Immune cell gene signatures for profiling the microenvironment of solid tumours

Immune cell gene signatures for profiling the microenvironment of solid tumours

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

[1]The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Edinburgh, EH5 9RG, UK.

[2]Applied Bioinformatics of Cancer, Edinburgh Cancer Research Centre, Institute of Genetics and Molecular Medicine, University of Edinburgh, Crewe Road South, Edinburgh, EH4 2XU, UK.

[3]Mater Research-University of Queensland, Translational Research Institute, 37 Kent St, Woolloongabba, Qld 4160, Australia.

Running title: Immune cell gene signatures for profiling solid tumours

Keywords: Gene expression, tissue immune cells, immune signatures, network analysis

Financial support: AJN is a recipient of The Roslin Institute and CMVM scholarship and Edinburgh Global Research Scholarship. AHS is funded by Breast Cancer Now, TR, BJS and TCF are funded by MRC consortium grants (MR/M003833/1, MR/L014815/1) and TCF is funded by an Institute Strategic Grant from the Biotechnology and Biological Sciences Research Council (BBSRC) (BB/JO1446X/1).

Author contributions: AJN performed the majority of work described here with assistance from TR, & BJS. AJN, DAH, AHS and TCF wrote and edited the manuscript. TCF supervised the project.

Corresponding author

Tom C. Freeman,
Systems Immunology Group,
The Roslin Institute and Royal (Dick) School of Veterinary Studies,
University of Edinburgh,
Easter Bush, EH25 9RG.
T: +44 (0)131 651 9203
F: +44 (0)131 651 9105
tom.freeman@roslin.ed.ac.uk

Conflict of Interest Disclosure: The authors declare no potential conflicts of interest.

Word count: 6432
Total number of Figures and tables: 5 figures, 3 tables, 5 supplementary tables, 3 supplementary figures.

Abstract

The immune composition of the tumour microenvironment has been shown to regulate processes including angiogenesis, metastasis and the response to drugs or immunotherapy. To facilitate the characterisation of the immune component of tumours from transcriptomics data, a number of immune cell transcriptome signatures have been reported, i.e. lists of marker genes that together are 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 have been shown not to reflect the differentiation or activation state of similar cells within tissues, including tumours, and consequently perform poorly on tissue data. 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 they can be used for the quantitative estimation of the immune content of tumour and non-tumour tissue samples. The utility of *ImSig* is demonstrated through the stratification of melanoma patients into immuno-subgroups of prognostic significance and the identification of immune cells from single-cell RNA-Seq data of derived from tumours. *ImSig* is available as an R package ('imsig').

Introduction

Modulating the activity of the immune component of the tumour microenvironment holds great potential in the treatment of cancer. Checkpoint inhibitors are perhaps the most exciting advance in cancer therapy in the past decade, with anti-PD1 and CTLA4 antibodies, in particular, demonstrating remarkable therapeutic results in some patients (1). However, multiple factors within the tumour microenvironment are recognised to influence the response to immunotherapy, in particular, the immune infiltrate prior to treatment (2). Immunohistochemistry and flow cytometry have conventionally been used to study the immune status of tumours, but are limited by the fact that histological analyses are limited to small areas of tissue and a small numbers of markers, and flow cytometry 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 main approaches are currently used to infer the relative proportion of cell types from transcriptomic data. A first type of approach fits reference gene expression profiles from sorted cells to the data in question (4-7) and a second approach, employs 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 these genes are not necessarily cell type-specific in their expression and use supervised learning algorithms to leverage the additional power 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 report to detect varies considerably. For instance, collectively the published studies report gene signatures for 22 T cell subtypes. Among the signatures that define marker genes, numerous markers are used interchangeably to define different subtypes and many are expressed by non-immune cell types. Another, shortfall of these signatures is that they are all derived from cultured or blood-derived cells. The expression profiles of the same immune cell from blood (PBMC's) and tissues are significantly different (15) which compromises the predictive value of signatures (16).

Genes that contribute to a common biological process or define a given cell type are frequently co-regulated, i.e. coexpressed giving rise to expression modules (17,18). We have previously validated gene correlation network (GCN) analysis of large gene expression datasets from human (including cancer), mouse, pig and sheep, as a means to define such expression modules (19-21). Here we have analysed a broad range of human tissue transcriptomic data to identify a set of robustly co-expressed marker genes representing seven immune cell types and three cellular pathway processes present in many tissue data. We have named this set of signatures, *ImSig*. We demonstrate the advantages of *ImSig* over other reported signatures derived from the comparison of isolated blood cells and its utility in characterising the immune microenvironment of tumours.

Methods

Derivation of *ImSig*

Eight publically available expression datasets derived from human tissue were sourced from the Gene Expression Omnibus (GEO) database (22) (GSE11318, GSE50614, GSE75214, GSE38832, GSE23705, GSE24383, GSE58812, GSE65904), based on the criteria that the unprocessed data files were available, they included a variety of normal and diseased samples, represented a variety of array platforms and contained >20 samples (median size 114 samples). The datasets was chosen such as to include the diverse variety of immune cell types and differentiation states. Raw Affymetrix data was processed using oligo package (23) and Illumina data was processed using lumi package (24) in R. The signal intensities were normalised using the robust multi-array average (RMA) and genes with multiple probes were summarised into one by choosing the probe with maximum intensity across samples.

The resultant expression matrix was loaded into the network analysis tool Graphia Professional (Kajeka Ltd., Edinburgh, UK), previously known as BioLayout *Express*^{3D} (25,26). Within the tool, a correlation network was generated (an *r* value was chosen so as to include approximately 10,000 genes in the analysis) for each dataset and clustered using the Markov Clustering (MCL) algorithm (27). Clusters were manually annotated based on domain knowledge and with the help of Gene Ontology (GO) and Reactome pathway enrichment analyses (28,29). The gene modules representing immune cell types and biological processes were identified for each of the eight datasets. 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 processes, these genes were further refined/filtered using eight independent validation datasets (GSE9891, GSE14580, GSE38832, GSE14951, GSE15773, GSE7305, GSE22619, GSE52171) by the following procedure: Robust cell type/pathway signatures were identified by excluding genes that were poorly co-expressed using an unbiased approach. Each dataset was loaded into Graphia (*r* values were 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. Based on this model, we estimated the number of genes that might contribute to noise within the signatures, and should be filtered out. To facilitate such inverse estimation, the 'investr' package in R 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 datasets and the set of genes that survived the filtration process were defined as *ImSig*. In essence, the approach sought to identify the most robustly correlated genes across datasets 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 sourced from the literature (5,8,10-14). To visualise the concordance between the immune genes defined by the different studies, a chord diagram was built using circlize package (30) in R. Only genes reported as markers of immune cells were used – *ImSig* includes pathway signatures, other studies included signatures for other cells, e.g. fibroblast, endothelial cells etc. Due to the sheer variety of T cell subtype signatures, these were further explored to identify gene usage between them. 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 visualised 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. For the Newman *et.al* signature genes that were not common between cell types were only considered. For visualisation of the results, genes pertaining to cell subsets (Treg, Th1) were all pooled to represent the parent population (T cells) and the Jaccard similarity index was re-calculated.

