Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data

Running title: Enumeration of cancer and immune cell types.

Authors: Julien Racle[a,b], Kaat de Jonge[c], Petra Baumgaertner[c], Daniel E. Speiser[c] and David Gfeller[*,a,b]

[a]Ludwig Centre for Cancer Research, Department of Fundamental Oncology, University of Lausanne, CH-1066 Epalinges, Switzerland

[b]Swiss Institute of Bioinformatics (SIB), CH-1015 Lausanne, Switzerland

[c]Department of Fundamental Oncology, Lausanne University Hospital (CHUV), CH-1066 Epalinges, Switzerland

[*]Corresponding author:

David Gfeller
Computational Cancer Biology
Ludwig Centre for Cancer Research, UNIL
Ch. des Boveresses 155
CH-1066 Epalinges
Switzerland
Tel: +41 (0)21 692 59 83
Fax: +41 (0)21 692 59 95
E-mail: david.gfeller@unil.ch

Keywords: Tumor immune microenvironment, gene expression analysis, cell fraction predictions, computational biology.

Abstract

Immune cells infiltrating tumors can have important impact on tumor progression and response to therapy. We present an efficient algorithm to simultaneously estimate the fraction of cancer and immune cell types from bulk tumor gene expression data. Our method integrates novel gene expression profiles from circulating and tumor infiltrating cells for each major immune cell type, cell-type specific mRNA content and the ability to model uncharacterized, and possibly highly variable, cell types. Feasibility is demonstrated by validation with flow cytometry, immunohistochemistry and single-cell RNA-Seq analyses of human melanoma and colorectal tumor specimens. Altogether, our work not only improves accuracy but also broadens the scope of absolute cell fraction predictions from tumor gene expression data, and provides a unique novel experimental benchmark for immunogenomics analyses in cancer research.

Introduction

Tumors form complex microenvironments composed of various cell types such as cancer, immune and stromal cells (Hanahan & Weinberg, 2011; Joyce & Fearon, 2015). Immune cells infiltrating the tumor microenvironment play a major role in shaping tumor progression, response to (immuno-)therapy and patient survival (Fridman, Pagès, Sautès-Fridman, & Galon, 2012). Today, gene expression analysis is widely used to characterize tumors at the molecular level. As a consequence, tumor gene expression profiles from tens of thousands of patients are available across all major tumor types in databases such as Gene Expression Omnibus (GEO (Edgar, Domrachev, & Lash, 2002)) or The Cancer Genome Atlas (TCGA (Hoadley et al., 2014)). Unfortunately, flow cytometry or immunohistochemistry (IHC) measurements to quantify the number of both malignant and tumor infiltrating immune cells are rarely performed for samples analyzed at the gene expression level. Therefore, to correctly interpret these data in particular from an immuno-oncology point of view (Angelova et al., 2015; Gentles et al., 2015; Hackl, Charoentong, Finotello, & Trajanoski, 2016; B. Li et al., 2016; Linsley, Chaussabel, & Speake, 2015; Rooney, Shukla, Wu, Getz, & Hacohen, 2015; Şenbabaoğlu et al., 2016; Zheng, Zhang, Wu, & Wu, 2017), reliable and carefully validated bioinformatics tools are required to infer the fraction of cancer and immune cell types from bulk tumor gene expression data.

To this end, diverse bioinformatics methods have been developed. Some aim at estimating tumor purity based on copy number variation (Carter et al., 2012; B. Li & Li, 2014), or expression data (Quon et al., 2013; Yoshihara et al., 2013), but do not provide information about the different immune cell types. Others focus on predicting the relative proportions of immune cell types by fitting reference gene expression profiles from sorted immune cells (Gong & Szustakowski, 2013; B. Li et al., 2016; Newman et al., 2015) or with help of gene signatures (Becht et al., 2016; Zhong, Wan, Pang, Chow, & Liu, 2013). These approaches have been recently applied to cancer genomics data to investigate the influence of immune infiltrates on survival (Gentles et al., 2015; Şenbabaoğlu et al., 2016) or predict potential targets for cancer immunotherapy (Angelova et al., 2015; B. Li et al., 2016). However, none of these methods provides quantitative information about both cancer and immune cell type proportions directly from tumor gene expression profiles. In addition, reference gene expression profiles used in previous studies have been mainly obtained from circulating immune cells sorted from peripheral blood and were generally based on microarrays technology. Finally, several of these approaches have not been experimentally validated in solid tumors from human patients.

Here, we developed a robust approach to simultaneously Estimate the Proportion of Immune and Cancer cells (EPIC) from bulk tumor gene expression data. EPIC is based on a unique collection of RNA-Seq reference gene expression profiles from either circulating or tumor infiltrating immune cell types. To account for the high variability of cancer cells across patients and tissue of origin, we implemented in our algorithm the ability to consider uncharacterized, possibly highly variable, cell types. To validate our predictions in human solid tumors, we first analyzed melanoma samples with both flow cytometry and RNA-Seq. We then collected publicly available IHC and single-cell RNA-Seq data of colorectal and melanoma tumors. All three validation datasets showed that very accurate predictions of both cancer and immune cell type proportions could be obtained even in the absence of *a priori* information about cancer cell gene expression profiles.

Results

Reference gene expression profiles from circulating and tumor infiltrating immune cells

EPIC incorporates reference gene expression profiles from each major immune cell type to model bulk RNA-Seq data as a superposition of these reference profiles (Figure 1A,B). To tailor our predictions to recent gene expression studies, we first collected and curated RNA-Seq profiles of various human innate and adaptive circulating immune cell types (Hoek et al., 2015; Linsley, Speake, Whalen, & Chaussabel, 2014; Pabst et al., 2016) (T, B, NK, Monocytes and Neutrophils) from a diverse set of patients analyzed in different centers (see *Materials and Methods*). Principal component analysis (PCA) of these data (Figure 1C) showed that samples clustered first according to cell type and not according to experiment of origin, patient age, disease status or other factors, suggesting that they could be used as *bona fide* reference expression profiles across different patients. Reference gene expression profiles for each major immune cell type were built from these RNA-Seq samples based on the median normalized counts per gene and cell type. The variability in expression for each gene was also considered when predicting the various cell proportions based on these reference profiles (see *Materials and Methods* and Supplementary Files 2-3).

Immune cells differ in their gene expression profiles depending on their state and site of origin (e.g., blood or tumors). To study the potential effect of these differences on our predictions, we further established reference gene expression profiles of each major tumor infiltrating immune cell type (i.e., T, B, NK, macrophages). These were obtained as cell type averages from the single-cell RNA-Seq data of melanoma patients from Tirosh and colleagues (Tirosh et al., 2016), considering only samples from primary tumor and non-lymphoid tissue metastasis (see *Materials and Method* and Supplementary File 4). As for circulating immune cell data, principal component analysis of the tumor infiltrating immune cell gene expression profiles showed that samples clustered first according to cell type (Figure 1D and Figure 1-figure supplement 1, see also results in (Tirosh et al., 2016)).

Cancer and immune cell fraction predictions

Reference gene expression profiles from each of these immune cell types were then used to model bulk gene expression data as a linear combination of *m* different cell types (Figure 1B). To include cell types like cancer cells that show high variability across patients and tissues of origin, we further implemented in our algorithm the ability to consider an uncharacterized cell population. Mathematically this was done by taking advantage of the presence of gene markers of immune cells that are not expressed in cancer cells. Importantly, we do not require immune marker genes to be expressed in exactly one cell type, but only to show very low expression in non-immune cells. The mRNA proportion of each immune cell type was inferred using least-square regression, solving first our system of equations for the immune marker genes (green box in Figure 1B, see *Materials and Methods*). The fraction of cancer or other non-immune cells was then determined as one minus the fraction of all immune cell types. Immune cell markers used in this work were determined by differential expression analysis based on our reference immune cell gene expression profiles as well as gene expression data from non-hematopoietic tissues (see *Materials and Methods* and Supplementary File 1-table S1). Finally, to account for different amounts of mRNA in different cell types and enable meaningful comparison with flow cytometry and IHC data, we measured the mRNA content of all major immune cell types as well as of cancer cells (Figure 1-figure supplement 2) and used these values to renormalize our predicted mRNA proportions (see *Materials and* Methods).

Validation in blood

We first tested our algorithm using three datasets comprising bulk RNA-Seq data from PBMC (Hoek et al., 2015; Zimmermann et al., 2016) or whole blood (Linsley et al., 2014), as well as the corresponding proportions of immune cell types determined by flow cytometry (Figure 2A). These data were collected from various cancer-free human donors (see *Materials and Methods*). Overall, very accurate predictions were obtained by fitting reference profiles from circulating immune cells. When comparing with other widely used immune cell fraction prediction methods (Becht et al., 2016; Gong & Szustakowski, 2013; B. Li et al., 2016; Newman et al., 2015; Quon et al., 2013; Zhong et al., 2013), we observed a clear improvement (Figure 2B and Figure 2-figure supplement 1). Of note, the renormalization by mRNA content, which had not been considered in previous approaches, appeared to be important for predicting actual cell fractions (Figure 2C).

Validation in solid tumors

To validate our predictions in tumors, we collected single cell suspensions from lymph nodes of four metastatic melanoma patients (see *Materials and Methods*). A fraction of the cell suspension was used to measure the different cell type proportions with flow cytometry (T, B, NK, melanoma and other cells comprising mostly stromal and endothelial cells; Supplementary File 5A), and the other fraction was used for bulk RNA sequencing (Figure 3-figure supplement 1). EPIC was first run with reference profiles from circulating immune cells. We observed a remarkable agreement between our predictions and experimentally determined cell fractions (Figure 3A). Of note, the proportion of melanoma cells could be very accurately predicted even in the absence of *a priori* information about their gene expression.

As a second validation, we compared EPIC predictions with IHC data from colon cancer (Becht et al., 2016) and melanoma (Jönsson et al., 2010) (see *Materials and Method*). Although a limited number of immune cell types had been assayed in these two datasets, we observed a good agreement between cell proportions measured by IHC and our predictions (Figure 3B,C).

As a third validation, we used recent single-cell RNA-Seq data from 19 melanoma samples (Tirosh et al., 2016). We applied EPIC on the average expression profile over all single cells for each patient and compared the results with the actual cell fractions (see *Materials and Methods*). Here again, our predictions were consistent with the observed cell fractions, even for melanoma cells for which we did not assume any reference gene expression profile (Figure 3D).

We next compared these predictions to those obtained with reference profiles from tumor infiltrating immune cells (Figure 4). For the single-cell RNA-Seq data (Figure 4D), we applied a leave-one-out procedure, avoiding to use the same samples both to build the reference profiles and the bulk RNA-Seq data used as input for the predictions (see *Materials and Methods*). Overall, predictions did not change much compared to those based on circulating immune cell reference gene expression profiles (Figure 4). Interestingly however, we can observe some differences between the results obtained from circulating immune cell reference gene expression profiles and those from tumor infiltrating cell reference gene expression profiles, when considering the proportions from each cell type independently (Figures 3-4 and Figure 4-figure supplement 1): (i) predictions for NK cells improved in all datasets; (ii) predictions for macrophages improved in the datasets of primary tumors and non-lymph node metastases but were less accurate in the datasets of lymph node metastases; (iii) predictions for T cells based on the blood circulating reference cells show high accuracy in all datasets but predictions based on tumor infiltrating cells are only good in the datasets of primary tumors and non-lymph node metastases; and (iv) predictions for B cells display similar accuracy based on the circulating cells or tumor infiltrating cells profiles for all datasets.

Benchmarking of other methods

We took advantage of our unique collection of independent validation datasets to benchmark other methods for predictions of immune cell type fractions in human tumors. We first compared the results of EPIC and ISOpure, which is the only other method that can consider uncharacterized cell types and therefore predict the fraction of cancer and immune cell types based only on RNA-seq data. EPIC displayed improved accuracy in all three datasets (Figure 5A, and Figure 5-figure supplements 1-4). To benchmark other methods, we then restricted our analysis to the predictions of the different immune cell types (Figure 5B and Figure 5-figure supplements 1-4). Predictions from EPIC were in general more accurate, especially when considering all cell types together. Nevertheless, when restricting the comparisons to relative cell type proportions, some methods like MCPcounter and TIMER were quite consistent in their predictions across the various datasets and showed similar accuracy as EPIC (Figure 5-figure supplements 1-4). Of note, MCPcounter could not be included in the global prediction comparison as this method returns scores that are not comparable between different cell types. Predictions from DSA were also quite accurate when available, but in multiple cases some cell type proportions returned by the method were simply equal to 0 in all samples (Figure 5-figure supplements 1-4).

Immune and tumor purity scores (Yoshihara et al., 2013) based on gene set enrichment analysis also showed significant correlations with the total fraction of immune cells and the fraction of cancer cells. However, these correlations were significantly lower than those obtained with our approach (Figure 5C,D and Figure 5-figure supplement 5). Moreover, such scores are less quantitative and are thus more difficult to interpret with respect to actual cell type proportions.

Discussion

By combining RNA-Seq profiles of all major immune cell types established from both circulating and tumor infiltrating cells together with information about cell morphology (i.e., mRNA content) and algorithmic developments to consider uncharacterized and possibly highly variable cell types, EPIC overcomes several limitations of previous approaches to predict the fraction of both cancer and immune cell types from bulk tumor gene expression data. From an algorithmic point of view, EPIC takes advantage of the fact that cancer cells, in general, express no or only low levels of immune markers. Therefore the method can be broadly applied to most solid tumors, as confirmed by our validation in both melanoma and colorectal samples, but it will not be suitable for hematological malignancies like leukemia or lymphoma.

The accuracy of the predictions for some cell types might be sensitive to the origin or condition of the immune cells used for establishing reference profiles. For instance, we observed that macrophages from primary tumors and non-lymph node metastases samples were best predicted using the reference profiles from tumor infiltrating cells. This may be explained by the fact that the reference profiles from circulating cells corresponded to monocytes as no macrophages are circulating in blood. Interestingly however, it appears that the profiles based on circulating monocytes were better for the predictions in the lymph node metastasis samples, possibly due to the presence of some monocytes that are not differentiated to macrophages in the lymph nodes.

Overall, our results suggest that for primary tumors or non-lymphoid tissue metastases reference gene expression profiles from tumor infiltrating immune cells are more appropriate, while for lymph node metastases, profiles from circulating immune cells perform better.

