Title

The Proteogenomic Landscape of Curable Prostate Cancer

Permalink
https://escholarship.org/uc/item/35z033sq

Journal
Cancer Cell, 35(3)

ISSN
1535-6108

Authors
Sinha, Ankit
Huang, Vincent
Livingstone, Julie
et al.

Publication Date
2019-03-01

DOI
10.1016/j.ccell.2019.02.005

Peer reviewed

Published in final edited form as:

Cancer Cell. 2019 March 18; 35(3): 414–427.e6. doi:10.1016/j.ccell.2019.02.005.

The Proteogenomic Landscape of Curable Prostate Cancer

Ankit Sinha[1,†], **Vincent Huang**[2,†], **Julie Livingstone**[2,†], **Jenny Wang**[2,3,7], **Natalie S. Fox**[1,2], **Natalie Kurganovs**[2,4], **Vladimir Ignatchenko**[4], **Katharina Fritsch**[1,4], **Nilgun Donmez**[5], **Lawrence E. Heisler**[2], **Yu-Jia Shiah**[2], **Cindy Q. Yao**[2], **Javier A. Alfaro**[1,2], **Stas Volik**[5], **Anna Lapuk**[5], **Michael Fraser**[2], **Ken Kron**[4], **Alex Murison**[4], **Mathieu Lupien**[1,2,4], **Cenk Sahinalp**[5], **Colin C. Collins**[5,6], **Bernard Tetu**[9], **Mehdi Masoomian**[8], **David M. Berman**[3,7], **Theodorus van der Kwast**[4,8], **Robert G. Bristow**[1,4,10], **Thomas Kislinger**[1,4,*], and **Paul C. Boutros**[1,2,11,12,13,14,15,16,*]

[1]Department of Medical Biophysics, University of Toronto; Toronto, ON M5G 1L7, Canada

[2]Ontario Institute for Cancer Research; Toronto, ON M5G 0A3, Canada

[3]Department of Pathology and Molecular Medicine, Queen’s University; Kingston, ON K7L 3N6 Canada

[4]Princess Margaret Cancer Centre, University Health Network; Toronto, ON M5G 2M9, Canada

[5]Vancouver Prostate Centre; Vancouver, BC V6H 3Z6, Canada

[6]Department of Urologic Sciences, University of British Columbia; Vancouver, BC V5Z 1M9, Canada

[7]Queen’s Cancer Research Institute, Queen’s University, Kingston, ON K7L 3N6, Canada

[8]Department of Pathology, Laboratory Medicine Program, University Health Network, Toronto, ON M5G 2C4, Canada

[9]Department of Pathology and Research Centre of CHU de Québec-Université Laval, Québec City, QC G1V 4G2, Canada

[10]Department of Radiation Oncology, University of Toronto, Toronto, ON M5T 1P5, Canada

[*]Correspondence to: **Dr. Paul C. Boutros**, 12-109 CHS; 10833 Le Conte Avenue; Los Angeles, CA 90095, pboutros@mednet.ucla.edu, Phone: 310-794-7160, **Dr. Thomas Kislinger**, MaRS Centre, Princess Margaret Cancer Research Tower; 101 College Street, Toronto, Ontario, Canada, M5G 1L7, Thomas.Kislinger@utoronto.ca, Phone: 416-581-7627.

Authors’ contributions
Sample preparation and LC-MS analyses: AS, KF, MF, KK, AM
Pathology Analyses: MM, BT, TvdK
Performed statistical and bioinformatics analyses: AS, VH, JL, JW, NSF, NK, VI, ND, LEH, YJS, CQY, JAA, SV, AL
Wrote the first draft of the manuscript: VH, JL, AS, TK, PCB
Initiated the project: TvdK, RGB, TK, PCB
Supervised research: CS, CCC, DMB, ML, TvdK, RGH, TK, PCB
Approved the manuscript: all authors

[†]These authors contributed equally to this work

Declaration of Interest
All authors declare that they have no conflicts of interest.

[11]Department of Pharmacology and Toxicology, University of Toronto, Toronto, ON M5S 1A8, Canada

[12]Department of Human Genetics, University of California, Los Angeles, CA 90095, USA

[13]Department of Urology, University of California, Los Angeles, CA 90024, USA

[14]Jonsson Comprehensive Cancer Centre, University of California, Los Angeles, CA 90024, USA

[15]Institute for Precision Health, University of California, Los Angeles, CA 90024, USA

[16]Lead Contact

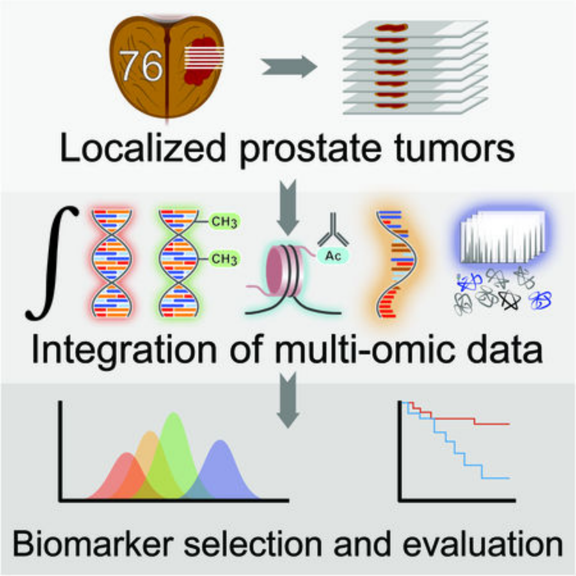
Summary

DNA sequencing has identified recurrent mutations that drive the aggressiveness of prostate cancers. Surprisingly, the influence of genomic, epigenomic and transcriptomic dysregulation on the tumor proteome remains poorly understood. We profiled the genomes, epigenomes, transcriptomes and proteomes of 76 localized, intermediate-risk prostate cancers. We discovered that the genomic subtypes of prostate cancer converge on five proteomic subtypes, with distinct clinical trajectories. ETS fusions, the most common alteration in prostate tumors, affect different genes and pathways in the proteome and transcriptome. Globally, mRNA abundance changes explain only ~10% of protein abundance variability. As a result, prognostic biomarkers combining genomic or epigenomic features with proteomic ones significantly outperform biomarkers comprised of a single data-type.

Abstract

Sinha et al. determine the proteogenomic landscape of localized, intermediate-risk prostate cancers and show that the presence of ETS gene fusions has one of the strongest effects on the proteome. Prognostic biomarkers that integrate multi-omics significantly outperform those comprised of a single data-type.

Graphical Abstract


76
Localized prostate tumors
CH3
CH3
Ac
Integration of multi-omic data
Biomarker selection and evaluation

Keywords

Prostate cancer; genome; epigenome; transcriptome; proteome; biomarker; multi-omic features

Introduction

Prostate cancer remains the most common non-skin malignancy in men world-wide. In many regions, its incidence is increasing because of demographic shifts in population structure and increased life-expectancy (Canadian Cancer Statistics Advisory Committee, 2017; Center et al., 2012; Torre et al., 2016). Prostate cancer is most often diagnosed while still localized, largely through screening with digital rectal exams and quantitation of serum levels of prostate specific antigen (PSA). As a result, ~75% of patients receive definitive local treatment with either surgery or radiotherapy. The combination of these morbid treatments and the large number of indolent tumors diagnosed has led to significant over-treatment, creating an urgent need for more appropriate prognostic assays (Cooperberg et al., 2009).

The genome and epigenome of prostate cancer have been well-studied. Large consortia have cataloged genomic and transcriptomic aberrations, including driver events (Baca et al., 2013; Fraser et al., 2017; Hopkins et al., 2017; The Cancer Genome Atlas Research Network, 2015). Subclonal reconstructions have discovered dramatic intra-tumoral heterogeneity and subclonal selection during disease evolution (Boutros et al., 2015; Cooper et al., 2015; Espiritu et al., 2018; Gundem et al., 2015). The epigenome of localized disease has been analyzed, both for CpG methylation and chromatin marks (Brocks et al., 2014; Kron et al., 2017). Candidate prognostic biomarkers have been developed using copy number and transcriptome data (Blume-Jensen et al., 2015; Cuzick et al., 2011; Den et al., 2015; Klein et al., 2014; Lalonde et al., 2014). Consequently, many mutations are known to drive the tumorigenesis and aggressivity of localized prostate cancer, with alteration of pathways including hypoxia response, androgen signaling and DNA repair (The Cancer Genome Atlas Research Network, 2015).

However, the ways in which the proteome of localized prostate cancer is shaped by genomic, epigenomic and transcriptomic aberrations is almost entirely unknown. Recent studies in breast, ovarian and colorectal cancer have suggested that the transcriptome is a poor proxy for the proteome, with only 10–20% of variation in protein abundance explained by mRNA abundance (Mertins et al., 2016; Zhang et al., 2014, 2016). Previous studies of the prostate cancer proteome have focused on cohorts of patients with inconsistent clinical features (Drake et al., 2016; Iglesias-Gato et al., 2016, 2018; Latonen et al., 2018; The Cancer Genome Atlas Research Network, 2015), and indeed no tumor type has yet integrated whole genome, epigenome and transcriptome data. Improved understanding of the dysregulation of the prostate cancer proteome can enhance interpretation of driver aberrations and facilitate development of rapidly-implementable clinical assays based on immunohistochemistry techniques. To fill this gap, we performed a proteogenomic analysis of a richly annotated cohort of localized prostate cancers.

Results

The proteome of curable prostate cancers

To understand the flow of biological information in prostate tumors, we assembled a clinically-homogeneous cohort of 76 patients diagnosed with sporadic, localized, treatment-naive intermediate-risk prostate cancer (Table 1). All patients were treated by radical prostatectomy, with a median follow-up of 6.8 years (Figure 1A; Table S1). The histologically most representative regions (i.e. the index lesion - used for initial diagnosis and treatment, see STAR Methods) were subject to array-based copy number aberration (CNA) profiling and whole-genome sequencing (WGS) to detect genomic rearrangements (GRs), single nucleotide variants (SNVs), chromothripsis and kataegis. The epigenome was evaluated with methylome profiling and for 35 cases, the *cis*-regulatory element landscape was assessed using histone H3K27Ac ChIP-Seq (Kron et al., 2017). The transcriptome was quantified with both RNA-Seq and microarrays. Finally, the proteome was quantified *via* mass spectrometry-based shotgun proteomics, with each sample analyzed in duplicate. Globally, replicate analyses demonstrated high correlation both in terms of detection (>85% detected in both replicates) and quantification (Pearson's correlation >0.95; data not shown). In Data S1, we provide a proteogenomic fingerprint for every analyzed patient tumor, including data for the replicate proteome analyses demonstrating high correlation. Tumors were sequenced to a mean coverage of 79x ± 28x and normal blood reference samples to a mean coverage of 46x ± 17x. To ensure detection of low-abundance transcripts and accurate quantitation of the full dynamic range, ultra-deep RNA-Seq was performed (median 382 M ± 138 M reads per tumor).

We detected 7,054 protein groups (Table S2), corresponding to 6,924 protein coding genes. Of these, 3,397 protein groups were detected and quantified in all 76 patients (Figure 1B, top), including those corresponding to classic prostate cancer-associated genes like the prostate serum antigen (PSA) gene *KLK3* and the DNA damage repair gene *ATM*, for which both germline polymorphisms and somatic SNVs are associated with patient outcome

(Fraser et al., 2017; Pritchard et al., 2016). We separated the 7,054 protein groups into deciles based on their median abundance (Figure 1B, bottom). As expected, high-abundance proteins were observed in a larger fraction of samples, replicating previous mass spectrometry results (Kislinger et al., 2006; Liu et al., 2004). Proteins encoded by most prostate cancer driver genes were detected in over 70% of the analyzed tumors, including MED12, FOXA1, NKX3–1 and PTEN, amongst others.

To understand the global proteomic patterns of primary localized prostate cancer, we performed subtype discovery, identifying four clusters of proteins (i.e. P1, P2, P3 and P4) and five clusters of patients (i.e. C1, C2, C3, C4 and C5; Figure 1C). Protein clusters P1 and P3 are enriched for products of immune-related genes (Sallari et al., 2016) whereas no significantly enriched pathways were detected in P2 and P4 (Table S3). Of the five distinct patient subgroups, C2 and C3 are associated with an increased rate of biochemical recurrence (BCR; Figure 1D). Because CNAs are tightly associated with patient outcome (Hieronymus et al., 2014), we compared these five proteomic clusters to our previously described genomic subtypes and to the TCGA prostate cancer subtypes (Lalonde et al., 2014; The Cancer Genome Atlas Research Network, 2015). Genomic and proteomic subtypes were largely independent (Adjusted Rand Index (ARI) = −0.004; Figure S1A; ARI = 0.037; Figure S1B). This suggests that nucleotide features are poor proxies for proteomic diversity. Proteomic subtypes were also independent of androgen receptor activity signatures (Stelloo et al., 2015), as expected for treatment-naïve, hormone-sensitive tumors (Figure S1C).