Comparative analysis of gene signatures in the context of a tissue dataset

Seven immune signatures were sourced from the literature (5,8,10-14). The LM22 signature (5) did not provide an absolute signature, i.e. same genes may represent multiple cell types and so only a subset of genes that were unique to cell types was used for this analysis. The median correlation of the signature genes was calculated within the context of a dataset (GSE20436) generated from swabs taken from the eyes of children with symptoms of trachoma or controls (32). The dataset contains transcriptomics data generated from samples taken 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* (possibly who were in the resolution stage); and 20 individuals with symptoms and active infections. This dataset was chosen due to the well documented immune infiltration associated with this disease and the presence of all immune populations defined by *ImSig*. To be able to directly compare with *ImSig*, genes pertaining to cell subsets were all pooled to represent the parent population. In addition, analysis of the median correlation of non-pooled signatures, i.e. marker sets representing sub-populations of cells, were also analysed in the context of these data.

To validate *ImSig* in tumours, transcriptomic data from single-cell suspensions from lymph nodes of four metastatic melanoma patients were analysed (GSE93722) for which cell type proportions (CD4 T cells, CD8 T cells, B cells, NK cells) measured with flow cytometry was available. In order to perform a direct comparison proportions of CD4 and CD8 T cells were summed to estimate total T cell content. The average expression of *ImSig* genes were calculated to determine the relative abundance of immune cells in each patient. The predicted and observed abundance were then scaled between 0 and 1 to be comparable. This analysis also served to validate the applicability of *ImSig* to RNA-Seq data. To assess the ability of *ImSig* to define known clinical differences between patient subgroups and to illustrate the explorative power of a network-based analysis, we used the trachoma dataset 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 wider context of the immune environment and extrapolate immune subsets, a GCN (r >0.7) was visualised 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 plotted.

Pan-cancer analysis of tumour data (TCGA)

Pre-normalised (level 3 data) transcriptomic data from 12 cancers were downloaded from the 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 co-expressed indicating the likely absence or low abundance of these cells and so were not included in the survival analysis. A univariate Cox-proportional hazard ratio analysis was performed for the rest using the survcomp package in R (33). P-values are based on the log-rank test.

Molecular subtyping (patient stratification) of melanoma

RNA-Seq data for the SKCM (human skin cutaneous melanoma) was 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-sample correlation plot, grouped the patients into 5 clusters. These groupings were mapped as a class-set onto the complete GCN to study the expression patterns of immune cells between groups. A univariate Cox-proportional analysis was also performed using the survcomp package (33) in R between the groups in various combinations. The P-value was calculated using the log-rank test.

An independent melanoma dataset- GSE65904 (51) was used for validation. The dataset 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 normalisation using the lumi package (24) in R, samples that were not present in the network graph (r ≥ 0.8) were also removed and the remaining samples (210) were processed as described above for the TCGA dataset.

Processing and analysis of single-cell RNA-Seq data

Single-cell transcriptomics data (log2 [(TPM/10)+1]) for melanoma (34) and head and neck cancer (HNSCC) (35) 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 the results. Therefore, to compensate for dropouts, a diffusion-based imputation method was used to impute missing values (36).

To validate the cell type specificity of *ImSig*, the average expression of B, T, NK cell and macrophage signature genes were calculated from the melanoma cell data dataset and compared to the average expression of the other immune-related *ImSig* genes. To evaluate the concordance between estimated abundance and measured number of cells, the average 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 visually 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 plotted as a stacked bar plot scaled to 100%.

In order to predict immune cell types in the HNSCC dataset using the SVM-based algorithm Cibersort, a reference matrix (*ImSig* as features) was first generated using the melanoma single-cell data as per the requirements. The algorithm was run with the generated reference matrix and HNSCC single-cell data, uploaded on to the Cibersort web portal (https://cibersort.stanford.edu). The output contained a score of 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, e.g. 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 of *ImSig*

We implemented *ImSig* as an R package called “imsig”. Users should call the “imsig” function, which takes a normalized gene expression matrix (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 network graph of *ImSig* genes and perform survival analysis using the package. A short 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, a set of co-expressed gene modules associated with human tissue immune cell populations and frequently observed biological processes were identified from eight independent tissue transcriptomics datasets. An illustrative example of a gene correlation network

(GCN) is shown in Fig. 1A. These initial gene signatures were further refined and validated by testing for co-expression of the genes associated with each signature across an additional eight independent datasets (Fig. 1B). The result was 569 marker genes representative of seven immune populations (B cells (37 genes), plasma cells (14), monocytes (37), macrophages (78), neutrophils (47), NK cells (20), T cells (85)) and three biological processes (Interferon response (66), translation (86), proliferation (99)), named collectively *ImSig* (Table 1,2 & Supplementary Table S1). The data-driven definition of each immune signature is internally-validated by the association of many well-known markers with the specific 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 extensive reference to the literature, supported the association of the majority of markers identified with the relevant cell types and processes. The top 5 enrichment terms for all signatures are listed in Supplementary Table S2 and the top term is given in Fig. 1C. In contrast to a number of the published immune gene signatures, we did not define signatures for immune cell sub-types, such as sub-populations of T cells or activation states of macrophages. Across the diversity of tissue datasets, we found no support for distinct modules of co-expressed markers describing T cell or macrophage subpopulations. This is consistent with previous analyses of isolated human macrophages responding to different stimuli, which did not support the existence of distinct activation states of macrophages but rather a continuum of difference states depending on the stimulus (37). Where present, 'activation-specific' transcripts such as receptors, cytokines or transcription factors, tend to form part of the overall cell expression module. By inference, if a particular gene is strongly co-expressed with a particular cell type-specific signature in the context of a particular dataset, one can conclude that either it is likely expressed by those cells or at least a sub-population of them.

Comparison between *ImSig* and published immune signatures

The gene content of seven published immune signatures, all derived from the comparison of isolated blood cells (5,8,10-14), were compiled and compared, excluding signatures for non-immune cell types, e.g. endothelial cells, fibroblast etc. When *ImSig* was added to the list it contained 3,658 genes (Supplementary Table S3). To compare these the gene signatures a Jaccard similarity index was calculated (Supplementary Table S4) and highlights the poor concordance between signatures (Supplementary Table S4 and Supplementary Fig. S1). The highest observed similarity was between *ImSig's* and Becht *et al.'s* B cell signature, Jaccard score = 0.26, which in itself is a not a high Jaccard score. Fig. 2A illustrates the lack of consensus between published signatures and *ImSig*, and highlights the fact that 76.3% of genes are only associated with a single study. Of these 2,794 genes, only a small proportion described unique populations, e.g. erythroblast (297 genes) and megakaryocyte (259) described by Watkins *et al.* The poor conservation of immune marker genes across studies is likely due to a number of technical and statistical artefacts. 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 (38) and are therefore poor markers of cell type. Notably, of all signatures proposed, *ImSig* contains the fewest unique genes (only 60 *ImSig* genes have not been previously been included in other signatures), suggesting a high degree of consensus with other studies overall, but not particularly with any previous signature alone.

