Supplementary Information

The prognostic landscape of genes and infiltrating immune cells across human cancers

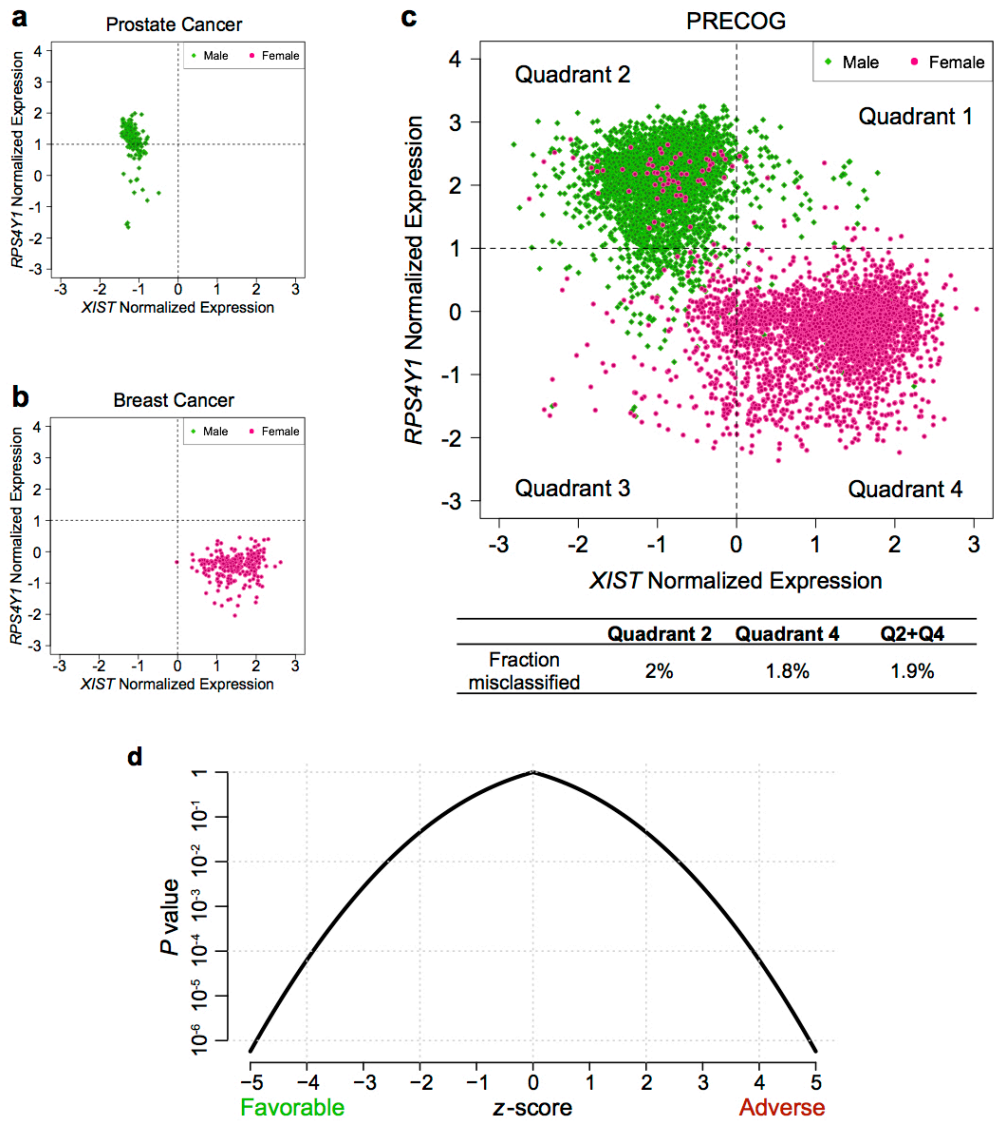
*Andrew J. Gentles, *Aaron M. Newman, Chih Long Liu, Scott V. Bratman, Weiguo Feng, Dongkyoon Kim, Viswam S. Nair, Yue Xu, Amanda Khuong, Chuong D. Hoang, Maximilian Diehn, Robert B. West, Sylvia K. Plevritis, Ash A. Alizadeh

