Cancer genetics-guided discovery of serum biomarker signatures for diagnosis and prognosis of prostate cancer

Igor Cima1,3, Ralph Schiess1,2, Peter Wild2, Martin Kaelin3, Peter Schüffler4, Vinzenz Lange5, Paola Picotti6, Reto Ossola6, Arnoud Templeton4, Olga Schubert4, Thomas Fuchs7, Thomas Leippold7, Stephen Wyler1, Jens Zehetner3, Wolfram Jochem8, Joachim Buhmann9, Thomas Cerny9, Holger Moch10, Silke Gillessen11, Ruedi Aebersold6,9,12, and Wilhelm Krek1,2

1Institute of Cell Biology and 2Institute of Molecular Systems Biology, Eidgenössische Technische Hochschule Zurich, 8093 Zurich, Switzerland; 3Institute of Surgical Pathology, University Hospital Zurich, 8091 Zurich, Switzerland; 4Department of Medical Oncology, 5Department of Urology, and 6Institute of Pathology, Kantonsstipital St. Gallen, 9007 St. Gallen, Switzerland; 7Department of Computer Science, Eidgenössische Technische Hochschule Zurich, 8092 Zurich, Switzerland; and 8Institute of Medical Informatics, University of Zurich, 8057 Zurich, Switzerland

Edited by Owen N. Witte, The Howard Hughes Medical Institute, University of California, Los Angeles, CA, and approved January 12, 2011 (received for review September 25, 2010)

A key barrier to the realization of personalized medicine for cancer is the identification of biomarkers. Here we describe a two-stage strategy for the discovery of serum biomarker signatures corresponding to specific cancer-causing mutations and its application to prostate cancer (PCa) in the context of the commonly occurring phosphatase and tensin homolog (PTEN) tumor-suppressor gene inactivation. In the first stage of our approach, we identified 775 N-linked glycoproteins from sera and prostate tissue of wild-type and Pten-null mice. Using label-free quantitative proteomics, we showed that Pten inactivation leads to measurable perturbations in the murine prostate and serum glycoproteome. Following bioinformatic prioritization, in a second stage we applied targeted proteomics to detect and quantify 39 human ortholog candidate biomarkers in the sera of PCa patients and control individuals. The resulting proteomic profiles were analyzed by machine learning to build predictive regression models for tissue PTEN status and diagnosis and grading of PCa. Our approach suggests a general path to rational cancer biomarker discovery and initial validation guided by cancer genetics and based on the integration of experimental mouse models, proteomics-based technologies, and computational modeling.

Results

In the first stage of our approach, we identified PCa candidate biomarkers by applying a large-scale quantitative proteomic screen to detect and quantify N-linked glycoproteins that differ in their amount in the prostate tissue and sera of prostatic Pten-deficient PbCre4-Ptenfl/fl (Pten CKO) and littermate control animals (9) (Fig. 1A and Fig. S1 A–C). The choice of an experimental mouse model as entry point for the identification of candidate biomarkers was guided by the possibility of collecting tissue samples from a genetically defined and homogeneous population in which variables such as environmental factors, age, and tumor type and stage are controlled and standardized. We selectively analyzed N-glycosylated proteins to maximize their subsequent detectability in the serum (10) and to focus on a subproteome that is enriched for validated serum biomarkers. In fact, 30 of the 38 protein biomarkers currently used in the clinic are glycosylated (11).

Molecular and genetic biomarkers play a paramount role in clinical oncology. They can help predict who will develop cancer or detect the disease at an early stage. Biomarkers also can guide treatment decisions and help identify subpopulations of patients who are most likely to respond to a specific therapy (1, 2). However, the noninvasive detection and prognostic evaluation of a specific tumor by the analysis of indicators in body fluids such as serum remains a formidable challenge. Novel biomarkers represent today an urgent and critical medical need.

Serum has long been considered a rich source for biomarkers (3). However, the discovery of serum biomarkers has been technically challenging and ineffectual for reasons that include the particular and variable composition of the serum proteome and its enormous complexity (4). As the genetic alterations that cause cancer are becoming better understood, one strategy to overcome the limitations of the traditional serum proteome comparisons is to use the knowledge about specific cancer-causing mutations and the underlying disrupted signaling pathways to guide the discovery of novel cancer serum biomarkers.

The tumor-suppressor gene phosphatase and tensin homolog (PTEN) is one of the most commonly inactivated genes in human cancer and has been identified as lost or mutated in several sporadic cancers, including endometrial carcinoma, glioblastoma, breast cancer, and prostate cancer (5). An established consequence of PTEN inactivation is the constitutive aberrant activation of the PI3K-signaling pathway that drives uncontrolled cell growth, proliferation, and survival (6, 7). It is expected that specific signaling pathway-activating mutations such as PTEN loss will produce changes in the surface and secreted proteomes of the affected tissue (8), and, in principle, these changes should be detectable as discrete biomarker signatures in the serum. Based on this conceptual consideration, we developed a two-stage strategy for the discovery and initial validation of serum biomarkers in humans based on a mouse model of prostate cancer (PCa) progression caused by Pten inactivation.

1073/pnas.1013699108/-/DCSupplemental.


This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

1I.C. and R.S. contributed equally to this work.