One known limitation of cell fraction predictions arises when some cell types are present at very low frequency (Shen-Orr & Gaujoux, 2013). Our results suggest that predictions of cell proportions are reliable within an absolute error of about 7%, as estimated by the root mean squared error (Figure 3 and Figure 4-figure supplement 1B). These estimates are consistent with the lower detection limit proposed by other groups (Becht et al., 2016; Zhong et al., 2013) and may explain why the relative proportions of NK cells, which are present at lower frequency in melanoma tumors (Balch et al., 1990; Sconocchia et al., 2012), could not be predicted with accuracy comparable to other cell types (Figure 4-figure supplement 1). While this may prevent applications of cell fraction predictions in some tumor types that are poorly infiltrated, many other tumors, like melanoma or colorectal cancer, display high level of infiltrating immune cells and the role of immune infiltrations on tumor progression and survival appears to be especially important in these tumors (Clemente et al., 1996; Fridman et al., 2012; Galon et al., 2006).

Our predictions for the fraction of uncharacterized non-immune cells may include stromal cells or endothelial cells from neighboring tissues, in addition to cancer cells. Compared to recent algorithms that first predict tumor purity based on exome sequencing data, and later infer the relative fraction of immune cell types (B. Li et al., 2016), the predictions of EPIC are likely more quantitative because they implicitly consider the presence of not only cancer cells but also other non-immune cells like stromal cells. Moreover, EPIC does not require both exome and RNA-Seq data from the same tumor samples, thereby reducing the cost and amount of experimental work for prospective studies, and broadening the scope of retrospective analyses of cancer genomics data to studies that only include gene expression data.

Recent technical developments in single-cell RNA-Seq technology enable researchers to directly access information about both the proportion of all cell types in a tumor and their gene expression characteristics (Efroni, Ip, Nawy, Mello, & Birnbaum, 2015; Jaitin et al., 2014; Singer et al., 2016; Stegle, Teichmann, & Marioni, 2015; Tirosh et al., 2016). Such data are much richer than anything that can be obtained with computational deconvolution of bulk gene expression profiles and this technology may eventually replace standard gene expression analysis of bulk tumors. Nevertheless, it is important to realize that, even when disregarding the financial aspects, single-cell RNA-Seq of human tumors is still logistically and technically very challenging due to high level of cell death upon sample manipulation (especially freezing and thawing) and high transcript dropout rates (Finak et al., 2015; Saliba, Westermann, Gorski, & Vogel, 2014; Stegle et al., 2015). Moreover, one cannot exclude that some cells may better survive the processing with microfluidics devices used in some single-cell RNA-Seq platforms, thereby biasing the estimates of cell type proportions. It is therefore likely that bulk tumor gene expression analysis will remain widely used for several years. Our work shows how we can exploit recent single-cell RNA-Seq data of tumor infiltrating immune cells obtained from a few patients to refine cell fraction predictions in other patients that could not be analyzed with this technology, thereby overcoming some limitations of previous computational approaches that relied only on reference gene expression profiles from circulating immune cells.

Unlike some previous computational approaches, we provide here a detailed biologically relevant validation of our predictions using actual tumor samples from human patients analyzed with flow cytometry, IHC and single-cell RNA-Seq. We note that the slightly lower agreement between our predictions and IHC data may be partly explained by the fact that the exact same samples could not be used for both gene expression and IHC analyses because of the incompatibility between the two techniques. Nevertheless, the overall high accuracy of our predictions indicates that infiltrations of major immune cell types can be quantitatively studied directly from bulk tumor gene expression data using computational approaches such as EPIC.

EPIC can be downloaded as a standalone R package (available upon publication) and can be used with reference gene expression profiles pre-compiled from circulating or tumor infiltrating immune cells, or provided by the user.

Materials and Methods

Code availability

EPIC has been written as an R package. It will be freely available upon publication for academic non-commercial research purposes. Version v1.0 of the package was used for our analyses.

Prediction of cancer and immune cell type proportions

In EPIC, the gene expression of a bulk sample is modeled as the sum of the gene expression profiles from the pure cell types composing this sample (Figure 1A,B). This can be written as (Venet, Pecasse, Maenhaut, & Bersini, 2001):

$$b = C \times p \qquad (1)$$

Where b is the vector of all n genes expressed from the bulk sample to deconvolve; C is a matrix (n x m) of the m gene expression profiles from the different cell types; and p is a vector of the proportions from the *m cell types* in the given sample (Figure 1B).

Matrix C consists of *m-1* columns corresponding to various reference non-malignant cell types whose gene expression profiles are known, and one column corresponding to uncharacterized cells (i.e. mostly cancer cells, but possibly also other non-malignant cell types not included in the reference profiles). EPIC assumes the reference gene expression profiles from the non-malignant cell types are well conserved between patients. Such a hypothesis is supported by the analysis in Figure 1C,D. The uncharacterized cells can be more heterogeneous between patients and EPIC makes no assumption on them.

EPIC finds the proportions of all cells in the sample by first performing a consistent normalization, similar to transcripts per million (TPM) normalization, both for the reference cells and for the bulk sample (see Supplementary File 1-method S1), and then by solving eq. (1) for a subset of n_S equations corresponding to the n_S signature genes (S) that are expressed by one or more of the normal cell types but only expressed at a negligible level in the other cells (Figure 1B). Importantly, such cell specific signature genes are well established and widely used in flow cytometry to sort immune cells. Thus, EPIC solves the following system of equations:

$$\bar{b}_i\big|_{i\in S} = (\bar{C}^{*}\times\bar{p}^{*})_i|_{i\in S} \tag{2}$$

where $\bar{C}^{*}$ and $\bar{p}^{*}$ are the matrix of the normalized profiles and vector of proportions from all the reference cell types, and the term corresponding to the *uncharacterized* cells proportions vanished thanks to the definition of the signature genes. The solution to eq. (2) can be estimated by a constrained least square optimization, forcing each proportion to be bigger than zero and their sum smaller than one.

When solving this constrained least square optimization, EPIC also takes advantage of the known variability for each gene in the reference profile: a weight, based on this variability, is given for the fit of each gene in order to force more precise fits of genes that display less variability in the reference profiles (see Supplementary File 1-method S2).

Finally, the proportion for the *uncharacterized* cells can be obtained by:

$$\bar{p}_{\text{unchar-cell}} = 1 - \sum_{j \in \text{normal-cells}} \bar{p}_j \qquad (3)$$

Since we used normalized gene expression data, values of $\bar{p}$ correspond actually to the fraction of mRNA coming from each cell type, rather than the cell proportions. As the mRNA content per cell type can vary significantly (Figure 1-figure supplement 2), the actual proportions of each cell type can be estimated as:

$$p_j = \alpha \cdot \frac{\bar{p}_j}{r_j} \qquad (4)$$

where r_j is the amount of RNA nucleotides in cell type j (or equivalently the total weight of RNA in each cell type) and α is a normalization constant to have $\sum p_j = 1$.

Flow cytometry and gene expression analysis of melanoma samples

Patients agreed to donate metastatic tissues upon informed consent, based on dedicated clinical investigation protocols established according to the relevant regulatory standards. The protocols were approved by the local IRB, i.e. the “Commission cantonale d’éthique de la recherche sur l’être humain du Canton de Vaud”. Lymph nodes (LN) were obtained from stage III melanoma patients, by lymph node dissection that took place before systemic treatment. The LN from one patient was from the right axilla and the

LNs from the other three patients were from the iliac and ilio-obturator regions (Supplementary File 1-table S2). Single cell suspensions were obtained by mechanical disruption and immediately cryopreserved in RPMI 1640 supplemented with 40% FCS and 10% DMSO. Single cell suspensions from four lymph nodes were thawed and used in parallel experiments of flow cytometry and RNA extraction. In order to limit the number of dead cells after thawing, we removed those cells using a dead cell removal kit (Miltenyi Biotech). Proportions of T (CD45$^+$/CD3$^+$/Melan-A$^-$), NK (CD45$^+$/CD56$^+$/CD3$^-$/CD33$^-$/Melan-A$^-$), B (CD45$^+$/CD19$^+$/CD3$^-$/CD33$^-$/Melan-A$^-$) and Melan-A expressing tumor cells (Supplementary File 5A) were acquired via flow cytometry using the following antibodies: anti-CD3 BV711 (clone: UCHT1, BD Biosciences), anti-CD56 BV421 (clone: HCD56, Biolegend), anti-CD19 APCH7 (clone: SJ25C1, BD Biosciences), anti-CD33 PE-Cy7 (clone: P67.6, BD Biosciences), anti-CD45 APC (clone: HI30, Biolegend), anti-Melan-A FITC (clone: A103, Santa Cruz Biotechnologies) and Fixable Viability Dye eFluor 455UV (eBioscience). Data was acquired on a BD LSR II SORP flow cytometry machine (BD Bioscience). Analysis was performed using FlowJo (Tree Star). Cell proportions were based on viable cells only. In parallel total RNA was extracted using the RNAeasy Plus mini kit (Qiagen) following the manufactures' protocol. Starting material always contained minimum 0.2x10^6 cells. RNA was quantified and integrity was analyzed using a Fragment Analyser (Advanced Analytical). Total RNA from all samples used for sequencing had an RQN ≥ 7. Libraries were obtained used the Truseq stranded RNA kit (Illumina). Single read (100bp) was performed using an Illumina HiSeq 2500 sequencer (Illumina).

Post processing of the sequencing was done using Illumina pipeline Casava 1.82. FastQC (version 0.10.1) was used for quality control. The reads obtained were mapped to the human genome, *hg19*, with *TopHat* (Kim et al., 2013) version 2.0.13 using default parameters and *Bowtie2* (Langmead & Salzberg, 2012) version 2.2.4, followed by sorting with *Samtools* (H. Li et al., 2009) version 1.2. Raw counts were then obtained with *HTSeq* (Anders, Pyl, & Huber, 2015), version 0.6.1, using the options "*-i gene_name –s no –t exon –m union*".

RNA-Seq data from this experiment will be deposited on a public database upon publication.

Amount of mRNA per cell type

Healthy donor peripheral blood was obtained through the blood transfusion center in Lausanne. PBMCs were purified by density gradient using Lymphoprep (Axis-Shieldy). Mononuclear cells were stained in order to sort monocytes, B, T and NK cells using the following antibodies: CD14 FITC (Clone: RMO52, Beckman Coulter), CD19 PE (clone: 89B, Beckman Coulter), CD3 APC (clone UCHT1, Beckman Coulter), CD56 BV421 (Clone: HCD56, Biolegend) and fixable live/dead near IR stain (ThermoFisher Scientific). 1 x 10^6 live cells from each cell type were sorted using the BD FACS ARIA III (BD Biosciences). Total RNA was extracted using the RNAeasy Plus mini kit (Qiagen) following the manufactures' protocol and quantified using a Fragment Analyser (Advanced Analytical). Values obtained are given in Figure 1-figure supplement 2A.

The mRNA content for the cancer cells was estimated from the flow cytometry data described in the previous section from the four patients with melanoma. For this we used the forward scatter width, which is a good proxy of cell size and mRNA content (Padovan-Merhar et al., 2015; Tzur, Moore, Jorgensen, Shapiro, & Kirschner, 2011), and observed that cancer cells had similar amount of mRNA than B, NK and T cells (Figure 1-figure supplement 2B). We thus used a value of 0.4 pg of mRNA per cancer cell.

Public external datasets used in this study

- Dataset 1 was obtained from Zimmermann and colleagues (Zimmermann et al., 2016), through ImmPort (http://www.immport.org), accession SDY67. It includes RNA-Seq samples from PBMC of healthy donors before and after influenza vaccination. In addition, the original flow cytometry results files were available, containing multiple immune cell markers. As an independent validation of EPIC, we used the data from 12 pre-vaccination samples of healthy donors and we computed the corresponding immune cell proportions from the flow cytometry files based on a similar gating than in Hoek et al. (Hoek et al., 2015) (obtaining B, NK, T cells and monocytes, Supplementary File 5B; Figure 2A)
- Dataset 2 was obtained from Hoek and colleagues (Hoek et al., 2015), GEO accession GSE64655. This corresponds to RNA-Seq samples from 2 different donors. Samples have been taken before an influenza vaccination and also 1, 3 and 7 days after the vaccination (56 samples in total). In their experiment, the authors measured RNA-Seq from bulk PBMC samples and also from sorted immune cells (B, NK, T cells, myeloid dendritic cells, monocytes and neutrophils). In addition, flow cytometry was performed to measure the proportion of these cell types in PBMC before the influenza vaccination (personally communicated by the authors; Supplementary File 5C).
- Dataset 3 was obtained from Linsley and colleagues (Linsley et al., 2014), GEO accession GSE60424. This dataset includes 20 donors (healthy donors and other donors with amyotrophic lateral sclerosis, multiple sclerosis, type 1 diabetes or

sepsis), for a total of 134 samples. RNA-Seq from these donors has been extracted from whole blood and sorted immune cells (B, NK cells, monocytes, neutrophils, T CD4 and T CD8 cells). In our analyses, T CD4 and T CD8 cells were taken together as T cells. In addition to RNA-Seq data, complete blood counts data was available for 5 of these donors (personally communicated by the authors; Supplementary File 5D).

- Dataset 4 was obtained from Pabst and colleagues (Pabst et al., 2016), GEO accession GSE51984. This includes RNA-Seq from 5 healthy donors. Samples are from total white blood cells and sorted immune cells (B cells, granulocytes, monocytes and T cells).
- Colon cancer dataset was obtained from Becht and colleagues (Becht et al., 2016), GEO accession GSE39582. This corresponds to microarrays of primary colon cancer tumors. In addition to gene expression data, immunohistochemistry data of CD3, CD8 and CD68 was available (personally communicated by the authors) for 33 patients.
- Melanoma dataset was obtained from Jönsson and colleagues (Jönsson et al., 2010), GEO accession GSE22153. This is data from 57 patients, with biopsies obtained mostly from subcutaneous metastases but also with some coming from lymph nodes. Gene expression was measured by microarrays (we downloaded the non-normalized gene expression data). Immunohistochemistry data of CD3 and CD20 was also available (Supplementary File 5E).
- Single-cell RNA-Seq data from tumor infiltrating cells were obtained from Tirosh and colleagues (Tirosh et al., 2016), GEO accession GSE72056. This corresponds to 19 donors and comprises primary tumors, lymph node metastasis or other lesions. It includes 4,645 cells. Cell type identity was taken from Tirosh et al.

(Tirosh et al., 2016) (B, NK, T cell, macrophage, cancer-associated fibroblast, endothelial cell, cancer cell as well as cell not assigned a specific cell type). The data is given as TPM counts. In *silico* reconstructed bulk samples from each donor were obtained as the average per gene from all samples of the given donor. The corresponding cell fractions from these bulk samples are obtained as the number of cells from each cell type divided by the total number of cells (Supplementary File 5F). In the results, we split this dataset in two depending on the origin of the biopsies: lymphoid tissues for samples obtained from lymph node and spleen metastases, vs. the rest of samples, which were obtained from primary tumor and other metastases.

