Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade

Pornpimol Charoentong[†], Francesca Finotello[†], Mihaela Angelova[†], Clemens Mayer, Mirjana Efremova, Dietmar Rieder, Hubert Hackl, Zlatko Trajanoski*

[1]Biocenter, Division of Bioinformatics, Medical University of Innsbruck, Innsbruck, Austria

[†]Equal contribution

*To whom correspondence should be addressed:

zlatko.trajanoski@i-med.ac.at

Zlatko Trajanoski, PhD

Biocenter, Division of Bioinformatics

Medical University of Innsbruck

Innrain 80, 6020 Innsbruck Austria

Email: zlatko.trajanoski@i-med.ac.at

ABSTRACT:

Current major challenges in cancer immunotherapy include identification of patients likely to respond to therapy and development of strategies to treat non-responders. To address these problems and facilitate understanding of the tumor-immune cell interactions we inferred the cellular composition and functional orientation of immune infiltrates, and characterized tumor antigens in 19 solid cancers from The Cancer Genome Atlas (TCGA). Decomposition of immune infiltrates revealed prognostic cellular profiles for distinct cancers, and showed that the tumor genotypes determine immunophenotypes and tumor escape mechanisms. The genotype-immunophenotype relationships were evident at the high-level view (mutational load, tumor heterogenity) and at the low-level view (mutational origin) of the genomic landscapes. Using random forest approach we identified determinants of immunogenicity and developed an immunophenoscore based on the infiltration of immune subsets and expression of immunomodulatory molecules. The immunophenoscore predicted response to immunotherapy with anti-CTLA-4 and anti-PD-1 antibodies in two validation cohorts. Our findings and the database we developed (TCIA-The Cancer Immunome Atlas, http://tcia.at) may help informing cancer immunotherapy and facilitate the development of precision immuno-oncology.

Recently approved anticancer drugs which augment T cell activity by blocking cytotoxic T lymphocyte antigen-4 (CTLA-4), programmed cell death protein 1 (PD-1), or PD-1 ligand (PD-L1) show remarkable clinical effects. Analysis of long-term data of patients who received anti-CTLA-4 antibodies in unresectable or metastatic melanoma shows a plateau in the survival curve after 3 years[1], suggesting durable benefits or even curative potential. Over and above, efficacy of anti-PD-1 antibodies has been shown not only in melanoma, but also in nine different tumor types, as diverse as non-small cell lung cancer, liver cancer, kidney cancer and lymphoma[2]. We are currently witnessing a mind-blowing pace of development of checkpoint blockers evident from more than 150 clinical trials with monotherapies or combination therapies[2]. However, only a fraction of the patients is responsive to monotherapies with checkpoint blockers and the identification of the precise mode of action and predictive markers is a subject of intense research.

With the development of the immunotherapies with checkpoint blockers as well as other immunotherapeutic strategies including therapeutic vaccines and engineered T cells[3], the tumor-immune cell interaction came into focus. The investigation of tumor-immune cell interaction poses considerable challenges, due to the evolving nature of this two ecosystems: the development of cancer, which can be seen as evolutionary process, and the immune system, with a number of innate and adaptive immune cell subpopulations, some of which show phenotypic plasticity and possess memory. Using next-generation sequencing (NGS) technologies and bioinformatics tools it is now possible to computationally dissect tumor-immune cell interactions[4]. These immunogenomic analyses can provide information on the two crucial characteristics of the tumour microenvironment: 1) composition and functional orientation of the infiltrated immune cells, and 2) expression of the cancer antigenome, i.e. the repertoire of tumor antigens including two classes of antigens: neoantigens which arise from somatic mutations, and cancer-germline antigens (CGA)[5].

Previous studies have used genomic data from the The Cancer Genome Atlas (TCGA) to characterize neoantigens and their association with survival[6] or with cytolytic activity[7]. However, to the best of our knowledge, comprehensive decomposition of the immune infiltrates from these genomic data has not been performed so far. We have recently developed an analytical strategy to characterize the cellular composition of the immune infiltrates in tumors using gene set enrichment analyses (GSEA) and examined colorectal cancer (CRC) datasets from TCGA[8]. The analyses revealed that different immunophenotypes employ distinct tumor escape mechanisms and pinpointed novel targets for immunotherapy of CRC.

In this study, we carried out comprehensive immunogenomic analyses of the TCGA data for 19 solid cancers comprising 8243 tumor samples, by using analytical pipelines to decompose the immune infiltrates and characterize the neoantigens and CGAs. Decomposition of immune infiltrates revealed genotype-immunophenotype relationships and cellular profiles that were predictors of survival for distinct cancers. We then identified determinants of immunogenicity and developed an immunophenotype score based on the infiltration of immune subsets and expression of immunomodulatory molecules. The immunophenoscore predicted response to immunotherapy with anti-CTLA-4 and anti-PD-1 antibodies in two validation cohorts. Finally, we developed a web-accessible database TCIA (The Cancer Immunome Atlas) comprising the immunogenomic landscapes of solid cancers (http://tcia.at).

RESULTS

High-resolution genomic analyses of the tumor-immune interface

An overview of the strategy for the immunogenomic analyses and the used methods is shown in Figure 1a (for details see Methods and Supplementary Material). Briefly, we first built a compendium of genes (1625) related to specific immune cells using gene expression profiles from sorted immune cell subpopulations from 37 studies comprising 366 microarrays (Figure 1b). A subset of these genes, which are representative for specific immune cell subpopulations and are neither expressed in cancer cell lines nor in normal tissue were then selected based on correlation analysis (782 in total). We then used gene set enrichment analysis (GSEA)[9] with this gene set to decompose cellular profiles from RNA sequencing data from bulk tissue. Additionally, we adapted a deconvolution approach (CIBERSORT[10]) to RNA sequencing (RNA-seq) data to identify fractions of immune subpopulations. Since several important subpopulations were missing in the deconvolution method, we present here only the results from the GSEA method.

Exome sequencing data and SNP array data was used to estimate cancer cell fractions (CCF) and subsequently tumor heterogeneity[8]. RNA-sequencing data was used for HLA typing, and filtering of expressed genes with somatic mutations. The predicted 4-digit HLA class I alleles and the mutated expressed peptides were used as input for the algorithm netMHCpan[11] to estimate their binding affinities and predict neoantigens. RNA-sequencing data was used to derive expression levels of CGAs. Finally, exome sequencing data and SNP arrays data was used to estimate cancer cell fractions (CCF) and subsequently tumor heterogeneity[8]. Driver genes were assigned from a recently published study[12] whereas clonal and subclonal origin of the neoantigens was estimated as described[13].

TCGA data were analyzed for 8243 samples and 19 solid tumors. Additionally, available exome-seq and RNA-seq data from recent clinical studies using antibodies against CTLA-4[14] and PD-1[15] were analyzed using the same procedures. We then developed a web-accessible relational database TCIA (http://tcia.at) and deposited the results of the analyses including CCFs, HLA alleles, neoantigens, CGAs, expression of predefined immune subsets, and immune infiltrates calculated with both, GSEA and the deconvolution method.

Decomposition of immune infiltrates reveals prognostic cell profiles

Using our analytical strategy we estimated 28 subpopulations of TILs including major types related to adaptive immunity: activated, central, and effector memory CD4 and CD8 T cells, gamma delta T cells (Tγδ), T helper 1 (Th1), Th2, Th17, regulatory T cells (Treg), follicular helper T cells (Tfh), activated, immature, and memory B cells, as well as cell types related to innate immunity: macrophages, monocytes, mast cells, eosinophils, neutrophils, activated, plasmocytoid and immature dendritic cells (DCs), natural killer cells (NK), natural killer T cells (NKT), and myeloid-derived suppressor cells (MDSC).

As expected, the results of the decomposition of the immune infiltrates using GSEA revealed cellular heterogeneity across cancers and within individual cancer entities, and showed associations of specific TIL subpopulations with survival (Figure 2a). The cellular profiles associated with survival differed between cancers. As reported in a number of studies, the infiltration of many TIL subpopulations related to adaptive immunity was associated with good prognosis including activated CD8, effector memory and central memory CD8 cells, and effector memory CD4 cells, whereas MDSCs and Tregs were associated with bad prognosis. Using dimensional reduction technique we observed separation of the subpopulation associated with bad prognosis (MDSCs, Tregs) from the subpopulations associated with good prognosis. An example of this is shown in Figure 2b for lung adenoma (LUAD) for all cell types in the LUAD cohort as well as for individual tumors for selected subpopulations.

We then classified the tumors into three groups based on the mutational load: tumors with high mutational load (upper quartile), intermediate mutational load (two intermediate quartiles), and low mutational load (lower quartile) (Figure 2a). Tumors with high mutational load were enriched with activated T cells and effector memory T cells and were depleted with immunosuppressive cells like

Tregs and MDSCs, whereas tumors with low mutational load showed opposite enrichments and depletions (Figure 2c). As expected, patients bearing tumors with high mutational load and hence, tumors enriched with effector T cells and depleted with immunosuppressive cells had better outcome (Figure 2d).

The progression of the tumor across cancers was characterized by distinct immune cell patterns. For example, effector memory CD8 cells were enriched in stage I and stage II tumors and depleted in stage III and stage IV tumors (Figure 2e). In contrast, Tregs and MDSCs were depleted in early stage tumors and enriched in late stage tumors. In general, the enrichment of the TIL subpopulations related to adaptive immunity was decreasing whereas enrichment of TILs related to innate immunity was increasing. This was also evident at the level of individual markers (data not shown). Hence, the cellular composition of the immune infiltrates across solid cancers is shifting towards less favorable outcome during progression, also evident in the survival analysis (Figure 2f).

These data unveiled the cellular heterogeneity of the tumor infiltrates across and within solid cancers, and further supports the notion of the evolving nature of the cellular composition of the infiltrates during tumor progression.

Neoantigen landscapes are heterogeneous and sparse

The antigenome includes several classes of antigens which can be recognized by T cells[16]. Of these, two classes are of relevance for therapeutic cancer vaccination: 1) neoantigens, and 2) CGAs, i.e. proteins that are normally expressed by germline cells, but have aberrant expression in tumor cells. We therefore characterized these two antigens classes in 19 solid cancers.