2To whom correspondence may be addressed. E-mail: holger.moch@usz.ch, silke.gillessen@kssg.ch, aebersold@imsb.biol.ethz.ch, or wilhelm.krek@cell.biol.ethz.ch.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1013699108/-/DCSupplemental.
Fig. 1. Translational approach for biomarker discovery and validation workflow. (A) Candidate biomarkers are discovered using a genetic mouse model by enriching N-linked glycoproteins to sera and freshly isolated perfused prostates from wild-type and Pten cKO mice. Trypsic N-glycosites then are measured by LC-MS/MS. Identification and quantitation of proteins is performed as described. After a filtering process, candidate biomarkers are selected for the verification phase. (B) Verification phase. Highly standardized biobanking and clinical data collection are used for collecting serum and matching tissue samples from patients harboring localized PCa and control patients with BPH. N-linked glycoproteins are extracted as in A, and selected candidates from the discovery phase are measured by targeted proteomics and ELISA. At the same time, tissues are spotted as microarray and stained for the indicated antigens. Feature selection and modeling then is performed to find novel biomarkers for diagnosis, patient stratification by Gleason score, and PTEN status.

Fig. 2. Murine prostate and serum N-linked glycoproteome. (A) Venn diagrams of the mouse prostate and serum glycoproteome identified in the wild-type and Pten cKO mice indicating proteins commonly detected or detected only in the respective genotypes/organisms. (B) Label-free quantification of the proteins by means of SuperHirn plotted for prostate and serum. Dots indicate the ratio for each protein between the Pten cKO and wild-type prostates or sera and indexed from the most down-modulated to the most up-regulated. (C and D) Previously unknown Pten-dependent changes in protein expression are verified by standard cell biology techniques such as Western blot (C) and immunofluorescence (D).
chromosome 10 copy numbers per cell using commercially available fluorescently labeled DNA probes for cytoband 10p23.3 and region 10p11.1–11q11.1, respectively. We also stained sections from the TMA-P92 with antibodies reporting the activation state of the PI3K-pathway including phospho-serine (pSer)-473-Akt and stathmin (18). Seventy-two percent of prostate cancers displayed focal loss of PTEN gene copy numbers compared with the control group, indicating deletion of one or both alleles of PTEN (Fig. 3A, and Table S1) in at least 20% of the cells analyzed. This result is in agreement with previous reports using the same technique (19, 20). As expected, a significant fraction of these cancers demonstrated PTEN-pathway activation, as evidenced by the increased staining of pSer-473-Akt and stathmin (Fig. 3B and C).

Next, we analyzed serum samples from these patients by using N-glycosylation extraction followed by targeted quantitative MS via selected reaction monitoring (SRM) (21). To this end, we used a hybrid quadrupole/linear ion trap mass spectrometer (22) to detect and quantify 57 N-glycotypes, corresponding to 49 candidate protein biomarkers present on the list of prioritized candidates. The absolute serum concentrations of these proteins were determined using stable isotope-labeled reference peptides as external standards. Of the 57 targeted peptides, 36 peptides representing 33 different proteins were detected consistently and quantified in 80–105 patients (Dataset S3). The median concentration of the measured proteins varied from 320 μg/mL to 5.5 ng/mL, indicating that our approach allows the quantification of protein concentrations in sera along six orders of magnitude. The median concentration of various measured proteins was in the concentration range of prostate-specific antigen (PSA), a widely used diagnostic serum biomarker for prostate cancer. For nine proteins, we also established ELISAs, which confirmed the validity of the SRM approach for two proteins and provided independent quantitative data for the other proteins that were not detected by SRM, thus resulting in a total of 39 proteins that were quantified consistently (Dataset S3). Next, we used this dataset to select the best candidate biomarkers and to build predictive models for the discrimination between normal and aberrant PTEN status. First, we applied the random forest (RF) classifier algorithm (23) for variable ranking and subsequent selection. RF is particularly well-suited in this regard, because it does not assume that the data are linearly separable. Moreover, the selection of the top-ranked variables reduces the dimensionality of the feature space and the computing time, thus allowing a subsequent exhaustive screening of the best models.

We selected the 20 top-ranked variables resulting from 100 RFs and screened for all logistic regression models to predict focal loss of PTEN by combining one to five serum proteins. This screening resulted in 21,699 different models, which were validated by 100-fold bootstrapping (24). For each model we calculated the median area under the receiver operating characteristic (ROC) curve (AUC), thereby identifying the best regression models that are able to predict significantly aberrant PTEN status from an overlapping data set comprising 54 patients derived from the PTEN FISH analysis of 82 patients and the SRM and ELISA quantification of sera from 105 patients (PTEN focal loss <20%; n = 26; PTEN focal loss ≥20%; n = 28) (Fig. S2B). The signature comprising thrombospondin-1 (THBS1), metalloproteinase inhibitor 1 (TIMP-1), complement factor H (CFH), and prolow-density lipoprotein receptor-related protein 1 (LRP-1) could predict correctly 78% of cases belonging to patients having aberrant or normal PTEN status with a sensitivity of 79.2% and specificity of 76.5% [AUC = 0.82; 95% confidence: P = 5.49 × 10E-5; 95% confidence interval (CI) = 0.704–0.936] (Fig. 3E). Taken together, these results suggest that the reduction of PTEN gene copy number in prostate cancer led to a measurable perturbation of the serum proteome. Moreover, they demonstrate the usefulness of computational variable selection using RF followed by exhaustive regression model screening as a valid approach to extract information on candidate biomarkers.

