cell RNA-seq identifies recurring programs of cellular heterogeneity. *Nature genetics* **52**, 1208-1218 (2020).

75. Cheng Q*, et al.* Pre-existing Castration-resistant Prostate Cancer-like Cells in Primary Prostate Cancer Promote Resistance to Hormonal Therapy. *European urology* **81**, 446-455 (2022).

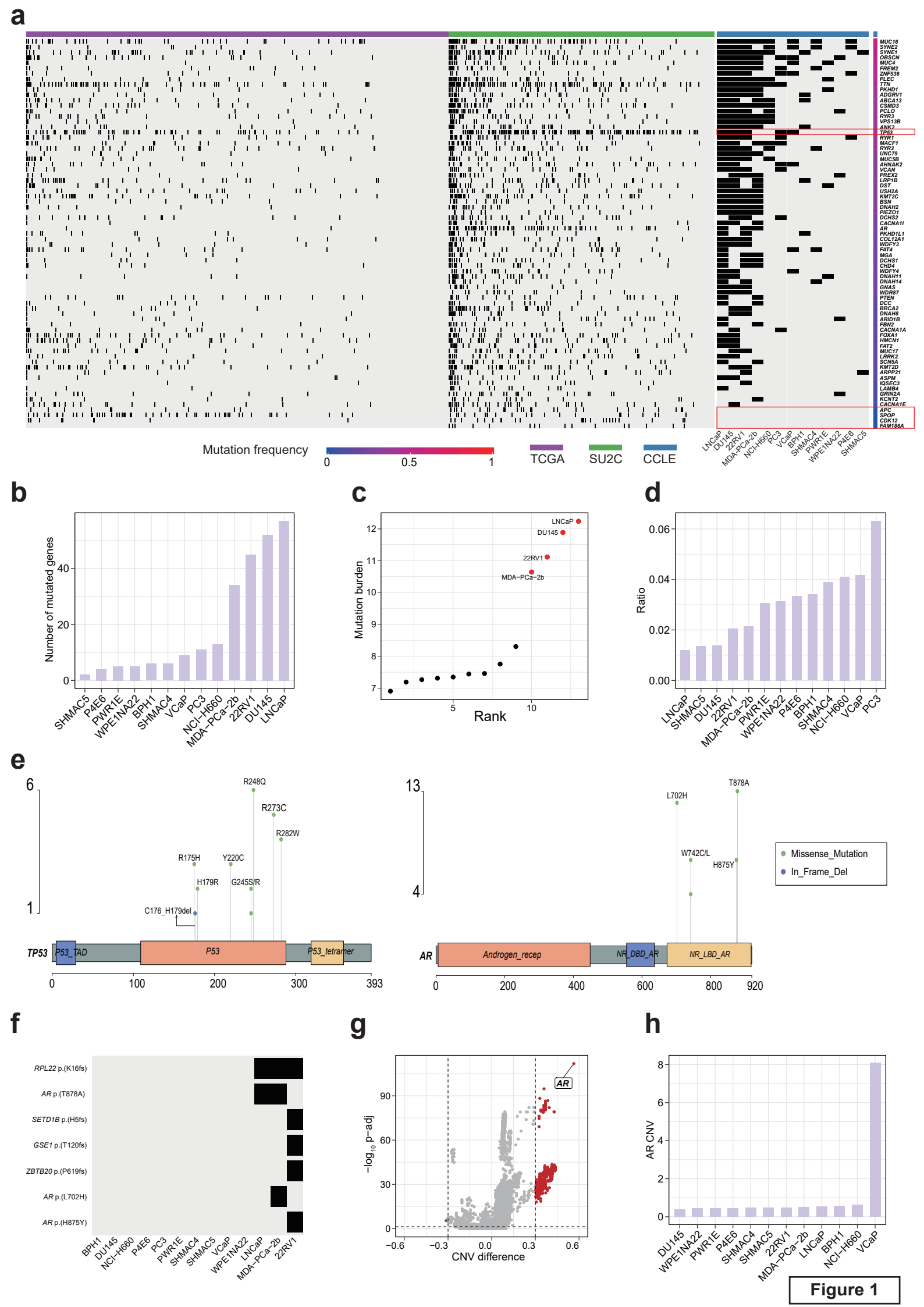

a
Mutation frequency
0
0.5
1
TCGA
SU2C
CCLE
LNCaP
DU145
22RV1
MDA-PCa-2b
NCI-H660
PC3
VCaP
BPH1
SHMAC4
PWR1E
WPE1NA22
P4E6
SHMAC5
b
Number of mutated genes
c
Mutation burden
Rank
d
Ratio
e
TP53
P53_TAD
P53
P53_tetramer
R248Q
R273C
R282W
R175H
Y220C
H179R
G245S/R
C176_H179del
AR
Androgen_recep
NR_DBD_AR
NR_LBD_AR
L702H
T878A
W742C/L
H875Y
Missense_Mutation
In_Frame_Del
f
RPL22 p.(K16fs)
AR p.(T878A)
SETD1B p.(H5fs)
GSE1 p.(T120fs)
ZBTB20 p.(P619fs)
AR p.(L702H)
AR p.(H875Y)
g
−log10 p-adj
CNV difference
h
AR CNV

Figure 1