It is also interesting to note the association of certain genes 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. As Fig. 2B illustrates the assignment of markers to cell types across studies is highly complicated. For example, *LRRN3*, was used to define resting cytotoxic T cells by Abbas *et al.* and as a Th1 marker by Bindea *et al. CTLA4* is annotated as either a marker of Tregs, Th1 and CD4 T cells and by Angelova *et al.*, Bindea *et al.,* and Watkins *et al.*, respectively. *CTLA4* can also be expressed on CD8+ T cells (39). 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 co-expression module. To compare the co-expression of the *ImSig* signatures to previous signatures, the median correlation of each set of signature genes were calculated within the context of a dataset derived trachoma patients. This was selected as one of the few examples we could find of a dataset derived from a tissue, where all immune cell types defined by *ImSig* are present, these being recruited in response to a bacterial infection. For comparison with previous signatures, those modules representing sub-populations, e.g. T cell subsets were collated into one, e.g. T cells. Their median correlation in the context of the trachoma dataset is shown in Fig. 2C. A non-collated version of the results is provided in Supplementary Table S5. Regardless of whether they were aggregated by broad cell type, or considered separately; none of the blood-derived modules were strongly co-expressed across the set of trachoma patient samples. In contrast, all of the *ImSig* signatures displayed a high median correlation (co-expression) value. Of the other signatures examined, Becht *et al.* (8) performed next best. The bacterial infection that gives rise to the pathology of trachoma leads a significant increase in the 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 dataset. As seen in Fig. 2D, a significant increase in all immune populations is associated with patient groups relative to controls, particularly in those patients with an active infection.

Finally, to validate the applicability of *ImSig* on RNA-Seq data and in the context of tumour biology, we computed the relative abundance of immune cells in four metastatic melanoma patients for which single-cell suspensions were collected from lymph nodes. A fraction of the cell suspension was used to measure cell type proportions by flow cytometry and the other fraction was used for bulk RNA sequencing. We observed a good agreement (r = 0.91, RMSE = 0.1 and P value = 2.74E-05) between predictions of relative cell number made using *ImSig* and experimentally determined cell numbers (see also Supplementary Fig. S2). This indicates that the relative cell numbers were accurately predicted 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 co-expressed 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*, *IL17A* etc. 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 classical M1 markers. Network analysis also supports the wider appreciation of the transcriptional signatures of other cell types present in clinical samples, i.e. when examining the dataset as a whole, many other GCN clusters can be assigned to other cell populations or processes.

Satisfied with the performance of *ImSig* in the context of tissue transcriptomics data in general, we set out to explore 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 datasets, however, this analysis was not correlated with outcomes (20). To test the use of *ImSig* in the study of the tumour microenvironment, the twelve largest TCGA cancer datasets were examined and hazard ratios were computed between high and low immune cell infiltrate groups (Fig. 4A). Whilst the survival analysis was not adjusted for potentially confounding variables (such as tumour stage, grade, age or treatment), the findings were largely consistent with the literature. In melanoma (SKCM), we reaffirmed the known association between tumour infiltrating lymphocytes (TIL) and a good prognosis (40,41). Breast cancer (BRCA) is not as immunogenic as melanoma, but several studies have associated TIL's with a good prognosis as observed here (42). A negative association between TIL's and prognosis was evident in low-grade glioma (LGG) (43,44) and lung squamous cell carcinoma (LUSC) (45,46) in accordance with the previous literature. A novel finding was of the potential prognostic value of the interferon response in low-grade glioma. Another surprising observation was that a high rate of proliferation is associated with a good prognosis in LUSC and colorectal cancers (COAD). This observation has been reported previously in colorectal cancer (47), but not in LUSC. Analysis of individual proliferation-related genes in LUSC also 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 counter-intuitive observation is that the mitotic signal in these tissues originates from proliferating immune cells, not from cancer itself (48,49).

Extending the analysis above, a molecular subgrouping of melanoma based on *ImSig* was performed i.e. only the signature genes were used in the grouping of patient samples. Unsupervised clustering based on the immune profile revealed five groups of patient samples (Fig. 4B). Clinical features such as the tissue of origin and tumour type (metastatic or primary) did not affect the subtyping. Nearly half the patients were in cluster-1, characterised by a low level of immune infiltrate (Fig. 4C). Hazard ratio (HR) analysis between these low immune (cluster-1) and high immune infiltrate (clusters-2 and -3) tumours revealed a significant difference in survival (HR: 0.38, p = 3E-9). The median survival of patients in the high immune group was 10 years greater than that of patients in the low immune subgroup (Fig. 4D). Within the high immune subgroup, cluster-2 appeared to have a higher level of B cells and plasma cells in contrast to cluster-3 (Fig. 4C) but overall survival (HR) was not significantly different between the two groups (Fig. 4D). Cluster-4 samples displayed higher levels of the interferon response genes and also showed improved survival compared to the low immune group (Fig. 4D). Finally, patients in cluster-5 had a low immune infiltrate but were enriched for keratin related genes and presented the worst survival rates (median survival = 2.34 yr). Whilst patients in clusters-2 and cluster-4 did not show a significant difference in hazard ratio compared to those in cluster-3, they could potentially show other features, such as differing responses to treatment. Following an analogous analysis, we were able to reproduce the five patient groupings on an independent validation dataset (GSE65904) which showed a similar infiltration pattern (Supplementary Fig. S3A) and survival analysis on the same exhibited similar prognostic pattern (Supplementary Fig. S3B). High immune and keratin subgroups have been identified and described previously in melanoma (50,51) but these studies did not describe the type and variation in the immune infiltrate in melanomas. Our analysis provides a greater degree of granularity as to the exact nature of the immune landscape of these tumours and consequently improved the prognostic power.

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

To extend these analyses and further validate the *ImSig* signatures in the context of single-cell data, we examined single-cell data derived from melanomas (34). 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 to the true proportion. The estimated proportion displayed a high degree of concordance with the measured number of cells (p < 0.05), with the poorest observed correlation being *r* = 0.97. Randomised permutation analysis with the same sized gene sets produced no significant correlation (Fig. 5B). Fig. 5C illustrates the concordance between the measured and estimated number of cells.

The single-cell community depends on gene markers/signatures and clustering algorithms, to define cell types. Here we have attempted to show the utility of *ImSig* when used in association of classification algorithms, such as support vector machine (SVM), to predict cell types from single-cell RNA-Seq data. To demonstrate such 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 dataset from head and neck tumours (HNSCC) (35). The immune component of the HNSCC dataset contained 1,473 cells. Prediction using *ImSig* yielded a high degree of accuracy for B cells (88.4%), macrophages (98.8%) and T cells (99.8%) (Table 3). 63 immune cells failed to be categorised into one of the cell types described above (p-value > 0.05). With respect to the other 4,087 cells, i.e. myocytes, mast cells, malignant cells, fibroblast, dendritic cells and endothelial cells, only 2.2% of cells were misclassified as macrophages, B or T cells. In contrast, Cibersort's default blood-derived signature (LM22) showed limited ability to identify immune cell types in these data, with an accuracy rate for B cells of 15.2%, macrophages, 0% and T cells, 75.3%. However, LM22 signature was not designed to deconvolute single-cell data and its poor performance is likely a cumulative outcome of using a blood-derived signature and a reference gene matrix based on microarrays.

Discussion

Cellular heterogeneity is a hallmark of cancer, both in terms of the tumours themselves and the normal cells that both support and control their growth. There is now a wealth of transcriptomics data generated from cancer samples and there have been a number of previous studies that report approaches to deconvolute these data in an attempt 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 (Fig. 1A) when large collections of transcriptomics data are subjected to network-based correlation analysis (18,52). Whilst 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 co-expressed genes for each cell immune cell type directly 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 signalling, proliferation (mitosis) and translation, as these are commonly observed co-expression modules in tissue and act as additional controls. Validatory analysis of the resultant *ImSig* signatures showed the gene signatures to be highly enriched with appropriate GO terms (Fig. 1C) 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 co-expression of the *ImSig* signatures across a wide range of datasets not used for their derivation and their average expression following changes in immune cell numbers, where known, e.g. in trachoma.