To corroborate this analysis, we determined the occurrence of the 15 RF-selected top-ranked variables in the best 50 bootstrapped models (Fig. 3D). All proteins have been selected in more than half of all highly predictive models. This approach thus provided the theoretical robustness of discrimination of individual candidate biomarkers described in Fig. 3E. To determine whether our signature is significantly linked to the PTEN network, we sought curated knowledge-based connections between our signature and the PTEN network. When tested against 50 random signatures, the PTEN signature identified here showed significantly more direct and indirect connections to the PTEN signaling network, thus providing independent support for our serum signature as a predictor of tissue PTEN status (Fig. S3A and B). Because PTEN loss is causally associated with accelerated PCA progression and aggressiveness, as exemplified by the association between PTEN loss of function and Gleason sum score (25, 26), we next asked whether the bioinformatic approach also could extract serum protein signatures reflecting tumor grading. Examination of our TMA-P92 revealed a corre-
lation between PTEN loss and Gleason score sum (Fig. S4), confirming previous reports (25, 26). The Gleason grading available for 69 tumors and the corresponding quantitative SRM serum analysis served as the basis for applying the bioinformatics approach outlined above (Fig. S2B). Intriguingly, we identified a five-protein signature from an overlapping dataset of 54 patients comprising polyepptide GalNAc transferase-like protein 4 (GALNTL4), fibronectin (FN), zinc-α2-glycoprotein (AZGP1), biglycan (BGN), and extracellular matrix protein 1 (ECM1) that predicted patients having tumors with a Gleason score <7 or ≥7 with an AUC = 0.788 (P = 3.1*10E-4; 95% CI = 0.668–0.907; sensitivity (sens.) = 60.9%; specificity (spec.) = 67.8%) (Fig. 4B). The predictive relevance ranking corroborated the composition of the best signature (Fig. 4A). These results imply a potential link between aberrant PTEN status and the emergence of protein signatures in the serum reporting on tumor grading. Taken together, because only few reports describe serum biomarkers for the stratification of patients based on the grading of the tumors (2, 27, 28), the data suggest an application of our biomarker discovery platform for the prognosis of locPCa, wherein patients with clinically significant or insignificant prostate cancer can undergo stratification for therapy or watchful waiting, respectively (29). Finally, we assessed whether our approach can reveal signatures for PCa diagnosis. As reported previously, we note that the vast majority of the tumors show aberrant focal PTEN loss (Fig. 3A and Table S1) and altered PI3K signaling (Fig. 3 B and C). The current method of choice for noninvasive screening of PCa is the blood-based quantification of PSA together with digital rectal examination (DRE). Recent studies showing that PSA, alone or in combination with DRE, is prone to overdiagnosis and has no or very limited beneficial effects on overall survival (30, 31) suggest a strong need for better diagnostic signatures. We thus analyzed a total of 143 sera from 77 PCa patients and 66 controls. Sera from 105 patients were selected as training-validation set (Fig. S2B). Machine learning analysis applied to a quantitative data set derived from SRM analysis of the sera of 82 patients identified a four-protein signature comprised of hypoacetylated protein 1 (HYOU1), asporin (ASPN), cathepsin D (CTSD), and olfactomedin-4 (OLFM4) (32). This signature discriminated between BPH and PCa with an AUC = 0.726 (P = 0.01; 95% CI = 0.614–0.838; sens. 81%, spec. 57%). PSA measurements resulted in a similar AUC = 0.730 (P = 1*10E-6; 95% CI = 0.693–0.871; sens. 78%, spec. 63%). Strikingly, the combination of the four-protein signature with PSA resulted in an AUC = 0.840 (P = 5*10E-9; 95% CI = 0.824–0.964; sens. 85%, spec. 79%) (Fig. 4 C and D). With the aim of testing the reproducibility of our approach, we measured the four-biomarker signature by SRM in an independent test set comprising 38 patients that were not included in the training-validation set. In the test set, the diagnostic signature from an overlapping dataset of 37 patients performed equally as well as the training-validation set, indicating reproducibility and robustness of the test as well as of the measurement procedure (Fig. 4E). To exclude confounding variables such as inflammatory conditions as the origin of eventual bias (33) in our analysis, we correlated the single biomarkers comprised in the prognostic and diagnostic signatures mentioned above with clinical parameters of inflammatory state in a subset of patients, independently of the disease status. Specifically, we correlated C-reactive protein (CRP) and the leukocyte count. All the selected variables failed to correlate with either parameter, thus excluding a bias derived from the inflammatory status of the patient at the time of diagnosis (Fig. S5).

**Discussion**

The present study provides a general framework for rational cancer biomarker discovery. The underlying concept is that activation of cancer-signaling pathways caused, for example, by the inactivation of a defined tumor-suppressor gene is associated with specific protein signatures that can be measured in serum and potentially used to detect disease at an early stage or to derive information about the tumor grade and thus guide treatment decisions. In the past the discovery of serum biomarkers has been technically challenging because of the enormous complexity of the serum proteome and the lack of sensitive discovery-driven measurement technologies (4). Based on these considerations, we implemented a two-stage strategy for biomarker discovery. In the first stage, we generated a list of candidate biomarkers based on information derived from large-scale screens of the glycosylated proteome of a mouse model of PCa progression caused by prostate-specific Pten inactivation. This
approach identified multiple proteins differentially regulated upon Pen inactivation. In the second, hypothesis-testing stage, the candidate proteins were quantified simultaneously by SRM-based targeted MS in sera of PCa patients and integrated with information derived from matched PCa tissue characterized with respect to PTEN status and PI3K-pathway activation. Using machine learning algorithms, we then extracted robust patterns suitable for predicting tissue PTEN status and for the diagnosis and grading of PCa.

The current standard biomarker for early detection of PCa is PSA. However, the effectiveness of systematic PCa screening with PSA testing remains controversial, in part because a lack of sensitivity and specificity results in considerable overdiagnosis and overtreatment (30, 31, 34). The ability of our four-protein signature for prostate cancer diagnosis to distinguish accurately between locPCa and BPH makes it potentially suited for screening tests by reducing false-positive outcomes and therefore avoiding anxiety and biopsies in men who have an elevated PSA but do not harbor cancer.

Another potential drawback of PSA testing relates to the detection of clinically insignificant prostate cancers in asymptomatic men. Therefore, overdiagnosis, in this context meaning detection of cancer that has no clinical impact on an individual during his lifetime, is a major problem. There are several definitions of insignificant prostate cancer (29); most definitions exclude patients with any Gleason pattern 4 prostate cancer. Our five-protein signature predictive for Gleason score therefore may be suited for improving screening efficacy by reporting which men might harbor insignificant cancers. Such patients might be offered active surveillance instead of immediate treatment. If active surveillance is chosen as the treatment option, repetitive biopsies for detection of prostate cancer progression ultimately could be replaced by a serum test.

