Published OnlineFirst September 28, 2018; DOI: 10.1158/2326-6066.CIR-18-0342

Research Article

Cancer Immunology Research

Immune Cell Gene Signatures for Profiling the Microenvironment of Solid Tumors

Ajit J. Nirmal[1], Tim Regan[1], Barbara B. Shih[1], David A. Hume[1,3], Andrew H. Sims[2], and Tom C. Freeman[1]

Abstract

The immune composition of the tumor microenvironment regulates processes including angiogenesis, metastasis, and the response to drugs or immunotherapy. To facilitate the characterization of the immune component of tumors from transcriptomics data, a number of immune cell transcriptome signatures have been reported that are made up of lists of marker genes indicative of the presence a given immune cell population. The majority of these gene signatures have been defined through analysis of isolated blood cells. However, blood cells do not reflect the differentiation or activation state of similar cells within tissues, including tumors, and consequently markers derived from blood cells do not necessarily transfer well to tissues. To address this issue, we generated a set of immune gene signatures derived directly from tissue transcriptomics data using a network-based deconvolution approach. We define markers for seven immune cell types, collectively named *ImSig*, and demonstrate how these markers can be used for the quantitative estimation of the immune cell content of tumor and nontumor tissue samples. The utility of *ImSig* is demonstrated through the stratification of melanoma patients into subgroups of prognostic significance and the identification of immune cells with the use of single-cell RNA-sequencing data derived from tumors. Use of *ImSig* is facilitated by an R package (imsig). *Cancer Immunol Res; 6(11); 1388–400. ©2018 AACR.*

Introduction

Modulating the activity of the immune component of the tumor microenvironment holds potential in the treatment of cancer. Checkpoint inhibitors, particularly anti-PD1 and CTLA4, have advanced therapeutic options in the past decade producing benefit for some patients (1). However, multiple factors within the tumor microenvironment, including the immune infiltrate prior to treatment (2), influence the response to immunotherapy. IHC and flow cytometry are often used to study the immune status of tumors. However, the former analyses are limited to small areas of tissue and a few markers, and the latter requires tissue disaggregation, which may not always be practical. To overcome these limitations, computational methods have been developed to estimate the immune content of blood and tissue samples from transcriptomic data (3). Two approaches can be used to infer the relative proportion of cell types from transcriptomic data: (i) fitting reference gene-expression profiles from sorted cells to the data in question (4–7) and (ii) following cell type–specific genes to indicate the presence of certain cell populations (8–11). Both approaches rely on sets of gene markers (gene signatures); however, in the first case, the gene signature is not necessarily cell type–specific, and supervised learning algorithms are needed to distinguish between cell types.

A number of computational frameworks leveraging these approaches have been described to estimate the contribution of different immune cell types to the tissue transcriptome (5, 10–14). Across these studies, the range of immune cell types that each method detects varies. For instance, collectively, the published studies report gene signatures for 22 different T-cell subtypes, but with many "marker genes" expressed by nonimmune cell types and others used interchangeably to define different T-cell subtypes. Another shortfall is that these signatures are based on gene-expression data gathered from primary blood-derived cells generally collected from healthy donors. When the expression profiles of the same immune cell either from blood (peripheral blood mononuclear cell) or from tissue can differ (15), the predictive value of signatures is compromised (16).

Genes that contribute to a common biological process or define a given cell type are frequently coregulated and coexpressed, giving rise to expression modules (17, 18). We have previously validated gene correlation network (GCN) analysis of gene-expression data sets from human (including human cancers), mouse, pig, and sheep, as a means to define such expression modules (19–21). Here, we have analyzed human tissue transcriptomic data to identify coexpressed marker genes representing seven immune cell types and three cellular pathways present in data from many tissues. We have named this set of signatures *ImSig*. We demonstrate the advantages of *ImSig* over other reported signatures derived from the

[1]The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Edinburgh, United Kingdom. [2]Applied Bioinformatics of Cancer, Edinburgh Cancer Research Centre, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, United Kingdom. [3]Mater Research Institute, University of Queensland, Queensland, Australia.

Note: Supplementary data for this article are available at Cancer Immunology Research Online (http://cancerimmunolres.aacrjournals.org/).

Corresponding Author: Tom C. Freeman, University of Edinburgh, Edinburgh EH25 9RG, UK. Phone: 44-131-651-9203; Fax: 44-131-651-9105; E-mail: tfreeman@roslin.ed.ac.uk

doi: 10.1158/2326-6066.CIR-18-0342

Published OnlineFirst September 28, 2018; DOI: 10.1158/2326-6066.CIR-18-0342

comparison of isolated blood cells and characterize *ImSig's* utility for analyzing the immune microenvironment of tumors.

Materials and Methods

Derivation of *ImSig*

Eight publicly available expression data sets derived from human tissue were extracted from the Gene-Expression Omnibus (GEO) database (ref. 22; GSE11318, GSE50614, GSE75214, GSE38832, GSE23705, GSE24383, GSE58812, and GSE65904). Prerequisites were that the unprocessed data files were available, the data set included a variety of normal and diseased samples, represented a variety of array platforms, and contained >20 samples (median size, 114 samples). The data sets were chosen to include the variety of immune cell types and differentiation states. Data sets were subjected to standard processing (i.e., conversion of raw platform-specific files into expression matrix and normalization) with the help of R packages such as "oligo" (23) and "lumi" (24) for Affymetrix and Illumina data, respectively. The signal intensities were normalized using the robust multiarray average. The expression values for genes with multiple probes were reduced to one probe per gene by choosing the probe with maximum intensity across samples.

The resultant expression matrix was loaded into the network analysis tool Graphia Professional (Kajeka Ltd.), previously known as BioLayout *Express*^{3D} (25, 26). Within the tool, a Pearson correlation matrix was generated, i.e., an all versus all comparison of expression profiles with genes exhibiting a similar expression pattern across the samples scoring highly (with a maximum correlation value equal to 1). A GCN was then generated using a correlation threshold value so as to include approximately 10,000 genes in the analysis for each data set. In the context of a GCN, nodes represent genes/transcripts and edges, correlations above the threshold. The optimal correlation threshold is data set–specific, as generally smaller data sets exhibit a higher overall correlation and all threshold values used also minimized chance associations. The GCN for each data set was then clustered using the Markov clustering (MCL) algorithm (27), an algorithm analyzes a graph's structure to define gene clusters of nodes, in this case coexpression modules. Clusters were manually annotated based on domain knowledge with the help of Gene Ontology (GO) and Reactome pathway enrichment analyses (28, 29). Gene modules representing immune cell types and biological processes were identified for each of the eight data sets. The genes within the modules were consolidated into a list of genes for seven immune cell types and three biological processes. In order to identify the core set of genes that represents each cell type or process, these genes were further refined using eight independent validation data sets (GSE9891, GSE14580, GSE38832, GSE14951, GSE15773, GSE7305, GSE22619, and GSE52171) by the following procedure: Robust signatures were identified by excluding genes that were poorly coexpressed using an unbiased approach. Each data set was loaded into Graphia (r values were again selected so as to include approximately 10,000 genes in the analysis) and clustered using the MCL algorithm. To model the contribution of noise by random genes within signatures, 0 to 100% of genes within every MCL cluster were replaced with random genes (using the R function "sample") in a stepwise manner, in 2% increments. For each of these replacements, the resultant median correlation of every cluster was noted. The combined data points were fitted to a sigmoidal curve using the nonlinear least squares method. On the basis of this model, we estimated the number of genes that might contribute to noise within the signatures and should be filtered out. To facilitate such estimation, the R package "investr" was used. For example, based on the median correlation of signature genes, if the model suggested 30% of genes represented noise, then 30% of genes exhibiting the poorest median correlation were discarded. This process was repeated for each signature across the eight validation data sets. The set of genes that survived the filtration process were defined as *ImSig*. Our approach sought to identify the genes most correlated across data sets to arrive at the final list of genes for the individual *ImSig* signatures. TopGo was used to identify the five most enriched GO Biological Process (GO_BP) terms associated with each gene set (28), and P values were generated using the Fisher exact test.

Comparison of *ImSig* with other published signatures

Seven published immune signatures were taken from the literature (5, 8, 10–14). To visualize the concordance between the immune genes defined by the different studies, a chord diagram was built using the circlize package (30) in R. We used only genes reported as markers of immune cells, and signatures of nonimmune cells such as fibroblast or endothelial cells were omitted from this analysis. Due to the great variety of T cell–subtype signatures reported, these were further explored to identify how genes were used to define the different subtypes. Genes that were present in two or more studies and ascribed to a T cell or one of its subtypes were identified. Using these genes, a graph was constructed using Cytoscape (31) and visualized with a circular layout. The size of nodes representing individual signatures was adjusted according to the number of connections each signature had with others. A Jaccard similarity index was also calculated between all signatures. The LM22 signature (5) did not provide an absolute signature, that is, the same genes may represent multiple cell types and only a subset of genes that were unique to cell types were used for our analysis. For visualization of the results, genes pertaining to cell subsets [regulatory T cell (Treg) and Th1] were pooled to represent the parent population (T cells) and the Jaccard similarity index was recalculated.

Comparative analysis of gene signatures in the context of a tissue data set

The median correlation of the signature genes from the same seven published immune signatures (5, 8, 10–14) was calculated within the context of a trachoma data set (GSE20436; ref. 32). The transcriptomics data set was generated from swabs taken from the eyes of children with symptoms of trachoma or controls and contained samples from three patient subgroups; 20 controls with normal conjunctivas; 20 individuals with clinical signs of trachoma but that tested negative for the bacteria *C. trachomatis* (these patients may have been in the resolution stage); and 20 individuals with symptoms and active infections. This data set was chosen due to the immune cell infiltration associated with this disease. The presence of all immune cell populations was confirmed by *ImSig*. To facilitate comparison with *ImSig*, genes pertaining to cell subsets were pooled to represent the parent cell population. In addition, median correlations of nonpooled signatures (i.e., marker sets representing subpopulations of cells) were also analyzed.

