SUPPLEMENTARY MATERIAL

SUPPLEMENTARY METHODS

1.Computing the immune infiltration pattern of a tumor

Several computational methods have been developed to quantify the infiltration of immune cell populations in a solid tumor from the RNA-seq of the tumor bulk sample. These methods fall into two broad categories: (i) sample level **gene set enrichment** methods compute the relative abundance of each immune cell population in each tumor of a cohort [1–3], whereas (ii) **deconvolution** methods [4,5] estimate the proportion of each immune cell population within the sample admixture.

State-of-the-art deconvolution methods rely on reference expression matrices that either (i) have not been validated for RNA-seq data; (ii) are only available in a non-customizable setting, as the populations and genes in the matrix are pre-defined; and/or (iii) must be used in conjunction with another method/estimate to assess the overall abundance of infiltration within the samples. On the other hand, gene set enrichment methods (i) can be customized (defining the set of genes whose collective expression identifies a given immune cell population); (ii) are suitable for RNA-seq data; and (iii) provide a score that takes into account the level of infiltration of each cell type in each tumor [6].

In section 4 we evaluate the performance of a widely employed deconvolution-based method to analyze RNA-seq data and compare the immune infiltration pattern computed using this methodology with that estimated using a gene set enrichment. Considering the *pros* and *cons* of each type of approach (described in that section), we decided to use a gene set enrichment-based method in our study.

2.Selection of the sample level enrichment method

The most widely used methods that carry out sample level enrichment analysis are the Gene Set Variation Analysis (GSVA) [7] and the single sample Gene Set Enrichment Analysis (ssGSEA) [8]. Both are unsupervised gene set enrichment methods that compute an enrichment score integrating the collective expression of a given gene set relative to the other genes in the sample.

We compared the results produced by the GSVA and the ssGSEA using the 23 immune populations gene signatures proposed by Bindea et al. (2013) in 9,174 tumors across 28 cancer types (triple-negative tumors were considered part of the breast cancer cohort in this analysis). We executed the ssGSEA and GSVA implemented in the R Bioconductor package *gsva* (v.3.5) with default parameters [9]. The gene set enrichment scores computed through ssGSEA and GSVA exhibited a high positive correlation (Pearson's correlation coefficient: 0.87, P-value < 0.05; Figure 1).

We decided to carry out our analysis with the GSVA method because it includes a normalization step of the gene expression aimed at reducing the noise of the data, which has been shown to outperform ssGSEA when measuring the signal-to-noise ratio in differential gene expression and differential pathway activity identification analyses [10].

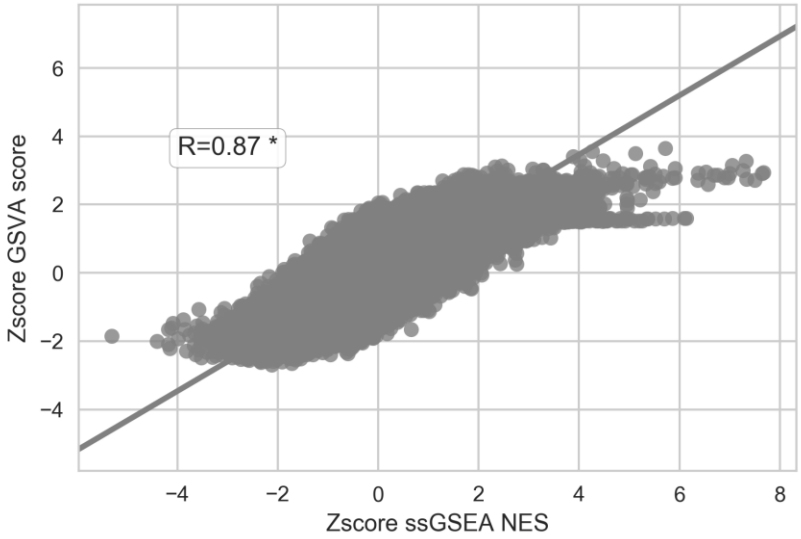


Figure 1. Comparison of gene set enrichment (relative abundance) scores obtained with the ssGSEA and GSVA methods for 23 gene signatures representing different immune cell populations.

Pan-cancer (9,174 tumors) regression plot between GSVA and ssGSEA scores of 23 immune populations (Pearson correlation coefficient, 0.87; P value <0.05). The gene set enrichment scores were normalized across the samples of each cancer type to a standard Z-score to make the scales comparable.

3.Selection of immune signatures for the gene set enrichment analysis

We obtained gene signatures identifying immune cell populations from the supplemental material of three different articles: (1) Bindea et al. (2013) [11], (2) Angelova et al. (2015) [12] and (3) Charoentong et al. (2017) [13]. The three studies used similar methodologies to build the gene signatures to quantify cell populations. The only dataset with experimental validation was (1). In addition, Senbabaoglu et al (2016) [14] validated the ssGSEA enrichment scores obtained for five immune populations (natural killer cells, CD8+ T cells, CD4+ T cells, regulatory T cells and Macrophages) using the genes sets of (1) with fluorescence-activated cell sorting and/or immuno-fluorescence.

Using the gene sets obtained from these three studies, we computed the GSVA scores of all cell populations across the pan-cancer cohort of 9,174 tumors. Then, we tested the correlation of the GSVA scores computed from pairs of gene sets of each source, with the aim of studying their capability to discriminate cell populations. GSVA scores computed for most cell populations using the gene sets from (1) exhibited a mild positive correlation (Figure 2, first panel). GSVA scores computed for most cell populations using the gene sets from (2) in most of the cases were either highly correlated or mildly anti-correlated (Figure 2, second panel). GSVA scores computed for most cell populations using the gene sets from (3) are highly correlated (Figure 2, third panel).

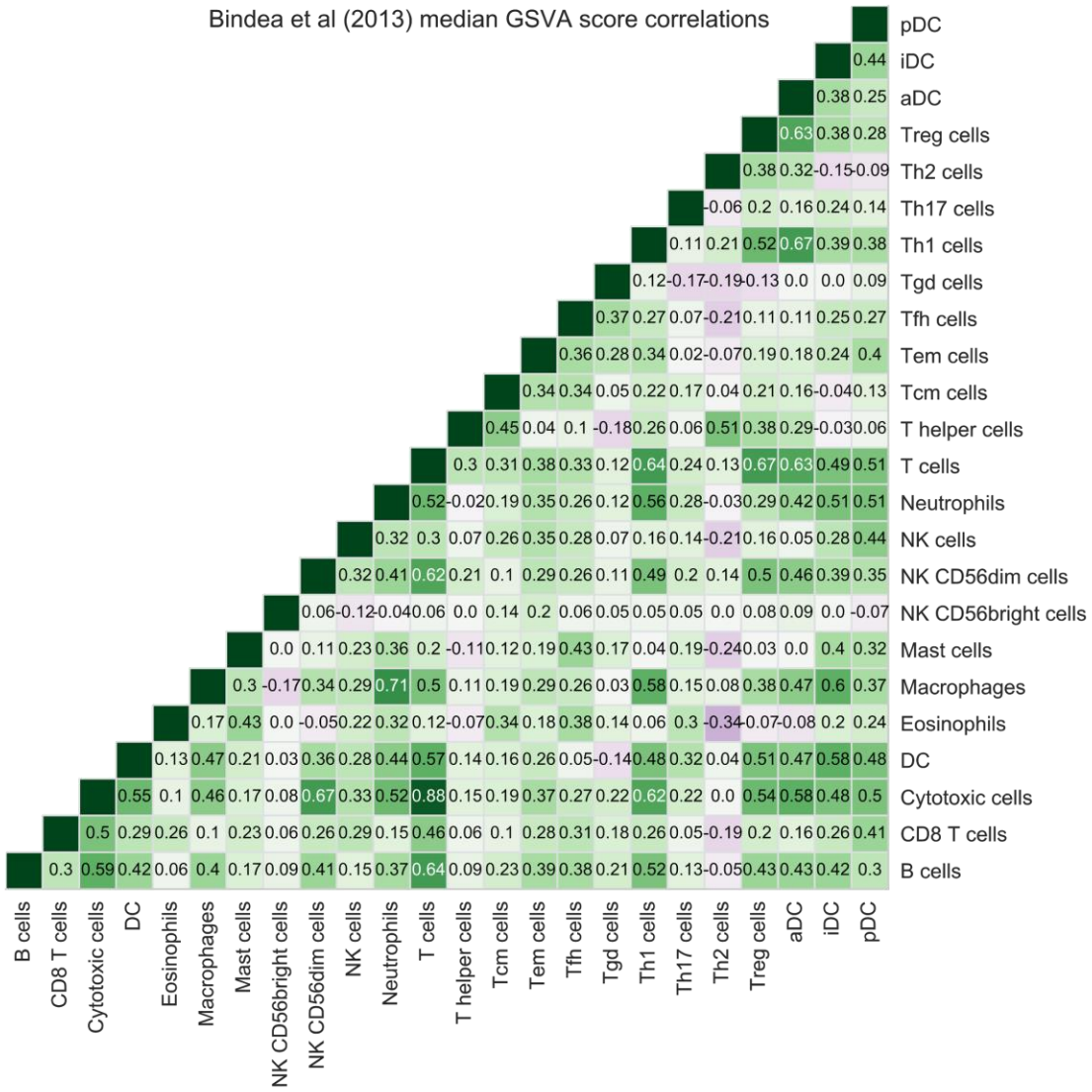


Figure 2 (1 out of 3)

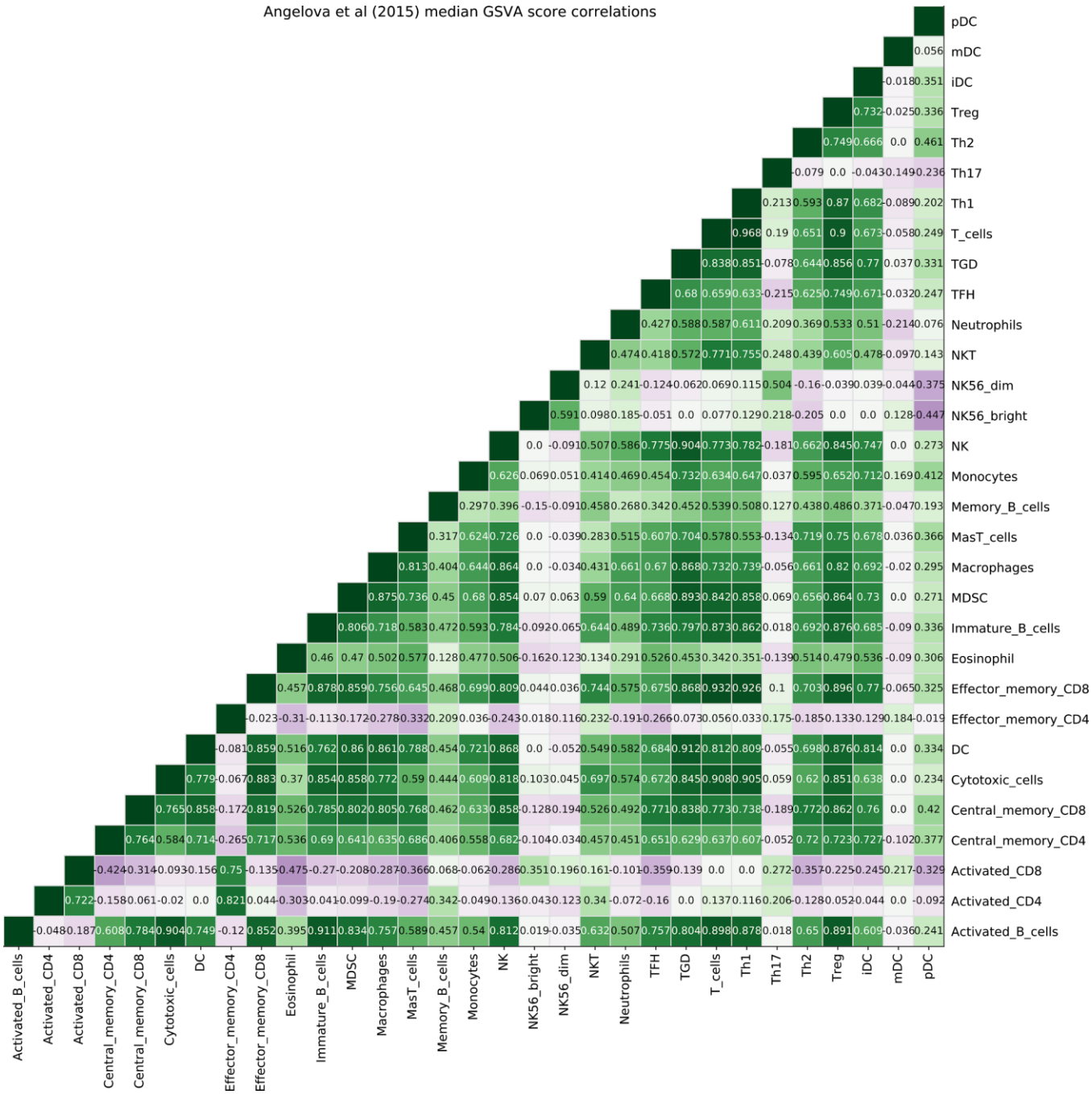

Angelova et al (2015) median GSVA score correlations
pDC
mDC
iDC
Treg
Th2
Th17
Th1
T_cells
TGD
TFH
Neutrophils
NKT
NK56_dim
NK56_bright
NK
Monocytes
Memory_B_cells
MasT_cells
Macrophages
MDSC
Immature_B_cells
Eosinophil
Effector_memory_CD8
Effector_memory_CD4
DC
Cytotoxic_cells
Central_memory_CD8
Central_memory_CD4
Activated_CD8
Activated_CD4
Activated_B_cells

Figure 2 (2 out of 3)

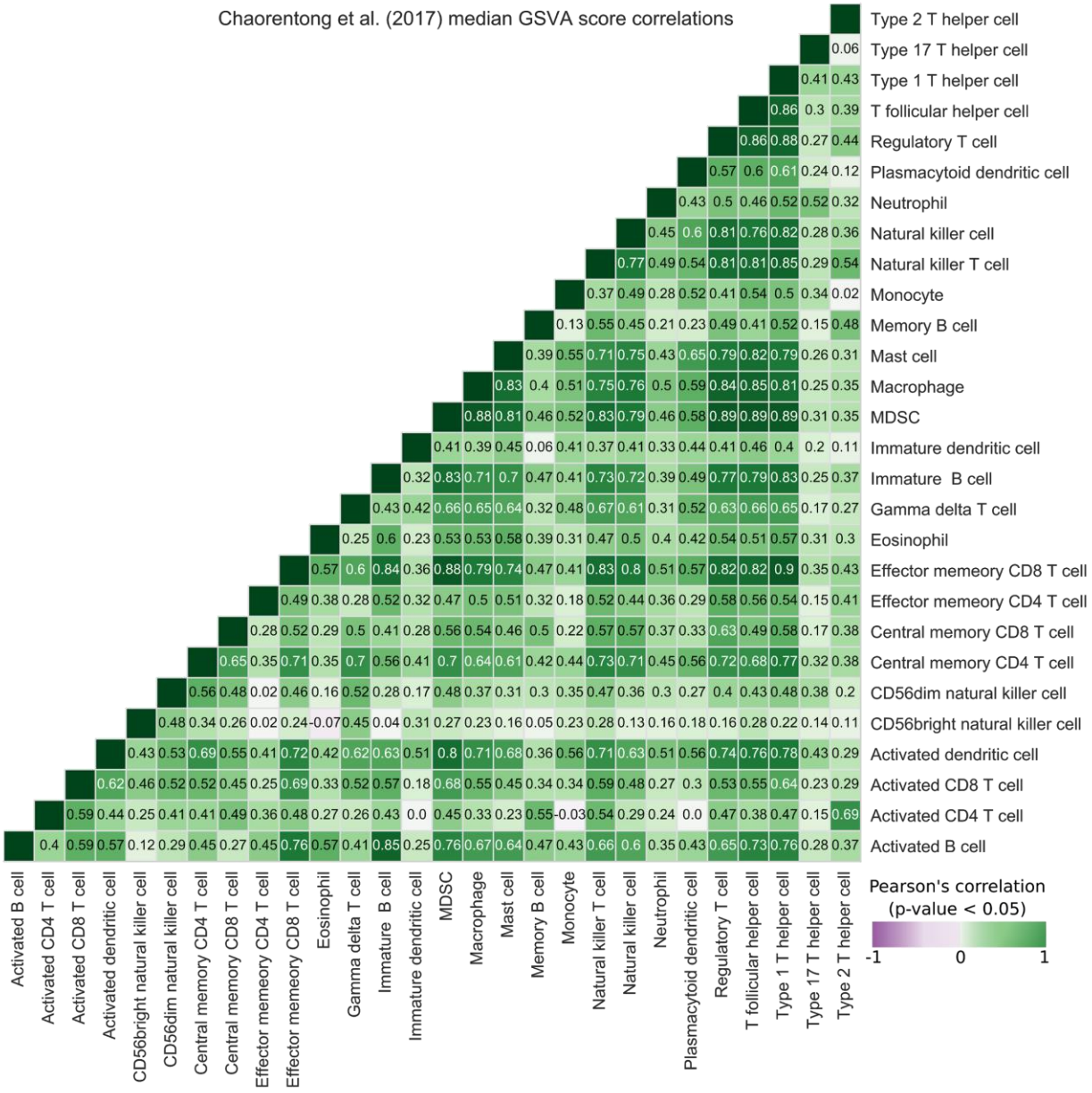


Figure 2. Correlation of relative abundance (GSVA scores) of immune cell populations computed using distinct groups of gene sets show different degrees of internal correlation.

The strong positive correlation observed among between most cell populations computed using the gene sets in group (2) and the fact that the observed anti-correlations contradicts biological knowledge led us to discard them. We next investigated the degree of agreement between the GSVA scores obtained using the gene sets obtained from (1) and (3) with genes (or combinations thereof) that constitute known markers of certain cell populations. Specifically, we compared the GSVA scores of (a) macrophages and monocytes with the expression of the myeloid marker *CD68*; and (b) effector immune populations (T cells CD8, NK cells and gamma delta T cells) with the total cytolytic activity of the infiltrate, computed as described by [15]; Figure 3).

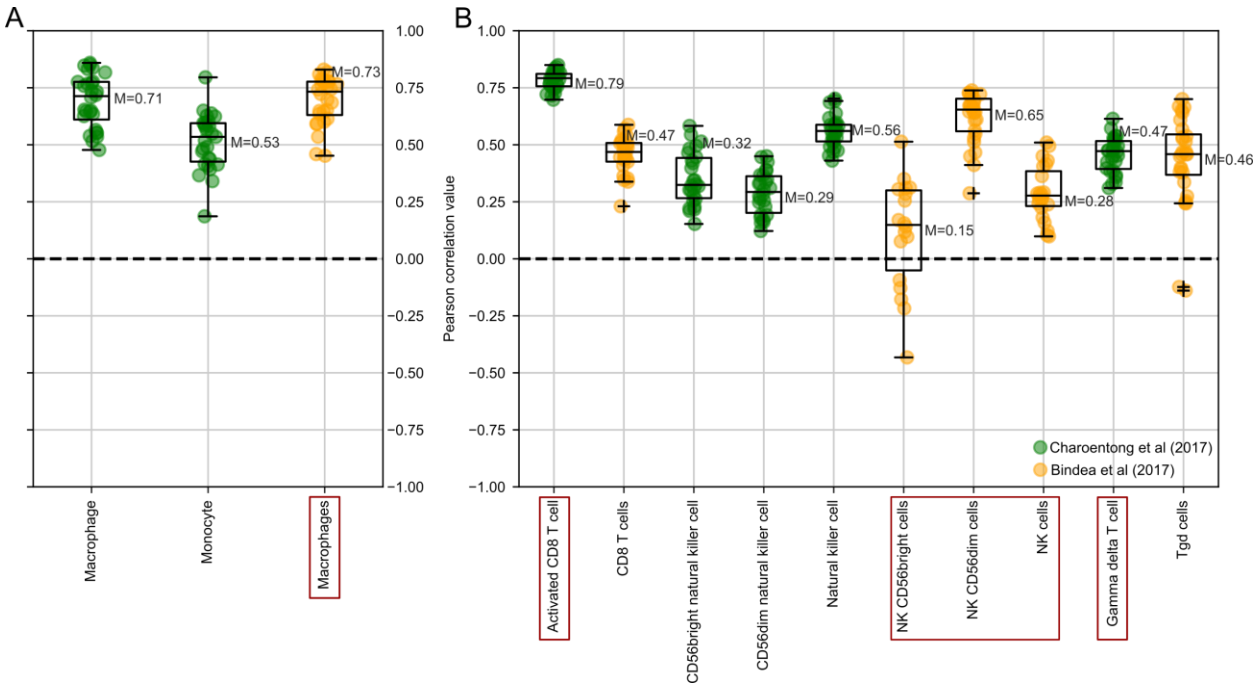


Figure 3. Agreement between GSVA scores and known markers of cell populations.

The graphs show the degree of agreement between GSVA scores computed using gene sets provided by Bindea et al (2013) and Charoentong et al (2017) for several immune cell populations and known markers of these cells. Panel A: macrophages and monocytes and the expression of the myeloid marker *CD68*. Panel B: effector immune cells (T cells CD8, NK cells and gamma delta T cells) and the cytolytic activity of the infiltrate, measured as the geometric mean of the expression of genes *PRF1* and *GZMA* [16]. Boxplots represent the distribution of Pearson's correlation coefficients across cancer types (each dot depicts the results in each

cancer cohort; in this analysis BRCA subtypes have been considered as a single cohort. Only dots that entail significant correlations are shown (corrected P-value < 0.1). Beside each boxplot the median of the Pearson correlation scores across cancer cohorts is shown. The gene set that we finally selected to represent the immune population is highlighted with a red square.

GSVA scores computed for macrophages, monocytes and natural killer cells using the gene set of (1) correlated better with their respective markers, whereas CD8 T cells and gamma delta T cells of (3) exhibited better performance. Thus, we decided to use CD8 T cells and gamma delta T cells from (3), and NK cell and macrophage gene signatures from (1). In addition, we attempted to further characterize the macrophage populations by exploring the immune population data matrix used by CIBERSORT, which distinguishes between M1 and M2 cells. However, the collective expression of the gene sets employed by this method to estimate the abundance of M1 and M2 cells (the top-ranked genes in the LM22 matrix) exhibited no correlation with the expression of *CD68*. This may be because the CIBERSORT framework was not primarily developed to analyze RNA-seq data (see next section). In addition, we found that the gene signature we selected to represent macrophages correlated strongly (Pearson's coefficient of 0.74) with *CD163,* a known marker of the M2 polarized macrophages, indeed it is part of the macrophage gene signature. This suggests that this gene set may predominantly represent the relative abundance of M2 macrophages in the infiltrate.