For the above datasets 1 and 2, we obtained raw *fastq* files. These *fastq* files were mapped to the human genome, *hg19*, with *TopHat* (Kim et al., 2013) version 2.0.13 using default parameters and *Bowtie2* (Langmead & Salzberg, 2012) version 2.2.4, followed by sorting with *Samtools* (H. Li et al., 2009) version 1.2. Raw counts were then obtained with *HTSeq* (Anders et al., 2015), version 0.6.1, using the options "*-i gene_name –s no –t exon –m union*".

For the other datasets, we directly obtained the summary counts data from these studies without mapping the reads by ourselves.

Reference gene expression profiles from circulating cells

Reference gene expression profiles of sorted immune cells from peripheral blood were built from the datasets 2, 3 and 4 described in previous section. We verified no experimental biases were present in these data through unsupervised clustering of the samples, with help of a principal component analysis based on the normalized expression from the 1000 most variable genes (Figure 1C).

Each sample was independently normalized as described in Supplementary File 1-method S1 and the median value of normalized counts was computed per cell type and per gene. Similarly, the interquartile range of the normalized counts was computed per cell type and gene, as a measure of the variability of each gene expression in each cell type. Values of these reference profiles are given in Supplementary File 2). Granulocytes from dataset 4 and neutrophils from datasets 2 and 3 were combined to build the reference profile for neutrophils (neutrophils constitute more than 90% of granulocytes). No reference profile was built for the myeloid dendritic cells as only few samples of these sorted cells existed and they were all from the same experiment. Monocytes are not found in tumors but instead there are macrophages, mostly from monocytic lineage, that are infiltrating tumors and that are not found in blood. For this reason, we also used the monocyte reference gene expression profile as a proxy for macrophages when applying EPIC to tumor samples. Such an assumption gave coherent results as observed in the results.

In addition to these *raw counts* based white blood cell reference profiles, we also built TPM based reference profiles from the same datasets (using *RSEM* (B. Li & Dewey, 2011) v.1.2.19 and *Bowtie2* (Langmead & Salzberg, 2012) version 2.2.4 instead of the *HTSeq* (Anders et al., 2015) based steps when mapping the raw reads to the human genome). This TPM based reference profile (Supplementary File 3) was used with EPIC to predict immune cell proportions for the single-cell RNA-Seq dataset (Tirosh et al., 2016) as counts in this dataset were only available as TPM.

Reference profiles from tumor infiltrating cells

We also built gene expression reference profiles from tumor infiltrating immune cells. These are based on the single-cell RNA-Seq data from Tirosh and colleagues (Tirosh et al., 2016) described above. We only used the non-lymphoid tissue samples to build these tumor infiltrating cell's profiles, avoiding in this way potential "normal immune cells" present in the lymph nodes and spleen. These reference profiles (Supplementary File 4) were built in the same way as described above for the reference profiles of circulating immune cells, but based on the mean and standard deviation instead of median and interquartile range respectively, due to the nature of single-cell RNA-Seq data and gene dropout present with such technique.

When testing EPIC with these profiles for the single-cell RNA-Seq datasets, for the samples of primary tumor and other non-lymph node metastases, a leave-one-out procedure was applied: for each donor we built reference immune cell profiles based only on the data coming from the other donors.

Immune marker gene identification

EPIC relies on signature genes that are expressed by the reference cells but not by the uncharacterized cells (e.g., cancer cells). For each reference immune cell type, we thus built a list of 20 signature genes through the following steps:

1) The samples from this immune cell type were tested for overexpression against:
   a) the samples from each other immune cell (1 cell type vs. 1 other cell type at a time);
   b) the samples of the Illumina Human Body Map 2.0 Project (ArrayExpress ID: E-MTAB-513) considering all non-immune related tissues;

   c) the samples from GTEx (GTEx Consortium, 2015) from each of the following tissues (1 tissue at a time): adipose subcutaneous; bladder; colon-transverse; ovary; pancreas; testis (data version V4).

2) Only genes overexpressed in the given cell type with an adjusted p-value < 0.01 for all these tests were kept. Conditions 1b) and 1c) are there to ensure signature genes are expressed in the immune cells and no other tissues.

3) The genes that passed 2) were then ranked by the mean fold change from the overexpression tests of 1) and the top twenty genes were selected as signature genes of the given immune cell.

4) For neutrophils, *CSF3R* was expressed at a level much higher than other genes and thus was totally biasing the cell proportion predictions towards this gene. For this reason, we removed it from neutrophils signature genes, in order to keep genes expressed at levels that are more similar.

5) For NK cells, we observed the signature genes included *GZMA, GZMB, GZMH, IFNG* and *PRF1*, which are expressed constitutively by NK cells but not by resting T cells, which comprise the majority of T cells used to build the reference profiles from circulating cells. However, these genes are also highly expressed by activated T cells, which may affect the NK and T cell fraction predictions in tumor samples. For this reason, we also removed these genes from the immune signature and replaced them by the next five best NK cell signature genes.

All the differential expression tests were performed with *DESeq2* (Love, Huber, & Anders, 2014).

Supplementary File 1-table S1 summarizes the full list of signature genes per cell type.

Prediction of cell proportions in bulk samples with other tools

We compared EPIC's predictions with those from various cell fraction prediction methods. These other methods were run with the following packages (using the default options when possible):

- CIBERSORT (Newman et al., 2015) (R package version 1.03) was run with their *LM22* reference profiles. For comparison with the experimentally measured cell proportions, we summed together the sub-types predictions of CIBERSORT within each major immune cell type.
- DeconRNASeq (v1.12) (Gong & Szustakowski, 2013) does not contain immune cell reference profiles and we used the reference profiles built in this work. We present the results with "use.scale" parameter set to FALSE, which usually gave better results.
- DSA (Zhong et al., 2013) only needs a gene signature per cell type to estimate the proportion of cells in multiple bulk samples together. We used the implementation of DSA found in CellMix (Gaujoux & Seoighe, 2013) R package (version 1.6.2). As DSA needs many samples to estimate simultaneously the proportions of cells in these samples, we considered all the PBMC samples from Hoek et al. data when fitting this dataset (8 samples) and all whole blood samples from Linsley et al. data when fitting this other dataset (20 samples), even though the cell proportions have been measured experimentally only for 2 and 5 samples respectively. For the gene signature, we used the same genes as those used for EPIC (Supplementary File 1-table S1).
- ISOpure (Quon et al., 2013) estimates the profile and proportion of cancer cells by comparing many bulk samples containing cancer cells and many healthy bulk samples of the same tissue. Although the primary goal is not to compute the

proportions of the different cell types composing a sample, cell fractions can still be obtained with this method. In particular, one output of ISOpure is how much each of the healthy reference samples is contributing to a given bulk sample. Instead of using bulk healthy samples, we used as input our immune cell reference profiles, so that each "reference sample" corresponded to a different cell type. The contribution of each cell type was taken as the relative contribution outputted by ISOpure from each of the reference cell sample. The R implementation ISOpureR (Anghel et al., 2015) version 1.0.20 was used.

- MCP-counter (Becht et al., 2016) (R package version 1.1.0) was run with the "*HUGO_symbols*" chosen as features.
- TIMER (B. Li et al., 2016) predictions were obtained by slightly adapting the available source code. The reference profiles from this method were used and predictions for T cells were defined as the sum of CD4 and CD8 T cells. In addition to bulk gene expression, tumor purity estimates based on DNA copy number variation are needed in TIMER to refine the gene signature. As this information is not available in our benchmarking datasets, we kept all the original immune gene signatures for predictions in blood. For the tumor datasets, we used the gene signatures obtained from the TCGA data for melanoma or colorectal cancer depending on the origin of cancer.
- ESTIMATE (Yoshihara et al., 2013) was run with their R package version 1.0.11.

List of abbreviations

EPIC: acronym for our method to "Estimate the Proportion of Immune and Cancer cells"; GEO: Gene Expression Omnibus; IHC: immunohistochemistry; PCA: principal component analysis; RMSE: root mean squared error; TCGA: The Cancer Genome Atlas.

Acknowledgements

We are grateful to Hélène Maby-El Hajjami for compiling the clinical data. We thank Kristen L. Hoek, Andrew Link and their colleagues (Hoek et al., 2015), Cate Speak, Scott Presnell and their colleagues (Linsley et al., 2014), Aurélien De Reynies, Etienne Becht and their colleagues (Becht et al., 2016), for providing us with additional data relating to their published studies. Computations were performed at the Vital-IT (http://www.vital-it.ch) Center for high-performance computing of the Swiss Institute of Bioinformatics.

Funding

DG and JR acknowledge the financial support of CADMOS.

Authors' contributions

J.R. and D.G. designed the study, performed the research, analyzed the data and wrote the manuscript. J.R. wrote the code. K.D.J., P.B. and D.E.S. performed and analyzed the experiments.

Competing interests

The authors declare that they have no competing interests.

References

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31*(2), 166–169. https://doi.org/10.1093/bioinformatics/btu638

Angelova, M., Charoentong, P., Hackl, H., Fischer, M. L., Snajder, R., Krogsdam, A. M., … Trajanoski, Z. (2015). Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy. *Genome Biology*, *16*(1), 64. https://doi.org/10.1186/s13059-015-0620-6

Anghel, C. V., Quon, G., Haider, S., Nguyen, F., Deshwar, A. G., Morris, Q. D., & Boutros, P. C. (2015). ISOpureR: an R implementation of a computational purification algorithm of mixed tumour profiles. *BMC Bioinformatics*, *16*(1), 156. https://doi.org/10.1186/s12859-015-0597-x

Balch, C. M., Riley, L. B., Bae, Y. J., Salmeron, M. A., Platsoucas, C. D., Eschenbach, A. von, & Itoh, K. (1990). Patterns of Human Tumor-Infiltrating Lymphocytes in 120 Human Cancers. *Archives of Surgery*, *125*(2), 200–205. https://doi.org/10.1001/archsurg.1990.01410140078012

Becht, E., Giraldo, N. A., Lacroix, L., Buttard, B., Elarouci, N., Petitprez, F., … de Reyniès, A. (2016). Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biology*, *17*, 218. https://doi.org/10.1186/s13059-016-1070-5

Carter, S. L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., … Getz, G. (2012). Absolute quantification of somatic DNA alterations in human cancer. *Nature Biotechnology*, *30*(5), 413–421. https://doi.org/10.1038/nbt.2203

Clemente, C. G., Mihm, M. C., Bufalino, R., Zurrida, S., Collini, P., & Cascinelli, N. (1996). Prognostic value of tumor infiltrating lymphocytes in the vertical growth phase of primary cutaneous melanoma. *Cancer*, *77*(7), 1303–1310. https://doi.org/10.1002/(SICI)1097-0142(19960401)77:7<1303::AID-CNCR12>3.0.CO;2-5

Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, *30*(1), 207–210. https://doi.org/10.1093/nar/30.1.207

Efroni, I., Ip, P.-L., Nawy, T., Mello, A., & Birnbaum, K. D. (2015). Quantification of cell identity from single-cell gene expression profiles. *Genome Biology*, *16*(1), 9. https://doi.org/10.1186/s13059-015-0580-x

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., … Gottardo, R. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology*, *16*(1), 278. https://doi.org/10.1186/s13059-015-0844-5

Fridman, W. H., Pagès, F., Sautès-Fridman, C., & Galon, J. (2012). The immune contexture in human tumours: impact on clinical outcome. *Nature Reviews Cancer*, *12*(4), 298–306. https://doi.org/10.1038/nrc3245

Galon, J., Costes, A., Sanchez-Cabo, F., Kirilovsky, A., Mlecnik, B., Lagorce-Pagès, C., … Pagès, F. (2006). Type, Density, and Location of Immune Cells Within Human Colorectal Tumors Predict Clinical Outcome. *Science*, *313*(5795), 1960–1964. https://doi.org/10.1126/science.1129139

Gaujoux, R., & Seoighe, C. (2013). CellMix: a comprehensive toolbox for gene expression deconvolution. *Bioinformatics*, *29*(17), 2211–2212. https://doi.org/10.1093/bioinformatics/btt351

Gentles, A. J., Newman, A. M., Liu, C. L., Bratman, S. V., Feng, W., Kim, D., … Alizadeh, A. A. (2015). The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nature Medicine*, *21*(8), 938–945. https://doi.org/10.1038/nm.3909

Gong, T., & Szustakowski, J. D. (2013). DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data.

Bioinformatics, *29*(8), 1083–1085. https://doi.org/10.1093/bioinformatics/btt090

GTEx Consortium. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, *348*(6235), 648–660. https://doi.org/10.1126/science.1262110

Hackl, H., Charoentong, P., Finotello, F., & Trajanoski, Z. (2016). Computational genomics tools for dissecting tumour-immune cell interactions. *Nature Reviews Genetics*, *17*(8), 441–458. https://doi.org/10.1038/nrg.2016.67

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of Cancer: The Next Generation. *Cell*, *144*(5), 646–674. https://doi.org/10.1016/j.cell.2011.02.013

Hoadley, K. A., Yau, C., Wolf, D. M., Cherniack, A. D., Tamborero, D., Ng, S., … Stuart, J. M. (2014). Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin. *Cell*, *158*(4), 929–944. https://doi.org/10.1016/j.cell.2014.06.049

Hoek, K. L., Samir, P., Howard, L. M., Niu, X., Prasad, N., Galassie, A., … Link, A. J. (2015). A Cell-Based Systems Biology Assessment of Human Blood to Monitor Immune Responses after Influenza Vaccination. *PLoS ONE*, *10*(2), e0118528. https://doi.org/10.1371/journal.pone.0118528

Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., … Amit, I. (2014). Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. *Science*, *343*(6172), 776–779. https://doi.org/10.1126/science.1247651

Jönsson, G., Busch, C., Knappskog, S., Geisler, J., Miletic, H., Ringnér, M., … Lønning, P. E. (2010). Gene Expression Profiling–Based Identification of Molecular Subtypes in Stage IV Melanomas with Different Clinical Outcome. *Clinical Cancer Research*, *16*(13), 3356–3367. https://doi.org/10.1158/1078-0432.CCR-09-2509

Joyce, J. A., & Fearon, D. T. (2015). T cell exclusion, immune privilege, and the tumor microenvironment. *Science*, *348*(6230), 74–80. https://doi.org/10.1126/science.aaa6204

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, *14*(4), R36. https://doi.org/10.1186/gb-2013-14-4-r36

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*(1), 323. https://doi.org/10.1186/1471-2105-12-323

Li, B., & Li, J. Z. (2014). A general framework for analyzing tumor subclonality using SNP array and DNA sequencing data. *Genome Biology*, *15*, 473. https://doi.org/10.1186/s13059-014-0473-4

Li, B., Severson, E., Pignon, J.-C., Zhao, H., Li, T., Novak, J., … Liu, X. S. (2016). Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome Biology*, *17*, 174. https://doi.org/10.1186/s13059-016-1028-7