SI Guide

Nature Medicine: doi:10.1038/nm.3909

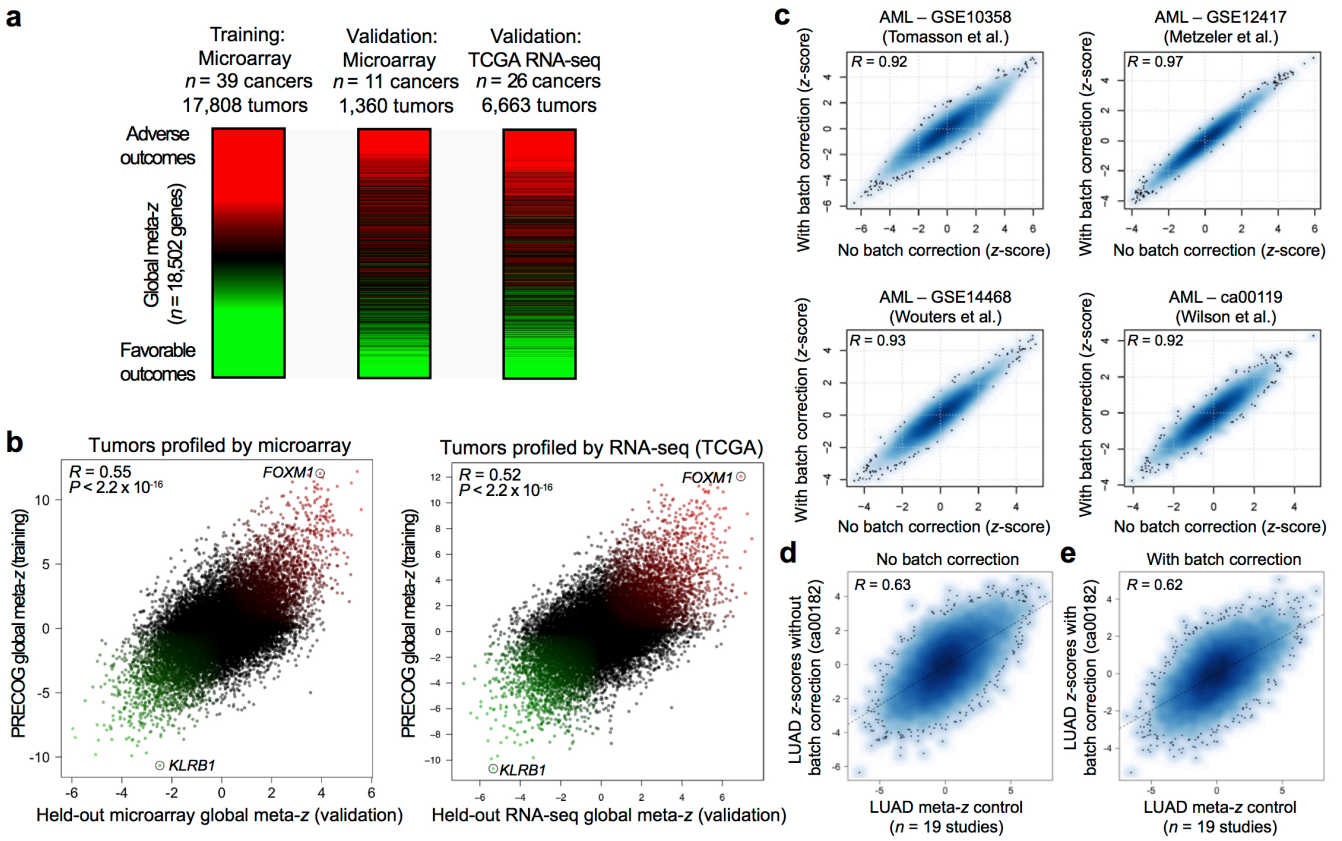
Supplementary Table 7	Lung adenocarcinoma TMA analyses, including clinical data and marker quantification, multivariate survival analysis with clinical covariates, and comparison of TAL levels with circulating leukocytes

Nature Medicine: doi:10.1038/nm.3909



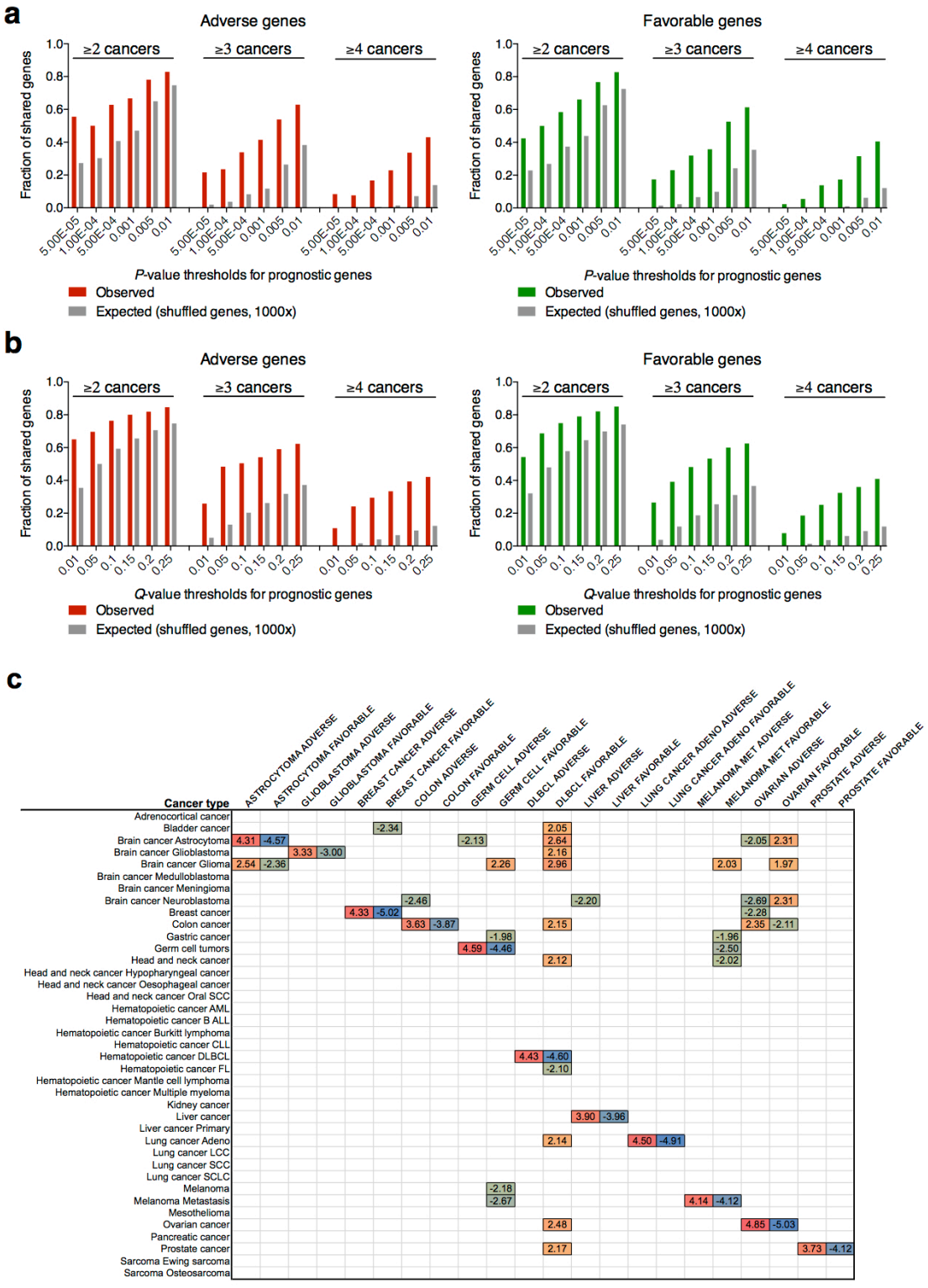
Supplementary Figure 1: PRECOG quality control and relationship between *z* scores and *P* values. Normalized gene expression values for the gender-dimorphic genes, *RPS4Y1* (males) and *XIST* (females), are plotted from prostate cancer (**a**), breast cancer (**b**), and from 7,053 tumors in PRECOG with available annotation of patient gender (**c**): green points, males; fuchsia-colored points, females. The mutually exclusive expression values of these two genes naturally partition this plot into 4 quadrants. Discordance between clinically-annotated and molecularly-inferred patient gender (i.e., fuschia points in quadrant 2 and green points in quadrant 4) indicate potential technical errors in PRECOG, however, such errors are estimated to account for <2% of integrated datasets. All arrays used in this analysis were MAS5-normalized, quantile normalized, and unit variance normalized. (**d**) The *z*-score is a measurement of statistical significance, specifically the number of standard deviations that a calculated or observed quantity is away from its mean value. Assuming that the underlying distribution is Gaussian (true to high fidelity for PRECOG; also see **Supplementary Fig. 6b–d** relating to immune-PRECOG), *P* values can be computed directly from *z*-scores. This relationship is illustrated here for $|Z| < 5$. Two-sided *P* values corresponding to $Z = 2, 3, 4, 5$ are 0.04, 0.003, 6.3×10^{-5}, and 5.7×10^{-7} respectively. Unlike the *P* value, the *z*-score conveniently encodes directionality of the association. In this work, a positive *z*-score indicates an adverse prognostic association whereas a negative *z*-score indicates a favorable prognostic association.