Figure 1: Systematic comparison of genomic profiles

(a) Somatic mutation landscape of the 73 selected genes across TCGA (primary), SU2C (metastatic), and CCLE (cell line) samples. The top annotation bar indicates the data source, and the right bar shows the mutation frequency of each gene. *TP53*, *APC*, *SPOP*, *CDK12*, and *FAM186A* are highlighted with red boxes.

(b) Ranking CCLE prostate cancer cell lines based on the total number of mutations in the 73 genes.

(c) Distribution of mutation burden in CCLE prostate cancer cell lines, revealing a bimodal pattern.

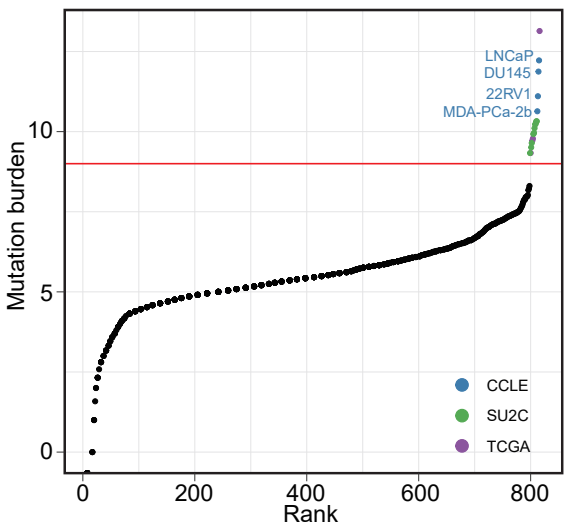
(d) Ranking CCLE prostate cancer cell lines based on normalized ratio (mutation count adjusted by mutation burden).

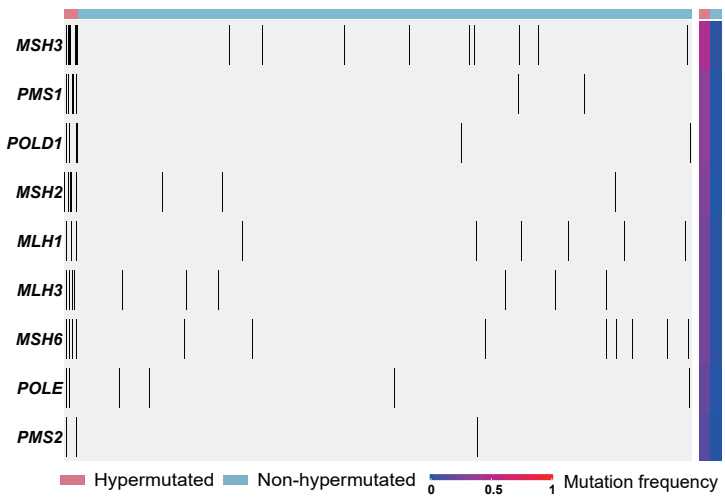
(e) Lollipop plots showing mutation hotspots in *TP53* (left) and *AR* (right). The protein domains are annotated according to PFAM.