The expression of CGAs was rather homogeneous across different cancer types (Figure 3a) as well as within individual cancers. The fraction of CGAs shared in at least 5% of the patients with expression above 2 TPM ranged from 22% (KIRC) to 81% (STAD) (Supplementary material). This result suggests that the expression of CGA is independent of both, the molecular phenotype and the immunophenotype of the tumors. The CGAs with the highest median expression in most of cancer types was KIAA0100 (Figure 3a), followed by PRAME (SKCM, OV, and UCEC), CTAGE5 (LIHC), XAGE1D (LUAD) and TMEFF2 (PRAD). The most shared expressed genes (TPM>2) that can be potentially used for vaccination were CAGE5 (BLCA, KIRC, KIRP, LIHC, and LUAD) and ATAD2 (LUSC, OV, BRCA, CESC, and HNSC). Both show also high median expression through all cancer types (Figure 3a). It is noteworthy that many CGA were also expressed in normal tissues, which can be a confounding factor for the development of a therapeutic vaccine (Supplementary material).

The neoantigen landscape in solid tumors was composed of 933,954 expressed neoantigens (911,548 unique) originating from 893,960 somatic point mutations. As expected, the number of neoantigens correlated with the mutational load (Figure 3b). Interestingly, the neoantigen frequency (number of neoantigens per mutation), which can be seen as surrogate marker for the antigenicity of the tumors was inversely correlated with the mutational load and was significantly different between the tumors with low, medium, and high mutational load (all adjusted pairwise p-values<0.0001, two-sided Dunn's post hoc tests on ranked sums; Figure 3c). Tumors with the highest neoantigen frequencies belonged to TCHA, BRAC, and PRAD. The neoantigen frequencies varied across cancers and ranged from 0.98 (SKCM and KICH) to 1.49 (KIRC) (Figure 3d). There was a slight increase in neoantigen burden from stage I to IV (all adjusted pairwise p-values<0.032; two-sided Dunn's post hoc tests on ranked sums; Figure 3e). Hence, with increasing tumor stage, tumor antigenicity appears to increase.

The fraction of neoantigens derived from driver genes for all solid tumors was 7.6% (ranging between 7.0% for LUSC and 10.6% for GBM) (Figure 3f). Hence, the bulk of neoantigens had its origin in passenger genes. On average, 56% of the neoantigens were of clonal origin with varying proportions of clonal to subclonal fractions across cancers (Figure 3f). The fractions of neoantigens with clonal origin ranged from 31% (KICH) to 71% (LUAD) (Supplementary material).

In strong contrast to the CGAs, the neoantigens were infrequently shared between patients (Figure 3g). From the total of 911,548 unique neoantigens only 24 were shared in at least 5% of patients in one or more cancer types. These shared neoantigens represent identical peptides originating from one or more

genes. As expected, the most frequent neoantigens were induced by mutations in driver genes like BRAF, RAS, and PIK3CA. Among these, only two peptides were shared in more than 15% of patients in one cancer type: KIGDFGLAT**E**K was shared in THCA and SKCM, and KLVVVGA**D**GV was shared in PAAD. The neoantigen KIGDFGLAT**E**K originates from $BRAF^{V600E}$ mutation, which is present in a large fraction of THCA[17] and SKCM[18] tumors. The neoantigen KLVVVGA**D**GV originates from the p.G12D mutation of KRAS, which is shared across a large fraction of PAAD patients[19].

Thus, the antigenome landscape is rather homogeneous with respect to CGAs and highly diverse and sparse with respect to the neoantigens.

Genotypes of the tumors determine immunophenotypes and tumor escape mechanisms

The immunogenomic analyses of the CRC data in our previous study revealed that the immunophenotypes in the hypermutated compared to the non-hypermutated tumors were characterized by increased enrichment of effector T cells[8], likely to be a consequence of the higher neoantigen burden. We asked the question how are the immunophenotypes related to other genomic features describing the complexity of the tumor genome, like tumor heterogeneity (high vs. low) and antigenicity (high vs. low). Tumor heterogeneity was estimated using exome sequencing and SNP array data as previously described[8]. To validate the estimation of the tumor heterogeneity, two different pipelines were applied on the CRC data showing similar results[8]. Tumor antigenicity was approximated with the neoantigen frequencies.

Tumors with high heterogeneity were enriched with activated T cells and effector memory T cells, and depleted of immunosuppressive cells (Figure 4a), but there was no difference in the prognosis (Figure 4b). Tumors with high antigenicity showed similar cellular patterns of immune infiltrates (Figure 4c), but were associated with better prognosis (logrank $p<0.005$) (Figure 4d).

We hypothesized that the genotype of the tumor determines the immunophenotype also in tumors with similar mutational loads. We selected two cancers with the lowest and highest mutational load, i.e. THCA and SKCM, and analyzed the immunophenotypes for their distinct genotypes: BRAF and RAS subtypes for THCA, and BRAF, RAS, NF1, and triple negative wild-type for SKCM. Additionally, we examined the expression of three classes of molecules that are involved in tumor escape mechanisms: MHC molecules (class I, class II and non-classical) which may be downregulated to avoid recognition by T cells, immunostimulators (e.g. OX40), which may be down-regulated to avoid immune destruction, and immunoinhibitory genes (e.g. CTLA-4) which may be up-regulated to enable tumor escape.

The different THCA genotypes were associated with specific immunophenotypes (Figure 4e) and showed distinct cellular patterns of immune infiltrates (Figure 4f) despite of comparable neoantigen burden ($p>0.05$, two-sided Wilcoxon rank sum test; Figure 4g). Analysis of the differentially expressed genes with respect to the immune-related Gene Ontology (GO) terms highlighted the pathways, which might explain the different effects on the immune system (Figure 4h). The expression levels of MHC molecules, immunostimulatory and immunoinhibitory molecules were also associated with the genotypes (Figure 4i). These results suggest that BRAF mutated THCA tumors employ different tumor escape mechanisms compared to RAS mutated tumors in THCA: BRAF tumors were infiltrated with immunosuppressive cells, whereas RAS tumors downregulated MHC molecules as well as immunomudulatory molecules.

Similarly, the SKCM genotypes were also associated with distinct immunophenotypes (Figure 4j), with BRAF tumors enriched with effector T cells, whereas other genotypes were enriched with immunosuppressive cells. It is noteworthy that for the SKCM cohort the neoantigen burden differed (all adjusted pairwise p-values ≤ 0.002; two-sided Dunn's post hoc tests on ranked sums; Figure 4k) and the genotypes had varying prognosis (Figure 4l). The four SKCM genotypes were also associated with varying expression levels of MHC and immunomodulatory molecules (Figure 4d). Analysis of the differentially expressed genes with respect to the immune-related GO terms is shown in the Supplementary material.

The results of this study suggest that the genotypes of the tumor determine the immunophenotypes and the tumor escape mechanisms. This was evident at high-level view (e.g. mutational load, tumor heterogenity) and at low-level view (e.g. mutational origin like BRAF or RAS mutated tumors) of the genomic landscapes.

Random forest classification identifies major determinants of tumor immunogenicity in solid cancers

The results of this work showed not only highly heterogeneous TILs but also varying ratios of different T cell subsets including suppressive ones. These observations raise the question of the underlying molecular mechanisms that explain the differences in immunogenicity of the tumors. The question can be reduced to the notion of sources of immunogenic differences, which can be divided into two categories: tumor intrinsic factors and tumor extrinsic factors. Tumor intrinsic factors include the mutational load, the neoantigen load, the neoantigen frequency, the expression of immunoinhibitors and immunostimulators (e.g. PD-L1[16]), and HLA class I molecule alterations[20]. Tumor extrinsic factors include chemokines which regulate T cell trafficking[21], infiltration of effector TILs and immunosuppressive TILs, and soluble immunomodulatory factors (cytokines)[21]. We reasoned that the immunogenicity of the tumor can be represented by the cytolytic activity (expression of granzyme A *GZMA* and perforin *PRF1*) and examined major tumor intrinsic and extrinsic factors. For each cancer type we used a random forest classification approach, which is based on a multitude of decision trees, including 127 parameters (Supplementary material) to separate tumors with high cytolytic activity from tumors with low cytolytic activity. For individual cancer types the most predictive features were identified using the mean decrease of accuracy over all cross validated predictions (Figure 5a). For each of the studied cancers the analysis revealed only immune-related factors which we classified in four classes: 1) infiltration of activated CD8/CD4 T cells and effector memory CD8/CD4 T cells, 2) infiltration of immunosuppressive cells (Tregs and MDSCs), 3) expression of MHC class I, class II, and non-classical molecules, and 4) expression of certain co-inhibitory and co-stimulatory molecules (Figure 5a).

To visualize the information about the immunophenotypes of the tumors based on the identified major determinants we constructed an immunophenogram that includes these four categories (Figure 5b). We then calculated an aggregated score – immunophenoscore – based on the expression of the representative genes or gene sets (see online Methods). The immunophenoscore was significantly associated with survival in 14 solid cancers (Figure 5c). The immunophenogram enables the visualization of the immunophenotypes of the tumor also at the levels of individual tumors as seen for CRC (Figure 4d) and BRCA (Figure 4e). As can be seen, subgroups of tumors like microsatellite-instable (MSI) and triple-negative breast cancer (TNBC) are grouped with respect to distinct TILs like efector memory CD8 T cells and Tregs, respectively. We also implemented an interactive version of the tool (see http://tcia.at), which enables visualization and scoring of tumor samples using expression data for the markers we identified using the random forest approach.

Hence, using a data-driven approach, a large number of tumors (>8000), and both, genomic and immunological features, we were able to identify the major determinants of tumor immunogenicity. Based on these results we propose a visualization method –the immunoiphenogram, and a scoring scheme –the immunophenoscore, for solid tumors.

Immunophenoscore predicts response to immunotherapy with CTLA-4 and PD-1 blockers

The overall mutational load has been associated with responses to anti-PD-1 therapy[22]. However, it was also shown that tumors with low mutational load are also responsive to therapy with checkpoint blockers[23]. More recently, it was proposed that neoantigens from clonal origin elicit T cell reactivity and sensitivity to checkpoint blockade[24].

We reasoned that the determinants of immunogenicity identified using the random forest approach might have also a predictive value and analyzed two genomic and transcriptomic datasets from patients with melanoma treated with anti-CTLA-4[14] and anti-PD-1 antibodies[15]. Using RNA-seq data

and GSEA we reconstructed the TIL landscape and scored the patients using the immunophenoscore. The immunophenograms of the individual patients treated with anti-CTLA-4 antibodies are shown in Figure 6a. Tumors of the responders were enriched with cytotoxic cells (CD8, Tγδ, NK cells) and depleted of MDSCs and Tregs (Figure 6b). More importantly, the immunophenogram and the score derived from the analyses enabled stratification of patients to responders and non-responders (Figure 6c) with an improved predictive power compared to the expression of checkpoint molecules as can be seen in the receiver operating characteristics (ROC) curve (Figure 6d).