The biomarker discovery platform presented here provides an approach for the discovery and validation of biomarkers with the aim of improving the effectiveness of PCa testing and non-invasive diagnostic of prostate cancer. Ideally this approach could avoid overdiagnosis and overtreatment and guide treatment decision.

The simultaneous analysis of a large number of candidate biomarkers by SRM allows the discovery of new potential biomarkers independently from the availability of established immunoassays. However, this emerging technology does not yet allow the analysis of large cohorts of patients. The identified candidate biomarkers thus must be validated further in larger, prospective studies, preferably using standardized immunooassays, which are limited in the amount of proteins analyzed per sample but allow the analysis of much larger cohorts. External independent data sets must be added as well to confirm further the clinical usefulness of the reported signatures.

Because we now are entering an era in which the genetic and epigenetic abnormalities responsible for specific forms of cancer guide the design of molecularly targeting drugs, efficient strategies to evaluate such targeted therapies in patients are critical, especially as more such compounds become available. Biomarkers able to identify reliably the patients who are most likely to benefit from a specific molecularly targeted therapy therefore would have significant clinical benefit. In this regard, an increasing armamentarium of targeting agents that inhibit key components of the PI3K pathway exists, and many of these inhibitors already are in clinical testing (35–37). It is conceivable that signatures reflecting tissue PTEN status may aid in selecting suitable patients and provide proof of target modulation by these inhibitors. Thus, a cancer genetics-guided path to biomarker discovery, as described here, may hold the promise for the realization of personalized cancer medicines.

Method

Experimental Mice. PTEN cKO mice were generated as described in ref. 9. The Zurich cantonal veterinary office approved all animal studies. Details are provided in SI Methods.

Glycoprotein Enrichment from Murine Serum. Glycoproteins were enriched from sera and tissue of mice using the protocol published by Zhang et al. (12). Details about the isolation of sera and tissues from mice and the glycoprotein enrichment method are explained in SI Methods.

Mass Spectrometry Analysis. Samples were analyzed on a hybrid LTQ-FT mass spectrometer (Thermo Electron) equipped with a nanoelectrospray ion source. Chromatographic separation of peptides was performed on an Agilent 1100 micro HPLC system equipped with a 15-cm fused silica emitter, 150-μm inner diameter, packed with a Magic C18 AQ 5 μm resin (Michrom Bioresources). Further details are provided in SI Methods.

Protein Identification. Proteins were identified following protocols described in refs. 12 and 36–38. Further details are provided in SI Methods.

Label-Free Quantification of Peptide and Protein Ratios. Data from LC-MS runs were converted from raw to the mzXML data format (38) and processed by the software tool SuperHirn as described previously (19). Rratio was used for the calculation and visual assessment of peptide and protein ratios (39). After examining the distribution of the data and assuming normality (Fig S1G), we applied a two-tailed Student’s t test with unequal variances statistics to assess the significance of a protein fold-change. Protein fold-changes with a nonstringent P value ≤0.15 were selected for further analyses. To verify the results obtained by SuperHirn and Rratio, we performed spectral counting analysis (SI Methods). Prostate specificity was calculated from gene-expression profiles obtained for the BioGPS database (http://biogps.org) (40) by calculating the ratio of average gene expression from prostate and average gene expression of the remaining tissues. Gene ratios of 10 or more were considered prostate specific.

SDS/PAGE and Western Blotting. Perfused fresh-frozen prostate tissues were solubilized in RIPA buffer (150 mM NaCl, 10 mM Tris, 0.1% SDS, 1% Triton X-100, 1% deoxycholate, 5 mM EDTA) plus protease inhibitors (1 mM phenylmethylsulfonyl fluoride, 10 mM benzamidine, 10 μg/ml aprotinin). Fifty-microgram protein extracts were resolved on 8–12% SDS/PAGE, blotted on nitrocellulose, and visualized by immunoblotting with the following primary antibodies: anti–phospho-Ser473 AKT (#4058; Cell Signaling Technology), anti-complement factor B (#HPA001817; Sigma Aldrich), anti-KDEL (#ab12223; Abcam), anti–Niemann-Pick C1 (#NB400-14855; Novus Biologicals), anti–LAMP-1 (clone 1D4B; Developmental Studies Hybridoma Bank), anti–α-tubulin (clone YL1/2, #ab6160; Abcam).

Immunofluorescence. Five-micrometer cryostat sections on poly-L-lysine slides were fixed in PBS/4% paraformaldehyde for 10 min, washed in PBS, and stained using antibodies against the indicated proteins. Further details are provided in SI Methods.

Real-Time PCR Analysis. Prostate tissues from three wild-type and three PTEN cKO animals were isolated as described. Total RNA was prepared from powdered tissue using the RNeasy Mini Kit (Qiagen), and cDNA was prepared using random hexanucleotide primers and Ready-to-go you-prime first-strand beads (GE Healthcare). Real-time PCR analysis of cDNA was performed using LightCycler 480 SYBR Green I Master from Roche and specific primers reported in SI Methods.