Nature Medicine: doi:10.1038/nm.3909



Supplementary Figure 2: Validation of PRECOG using external data sets, and analysis of batch effects. (**a**) Left: Heat map (*Training*) depicting prognostic *z*-scores in PRECOG ranked by decreasing global meta-*z*, with the most adversely prognostic genes on the top and the most favorable prognostic genes on the bottom. Center/Right: Heat map (*Validation*) showing the same genes in the same order as PRECOG *training*, but with global meta *z*-scores from two held-out datasets, compiled from independent tumors profiled by microarray (center) and RNA-seq (right) (**Supplementary Table 1**). The color scale ranges from –1 (green) to +1 (red). (**b**) Comparison of whole-transcriptome global meta *z*-scores between PRECOG and both the microarray validation dataset (left) and TCGA RNA-seq validation data set (right) (Pearson's correlation coefficients are shown). *FOXM1* and *KLRB1* are indicated. TCGA global meta-*z* scores are based on 26 cancer types with available RNA-seq data and overall survival outcomes (ACC, BLCA, CESC, COAD, DLBC, GBM, HNSC, KICH, KIRC, KIRP, LAML, LGG, LIHC, LUAD, LUSC, OV, PAAD, PCPG, PRAD, READ, SARC, SKCM, THCA, UCEC, UCS. (**c**) Comparison of whole-transcriptome *z*-scores of four AML datasets before and after batch effect correction (applied to microarray processing dates encoded in CEL files) using COMBAT[1]. (**d**) Comparison of pre-batch-corrected *z*-scores from the NCI director's challenge lung adenocarcinoma (LUAD) dataset (ca00182) with corresponding meta-*z* scores from all 19 remaining LUAD datasets in PRECOG. (**e**) Same as **d** but after batch correction was applied to arrays from different study sites in ca00182. (**c–e**) Pearson's correlation coefficients shown in the upper left corner of each plot are all significant ($P < 2.2 \times 10^{-16}$).

