GEM-DeCan: Improving tumor immune microenvironment profiling by the integration of novel gene expression and DNA methylation deconvolution signatures

Ting Xie[1,2*], Julien Pernet[1,2], Nina Verstraete[1,2], Miguel Madrid-Mencía[1,2,4], Mei-Shiue Kuo[3], Alexis Hucteau[1,2], Alexis Coullomb[1,2], Jacobo Solórzano[1,2], Olivier Delfour[3], Francisco Cruzalegui[3], Vera Pancaldi[1,2,4*]

1 Centre de Recherches en Cancérologie de Toulouse (CRCT), INSERM U1037, Toulouse 31037, France

2 Université Paul Sabatier III, Toulouse 31400, Toulouse, France

3 Translational Medicine, Institut de Recherche Pierre Fabre, Toulouse, France

4 Barcelona Supercomputing Center, Barcelona, 08034, Spain

Keywords: Deconvolution, Tumour microenvironment, DNA methylation, gene expression, EpiDISH, deconRNAseq, quanTIseq, MCP-Counter, Promoter Capture Hi-C, chromatin network

Correspondence to**:** ting.xie@inserm.fr, vera.pancaldi@inserm.fr

ABSTRACT

Quantifying the proportion of the different cell types present in tumor biopsies remains a priority in cancer research. So far, a number of deconvolution methods have emerged for estimating cell composition using reference signatures either based on gene expression or on DNA methylation from purified cells. These two deconvolution approaches could be complementary to each other, leading to even more performant signatures, in cases where both data types are available. However, the potential relationship between signatures based on gene expression and those based on DNA methylation remains underexplored.

Here we present five new deconvolution signature matrices, based on DNA methylation or RNAseq data, which can estimate the proportion of immune cells and cancer cells in a tumour sample. We test these signature matrices on available datasets for in-silico and in-vitro mixtures, peripheral blood, cancer samples from TCGA, bone marrow from multiple myeloma patients and a single-cell melanoma dataset. Cell proportions estimates based on deconvolution performed using our signature matrices, implemented within the EpiDISH framework, show comparable or better correlation with FACS measurements of immune cell-type abundance and with various estimates of cancer sample purity and composition than existing methods.

Finally, using publicly available data of 3D chromatin structure in haematopoietic cells, we expanded the list of genes to be included in the RNAseq signature matrices by considering the presence of methylated CpGs in gene promoters or in genomic regions which are in 3D contact with these promoters. Our expanded signature matrices have improved performance compared to our initial RNAseq signature matrix. Finally, we show the value of our signature matrices in predicting patient response to immune checkpoint inhibitors in three melanoma and one bladder cancer cohort, based on bulk tumour sample gene expression data.

We also provide GEM-DeCan: a snakemake pipeline, able to run an analysis from raw sequencing data to deconvolution based on various gene expression signature matrices, both for bulk RNASeq and DNA methylation data. The code for producing the signature matrices and reproducing all the figures of this paper is available on GitHub: https://github.com/VeraPancaldiLab/GEMDeCan.

INTRODUCTION

The tumor microenvironment (TME) is defined as the collection of cells and extracellular matrix that surround cancer cells inside a tumor. It affects tumor development through interactions between the different cells, which impact the probability for cancer cells to escape immune-control, grow and metastasize, and plays an important role in therapy response and resistance [1].

Much of the recent progress in cancer treatment derives from the exploitation and reactivation of immune cells such as lymphocytes that infiltrate the TME and fight cancer cells. Despite the great potential of immuno-oncology, there is a considerable difference in efficacy of these therapies across tumor types and patients. It is thus of paramount importance to develop tools to identify the different types of immune cells present in biopsy samples, as this could aid personalised therapy approaches.

More specifically, recent findings about the importance of myeloid cells in hampering the response to immunotherapies make the development of macrophage signature matrices an important goal for immuno-oncology [2]. Traditional immunotherapies rely on the concept that CD8+ T cells, that are normally responsible for killing cancer cells, are already infiltrated in the tumor region, despite being inactivated by their interaction with the cancer cells that disable their cytotoxic activity [3]. The main mechanisms that block the killing of cancer cells were found to rely on the presence of PD-1 and CTLA-4 receptors on the T cell surface, which are inhibited by binding with their specific ligands expressed on cancer cells and antigen-presenting cells respectively [4]. Most current approaches aim to interrupt this interaction using antibodies directed at PD-1 or PDL-1 and CTLA-4, reactivating the killing action of CD8+ T cells against cancer cells. Unfortunately, a high percentage of patients [5] do not seem to respond durably to these treatments and recent findings point to the presence of myeloid cells in the TME of these patients that can potentially prevent infiltration of CD8+ T cells inside the tumor and even protect cancer cells in some cases [6].

Tumor-associated macrophages (TAMs) can be found in the microenvironment of solid tumors in high numbers. Depending on their phenotypes, these myeloid cells can promote tumor progression, by suppressing antitumor immunity, or directly protecting cancer cells. TAMs are already becoming important treatment targets in cancer, especially in tumors which are not presenting high lymphocyte infiltration [7,8].

Some of these TAMs are differentiated from circulating monocytes upon entering the tumor area while others might be pre-existent residing in the tissue. Macrophages acquire different phenotypes that can be indicatively distributed along a spectrum of polarization going from M1 polarization, in which they promote inflammation, to M2 polarization state, which involves a tumor protective behaviour generated by their role in tissue reconstruction, promoting cancer cell survival [9]. However, these cells are extremely plastic and our knowledge on their behaviour has been mostly generated through in-vitro experiments that probably remain quite distant from the situations they encounter inside tumors [9,10].

Identifying the presence of different macrophage types in bulk samples has proven particularly challenging due to two main reasons: first, macrophages are not found in the circulating blood, and obtaining them in-vitro requires their stimulation with cytokines that can either produce an M1 or an M2 polarization state in an artificial way, which leaves no guarantee that their phenotype will be comparable to that of TAMs found inside tumors; second, the two macrophage states are very plastic and unstable, while also being quite similar, such that distinguishing between the two states in a bulk dataset is particularly difficult [9,10].

Fluorescence Activated Cell Sorting (FACS) is an experimental technique generally used to determine cellular composition after cell separation from tissues or cultures. However, it is difficult to use FACS to identify cells which are poorly characterised due to a lack of well-defined surface markers, and it can become heavy when the samples have multiple cell types and each cell type has to be sorted through multiple markers. Exploiting the wide availability of transcriptomics and methylomics data for bulk samples, multiple algorithms have recently become available to estimate cell type proportions [11] or to estimate tumor purity in complex cellular mixtures [12]. Deconvolution methods can be classified as "reference-free" [13,14] or "reference-based" [15] depending on whether a specific signature matrix is used to identify the cell types, or clustering is used to simply infer the different cell types present [16].

DNA methylation (DNAm) profiles are cell-type specific and an excellent alternative to transcriptomes to perform cell-type deconvolution [17]. This is due mostly to the fact that the methylome can be thought of as a record of the cell's past history and is less affected by transient perturbations of the cell's environment [18]. Traditionally, DNAm-based deconvolution was developed for studying different cell types present in blood, specifically in studies regarding the effects of the ageing process. It was quickly realised that the differences observed in blood across ages could be potentially related to changes in cell-type composition [19].

As for "reference-based" deconvolution methods based on DNAm, the procedure usually involves constructing a signature matrix which is specific and important to the problem of interest. So far, available methods have only used references based on Illumina human 450k or MethylationEPIC (also named 850k array since it features 850,000 probes along the genome), and are usually estimating absolute fractions referred to the immune cells from blood, such as T cells (CD4+, CD8+), neutrophils (Neu), B cells, natural killer (NK) cells, and monocytes (Mono), but often do not consider immune cells that are important for cancer immunology like regulatory T cells (T_{reg}) and macrophages (M).

In this paper, we exploited a large collection of haematopoietic epigenomes of reference produced by the BLUEPRINT project [20], and published data sets (**Supplementary file 1**) to establish a series of novel signature matrices for a gene expression and methylation based deconvolution approach. In addition to novel signatures for expression or methylation based deconvolution of immune cell types in blood (BPmet, BPRNA), we generated specific signature matrices to quantify the proportion of cancer cells as well as of specific immune cell types infiltrated in tumours (BPmetCan, BPRNACan).

We further hypothesized that genes that are not passing the filter of inclusion in the gene expression-based signature matrix, but are associated with CpGs that are in the methylation signature matrix, should also be included in the gene-expression deconvolution signature matrix. We exploited the available chromatin structure in haematopoietic cells [21] to identify genes whose expression might be impacted by CpGs in the DNAm signature matrix, locally or through 3D contacts and were able to show that our expression-based signature matrix is improved when including these genes.

Finally, we provide a pipeline named GEM-DeCan to allow for gene expression and DNA methylation data processing, in order to run different deconvolution methods with multiple provided signature matrices. Along with the pipeline, a script to guide users into the generation of their own signature matrix is supplied.

METHODS

Collection of WGBS data

To build the Blueprint signature matrix, we collected 52 samples generated from whole-genome bisulfite sequencing (WGBS) of 10 purified blood-derived immune cells on GRCh38 (Homo sapiens genome) (**Supplementary file 1: Table S1**). Bigwig files including methylation signal and coverage of methylation signal were downloaded from Blueprint epigenome level 3 data (http://www.blueprint-epigenome.eu). In addition, we also downloaded 7 cancer datasets, 4 WB and 11 normal tissues (**Supplementary file 1: Table S2**) via Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/) and Brinkman et al. [22]**.**

WGBS data processing

The files were parsed into R data structures, we then discarded bases that have coverage below 10X [23] and also have more than 99.9th percentile of coverage in each sample. Methylation signal ($WGBS_{\beta}$) was calculated with the following formula:

$$WGBS_{\beta} = methylatedCounts / (methylatedCounts + unmethylatedCounts)$$

The hg19 coordinates of WGBS from GEO were converted to GRCh38 with liftOver (https://www.bioconductor.org/help/workflows/liftOver/) R package. To map the common methylated cytosines from GEO and Blueprint datasets by genomic position to the Illumina 850K Methylation Array CpG sites we used the Infinium MethylationEPIC v1.0 B5 Manifest (https://support.illumina.com/downloads). After the coordinate transformation, the missing values on the beta matrix were imputed with impute R package using default parameters to generate the final beta matrix without missing values.

A beta-value matrix was generated from the WGBS dataset, including 409,103 CpGs (48% overlap with EPIC's 850k CpGs) measured across a total 107 samples (Normal=17, WB=4, Cancer=34, Immune=52).

Deconvolution BluePrint (BP) signature matrices generation procedure

The signature matrices were established through the following steps:

1. The limma [24] R package was used to determine differentially methylated CpGs (DM-CpGs) or differentially expressed genes (DEGs) for all pairwise comparisons between cell types: for GE signature matrices, the voom function was used to remove heteroscedasticity for RNA-seq data using log2(TPM + 1).

2. We preselected significant candidate DM-CpGs/DEGs for the signature matrices using Benjamini-Hochberg corrected p-values . For GE signature matrices, the false discovery rate (FDR) cutoff 0.05 was selected, while for the DM-CpGs, 2 different FDR values were selected (**Supplementary file 2: Method S1**) to generate their signature matrices:
    a. FDR < 0.05 was used to generate BPmet in blood.
    b. FDR < 1e-05 was used to generate BPmetCan.

3. The significant candidate genes/CpGs that were obtained from the previous step (point 2) were selected according to specific thresholds based on their absolute logFC for pairwise comparisons.(**Supplementary file 2: Method S1**). We filtered the DEGs with 2 different abs(logFC) cutoffs to obtain our GE signature matrices:
    a. abs(logFC) > 2.5 was used to generate BPRNA signature matrix.
    b. abs(logFC) > 2 was used to generate BPRNACan, while abs(logFC) > 3 was used to obtain cancer cells when generating BPRNACan.

    We then sorted the genes obtained from point 3 and ranked them by decreasing fold change (to obtain positive variable genes). The number of genes to be included in each signature was chosen depending on their logFC density distributions (**Supplementary file 2: Method S1**).

Generation of the DNA methylation signature matrix (BPmet)

We started from WGBS public datasets from the Blueprint (BP) project to generate a methylation based signature matrix (BPmet), which we used for immune cell decomposition in blood from healthy donors. Datasets for purified immune cells of 6 types were considered: CD4 (N=8), CD8 (N=8), B cells (N=9), Monocytes (N=4), NK (N=2) and Neutrophils (N=7) [20]. The signature matrix includes only CpGs with false discovery rate (FDR) < 0.05, absolute log fold change (abs(lfc)) > 0.2 and the top 100 CpGs (**Supplementary file 3: Table S1**).

Generation of the enhanced DNA methylation signature matrix for cancer samples (BPmetCan)

To identify cancer, healthy and immune cells in cancer samples, we used firstly a set of Normal (tissues), Whole Blood (WB) and cancer (solid tumours) WGBS samples (Normal: N=17, WB: N=4, Cancer: N=34) via GEO and Brinkman et al. [22] to generate a signature matrix that can recognize the three groups of cells. To identify the CpGs we needed to include in this signature matrix, we chose CpGs that were highly methylated in cancer cells

compared to Normal as well as WB samples with cutoffs at FDR < 0.05, lfc > 0.5, selecting a maximum of 100 CpGs for each pairwise comparison between the 3 cell group.

In the tumor microenvironment, three states of macrophages can be found (M0, M1, M2) [25] as well as one specific type of T cell (Treg) [26]. For this reason, we extended the BPmet signature matrix with those 4 new immune cell types to generate a new BPmet signature matrix following the signature matrix process section with FDR threshold of 1e-05, lfc difference of 0.3 and with a maximum of 300 CpGs for each pairwise comparison.

To build the final BPmetCan signature matrix which can identify cancer cells and also specific immune types, we merged the above described signature matrix for cancer/normal/blood cells with the BPmet signature matrix extended with macrophages and Tregs. The final significant CpGs are the union of CpGs presented in those two signature matrices, respectively, whereas the profiles of CpGs in each cell type were calculated through the median of methylated values over all samples belonging to that cell type. The BPmetCan signature matrix includes 1896 CpGs (**Supplementary file 3: Table S2**).

RNAseq data processing

The TCGA expression data normalized as Fragments Per Kilobase of transcript per Million (FPKM) were taken from public datasets [27]. The transcripts per millions (TPM) expression datasets for WB were downloaded from GTEx portal (https://gtexportal.org/home/). 9 purified blood-derived immune cells TPM expression datasets on GRCh38 (**Supplementary file 1: Table S3**) were collected from the Blueprint project portal [20]. The expression value FPKM was first converted to TPM using the following formula:

$$TPM_i = (FPKM_i / \sum FPKM_j) * 10^6 \quad (1)$$

The final expression value for each $gene_i$ in the $sample_j$ was normalized from TPM with the following formula:

$$TPM_{ij} = TPM_{ij} * 10^6 / \sum_t TPM_{tj} \quad (2)$$

Generation of the BPRNA signature matrix

We generated the Blueprint RNAseq based signature matrix (BPRNA) using the same approach and the same 6 immune cell types as for the BPmet signature matrix. We selected the signature CpGs as the top 200 genes (ranked by decreasing log fold change, LFC), among those which have false discovery rate (FDR) < 0.05 and lfc > 2.5(**Supplementary file 3: Table S3**).