Patients, Sampling, and Handling of Human Sera and Glycoprotein Enrichment. The Ethics Committee of the Kanton St. Gallen, Switzerland, approved all procedures involving human material, and all patients signed an informed consent. For the study we included patients with locPCa and BPH. We excluded from the analysis patients with advanced prostate cancer, infectious or inflammatory diseases, or other malignancies. Eight milliliters of blood were drawn and collected in a serum separator tube containing clot activator and gel (Vacutainer, SSTTM II Advance, REF 367953; Becton Dickinson). Tubes were inverted eight times and centrifuged within 4 h of collection at 4 °C for 10 min at 1,428 g. The serum was divided into five aliquots of 500 μl each and stored at −60 °C or lower until use. Glycoprotein extraction was performed exactly as described for the murine serum.
Targeted MS Analysis Using SRM. We used the absolute quantification of proteins (AQUA) strategy introduced by Gerber et al. (41). Further details are reported in SI Methods.

Tissue Microarray Preparation. A tissue microarray was constructed as described (42) using formalin-fixed, paraffin-embedded tissue samples derived from 92 patients (BPH, n = 40; locPCa, n = 52) with matched serum samples that were used for SRM or ELISA. Details are provided in SI Methods.

Immunohistochemistry. Immunohistochemistry was performed using a Ventana Benchmark automated staining system (Ventana Medical Systems) and the following primary antibodies: anti-phospho-Ser473 AKT (dilution 1:150; #ab8932; Abcam) and anti-phostathmin (dilution 1:50; #3352; Cell Signaling Technology).

FISH. To assess PTEN deletion, we performed dual-color FISH on paraffin-embedded tissue using commercially available fluorescently labeled DNA probes for cytoband 10q22.3 (SpectrumOrange, PTEN locus-specific probe) and chromosome 10 (Spectrum-Green centromere of chromosome 10 probe; LSI PTEN/CEP 10; Abbott Laboratories) according to the manufacturer's instructions. Details are provided in SI Methods.

Bioinformatic Analysis of SRM Data. SRM data were normalized and featured selection using random forest followed by signature modeling using brute force search for all logistic models. AUCs for every model were calculated by bootstrapping to avoid overfitting. Details are provided in SI Methods.

ELISA. The concentration of selected candidate biomarkers (Dataset 53) was measured by sandwich or competitive ELISA following the manufacturer's instructions. Details are given in SI Methods.

ACKNOWLEDGMENTS. We thank all members of our laboratory for helpful discussions and Corina Fusaro-Graf, Doris Kradolfer, Melanie Sticker, and Manuela Hitz for technical assistance. The project was supported by Grant 3100A0-107679 from the Swiss National Science Foundation (to R.A.), Grant KFP OCS 1262-06-2002 from the Collaborative Cancer Research Project from Oncosuisse (to T.C., S.G., W.K., and H.M.), Smart Inactivity Monitor using Array Based Detectors Grant 213250 from the Future and Emerging Technologies Programme of the European Commission (to P.S., T.F., and J.B.), and by a grant from the Vontobel-Stiftung (to W.K.).

Supporting Information

Cima et al. 10.1073/pnas.1013699108

SI Methods

Experimental Mice. To generate phosphatase and tensin homolog (PTEN) cKO mice with specific deletion of PTEN in the prostate epithelium (Ptenfl/fl;Cre+), ARR2Probasin-Cre transgenic line, PB-Cre4 on C57BL/6xDBA2 background (1) [Mouse Models of Human Cancers Consortium (MMHCC) Repository] were crossed to Ptenfl/fl mice on a 129/BALB/c background (Jackson Laboratory). The male offspring with Ptenfl/fl;Cre+ genotype then were crossed to Ptenfl/fl;Cre− females. As control, normal littermates Ptenfl/fl;Cre− were used for the study.

Procedures for Tissue and Blood Extraction from Mice. Mice were anesthetized, and blood was extracted by pinning the left heart ventricle. Mice subsequently were heart-perfused using ice-cold PBS containing 10 µg/mL aprotinin. Tissue samples were dissected, and pure prostate tissue was quickly snap-frozen and pulverized by using a mortar and pestle in the presence of liquid nitrogen. Serum was extracted from the blood after complete clotting on ice and was stored at ~80 °C until use. Intact prostate samples were embedded in optical cutting temperature (OCT) medium (Medite) for cryosections and immunofluorescence or in paraffin blocks for immunohistochemical staining.

Glycoprotein Enrichment from Murine Tissue. Murine prostate tissue was subjected to glycoprotein extraction using 50% trifluoroethanol (TFE) for 2 h at 60 °C. Proteins were solubilized using RapigestTM (Waters) and proteolyzed with trypsin. Solid-phase extraction of glycopeptides (SPEG) was performed as follows: Glycan moieties of glycopeptides were oxidized and coupled to N-linked glycated peptides and tissue debris were subsequently washed away. N-linked glycopeptides then were released via N-Glycosidase F (PNGase F) (BioConcept), and the recovered peptides were analyzed by liquid chromatography tandem MS (LC-MS/MS).

Glycoprotein Enrichment from Murine Serum. Glycoproteins were oxidized by adding sodium periodate, and then the sample was conjugated to the hydrazide gel. Nonglycoproteins and tissue debris were subsequently washed away. Glycopeptides were released via PNGase F (BioConcept), and the recovered peptides were analyzed by liquid chromatography tandem MS (LC-MS/MS).

Mass Spectrometry Analysis. Peptides were loaded on the column from a cooled (4 °C) Agilent autosampler and separated with a linear gradient of acetonitrile/water, containing 0.1% formic acid, at a flow rate of 1.2 µL/min. We used a linear gradient from 2–40% acetonitrile over 60 min that was optimized for the number of peptide features detected. The MS instrument was operated to maximize the quality of LC-MS feature maps as opposed to maximizing the number of identifications. Therefore, for each peptide sample, a standard data-dependent acquisition (DDA) on the three most intense ions per MS-scan was performed. Three MS/MS spectra were acquired in the linear ion trap per Fourier transformed (FT)-MS scan, acquired at 100,000 FWHM (at 350 m/z) nominal resolution, resulting in an overall cycle time of ~1 s. Charge-state screening was used, allowing fragmentation of doubly and higher charged ions and rejecting ions of single and unknown charge state. A threshold of 200 ion counts was set to trigger an MS/MS attempt.