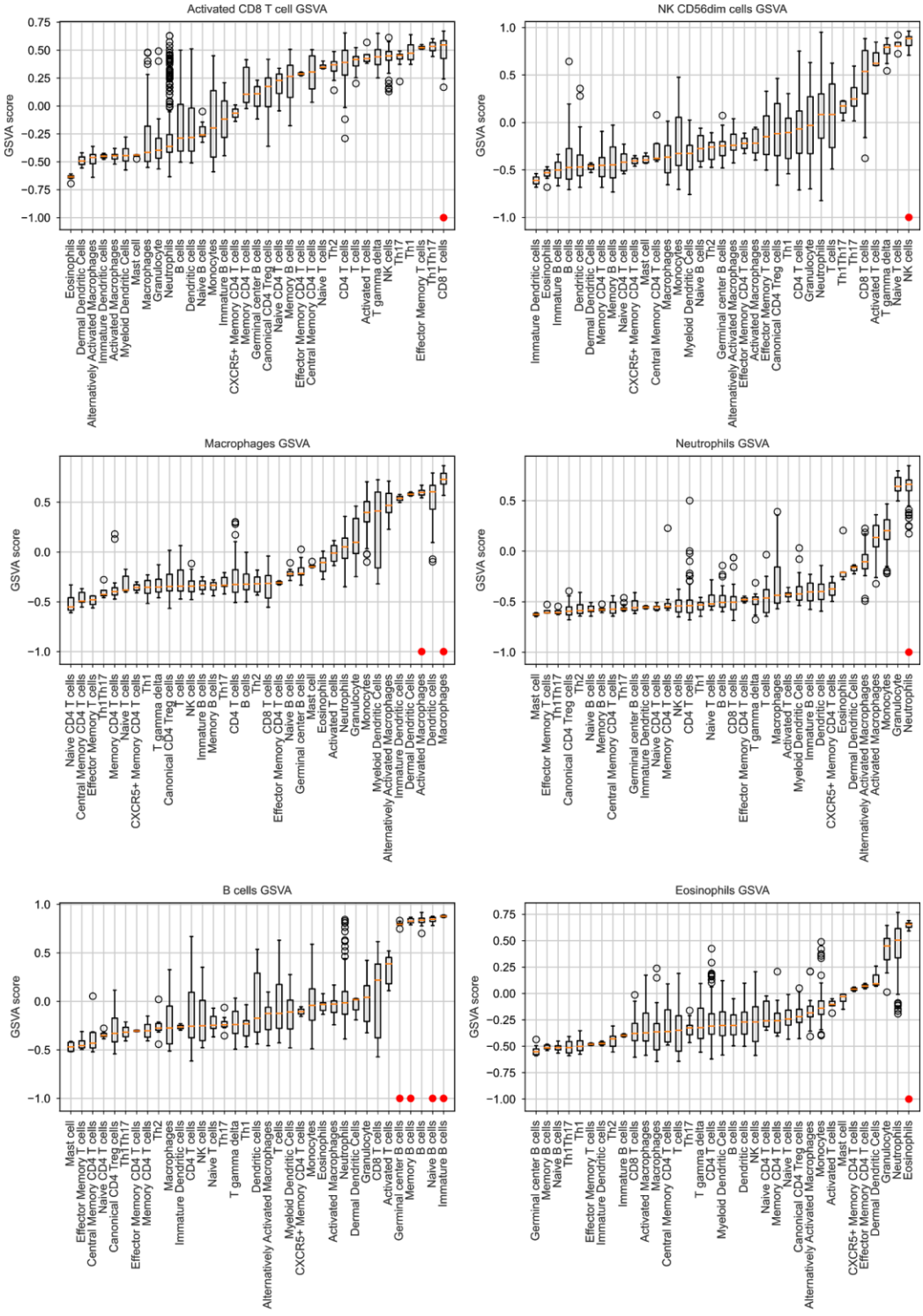
We also chose the gene set representing Regulatory T cells of (3), since the one provided in (1) contains a single gene (and was therefore unsuitable for a gene set enrichment analysis). For the same reason, we discarded the signature of plasmacytoid dendritic cells (pDC) of (1). Finally, when gene sets representing cells populations with different specificity (e.g. NK cells or NK CD56 bright and NK CD56 dim cells) we favored the set representing the most specific cell populations. An exception was the T helper cells, in which we did not used the gene sets available for subpopulations Th1, Th2 and Th17, since we observed that the relative abundance of these cells anti-correlated with that of other immune populations without any biological explanation. We included as independent populations the regulatory T cells and follicular helper T cells.

In summary, we collected 16 gene sets, three from (3) and thirteen from (1). Of note, gene signatures of (1) were preferred over those of (3) because they exhibited weaker correlations between them the immune populations represented by them. We reasoned that this implied that they were more capable to discriminate between immune populations. The 16 selected gene sets represented B cells, eosinophils, macrophages, mast cells, NK CD56 bright cells (NKbright), NK CD56 dim cells (NKdim), neutrophils, T helper cells, central memory T cells (Tcm), effector memory T cells (Tem), follicular helper T cells (Tfh), activated dendritic cells (aDC), immature dendritic cells (iDC), activated CD8 T cell (CD8+ T), gamma delta T cells (Tgd) and regulatory T cells (Treg). We also constructed a seventeenth gene set representing the relative abundance of all cytotoxic cell populations in the infiltrate, which we used to drive the definition of the immune-phenotypes of each cancer type (see Methods).

To further assess the quality of the selected gene sets we carried out two additional analyses: First, we corroborated that the selected immune population gene sets (immune genes) were not expressed in tumor cells. This was originally addressed by the authors that identified the signatures[2,3], but only for a limited set of cancer types in the former case. To carry out this analysis in a comprehensive manner, we downloaded RNA-seq data (RPKM) from the Cancer Cell Line Encyclopedia [17] (18-02-14 release). When there was more than one expression level per gene we kept the one with the highest median expression level across cell lines. We were able to map 367/401 of the immune genes and considered data for all solid tumor cell lines (n=816). We set the absolute threshold to consider a gene expressed at log2 RPKM=6. We found that most immune genes (358/367) were below this threshold (median=-0.87 log2 RPKM) and they could thus be considered as not expressed in solid tumor cell lines. These genes are therefore suitable candidates to identify immune cell-specific patterns.

Second, we corroborated the validity of the gene sets in an independent set of microarray experiments. We computed the GSVA scores of six cell populations (T cell CD8, NK dim, Neutrophils, Macrophages, Eosinophils and B cells) in microarray expression profiles from purified immune populations. We chose gene sets representing both cytotoxic and regulatory cell populations. We obtained the data of these experiments from a GEO super series dataset (GSE86362) collecting them [5]. We downloaded the harmonized annotation of the microarray dataset and the expression profiles matrix. We mapped the Affymetrix 133 Plus 2.0 probe

identifiers to gene symbols using Ensembl Biomart (v91). In cases in which more than one probe mapped the one gene we kept the probe with the highest median expression across all expression profiles. We thus obtained a matrix with 1,189 expression profiles of 21,880 genes. We next computed the GSVA scores of the immune populations across all expression profiles of immune cells in the GSE86362 dataset. To do this, we mapped the six populations under analysis to the corresponding populations represented in the harmonized annotation proposed by Bech et al (2016)[5], whose work collected the dataset. For these gene signatures we observed that the GSVA scores of the matched immune populations (e.g. activated CD8 T cells gene signature vs GSVA score of CD8 T cells) were always in the top-ranking (Figure 4).



Cell type	Experiments
Neutrophils	299
CD4 T cells	135
Monocytes	106
Canonical CD4 Treg cells	65
CD8 T cells	53
Granulocyte	44
Macrophages	43
T cells	37
Myeloid Dendritic Cells	36
NK cells	30
Dendritic cells	28
B cells	26

Cell type	Experiments
Alternatively Activated Macrophages	20
Th17	18
Memory CD4 T cells	16
Th1	15
T gamma delta	14
Germinal center B cells	13
Activated Macrophages	12
Th2	12
Memory B cells	8
Naive B cells	8
Central Memory CD4 T cells	7
Eosinophils	7

Cell type	Experiments
Naive CD4 T cells	6
Activated T cells	5
Naive T cells	5
Effector Memory T cells	4
Mast cell	4
Th1Th17	4
Dermal Dendritic Cells	3
CXCR5+ Memory CD4 T cells	2
Effector Memory CD4 T cells	2
Immature B cells	2
Immature Dendritic cells	2

Figure 4. Distribution of GSVA scores computed from gene sets representing immune populations on microarray expression experiments of purified immune populations.
The boxplots represent the distribution of GSVA scores computed from the gene sets employed in our study and spelled at the title of each panel using microarray expression data collected at the GSE86362 GEO entry. Each boxplot groups the GSVA scores computed with the expression of experiments of a given purified cell population (marked at the corresponding x axis label). The purified cell population (or populations) that most closely resembles that represented by the gene set assayed at each graph is marked by a red dot. At the bottom, the number of experiments probing the expression of each cell type

4.Comparison of performance of a deconvolution method and a gene set enrichment method

We compared the performance of the GSVA with that of a deconvolution method, CIBERSORT [18,19]. This method receives as input the expression of genes in a bulk sample and estimates the fraction of 22 cell types (using the gene weights included in the LM22 matrix) within the overall immune infiltrate. CIBERSORT is implemented as a linear support vector regression trained with microarray immune cell population-specific expression data. Since the gene expression data measured with microarray and RNA-seq technologies follow different distributions, the analysis of RNA-seq samples with CIBERSORT is bound to provide an incorrect LM22 matrix fitting, and its performance has been not validated. Therefore, CIBERSORT may not be suitable for the use of the RNA-seq data of TCGA cancer cohorts. Moreover, the expression of some of the genes in the LM22 matrix is absent from the TCGA RNA-seq expression data (see below), which probably may further affect the performance of the method.

To evaluate the performance of CIBERSORT on RNA-seq data, we first compared the results obtained by the method for samples in which both RNA-seq and microarray data are available in TCGA. These include a subset of ovarian cystadenocarcinomas (OV), glioblastomas (GBM) and lung squamous carcinomas (LUSC). First, we downloaded microarray HG platform (level 2) data on probe-sample-intensities from the GEO repository. If more than a probe mapped to a gene, the probe providing the highest mean intensity across samples was kept. Probe-symbol mapping was done with Ensembl v79, followed by manual mapping to guarantee maximum coverage of the LM22 matrix. Overall, we retrieved intensity values for 533 out of the 548 genes

in the LM22 matrix. Additionally, we performed a quantile-normalization of microarray data (microarray q-normalized data) following standard procedures. Figure 5 shows the weight that the CIBERSORT algorithm assigns (for the quantification of immune cell populations) to each of the 15 genes that we failed to recover from the TCGA microarray data. Some of these missing genes, such as *TRBC1*, receive a high weight for the estimation of the proportion of certain immune cell populations. The absence of these genes from the input expression data may thus affect considerably the quantification of the corresponding cell types.

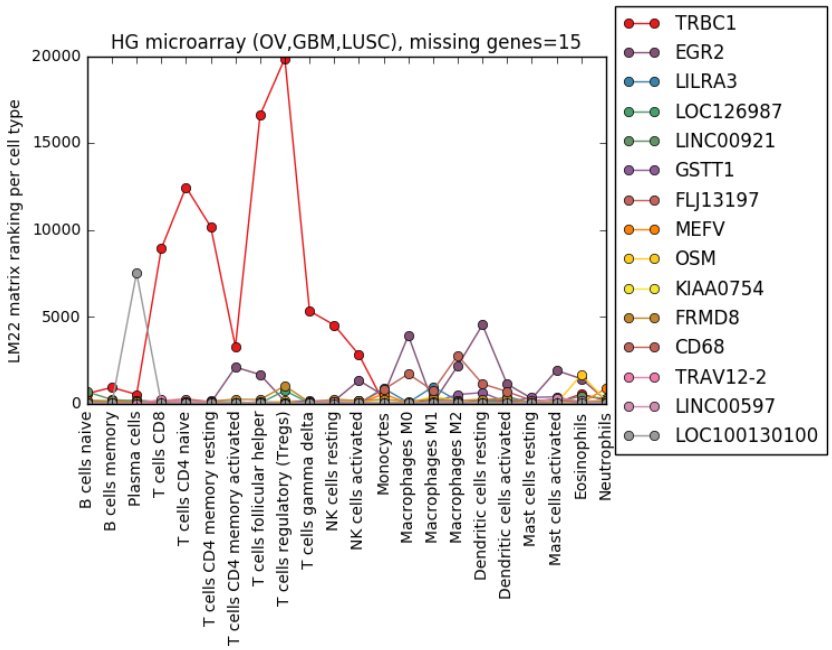


Figure 5. Weight in CIBERSORT of genes absent from TCGA microarray data across immune cell populations.

CIBERSORT estimates of the immune cell proportions shows a moderate to low correlation (e.g. Pearson correlation of 0.32 for the evaluated samples in OV) between 'paired' RNA-seq and microarray data (i.e. data from the same tumors), which varies depending on the immune cell population (data not shown). Next, we reasoned that this correlation could improve by transforming the RNA-seq data to a microarray-like distribution. To do this, we followed the procedure described by Charoentong et al. (2017) to transform the RNA-seq data of the OV, GBM and LUSC cohorts. Then, we built a univariate cubic smoothing spline (CSS) model with four degrees of freedom of the TCGA microarray data using the class *interp1d* from the

scipy.interpolate Python package [20]. With this model, we transformed the RNA-seq data of the OV, GBM and LUSC cohorts to obtain a microarray-like distribution of values. We next fed the transformed RNA-seq expression matrix to CIBERSORT and compared these results with those of the paired microarray expression data in each cohort using a leave-one-out cross-validation approach. Joining the results of the three cohorts, we observed an increase of the overall correlation (Pearson's correlation 0.79, P-value < 0.05; Figure 6) of the immune cells proportions estimated using both input data types.

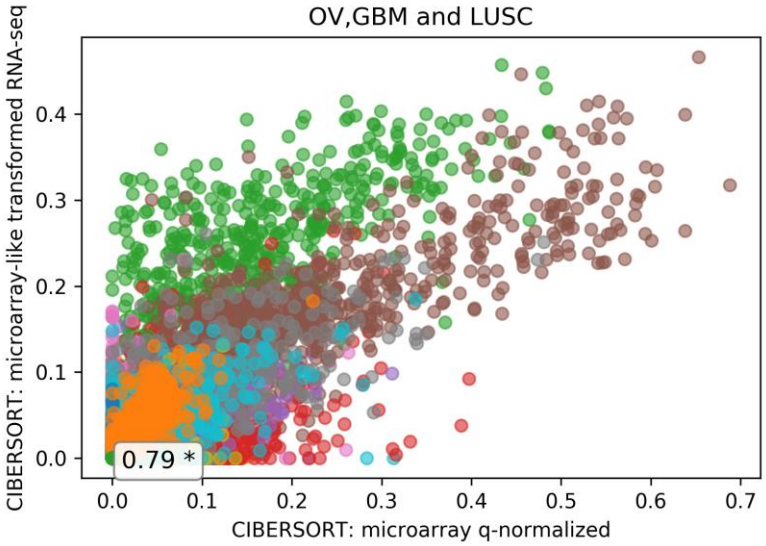


Figure 6. Correlation of CIBERSORT results obtained with microarray expression data and RNA-seq microarray-like transformed expression data in the three cohorts of tumors (OV, GBM and LUSC). Each circle represents the CIBERSORT output value of a given immune population (colored in distinct colors) in a sample, with coordinates representing the proportion of that cell type within the bulk sample admixture. The Pearson's correlation coefficient of the full set of cell populations is displayed.

However, the correlation varies across the immune cell populations computed by the method (Figure 7). On one end of the spectrum, macrophages (both M1 and M2), CD8 T cells, neutrophils, plasma cells and activated dendritic cells exhibited a good correlation (Pearson's correlation > 0.6); on the other, eosinophils, T cells CD4 naive, B cells memory, dendritic cells

resting, NK resting, monocytes, T gamma delta and B cells naive cells exhibited Pearson's correlation coefficient values below 0.4.

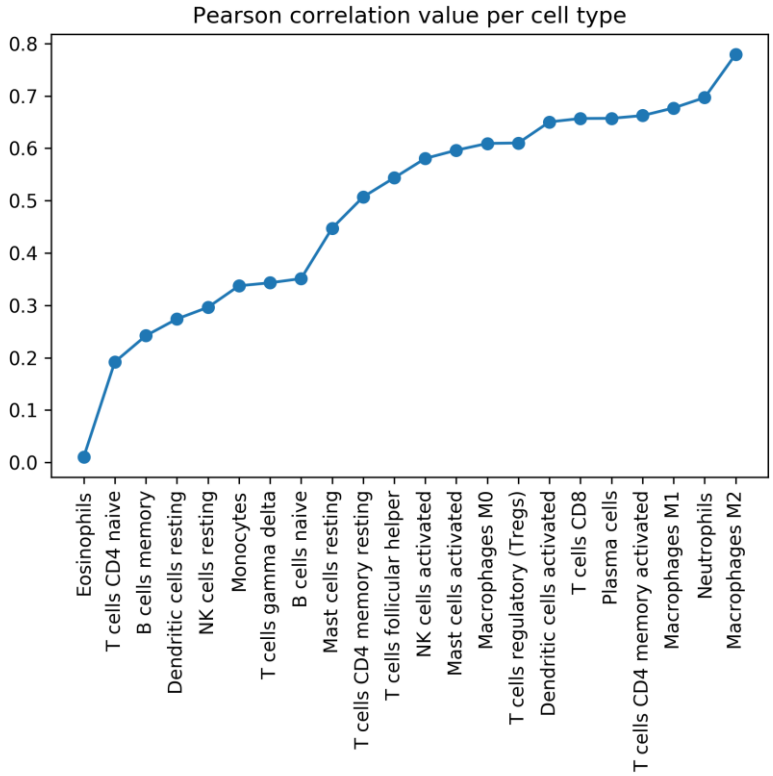


Figure 7. Correlation of CIBERSORT computed abundance of immune cell populations on transformed RNA-seq data with real microarray data sorted by increasing Pearson's correlation coefficient.

In conclusion, CIBERSORT results with RNA-seq input data largely differ from those obtained with microarrays. The attempt to transform the distribution of RNA-seq data to a microarray-like distribution improved the overall correlation, but major differences were still observed for certain individual immune cell populations. This result highlights the caveats that the use of this method to analyze TCGA RNA-seq data would present.

Next, we evaluated to which extent CIBERSORT results, which estimate the proportion of each immune cell population in the sample independent of its degree of overall infiltration, differ from those obtained with a gene set enrichment approach, which does take into consideration the

dimension of the immune infiltrate. First, we compared the results of CIBERSORT and GSVA with the RNA-seq data available for the 9,174 TCGA tumors (previously transforming them to fit a microarray-like distribution as explained before). To that end, we mapped the immune cell populations considered by CIBERSORT to the ones represented by the gene sets used in our GSVA approach (see section 3). Of note, we observed weak correlations for most cell types. Moreover, we found anti-correlations for some cell types (e.g. NK cells activated from CIBERSORT with Cytotoxic cells from GSVA; Figure 8, panel A).

Then, we evaluated whether these correlations could increase when incorporating to the CIBERSORT results the component of the overall immune infiltration. To do this, we used the Immune score computed by the ESTIMATE method [21]. When multiplying the CIBERSORT values by this estimation of the overall immune infiltration, we observed that the correlations increased (Figure 8, panel B).

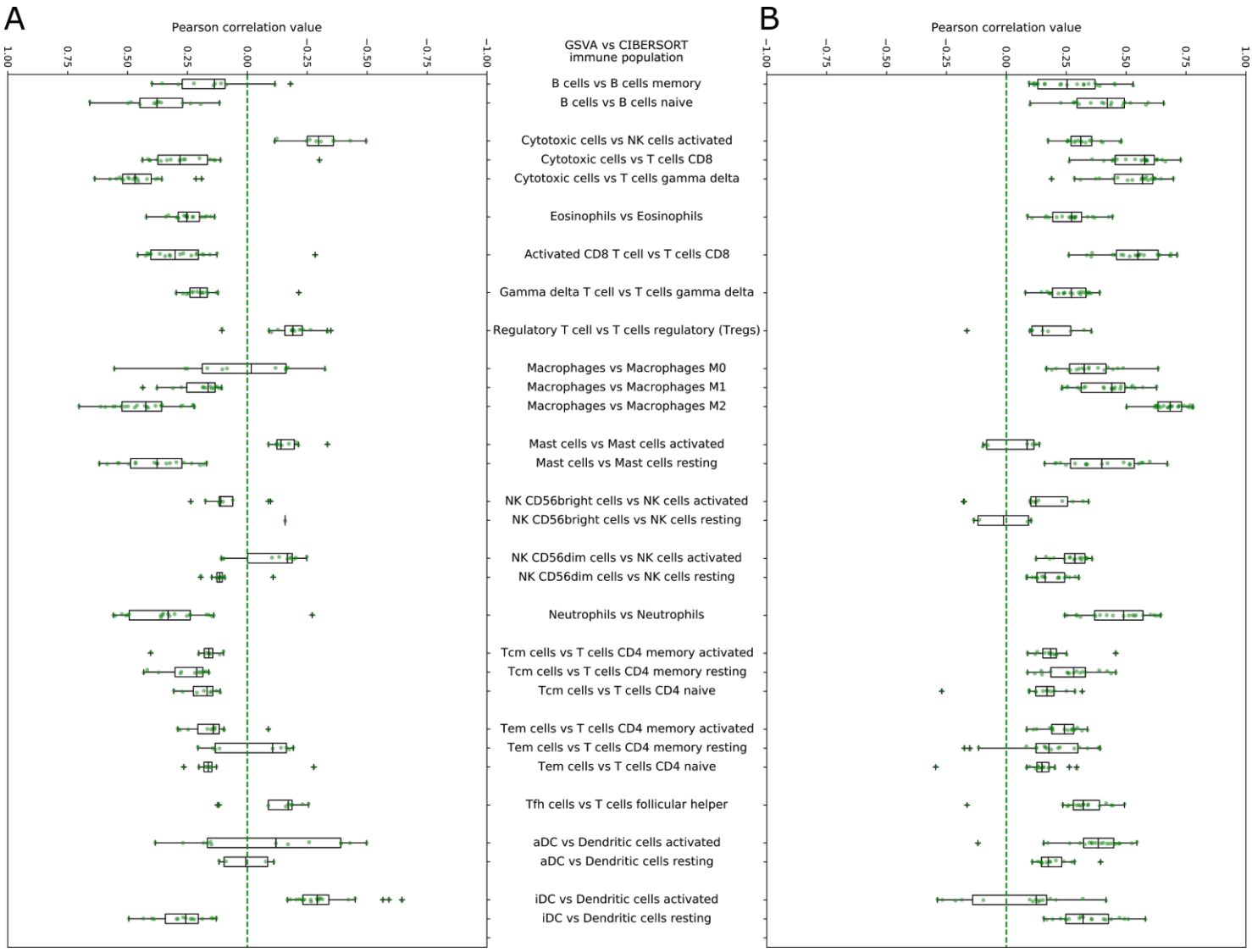


Figure 8. Correlation of CIBERSORT and GSVA estimations of immune cell populations infiltration 9,174 tumors in the TCGA pan-cancer cohort.