Published OnlineFirst September 28, 2018; DOI: 10.1158/2326-6066.CIR-18-0342

To assess the ability of *ImSig* to define known clinical differences between patient subgroups and to illustrate the explorative power of network-based analysis, we used the trachoma data set described above. In order to estimate the relative abundance of immune cells across patient groups, the average expression of the *ImSig* signature genes was computed. A two-tailed, unequal variance *t* test was conducted between groups to obtain *P* values. To explore the immune environment and extrapolate immune cell subsets, a GCN ($r > 0.7$) was visualized in Graphia. By visual inspection of the network graph, immunologically relevant genes (subtype/differentiation-specific) were identified in the vicinity of the *ImSig* modules, and their average expression profile across patient groups was plotted.

To validate *ImSig* in the context of tumor-derived samples, transcriptomic data from single-cell suspensions from lymph nodes of four patients with metastatic melanoma were analyzed (GSE93722). Here, the relative proportion of immune cells, $CD4^+$ T cells, $CD8^+$ T cells, B cells, natural killer (NK) cells, had been measured with flow cytometry. To facilitate direct comparison, proportions of $CD4^+$ and $CD8^+$ T cells were summed to estimate total T-cell content. The average expression of *ImSig* genes was calculated to determine the relative abundance of immune cells in each patient. Predicted and observed abundances were normalized between 0 and 1 to facilitate comparisons. This analysis also served to validate the applicability of *ImSig* to RNA-seq data.

Pan-cancer analysis of tumor data (TCGA)

Prenormalized (level 3 data: the calculated expression signal of a gene per sample) transcriptomic data from 12 cancers were downloaded from The Cancer Genome Atlas (TCGA) database. For each cancer type, the patients were ordered based on the average expression of the individual *ImSig* signatures and split into two groups based on the median expression value of the signature genes. In cases such as brain lower grade glioma (LGG), kidney renal clear cell carcinoma (KIRC), and uterine corpus endometrial carcinoma (UCEC), B-cell signature genes were not coexpressed, indicating the absence or low abundance of these cells, and so were not included in the survival analysis. A univariate Cox-proportional hazard ratio (HR) analysis was performed for the rest using the R package "survcomp" (33). *P* values are based on the log-rank test.

Molecular subtyping (patient stratification) of melanoma

RNA-seq data for human skin cutaneous melanoma (SKCM) were downloaded from the TCGA data portal. Using the expression data of *ImSig* genes, a sample-to-sample correlation plot ($r > 0.85$) was generated. MCL clustering (inflation value 1.7) of the sample-to-sample correlation plot grouped the patients into five clusters. These groupings were mapped onto the GCN to study the differences in expression patterns of immune cells between groups. A univariate Cox-proportional analysis was also performed using the R package survcomp (33) between the groups in various combinations. The *P* value was calculated using the log-rank test.

An independent melanoma data set GSE65904 (34) was used for validation. The data set was produced on the Illumina HumanHT-12 V4.0 microarrays and composed of samples from 214 melanoma patients. Samples that did not contain necessary information such as disease-specific survival, gender, and sample type were removed. After processing and normalization using the "lumi" package (24) in R, samples that were not present in the network graph ($r \geq 0.8$) were also removed, and the remaining samples (210) were processed as described above for the TCGA data set.

Processing and analysis of single-cell RNA-seq data

Single-cell transcriptomics data for melanoma (35) and head and neck squamous cell cancer (HNSCC; ref. 36) were downloaded from The Broad Institute single-cell portal (https://portals.broadinstitute.org/single_cell). As computation of the relative abundance of cell types is based on the average expression of *ImSig* genes, missing values in single-cell data can affect results. Therefore, to compensate for dropouts, a diffusion-based method was used to impute missing values (37).

To validate the cell-type specificity of *ImSig*, the average expression of B cell, T cell, NK cell, and macrophage signature genes was calculated from the melanoma cell data set and compared with the average expression of the other immune-related *ImSig* genes. To evaluate the concordance between estimated abundance and measured number of cells, the averages for expression of signature genes for 10 patients were computed (estimated abundance). Correlation between estimated abundance and measured number of cells was calculated, and *P* values were attained by building a linear regression model. To illustrate the concordance of relative proportions, both the estimated abundance and measured number of cells were scaled using the formula $[x - \min(x)/\max(x) - \min(x)$, where x is the cell abundance value], and results were plotted as a stacked bar plot normalized to 100%.

In order to predict immune cell types in the HNSCC data set using the SVM-based algorithm Cibersort, a reference matrix (*ImSig* as features) was generated using the melanoma single-cell data. The algorithm was run with the generated reference matrix and HNSCC single-cell data by uploading it to the Cibersort's web portal (https://cibersort.stanford.edu). The portal computes a score for B cell, T cell, and macrophage for each sample and an associated *P* value. *P* values of <0.05 and a score of >0.75 (upper quartile) were set as defining correct predictions, that is, a T-cell score of >0.75 in a T cell with a *P* value of <0.05 was judged as a correct prediction.

R implementation and availability of *ImSig*

We implemented *ImSig* as an R package called "imsig." Users should call the "imsig" function, which takes a normalized gene-expression matrix made up of HUGO symbols in rows and samples in columns as its first argument, and a correlation threshold (r) as its second argument. Users can also generate a network graphic of *ImSig* genes and perform survival analysis using the package. A tutorial is available at https://github.com/ajitjohnson/imsig.

This package is available at CRAN (https://cran.r-project.org/web/packages/imsig/).

Results

Derivation of *ImSig*

Using a network-based approach, we identified a set of co-expressed gene modules associated with human tissue immune cell populations and frequently observed biological processes, from eight independent tissue transcriptomics data sets. An illustrative example of a GCN is shown in Fig. 1A. These initial

Published OnlineFirst September 28, 2018; DOI: 10.1158/2326-6066.CIR-18-0342

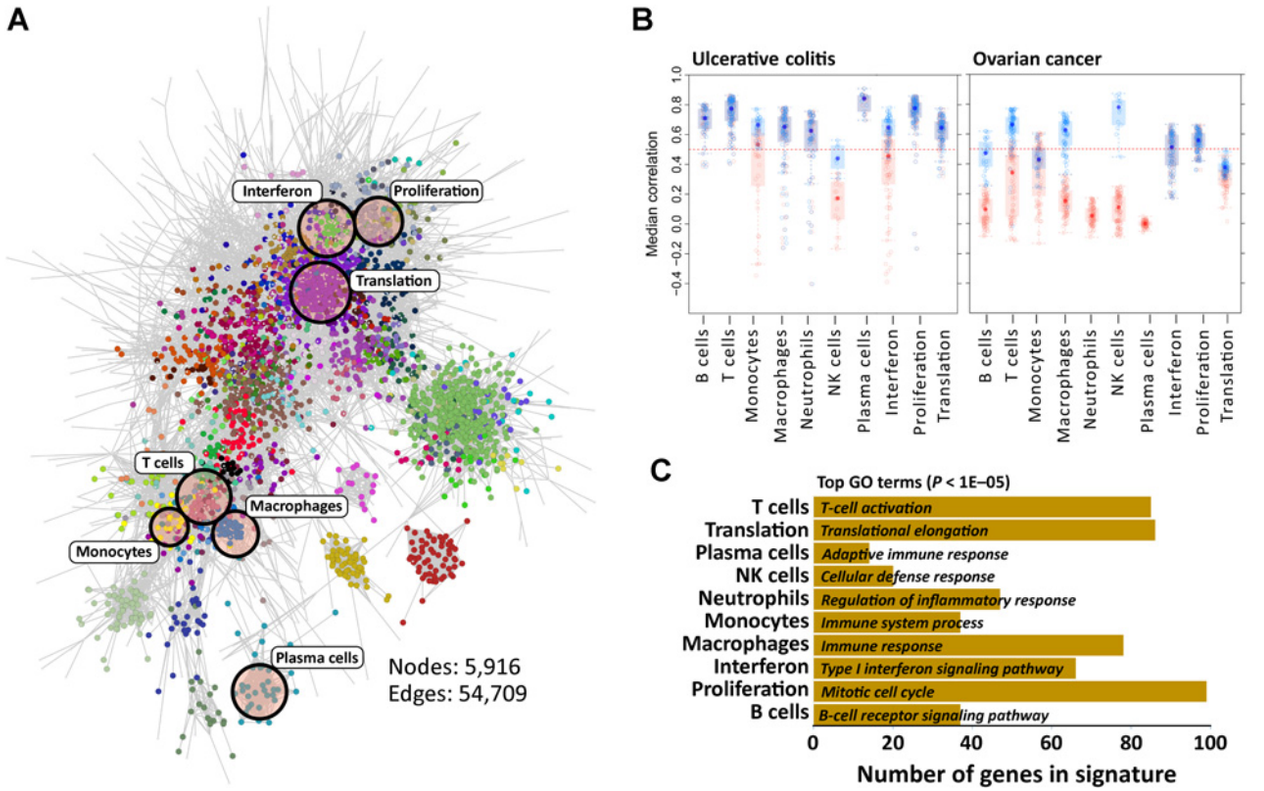

Figure 1.