Protein Identification. The raw data acquired by the LTQ-FT (software: Xcalibur 2.0 SR1) were converted to mzXML using ReAdW 3.5.1 applying default parameters. MS/MS scans then were exported as .dta files without further processing using the program mzXML2Other (2). MS/MS spectra were searched against the International Protein Index (IPI) murine protein database (version 3.26) using SEQUEST v.27 (3). The identified peptides were processed and analyzed through the mass spectrometry Trans-Proteomic Pipeline 3.5 (TPP) (4). The peptides then were assigned for protein identification using the Protein-Prophet software (5) using a PeptideProphet probability score ≥0.9, and a ProteinProphet probability score ≥0.9, resulting in an overall false-positive error rate below 1% as determined by ProteinProphet. The details of the SEQUEST and database search are explained in the following section.

SEQUEST and TPP Database Search. The SEQUEST database search criteria included static modifications of 57.02146 Da for cysteines (for the alkylation with iodoacetamide), variable modifications of 15.99491 Da for methionines (for oxidation), and modifications of 0.98406 Da for potential formerly N-glycosylated asparagines (which are converted to aspartic acid by PNGase F release). The following additional search constraints were applied: mono-isotopic parent and fragment masses precursor-ion mass tolerance of 0.05 Da; fragment-ion mass tolerance of 0.5 Da; at least one tryptic terminus; and one missed cleavage. In the TPP, the database search results were validated using the PeptideProphet software (6), which uses various SEQUEST scores (XCorr, ΔCn, Sp) to calculate a probability score for each identified peptide by linear discriminant analysis. N-glycosylation motif information and accurate mass binning were used in PeptideProphet.

Spectral Counting Analysis. For every LC-MS/MS run, the ratio of total peptides for the protein to be quantified to the total peptides detected in the run was calculated. The average ratio from all LC-MS/MS runs from all mice analyzed was calculated. The averaged ratio from Pten cKO mice was subtracted from the averaged ratio from Pten wild-type mice. A positive ratio was considered as up-regulation for the selected protein. A two-tailed Student’s t test with unequal variances was applied to calculate the level of significance of the detected differences. The null hypothesis was rejected using a nonstringent P value ≤0.15.

Immunofluorescence. The following primary antibodies were used for immunofluorescence staining: anti-phospho-Ser473 AKT (1:200; #4058; Cell Signaling Technology), anti-CD133 (1:200; #14–1331; eBioscience), anti-LAMP2 (1:50; clone GL2A7; Developmental Studies Hybridoma Bank), anti-cytokeratin 19 (1:100; clone TROMA-III; Developmental Studies Hybridoma Bank), anti-CD44 (1:200; #14–0441; eBioscience), anti-cytokeratin 14 (1:500; polyclonal rabbit; gift from S. Werner, ETH Zurich, Zurich), and anti-CD49f, (1:200; #14–0495; eBioscience). For BrdU staining, animals were injected twice with BrdU at 50 mg/kg at time points t = 0 h and t = 2 h. Animals were killed at time point t = 6 h, and organs were isolated as described. Tissue sections were incubated with 2N HCl for 45 min for DNA denaturation followed by visualization of BrdU using anti-BrdU antibodies (1:1,000; #ab6326; Abcam).
Primers for Real-Time PCR Analysis. The following primers were used to amplify the respective transcripts: Cd44: forward primer (fwd) 5'-TCGATTGTAATGACCTGGCC-3', reverse primer (rev) 5'-CAGTCCGGGAGATGACTGTGGC-3'; Smo: fwd 5'-GAGCGTACGCTCAGGCACTA-3', rev 5'-CAGCGGCGAT-TCCTGATCC-3'; Bcl-2: fwd 5'-ATGCGCTTGTGGGACCAT- TATTCGGC-3', rev 5'-GGTTAGTACCCCAAGGTATGC-3'; Cd24: fwd 5'-GTTCAGCCTGTTCCGGGATTA-3', rev 5'-CC- CCTTGGTGTTACGCGT-3'.

Targeted Mass Spectrometry Analysis Using Selected Reaction Monitoring. As internal standards we used at least one heavily-labeled form of an N-glycosylpeptide from protein from Sigma-Aldrich and Thermo Fischer Scientific, designated the “absolute quantification of proteins” (AQUA) peptide. Selected reaction monitoring (SRM) analyses were performed on a hybrid triple-quadrupole/linear ion trap mass spectrometer 4000 Q TRAP operated with Analyst 1.4.2 supporting scheduled experiments (Applied Biosystems/MDS Scieix). The instrument was coupled to a Tempo Nano LC system (Applied Biosystems/MDS Scieix) for peptide separation using a 30-min gradient from 5–30% acetonitrile (0.1% formic acid) at a flow rate of 300 nL/min. A fused silica emitter with a 75-μm inner diameter was packed in-house with 15 cm Reprosil-Pur 120 ODS-3.3 μm (Dr. Maisch GmbH). Quantitative analyses in SRM mode were performed with Q1 and Q3 operated in unit resolution (0.7 m/z half maximum peak width). For quantification, peak height was determined with Multiquant software v. 1.1.0.16 (Applied Biosystems/MDS Scieix) after confirming the coelution of all transitions for each peptide. Low-abundant signals of endogenous peptides were checked for correspondence of the fragment pattern compared with the internal AQUA standard. Peptides with unfavorable elution profiles (bad resolution) or interfering noise in the heavy or light transitions were excluded from further data analysis and from the transition table. No individual outlier data points were removed. To correct for spray efficiency and ionization differences between runs, the ratio of the peak height for the light transition to the peak height for the heavy transition was calculated and used for normalization and quantification of absolute protein concentrations.