As the current study is by no means the first to attempt to define sets of signatures for immune cells, we sought to compare *ImSig* with other published signatures, both in terms of gene content and performance. Definition of cell signatures is not trivial, nor is simple to compare signatures across studies. In the first instance, the published gene signatures all vary in terms of the number of genes they include and the cell populations and sub-populations they seek to define. Secondly, there is no benchmark dataset 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, and where common to more than one study, there was a highly complex relationship between the assignation of genes to cells across studies; in other words, there is little consensus across published immune marker lists (Figs. 2A&B). What was apparent is that of all the signatures, *ImSig* contained the fewest unique genes (65), suggesting that rather than the gene content of *ImSig* being particularly novel, it represents more of a consensus view of other studies, despite being derived independently from them. The comparison of the performance of signatures again represented a challenge. Where multiple subtypes of cells were defined, the genes associated with subtypes were either analysed separately or collapsed into a single signature. We chose to compare the performance of these summarised signatures in the context of the trachoma dataset, where we knew all immune cell types to be present and that their relative level increases during active infection (32). In this context, the degree of co-expression between genes associated with individual *ImSig* signatures was in many cases dramatically better than others (Fig. 2C). Furthermore, the average expression of *ImSig* signatures mirrored the known increase in immune cell infiltrate during across patient groups (32) (Fig. 2D).

Ever since the first description of major types of immune cells, researchers have sought to define sub-types, i.e. sub-populations and activation states associated with different tissues, developmental stages and pathologies. Whilst heterogeneity amongst immune cell populations undoubtedly exists, the number of markers that definitively identify them outside of the context of flow cytometry and immunohistochemical experiments or comparison of isolated populations, is limited. For instance, tissue macrophages are named differently 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). Across the previous studies referred to here, 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 addition, in a given pathological state multiple cellular subtypes or populations whose biology is adapted to different niches are likely to be present. We would argue that it is unrealistic to expect to be able 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 (37). Even amongst different myeloid populations, i.e. monocytes, macrophages and neutrophils, we have found very few markers that are entirely specific to one population or another, and the markers selected to define the presence of these populations, do so more by their co-expression than absolute expression in the context of tissue.

Whilst 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, network analysis can provide a comprehensive picture of the immune microenvironment. By examination of the genes that closely correlate with the core signature genes (Fig. 3B), even if one cannot with any degree of certainty assign their expression to one cell type or another, it is possible to capture the overall profile the immune microenvironment of a tissue in health or disease. It may after all be the sum of the individual parts that matter. How one translates these finding into immune subset identification we leave to the individual analyst, with the cellular subtypes they recognise and the marker genes that define them.

After satisfying ourselves of the validity of *ImSig* and its superiority over other signatures in defining immune populations in tissue data, we used it to analyse a broad spectrum of large transcriptomics datasets derived from 12 cancer types. In each case, the majority of signature genes were tightly co-expressed, apart from instances where we believe the target cell was not present or there in low abundance. When the samples for each tumour type were ranked according to their immune cell content (as defined by the average expression of the signature genes), we were able to demonstrate a clear variation in the immune microenvironment between tumours and the association of specific immune cell populations with a good or poor prognoses (Fig. 4A). Despite an established association between the immune system and survival in melanoma (56), there has been little effort to subgroup patients based upon specific immune cell types present, previous studies merely defining tumours as having a high or low immune content (51,57). 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 (50,51) with tumours that have: high levels of T cells, macrophages (cluster 3); a high interferon enrichment (cluster 4); and tumours with high B cell infiltration (cluster 2). This analysis highlights the fact that by treating the immune infiltrate of tumours as an overall signature, loses 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, e.g. if T cells were associated with a good prognosis and macrophages a bad prognosis. Understanding the immune heterogeneity tumours may also be key in predicting their response 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 tumour 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 rapidly and may in the future answer many of the questions about immune cell heterogeneity. However, at the present time, the data available is limited and the droplet-based RNA sequencing 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 define the type and relative abundance of immune cells in single-cell data derived from melanoma, and head and neck cancer with a high degree of accuracy. This both further validates the signatures and demonstrates how they may be used in this context. As the quantity and quality of single-cell cancer datasets improve and we understand the expression profile of these cells in many contexts is better appreciated, perhaps then reliable markers may be defined that are able to differentiate between immune subtypes or activation states, specifically in the context of the tumour microenvironment.

ImSig is the first immune signature to be directly derived from tissue data. Although its gene content is not necessarily novel in the context of those reported previously, we believe it to be superior to published immune signatures in terms of being a robust, subtype agnostic means to estimate the relative abundance of these cells across tissue samples. We also demonstrate the ability of *ImSig* to be a powerful companion for the identification of novel biomarkers when applied in the context of network co-expression analyses. We anticipate that *ImSig* will prove to be a valuable resource for studying immune cell variation in tumour samples and how they respond to therapy, aiding in the discovery of novel predictive biomarkers.

References

1. Postow MA, Callahan MK, Wolchok JD. Immune Checkpoint Blockade in Cancer Therapy. Journal of Clinical Oncology 2015;33(17):1974-82.
2. Denkert C, von Minckwitz G, Darb-Esfahani S, Lederer B, Heppner BI, Weber KE, et al. Tumour-infiltrating lymphocytes and prognosis in different subtypes of breast cancer: a pooled analysis of 3771 patients treated with neoadjuvant therapy. The Lancet Oncology 2018;19(1):40-50.
3. Hackl H, Charoentong P, Finotello F, Trajanoski Z. Computational genomics tools for dissecting tumour-immune cell interactions. Nat Rev Genet 2016;17(8):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 Biology 2016;17(1):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(12):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 Biology 2016;17(1):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 Bioinformatics 2013;14(1):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(4):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(7):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 Biology 2015;16(1):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 tumour tissue using single-cell RNA-seq data. Nature Communications 2017;8(1):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(1):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(5643):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(7493):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 tumour microenvironment. BMC Genomics 2013;14(1):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 Biology 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(19):2363-67.

24. Du P, Kibbe WA, Lin SM. lumi: a pipeline for processing Illumina microarray. Bioinformatics 2008;24(13):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(19):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 Research 2003;13(11):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. Infection and Immunity 2010;78(11):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(22):3206-08.

34. 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 (New York, NY) 2016;352(6282):189-96.

35. 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(7):1611-24.e24.

36. 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.

37. 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(2):274-88.

38. 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.

39. McCoy KD, Le Gros G. The role of CTLA-4 in the regulation of T cell immune responses. Immunology And Cell Biology 1999;77:1.

40. Ladanyi A. Prognostic and predictive significance of immune cells infiltrating cutaneous melanoma. Pigment Cell & Melanoma Research 2015;28(5):490-500.

41. 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. Journal of Investigative Dermatology 2013;133(2):509-17.

42. West NR, Kost SE, Martin SD, Milne K, deLeeuw RJ, Nelson BH, et al. Tumour-infiltrating FOXP3+ lymphocytes are associated with cytotoxic immune responses and good clinical outcome in oestrogen receptor-negative breast cancer. Br J Cancer 2013;108(1):155-62.

43. 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. Clinical Cancer Research 2016;22(11):2778.

44. 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).

45. 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 Science 2008;99(8):1595-602.