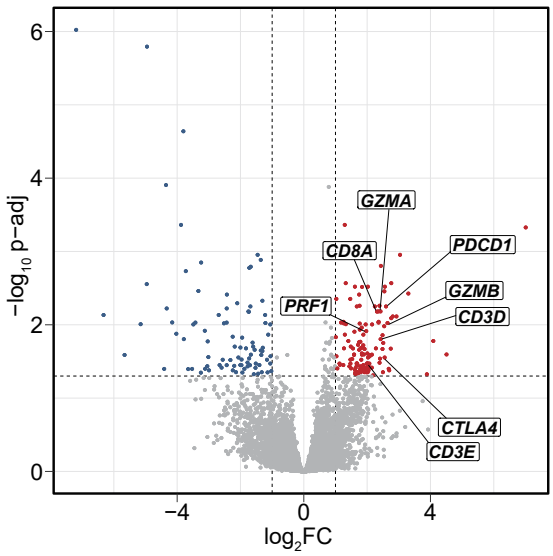
(f) Presence (or absence) of mutation hotspots across CCLE prostate cancer cell lines.

(g) Volcano plot comparing copy number variations (CNVs) between SU2C and TCGA samples. Each point represents a gene; the x-axis denotes the difference in median CNV values (SU2C – TCGA), and the y-axis shows statistical significance ($-\log_{10}$(adjusted *P*-value)). Dashed vertical lines represent CNV difference thresholds (±0.3), and the horizontal dashed line indicates an adjusted *P*-value cutoff of 0.05.

(h) CNV value of the *AR* gene across CCLE prostate cancer cell lines.


a
Mutation burden
Rank
LNCaP
DU145
22RV1
MDA-PCa-2b
CCLE
SU2C
TCGA


b
MSH3
PMS1
POLD1
MSH2
MLH1
MLH3
MSH6
POLE
PMS2
Hypermutated
Non-hypermutated
Mutation frequency


c
−log10 p−adj
log2FC
GZMA
CD8A
PDCD1
GZMB
PRF1
CD3D
CTLA4
CD3E

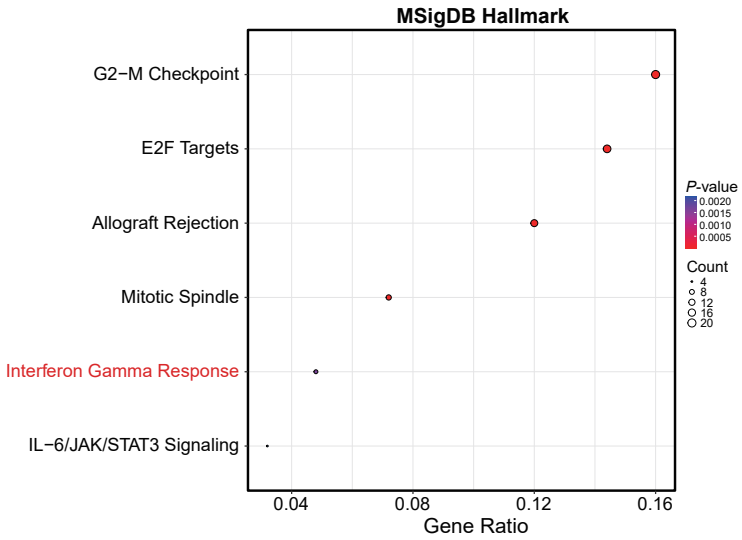

d
MSigDB Hallmark
G2−M Checkpoint
E2F Targets
Allograft Rejection
Mitotic Spindle
Interferon Gamma Response
IL−6/JAK/STAT3 Signaling
Gene Ratio
P-value
Count

Figure 2

Figure 2: Hypermutated prostate cancer has increased cytotoxic $CD8^+$ T cell infiltration

(a) Identification of hypermutated samples. Samples from TCGA, SU2C, and CCLE are combined and ranked by total mutation burden. The red line denotes the threshold used to define hypermutation. Hypermutated samples are color-coded by dataset origin, and hypermutated CCLE cell lines are labeled.