Generation of the CCLE_TIL10 signature matrix

To develop this signature matrix we started from the TIL10 (170 genes) one proposed in the quanTIseq method[28], which includes 10 immune cell types (B cells, M1 and M2 macrophages, monocytes (Mono), neutrophils (Neu), natural killer (NK) cells, non-regulatory CD4+ T cells, CD8+ T cells, Treg cells, and myeloid dendritic cells (DC)).

FASTQ files of all samples used for generating the TIL10 signature matrix (170 genes [28]) were downloaded, preprocessed and gene expression was quantified as described in [28]. An expression matrix, for 10 immune cell types, was constructed, consisting of 19,423 genes and 51 samples. We then constructed a cancer signature matrix ("CCLE") by considering differential expression between cancer cell line samples from CCLE (eliminating blood cancer cell lines) [29] and healthy tissues and blood samples from GTEx [30]. Briefly, we used a bootstrap approach by which 50 samples were randomly taken from each of three datasets to construct the complete dataset (150 samples in total) for the analysis of differential expression by limma. The mean-variance relationship was modeled with the voom function and the Benjamini-Hochberg method was used for multiple hypothesis testing. lfc > 2.5 and FDR < 0.005 were used to select differentially expressed genes. Only genes which are highly expressed in cancer cells compared to normal tissues and compared to blood samples were selected as "UP" genes. This procedure was repeated 30 times and only the UP genes which are present in each iteration are selected as cancer cell specific genes (Fig. 2b). The expression profile of cancer cells was computed as the median of the expression values over all samples for all UP genes in the CCLE dataset, resulting in the CCLE signature matrix (138 genes).

To build the combined CCLE_TIL10 signature matrix, we simply considered the union of genes from the TIL10 and CCLE signature matrices (Fig. 2b). The expression profiles in the matrix were computed as the median of the expression values over all samples belonging to each cell type (**Supplementary file 3: Table S4**).

Generation of the BPRNACan signature matrix

We generated first an extended BPRNA signature matrix, based on Blueprint expression data [20], selecting samples for CD4 (N=12), CD8 (N=3), B cells (N=5), Monocytes (N=7), M0 (N=4), M1(N=4), M2(N=5), NK (N=2) and Neutrophils (N=10) and filtering out Treg cells due to low number of samples (N=1). We then selected genes with an FDR < 0.05, and lfc > 2, including the 150 genes with highest lfc to identify differentially expressed genes following the procedure to create the signature matrix. Second, we built a cancer signature matrix using normal and cancer samples collected from TCGA for different cancer types with normal adjacent tissue and WB from GTEx, using FDR < 0.05, lfc > 3, including the top 100 genes with highest lfc. Samples from TCGA were from 18 cancer types: Bladder, Cholangiocarcinoma, Thyroid carcinoma, Liver hepatocellular carcinoma, Colon adenocarcinoma, Kidney Chromophobe, Kidney Renal Clear Cell Carcinoma, Kidney renal papillary cell carcinoma, Lung Squamous Cell Carcinoma, Lung Adenocarcinoma, Stomach adenocarcinoma, Cervical squamous cell carcinoma and endocervical adenocarcinoma, Uterine Corpus Endometrial Carcinoma, Head-Neck Squamous Cell Carcinoma, Breast invasive carcinoma, Rectum adenocarcinoma, Esophageal carcinoma. The final BPRNACan signature matrix was generated by merging the two signature matrices described above, and includes 1403 genes (**Supplementary file 3: Table S5**).

Generation of gene expression signature matrices expanded according to the methylation signature matrix (BPRNACanProMet) and 3D chromatin contact maps (BPRNACan3DProMet)

Our BPRNACan signature matrix contains 1403 genes which we call *sig genes*, whereas the 1896 CpGs from the BPMetCan signature matrix are denoted as *sig CpGs.* To take into account the potential involvement of genes that are important for each cell type, as evidenced by methylation, but not sufficiently differentially expressed to be *sig genes,* we created a set of expanded gene expression signature matrices.

We considered 3D chromatin contact networks for all immune cells included in Javierre et al. [21] , detected by the Promoter Capture Hi-C technique and filtered using CHiCAGO [31]. First, we combined BPRNACan *sig genes* with the genes that have a BPMetCan *sig CpG* in their promoter (according to promoter definitions based on promoter capture libraries in [21]), leading to the "BPRNACanProMet" signature matrix (**Supplementary file 3: Table S6**). Additionally, we generated the "BPRNACan3DProMet" signature matrix by appending to BPRNACan signature the genes that respect the two characteristics mentioned above: having a *sig CpG* in their promoter and having a 3D contact with a fragment (promoter or not) containing a *sig CpG*, (**Supplementary file 3: Table S7**). Finally we constructed a further expanded signature matrix adding genes whose promoters only have a 3D contact with a fragment containing a *sig CpG,* leading to the "BPRNACan3DMet" signature matrix.

Validation datasets

Whole blood methylation datasets

For validating and assessing the performance of the BPmet signature matrix, we used two independent public datasets (**Supplementary file 2: Table S1**): 100 WB samples from the Grady Trauma Project (GSE132203) profiled using IlluminaHumanMethylationEPIC and another 6 WB samples using the 450k methylation array from Koestler et al. (GSE77797) [32]. Flow-cytometry estimates of the proportion of blood cell types were available for the two WB datasets. The estimated fraction of cells obtained by deconvolution using the EpiDISH (RPC: robust partial correlation) method [33] was compared to the flow-cytometry estimated proportions using Pearson Correlation.

Methylation datasets from Peripheral Blood Mononuclear Cells (PBMC)

To test our BPRNA deconvolution signature matrix on whole blood we used 13 PBMC samples with corresponding flow cytometry data [34] (**Supplementary file 2: Table S1**).

In-silico mixtures

To test the BPRNACan signature matrix, we used a simulated RNA-seq data from quanTIseq [11], consisting of 1700 samples created by in-silico mixing of reads from immune-cells and cancer cell lines in different proportions, simulating different tumor purity (0 to 100%)(**Supplementary file 2: Table S1**).

Melanoma cancer samples

To assess the performance of our GE signature matrices, we used 4 metastatic melanoma lymph nodes from Racle et al. (GSE93722) [35] and another 19 primary tumor samples from non-metastatic patients from Tirosh et al. (GSE72056) [36] (**Supplementary file 2: Table S1**).

Cancer samples with methylation and gene expression profiles

To test the BPmetCan and BPRNACan signature matrices, 495 methylation profiles from LUAD samples in TCGA were downloaded in level 3 (beta-value, 450k Illumina array) as well as their corresponding gene expression profiles measured by RNAseq (RNASeq2GeneNorm) using the RTCGAToolbox [37] R package. Finally, we also considered 59 RNA-seq datasets from multiple myeloma bone marrow samples after removal of cancer cells [38] (**Supplementary file 2: Table S1**).

Estimating accuracy in cancer cell proportion deconvolution

To estimate the proportion of cancer cells inside tumor samples, different methods have been proposed. For TCGA samples, estimates based on the ABSOLUTE, ESTIMATE, LUMP and IHC methods were available in [12] as described in **Supplementary file 2: Table S2**.

Comparison between proportions estimated by deconvolution and other methods

We used Pearson's correlation coefficients to compare our estimates of deconvolved proportions to either FACS data or alternatively estimated proportions (**Supplementary file 2: Table S2**).

Regression models to predict response to immunotherapy

Three public melanoma datasets and one bladder cancer dataset with response to anti-PD1 [39–42] were considered. ElasticNet [43] penalized logistic regression models were run using the results from different deconvolution methods and signature matrices as features.

For each combination of signature matrix and deconvolution method, 5 models were trained, including 4 models trained by leave-one-dataset-out (lodo) and one model trained by 5-fold cross-validation (standard CV). The training includes a hyperparameter search for the l1 ratio and penalty strength. For the lodo training this search is performed by 5-fold CV on training datasets, and models are evaluated on the remaining test dataset. For the standard CV, a fourth of samples is kept as a hold-out test set, and the hyperparameter search is performed by 5-fold CV on the remaining samples, the model is then evaluated on the hold-out test set.

RESULTS

RNAseq processing and deconvolution pipeline

In order to test the different signature matrices we present in this paper, we provide an RNAseq analysis pipeline, built with snakemake [44] and conda [45]. It allows the user to choose from various tools and options (Fig. 1). The pipeline can start from raw Illumina sequencing data (.bcl format) or from already processed and normalized TPM data and runs a selection of tools performing deconvolution with the methods used in this paper (quanTIseq, MCPCounter, deconRNASeq, and EpiDISH) and all signature matrices mentioned in the paper. It is freely available on GitHub : https://github.com/VeraPancaldiLab/GEMDeCan

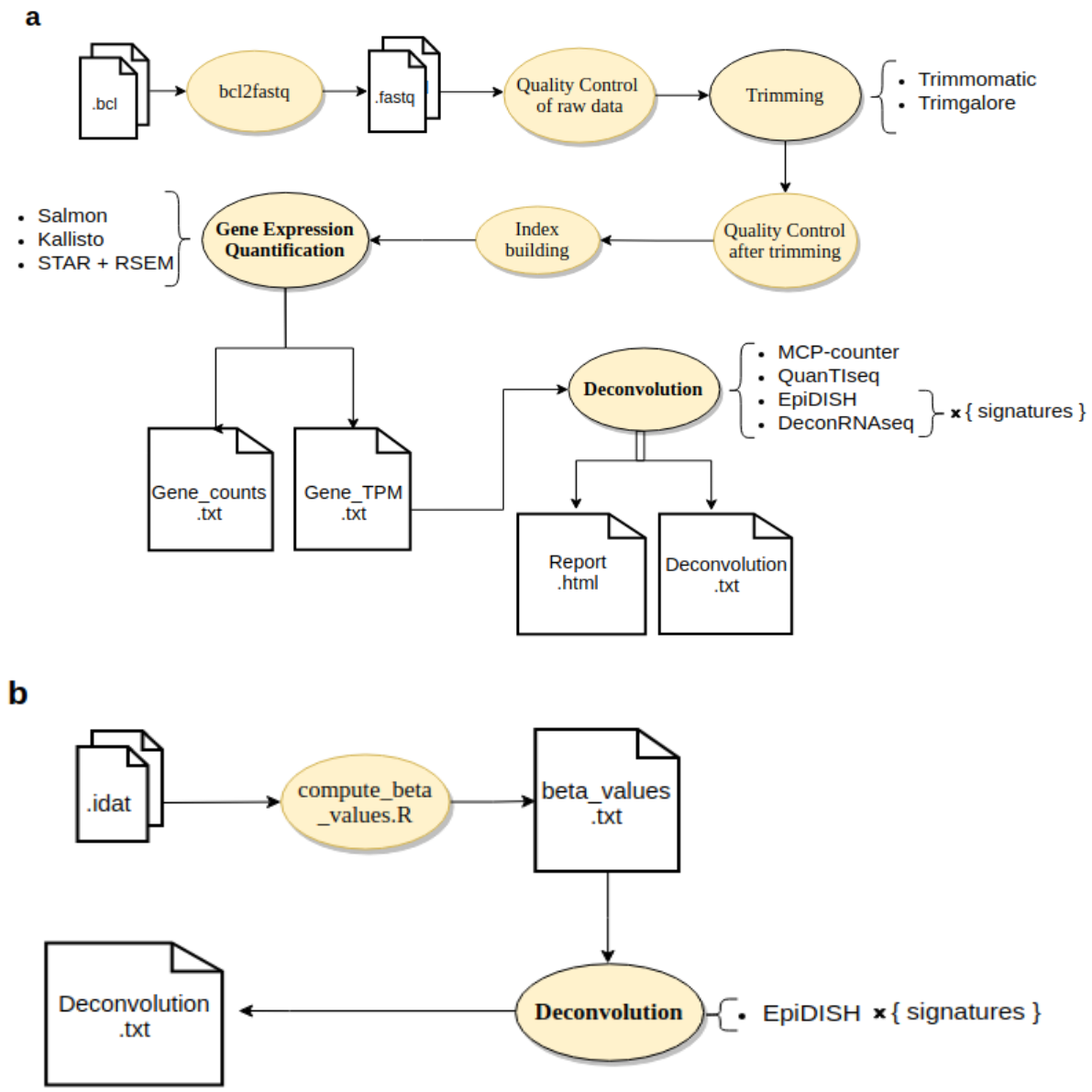
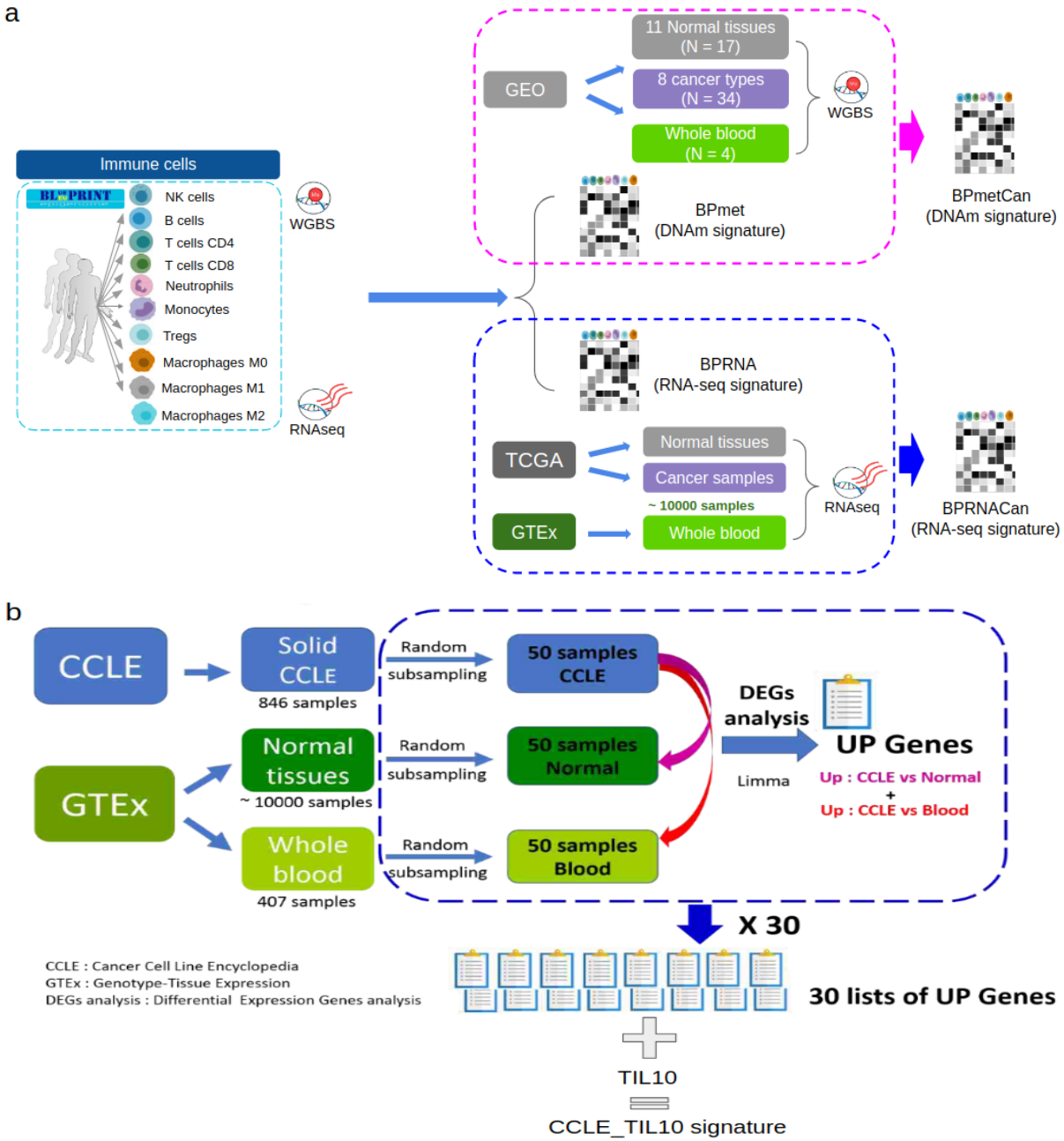


Fig. 1 : Deconvolution pipeline workflow. a) RNAseq processing and deconvolution pipeline. **b)** DNA methylation deconvolution workflow

BPmet: A Novel DNA methylation-based signature matrix for immune cell deconvolution

We exploited the WGBS methylation datasets that were produced for bulk samples of purified cells as part of the Blueprint project [46] to generate a signature matrix of 502 CpGs which allows us to identify 6 major immune cell types in blood (see methods, Fig. 2a and **Supplementary File 1: Table S1**). We named this signature matrix BPmet (**Supplementary file 3: Table S1**). To test this newly generated signature matrix, we performed deconvolution with the EpiDISH R package [33], using our BPmet signature matrix, and choosing the RPC method as well as other available methods (**Supplementary file 2: Table S2**). We then proceeded to test our BPmet new signature matrix on various datasets (**Supplementary file 2: Table S1**).