Similar observation was made also for the patients treated with anti-PD-1 antibody. Visualization of the determinants of the immunogenicity with the immunophenogram for responders and nonresponders showed distinct expression patterns in the two groups (Figure 6e). Again, tumors of the responders were enriched with cytotoxic cells (CD8, Tγδ, NK cells) and depleted of MDSCs and Tregs, as evident in the volcano plot (Figure 6f). Finally, the immunophenoscore (Figure 6g), and the ROC curve (Figure 6h) showed the predictive value also for patients treated with anti-PD-1 antibodies.

Hence, the immunophenoscore developed from a panel of immune-related genes belonging to the four classes: effector cells, immunosuppressive cells, MHC molecules, and selected immunomodulators has predictive value in melanoma patients treated with the CTLA-4 and PD-1 blockers.

DISCUSSION

We used an analytical strategy to comprehensively characterize the immunophenotypes and the antigenomes of solid human cancers and developed a valuable resource for cancer geneticists and immunologists. Our approach to deeply mine cancer genomic datasets revealed cellular profiles of immune infiltrates that were predictors of survival, genotype-immunophenotype relationships, and a panel of immune genes with predictive value for checkpoint blockade with CTLA-4 and PD-1 antibodies, which was subsequently validated in two independent cohorts. The computational dissection of the complex tumor microenvironment suggests several important biological conclusions.

First, our intriguing observation of the association of the genotypes and the immunophenotypes and the distinct tumor escape mechanisms determined by the genotypes implicates that the interaction with the immune system is predetermined by the genetic basis of the tumors. This genotype-immunophenotype relationship was evident at the high-level view of the genomic landscape including hypermutated vs. non-hypermutated, heterogenous vs. homogenous, and antigenic vs. less antigenic tumors. Strikingly, this association was observed also at the low-level view, i.e. at the level of mutational origin of the tumors like BRAF or RAS, as well as for tumors at both ends of the mutational load: THCA and SKCM. This raises the question of the underlying molecular mechanisms. Currently we can only speculate on the possible mechanisms, which can include modifications of different signaling pathways due to mutational changes. For example, evidence from clinical and experimental studies has demonstrated that signaling by BRAF and RAS oncogenes can present similarities but also differential effects[25]. The analysis of the genes differentially expressed between the two genotypes in THCA indicates several processes that might be involved in the sculpting of the immune landscape including chemotaxis, T cell differentiation, T cell proliferation, T cell activation or B cell receptor signaling pathway.

Second, charting of the antigenome showed that the number and the expression levels of CGAs were similar across tumors with different molecular profiles and with different TIL enrichments, whereas the neo-antigen landscape was highly diverse. This further supports the notion that the T cell responses are primarily directed against neoantigens rather than against CGAs. It is noteworthy that the quality of the T cell responses directed against neoantigens and CGAs has not been compared so far and future experimental studies are needed in order to validate our observation.

Beyond these biological insights, the results from this study have also important implications for cancer immunotherapy with checkpoint blockers and for therapeutic vaccination. Most importantly, we propose an immunophenoscore, i.e. a panel of immune genes for classification of patients likely to respond to therapy with antibodies targeting CTLA-4 and PD-1. The immunophenogram we developed to visualize the determinants of immunogenicity was derived from the cancer genomic data from >8000 patients in an unbiased manner, and is similar to the conceptual immunogram that was recently proposed[26]. Here we used RNA expression data but the method can be also used with other techniques for gene expression profiling like microarrays and qPCR. Notably, the method can be further improved by optimizing the immunophenoscore for specific cancers using larger datasets.

With respect to therapeutic vaccination, the sparsity of the neo-antigen space advocates against the development of off-the-shelf vaccines. A notable exception represents vaccine for THCA, SKCM or PAAD, which could target 20% of the patients. Thus, personalized cancer vaccination strategy is required in which whole-exome NGS is performed to identify somatic mutations, followed by bioinformatics analyses to identify neoantigens, and synthesis of peptide- or DNA/RNA-based vaccines. Viability of such personalized cancer vaccination strategy was recently demonstrated in clinical studies in melanoma patients[27,28].

Finally, with the large number of ongoing studies with checkpoint blockers either as monotherapy or combination therapy, we expect that the immunogenomic data amount will continuously increase. Thus, we strongly believe that the TCIA in its current form and future incorporation of additional datasets represents an important contribution to the field and will enhance the identification of novel mechanistic insights of the complex tumor-immune cell interactions.

Online METHODS

Identification of immune-related genes

For the identification of immune-related genes we used Affymetrix HG-U133A microarray data from a number of different studies (Supplementary material) according to our recently developed approach for TILs in CRC[8], and included expression profiles from normal tissues[29] and from relevant cancer cell lines[29]. The so defined immune-related gene expression signatures comprised 1625 genes. For the identification of subsets of genes representative for specific immune cell types, we selected genes with an average correlation $r \geq 0.4$ ($p < 0.01$) between all specific immune genes in the same cell type. This threshold was chosen to satisfy two goals: selection of genes with relatively high correlation such that their correlation could not be considered a chance event; and selection of a reasonable number of genes (at least 10 genes per subpopulations). Furthermore, using a set of genes instead of individual markers for specific TILs ensures robust estimation and is less susceptible to noise arising from the expression of the genes in tumor or stromal cells. The immune genes (782) are listed in the Supplementary material.

Genomic and clinical data

Genomic and clinical data for 19 solid tumors from The Cancer Genome Atlas (TCGA) were downloaded via the TCGA data portal or queried via Broad GDAC firehose/firebrowse and include clinical information (n=9151), SNP arrays (n=3377), microarrays (n=550), and RNA-seq expression profiles (n=8243), as well as curated MAF files of somatic mutations[30] (Supplementary material). All curated MAF files were re-annotated with Oncotator[31] in order to have a common annotation resource and file format. FASTQ files of RNA-seq reads (n=8,398) were downloaded from CGHub (https://cghub.ucsc.edu) and analyses were done on the fly. RNA-seq expression levels, available as Transcripts-Per-Millions (TPM) were transformed to $\log_2$(TPM+1). For RNA-seq expression data, we considered data from primary solid tumor and normal solid tissue for all cancer types. In case of subjects with multiple RNA-seq samples available, the sample with the highest sequencing depth was considered for GSEA and deconvolution.

Identification of TILs subpopulations

We used single sample gene set enrichment analysis (ssGSEA)[32] to identify immune cell types that are over-represented in the tumor microenvironment. The expression levels of each gene were z-score normalized across all patients. For each patient (or group of patients) genes were then ranked in descending order according to their z-scores (mean of z-scores). The association was represented by a normalized enrichment score (NES). An immune cell type was considered enriched in a patient or group of patients when FDR (q-value) $\leq$ 10%. Clustering and visualization was done with the software Genesis[33]. Enriched immune celltypes (NES>0) are illustrated as bubble plots (similar to[34]) where the size represents percentage of patients with enriched cell type and color is encoded by hazard ratio from overall survival analysis. The similarity of the enrichment of immune infiltrates (averaged NES) were calculated using multidimensional scaling (MDS). The distribution of selected cell types for individual patients were analyzed with t-distributed stochastic neighbor embedding (t-SNE)[35] using the Matlab toolbox t-SNE.

Additionally, a deconvolution approach was applied using the tool CIBERSORT[10]. At the time of analysis, CIBERSORT was only optimized for microarray data, since its signature matrix was built from a microarray compendium. Therefore, we implemented a strategy to modify RNA-seq data from TCGA to be used as input for CIBERSORT. To build the model, we considered only tumor samples for which both Affymetrix microarrays and Illumina RNA-seq data were available (131 samples for LUSC, 266 for OV and 153 for GBM). We modeled the RNA-Seq-to-microarrays mapping with a gene-specific, cubic smoothing spline and used the model to transform RNA-seq data into microarrays-like data. Transformed data were then processed with the R-based version of CIBERSORT, with default parameter settings. The performance of the model was tested with leave-one-out cross-validation, confirming a high agreement between the cell fractions estimated from the modified RNA-seq data and those computed from the corresponding microarray data (0.88 Pearson's correlation, data not shown).

Cancer-germline antigens

The list of cancer germline antigens (CGA) was extracted from the Cancer-Testis database[36] (Supplementary material). The heatmap of CGA expression per cancer type was computed on the median $\log_2$(TPM+1) values per cancer type. Furthermore, only expressed CGAs were selected, by considering a median TPM per cancer type higher than 2. The heatmaps of log-fold-changes between tumor and normal samples were computed as the difference between the median $\log_2$(TPM+1) in tumor versus normal samples.

Characterization of neoantigens

HLA alleles were called from RNA-seq FASTQ files using Optitype[37], selected for its high performance and for its applicability to RNA-seq data. For each subject, the HLA alleles estimated from the sample with the highest coverage over the HLA locus were considered. To estimate the mutated proteins, we focused on non-synonymous missense mutations, and selected mutations associated to Uniprot protein identifiers. The protein sequence retrieved from Uniprot was changed according to the non-synonymous, missense mutations reported in the MAF file, and truncated in case of stop codons. We removed candidate proteins affected by annotation inconsistencies between protein identifiers and predicted effect. Peptides of 8-11 amino acids in length, covering the mutated region of the protein, were analyzed with NetMHCpan[11] (Version 2.8) to estimate their binding affinity to the HLA alleles. Self-antigens mapping to human Uniprot proteins were identified with BLAST and filtered out. Amongst the candidate antigenic peptides, we selected strong binders with binding affinity < 500 nM as in[7], and considered peptides arising from expressed genes. We identified expressed genes as those having median TPM greater than 2 in a given cancer type.

Estimation of tumor heterogeneity and clonality of mutations

The ABSOLUTE algorithm[38] was used to integrate the copy number data together with the somatic mutations in order to estimate the purity and ploidy, and measure the fraction of cancer cells per mutation (CCF). The SNP data were downloaded from the TCGA portal and analyzed with HAPSEG[39]. The tumor heterogeneity was estimated as the area under the curve (AUC) of the cumulative density function from all cancer cell fractions per tumor. A mutation was classified as clonal if the CCF was > 0.95 with probability > 0.5, and subclonal otherwise[13]. Since at the time of our data freeze, only hg18 was available for OV and the SNP array data for STAD were not processed, these cancers were excluded from the analysis of the clonal/subclonal origin of neoantigens.