(b) Somatic mutation profile of mismatch repair (MMR) genes in hypermutated versus non-hypermutated samples. The top annotation bar indicates sample classification (hypermutated vs. non-hypermutated), while the right annotation bar shows the mutation frequency of each MMR gene.

(c) Volcano plot of differentially expressed genes between hypermutated and non-hypermutated TCGA samples. Red and blue dots indicate significantly upregulated and downregulated genes, respectively (adjusted *P*-value < 0.05, $|\log_2 FC| > 1$). The x-axis shows $\log_2$ fold change; the y-axis shows $-\log_{10}$(adjusted *P*-value). Dashed vertical lines indicate $|\log_2 FC| = 1$, and the horizontal line marks the significance threshold of *P*-value = 0.05.

(d) Enrichment analysis of the 125 genes significantly upregulated in hypermutated TCGA samples using hallmark gene sets from the MSigDB database.

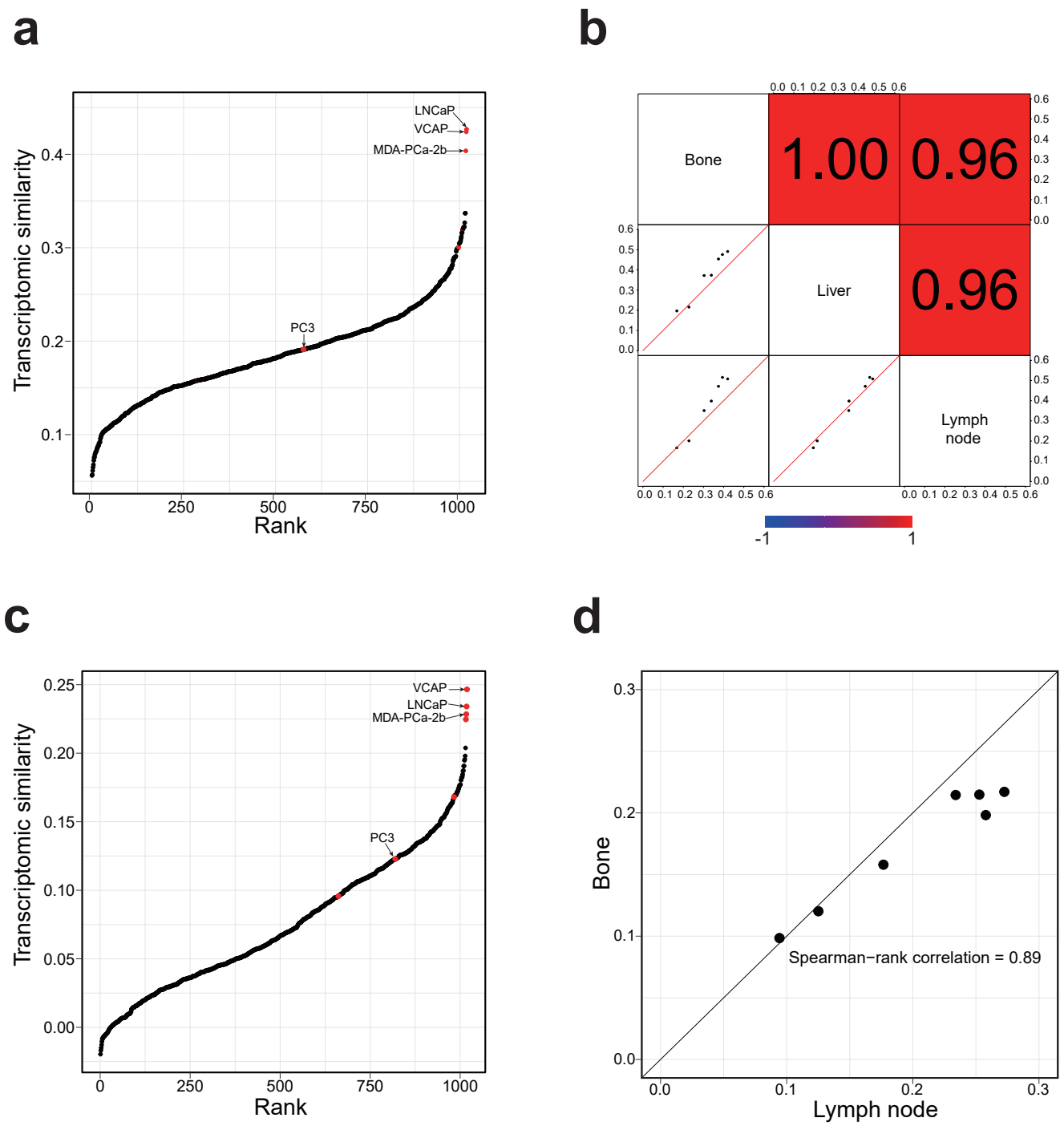

a
Transcriptomic similarity
Rank
LNCaP
VCAP
MDA-PCa-2b
PC3
b
Bone
Liver
Lymph node
1.00
0.96
0.96
-1
1
c
Transcriptomic similarity
Rank
VCAP
LNCaP
MDA-PCa-2b
PC3
d
Bone
Lymph node
Spearman−rank correlation = 0.89

Figure 3

Figure 3: Correlating cell lines with metastatic prostate cancer using bulk and single-cell RNA-seq data

(a) Ranking 1,019 CCLE cell lines based on their transcriptomic similarity to MET500 prostate cancer samples. Each dot represents a CCLE cell line, and the prostate cancer cell lines are highlighted in red.

(b) Pair-wise comparison of site-specific TC analysis results. In the lower-left plots, each dot is a CCLE prostate cancer cell line, with the two-axis representing transcriptomic similarity to MET500 prostate cancer samples of the two intersecting sites. The upper-right shaded values are the corresponding pair-wise Spearman rank correlation values of each pair.

(c) Ranking 1,019 CCLE cell lines based on their transcriptomic similarity to the malignant cells from a metastatic prostate cancer scRNA-seq dataset. Each dot represents a CCLE cell line, and the prostate cancer cell lines are highlighted in red.

(d) TC analysis results using single malignant cells from different metastatic sites are highly correlated. Each dot represents a CCLE prostate cancer cell line, with x-axis and y-axis representing transcriptomic similarity to the malignant cells from lymph node and bone metastases, respectively.