Derivation of *ImSig*. **A,** An example of a correlation network generated from a tissue data set where nodes represent unique genes and edges represent correlations between genes above a defined threshold. Groups of nodes sharing the same color represent gene modules (obtained by MCL clustering), those highlighted being associated with a given immune cell type or biological process. **B,** Example plots from the approach used to refine the gene signatures. Blue points represent genes that were kept, i.e., they were highly correlated with other genes in the preliminary signature. Red represents genes that were discarded. This approach was applied to eight tissue data sets (only two are shown here). The most robustly coexpressed genes across the data sets were used to define *ImSig*. **C,** Bar plot depicting the number of genes within each marker gene signature comprising *ImSig* and the top GO enrichment term for each signature.

gene signatures were refined and validated by testing for coexpression of the genes associated with each signature across an additional eight independent data sets (Fig. 1B). The result was 569 marker genes representative of seven immune populations [B cells (37 genes), plasma cells (14 genes), monocytes (37 genes), macrophages (78 genes), neutrophils (47 genes), NK cells (20 genes), T cells (85 genes)] and three biological processes [Interferon response (66 genes), translation (86 genes), proliferation (99 genes)]. We named this set of genes collectively *ImSig* (Tables 1 and 2; Supplementary Table S1). The data-driven definition of each immune signature is internally validated by association of known markers with the specific gene signatures, e.g., *CD3D* and *CD3E* (T cells), *CD19*, *CD22*, and *CD79* (B cells), *CD14* (monocytes), *CD68* and *CD163* (macrophages), KIR family (NK cells) and immunoglobulin family members (plasma cells). Furthermore, GO enrichment analysis of the gene signatures and data from the published literature supported the association of markers with relevant cell types and processes. The top five significant enrichment terms for all signatures are listed in Supplementary Table S2 and the top significant term is given in Fig. 1C. Unlike other published immune gene signatures, our gene signatures do not distinguish immune cell subtypes, such as subpopulations of T cells or activation states of macrophages. We found no support for distinct modules of coexpressed markers describing T cell or macrophage subpopulations. Indeed, analysis of isolated human macrophages responding to different stimuli did not support the existence of distinct activation states of macrophages but rather indicated a continuum of states depending on the stimulus (38). Where present, "activation-specific" transcripts, such as receptors, cytokines, or transcription factors, tend to form part of the overall cell expression module. By inference, coexpression of a gene with a particular cell type–specific signature in a particular data set indicated that the gene is likely expressed by those cells or at least a subpopulation of them.

Comparison between *ImSig* and published immune signatures

The gene content of seven published immune signatures, all derived from comparisons of isolated blood cells (5, 8, 10–14), was compiled and compared. We excluded signatures for non-immune cell types, e.g., endothelial cells, fibroblasts etc. When *ImSig* was added to the list, the list contained a total of 3,658 genes (Supplementary Table S3). To compare these gene signatures, we calculated a Jaccard similarity index

Published OnlineFirst September 28, 2018; DOI: 10.1158/2326-6066.CIR-18-0342

Table 1. Table of *ImSig* genes (immune signatures)

Signature	Genes
B cells	*AFF3, BANK1, BLK, BTLA, CCR6, CD180, CD19, CD22, CD37, CD72, CD79A, CD79B, CR2, EBF1, FAM129C, FCRL1, FCRL2, FCRL3, FCRL5, FCRLA, HLA-DOB, IGHV5-78, KIAA0125, LINC00926, LOC100507616, LY9, MS4A1, P2RX5, PAX5, PNOC, POU2F2, S1PR4, SNX22, STAP1, TCL1A, TLR10, VPREB3*
T cells	*AMICA1, APBB1IP, ARHGAP15, ARHGAP25, ARHGAP9, BIN2, BTK, C1orf162, CCL19, CCR7, CD2, CD27, CD28, CD3D, CD3E, CD3G, CD48, CD52, CD6, CD8A, CD96, CORO1A, CRTAM, CXCL9, CXCR6, CYTIP, DOCK10, DOCK2, DOCK8, DPEP2, EVI2A, EVI2B, FAM26F, FLI1, FYB, FYN, GAB3, GIMAP2, GIMAP4, GIMAP5, GIMAP6, GIMAP7, GMFG, GPR171, GPR18, GZMK, HCST, HMHA1, HVCN1, ICOS, IL10RA, IL16, IL23A, IL7R, ITGAL, ITK, KLHL6, KLRB1, LCP1, LY86, NCF1B, NLRC3, PARVG, PRKCH, PSTPIP1, PTPRCAP, PVRIG, RASSF5, RCSD1, RGS18, RHOH, SASH3, SH2D1A, SIRPG, SLA, SP140, TARP, TBC1D10C, TNFRSF9, TRAC, TRAF3IP3, TRAT1, TRGC2, TRGV9, UBASH3A*
Macrophages	*ADAMDEC1, ADORA3, AOAH, ARRB2, ATP8B4, BCL2A1, C1orf54, C1QA, C1QB, C2, C3AR1, C5AR1, CCR1, CCRL2, CD163, CD300A, CD4, CD68, CD74, CD86, CECR1, CLEC7A, CMKLR1, CSF1R, CTSB, CTSS, CYBB, CYTH4, DPYD, EMR2, FCER1G, FCGR1A, FCGR1B, FCGR2A, FCGR3B, FPR3, GPNMB, HK3, HLA-DRB6, IFI30, IGSF6, ITGAM, ITGAX, ITGB2, LAIR1, LAPTM5, LILRB4, LIPA, LY96, MAN2B1, MFSD1, MNDA, MS4A4A, MS4A7, MSR1, MYO1F, NCKAP1L, NPL, NR1H3, PLA2G7, PLEKHO2, SCPEP1, SLAMF8, SLC15A3, SLC31A2, SLCO2B1, SNX10, SPI1, TBXAS1, TLR8, TMEM140, TNFAIP2, TNFRSF1B, TNFSF13B, TRPV2, TYMP, TYROBP, VSIG4*
Monocytes	*AGTRAP, AIF1, C10orf54, CD14, CD300LF, CD33, CD93, CTSD, EMILIN2, FCN1, FES, FGR, GNS, GRN, HCK, HMOX1, KIAA0930, LILRA6, LILRB2, LILRB3, LRRC25, LST1, NFAM1, NOTCH2, PILRA, PLXDC2, PRAM1, PSAP, PYCARD, RHOG, SERPINA1, SLC7A7, TGFBI, THEMIS2, TIMP2, TPP1, VCAN*
Neutrophils	*ACSL1, ALPK1, AQP9, BASP1, BCL6, CD97, CEP19, CFLAR, CSF3R, CXCR2, DENND5A, DYSF, FAM65B, FCGR2C, FPR1, GLT1D1, GPR97, IFITM2, IL17RA, KCNJ2, KIAA0247, LILRA2, LIMK2, LINC01002, MGAM, MOB3A, NAMPT, NCF4, PADI2, PHC2, PHF21A, PLXNC1, PREX1, RALB, RNF149, S100A8, S100A9, SLC25A37, SNORD89, SSH2, STAT3, STAT5B, THBD, TLR2, TLR4, TMEM154, TNFRSF1A*
NK cells	*KIR2DL1, KIR2DL2, KIR2DL3, KIR2DL4, KIR2DL5A, KIR2DS1, KIR2DS2, KIR2DS3, KIR2DS5, KIR3DL1, KIR3DL2, KIR3DL3, KLRC2, KLRC3, KLRC4, KLRD1, PRF1, SAMD3, SH2D1B, TBX21*
Plasma cells	*GUSBP11, IGH, IGHG3, IGJ, IGKC, IGKV1D-13, IGLC1, IGLJ3, IGLL3P, IGLV@, IGLV1-44, MZB1, TNFRSF17, TXNDC5*

(Supplementary Table S4), which highlights the poor concordance between signatures (Supplementary Table S4; Supplementary Fig. S1). The highest observed similarity was between the B-cell signatures from *ImSig* and from (8), with a Jaccard score of 0.26, which is a not a high Jaccard score. Figure 2A illustrates the lack of consensus between published signatures and *ImSig*. Most (76.3%, or 2,794 genes) of the genes are associated with only a single study, and fewer than 10% of those genes described unique populations, e.g., erythroblast (297 genes; ref. 13) and megakaryocyte (259 genes; ref. 13). The poor conservation of immune marker genes across studies is likely due to various technical and statistical artifacts. For example, proliferation-related genes were identified as part of the signature for activated CD4 (12) and T cells (10). The mitotic index of resting versus activated T cells may be a true difference between them, but cell-cycle genes are expressed by all proliferating cells (39) and are therefore poor markers of cell type. Of all signatures proposed, *ImSig* contains the fewest unique genes; of the 318 immune-related markers defined by *ImSig*, only 60 genes have not been previously reported in other signatures, suggesting a good consensus overall with other studies but not with an individual published signature.

