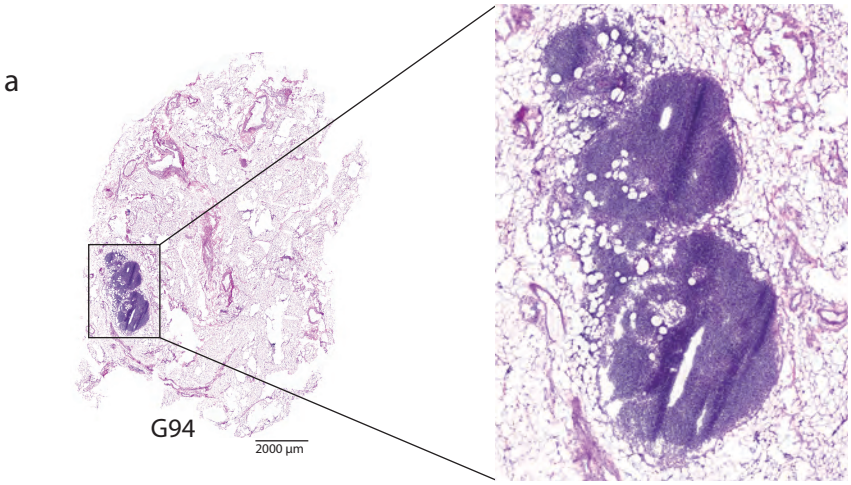
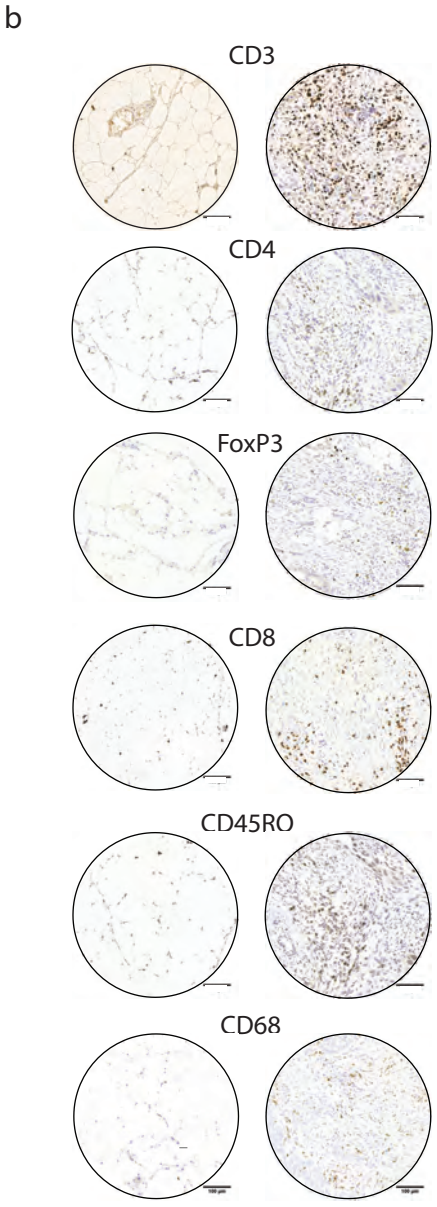
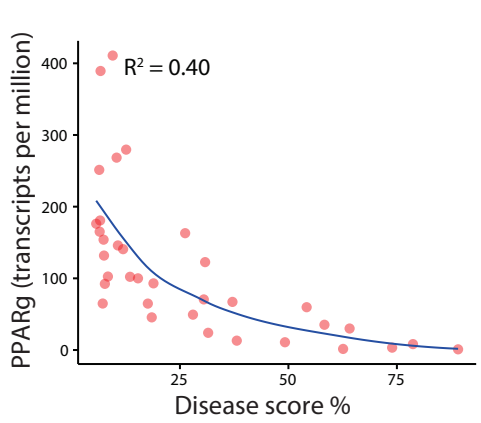
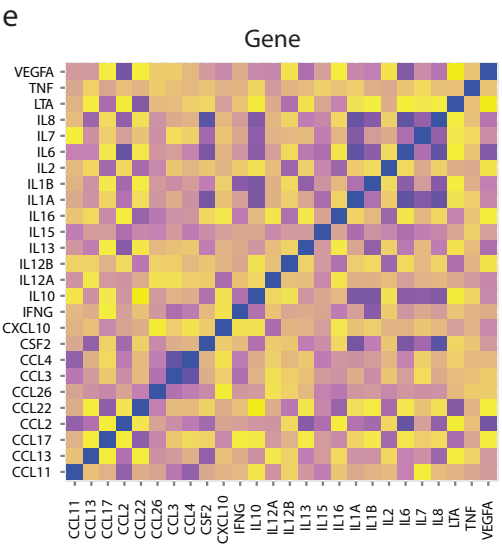

a
G94
2000 µm


b
CD3
CD4
FoxP3
CD8
CD45RO
CD68


c
PPARg (transcripts per million)
R² = 0.40
Disease score %


e
Gene

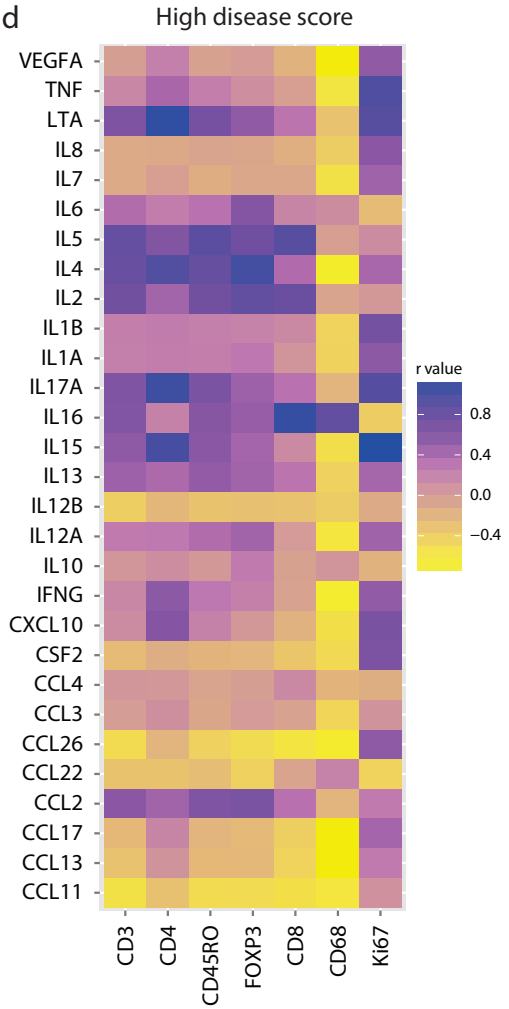
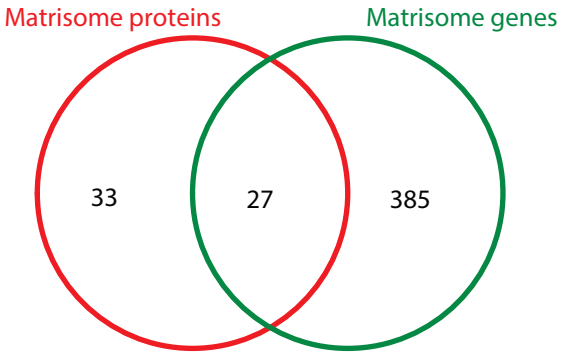
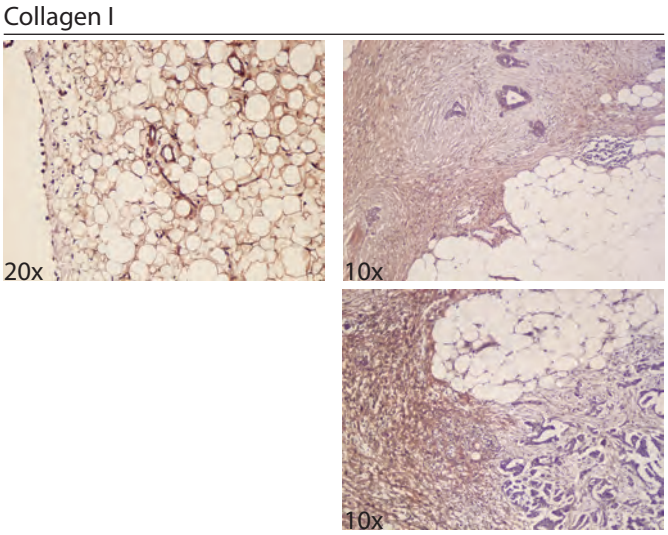