The box plots represent the distribution of Pearson's correlations coefficients between GSVA scores and CIBERSORT fractions of a given immune cell population. The name of the immune populations evaluated by GSVA and their CIBERSORT counterparts is shown at the center. In panel A, CIBERSORT fractions correspond to CIBERSORT output values, while in panel B, they are the product of CIBERSORT fractions multiplied by the ESTIMATE Z-score values. Each dot depicts the correlation observed in each cancer cohort.

In summary, another caveat of CIBERSORT is that it does not consider the overall abundance of immune infiltrate in the tumor, and thus the results need to be combined with an additional proxy of the overall immune infiltrate in the tumor sample. Of note, as of August 2017, CIBERSORT website announced a release of a beta version of the tool that scales the proportions of different cell populations by the overall immune infiltrate, calculated as the median expression of LM22 genes divided by the median expression of all genes in the sample. While new versions of CIBERSORT may be released in the future that are ready to process RNA-seq data, to the best of our knowledge, at the moment of writing this manuscript no such version of the method that is able to solve the caveats of analyzing RNA-seq data described here is available. Therefore, and as mentioned in the first section, we decided to use a sample level gene set enrichment method (GSVA) to calculate the immune cell infiltration patterns.

5. Hierarchical clustering of tumor immune infiltration patterns

Hierarchical agglomerative clustering of the tumor samples represented as 16-component vectors of GSVA scores was carried out using the Euclidean distance and Ward's linkage function [22]. To this end, we used the class *hierarchy* implemented within the Python's clustering package *scipy.cluster* [23].

For a given integer *n*, the method provides a partition of the set of tumor samples into *n* clusters. To determine the most appropriate value of *n*, we computed the proportion of inter-tumor variance explained (VAR) by the clustering for different values of *n*.

$$VAR = 1 - (SSE/SST) \qquad SST = \sum_{j=1}^{m} d(x_j, \bar{x})^2 \qquad SSE = \sum_{i=1}^{n} SSE_i \qquad SSE_i = \sum_{x \in C_i} d(x, \bar{x}_i)^2$$

In this equation, *m* is the number of samples, *i* is the index of each cluster and C_i represents the *i*-th cluster; *d* stands for the Euclidean distance; and $\underline{x}, \underline{x}i$ represent the cohort and cluster centroids, respectively.

In the case of the pan-cancer pooled analysis, we observed that although the rate of increase of the VAR diminished for values of *n* greater than 9, it did not reach a plateau --at least for values of *n* up to 30. A clear cutoff was therefore not observed here (Figure 9).

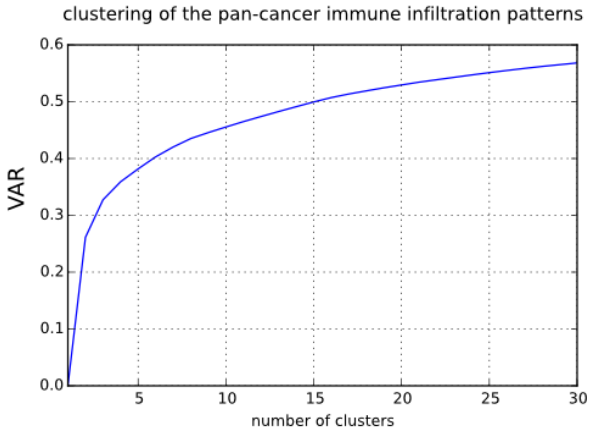


Figure 9. Proportion of explained variance (VAR) when grouping the immune infiltrate patterns of the pan-cancer cohort into *n* immune-clusters, with *n* ranging from 1 to 30.

Therefore, and as an additional criterion to choose the optimum value of *n*, we evaluated how the distribution of tumors of each cancer type across the pan-cancer immune clusters changed as a function of *n* for values greater than 9. To address this question, we calculated the amount of information of the distribution of tumors of each cancer type across the *n* pan-cancer clusters. Namely, for each cancer type *i*, we computed an entropy score H_i as follows:

$$H_i = -\sum_{k=1}^{n} p_{ik} \log_2 p_{ik}$$

where p_{ik} is the proportion of tumors of cohort *i* grouped in cluster *k*. We observed that the median entropy of cancer types with lower values of H_i plateaued approximately at *n* equal to 17 (Figure 10). Therefore, we clustered the pan-cancer immune infiltrate pattern into 17 groups.

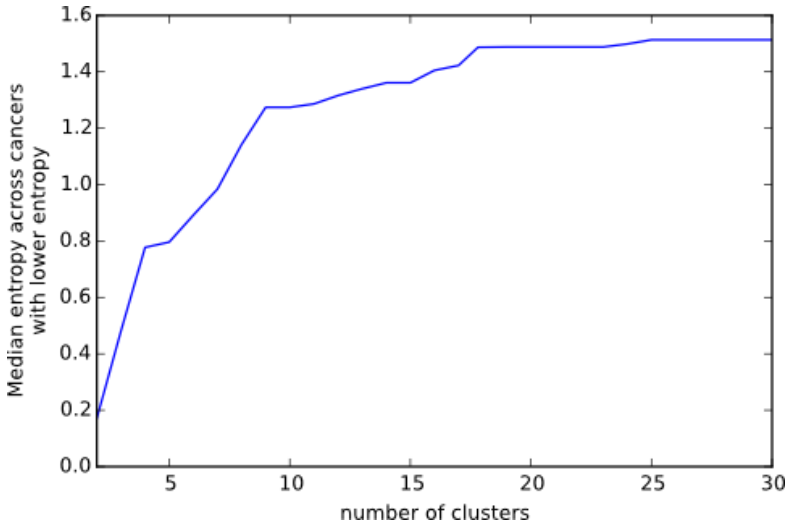


Figure 10. Median H_i values as a function of the number of clusters, computed for the 10 cancer types with more homogeneous infiltration pattern.

We then wanted to define per-cancer immune-phenotypes (clusters of the immune infiltration pattern of the tumors in each cancer type) that reflected the selective pressure applied by the immune infiltrate on the tumor. To this end, we incorporated to the immune infiltration pattern a gene set representing the relative abundance of all cytotoxic cells. We then over weighted the GSVA scores of this population in the hierarchical clustering algorithm. We assayed several weights for the cytotoxic cell population and found that the samples included in each cluster remained stable (Jaccard index > 0.8) for weights larger than 3.5 (Figure 11). We thus decided to overweight the cytotoxic cells population by a factor of four, thus giving it 25% of the total contribution to the clustering.

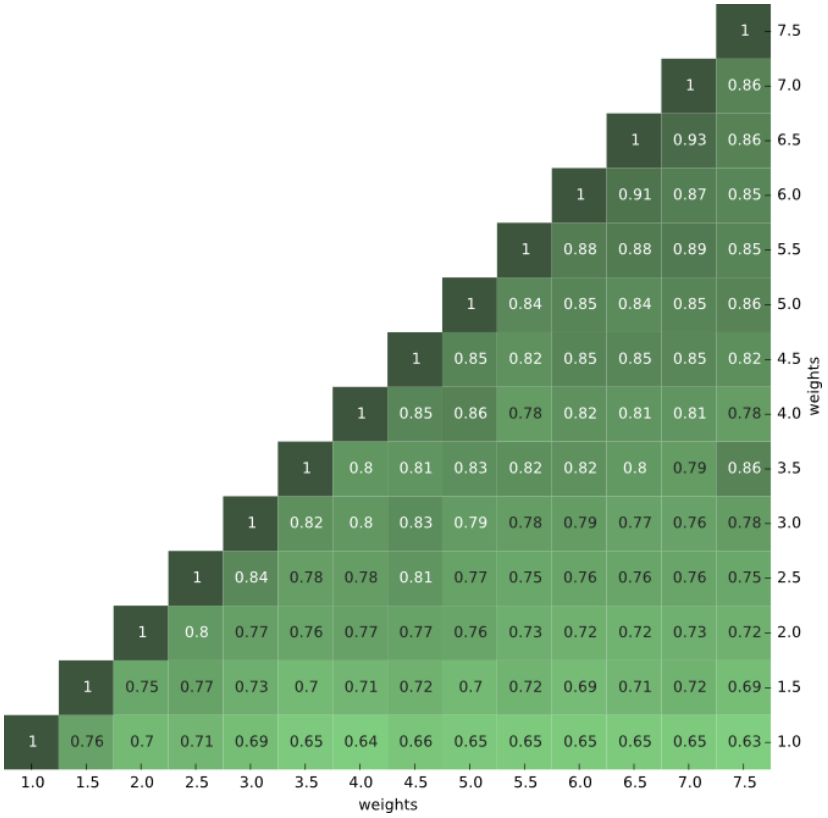


Figure 11. Effect of different weights of cytotoxic cell populations on the structure of immune-phenotypes.

The heatmap represents the overlap (Jaccard index) between different groupings of tumor samples of each cohort that is obtained with different weights (ranging from 1 to 7.5) of the cytotoxic cells population.

Finally, to determine the optimum number of clusters (*n*) from the GSVA score matrices of each cancer cohort (i.e. the number of immune-phenotypes), we measured the fraction of inter-tumor variance explained as a function of *n* (as explained for pan-cancer clusters above). For most cohorts, the function began to plateau at $n = 3$ or $n = 4$ and reached a plateau at $n > 6$ (Figure 12). To facilitate the comparison across cancer types, we decided to define the same number of groups in all cohorts. Therefore, we grouped the immune infiltration patterns of tumors in each cohort into six groups (named immune-phenotypes).

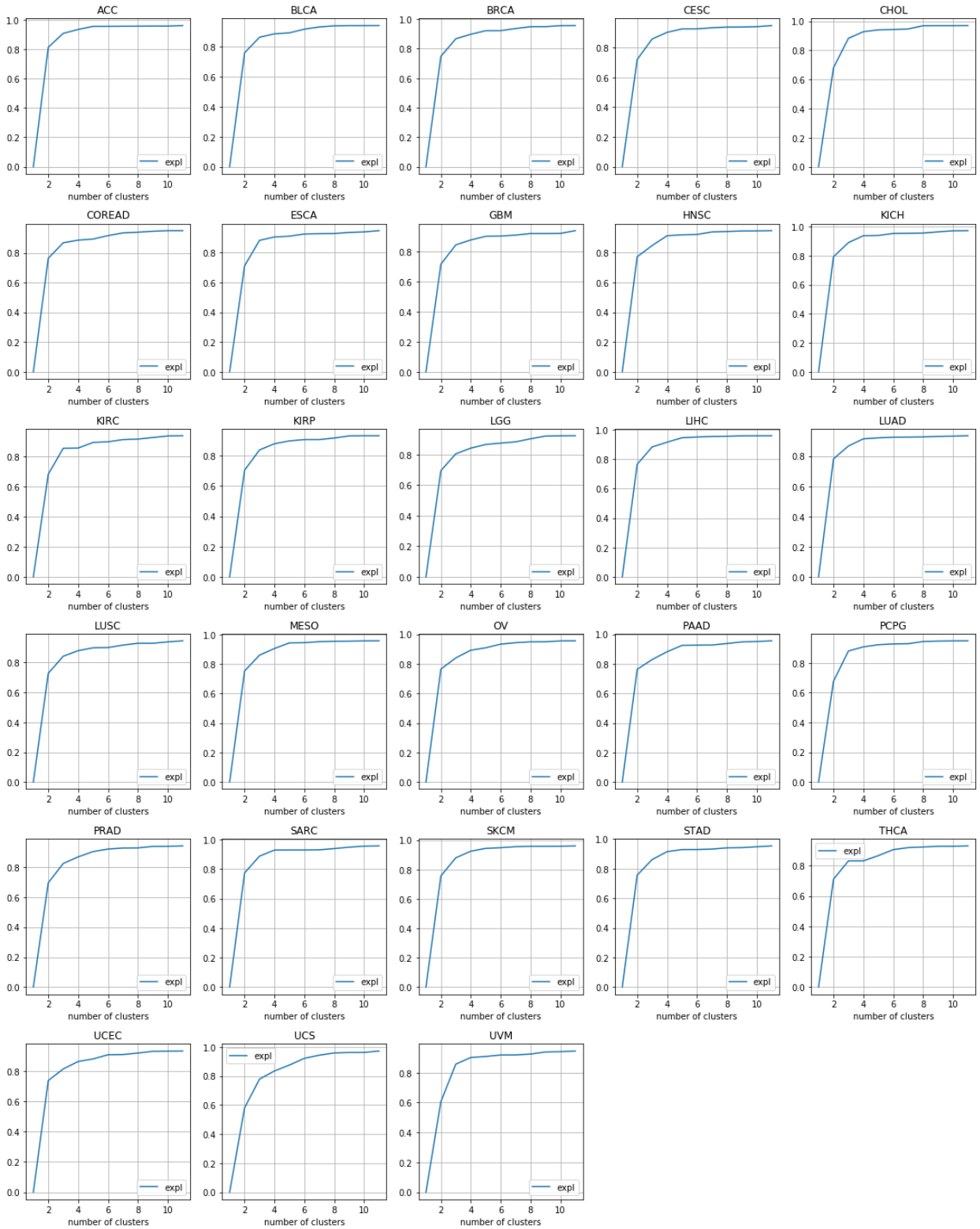


Figure 12. Computing the optimal number of immune-phenotypes across cohorts.

Each plot presents the fraction of inter-tumor variance (vertical axis) of the immune infiltration pattern explained by immune-phenotypes as a function of their number (horizontal axis), *n*. The breast cancer (BRCA) cohort includes triple-negative tumors in this analysis. The values of *n* assayed range from 1 to 12.

6. Immune infiltration of healthy tissues

The expression of genes across tissues of healthy donors was downloaded from the GTEx portal (https://gtexportal.org/home/) (sample level RPKM matrix, v6). In cases of duplicated gene symbols, we kept the entry with the highest median expression. Twenty-two healthy tissues (6,544 donors) represented in the GTEx dataset were mapped to their corresponding TCGA cancer type (e.g. Lung was mapped to LUAD and LUSC) (see Table 1 below).

Healthy tissue	Matching tumor tissue(s)	Number of healthy donors
Adipose Tissue	SARC	577
Adrenal Gland	ACC,PCPG	145
Bladder	BLCA	11
Brain	LGG, GBM	1259
Breast	BRCA	214
Cervix Uteri	CESC	11
Colon	COADREAD	345
Esophagus	ESCA	686
Kidney	KICH, KIRC, KIRP	32
Liver	LIHC	119
Lung	LUAD, LUSC	320
Muscle	SARC	430
Nerve	SARC	304
Ovary	OV	97
Pancreas	PAAD	171
Prostate	PRAD	106
Salivary Gland	HNSC	57
Skin	SKCM, HNSC	890
Stomach	STAD	192
Testis	TGCT	172
Thyroid	THCA	323
Uterus	UCS, UCEC	83

Table 1. Healthy donors sample collection. Table with the mapping of the 22 healthy GTEx tissues to TCGA cancer types. Note that a healthy tissue may be mapped to more than one cancer type. The number of healthy donors with available RNA-seq data available for every tissue is shown.

Of note, since the expression of genes across TCGA tumors and samples from GTEx healthy donors were processed with different pipelines, we could not directly compare them. Instead, we compared the median relative GSVA values of each immune cell population across the tumors of a given cancer cohort (computed by the GSVA analysis of the 9,174 tumors together) with that of the samples of the matching healthy tissue (computed via GSVA analysis of the 6,544 healthy donors together).

7. Identification of immune-phenotype associated genomic drivers

With the aim of identifying genes whose driver alterations appear enriched for immune-phenotypes, we implemented a regularized logistic regression analysis. We selected this approach given the sparsity of the data, as it reduces spurious fitting through the introduction of a tolerable degree of bias in exchange of reducing the variability of parameter estimates. Specifically, the method assumes weakly-informative prior distributions for the parameters to enhance the fitting stability [24]. The method provides enrichment effect-sizes and empirical significance scores for the association observed between the response variable and the clustering.

We considered a dataset with the following features defined for each sample: i) a variable y taking values in {0,1}, representing whether a certain event has been observed in the sample; ii) a categorical covariate C with levels $\{C_0,\ldots,C_k\}$ encoding the immune-phenotype of the sample; iii) a set of additional covariates associated to each sample denoted $Z=[Z_1,\ldots,Z_m]$ (and representing the burden of mutations and copy number alterations). We wanted to determine the enrichment of observed events ($y=1$) across the distinct levels of C in the dataset. To do this, for each possible level C_j of C we estimated the conditional probability $p(y=1|C=C_j,Z)$, i.e. corrected for the influence of the covariates Z.

After encoding the *k*+1 levels of *C* into *k* new dummy variables $W = [W_1,\ldots,W_k]$, we denoted $X = [1, W, Z]$ the full set of covariates, including an intercept column. We thus proceeded to fit the following logistic regression model:

$$\log\left(\frac{p}{1-p}\right) = X \cdot \beta$$

where $p = p(y=1|X)$ and $\beta = (\beta_0,\ldots,\beta_{k+m})$

Under particular circumstances, e.g., when the data contains a large proportion of zeros, it is possible that the maximum likelihood estimator for the parameters *β* does not exist or converges slowly at high values, which poses difficulties for the interpretation [25]. To prevent this artifact, we added a penalty term to the logistic log-likelihood of the logistic model, denoted: . In other words, $\tilde{l}(\beta, X)$, we conducted a maximum likelihood estimation (MLE) of *β* with the following penalized log-likelihood objective function:

$$\tilde{l}(\beta, X) = \sum_{i=1}^{N} y_i \log(p_i) + (1 - y_i)\log(1 - p_i) - \frac{1}{2}\bar{\beta}^t \Sigma^{-1}\bar{\beta},$$

where

$$p_i = p(y_i = 1|X_i) = \frac{\exp(X_i\beta)}{1+\exp(X_i\beta)}$$

with $\underline{\beta} = \beta - \mu$ the parameters centered in the prior means *μ* and $\sum = diag(\sigma_0^2,\ldots,\sigma_{k+m}^2)$ the prior covariance matrix. This penalty has a Bayesian interpretation, in the sense that the optimization procedure provides a Bayesian update of the joint Gaussian distribution $N(\mu, \sum)$ taken as prior distribution of the covariate parameters. We then set as weakly-informative prior distribution a joint Gaussian distribution with mean vector $\mu=(0,\ldots,0)$ and prior standard deviations 10 for the intercept coefficient and 2.5 for the rest of the coefficients, following the default setting described in [26]. We implemented this methodology using the Python package *bayes_logistic* [27]. Using this MLE approach we provided the mean of the posterior distribution for the parameters *β*, and the posteriori covariance matrix $\hat{\Sigma}_X$ from which we can derive confidence intervals.

To assess the strength of association between the response variable and the clusters covariate C, we computed the differences of the log-likelihoods of two penalized logistic regression models evaluated at their respective MLE estimates. Specifically, we defined a LOD score as:

$$LOD = \tilde{l}(\hat{\beta}_X, X) - \tilde{l}(\hat{\beta}_Z, Z).$$

Each log-likelihood was computed after fitting a different penalized logistic regression model, namely, one based on the full set of covariates (*X*), another based only the adjustment covariates (*Z*, i.e. the number of copy number alterations, point mutations and small indels), with $\hat{\beta}_X$ and $\hat{\beta}_Z$ being the respective MLE estimates derived for each model.

To provide an empirical significance score for the observed strength of association between the response variable and the clustering, a null collection of LOD scores was generated by resampling the levels of *C*. Using these resampled levels of *C*, we generated a null set of LOD scores (n=10,000). We then used this null set to derive an empirical p-value for each LOD score. Thus, for each response variable (i.e. genomic event) we obtained a LOD score, and a significance.

8. Adjustment of tumors expression for its immune component

To adjust the expression of genes in tumors for its immune component we followed the rationale described by Aran et al. (2016) [28] (Figure S1A). To carry out this adjustment, we used the expression of *CD45* (taken as a proxy of the overall leukocyte infiltrate) across the same set of samples. Briefly, from GTEx samples of matching healthy tissue of each cancer cohort, we first regressed the expression of each gene as a function of the expression of *CD45*, fitting the data into a degree one polynomial, with the Python Numpy *polyfit* implementation [29]. We then used the equation of the regression line to subtract from each tumor, of the cohort, the contribution of the infiltrate to the overall expression of the gene. We were not able to adjust the expression of genes in tumors of UVM, CHOL and MESO for lack of matching healthy tissue in GTEx.

To assess the effect of the adjustment we looked at specific genes described to be expressed (i) mostly in immune but not tumor cells (e.g., *PD1*), (ii) in both immune and tumor cells (e.g., *HLA-A*), and (iii) mostly in tumor but not immune cells (e.g., *NOTCH1*). While in tumors the expression of *PD1* showed an important decrease upon adjustment -as expected from its strong

correlation with *CD45* expression-, the expression of *HLA-A* was only partially lowered, and that of *NOTCH1* remained virtually unchanged (Figure S1B).

Moreover, we explored the effect of the expression adjustment on the enrichment of gene sets. Again, as expected, we observed that the collective overexpression of several immune-related pathways was appreciably lowered upon correction. On the other hand, the collective overexpression of pathways related with tumorigenesis remained unchanged after adjustment (Figure S1C).

9. Pathway enrichment analysis across immune-phenotypes

We ran a gene set enrichment analysis (GSEA) [30] using the code available in http://software.broadinstitute.org/ on the adjusted expression of the genes in pathways of interest (Table S1B). We added small modifications to adapt the GSEA to R version 3. The expression of the genes of a given pathway in each immune-phenotype was compared against that in all other immune-phenotypes. Each TCGA cohort was analyzed separately. If a pathway was found significantly up-regulated in more than one immune-phenotype of a cohort, we report only the result with the highest Normalized Enrichment Score (NES).

