iScience, Volume *July 24*

Supplemental Information

Immune Cell Associations with Cancer Risk

Luis Palomero, Ivan Galván-Femenía, Rafael de Cid, Roderic Espín, Daniel R. Barnes, CIMBA, Eline Blommaert, Miguel Gil-Gil, Catalina Falo, Agostina Stradella, Dan Ouchi, Albert Roso-Llorach, Concepció Violan, María Peña-Chilet, Joaquín Dopazo, Ana Isabel Extremera, Mar García-Valero, Carmen Herranz, Francesca Mateo, Elisabetta Mereu, Jonathan Beesley, Georgia Chenevix-Trench, Cecilia Roux, Tak Mak, Joan Brunet, Razq Hakem, Chiara Gorrini, Antonis C. Antoniou, Conxi Lázaro, and Miquel Angel Pujana

Supplementary figures legends

Fig. S1. Evaluation of immune/stromal cell tissue content estimates in relation to two other methods. Related to Figure 1.

(A) Heatmap showing the correlations (Spearman's ρ) between Consensus^{TME}-based values and analogous TIMER cell type estimates.

(B) Heatmap showing the correlations (Spearman's ρ) between Consensus^{TME}-based values and analogous MCP-counter cell type estimates.

Fig. S2. Evaluation of immune/stromal cell tissue content estimates in relation to independent leukocyte estimates. Related to Figure 1. Heatmap showing the correlations (Spearman's ρ) between Consensus^{TME}-based values and independent leukocyte estimates using the approach of Taylor et al. (2018).

Fig. S3. Evaluation of immune/stromal cell tissue content estimates in relation to aneuploidy scores. Related to Figure 1.

(A) Heatmap showing the correlations (Spearman's ρ) between Consensus^{TME}-based values and aneuploidy scores (Taylor et al., 2018) across major cancer types.

(B) Heatmap showing the correlations (Spearman's ρ) between Consensus^{TME}-based values and aneuploidy scores (Taylor et al., 2018) across in BRCA subtypes, which show positive correlations in claudin-low.

Fig. S4. Differences of inferred immune/stromal cell content between primary tumors with low and high levels of *CD274/PDL1* expression.

Related to Figure 1. The graphs show the median cell content value in each group and the significance of the difference (Wilcoxon test *P* value).

Fig. S5. Differences of inferred immune/stromal cell content between primary tumors with low and high levels of *CD279/PDCD1* expression. Related to Figure 1. The graphs show the median cell content value in each group and the significance of the difference (Wilcoxon test *P* value).

Fig. S6. Correlations between inferred blood immune cell contents and measures from fluorescence-activated cell sorting in blood samples. Related to Figure 1. Forest plot showing correlation estimates and 95% CIs of each inferred cell type (data from whole blood samples of healthy adults; n = 12, GEO GSE127813).

Fig. S7. Correlations between inferred immune/stromal cell tissue contents and single cells used to generate pseudo-bulk breast tumors. Related to Figure 1. Each panel shows the correlation between immune cell signature scores (Y-axis) and the number of cells (X-axis) used to generate 100 pseudo-bulk breast tumors (data from Gene Expression Omnibus reference GSE75688). The trend lines, Spearman's correlations and *P* values are shown.

Fig. S8. Correlations between immune/stromal cell tissue contents and expression of immune benchmark genes. Related to Figure 1. Top panel, distribution of PCCs using data from normal TCGA tissue. Bottom panel,

distribution of PCCs using data from primary tumors of TCGA. Mean PCCs and 95% CIs are shown.

Fig. S9. Correlations between immune cell signatures and pathway signaling-inferred activities. Related to Figure 1. Unsupervised clustering of the correlation coefficients between inferred cell contents (Y-axis) and KEGG pathway activities (X-axis). Differentiated clusters in normal tissue are marked by red-outlined rectangles.

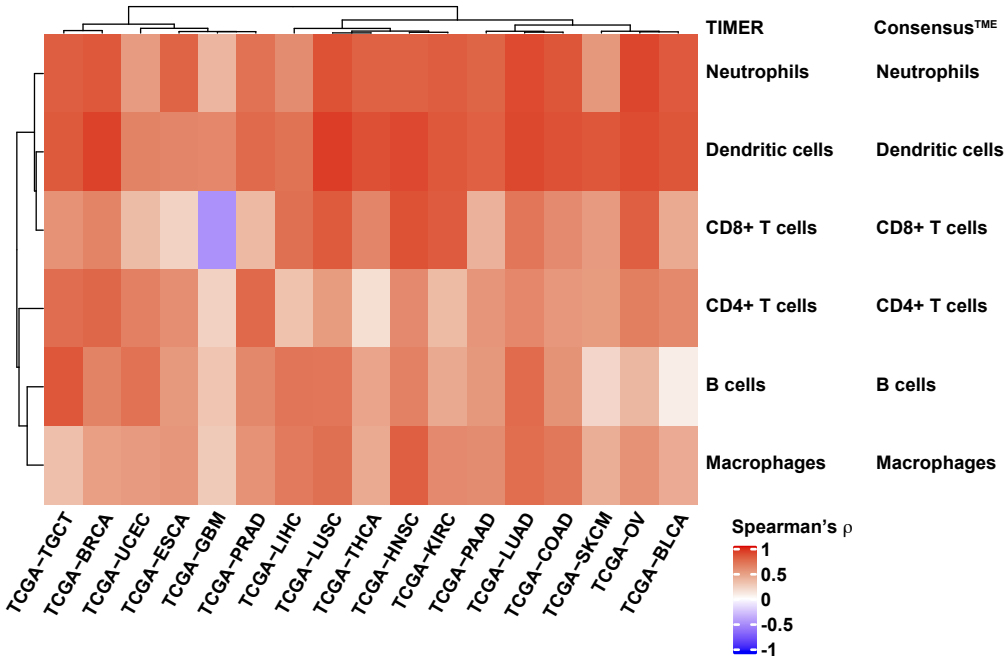
Fig. S10. Gene targets of eQTL recognized in isQTLs are frequently correlated with the corresponding immune/stromal cell signatures. Related to Figure 2. Distributions of random gene sets (same gene set size and equivalent comparisons for each signature and TCGA setting) relative to the number of significant correlations between eQTL-target and immune/stromal signatures. Left- and right-hand panels show results for the first and second isQTL sets presented in the main text, respectively. Empirical test probabilities are shown.

Fig. S11. Minimal correlation estimates to detect significant signature-PRS associations. Related to Figure 3. Left and right panels show the lowest correlations required in each normal and primary tumor setting, respectively, to detect nominal ($P < 0.05$) associations given the TCGA sample sizes.

Fig. S12. LUAD and LUSC PRS correlations with NK cell content. Related to Figure 3. Top panels, positive correlations between NK cell content in primary tumors of LUAD and LUSC, and the corresponding PRSs. The adjusted-R^2 and *P* values of the linear regression model are shown. Bottom panels, correlation trends of patients stratified by smoking status, as depicted in the insets. The estimate for LUAD cases classified as current smokers was found to be significantly less than zero (r = -0.12, *P* = 0.012).

Figure S1

A


TIMER
Consensus^TME
Neutrophils
Neutrophils
Dendritic cells
Dendritic cells
CD8+ T cells
CD8+ T cells
CD4+ T cells
CD4+ T cells
B cells
B cells
Macrophages
Macrophages
TCGA-TGCT
TCGA-BRCA
TCGA-UCEC
TCGA-ESCA
TCGA-GBM
TCGA-PRAD
TCGA-LIHC
TCGA-LUSC
TCGA-THCA
TCGA-HNSC
TCGA-KIRC
TCGA-PAAD
TCGA-LUAD
TCGA-COAD
TCGA-SKCM
TCGA-OV
TCGA-BLCA
Spearman's ρ
1
0.5
0
-0.5
-1