d
High disease score
r value

Figure S1. Immune cells and cytokines of the tumor microenvironment. Related to Figures 1 and 2. a) Microscope image showing H&E stained section from G94 indicating the lymphoid structured present in that sample. Scale-bars correspond to 2000μm. b) Representative IHC images of low (left) and high (right) disease score tissue sections stained for the indicated markers. Scale-bars correspond to 100μm. c) Scatterplot illustrating the correlation of PPARγ gene expression (tpm) against disease score (polynomial regression, $N = 35$, $R^2 = 0.40$, $p < 0.0001$). d) Heatmap of pairwise Pearson's correlation coefficients of MSD-quantified cytokine/chemokine gene expression (tpm). e) Heatmap of pairwise Pearson's correlation coefficients of MSD-quantified cytokine/chemokine correlations against immune cell counts in the top 10 highest disease score samples.

a


Matrisome proteins
Matrisome genes
33
27
385

b


Collagen I
20x
10x
10x

c

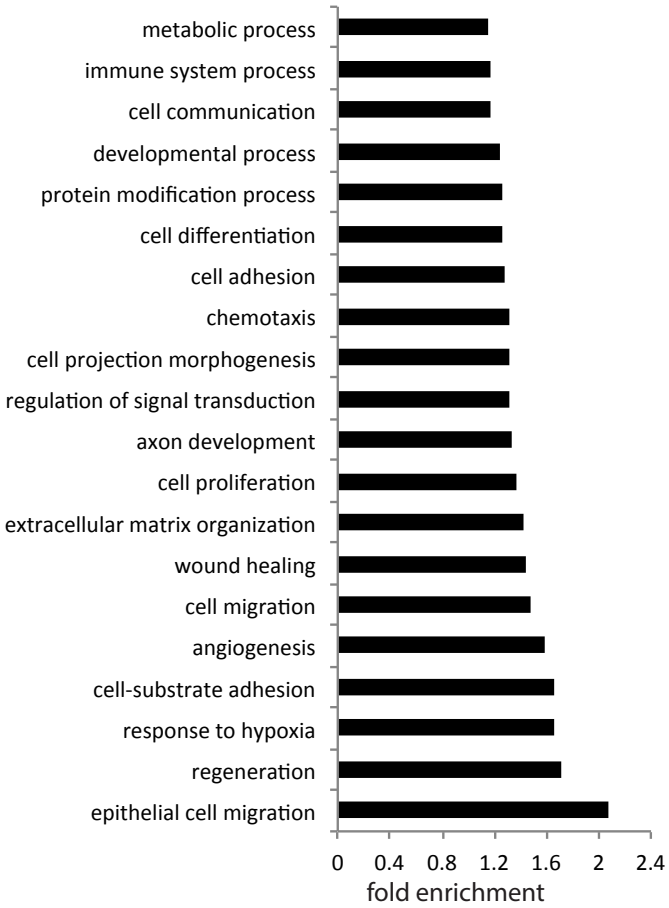

metabolic process
immune system process
cell communication
developmental process
protein modification process
cell differentiation
cell adhesion
chemotaxis
cell projection morphogenesis
regulation of signal transduction
axon development
cell proliferation
extracellular matrix organization
wound healing
cell migration
angiogenesis
cell-substrate adhesion
response to hypoxia
regeneration
epithelial cell migration
0
0.4
0.8
1.2
1.6
2
2.4
fold enrichment

Figure S2. Analysis of PLS-identified matrisome proteins and genes. **Related to Figure 3.** a) Venn diagram showing the overlap of matrisome proteins and genes identified by PLS regression models as significantly associated with disease score. b) IHC staining for Collagen I in low disease score (left) and high disease score (right) samples. c) Significantly enriched Biological Process Gene Ontology terms in PLS-identified protein coding genes (7,380) correlative to disease score ($p < 0.05$).

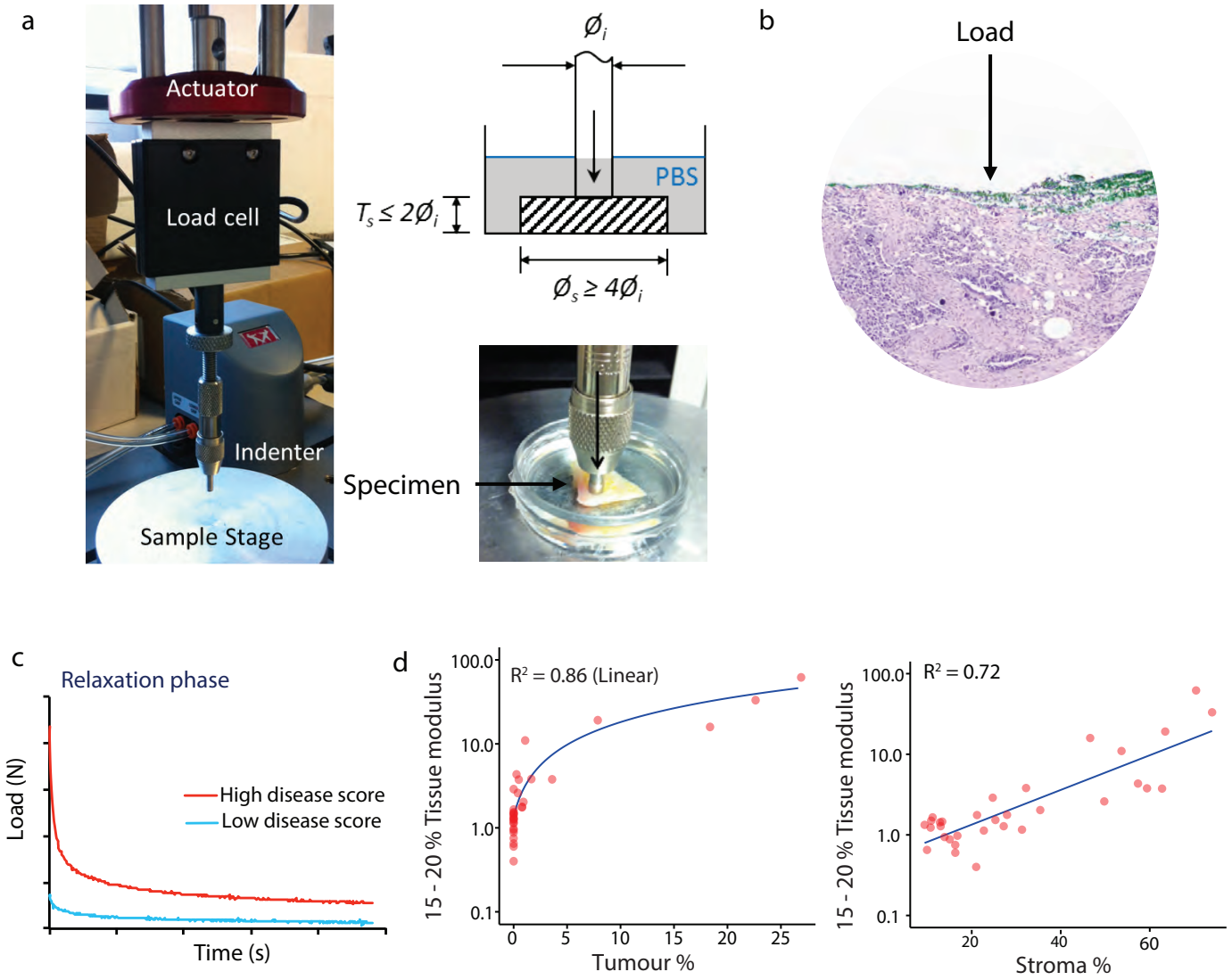
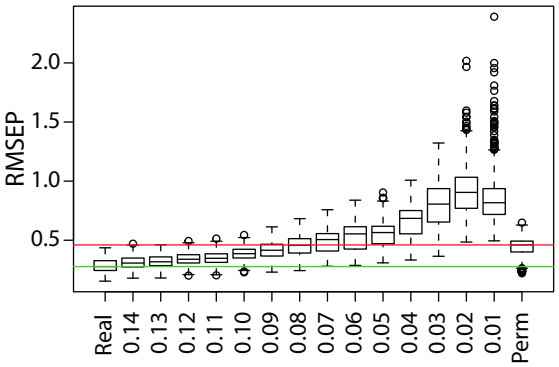
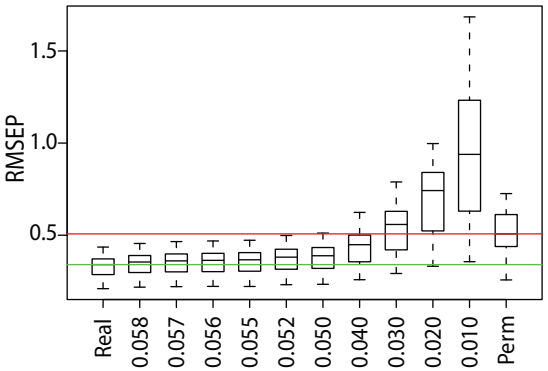