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Linsley, P. S., Chaussabel, D., & Speake, C. (2015). The Relationship of Immune Cell Signatures to Patient Survival Varies within and between Tumor Types. *PLoS ONE*, *10*(9), e0138726. https://doi.org/10.1371/journal.pone.0138726

Linsley, P. S., Speake, C., Whalen, E., & Chaussabel, D. (2014). Copy Number Loss of the Interferon Gene Cluster in Melanomas Is Linked to Reduced T Cell Infiltrate and Poor Patient Prognosis. *PLoS ONE*, *9*(10), e109760. https://doi.org/10.1371/journal.pone.0109760

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*, 550. https://doi.org/10.1186/s13059-014-0550-8

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., … Alizadeh, A. A. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods*, *12*(5), 453–457. https://doi.org/10.1038/nmeth.3337

Pabst, C., Bergeron, A., Lavallée, V.-P., Yeh, J., Gendron, P., Norddahl, G. L., … Barabé, F. (2016). GPR56 identifies primary human acute myeloid leukemia cells with high repopulating potential in vivo. *Blood*, *127*(16), 2018–2027. https://doi.org/10.1182/blood-2015-11-683649

Padovan-Merhar, O., Nair, G. P., Biaesch, A. G., Mayer, A., Scarfone, S., Foley, S. W., … Raj, A. (2015). Single Mammalian Cells Compensate for Differences in Cellular Volume and DNA Copy Number through Independent Global Transcriptional Mechanisms. *Molecular Cell*, *58*(2), 339–352. https://doi.org/10.1016/j.molcel.2015.03.005

Quon, G., Haider, S., Deshwar, A. G., Cui, A., Boutros, P. C., & Morris, Q. (2013). Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction. *Genome Medicine*, *5*(3), 29. https://doi.org/10.1186/gm433

Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G., & Hacohen, N. (2015). Molecular and Genetic Properties of Tumors Associated with Local Immune Cytolytic Activity. *Cell*, *160*(1–2), 48–61. https://doi.org/10.1016/j.cell.2014.12.033

Saliba, A.-E., Westermann, A. J., Gorski, S. A., & Vogel, J. (2014). Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Research*, *42*(14), 8845–8860. https://doi.org/10.1093/nar/gku555

Sconocchia, G., Arriga, R., Tornillo, L., Terracciano, L., Ferrone, S., & Spagnoli, G. C. (2012). Melanoma Cells Inhibit NK Cell Functions—Letter. *Cancer Research*, *72*(20), 5428–5429. https://doi.org/10.1158/0008-5472.CAN-12-1181

Şenbabaoğlu, Y., Gejman, R. S., Winer, A. G., Liu, M., Van Allen, E. M., de Velasco, G., … Hakimi, A. A. (2016). Tumor immune microenvironment characterization in clear cell renal cell carcinoma identifies prognostic and immunotherapeutically relevant messenger RNA signatures. *Genome Biology*, *17*, 231. https://doi.org/10.1186/s13059-016-1092-z

Shen-Orr, S. S., & Gaujoux, R. (2013). Computational deconvolution: extracting cell type-specific information from heterogeneous samples. *Current Opinion in Immunology*, *25*(5), 571–578. https://doi.org/10.1016/j.coi.2013.09.015

Singer, M., Wang, C., Cong, L., Marjanovic, N. D., Kowalczyk, M. S., Zhang, H., … Anderson, A. C. (2016). A Distinct Gene Module for Dysfunction Uncoupled from Activation in Tumor-Infiltrating T Cells. *Cell*, *166*(6), 1500–1511.e9. https://doi.org/10.1016/j.cell.2016.08.052

Stegle, O., Teichmann, S. A., & Marioni, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, *16*(3), 133–145. https://doi.org/10.1038/nrg3833

Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, M. H., Treacy, D., Trombetta, J. J., … Garraway, L. A. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*, *352*(6282), 189–196. https://doi.org/10.1126/science.aad0501

Tzur, A., Moore, J. K., Jorgensen, P., Shapiro, H. M., & Kirschner, M. W. (2011). Optimizing Optical Flow Cytometry for Cell Volume-Based Sorting and Analysis. *PLoS ONE*, *6*(1), e16053. https://doi.org/10.1371/journal.pone.0016053

Venet, D., Pecasse, F., Maenhaut, C., & Bersini, H. (2001). Separation of samples into their constituents using gene expression data. *Bioinformatics*, *17*(suppl 1), S279–S287. https://doi.org/10.1093/bioinformatics/17.suppl_1.S279

Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., … Verhaak, R. G. W. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. *Nature Communications*, *4*, 2612. https://doi.org/10.1038/ncomms3612

Zheng, X., Zhang, N., Wu, H.-J., & Wu, H. (2017). Estimating and accounting for tumor purity in the analysis of DNA methylation data from cancer studies. *Genome Biology*, *18*, 17. https://doi.org/10.1186/s13059-016-1143-5

Zhong, Y., Wan, Y.-W., Pang, K., Chow, L. M., & Liu, Z. (2013). Digital sorting of complex tissues for cell type-specific gene expression profiles. *BMC Bioinformatics*, *14*(1), 1. https://doi.org/10.1186/1471-2105-14-89

Zimmermann, M. T., Oberg, A. L., Grill, D. E., Ovsyannikova, I. G., Haralambieva, I. H., Kennedy, R. B., & Poland, G. A. (2016). System-Wide Associations between DNA-Methylation, Gene Expression, and Humoral Immune Response to Influenza Vaccination. *PLOS ONE*, *11*(3), e0152034. https://doi.org/10.1371/journal.pone.0152034

Figures

Figure 1. Estimating the Proportion of Immune and Cancer cells. (**A**) Schematic description of our method. (**B**) Matrix formulation of our algorithm, including the uncharacterized cell types (red box) with no or very low expression of immune signature genes (green box). (**C**) Low dimensionality representation (PCA based on the 1000 most variable genes) of the samples used to build the reference gene expression profiles from circulating immune cells (study 1 (Hoek et al., 2015), study 2 (Linsley et al., 2014), study 3 (Pabst et al., 2016)). (**D**) Low dimensionality representation (PCA based on the 1000 most variable genes) of the tumor infiltrating immune cell gene expression profiles from different patients. Each point corresponds to cell-type averages per patient of the single-cell RNA-Seq data of Tirosh et al. (Tirosh et al., 2016) (requiring at least 3 cells of a given cell type per patient). Only samples from primary tumors and non-lymphoid tissue metastases were considered. Projection of the original single-cell RNA-Seq data can be found in Figure 1-figure supplement 1.

Figure supplement 1. Low dimensionality representation of the tumor infiltrating immune cell samples.

Figure supplement 2. Cell type mRNA content.

Figure 2. Predicting cell fractions in blood samples. (**A**) Predicted vs. measured immune cell proportions in PBMC (dataset 1 (Zimmermann et al., 2016), dataset 2 (Hoek et al., 2015)) and whole blood (dataset 3 (Linsley et al., 2014)); predictions are based on the reference profiles from circulating immune cells. (**B**) Performance comparison with other methods. (**C**) Predicted immune cells' mRNA proportions (i.e., without mRNA renormalization step) vs. measured values in the same datasets. Correlations are based on Pearson correlation; RMSE: root mean squared error.

Figure supplement 1. Comparison of multiple cell fraction prediction methods in blood datasets.

Figure 3. Predicting cell fractions in solid tumors. (**A**) Comparison of EPIC predictions with our new flow cytometry analysis of lymph nodes from metastatic melanoma patients. (**B**) Comparison with immunohistochemistry data from colon cancer primary tumors (Becht et al., 2016). (**C**) Comparison with immunohistochemistry data from melanoma samples (mostly from primary tumors) (Jönsson et al., 2010). This study only reported absence of the given marker, a non-brisk expression of it (i.e. low expression) or a brisk expression (i.e. high). One-sided Wilcoxon rank-sum tests was used to determine p-values (* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$). (**D**) Comparison with single-cell RNA-Seq data (Tirosh et al., 2016) from melanoma samples either from lymphoid tissues or primary and non-lymphoid metastatic tumors. Correlations are based on Pearson correlation.

Figure supplement 1. Sketch of the experiment designed to validate EPIC predictions starting from *in vivo* tumor samples.

Figure 4. Predictions with reference profiles from tumor infiltrating immune cells. Same as Figure 3 but based on reference profiles built from the single-cell RNA-Seq data of primary tumor and non-lymphoid metastatic melanoma samples from Tirosh et al. (Tirosh et al., 2016). (**A**) Comparison with flow cytometry data of lymph nodes from metastatic melanoma patients. (**B**) Comparison with IHC from colon cancer primary tumors (Becht et al., 2016). (**C**) Comparison with IHC from melanoma (Jönsson et al., 2010) (* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$, see Figure 3C). (**D**) Comparison with single-cell RNA-Seq data (Tirosh et al., 2016).

Figure supplement 1. Comparison of EPIC results per cell type for gene expression reference profiles from circulating or tumor infiltrating immune cells.

Figure 5. Performance comparison with other methods in tumor samples. (**A**) Pearson correlation R-values between the cell proportions predicted by EPIC and ISOpure and the observed proportions measured by flow cytometry or single-cell RNA-Seq (Tirosh et al., 2016), considering all cell types together (i.e., B, NK, T, macrophages, cancer + other cells). (**B**) Same analysis as in Figure 5A but considering only immune cell types (i.e., B, NK, T, macrophages) in order to include more methods in the comparison. The star (*) indicates a case where DSA could not predict the cell type proportions. (**C**) Analysis of ESTIMATE predictions in the single-cell RNA-Seq dataset for the sum of all immune cells and the proportion of cancer cells (cells identified as melanoma cells by Tirosh and colleagues). (**D**) Same as in Figure 5C but for EPIC predictions of immune and non-immune cells.

Figure supplement 1. Comparison of multiple cell fraction prediction methods in tumor datasets.

Figure supplement 2. Comparison of cell fraction prediction methods with flow cytometry data of melanoma tumors.

Figure supplement 3. Comparison of cell fraction prediction methods with immunohistochemistry data in colon cancer data.

Figure supplement 4. Comparison of cell fraction prediction methods with single-cell RNA-Seq data from melanoma tumors.

Figure supplement 5. Comparison between ESTIMATE scores and EPIC predictions in our new flow cytometry dataset.

Figure 1-figure supplement 1. Low dimensionality representation of the tumor infiltrating immune cell samples. Principal component analysis of the samples used to build the reference gene expression profiles from tumor infiltrating immune cells, based on the data from Tirosh et al. (Tirosh et al., 2016), considering only the primary tumor and non-lymphoid tissue metastasis samples.

Figure 1-figure supplement 2. Cell type mRNA content. (**A**) mRNA content per cell type obtained for cell types sorted from blood. Values for B, NK, T cells and monocytes were obtained as described in *Materials and Methods*. Values for Neutrophils are from (Subrahmanyam et al., 2001). (**B**) Width of the forward scatter values for the different immune and cancer cells from flow cytometry data of melanoma metastatic lymph nodes. Data was first normalized by the mean FSC-W for each donor. Error bars represent the standard deviation from data of 4 patients.

Figure 2-figure supplement 1. Comparison of multiple cell fraction prediction methods in blood datasets (dataset 1 (Zimmermann et al., 2016), dataset 2 (Hoek et al., 2015), dataset 3 (Linsley, Speake, Whalen, & Chaussabel, 2014)). Heatmaps show (**A**) the Pearson R correlation and (**B**) the root mean squared error, between the cell fractions predicted by each method and the experimentally measured fractions. Results are based either on all cell types together (noted as "All cells") or for each individual cell type measured experimentally. *NA's* indicate cases where the cell type could not be predicted by a method. The "All cells" boxes are hatched when a method could not predict all the cell types so that the values computed there correspond to less cell types than for the other methods. For the dataset 2, as there are only 2 donors data, the results are only presented with all cells together (includes 8 data points). In (**A**) the significance of the Pearson correlation is indicated by stars: * p.value < 0.1, ** p.value < 0.05, *** p.value < 0.01, while not significant values are inside parentheses.

Figure 3-figure supplement 1. Sketch of the experiment designed to validate EPIC predictions starting from *in vivo* tumor samples.

Figure 4-figure supplement 1. Comparison of EPIC results per cell type for gene expression reference profiles from circulating or tumor infiltrating immune cells. (**A**) Pearson R correlation, (**B**) RMSE and (**C**) one-sided Wilcoxon rank-sum tests p-values between the cell fractions predicted and the experimentally measured fractions (from flow cytometry (this study), colorectal cancer IHC (Becht et al., 2016), single-cell RNA-Seq data (Tirosh et al., 2016) and melanoma immunohistochemistry data (Jönsson et al., 2010)). *NA's* indicate cases where the cell type could not be predicted by a method. The "*Cancer + other cells*" correspond to cancer cells and other stromal cells, for which no reference profile. No RMSE value can be computed for the IHC data in (**B**) as the measured values are not for all cells and do not reflect cell proportions. In (**A**) the significance of the Pearson correlation is indicated by stars: * p.value < 0.1, ** p.value < 0.05, *** p.value < 0.01, while not significant values are inside parentheses.

Figure 5-figure supplement 1. Comparison of multiple cell fraction prediction methods in tumor datasets. (**A**) Pearson R correlation, (**B**) root mean squared error and (**C**) one-sided Wilcoxon rank-sum tests p-values between the cell fractions predicted by each method and the experimentally measured fractions (from flow cytometry (this study), colorectal cancer immunohistochemistry (Becht et al., 2016), single-cell RNA-Seq data (Tirosh et al., 2016) and melanoma immunohistochemistry data (Jönsson et al., 2010)). Results are based either on cell types grouped together (noted as "All cells", including the immune and cancer + other cells, or "All immune cells", including only the immune cell types) or for each individual cell type measured experimentally. *NA's* indicate cases where the cell type could not be predicted by a method. The "*Cancer + other cells*" correspond to cancer cells and other stromal cells, for which no reference profile or gene signature is assumed. In (**A**) the significance of the Pearson correlation is indicated by stars: * p.value < 0.1, ** p.value < 0.05, *** p.value < 0.01, while not significant values are inside parentheses.

Figure 5-figure supplement 2. Comparison of cell fraction prediction methods with flow cytometry data of melanoma tumors. (**A**) Comparison directly of all cell types together. When a cell type could not be predicted by a given method, this cell type is absent from the subfigure. (**B**) Comparison per cell type for MCP-counter as the predictions are not comparable across different cell types. Correlation and RMSE values are available in Figure 5-figure supplement 1.