46. 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 Journal of Gastroenterology : WJG 2015;21(34):9966-73.

47. 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. Journal of Clinical Pathology 2016;69(3):209-14.

48. 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(7648):105-09.

49. Kallinikos-Maniatis A. Megakaryocytes and Platelets in Central Venous and Arterial Blood. Acta Haematologica 1969;42(6):330-35.

50. Network TCGA. Genomic Classification of Cutaneous Melanoma. Cell 2015;161(7):1681-96.

51. 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(14):12297-309.

52. 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. The Journal of Pathology 2017:n/a-n/a.

53. Hume DA. The Many Alternative Faces of Macrophage Activation. Frontiers in Immunology 2015;6:370.

54. Hume DA, Mabbott N, Raza S, Freeman TC. Can DCs be distinguished from macrophages by molecular signatures? Nature Immunology 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. The Journal of Immunology 2018;200(1):336-46.

56. Rangwala S, Tsai KY. Roles of the Immune System in Skin Cancer. The British journal of dermatology 2011;165(5):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(7):1681-96.

58. Mignogna C, Scali E, Camastra C, Presta I, Zeppa P, Barni T, et al. Innate immunity in cutaneous melanoma. Clinical and Experimental Dermatology 2017;42(3):243-50.

59. Bender C, Hassel JC, Enk A. Immunotherapy of Melanoma. Oncology Research and Treatment 2016;39(6):369-76.

60. Saadatpour A, Lai S, Guo G, Yuan G-C. Single-cell analysis in cancer genomics. Trends in genetics : TIG 2015;31(10):576-86.

Tables

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*

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*

Table-3: Identification of immune cells within single-cell data. *ImSig* was used in conjunction with the SVM based classifier Cibersort, to identify immune cells within the head and neck tumour (HNSCC) single-cell data. The table shows the accuracy of identification. 63 immune cells were unassigned as its p-value was greater than 0.05.

Cells	Correct prediction	Wrong prediction	Accuracy (%)	Error (%)
B cells	122	16	88.4	11.6
Macrophages	84	1	98.8	1.2
T cells	1185	2	99.8	0.2
Other cells (4087 cells)		93		2.3

Figure Legends

Figure 1: Derivation of *ImSig*. (A) An illustrative example of a correlation network generated from a tissue dataset where nodes represent unique genes and edges represent correlations between them above a defined threshold. Groups of nodes sharing the same colour 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 and red represents genes that were discarded. This approach was applied to eight tissue datasets (only 2 shown here), the most robustly coexpressed genes across the datasets being 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.

Figure 2: Comparison of *ImSig* with other published signatures. (A) Chord diagram showing the overlap between marker genes across studies. In most studies, a significant proportion of genes were unique to the signatures defined by them, while *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 (98 genes i.e. 13.4%) for this plot. Nodes are sized relative to the number of shared genes between one signature and others. *ImSig* was found to be inclusive of genes describing various subtypes and was the most conserved set among all studies compared. **(C)** Heatmap of the median correlation between genes from published signatures as assessed in the context of the trachoma dataset (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 also available (Supplementary Table S4). **(D)** Bar plots of the average expression of signature genes (estimated relative abundance) across the dataset, each bar representing the average expression of signature genes in an individual patient sample. Samples are ordered according to T cell content, low-high, (left-right) and this order is maintained for other plots.

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 coloured according to the different immune cell types they represent, while the genes co-clustering with the *ImSig* immune genes are shown as nodes without colour 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.

Figure 4: Application of *ImSig* to tumour 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/tumour type. Samples were then ordered according to each signature (low-high, black plot in each square) and the hazard ratio 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-value < 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 plot based on expression of *ImSig* genes in melanoma patients 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 are the patient groupings as shown in B. **(D)** Univariate Cox proportional analysis between the patient groups as defined in B.

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 to 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 ten patients. For five patients (P1, P3, P5, P7, 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.