SUPPLEMENTARY NOTES

Metastatic melanoma patients treated with immune checkpoint inhibitors

We analyzed the transcriptomic data of tumors of two cohorts of patients with metastatic melanomas that received anti-PD1 (n=28) or anti-CTLA4 (n=42) checkpoint blockers [31,32]. First, we assessed whether the response to each drug correlated with the immune infiltration pattern. Following the same approach employed for TCGA tumors, we computed the immune infiltration pattern of tumors as the GSVA scores of 16 immune cell populations and cytotoxic cells across both cohorts. We observed that the tumors of patients with weaker response to the anti-CTLA4 inhibitor were enriched for lowly cytotoxic immune infiltration pattern (Figure S15A). On the other hand, the tumors of the patients that exhibited a strong clinical response to anti-PD1 therapy exhibited a heterogeneous immune infiltrate pattern (Figure S15B).

However, tumors of patients who responded to the anti-PD1 therapy exhibited enrichment for macrophages and -to a lesser extent- Tregs (Mann Whitney test p-values 0.018 and 0.076, respectively) in their microenvironment (Figure S15C). In line with this observation, we found that genes involved in angiogenesis and the remodeling of the extracellular matrix, two processes associated with the recruitment of immune-suppressive cells (see main paper), were up-regulated among anti-PD1 non-responder tumors (GSEA p-values of 0.02 and 0.04, respectively; data not shown), in agreement with the results reported by [33,34]. In addition, the GSEA analysis of expression values corrected by the leukocyte content of the sample also identified the up-regulation of the WNT-β-catenin signaling enriched among anti-PD1 non-responders (p-value=0.02). This observation may be supported by the role of this pathway in driving the escape of melanomas from the immune surveillance by mechanisms independent of the ectopic expression of negative immune-checkpoints.

SUPPLEMENTARY FIGURES

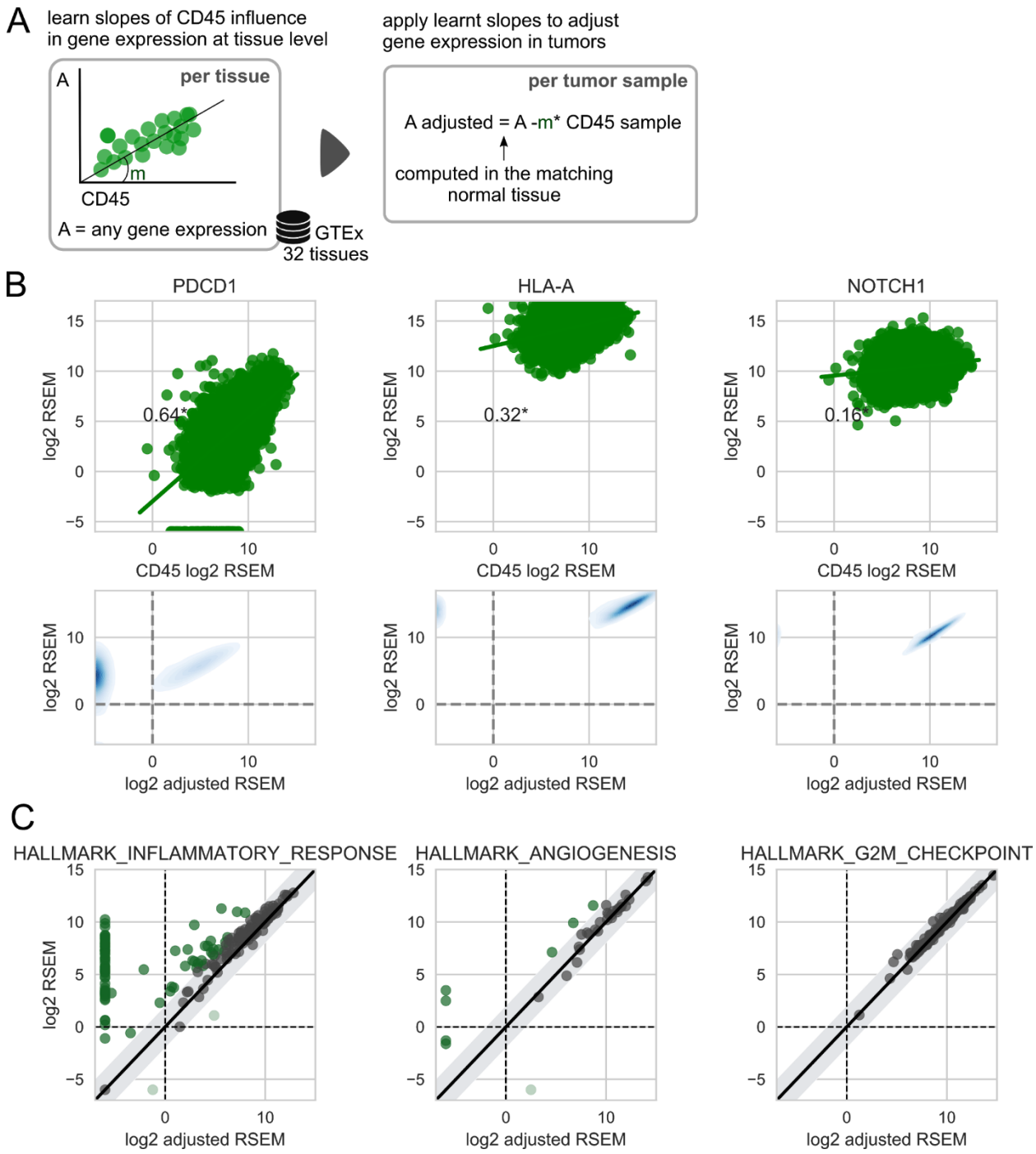


Figure S1. Adjustment of the expression in tumors to account for their immune component

(A) Graphical summary of the approach to adjust the gene expression for the immune content of the tumor sample. First, we regressed the expression of each gene as a function of the *CD45* expression across samples of the matching healthy tissue. Then, to obtain the adjusted expression value of a gene in a tumor, we interpolated the *CD45* expression value measured in the tumor sample in the regression line constructed for the gene in the matching healthy tissue (see Methods).

(B) Top panels: pan-cancer correlation of the expression of *PDCD1* (*PD-1*), *HLA-A* and *NOTCH1* genes and *CD45* expression across all healthy tissues. Bottom panels depict the pan-cancer raw (y-axis) versus adjusted (x-axis) expression values (in log2 RSEM) of these genes. The expression of *PDCD1*, mostly contributed by immune cells (high correlation with *CD45)* was reduced in many tumor samples upon adjustment for their immune content. The expression of *HLA-A*, contributed by both immune and tumor cells (intermediate correlation with *CD45*) decreased in tumor samples in a smaller proportion upon adjustment. Finally, the expression of *NOTCH1,* mostly contributed by tumor cells (low correlation with *CD45*), did not change appreciably upon adjustment.

(C) Median expression changes upon adjustment for immune infiltration across pan-cancer tumors of genes of three selected pathways. The median pan-cancer expression of each gene in a pathway is represented as a circle both before (y-axis) and after adjustment (x-axis). Genes outside the diagonal grey area exhibit significantly larger expression changes after adjustment (Mann Whitney p values<0.05 and differences larger than 2 log2 RSEM in absolute value), leading to gene expression decrease (circles at the left of the diagonal) or increase (circles at the right of the diagonal). While the expression of genes involved in the inflammatory response pathway (left panel) exhibit large collective changes (and thus the up-regulation of the pathway would be overestimated in infiltrated tumors), the expression of genes involved in angiogenesis (center panel) and G2M checkpoint (right panel), which are mostly contributed by tumor cells, change little upon adjustment.

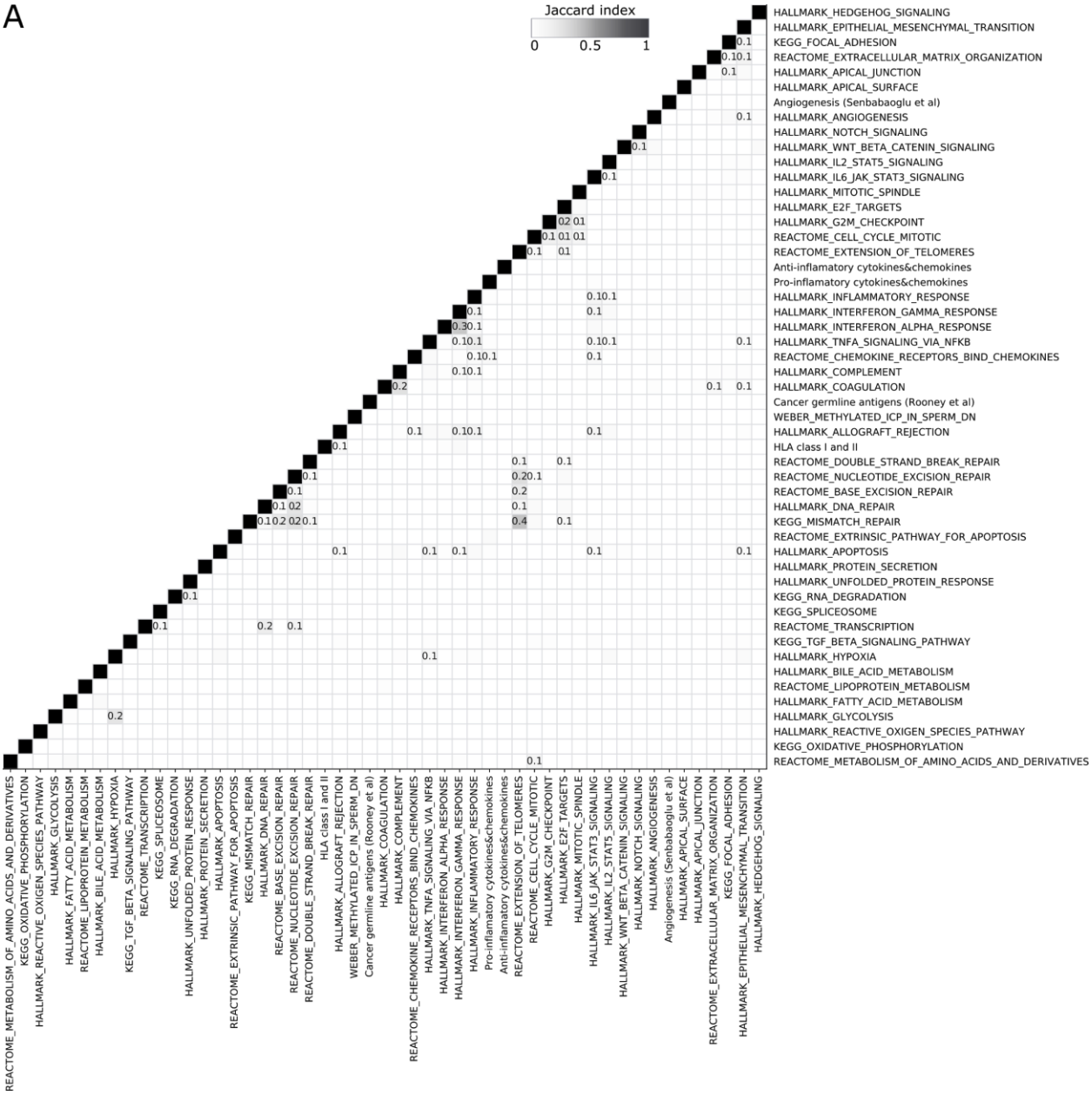
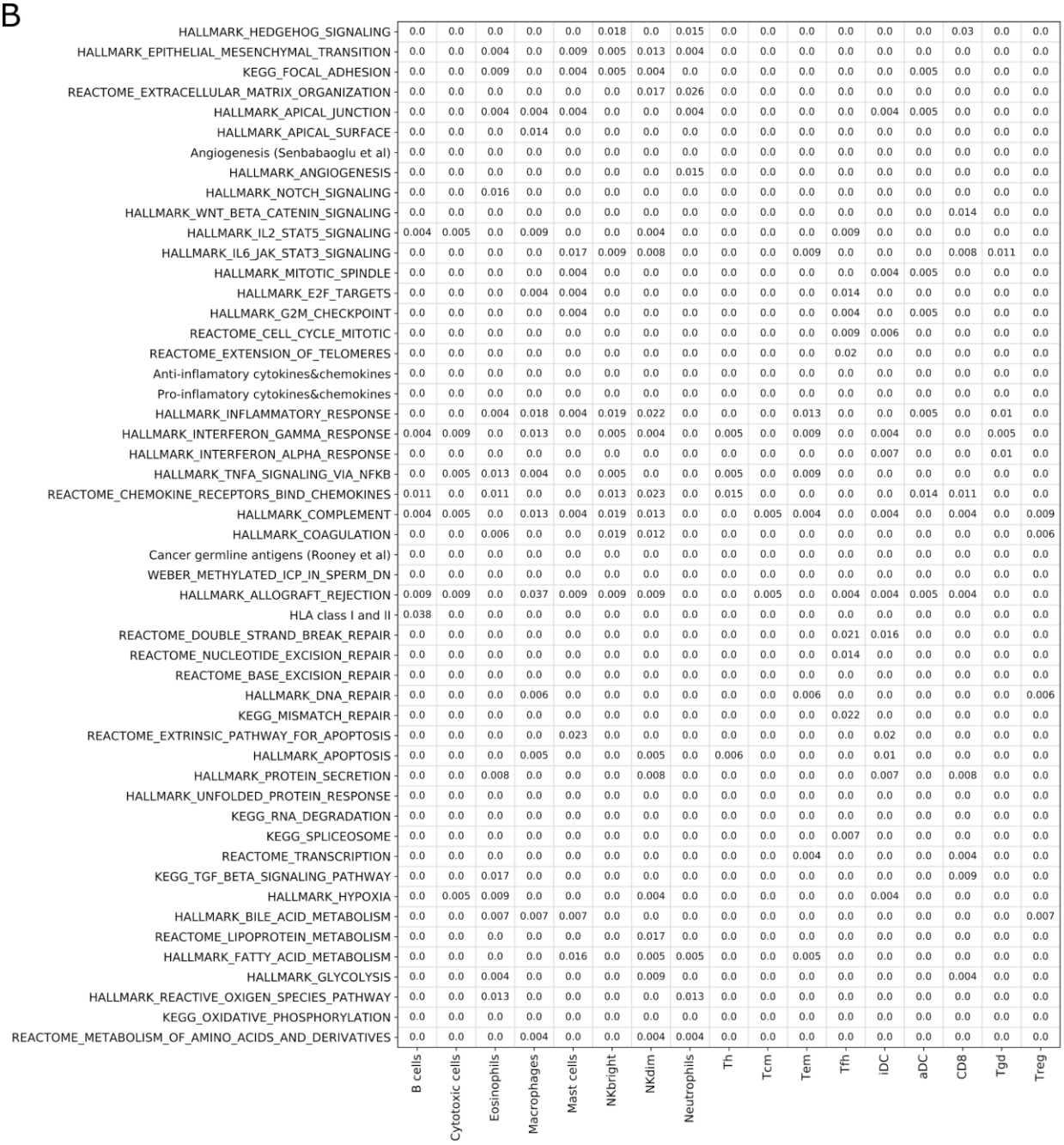


Figure S2 (1 out of 2)

B



	B cells	Cytotoxic cells	Eosinophils	Macrophages	Mast cells	NKbright	NKdim	Neutrophils	Th	Tcm	Tem	Tfh	iDC	aDC	CD8	Tgd	Treg
HALLMARK_HEDGEHOG_SIGNALING	0.0	0.0	0.0	0.0	0.0	0.018	0.0	0.015	0.0	0.0	0.0	0.0	0.0	0.0	0.03	0.0	0.0
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	0.0	0.0	0.004	0.0	0.009	0.005	0.013	0.004	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
KEGG_FOCAL_ADHESION	0.0	0.0	0.009	0.0	0.004	0.005	0.004	0.0	0.0	0.0	0.0	0.0	0.0	0.005	0.0	0.0	0.0
REACTOME_EXTRACELLULAR_MATRIX_ORGANIZATION	0.0	0.0	0.0	0.0	0.0	0.0	0.017	0.026	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
HALLMARK_APICAL_JUNCTION	0.0	0.0	0.004	0.004	0.004	0.0	0.0	0.004	0.0	0.0	0.0	0.0	0.004	0.005	0.0	0.0	0.0
HALLMARK_APICAL_SURFACE	0.0	0.0	0.0	0.014	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Angiogenesis (Senbabaoglu et al)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
HALLMARK_ANGIOGENESIS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.015	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
HALLMARK_NOTCH_SIGNALING	0.0	0.0	0.016	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
HALLMARK_WNT_BETA_CATENIN_SIGNALING	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.014	0.0	0.0
HALLMARK_IL2_STAT5_SIGNALING	0.004	0.005	0.0	0.009	0.0	0.0	0.004	0.0	0.0	0.0	0.0	0.009	0.0	0.0	0.0	0.0	0.0
HALLMARK_IL6_JAK_STAT3_SIGNALING	0.0	0.0	0.0	0.0	0.017	0.009	0.008	0.0	0.0	0.0	0.009	0.0	0.0	0.0	0.008	0.011	0.0
HALLMARK_MITOTIC_SPINDLE	0.0	0.0	0.0	0.0	0.004	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.004	0.005	0.0	0.0	0.0
HALLMARK_E2F_TARGETS	0.0	0.0	0.0	0.004	0.004	0.0	0.0	0.0	0.0	0.0	0.0	0.014	0.0	0.0	0.0	0.0	0.0
HALLMARK_G2M_CHECKPOINT	0.0	0.0	0.0	0.0	0.004	0.0	0.0	0.0	0.0	0.0	0.0	0.004	0.0	0.005	0.0	0.0	0.0
REACTOME_CELL_CYCLE_MITOTIC	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.009	0.006	0.0	0.0	0.0	0.0
REACTOME_EXTENSION_OF_TELOMERES	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.02	0.0	0.0	0.0	0.0	0.0
Anti-inflamatory cytokines&chemokines	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Pro-inflamatory cytokines&chemokines	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
HALLMARK_INFLAMMATORY_RESPONSE	0.0	0.0	0.004	0.018	0.004	0.019	0.022	0.0	0.0	0.0	0.013	0.0	0.0	0.005	0.0	0.01	0.0
HALLMARK_INTERFERON_GAMMA_RESPONSE	0.004	0.009	0.0	0.013	0.0	0.005	0.004	0.0	0.005	0.0	0.009	0.0	0.004	0.0	0.0	0.005	0.0
HALLMARK_INTERFERON_ALPHA_RESPONSE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.007	0.0	0.0	0.01	0.0
HALLMARK_TNFA_SIGNALING_VIA_NFKB	0.0	0.005	0.013	0.004	0.0	0.005	0.0	0.0	0.005	0.0	0.009	0.0	0.0	0.0	0.0	0.0	0.0
REACTOME_CHEMOKINE_RECEPTORS_BIND_CHEMOKINES	0.011	0.0	0.011	0.0	0.0	0.013	0.023	0.0	0.015	0.0	0.0	0.0	0.0	0.014	0.011	0.0	0.0
HALLMARK_COMPLEMENT	0.004	0.005	0.0	0.013	0.004	0.019	0.013	0.0	0.0	0.005	0.004	0.0	0.004	0.0	0.004	0.0	0.009
HALLMARK_COAGULATION	0.0	0.0	0.006	0.0	0.0	0.019	0.012	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.006
Cancer germline antigens (Rooney et al)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
WEBER_METHYLATED_ICP_IN_SPERM_DN	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
HALLMARK_ALLOGRAFT_REJECTION	0.009	0.009	0.0	0.037	0.009	0.009	0.009	0.0	0.0	0.005	0.0	0.004	0.004	0.005	0.004	0.0	0.0
HLA class I and II	0.038	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
REACTOME_DOUBLE_STRAND_BREAK_REPAIR	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.021	0.016	0.0	0.0	0.0	0.0
REACTOME_NUCLEOTIDE_EXCISION_REPAIR	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.014	0.0	0.0	0.0	0.0	0.0
REACTOME_BASE_EXCISION_REPAIR	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
HALLMARK_DNA_REPAIR	0.0	0.0	0.0	0.006	0.0	0.0	0.0	0.0	0.0	0.0	0.006	0.0	0.0	0.0	0.0	0.0	0.006
KEGG_MISMATCH_REPAIR	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.022	0.0	0.0	0.0	0.0	0.0
REACTOME_EXTRINSIC_PATHWAY_FOR_APOPTOSIS	0.0	0.0	0.0	0.0	0.023	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.02	0.0	0.0	0.0	0.0
HALLMARK_APOPTOSIS	0.0	0.0	0.0	0.005	0.0	0.0	0.005	0.0	0.006	0.0	0.0	0.0	0.01	0.0	0.0	0.0	0.0
HALLMARK_PROTEIN_SECRETION	0.0	0.0	0.008	0.0	0.0	0.0	0.008	0.0	0.0	0.0	0.0	0.0	0.007	0.0	0.008	0.0	0.0
HALLMARK_UNFOLDED_PROTEIN_RESPONSE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
KEGG_RNA_DEGRADATION	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
KEGG_SPLICEOSOME	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.007	0.0	0.0	0.0	0.0	0.0
REACTOME_TRANSCRIPTION	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.004	0.0	0.0	0.0	0.004	0.0	0.0
KEGG_TGF_BETA_SIGNALING_PATHWAY	0.0	0.0	0.017	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.009	0.0	0.0
HALLMARK_HYPOXIA	0.0	0.005	0.009	0.0	0.0	0.0	0.004	0.0	0.0	0.0	0.0	0.0	0.004	0.0	0.0	0.0	0.0
HALLMARK_BILE_ACID_METABOLISM	0.0	0.0	0.007	0.007	0.007	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.007
REACTOME_LIPOPROTEIN_METABOLISM	0.0	0.0	0.0	0.0	0.0	0.0	0.017	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
HALLMARK_FATTY_ACID_METABOLISM	0.0	0.0	0.0	0.0	0.016	0.0	0.005	0.005	0.0	0.0	0.005	0.0	0.0	0.0	0.0	0.0	0.0
HALLMARK_GLYCOLYSIS	0.0	0.0	0.004	0.0	0.0	0.0	0.009	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.004	0.0	0.0
HALLMARK_REACTIVE_OXIGEN_SPECIES_PATHWAY	0.0	0.0	0.013	0.0	0.0	0.0	0.0	0.013	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
KEGG_OXIDATIVE_PHOSPHORYLATION	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
REACTOME_METABOLISM_OF_AMINO_ACIDS_AND_DERIVATIVES	0.0	0.0	0.0	0.004	0.0	0.0	0.004	0.004	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure S2. Overlap between the gene sets representing selected pathways and immune populations.