Figure 5-figure supplement 3. Comparison of cell fraction prediction methods with immunohistochemistry data in colon cancer data (Becht et al., 2016) for T cell and macrophage infiltration values. Observed values are in number of cells/mm^2. Correlation values are available in Figure 5-figure supplement 1.

Figure 5-figure supplement 4. Comparison of cell fraction prediction methods with single-cell RNA-Seq data from melanoma tumors (Tirosh et al., 2016). (**A**) Comparison directly of all cell types together. When a cell type could not be predicted by a given method, this cell type is absent from the subfigure. (**B**) Results for

MCP-counter, splitting the different cell types as the predictions are not comparable across different cell types. Correlation and RMSE values are available in Figure 5-figure supplement 1.

Figure 5-figure supplement 5. Comparison between ESTIMATE scores (A) and EPIC predictions (B) in our new flow cytometry dataset. The predictions are compared to the observed cell proportions. ESTIMATE returns a score of global immune infiltration and thus the sum of all observed immune cells has been taken for the comparison. For cancer cells in (**A**), one minus the fraction of cancer cells is plotted as ESTIMATE score is inversely correlated to the fraction of cancer cells in a sample. The observed cancer cells correspond to the melan-A+ cells. Correlations between observed fractions and predictions are based on Spearman correlations.

Additional files

Supplementary File 1. File contains the supplementary information with supplementary methods, Table S1 and Table S2.

Supplementary File 2. Gene expression reference profiles built from RNA-Seq data of immune cells sorted from blood as described in *Materials and Methods*: "*Reference gene expression profiles from circulating cells*". The file includes two sheets: (**A**) the reference gene expression values; (**B**) the gene variability relating to the reference profile. Columns indicate the reference cell types; rows indicate the gene names. (This file will be available upon publication).

Supplementary File 3. Gene expression reference profiles, built from TPM (transcript per million) normalized RNA-Seq data of immune cells sorted from blood as described in the *Materials and Methods*: "*Reference gene expression profiles from circulating cells*". The file includes two sheets: (**A**) the reference gene expression values; (**B**) the gene variability relating to the reference profile. Columns indicate the reference cell types; rows indicate the gene names. (This file will be available upon publication).

Supplementary File 4. Gene expression reference profiles built from tumor infiltrating immune cells obtained from TPM normalized single-cell RNA-Seq data as described in the *Materials and Methods*: "*Reference profiles from tumor infiltrating cells*". The file includes two sheets: (**A**) the reference gene expression values; (**B**) the gene variability relating to the reference profile. Columns indicate the reference cell types; rows indicate the gene names. (This file will be available upon publication).

Supplementary File 5. Proportion of cells measured in the different datasets: (**A**) this study; (**B**) dataset 1 (Zimmermann et al., 2016); (**C**) dataset 2 (Hoek et al., 2015); (**D**) dataset 3 (Linsley et al., 2014); (**E**) melanoma immunohistochemistry dataset (Jönsson et al., 2010) and (**F**) single-cell RNA-Seq dataset (Tirosh et al., 2016). The "Other cells" type correspond always to the rest of the cells that were not assigned to one of the given cell types from the tables. (This file will be available upon publication).

Figure 1

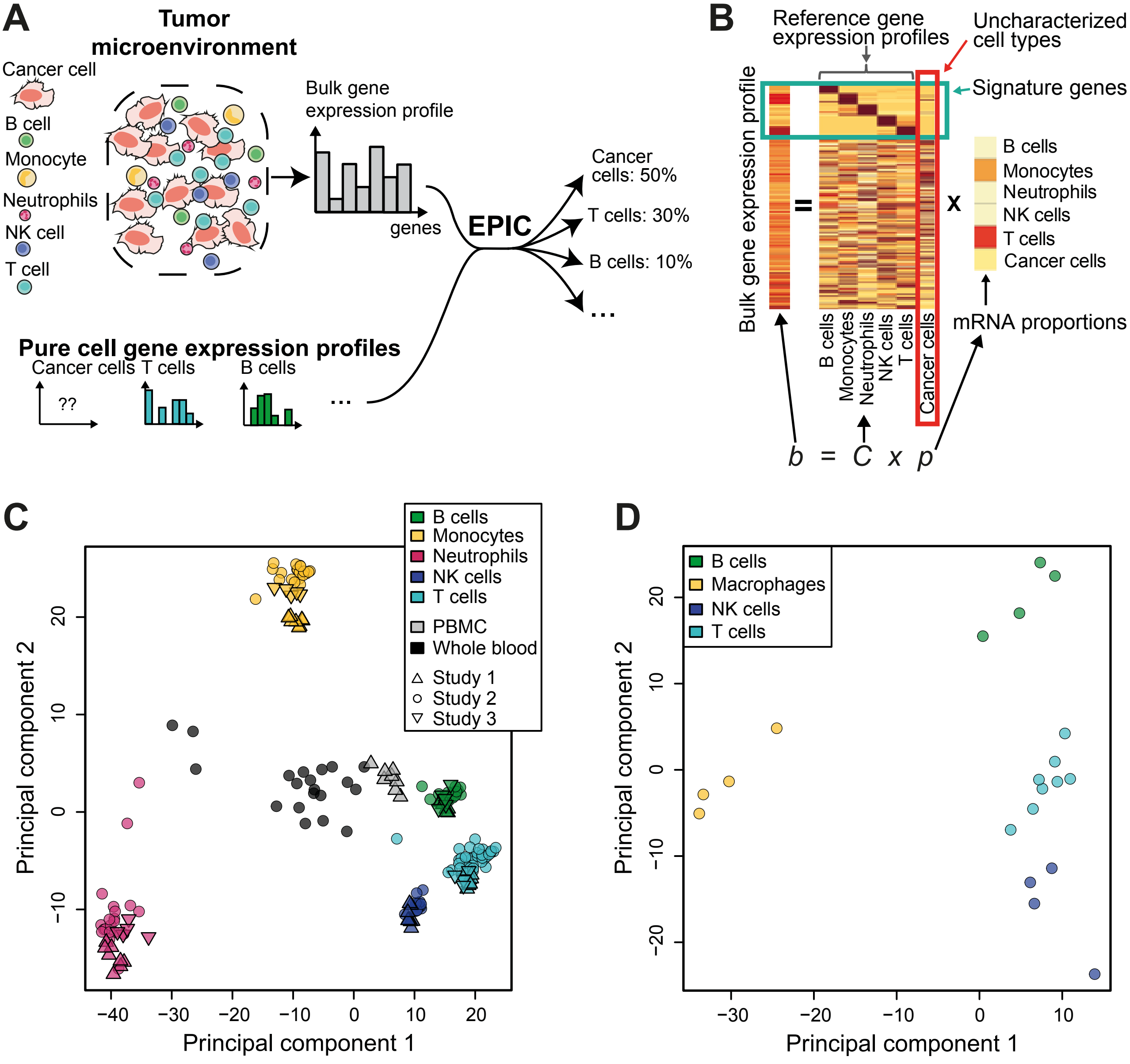

A
Tumor microenvironment
Cancer cell
B cell
Monocyte
Neutrophils
NK cell
T cell
Bulk gene expression profile
genes
EPIC
Cancer cells: 50%
T cells: 30%
B cells: 10%
...
Pure cell gene expression profiles
Cancer cells
??
T cells
B cells
...
B
Reference gene expression profiles
Uncharacterized cell types
Signature genes
Bulk gene expression profile
=
x
B cells
Monocytes
Neutrophils
NK cells
T cells
Cancer cells
mRNA proportions
b = C x p
C
B cells
Monocytes
Neutrophils
NK cells
T cells
PBMC
Whole blood
Study 1
Study 2
Study 3
Principal component 1
Principal component 2
−40
−30
−20
−10
0
10
20
−10
0
10
20
D
B cells
Macrophages
NK cells
T cells
Principal component 1
Principal component 2
−30
−20
−10
0
10
−20
−10
0
10
20

Figure 2

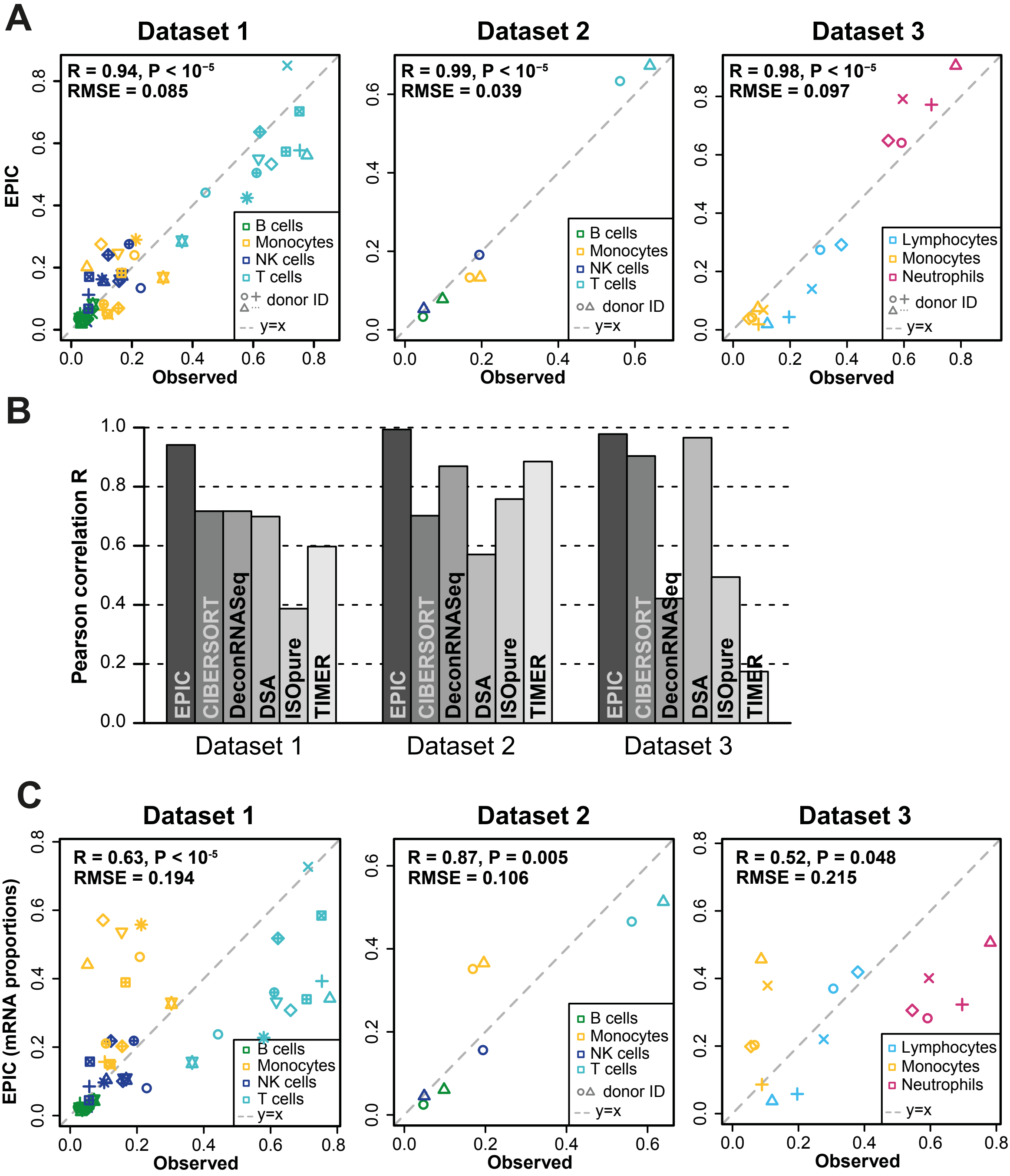

A
Dataset 1
R = 0.94, P < 10⁻⁵
RMSE = 0.085
EPIC
Observed
B cells
Monocytes
NK cells
T cells
donor ID
y=x
Dataset 2
R = 0.99, P < 10⁻⁵
RMSE = 0.039
Dataset 3
R = 0.98, P < 10⁻⁵
RMSE = 0.097
Lymphocytes
Neutrophils
0.0
0.2
0.4
0.6
0.8
B
Pearson correlation R
1.0
EPIC
CIBERSORT
DeconRNASeq
DSA
ISOpure
TIMER
Dataset 1
Dataset 2
Dataset 3
C
Dataset 1
R = 0.63, P < 10⁻⁵
RMSE = 0.194
EPIC (mRNA proportions)
Dataset 2
R = 0.87, P = 0.005
RMSE = 0.106
Dataset 3
R = 0.52, P = 0.048
RMSE = 0.215