Certain genes are associated with different cell types in different studies. Of the 729 genes proposed to represent distinct T-cell states, none were common to all seven studies and only 98 were listed by two or more studies. The assignment of markers to cell types varies across studies (Fig. 2B). For example, *LRRN3* was used to define resting cytotoxic T cells by (11) and as a Th1 marker by (14), *CTLA4* is annotated as either a marker of Tregs, Th1, and $CD4^+$ T cells and by (12, 14) and (13), respectively. *CTLA4* can also be expressed on $CD8^+$ T cells (40). There are many such examples of discordance between marker gene/cell-type assignations. The *ImSig* T-cell signature, which was designed to be subtype agnostic, exhibited the greatest overlap between all T-cell signatures (displayed by the relative node size in Fig. 2B) and includes genes defined as subtype-specific by other studies but for which we found no support as a separate coexpression module. To compare the coexpression of the *ImSig* signatures to previous signatures, the median correlation of each set of signature genes was calculated within the context of a data set derived from trachoma patients. We used this data set because it was derived from a tissue in which all immune cell types defined by *ImSig* were present, these being recruited in response to a bacterial infection. For comparison with previous signatures, the modules representing subpopulations, such as T-cell subsets, were

Table 2. Table of *ImSig* genes (pathways signatures)

Interferon	*APOL1, APOL6, BATF2, BST2, C19orf66, C5orf56, CMPK2, DDX58, DDX60, DHX58, DTX3L, EPSTI1, FBXO6, GBP1, GBP4, HELZ2, HERC5, HERC6, HSH2D, IFI16, IFI35, IFI44, IFI44L, IFI6, IFIH1, IFIT1, IFIT2, IFIT3, IFIT5, IFITM1, IRF7, IRF9, ISG15, LAMP3, LAP3, MX1, MX2, OAS2, OAS3, OASL, PARP10, PARP12, PARP14, PARP9, PHF11, PML, PSMB9, RNF213, RSAD2, RTP4, SAMD9, SAMD9L, SHISA5, SIGLEC1, SP110, STAT1, STAT2, TAP1, TRAFD1, TRIM21, TRIM22, TRIM5, UBE2L6, USP18, XAF1, ZNFX1*
Proliferation	*ANLN, ASPM, AURKA, AURKB, BIRC5, BUB1, BUB1B, CASC5, CCNA2, CCNB1, CCNB2, CCNE2, CDC20, CDC6, CDCA2, CDCA3, CDCA5, CDCA7, CDCA8, CDK1, CDKN3, CDT1, CENPA, CENPE, CENPF, CENPL, CEP55, CKS1B, DEPDC1, DEPDC1B, DLGAP5, DONSON, DTL, E2F8, ECT2, EZH2, FAM72C, FANCI, FBXO5, FOXM1, GINS1, GINS2, GMNN, HJURP, HMGB3, HMMR, KIAA0101, KIF11, KIF14, KIF15, KIF18B, KIF20A, KIF2C, KIF4A, MAD2L1, MCM10, MCM2, MCM4, MCM6, MELK, MKI67, MND1, MTFR2, NCAPG, NCAPG2, NDC80, NEK2, NUF2, NUSAP1, OIP5, PARPBP, PBK, PCNA, PLK4, POLE2, POLQ, PTTG1, RACGAP1, RAD51, RAD51AP1, RRM1, RRM2, SHCBP1, SKA1, SMC2, SPC25, STIL, STMN1, TCF19, TK1, TOP2A, TPX2, TRIP13, TTK, TYMS, UBE2C, UHRF1, ZWILCH, ZWINT*
Translation	*EEF1A1, EEF1B2, EEF1D, EEF1G, EIF3D, EIF3E, EIF3F, EIF3G, EIF3H, EIF3K, FAU, GNB2L1, NACA, PFDN5, RPL10, RPL10L, RPL11, RPL12, RPL13, RPL13A, RPL14, RPL15, RPL17, RPL18, RPL18A, RPL19, RPL21, RPL22, RPL23, RPL23A, RPL24, RPL27, RPL27A, RPL28, RPL29, RPL3, RPL30, RPL31, RPL32, RPL34, RPL35, RPL35A, RPL36A, RPL37, RPL37A, RPL38, RPL39, RPL4, RPL5, RPL6, RPL7, RPL7A, RPL8, RPL9, RPLP0, RPLP2, RPS10, RPS11, RPS13, RPS14, RPS15, RPS15A, RPS16, RPS17, RPS18, RPS19, RPS2, RPS20, RPS21, RPS23, RPS25, RPS27A, RPS28, RPS29, RPS3, RPS3A, RPS5, RPS6, RPS7, RPS8, RPS9, RPSA, SNHG6, SNHG8, SNRPD2, UXT*

Published OnlineFirst September 28, 2018; DOI: 10.1158/2326-6066.CIR-18-0342

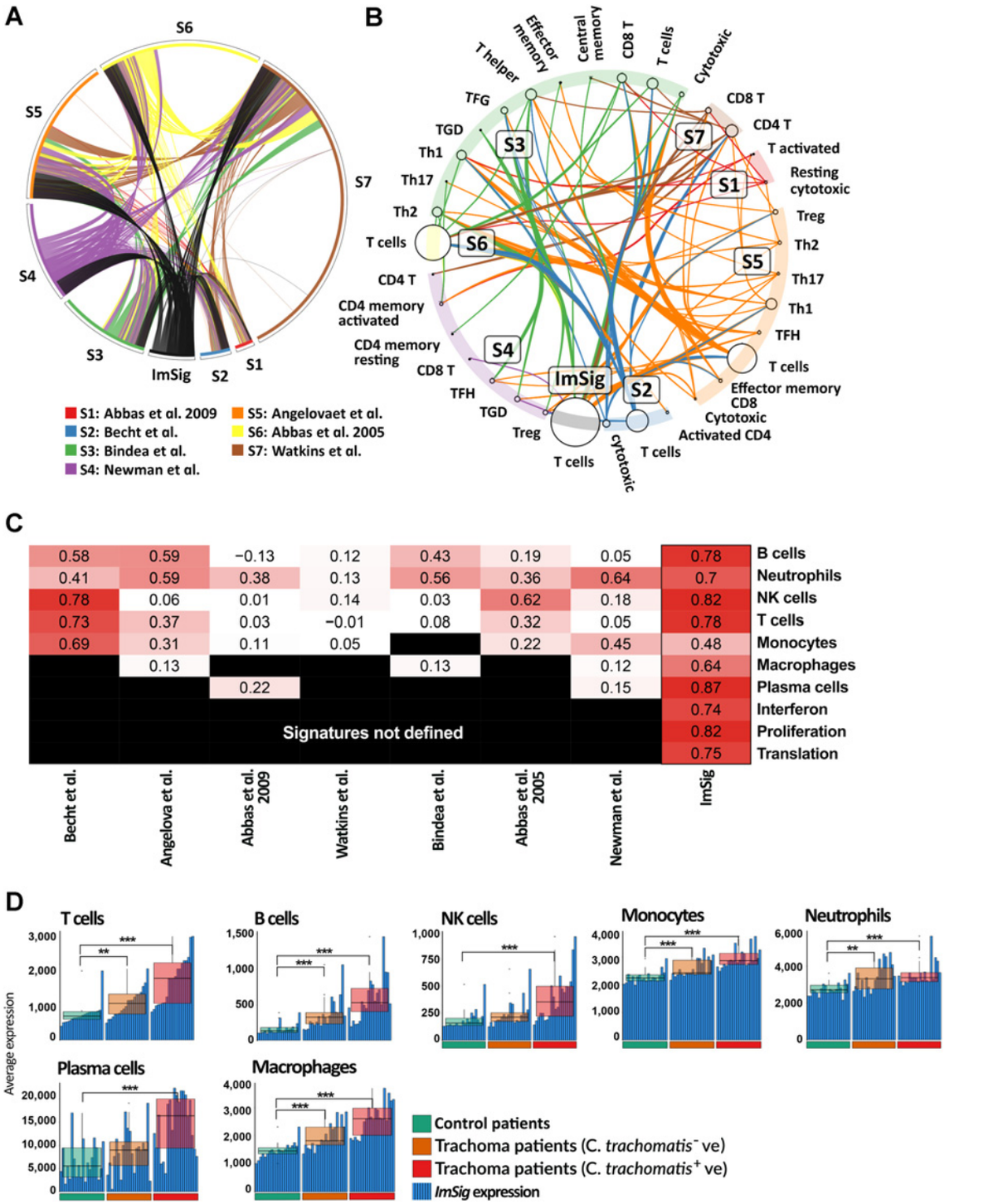

Figure 2.

Comparison of *ImSig* with other published signatures. **A,** Chord diagram showing the overlap between marker genes across studies. In most studies, a large proportion of genes were unique to the signatures defined by them. *ImSig* showed the best overlap (81%) with other studies. **B,** Network diagram showing the relationship between T-cell subtype–specific genes among six studies and *ImSig*. Only genes that were present in two or more studies are represented in this plot (98 genes representing 13.4% of all T-cell marker genes). Nodes are sized relative to the number of shared genes between one signature and others. *ImSig* included genes describing various subtypes and was the most conserved set among all studies compared. **C,** Heat map of the median correlation between genes from published signatures as assessed in the context of the trachoma data set (GSE20436). Where a cell-type signature was split into subsets, subset signatures were combined to represent the parent population. The median correlation values of signatures without combining them into parent population is available (Supplementary Table S5). **D,** Bar plots of the average expression of signature genes (estimated relative abundance) across the data set, each bar representing the average expression of signature genes in an individual patient sample. Samples are ordered according to T-cell content, low to high (left to right), and this order is maintained for other plots. **, $P = 0.01$; ***, $P = 0.001$.

Published OnlineFirst September 28, 2018; DOI: 10.1158/2326-6066.CIR-18-0342

subsumed into one module, such as T cells. Their median correlation in the context of the trachoma data set is shown in Fig. 2C. A noncollated version of the results is provided in Supplementary Table S5. Regardless of whether data were aggregated by broad cell type, or by subtype, none of the blood-derived modules were strongly coexpressed across the set of trachoma patient samples. In contrast, all of the *ImSig* signatures displayed a high median correlation (coexpression) value. The gene signatures from ref. 8 performed next best. The bacterial infection that gives rise to the pathology of trachoma leads to recruitment of immune cells to the site of infection (32). In order to evaluate the ability of *ImSig* to estimate the relative abundance of immune cells, the average expression of each gene signature was used as a proxy for immune cell number in the trachoma data set. All immune cell populations increased in patient groups relative to controls, with greater increases seen in patients with an active infection (Fig. 2D).