Figure 1

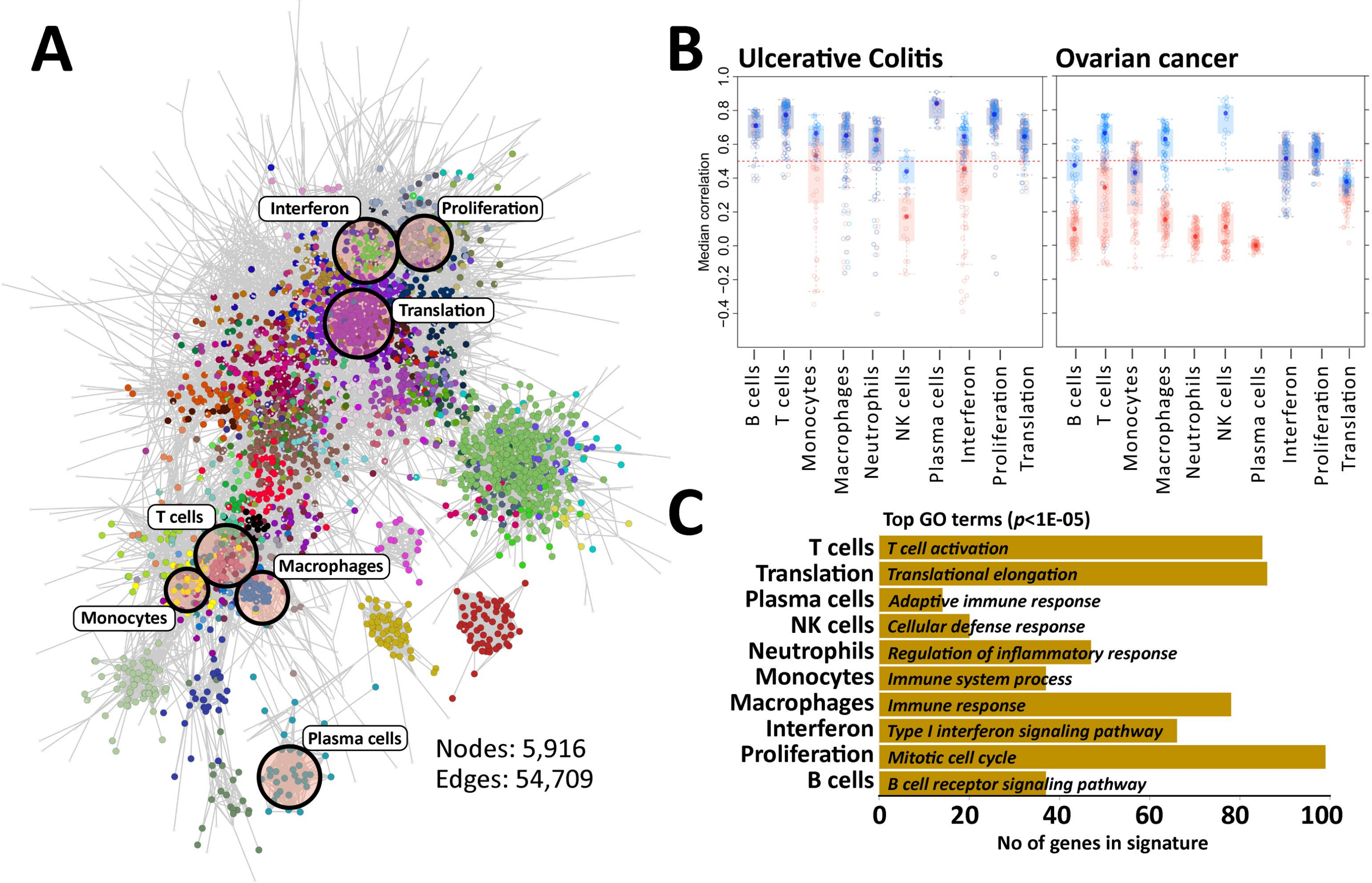

A
Interferon
Proliferation
Translation
T cells
Macrophages
Monocytes
Plasma cells
Nodes: 5,916
Edges: 54,709
B
Ulcerative Colitis
Ovarian cancer
Median correlation
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
B cells
T cells
Monocytes
Macrophages
Neutrophils
NK cells
Plasma cells
Interferon
Proliferation
Translation
C
Top GO terms (p<1E-05)
T cells
T cell activation
Translation
Translational elongation
Plasma cells
Adaptive immune response
NK cells
Cellular defense response
Neutrophils
Regulation of inflammatory response
Monocytes
Immune system process
Macrophages
Immune response
Interferon
Type I interferon signaling pathway
Proliferation
Mitotic cell cycle
B cells
B cell receptor signaling pathway
0
20
40
60
80
100
No of genes in signature

Figure 2

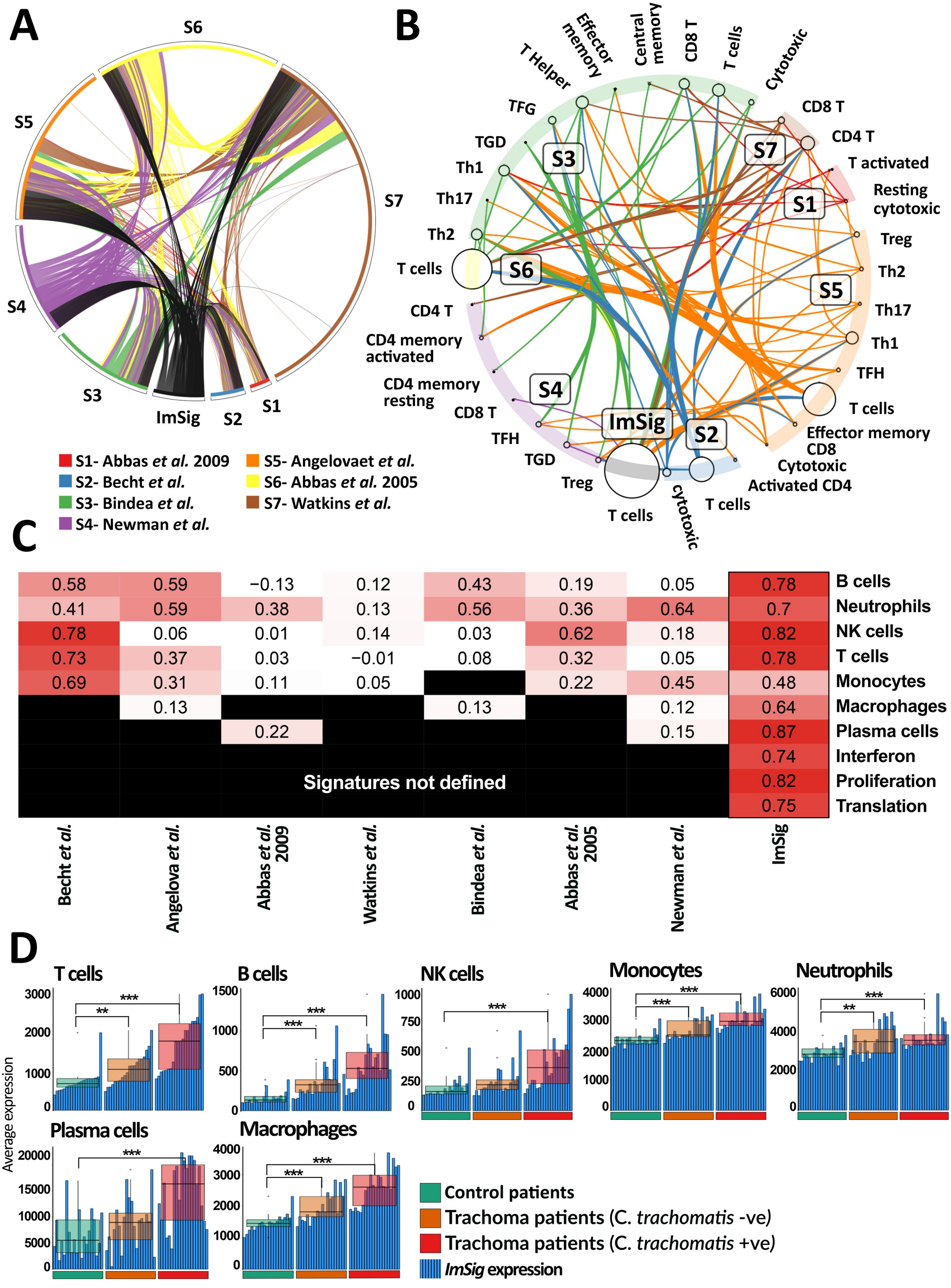

A
S6
S5
S7
S4
S3
ImSig
S2
S1
S1- Abbas et al. 2009
S2- Becht et al.
S3- Bindea et al.
S4- Newman et al.
S5- Angelovaet et al.
S6- Abbas et al. 2005
S7- Watkins et al.
B
T Helper
Effector memory
Central memory
CD8 T
T cells
Cytotoxic
TFG
TGD
Th1
Th17
Th2
T cells
CD4 T
CD4 memory activated
CD4 memory resting
CD8 T
TFH
TGD
Treg
T cells
cytotoxic
T cells
Activated CD4
Cytotoxic
Effector memory CD8
T cells
TFH
Th1
Th17
Th2
Treg
Resting cytotoxic
T activated
CD4 T
CD8 T
S3
S7
S1
S6
S5
S4
ImSig
S2
C
Becht et al.: 0.58, 0.41, 0.78, 0.73, 0.69
Angelova et al.: 0.59, 0.59, 0.06, 0.37, 0.31, 0.13
Abbas et al. 2009: −0.13, 0.38, 0.01, 0.03, 0.11, 0.22
Watkins et al.: 0.12, 0.13, 0.14, −0.01, 0.05
Bindea et al.: 0.43, 0.56, 0.03, 0.08, 0.13
Abbas et al. 2005: 0.19, 0.36, 0.62, 0.32, 0.22
Newman et al.: 0.05, 0.64, 0.18, 0.05, 0.45, 0.12, 0.15
ImSig: 0.78, 0.7, 0.82, 0.78, 0.48, 0.64, 0.87, 0.74, 0.82, 0.75
B cells
Neutrophils
NK cells
T cells
Monocytes
Macrophages
Plasma cells
Interferon
Proliferation
Translation
Signatures not defined
D
T cells
B cells
NK cells
Monocytes
Neutrophils
Plasma cells
Macrophages
Average expression
0
1000
2000
3000
500
1500
250
750
4000
6000
5000
10000
15000
20000
***
**
Control patients
Trachoma patients (C. trachomatis -ve)
Trachoma patients (C. trachomatis +ve)
ImSig expression