a
Actuator
Load cell
Indenter
Sample Stage
$\emptyset_i$
PBS
$T_s \leq 2\emptyset_i$
$\emptyset_s \geq 4\emptyset_i$
Specimen
b
Load
c
Relaxation phase
Load (N)
High disease score
Low disease score
Time (s)
d
15 - 20 % Tissue modulus
R^2 = 0.86 (Linear)
Tumour %
R^2 = 0.72
Stroma %

Figure S3. Overview of the biomechanical approach taken to quantify tissue modulus. Related to Figure 4. a) Setup of flat-punch indentation technique; left panel shows image of actuator driven flat-punch indenter connected to a load cell; top right panel shows a schematic of the relationship between the indenter diameter, **Øi**, and the test specimen thickness, **Ts**, and diameter, **Øs**, while loaded (direction indicated by vertical arrow) in phosphate buffered saline (PBS); bottom right panel shows a test in progress. b) A representative H&E cross-section taken from a test specimen cut perpendicular to the direction of load (arrow) under the area of flat-punch contact marked by green tissue dye. c) Representative load-displacement curve from relaxation phase obtained from high and low disease score samples. d) Optimal tissue modulus correlated against % tumor and % stroma $N = 32$, $p < 0.05$).

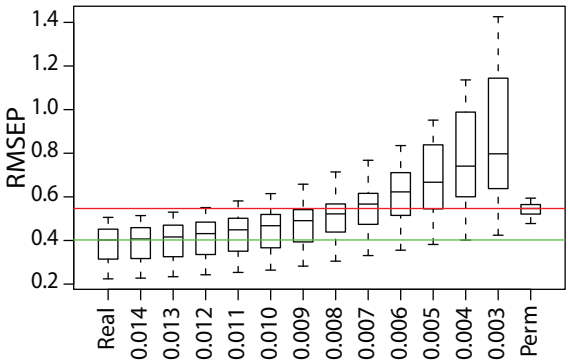
a


RMSEP
2.0
1.5
1.0
0.5
Real
0.14
0.13
0.12
0.11
0.10
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
Perm

b


RMSEP
1.5
1.0
0.5
Real
0.058
0.057
0.056
0.055
0.052
0.050
0.040
0.030
0.020
0.010
Perm

c


RMSEP
1.4
1.2
1.0
0.8
0.6
0.4
0.2
Real
0.014
0.013
0.012
0.011
0.010
0.009
0.008
0.007
0.006
0.005
0.004
0.003
Perm

d

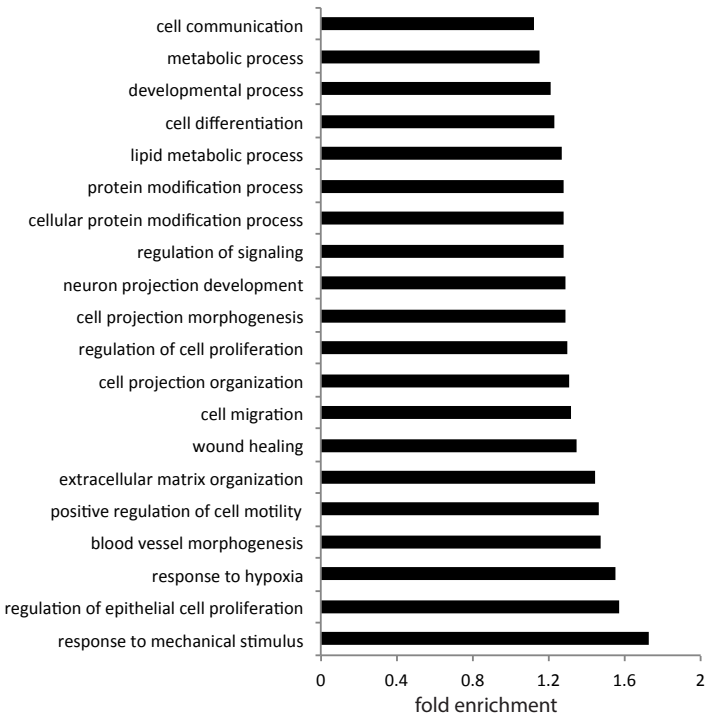
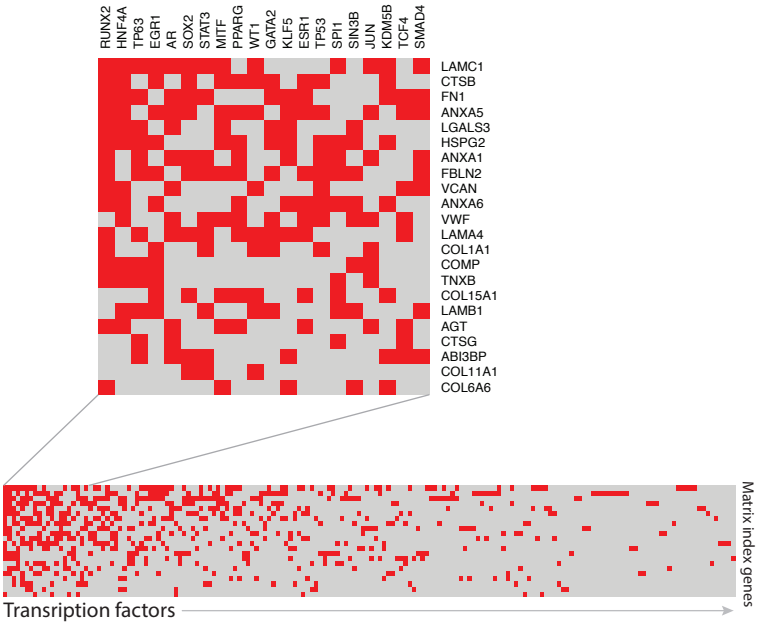

cell communication
metabolic process
developmental process
cell differentiation
lipid metabolic process
protein modification process
cellular protein modification process
regulation of signaling
neuron projection development
cell projection morphogenesis
regulation of cell proliferation
cell projection organization
cell migration
wound healing
extracellular matrix organization
positive regulation of cell motility
blood vessel morphogenesis
response to hypoxia
regulation of epithelial cell proliferation
response to mechanical stimulus
0
0.4
0.8
1.2
1.6
2
fold enrichment

Figure S4. Analysis used to identify components associated with tissue modulus. Related to Figure 4. a-c) Permutation-derived threshold for determining sets of molecular components significantly associated with tissue modulus. Boxplots illustrate bootstrapped RMSEP values on cross-validated PLS regression models of a) ECM associated protein versus tissue modulus b) Matrisome genes versus tissue modulus, c) all coding genes versus tissue modulus. In each case, bootstrapped RMSEP of the complete dataset as well as following exclusion of variables in order of weight and of a permuted dataset is illustrated. Green line denotes median RMSEP of the complete dataset; red line denotes median RMSEP of the permuted dataset and was used as a cutoff value. d) Significantly enriched Biological Process Gene Ontology terms in PLS identified protein coding genes (7,287) correlative to tissue modulus ($p < 0.05$).