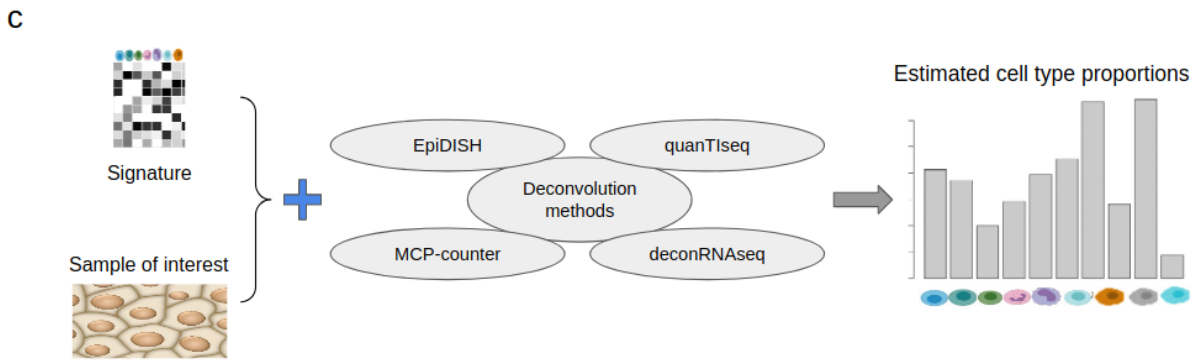


Fig. 2: Schematic description of the proposed deconvolution approach. (**a**) Workflow to generate BPmetCan and BPRNACan signature matrices, through combining BPmet and BPRNA immune signature matrices, and cancer signature genes and CpGs. (**b**) To deconvolve the exact proportion of cancer cells as well as immune cells, the TIL10 signature matrix [28] was combined with a list of genes that differs between cancer cell lines, normal tissues and whole blood, generating the CCLE_TIL10 signature matrix. (**c**) Using these signature matrices with deconvolution methods (based on DNAm or GE) we can estimate cancer and immune cell type proportion from bulk samples.

Testing the BPmet signature matrix on peripheral blood samples

To evaluate the performance of the BPmet signature matrix using the EpiDISH method, we first applied it to an independent publicly available Illumina EPIC array (850k CpGs) dataset choosing 100 whole blood samples from healthy donors with the true cell composition verified by FACS (gold standard) (GSE132203, Grady Trauma Project). We observed extremely high Pearson correlations between the estimated cell compositions and FACS fractions for all included samples (Pearson's R = 0.993, $p < 2.2e-16$) and each cell subtype (**Fig. 3a and Supplementary file 2: Figure S1a**).

We then compared the results from using the BPmet signature matrix in these 100 whole blood samples from the previously mentioned gold standard dataset (GSE132203) to the results using other signature matrices. The signature matrices we compared were the MethylCIBERSORT signature matrix [47], and the default signature matrix for the EpiDISH method [33], based on DNAse hypersensitivity sites (DHS), which are highly cell-type specific regions of open chromatin [48]. Altogether, correlations between the different signature matrix estimates and the FACs fractions were similar for EpiDISH-DHS and BPmet, while MethylCIBERSORT had a worse performance. Only NK cells estimate was worse correlated with BPmet compared to EpiDISH-DHS and MethylCIBERSORT, probably due to the few samples included for this cell type in creating our signature matrix (NK: n = 2) (**Supplementary file 2: Figure S1a-c**)

Since the MethylCIBERSORT and EpiDISH-DHS signature matrices were generated starting from the Illumina 450k platform, we tested them on another set of 6 whole blood samples analysed by the same technology and compared results to the flow cytometry measurements provided for that dataset [32]. We were able to obtain better correlations (Pearson's R = 0.93, P < 2.2e-16) for all cell types confounded compared to either methylCIBERSORT or EpiDISH-DHS. We obtained correlations above 0.94 for each cell

subtype except for NK cells (**Fig. 3b and Supplementary file 2: Figure S2**). We thus conclude that our BPmet signature matrix can correctly capture cell type composition as well or better than other available methods in whole blood.

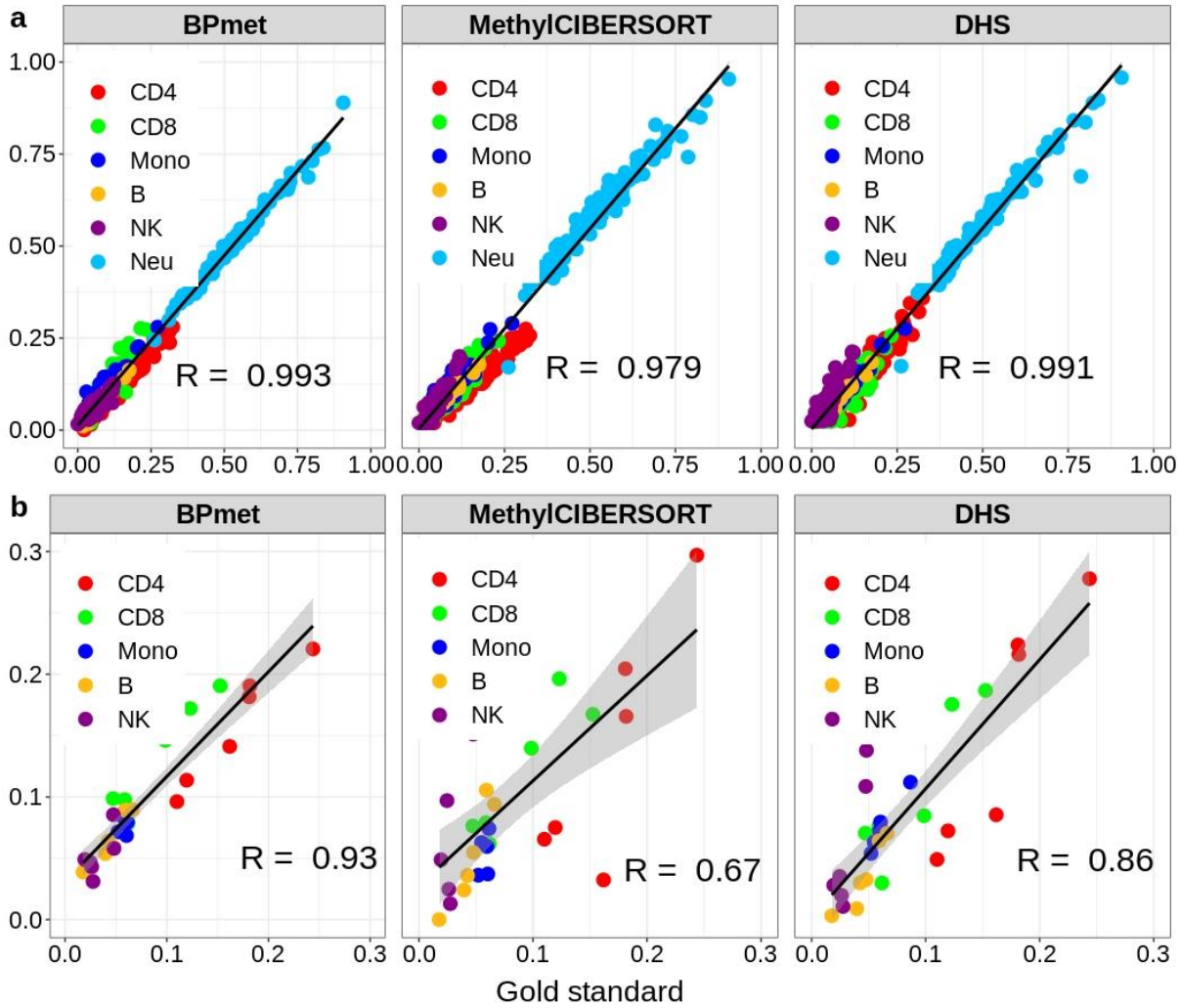


Fig. 3: Correlation of estimated cell fractions using the 3 signatures for EpiDISH vs cell fractions estimated by FACS (gold standard) (**a**) Comparing our DNA methylation signature matrix BPmet and available signature matrices (MethylCIBERSORT, EpiDISH-DHS) in 100 whole blood data with FACS data (GSE132203) and DNA Methylation data (Illumina EPIC 850k) from the Grady Trauma Project using the EpiDISH RPC method. (**b**) Comparing the DNA methylation signature matrices in 6 WB data analysed by Illumina 450k platform and by FACS (GSE77797) [32].

Extending the BPmet signature matrix to estimate cancer and immune cell proportion in tumor samples

Confident of the satisfying performance of the BPmet signature matrix in identifying cell type proportions in whole blood samples, we turned towards applying deconvolution to cancer tissues, extending the BPmet signature matrix to also estimate the proportion of cancer cells.

To expand our BPmet signature matrix to include immune cell types in the TME and also cancer and normal cells of different types, we combined data from the previously used 52 immune samples with 17 normal and 34 cancer samples (see Methods for details). This signature matrix, called BPmetCan (see Methods, **Fig. 2a and Supplementary file 3: Table S2**), allowed us to estimate the proportion of cancer cells as well as of 10 immune cell types in bulk tumor samples.

To test the performance of BPmetCan in a realistic scenario, we used it to estimate the proportion of cancer cells in 495 Lung Adenocarcinoma (LUAD) DNA methylation datasets from TCGA. Firstly, we retrieved publicly available estimates of the proportion of cancer cells in these samples (purity) produced using the ABSOLUTE method [49], which involves analysing somatic DNA alterations to estimate ploidy and hence tumor purity. When comparing purity estimates using BPMetCan with ABSOLUTE results, we obtained a Pearson correlation of $R = 0.79$ and $P < 2.2e-16$, proving that our signature matrix is able to identify the proportion of cancer cells in a tumor sample (Fig. 4a). We then compared the results obtained with BPmetCan to other previously published purity estimation methods from ref [12,47]. These include a method based on gene expression of immune and stromal genes (ESTIMATE [50]), a method based on CpGs that are unmethylated in immune cells (LUMP [12]), a methylation based deconvolution approach (MethylCIBERSORT [47]) and a method that estimates purity based on images of haematoxylin and eosin stain slides (IHC [12]).

We found that BPmetCan displayed a higher correlation with ABSOLUTE than either ESTIMATE or IHC, and a very similar one to MethylCIBERSORT and LUMP in TCGA-LUAD (Fig. 4a). In order to further verify whether our signature matrix is accurate in calculating the proportion of tumor purity in other cancer types, we also applied it to breast cancer samples in TCGA (TCGA-BRAC). Once again, we observed a high correlation of estimated tumor purity between BPmetCan and ABSOLUTE, and our method had the second best concordance with ABSOLUTE after MethylCIBERSORT (**Supplementary file 2: Figure S3**).

In order to assess our ability to determine the proportions of other immune cell types in the sample, the estimated fraction of immune cells using BPmetCan or MethylCIBERSORT signature matrix was compared to quantification of lymphocytes, macrophages and neutrophils estimated from H&E images also in the same samples from TCGA-LUAD [51]. We observed that BPmetCan and MethylCIBERSORT have very similar performances, as measured by the correlation of deconvolved immune cell types compositions with H&E estimates (Fig. 4b). Investigating macrophages (Monocytes/Macrophages lineage) showed that the prediction of macrophage abundance derived from MethylCIBERSORT was more accurate than that derived using the BPmetCan signature matrix ($R = 0.48$ against $R = 0.33$) (Fig. 4b). Nevertheless, it appears that the MethylCIBERSORT reference profiles do not capture all macrophages (some samples are on the x-axis indicating no estimated abundance while non-zero values are clearly detected by H&E images). This could be due to the classification of macrophages by the MethylCIBERSORT signature matrix that only includes monocytes (CD14+ cells), whereas our signature matrix can potentially classify monocytes, M0, M1 and M2 separately (all these types were merged for making this comparison).