B

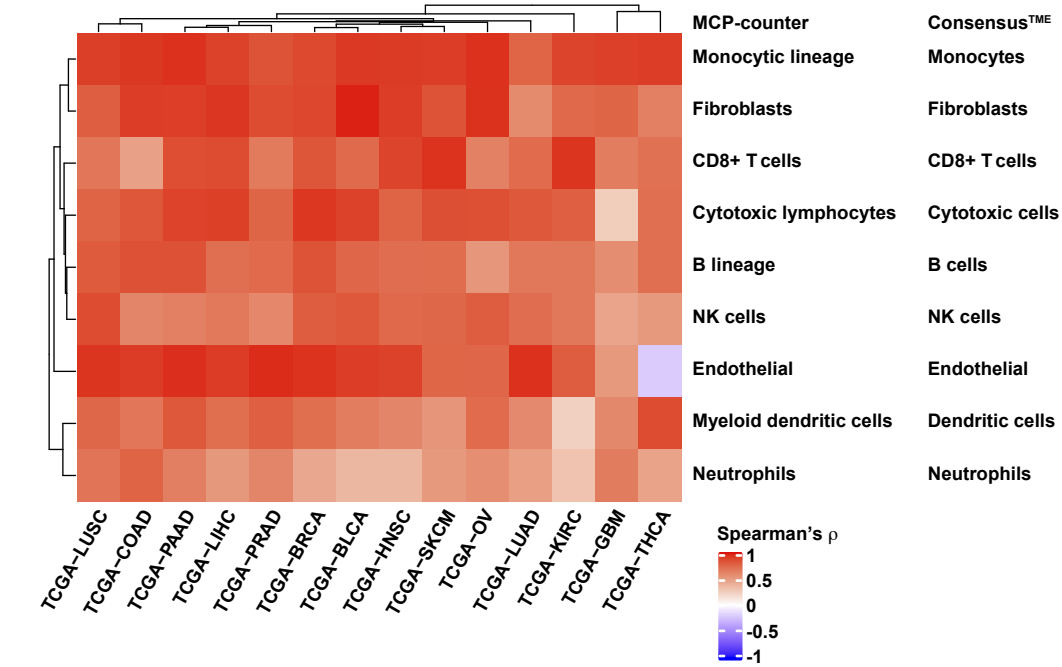

MCP-counter
Consensus^TME
Monocytic lineage
Monocytes
Fibroblasts
Fibroblasts
CD8+ T cells
CD8+ T cells
Cytotoxic lymphocytes
Cytotoxic cells
B lineage
B cells
NK cells
NK cells
Endothelial
Endothelial
Myeloid dendritic cells
Dendritic cells
Neutrophils
Neutrophils
TCGA-LUSC
TCGA-COAD
TCGA-PAAD
TCGA-LIHC
TCGA-PRAD
TCGA-BRCA
TCGA-BLCA
TCGA-HNSC
TCGA-SKCM
TCGA-OV
TCGA-LUAD
TCGA-KIRC
TCGA-GBM
TCGA-THCA
Spearman's ρ
1
0.5
0
-0.5
-1

Figure S2

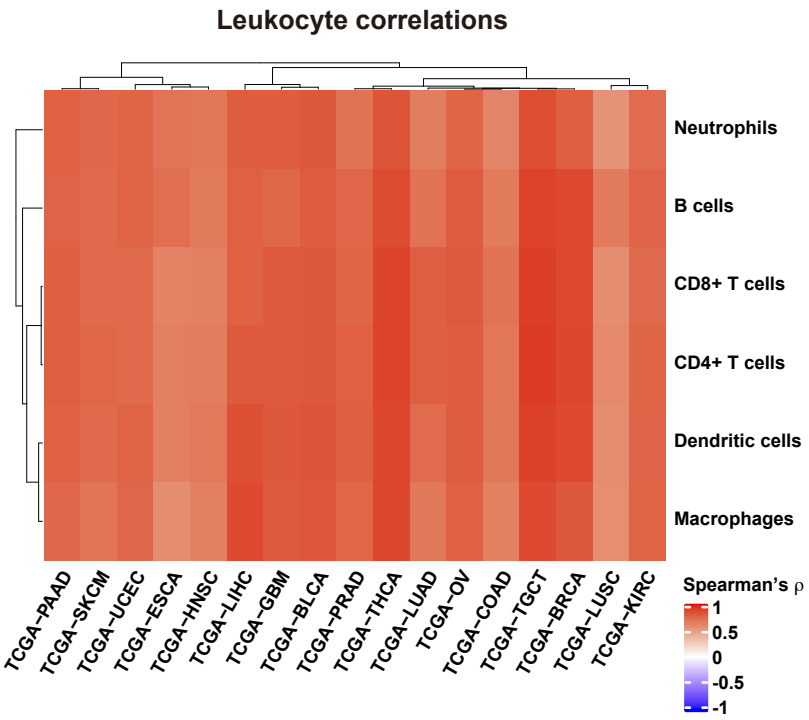

Leukocyte correlations
Neutrophils
B cells
CD8+ T cells
CD4+ T cells
Dendritic cells
Macrophages
TCGA-PAAD
TCGA-SKCM
TCGA-UCEC
TCGA-ESCA
TCGA-HNSC
TCGA-LIHC
TCGA-GBM
TCGA-BLCA
TCGA-PRAD
TCGA-THCA
TCGA-LUAD
TCGA-OV
TCGA-COAD
TCGA-TGCT
TCGA-BRCA
TCGA-LUSC
TCGA-KIRC
Spearman's ρ
1
0.5
0
-0.5
-1

Figure S3

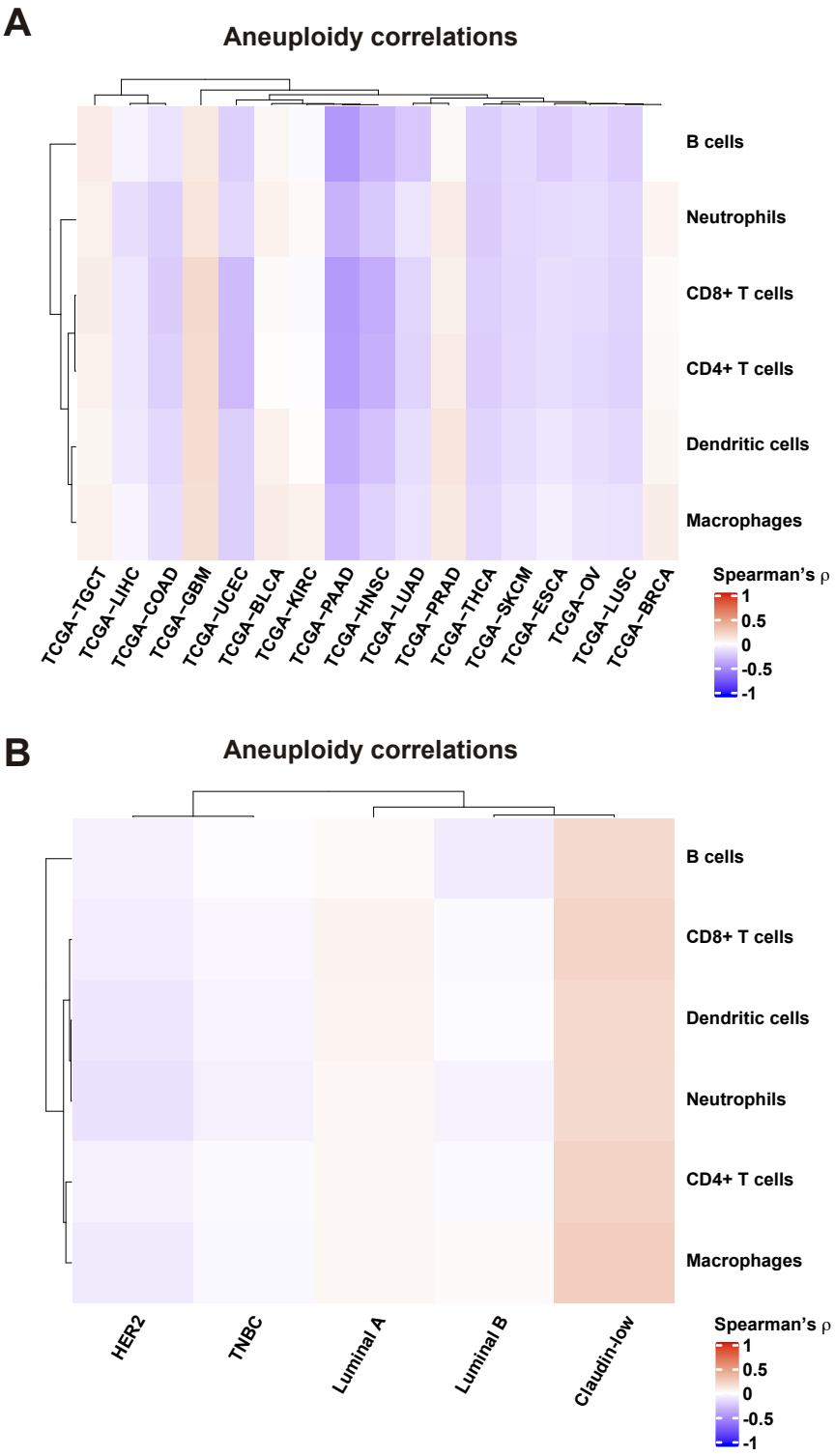

A
Aneuploidy correlations
B cells
Neutrophils
CD8+ T cells
CD4+ T cells
Dendritic cells
Macrophages
TCGA-TGCT
TCGA-LIHC
TCGA-COAD
TCGA-GBM
TCGA-UCEC
TCGA-BLCA
TCGA-KIRC
TCGA-PAAD
TCGA-HNSC
TCGA-LUAD
TCGA-PRAD
TCGA-THCA
TCGA-SKCM
TCGA-ESCA
TCGA-OV
TCGA-LUSC
TCGA-BRCA
Spearman's ρ
1
0.5
0
-0.5
-1
B
Aneuploidy correlations
B cells
CD8+ T cells
Dendritic cells
Neutrophils
CD4+ T cells
Macrophages
HER2
TNBC
Luminal A
Luminal B
Claudin-low
Spearman's ρ
1
0.5
0
-0.5
-1