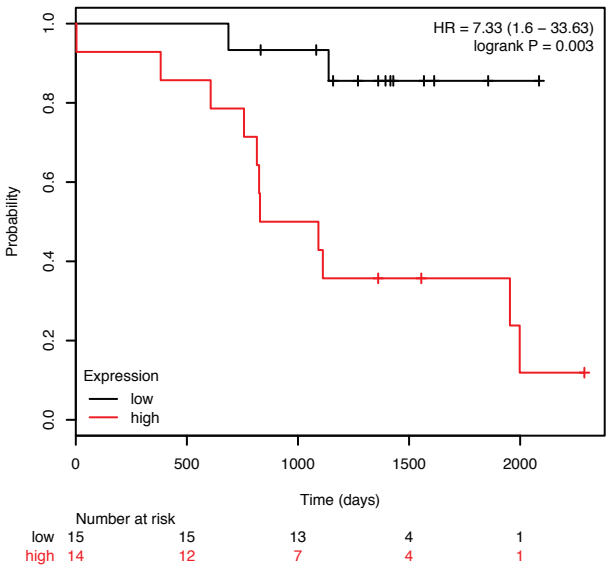
a

Symbol	Description	Category	Class
COL11A1	collagen type XI alpha 1 chain	Core matrisome	Collagen
COMP	cartilage oligomeric matrix protein	Core matrisome	ECM Glycoprotein
FN1	fibronectin 1	Core matrisome	ECM Glycoprotein
VCAN	versican	Core matrisome	Proteoglycan
CTSB	cathepsin B	Matrisome-associated	ECM Regulator
COL1A1	collagen type I alpha 1	Core matrisome	Collagen
AGT	angiotensinogen	Matrisome-associated	ECM Regulator
ANXA5	annexin A5	Matrisome-associated	ECM-affiliated Protein
ANXA6	annexin A6	Matrisome-associated	ECM-affiliated Protein
LAMB1	laminin subunit beta 1	Core matrisome	ECM Glycoprotein
FBLN2	fibulin 2	Core matrisome	ECM Glycoprotein
LAMC1	laminin subunit gamma 1	Core matrisome	ECM Glycoprotein
LGALS3	lectin, galactoside binding soluble 3	Matrisome-associated	ECM-affiliated Protein
CTSG	cathepsin G	Matrisome-associated	ECM Regulator
HSPG2	heparan sulfate proteoglycan 2	Core matrisome	Proteoglycan
COL15A1	collagen type XV alpha 1 chain	Core matrisome	Collagen
ANXA1	annexin A1	Matrisome-associated	ECM-affiliated Protein
LAMA4	laminin subunit alpha 4	Core matrisome	ECM Glycoprotein
COL6A6	collagen type VI alpha 6	Core matrisome	Collagen
VWF	von Willebrand factor	Core matrisome	ECM Glycoprotein
ABI3BP	ABI family member 3 binding protein	Core matrisome	ECM Glycoprotein
TNXB	tenascin XB	Core matrisome	ECM Glycoprotein

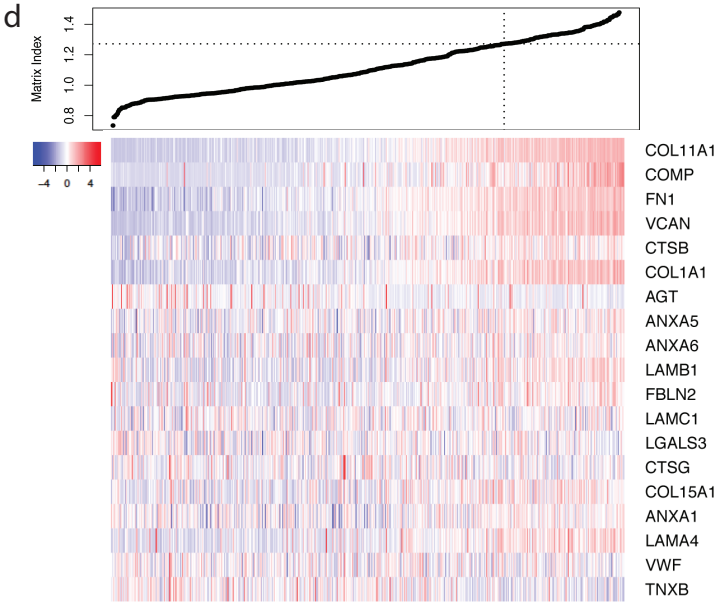
b



c

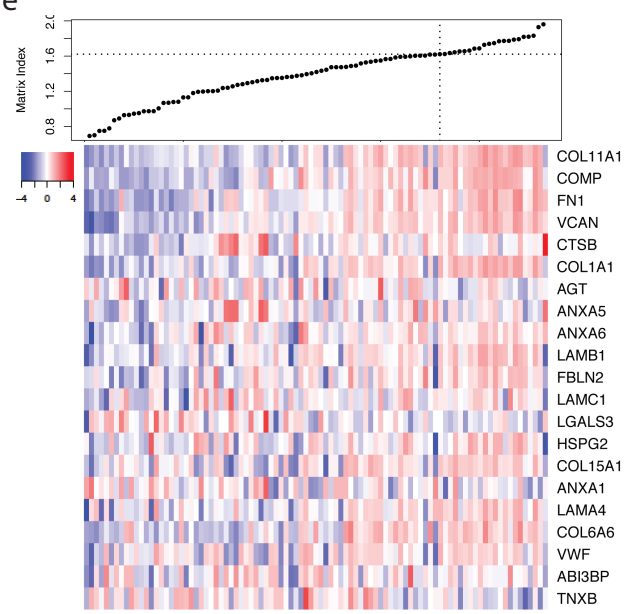


d



TCGA OV Affy u133a 564 samples

e



ICGC OV RNA-seq 93 samples

f

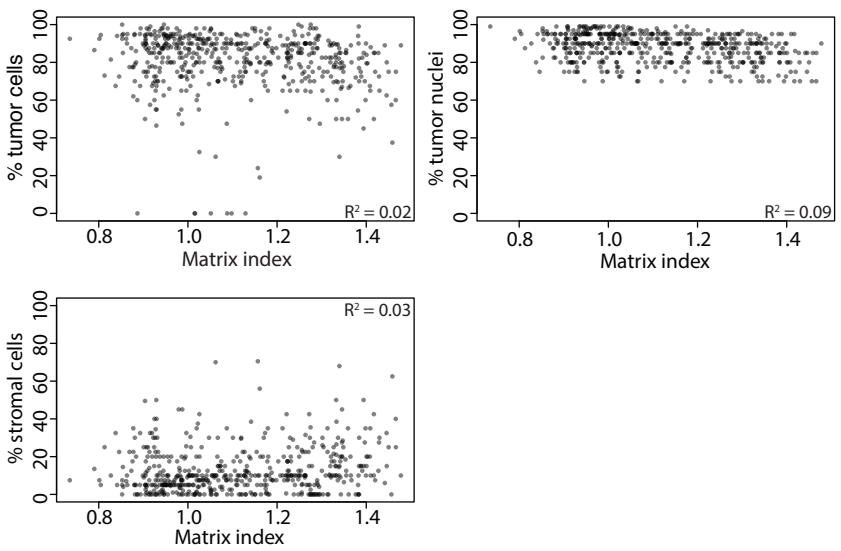


Figure S5. The matrix index signature. Related to Figure 5. a) Description of gene, matrisome category and class of the 22-matrix molecules. b) Transcription factors with experimental evidence of binding to the promoter region of matrix index genes were extracted from the ChEA database. Clustergraphs indicate transcription factor binding in red. Rows correspond to matrix index genes and columns to transcription factors. Top image shows transcription factors binding to at least 7 genes from matrix index. Rows and columns are ordered by decreasing number of hits. c) Kaplan-Meier survival curve with overall survival divided by high or low matrix index derived from the present study's transcriptomic dataset. d, e) Matrix index values and expression heatmap of matrix index genes detected across patient samples of the d) TCGA OV Affy u133a and e) ICGC OV RNA-seq datasets. Dotted lines denote the cut-off value of high and low index patient groups. f) Scatterplots of the percent tumor cells, tumor nuclei and stromal cells from the TCGA OV Affy u133a clinical data (the average of TOP and BOTTOM specimen values were used) versus matrix index (N = 564).