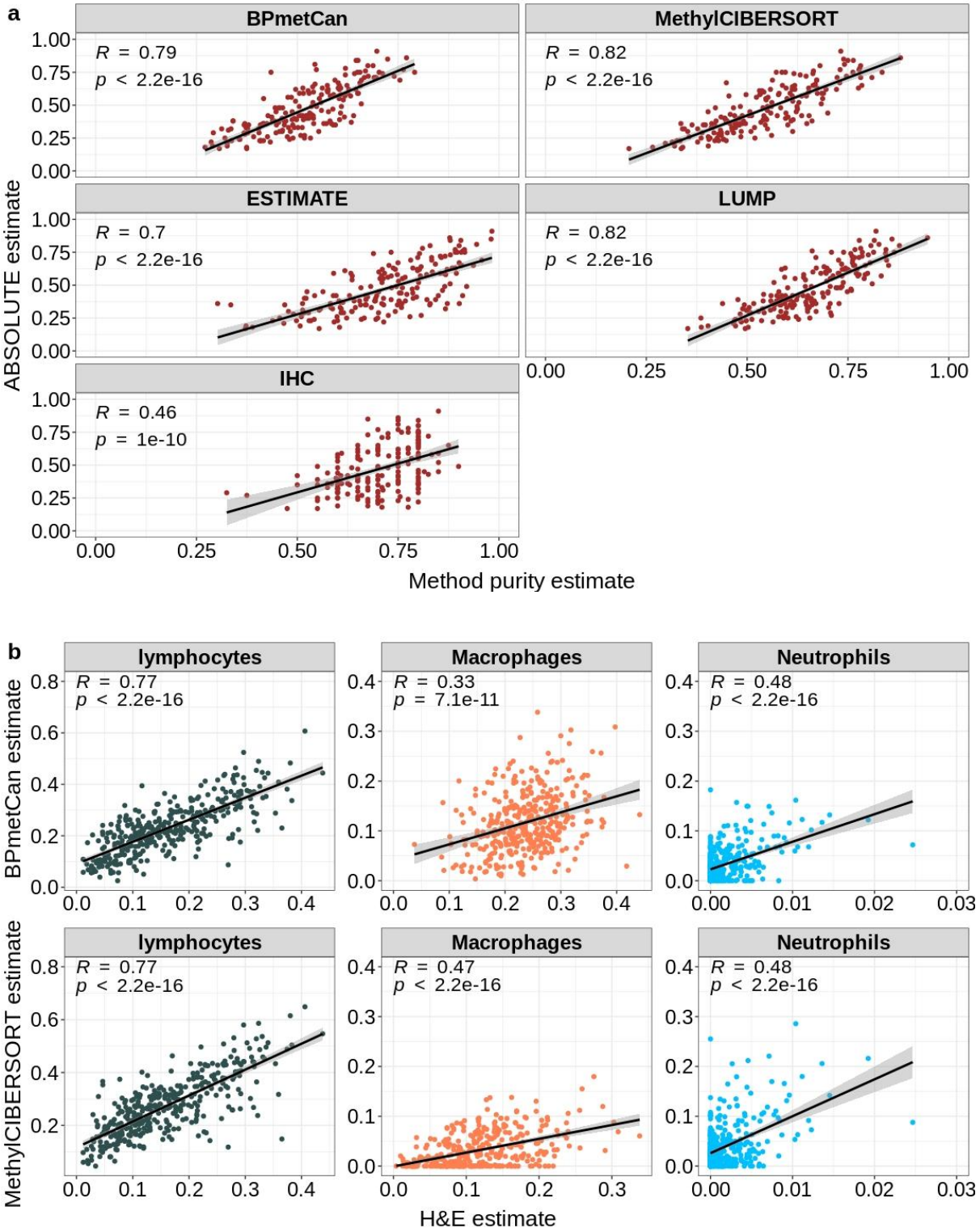


Fig. 4: Deconvolution using different DNA methylation signature matrices in tumor samples. (**a**) Correlations between TCGA-LUAD tumor purity proportion estimated by BPmetCan and other previously published purity estimation methods compared to estimates by ABSOLUTE using Pearson correlations [12]. (**b**) Scatterplots showing correlation between estimates of proportions of different immune cell types and H&E image estimates from [51].

BPRNA: a signature matrix for immune cell deconvolution in blood

We started by developing a novel immune cell type deconvolution signature matrix based on RNAseq expression data from primary samples for 6 immune cell types (see BPmet part, including CD4 T cells, CD8 T cells, Monocytes, B cells, NK cells and Neutrophils) in blood [20], which we called BPRNA (see methods, **Fig. 2a and Supplementary file 3: Table S3**). We then proceeded to test the BPRNA signature matrix using two deconvolution methods for which signature matrices can be specified by the user (EpiDISH [33] and deconRNAseq [52]) on peripheral blood mononuclear cell (PBMC) samples (**Supplementary file 2: Table S1**).

Testing the BPRNA signature matrix on PBMC samples

We tested our BPRNA signature matrix using PBMC mixtures where both RNA-seq and flow cytometry data was available [34]. We observe lower accuracy of estimated total cell fractions for this data set, especially for NK and CD8 cells (Fig 5a). This is probably due to the limited number of samples available for NK (N=2) and to a possible discrepancy between the CD8 activation state in these PBMC samples and the ones used in the signature matrix (which exclude naive CD8 cells).

Multiple gene expression reference-based methods have demonstrated a high accuracy in estimating cell proportions in blood and immune infiltrates from bulk RNA-seq data, amongst which we chose two for comparison to estimates produced by BPRNA: we first compared to two methods that come with their own signature: MCP-counter [53], which is a scoring method based on marker genes, and quanTIseq [28], which is based on constrained least squares regression and can estimate immune cell fractions and fractions of unknown cells with high accuracy. Finally, we tested the deconRNAseq [52] method using different signature matrices to compare its performance with EpiDISH. The results of running deconvolution with all different methods and signature matrices compared to FACS estimates in PBMC are summarised in Figure 5b.

Considering all cell types together or sub-cell types, we observed the combination of BPRNA with EpiDISH compared very favorably relative to the combination of BPRNA with deconRNAseq which overall showed the weakest performance (Fig 5b). However, we note that MCP-counter or quanTIseq outperformed the combination of BPRNA with EpiDISH in several sub-cell types (CD8 T cells, Monocytes, B cells, NK cells, and Neutrophils).

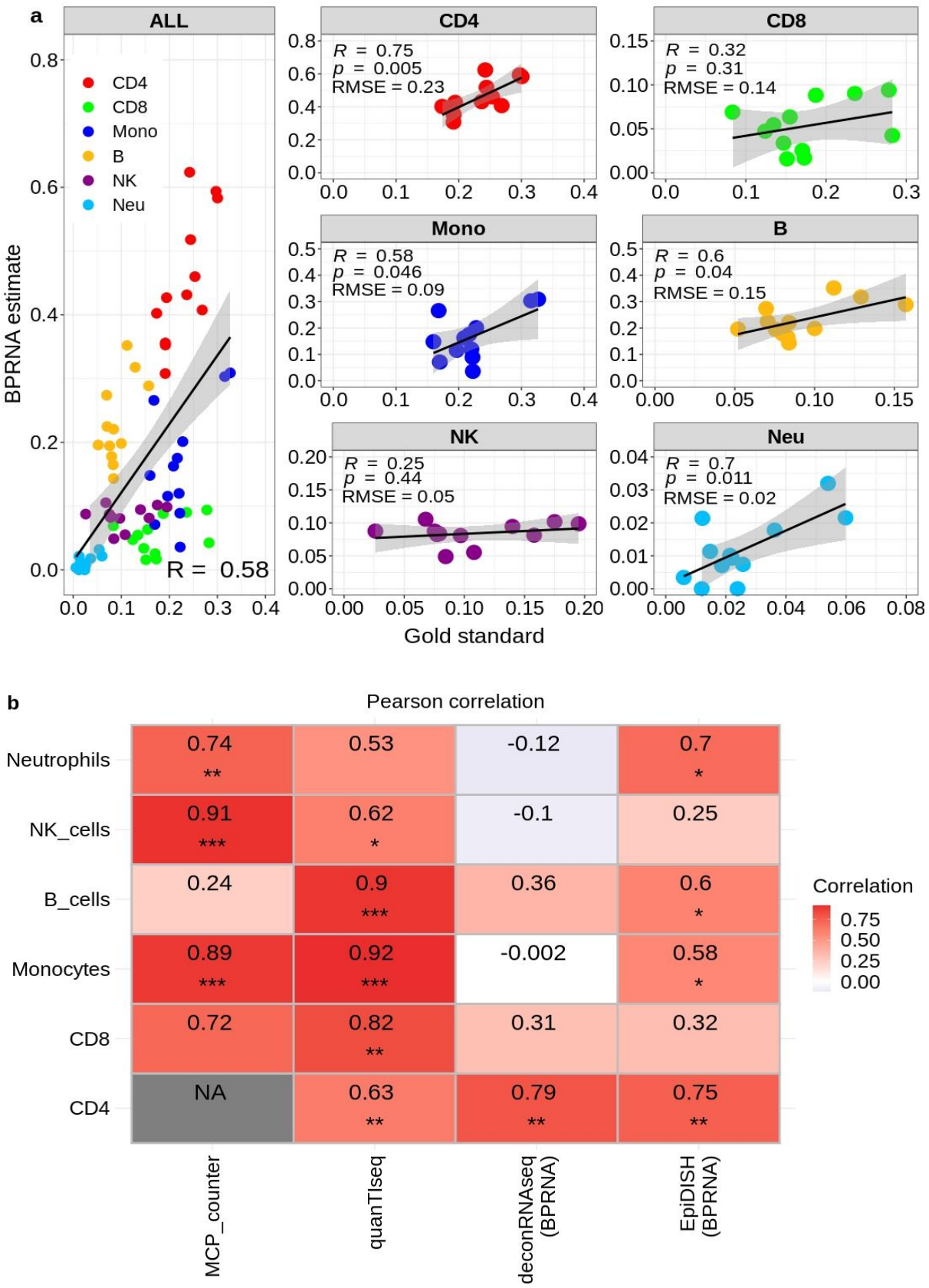


Fig. 5: Deconvolution using **(a)** BPRNA signature matrices and **(b)** Pearson Correlation between the cell fractions estimated by each method Predicted vs. FACS proportions in PBMC (GSE107011) [34]. The significance of the Pearson Correlation is indicated by stars: *p<0.05, **p<0.01, ***p<0.001

BPRNACan and CCLE_TIL10: new immune-cell and cancer gene expression signature matrices

To exploit the wide availability of expression data for cancer samples, we generated two gene expression-based signature matrices to identify immune and cancer cells. Similar to what was done for the methylation based signature matrix, we extended the BPRNA signature matrix to detect cancer cells as well as specific immune types.

We considered two ways of designing a signature matrix that would estimate tumor purity based on bulk RNAseq expression data. In the first case, we constructed a signature matrix named CCLE_TIL10 (see methods and **Supplementary file 3: Table S4**) based on RNAseq data for over a thousand cancer cell lines [29] and a large number of healthy tissue and blood RNAseq samples from the GTEx [30].

Aware of the difference between cancer cell lines and cancer cells, we also developed a signature matrix for detecting cancer and specific immune cells starting from expression data in cancer, adjacent normal tissues and immune cells. To this end, similarly to what was done in the case of DNA methylation, we looked for samples with RNAseq data for tumor and non-tumor tissues and whole blood, which are readily available through the TCGA and GTEx, and integrated it with the BPRNA immune cell signature matrix, which was based on 9 immune cell types [20]. This new signature matrix, which we called BPRNACan consists of 1403 genes (see Methods, Fig. 2a and **Supplementary file3: Table S5**). We used the CCLE_TIL10 and BPRNACan signature matrices with the EpiDISH method to estimate cell composition in different samples with gene expression datasets (**Supplementary file 2: Table S1**).

Validation of the CCLE_TIL10 and BPRNACan signature matrices on in-silico and in-vivo tumor samples

As a first step to validate our method on samples containing cancer cells, we considered an in-silico simulated cancer RNAseq dataset, composed of reads from purified samples from 10 immune cells (B-cells, NK-cells, CD4+ T-cells, CD8+ T-cells, Monocytes, NK, Neutrophils, M1, M2 and Dendritic cells), which were added to reads from a sample of MCF10 cancer cell lines in different proportions [28]. The results of using the CCLE_TIL10 signature matrix on this in-silico mixture are in excellent agreement with true mRNA proportions (Pearson R > 0.9), for 11 cell types (**Supplementary file 2: Figure S4a**), as expected since the TIL10 signature matrix was derived from this data and CCLE_TIL10 signature matrix is an extension of it.

The proportions estimated using our newly developed BPRNACan signature matrix (with the EpiDISH method) were also in very good agreement with the true mRNA proportions (Pearson R = 0.96 for cancer cells and Pearson R > 0.74 for other immune cells fractions, **Supplementary file 2: Figure S4b**), except for macrophages M2 (R = 0.35). This lower performance in the detection of M2 macrophages could be due to BPRNACan missing M2s in some samples, suggesting that the signature matrix potentially does not capture all M2 phenotypes that are present. Our signature matrix might also have similar issues with M1

macrophages, as we can clearly see two groups of samples, one of which has estimated M1 proportions quite discordant from the true cell fractions (**Supplementary file 2: Figure S4b**).

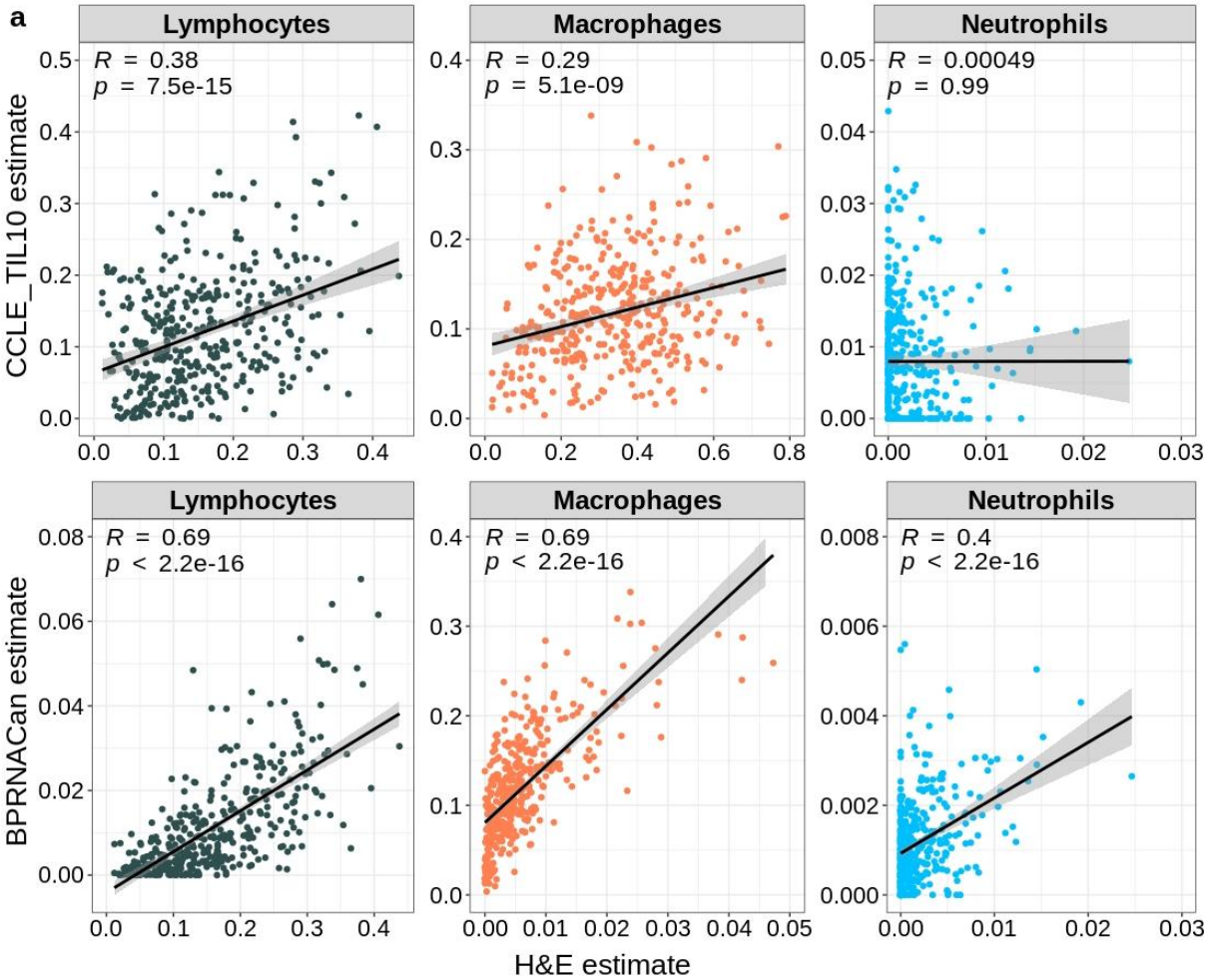
We then investigated whether these deconvolution signature matrices would be able to estimate tumor purity from real biological samples, namely tissue samples from TCGA. We therefore analysed the results obtained with the 2 signature matrices on TCGA samples (**Supplementary File 2, Figure S5**)**.** We first compared the estimation of cancer cell purity in the samples to ABSOLUTE, ESTIMATE, LUMP and IHC results, as was done for the methylation analysis. We found that the tumor purity estimates derived using BPRNACan were better than that derived from CCLE_TIL10 on the TCGA-LUAD dataset, with correlations with the ESTIMATE method reaching Pearson's R = 0.72 (**Supplementary File 2, Figure S5b**).