Figure S4

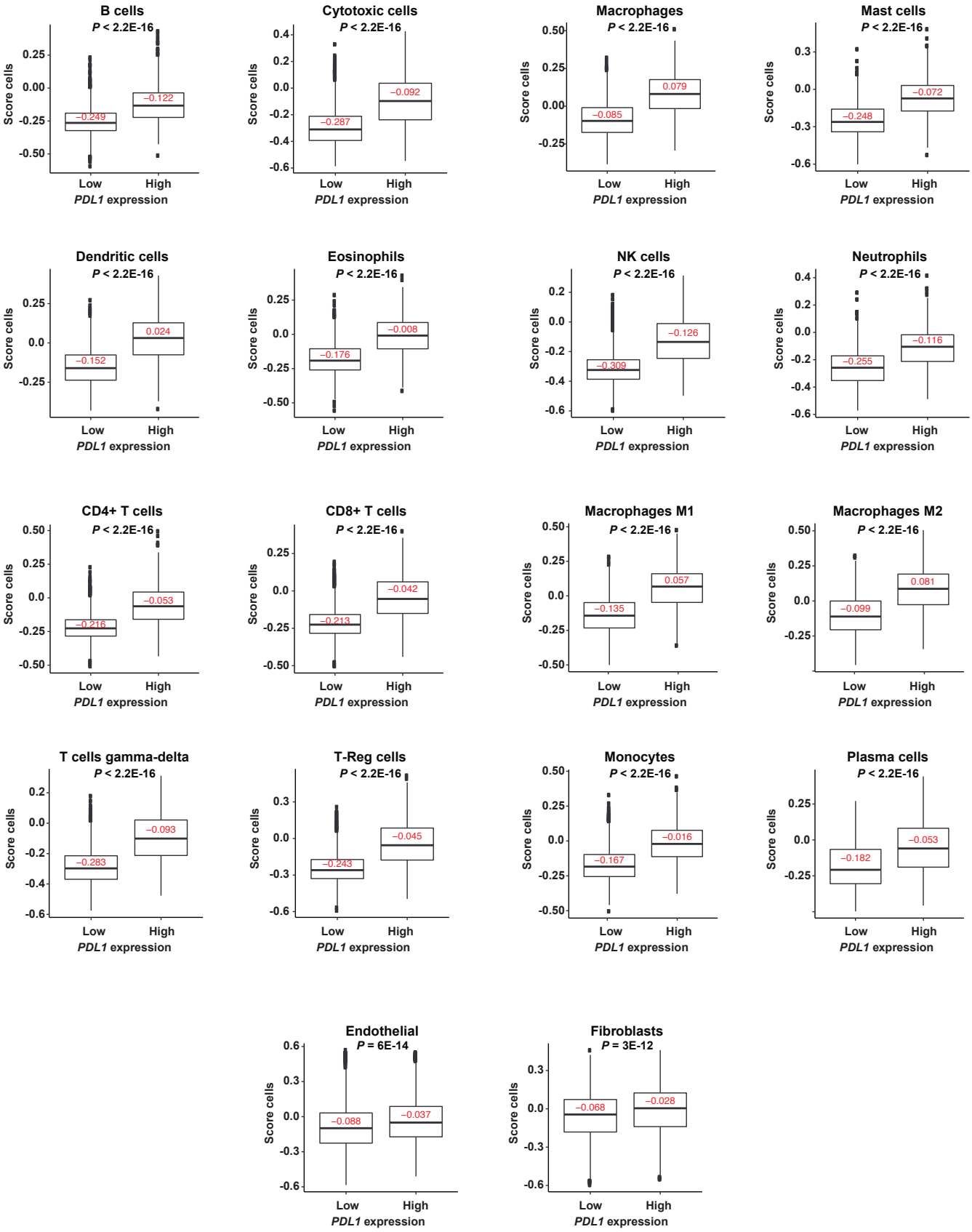

B cells
P < 2.2E-16
−0.249
−0.122
Cytotoxic cells
P < 2.2E-16
−0.287
−0.092
Macrophages
P < 2.2E-16
−0.085
0.079
Mast cells
P < 2.2E-16
−0.248
−0.072
Dendritic cells
P < 2.2E-16
−0.152
0.024
Eosinophils
P < 2.2E-16
−0.176
−0.008
NK cells
P < 2.2E-16
−0.309
−0.126
Neutrophils
P < 2.2E-16
−0.255
−0.116
CD4+ T cells
P < 2.2E-16
−0.216
−0.053
CD8+ T cells
P < 2.2E-16
−0.213
−0.042
Macrophages M1
P < 2.2E-16
−0.135
0.057
Macrophages M2
P < 2.2E-16
−0.099
0.081
T cells gamma-delta
P < 2.2E-16
−0.283
−0.093
T-Reg cells
P < 2.2E-16
−0.243
−0.045
Monocytes
P < 2.2E-16
−0.167
−0.016
Plasma cells
P < 2.2E-16
−0.182
−0.053
Endothelial
P = 6E-14
−0.088
−0.037
Fibroblasts
P = 3E-12
−0.068
−0.028
Score cells
Low
High
PDL1 expression

Figure S5

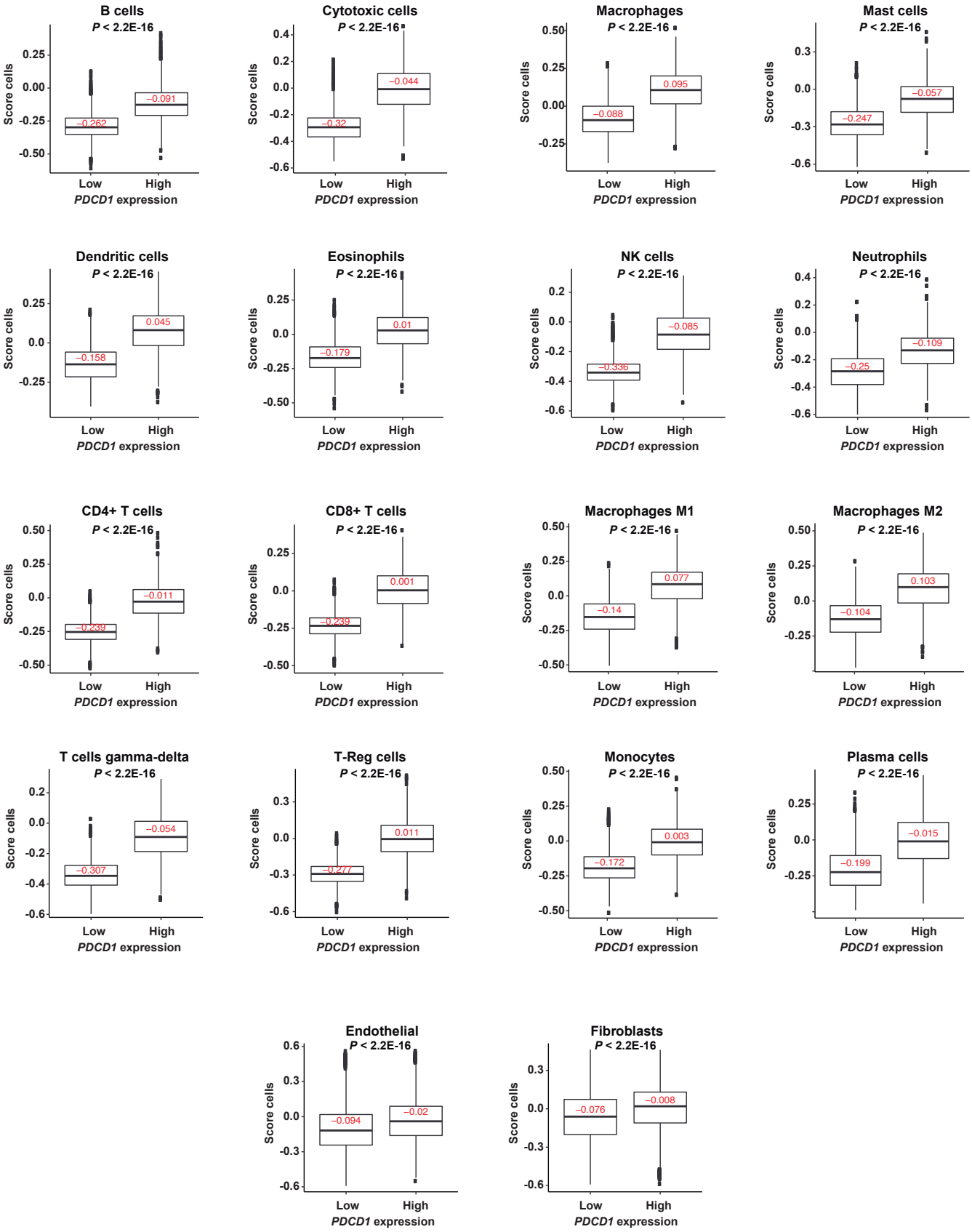

B cells
P < 2.2E-16
−0.262
−0.091
Cytotoxic cells
P < 2.2E-16
−0.32
−0.044
Macrophages
P < 2.2E-16
−0.088
0.095
Mast cells
P < 2.2E-16
−0.247
−0.057
Dendritic cells
P < 2.2E-16
−0.158
0.045
Eosinophils
P < 2.2E-16
−0.179
0.01
NK cells
P < 2.2E-16
−0.336
−0.085
Neutrophils
P < 2.2E-16
−0.25
−0.109
CD4+ T cells
P < 2.2E-16
−0.239
−0.011
CD8+ T cells
P < 2.2E-16
−0.239
0.001
Macrophages M1
P < 2.2E-16
−0.14
0.077
Macrophages M2
P < 2.2E-16
−0.104
0.103
T cells gamma-delta
P < 2.2E-16
−0.307
−0.054
T-Reg cells
P < 2.2E-16
−0.277
0.011
Monocytes
P < 2.2E-16
−0.172
0.003
Plasma cells
P < 2.2E-16
−0.199
−0.015
Endothelial
P < 2.2E-16
−0.094
−0.02
Fibroblasts
P < 2.2E-16
−0.076
−0.008
Score cells
Low
High
PDCD1 expression