In the beginning, we used AQUA peptides to optimize the SRM transitions specific for each N-glycosylpeptide that formerly were generated in silico using TQIAM software (7). For each AQUA-peptide precursor we calculated the transitions with precursor charges 2+ and 3+ and the four smallest y-ions with m/z > precursor m/z + 30. Collision energies were calculated according to the formulas: CE = 0.044 * m/z + 5.5 (2+) and CE = 0.051 * m/z + 0.5 (3+). Then we optimized the collision energy and declustering potential for all transitions as described by Picotti et al. (8). Results were imported into TQIAM, and the three transitions with the best signal/noise ratio were selected for quantitative analysis. For quantitative analysis we spiked heavy peptides as an internal standard to each serum sample. For each AQUA peptide transition we calculated the corresponding transition of the endogenous peptide, resulting in six transitions per peptide. Quantitative measurements were performed via time-scheduled SRM acquisition using a 2–to-3-min retention time window and a cycle time of 3 s on average.

Tissue Microarray Preparation. Briefly, suitable areas for prostate tissue retrieval were marked on routine H&E-stained sections. These areas were punched out of the paraffin block and inserted into the recipient block using a tissue arrayer (Beecher Instruments). The punch diameter was 0.6 mm. The whole tissue microarray preparation was accomplished on the paraffin blocks using cores from two different areas of each tumor or control tissue to account for tissue heterogeneity. After the block construction was completed, 4-μm sections of the resulting tissue microarray block were cut for further FISH or immunohistochemical analyses.

FISH. For each case, 60 nuclei were selected for scoring according to morphological criteria using DAPI staining. All hybridizations were evaluated by two investigators (P.W. and W.J.). We scored only nonoverlapping intact interphase nuclei from epithelial cells. Each cell was analyzed simultaneously for the centromeric signals of chromosome 10 and the PTEN locus on 10q23.3, using the Vysis PTEN/CEP 10 FISH Probe Kit from Abbott. 4'-6-Diamidino-2-phenylindole, dihydrochloride staining of nuclei with reference to the corresponding H&E-stained tissue identified the areas of adenocarcinoma. Based on hybridization in 10 nonepithelial control cores, deletion of PTEN was defined as cores having ≥20% (mean ± SD) of cells that had lost at least one copy number of PTEN.

ELISA. For the sandwich ELISA, Capture antibody was coated overnight at 4 °C in 50 mM Tris buffer either at pH 7.2 or pH 9.4. Plates subsequently were blocked using 1% BSA in PBS. Samples were diluted 1:5–500 in diluent buffer [1/3 LowCross Buffer (Candor Bioscience), 0.1% goat serum, 1% BSA in PBS] for 2 h at room temperature. The corresponding biotinylated detection antibody was diluted in diluent buffer and incubated for 2 h at room temperature, followed by 20-min incubation with Streptavidin–HRP and substrate detection. For competitive ELISA, human serum was diluted 1:200–1,000 in 1% BSA and incubated overnight with the suitable antibody. The antibody–antigen complex then was transferred to a plate coated with recombiant protein for 2 h to detect the endogenous antigen–antibody fraction. Signals were detected by anti-mouse antibodies conjugated with HRP and a suitable substrate according to the manufacturer's protocol.

Bioinformatic Analysis of SRM Data. Feature selection. To obtain stable predictions for the most important candidates, a series of 100 random forest classifiers (9) was grown for each of the following target variables: focal PTEN loss ≥20%, Gleason score ≤7, and diagnosis of localized prostate cancer. The 20 most important predictor variables were selected for each of the random forest classifiers. The number density for each selected predictor derived from this approach was assessed, and the 20 predictors with the highest density among the random forests were chosen for signature modeling.

Signature modeling. The selected predictors were subjected to exhaustive (or brute-force) search of all logistic regression models comprising one to five predictors. To avoid overfitting, each model was validated using 100-fold bootstrapping with replacement. The resulting median area under the receiver operating characteristic curve on the target was calculated for all of the models. The best model for each target then was chosen and described in the paper. To establish the importance of every single predictor variable, we calculated the number density of the 20 predictors among the 50 best regression models. The reported diagnostic model was validated further in an independent test set, as described in the paper, by binary logistic regression using leave-one-out cross validation to avoid overfitting.

PTEN Network Connections Analysis. The number of connections to the PTEN network shown in Fig. S3A were counted using the “explore” function of the Ingenuity Systems pathway-analysis software.


Characterization of the Pten cKO mouse model and cellular localization of the extracted glycoproteome. Pten cKO mice develop early-stage prostate tumors at complete penetrance, characterized by high-grade prostatic intraepithelial neoplasia (PIN) with histological characteristics of human in situ carcinoma with enlarged nuclei, loss of cell polarity, prominent nucleoli, and intraglandular cribriform growth. Moreover, loss of Pten in the prostate leads to defects in the differentiation of cells along the basal–luminal axis. (A) Dorsolateral prostate histology from 4-, 8-, 12-, and 16-wk-old wild-type and Pten cKO mice with enlarged views of areas indicated by squares. (B) Immunohistochemical staining of wild-type and Pten cKO prostate from 16-wk-old mice for PTEN indicating Pten loss in the epithelium. Residual PTEN staining indicates stromal cells. Downstream signaling events for PTEN-pathway activation are indicated by increased staining for pAkt (Ser473) and reduced nuclear p27 staining in the epithelium of the Pten cKO mice compared with wild-type mice. (C) Prostate tissues from three wild-type and three Pten cKO mice were lysed and probed for the indicated proteins by Western blot. (D) Immunofluorescence staining shows differentiation defects along the basal–luminal axis with the expansion of the basal cell compartment as shown by cytokeratin 14 (CK14) and cytokeratin 19 (CK19) staining, uncontrolled proliferation in the basal and luminal compartment as shown by BrdU staining, and the up-regulation of markers associated with the stem cell phenotype such as integrin α6 (CD49f). (E) Expression of stem cell markers in vivo. Asterisks indicate significant differences in the mRNA levels.
expression of Cd44 \( (P = 0.03) \), Smo \( (P = 0.02) \), and Bcl-2 \( (P = 0.046) \) in wild-type and Pten cKO animals (two-tailed Student’s \( t \) test with unequal variance; \( n = 3 \)).