Nature Medicine: doi:10.1038/nm.3909

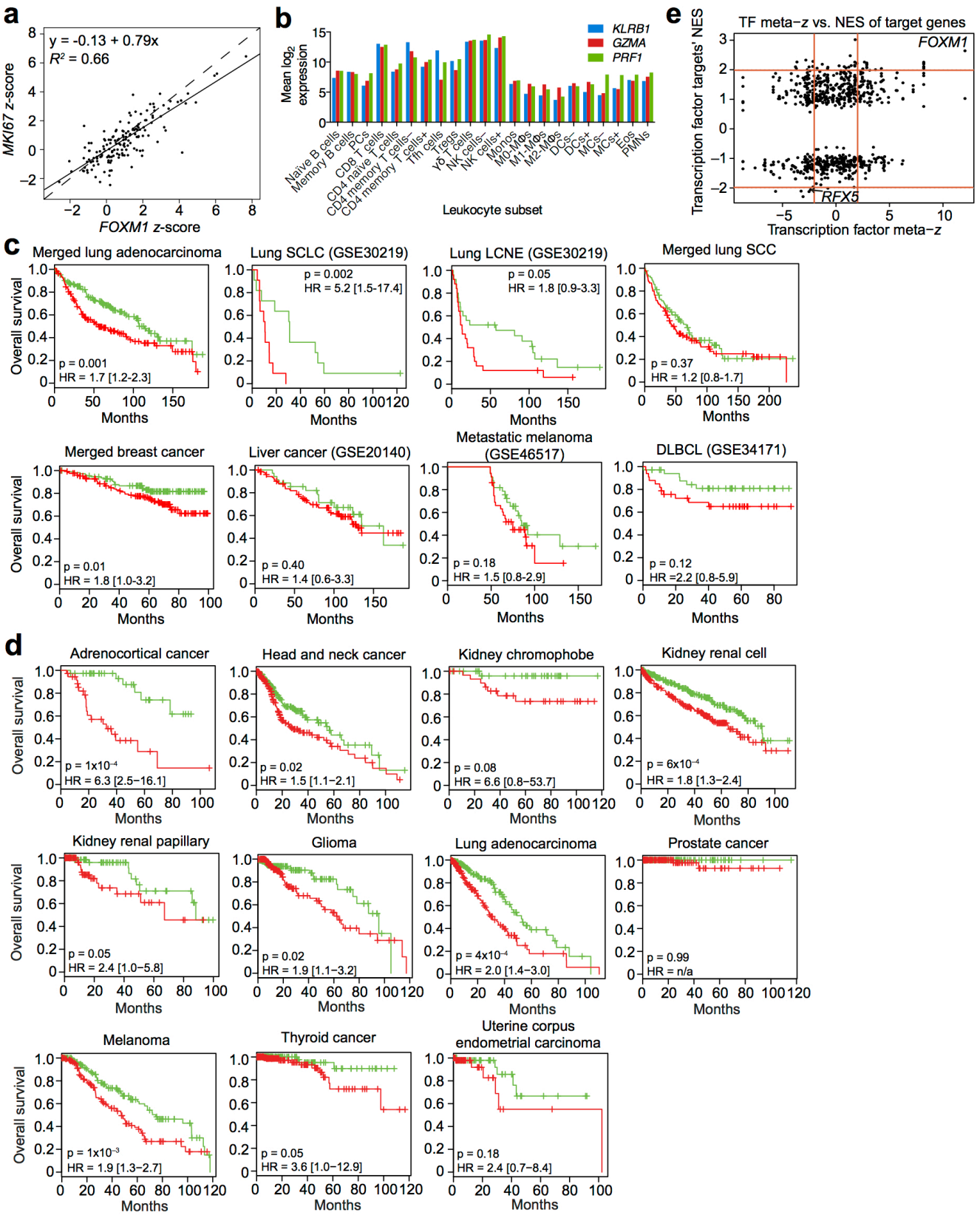


Supplementary Figure 3: Significance of shared and cancer-specific prognostic genes. Analysis of the significance of shared prognostic genes across a broad range of (**a**) *P* value thresholds and (**b**) *Q* value thresholds assessed by Monte Carlo simulation. Analyses were divided into adverse ($Z > 0$) and favorable ($Z < 0$) prognostic genes, and one-sided *P* values were determined from cancer-specific meta-*z* scores. Cancer-level *P* values were converted into *Q* values as described in Methods. To calculate the median fraction of shared prognostic genes, gene labels were randomized separately for each cancer and evaluated genes were required to be shared by at least 2, 3, or 4 cancers in PRECOG. (Notably, all *P*

Nature Medicine: doi:10.1038/nm.3909

values and *Q* values were determined for each cancer separately, and do not reflect pan-cancer values.) For all combinations, observed fractions in PRECOG were compared to expected fractions determined by 1,000 Monte Carlo trials. Significantly more shared prognostic genes were found than would be expected by random chance ($P < 0.05$), regardless of the statistical threshold or minimum number of cancers considered. (**c**) Significant prognostic genes found only in one cancer type are generally not borderline significant in other cancers. Cancer-specific lists of adversely and favorably prognostic genes in PRECOG (**Supplementary Table 2**) were tested as gene sets against ranked meta-z scores in other cancers, using GSEA[2]. Shown are the normalized enriched scores (NES) of the cancer-specific sets (columns) across PRECOG (rows). With few exceptions, the cancer-specific gene lists are not enriched in malignancies other than the one from which they were derived. Moreover, among elements with $|NES| > 1.96$ (shown), there is a large gap between NES scores of matching cancers (mean $|NES| = 4.2$) versus non-matching cancers (mean $|NES| = 2.3$). Empirically, $|NES| > 1.96$ corresponds approximately to FDR < 0.05.