Figure 3

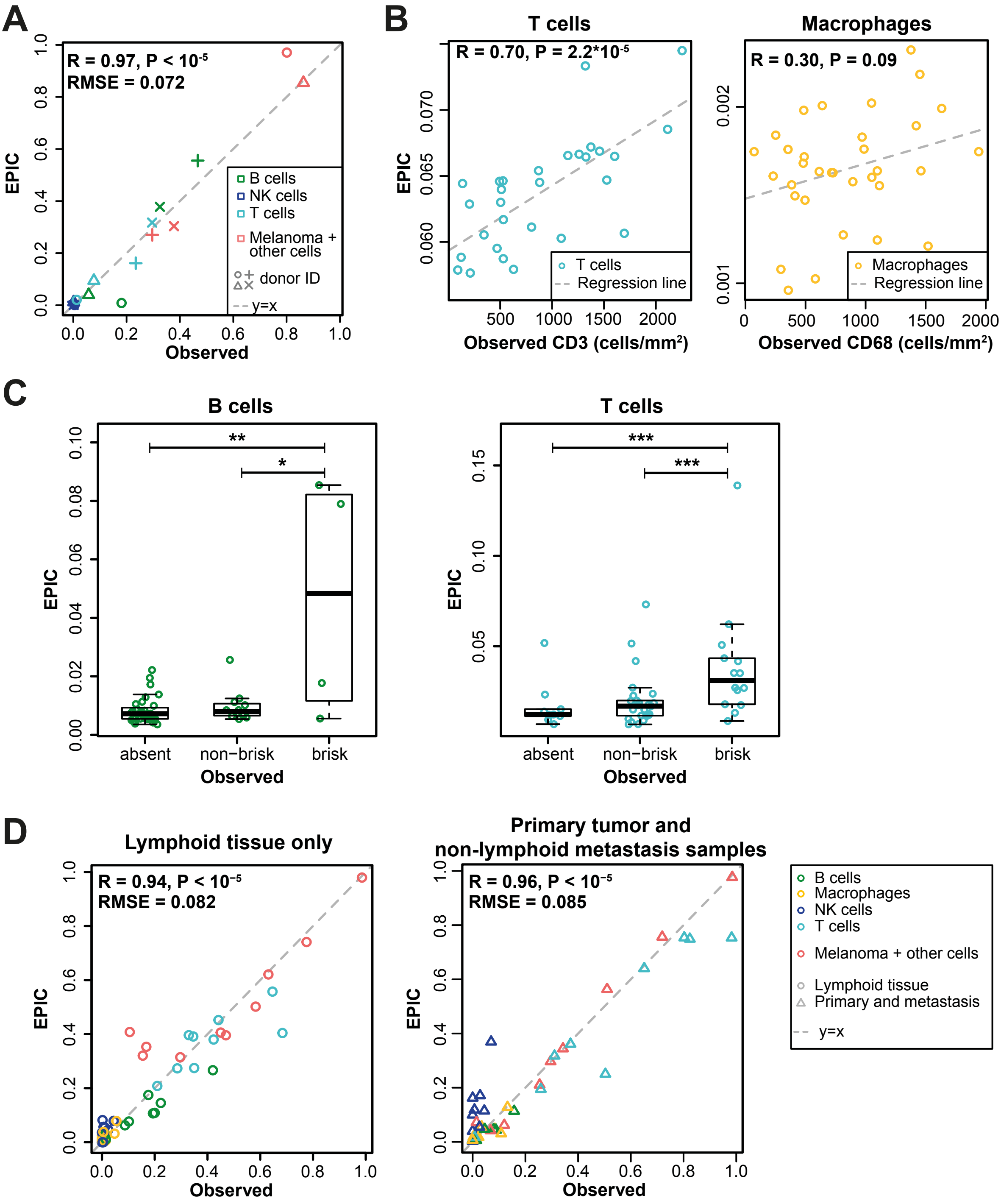

A
R = 0.97, P < 10-5
RMSE = 0.072
B cells
NK cells
T cells
Melanoma + other cells
donor ID
y=x
EPIC
Observed
0.0
0.2
0.4
0.6
0.8
1.0
B
T cells
R = 0.70, P = 2.2*10-5
T cells
Regression line
EPIC
0.060
0.065
0.070
500
1000
1500
2000
Observed CD3 (cells/mm²)
Macrophages
R = 0.30, P = 0.09
Macrophages
Regression line
0.001
0.002
0
500
1000
1500
2000
Observed CD68 (cells/mm²)
C
B cells
**
*
EPIC
0.00
0.02
0.04
0.06
0.08
0.10
absent
non−brisk
brisk
Observed
T cells
***
***
EPIC
0.05
0.10
0.15
absent
non−brisk
brisk
Observed
D
Lymphoid tissue only
R = 0.94, P < 10−5
RMSE = 0.082
EPIC
Observed
0.0
0.2
0.4
0.6
0.8
1.0
Primary tumor and non-lymphoid metastasis samples
R = 0.96, P < 10−5
RMSE = 0.085
EPIC
Observed
B cells
Macrophages
NK cells
T cells
Melanoma + other cells
Lymphoid tissue
Primary and metastasis
y=x

Figure 4

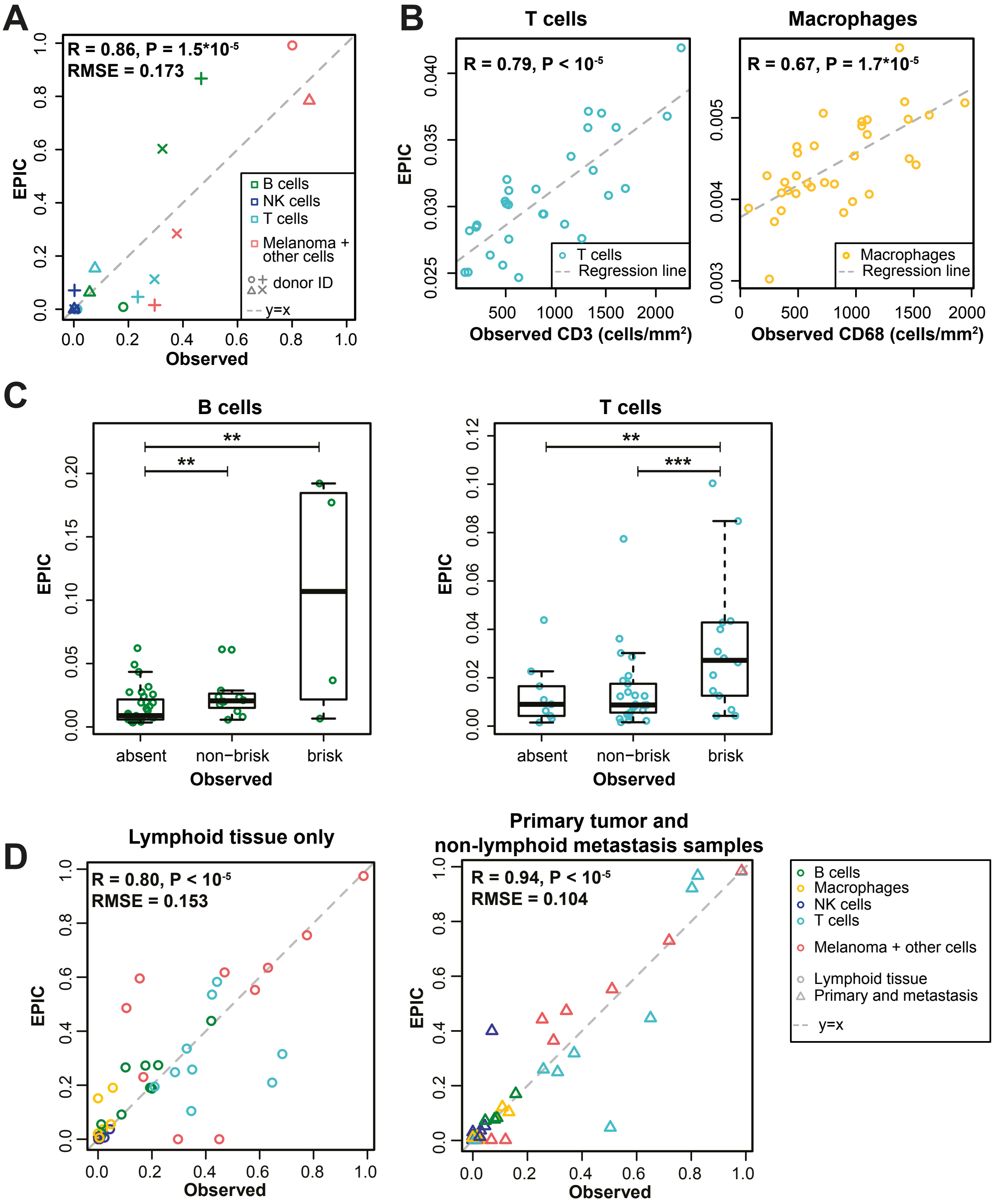

A
R = 0.86, P = 1.5*10^-5
RMSE = 0.173
B cells
NK cells
T cells
Melanoma + other cells
donor ID
y=x
EPIC
Observed
0.0
0.2
0.4
0.6
0.8
1.0
B
T cells
R = 0.79, P < 10^-5
T cells
Regression line
EPIC
0.025
0.030
0.035
0.040
500
1000
1500
2000
Observed CD3 (cells/mm²)
Macrophages
R = 0.67, P = 1.7*10^-5
Macrophages
Regression line
0.003
0.004
0.005
0
500
1000
1500
2000
Observed CD68 (cells/mm²)
C
B cells
**
**
EPIC
0.00
0.05
0.10
0.15
0.20
absent
non−brisk
brisk
Observed
T cells
**
***
EPIC
0.00
0.02
0.04
0.06
0.08
0.10
0.12
absent
non−brisk
brisk
Observed
D
Lymphoid tissue only
R = 0.80, P < 10^-5
RMSE = 0.153
EPIC
Observed
Primary tumor and non-lymphoid metastasis samples
R = 0.94, P < 10^-5
RMSE = 0.104
EPIC
Observed
B cells
Macrophages
NK cells
T cells
Melanoma + other cells
Lymphoid tissue
Primary and metastasis
y=x

Figure 5

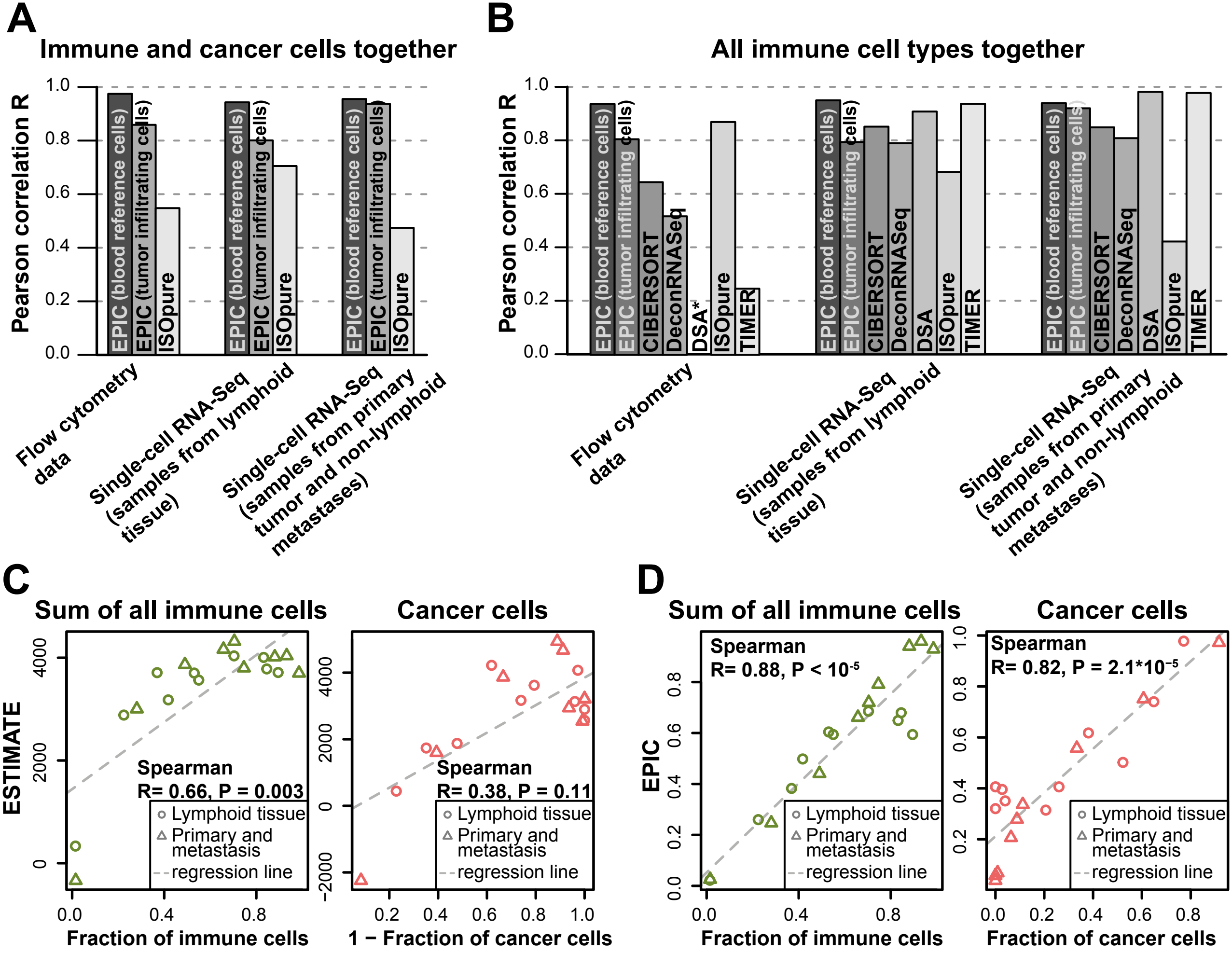

A
Immune and cancer cells together
Pearson correlation R
1.0
0.8
0.6
0.4
0.2
0.0
EPIC (blood reference cells)
EPIC (tumor infiltrating cells)
ISOpure
Flow cytometry data
Single-cell RNA-Seq (samples from lymphoid tissue)
Single-cell RNA-Seq (samples from primary tumor and non-lymphoid metastases)
B
All immune cell types together
Pearson correlation R
EPIC (blood reference cells)
EPIC (tumor infiltrating cells)
CIBERSORT
DeconRNASeq
DSA*
DSA
ISOpure
TIMER
Flow cytometry data
Single-cell RNA-Seq (samples from lymphoid tissue)
Single-cell RNA-Seq (samples from primary tumor and non-lymphoid metastases)
C
Sum of all immune cells
ESTIMATE
0
2000
4000
Spearman
R= 0.66, P = 0.003
○ Lymphoid tissue
△ Primary and metastasis
-- regression line
0.0
0.4
0.8
Fraction of immune cells
Cancer cells
−2000
0
2000
4000
Spearman
R= 0.38, P = 0.11
○ Lymphoid tissue
△ Primary and metastasis
-- regression line
0.2
0.4
0.6
0.8
1.0
1 − Fraction of cancer cells
D
Sum of all immune cells
EPIC
0.0
0.2
0.4
0.6
0.8
Spearman
R= 0.88, P < 10⁻⁵
○ Lymphoid tissue
△ Primary and metastasis
-- regression line
0.0
0.4
0.8
Fraction of immune cells
Cancer cells
0.2
0.4
0.6
0.8
1.0
Spearman
R= 0.82, P = 2.1*10⁻⁵
○ Lymphoid tissue
△ Primary and metastasis
-- regression line
0.0
0.2
0.4
0.6
0.8
Fraction of cancer cells