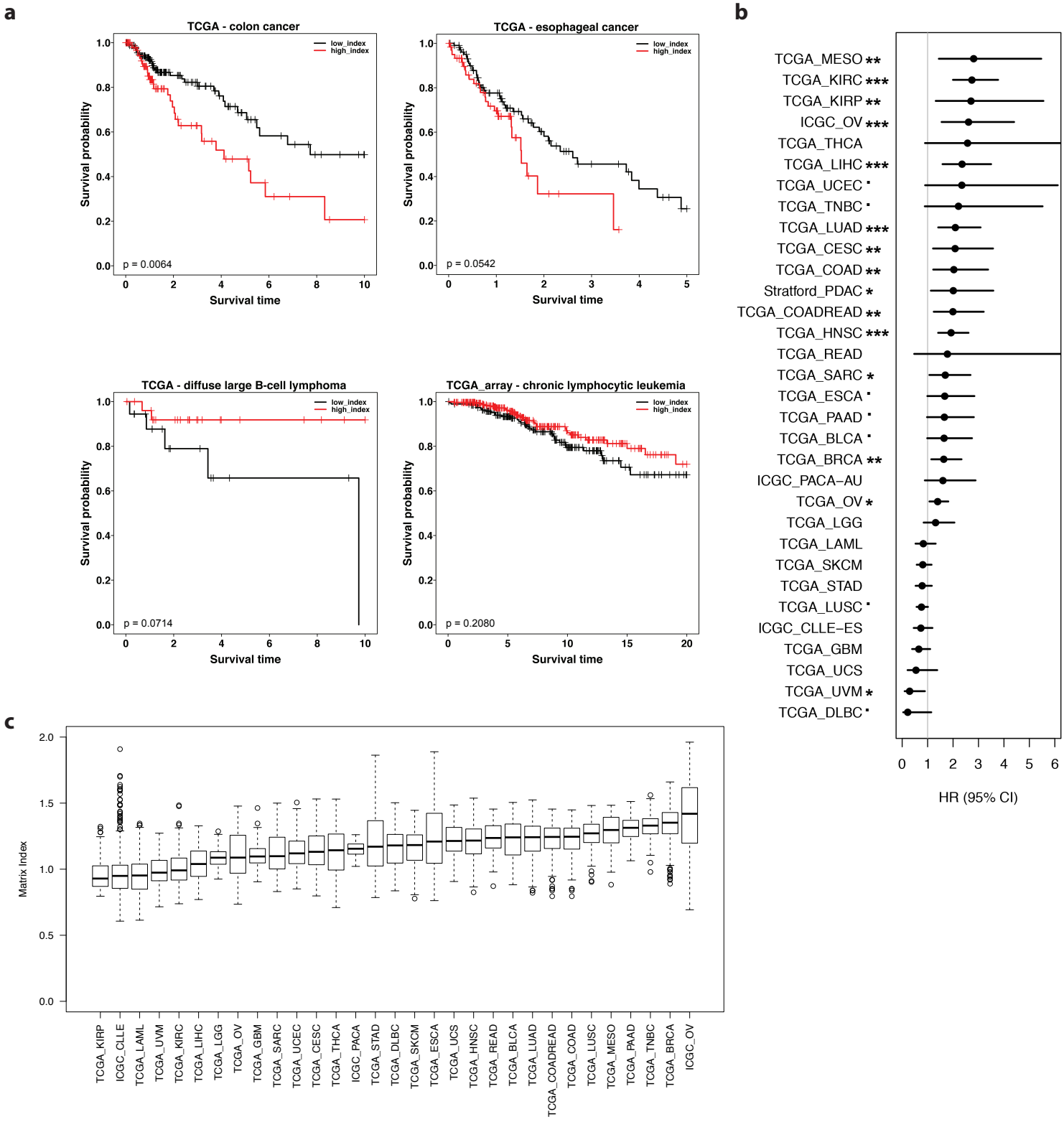

a
TCGA - colon cancer
low_index
high_index
Survival probability
p = 0.0064
Survival time
TCGA - esophageal cancer
p = 0.0542
TCGA - diffuse large B-cell lymphoma
p = 0.0714
TCGA_array - chronic lymphocytic leukemia
p = 0.2080
b
TCGA_MESO **
TCGA_KIRC ***
TCGA_KIRP **
ICGC_OV ***
TCGA_THCA
TCGA_LIHC ***
TCGA_UCEC .
TCGA_TNBC .
TCGA_LUAD ***
TCGA_CESC **
TCGA_COAD **
Stratford_PDAC *
TCGA_COADREAD **
TCGA_HNSC ***
TCGA_READ
TCGA_SARC *
TCGA_ESCA .
TCGA_PAAD .
TCGA_BLCA .
TCGA_BRCA **
ICGC_PACA-AU
TCGA_OV *
TCGA_LGG
TCGA_LAML
TCGA_SKCM
TCGA_STAD
TCGA_LUSC .
ICGC_CLLE-ES
TCGA_GBM
TCGA_UCS
TCGA_UVM *
TCGA_DLBC .
HR (95% CI)
c
Matrix Index
TCGA_KIRP
ICGC_CLLE
TCGA_LAML
TCGA_UVM
TCGA_KIRC
TCGA_LIHC
TCGA_LGG
TCGA_OV
TCGA_GBM
TCGA_SARC
TCGA_UCEC
TCGA_CESC
TCGA_THCA
ICGC_PACA
TCGA_STAD
TCGA_DLBC
TCGA_SKCM
TCGA_ESCA
TCGA_UCS
TCGA_HNSC
TCGA_READ
TCGA_BLCA
TCGA_LUAD
TCGA_COADREAD
TCGA_COAD
TCGA_LUSC
TCGA_MESO
TCGA_PAAD
TCGA_TNBC
TCGA_BRCA
ICGC_OV

Figure S6. The matrix index in other cancers. Related to Figure 6. a) Kaplan-Meier survival curves with overall survival from the indicated datasets divided by high or low matrix index. The x-axis is in the unit of years. b) Univariate hazard ratio (HR, with 95% CI) derived from a Cox proportional hazards model across cancer types using the matrix index. In each cancer, patients were split into high and low index groups, and their association with the overall survival (OS) was tested. The asterisks represent the significance in the KM analysis between the high- and low-index groups (***$p < 0.001$, **$p < 0.01$, *$p < 0.05$ and ▪$0.05 < p < 0.1$). HR > 1 means that high index is inversely correlated with OS, while HR < 1 means high index positively correlated OS. c) Distribution of matrix index across cancer datasets by boxplots.

Supplementary Methods

RNA Sequencing and analysis

RNA-Seq was performed by Oxford Gene Technology (Benbroke, UK) to ~42x mean depth on the Illumina HiSeq2500 platform, strand-specific, generating 101bp paired-end reads, as previously described (1). RNA-Seq reads were mapped to the human genome (hg19, Genome Reference Consortium GRCh37) using RSEM version 1.2.4 (2) in dUTP strand-specific mode. Bowtie version 0.12.7 (3) was used to perform the mapping as part of the RSEM pipeline. The number of reads aligned to the exonic region of each gene was counted based on Ensembl annotations. Only genes that achieved at least 10 reads per sample were kept. Log_2 counts per million (cpm) were calculated using the edgeR package (version 3.8.6) (4). RNA-Seq data have been deposited in Gene Expression Omnibus (GEO) under the accession number GSE71340.