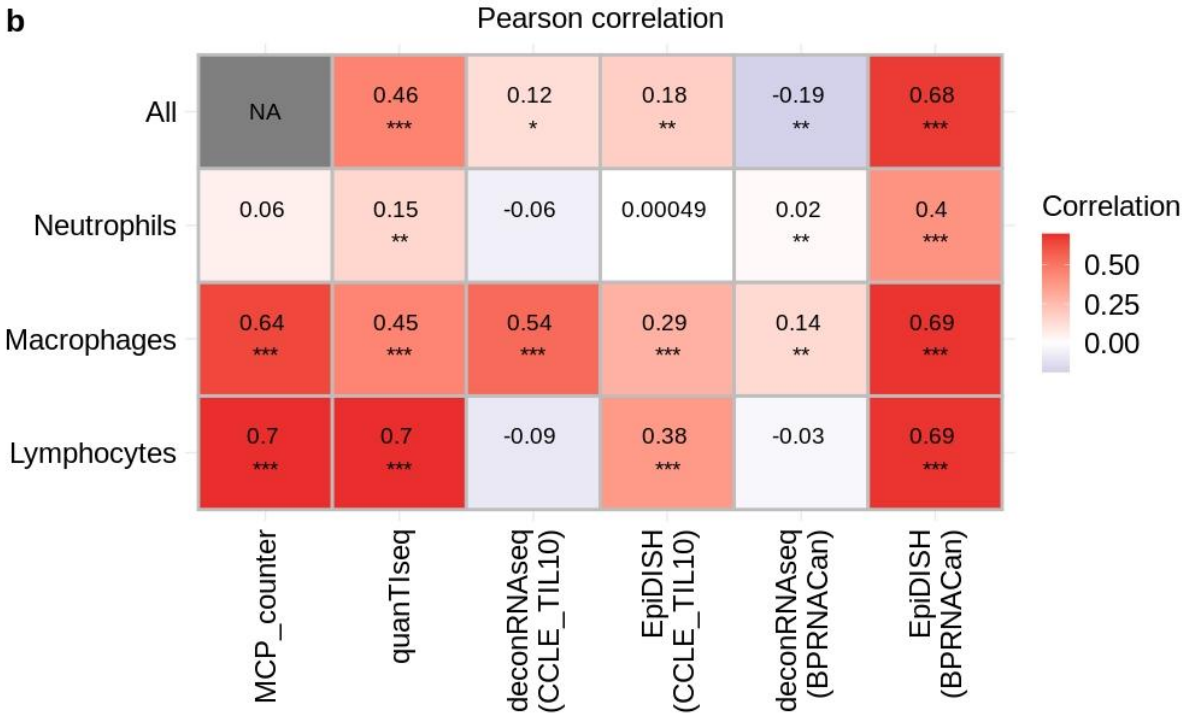
As far as different cell types are concerned, we again compared our deconvolved proportions to estimations based on H&E images [51] (**Supplementary file 2: Table S1**). Interestingly, once again, the BPRNACan signature matrix performed better than CCLE_TIL10 on biological samples (Fig. 6a). However, we observed the correlation for neutrophils was lower than expected, probably due to the presence of neutrophils inside tumors with a specific phenotype not captured by our signature matrix. In order to further test our signature matrices using a different method from EpiDISH, we again compared results to the ones using deconRNAseq with our two signature matrices, MCP-counter and quanTIseq. We observed that the deconRNAseq method using both CCLE_TIL10 or BPRNACan signature matrices produced estimates in weak agreement with H&E for all immune cell types. On the contrary, MCP-counter and quanTIseq estimates showed high correlations with H&E estimates for lymphocytes, but strikingly low correlations for neutrophils (**Fig. 6b and Supplementary file2: Figure S6**). In summary, we observed that none of the methods manage to achieve a high correlation with H&E for neutrophil proportions. However, in terms of accuracy, BPRNACan with EpiDISH emerged as the best performing method for this dataset (**Fig. 6b and Supplementary file2: Figure S6b**).

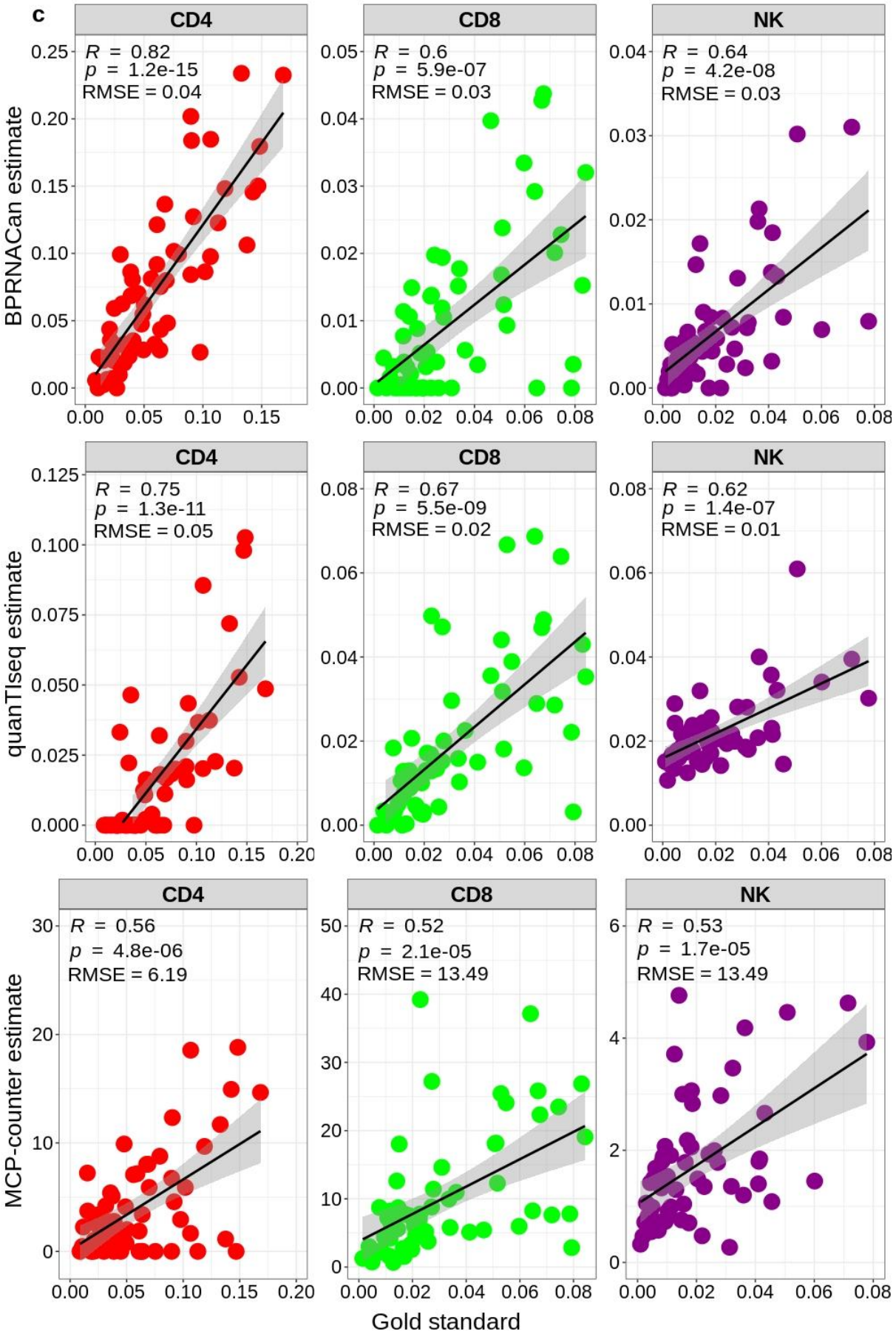
These analyses demonstrated that CCLE_TIL10 signature matrix was generated based on in-silico samples, making it accurate for in-vitro and in-silico mixtures, whereas BPRNACan performs better on biological samples, in which cells are in a biologically relevant context, including the exchange of signals between them. Thus we next tested our two signature matrices on additional in-vivo cancer samples.

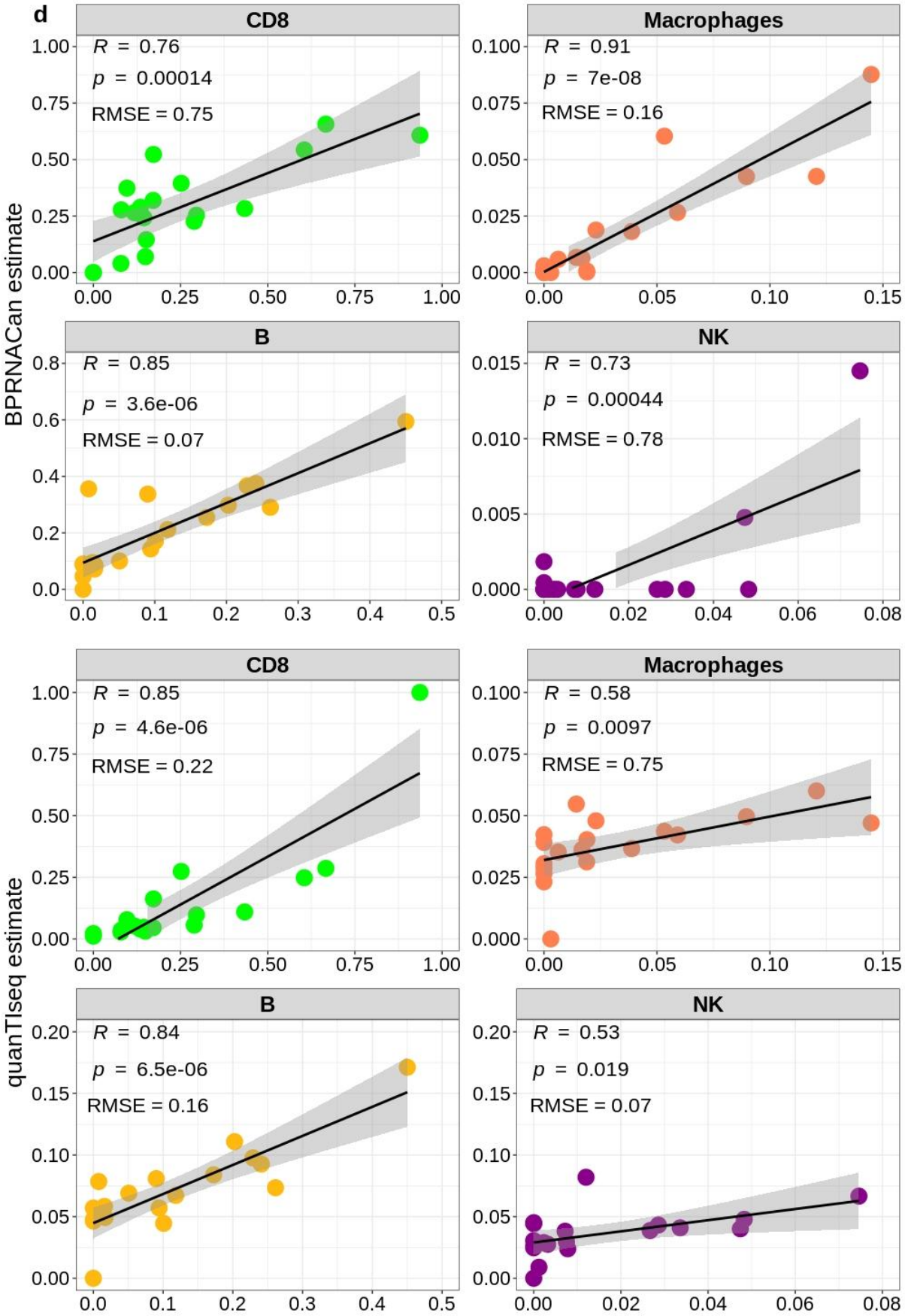
To further benchmark the methods, we used a series of 59 multiple myeloma patient TME samples, generated by removal of cancer cells from bone marrow samples [38]. For this dataset FACS analysis was performed to quantify NK, CD4 and CD8 T cells so we could compare our estimations for these cell types from the different methods to FACS results. Although only 3 immune cell types had been assayed in this study, BPRNACan produced higher correlations with FACS for CD4 T and NK cells, and lower correlations for CD8 T cells compared to quanTIseq (Fig. 6c), while for all three cell types it gives higher correlations than MCP-counter (Fig. 6c).