While patient clusters were generally not significantly associated with mutational burden in either SNVs or GRs (Table S3), the abundances of specific proteins were associated with clinical phenotypes. Percent genome altered (PGA), a biomarker of aggressive disease (Lalonde et al., 2014) was found to be associated with 421 proteins, where the most significant associations included a protein involved in Wnt signaling, FZD7 (Spearman's $\rho = -0.59$; FDR = 6.18×10^{-4}), and a deubiquitinase, USP11 (Spearman's $\rho = -0.58$; FDR = 6.18×10^{-4}) (Figure S1D, Table S4). Tumor size was associated with the abundance of eight proteins, while presence of the aggressive intraductal carcinoma/cribriform architecture sub-pathology was associated with seven, including *PTEN* ($\Delta_{median\_protein} = -12.96$; FDR = 0.227), recapitulating previous findings (Bhandari et al., 2019; Chua et al., 2017).

ETS gene fusions are linked to cell migration and lipid metabolism

One of the strongest effects on the proteome was the presence of ETS gene fusions. These fusions are the most frequent somatic aberration in prostate cancer, and are not associated with clinical outcome (Dal Pra et al., 2013; Minner et al., 2011). We focused on 245 mRNAs and 68 proteins significantly associated with ETS gene fusion status (Table S4; mRNA $Q < 0.01$; protein $Q < 0.05$; 36 overlapping genes, 277 genes in total). To be conservative we excluded 22 genes with a high proportion of missing protein abundance measurements. Overall changes in mRNA and protein abundances were well correlated (Spearman's $\rho = 0.72$, $p < 2.2 \times 10^{-16}$; Figure 2A), but protein abundances showed larger dynamic range than mRNA abundances. The median differentially abundant mRNA differed 1.50-fold between ETS fusion-positive and ETS fusion-negative tumors, while the median protein differed

1.66-fold ($p = 4.63 \times 10^{-6}$; paired Mann-Whitney U test). The many genes showing only mRNA or protein abundances associations with ETS fusions, but not both, may be attributed to biological factors like translational and post-translational regulation, as well as to technical factors.

For some individual genes, mRNA and protein abundances diverged dramatically. The transcription factor EB (TFEB) was almost unchanged at the RNA level (1.46 fold higher in tumors with an ETS gene fusion) but was 1,012-fold higher at the protein level. Similarly, Lysyl Oxidase (LOX) was 1.88-fold higher at the RNA level in tumors with an ETS gene fusion, but had 21,031-fold higher protein abundance. While relative quantification by label-free proteomics is well-established, the presence of missing values (i.e. protein not detected and quantified in a sample) are caveats for binary comparisons. More accurate, absolute quantitation *via* targeted proteomics assays and stable isotope labeled standards, are needed to better understand these divergences.

To better understand the differences in ETS fusion associated genes in the transcriptome and proteome, we expanded our analysis to methylation, histone status (H3K27Ac) and copy number (CN) data (Figure 2B). Only a single gene, *ARHGDIB* was associated with ETS gene fusions at the protein, mRNA, methylation and acetylation levels. One gene contained in the deletion region between *TMPRSS2* and *ERG* on chromosome 21, *FAM3B*, showed correlated copy number aberrations, methylation changes and mRNA abundance changes. By contrast, 630 genes showed differential methylation associated with ETS gene fusions, while 124 showed differential H3K27 acetylation. These interactions do not fully explain the modest overlap between ETS-associated proteins and ETS-associated mRNAs, and highlights the importance of post-transcriptional regulatory factors not easily quantified by -omic studies.

To determine if functional inference from RNA and protein data would yield similar conclusions, we performed pathway analysis separately on ETS fusion-associated genes identified at each biological level (i.e. CNAs, methylation, H3K27Ac, RNA abundance and protein abundance). No pathways were associated with CNAs, and only one with H3K27Ac (Figure 2C), although genes associated with differential H3K27 acetylation were enriched for ETS binding motifs (Hypergeometric test; $p = 3.5 \times 10^{-2}$). Genes associated with carboxylic acid metabolism were enriched at the mRNA, protein and methylation level, corroborating links between *ERG* fusions and lipid metabolism (Hansen et al., 2016; Wu et al., 2014). Genes associated with intra- and extracellular vesicles were enriched in the mRNAs and proteins associated with ETS gene fusions. At the mRNA level, we identified enrichment in cell migration, actin binding and phospholipid binding while at the protein level, there was an enrichment in lysosomal genes. These data suggest that a myriad of genomic mechanisms may differentiate the ETS-associated transcriptome from the ETS-associated proteome.

Interestingly, one patient showed ERG over-expression through immunohistochemistry with ERG antibodies, but no ETS gene fusion was detectable by either WGS or RNA Seq (Figure S2A). This tumor exhibited neither the mRNA nor protein signatures of ETS gene fusions (Figure S2B), suggesting not all cases of ERG over-expression validated by IHC will impact a tumor's transcriptional and proteomic repertoire, or potentially reflecting the large spatial heterogeneity of prostate tumor genomes (Boutros et al., 2015). Critically, divergences in signaling pathway detected by transcriptome and proteome data suggest caution when interpreting the effects of genomic aberrations on the basis of the transcriptome alone. Enhanced study of protein abundances in such analyses is key to fully understand the effects of genomic aberrations.

Quantifying transcriptome-proteome discordance

These large discordances between differences in ETS gene fusion-associated mRNA abundance and protein abundance led us to systematically quantify their relationship across all genes. Globally, mRNA and protein abundances are weakly correlated (median Spearman's $\rho = 0.21$) indicating that mRNA abundance is poorly predictive of protein abundance (Figure 3A). RNA-protein correlations varied with protein abundance – the 10% most abundant proteins were much better correlated (median: 0.32) than the bottom 10% (median: 0.07).

One specific example of this phenomenon is ATM, where we detected relatively low protein abundance (8th decile), but higher RNA abundance (3rd decile) and a weak correlation between them (Spearman's $\rho = 0.10$; Figure S3A). *ATM* SNVs are associated with patient outcome, but there was no significant difference in biochemical relapse rate between patients with and without *ATM* loss (Figure S3B). This poor correlation between transcript and protein abundance may reflect differing rates of mRNA degradation, translation or protein degradation, and mirrors recent reports in breast, ovarian and colorectal cancers (Mertins et al., 2016; Zhang et al., 2014, 2016).

RNA-protein correlations were generally less-dependent on transcript abundance than on protein abundance (Figure S3C). This may reflect some combination of larger translational regulation or increased measurement error for low abundance biomolecules (i.e. both transcripts and proteins). Indeed, the 10% most abundant proteins were enriched for localization to membrane-bound organelles and extracellular proteins, and these trends held in an independent, clinically diverse validation cohort (Figure S3D) (Iglesias-Gato et al., 2016).

In extreme cases, either protein or RNA for a gene is detected but the other is not. We examined these cases, focusing on the 10% most abundant transcripts and proteins to minimize the possibility of technical false-negatives. Of the 4,694 most abundant transcripts, 1,342 did not have a detected protein, including 1,070 from known coding genes. By contrast, only 68 of the most highly-abundant proteins had low or undetected transcript abundance (Table S5). Coding transcripts without a detected protein represent a diverse collection of genes preferentially localized to the nucleus, while proteins without detected transcripts are enriched for immune-related genes (Table S5).

cis and *trans* effects of genomic and transcriptomic changes

Prostate cancer is driven by CNAs more than single nucleotide variants: it is a C-class tumor (Fraser et al., 2017). We therefore investigated the role recurrent CNAs play in modulating mRNA and protein abundance, both in *cis* and in *trans* (Figure 3B–C). While genes may be lost or gained as part of a larger segment, the effect on mRNA or protein abundance may vary per gene, therefore we investigate differences per gene and not per segment. To increase statistical power, we generated 210 array-based transcriptomes from localized prostate tumors. To identify *cis* effects, for each gene we compared mRNA and protein abundances between tumors with and without a CNA at that gene (6,607 genes with all three types of data available). We detected strong *cis* effects in RNA, which can be seen along the diagonal of Figure 3C as influencing ~10% of all genes (592/6,607; FDR < 0.05; t-test). These effects were present, but weaker at the protein level, with ~2% of proteins having their abundance associated to CNAs (133/6,607; Figure S4A). We validated this result in TCGA data, which contained 491 samples with matching RNA and CNA data and found that 35% of genes show *cis* effects, highlighting the importance of sample size and, potentially, clinical diversity of the patient cohort, in multi-omic studies (Figure S4B).

Next, to identify *trans* effects where a CNA on one chromosome is associated with mRNA changes of a gene on another chromosome, we repeated our earlier analysis transcriptome- and proteome-wide. For each of the 23,068 genes with copy number information, we identified which of the 6,636 genes with both mRNA and protein abundance data and changed abundance with CNA status. On average, each gene-specific CNA had 593 ± 528 *trans* effects, where it was associated with statistically significant changes in RNA abundance. By contrast, *trans* effects were rarer for protein abundance, influencing 10 ± 31 genes (Figure S4A). For example, deletion of *PTEN* alters abundance in 52% of the genes investigated at the RNA level (3,416/6,607) but only 2.7% at the protein level (179/6,607), all of which showed RNA changes. To be conservative, we removed genes that were themselves frequently copy number altered (>5% of samples) to exclude confounding effects of genomic subtypes. Even after this control, *PTEN* showed *trans* effects on 54% of genes at the RNA level and 2.8% at the protein level (113/4,086; Figure 3D).

Other genes with substantial *tran*s effects included *NXK3–1*, a tumor suppressor deleted in almost half of prostate tumors and *CD68*, which mediates macrophage recruitment (Figure 3D). Overall, 694 genes had large *trans* effects: defined as influencing the RNA levels of at least 10% of all genes. Of these, 67.4% (468/694) also exhibited *trans* effects at the protein level (affecting 0.2% - 40% of proteins); the smaller effect-sizes reflects the small number of samples with protein abundance measurements. Genes with large *trans* effects included *CMAS*, an immune related gene, *ATAD1*, a gene related to ATPase activity, and *MINPP1*, a gene previously implicated in cancer. Interestingly, these three genes were associated with poor prognosis at both the CNA and RNA levels (Figure S4C). Consistent with our observations of ETS gene fusion-associated genes, protein abundances showed a much higher dynamic range for *trans* effects than transcript abundances (e.g. Spearman's ρ = 0.87 for *PTEN*; Figure 3D).

Not all genes were influenced by CNA *trans* effects at the same frequency. For example *CRISP3*, which plays a role in sperm function and is upregulated in prostate tumors (Ribeiro et al., 2011), shows RNA *trans* effects with 9.2% (2,123/23,068) of genes and protein *trans* effects with 0.4% (85/23,068) – in some cases with large magnitudes. *CRISP3* RNA and protein are both more abundant in samples with either *CD68* deletion or *PTEN* deletion (Figure 3D). Thus, a large network of *trans* CNA effects exists, highlighting interconnections between specific somatic mutations and consequent transcriptome and proteome dysregulation.

Multi-layer information flow in prostate cancer

To better quantify this complex flow of information from the cancer genome to its proteome, we performed an information-content analysis. For each gene, we calculated the mutual information (MI) between the five classes of molecular data in our prostate cancer study: CNAs, methylation, histones (H3K27 acetylation), RNA abundance and protein abundance (Figure 4A). MI measures, in bits, the knowledge of one variable when a second variable becomes known. MI values of zero indicate independent variables: knowing one variable gives no information on the other. MI is related to classic correlations, but lacks some of their assumptions about linearity and ordering, making MI useful for complex relationships. To standardize analyses, we median-normalized MI separately for each dataset.

As expected, different pairs of molecules have varying amounts of redundancy in their information content. For example, CNAs were weakly predictive compared to other molecular datatypes. CNAs were modestly more tightly associated with protein than with mRNA abundance (median $MI_{\text{CNA-Protein}} = 0.055$ *vs.* median $MI_{\text{CNA-RNA}} = 0.048$). Similarly, methylation status was more strongly linked to protein than mRNA abundance (median $MI_{\text{Methylation-Protein}} = 0.43$ *vs.* median $MI_{\text{Methylation-RNA}} = 0.32$). Intriguingly, the highest mutual-information across genomic regions was between H3K27Ac and RNA (median $MI_{\text{H3K27Ac-RNA}} = 0.652$), while the lowest was between CNAs and methylation (median $MI_{\text{CNA-Methylation}} = 0.032$). This may suggest a prominent role for epigenomic features, independent of the frequent subclonal CNAs (Espiritu et al., 2018).