Quantitative Proteomics

Enrichment for ECM-component. The ECM component was enriched from frozen whole tissue sections (20 x 30 μm sections, approximately 40-50 mg of tissue) as previously described (5) using a CMNCS extraction kit (Stratech). Briefly, tissue sections were homogenized in buffer C (250 μL per sample) by vortexing for 2 min per sample then incubating for 20 min, 4^{o}C, with agitation. The samples were centrifuged at 18000 g for 20 min at 4^{o}C and the supernatants were stored at -20^{o}C. This fraction was analyzed for cytokine and chemokine content using the mesoscale discovery platform (see separate method section below)**.** The samples were then washed with buffer W (300 μL per sample), quickly vortexed and then centrifuged at 18000 g for 20 min, 4^{o}C. The supernatants were removed and the pellets resuspended in buffer N (150 μL per sample), incubated for 20 min, 4^{o}C, with agitation and centrifuged at 18000 g for 20 min, 4^{o}C. Supernatants were discarded and this step was repeated. Pellets were then resuspended and well-mixed in buffer M (100 μL per sample), incubated for 20 min, 4^{o}C, with agitation and then centrifuged at 18000 g for 20 min, 4^{o}C. The supernatants were discarded and the pellets were then resuspended and well-mixed in buffer CS (200 μL per sample, pre-heated at 37^{o}C), incubated for 20 min at room temperature, with agitation and centrifuged at 18000 g for 20 min, 4^{o}C. The supernatants were discarded and the pellets resuspended and well-mixed in buffer C (150 μL per sample), incubated for 20 min, 4^{o}C, with agitation and centrifuged at 18000 g for 20 min, 4^{o}C. The pellets that remained at the end of this process were enriched for extracellular matrix (ECM) proteins and stored at -80^{o}C.

Peptide preparation. ECM enriched pellets were solubilised in 250 μL of an 8 M Urea in 20 mM HEPES (pH8) solution containing Na_3VO_4 (100 mM), NaF (0.5 M), β-Glycerol Phosphate (1 M), $Na_2H_2P_2O_7$ (0.25 M). Samples were vortexed for 30 sec and left on ice prior to sonication at 50 % intensity, 3 times for 15 sec, on ice. Tissue lysate suspensions were centrifuged at 20000 g for 10 min, 5^{o}C, and the supernatant recovered to protein low-bind tubes. BCA assay for total protein was then performed and 80 μg of protein was carried forward to the next step in urea (8 M, 200 μL per sample). Prior to trypsin digestion disulphide bridges were reduced by adding 500 mM Dithiothreitol (DTT, in 10 μL) to samples, which were then incubated at room temperature for 1 h with agitation in the dark. Free cysteines were then alkylated by adding 20 μL of a 415 mM iodacetamide solution to samples, which were again

incubated at room temperature for 1 h with agitation in the dark. The samples were then diluted 1 in 4 with 20 mM HEPES. Removal of N-glycosylation was then achieved by addition of 1500U PNGaseF (New England Biolabs), then vortexing, and incubation at 37°C for 2 h. 2 µL of a 0.8 µg/µL LysC (Pierce) per sample was then added, gently mixed and then incubated at 37°C for 2 h. Protein digestion was achieved with the use of immobilized Trypsin beads (40 µL of beads per 250 µg of protein) incubated with the derivitised protein lysate for 16 h at 37°C with shaking. Peptides were then de-salted using C-18 tip columns (Glygen). Briefly, samples were acidified with trifluoroacetic acid (1% v/v), centrifuged at 2000 g, 5 min, 5°C, before transferring the supernatant to a new microcentrifuge tube on ice. Glygen TopTips were washed with 100 % ACN (LC-MS grade) followed by 99 % H_2O (+ 1 % ACN, 0.1 % TFA) prior to loading the protein digest sample. The sample was washed with 99 % H_2O (+ 1 % ACN, 0.1 % TFA), and the desalted peptides eluted with 70/30 ACN/H2O + 0.1 % FA. The samples were dried and stored at -20 °C.

Mass Spectroscopy analysis and bioinformatics. Dried samples were dissolved in 0.1 % TFA (0.5 µg/µl) and run in a LTQ-Orbitrap XL mass spectrometer (Thermo Fisher Scientific) connected to a nanoflow ultra-high pressure liquid chromatography (UPLC, NanoAcquity, Waters). Peptides were separated using a 75 µm × 150 mm column (BEH130 C18, 1.7 µm Waters) using solvent A (0.1 % FA in LC–MS grade water) and solvent B (0.1 % FA in LC–MS grade ACN) as mobile phases. The UPLC settings consisted of a sample loading flow rate of 2 µL/min for 8 min followed by a gradient elution starting with 5 % of solvent B and ramping up to 35 % over 220 min followed by a 10 min wash at 85 % B and a 15 min equilibration step at 1 % B. The flow rate for the sample run was 300 nL/min with an operating back pressure of about 3800 psi. Full scan survey spectra (*m*/*z* 375–1800) were acquired in the Orbitrap with a resolution of 30000 at *m*/*z* 400. A data dependent analysis (DDA) was employed in which the five most abundant multiply charged ions present in the survey spectrum were automatically mass-selected, fragmented by collision-induced dissociation (normalized collision energy 35 %) and analysed in the LTQ. Dynamic exclusion was enabled with the exclusion list restricted to 500 entries, exclusion duration of 30 sec and mass window of 10 ppm.
MASCOT search was used to generate a list of proteins. Peptide identification was performed by searching against the SwissProt database (version 2013-2014) restricted to human entries using the Mascot search engine (v 2.5.0, Matrix Science, London, UK). The parameters included trypsin as the bdigestion enzyme with up to two missed cleavages permitted, carbamidomethyl (C) as a fixed modification and Pyro-glu (*N*-term), Oxidation (M) and Phospho (STY) as variable modifications. Datasets were searched with a mass tolerance of ±5 ppm and a fragment mass tolerance of ±0.8 Da.
A MASCOT score cut-off of 50 was used to filter false-positive detection to a false discovery rate below 1 %. PESCAL was used to obtain peak areas in extracted ion chromatograms of each identified peptide (6) and protein abundance determined by the ratio of the sum of peptide areas of a given protein to the sum of all peptide areas. This approach for global protein quantification absolute quantification, described in (6), is similar to intensity based protein quantification (iBAQ) (7), and total protein abundance (TPA)(8). Proteomic data are available via the PRIDE database accession number PXD004060.

Mechanical characterization

Flat-punch Indentation. Mechanical characterisation was performed using a previously published methodology in order to measure the modulus of the tissue samples (9). The modulus provides a measure of the stiffness of the material that is independent of specimen geometry. Frozen tissue specimens (n = 32) were fully thawed at room temperature in PBS for 1 hour before testing. Indentation was performed using an Instron ElectroPuls E1000 (Instron, UK) equipped with a 10 N load cell (resolution = 0.1 mN) (Figure S1A). Specimens were indented using a stainless steel plane-ended cylindrical punch with a diameter ($Ø_i$) of 2 or 3 mm. Specimen thickness (T_s) was measured as the distance between the base of the test dish and top of the sample, each detected by applying a pre-load of 0.3-5 mN. Specimen diameter ($Ø_s$) was measured using callipers. In order to minimise errors in calculations of mechanical parameters, specimen to indenter ratios were $Ø_s$:$Ø_i \geq 4$:1 and T_s: $Ø_i \leq 2$:1 (9). Indentation was performed at room temperature with specimens fully submerged in PBS throughout testing. Tests were performed using two consecutive displacement-controlled static loading regimes on each specimen with a recovery period of 20 min between tests. Specimens were displaced to 20 % or 30 % of their measured thickness at a rate of 1 %.s^{-1} followed by a displacement-hold period to allow full sample stress-relaxation, and then an unloading phase to 0 % specimen strain. The resulting load detected from the sample was recorded. Green tissue dye was used to mark the surface area of tissue-indenter contact for later correlation of mechanics with tissue architecture (Figure S1B). After testing, specimens were snap frozen in LN_2 and stored at -80 C until further processing.

Mechanical quantification. Tissue modulus, *E*, was calculated from the obtained load-displacement experimental data with the aid of a mathematical model derived from the solution of Sneddon for the axisymmetric Boussinesq problem as shown in equation 1. Full details of this model and its validation are given in our previous study (9)

$$E = \frac{S}{2a}(1 - \nu^2) \qquad \text{(Eq. 1)}$$

The indentation stiffness, *S*, was calculated from the tangent of the slope representing 15-20% sample strain on the load-displacement curve and '*a*' is the radius of the flat-punch indenter. Poisson's ratio, *v*, was assumed to be 0.5 for all samples. Mechanical values were plotted against disease scores determined from tissue architecture analysis.