The CCFs in the CRC samples were also estimated using an alternative analysis workflow. SNP 6.0 arrays downloaded from the TCGA portal were normalized and preprocessed using Affymetrix Power Tools. LogR and B-allele frequencies were obtained using PennCNV v1.0.3[40]. ASCAT v2.4[41] was used to generate allele-specific copy number alterations, as well as tumor purity and ploidy estimates. Samples that failed to be fit by ASCAT were discarded from the analysis. The ASCAT-derived copy number and purity estimates together with the variant allele frequency of each point mutation were integrated and used as an input to PyClone v0.12.7[42] to infer the clonal composition of each sample and the CCF of each mutation. As a complementary approach, we also inferred the clonal subpopulations with EXPANDS v1.6.1[43] using the copy number estimates from ASCAT and the variant allele frequencies of the mutations.

Identification of determinants of tumor immunogenicity, immunophenogram, and immunophenosocore

For each patient the cytolytic activity was calculated as the mean of the *GZMA* and *PRF1* expression levels (log2 (TPM+1)) as previously used[7]. For each cancer type, patients were divided into two groups based on median cytolytic activity. A random forest classifier[44] separating the group of patients with TILs exhibiting higher cytolytic activity from the group of patients with TILs exhibiting lower cytolytic activity was trained using the R package *randomForest* with 10,000 trees and included mutational load per megabase, number of neoantigens, fraction of neoantigens per mutations, expression of MHC related molecules, expression of immunomodulatory factors, and mean expression of the respective immune genes for each of the 28 immune celltypes as independent variables. The mean decrease of accuracy over all out-of-bag cross validated predictions was used to rank predictors.

The immunophenogram was constructed similar to the recently proposed metabologram[45]. Samplewise z-score from gene expression of all factors (cell types) included in any of the ten best predictors within each cancer type are color coded and divided into four categories (effector cells, suppressive cells, MHC related molecules, immune modulators). The outer part of the wheel includes individual factors, whereas the inner wheel illustrates the weighted average z-scores of the factors included in the particular category.

The immunophenscore (IPS) was calculated on a 0-10 scale based on the expression of the representative genes or gene sets of the immunophenogram. Samplewise z-scores are positively weighted according to stimulatory factors (cell types) and negatively weighted according to inhibitory factors (cell types) and averaged. Z-scores≥3 were designated as IPS10 and z-scores≤0 are designated as IPS0. To determine the predictive power patients were stratified into responder and non-responder and a univariate logistic regression analyses was performed using leave-one-out cross validation including the IPS, CTLA-4 expression, PD1 expression, or PD-L1 expression as independent variable. As predictive value the area under the curve (AUC) from receiver operating characteristics (ROC) analyses (R package *ROCR)* was used.

Statistical analyses

Sample sizes from available TCGA data were considered adequate as sufficient power using equivalent tests was observed in a previous study[8]. To test for differential expression across two groups (tumor and normal) we used the R package *DESeq2* on raw count data. The p-values were adjusted for multiple testing based on the false discovery rate (FDR) according to the Benjamini-Hochberg approach. For comparison of two patient groups two-sided Student's t-test was used where stated, otherwise the none-parametric two-sided Wilcoxon-rank sum test was used. For comparisons among multiple patient groups one-way analysis of variance (ANOVA) and Tukey's HSD post-hoc tests were used where stated, otherwise the none parametric Kruskal-Wallis test followed by two-sided Dunn's pairwise post hoc tests on rank sums with Benjamini-Hochberg adjustment of p-values using the R package *PMCMR* were used. Normality of the distributions was tested with Shapiro-Wilk test and for normal distributed data the variance within each group of data was estimated and tested for equality between groups by a two-sided F-test. Distributions of data are shown either as individual data points, as box-and-whisker plots, or as violin plots.

Overall survival analysis were performed using the R package *survival* and the patients were dichotomized based on median expression (NES) or divided in two or more groups by specified parameters. Kaplan Meier estimator of survival was used to construct the survival curves. Logrank tests (corresponding to a two-sided z-test) were used to compare overall survival between patients in different groups and hazard ratio (HR) (95% confidence interval) was provided for comparison of two groups. Patients for each cancer were divided in two groups based on median immunophenoscore and univariate cox regression analysis were performed and illustrated as forest plot showing $\log_2$(HR) and 95% confidence interval. Proportional hazard assumptions were tested. Analysis and visualization of Gene Ontology terms associated to differentially expressed genes was performed with ClueGO[46].

TCIA database

The web application TCIA is based on the MEAN Stack, which refers to MongoDB, ExpressJS, AngularJS and NodeJS and is completely written in JavaScript. As scaffolding tool we used the AngularJS Full-Stack Generator. AngularJS in combination with Bootstrap as front-end framework uses on the server side asynchronous data access through a RESTful Node API, built with ExpressJS. The data is stored in MongoDB, which is an open-source NoSQL database that provides a dynamic schema design. The Highcharts library is used for charting and visualization of the data. TCIA is supported by JavaScript capable browsers i.e. Google Chrome, Mozilla Firefox, Safari, Microsoft Internet Explorer, Microsoft Edge.

ACKNOWLEDGMENTS

The results shown here are in part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov. This work was supported by the Austrian Science Fund (DK Molecular Cell Biology and Oncology), the Tiroler Standortagentur (Bioinformatics Tyrol), the European Commission (Horizon2020 project APERIM: Advanced bioinformatics tools for personalized cancer immunotherapy), and the Austrian National Bank (Jubiläumsfondsprojekt Nr. 16534).

AUTHOR CONTRIBUTIONS

Z.T conceived the project. P.C. developed the GSEA method and analyzed the data. F.F. and M.A. analyzed the neoantigens and CGAs. M.A. and M.E. estimated tumor heterogeneity and clonality of mutations. D.R. organized and managed the data transfer and storage. C.M. and D.R. developed the database. P.C. and H.H. developed the random forest approach and the immunophenogram, and analyzed the data. H.H. and Z.T interpreted the results. Z.T. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

REFERENCES

1. Schadendorf, D. *et al.* Pooled Analysis of Long-Term Survival Data From Phase II and Phase III Trials of Ipilimumab in Unresectable or Metastatic Melanoma. *J Clin Oncol* **33**, 1889-94 (2015).

2. Wolchok, J.D. PD-1 Blockers. *Cell* **162**, 937 (2015).

3. Schumacher, T.N. & Schreiber, R.D. Neoantigens in cancer immunotherapy. *Science* **348**, 69-74 (2015).

4. Hackl, H., Charoentong, P., Finotello, F. & Trajanoski, Z. Computational genomic tools for dissecting tumor-immune cell interactions. *Nat Rev Genet*, (in press) (2016).

5. Heemskerk, B., Kvistborg, P. & Schumacher, T.N. The cancer antigenome. *EMBO J* **32**, 194-203 (2013).

6. Brown, S.D. *et al.* Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival. *Genome Res* **24**, 743-50 (2014).

7. 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).

8. 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).

9. 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**, 15545-50 (2005).

10. Newman, A.M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* **12**, 453-7 (2015).

11. Nielsen, M. *et al.* NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *PLoS One* **2**, e796 (2007).

12. Rubio-Perez, C. *et al.* In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities. *Cancer Cell* **27**, 382-96 (2015).

13. Landau, D.A. *et al.* Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* **152**, 714-26 (2013).

14. Van Allen, E.M. *et al.* Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* **350**, 207-11 (2015).

15. Hugo, W. *et al.* Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. *Cell* (2016).

16. Coulie, P.G., Van den Eynde, B.J., van der Bruggen, P. & Boon, T. Tumour antigens recognized by T lymphocytes: at the core of cancer immunotherapy. *Nat Rev Cancer* **14**, 135-46 (2014).

17. Cancer Genome Atlas Research, N. Integrated genomic characterization of papillary thyroid carcinoma. *Cell* **159**, 676-90 (2014).

18. Cancer Genome Atlas, N. Genomic Classification of Cutaneous Melanoma. *Cell* **161**, 1681-96 (2015).

19. Witkiewicz, A.K. *et al.* Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets. *Nat Commun* **6**, 6744 (2015).

20. Garrido, F., Cabrera, T. & Aptsiauri, N. "Hard" and "soft" lesions underlying the HLA class I alterations in cancer cells: implications for immunotherapy. *Int J Cancer* **127**, 249-56 (2010).

21. Gajewski, T.F. *et al.* Immune resistance orchestrated by the tumor microenvironment. *Immunol Rev* **213**, 131-45 (2006).

22. Rizvi, N.A. *et al.* Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* **348**, 124-8 (2015).

23. Tran, E. *et al.* Immunogenicity of somatic mutations in human gastrointestinal cancers. *Science* **350**, 1387-90 (2015).

24. McGranahan, N. *et al.* Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science* **351**, 1463-9 (2016).

25. Oikonomou, E., Koustas, E., Goulielmaki, M. & Pintzas, A. BRAF vs RAS oncogenes: are mutations of the same pathway equal? Differential signalling and therapeutic implications. *Oncotarget* **5**, 11752-77 (2014).

26. Blank, C.U., Haanen, J.B., Ribas, A. & Schumacher, T.N. CANCER IMMUNOLOGY. The "cancer immunogram". *Science* **352**, 658-60 (2016).

27. Robbins, P.F. *et al.* Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells. *Nat Med* **19**, 747-52 (2013).

28. van Rooij, N. *et al.* Tumor exome analysis reveals neoantigen-specific T-cell reactivity in an ipilimumab-responsive melanoma. *J Clin Oncol* **31**, e439-42 (2013).

29. Petryszak, R. *et al.* Expression Atlas update--an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Res* **44**, D746-52 (2016).

30. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333-9 (2013).

31. Ramos, A.H. *et al.* Oncotator: cancer variant annotation tool. *Hum Mutat* **36**, E2423-9 (2015).

32. Barbie, D.A. *et al.* Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **462**, 108-12 (2009).

33. Sturn, A., Quackenbush, J. & Trajanoski, Z. Genesis: cluster analysis of microarray data. *Bioinformatics* **18**, 207-8 (2002).

34. Spinelli, L., Carpentier, S., Montanana Sanchis, F., Dalod, M. & Vu Manh, T.P. BubbleGUM: automatic extraction of phenotype molecular signatures and comprehensive visualization of multiple Gene Set Enrichment Analyses. *BMC Genomics* **16**, 814 (2015).

35. van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9**, 2579-2605 (2008).

36. Almeida, L.G. *et al.* CTdatabase: a knowledge-base of high-throughput and curated data on cancer-testis antigens. *Nucleic Acids Res* **37**, D816-9 (2009).

37. Szolek, A. *et al.* OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics* **30**, 3310-6 (2014).

38. Carter, S.L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol* **30**, 413-21 (2012).

39. Carter, S.L., Meyerson, M. & Getz, G. Accurate estimation of homologue-specific DNA concentration-ratios in cancer samples allows long-range haplotyping. (2011).

40. Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* **17**, 1665-74 (2007).

41. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A* **107**, 16910-5 (2010).

42. Roth, A. *et al.* PyClone: statistical inference of clonal population structure in cancer. *Nat Methods* **11**, 396-8 (2014).

43. Andor, N., Harness, J.V., Muller, S., Mewes, H.W. & Petritsch, C. EXPANDS: expanding ploidy and allele frequency on nested subpopulations. *Bioinformatics* **30**, 50-60 (2014).

44. Breiman, L. Random forests. *Machine Learning* **45**, 5-32 (2001).

45. Hakimi, A.A. *et al.* An Integrated Metabolic Atlas of Clear Cell Renal Cell Carcinoma. *Cancer Cell* **29**, 104-16 (2016).

46. Bindea, G. *et al.* ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25**, 1091-3 (2009).

FIGURE LEGENDS:

Figure 1 Pan-cancer immunogenomic analyses. (**a**) The scheme shows the immunogenomic analyses and the types of data used for the analyses. The results are deposited in a web-accessible database The Cancer Immunome Atlas (TCIA) (http://tcia.at). (**b**) Immune-related signatures are derived from expression profiles of purified immune cells, normal cells, and cancer cell lines, and used for the gene set enrichment analysis (GSEA) of the TCGA RNA-seq data.

Figure 2 Decomposition of immune infiltrates in solid cancers. (**a**) Immune subpopulations across 19 solid cancers. Cancers are sorted according to the mutational load and the immune cell subpopulations alphabetically for adaptive and innate immunity. Shown are the fractions of tumors with high, medium, and low mutational load. (**b**) Visualization of the immune infiltrates (averaged normalized enrichment score (NES)) in lung adenocarcinoma (LUAD) for all patients using two dimensional coordinates from multidimensional scaling (MDS) (upper panel) and for individual patients and selected cell types based on two dimensional coordinates from t-distributed stochastic neighbor embedding (t-SNE) (lower panel). (**c**) Volcano plots for the enrichment (blue) and depletion (yellow) of immune cell types across cancers for tumors with high, intermediate, and low mutational load calculated based on the NES score from the GSEA. (**d**) Survival analysis of tumors with high, intermediate, and low mutational load (logrank test). (**e**) Volcano plots for the enrichment (blue) and depletion (yellow) of immune cell types across cancers for tumor stage I to IV calculated based on the NES score from the GSEA. (**f**) Survival analysis across cancers for tumor stage I to IV (logrank test).

Figure 3 Antigenomes in solid cancers. (**a**) Expression of cancer-germline antigens (CGA) in solid tumors. (**b**) Pearson's correlation of the the neoantigen burden and mutational load (low, medium, high) across cancers. (**c**) Pearson's correlation of the neoantigen frequencies and mutational load (low, medium, high) across cancers (Kruskal Wallis test followed by two-sided Dunn's pairwise post hoc tests on rank sums with Benjamini-Hochberg adjustment of p-values). Gray box: upper quartile of the neoantigen frequencies for the low-mutation group and the fraction of tumors represented in this box. (**d**) Neoantigen frequencies for solid cancers. (**e**) Neoantigen load for different tumor stages (Kruskal Wallis test followed by two-sided Dunn's pairwise post hoc tests on rank sums with Benjamini-Hochberg adjustment of p-values). (**f**) Fractions of neoantigens and their origin. (**g**) Shared neoantigens in solid tumors. Shown are only neoantigen shared in at least 5% of the tumors.

Figure 4 Genotypes and immunophenotypes in solid cancers. (**a**) Tumor heterogeneity and immune infiltrates. Shown is a volcano plot for tumors with high and low heterogeneity calculated based on the NES score from the GSEA. (**b**) Survival analyses for tumors with high and low heterogeneity (logrank test). Insert shows mutational load (two-sided Wilcoxon rank sum test). (**c**) Neoantigen frequencies and immune infiltrates. Shown is a volcano plot for tumors with high and low antigenicity calculated based on the NES score from the GSEA. (**d**) Survival analyses for tumors with high and low neoantigen frequency (logrank test). Insert shows mutational load (two-sided Wilcoxon rank sum test). (**e**) Hierarchical clustering of immune cell composition for BRAF and RAS mutated THCA tumors. (**f**) Volcano plot for BRAF and RAS mutated TCHA tumors calculated based on the NES score from the GSEA. (**g**) Mutational load for BRAF and RAS mutated TCHA tumors (two-sided Wilcoxon rank sum test). (**h**) Gene ontology (GO) analysis of the differentially expressed genes for BRAF and RAS mutated TCHA tumors using ClueGO[46]. (**i**) Expression of MHC and immunomodulatory molecules in BRAF and RAS mutated TCHA tumors. Expression values were compared to normal tissue ($\log_2$-fold changes are color coded according to the legend). (**j**) Volcano plots for SKCM genotypes calculated based on the NES score from the GSEA. (**k**) Mutational load for SKCM genotypes (Kruskal Wallis

test followed by two-sided Dunn's pairwise post hoc tests on rank sums with Benjamini-Hochberg adjustment of p-values). (**l**) Survival analysis for SKCM genotypes (logrank test). (**m)** Expression of MHC and immunomodulatory molecules for SKCM genotypes. Expression values are represented by z-score calculated across all SKCM tumors and color coded according to the legend

Figure 5 Determinants of immunogenicity in solid cancers. (**a**) Major parameters determining immunogenicity in solid cancers revealed using random forest approach. (**b**) Immunophenogram based on the results of the random forest analyses. (**c**) Survival analyses using the immunophenoscore for all solid cancers. Forest plots showing log2 hazard ratio (95% confidence interval), * indicates $p<0.05$ (**d**) Visualization based on two dimensional coordinates from multidimensional scaling (MDS) of expression profiles from the genes and the gene sets as used in the immunophenogramm in colorectal cancer (CRC). Mean expression (z-score) of immune genes of Tem CD8 cells as well as MSI and MSS samples for CRC are indicated by colors according to the legend and the mean expression (z-scores). (**e**) Visualization based on two dimensional coordinates from MDS in breast cancer (BRCA). Tregs and triple negative breast cancer (TNBC) samples for BRCA are indicated by colors according to the legend. Representative immunophenograms for selected patients are shown.

Figure 6 Immunophenoscores (IPS) and response to checkpoint blockade. (**a**) IPS and response to blockade with anti-CTLA-4 antibody (Data from van Allen et al.[14]). Shown are immunophenograms for individual patients. (**b**) Volcano plot for the enrichment and depletion of immune subsets in the tumor calculated based on the NES score from the GSEA. (**c**) IPSs for the cohort. (**d**) Receiver operating characteristics for the logistic regression classification (using leave-one-out cross validation) of responders and none-responders using for IPS, PD-1, PD-L1, or CTLA4 expression. (**e)** IPS and response to blockade with anti-PD-1 antibody (Data from Hugo et al.[15]). Shown are immunophenograms for individual patients. (**f**) Volcano plot for the enrichment and depletion of immune subsets in the tumor calculated based on the NES score from the GSEA. (**g**) IPSs for the cohort. (**h**) Receiver operating characteristics for the logistic regression classification (using leave-one-out cross validation) of responders and none-responders for IPS, PD-1, PD-L1, or CTLA4 expression

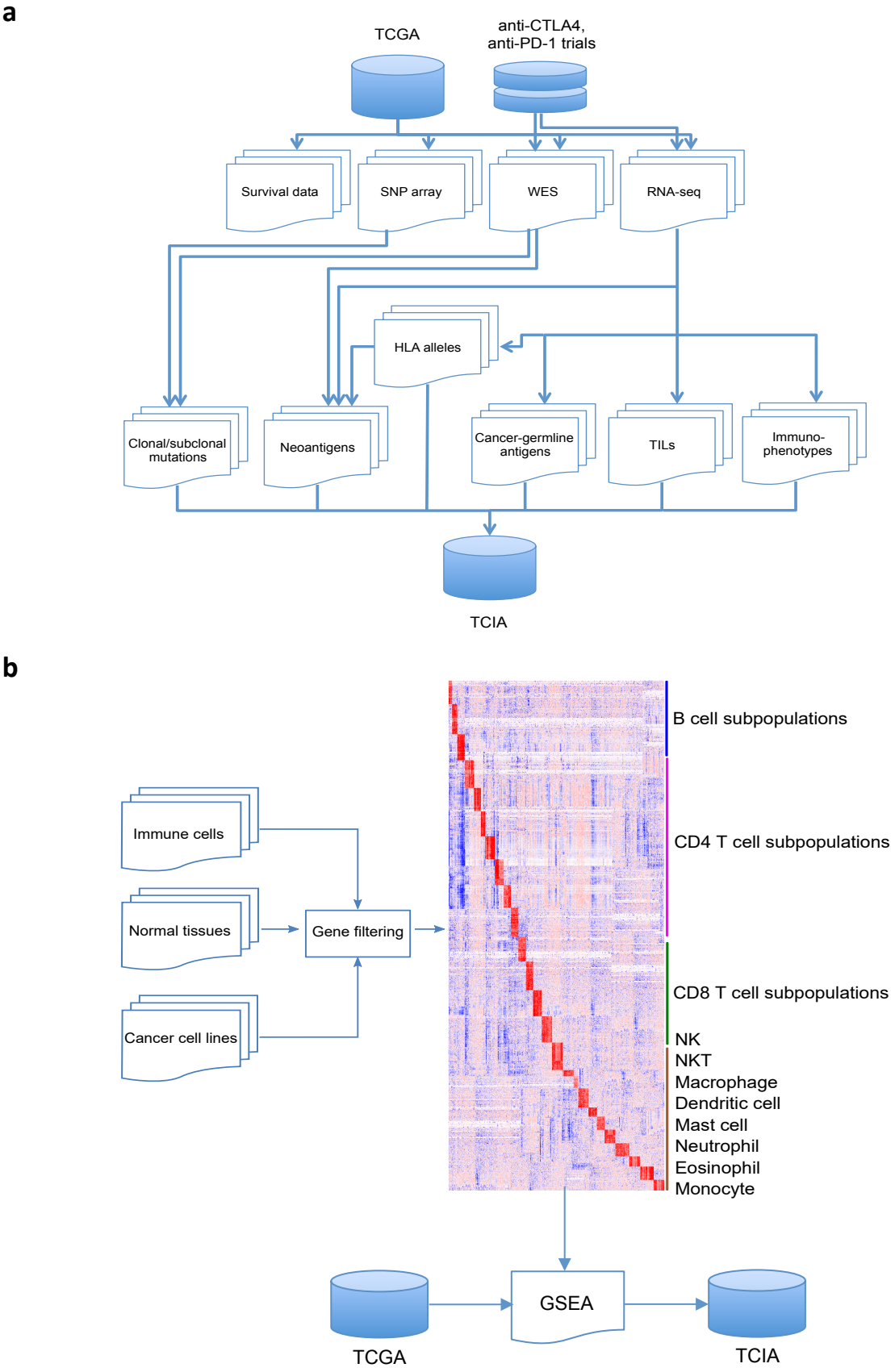
a
TCGA
anti-CTLA4,
anti-PD-1 trials
Survival data
SNP array
WES
RNA-seq
HLA alleles
Clonal/subclonal
mutations
Neoantigens
Cancer-germline
antigens
TILs
Immuno-
phenotypes
TCIA
b
Immune cells
Normal tissues
Cancer cell lines
Gene filtering
B cell subpopulations
CD4 T cell subpopulations
CD8 T cell subpopulations
NK
NKT
Macrophage
Dendritic cell
Mast cell
Neutrophil
Eosinophil
Monocyte
GSEA
TCGA
TCIA