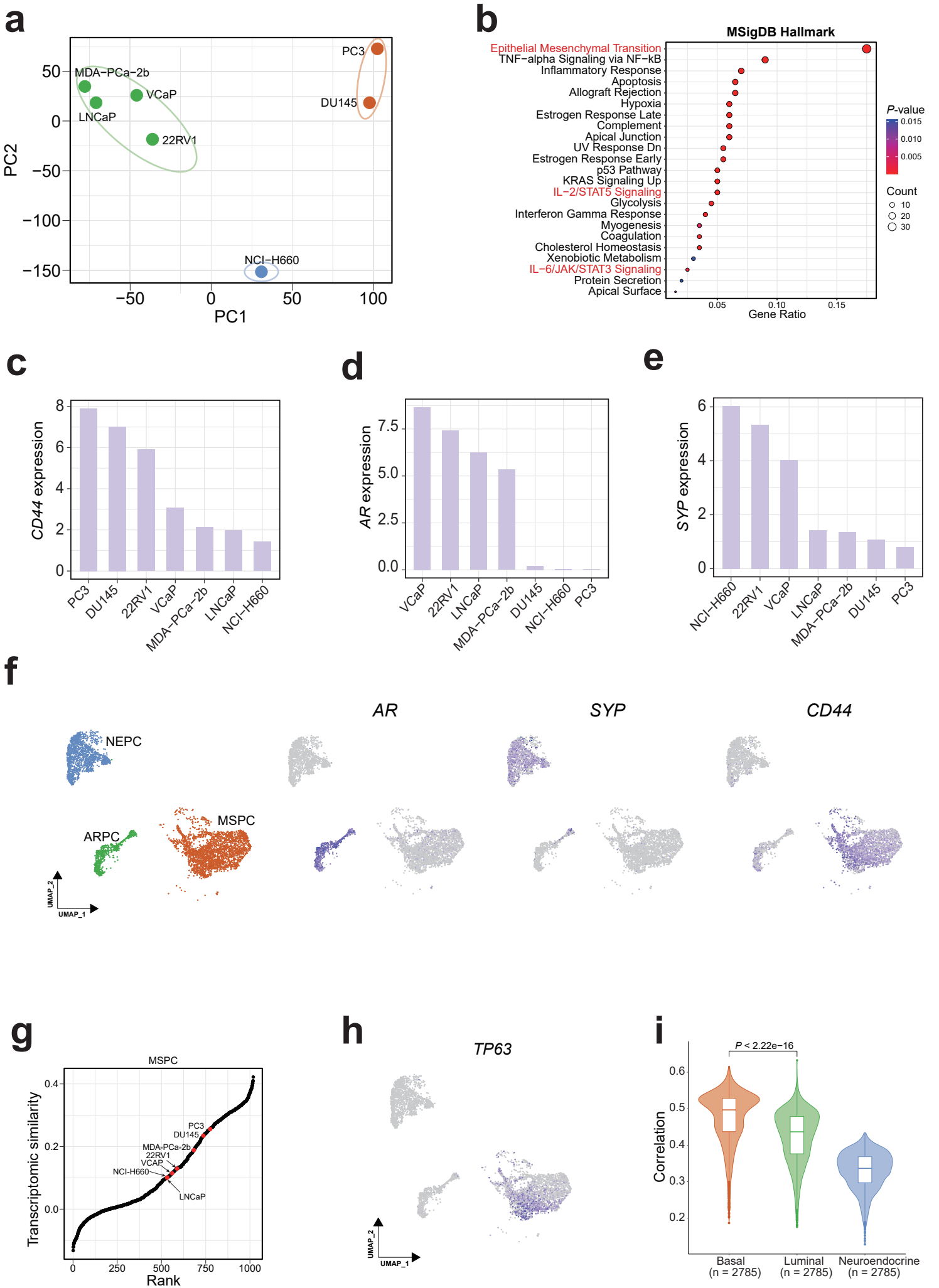

a
PC3
MDA-PCa-2b
VCaP
DU145
LNCaP
22RV1
NCI-H660
PC2
PC1
b
MSigDB Hallmark
Epithelial Mesenchymal Transition
TNF-alpha Signaling via NF-kB
Inflammatory Response
Apoptosis
Allograft Rejection
Hypoxia
Estrogen Response Late
Complement
Apical Junction
UV Response Dn
Estrogen Response Early
p53 Pathway
KRAS Signaling Up
IL-2/STAT5 Signaling
Glycolysis
Interferon Gamma Response
Myogenesis
Coagulation
Cholesterol Homeostasis
Xenobiotic Metabolism
IL-6/JAK/STAT3 Signaling
Protein Secretion
Apical Surface
Gene Ratio
P-value
Count
c
CD44 expression
d
AR expression
e
SYP expression
f
NEPC
ARPC
MSPC
AR
SYP
CD44
UMAP_1
UMAP_2
g
MSPC
Transcriptomic similarity
Rank
PC3
DU145
MDA-PCa-2b
22RV1
VCAP
NCI-H660
LNCaP
h
TP63
i
P < 2.22e-16
Correlation
Basal (n = 2785)
Luminal (n = 2785)
Neuroendocrine (n = 2785)

Figure 4

Figure 4: Limitation of the PC3 cell line

(a) Principal Component Analysis (PCA) of CCLE prostate cancer cell lines based on gene expression profiles. Three distinct clusters are identified and color-coded.

(b) Enrichment analysis of the top 200 genes with the highest positive loadings along PC1 using hallmark gene sets from the MSigDB database.

(c-e) Expression of key lineage markers *CD44* (c), *AR* (d) and *SYP* (e) across CCLE prostate cancer cell lines.

(f) UMAP visualization of the malignant cells from a castration-resistant prostate cancer (CRPC) scRNA-seq dataset. The left panel shows subtype annotations; the right three panels depict expression levels of *AR*, *SYP*, and *CD44*, respectively, and color represents expression level, from gray (low) to dark blue (high).