To determine if the regulatory relationships between pairs of biomolecules distinguish specific functional groups of genes, we performed consensus clustering and identified six subgroups (labeled MI1–6; Figure 4B). Individual subgroups were not enriched for MSigDB hallmark gene sets, but rather more specific features (Liberzon et al., 2015) (Table S6). Subgroup MI6 was characterized by genes with higher CNA-H3K27Ac, CNA-protein, CNA-RNA and CNA-Methylation links, and was enriched for genes related to cellular response to stress, suggesting tight regulatory networks (FDR = 0.005; Table S6; Figure 4B). By contrast, MI1 harbored genes with strong H3K27Ac-Protein, RNA-Protein, and Methylation-Protein links and are enriched in extracellular exosomes ($FDR = 3.87 \times 10^{-15}$; Table S6). These results are compelling, but further exploration will be required to fully elucidate the biological mechanisms and implications underlying these links in mutual information.

To validate our mutual information findings in an independent dataset, we calculated normalized MI in 245 intermediate risk TCGA samples. For each pair of molecular datatypes, we considered genes with significant MI in the discovery cohort (defined as $MI > 0.05$). We then calculated the MI for these genes in the TCGA cohort, if the same molecular datatype was collected within that cohort. MI values validated strongly, with 99% of genes with significant Methylation-RNA and 75% of genes with significant CNA RNA MI validating (Figure 4C). To quantify the validation of the MI analyses, we created a ROC curve for each molecular datatype by iteratively increasing the MI threshold used for significance in both datasets (Figure S4D).

We followed by calculating the percent variance explained (PVE) by upstream *cis*-information from CNA, methylation and RNA. This analysis was performed on genes present in most samples with a known link to prostate cancer: *TGM2*, *NDRG3*, *KLK3*, *AKT1*, *PTEN*, *NKX3–1*, *KRAS* and *ATM* (Figure 4D). A strong association was detected between methylation and protein abundances for *TGM2* and *AKT1* in which almost 40% and 30% of the protein variance, respectively, can be explained by methylation. Both *NDRG3* and *PTEN* show relatively high PVE by CNAs when compared to the other genes examined (17% and 6.2% respectively), but while 60% of variability *NDRG3* protein abundance can be explained by CNA, methylation and RNA, less than 10% of variability in *PTEN* abundance was captured by the model. Curiously, despite its high abundance, only 33% of the variance in *KLK3* was explained by RNA (21%) and methylation (11%). *KLK3* is generally copy number neutral (Figure 4E) and protein abundance is univariately correlated with RNA ($\rho = 0.48$; $p = 2.53 \times 10^{-4}$) and methylation ($\rho = -0.34$; $p = 3.3 \times 10^{-3}$). In contrast, *PTEN* is dominated by CN losses (42%; 31/74, 2 missing; Figure 4F) and its protein abundance was univariately correlated with RNA ($\rho = 0.29$; $p = 2.6 \times 10^{-2}$) and methylation ($\rho = -0.29$; $p = 1.3 \times 10^{-2}$), but its low RNA values ($\mathrm{median}_{\mathrm{PTEN}} = 4.76$; $\mathrm{median}_{\mathrm{GAPDH}} = 11.8$) and low and narrow methylation values ($Q1_{,}Q3_{\mathrm{PTEN}} = 0.075{,}0.086$; $Q1_{,}Q3_{\mathrm{KLK3}} = 0.12{,}0.13$) may explain the low PVE.

These data provided detailed maps of the complex ways in which information flows from the germline and somatic genome, epigenome and transcriptome and finally to the proteome. By improving quantitation of regulatory data, as by miRNA-profiling or histone ChIP-Seq, specific functional classes of genes can be delineated.

Protein abundances may predict prostate cancer relapse

Finally, to evaluate the potential clinical importance of proteomic profiling of primary prostate tumors, we quantified the association of each gene with disease relapse after definitive local therapy with curative intent. We used time to biochemical relapse (BCR) as our outcome, which reflects rising serum PSA levels, which can trigger administration of salvage therapy. For each gene, we fit Cox proportional hazards (Cox PH) models to patient groups dichotomized by both median protein and median mRNA abundance. Hazard ratios (HRs) from protein abundances were weakly correlated to those from mRNA abundances for the same genes (Spearman's $\rho = 0.25$; Figure 5A). Thus, some individual genes were associated with aggressive disease at the RNA level, others at the protein level, and a subset of 53 at both. Proteins exhibited a wider dynamic range of HRs (range: 0.22 – 4.23) than mRNAs (range: 0.33 – 2.73).

In some cases, mRNA and protein abundances showed divergent associations with patient outcome. For example, increased abundance of *PUS1*, a gene not previously implicated in cancer, was associated with increased risk of BCR, but unexpectedly, increased mRNA abundance was associated with a reduced risk (Figure 5B). In total, six genes showed divergent mRNA-protein associations with patient outcome, which may represent complex regulatory loops, translational dysregulation, post-translational modifications, or post-transcriptional processes that participate in driving aggressive disease. For validation, we focused on the 53 genes whose mRNA and protein abundances were both associated with disease aggressivity. We first considered *ACAD8* (Figure 5C) as it has high protein abundance (3rd decile) and has not been strongly linked to prostate cancer previously. We validated the association of low *ACAD8* protein abundance with poor outcome using immunohistochemistry on a tissue microarray of 73 intermediate-risk prostate tumors (Figure 5D; Figure S5A). Validating these candidate prognostic markers in larger cohorts will be key.

Multi-omic integration improves prediction of patient outcome

Clinically-used biomarkers are derived from many different classes of biomolecules, with DNA- and RNA-based assays being particularly prominent in prostate cancer (Fraser et al., 2015). It is unknown whether a particular class of biomolecules is generally superior for a given biomarker question. To quantitatively address this question, we again focused on prediction of BCR. We performed a null distribution (information content) analysis, generating 10-million gene-sets, each comprising 100 genes randomly selected without replacement (Boutros et al., 2009; Lalonde et al., 2014). This gene-set size was chosen to match that of several validated prognostic biomarkers (Lalonde et al., 2014). For each gene-set, we used supervised machine-learning (random forests) to train and validate CNA, methylation, RNA and protein biomarkers, resulting in 40-million trained biomarkers. For each biomarker, we assessed their accuracy assessed via the area under the receiver-operating characteristic curve (AUC).

The resulting null distributions showed that random biomarkers generated from CNA or methylation data had similar performance (Figure 5E; blue curves, median AUC = 0.60). By contrast, biomarkers generated from mRNA and protein abundances were significantly superior, improving mean AUC by 0.03 for mRNA and 0.04 for protein ($p < 2.2 \times 10^{-16}$; t-test). These results were independent of gene-set size (range: 5–100; Figure S5B). Thus, proteomic features are significantly more informative for BCR prediction than genomic, epigenomic or transcriptomic ones.

These data provided us a unique opportunity to consider the synergy of constructing biomarkers from multiple distinct datatypes. Pair-wise comparison of matched mRNA and protein biomarkers show a classic long-tail distribution, suggestive of potential synergy (Figure S5C). To test this explicitly, we created biomarkers from pairwise combinations of biomolecules and evaluated them using the AUC (Figure S5D). Pairs of data-types produced biomarkers significantly better than genomic-features only. For example, methylation-protein biomarkers were on-average the best combination, with a median AUC of 0.66 (Figure 5E), reflecting their low MI (Figure 4A). Thus, low mutual information amongst different biomolecules may facilitate explicit complexity-accuracy trade-offs in the construction of multi-modal biomarkers, but these are incremental in magnitude and require further validation in the future.

Discussion

Cancer is a disease of the genome. The accumulation of point and structural mutations, along with epigenomic changes, which dysregulates the transcriptome, leads to a dysregulated proteome and ultimately to the hallmark phenotypes of cancer cells (Hanahan and Weinberg, 2011). Proteins are the most abundant class of functional molecules in the cell, and the central dogma guides information flow from the genome and epigenome to the proteome. We therefore created a unique cohort of the most commonly treated form of prostate cancer: localized, intermediate-risk disease. Our compendium of whole genome, epigenome, transcriptome and proteome profiles reveals patterns of information exchange across levels of the central dogma whose biological implications are uncertain, but intriguing.

Previous studies in other tumor types have shown low correlations between the abundances of specific RNA transcripts and the abundances of the resulting proteoform (Mertins et al., 2016; Zhang et al., 2014, 2016). We confirm that the weak transcriptome-proteome relationships in other tumor types (Shao et al., 2017) also exist in prostate cancer, and generalize this to a broad range of genomic and epigenomic features, including *cis*-regulatory elements. These weak correlations are reflected in a large network of *trans* effects across data-types, which differentially affect RNA and protein abundances and are correlated to specific functional sets of genes. These networks may provide an avenue for understanding the influences of specific genomic features on specific aspects of the proteome, and subsequently on downstream pathways, cellular and clinical phenotypes.

The proteomic subtypes of prostate tumors are only weakly related to their genomic ones. This suggests strong post-transcriptional regulatory mechanisms that are not easily detected in genomic data. This observation is mirrored by the differences in specific genes and pathways associated with ETS gene fusion status at the RNA and protein levels. The proteomic characteristics of ETS-positive tumors indicate an extensive dysregulation of their metabolic profile (Bose et al., 2017), which is not reflected in the transcriptional changes seen. Indeed, these data highlight the drawbacks of studies that implicitly infer changes in functional protein directly from mRNA abundance data. The assumption that transcriptional profiles are a reliable surrogate for proteomic ones is incorrect for most genes. Only ~10% of variation in protein abundances is explained by changes in the transcriptome, suggesting an urgent need for statistical models that better predict protein abundances from more readily available nucleotide-based data.

The clinical potential of proteomic biomarkers is high – proteins harbored more information on patient relapse than any other data type, and multi-modal biomarkers consistently out-performed those generated from individual data-types. Yet multi-modal biomarkers are inherently complex, highlighting the need for improved technologies to accurately profile multiple analytes from individual tumors, especially considering that high-throughput proteomics clinical tests have not been implemented in routine practice to date. Larger patient cohorts and complementary validation studies will be key to reaching the translational potential of multi-modal data. This suggests an opportunity for expansion of existing cancer genomics consortia to pair their high-quality, deeply-analyzed genomes and transcriptomes to unbiased proteomic surveys.

STAR METHODS

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Paul C. Boutros (pboutros@mednet.ucla.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Selection of patient cohort and tumor sections—Patient selection, tissue collection and sample processing was performed as previously described (Fraser et al., 2017). Informed consent, consistent with local Research Ethics Board (REB) and International Cancer Genome Consortium (ICGC) guidelines, was obtained at the time of clinical follow-up. Previously collected tumor tissues were used, following University Health Network REB-approved study protocols (UHN 06–0822-CE, UHN 11–0024-CE, CHUQ 2012-913:H12-03-192). All patients were treated surgically via radical prostatectomy (RadP). Primary treatment failure was defined as PSA levels at or above 0.2 ng/mL three months after surgery; no patients in this cohort experienced primary treatment failure. Biochemical recurrence (BCR) after RadP was defined as two consecutive measurements of PSA > 0.2 ng/mL or the administration of salvage radiotherapy. All patients were National Comprehensive Cancer Network (NCCN) intermediate-risk based on pre-surgical parameters. ISUP Scores and tumor cellularity were evaluated by two genitourinary pathologists (T.v.d.K., and B.T.) on scanned haematoxylin- and eosin-stained slides as described previously (Fraser et al., 2017). Adjacent 10 serial sections (10 µm thickness of each section) from each patient tumor were used for acquiring each of the multi-omic dataset. All sections were pathologically inspected by our co-author (T.v.d.K.) and tumor tissue was macro-dissected to reach ~70% cellularity. As a validation, cellularity was calculated *in silico* from OncoScan array data using qpure (v1.1) (Song et al., 2012). Additionally, for all multi-omic analyses, we analyzed the index lesion - the lesion that led to the initial diagnosis and treatment of the patient. Given our focus on biomarkers, this allowed a consistent analysis that avoids conflating information available post-operatively with that available during initial staging and management.

METHOD DETAILS.

Tumor tissue preparation for shotgun proteomics—Fresh-frozen RadP specimens were obtained from the University Health Network (UHN) Pathology BioBank or from the Genito-Urinary BioBank of the Centre Hospitalier Universitaire de Québec (CHUQ). Ten pathologically inspected, optimal cutting temperature compound (OCT) embedded tissue sections (10 µm each) were processed for each prostate cancer sample (i.e. sections are in close proximity to previously used sections for genomics, epigenomics and transcriptomics, but not identical). Tissues were scraped from the glass slides and transferred to a 1.5 mL conical tube. Removal of the OCT compound was performed using various dilutions of ethanol with water as follows. Initially, 1 mL of 70% (v/v) ethanol was added to each conical tube, followed by 30 s of vortex at high speed. The tubes were centrifuged at 14,000-

rcf for 3 min, and the supernatant was discarded. Subsequently, tissue pellets were treated with 100% ethanol, 70% ethanol, 85% ethanol and lastly in 100% ethanol with an additional 5 min incubation at room temperature and rigorous high-speed vortexing. The obtained tissue pellet was then processed for shotgun proteomics.