Figure 1

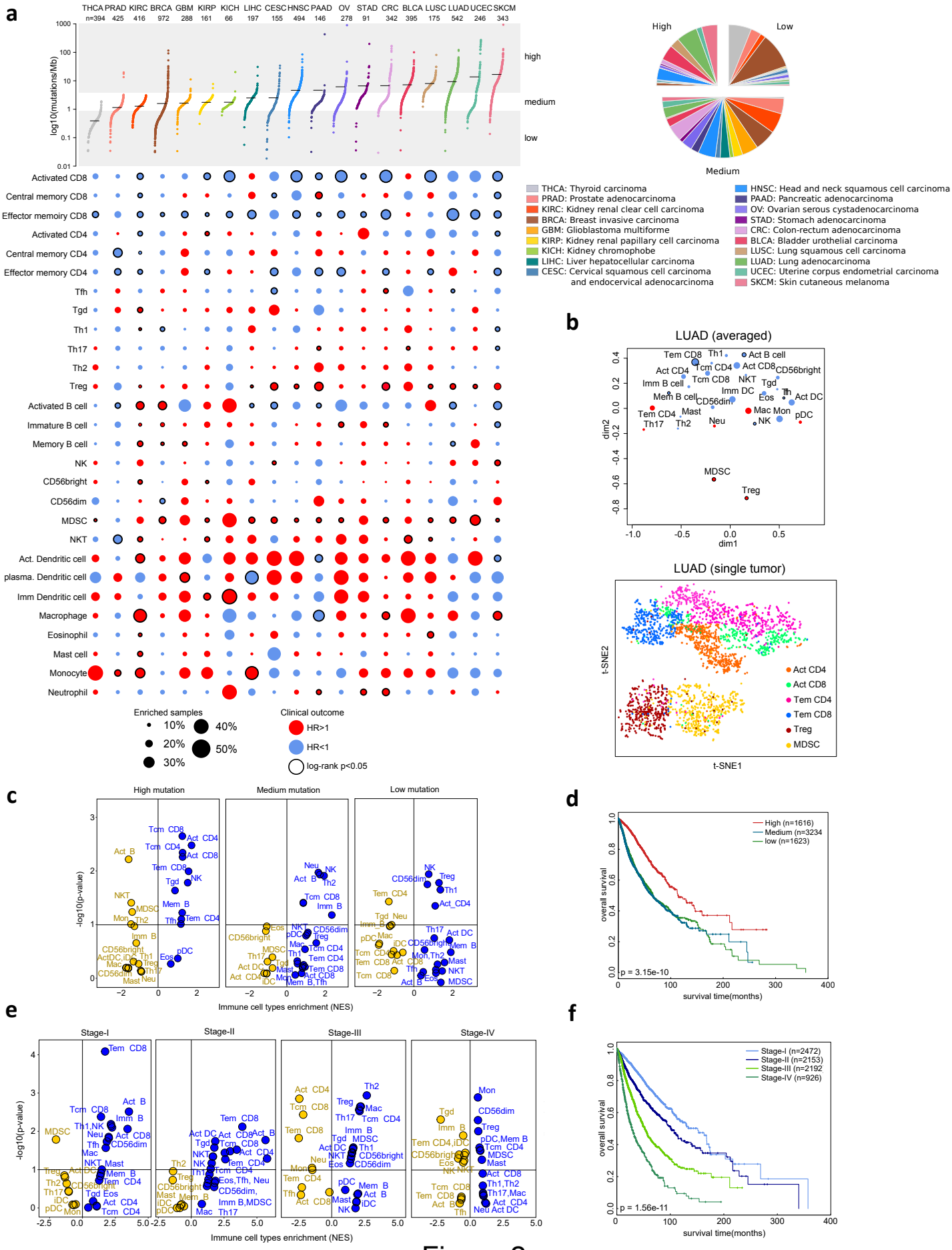
a
THCA PRAD KIRC BRCA GBM KIRP KICH LIHC CESC HNSC PAAD OV STAD CRC BLCA LUSC LUAD UCEC SKCM
n=394 425 416 972 288 161 66 197 155 494 146 278 91 342 395 175 542 246 343
log10(mutations/Mb)
high
medium
low
Activated CD8
Central memory CD8
Effector memoiry CD8
Activated CD4
Central memory CD4
Effector memory CD4
Tfh
Tgd
Th1
Th17
Th2
Treg
Activated B cell
Immature B cell
Memory B cell
NK
CD56bright
CD56dim
MDSC
NKT
Act. Dendritic cell
plasma. Dendritic cell
Imm Dendritic cell
Macrophage
Eosinophil
Mast cell
Monocyte
Neutrophil
Enriched samples
10%
20%
30%
40%
50%
Clinical outcome
HR>1
HR<1
log-rank p<0.05
High
Low
Medium
THCA: Thyroid carcinoma
PRAD: Prostate adenocarcinoma
KIRC: Kidney renal clear cell carcinoma
BRCA: Breast invasive carcinoma
GBM: Glioblastoma multiforme
KIRP: Kidney renal papillary cell carcinoma
KICH: Kidney chromophobe
LIHC: Liver hepatocellular carcinoma
CESC: Cervical squamous cell carcinoma and endocervical adenocarcinoma
HNSC: Head and neck squamous cell carcinoma
PAAD: Pancreatic adenocarcinoma
OV: Ovarian serous cystadenocarcinoma
STAD: Stomach adenocarcinoma
CRC: Colon-rectum adenocarcinoma
BLCA: Bladder urothelial carcinoma
LUSC: Lung squamous cell carcinoma
LUAD: Lung adenocarcinoma
UCEC: Uterine corpus endometrial carcinoma
SKCM: Skin cutaneous melanoma
b
LUAD (averaged)
dim1
dim2
LUAD (single tumor)
t-SNE1
t-SNE2
Act CD4
Act CD8
Tem CD4
Tem CD8
Treg
MDSC
c
High mutation
Medium mutation
Low mutation
-log10(p-value)
Immune cell types enrichment (NES)
d
High (n=1616)
Medium (n=3234)
low (n=1623)
overall survival
survival time(months)
p = 3.15e-10
e
Stage-I
Stage-II
Stage-III
Stage-IV
-log10(p-value)
Immune cell types enrichment (NES)
f
Stage-I (n=2472)
Stage-II (n=2153)
Stage-III (n=2192)
Stage-IV (n=926)
overall survival
survival time(months)
p = 1.56e-11