Figure 3

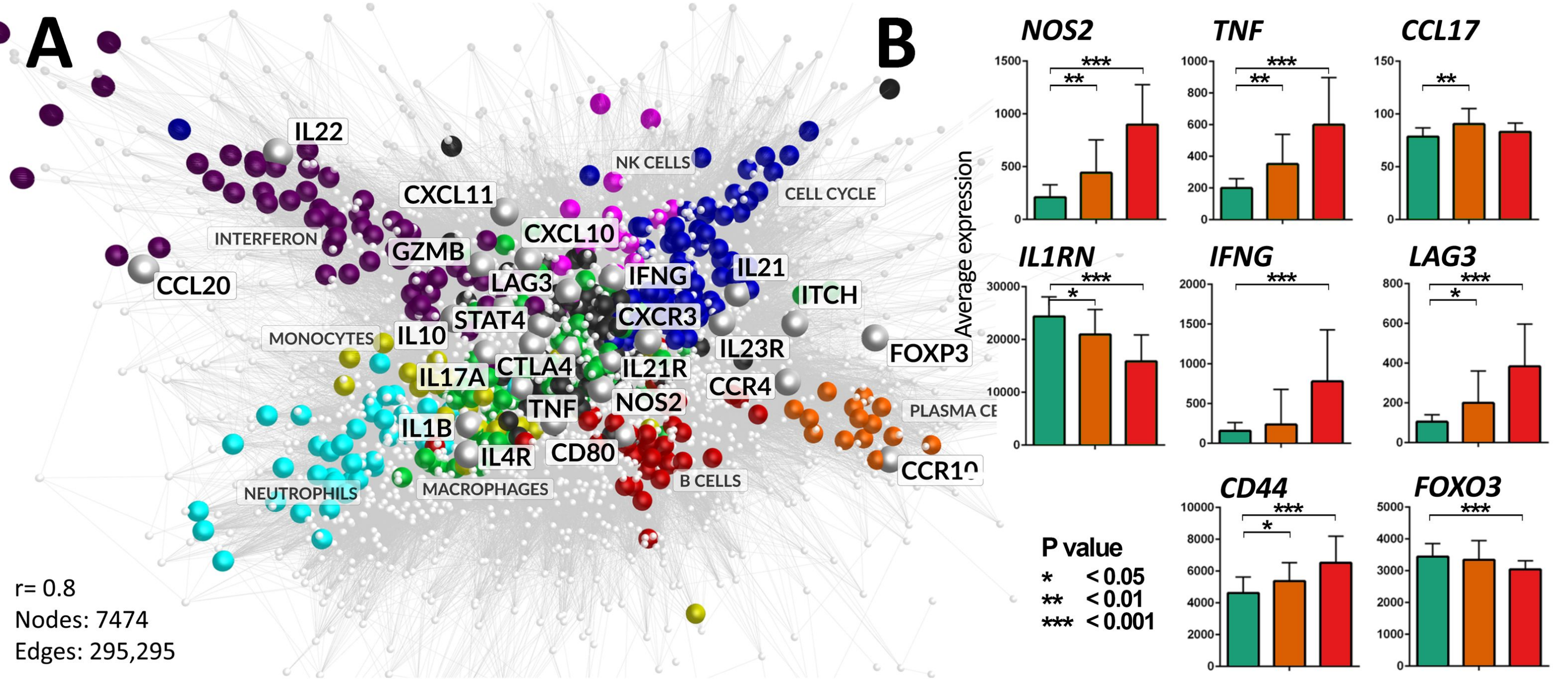
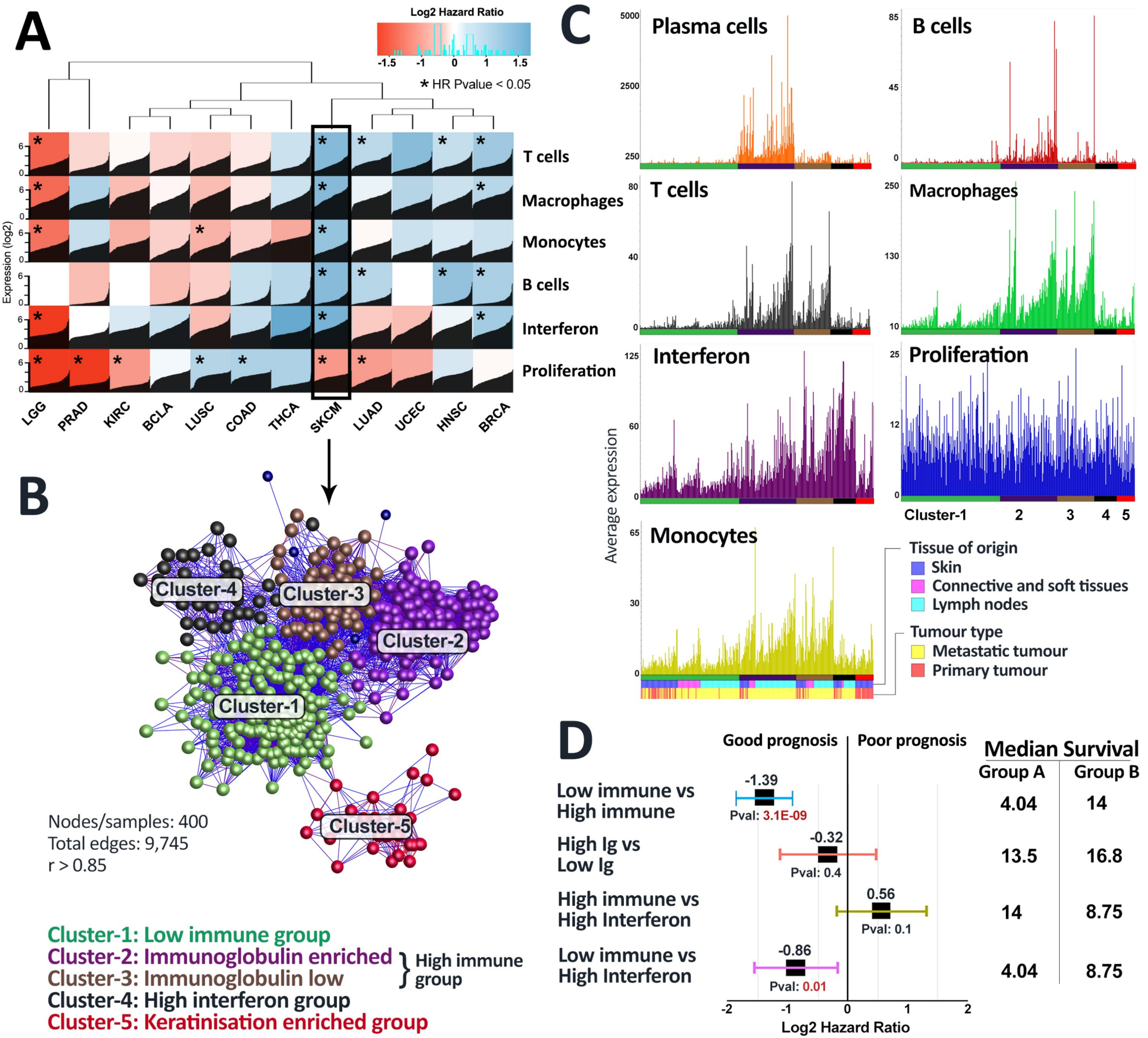
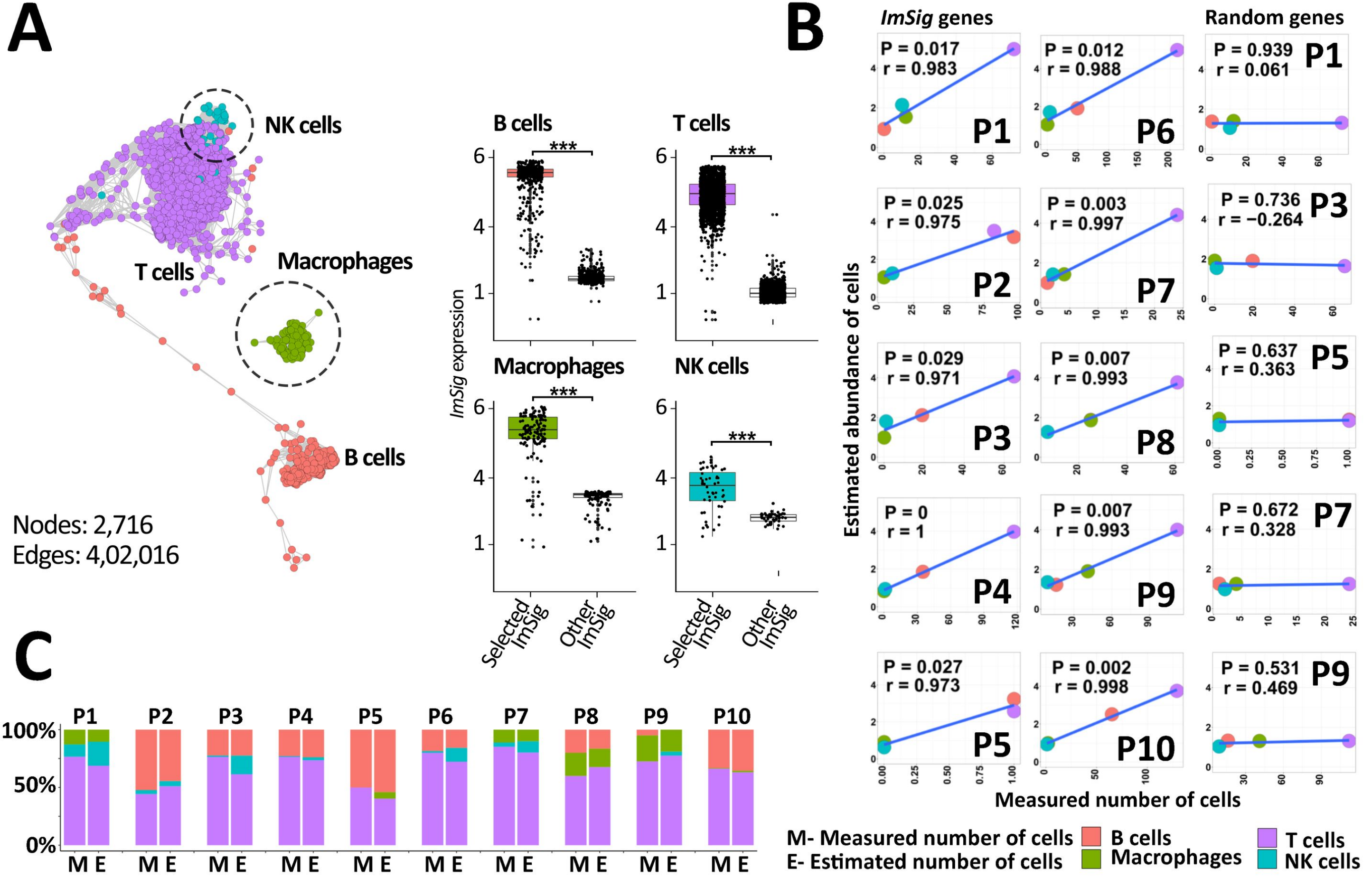