Shotgun proteomics—Tissue pellets were resuspended in 100 μL of 50% (v/v) 2,2,2-Trifluoroethanol in phosphate buffered saline (pH 7.4) and incubated for one hr at 60°C. Subsequently disulphide bonds were reduced by incubation for 30 min at 60°C with 5-mM dithiothreitol. Afterward, alkylation of reduced disulphide-bridges was performed using 25 mM iodoacetamide for 30 min at room temperature in the dark. Samples were diluted 1:5 using 100 mM ammonium bicarbonate with 2 mM CaCl (pH 8.0). Proteins were digested with 5 μg of trypsin at 37°C overnight. Pe ptides were desalted using C18-based solid phase extraction. Subsequently, eluted peptides were lyophilized and solubilized in mass spectrometry-grade water with 0.1% formic acid. Peptide concentration was quantified using a NanoDrop Lite (at 280 nm) and a constant aliquot of 2 μg of peptides were injected onto the column for chromatography and proteomics analysis. LC-MS/MS data was acquired using an Easy nLC 1000 (Thermo) nano-flow liquid chromatography system with a 50 cm EasySpray (Thermo) column coupled to a Q Exactive (Thermo) tandem mass spectrometer. Data was acquired using a four hr chromatographic gradient with the mass spectrometer operating in data dependent mode. MS^1 data was acquired at resolution of 70,000, while MS^2 data was acquired at resolution of 17,500 with a top 15 method (Michalski et al., 2011; Sinha et al., 2014). The acquired data was searched using MaxQuant (v1.6.1.0) (Cox and Mann, 2008) and a UniProt protein sequences database (complete human proteome; v1-27-2015, number of sequences 42,041). Searches were performed with a maximum of two missed cleavages, cabamidomethylation of cysteine as fixed modification and oxidation of methionine as variable modification. False discovery rate (FDR) was set to 1% for peptide spectral matches and protein identification using a target-decoy strategy (Kislinger et al., 2006). The ProteinGroup.txt file was used for all subsequent analysis. Proteins identified with two or more peptides were carried forward. Relative quantification was performed using MS^1 signal intensity for label-free quantification, following a standard MaxQuant analysis strategy MaxLFQ (Cox et al., 2014), which uses an aggregate of all MS^1 peptide intensities of a reported protein.

Proteomic data batch correction and missing value imputation—Four batch correction methods were evaluated: ComBat (v3.20.0) (Johnson et al., 2007), limma (v3.28.21) (Ritchie et al., 2015) and the removal of one and two surrogate variables using sva (v3.20.0) (Leek et al., 2012). ComBat batch correction was performed using the null model. Correction with limma was performed using the removeBatchEffect command with no additional covariates included. For surrogate variable analysis, biochemical recurrence was used as the endpoint of the model to preserve associated variance. The number of nuisance variables was automatically estimated using the num.sv command, and a data matrix was regenerated following removal of one or two nuisance variables. Metrics used to evaluate the correction methods included examining the variance of *GAPDH*, *SDHA* and *GPI* post correction, fitting a linear model between batch and the protein abundance for each gene as the response and calculating the 90th percentile of percent variance explained by the batch term, and by calculating Spearman's correlation between a duplicated sample in two of the batches. We ranked each method based on these criteria, calculated the rank product and arrived at ComBat as the highest ranked batch correction method for this dataset. Missing values were imputed from the lower half of a Gaussian distribution around a mean of the protein intensities from the 0.01th percentile of all protein intensities. The imputation of missing values can potentially lead to an overestimation for binary comparisons (i.e. ratios).

DNA and RNA-Sequencing—Whole genome sequencing of DNA was performed as previously described (Fraser et al., 2017). RNA samples were sent to BGI Americas and underwent QC and DNAse treatment. For each sample, 200 ng of total RNA was used to construct a TruSeq strand specific library with the Ribo-Zero protocol (Illumina), and all samples were sequenced on a HiSeq2000v3 to a minimal target of 180 million paired-end reads. Reads were mapped using the STAR aligner (v2.5.3a) (Dobin et al., 2013) to GRCh17 with Gencode v24lift37 (Harrow et al., 2012). RSEM (v1.2.29) was used to quantify gene abundance (Li and Dewey, 2011).

mRNA microarray data generation—Total RNA was extracted using the mirVana miRNA Isolation Kit (Life Technologies), according to the manufacturer's instructions and assayed on Affymetrix transcriptome arrays as previously described (Fraser et al., 2017). All mRNA analysis was performed using R (v3.2.1). Background correction, normalization algorithms and annotation were implemented in the oligo (v1.32.0) package (Carvalho and Irizarry, 2010) from the BioConductor (v3.0) open-source project. The Robust multichip average (RMA) algorithm was applied to the raw intensity data. Annotations were performed using hugene20sttranscriptcluster.db (v2.13.0) and hta20sttranscriptcluster.db (v8.3.1). The sva package (v3.14.0) was used to correct for batch effects between different arrays. Annotated data from HuGene 2.0 ST and HTA 2.0 were combined into one data set based on Entrez Gene IDs. The mRNA abundance values were averaged amongst duplicated Entrez Gene IDs.

SNP microarray data generation and CNA calling—SNP microarrays were performed with 200 ng of DNA on Affymetrix OncoScan FFPE Express 3.0 arrays as previously described (Fraser et al., 2017). BioDiscovery's Nexus Express™ for OncoScan 3 Software was used to call CNAs using the SNP-FASST2 algorithm with default parameters except that the minimum number of probes per segment was changed from 3 to 20. When necessary, samples were re-centred using the Nexus Express™ software, choosing regions that showed diploid $\log_2$ ratios and B allele frequency profiles. Gene level CNAs for each patient were identified by overlapping copy number segments, with RefGene (2014-07-15) annotation, using BEDTools (v2.17.0) (Quinlan and Hall, 2010). To account for technical noise, a gene level CNV blacklist was created from matched normal blood samples. Genes were added to the blacklist if they were seen in at least 75% of normal samples and filtered from downstream analyses. Percent genome altered (PGA) was calculated for each sample by dividing the number of base pairs that were involved in all copy number segments by the total length of the genome.

Somatic variant calling—Single nucleotide variants (SNVs) and genomic rearrangements (GRs) were called using pipelines that have been described in detail elsewhere (Fraser et al., 2017). Briefly, lane-level WGS reads for blood normal and tumor samples were aligned against human reference build hg19 with BWA (v0.5.7) (Li and Durbin, 2009) before being merged. SNVs were called using SomaticSniper (v1.0.2) (Larson et al., 2012) and annotated using ANNOVAR (v2015-06-17) (Wang et al., 2010) with the RefGene database. Somatic GRs were called using Delly (v0.5.5) (Rausch et al., 2012) and filtered for mapping quality (>20) or pair-end evidence (>4 reads) before being filtered against their corresponding normal sample and a consolidated set of normal calls. Kataegis was called using the SeqKat (v0.0.1) (Yousif et al., 2018) R package. Chromothriptic regions were identified using Shatterproof (v0.14) (Govind et al., 2014) with default settings.

Methylation microarray data generation—Illumina Infinium HumanMethylation 450k BeadChip kits were used to assess global methylation, using 500 ng of input genomic DNA, at McGill University and the Genome Quebec Innovation Centre (Montreal, QC). All samples used in this study (n = 54) were processed from fresh-frozen prostate cancer tissue and can be found on GEO under the accession GSE107298. Methylation pre-processing were performed in R statistical environment (v3.4.0). The IDAT files were loaded and converted to raw intensity values with the use of wateRmelon package (v1.15.1) (Pidsley et al., 2013). Quality control was conducted using the minfi package (v1.22.1) (Aryee et al., 2014). No outlier samples were detected. Raw methylation intensity levels were then pre-processed using Dasen (Pidsley et al., 2013). Probe filtering was conducted after the normalization, as previously described (Fraser et al., 2017). Annotation to chromosome location, probe position, and gene symbol was conducted using the IlluminaHumanMethylation450kanno.ilmn12.hg19 package (v0.6.0).

Epigenetic data annotation—H3K27Ac peaks were annotated to the closest genes using the *annotatePeak* function from the ChIPseeker (v1.12.1) (Yu et al., 2015) R package. The *tssRegion* was set as: - 5000 to +5000) as proximal promoters can be up to 5 kbp away from transcription start site (Woo and Li, 2012).

Rather than using the full complement of CpG methylation sites, the probes with the greatest negative correlation to their corresponding mRNA abundance from the TCGA prostate cancer methylation dataset (Broad Institute TCGA Genome Data Analysis Center, 2016) were used, irrespective of their proximity to the gene. If there were no correlated probes associated with the gene, but the gene had annotated probes, a probe with the greatest variance was selected with the following priority: proximity to transcription start site, 5' UTR, 3' UTR and gene body. If a probe was not annotated to the gene, we retrieved the closest probe within 10 kbp. Genes that were not assigned a probe were denoted as missing (NA). Gene names from each biomolecule type were intersected to identify genes present in all data types.

Consensus clustering of proteomic data—Consensus clustering (*max_k* = 20; Spearman's ρ as the similarity metric; *pItem* = 0.8, *pFeature* = 0.8; *seed* = 17; *reps* = 1000; ConsensusClusterPlus v1.38.0) (Wilkerson and Hayes, 2010) was performed using a divisive algorithm on the 25% most variable proteins to cluster the patients and the proteins. Adjusted Rand Index (ARI) was calculated on patient classification using the protein subtypes and subtypes defined by copy number aberrations (Fraser et al., 2017) to determine if there is an overlap. Associations between patient subgroups and mutation burden were performed using a Mann-Whitney U test. Survival analysis was performed on the protein subtypes with the R package Survival (v2.40–3) by fitting a Cox PH model between patients in subtypes two, three, four and five against subtype one as the baseline, which was the largest group with BCR as the end point. Proportional hazards assumptions were evaluated using the *cox.zph* function ($p < 0.1$). Androgen receptor signature scores were created by identifying the top 100 genes that are positively correlated to *AR* (Spearman's ρ), converting their abundances to z-score before taking the mean. For the signature from the literature, the abundances of the genes used in the signature were retrieved, converted to z-scores, and then averaged. ANOVA was used to test for an association between the subtypes and the scores.

Clinical associations and ETS analysis—To identify proteins that may be associated with clinical features, univariate association tests (Spearman's ρ for continuous values, Mann-Whitney U test for binary values) were performed with each protein group against the following clinical covariates: percent genome altered (PGA), ETS gene fusion status, clinical T-category, presence of intraductal carcinoma or cribriform architecture, biochemical recurrence (BCR), age at treatment, pre-treatment prostate specific antigen levels, kataegis score, chromothripsis score and ISUP scores (dichotomized by ISUP 1 and 2 *vs*. ISUP 3). P values were adjusted for multiple comparisons using FDR. Mann-Whitney U tests were performed on mRNA abundances and ETS gene fusion status to identify mRNAs that are associated with the presence of ETS gene fusions and p values were corrected using FDR. Spearman's ρ was calculated on the difference in fold-change between mRNA and protein abundances in genes that were significantly associated with ETS gene fusion status at either the mRNA or protein level.

ETS fusion associated gene intersection—Mann-Whitney U tests were performed to identify H3K27Ac peaks, methylation and copy number aberrations associated with ETS gene fusions. P values were corrected using FDR. The VennDiagram (v1.6.19) (Chen and Boutros, 2011) package was used to identify the genes and number of genes found at all possible intersections amongst the genes significantly associated with ETS gene fusions in the protein, mRNA, methylation, H3K27 acetylation, and copy number data.

Pathway enrichment analysis—Gene sets of interest were processed using g:Profiler (Reimand et al., 2011) (v r1732_e88_eg35; significant only; query ordered by significance when applicable; the list of all proteins detected as the background; significance threshold set to FDR; output set to generic enrichment map; gene ontology and REACTOME databases), which was subsequently visualized in Cytoscape (v3.6.1) (Shannon et al., 2003) using the Enrichment Map App (Merico et al., 2010). For the ETS associated genes (at FDR < 0.05) g:Profiler was ran on all gene sets separately, but visualized in the same instance to better show potential overlaps in pathways.

Correlation analysis between mRNA and protein abundances—To determine the strength of the correlation between mRNA and protein abundances for each gene, overlapping genes (n = 6,946) were identified between the two data types using 55 matched samples. Correlation between the mRNA and protein abundance values for each of these gene was determined using Spearman's ρ.