(A) The heatmap summarizes the degree of overlap between the gene sets representing pathways (n=51) selected to compute the enrichment of their up-regulation across immune-

phenotypes. We used the Jaccard index to compute the overlap between the genes of each pair of pathways. Note that most pairs of pathways share zero genes (Jaccard index= 0, empty cells). The highest overlap corresponded to the extension of telomeres (Reactome database) and mismatch repair (KEGG database) pathways (Jaccard index = 0.4).

(B) Heatmap summarizing the degree of overlap between the gene sets of the selected 51 pathways for the enrichment analysis and the gene signatures of the sixteen immune cells populations (and the cytotoxic cells). Again, the overlaps are negligible.

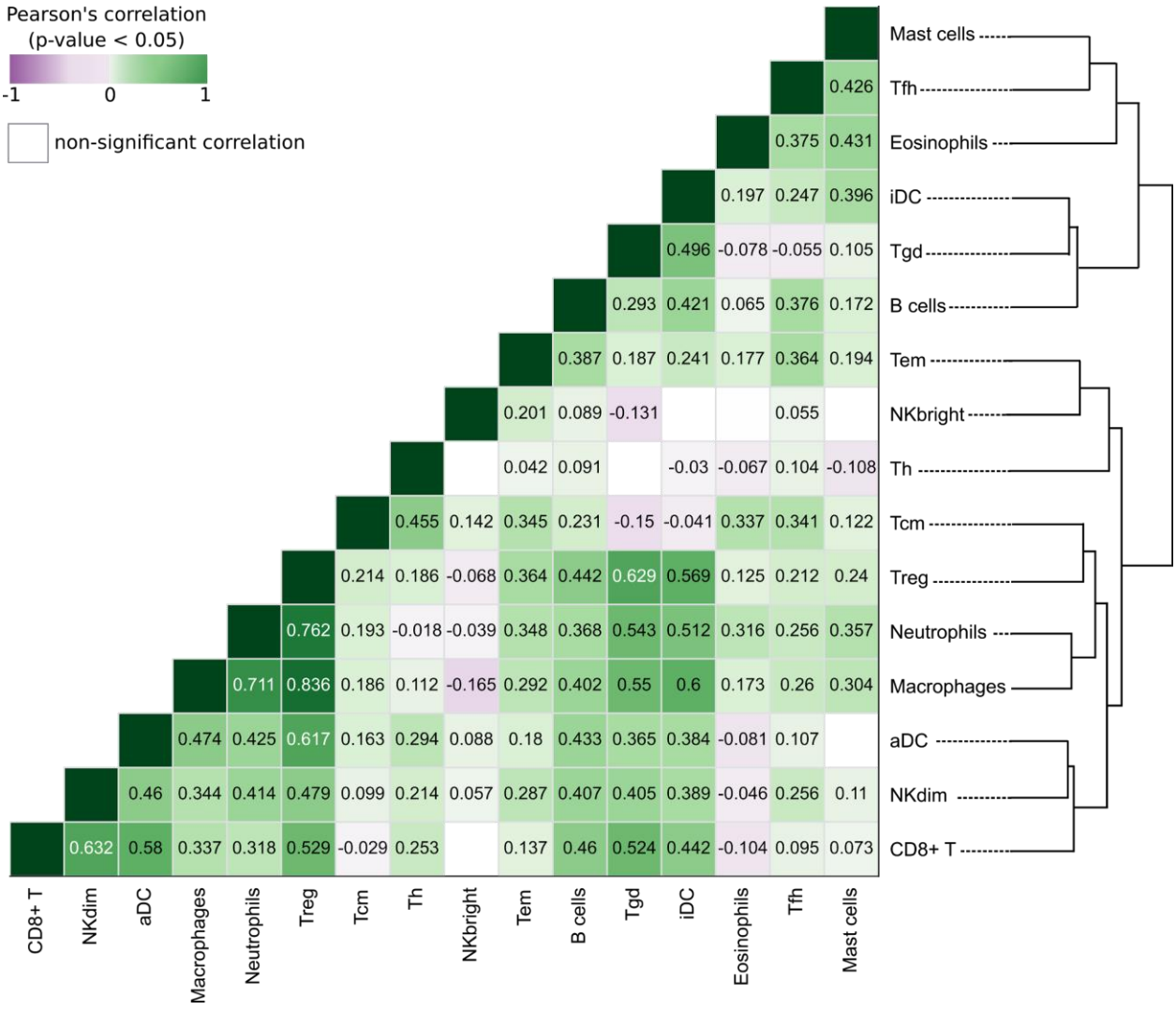


Figure S3. Correlation of GSVA scores of 16 immune cell populations across the pan-cancer cohort

In the heatmap, all significant correlations of the relative abundance of pairs of cell populations (p-value < 0.05) are colored following their Pearson's correlation coefficient. The dendrogram in the right groups the cell populations following their degree of co-infiltration of the tumors across the pan-cancer cohort (see Methods).

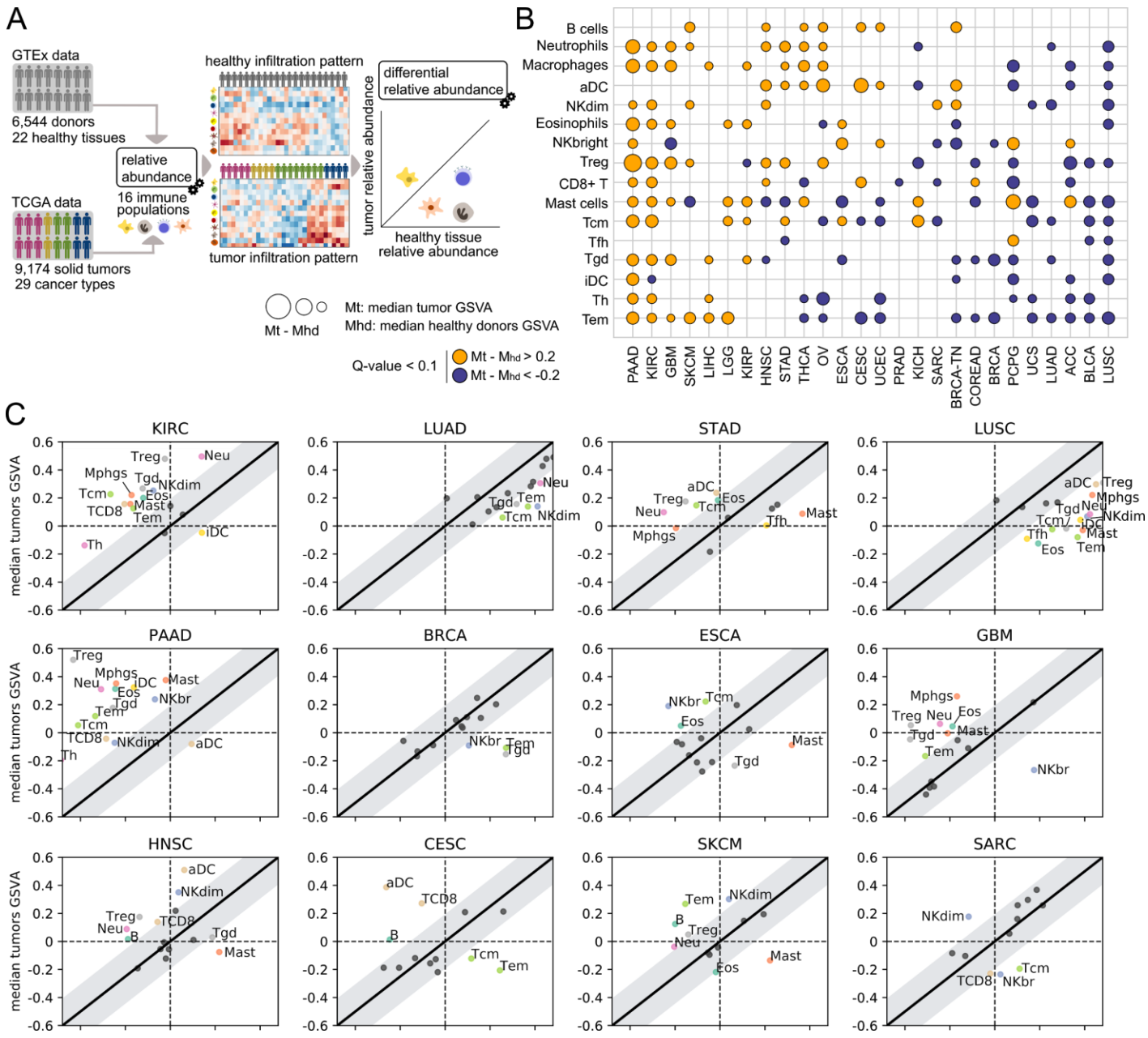


Figure S4 (1 out of 2)

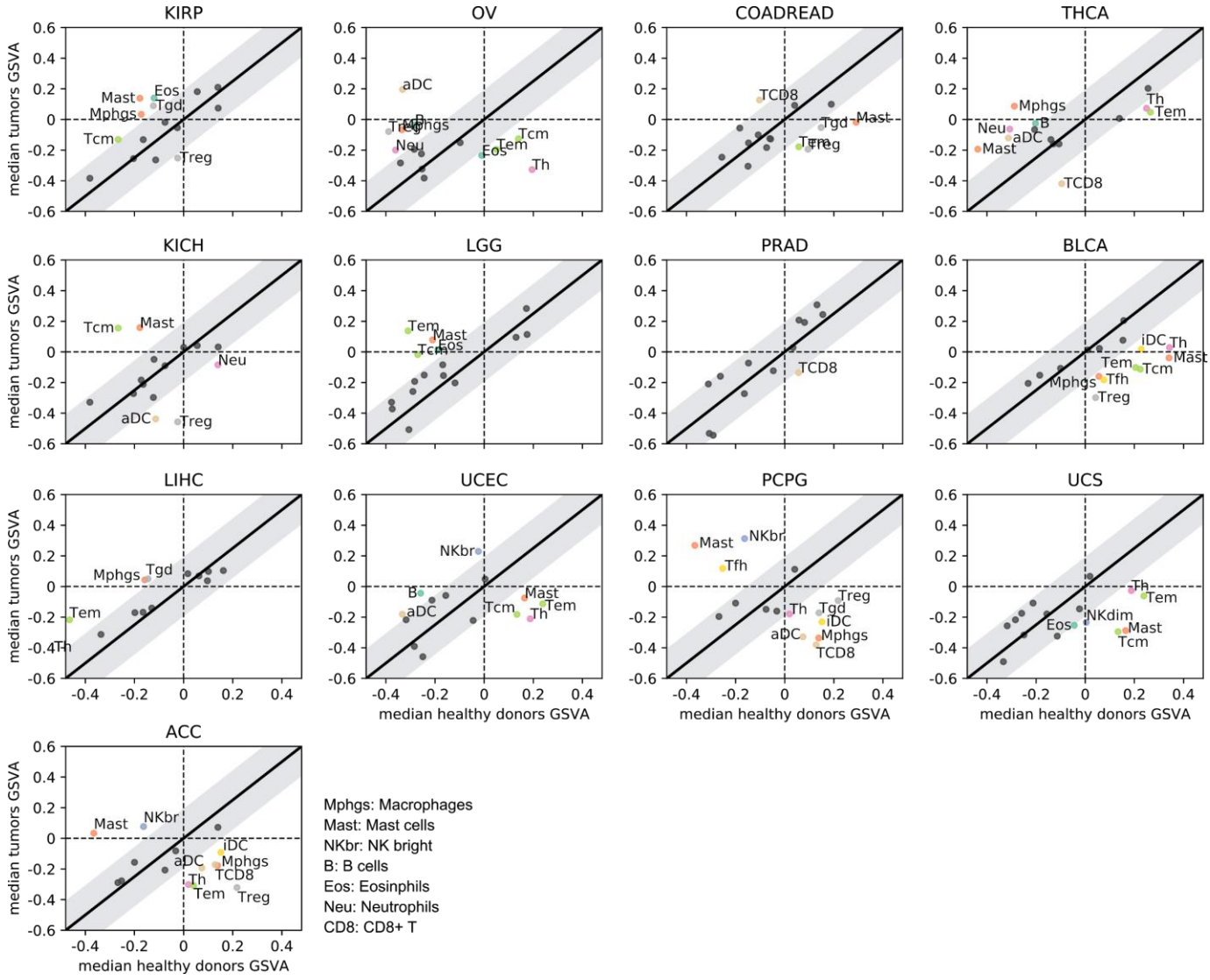


Figure S4. Comparison of the relative abundance of immune cell populations in tumors and matching healthy tissues.

(A) Schematic representation of the computation of the relative abundance of immune cell populations in tumors and matching healthy tissues (see Methods).

(B) Summary of the comparison of GSVA scores of the 16 immune cell populations in each cancer cohort to that of its matching healthy tissue. The size of each circle represents the difference (D) between the GSVA enrichment scores of a cell population in tumors (Mt) and

healthy donors (Mhd). Only significant comparisons (Mann-Whitney Q-value<0.1 and absolute D greater than 0.2) are shown. Circles are colored orange if Mt>Mh and violet if Mh>Mt. Cancer types are sorted in ascending order of accumulated D across the 16 immune cell populations.
(C) Comparison of the relative abundance (median GSVA scores) of the sixteen immune cell populations in tumors (y-axis) versus samples of matching healthy tissue (x-axis) in each cohort. (Triple-negative breast tumors are included in the BRCA cohort for this analysis). Each dot represents the median GSVA score of an immune cell population in a tumor-normal tissue pair. Dots outside the grey-colored area (labeled with the name of the immune cell they represent) constitute cases that exhibit major differences of relative abundance between tumors and matching normal tissue (absolute D greater than 0.2). Each panel corresponds to a cohort and are sorted in decreasing order of overall leukocyte infiltration (measured as the expression of *CD45* in tumors), from KIRC to ACC (see Fig. 1C of the main paper).

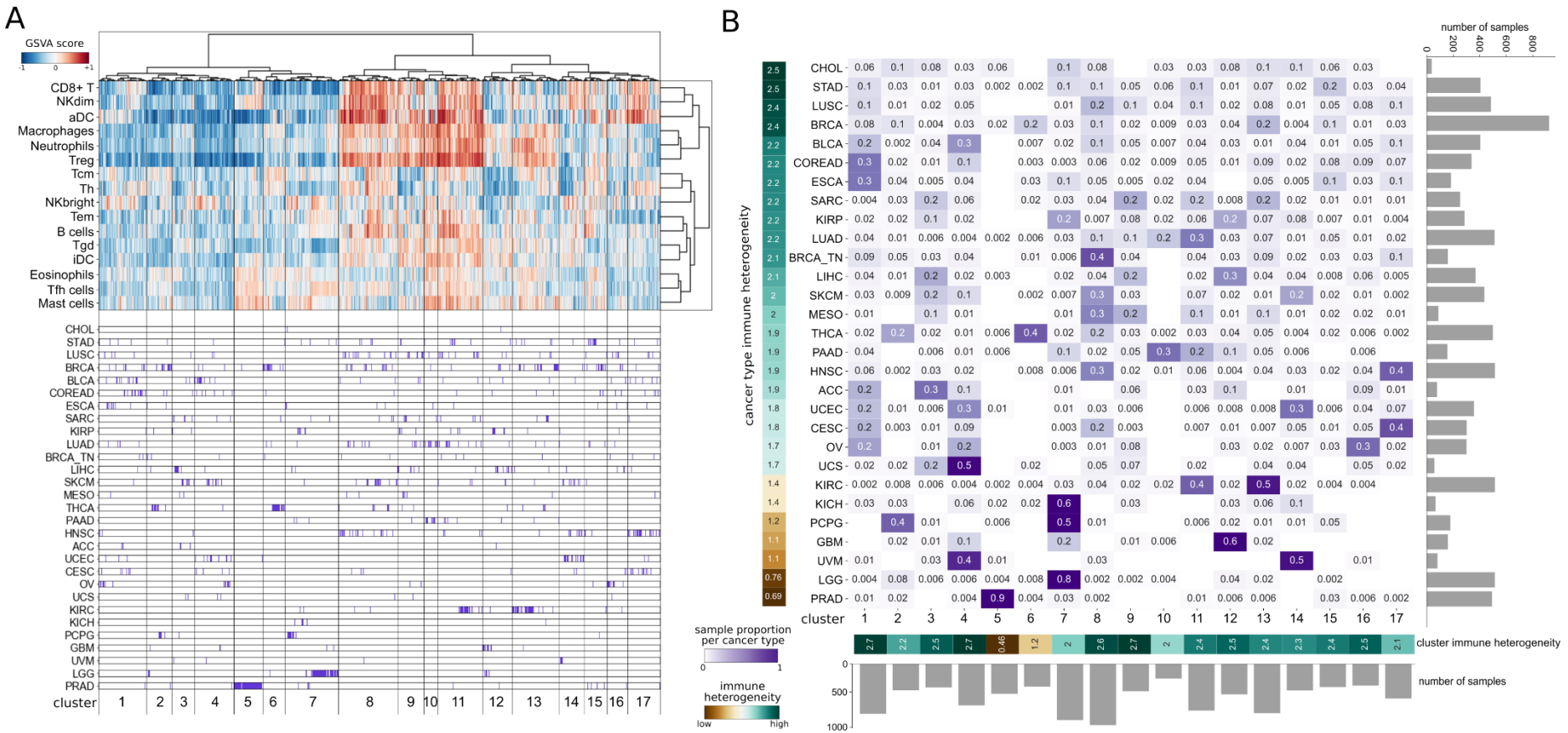


Figure S5 Pan-cancer comparison of the immune infiltration pattern

(A) Pan-cancer clusters of immune infiltration pattern of 9,174 tumors of 29 solid cancers obtained via hierarchical clustering. We defined 17 clusters (see Methods) that represent distinct arrangements of immune infiltration of solid tumors. The panel below the

heatmap with the immune infiltration pattern of the tumors in the pan-cancer cohort depicts how the tumors of each cohort distribute across the clusters.

(B) Heatmap of the distribution of the tumors of each malignancy across the pan-cancer clusters of immune infiltration pattern. Numbers in the matrix represent the proportion of tumors of each cohort grouped in each cluster. One-column heatmap at the left: metric based on the entropy score (see Methods) representing the heterogeneity in the distribution of tumors of each cohort across clusters. One-column heatmap at the bottom: metric based on the entropy score (see Methods) representing the heterogeneity of each cluster in terms of the contribution of each cohort.

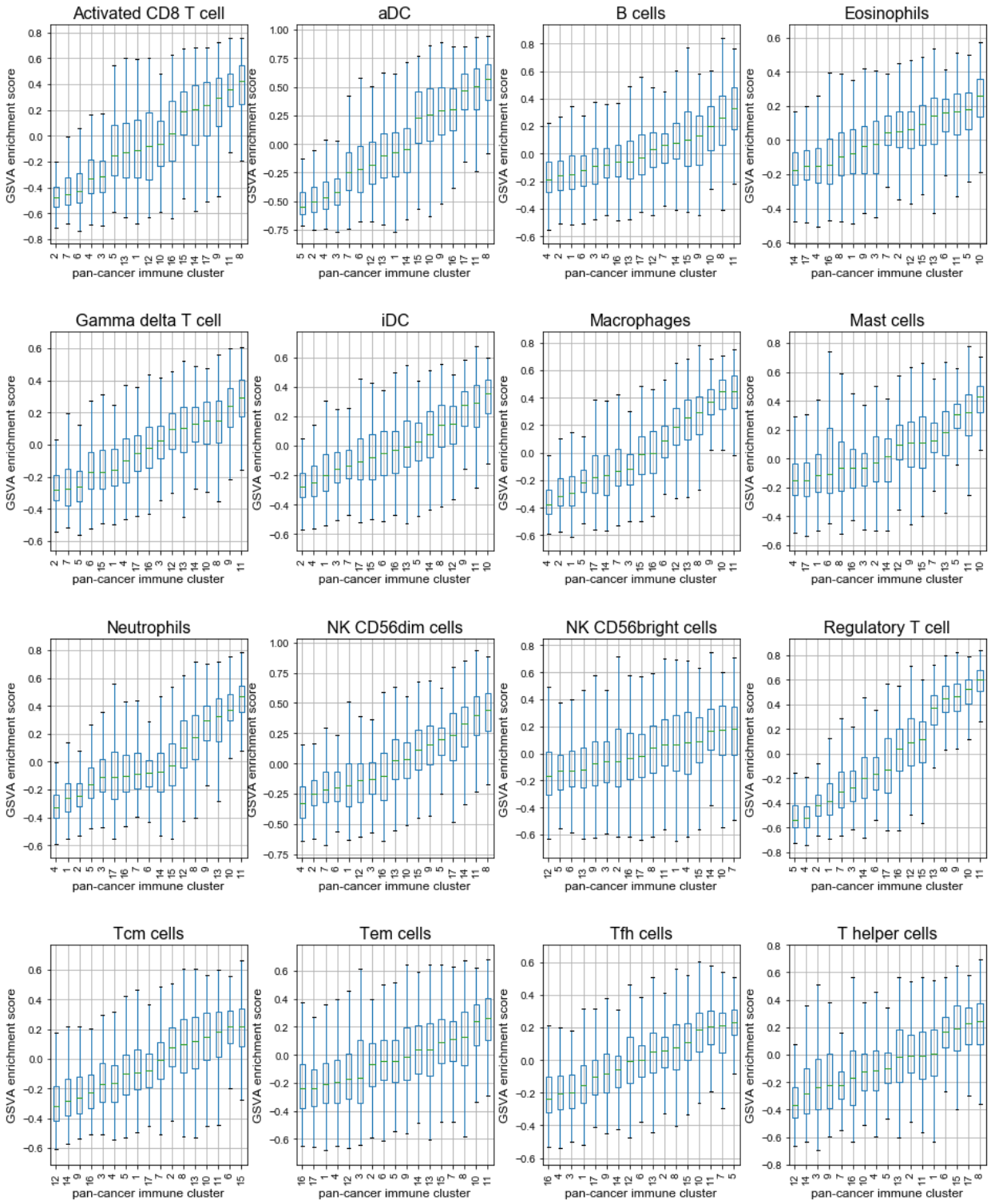

Activated CD8 T cell
aDC
B cells
Eosinophils
Gamma delta T cell
iDC
Macrophages
Mast cells
Neutrophils
NK CD56dim cells
NK CD56bright cells
Regulatory T cell
Tcm cells
Tem cells
Tfh cells
T helper cells
GSVA enrichment score
pan-cancer immune cluster

Figure S6. Enrichment of immune populations across pan-cancer immune clusters

Distribution of the relative abundance (GSVA enrichment scores) of each immune cell population across the tumors in the 17 pan-cancer clusters of immune infiltration. Boxplots are ordered in ascending order of the median GSVA score observed in each cluster (see x- axis). As displayed in Figure S5, some of the clusters group tumors of several malignancies that exhibit an overall depletion of immune cells (e.g. clusters 3 and 4) or an overall enrichment of immune cells (e.g. clusters 8 and 11). Some other clusters are more cancer-specific, and show particular immune infiltrate patterns such as: cluster 5, formed almost only by prostate adenocarcinomas, exhibits a very low relative abundance of regulatory T cells and activated dendritic cells but a high relative abundance of follicular helper T cells, eosinophils and mast cells; cluster 6, which groups most of the thyroid carcinomas and some breast cancers, and cluster 7, which groups most of the low-grade gliomas and some paragangliomas, both cluster 6 and 7 exhibiting low relative abundance of *bona fide* effector immune cells (activated CD8 T cells, NK-dim and activated dendritic cells), the latter cluster showing also a low abundance of T gamma delta cells.