A
IL22
CXCL11
CXCL10
NK CELLS
CELL CYCLE
INTERFERON
GZMB
CCL20
LAG3
IFNG
IL21
ITCH
STAT4
CXCR3
MONOCYTES
IL10
IL23R
FOXP3
CTLA4
IL17A
IL21R
CCR4
NOS2
TNF
PLASMA CE
IL1B
CD80
IL4R
CCR10
NEUTROPHILS
MACROPHAGES
B CELLS
r= 0.8
Nodes: 7474
Edges: 295,295
B
Average expression
NOS2
0
500
1000
1500
**
***
TNF
0
200
400
600
800
1000
**
***
CCL17
0
50
100
150
**
IL1RN
0
10000
20000
30000
*
***
IFNG
0
500
1000
1500
2000
***
LAG3
0
200
400
600
800
*
***
CD44
0
2000
4000
6000
8000
10000
*
***
FOXO3
0
1000
2000
3000
4000
5000
***
P value
* < 0.05
** < 0.01
*** < 0.001

Figure 4



A
Log2 Hazard Ratio
-1.5
-1
0
1
1.5
* HR Pvalue < 0.05
Expression (log2)
6
0
T cells
Macrophages
Monocytes
B cells
Interferon
Proliferation
LGG
PRAD
KIRC
BCLA
LUSC
COAD
THCA
SKCM
LUAD
UCEC
HNSC
BRCA
B
Cluster-1
Cluster-2
Cluster-3
Cluster-4
Cluster-5
Nodes/samples: 400
Total edges: 9,745
r > 0.85
Cluster-1: Low immune group
Cluster-2: Immunoglobulin enriched
Cluster-3: Immunoglobulin low
High immune group
Cluster-4: High interferon group
Cluster-5: Keratinisation enriched group
C
Plasma cells
5000
2500
250
B cells
85
45
0
T cells
80
40
0
Macrophages
250
130
10
Interferon
125
60
0
Proliferation
25
12
0
Cluster-1
2
3
4
5
Average expression
Monocytes
65
30
0
Tissue of origin
Skin
Connective and soft tissues
Lymph nodes
Tumour type
Metastatic tumour
Primary tumour
D
Good prognosis
Poor prognosis
Median Survival
Group A
Group B
Low immune vs High immune
-1.39
Pval: 3.1E-09
4.04
14
High Ig vs Low Ig
-0.32
Pval: 0.4
13.5
16.8
High immune vs High Interferon
0.56
Pval: 0.1
14
8.75
Low immune vs High Interferon
-0.86
Pval: 0.01
4.04
8.75
-2
-1
0
1
2
Log2 Hazard Ratio

Figure 5


A
NK cells
T cells
Macrophages
B cells
Nodes: 2,716
Edges: 4,02,016
B cells
T cells
Macrophages
NK cells
***
ImSig expression
6
4
1
Selected ImSig
Other ImSig
B
ImSig genes
Random genes
P = 0.017
r = 0.983
P1
P = 0.012
r = 0.988
P6
P = 0.939
r = 0.061
P1
P = 0.025
r = 0.975
P2
P = 0.003
r = 0.997
P7
P = 0.736
r = −0.264
P3
P = 0.029
r = 0.971
P3
P = 0.007
r = 0.993
P8
P = 0.637
r = 0.363
P5
P = 0
r = 1
P4
P = 0.007
r = 0.993
P9
P = 0.672
r = 0.328
P7
P = 0.027
r = 0.973
P5
P = 0.002
r = 0.998
P10
P = 0.531
r = 0.469
P9
Estimated abundance of cells
Measured number of cells
C
P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
100%
50%
0%
M E
M- Measured number of cells
E- Estimated number of cells
B cells
Macrophages
T cells
NK cells