Figure S6

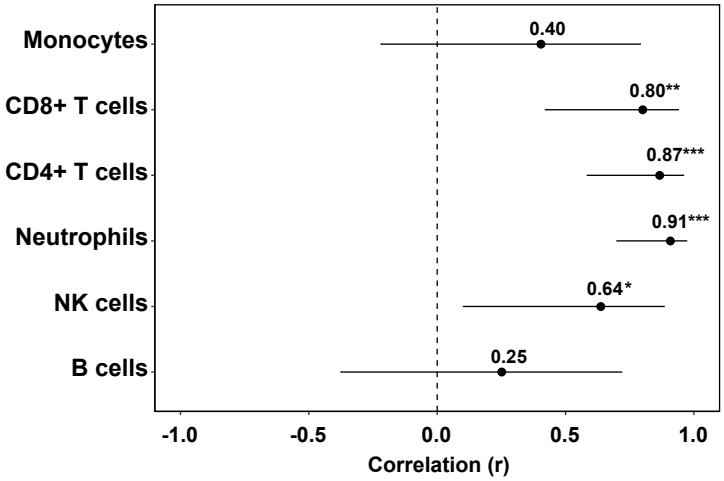

Monocytes
CD8+ T cells
CD4+ T cells
Neutrophils
NK cells
B cells
0.40
0.80**
0.87***
0.91***
0.64*
0.25
-1.0
-0.5
0.0
0.5
1.0
Correlation (r)

Figure S7

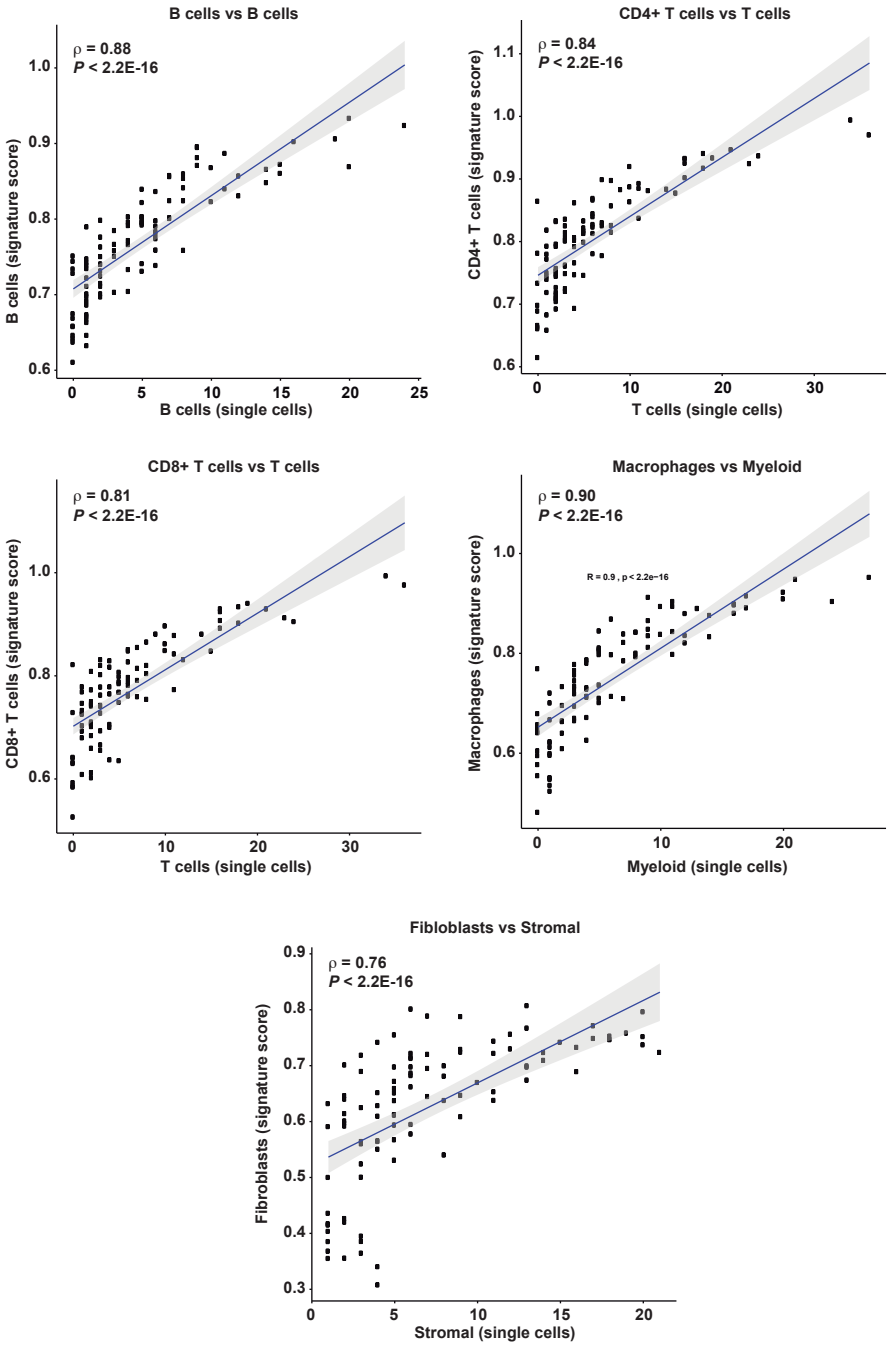

B cells vs B cells
ρ = 0.88
P < 2.2E-16
B cells (signature score)
B cells (single cells)
CD4+ T cells vs T cells
ρ = 0.84
P < 2.2E-16
CD4+ T cells (signature score)
T cells (single cells)
CD8+ T cells vs T cells
ρ = 0.81
P < 2.2E-16
CD8+ T cells (signature score)
T cells (single cells)
Macrophages vs Myeloid
ρ = 0.90
P < 2.2E-16
R = 0.9 , p < 2.2e−16
Macrophages (signature score)
Myeloid (single cells)
Fibloblasts vs Stromal
ρ = 0.76
P < 2.2E-16
Fibroblasts (signature score)
Stromal (single cells)

Figure S8

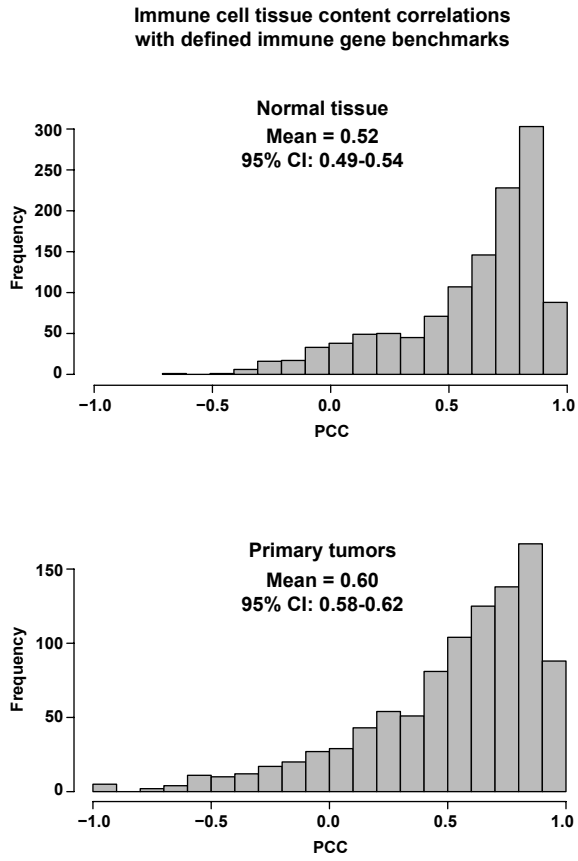

Immune cell tissue content correlations
with defined immune gene benchmarks
Normal tissue
Mean = 0.52
95% CI: 0.49-0.54
Frequency
300
250
200
150
100
50
0
−1.0
−0.5
0.0
0.5
1.0
PCC
Primary tumors
Mean = 0.60
95% CI: 0.58-0.62
Frequency
150
100
50
0
−1.0
−0.5
0.0
0.5
1.0
PCC