To validate the applicability of *ImSig* on RNA-seq data and in the context of tumor biology, we computed the relative abundance of immune cells in four metastatic melanoma patients from whom samples were collected from lymph nodes. A fraction of the single-cell suspension was used to measure cell-type proportions by flow cytometry and the other fraction was used for RNA-seq analysis. We observed good agreement ($r = 0.91$, RMSE = 0.1, and $P = 2.74E{-}05$) between predictions of relative cell number made using *ImSig* and experimentally determined cell numbers (see also Supplementary Fig. S2). Thus, *ImSig* accurately predicted relative cell numbers for all cell types, as confirmed by the low root-mean-square error (RMSE).

Deconvolution of tissue data

In the context of GCN analyses, the *ImSig* signatures can be used to identify other context-specific genes expressed by immune populations. For example, the T cell and macrophage signatures were correlated with each other, consistent with an immune-mediated inflammatory process, and many immune-related genes were coexpressed with *ImSig* genes in the context of the trachoma data (Fig. 3A). The expression profile of genes such as *IFNG*, *LAG3*, *CD44*, *FOX03*, *FOXP3*, *CD80*, *IL20*, *STAT4*, and *IL17A* was correlated with T-cell signature genes, indicating that the T-cell population included Th17, Treg, and Th1 subtypes (Fig. 3B). Similarly, genes associated with the macrophage signature contained many M1 markers. Performing a network analysis such as this can also provide a broader perspective of the transcriptional signatures of other cell types present in clinical samples. When the data set is examined as a whole, many GCN clusters can be assigned to other cell populations or processes (41).

Satisfied with the performance of *ImSig* in the context of tissue transcriptomics data, we explored its utility in the analysis of transcriptomics data derived from cancer.

Analysis of immune infiltrates in cancer

Our previous analysis of the cancer transcriptome showed that expression signatures of immune cells can be extracted from large cancer transcriptomic data sets, but we did not at that time correlate gene-expression signatures with patient outcomes (20). To test the use of *ImSig* in the study of the tumor microenvironment, the 12 largest TCGA cancer data sets were examined and HRs were computed between high and low-immune cell infiltrate groups (Fig. 4A). Although the survival analysis was not adjusted for potentially confounding variables (such as tumor stage, grade, age, or treatment), the findings were consistent with the literature. In melanoma (SKCM), we reaffirmed the known association between tumor-infiltrating lymphocytes (TIL) and a good

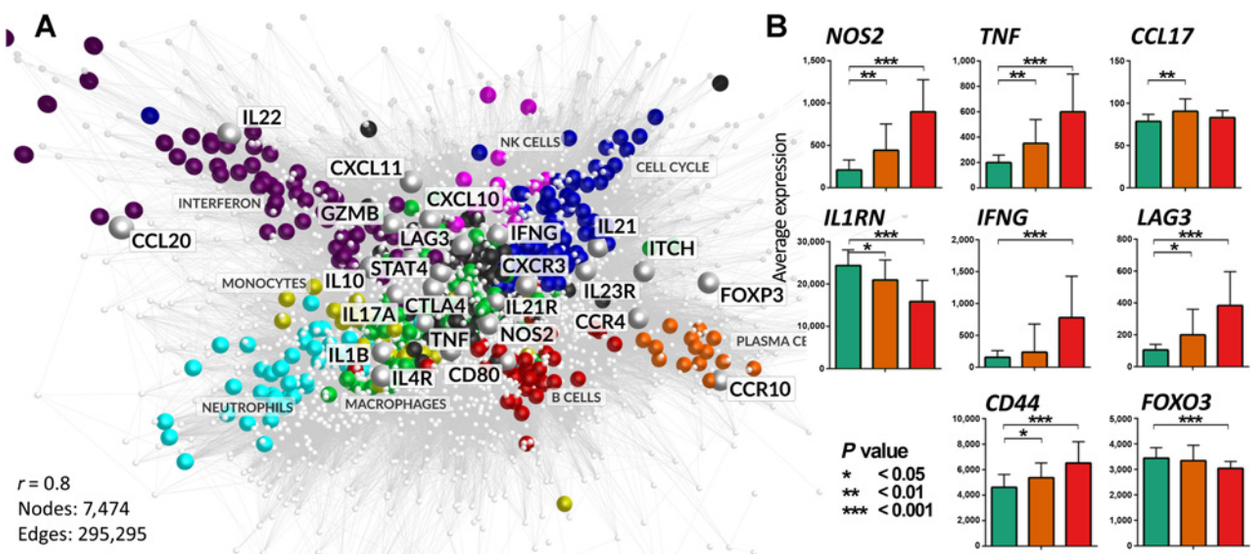

Figure 3.

Coexpression of other immune genes with *ImSig* core signatures. **A,** Correlation network of genes associated with the immune clusters during trachomatis infection. *ImSig* genes are colored according to the different immune cell types they represent, whereas the genes coclustering with the *ImSig* immune genes are shown as nodes without color and reduced in size. Highlighted with a greater node size and label are a few well-known immune-modulatory genes present in the immediate vicinity of the signature genes. **B,** Bar plots of the average expression intensity of a few well-known immune-modulatory genes across the three patient groups.

Published OnlineFirst September 28, 2018; DOI: 10.1158/2326-6066.CIR-18-0342

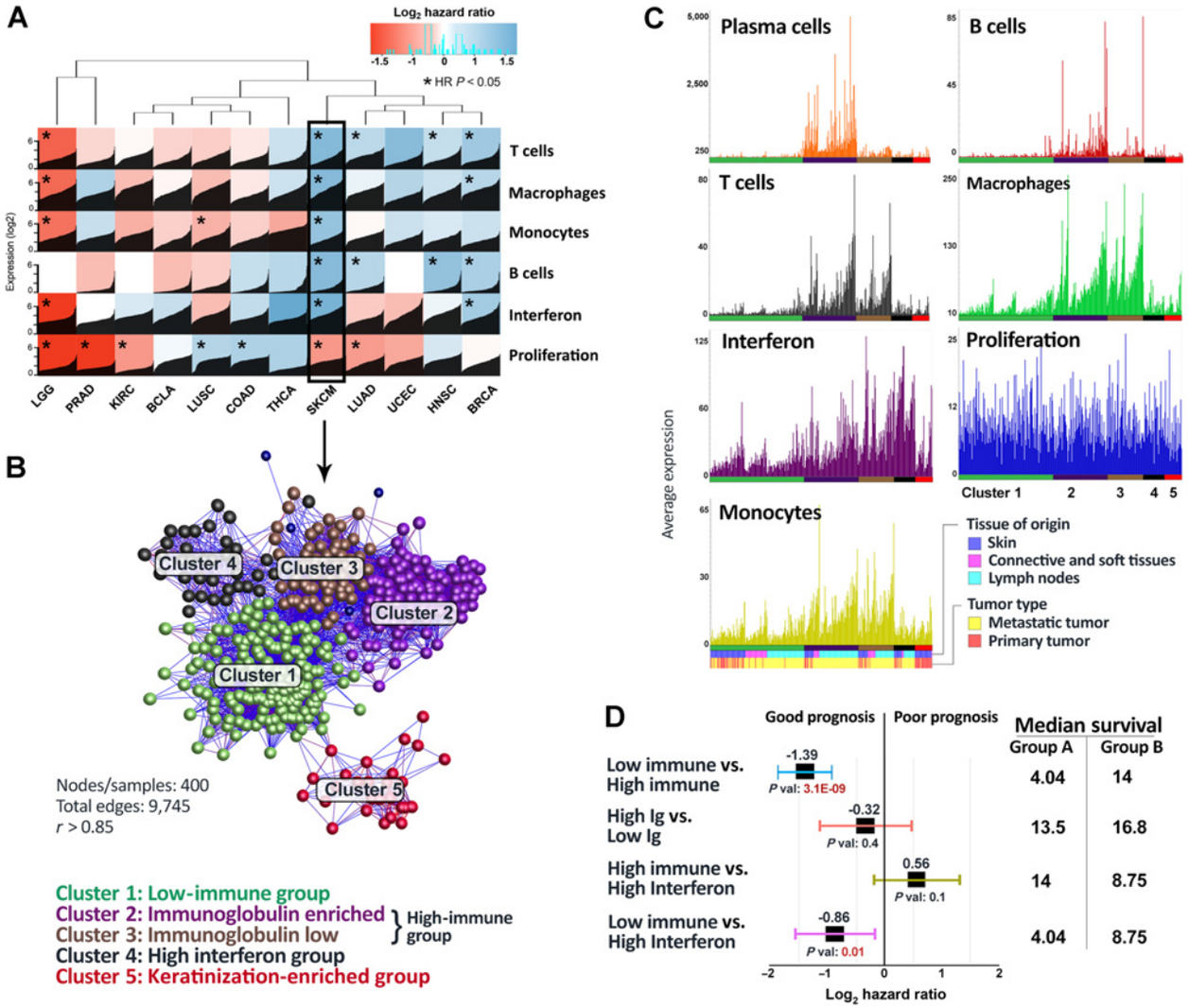

Figure 4.