Finally, we examined a single-cell RNAseq dataset including 19 reconstructed primary melanoma non metastatic samples [36] and aggregated reads to produce an *in-silico* bulk sample for each patient to compare deconvolved proportions to the corresponding cell fractions estimated by counting single cells of each type from Racle et al. [35] (CD8 T cells, macrophages, B cells and NK cells). We ran EpiDISH deconvolution using the BPRNACan signature matrix and measured high correlations with the cell-type proportions estimated in the publication (Pearson's R > 0.7 for all 4 cell types, **Fig. 6d and Supplementary file 2: Figure S6c**). Moreover, we also explored four metastatic melanoma patients with available flow cytometry data (for CD4 T cells, CD8 T cells, B cells, NK cells, melanoma cells) provided in the same publication [35]. Upon comparing the cell proportion estimates derived using BPRNACan with estimates derived from MCP-counter and quanTIseq, we observed that BPRNACan has the most robust performance for cancer cells, CD4 and CD8. In addition, we once again observed the combination of BPRNACan with EpiDISH performed better than the combination of BPRNACan with deconRNAseq (**Supplementary file 2: Figure S6d**).


a
Lymphocytes
Macrophages
Neutrophils
R = 0.38
p = 7.5e-15
R = 0.29
p = 5.1e-09
R = 0.00049
p = 0.99
R = 0.69
p < 2.2e-16
R = 0.69
p < 2.2e-16
R = 0.4
p < 2.2e-16
CCLE_TIL10 estimate
BPRNACan estimate
H&E estimate


b
Pearson correlation
All
Neutrophils
Macrophages
Lymphocytes
NA
0.46 ***
0.12 *
0.18 **
-0.19 **
0.68 ***
0.06
0.15 **
-0.06
0.00049
0.02 **
0.4 ***
0.64 ***
0.45 ***
0.54 ***
0.29 ***
0.14 **
0.69 ***
0.7 ***
0.7 ***
-0.09
0.38 ***
-0.03
0.69 ***
MCP_counter
quanTIseq
deconRNAseq (CCLE_TIL10)
EpiDISH (CCLE_TIL10)
deconRNAseq (BPRNACan)
EpiDISH (BPRNACan)
Correlation
0.50
0.25
0.00


c
CD4
CD8
NK
R = 0.82
p = 1.2e-15
RMSE = 0.04
R = 0.6
p = 5.9e-07
RMSE = 0.03
R = 0.64
p = 4.2e-08
RMSE = 0.03
R = 0.75
p = 1.3e-11
RMSE = 0.05
R = 0.67
p = 5.5e-09
RMSE = 0.02
R = 0.62
p = 1.4e-07
RMSE = 0.01
R = 0.56
p = 4.8e-06
RMSE = 6.19
R = 0.52
p = 2.1e-05
RMSE = 13.49
R = 0.53
p = 1.7e-05
RMSE = 13.49
BPRNACan estimate
quanTIseq estimate
MCP-counter estimate
Gold standard


d
CD8
R = 0.76
p = 0.00014
RMSE = 0.75
Macrophages
R = 0.91
p = 7e-08
RMSE = 0.16
B
R = 0.85
p = 3.6e-06
RMSE = 0.07
NK
R = 0.73
p = 0.00044
RMSE = 0.78
BPRNACan estimate
CD8
R = 0.85
p = 4.6e-06
RMSE = 0.22
Macrophages
R = 0.58
p = 0.0097
RMSE = 0.75
B
R = 0.84
p = 6.5e-06
RMSE = 0.16
NK
R = 0.53
p = 0.019
RMSE = 0.07
quanTIseq estimate

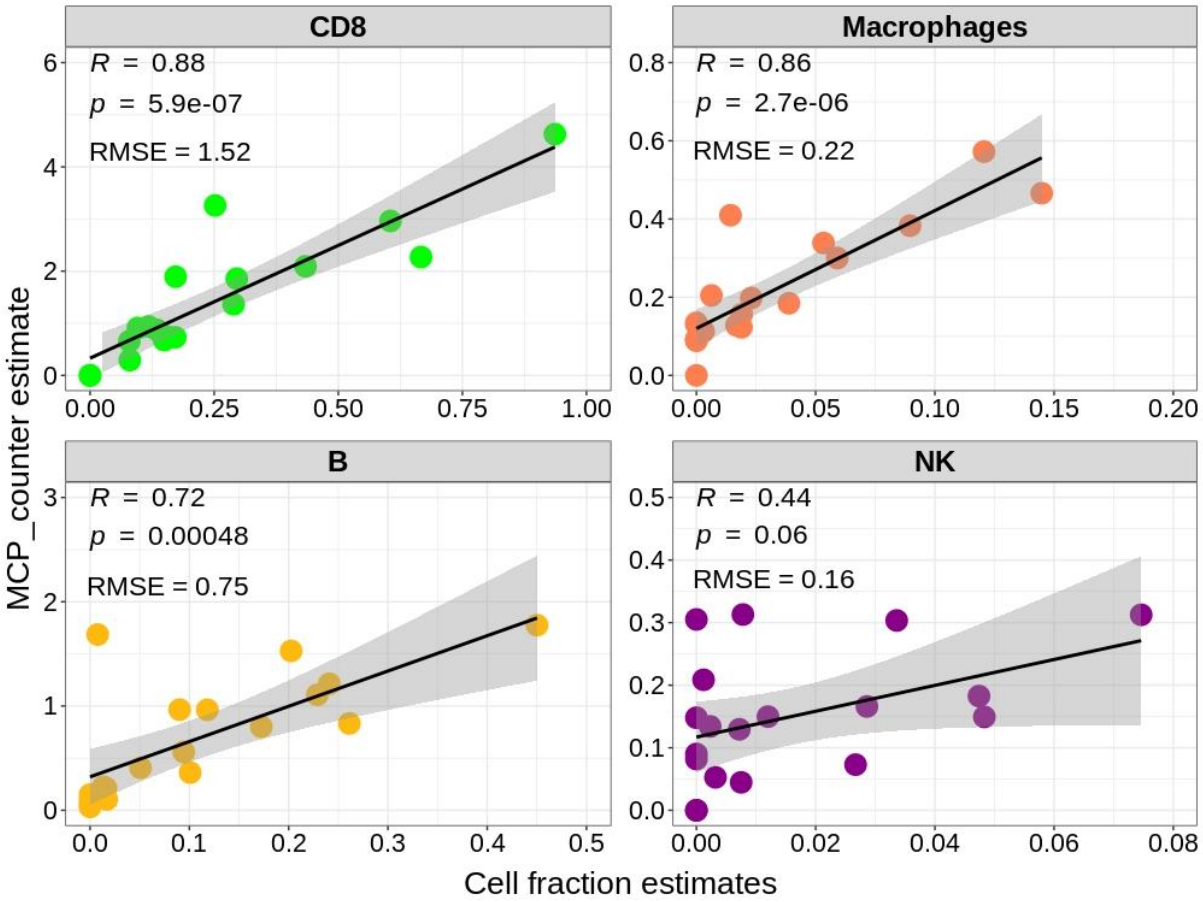


Fig. 6 Deconvolution in tumor samples

(**a**) Comparison of the proportion of immune between CCLE_TIL10 (**Top**) and BPRNAcan (**Bottom**) and H&E staining images in TCGA-LUAD [51]**.** (**b**) The correlation with cell type proportions measured on different datasets with samples containing cancer cells by different RNAseq reference-based deconvolution methods and experimental estimates based on H&E analysis [51]. The significance of the Pearson Correlation is indicated by stars: *p<0.05, **p<0.01, ***p<0.001. (**c**) Comparing cell type proportions against FACS data in 59 bone marrow samples [38]. (**d**) Scatter plots and Pearson correlations of each cell subtype proportion predicted by the BPRNACan, quanTIseq and MCP-counter and the true cell fractions from scRNAseq data from melanoma samples [36].

Gene expression and methylation based cell type signatures are significantly associated

In this part we investigated what could be the relation between the genes that are included in the BPRNACan signature matrix gene set (1403 *sig genes*) and the CpGs from the BPMetCan signature matrix (1896 *sig CpGs*), assuming that comparing both methylome and transcriptome of purified cells would yield similar genes as important in defining the cell type. We then calculated the number of *sig genes* that are associated with the *sig CpGs*. To associate CpGs to genes we first used Illumina annotation, which provides gene associations for each CpG on their array, based on genomic proximity between the gene and the CpG. Out of 169 genes associated with *sig CpGs*, 24 were included in the 1403 *sig genes* (Fisher's exact test p-value=1e-5) (**Supplementary File 2: Table S4a**).

Combining the methylation and expression based signature matrices

Inspired by the availability of 3D chromatin contact maps for haematopoietic cells produced by the Blueprint project [21] and by known relations between 3D conformation, methylation and expression levels [54], we decided to re-evaluate CpGs annotation to genes and consider 3D chromatin contacts between these CpGs and gene promoters, as identified by Promoter-Capture Hi-C (PCHi-C). Briefly, this method involves a step of hybridization to a promoter library during a traditional Hi-C experiment, resulting in chromatin contact maps that contain either contacts amongst promoters or between a promoter and a regulatory region (denoted as OE for Other end or PIR for Promoter Interacting Region) [55]. The total PCHi-C network for haematopoietic cells consists of 249,511 chromatin fragments, of which 20,582 are promoter fragments (including 1127 fragments of *sig gene* promoters) and 228,929 other ends, (including 1131 fragments containing *sig CpGs*) (**Supplementary file 2: Table S3**).

We first mapped *sig CpGs* to their corresponding gene promoters, according to the definition of promoters used in the PCHi-C datasets. We found 645 genes that have *sig CpGs* in their promoter, of which 52 were already identified in *sig genes* and 314 were not in *sig genes* and had expression profiles in our reference datasets. Moreover, the improvement of the mapping between *sig CpGs* and genes harboring *sig CpGs* in their promoter shows that the Blueprint project annotation is more accurate than Illumina annotation of promoter, Fisher's exact test p-value = 7.34e-08 (**Supplementary File 2: Table S4b**).

We reasoned that genes whose promoter methylation is cell-type specific should be important to include in the signature, even if they are not included in the gene expression signature. We therefore created an expanded gene expression deconvolution signature matrix by adding these 314 genes to BPRNACan signature, leading to the generation of BPRNACanProMet.

Given the importance of gene regulation by genomically distal regulatory elements that are brought into 3D contact with the promoter, having *sig CpGs* in both the promoter and a distal interacting fragment could strengthen the benefit of including the gene in the gene expression deconvolution signature. We therefore considered a more stringent expansion of our BPRNACan signature matrix to include the genes that have *sig CpGs* both inside their promoter and in regions that contact their promoter in 3D (BPRNACan3DProMet).

Finally, we considered expanding our BPRNACan signature with genes that have sig CpGs only in regions that are contacting their promoters (and not directly in their promoter), giving rise to the BPRNACan3DMet signature matrix, which included a large number of genes (N=3999).

The principle of gene inclusion in these 3 signatures is detailed in Fig. 7. We found the BPRNACanProMet and the BPRNACan3DProMet signature matrices to have better performance compared to the original BPRNACan signature matrix. (**Supplementary File 3: Table S6 and S7 and Supplementary File 2: Figure S7**). On the contrary BPRNACan3DMet included too many genes and did not improve performance.

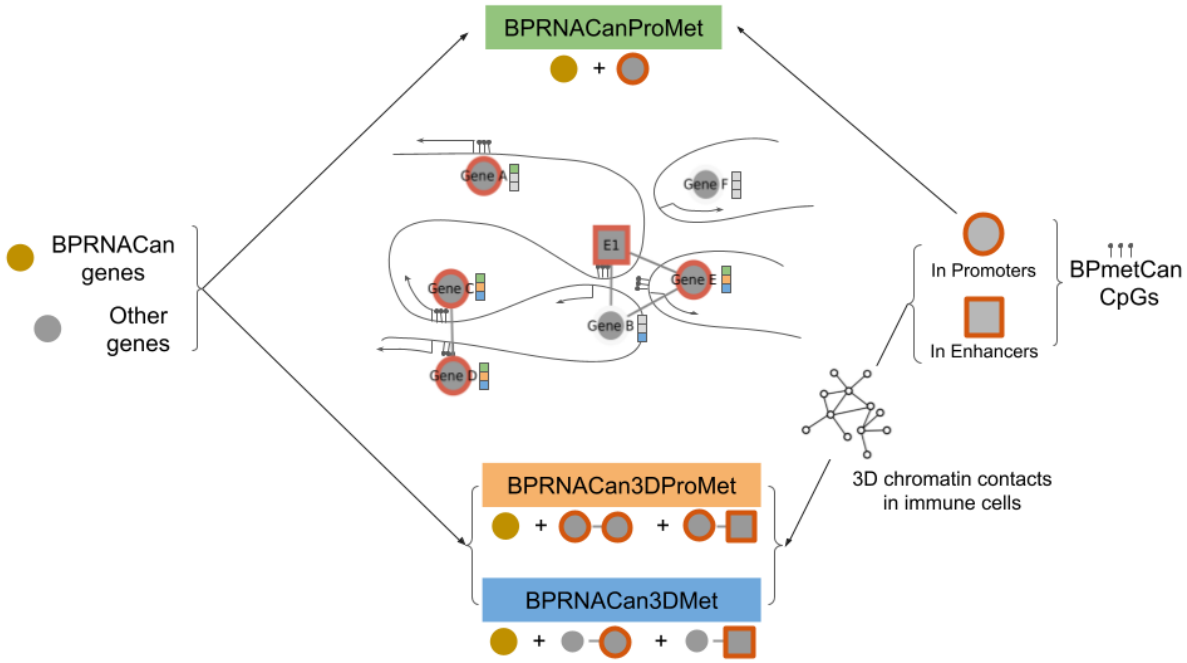


Fig 7: Strategy of integration of BPRNACan and BPmetCan signature matrices via 3D chromatin contact networks. The BPRNACan signature is expanded including genes that have *sig CpGs* in their promoters (BPRNACanProMet), genes that have *sig CpGs* only in 3D regions in contact with their promoters (BPRNACan3DMet) and genes that have *sig CpGs* in both promoter and 3D contacting regions (BPRNACan3DProMet).

Using deconvolution to predict response to immune checkpoint inhibitors

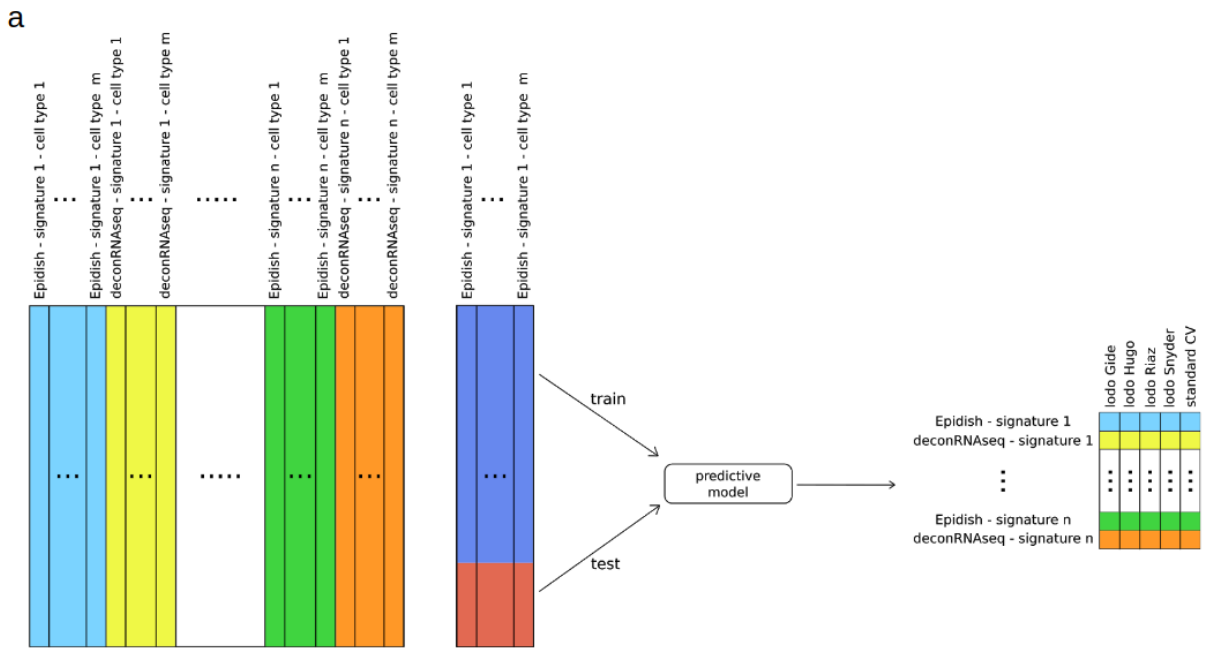
We have so far compared performances of different deconvolution methods and signatures based on cell type composition estimates provided by FACS, H&E image quantification of specific markers or single-cell data. All of these methods, however, carry their own biases in estimating composition. One of the main reasons that we are interested in quantifying cell types inside the TME is to better understand and predict response to immune check-point inhibitors in a personalized approach. This remains challenging and biomarkers of response are needed to achieve better outcomes in immunotherapy. We therefore decided to test the pertinence of our signature matrices by estimating the value that they carry in models that predict response to immunotherapy based on gene expression (Fig. 8a).