Identification of proteins with undetected transcripts and *vice versa*—High abundance mRNA transcripts were identified by filtering out transcripts that appeared in two or fewer samples and then selecting transcripts with a median abundance in the top 10%. The high abundance transcripts were filtered against the detected proteins to identify RNAs without a protein counterpart. Similarly, to identify proteins without an RNA transcript, proteins were intersected with transcripts that have a median abundance of zero in the cohort. Pathway enrichment analysis was performed on both sets of genes using g:Profiler. To quantify number of coding and non-coding transcripts, the transcript sets were annotated using information from Gencode v24lift37 (Harrow et al., 2012).

Association analyses of CNAs on mRNA and protein abundances—FDR adjusted p values from a two-sided t-test and fold changes were calculated for each gene (n = 6,607, microarray) by CNA locus (n = 23,068) testing the difference in mean mRNA abundance between samples with a copy number aberration against those without (n = 210 samples). CNA status was quantified from OncoScan SNP arrays and mRNA abundance data was measured from Affymetrix transcriptome arrays. The same analysis was performed using protein abundance data (n = 55).

Mutual information analysis—For each pairwise combination of mRNA abundance, protein abundance, CNA state, methylation β value and H3K27Ac score, mutual information was calculated in bits for each gene using 21 bins, and the entropy function from the Entropy R package (v1.2.1) (Hausser and Strimmer, 2008). $I(X;Y) = H(X) + H(Y) - H(X,Y)$, where $H(X)$ and $H(Y)$ are the marginal entropies and $H(X,Y)$ is the joint entropy. MI was normalized over the mean entropy of the two input vectors. Consensus clustering was performed on the z-scored normalized MI using a maxK of 15, Spearman's correlation as the similarity metric and with 1,000 replicates. Hallmark enrichment analysis was performed using a hypergeometric test. P values were FDR-adjusted to control for multiple comparisons. Agreement of normalized MI between our dataset and TCGA was assessed using the area under the receiver operating characteristic curve. For each gene, normalized MI was binarized within each dataset based on whether it was above or below a threshold. True positive and false positive rates were calculated using whether normalized MI was greater than the threshold in our dataset as a True Positive.

Percent variance analysis—Sum of squares were extracted from an ANOVA on a linear model of protein abundances to CNA, RNA, and methylation R for the following genes: *TGM2*, *NDRG3*, *KLK3*, *AKT1*, *PTEN*, *NKX3–1*, *KRAS* and *ATM*. Percent variance explained was calculated as the sum of squares for each of the input variables divided by the total sum of squares. For *KLK3* and *PTEN*, Spearman ρ was used to determine if protein abundances are correlated with their corresponding mRNA abundances and methylation. For

PTEN, a Mann-Whitney U test was used to determine if the CNAs are associated with protein abundances.

Univariate survival analysis—Survival analysis was performed on the top 25% of proteins whose abundance had the highest variance, and their mRNA counterparts, to determine if the protein abundance or the mRNA abundance of each gene were univariately associated with biochemical recurrence. Hazard ratios were calculated by fitting Cox PH models to patient groups dichotomized using the median of the protein intensity or mRNA abundance against BCR as the endpoint. Assumptions for the Cox PH models were tested using the *cox.zph* function in the R Survival package (v2.41–3). Genes were considered to have divergent association with outcome if they were significantly associated with BCR, but $\log_2$ hazard ratios had opposite signs. For genes that failed the assumptions for a Cox PH model ($p < 0.1$), a log-rank test was used and p values were adjusted for multiple comparisons using FDR. Spearman's ρ was used to calculate the relationship of hazard ratios between mRNA and proteins. For proteins that were detected in 15% to 85% of the 75 samples (one sample was removed due to lack of clinical information), patients were dichotomized based on protein presence and absence before being fitted with a Cox PH model to determine if presence or loss of that protein group was associated with biochemical recurrence. P values were adjusted for multiple comparisons using FDR. Kaplan-Meier curves were generated for specific genes of interest. Validation of prognostic proteins was performed through immunohistochemistry on 79 additional prostate cancer samples in tissue microarrays. Antibodies for ACAD8 were obtained from Sigma-Aldrich (Prestige Antibodies - HPA040689 for ACAD8; Lot #A114184). A pathologist (M.M.) scored each core in comparison with the internal positive control using a semi-quantitative scoring system based on the intensity of the staining (cytoplasmic for ACAD8): 0 (No reactivity), 1 (mild intensity), 2 (moderate) and 3 (high intensity, equivalent to positive control cells). Staining intensity was recorded per TMA core. A Cox PH model for the validation set was fit to patients' groups, dichotomized by whether they have any cores scored with a '3 - high intensity staining'.

Biomarker null distribution analysis—To assess the performance of biomolecules at predicting patients with BCR at 10 years, a null distribution analysis was performed using 10 million areas under the receiver-operating characteristic curves (AUC) for each biomolecule (40 million AUCs total). To calculate each AUC, 100 genes were randomly selected without replacement from the intersection of the genes present in each biomolecule data matrix. A random forest classification model with 4-fold cross validation (randomForest v4.6–12) (Svetnik et al., 2003) was then built for each gene set in each biomolecule. The hyper-parameters *mtry* and *sampsize* were tuned through a grid search based on lowest out-of-bag errors while *nTrees* was set to 10,000 to reduce grid search time since having too few trees will negatively impact model performance, but having more trees only incurs more computational time (Huang and Boutros, 2016). For protein, mRNA and CNA, there were 7,042 matched protein-groups to genes. Methylation data was set to one probe per gene as described above. For both the RNA-Seq and protein, random gene sets of 5, 10, 25, 50 and 100 were used for the random forest model to determine if gene set numbers will change the conclusions, whereas the CNA and methylation null distributions was generated using a gene set size of 100. To evaluate the predictive power of using two biomolecules at the same time, 100 genes were randomly selected and values for both biomolecules were used as features in the same random forest model.

QUANTIFICATION AND STATISTICAL ANALYIS

The specific statistical tests used are indicated in the figure legends or appropriate methods section and were performed within the R statistical environment (v3.3.1).

Visualization in R was performed through the BoutrosLab.Plotting.General package (v5.9.2) (P'ng et al., 2019). Pathway network graphs were generated using Cytoscape (v3.6.1) with the Enrichment Map App (Merico et al., 2010). Study outline was produced with Inkscape (v0.48) for Ubuntu.

DATA AND SOFTWARE AVAILABILITY

MS data was deposited in UCSD's MASSive database under the accession MSV000081552 at ftp://massive.ucsd.edu/MSV000081552. Oncoscan CNA microarray data can be found in the European Genome-Phenome Archive (EGA) at https://www.ebi.ac.uk/ega/studies/EGAS00001000900. Whole genome DNA sequencing and RNA-Seq data can also be found on EGA, under accession EGAS00001000900. H3K27Ac CHiP-Seq data are in the Gene Expression Omnibus under GSE96652. RNA microarray data is available under the accession GSE107299. Methylation data is available under the accession GSE107298.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors thank all members of the Boutros and Kislinger labs for helpful suggestions. PCB and TK were funded through a Movember Discovery grant from Prostate Cancer Canada (D2013–21), an operating grant from the National Cancer Institute Early Detection Research Network (1U01CA214194–01), a CIHR Project Grant (PJT 156357) and a Collaborative Personalized Cancer Medicine Team Grant from the Princess Margaret Cancer Centre. P.C.B. was supported by a Terry Fox Research Institute New Investigator Award, a CIHR New Investigator Award and a CIHR Project Grant. A.S. was supported by a CIHR Banting and Best doctoral research scholarship. This research was funded in part by the Ontario Ministry of Health and Long Term Care. This work was supported by the Discovery Frontiers: Advancing Big Data Science in Genomics Research program, which is jointly funded by the Natural Sciences and Engineering Research Council (NSERC) of Canada, the Canadian Institutes of Health Research (CIHR), Genome Canada, and the Canada Foundation for Innovation (CFI). This work was supported by Prostate Cancer Canada and is proudly funded by the Movember Foundation - Grant #RS2014–01. This work was funded by the Government of Canada through Genome Canada and the Ontario Genomics Institute (OGI-125). This research is funded by the Canadian Cancer Society (grant #705649).

References

Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, and Irizarry RA (2014). Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics 30, 1363–1369. [PubMed: 24478339]

Baca SC, Prandi D, Lawrence MS, Mosquera JM, Romanel A, Drier Y, Park K, Kitabayashi N, MacDonald TY, Ghandi M, et al. (2013). Punctuated evolution of prostate cancer genomes. Cell 153, 666–677. [PubMed: 23622249]

Bhandari V, Hoey C, Liu LY, Lalonde E, Ray J, Livingstone J, Lesurf R, Shiah Y-J, Vujcic T, Huang X, et al. (2019). Molecular landmarks of tumor hypoxia across cancer types. Nat. Genet doi:10.1038/s41588-018-0318-2

Blume-Jensen P, Berman DM, Rimm DL, Shipitsin M, Putzi M, Nifong TP, Small C, Choudhury S, Capela T, Coupal L, et al. (2015). Development and Clinical Validation of an In Situ Biopsy-Based Multimarker Assay for Risk Stratification in Prostate Cancer. Clin. Cancer Res. 21, 2591–2600. [PubMed: 25733599]

Bose R, Karthaus WR, Armenia J, Abida W, Iaquinta PJ, Zhang Z, Wongvipat J, Wasmuth EV, Shah N, Sullivan PS, et al. (2017). ERF mutations reveal a balance of ETS factors controlling prostate oncogenesis. Nature 546, 671–675. [PubMed: 28614298]

Boutros PC, Lau SK, Pintilie M, Liu N, Shepherd FA, Der SD, Tsao M-S, Penn LZ, and Jurisica I (2009). Prognostic gene signatures for non-small-cell lung cancer. Proc. Natl. Acad. Sci. U. S. A 106, 2824–2828. [PubMed: 19196983]

Boutros PC, Fraser M, Harding NJ, de Borja R, Trudel D, Lalonde E, Meng A, Hennings-Yeomans PH, McPherson A, Sabelnykova VY, et al. (2015). Spatial genomic heterogeneity within localized, multifocal prostate cancer. Nat. Genet 47, 736–745. [PubMed: 26005866]

Broad Institute TCGA Genome Data Analysis Center (2016). Correlation between mRNA expression and DNA methylation. Broad Inst. MIT Harvard.

Brocks D, Assenov Y, Minner S, Bogatyrova O, Simon R, Koop C, Oakes C, Zucknick M, Lipka D, Weischenfeldt J, et al. (2014). Intratumor DNA Methylation Heterogeneity Reflects Clonal Evolution in Aggressive Prostate Cancer. Cell Rep. 8, 798–806. [PubMed: 25066126]

Canadian Cancer Statistics Advisory Committee (2017). Canadian Cancer Statistics.

Carvalho BS, and Irizarry RA (2010). A framework for oligonucleotide microarray preprocessing. Bioinformatics 26, 2363–2367. [PubMed: 20688976]

Center MM, Jemal A, Lortet-Tieulent J, Ward E, Ferlay J, Brawley O, and Bray F (2012). International Variation in Prostate Cancer Incidence and Mortality Rates. Eur. Urol 61, 1079–1092. [PubMed: 22424666]

Chen H, and Boutros PC (2011). VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. BMC Bioinformatics 12, 35. [PubMed: 21269502]

Chua MLK, Lo W, Pintilie M, Murgic J, Lalonde E, Bhandari V, Mahamud O, Gopalan A, Kweldam CF, van Leenders GJLH, et al. (2017). A Prostate Cancer “Nimbosus”: Genomic Instability and SChLAP1 Dysregulation Underpin Aggression of Intraductal and Cribriform Subpathologies. Eur. Urol 71, 183–192. [PubMed: 27451135]

Cooper CS, Eeles R, Wedge DC, Van Loo P, Gundem G, Alexandrov LB, Kremeyer B, Butler A, Lynch AG, Camacho N, et al. (2015). Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. Nat. Genet 47, 367–372. [PubMed: 25730763]

Cooperberg MR, Broering JM, and Carroll PR (2009). Risk assessment for prostate cancer metastasis and mortality at the time of diagnosis. J. Natl. Cancer Inst. 101, 878–887. [PubMed: 19509351]

Cox J, and Mann M (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol 26, 1367–1372. [PubMed: 19029910]

Cox J, Hein MY, Luber CA, Paron I, Nagaraj N, and Mann M (2014). Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. Mol. Cell. Proteomics 13, 2513–2526. [PubMed: 24942700]

Cuzick J, Swanson GP, Fisher G, Brothman AR, Berney DM, Reid JE, Mesher D, Speights V, Stankiewicz E, Foster CS, et al. (2011). 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. [PubMed: 21310658]

Dal Pra A, Lalonde E, Sykes J, Warde F, Ishkanian A, Meng A, Maloff C, Srigley J, Joshua AM, Petrovics G, et al. (2013). TMPRSS2-ERG Status Is Not Prognostic Following Prostate Cancer Radiotherapy: Implications for Fusion Status and DSB Repair. Clin. Cancer Res. 19, 5202–5209. [PubMed: 23918607]

Den RB, Yousefi K, Trabulsi EJ, Abdollah F, Choeurng V, Feng FY, Dicker AP, Lallas CD, Gomella LG, Davicioni E, et al. (2015). Genomic classifier identifies men with adverse pathology after radical prostatectomy who benefit from adjuvant radiation therapy. J. Clin. Oncol 33, 944–951. [PubMed: 25667284]

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, and Gingeras TR (2013). STAR: Ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21. [PubMed: 23104886]

Drake JM, Paull EO, Graham NA, Lee JK, Smith BA, Titz B, Stoyanova T, Faltermeier CM, Uzunangelov V, Carlin DE, et al. (2016). Phosphoproteome Integration Reveals Patient-Specific Networks in Prostate Cancer. Cell 166, 1041–1054. [PubMed: 27499020]

Espiritu SMG, Liu LY, Rubanova Y, Bhandari V, Holgersen EM, Szyca LM, Fox NS, Chua MLK, Yamaguchi TN, Heisler LE, et al. (2018). The Evolutionary Landscape of Localized Prostate Cancers Drives Clinical Aggression. Cell.