Nature Medicine: doi:10.1038/nm.3909

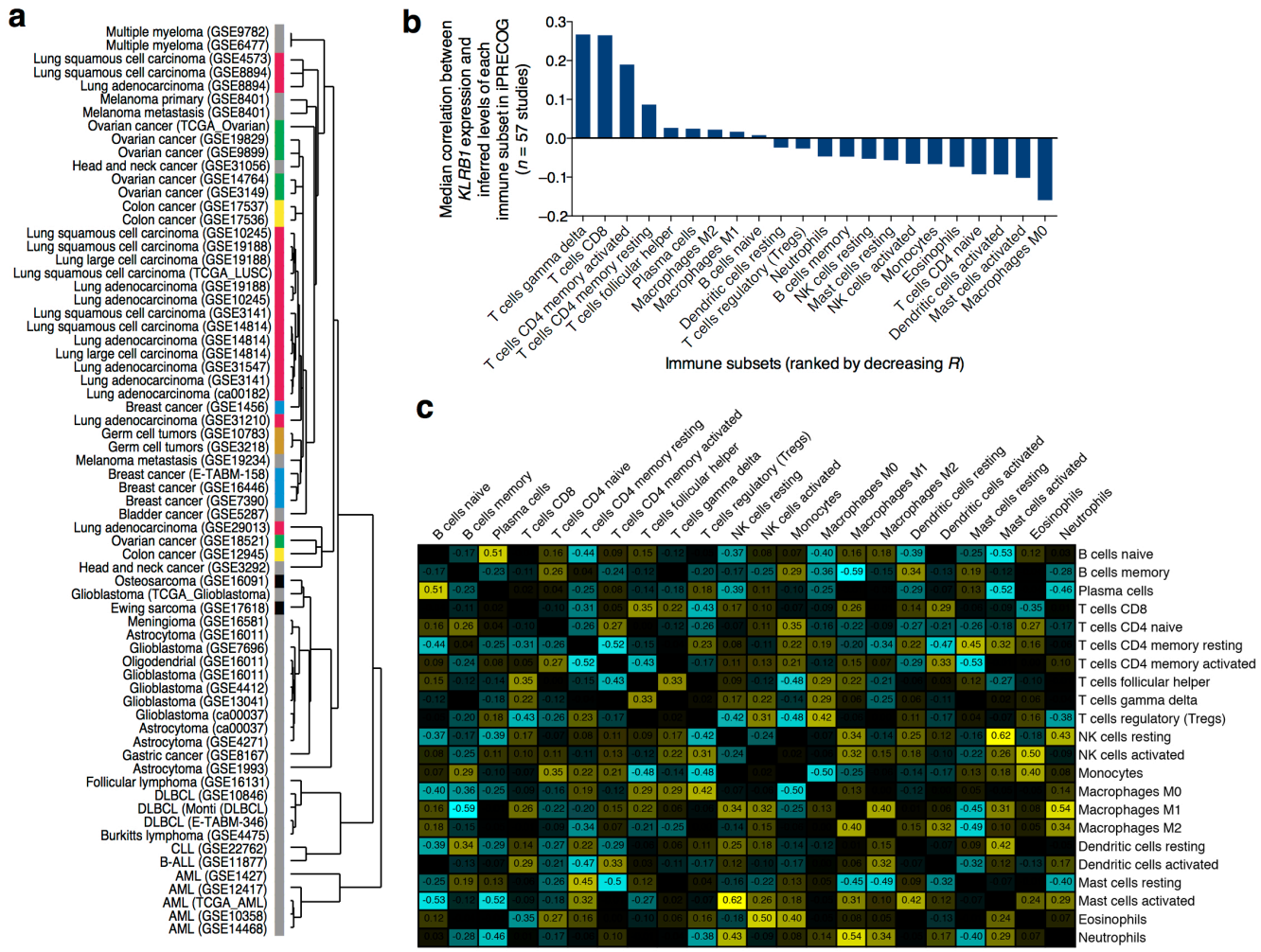


Supplementary Figure 4: Analyses related to *FOXM1* and *KLRB1*, including assessment of a *FOXM1-KLRB1* prognostic score in validation datasets. (**a**) The survival *z*-score of *FOXM1* is shown compared to that for *MKI67*. Each point represents one dataset in PRECOG. The slope of the regression (solid line) illustrates that in general, *FOXM1* tends towards larger *z*-scores than *MKI67* indicating more