To this end, we made predictors of response to anti-PD1 antibodies using different deconvolution results as features and evaluated their performance on 3 public melanoma datasets and one bladder cancer dataset [39–42] with response to anti-PD1 through ElasticNet penalized logistic regression [62] (see Methods and Fig. 8b).

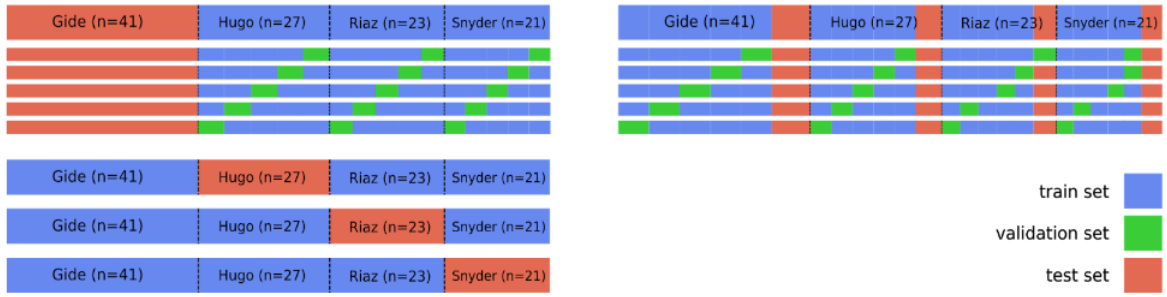
Our signature matrices in combination with Epidish or DeconRNA allowed us to train models with a performance that was better than random in a classification task (ROC AUC score above 0.5 on the standard Cross Validation (CV) training for 10 models among 12) and better than using quanTIseq or MCPCounter. While training on all samples (standard CV), the BPRNACanProMet signature used by Epidish was the top performing model, with a ROC

AUC score of 0.703 (Fig. 8c). The BPRNACan signature in combination with deconRNAseq is the only one that could produce models with a ROC AUC score always above 0.5 (Fig. 8c), even when predicting on datasets that had not been used for training. When excluding it from training, the Snyder bladder cancer dataset is the one with the overall worst performing models, which is probably because it is the only dataset with bladder cancer samples, all the other ones corresponding to melanoma [39–41].

Looking at the coefficients in our regression models we can estimate which variables are associated with either progressive diseases or response to therapy. The coefficients of the model trained with the combination of the BPRNACan3DProMet signature and the Epidish method show that B cells [56–58] (presumably indicative of the presence of tertiary lymphoid structures) and M0 macrophages proportions are associated with response of patients, whereas M2 macrophages [59], CD4, CD8 and NK cells associated with progressive disease in this predicting model (Fig. 8d). Proportions of cancer cells, M1 macrophages, monocytes and neutrophils were not significantly associated with either of the two outcomes. We can also consider the combination of the BPRNACan signature with deconRNAseq that always produced models performing better than a random predictor, and that is the second best when Snyder dataset was used as a test set. The coefficients of this model indicate that B cells, M0 and M1 macrophages [60] proportions are associated with response of patients, whereas CD4 and CD8 cells, Nk cells, monocytes and M2 macrophages are associated with progressive disease in this model (Fig. 8d). Like the previous model, proportions of cancer cells and neutrophils were not strongly associated with any of the two outcomes. Thus, the two models differ in the importance of M1 macrophages and monocytes for predicting response or progressive disease.



b



c

Signature	Gide	Hugo	Riaz	Snyder	CV all datasets
quanTIseq	0.643	0.500	0.568	0.333	0.510
MCP-counter	0.675	0.467	0.477	0.639	0.552
Epidish_BPRNACan	0.500	0.478	0.773	0.491	0.646
deconRNAseq_BPRNACan	0.655	0.550	0.667	0.509	0.594
Epidish_BPRNACanProMet	0.663	0.606	0.629	0.472	0.703
deconRNAseq_BPRNACanProMet	0.620	0.617	0.606	0.481	0.568
Epidish_BPRNACan3DProMet	0.575	0.550	0.652	0.500	0.688
deconRNAseq_BPRNACan3DProMet	0.645	0.606	0.606	0.463	0.568
Epidish_BPRNACan3DMet	0.500	0.544	0.773	0.426	0.490
deconRNAseq_BPRNACan3DMet	0.500	0.567	0.659	0.426	0.427

d

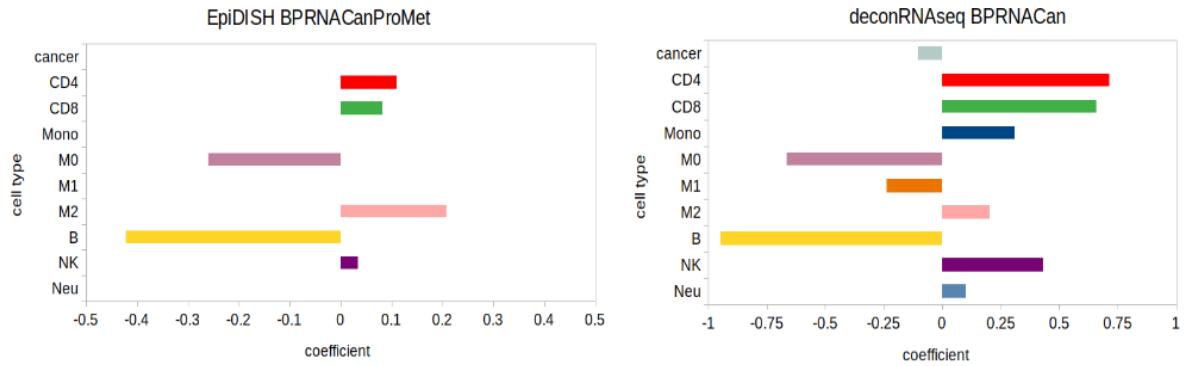


Fig 8: Evaluation of predictive power of signatures in combinations with deconvolution methods. (**a**) Each combination of signature and deconvolution method was used alone to estimate cell types proportions for all samples. This data was used to train and test an elasticnet [43] penalized logistic regression model to predict response to immunotherapy. Performance of models can then be compared across signatures and training methods. (**b**) **Left:** One training method is leave-one-dataset-out (lodo), where one dataset is used for testing (red) and the other ones are used for training (blue). During each training phase, hyperparameters for the l1 ratio and regularization strength were searched for by 5-fold cross validation (CV), where the training set is subdivided in 5 parts, each one of them being used as a validation set (green) while the other ones are used for training. At the end of the CV search the model is re-trained on all training samples, and tested on the test set. **Right:**

The last training method is standard cross validation, during which a fourth of all samples is held-out as a test set. The remaining samples are used for training, with a 5-fold CV search for the hyperparameters that includes successive training sets (blue) and validation sets (green). (**c**) Table of models performances by combinations of signatures and deconvolution methods and by training methods. The performance is assessed with the ROC AUC score computed on test sets, which is indicated by the name of the dataset for the lodo training. For the standard CV, a fifth of samples was held out as a test set. (**d**) Coefficients of logistic regression models for the BPRNACanProMet signature with Epidish (left) and for the BPRNACan signature with deconRNAseq (right). Variables with negative coefficients are associated with patient response, the amplitude of coefficients is related to variable importance.

DISCUSSION

Numerous reference-based deconvolution methods using DNA methylation or gene expression can be used to estimate the proportion of cell types in bulk datasets from cell mixtures, such as EpiDISH, MethylCibersort (for DNAm) and CIBERSORT, MCP-counter, quanTIseq, DeconRNASeq (for GE). Moreover, there are methods that can accurately predict the proportion of cancer cells in tumor samples (purity). However, for application in immuno-oncology it can be important to estimate the proportions of cancer and immune cells at the same time.

Recently, Chakravarthy *et al.*[47] have reported that DNAm-based or GE-based deconvolution methods could be complementary to each other in cases where both data types are available. Despite this, they did not further explore if the reference signature matrices for each type of data can be improved by combining them. Here we present five novel signature matrices for reference-based deconvolution:

BPmetCan: which is able to deconvolve the proportion of cancer and immune cell types from DNA methylation data (Arrays or WGBS), based on WGBS signature matrices, which we validate against MethylCIBERSORT, EpiDISH and various methods to estimate tumor purity in blood and especially in tumor samples.

CCLE_TIL10: combines the TIL10 signature matrix [28] for immune cell types with a new list of genes identified to be cancer-cell specific using data from GTEx and CCLE, which displays excellent performance on in-vitro and in-silico mixtures of cancer cell lines and immune cells.

BPRNACan: combines our Blueprint derived immune cell signature matrix (BPRNA) with genes that are discriminant of cancer tissues compared to normal and outperforms or equals quanTIseq and MCP-Counter on cancer samples for many of the cell types.

BPRNACanProMet: is an enhancement of the BPRNACan signature matrix by adding genes that have a *sig CpG* contained in their promoter.

BPRNACan3DProMet: is an enhancement of the BPRNACan signature matrix by adding genes that have a *sig CpG* in their promoter and whose promoters also have a 3D contact with a fragment containing a *sig CpG.*

We performed extensive validation using previous studies, such as whole blood mixtures, solid tumor-TCGA, PBMC, multiple myeloma patient TME bone marrow samples and melanoma non metastatic and metastatic samples. We compared the available DNAm or GE signature matrices to demonstrate that our novel signature matrices could faithfully estimate the fraction of cancer and specific immune cell compositions from DNA methylation and bulk gene expression data. Our signature matrices can be applied to solid tumors, as confirmed by the validations presented above, but they are likely to have limited use for hematological malignancies, in which the presence of cancerous immune cells could confound the estimations, and for which more targeted signature matrices should be developed. We also showed that our DNAm signature matrix BPmet is more robust than others for estimating the proportions of cell types in Whole Blood samples. Moreover, our new gene expression signature matrix BPRNACan displayed higher accuracy on the predicted cell fractions on in-vivo cancer samples. Additionally, application of our GE signature matrices to publicly available data using our predicting model in this study revealed several important biological insights in response to immunotherapy.

Despite the overall accuracy of our signature matrices, we found that the performance in the estimation for some cell types is lower than what we expected, probably due to the number or condition of the cell types used for establishing the reference profiles. For instance, we observed that the correlation of NK cells was lower than expected using both BPmet and BPRNA (**Fig. 5a-b, and Supplementary file 2: Figure S1a, S2a**). This may be explained as we had only n = 2 samples to create the NK reference expression profile. Moreover, we observed a very low performance in predicting M2 macrophages in the *in-silico* tumor samples using BPRNACan (**Supplementary file 2: Figure S4b**). This may be explained by the fact that this M2 state in-vitro is artificially induced by cytokines to mimic the context of the TME, thereby the GE reference profiles corresponding to M2 cells might not capture the M2s that are found, albeit at low frequency [61] in artificial *in-silico* mixtures of purified cells. As opposed to NK and M2 cells, CD8 T cells and Macrophages were better predicted using GE signature matrices on cancer samples (**Fig. 6d and Supplementary file 2: Figure S4 and S6**) rather than on PBMCs (Fig. 5a). This could be explained by the differences in cell proportions and cell states between the tumor microenvironment and circulating blood. For example, only a few activated CD8 T cells and no macrophages can be found in circulating blood, but their detection in tumors is key.

We also found a low performance in detecting neutrophils according to proportions derived by H&E (Fig. 4b, 6a). This may be explained by the fact that the reference profiles of neutrophils used to build the signature matrix are unlikely to capture all the phenotypes that neutrophils can display, especially inside tumors. Like macrophages, Tumor Associated Neutrophils (TANs) can display at least two different phenotypes - one characterised by proinflammatory programs and antitumorigenic functions and the second characterised by a protumorigenic activity [62,63]. Indeed neutrophils' gene expression and methylation were found

to be extremely variable across individuals [64], time of the day, as well as across the different parts of the body in which they are found, highlighting their extreme plasticity [65,66].

We also demonstrated that our GE signature matrices often perform better at estimating cell types proportions with EpiDISH rather than with deconRNASeq. This result is generally consistent between blood or tumor samples, except for a higher correlation of CD4 T cells while using deconRNAseq in PBMC data (Fig. 5b).

Another important novel insight of our study is that the GE signature matrix can be improved by incorporating specific genes that are associated with CpGs which are included in the DNA methylation signature matrix. This can be done by either identifying genes that harbor signature matrix CpGs (*sig CpGs)* in their promoter, or those genes whose promoters also have a 3D contact with a fragment containing a *sig CpG*. We are thus able to expand the list of genes to be used in performing RNAseq-based deconvolution using information gathered from the DNA methylation signature matrix we have generated (**Supplementary file 2: Table S2**). This expanded GE signature matrix (**Supplementary file 2: Table S6-S7**) can be applied to deconvolve samples for which only RNAseq data is available.

Despite our DNAm and RNAseq reference-based signature matrices displaying comparable to or better performance than existing methods in whole blood or cancer samples, several issues will require further investigation. We are limited by the number of samples available for specific cell types (such as NK cells) and by the fact that these profiles are generated from purified cells that are isolated from their natural environment. This is especially true for cells that acquire specific phenotypes in a TME context, such as TAMs and TANs. The availability of single-cell RNAseq datasets from cancer samples, especially from technologies that also provide protein marker quantification such as CITEseq, will greatly improve our chances of generating relevant signature matrices for any cell type of interest. For this reason we provide code for the generation of new reference signature matrices in our openly available repository.

Importantly, one of the main applications of deconvolution in immuno-oncology will be the prediction of response to immunotherapy, which can be made based on the inferred cell type proportions. We therefore benchmarked our new signature matrices and methods by evaluating their accuracy in predicting response to anti-PD1 agents in 4 public datasets. This type of exercise is aimed at identifying which signature matrices and methods uncover the presence of specific cell subtypes that can impact immune checkpoint blocker response. These important subtypes might not correspond easily to literature definitions or FACS derived populations. If M2 macrophages are known to impair response to immune checkpoint blockade [59], interestingly, B cells and M1 macrophages have recently been proposed as potential predictors of response to immunotherapies [56–58,60]. To our knowledge, the proportion of M0 (naïve) macrophages have not been reported as predictors to response to immunotherapies, but both of our models suggest it is a potential relevant measurement to predict patient response. If the 2 models also suggest that proportions of CD4, CD8 and NK cells are associated with progressive disease, this has to be considered with caution, as response to immunotherapy has been shown to depend on the proportion of sub-cell types

like memory, effector, and senescent phenotypes or defined by the presence or absence of several proteins like PD-1, PD-L1, CTLA-4, LAG-3 or TIGIT on these cells' surfaces.