Fraser M, Berlin A, Bristow RG, and van der Kwast T (2015). Genomic, pathological, and clinical heterogeneity as drivers of personalized medicine in prostate cancer. Urol. Oncol. Semin. Orig. Investig 33, 85–94.

Fraser M, Sabelnykova VY, Yamaguchi TN, Heisler LE, Livingstone J, Huang V, Shiah Y-J, Yousif F, Lin X, Masella AP, et al. (2017). Genomic hallmarks of localized, non-indolent prostate cancer. Nature 541, 359–364. [PubMed: 28068672]

Govind SK, Zia A, Hennings-Yeomans PH, Watson JD, Fraser M, Anghel C, Wyatt AW, van der Kwast T, Collins CC, McPherson JD, et al. (2014). ShatterProof: operational detection and quantification of chromothripsis. BMC Bioinformatics 15, 78. [PubMed: 24646301]

Gundem G, Van Loo P, Kremeyer B, Alexandrov LB, Tubio JMC, Papaemmanuil E, Brewer DS, Kallio HML, Högnäs G, Annala M, et al. (2015). The evolutionary history of lethal metastatic prostate cancer. Nature 520, 353–357. [PubMed: 25830880]

Hanahan D, and Weinberg RA (2011). Hallmarks of Cancer: The Next Generation. Cell 144, 646–674. [PubMed: 21376230]

Hansen AF, Sandsmark E, Rye MB, Wright AJ, Bertilsson H, Richardsen E, Viset T, Bofin AM, Angelsen A, Selnaes KM, et al. (2016). Presence of TMPRSS2-ERG is associated with alterations of the metabolic profile in human prostate cancer. Oncotarget 7, 42071–42085. [PubMed: 27276682]

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 22, 1760–1774. [PubMed: 22955987]

Hausser J, and Strimmer K (2008). Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. ArXiv.

Hieronymus H, Schultz N, Gopalan A, Carver BS, Chang MT, Xiao Y, Heguy A, Huberman K, Bernstein M, Assel M, et al. (2014). Copy number alteration burden predicts prostate cancer relapse. Proc. Natl. Acad. Sci 111, 11139–11144. [PubMed: 25024180]

Hopkins JFJ, Sabelnykova VYV, Weischenfeldt J, Simon R, Aguiar JAJ, Alkallas R, Heisler LELE, Zhang J, Watson JDJ, Chua MLKM, et al. (2017). Mitochondrial mutations drive prostate cancer aggression. Nat. Commun 8, 656. [PubMed: 28939825]

Huang BFF, and Boutros PC (2016). The parameter sensitivity of random forests. BMC Bioinformatics 17, 331. [PubMed: 27586051]

Iglesias-Gato D, Wikström P, Tyanova S, Lavallee C, Thysell E, Carlsson J, Hägglöf C, Cox J, Andrén O, Stattin P, et al. (2016). The Proteome of Primary Prostate Cancer. Eur. Urol 69, 942–952. [PubMed: 26651926]

Iglesias-Gato D, Thysell E, Tyanova S, Crnalic S, Santos A, Lima TS, Geiger T, Cox J, Widmark A, Bergh A, et al. (2018). The Proteome of Prostate Cancer Bone Metastasis Reveals Heterogeneity with Prognostic Implications. Clin. Cancer Res. 24, 5433–5444. [PubMed: 30042207]

Johnson WE, Li C, and Rabinovic A (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8, 118–127. [PubMed: 16632515]

Kislinger T, Cox B, Kannan A, Chung C, Hu P, Ignatchenko A, Scott MS, Gramolini AO, Morris Q, Hallett MT, et al. (2006). Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling. Cell 125, 173–186. [PubMed: 16615898]

Klein EA, Cooperberg MR, Magi-Galluzzi C, Simko JP, Falzarano SM, Maddala T, Chan JM, Li J, Cowan JE, Tsiatis AC, et al. (2014). 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. [PubMed: 24836057]

Kron KJ, Murison A, Zhou S, Huang V, Yamaguchi TN, Shiah Y-J, Fraser M, van der Kwast T, Boutros PC, Bristow RG, et al. (2017). TMPRSS2–ERG fusion co-opts master transcription factors and activates NOTCH signaling in primary prostate cancer. Nat. Genet 49, 1336–1345. [PubMed: 28783165]

Lalonde E, Ishkanian AS, Sykes J, Fraser M, Ross-Adams H, Erho N, Dunning MJ, Halim S, Lamb AD, Moon NC, et al. (2014). Tumour genomic and microenvironmental heterogeneity for integrated prediction of 5-year biochemical recurrence of prostate cancer: a retrospective cohort study. Lancet. Oncol 15, 1521–1532. [PubMed: 25456371]

Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, Ley TJ, Mardis ER, Wilson RK, and Ding L (2012). SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics 28, 311–317. [PubMed: 22155872]

Latonen L, Afyounian E, Jylhä A, Nättinen J, Aapola U, Annala M, Kivinummi KK, Tammela TTL, Beuerman RW, Uusitalo H, et al. (2018). Integrative proteomics in prostate cancer uncovers robustness against genomic and transcriptomic aberrations during disease progression. Nat. Commun 9, 1176. [PubMed: 29563510]

Leek JT, Johnson WE, Parker HS, Jaffe AE, and Storey JD (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics 28, 882–883. [PubMed: 22257669]

Li B, and Dewey CN (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12, 323. [PubMed: 21816040]

Li H, and Durbin R (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760. [PubMed: 19451168]

Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, and Tamayo P (2015). The Molecular Signatures Database Hallmark Gene Set Collection. Cell Syst. 1, 417–425. [PubMed: 26771021]

Liu H, Sadygov RG, and Yates JR (2004). A Model for Random Sampling and Estimation of Relative Protein Abundance in Shotgun Proteomics. Anal. Chem 76, 4193–4201. [PubMed: 15253663]

Merico D, Isserlin R, Stueker O, Emili A, and Bader GD (2010). Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. PLoS One 5, e13984. [PubMed: 21085593]

Mertins P, Mani DR, Ruggles KV, Gillette MA, Clauser KR, Wang P, Wang X, Qiao JW, Cao S, Petralia F, et al. (2016). Proteogenomics connects somatic mutations to signalling in breast cancer. Nature 534, 55–62. [PubMed: 27251275]

Michalski A, Damoc E, Hauschild J-P, Lange O, Wieghaus A, Makarov A, Nagaraj N, Cox J, Mann M, and Horning S (2011). Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. Mol. Cell. Proteomics 10, M111.011015.

Minner S, Enodien M, Sirma H, Luebke AM, Krohn A, Mayer PS, Simon R, Tennstedt P, Müller J, Scholz L, et al. (2011). ERG Status Is Unrelated to PSA Recurrence in Radically Operated Prostate Cancer in the Absence of Antihormonal Therapy. Clin. Cancer Res. 17, 5878–5888. [PubMed: 21791629]

P'ng C, Green J, Chong LC, Waggott D, Prokopec SD, Shamsi M, Nguyen F, Mak DYF, Lam F, Albuquerque MA, et al. (2019). BPG: Seamless, automated and interactive visualization of scientific data. BMC Bioinformatics 20, 42. [PubMed: 30665349]

Pidsley R, Y Wong CC, Volta M, Lunnon K, Mill J, and Schalkwyk LC (2013). A data-driven approach to preprocessing Illumina 450K methylation array data. BMC Genomics 14, 293. [PubMed: 23631413]

Pritchard CC, Mateo J, Walsh MF, De Sarkar N, Abida W, Beltran H, Garofalo A, Gulati R, Carreira S, Eeles R, et al. (2016). Inherited DNA-Repair Gene Mutations in Men with Metastatic Prostate Cancer. N. Engl. J. Med 375, 443–453. [PubMed: 27433846]

Quinlan AR, and Hall IM (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842. [PubMed: 20110278]

Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, and Korbel JO (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28, i333–i339. [PubMed: 22962449]

Reimand J, Arak T, and Vilo J (2011). G:Profiler - A web server for functional interpretation of gene lists (2011 update). Nucleic Acids Res. 39, 307–315.

Ribeiro FR, Paulo P, Costa VL, Barros-Silva JD, Ramalho-Carvalho J, Jerónimo C, Henrique R, Lind GE, Skotheim RI, Lothe RA, et al. (2011). Cysteine-Rich Secretory Protein-3 (CRISP3) Is Strongly Up-Regulated in Prostate Carcinomas with the TMPRSS2-ERG Fusion Gene. PLoS One 6, e22317. [PubMed: 21814574]

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, and Smyth GK (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43, e47–e47. [PubMed: 25605792]

Sallari RC, Sinnott-Armstrong NA, French JD, Kron KJ, Ho J, Moore JH, Stambolic V, Edwards SL, Lupien M, and Kellis M (2016). Convergence of dispersed regulatory mutations predicts driver genes in prostate cancer. BioRxiv 097451.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, and Ideker T (2003). Cytoscape: A software Environment for integrated models of biomolecular interaction networks. Genome Res. 13, 2498–2504. [PubMed: 14597658]

Shao W, Guo T, Toussaint NC, Wagner U, Li L, Charmpi K, Zhu Y, Beyer A, Ratsch G, Wild P, et al. (2017). Comparative analysis of mRNA degradation and protein degradation in 68 pairs of adjacent prostate tumor tissue samples In Proceedings of the 65th ASMS Conference on Mass Spectrometry and Allied Topics, (Indianapolis, IN), p. MP 114.

Sinha A, Ignatchenko V, Ignatchenko A, Mejia-Guerrero S, and Kislinger T (2014). In-depth proteomic analyses of ovarian cancer cell line exosomes reveals differential enrichment of functional categories compared to the NCI 60 proteome. Biochem. Biophys. Res. Commun 445, 694–701. [PubMed: 24434149]

Song S, Nones K, Miller D, Harliwong I, Kassahn KS, Pinese M, Pajic M, Gill AJ, Johns AL, Anderson M, et al. (2012). qpure: A Tool to Estimate Tumor Cellularity from Genome-Wide Single-Nucleotide Polymorphism Profiles. PLoS One 7, e45835. [PubMed: 23049875]

Stelloo S, Nevedomskaya E, van der Poel HG, de Jong J, van Leenders GJLH, Jenster G, Wessels LFA, Bergman AM, and Zwart W (2015). Androgen receptor profiling predicts prostate cancer outcome. EMBO Mol. Med 7, 1450–1464. [PubMed: 26412853]

Svetnik V, Liaw A, Tong C, Christopher Culberson J, Sheridan RP, and Feuston BP (2003). Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. J. Chem. Inf. Comput. Sci 43, 1947–1958. [PubMed: 14632445]

The Cancer Genome Atlas Research Network (2015). The Molecular Taxonomy of Primary Prostate Cancer. Cell 163, 1011–1025. [PubMed: 26544944]

Torre LA, Siegel RL, Ward EM, and Jemal A (2016). Global Cancer Incidence and Mortality Rates and Trends—An Update. Cancer Epidemiol. Prev. Biomarkers 25, 16–27.

Wang K, Li M, and Hakonarson H (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38, e164–e164. [PubMed: 20601685]

Wilkerson MD, and Hayes DN (2010). ConsensusClusterPlus: A class discovery tool with confidence assessments and item tracking. Bioinformatics 26, 1572–1573. [PubMed: 20427518]

Woo YH, and Li W-H (2012). Evolutionary conservation of histone modifications in mammals. Mol. Biol. Evol 29, 1757–1767. [PubMed: 22319170]

Wu X, Daniels G, Lee P, and Monaco ME (2014). Lipid metabolism in prostate cancer. Am. J. Clin. Exp. Urol 2, 111–120. [PubMed: 25374912]

Yousif F, Prokopec S, Sun RX, Fan F, Lalansingh CM, Park DH, Szyca L, Network P, and Boutros PC (2018). The Origins and Consequences of Localized and Global Somatic Hypermutation. BioRxiv 287839.