(g) Ranking 1,019 CCLE cell lines based on their transcriptomic similarity to the MSPC malignant cells. Each dot represents a CCLE cell line, and the prostate cancer cell lines are marked highlighted in red.

(h) Expression of *TP63* in malignant cells from the CRPC scRNA-seq dataset. Color represents expression level, from gray (low) to dark blue (high).

(i) Transcriptomic correlation between MSPC malignant cells and three normal prostate epithelial cell types. In each box, the central line represents the median value and the bounds represent the 25th and 75th percentiles (interquartile range). The whiskers encompass 1.5 times the interquartile range. Outliers are shown as individual points. *P*-value was calculated using the two-sided Wilcoxon signed-rank test.

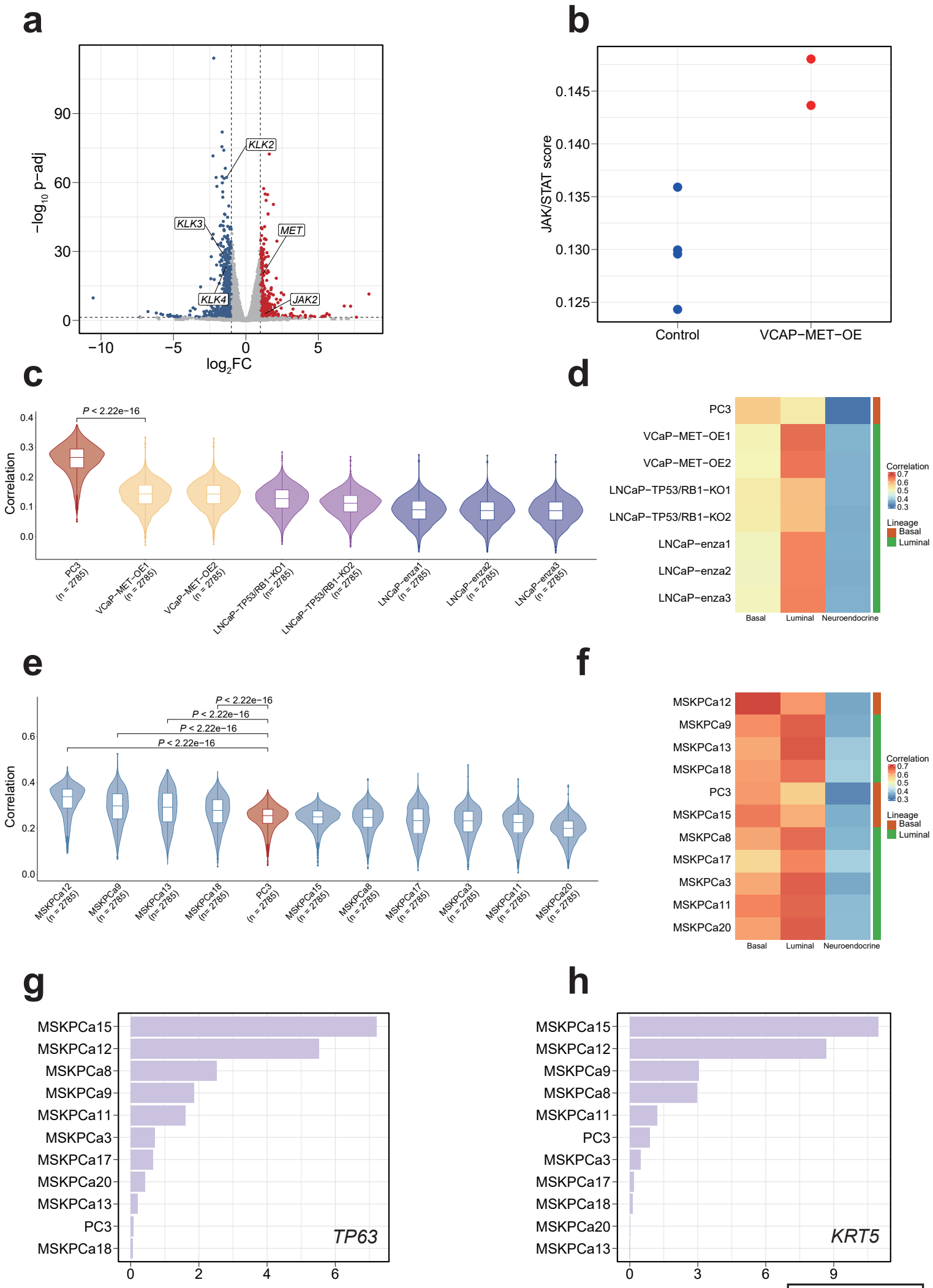

a
KLK2
KLK3
MET
KLK4
JAK2
−log10 p−adj
log2FC
b
JAK/STAT score
Control
VCAP−MET−OE
c
P < 2.22e−16
Correlation
PC3 (n = 2785)
VCaP−MET−OE1 (n = 2785)
VCaP−MET−OE2 (n = 2785)
LNCaP−TP53/RB1−KO1 (n= 2785)
LNCaP−TP53/RB1−KO2 (n= 2785)
LNCaP−enza1 (n = 2785)
LNCaP−enza2 (n = 2785)
LNCaP−enza3 (n = 2785)
d
PC3
VCaP−MET−OE1
VCaP−MET−OE2
LNCaP−TP53/RB1−KO1
LNCaP−TP53/RB1−KO2
LNCaP−enza1
LNCaP−enza2
LNCaP−enza3
Basal
Luminal
Neuroendocrine
Correlation
Lineage
Basal
Luminal
e
P < 2.22e−16
Correlation
MSKPCa12 (n = 2785)
MSKPCa9 (n = 2785)
MSKPCa13 (n= 2785)
MSKPCa18 (n= 2785)
PC3 (n = 2785)
MSKPCa15 (n = 2785)
MSKPCa8 (n = 2785)
MSKPCa17 (n = 2785)
MSKPCa3 (n = 2785)
MSKPCa11 (n = 2785)
MSKPCa20 (n = 2785)
f
MSKPCa12
MSKPCa9
MSKPCa13
MSKPCa18
PC3
MSKPCa15
MSKPCa8
MSKPCa17
MSKPCa3
MSKPCa11
MSKPCa20
Basal
Luminal
Neuroendocrine
Correlation
Lineage
Basal
Luminal
g
MSKPCa15
MSKPCa12