Figure 2

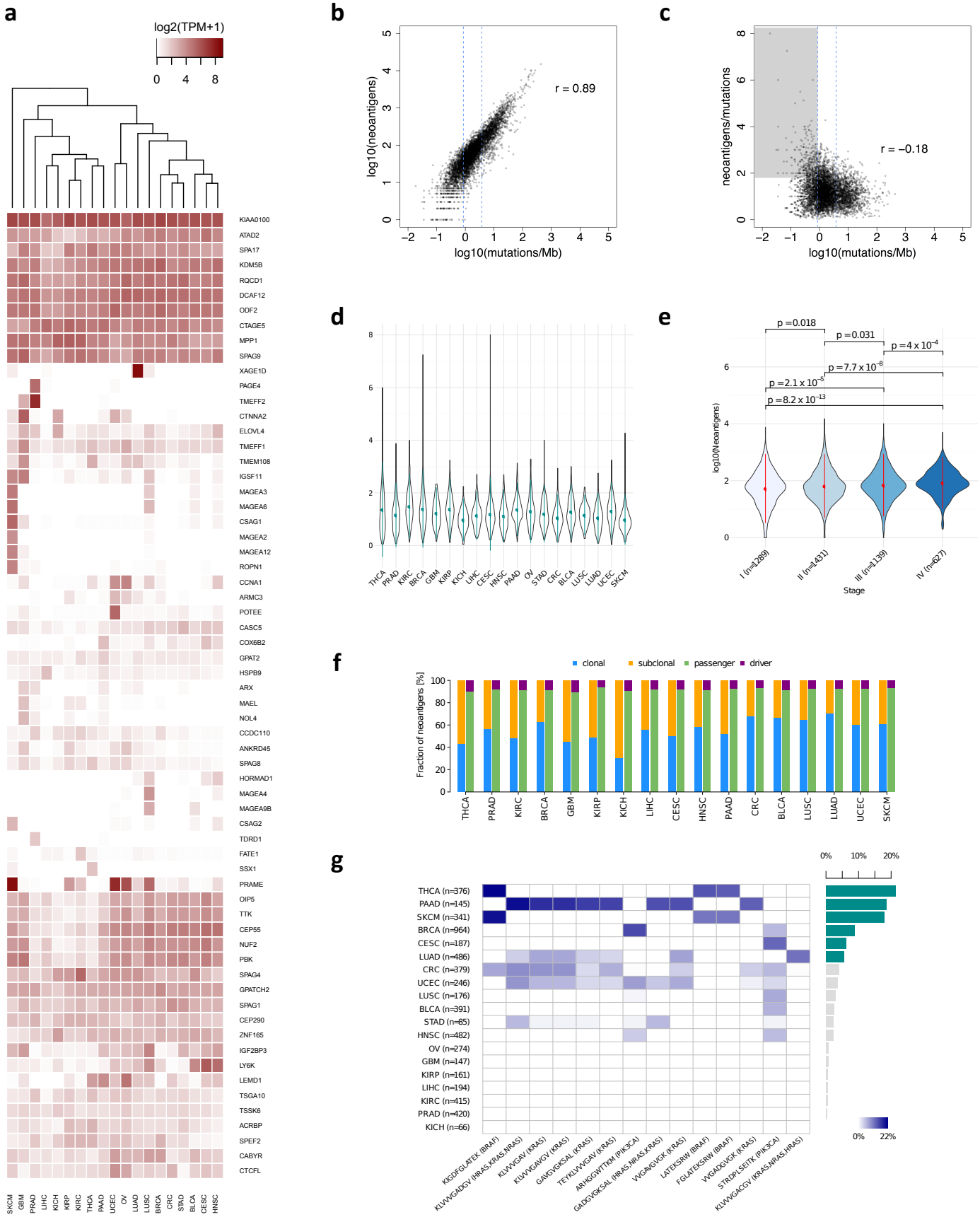
a
log2(TPM+1)
0 4 8
KIAA0100
ATAD2
SPA17
KDM5B
RQCD1
DCAF12
ODF2
CTAGE5
MPP1
SPAG9
XAGE1D
PAGE4
TMEFF2
CTNNA2
ELOVL4
TMEFF1
TMEM108
IGSF11
MAGEA3
MAGEA6
CSAG1
MAGEA2
MAGEA12
ROPN1
CCNA1
ARMC3
POTEE
CASC5
COX6B2
GPAT2
HSPB9
ARX
MAEL
NOL4
CCDC110
ANKRD45
SPAG8
HORMAD1
MAGEA4
MAGEA9B
CSAG2
TDRD1
FATE1
SSX1
PRAME
OIP5
TTK
CEP55
NUF2
PBK
SPAG4
GPATCH2
SPAG1
CEP290
ZNF165
IGF2BP3
LY6K
LEMD1
TSGA10
TSSK6
ACRBP
SPEF2
CABYR
CTCFL
SKCM GBM PRAD LIHC KICH KIRP KIRC THCA PAAD UCEC OV LUAD LUSC BRCA CRC STAD BLCA CESC HNSC
b
log10(neoantigens)
log10(mutations/Mb)
r = 0.89
c
neoantigens/mutations
log10(mutations/Mb)
r = −0.18
d
THCA PRAD KIRC BRCA GBM KIRP KICH LIHC CESC HNSC PAAD OV STAD CRC BLCA LUSC LUAD UCEC SKCM
e
p = 0.018
p = 0.031
p = 4 x 10^-4
p = 7.7 x 10^-8
p = 2.1 x 10^-5
p = 8.2 x 10^-13
log10(Neoantigens)
I (n=1289)
II (n=1431)
III (n=1139)
IV (n=627)
Stage
f
clonal
subclonal
passenger
driver
Fraction of neoantigens [%]
THCA PRAD KIRC BRCA GBM KIRP KICH LIHC CESC HNSC PAAD CRC BLCA LUSC LUAD UCEC SKCM
g
THCA (n=376)
PAAD (n=145)
SKCM (n=341)
BRCA (n=964)
CESC (n=187)
LUAD (n=486)
CRC (n=379)
UCEC (n=246)
LUSC (n=176)
BLCA (n=391)
STAD (n=85)
HNSC (n=482)
OV (n=274)
GBM (n=147)
KIRP (n=161)
LIHC (n=194)
KIRC (n=415)
PRAD (n=420)
KICH (n=66)
KIGDFGLATEK (BRAF)
KLVVVGADGV (HRAS,KRAS,NRAS)
KLVVVGAV (KRAS)
KLVVVGAVGV (KRAS)
GAVGVGKSAL (KRAS)
TEYKLVVVGAV (KRAS)
ARHGGWTTKM (PIK3CA)
GADGVGKSAL (HRAS,NRAS,KRAS)
VVGAVGVGK (KRAS)
LATEKSRW (BRAF)
FGLATEKSRW (BRAF)
VVGADGVGK (KRAS)
STRDPLSETTK (PIK3CA)
KLVVVGACGV (KRAS,NRAS,HRAS)
0% 10% 20%
0% 22%

Figure 3

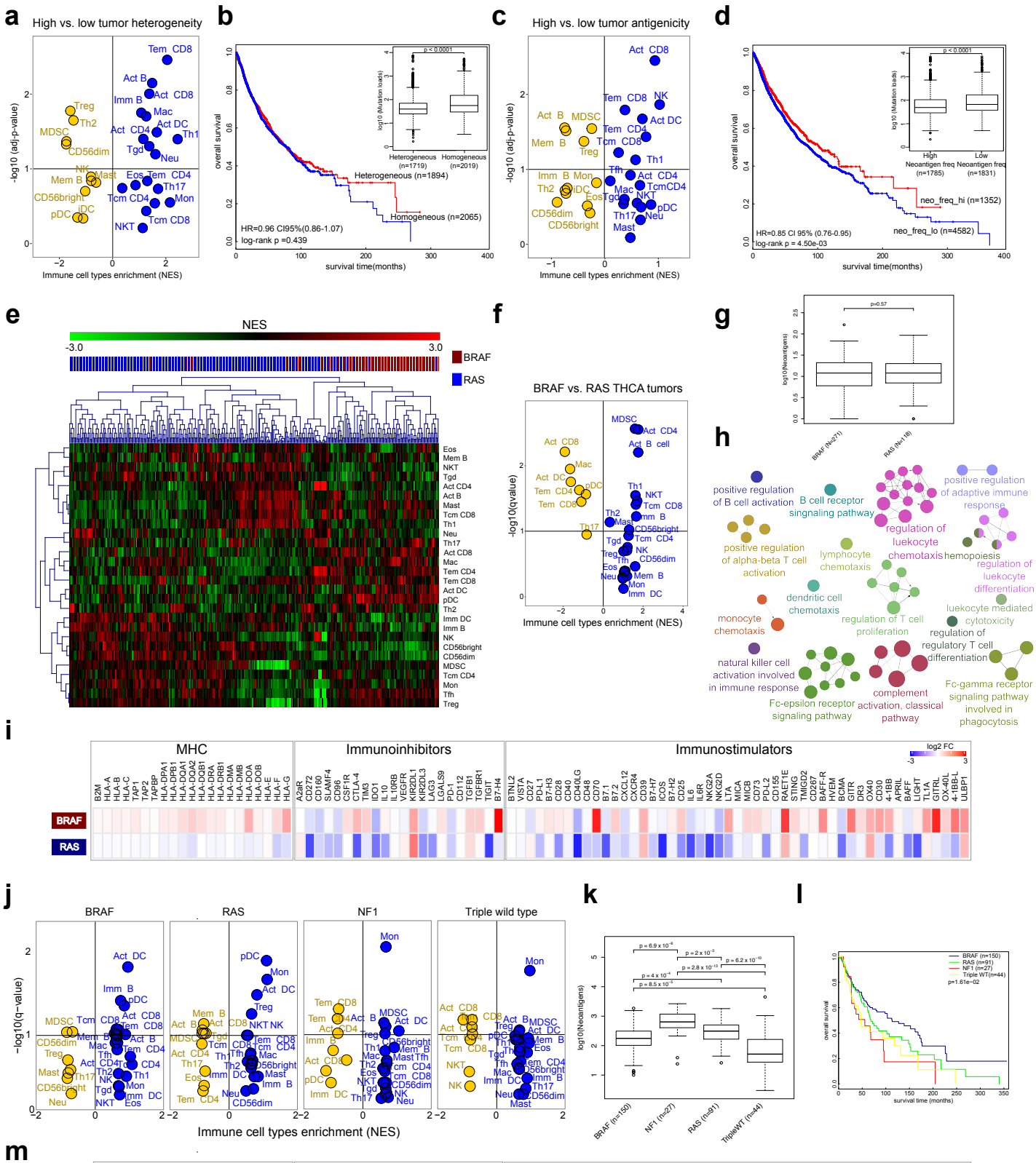
a
High vs. low tumor heterogeneity
-log10 (adj-p-value)
Immune cell types enrichment (NES)
b
overall survival
survival time(months)
Heterogeneous (n=1894)
Homogeneous (n=2065)
HR=0.96 CI95%(0.86-1.07)
log-rank p =0.439
p < 0.0001
log10 (Mutation loads)
Heterogeneous (n=1719)
Homogeneous (n=2019)
c
High vs. low tumor antigenicity
-log10 (adj-p-value)
Immune cell types enrichment (NES)
d
overall survival
survival time(months)
neo_freq_hi (n=1352)
neo_freq_lo (n=4582)
HR=0.85 CI 95% (0.76-0.95)
log-rank p = 4.50e-03
p < 0.0001
log10 (Mutation loads)
High (n=1785)
Low (n=1831)
e
NES
-3.0
3.0
BRAF
RAS
f
BRAF vs. RAS THCA tumors
-log10(qvalue)
Immune cell types enrichment (NES)
g
p=0.57
log10(Neoantigens)
BRAF (N=271)
RAS (N=118)
h
positive regulation of B cell activation
B cell receptor signaling pathway
regulation of leukocyte chemotaxis
positive regulation of adaptive immune response
positive regulation of alpha-beta T cell activation
lymphocyte chemotaxis
hemopoiesis
regulation of leukocyte differentiation
dendritic cell chemotaxis
monocyte chemotaxis
regulation of T cell proliferation
leukocyte mediated cytotoxicity
regulation of regulatory T cell differentiation
natural killer cell activation involved in immune response
Fc-epsilon receptor signaling pathway
complement activation, classical pathway
Fc-gamma receptor signaling pathway involved in phagocytosis
i
MHC
Immunoinhibitors
Immunostimulators
log2 FC
-3
3
BRAF
RAS
j
BRAF
RAS
NF1
Triple wild type
−log10(q-value)
Immune cell types enrichment (NES)
k
log10(Neoantigens)
BRAF (n=150)
NF1 (n=27)
RAS (n=91)
TripleWT (n=44)
l
overall survival
survival time (months)
BRAF (n=150)
RAS (n=91)
NF1 (n=27)
Triple WT(n=44)
p=1.61e-02
m
MHC
Immunoinhibitors
Immunostimulators
gene expression (z-score)
-3
3
BRAF
RAS
NF1
Triple WT