Yu G, Wang L-G, and He Q-Y (2015). ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics 31, 2382–2383. [PubMed: 25765347]

Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z, Chambers MC, Zimmerman LJ, Shaddox KF, Kim S, et al. (2014). Proteogenomic characterization of human colon and rectal cancer. Nature 513, 382–387. [PubMed: 25043054]

Zhang H, Liu T, Zhang Z, Payne SH, Zhang B, McDermott JE, Zhou J-Y, Petyuk VA, Chen L, Ray D, et al. (2016). Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. Cell 166, 755–765. [PubMed: 27372738]

Highlights

- A comprehensive proteomic analyses of localized prostate cancers
- Integration of all levels of the central dogma (DNA-> RNA-> protein)
- ETS fusions have divergent effects on transcriptome and proteome
- Combining genomics and proteomics improves biomarker performance

Significance

Our data demonstrate that the prostate cancer proteome is shaped by the complex interplay of genomic, epigenomic, transcriptomic and post-transcriptional dysregulation. Integration of data along the central dogma enables both a deeper biological insight and the development of multi-omic biomarkers with improved performance.

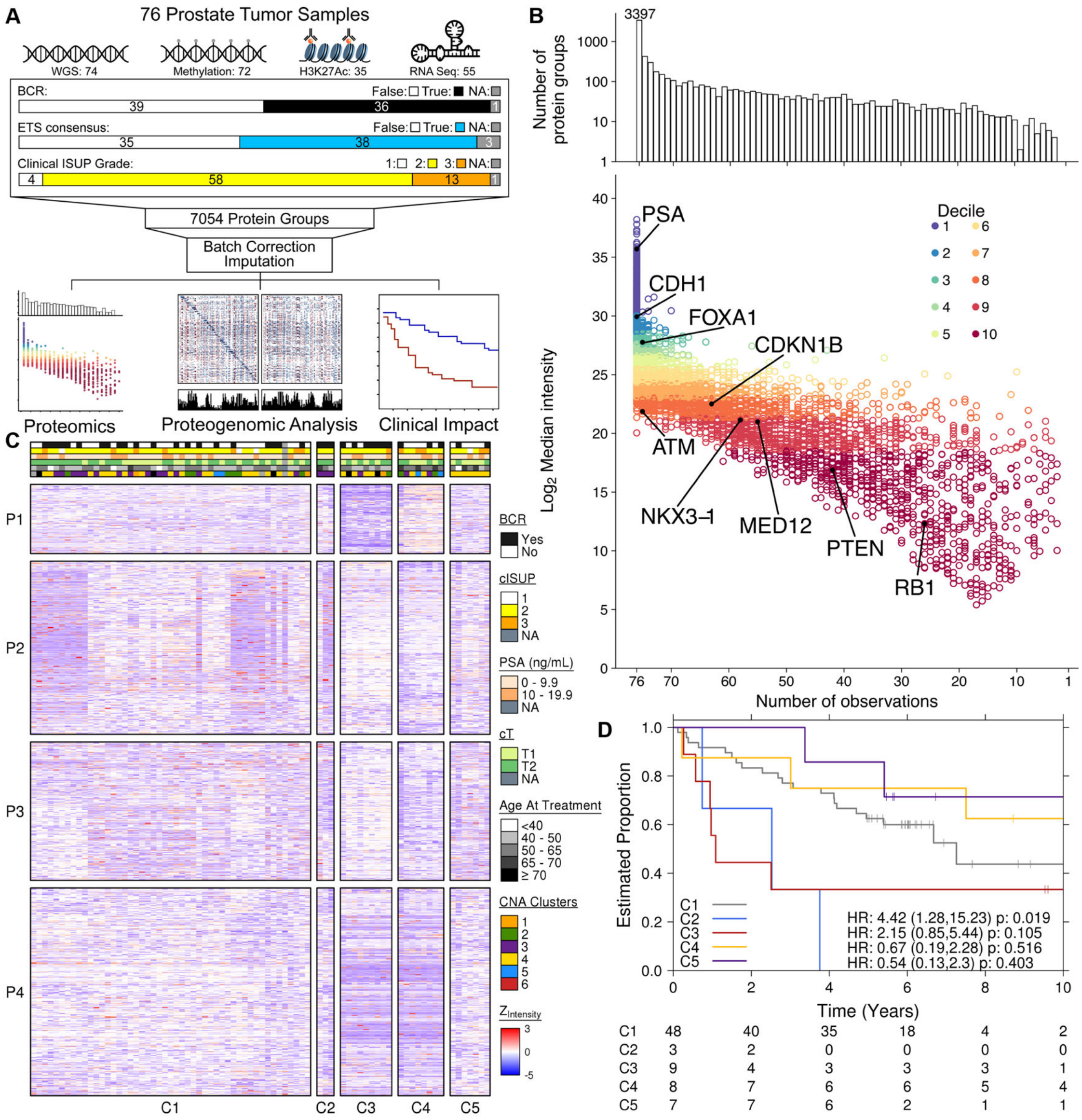


Figure 1. Proteomic landscape of curable prostate cancer

(A) Study overview showing the clinical characteristics of the cohort (n = 76) and the number of samples with whole genome sequencing, RNA-Sequencing, methylation data and CHiP-Seq data. Mass spectrometry yielded 7,054 protein groups, whose abundance was corrected for batch effects, and missing values were imputed prior to downstream analyses.

(B) Distribution of protein quantitation measured as median intensity by the number of samples they are detected in. Bar plot on top shows the total counts of proteins quantified in various number of samples. Missing values were omitted when calculating the median.

(C) Consensus clustering of 76 patients (K=5) using the top 25% most variable genes (n=1,800, K=5). Clinical covariates are shown in the heatmap above, indicating for each patient; biochemical relapse (BCR), clinical ISUP grade (cISUP), PSA levels, clinical T category (cT), and age at treatment (years).

(D) Subtypes identified from consensus clustering were evaluated to determine their association with BCR. A Cox PH model was fitted for subtype C2, C3, C4 and C5 against the baseline group of subtype C1. Hazard ratios and p values are shown with confidence intervals in parentheses.

Abbreviations: International Society for Urological Pathology (ISUP), prostate specific antigen (PSA), hazard ratio (HR)

See also Figure S1, Tables S1–S4, and Data S1.

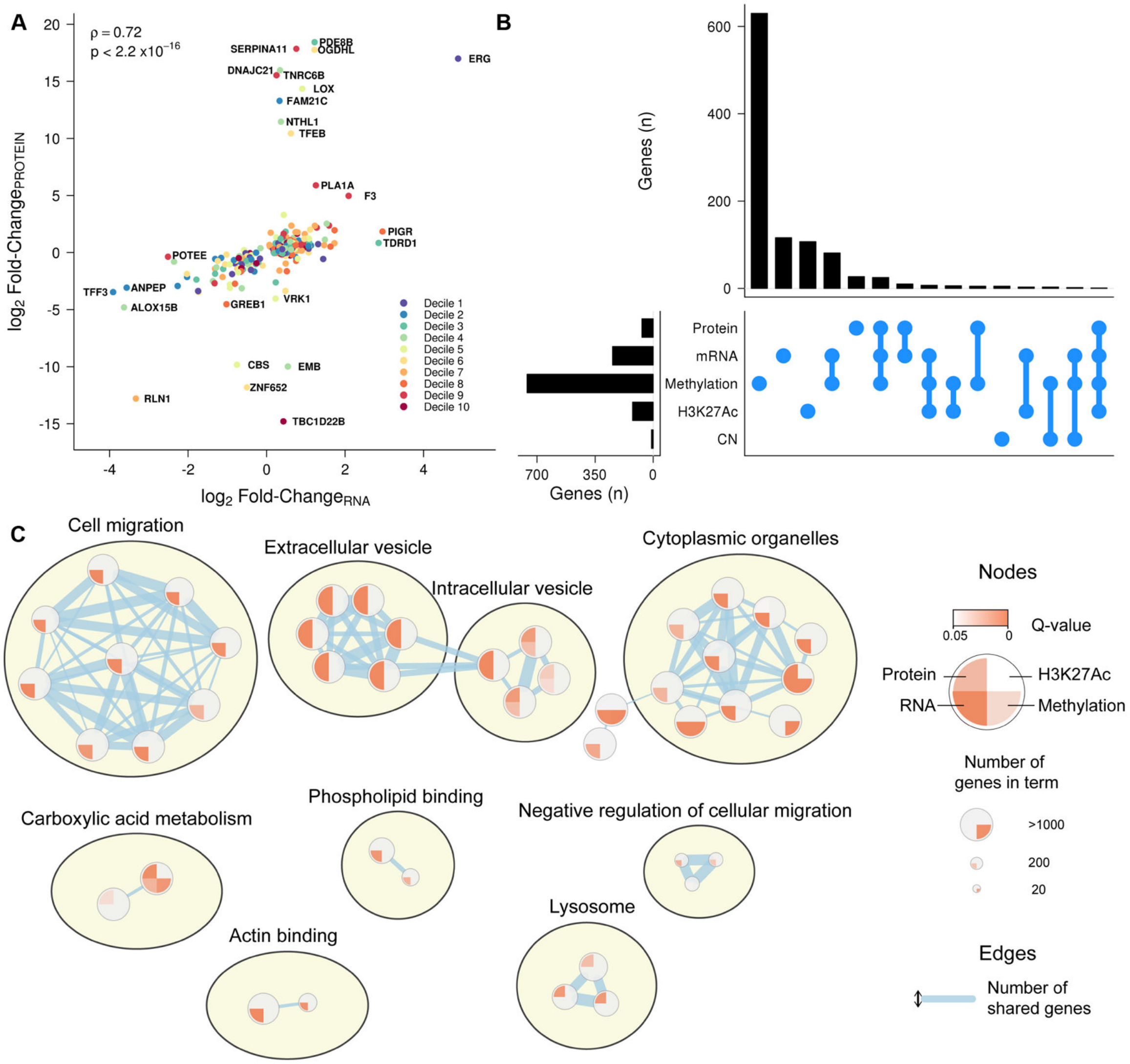


Figure 2. Transcriptomic and Proteomic Consequences of ETS fusions

(A) Comparison of the difference in protein and mRNA abundance observed between samples with an ETS gene fusion and those without. Analysis includes 55 samples with matched RNA-Seq and protein data in 255 genes as 22 genes were removed due to a high proportion of missing protein data. Color indicates which protein abundance decile the gene is in, where purple indicates the most abundant.

(B) Number of overlapping ETS gene fusion associated genes between protein, mRNA, methylation, H3K27Ac, and copy number status. Barplot on the left indicates the total number of associated genes in that data type. Barplot on top shows number of genes in the singleton or intersection groups as indicated by the dots below.

(C) Pathway enrichment analysis performed using g:Profiler on the five sets of genes associated with ETS gene fusions in the different data types. Large clusters of similar pathways are outlined in yellow and labeled. Singleton nodes were omitted. No pathway enrichment was detected in copy number changes associated with ETS gene fusions.

See also Figure S2.