Nature Medicine: doi:10.1038/nm.3909

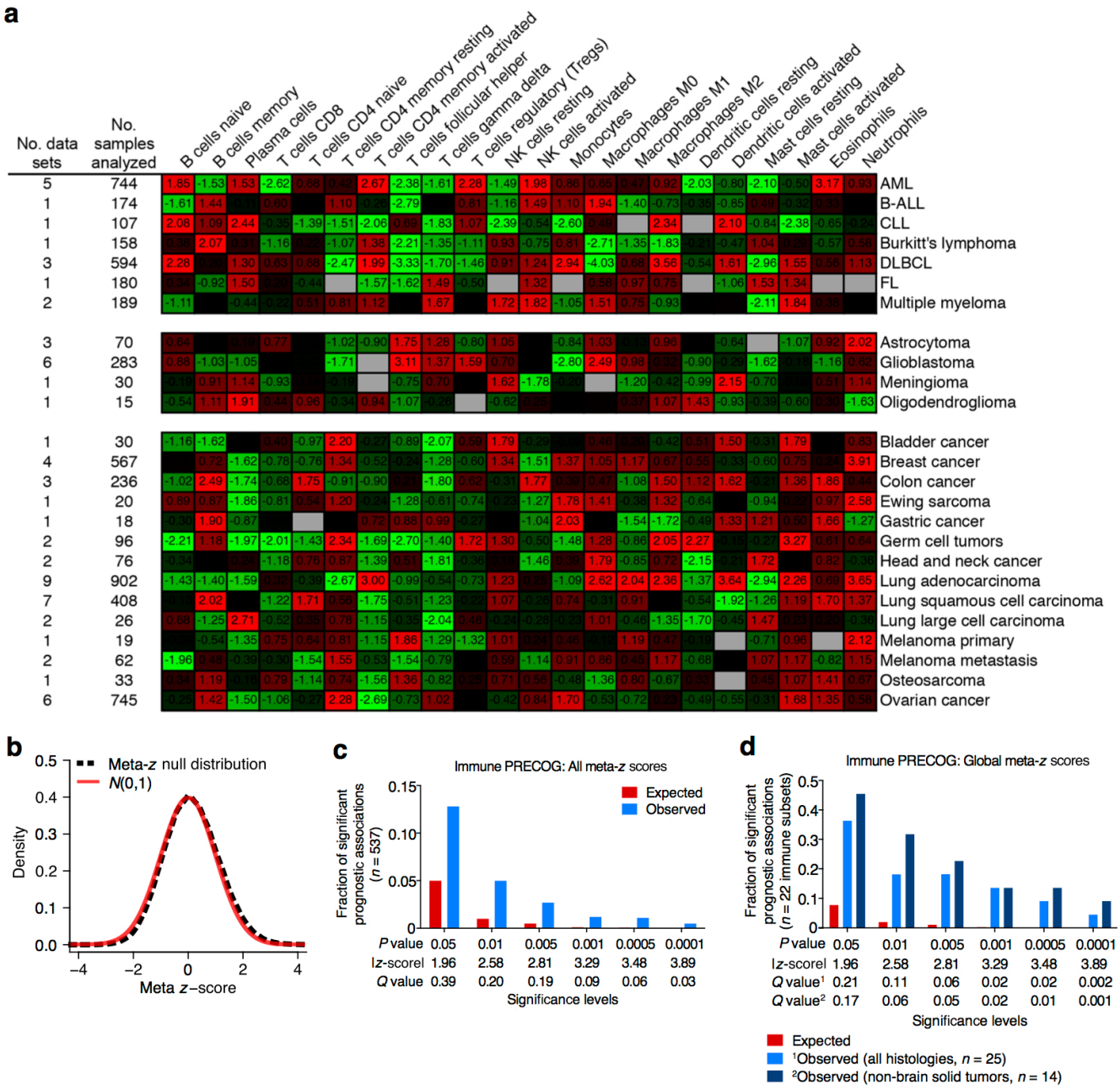
robust adverse prognostic capacity. (**b**) Expression of *KLRB1*, *GZMA*, and *PRF1* in the 22 leukocyte subsets analyzed in this work. While *KLRB1* (killer cell lectin-like receptor subfamily B, member 1), *GZMA* (Granzyme A), and *PRF1* (Perforin 1) are highly expressed on NK cells and cytotoxic CD8 T-cells, they are also widely expressed on other immune types. Shown are the mean levels of expression of these three genes across the 22 immune cell types in CIBERSORT's LM22 signature matrix[3]. The high expression of these genes (e.g., by γδ T-cells) motivates the use of gene expression deconvolution, as assignment of prognostic function to specific tumor-infiltrating immune cells based solely on these markers could lead to inaccurate conclusions. (**c**,**d**) Survival analysis of the *FOXM1-KLRB1* score applied to held-out microarray and RNA-seq datasets. As in **Fig. 2c**, the *FOXM1-KLRB1* composite score was used to stratify datasets in the microarray validation cohort (**c**) and in the TCGA RNA-seq validation cohort (**d**). Datasets corresponding to the same cancer type in panel **c** were merged (see **Supplementary Table 1**). Only datasets/cancers with significant or trending prognostic associations are shown. Dataset pre-processing and survival analysis for the *FOXM1-KLRB1* score are described in **Methods** and **Fig. 2c**. Of note, the relative weightings of *FOXM1* and *KLRB1* were not optimized for RNA-seq, further supporting the robustness of their pan-cancer association with overall survival. (**e**) The survival meta-*z* of expression levels of individual transcription factors (*x*-axis) was compared to the meta-*z* of their target genes (*y*-axis). The latter was assessed as the Gene Set Enrichment Analysis (GSEA)[2] Normalized Enrichment Score (NES) of transcription factor target gene sets (from ENCODE, ChEA, mSigDB) against the list of all genes ranked by their average meta-*z* across all cancers. *FOXM1* emerged as individually strongly prognostic of poor outcome, while also binding (according to ChIP-seq data) to many genes whose high expression level portends poor survival (**Supplementary Table 5**). Of note, RFX5 (regulatory factor X 5) binds to the promoters of MHC class II genes and activates their transcription. High expression of many of these genes correlate with good outcomes, suggesting that tumors expressing them are more likely to be subject to immune surveillance. Conversely, down-regulation of RFX5 and target MHC genes may provide a pro-survival immunoediting strategy for tumors.