Application of *ImSig* to tumor data. **A,** Prognostic map of 12 cancer types based on immune cell content. The average expression of each *ImSig* signature was calculated for each sample/tumor type. Samples were then ordered according to each signature (low–high, black plot in each square), and the HR was calculated between the lowest and highest expressing samples. Blue represents a good prognosis with increased expression of the signature genes and red a poor prognosis. *, a HR $P < 0.05$. BCLA-bladder urothelial carcinoma; BRCA, breast invasive carcinoma; COAD, colon adenocarcinoma; HNSC, head and neck squamous cell carcinoma; KIRC, kidney renal clear cell carcinoma; LGG, brain lower grade glioma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; PRAD, prostate adenocarcinoma; SKCM, skin cutaneous melanoma; THCA, thyroid carcinoma; UCEC, uterine corpus endometrial carcinoma. **B,** Sample–sample correlation graph of melanoma patient samples based on expression of *ImSig* genes in and clustered using MCL algorithm. Here, every node is a patient, and the edges correspond to the correlation between them. **C,** Expression profile of *ImSig*-related genes within the various clusters/grouping as defined in **B**. Here, the y-axis is the average expression of the signature genes, and x-axis describes patient groupings as shown in **B**. **D,** Univariate Cox-proportional analysis between the patient groups as defined in **B**.

prognosis (42, 43). Breast cancer (BRCA) is not as immunogenic as melanoma, but several studies have associated TILs with a good prognosis as observed here (44). A negative association between TILs and prognosis was evident in LGG (45, 46) and lung squamous cell carcinoma (LUSC; refs. 47, 48) in accordance with previous literature. We did find prognostic value of the interferon response in LGG. We confirmed an association between the proliferation signature and a good prognosis in colorectal cancers (COAD; as shown in ref. 49) and also in LUSC. Analysis of individual proliferation-related genes in LUSC supported this observation (log2HR: *G2E3*, 0.66; *MND1*, 0.56; *CHEK2*, 0.53; *RFC4*, 0.51; *CEP192*, 0.48; *CDKN3*, 0.47; *CENPA*, 0.47; *CCND2*, 0.47; *CDC7*, 0.46: $P < 0.05$). One possible explanation for this counterintuitive observation is that the mitotic signal in these tissues originates from proliferating immune cells, not from cancer cells (50, 51).

We performed a molecular subgrouping of melanoma based on *ImSig,* using the signature genes only to group patient samples. Unsupervised clustering based on the immune

Published OnlineFirst September 28, 2018; DOI: 10.1158/2326-6066.CIR-18-0342

profile revealed five groups of patient samples (Fig. 4B). Clinical features such as the tissue of origin and tumor type (metastatic or primary) did not affect the subtyping. Nearly half the patients were in cluster-1, characterized by little immune infiltrate (Fig. 4C). HR analysis between these low-immune (cluster 1) and high-immune infiltrate (clusters 2 and 3) tumors revealed a significant difference in patient survival (HR: 0.38, $P = 3E{-}9$). The median survival of patients in the group with high-immune infiltrate was 10 years greater than that of patients with low-immune infiltrate (Fig. 4D). Within the high-immune infiltration subgroup, cluster 2 appeared to have more B cells and plasma cells than cluster 3 (Fig. 4C), but overall survival (HR) was not significantly different between the two groups (Fig. 4D). Cluster-4 samples displayed higher expression of interferon response genes and also showed improved survival compared with the low-immune infiltrate group (Fig. 4D). Finally, patients in cluster 5 had a low-immune infiltrate, showed greater expression of keratin-related genes, and presented the worst survival rates (median survival = 2.34 years). Although patients in clusters 2, 3, and 4 did not differ in HR, they could differ in other ways, such as responses to treatment. Following an analogous analysis, we reproduced the five patient groupings on an independent validation data set (GSE65904), which showed a similar infiltration pattern (Supplementary Fig. S3A), survival analysis, and prognostic pattern (Supplementary Fig. S3B). High-immune and keratin subgroups have been identified and described in melanoma (52, 53) but these studies did not describe the type and variation in the immune infiltrate in melanomas. Our analysis reveals the nature of the immune landscape of these tumors and differences in their survival.

Use of *ImSig* in identifying immune cells in single-cell data

To extend these analyses and validate the *ImSig* signatures in the context of single-cell data, we examined single-cell data derived from melanoma (35). The immune component of the melanoma single-cell analysis included 515 B cells, 126 macrophages, 52 NK cells, and 2,069 T cells. Cell type–specific expression of *ImSig* markers was observed ($P < 7E{-}15$) as illustrated in Fig. 5A. For each patient, the estimated proportion of immune cells was compared with the true proportion. The estimated proportion was concordant with the measured number of cells ($P < 0.05$), with the poorest observed correlation being $r = 0.97$. Randomized permutation analysis with the same-sized gene sets produced no significant correlation (Fig. 5B). Figure 5C illustrates the concordance between the measured and estimated number of cells.

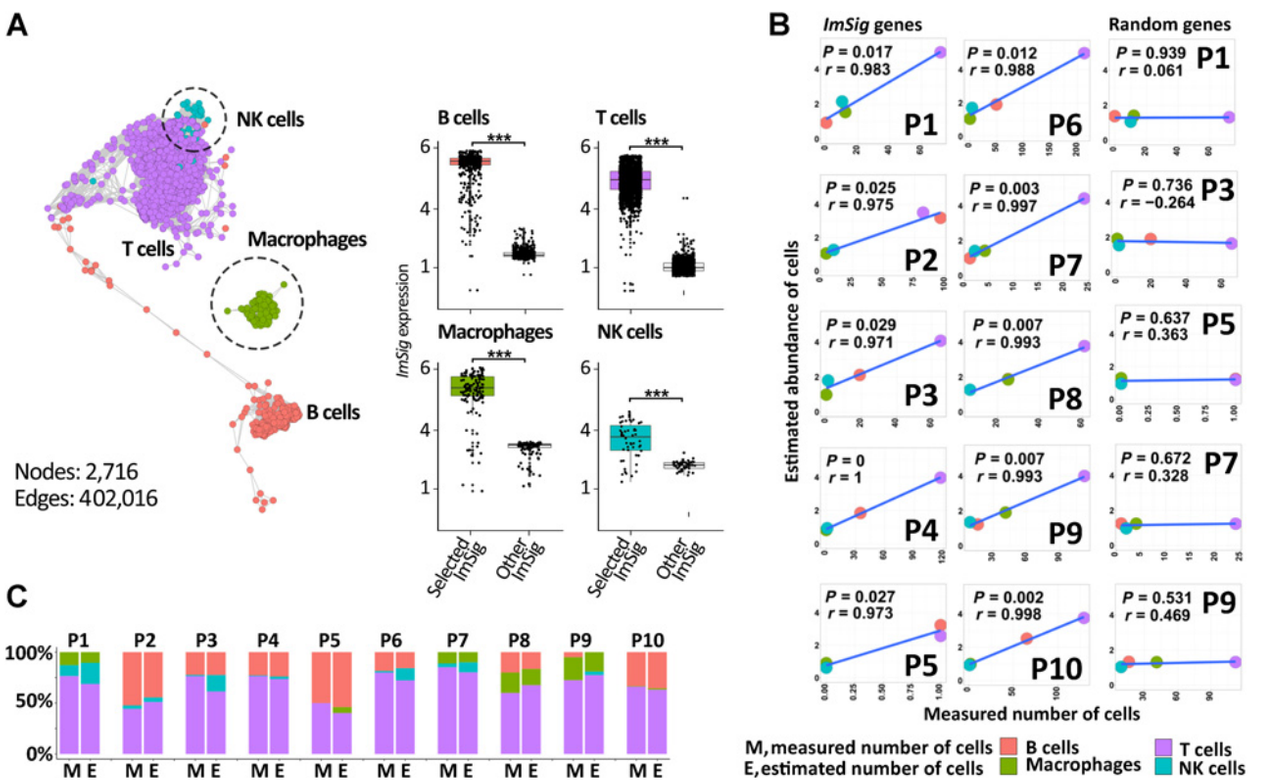

Figure 5.

Validation of *ImSig* using single-cell RNA-seq data from melanoma samples. **A,** The immune component of the melanoma single-cell data displayed as a correlation network, each node representing a cell from melanoma. Box plots display the average expression of cell type–specific *ImSig* genes in their respective cell types compared with the average expression of other *ImSig* genes. Process-specific *ImSig* signature genes (proliferation, interferon, and translation) were omitted in this analysis. **B,** Linear regression plots showing the concordance between the estimated and measured abundance of immune cells in 10 patients. For five patients (P1, P3, P5, P7, and P9), the regression line was also calculated using a random set of genes to highlight the specificity of *ImSig* genes. **C,** Stacked bar plots showing the concordance between measured and estimated proportions of immune cells. ***, $P = 0.001$.

Published OnlineFirst September 28, 2018; DOI: 10.1158/2326-6066.CIR-18-0342

Table 3. Identification of immune cells within single-cell data

Cells	Correct prediction	Wrong prediction	Accuracy (%)	Error (%)
B cells	122	16	88.4	11.6
Macrophages	84	1	98.8	1.2
T cells	1,185	2	99.8	0.2
Other cells (4,087 cells)		93		2.3

NOTE: *ImSig* was used in conjunction with the SVM-based classifier Cibersort, to identify immune cells within the head and neck tumor (HNSCC) single-cell data. The table shows the accuracy of identification. Sixty-three immune cells were unassigned as their *P* value was greater than 0.05.

The single-cell community depends on gene markers, gene signatures, and clustering algorithms to define cell types. Here, we show the utility of *ImSig* when used in association with classification algorithms, such as support vector machine (SVM), to predict cell types from single-cell RNA-seq data. To demonstrate the potential for automation, we used the SVM-based deconvolution tool Cibersort (5) with a reference profile generated with *ImSig* to predict immune cells within a single-cell data set from head and neck tumors (HNSCC; ref. 36). The immune component of the HNSCC data set contained 1,473 cells. Prediction using *ImSig* yielded a high degree of accuracy for B cells (88.4% correct prediction), macrophages (98.8%), and T cells (99.8%; Table 3). Only 63 immune cells remained uncategorized ($P > 0.05$). With respect to the other 4,087 cells, which consisted of myocytes, mast cells, malignant cells, fibroblast, dendritic cells, and endothelial cells, only 2.2% of cells were misclassified as macrophages, B cells, or T cells. In contrast, Cibersort's (5) default blood-derived signature (LM22) showed an accuracy rate for B cells of 15.2%; macrophages, 0%; and T cells, 75.3%. However, the LM22 signature was not designed to deconvolute single-cell data, and its poor performance is likely a result of using a blood-derived signature and a reference gene matrix based on microarrays.