Figure S9

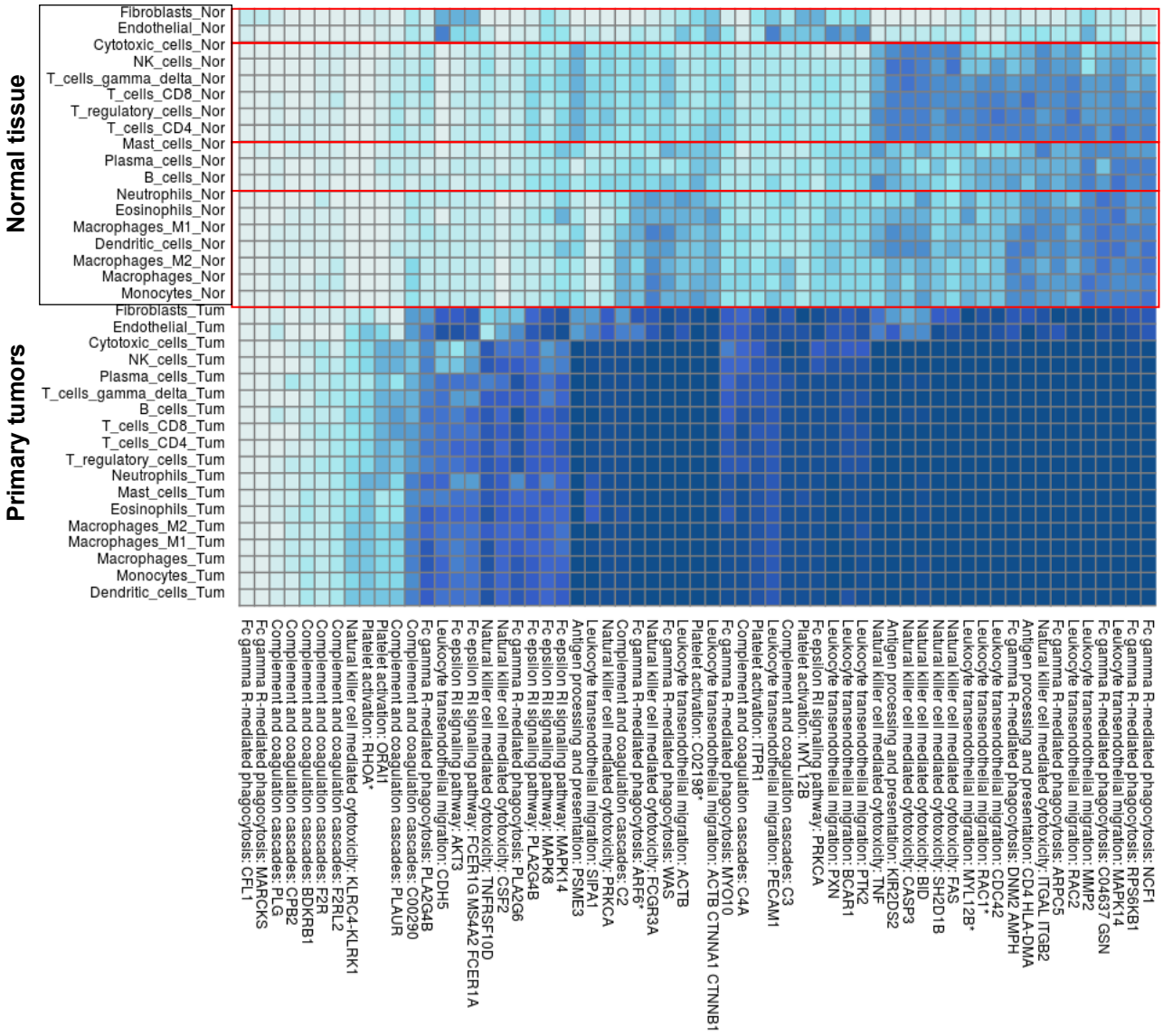

Normal tissue
Primary tumors
Fibroblasts_Nor
Endothelial_Nor
Cytotoxic_cells_Nor
NK_cells_Nor
T_cells_gamma_delta_Nor
T_cells_CD8_Nor
T_regulatory_cells_Nor
T_cells_CD4_Nor
Mast_cells_Nor
Plasma_cells_Nor
B_cells_Nor
Neutrophils_Nor
Eosinophils_Nor
Macrophages_M1_Nor
Dendritic_cells_Nor
Macrophages_M2_Nor
Macrophages_Nor
Monocytes_Nor
Fibroblasts_Tum
Endothelial_Tum
Cytotoxic_cells_Tum
NK_cells_Tum
Plasma_cells_Tum
T_cells_gamma_delta_Tum
B_cells_Tum
T_cells_CD8_Tum
T_cells_CD4_Tum
T_regulatory_cells_Tum
Neutrophils_Tum
Mast_cells_Tum
Eosinophils_Tum
Macrophages_M2_Tum
Macrophages_M1_Tum
Macrophages_Tum
Monocytes_Tum
Dendritic_cells_Tum
Fc gamma R-mediated phagocytosis: NCF1
Fc gamma R-mediated phagocytosis: RPS6KB1
Leukocyte transendothelial migration: MAPK14
Fc gamma R-mediated phagocytosis: C04637 GSN
Leukocyte transendothelial migration: MMP2
Leukocyte transendothelial migration: RAC2
Fc gamma R-mediated phagocytosis: ARPC5
Natural killer cell mediated cytotoxicity: ITGAL ITGB2
Antigen processing and presentation: CD4 HLA-DMA
Fc gamma R-mediated phagocytosis: DNM2 AMPH
Leukocyte transendothelial migration: CDC42
Leukocyte transendothelial migration: RAC1*
Fc gamma R-mediated phagocytosis: ARPC5
Leukocyte transendothelial migration: MYL12B*
Natural killer cell mediated cytotoxicity: FAS
Natural killer cell mediated cytotoxicity: SH2D1B
Natural killer cell mediated cytotoxicity: BID
Natural killer cell mediated cytotoxicity: CASP3
Antigen processing and presentation: KIR2DS2
Natural killer cell mediated cytotoxicity: TNF
Leukocyte transendothelial migration: PTK2
Leukocyte transendothelial migration: BCAR1
Leukocyte transendothelial migration: PXN
Fc epsilon RI signaling pathway: PRKCA
Platelet activation: MYL12B
Complement and coagulation cascades: C3
Leukocyte transendothelial migration: PECAM1
Platelet activation: ITPR1
Complement and coagulation cascades: C4A
Fc gamma R-mediated phagocytosis: MYO10
Leukocyte transendothelial migration: ACTB CTNNA1 CTNNB1
Platelet activation: C02198*
Leukocyte transendothelial migration: ACTB
Fc gamma R-mediated phagocytosis: WAS
Natural killer cell mediated cytotoxicity: FCGR3A
Fc gamma R-mediated phagocytosis: ARF6*
Complement and coagulation cascades: C2
Natural killer cell mediated cytotoxicity: PRKCA
Leukocyte transendothelial migration: SIPA1
Antigen processing and presentation: PSME3
Fc epsilon RI signaling pathway: MAPK14
Fc epsilon RI signaling pathway: MAPK8
Fc epsilon RI signaling pathway: PLA2G4B
Fc gamma R-mediated phagocytosis: PLA2G6
Natural killer cell mediated cytotoxicity: CSF2
Natural killer cell mediated cytotoxicity: TNFRSF10D
Fc epsilon RI signaling pathway: FCER1G MS4A2 FCER1A
Fc epsilon RI signaling pathway: AKT3
Leukocyte transendothelial migration: CDH5
Fc gamma R-mediated phagocytosis: PLA2G4B
Complement and coagulation cascades: C00290
Complement and coagulation cascades: PLAUR
Platelet activation: ORAI1
Platelet activation: RHOA*
Natural killer cell mediated cytotoxicity: KLRC4-KLRK1
Complement and coagulation cascades: F2RL2
Complement and coagulation cascades: F2R
Complement and coagulation cascades: BDKRB1
Complement and coagulation cascades: CPB2
Complement and coagulation cascades: PLG
Fc gamma R-mediated phagocytosis: MARCKS
Fc gamma R-mediated phagocytosis: CFL1

Figure S10

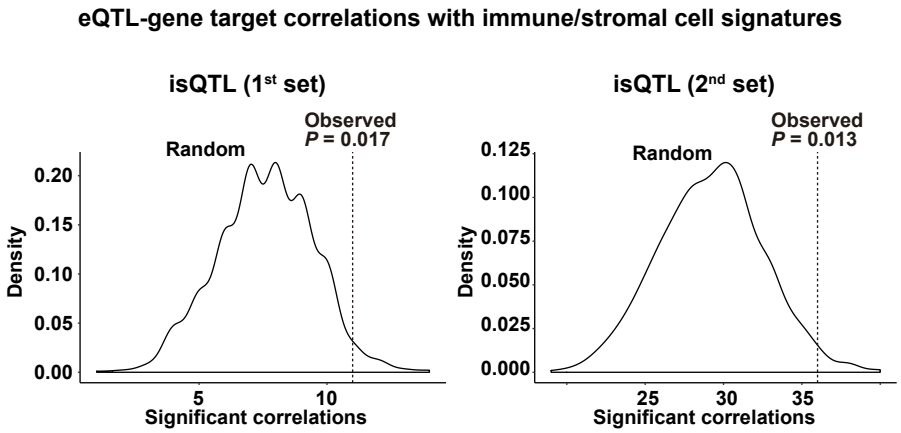

eQTL-gene target correlations with immune/stromal cell signatures
isQTL (1st set)
Observed
P = 0.017
Random
Density
0.20
0.15
0.10
0.05
0.00
5
10
Significant correlations
isQTL (2nd set)
Observed
P = 0.013
Random
Density
0.125
0.100
0.075
0.050
0.025
0.000
25
30
35
Significant correlations