Nature Medicine: doi:10.1038/nm.3909



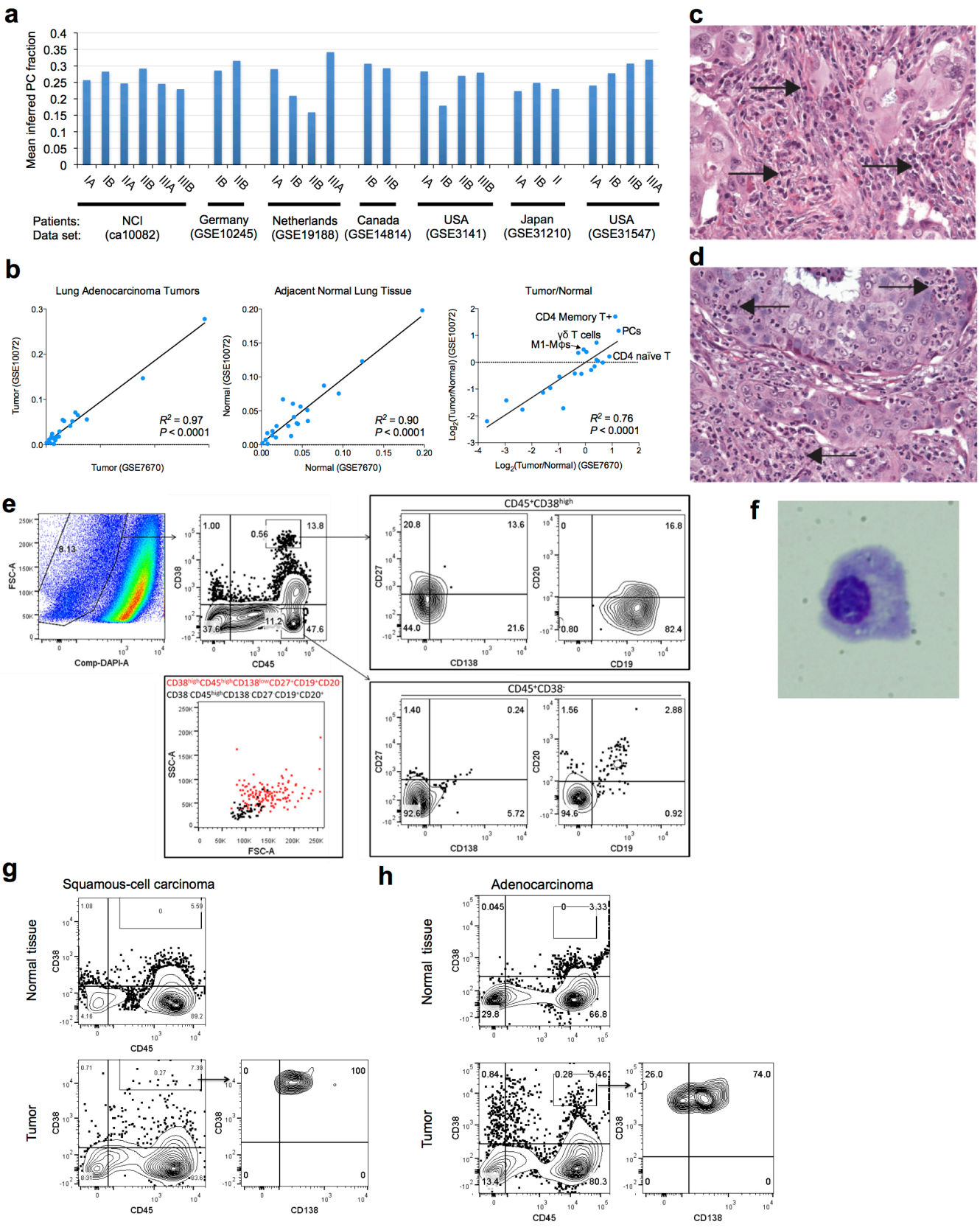
Supplementary Figure 5: Correlation analyses of estimated leukocyte fractions across cancer types and datasets. (**a**) Dendrogram showing results of centroid hierarchical clustering applied to leukocyte composition vectors (n = 22 subsets per dataset) from **Supplementary Table 6**. Centered correlation was used as the distance metric. Color-code: pink, lung cancer; green, ovarian cancer; blue, breast cancer; yellow, colon cancer; brown, germ cell tumors; black, sarcomas. Clustering of cancers of the same type from independent studies illustrates the reproducibility of CIBERSORT's estimation of relative immune infiltration levels. (**b**) Pearson correlation coefficients between *KLRB1* expression and inferred levels of each immune subset across all 57 studies analyzed in immune PRECOG (**Supplementary Table 6**). Data are presented as medians. (**c**) Cross-correlation analysis of leukocyte prognostic associations across cancers. All pairwise Pearson correlations between the meta-z scores of immune populations in immune PRECOG (**Supplementary Fig. 6a**), illustrated as a heat map.

Nature Medicine: doi:10.1038/nm.3909



Supplementary Figure 6: Prognostic associations between 22 leukocyte subsets and 25 cancer histologies. (**a**) Heat map depicting relationships between hematopoietic subsets and survival, represented as a meta *z*-score matrix. Red cells denote adverse outcomes and green cells denote favorable outcomes. (**b**) False discovery rates of leukocyte prognostic associations. Comparison of null distribution of *z*-scores obtained from shuffling cell type fractions in immune-PRECOG (dashed black line) to a standard normal distribution shows high concordance. (**c**) Expected versus observed fractions of statistically significant associations between cell type proportions and outcome obtained by filtering results in **6a** at various *z*-score cutoffs. *P* values and estimated FDRs are shown for each *z*-score value. The more stringent the cutoff, the higher the ratio of observed to expected significant associations (3-fold at $P < 0.05$, 5-fold at $P < 0.01$), indicating that immune-PRECOG captures statistically robust associations. (**d**) Similarly to panel **b**, but applied to global meta-*z* scores obtained from combining the individual cancer meta-*z*-scores across 25 cancer histologies or non-brain solid tumors (related to **Fig. 3c**). Details for **b–d** are provided in Methods.