(F) Pie charts indicating the cellular localization of the identified proteins on \( N \)-linked glycoprotein capture with a false-discovery rate <1\%. The identification of proteins indicates an enrichment of proteins belonging to compartments related to the extracellular space or the secretory pathway. (G) Normal quartile–quartile (QQ) plots and respective frequency histograms (with normal overlay) for the quantified peptide datasets of the mouse glycoproteome. The \( t \) test was applied under normality assumption following the inspection of the sample distributions from the indicated mouse genotypes.

Fig. S2. Patient availability for the hypothesis-testing stage and signature modeling. (A) Number of patients and available sera and tissues are displayed as a Venn diagram. (B) Sample numbers for the different signatures modeled are displayed as a workflow diagram.
**Fig. S3.** Knowledge-based link analysis of the reported PTEN signature with the PTEN signaling network. The predictive PTEN signature shows an increased number of connections with the PTEN signaling network compared with random signatures. (A) Representative direct and indirect connections between the PTEN signature and the PTEN signaling network. Proteins belonging to the PTEN signature are indicated in red. Nodes connecting the PTEN signature to the PTEN network are shown in blue. (B) Histograms show the distribution of direct and indirect connections between 50 random 4-glycoprotein signatures and the PTEN network. The black dashed line marks the median of the random distribution. The red dashed line indicates the number of the PTEN serum signature connections. Indirect connections are links between the four-protein signature bridged respectively by one (“indirect+1”) or two (“indirect+2”) further nodes connected to the PTEN signaling network.

**Fig. S4.** PTEN genetic status correlates with Gleason score. PTEN loss is monitored by FISH. Normal (absent), low, and high levels of PTEN deletion are defined by the percent of cells with loss of at least one PTEN allele in <15%, 15–30%, or >30% of analyzed cells, respectively. Focal PTEN loss data are plotted in a 100%-stacked bar chart for benign prostatic hyperplasia (BPH), localized prostate cancer (locPCa) Gleason score <7, and locPCa Gleason score >7 (n = 82).
Fig. S5. Single candidate biomarkers selected for the diagnostic and prognostic signatures described in Figs. 3 and 4 were tested for their correlation with clinical parameters for systemic inflammation such as C-reactive protein (CRP) (A) and leukocyte count (WBCs) (B). Data are shown as scatter plots. The y axis represents the concentration in ng/mL of the independent variables measured by SRM. Data were analyzed for their correlation with CRP and WBC (n = 52), and the Pearson’s coefficient (r) is shown in parentheses near the name of each variable. ASPN, asporin; BGN, biglycan; CFH, complement factor H; CTSD, cathepsin D; ECM1, extracellular matrix protein 1; FN-1, fibronectin 1; GALNTL4, GalNAc transferase-like protein 4; HYOU1: hypoxia up-regulated protein 1; OLFM4, olfactomedin-4; PSA, prostate-specific antigen; THBS1, thrombospondin-1; TIMP-1, metalloproteinase inhibitor 1.
Table S1. Clinical parameters collected from patients participating in the study

<table>
<thead>
<tr>
<th>Variable</th>
<th>Localized prostate cancer</th>
<th>Benign prostatic hyperplasia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total # patients</td>
<td>77</td>
<td>66</td>
</tr>
<tr>
<td>Age (y)</td>
<td>67.7 (49.1–89.4)</td>
<td>65.7 (50.8–90.26)</td>
</tr>
<tr>
<td>Body mass index</td>
<td>25.2 (19.6–41.3)</td>
<td>25.7 (22.6–44.3)</td>
</tr>
<tr>
<td>C-reactive protein (mg/L)</td>
<td>2 (0.2–36)</td>
<td>2.5 (0.2–32)</td>
</tr>
<tr>
<td>Leukocyte count (&gt;10^9/L)</td>
<td>6.15 (3.6–9.6)</td>
<td>6.75 (3.7–11.9)</td>
</tr>
<tr>
<td>Prostate-specific antigen (ng/mL)</td>
<td>8.6 (0.4–895)</td>
<td>4 (0.2–14.3)</td>
</tr>
<tr>
<td>&lt;4 ng/mL</td>
<td>21.3%</td>
<td>48.4%</td>
</tr>
<tr>
<td>&gt;4 ng/mL</td>
<td>78.7%</td>
<td>51.6%</td>
</tr>
<tr>
<td>Gleason score</td>
<td>6 (3–9)</td>
<td></td>
</tr>
<tr>
<td>3–6</td>
<td>64.7%</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>26.5%</td>
<td></td>
</tr>
<tr>
<td>8–9</td>
<td>8.8%</td>
<td></td>
</tr>
<tr>
<td><strong>PTEN FISH</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Epithelial cells with loss of at least one <strong>PTEN</strong> allele (range)</td>
<td>20% (6.7–98.3%)</td>
<td>7.5% (0–13.3%)</td>
</tr>
<tr>
<td>≥ 20%</td>
<td>73%</td>
<td>0%</td>
</tr>
<tr>
<td>&lt; 20%</td>
<td>27%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Data are represented as number, medians (range), or percentage.

Other Supporting Information Files

- Dataset S1 (XLS)
- Dataset S2 (XLS)
- Dataset S3 (XLS)