Figure 4

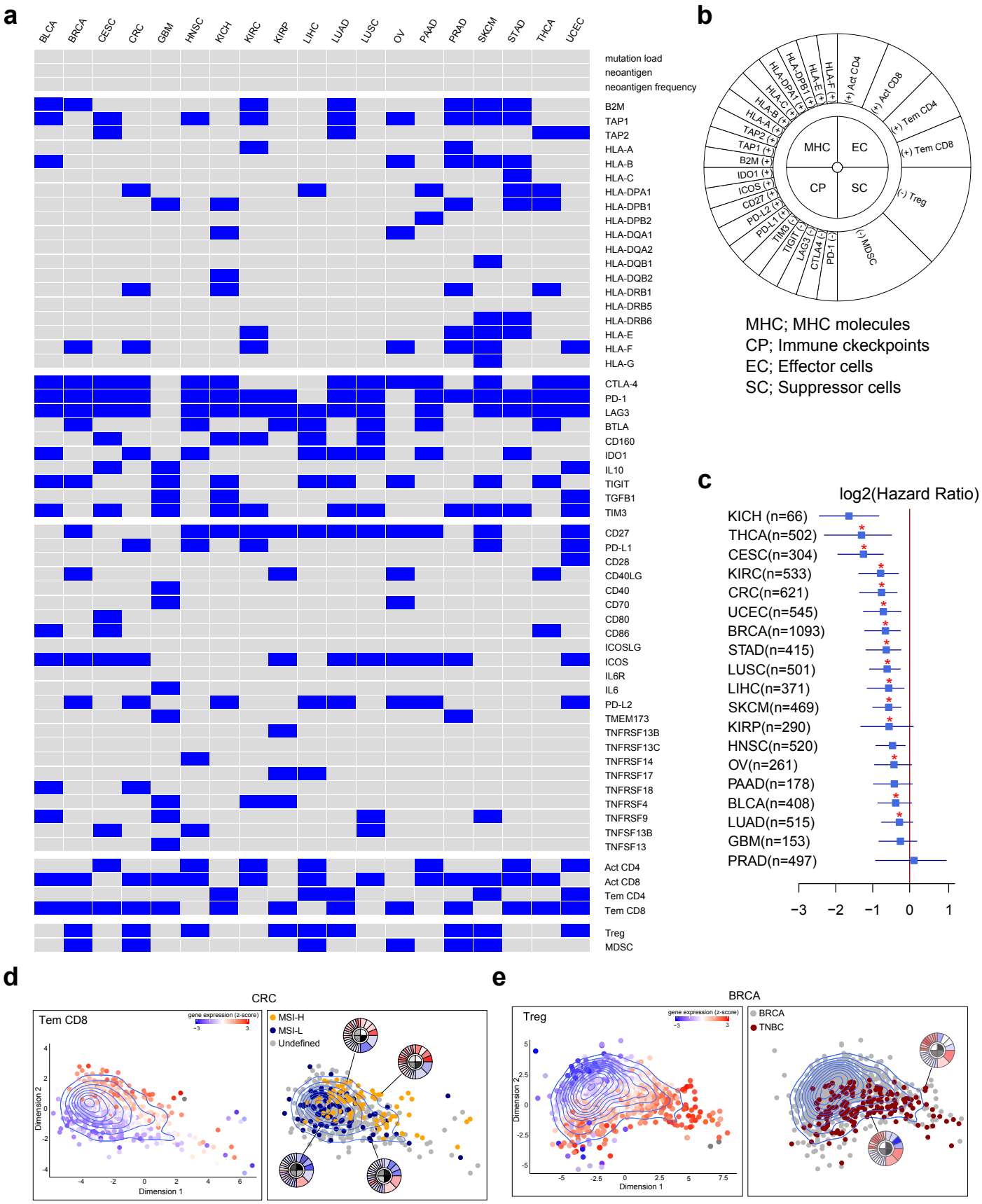
a
BLCA
BRCA
CESC
CRC
GBM
HNSC
KICH
KIRC
KIRP
LIHC
LUAD
LUSC
OV
PAAD
PRAD
SKCM
STAD
THCA
UCEC
mutation load
neoantigen
neoantigen frequency
B2M
TAP1
TAP2
HLA-A
HLA-B
HLA-C
HLA-DPA1
HLA-DPB1
HLA-DPB2
HLA-DQA1
HLA-DQA2
HLA-DQB1
HLA-DQB2
HLA-DRB1
HLA-DRB5
HLA-DRB6
HLA-E
HLA-F
HLA-G
CTLA-4
PD-1
LAG3
BTLA
CD160
IDO1
IL10
TIGIT
TGFB1
TIM3
CD27
PD-L1
CD28
CD40LG
CD40
CD70
CD80
CD86
ICOSLG
ICOS
IL6R
IL6
PD-L2
TMEM173
TNFRSF13B
TNFRSF13C
TNFRSF14
TNFRSF17
TNFRSF18
TNFRSF4
TNFRSF9
TNFSF13B
TNFSF13
Act CD4
Act CD8
Tem CD4
Tem CD8
Treg
MDSC
b
MHC
EC
CP
SC
(+) Act CD4
(+) Act CD8
(+) Tem CD4
(+) Tem CD8
(-) Treg
(-) MDSC
HLA-F (+)
HLA-E (+)
HLA-DPB1 (+)
HLA-DPA1 (+)
HLA-C (+)
HLA-B (+)
HLA-A (+)
TAP2 (+)
TAP1 (+)
B2M (+)
IDO1 (+)
ICOS (+)
CD27 (+)
PD-L2 (+)
PD-L1 (+)
TIM3 (-)
TIGIT (-)
LAG3 (-)
CTLA4 (-)
PD-1 (-)
MHC; MHC molecules
CP; Immune ckeckpoints
EC; Effector cells
SC; Suppressor cells
c
log2(Hazard Ratio)
KICH (n=66)
THCA(n=502)
CESC(n=304)
KIRC(n=533)
CRC(n=621)
UCEC(n=545)
BRCA(n=1093)
STAD(n=415)
LUSC(n=501)
LIHC(n=371)
SKCM(n=469)
KIRP(n=290)
HNSC(n=520)
OV(n=261)
PAAD(n=178)
BLCA(n=408)
LUAD(n=515)
GBM(n=153)
PRAD(n=497)
−3
−2
−1
0
1
d
CRC
Tem CD8
gene expression (z-score)
Dimension 1
Dimension 2
MSI-H
MSI-L
Undefined
e
BRCA
Treg
gene expression (z-score)
Dimension 1
Dimension 2
BRCA
TNBC

Figure 5

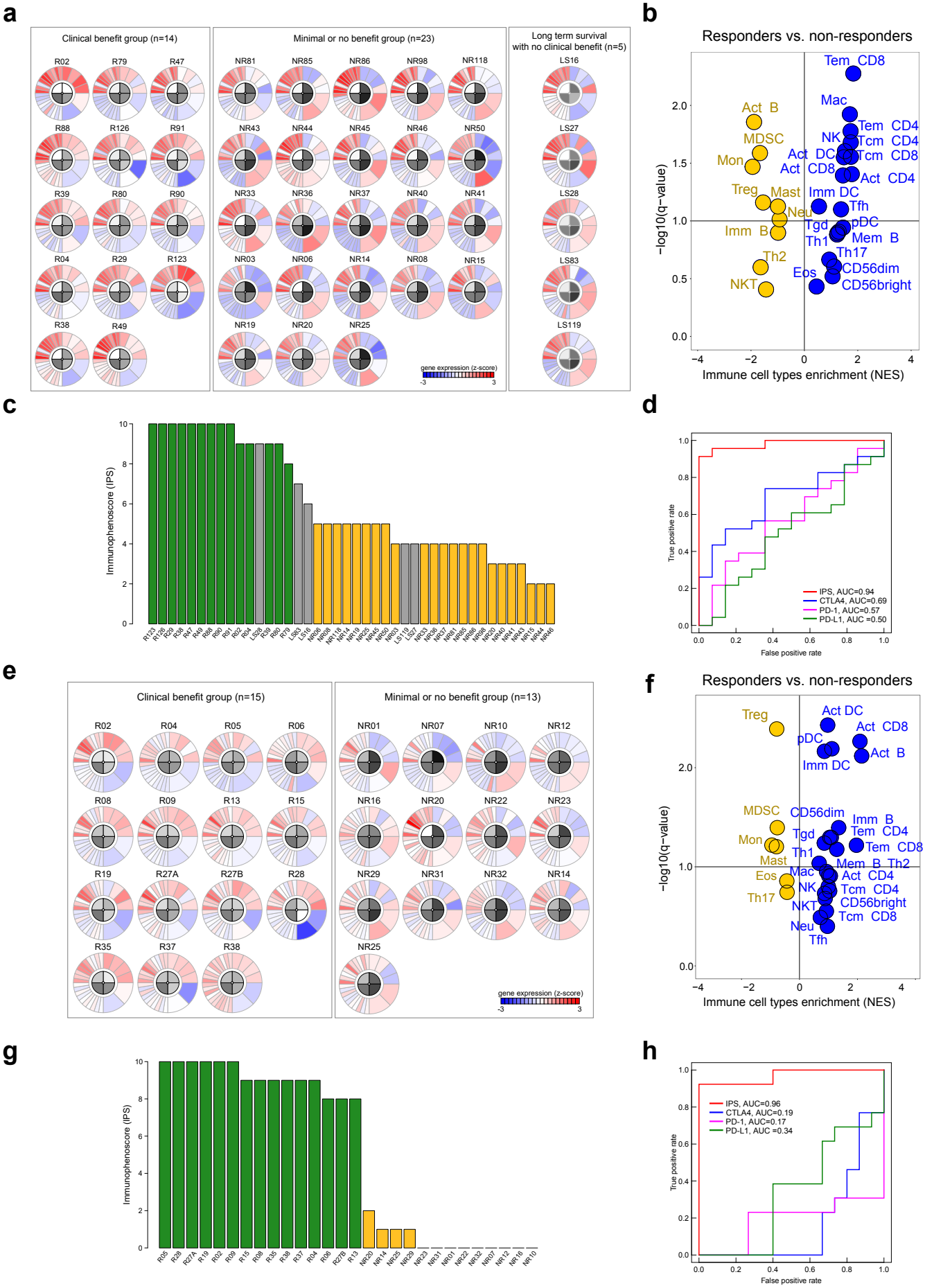
a
Clinical benefit group (n=14)
Minimal or no benefit group (n=23)
Long term survival with no clinical benefit (n=5)
gene expression (z-score)
b
Responders vs. non-responders
Immune cell types enrichment (NES)
−log10(q−value)
c
Immunophenoscore (IPS)
d
True positive rate
False positive rate
IPS, AUC=0.94
CTLA4, AUC=0.69
PD-1, AUC=0.57
PD-L1, AUC =0.50
e
Clinical benefit group (n=15)
Minimal or no benefit group (n=13)
gene expression (z-score)
f
Responders vs. non-responders
Immune cell types enrichment (NES)
−log10(q−value)
g
Immunophenoscore (IPS)
h
True positive rate
False positive rate
IPS, AUC=0.96
CTLA4, AUC=0.19
PD-1, AUC=0.17
PD-L1, AUC =0.34

Figure 6