Nature Medicine: doi:10.1038/nm.3909

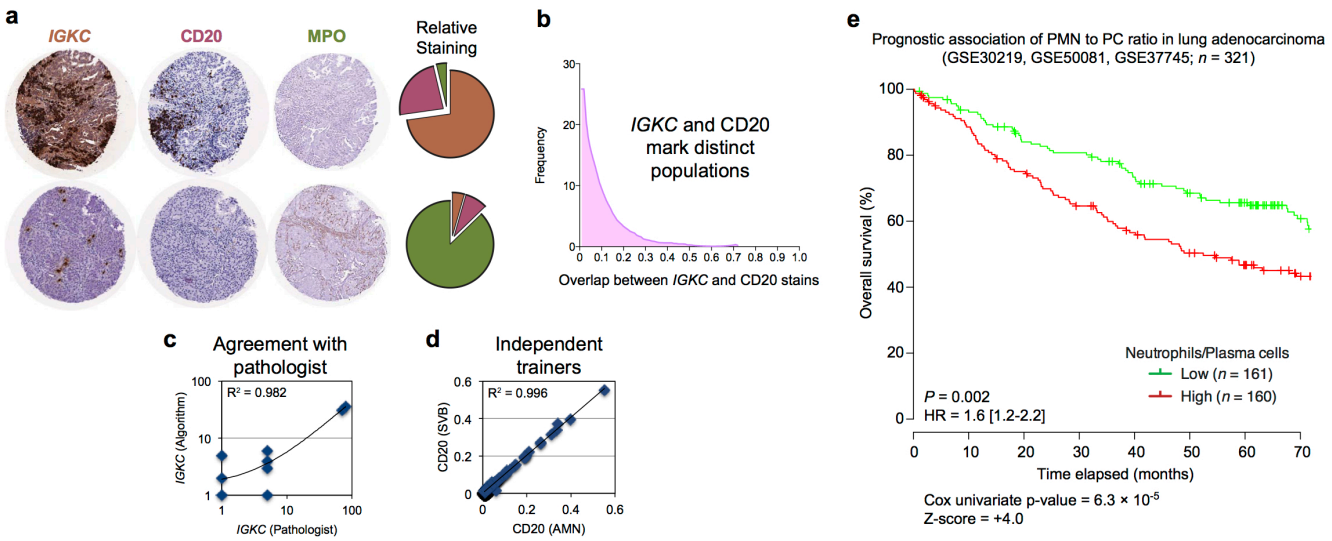


Supplementary Figure 7: Plasma cell levels in non-small cell lung cancer and adjacent normal tissues. (**a**) Relative RNA fractions of plasma cells inferred by CIBERSORT are independent of lung adenocarcinoma stage. (**b**) Relative fractions of 22 leukocyte subsets, as inferred by CIBERSORT, are

Nature Medicine: doi:10.1038/nm.3909

compared between two independent microarray datasets (GSE7670 and GSE10072) containing both lung adenocarcinoma tumor and adjacent normal specimens. For defined leukocyte abbreviations, see **Supplementary Table 6**. (**c**,**d**) Representative H&E stains of lung adenocarcinoma tissue specimens. Stained lung adenocarcinoma tumor sections showing cells (indicated by arrows) that morphologically resemble (**c**) plasma cells and (**d**) neutrophils. (**e–h**) Flow cytometric analysis and morphological assessment of plasmacytic cells in lung cancer. (**e**) Gating strategy for enrichment of $CD38^{high}/CD45^{high}/CD138^{low}/CD27^{+}/CD19^{+}/CD20^{-}$ cells from a lung adenocarcinoma tumor. As expected for plasmacytic cells, $CD38^{high}/CD45^{high}/CD138^{low}/CD27^{+}/CD19^{+}/CD20^{-}$ cells are larger than $CD38^{-}/CD45^{high}/CD138^{-}/CD27^{-}/CD19^{+}/CD20^{+}$ cells (B-cells) by forward and side scattering. (**f**) Using the gating strategy described in **e**, plasmacytic cells were sorted from a fresh lung adenocarcinoma tumor and isolated for microscopy by cytospinning. A representative cell with morphological features characteristic of plasmacytic cells is shown (100× oil objective lens). Representative flow cytometry results showing a considerable increase in plasmacytic cells in lung squamous cell carcinoma (**g**) and lung adenocarcinoma (**h**) tumors as compared to normal adjacent tissues.

Nature Medicine: doi:10.1038/nm.3909



Supplementary Figure 8: Assessment of TMA markers and staining quantification, and prognostic significance of inferred PMN/PC levels in held-out expression datasets. (**a**) Representative lung adenocarcinoma tissue sections, stained by an RNA in situ probe targeting *IGKC*, or antibodies targeting CD20 or MPO. Top: Serial sections in which *IGKC* and CD20 are high and MPO is low. Bottom: Serial sections in which MPO is high and *IGKC*/CD20 are low. Staining was quantified by GemIdent image analysis software[4] and post-processing (Methods). (**b**) Histogram of the spatial overlap between *IGKC* and CD20 stains in adjacent lung adenocarcinoma tissue sections (median overlap of ~4.8%). (**c**) Concordance between *IGKC* staining assessment by a pathologist (R.W.) and by GemIdent (Methods) for 10 randomly selected lung adenocarcinoma specimens. (**d**) GemIdent was trained to recognize CD20 staining by two different operators, and the results are plotted for all lung adenocarcinoma specimens. (**e**) Survival analysis of the ratio of PMNs to PCs in held-out lung adenocarcinoma datasets. Plasmacytic cell and neutrophil fractions estimated by CIBERSORT were used to compute their ratio in three lung cancer datasets not included in PRECOG. Patients were stratified into high or low groups based on the median value of the PMN:PC ratio in each dataset. This permitted merging of the three cohorts into one combined dataset of sufficient size for survival analysis. Hazard ratio (HR) with 95% confidence interval is shown along with *P* value in Cox regression (log-rank test).

Nature Medicine: doi:10.1038/nm.3909

Supplementary References

1. Johnson, W.E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118-127 (2007).
2. Mootha, V.K.*, et al.* PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267-273 (2003).
3. Newman, A.M.*, et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453-457 (2015).
4. Holmes, S., Kapelner, A. & Lee, P.P. An interactive Java statistical image segmentation system: Gemident. *J. Stat. Softw.* **30**, i10 (2009).