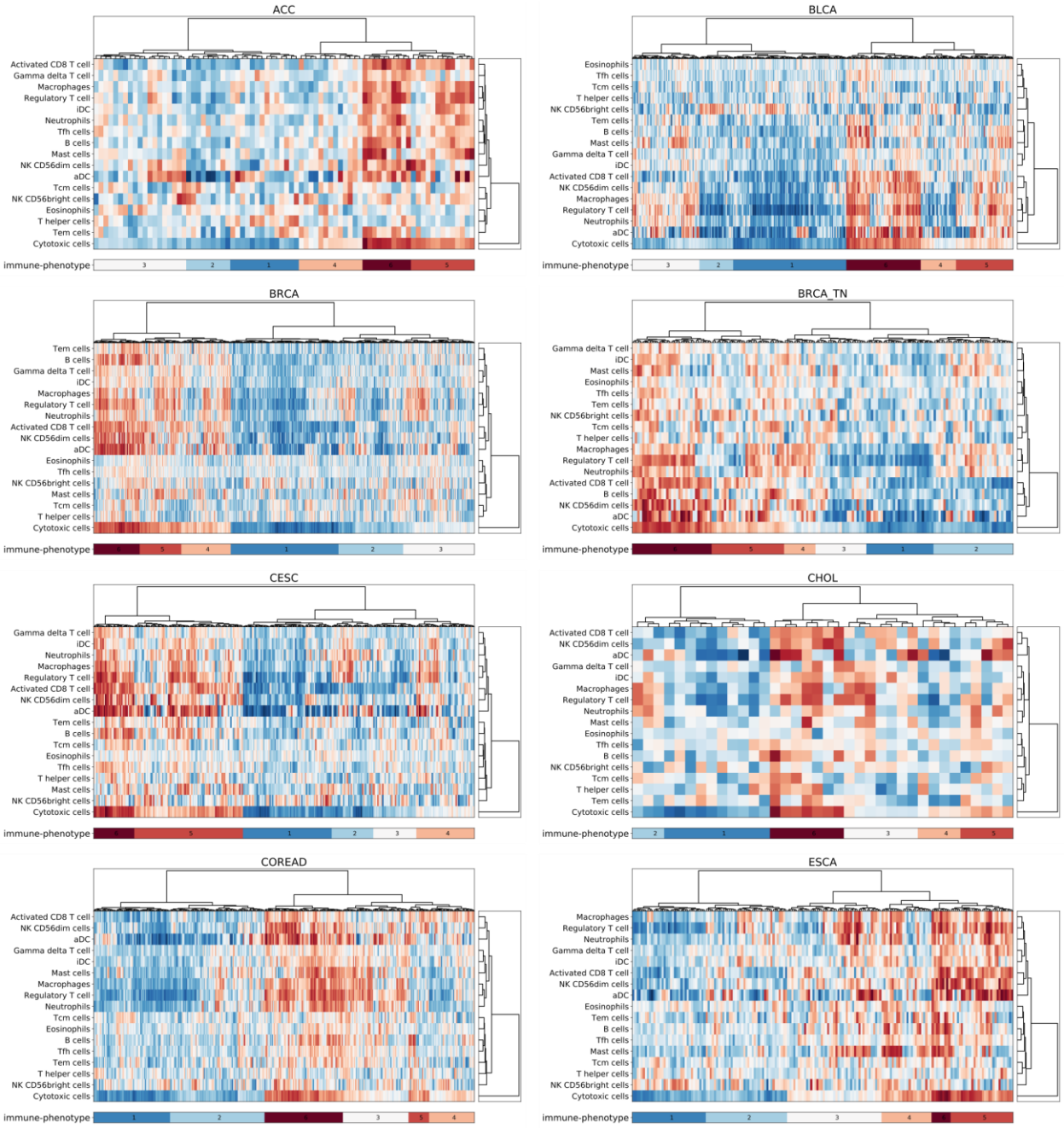


Figure S7 (1 out of 4)

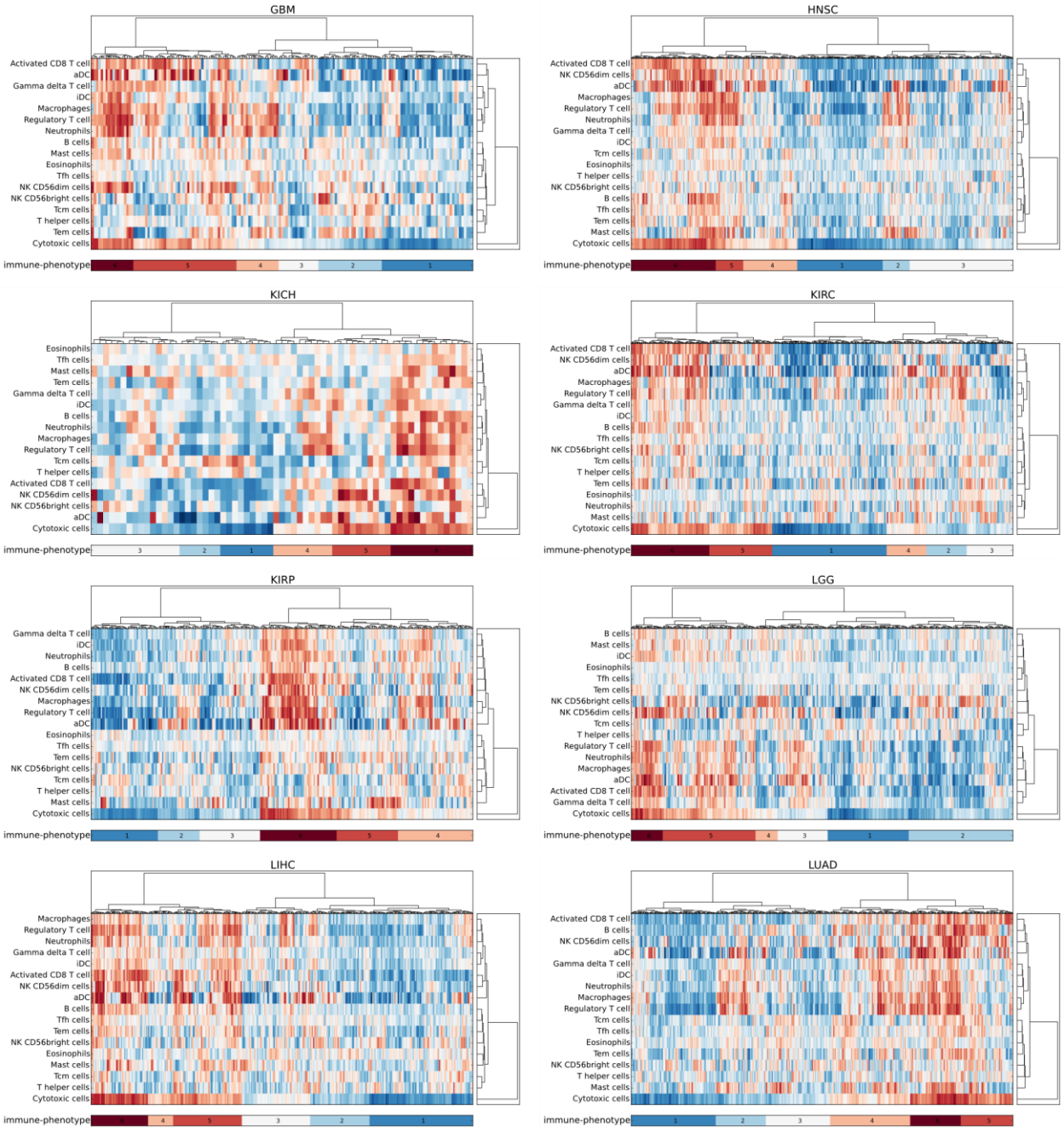


Figure S7 (2 out of 4)

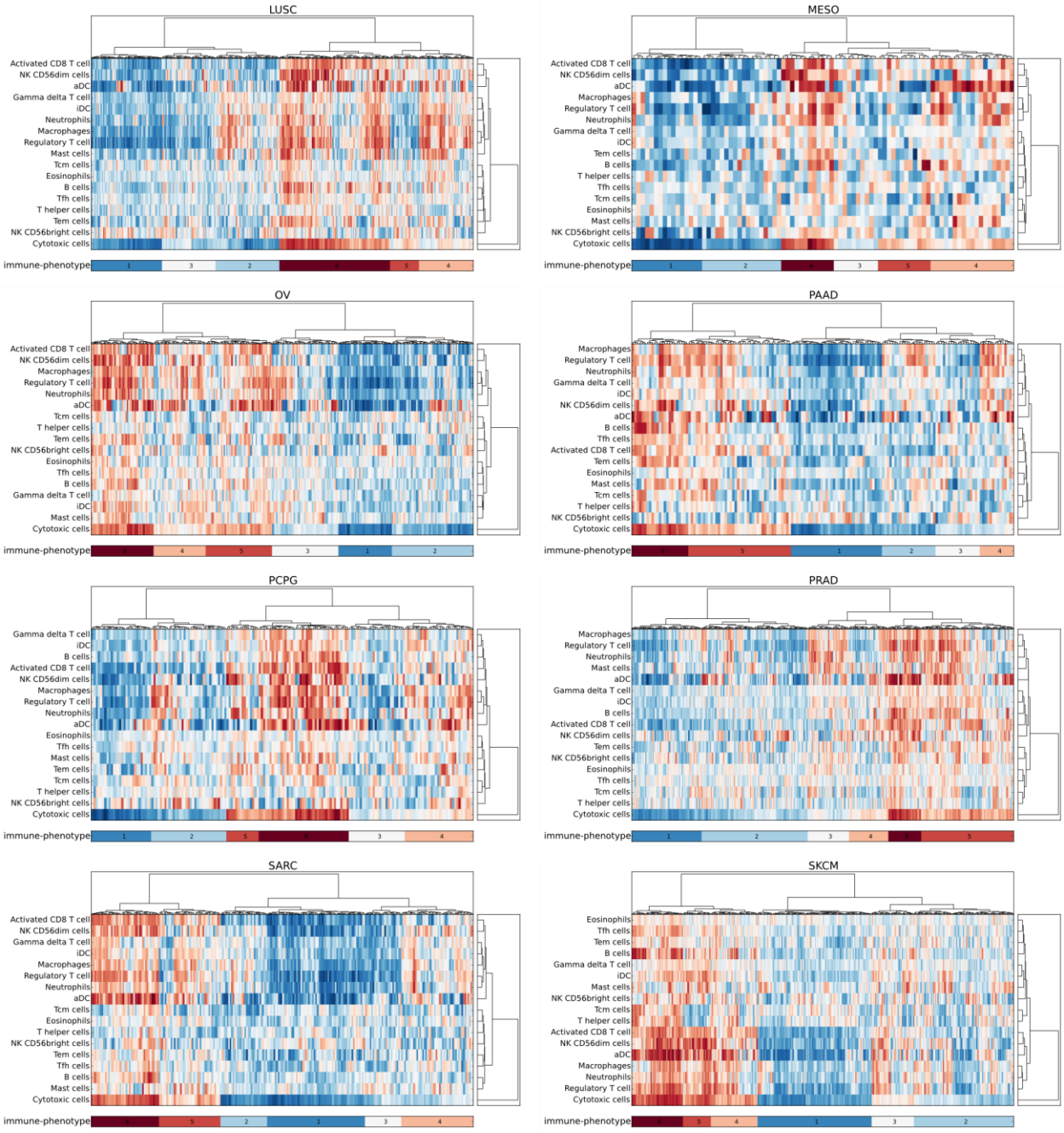


Figure S7 (3 out of 4)

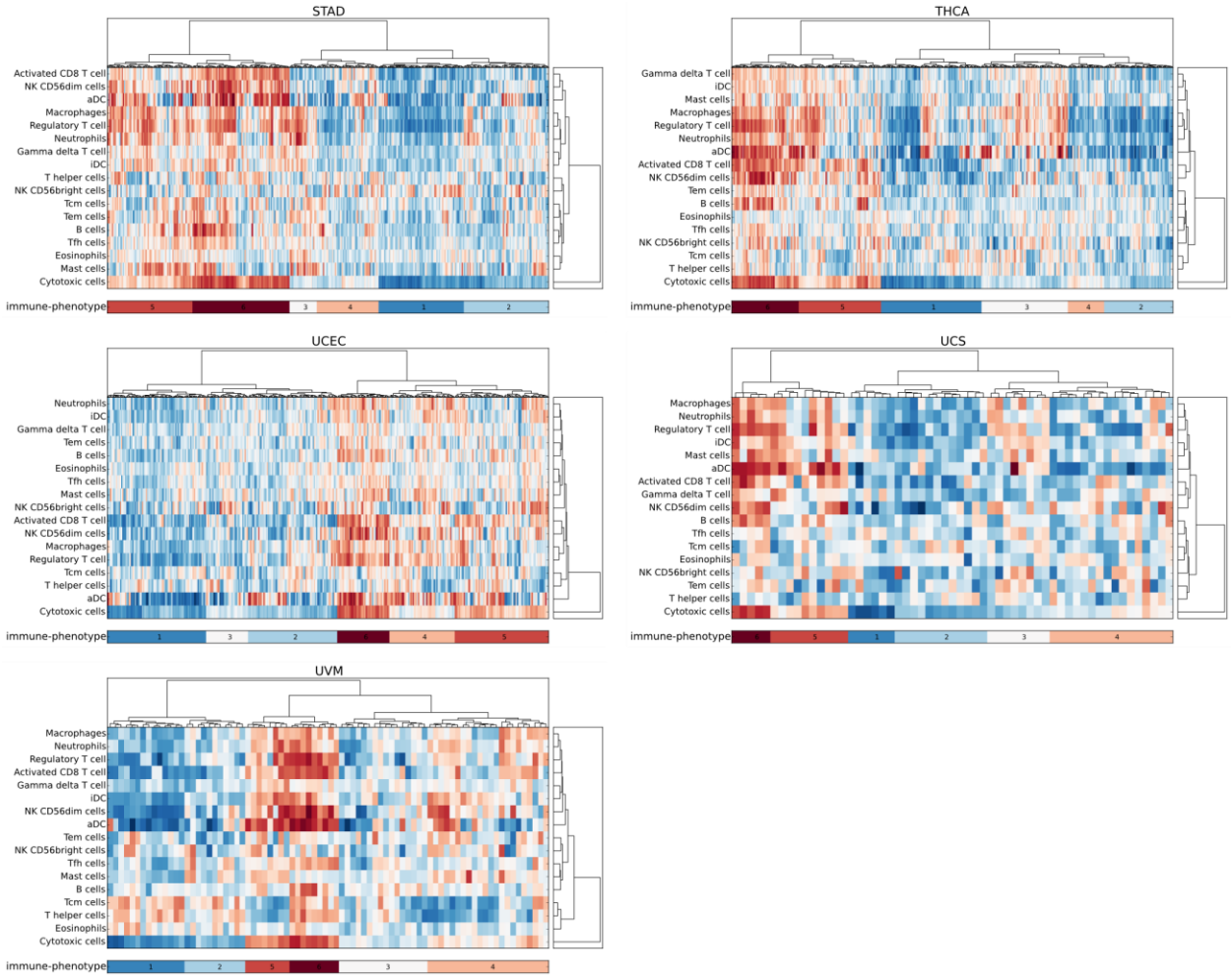


Figure S7. Immune-phenotypes of all cancer cohorts

Six immune-phenotypes were constructed in each cohort clustering the immune infiltration profile of the tumors of each cancer cohort with an overweight of the relative abundance of cytotoxic cells (see Methods). The bars below the heatmaps identify the six immune-phenotypes with colors ranging from dark blue for the immune-phenotype with the lowest relative abundance of cytotoxic cells to dark red for the immune-phenotype with the highest relative abundance of cytotoxic cells.

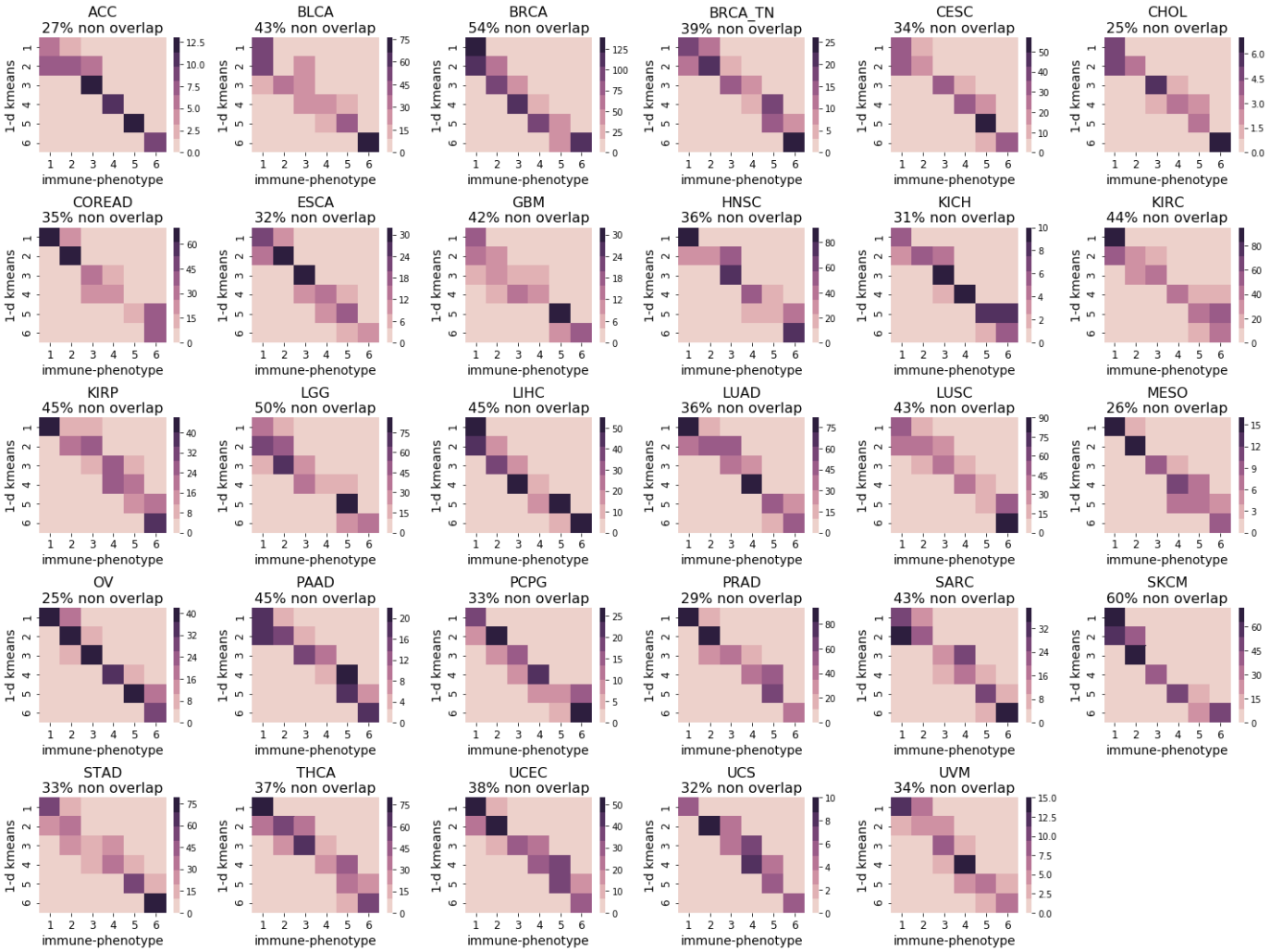


Figure S8. Agreement between immune-phenotypes and cytotoxicity-based clusters

To clarify the contribution of non-cytotoxic cell populations to the immune-phenotypes we grouped the tumors of each cohort into 6 clusters based only on their cytotoxic cell infiltration (using a one-dimensional k-means), called cyt-clusters for short. We then carried out a 6x6 way comparison of immune-phenotypes and cyt-clusters, computing the overlap of each combination in the 36-cells matrix. The heatmaps display the results of these analysis for each cohort (following the color scale by each heatmap). In both immune-phenotypes and cyt-clusters, 1 represents the group of tumors with lowest cytotoxic content and 6 the group with highest cytotoxic content. While the highest degree of overlap in all cohorts happens along the diagonal (determined by the overweight of cytotoxic populations in the construction of immune-

phenotypes), the identity between immune-phenotypes and cyt-clusters is far from perfect. The percentage of tumors in each cohort that do not overlap along the diagonal is shown in the title of each heatmap: it ranges from one quarter to half of the tumors depending on the cohort.