Figure 1 - figure supplement 1

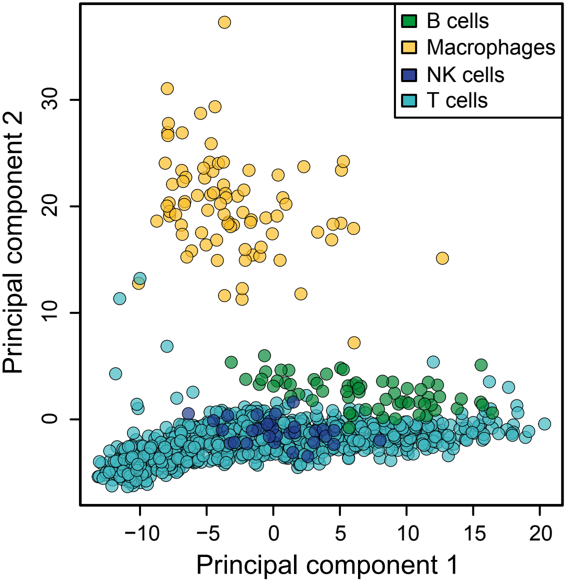


Figure 1 - figure supplement 2

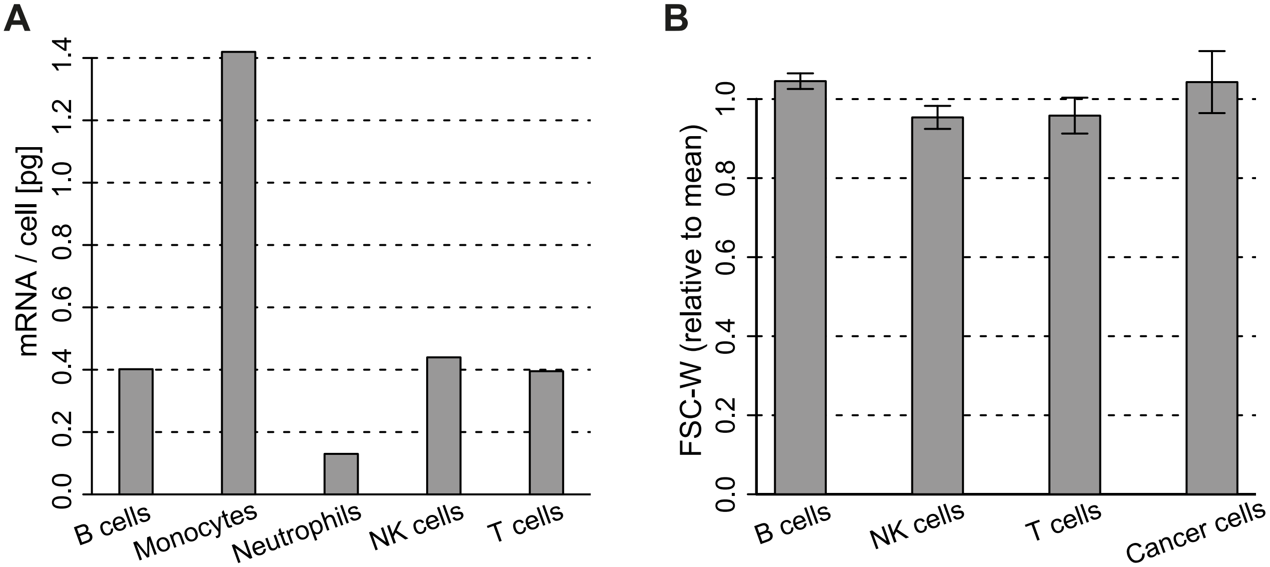
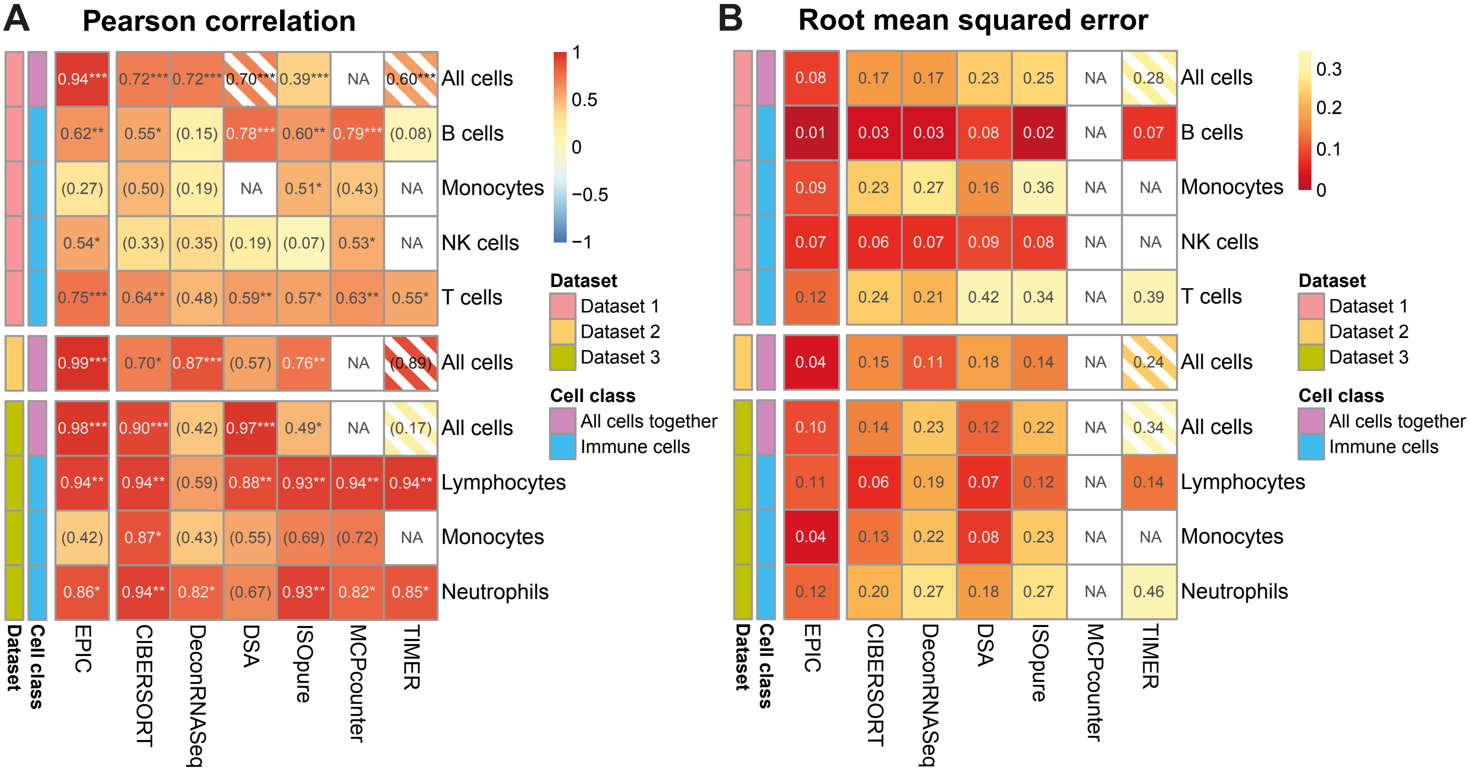