Confocal microscopy

Second harmonic generation. Paraffin embedded TMAs containing 3-6 x 1 mm tissue cores per sample were mounted in Fluoromount (Sigma, UK) and samples (n = 13) were imaged via two-photon confocal microscopy to collect second harmonic generation (SHG) illumination. Images were captured on an inverted Leica laser-scanning confocal TCS SP2 microscope (Leica) equipped with a tunable Ti:Sapphire femto-second multiphoton laser (Spectra-Physics). Specimens were illuminated at 820 nm and the resulting signal was collected in the backward scattering direction (epi), after filtration through a SP700 dichroic, using a photo-multiplier tube (PMT) set to collect SHG between 405-415 nm. The laser passed through a 63 x 1.4NA oil immersion objective with the pinhole set to maximum resulting in a laser excitation power at the specimen of 20 mW. Specimen images were acquired with a frame average of 2 and a line average of 16 at intervals of 1 µm in the z-direction each with a field of view equal to 238.1 x 238.1 µm containing 1024x1024 pixels. At least three

x 5 μm z-stacks were collected from each individual tissue core and then analysed using Image J to measure fibre orientation.

Immunohistochemical analysis

Quantification of Immune cells, α-SMA positive cells, and adipocyte diameters. TMA cores were used for immune cell counts and quantification of α-SMA positive cells and adipocyte diameters. Paraffin embedded TMAs were heated at 60 ^{0}C for 5 min followed by 2 x 5 min submersion in xylene and then a series of ethanol washes of decreasing concentration for 2 x 2 min each (100 %, 90 %, 70 %, and 50 %). Antigen retrieval was performed for 10 min using vector antigen unmasking buffer and a pressure cooker. TMAs were then washed with DAKO wash buffer followed by application of H_2O_2 for 5 min. Blocking was performed using 5 % BSA for 20 min at RT followed by incubation with primary antibody in biogenex antibody diluent for 30 min. After 3 x washes, biogenex super enhancer was added for 20 min and then washed off before addition of biogenex ss label poly-HRP for 30 min. Tissue was washed three times before addition of DAB chromagen for 3 min followed by washing to stop further DAB development. TMAs were counterstained with haematoxylin followed by washing with H_2O and ethanol solutions of increasing concentration for 2 min each (50 %, 70 %, 90 %, 100 %) and then 2 x xylene. Samples were then mounted and scanned using the 3DHISTECH Panoramic digital slide scanner. Immune cells were counted manually using Image J. The population of α-SMA positive cells was determined using Definiens software, firstly by setting a threshold and then quantifying the area of tissue expressing α-SMA to give a % SMA+ area. Adipocyte diameter was quantified on α-SMA stained TMAs using Panoramic Viewer software (3DHISTECH, Hungary) by measuring at least 100 adipocytes per sample (n = 16) to get the population mean. For samples with tumour and stromal remodelling, adipocytes that were either in contact with stroma or totally surrounded by stroma were measured. All cell analysis was plotted versus disease score determined using Definiens software analysis of haematoxylin and eosin stained TMAs.

Matrix staining. Immunohistochemical staining for ECM proteins was performed on 4 μm slides of FFPE human omentum tissue as described above.

Antibodies. The following antibodies were used for immunohistochemical analyses: anti-FOXP3 (clone 263A/E7, ab20034) from Abcam, UK; anti-CD3 (clone F7.2.38, M7254), anti-CD4 (clone 4B12, M7310), anti-CD8 (clone C8/144B, M7103), anti-CD68 (clone KP1, F7135), anti-CD45RO (clone UCHL1, M0742), anti-Ki67 (cloneMIB-1, M7240), all from Dako, UK; anti-VCAN (polyclonal, HPA004726), anti-SFRP4 (polyclonal, HPA009712), anti-COL11A1 (polyclonal, HPA052246) anti-TNC (polyclonal, HPA004823), anti-COL1A1 (polyclonal, HPA011795), anti-FN1 (polyclonal, F3648), anti-IL16 (polyclonal, HPA018467), anti-actin, α-smooth muscle (clone 1A4, A2547), all from Sigma, UK. Anti-CTSB (ab125067), and anti-COMP (ab11056), both from Abcam.

Tissue arrays. All tissues were obtained from patients with full written informed consent. Breast tissues were obtained through the Breast Cancer Campaign (now Breast Cancer Now) Tissue Bank (NRES Cambridgeshire 2 REC 10/H0308/48), and Barts Cancer Institute Breast Tissue Bank (NRES East of England 15/EE/0192). DLBCL lymph node tissues were obtained through the Local Regional Ethics

Boards (05/Q0605/140). Pancreatic tissues were obtained through the City and East London REC 07/H0705/87. Tissue microarrays (TMA) were prepared from paraffin blocks with triplicate 1mm cores taken from each biopsy material.

RNA *in situ* hybridization

Four µm sections of FFPE human omentum samples were heated at 60^{o}C for 1 h before deparaffinization in two changes of xylene for 5 min, followed by two changes of 100 % ethanol for 1 min. Slides were then treated with the pre-packaged hydrogen peroxide for 10 min and boiled for 15 min in the target retrieval reagent. The tissue was then dried in ethanol, outlined using a hydrophobic barrier pen and left at room temperature overnight. Slides were then incubated in the protease reagent at 40^{o}C in a HyBEZ Hybridization System (Advanced Cell Diagnostics Inc. USA) for 30 min, before a 2 h incubation at 40^{o}C with the gene-specific probe. The AMP 1-6 reagents were all subsequently hybridized at 40^{o}C or RT, 30 or 15 min as specified in the manufacturer's instructions. Labelled mRNAs were visualized using the included DAB reagent for 10 min, then counterstained for 2 min using 50 % Gill's haematoxylin followed by 3 dips in 0.02 % ammonia water. Counterstained slides were dehydrated using 70 % and 95 % ethanol then cleared in xylene before mounting coverslips using DPX.

PLS regression

Model fitting. PLS regression was implemented using the R package pls (version 2.4-3) (10). Briefly, the PLS algorithm consists of the following steps: first, the data is standardized by centering to column mean zero and scaled to unit variance (dividing columns by their standard deviation), resulting in a matrix $\boldsymbol{X}$ (genes or proteins) and vector $\boldsymbol{y}$ (disease score or tissue modulus). Second, using the linear dimension reduction $\boldsymbol{t} = \boldsymbol{Xw}$, the p predictors (genes or proteins) in $\boldsymbol{X}$ are mapped onto latent components in $\boldsymbol{t}$. The weights $\boldsymbol{w}$ are chosen with the response $\boldsymbol{y}$ explicitly taken into account, so that the predictive performance is maximal. Next, $\boldsymbol{y}$ is regressed by ordinary least squares against the latent components $\boldsymbol{t}$ (also known as X-scores) to obtain the loadings $\boldsymbol{q}$. Subsequently, the PLS estimate of the coefficients in $y = \beta X +$ error is computed from estimates of the weight matrix $\boldsymbol{w}$ and the $\boldsymbol{y}$-loadings via $\boldsymbol{\beta} = \boldsymbol{wq}$.

Prior to model fitting the data was randomly split into a "training" set of 18 samples (approximately 2/3 of data) leaving the remaining samples as a "test" set. Both training and test sets included samples ranging from low to high disease score. Using the training set a PLS model was initially fitted using 10 components with leave-one-out cross-validation. The validation results were expressed as root mean squared error of prediction (RMSEP).

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

where n is the total number of samples, y_i is the actual value of $\boldsymbol{y}$ (disease score or stiffness) for sample i and $\hat{y}_i$ the $\boldsymbol{y}$-value for sample i predicted with the model under evaluation.

The estimated RMSEPs were then plotted as functions of the number of components. The components that corresponded to the first local minimum RMSEP were chosen as

optimal for the model. The fitted model was then used to predict the response values of the test set of samples. Since we knew the true response values of the test data we were able to calculate the RMSEP, which was typically very similar to the cross-validated estimate of the training data.