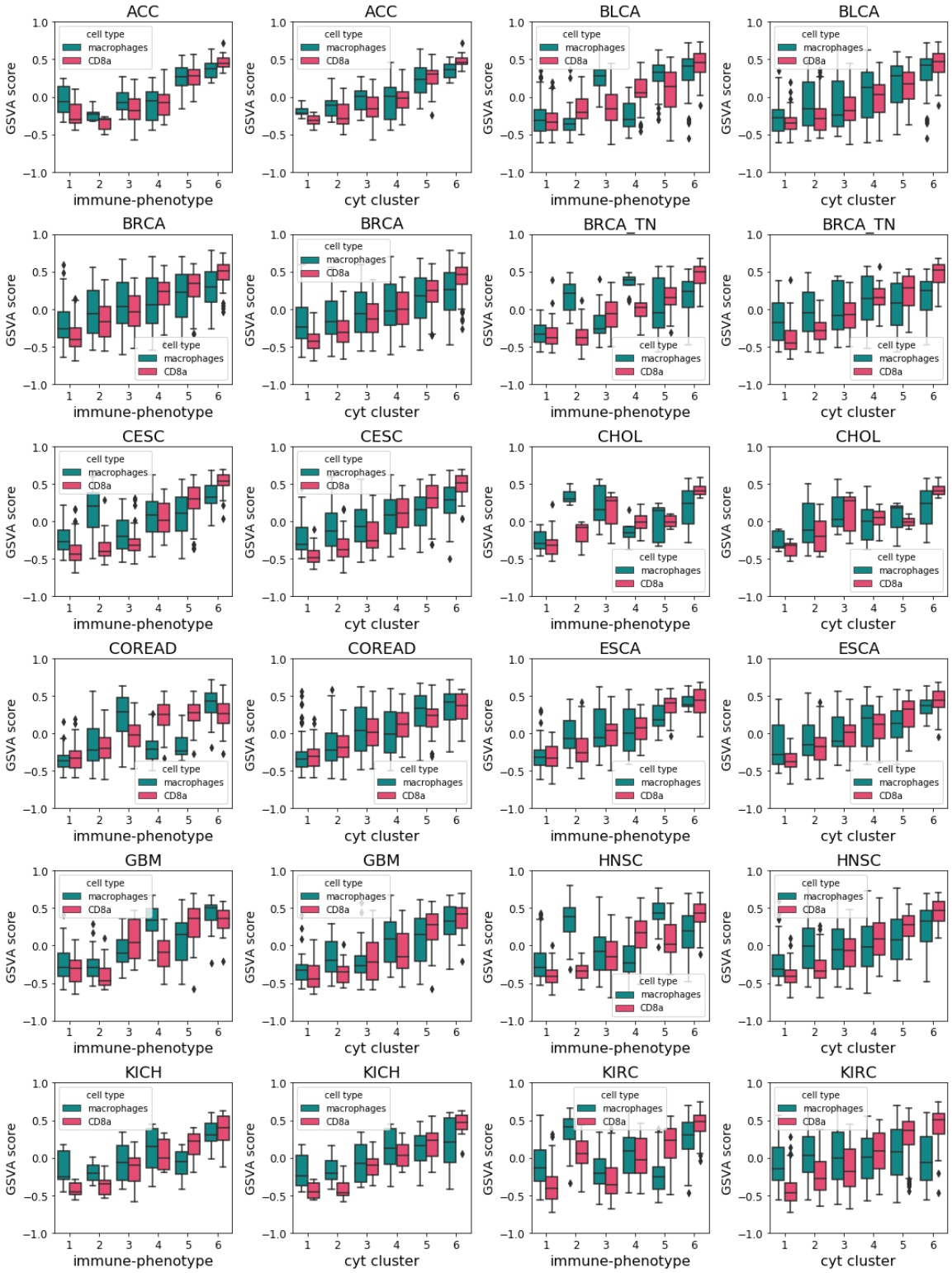


Figure S9 (1 out of 3)

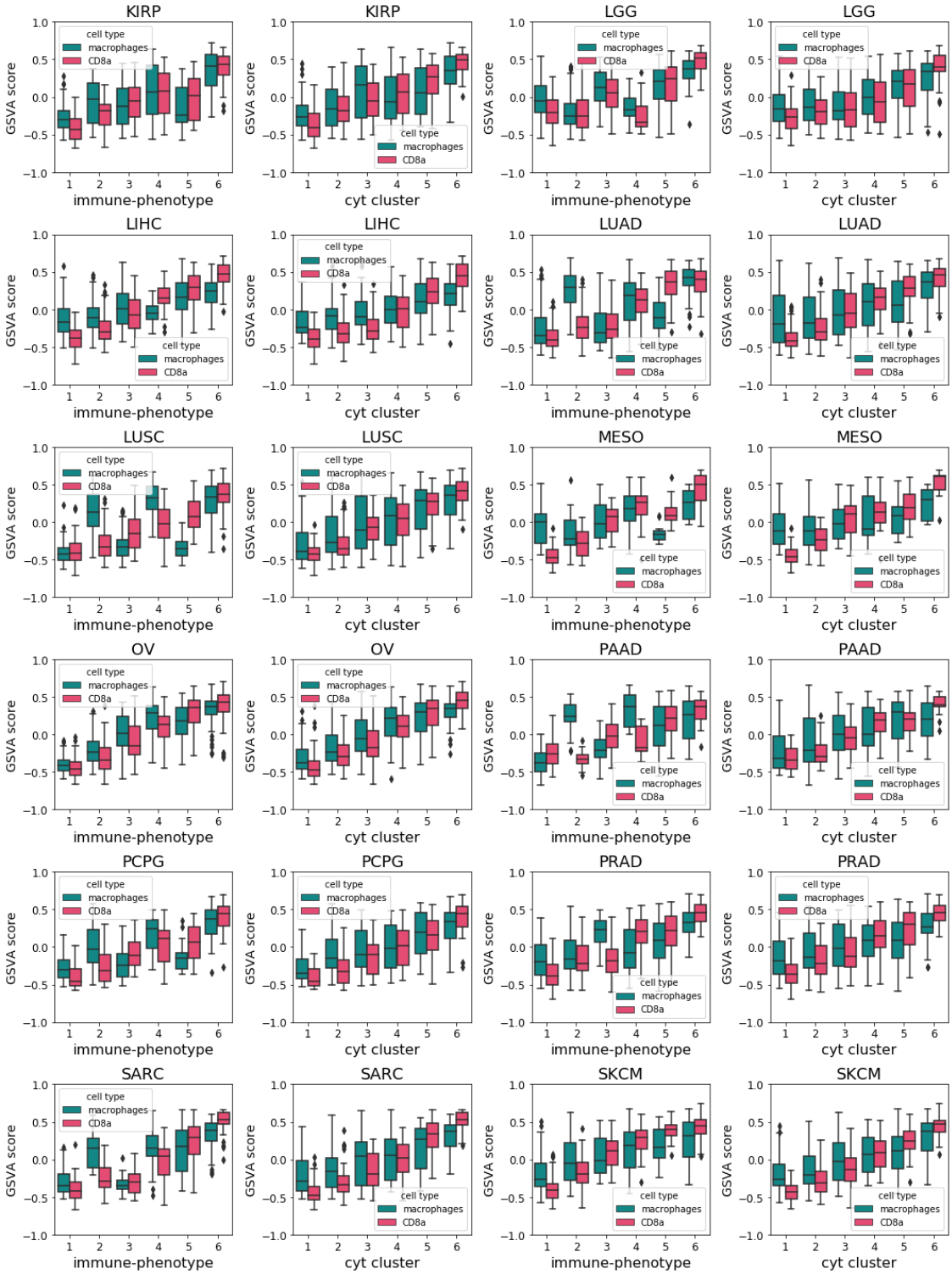


Figure S9 (2 out of 3)

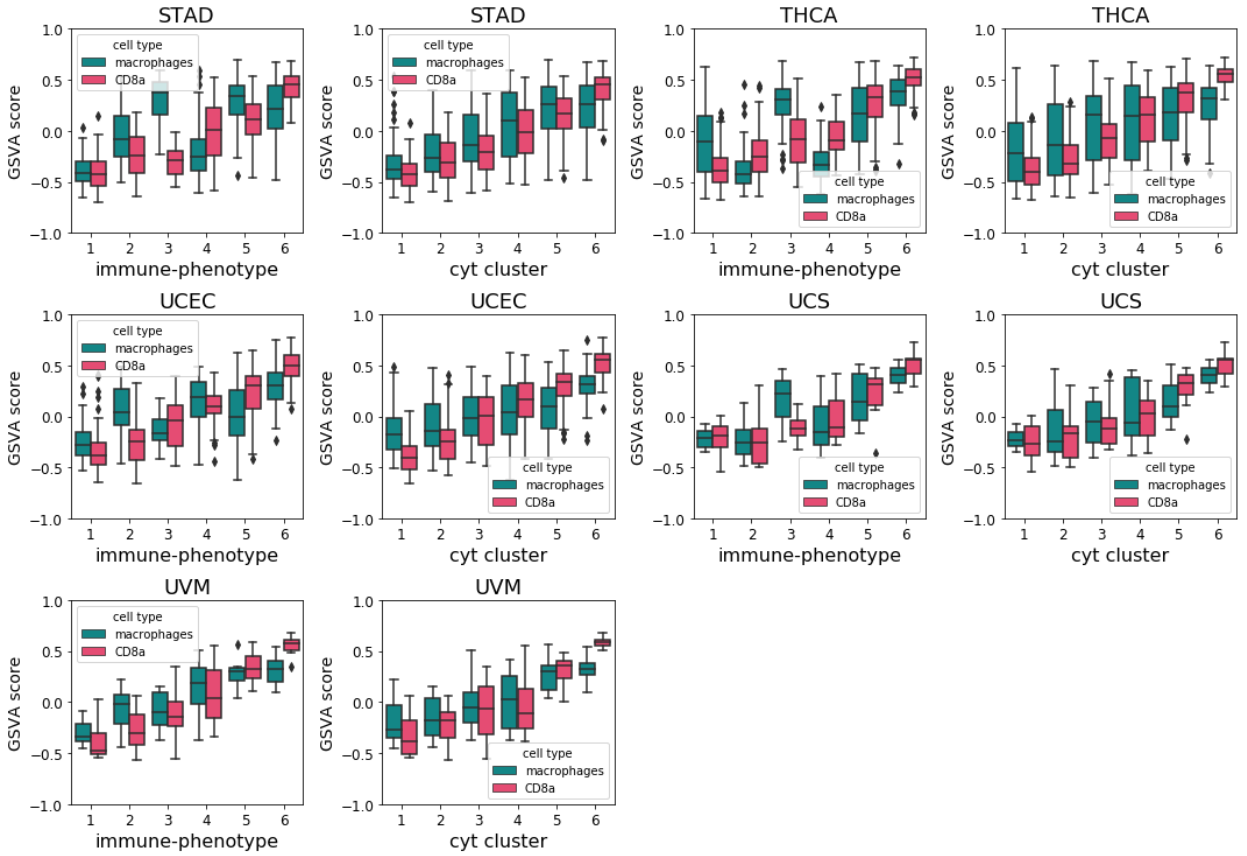


Figure S9. Relative abundance (GSVA score) distribution of selected immune cell populations across cyt-clusters and immune-phenotypes

To continue the study of the contribution of non-cytotoxic cell populations to the construction of immune-phenotypes, we compared the distribution of relative abundance of activated CD8+ T cells (red) and macrophages (green) across immune-phenotypes and cyt-clusters constructed in each cohort. Each pair of boxplots above (identified by the same cancer type acronym) presents the results of these comparison. In both immune-phenotypes and cyt-clusters the relative abundance of CD8+ T cells increases steadily, as expected from the definition of both types of clusters. However, in several cohorts we observed that while the relative abundance of macrophages increases steadily across cyt-clusters, this trend is not maintained for macrophages. In several instances, the latter exhibit higher relative abundance in intermediate immune-phenotypes than in highly cytotoxic ones. We observed a similar pattern when we compared other *bona fide* effector (e.g. NK cells) and suppressive (e.g. regulatory T cells and neutrophils) cell populations (data not shown).

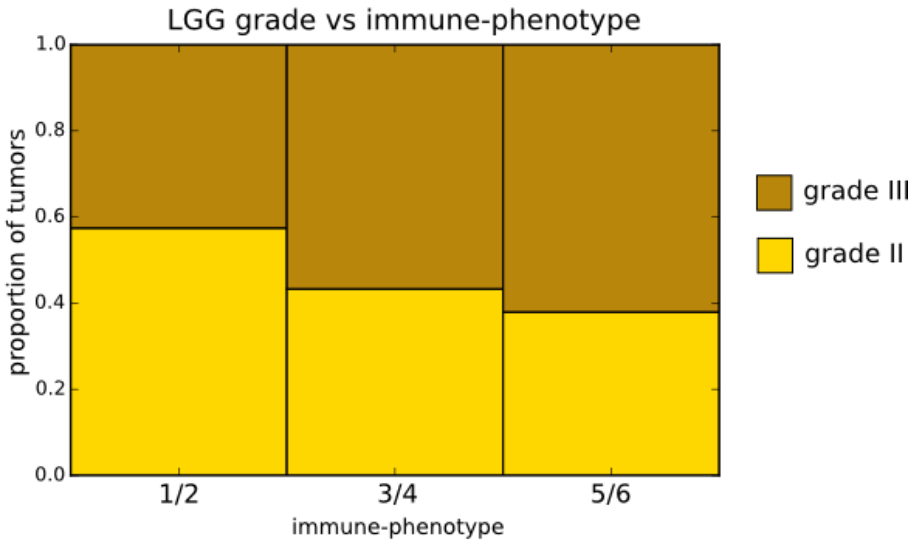


Figure S10. Grade of LGG tumors across immune-phenotypes

Proportion of patients with LGG grade II (yellow) or grade III (brown) across their immune-phenotypes. Immune-phenotypes of higher cytotoxicity are associated with tumors of higher grade (linear regression p-value= 6.4×10^{-5}).

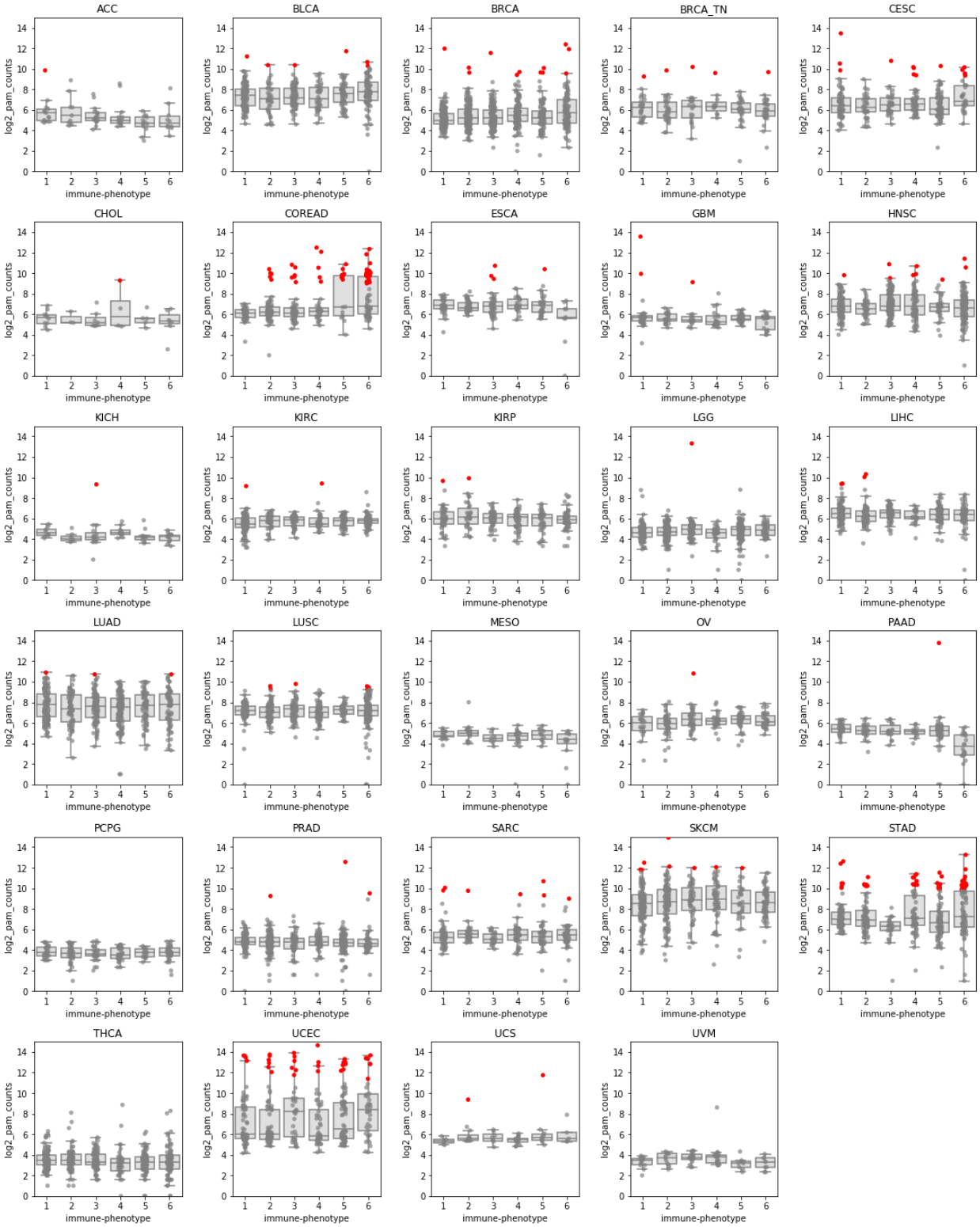

ACC
BLCA
BRCA
BRCA_TN
CESC
CHOL
COREAD
ESCA
GBM
HNSC
KICH
KIRC
KIRP
LGG
LIHC
LUAD
LUSC
MESO
OV
PAAD
PCPG
PRAD
SARC
SKCM
STAD
THCA
UCEC
UCS
UVM
log2_pam_counts
immune-phenotype

Figure S11. Distribution of tumor mutation burden across immune-phenotypes

Each panel presents the distribution of tumor mutation burden (protein-affecting mutations) across the six immune-phenotypes identified in each cancer type. Hypermutated tumors (defined based on their relative mutation burden, see Methods) are highlighted in red. Hypermutated tumors were over-represented in highly cytotoxic immune-phenotypes in the breast, colorectal, stomach and uterine corpus endometrial carcinoma cohorts (see Figure S12).

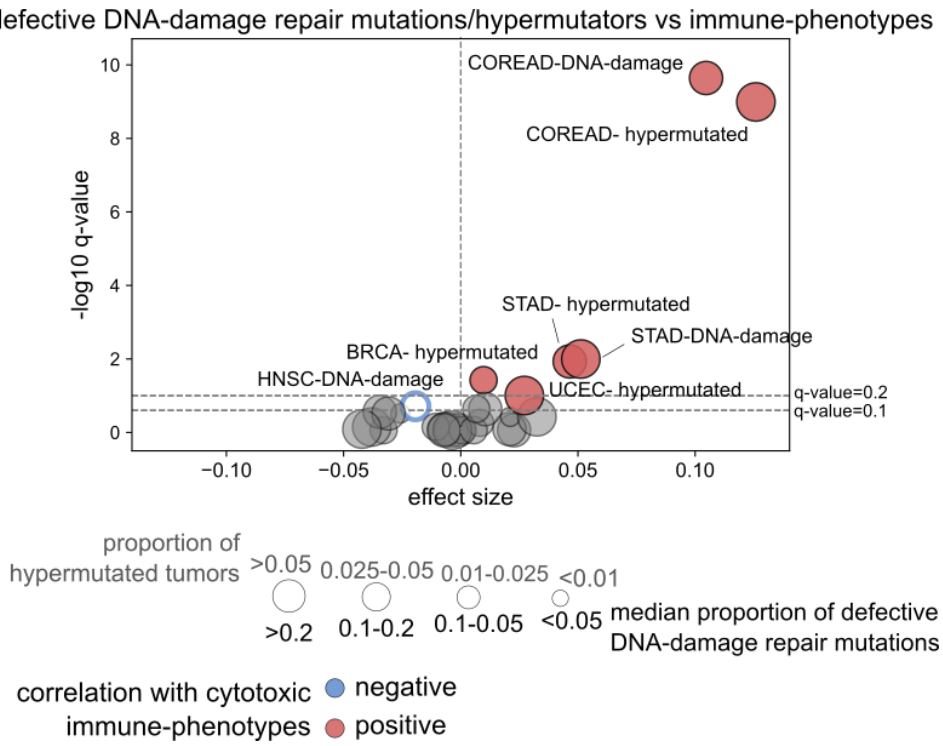


Figure S12. Significant associations between mutations caused by defective DNA repair mechanisms and hypermutated tumors with immune-phenotypes.

We considered all mutations fitting signatures 6, 10, 15, 18, 20, 21 and 26 (among the 30 available in the current version of http://cancer.sanger.ac.uk/cosmic/signatures) as likely caused by altered POLE, BRCA-1/2 and mismatch repair deficiency. We then assessed the differences in the distribution of these mutations across immune-phenotypes constructed for each cohort using a linear regression analysis. Only tumors with at least 50 mutations were included in this analysis. We followed the same approach to evaluate the association between hypermutated tumors (defined based on their relative mutation burden, see Methods) and immune-phenotypes.


PDL-2 (PDCD1LG2)
PDL-1(CD274)
LIHC
PRAD
ACC
KICH
BRCA
LGG
UCEC
PCPG
SARC
UCS
BRCA-TN
CESC
OV
PAAD
ESCA
HNSC
KIRP
GBM
KIRC
SKCM
THCA
STAD
LUAD
COREAD
LUSC
BLCA
1.0 0.8 0.6 0.4 0.2 0.0
0.0 0.2 0.4 0.6 0.8 1.0
proportion of tumors
proportion of tumors
linear regression
*** p-value < 10e-6
** 10e-6 < p-value < 10e-3
* 10e-3 < p-value < 0.05
expressed (log2 adjusted RSEM > 6)
expressed (log2 adjusted RSEM =< 6)
cancer type
5/6
3/4
1/2
immune-phenotype

Figure S13. Expression of immune-checkpoint genes across immune-phenotypes

Bar plots representing the fraction of tumors expressing the immune-checkpoint genes PDL-1 or PDL-2 across immune-phenotypes (grouped as 1/2, 3/4 and 5/6 from bottom to top in each graph). A gene was considered to be expressed in a tumor with log2 adjusted RSEM RNA-seq value greater than 6. For each cancer type, a p-value measuring the association between the checkpoint expression and the immune-phenotype (from 1 to 6) was computed using a linear regression.