Discussion

Cellular heterogeneity is a hallmark of cancer, in terms of both the tumors themselves and the normal cells that both support and control their growth. A wealth of transcriptomics data has been generated from cancer samples and a number of studies report approaches to deconvolute these data and to define the set of cell types present therein. However, we and others (16) found that immune signatures derived by comparing the expression profile of immune cells isolated from blood do not perform optimally when applied to tissue data.

The current work is based on the observation that genes associated with a specific cell population or biological process form highly connected cliques of nodes when large collections of transcriptomics data are subjected to network-based correlation analysis (18, 41). Although the main goal of this study was to define immune gene signatures for the deconvolution of cancer data, we have derived *ImSig* from a range of tissue pathologies and platforms to ensure its applicability across different data types and sources. Our aim in defining *ImSig* was to choose the most robustly coexpressed genes for each immune cell type from the analysis of tissue data, thereby defining a "core" or invariant cell type–specific signature.

In any given tissue, a gene may be expressed by multiple cell types present therein or a cell type may not be present, hence the need to explore a wide variety of tissue data. We also chose to include signatures for interferon signaling, proliferation (mitosis) and translation, as these are commonly observed coexpression modules in tissue and act as additional controls. Validatory analysis of the resultant *ImSig* signatures showed the gene signatures to be enriched with appropriate GO terms, and manual inspection of the lists with reference to the literature also supported the validity of the selected genes. This was further confirmed by the observed coexpression of the *ImSig* signatures across a range of data sets not used for their derivation and their average expression reflecting changes in immune cell numbers, where known, as in trachoma.

As the current study is not the first to attempt to define sets of signatures for immune cells, we sought to compare *ImSig* with other published signatures, in terms of both gene content and performance. Definition and comparison of cell signatures is complex. In the first instance, published gene signatures vary in terms of the number of genes they include and the cell populations and subpopulations they seek to define. Second, there is no benchmark data set where the number and nature of immune cells are known in the context of a tissue environment. Comparison of the signatures showed many to include gene markers only defined by that study. Where gene markers were common to more than one study, there was a complex relationship between the assignation of genes to cells across studies. In other words, there is little consensus across published immune marker lists. Of all the signatures, *ImSig* contained the fewest unique genes (60 genes), suggesting that the gene content of *ImSig* represents a consensus view of other studies, despite being derived independently. A comparison of the relative performance of signatures again represented a challenge. Where multiple subtypes of cells were defined, the genes associated with subtypes were either analyzed separately or collapsed into a single signature. We chose to compare the performance of these summarized signatures in the context of the trachoma data set, where we knew all immune cell types defined by *ImSig* to be present and that their numbers increase during an active infection (32). In this context, the degree of coexpression between genes associated with individual *ImSig* signatures was better in some cases than in others. Furthermore, the average expression of *ImSig* signatures mirrored the known increase in immune cell infiltrate across patient groups (32).

Researchers seek to define immune cell subtypes and activation states associated with different tissues, developmental stages, and pathologies. Although heterogeneity among immune cell populations exists, few markers can identify this heterogeneity outside of the context of flow cytometry and IHC. For instance, tissue macrophages are sometimes named depending on their tissue of origin (microglia, Kupffer cells, etc.) or activation state (M1, M2, etc.) and in other cases are referred to as dendritic cells (53, 54). In the literature we have cited, signatures for 22 T-cell subsets are reported, and this does not include all T-cell subsets that are defined in the literature (55). In a given pathologic state, multiple cellular subtypes or populations whose biology is adapted to different niches are likely to

Published OnlineFirst September 28, 2018; DOI: 10.1158/2326-6066.CIR-18-0342

be present. We would argue that it is unrealistic to categorically identify their individual signatures from bulk tissue data, especially when the differences between them are more likely to be a spectrum than a series of absolute states (38). Even among different myeloid populations, i.e., monocytes, macrophages, and neutrophils, we have found few markers that are specific to one population or another. Markers that define the presence of these populations do so more by their coexpression than by absolute expression in the context of tissue.

We suggest that many immune subtype markers are too poorly defined to reliably distinguish immune cell subsets in the context of transcriptomics data derived from tissue. However, network analysis can provide a comprehensive picture of the immune microenvironment. By examination of the genes that correlate with the core signature genes, even if those genes expression cannot be reliably assigned to one cell type or another, it is possible to capture the overall profile of the immune microenvironment of a tissue. It may, after all, be the sum of the individual parts that matter. How these findings are used to identify immune cell subtypes, we leave to the individual analyst.

After satisfying ourselves of the validity of *ImSig* and comparing it to other signatures for defining immune populations in tissue data, we used it to analyze transcriptomics data sets derived from 12 cancer types. In each case, the majority of signature genes were tightly coexpressed, apart from instances where we believe the target cell was not present or was in low abundance. When the samples for each tumor type were ranked according to their immune cell content (as defined by the average expression of the signature genes), we were able to demonstrate variation in the immune microenvironment between tumors and the association of specific immune cell populations with good or poor prognoses. Despite an established association between the immune system and survival in melanoma (56), there has been little effort to subgroup patients based upon what immune cell types are present in their tumors. Previous studies have merely defined tumors as having a high- or low-immune-cell content (34, 57). We, therefore, explored the use of *ImSig* in the molecular subtyping melanoma patients. The analysis demonstrated a greater heterogeneity in the immune infiltrate of melanoma than previously reported (52, 34). We distinguished tumors characterized by T cells and macrophages (cluster 3), interferon enrichment (cluster 4), or B-cell infiltration (cluster 2). Treating the immune infiltrate of tumors as an overall signature limits the potential to identify prognostically significant subgroups. In other cases, merging the immune infiltrate into one immuno-subgroup might result in opposing survival differences cancelling each other out, for example, if T cells were associated with a good prognosis and macrophages a bad prognosis. Understanding the immune heterogeneity of tumors may be key to predicting responses to immunotherapy (58, 59).

The advent of single-cell transcriptomics and its application to understanding the microenvironment of cancer promises to facilitate the profiling of all the cells of a tumor as never before possible (60) and may eventually circumvent the need to deconvolute tissue data, as described here. The technology to perform these analyses is improving and may in the future answer many questions about immune cell heterogeneity. However, at present, the data available are limited and the droplet-based RNA-seq methods being widely used may not provide a sufficient depth of sequencing to go beyond the identification of cell type. Here, we demonstrate how *ImSig* was able to accurately define the type and relative abundance of immune cells in single-cell transcriptomics data derived from melanoma, as well as head and neck cancer. As the quantity and quality of single-cell cancer data sets improve and our understanding of the expression profile of these cells improves, markers that are able to differentiate between immune subtypes or activation states, specifically in the context of the tumor microenvironment, may emerge.

ImSig is directly derived from tissue data. Although its gene content is not entirely novel, we believe *ImSig* performs better than previously published immune signatures as a subtype-agnostic means to estimate the relative abundance of immune cells across tissue samples. We also demonstrate the ability of *ImSig* to facilitate identification of biomarkers when applied in the context of network coexpression analyses. We anticipate that *ImSig* will aid studies of immune cell variation in tumors, responses to therapy, and predictive biomarkers.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors' Contributions

Conception and design: A.J. Nirmal, A.H. Sims, T.C. Freeman
Development of methodology: A.J. Nirmal, T. Regan, T.C. Freeman
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): A.J. Nirmal
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): A.J. Nirmal, B.B. Shih, D.A. Hume, A.H. Sims, T.C. Freeman
Writing, review, and/or revision of the manuscript: A.J. Nirmal, T. Regan, B.B. Shih, D.A. Hume, A.H. Sims, T.C. Freeman
Study supervision: T.C. Freeman

Acknowledgments

A.J. Nirmal is a recipient of The Roslin Institute and CMVM scholarship and Edinburgh Global Research Scholarship. A.H. Sims is funded by Breast Cancer Now; T. Regan, B.B. Shih, and T.C. Freeman are funded by MRC consortium grants (MR/M003833/1 and MR/L014815/1); T.C. Freeman is funded by an Institute Strategic Grant from the Biotechnology and Biological Sciences Research Council (BBSRC; BB/JO1446X/1); and D.A. Hume is supported by The Mater Foundation.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received May 23, 2018; revised July 21, 2018; accepted September 24, 2018; published first September 28, 2018.

References

1. Postow MA, Callahan MK, Wolchok JD. Immune checkpoint blockade in cancer therapy. J Clin Oncol 2015;33:1974–82.
2. Denkert C, von Minckwitz G, Darb-Esfahani S, Lederer B, Heppner BI, Weber KE, et al. Tumor-infiltrating lymphocytes and prognosis in different subtypes of breast cancer: a pooled analysis of 3771 patients treated with neoadjuvant therapy. Lancet Oncol 2018;19:40–50.
3. Hackl H, Charoentong P, Finotello F, Trajanoski Z. Computational genomics tools for dissecting tumor-immune cell interactions. Nat Rev Genet 2016;17:441–58.
4. Gong T, Szustakowski JD. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. Bioinformatics 2013;29.

5. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods 2015;12.
6. Li B, Severson E, Pignon J-C, Zhao H, Li T, Novak J, et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. Genome Biol 2016;17:174.
7. Qiao W, Quon G, Csaszar E, Yu M, Morris Q, Zandstra PW. PERT: a method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions. PLoS Comput Biol 2012;8: e1002838.
8. Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biol 2016;17:218.
9. Zhong Y, Wan Y-W, Pang K, Chow LM, Liu Z. Digital sorting of complex tissues for cell type-specific gene expression profiles. BMC Bioinformat 2013;14:89.
10. Abbas AR, Baldwin D, Ma Y, Ouyang W, Gurney A, Martin F, et al. Immune response in silico (IRIS): immune-specific genes identified from a compendium of microarray expression data. Genes Immun 2005;6:319–31.
11. Abbas AR, Wolslegel K, Seshasayee D, Modrusan Z, Clark HF. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. PLoS One 2009;4:e6098.
12. Angelova M, Charoentong P, Hackl H, Fischer ML, Snajder R, Krogsdam AM, et al. Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy. Genome Biol 2015;16:64.
13. Watkins NA, Gusnanto A, de Bono B, De S, Miranda-Saavedra D, Hardie DL, et al. A HaemAtlas: characterizing gene expression in differentiated human blood cells. 2009. e1–e9 p.
14. Bindea G, Mlecnik B, Tosolini M, Kirilovsky A, Waldner M, Obenauf AC. Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. Immunity 2013;39.
15. Schelker M, Feau S, Du J, Ranu N, Klipp E, MacBeath G, et al. Estimation of immune cell content in tumor tissue using single-cell RNA-seq data. Nat Commun 2017;8:2032.
16. Pollara G, Murray MJ, Heather JM, Byng-Maddick R, Guppy N, Ellis M, et al. Validation of immune cell modules in multicellular transcriptomic data. PLoS One 2017;12:e0169271.
17. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. Nature 1999;402:C47.
18. Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. Science 2003;302: 249–55.
19. Forrest ARR, Kawaji H, Rehli M, Kenneth Baillie J, de Hoon MJL, Haberle V, et al. A promoter-level mammalian expression atlas. Nature 2014;507: 462–70.
20. Doig TN, Hume DA, Theocharidis T, Goodlad JR, Gregory CD, Freeman TC. Coexpression analysis of large cancer datasets provides insight into the cellular phenotypes of the tumor microenvironment. BMC Genomics 2013;14:1–16.
21. Freeman TC, Ivens A, Baillie JK, Beraldi D, Barnett MW, Dorward D, et al. A gene expression atlas of the domestic pig. BMC Biol 2012;10:90–90.
22. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M. NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Res 2013;41.
23. Carvalho BS, Irizarry RA. A framework for oligonucleotide microarray preprocessing. Bioinformatics 2010;26:2363–67.
24. Du P, Kibbe WA, Lin SM. lumi: a pipeline for processing Illumina microarray. Bioinformatics 2008;24:1547–48.
25. Theocharidis A, van Dongen S, Enright AJ, Freeman TC. Network visualization and analysis of gene expression data using BioLayout express(3D). Nat Protoc 2009;4.
26. Freeman TC, Goldovsky L, Brosch M, Dongen S, Mazière P, Grocock RJ, et al. Construction, visualisation, and clustering of transcription networks from microarray expression data. PLoS Comput Biol 2007;3.
27. Enright AJ, Dongen SV, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 2002;30.
28. Alexa A RJ. topGO: enrichment analysis for gene ontology. R package 2016; version 2.26.0.
29. Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, et al. The reactome pathway knowledgebase. Nucleic Acids Res 2016;44.
30. Gu Z, Gu L, Eils R, Schlesner M, Brors B. Circlize implements and enhances circular visualization in R. Bioinformatics 2014;30:2811–12.
31. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13:2498–504.
32. Natividad A, Freeman TC, Jeffries D, Burton MJ, Mabey DCW, Bailey RL, et al. Human conjunctival transcriptome analysis reveals the prominence of innate defense in chlamydia trachomatis infection. Infect Immun 2010;78:4895–911.
33. Schröder MS, Culhane AC, Quackenbush J, Haibe-Kains B. survcomp: an R/Bioconductor package for performance assessment and comparison of survival models. Bioinformatics 2011;27:3206–08.
34. Cirenajwis H, Ekedahl H, Lauss M, Harbst K, Carneiro A, Enoksson J, et al. Molecular stratification of metastatic melanoma using gene expression profiling: prediction of survival outcome and benefit from molecular targeted therapy. Oncotarget 2015;6:12297–309.
35. Tirosh I, Izar B, Prakadan SM, Wadsworth MH, Treacy D, Trombetta JJ, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science 2016;352:189–96.
36. Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. Cell;171:1611–24.e24.
37. van Dijk D, Nainys J, Sharma R, Kathail P, Carr AJ, Moon KR, et al. MAGIC: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data. bioRxiv 2017.
38. Xue J, Schmidt SV, Sander J, Draffehn A, Krebs W, Quester I, et al. Transcriptome-based network analysis reveals a spectrum model of human macrophage activation. Immunity 2014;40:274–88.
39. Giotti B, Chen S-H, Barnett MW, Regan T, Ly T, Wiemann S, et al. Assembly of a parts list of the human mitotic cell cycle machinery. bioRxiv 2018.
40. McCoy KD, Le Gros G. The role of CTLA-4 in the regulation of T cell immune responses. Immunol Cell Biol 1999;77:1.
41. Shih BB, Nirmal AJ, Headon DJ, Akbar AN, Mabbott NA, Freeman TC. Derivation of marker gene signatures from human skin and their use in the interpretation of the transcriptional changes associated with dermatological disorders. J Pathol 2017:n/a-n/a.
42. Ladanyi A. Prognostic and predictive significance of immune cells infiltrating cutaneous melanoma. Pigment Cell Melanoma Res 2015;28: 490–500.
43. Mann GJ, Pupo GM, Campain AE, Carter CD, Schramm S-J, Pianova S, et al. BRAF Mutation, NRAS mutation, and the absence of an immune-related expressed gene profile predict poor outcome in patients with stage III melanoma. J Invest Dermatol 2013;133:509–17.
44. West NR, Kost SE, Martin SD, Milne K, deLeeuw RJ, Nelson BH, et al. Tumor-infiltrating FOXP3+ lymphocytes are associated with cytotoxic immune responses and good clinical outcome in oestrogen receptor-negative breast cancer. Br J Cancer 2013;108:155–62.
45. Yao Y, Ye H, Qi Z, Mo L, Yue Q, Baral A, et al. B7-H4(B7x)–Mediated Cross-talk between glioma-initiating cells and macrophages via the IL6/JAK/STAT3 pathway lead to poor prognosis in glioma patients. Clin Cancer Res 2016;22:2778.
46. Zhang C, Li J, Wang H, Wei Song S. Identification of a five B cell-associated gene prognostic and predictive signature for advanced glioma patients harboring immunosuppressive subtype preference. Oncotarget 2016; 7(45).
47. Hiraoka K, Zenmyo M, Watari K, Iguchi H, Fotovati A, Kimura YN, et al. Inhibition of bone and muscle metastases of lung cancer cells by a decrease in the number of monocytes/macrophages. Cancer Sci 2008;99:1595–602.
48. Shibutani M, Maeda K, Nagahara H, Ohtani H, Sakurai K, Yamazoe S, et al. Prognostic significance of the lymphocyte-to-monocyte ratio in patients with metastatic colorectal cancer. World J Gastroenterol 2015;21:9966–73.
49. Melling N, Kowitz CM, Simon R, Bokemeyer C, Terracciano L, Sauter G, et al. High Ki67 expression is an independent good prognostic marker in colorectal cancer. J Clin Pathol 2016;69:209–14.
50. Lefrançais E, Ortiz-Muñoz G, Caudrillier A, Mallavia B, Liu F, Sayah DM, et al. The lung is a site of platelet biogenesis and a reservoir for haematopoietic progenitors. Nature 2017;544:105–09.
51. Kallinikos-Maniatis A. Megakaryocytes and platelets in central venous and arterial blood. Acta Haematol 1969;42:330–35.
52. TCGA Network. Genomic classification of cutaneous melanoma. Cell 2015;161:1681–96.

Published OnlineFirst September 28, 2018; DOI: 10.1158/2326-6066.CIR-18-0342

53. Hume DA. The many alternative faces of macrophage activation. Front Immunol 2015;6:370.
54. Hume DA, Mabbott N, Raza S, Freeman TC. Can DCs be distinguished from macrophages by molecular signatures? Nat Immunol 2013;14:187.
55. Kunicki MA, Amaya Hernandez LC, Davis KL, Bacchetta R, Roncarolo M-G. Identity and diversity of human peripheral Th and T regulatory cells defined by single-cell mass cytometry. J Immunol 2018;200:336–46.
56. Rangwala S, Tsai KY. Roles of the immune system in skin cancer. Br J Dermatol 2011;165:953–65.
57. Akbani R, Akdemir Kadir C, Aksoy BA, Albert M, Ally A, Amin Samirkumar B, et al. Genomic classification of cutaneous melanoma. Cell 2015;161:1681–96.
58. Mignogna C, Scali E, Camastra C, Presta I, Zeppa P, Barni T, et al. Innate immunity in cutaneous melanoma. Clin Exp Dermatol 2017;42:243–50.
59. Bender C, Hassel JC, Enk A. Immunotherapy of melanoma. Oncol Res Treat 2016;39:369–76.
60. Saadatpour A, Lai S, Guo G, Yuan G-C. Single-cell analysis in cancer genomics. Trends Genet 2015;31:576–86.

Published OnlineFirst September 28, 2018; DOI: 10.1158/2326-6066.CIR-18-0342

Cancer Immunology Research

Immune Cell Gene Signatures for Profiling the Microenvironment of Solid Tumors

Ajit J. Nirmal, Tim Regan, Barbara B. Shih, et al.

Cancer Immunol Res 2018;6:1388-1400. Published OnlineFirst September 28, 2018.