Estimating confidence of model predictions and assessing the significance of model performance. In order to determine the performance of the constructed PLS models over multiple iterations of model building and testing, bootstrapping was carried out by iterating 1000 times through the whole process of random selection of training and test datasets, model fitting and recording predicted values and RMSEP. By this process, frequency distributions for the overall test accuracies (RMSEPs) and the predicted response values were obtained.

We then examined the statistical significance of the performance of the constructed PLS regression models compared to random chance using permutation testing. The data was randomly shuffled across samples within each variable. This process destroyed the correlations in the data while retaining the original variance of the variables. Then the process of model building, testing prediction accuracy by RMSEP and bootstrapping was repeated using the permuted datasets. Student's t-test was then used comparing the difference in model performance over RMSEP values obtained from permutation testing and RMSEP values obtained from the original datasets to determine whether the model was statistically significant. For all models that were used throughout the study $P_{realvspermuted} < 2.2x10^{-16}$.

PLS-ranking of variables and cut-off values. The loading weights of the first component, which explained >70 % of variance, were used to rank variables (genes or proteins) according to their contribution to the model (11,12). Inherently this vector is calculated to maximize covariance of Xw_1 with y. To determine which variables made a significant contribution to the model, variables were removed from the model in order of weight until the bootstrapped RMSEP exceeded that of permutation testing.

Matrix index and its clinical association across cancer types

Based on the 22 matrisome genes, we defined "matrix index" as the ratio of the mean expression of the genes positively correlated with disease score to that of the remaining negatively correlated genes. We first tested the clinical association and prognostic potential of this matrix index in two large ovarian cancer datasets from the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (13), as ICGC_OV and TCGA_OV. For the ICGC_OV set, raw read counts for all annotated Ensembl genes across 93 primary tumors were extracted from the exp_seq.OV-AU.tsv.gz file in the ICGC data repository Release 20 (http://dcc.icgc.org). Only genes that achieved at least one read count per million reads (cpm) in at least ten samples were selected, with these criteria producing 18,698 filtered genes in total. After applying scale normalization, read counts were converted to $\log_2$ (cpm) using the voom function (14). Clinical information (e.g., overall survival (OS)) was extracted from the donor.OV-AU.tsv.gz file. For the TCGA_OV set, the normalized gene expression data profiled by Affymetrix U133a 2.0 Array and clinical data were downloaded from UCSC Cancer Browser (http://genome-cancer.ucsc.edu/), version 2015-02-24. Only primary tumors were selected for further analysis, leading to 564 primary samples with both expression and OS data available.

Expression values for the matrisome genes were extracted and matrix index was calculated for each sample. For each dataset, the high and low index groups were determined using the method described previously (15). Briefly, each percentile of index between lower and upper quartiles was used in the Cox proportional hazards (Coxph) regression analysis and the best performing threshold of percentile associated with OS was determined. Survival modeling and Kaplan-Meier (KM) analysis was undertaken using R "survival" package. OS was defined as time from diagnosis to death, or to the last follow-up date for survivors. We further assessed the prognostic potential of matrix index using the multivariate analysis, accounting for age, tumor stage, grade and primary therapy outcome success. Note that for ICGC_OV set, only age and tumor stage information were available. Hazard ratio (HR) and 95% confidence interval (CI), as well as associated *p*-values for matrix index at the best performing threshold were derived from the Coxph regression model for both uni- and multivariate analyses.

We then benchmarked the performance of matrix index in prognostics against other existing ovarian cancer signatures (including the 193-gene signature from TCGA) and other relevant stroma and immune signatures extracted from literature on the TCGA_OV set (Table S22). For expression-based signatures, firstly consensus clustering, using ConsensusClusterPlus R package(16), was performed based on normalized expression values to split patients. After sample grouping, both uni- and multivariate survival analyses with OS were subsequently conducted using the Coxph regression. The prognostic value for the matrisome genes solely based on expression clustering was also assessed in this way.

We further expanded the survival analysis of matrix index into other cancer types and datasets, including additional 33 TCGA cancer sets and 2 ICGC sets (Table S23). For these TCGA sets, the gene expression Illumina HiSeqV2 RNA-seq normalized data were used, available from UCSC Cancer Browser. For the ICGC chronic lymphocytic leukemia dataset, ICGC_CLLE-ES, the expression array data was used. The two pancreatic cancer sets, ICGC_PACA-AU and Stratford_PDAC, were based on data previously described (17). In total, we assessed the prognostic values of matrix index in 38 cancer sets including the two ovarian sets. Six datasets were further excluded from our results due to the large HR 95% CI, resulting in final 32 valid datasets (Table S23). The same survival analysis protocol was applied for each dataset as above. For those datasets, pathogenic T-stage was used when tumor grade information was unavailable, and target molecular therapy or radiation therapy (in the "yes" or "no" category) was used if primary therapy outcome success information was not available.

Additional information on statistical analyses

All graphics and statistical analyses were performed in the statistical programming language R (version 3.1.3). For PLS regression models, a fourth square root transformation was applied to the proteomics and biomechanical data. Univariate correlations were calculated using spearman's correlation or pearson's correlation applied on linear, log or square-root transformed data. Overrepresented Gene Ontology annotations from the differentially expressed genes were identified by a modified Fisher's exact test using the web-based tool PANTHER (version 10) (18). Enrichment p-values were calculated with a modified Fisher's exact test and Bonferroni multiple testing correction. Identification of gene clusters with highly

correlative expression profiles was carried out by hierarchical clustering using the Pvclust R package (19).

Supplementary References

1. Bohm S, Montfort A, Pearce OM, Topping J, Chakravarty P, Everitt GL*, et al.* Neoadjuvant Chemotherapy Modulates the Immune Microenvironment in Metastases of Tubo-Ovarian High-Grade Serous Carcinoma. *Clin Cancer Res* **2016**;22:3025-36
2. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* **2011**;12:323
3. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **2009**;10:R25
4. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **2010**;26:139-40
5. Naba A, Clauser KR, Hoersch S, Liu H, Carr SA, Hynes RO. The matrisome: in silico definition and in vivo characterization by proteomics of normal and tumor extracellular matrices. *Molecular & cellular proteomics : MCP* **2012**;11:M111 014647
6. Cutillas PR, Vanhaesebroeck B. Quantitative profile of five murine core proteomes using label-free functional proteomics. *Mol Cell Proteomics* **2007**;6:1560-73
7. Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J*, et al.* Global quantification of mammalian gene expression control. *Nature* **2011**;473:337-42
8. Wisniewski JR, Ostasiewicz P, Dus K, Zielinska DF, Gnad F, Mann M. Extensive quantitative remodeling of the proteome between normal colon tissue and adenocarcinoma. *Mol Syst Biol* **2012**;8:611
9. Delaine-Smith RM, Burney S, Balkwill FR, Knight MM. Experimental validation of a flat punch indentation methodology calibrated against unconfined compression tests for determination of soft tissue biomechanics. *J Mech Behav Biomed Mater* **2016**;60:401-15
10. Mevik BH, Wehrens R. The pls package: Principal component and partial least squares regression in R. *Journal of Statistical Software* **2007**;18:1-23
11. Mehmood T, Liland KH, Snipen L, Saebo S. A review of variable selection methods in Partial Least Squares Regression. *Chemometr Intell Lab* **2012**;118:62-9
12. Johansson D, Lindgren P, Berglund A. A multivariate approach applied to microarray data for identification of genes with cell cycle-coupled transcription. *Bioinformatics* **2003**;19:467-73
13. TCGA. Integrated genomic analyses of ovarian carcinoma. *Nature* **2011**;474:609-15
14. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **2014**;15:R29
15. Mihaly Z, Kormos M, Lanczky A, Dank M, Budczies J, Szasz MA*, et al.* A meta-analysis of gene expression-based biomarkers predicting outcome after tamoxifen treatment in breast cancer. *Breast cancer research and treatment* **2013**;140:219-32
16. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **2010**;26:1572-3

17. Haider S, Wang J, Nagano A, Desai A, Arumugam P, Dumartin L*, et al.* A multi-gene signature predicts outcome in patients with pancreatic ductal adenocarcinoma. *Genome Med* **2014**;6:105
18. Mi H, Muruganujan A, Thomas PD. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res* **2013**;41:D377-86
19. Suzuki R, Shimodaira H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **2006**;22:1540-2