Figure S11

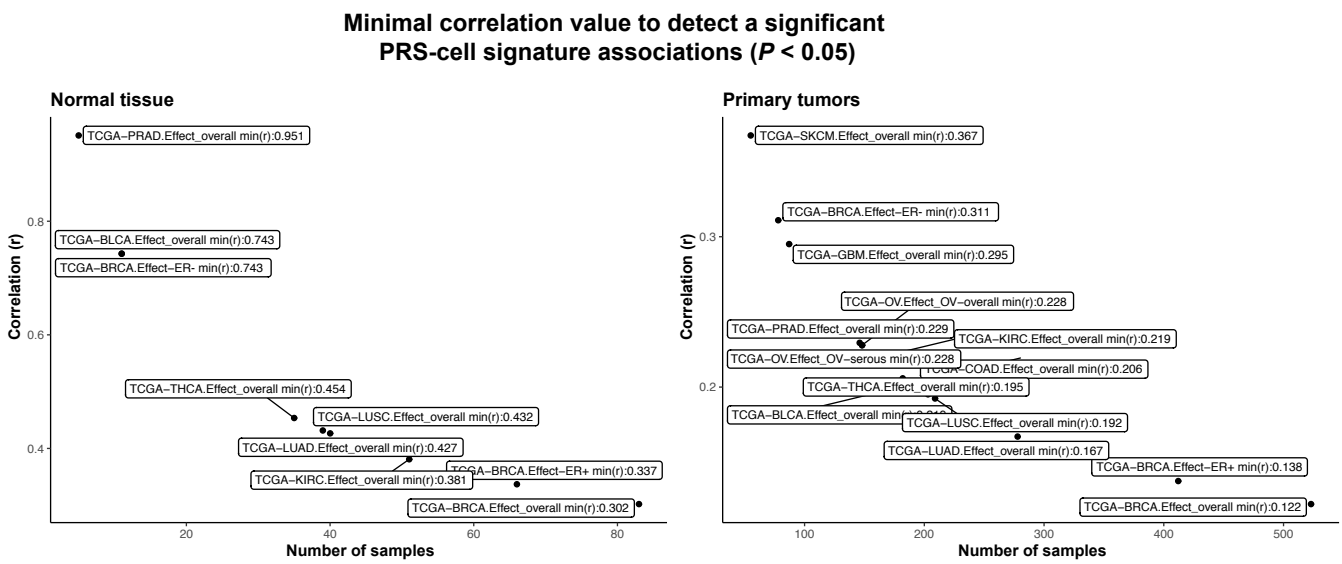
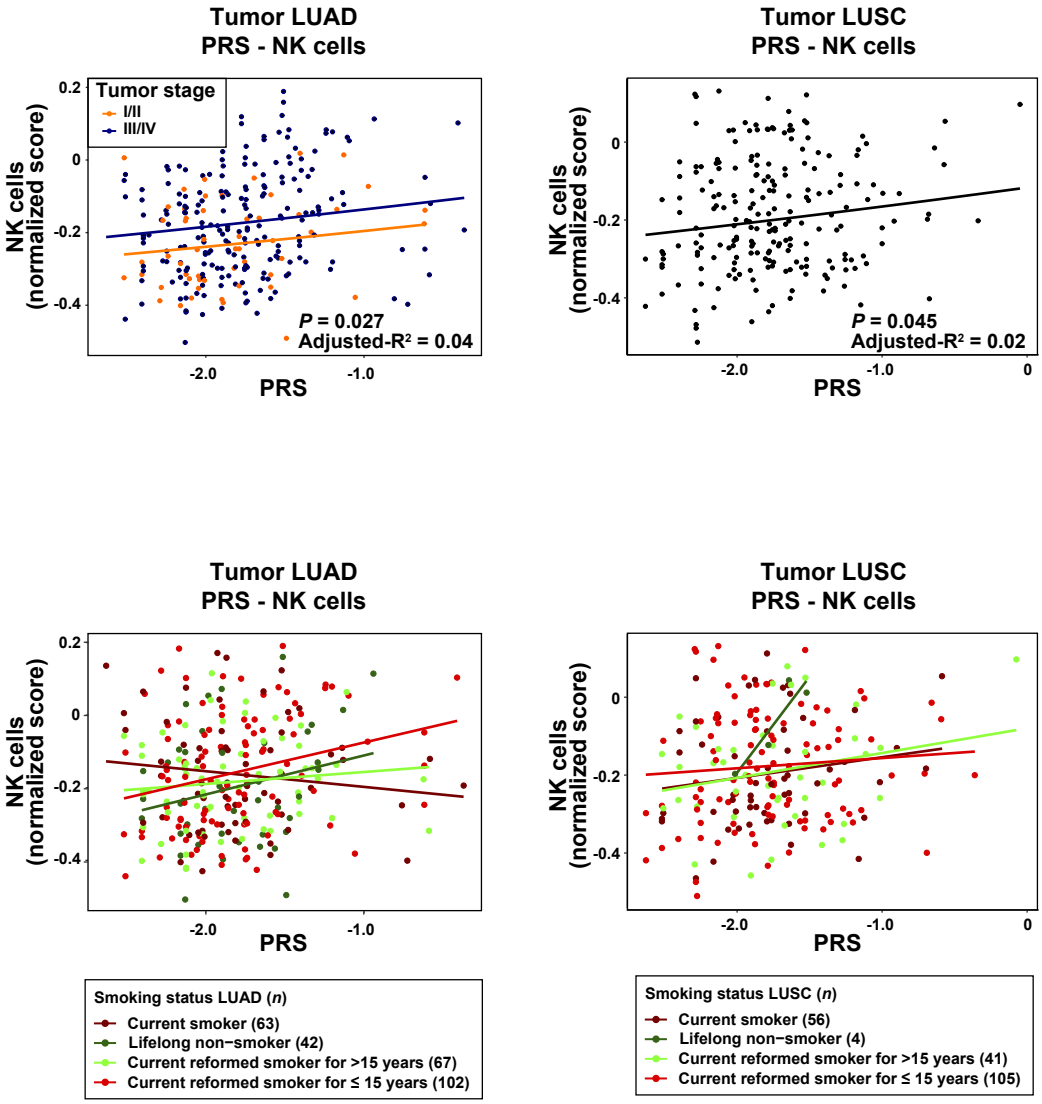

Minimal correlation value to detect a significant
PRS-cell signature associations (P < 0.05)
Normal tissue
TCGA−PRAD.Effect_overall min(r):0.951
TCGA−BLCA.Effect_overall min(r):0.743
TCGA−BRCA.Effect−ER- min(r):0.743
TCGA−THCA.Effect_overall min(r):0.454
TCGA−LUSC.Effect_overall min(r):0.432
TCGA−LUAD.Effect_overall min(r):0.427
TCGA−KIRC.Effect_overall min(r):0.381
TCGA−BRCA.Effect−ER+ min(r):0.337
TCGA−BRCA.Effect_overall min(r):0.302
Correlation (r)
0.8
0.6
0.4
20
40
60
80
Number of samples
Primary tumors
TCGA−SKCM.Effect_overall min(r):0.367
TCGA−BRCA.Effect−ER- min(r):0.311
TCGA−GBM.Effect_overall min(r):0.295
TCGA−OV.Effect_OV−overall min(r):0.228
TCGA−PRAD.Effect_overall min(r):0.229
TCGA−KIRC.Effect_overall min(r):0.219
TCGA−OV.Effect_OV−serous min(r):0.228
TCGA−COAD.Effect_overall min(r):0.206
TCGA−THCA.Effect_overall min(r):0.195
TCGA−LUSC.Effect_overall min(r):0.192
TCGA−LUAD.Effect_overall min(r):0.167
TCGA−BRCA.Effect−ER+ min(r):0.138
TCGA−BRCA.Effect_overall min(r):0.122
Correlation (r)
0.3
0.2
100
200
300
400
500
Number of samples