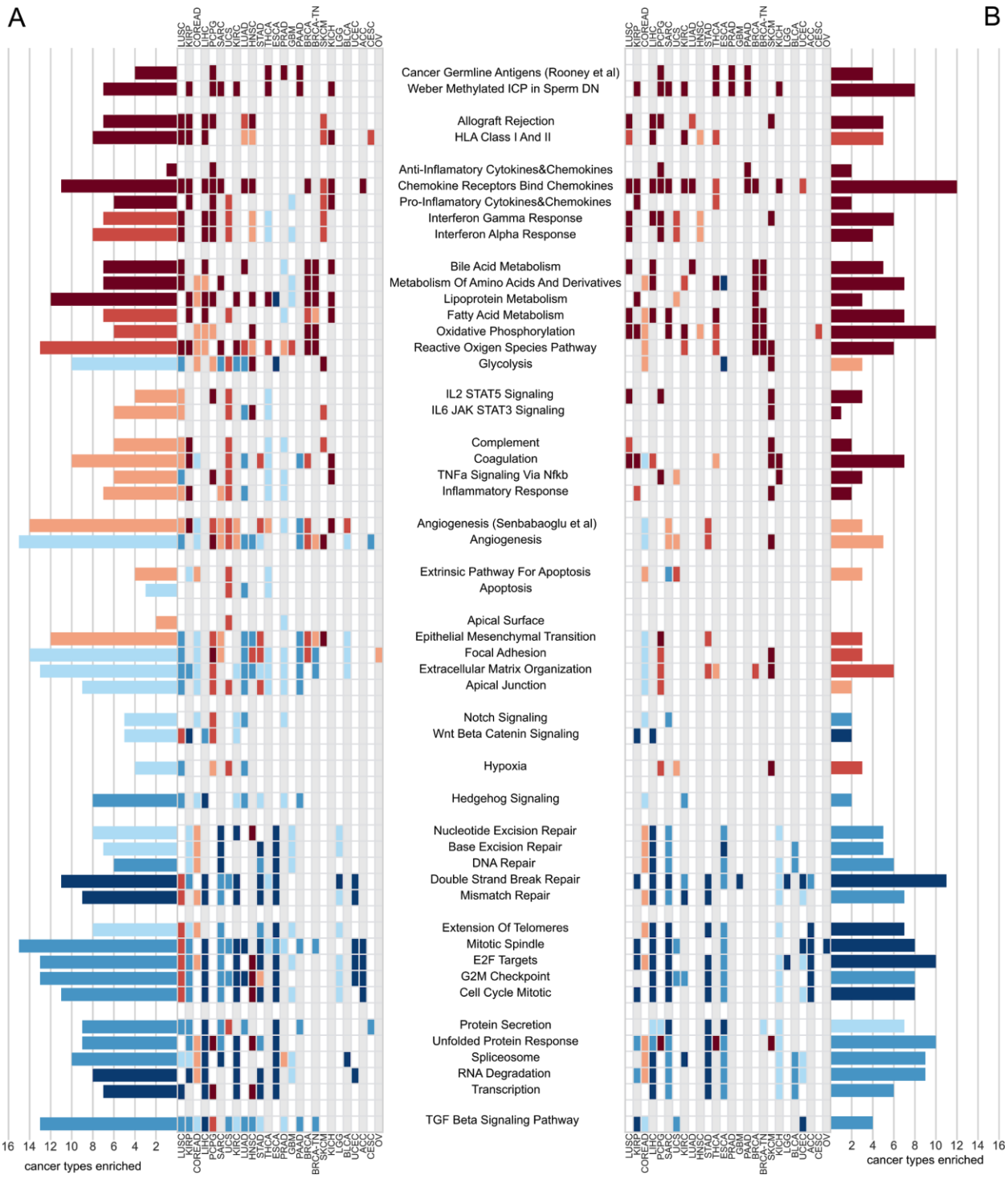

A
B
LUSC KIRP COREAD LIHC PCPG SARC UCS KIRC HNSC LUAD STAD THCA ESCA PRAD GBM PAAD BRCA BRCA-TN SKCM KICH LGG BLCA UCEC ACC CESC OV
Cancer Germline Antigens (Rooney et al)
Weber Methylated ICP in Sperm DN
Allograft Rejection
HLA Class I And II
Anti-Inflamatory Cytokines&Chemokines
Chemokine Receptors Bind Chemokines
Pro-Inflamatory Cytokines&Chemokines
Interferon Gamma Response
Interferon Alpha Response
Bile Acid Metabolism
Metabolism Of Amino Acids And Derivatives
Lipoprotein Metabolism
Fatty Acid Metabolism
Oxidative Phosphorylation
Reactive Oxigen Species Pathway
Glycolysis
IL2 STAT5 Signaling
IL6 JAK STAT3 Signaling
Complement
Coagulation
TNFa Signaling Via Nfkb
Inflammatory Response
Angiogenesis (Senbabaoglu et al)
Angiogenesis
Extrinsic Pathway For Apoptosis
Apoptosis
Apical Surface
Epithelial Mesenchymal Transition
Focal Adhesion
Extracellular Matrix Organization
Apical Junction
Notch Signaling
Wnt Beta Catenin Signaling
Hypoxia
Hedgehog Signaling
Nucleotide Excision Repair
Base Excision Repair
DNA Repair
Double Strand Break Repair
Mismatch Repair
Extension Of Telomeres
Mitotic Spindle
E2F Targets
G2M Checkpoint
Cell Cycle Mitotic
Protein Secretion
Unfolded Protein Response
Spliceosome
RNA Degradation
Transcription
TGF Beta Signaling Pathway
16 14 12 10 8 6 4 2
cancer types enriched
2 4 6 8 10 12 14 16
cancer types enriched

Figure S14. Comparison of the enrichment of up-regulated cell pathways across immune-phenotypes and cyt-clusters.

To understand how much the definition of immune-phenotypes (and their differences with the cyt-clusters shown in Figures S8 and S9) influences the transcriptional analysis presented in the paper, we computed the pathways that appear over-activated across the cyt-clusters. The two heatmaps above present the results of this comparison. Heatmap (A), also shown in Figure 3 of the main paper presents the pathways whose up-regulation is significantly enriched in each immune-phenotypes across cohorts. Heatmap (B) presents the same for cyt-clusters. The bar plots at the left side of (A) and right side of (B) present the number of cohorts where over-representation of the up-regulated pathway is significant (Q-value < 0.25, see Methods). The bars are colored according to the median immune-phenotype (left) or cyt-cluster (right; from one to six) number showing the enrichment in each cohort. Overall, we observed that fewer up-regulated pathways appeared enriched for cyt-clusters than for immune-phenotypes across cohorts, particularly in immune-phenotypes with intermediate cytotoxicity.

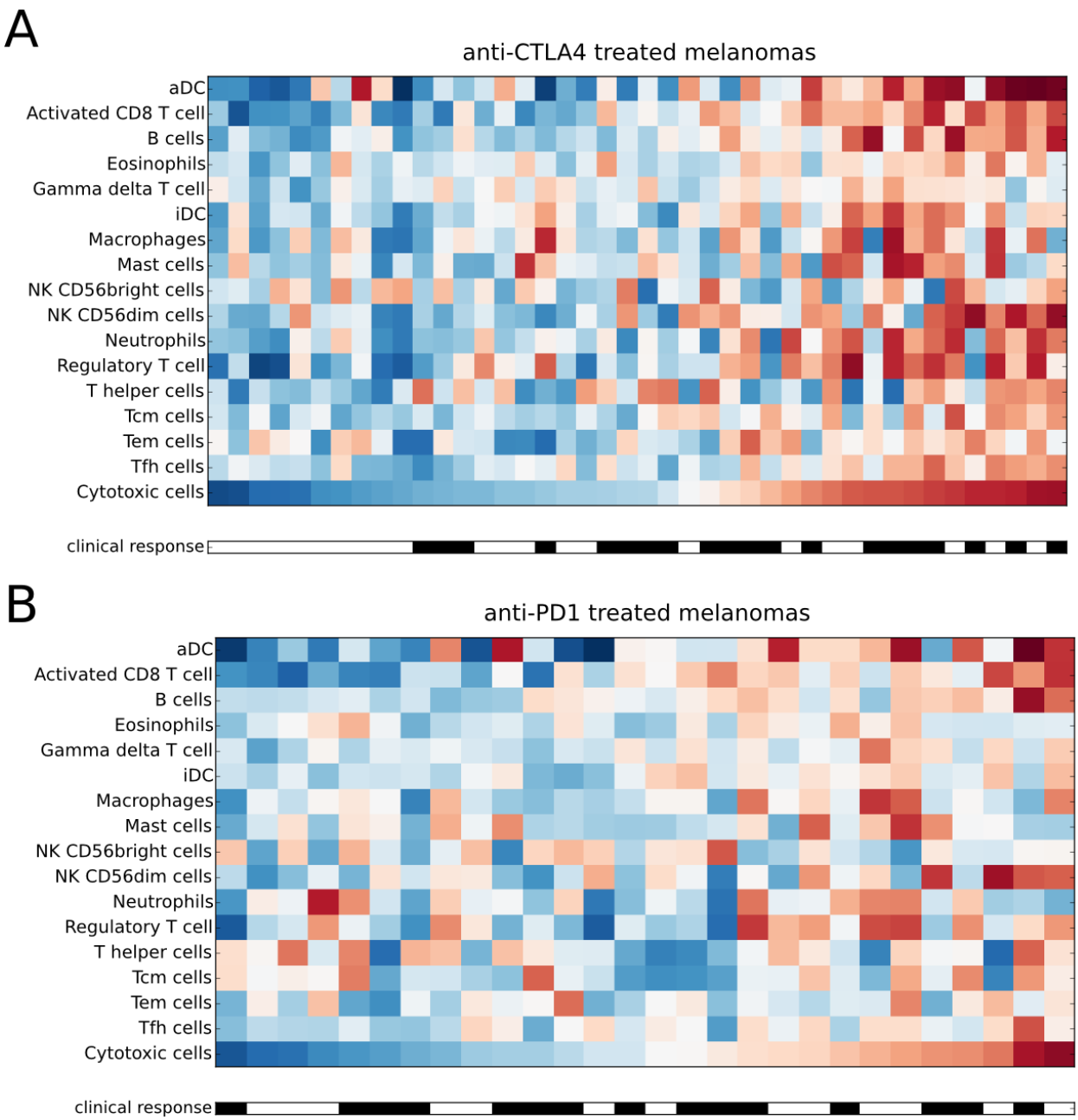


Figure S15 (1 out of 2)

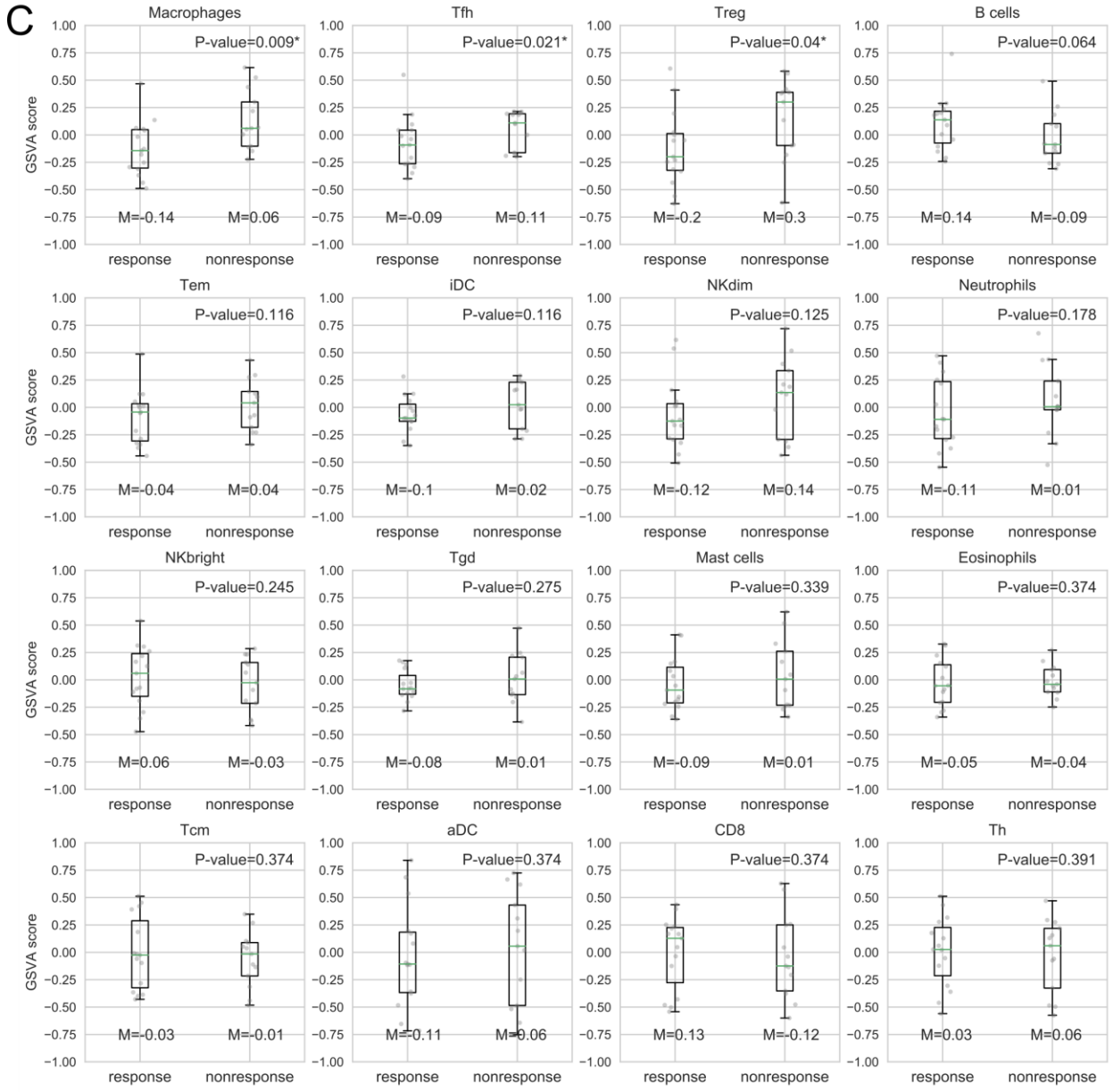


Figure S15. Immune infiltration pattern of two cohorts of metastatic melanomas treated with immune checkpoint inhibitors.

The heatmaps represent the relative abundance (GSVA scores) of the 16 immune cell populations comprised by the immune infiltration pattern (and the cytotoxic cells) across metastatic melanomas of patients treated with anti-CTLA4 [35] (n=42; panel A) or anti-PD-1 [36]

(n=28; panel B) therapies. In each heatmap, the samples are sorted by their relative abundance of cytotoxic cells. The bar below each heatmap identifies the patients that exhibited clinical response (partial and complete response; black color) to the treatment.
(C) Distribution of the relative abundance (GSVA enrichment scores) of each immune cell population in responders and non-responders to the PD-1 inhibitor. The significance of the differences was tested with a Mann Whitney U test (p-value shown).

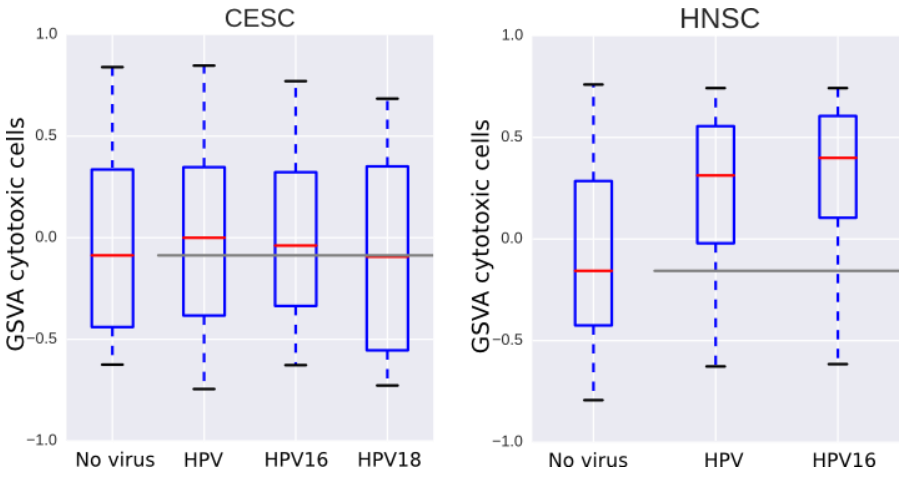


Figure S16. Distribution of the relative abundance (GSVA scores) of cytotoxic cells of human papillomavirus infected tumors

The boxplots represent the distribution of relative abundance (GSVA scores) of cytotoxic cells of tumors infected by different HPV subtypes (HPV16 and HPV18) and uninfected tumors (data from [37]).

SUPPLEMENTARY TABLES

Table S1. Gene sets representing immune cell populations and cell pathways.

Table S1A. Gene sets used to estimate the relative abundance of each of the 16 immune cell populations and the cytotoxic cells. The source of each gene set is also detailed.

Table S1B. Table describing the gene sets representing each pathway and their source.

Table S2. Details of the meta-processes employed (in Figure 4) in the description of tumor development in the three scenarios of immune infiltration

Table S3. Pan-cancer and per-cancer type GSVA scores for immune populations.

Zip file containing the relative abundance (GSVA enrichment scores) of the immune cell populations computed in tumors computed across the pan-cancer cohort (Table S3A), samples of healthy tissues (Table S3B), and in tumors of each cohort separately to build the immune-phenotypes (Tables S3C to S3ZE). Tumor samples are represented in rows and the immune cell types in columns. Each cell shows the GSVA enrichment score. The identifiers of the tables included in the zip file correspond to the following analyses:

Cancer type	Supp. table	Cancer type	Supp. table
Pan-cancer	TableS2A	UCEC	TableS3P
Healthy tissues	TableS2B	ESCA	TableS3Q
ACC	TableS3C	LIHC	TableS3R
CHOL	TableS3D	CESC	TableS3S
UCS	TableS3E	STAD	TableS3T
KICH	TableS3F	SKCM	TableS3U
UVM	TableS3G	LUSC	TableS3V
MESO	TableS3H	PRAD	TableS3W
PAAD	TableS3I	THCA	TableS3X
PCPG	TableS3J	LUAD	TableS3Y
SARC	TableS3K	HNSC	TableS3Z
KIRP	TableS3L	LGG	TableS3ZA
OV	TableS3M	KIRC	TableS3ZB
GBM	TableS3N	BLCA	TableS3ZC
COADREAD	TableS3O	BRCA	TableS3ZD
		BRCA-TN	TableS3ZE

Table S4. Immune-phenotypes

Immune-phenotype of each tumor sample included in the study.

Table S5. Enrichment for somatic driver alterations across tumor immune-phenotypes

Results of the logistic regression implemented to identify genes whose driver alterations are significantly enriched for a specific immune-phenotype, at pan-cancer and per-cancer type level (see Methods).

Table S6. Association of somatic driver alterations with immune populations

Results of the linear regression comparing the presence of driver alterations in genes associated with cancer immune-phenotypes (see Table S5) and the GSVA enrichment score of 16 immune cell populations. The model was adjusted by number of coding mutations and copy number alterations.

Table S7. Results of the GSEA enrichment

Each row corresponds to a pathway (see Methods). NES corresponds to the Normalized Enrichment Score and FDR Q-value corresponds to the P-value of the enrichment analysis adjusted by the gene set size and corrected by multiple testing, as provided by the GSEA software (see [38]).

SUPPLEMENTARY REFERENCES

1. Angelova, M. *et al.* Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy. *Genome Biol.* **16,** 64 (2015).
2. Bindea, G. *et al.* Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity* **39,** 782–795 (2013).
3. Charoentong, P. *et al.* Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. *Cell Rep.* **18,** 248–262 (2017).
4. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12,** 453–457 (2015).
5. Becht, E. *et al.* Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol.* **17,** (2016).
6. Şenbabaoğlu, Y. *et al.* Tumor immune microenvironment characterization in clear cell renal cell carcinoma identifies prognostic and immunotherapeutically relevant messenger RNA signatures. *Genome Biol.* **17,** 231 (2016).
7. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14,** 7 (2013).
8. Barbie, D. A. *et al.* Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **462,** 108–112 (2009).
9. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14,** 7 (2013).
10. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14,** 7 (2013).
11. Bindea, G. *et al.* Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity* **39,** 782–795 (2013).
12. Angelova, M. *et al.* Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy. *Genome Biol.* **16,** 64 (2015).
13. Charoentong, P. *et al.* Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. *Cell Rep.* **18,** 248–262 (2017).
14. Şenbabaoğlu, Y. *et al.* Tumor immune microenvironment characterization in clear cell renal cell carcinoma identifies prognostic and immunotherapeutically relevant messenger RNA signatures. *Genome Biol.* **17,** 231 (2016).
15. Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G. & Hacohen, N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* **160,** 48–61 (2015).
16. Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G. & Hacohen, N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* **160,** 48–61 (2015).
17. Cancer Cell Line Encyclopedia Consortium & Genomics of Drug Sensitivity in Cancer

Consortium. Pharmacogenomic agreement between two cancer cell line data sets. *Nature* **528,** 84–87 (2015).
18. Becht, E. *et al.* Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol.* **17,** (2016).
19. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12,** 453–457 (2015).
20. Jones, E. & Oliphant, T. SciPy: Open Source Scientific Tools for Python. Available at: http://www.scipy.org/. (Accessed: 03/2017)
21. Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4,** 2612 (2013).
22. Müllner, D. Modern hierarchical, agglomerative clustering algorithms. (2011).
23. Jones, E. & Oliphant, T. SciPy: Open Source Scientific Tools for Python. Available at: http://www.scipy.org/. (Accessed: 03/2017)
24. Murphy, K. P. *Machine Learning: A Probabilistic Perspective*. (MIT Press, 2012).
25. Heinze, G. A comparative investigation of methods for logistic regression with separated or nearly separated data. *Stat. Med.* **25,** 4216–4226 (2006).
26. Gelman, A., Jakulin, A., Pittau, M. G. & Su, Y.-S. A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat.* **2,** 1360–1383 (2008).
27. Maxpoint. Python package: bayes logistic. *GitHub* Available at: https://github.com/maxpoint/bayes_logistic. (Accessed: 31st August 2017)
28. Aran, D. *et al.* Widespread parainflammation in human cancer. *Genome Biol.* **17,** 145 (2016).
29. van der Walt, S., Chris Colbert, S. & Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. *Comput. Sci. Eng.* **13,** 22–30 (2011).
30. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 15545–15550 (2005).
31. Hugo, W. *et al.* Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. *Cell* **165,** 35–44 (2016).
32. Van Allen, E. M. *et al.* Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* **350,** 207–211 (2015).
33. Hugo, W. *et al.* Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. *Cell* **165,** 35–44 (2016).
34. Van Allen, E. M. *et al.* Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* **350,** 207–211 (2015).
35. Van Allen, E. M. *et al.* Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* **350,** 207–211 (2015).
36. Hugo, W. *et al.* Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. *Cell* **165,** 35–44 (2016).
37. Tang, K.-W., Alaei-Mahabadi, B., Samuelsson, T., Lindh, M. & Larsson, E. The landscape of viral expression and host gene fusion and adaptation in human cancer. *Nat. Commun.* **4,**

2513 (2013).
38. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 15545–15550 (2005).