MSKPCa8
MSKPCa9
MSKPCa11
MSKPCa3
MSKPCa17
MSKPCa20
MSKPCa13
PC3
MSKPCa18
TP63
h
MSKPCa15
MSKPCa12
MSKPCa9
MSKPCa8
MSKPCa11
PC3
MSKPCa3
MSKPCa17
MSKPCa18
MSKPCa20
MSKPCa13
KRT5

Figure 5

Figure 5: Patient-derived organoids better resemble the transcriptome of MSPC than PC3

(a) Volcano plot showing differentially expressed genes between VCaP-MET-OE and control groups. Red and blue dots indicate significantly upregulated and downregulated genes, respectively, in the VCaP-MET-OE group.

(b) Dot plot displaying ssGSEA scores for the JAK/STAT signaling pathway in VCaP-MET-OE and control groups.

(c) Violin plots showing the Spearman correlation between MSPC malignant cells and engineered prostate cancer cell lines. In each box, the central line represents the median value and the bounds represent the 25th and 75th percentiles (interquartile range). The whiskers encompass 1.5 times the interquartile range. Outliers are shown as individual points. The *P*-value was calculated using the two-sided Wilcoxon signed-rank test.

(d) Heatmap showing transcriptomic correlation between engineered prostate cancer cell lines and three normal prostate epithelial cell types. Lineage identity was assigned based on the highest correlation.

(e) Violin plots showing the Spearman correlation between MSPC malignant cells and patient-derived organoids. Plot elements and statistical analysis as in panel (c).

(f) Heatmap showing transcriptomic correlation between patient-derived organoids and three normal prostate epithelial cell types. Lineage identity was assigned based on the highest correlation.

(g-h) Expression of *TP63*(g) and *KRT5* (h) in patient-derived organoids and PC3 cell line.