Figure S12


Tumor LUAD
PRS - NK cells
Tumor stage
I/II
III/IV
NK cells (normalized score)
PRS
P = 0.027
Adjusted-R² = 0.04
Tumor LUSC
PRS - NK cells
P = 0.045
Adjusted-R² = 0.02
Tumor LUAD
PRS - NK cells
Tumor LUSC
PRS - NK cells
Smoking status LUAD (n)
Current smoker (63)
Lifelong non−smoker (42)
Current reformed smoker for >15 years (67)
Current reformed smoker for ≤ 15 years (102)
Smoking status LUSC (n)
Current smoker (56)
Lifelong non−smoker (4)
Current reformed smoker for >15 years (41)
Current reformed smoker for ≤ 15 years (105)

Transparent Methods

TCGA data

Clinical and gene expression (RNA-seq fragments per kilobase of transcript per million mapped reads (FPKM) upper quartile normalized (UQ)) data from The Cancer Genome Atlas (TCGA) projects were obtained from the Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov) and from the corresponding publications. Genetic data at the individual level were obtained following approval by the dbGaP Data Access Committee (project #11689). Metastases and recurrent tumors were excluded from this study, making normal tissue (blood or solid tissue) and primary tumor samples the focus of the analyses. The cancer types are named using the corresponding TCGA study abbreviations (https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/tcga-study-abbreviations). For normal tissue, according to the TCGA protocols, these samples were collected > 2 cm from the tumor margin and/or did not contain tumor identified by histopathological review. The protein expression measures of CD26 and TFCR corresponded to those obtained by TCGA using reverse-phase protein arrays (RPPAs; level 4 data, https://tcpaportal.org/tcpa/). The COAD subtypes were defined based genomic/genetic alterations (chromosomal instability (CIN), genomic stable (GS), and microsatellite instability (MSI) tumors) and on molecular features (consensus molecular subtypes, CMS1-4) (Guinney et al., 2015).

Cancer risk variants

The variants were compiled from the GWAS Central (Beck et al., 2014) and GWAS Catalog (Buniello et al., 2019) databases, and by literature searches using the PubMed MeSH terms “GWAS”, “association”, “cancer”, and “risk”. The variants are listed in **Table S1**. The UK Biobank GWAS results were taken from the public repository at http://www.nealelab.is/uk-biobank.

Benchmark immune genes

These genes were compiled from The Immunological Genome Project (ImmGen) (Shay and Kang, 2013) and CellMarker (Zhang et al., 2019) databases, and by a literature search using the MeSH terms corresponding to the specific immune cell types represented by the gene expression signatures. The benchmarks and their cell type assignments are included in **Table S3**.

Genotype data and imputation

Bulk genotyping data corresponding to the Affymetrix Genome-Wide Human SNP 6.0 Array were downloaded from the TCGA legacy archive (https://gdc-portal.nci.nih.gov/legacy-archive/). Of the initial normal tissue and primary tumor samples (n = 16,599), those corresponding to individuals of self-reported non-white origin (n = 4,770), and those of non-European origin based on principal component analysis using variants intersected in the 1000 Genome Project phase III (n = 2,598) were excluded from subsequent analyses; these filters were applied because summary

statistics of the GWASs used in this study are strongly biased towards populations of European origin. Normal and tumor samples were then examined separately for duplicates and up to third-degree relatives (kinship cutoff = 0.05), which resulted in the exclusion of an additional 672 samples. In the joint dataset, 765 samples were also excluded because they showed a gender mismatch in an analysis of pseudoautosomal genomic regions. Considering genetic variants, 108 samples that deviated by four or more standard deviations from the mean heterozygosity rate were also excluded. For imputation, variants were excluded if they fulfilled any of the following criteria: they mapped to chromosome Y, pseudoautosomal regions or the mitochondrial genome; they had a call rate < 100%; their minor allele frequency was < 0.01; they departed from Hardy–Weinberg equilibrium ($P < 5x10^{-6}$); or they mapped to AT-CG sites. Finally, 7,686 samples (4,154 normal, comprising 3,287 blood-derived and 867 solid-tissue samples; and 3,532 primary tumors, of which 94.4% were paired) and 589,101 variants were retained for subsequent analyses. Imputation was performed using the Shape-IT V2 (Delaneau et al., 2008) and IMPUTE2 (Howie et al., 2009) algorithms, and the 1000 Genome Project Phase III panel as reference. Poorly imputed variants (accuracy score < 0.7) were excluded from subsequent analyses. A standard cutoff dose was applied to calculate genotypes using a hard-call threshold of 0.1 (i.e., 0 – 0.1, 0.9 – 1.1, 1.9 – 2.0 for reference homozygote, heterozygous and alternative homozygous genotypes, respectively).

Immune/stromal cell signatures

Immune/stromal cell gene expression signatures for each TCGA cancer setting were computed using the Consensus^{TME} method (Jiménez-Sánchez et al., 2019), which was provided available as an R package (https://github.com/cansysbio/ConsensusTME). Ten single-cell breast cancer signatures (Azizi et al., 2018) were included in the TCGA BRCA analyses. Therefore, 18 signatures were examined in each normal tissue and primary tumor setting, except for normal breast and breast cancer tissue, for which a total of 28 signatures were analyzed. The signature scores were computed using the single-sample Gene Set Expression Analysis (ssGSEA) algorithm calculated within the Gene Set Variation Analysis (GSVA) software (Hänzelmann et al., 2013). These scores were calculated for normal tissue and primary tumors, but not for blood samples, since data from blood are limited to germline genotypes. Genes whose expression was uninformative in more than half the samples in a given setting were excluded from the signature calculations; otherwise, missing data were assigned the average value of the informative samples. Evaluation of signature scores computed by two different methods —ssGSEA and summing normalized gene expression Z-scores— revealed global coherence, whereby Pearson correlation coefficients (PCCs) were > 0.80 in 99% (571/578) of the score comparisons. To select independent signatures in each normal and cancer setting, we performed a principal component analysis using the *prcomp* function in R. Components with eigenvalues > 1 were retained to study quantitative trait loci (subsequent sections). Estimates of immune-related pathway activities were calculated using directed graphs from the Kyoto Encyclopedia of Genes and Genomes

(KEGG, https://www.genome.jp/kegg/). Briefly, gene expression profiles were converted into pathway module activity scores by taking into account the chain of reactions from a defined molecular input to a specific molecular output (Cubuk et al., 2018). The 84-gene signature linked to SH2B3 included the genes differentially expressed in *Sh2b3*-null cells and that participate in genetic and/or protein interactions to this gene/protein (Huan et al., 2015); *SH2B3* was excluded from this signature for subsequent analyses.

Pseudo-bulk breast tumors

To generate 100 pseudo-bulk breast tumors, we used the single-cell RNA-seq data from the Gene Expression Omnibus (GEO) reference GSE75688 (Chung et al., 2017) and aggregated read counts using the aggregateData function in R (https://github.com/HelenaLC/muscat). Each simulated sample of 100 cells was forced to include >50% tumor cells (average 75.3%, 95% CI 72.53 – 77.93%). For non-tumoral cells, 10% of them were fixed as stromal (bulk average 7.22%, 95% CI 6.22 – 8.36%), while the other 90% were a random combination of B cells (average 5.16%, 95% CI 4.28 – 6.28%), T cells (average 6.21%, 95% CI 5.05 – 7.48%), and myeloid cells (average 6.11%, 95% CI 5.05 – 7.39%). Most of the myeloid cells were originally assigned to macrophages (Chung et al., 2017).

Quantitative trait loci of immune/stromal cell tissue content

The germline genetic calls corresponded to genotype data obtained from blood or normal tissue samples. For cases with both types of sample, the variants with discordant

calls were excluded from subsequent analyses. As specified above, the somatic genetic calls corresponded to primary tumors only. The immune/stromal cell-content quantitative trait loci (isQTL) were analyzed using the *R/qtl2* package in R (Broman et al., 2019). These analyses included the covariates of gender (when informative), age at diagnosis, tumor stage and histology. The Haley–Knott regression method was used to compute the log odds (LOD) of the associations between genetic variants and immune/stromal cell scores. One thousand permutations were performed in each setting to obtain significance thresholds (Manichaikul et al., 2007) and the variant-signature associations with empirical values of $P < 0.05$ were considered significant isQTL. The gene targets were defined according to the genomic location of the identified variants. Additional targets were identified by analyzing all variants correlated ($r^2 > 0.8$, 1000 Genomes phase 3, version 5) with each isQTL and intersect them with various functional genomic data, including promoter capture Hi-C (Javierre et al., 2016), annotated enhancers (Hnisz et al., 2013, p.), and eQTL (Schmiedel et al., 2018) from B cells, monocytes, and CD4+ and CD8+ T cells. In addition, correlated variants were queried using the Ensembl Variant Effect Predictor (McLaren et al., 2016) for potential effects on protein coding sequences.