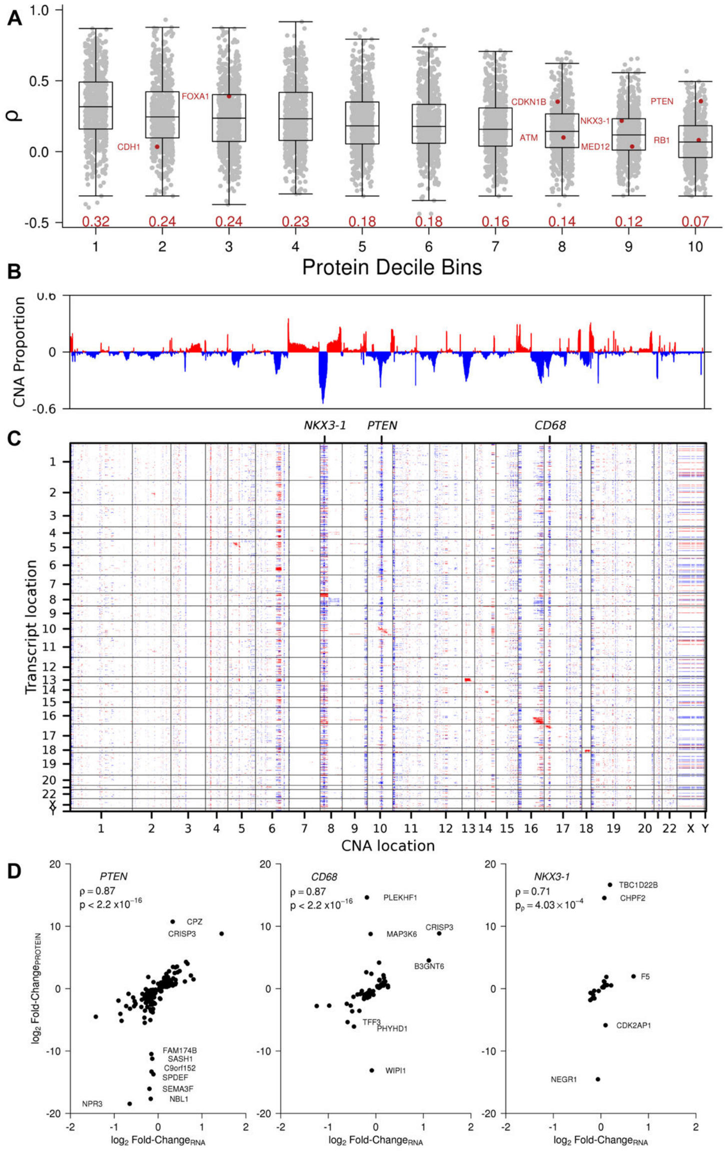


Figure 3. *Trans* proteomic effects of somatic CNAs

(A) Distribution of RNA-protein Spearman's ρ in each decile of protein abundance. Median correlations of each decile are indicated in red along the x-axis. Known genes of interest are highlighted and labeled in red. Boxplots depict the upper and lower quartiles, with the median shown as a solid line; whiskers indicate 1.5 times the interquartile range (IQR). Data points outside the IQR are shown.

(B) The proportion of samples that contain copy number amplifications (red) and deletions (blue) in 210 samples with mRNA data.

(C) The heatmap displays a global overview of the difference in mRNA abundance for each CNA locus comparing abundance from samples with a CNA to those without. Positive fold changes (i.e. higher abundance in samples with an amplification) are shown in red, negative fold changes (i.e. lower abundance in samples with a deletion) are shown in blue (FDR <

0.05). The x-axis plots 23,068 CNAs and the y-axis plots 6,636 mRNA genes. Genes are ordered by chromosome location on both axes.

(D) The fold change in mRNA and protein abundances in 55 matched samples (RNA Seq) comparing abundances in samples with a deletion and those without for *PTEN*, *CD68*, and *NKX3–1*. Only genes that show significant fold changes at the mRNA (Mann-Whitney U test; $p < 0.05$) and protein level (Mann-Whitney U test; $p < 0.05$) are plotted.

See also Figures S3–S4, and Table S5.

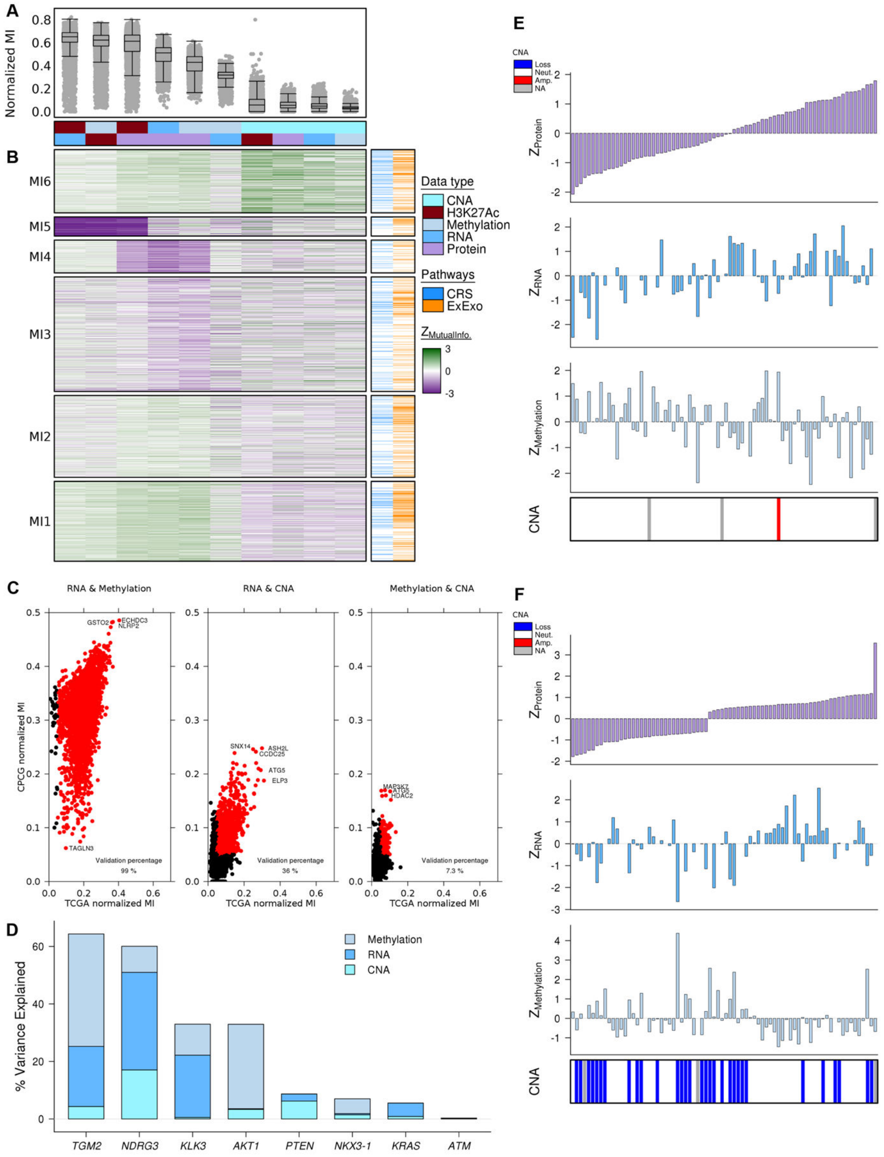


Figure 4. Integrated clustering of multi-omics data

(A) Distribution of the normalized mutual information (MI) for each data-type pair. Boxplots depict the upper and lower quartiles, with the median shown as a solid line; whiskers indicate 1.5 times the interquartile range (IQR). Data points outside the IQR are shown.

(B) Consensus clustering of normalized mutual information for each data-type pair. Biomolecules are indicated in the covariates along the top. Each row represents a gene (n = 6,484) comparison for which all data-types exist. Adjacent plots indicate if genes are known to be associated with the selected pathways. Normalized MI are plotted as z-scores for visualization purposes.

(C) Correlation of normalized mutual information between our cohort and TCGA in genes with MI above 0.05. Red dots indicate genes that had normalized MI about 0.05 in both our dataset and TCGA.

(D) Percent variance explained of protein abundance modeled using copy number status, methylation, and mRNA abundance for a select set of genes known to be associated with prostate cancer.

(E and F) Integrated distribution plots of *KLK3* (E) and *PTEN* (F) showing CN state and z-scored protein, mRNA, and methylation abundances for each of the 76 samples ordered by increasing protein abundance.

Abbreviations: Cellular Response to Stress (CRS), Extracellular Exosomes (ExExo)

See also Figure S4 and Table S6.

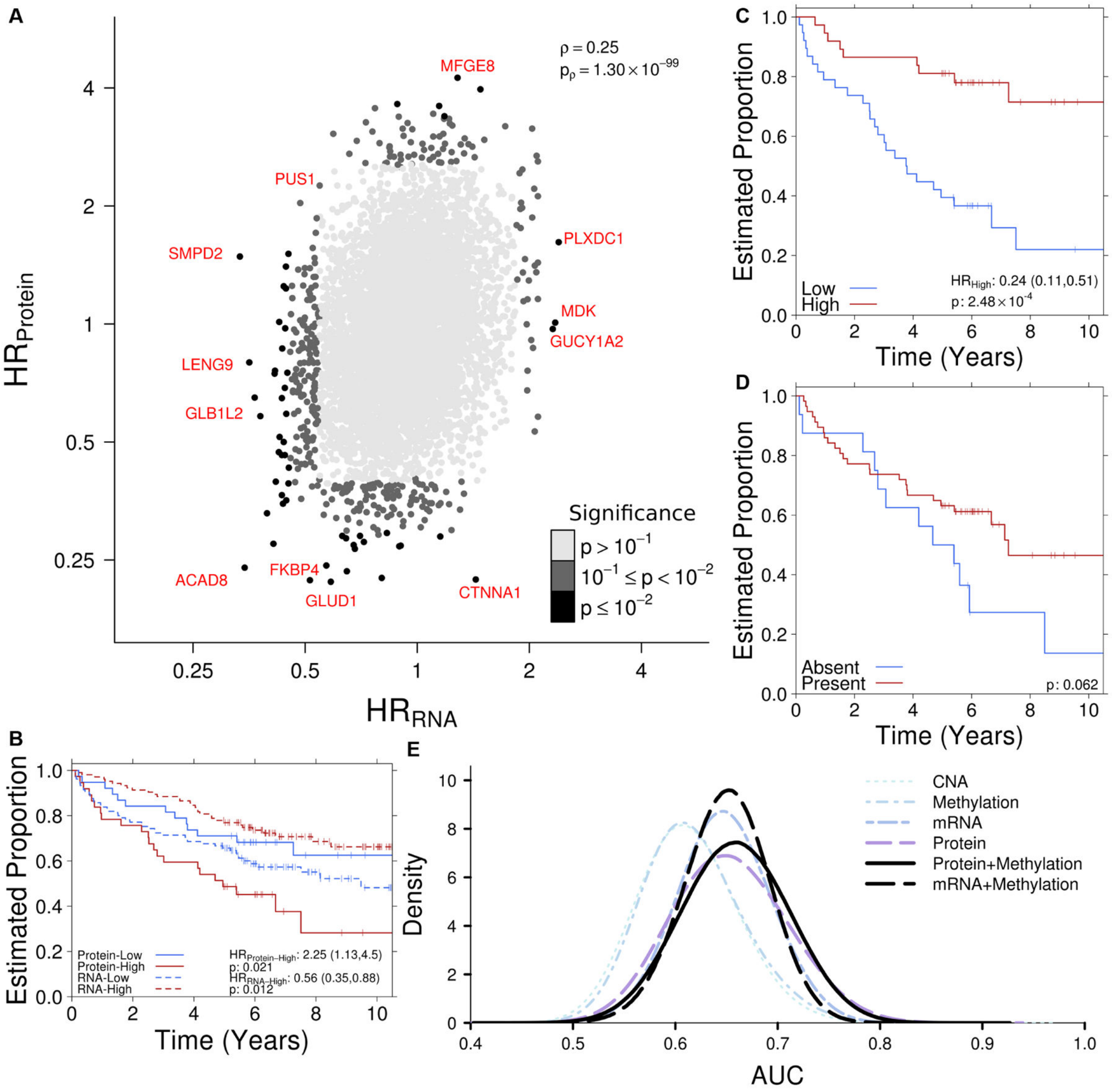


Figure 5. Protein abundance robustly predicts patient survival

(A) Hazard ratios were calculated using a Cox model on patient groups determined using median-dichotomized protein and RNA abundances. Shading of dots indicates statistical significance with selected genes labeled in red.

(B) Kaplan-Meier (KM) plot for PUS1 protein (solid lines) and mRNA (dashed lines). A Cox model was fit with patients stratified into high and low abundance of PUS1 protein (75 patients) and mRNA (209 patients).

(C) KM plot showing 10-year biochemical recurrence-free survival of patient groups as dichotomized by high and low protein abundance of ACAD8.

(D) KM plot for ACAD8 in 73 tissue microarrays. Three slides were evaluated per sample, and patients were grouped into 'low' protein abundance if at least two slides reported heterogeneous or faint staining. Significance of association was calculated using a log-rank test between high and low abundance patient groups.

(E) Null distribution of predictive accuracy for different biomolecules obtained from 10 million replicates of 100 randomly selected genes. For each replicate, a value for the area under the receiver-operator curve (AUC) was calculated using classification results from four-fold cross-validation in random forest.

See also Figure S5.

Table 1.

Clinical characteristics of patient cohort

	BCR	
	Yes (36)	**No (39)**
Clinical ISUP Group		
1	1	3
2	28	30
3	7	6
Age at treatment (years)		
40 – 50	2	2
50 – 65	24	25
65 – 70	8	7
≥ 70	2	5
Pre-Treatment PSA (ng/mL)		
< 10	27	27
≥ 10	9	12
Clinical T Category		
T1	12	21
T2	24	18
ETS-Fusion		
Present	21	17
Absent	15	22
Data Type		
WGS	36	38
CNA	36	37
H3K27AC	16	19
Methylation	34	38
RNA	25	30
Proteomics	36	39