In summary, we have shown the potential of our gene expression signature matrices to estimate the presence of immune populations that can be predictive of the response to checkpoint blockade, bringing us closer to personalized approaches and revealing resistance mechanisms.

CONCLUSION

We have presented and thoroughly validated five novel deconvolution signature matrices BPmetCan, CCLE_TIL10, BPRNACan, BPRNACanProMet and BPRNACan3DProMet, which show good performances in estimating the proportion of cell types from blood and tumor samples. Simultaneously, we have also shown that our signature matrices are more robust to estimate cell fractions compared to the other available signature matrices, and we have highlighted the relationship between genes in the expression signature matrix and CpGs in the methylation signature matrix. We also showed how a gene expression signature matrix can be improved by addition of genes that are recognized as being important through the creation of a methylation signature matrix, exploiting the knowledge of 3D chromatin contacts between promoters and regulatory elements now available through PCHi-C networks for 17 immune cell types [21]. Our signature matrices are particularly suitable for the analysis of tumor samples from an immuno-oncology perspective, as they provide accurate estimates of immune cells as well as cancer cell proportions and have predictive power for estimating response to anti-PD1 agents. We make the signature matrices and all the code available to the community through a user-friendly snakemake pipeline that can use any given reference signature matrix to apply a variety of reference-based deconvolution methods. Additionally, the code to generate a signature matrix following the method used in this paper is also available as well as a script to generate all the figures.

ABBREVIATIONS

BRCA: Breast invasive carcinoma

CCLE: Cancer Cell Line Encyclopedia

DHS: DNAse hypersensitivity sites

DNAm: DNA methylation

FACS: fluorescence-activated cell sorting

FPKM: Fragments Per Kilobase of transcript per Million

GE: gene expression

GEO: Gene expression omnibus

H&E: Hematoxylin and eosin

IHC: Immunohistochemistry

LUAD: Lung adenocarcinoma

M: macrophages

M1: Classically activated macrophages

M2: Alternatively activated macrophages

Mono: Monocytes

Neu: Neutrophils

NK: Natural killer cells

PBMC: Peripheral blood mononuclear cells

PCHi-C: Promoter-Capture Hi-C

R: Pearson's correlation

RPC: robust partial correlation

TANs: tumor Associated Neutrophils

TAMs: tumor Associated Macrophages

TCGA: The Cancer Genome Atlas

TME: The tumor microenvironment

TPM: Transcripts per millions

Treg : Regulatory T cells

WB: whole blood

WGBS: whole-genome bisulfite sequencing

REFERENCES:

1. DeBerardinis, R. J. Tumor Microenvironment, Metabolism, and Immunotherapy. *New England Journal of Medicine* vol. 382 869–871 (2020).
2. Engblom, C., Pfirschke, C. & Pittet, M. J. The role of myeloid cells in cancer therapies. *Nat. Rev. Cancer* **16**, 447–462 (2016).
3. O'Donnell, J. S., Teng, M. W. L. & Smyth, M. J. Cancer immunoediting and resistance to T cell-based immunotherapy. *Nature Reviews Clinical Oncology* vol. 16 151–167 (2019).
4. Wei, S. C., Duffy, C. R. & Allison, J. P. Fundamental Mechanisms of Immune Checkpoint Blockade Therapy. *Cancer Discov.* **8**, 1069–1086 (2018).
5. Sharma, P., Hu-Lieskovan, S., Wargo, J. A. & Ribas, A. Primary, Adaptive, and Acquired Resistance to Cancer Immunotherapy. *Cell* **168**, 707–723 (2017).
6. Anderson, K. G., Stromnes, I. M. & Greenberg, P. D. Obstacles Posed by the Tumor Microenvironment to T cell Activity: A Case for Synergistic Therapies. *Cancer Cell* vol. 31 311–325 (2017).
7. Mantovani, A., Marchesi, F., Malesci, A., Laghi, L. & Allavena, P. Tumour-associated macrophages as treatment targets in oncology. *Nat. Rev. Clin. Oncol.* **14**, 399–416 (2017).
8. Pathria, P., Louis, T. L. & Varner, J. A. Targeting Tumor-Associated Macrophages in Cancer. *Trends in Immunology* vol. 40 310–327 (2019).
9. Lawrence, T. & Natoli, G. Transcriptional regulation of macrophage polarization: enabling diversity with identity. *Nat. Rev. Immunol.* **11**, 750–761 (2011).
10. Murray, P. J. Macrophage Polarization. *Annu. Rev. Physiol.* **79**, 541–566 (2017).
11. Finotello, F., Rieder, D., Hackl, H. & Trajanoski, Z. Next-generation computational tools for interrogating cancer immunity. *Nat. Rev. Genet.* **20**, 724–746 (2019).
12. Aran, D., Sirota, M. & Butte, A. J. Systematic pan-cancer analysis of tumour purity. *Nat. Commun.* **6**, 8971 (2015).

13. Zou, J., Lippert, C., Heckerman, D., Aryee, M. & Listgarten, J. Epigenome-wide association studies without the need for cell-type composition. *Nat. Methods* **11**, 309–311 (2014).
14. Houseman, E. A., Molitor, J. & Marsit, C. J. Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics* **30**, 1431–1439 (2014).
15. Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 86 (2012).
16. Visvader, J. E. Cells of origin in cancer. *Nature* vol. 469 314–322 (2011).
17. Titus, A. J., Gallimore, R. M., Salas, L. A. & Christensen, B. C. Cell-type deconvolution from DNA methylation: a review of recent applications. *Hum. Mol. Genet.* **26**, R216–R224 (2017).
18. Cavalli, G. & Heard, E. Advances in epigenetics link genetics to the environment and disease. *Nature* **571**, 489–499 (2019).
19. Timp, W. *et al.* Large hypomethylated blocks as a universal defining epigenetic alteration in human solid tumors. *Genome Med.* **6**, 61 (2014).
20. Stunnenberg, H. G. & Hirst, M. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell* vol. 167 1897 (2016).
21. Javierre, B. M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167**, 1369–1384.e19 (2016).
22. Brinkman, A. B. *et al.* Partially methylated domains are hypervariable in breast cancer and fuel widespread CpG island hypermethylation. *Nat. Commun.* **10**, 1749 (2019).
23. Ziller, M. J., Hansen, K. D., Meissner, A. & Aryee, M. J. Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing. *Nat. Methods* **12**, 230–2, 1 p following 232 (2015).
24. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing

and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

25. Farha, M., Jairath, N. K., Lawrence, T. S. & El Naqa, I. Characterization of the Tumor Immune Microenvironment Identifies M0 Macrophage-Enriched Cluster as a Poor Prognostic Factor in Hepatocellular Carcinoma. *JCO Clin Cancer Inform* **4**, 1002–1013 (2020).
26. Binnewies, M. *et al.* Understanding the tumor immune microenvironment (TIME) for effective therapy. *Nat. Med.* **24**, 541–550 (2018).
27. Wang, Q. *et al.* Unifying cancer and normal RNA sequencing data from different sources. *Sci Data* **5**, 180061 (2018).
28. Finotello, F. *et al.* Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. *Genome Med.* **11**, 34 (2019).
29. Ghandi, M. *et al.* Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503–508 (2019).
30. Carithers, L. J. & Moore, H. M. The Genotype-Tissue Expression (GTEx) Project. *Biopreservation and Biobanking* vol. 13 307–308 (2015).
31. Cairns, J. *et al.* CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. *Genome Biol.* **17**, 127 (2016).
32. Koestler, D. C. *et al.* Improving cell mixture deconvolution by identifying optimal DNA methylation libraries (IDOL). *BMC Bioinformatics* vol. 17 (2016).
33. Teschendorff, A. E., Breeze, C. E., Zheng, S. C. & Beck, S. A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. *BMC Bioinformatics* **18**, 105 (2017).
34. Monaco, G. *et al.* RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types. *Cell Rep.* **26**, 1627–1640.e7 (2019).
35. Racle, J., de Jonge, K., Baumgaertner, P., Speiser, D. E. & Gfeller, D. Simultaneous

enumeration of cancer and immune cell types from bulk tumor gene expression data. *Elife* **6**, (2017).

36. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
37. Samur, M. K. RTCGAToolbox: a new tool for exporting TCGA Firehose data. *PLoS One* **9**, e106397 (2014).
38. Nakamura, K. *et al.* Dysregulated IL-18 Is a Key Driver of Immunosuppression and a Possible Therapeutic Target in the Multiple Myeloma Microenvironment. *Cancer Cell* vol. 33 634–648.e5 (2018).
39. Hugo, W. *et al.* Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. *Cell* **168**, 542 (2017).
40. Riaz, N. *et al.* Tumor and Microenvironment Evolution during Immunotherapy with Nivolumab. *Cell* **171**, 934–949.e16 (2017).
41. Gide, T. N. *et al.* Distinct Immune Cell Populations Define Response to Anti-PD-1 Monotherapy and Anti-PD-1/Anti-CTLA-4 Combined Therapy. *Cancer Cell* **35**, 238–255.e6 (2019).
42. Snyder, A. *et al.* Contribution of systemic and somatic factors to clinical response and resistance to PD-L1 blockade in urothelial cancer: An exploratory multi-omic analysis. *PLoS Med.* **14**, e1002309 (2017).
43. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33**, 1–22 (2010).
44. Köster, J. & Rahmann, S. Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics* **34**, 3600 (2018).
45. Anaconda Software Distribution. Computer software. Vers. 2-2.4.0. Anaconda, Nov. 2016. Web. .
46. Farlik, M. *et al.* DNA Methylation Dynamics of Human Hematopoietic Stem Cell

Differentiation. *Cell Stem Cell* **19**, 808–822 (2016).

47. Chakravarthy, A. *et al.* Author Correction: Pan-cancer deconvolution of tumour composition using DNA methylation. *Nat. Commun.* **9**, 4642 (2018).
48. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
49. Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nature Biotechnology* vol. 30 413–421 (2012).
50. Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4**, 2612 (2013).
51. Saltz, J. *et al.* Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images. *Cell Rep.* **23**, 181–193.e7 (2018).
52. Gong, T. & Szustakowski, J. D. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. *Bioinformatics* **29**, 1083–1085 (2013).
53. Becht, E. *et al.* Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol.* **17**, 218 (2016).
54. Madrid-Mencía, M., Raineri, E., Cao, T. B. N. & Pancaldi, V. Using GARDEN-NET and ChAseR to explore human haematopoietic 3D chromatin interaction networks. *Nucleic Acids Res.* **48**, 4066–4080 (2020).
55. Schoenfelder, S. *et al.* The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res.* **25**, 582–597 (2015).
56. Petitprez, F. *et al.* B cells are associated with survival and immunotherapy response in sarcoma. *Nature* **577**, 556–560 (2020).
57. Cabrita, R. *et al.* Tertiary lymphoid structures improve immunotherapy and survival in melanoma. *Nature* **577**, 561–565 (2020).

58. Helmink, B. A. *et al.* B cells and tertiary lymphoid structures promote immunotherapy response. *Nature* **577**, 549–555 (2020).

59. Ceci, C., Atzori, M. G., Lacal, P. M. & Graziani, G. Targeting Tumor-Associated Macrophages to Increase the Efficacy of Immune Checkpoint Inhibitors: A Glimpse into Novel Therapeutic Approaches for Metastatic Melanoma. *Cancers* **12**, (2020).

60. Zeng, D. *et al.* Macrophage correlates with immunophenotype and predicts anti-PD-L1 response of urothelial cancer. *Theranostics* **10**, 7002–7014 (2020).

61. Shen-Orr, S. S. & Gaujoux, R. Computational deconvolution: extracting cell type-specific information from heterogeneous samples. *Curr. Opin. Immunol.* **25**, 571–578 (2013).

62. Shaul, M. E. & Fridlender, Z. G. Tumour-associated neutrophils in patients with cancer. *Nat. Rev. Clin. Oncol.* **16**, 601–620 (2019).

63. Teijeira, Á. *et al.* CXCR1 and CXCR2 Chemokine Receptor Agonists Produced by Tumors Induce Neutrophil Extracellular Traps that Interfere with Immune Cytotoxicity. *Immunity* **52**, 856–871.e8 (2020).

64. Ecker, S. *et al.* Genome-wide analysis of differential transcriptional and epigenetic variability across human immune cell types. *Genome Biol.* **18**, 18 (2017).

65. Giese, M. A., Hind, L. E. & Huttenlocher, A. Neutrophil plasticity in the tumor microenvironment. *Blood* **133**, 2159–2167 (2019).

66. Sagiv, J. Y. *et al.* Phenotypic diversity and plasticity in circulating neutrophil subpopulations in cancer. *Cell Rep.* **10**, 562–573 (2015).

67. Chu, D. *et al.* Nanoparticle Targeting of Neutrophils for Improved Cancer Immunotherapy. *Adv. Healthc. Mater.* **5**, 1088–1093 (2016).

ACKNOWLEDGEMENTS

We thank members of the Pancaldi lab for critical reading of the manuscript and Sarah Djebali for help with testing the pipeline.

FUNDING

This work was funded by the Chair of Bioinformatics in Oncology of the CRCT (INSERM; Fondation Toulouse Cancer Santé and Pierre Fabre Research Institute).

Availability of data and materials

The published data used is indicated in Methods. The pipeline to generate all examples is available from GitHub: https://github.com/VeraPancaldiLab/GEMDeCan and we provide an R notebook to reproduce all figures in the paper.

Authors' contributions

TX developed the signature matrices from the Blueprint project and performed deconvolution analyses and wrote the manuscript with VP. MSK developed the CCLE_TIL10 signature matrix. AH contributed to improve the GE signature matrix by using the PCHi-C network. TX, JP and MMM implemented the pipeline. JP, MMM and NV tested and corrected the pipeline. NV wrote the pipeline tutorial. TX performed statistical analyses. AC predicted response to immune checkpoint inhibitors. JS revised the text and edited figures. FC, OD and VP supervised the project. All authors read and approved the final manuscript.

Supplementary files

Supplementary file 1:

List of datasets used in this study for DNA methylation and gene expression reference database construction

Supplementary file 2:

Method S1. The threshold of log2 fold change (logFC) chosen. **Figure S1.** Validation of BPmet signature matrix and comparing 3 different DNA methylation signature matrices on top of 100 whole blood data with FACS (GSE132203). **Figure S2.** Comparing 3 different DNA methylation signature matrices in 6 WB with FACS from [32]. **Figure S3.** Validation of BPmetCan and comparing previously published purity estimation methods from BRCA. **Figure S4.** In silico validation of CCLE_TIL10 and RPRNACan from [28]. **Figure S5.** Comparing TCGA-LUAD tumor purity proportion using CCLE_TIL10 and BPRNACan with other previously published purity estimation methods. **Figure S6.** Summary of the performance of our signatures. **Figure S7.** Summary of the performance of our expansion gene expression signatures. **Table S1.** Validation datasets used to evaluate our expansion signature matrix in this study. **Table S2.** Main reference-based deconvolution methods used. **Table S3.** PCHi-C network analysis. **Table S4.** Fisher's test.

Supplementary file 3:

Seven novel deconvolution signature matrices BPmet, BPmetCan, BPRNA, BPRNACan, CCLE_TIL10, BPRNACanProMet, BPRNACan3DProMet.