Computation of PRSs

The PRSs were compiled from the literature and computed by summing the products of the per-allele LOD ratio assigned to each risk variant, and the corresponding allele dosage, for the total number of variants initially defined for each PRS. There was no

previous evidence of significant interactions or deviations from a log-additive model in BRCA PRSs (Mavaddat et al., 2019), but it is not known for other cancers. In the analyses of BRCA, OV (no normal tissue data available), and PRAD PRSs, two sets were analyzed, both based on GWAS-identified variants: set #1 (hereafter PRSs-1), which corresponded to scores derived from large collections of GWAS cohorts and validated in independent studies (Mavaddat et al., 2019); and set #2 (hereafter PRSs-2), which corresponded to scores derived from a phenome-wide longitudinal study using electronic health records collected by the Michigan Genomics Initiative (Fritsche et al., 2018). In both sets, PRSs were developed for all BRCA patients, and separately for the estrogen receptor (ER)-positive and ER-negative subtypes. The number of initial variants in these BRCA PRSs and those included in our study, based on available genotypes and obtained imputations were 307 and 185 for PRSs-1, and 3,820 and 3,629 for PRSs-2. As expected, the PRSs from the two sets were found to be positively correlated using germline or primary tumor data: BRCA PRSs PCCs = 0.60 – 0.66, $P < 10^{-5}$; OV tumors PRSs PCC = 0.72, $P < 10^{-25}$ (serous PCC = 0.72); and PRAD PRSs PCCs = 0.23 – 0.99, $P < 0.01$. The Michigan Genomics Initiative also provided PRSs for seven other cancer types, and the number of variants originally included and analyzed in this study were, respectively: 103 and 21 for PRAD; 42 and 41 for COAD; 16 and 16 for BLCA and SKCM; 15 and 15 for OV; 9 and 9 for GBM, LUAD and LUSC; 8 and 7 for THCA; and 7 and 6 for KIRC.

Cell signature associations with PRSs

The *bestNormalize* package in R (https://github.com/petersonR/bestNormalize) was used to normalize the cell signature values. The transformation that produced the lowest value from the Pearson's statistic divided by the degrees of freedom was taken to indicate the best function. The error distributions of the models and Q-Q plots were examined individually. The parameters of each signature transformation are provided in **Table S10**. Outliers were identified using the interquartile range rule and excluded from subsequent analyses; these were < 5% in all settings. Normalized signature values were used as dependent variables in a linear regression analysis relative to the PRSs. Stepwise analyses including covariates of gender, tumor stage and histology were performed, and the best model was selected based on the Akaike information criterion (AIC). For normal tissue, only those studies with at least 50 informative samples were analyzed. The small number of samples in each setting meant that these analyses could only detect significant (nominal $P < 0.05$) correlation estimates > 0.27 and > 0.09 in normal breast tissue and BRCA, and stronger correlations would be required in all other settings if nominal significance were to be reached (**Fig. S11**). The significance of the associations was corrected for multiple testing using the false-discovery rate (FDR) method.

Cell signature associations with age at diagnosis

The associations between the cell signature scores (dependent variables) and age at diagnosis were evaluated by multiple linear regression, including gender and tumor

stage as covariates, the best model being determined from an AIC-based stepwise selection algorithm. The statistical significances of the associations were corrected for multiple testing separately in normal tissue and primary tumor analyses (since the expected effects were the opposite of what they proved to be) using the FDR method.

Breast cancer risk in *BRCA1/2*-mutation carriers

Analyses were performed using data from the OncoArray and Collaborative Oncological Gene-environment Study (iCOGS) consortiums with the participation of the Consortium of Investigators of BRCA1/2 Modifiers (CIMBA). The OncoArray and iCOGS designs, quality controls, and statistical analyses have been described previously (Milne et al., 2017). Summary statistics from the retrospective likelihood method are reported.

Analysis of blood cell parameters and age at diagnosis of breast cancer

Clinical and histopathological data from breast cancer patients were compiled through manual curation of hospital records of the Catalan Institute of Oncology, L'Hospitalet del Llobregat (Barcelona, Catalonia, Spain). Patients were randomly selected from health records collected between 2009 and 2014. The compiled data included date of birth, age, gender (only women selected), date at diagnosis, tumor stage, subtype and/or ER status, and date at initial-diagnostic blood test. The blood test parameters analyzed were the normalized numbers ($x10^9$/L) of basophils, eosinophils, leucocytes, lymphocytes, monocytes, neutrophils, and platelets. Linear regressions of each of these parameters

on age at diagnosis, including tumor stage and subtype as covariates, were performed.

The IDIBELL's Research Ethics Committee approved this study (reference PR066/20).

Supplemental References

Azizi, E., Carr, A.J., Plitas, G., Cornish, A.E., Konopacki, C., Prabhakaran, S., Nainys, J., Wu, K., Kiseliovas, V., Setty, M., et al. (2018). Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. Cell *174*, 1293-1308.e36.

Beck, T., Hastings, R.K., Gollapudi, S., Free, R.C., Brookes, A.J. (2014). GWAS Central: a comprehensive resource for the comparison and interrogation of genome-wide association studies. Eur. J. Hum. Genet. *22*, 949–952.

Broman, K.W., Gatti, D.M., Simecek, P., Furlotte, N.A., Prins, P., Sen, Ś., Yandell, B.S., Churchill, G.A. (2019). R/qtl2: software for mapping quantitative trait loci with high-dimensional data and multiparent populations. Genetics *211*, 495–502.

Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. *47*, D1005–D1012.

Chung, W., Eum, H.H., Lee, H.-O., Lee, K.-M., Lee, H.-B., Kim, K.-T., Ryu, H.S., et al. (2017). Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. Nat. Commun. *8*, 15081.

Cubuk, C., Hidalgo, M.R., Amadoz, A., Pujana, M.A., Mateo, F., Herranz, C., Carbonell-Caballero, J., Dopazo, J. (2018). Gene expression integration into pathway modules reveals a pan-cancer metabolic landscape. Cancer Res. *78*, 6059–6072.

Delaneau, O., Coulonges, C., Zagury, J.-F. (2008). Shape-IT: new rapid and accurate algorithm for haplotype inference. BMC Bioinformatics 9, 540.

Fritsche, L.G., Gruber, S.B., Wu, Z., Schmidt, E.M., Zawistowski, M., Moser, S.E., Blanc, V.M., Brummett, C.M., Kheterpal, S., Abecasis, G.R., Mukherjee, B. (2018). Association of polygenic risk scores for multiple cancers in a phenome-wide study: results from The Michigan Genomics Initiative. Am. J. Hum. Genet. *102*, 1048–1061.

Guinney, J., Dienstmann, R., Wang, X., de Reyniès, A., Schlicker, A., Soneson, C., Marisa, L., Roepman, P., Nyamundanda, G., Angelino, P., et al. (2015). The consensus molecular subtypes of colorectal cancer. Nat. Med. *21*, 1350–1356.

Hänzelmann, S., Castelo, R., Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics *14*, 7.

Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-André, V., Sigova, A.A., Hoke, H.A., Young, R.A. (2013). Super-enhancers in the control of cell identity and disease. Cell *155*, 934–947.

Howie, B.N., Donnelly, P., Marchini, J. (2009). A flexible and accurate genotype

imputation method for the next generation of genome-wide association studies. PLoS Genet. *5*, e1000529.

Huan, T., Meng, Q., Saleh, M.A., Norlander, A.E., Joehanes, R., Zhu, J., Chen, B.H., Zhang, B., Johnson, A.D., Ying, S., et al. (2015). Integrative network analysis reveals molecular mechanisms of blood pressure regulation. Mol. Syst. Biol. *11*, 799.

Javierre, B.M., Burren, O.S., Wilder, S.P., Kreuzhuber, R., Hill, S.M., Sewitz, S., Cairns, J., Wingett, S.W., Várnai, C., Thiecke, M.J., et al. (2016). Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. Cell *167*, 1369-1384.e19.

Jiménez-Sánchez, A., Cast, O., Miller, M.L. (2019). Comprehensive benchmarking and integration of tumor microenvironment cell estimation methods. Cancer Res. *79*, 6238–6246.

Manichaikul, A., Palmer, A.A., Sen, S., Broman, K.W. (2007). Significance thresholds for quantitative trait locus mapping under selective genotyping. Genetics *177*, 1963–1966.

Mavaddat, N., Michailidou, K., Dennis, J., Lush, M., Fachal, L., Lee, A., Tyrer, J.P., Chen, T.-H., Wang, Q., Bolla, M.K., et al. (2019). Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. Am. J. Hum. Genet. *104*, 21–34.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*, 122.

Milne, R.L., Kuchenbaecker, K.B., Michailidou, K., Beesley, J., Kar, S., Lindström, S., Hui, S., Lemaçon, A., Soucy, P., Dennis, J., et al. (2017). Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. Nat. Genet. *49*, 1767–1778

Schmiedel, B.J., Singh, D., Madrigal, A., Valdovino-Gonzalez, A.G., White, B.M., Zapardiel-Gonzalo, J., Ha, B., Altay, G., Greenbaum, J.A., McVicker, G., et al. (2018). Impact of genetic polymorphisms on human immune cell gene expression. Cell *175*, 1701-1715.e16.

Shay, T., Kang, J. (2013). Immunological Genome Project and systems immunology. Trends Immunol. *34*, 602–609.

Zhang, X., Lan, Y., Xu, J., Quan, F., Zhao, E., Deng, C., Luo, T., Xu, L., Liao, G., Yan, M., et al. (2019). CellMarker: a manually curated resource of cell markers in human and mouse. Nucleic Acids Res. *47*, D721–D728.