Figure 2 - figure supplement 1



A Pearson correlation

B Root mean squared error

Figure 3 - figure supplement 1

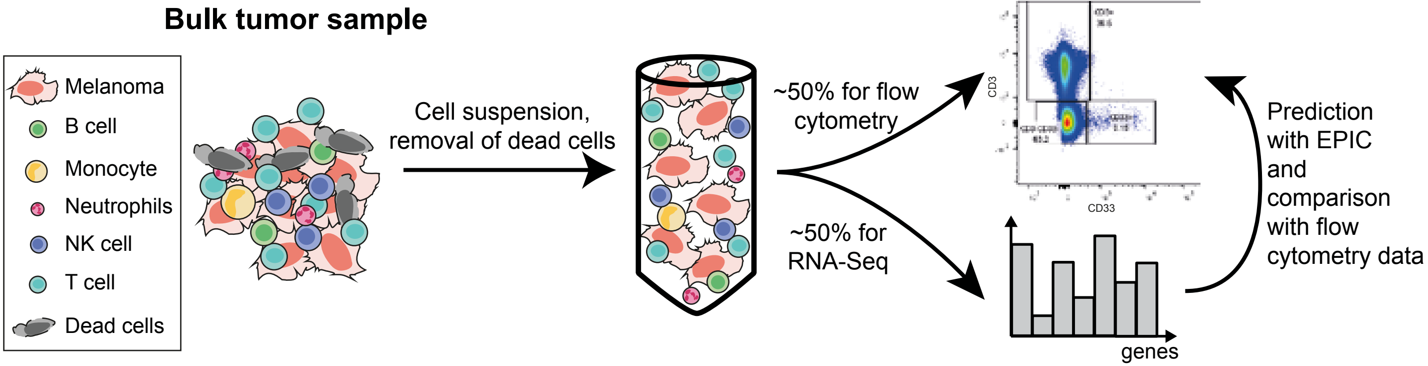


Figure 4 - figure supplement 1

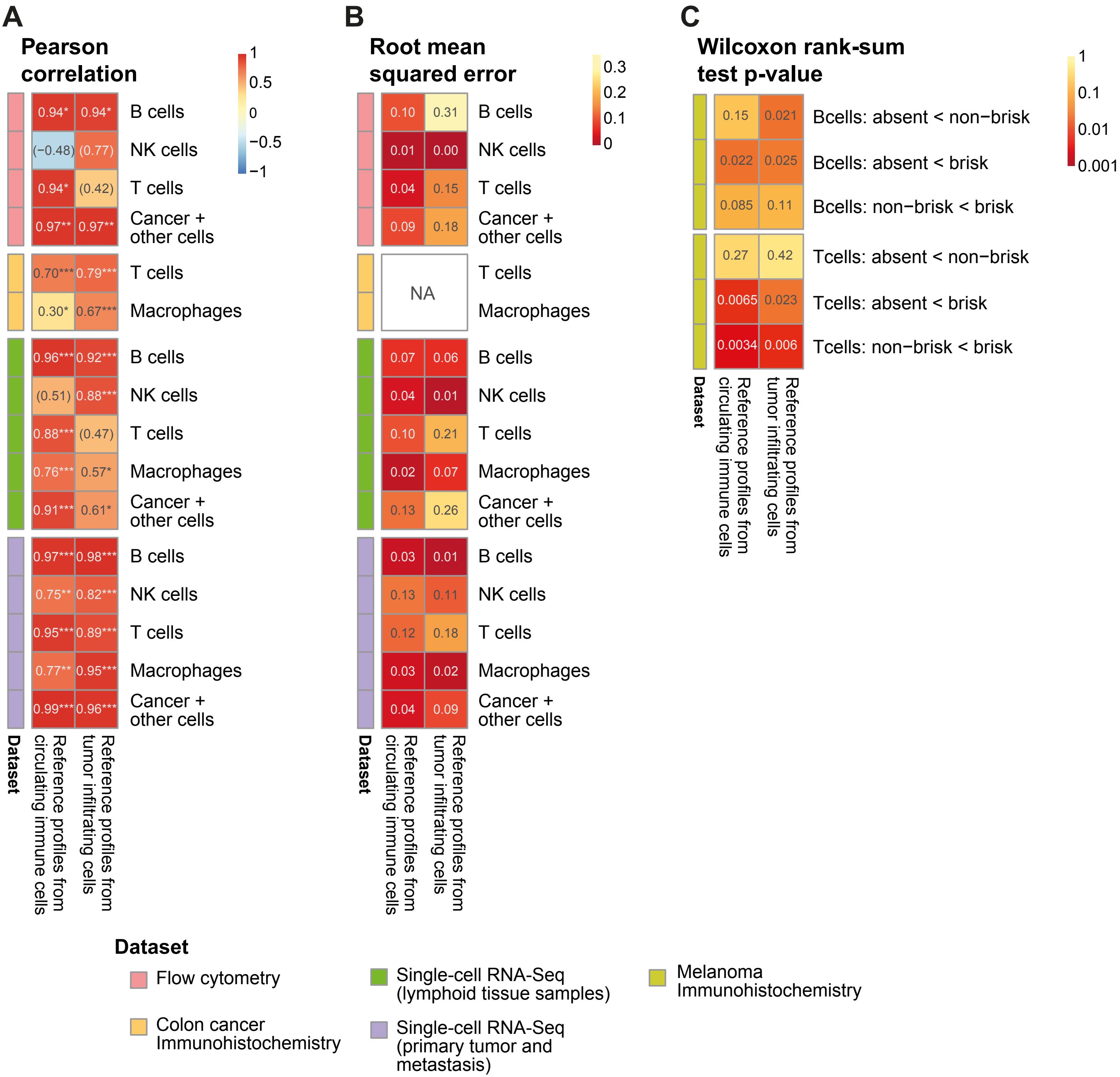


Figure 5 - figure supplement 1

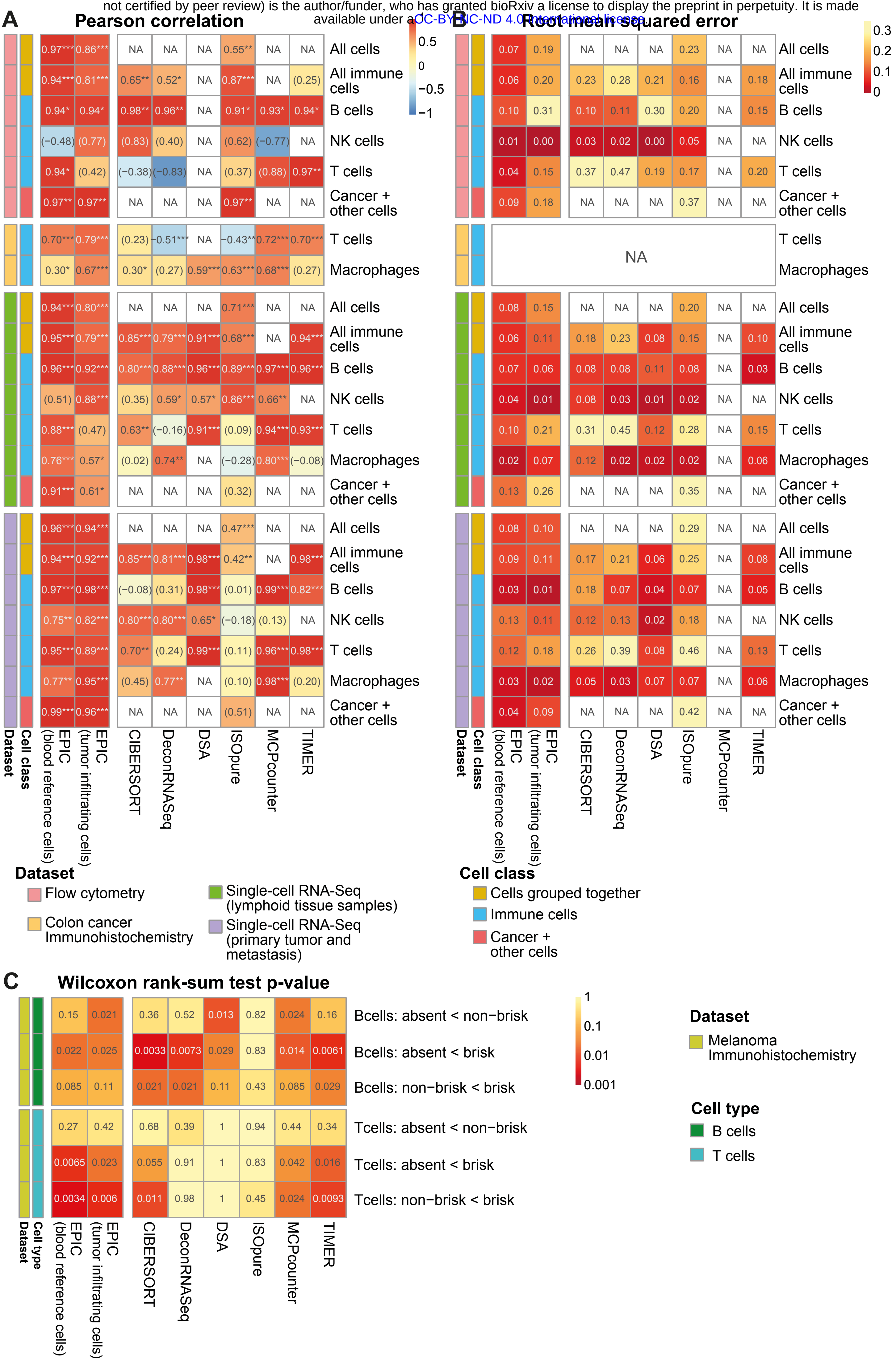


Figure 5 - figure supplement 2

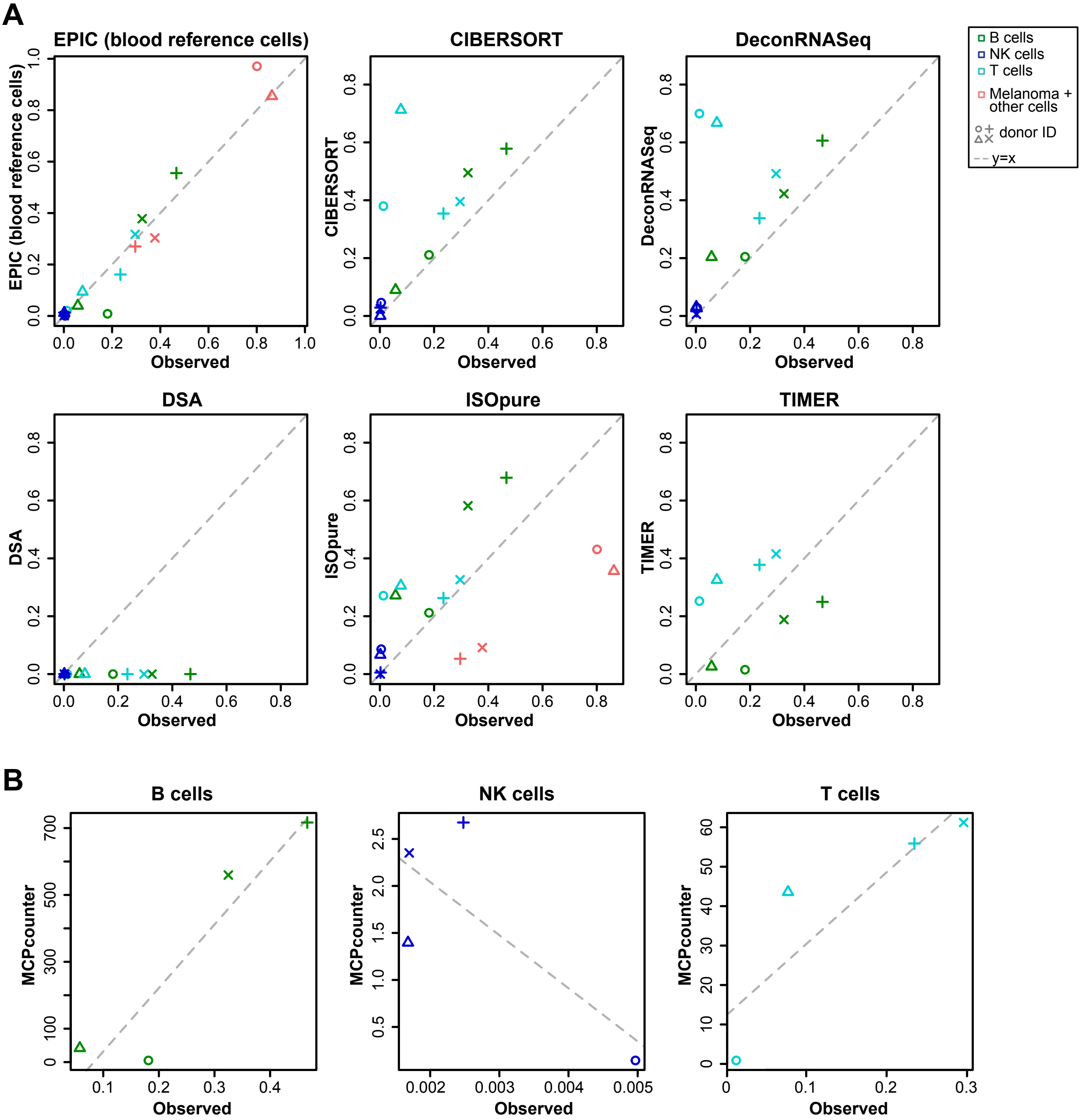


Figure 5 - figure supplement 3

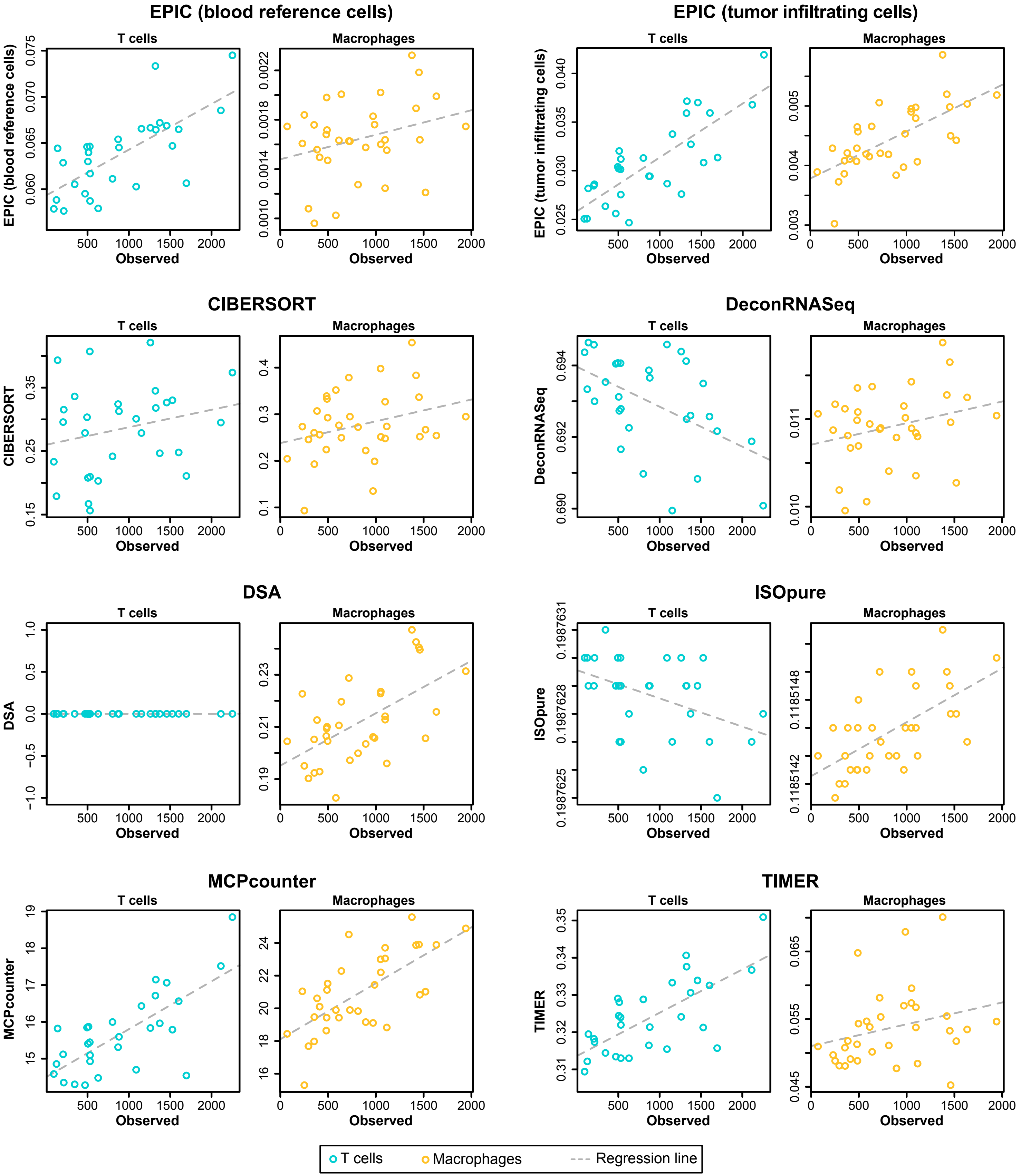


Figure 5 - figure supplement 4

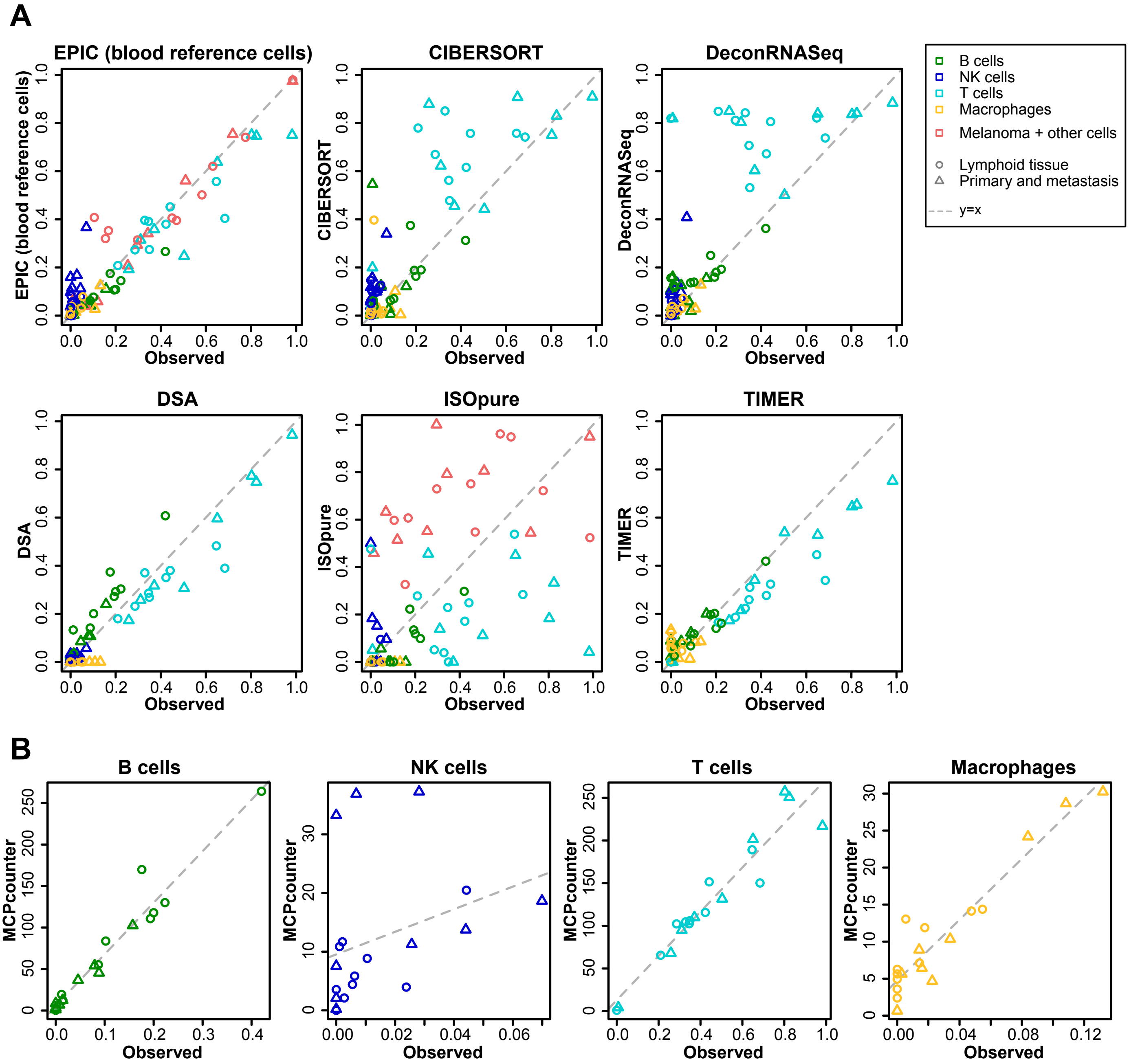
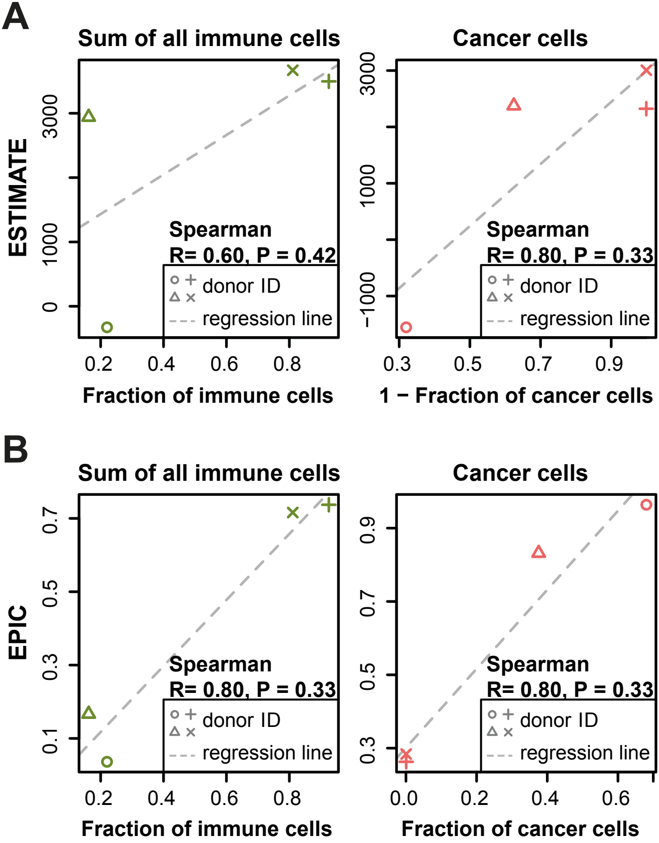


Figure 5 - figure supplement 5


A
Sum of all immune cells
ESTIMATE
Spearman
R= 0.60, P = 0.42
○ + donor ID
△ ×
regression line
Fraction of immune cells
Cancer cells
Spearman
R= 0.80, P = 0.33
○ + donor ID
△ ×
regression line
1 − Fraction of cancer cells
B
Sum of all immune cells
EPIC
Spearman
R= 0.80, P = 0.33
○ + donor ID
△ ×
regression line
Fraction of immune cells
Cancer cells
Spearman
R= 0.80, P = 0.33
○ + donor ID
△ ×
regression line
Fraction of cancer cells