

Robust estimation of cancer and immune cell-type proportions from bulk tumor ATAC-Seq data

Reviewed Preprint

v2 • August 7, 2024

Revised by authors

Reviewed Preprint

v1 • April 3, 2024

Aurélié AG Gabriel, Julien Racle, Maryline Falquet, Camilla Jandus, David Gfeller 

Department of Oncology, Ludwig Institute for Cancer Research, University of Lausanne, Lausanne, Switzerland • Agora Cancer Research Centre, Lausanne, Switzerland • Swiss Cancer Center Leman (SCCL), Switzerland • Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland • Ludwig Institute for Cancer Research, Lausanne Branch, Lausanne, Switzerland • Department of Pathology and Immunology Faculty of Medicine, University of Geneva, Geneva, Switzerland • Geneva Center for Inflammation Research, Geneva, Switzerland

 https://en.wikipedia.org/wiki/Open_access

 Copyright information

Abstract

Assay for Transposase-Accessible Chromatin sequencing (ATAC-Seq) is a widely used technique to explore gene regulatory mechanisms. For most ATAC-Seq data from healthy and diseased tissues such as tumors, chromatin accessibility measurement represents a mixed signal from multiple cell types. In this work, we derive reliable chromatin accessibility marker peaks and reference profiles for most non-malignant cell types frequently observed in the tumor micro-environment. We then integrate these data into the EPIC deconvolution framework (Racle *et al.*, 2017) to quantify cell-type heterogeneity in bulk ATAC-Seq data. Our EPIC-ATAC tool accurately predicts non-malignant and malignant cell fractions in tumor samples. When applied to a breast cancer cohort, EPIC-ATAC accurately infers the immune contexture of the main breast cancer subtypes.

eLife assessment

This study presents an **important** computational tool for the quantification of the cellular composition of human tissues profiled with ATAC-seq. The methodology is **solid** and its application results on breast cancer tumor tissues are **convincing**. It advances existing methods by utilizing a comprehensive reference profile for major cancer-relevant cell types, compatible with a widely-used cell type deconvolution tool.

<https://doi.org/10.7554/eLife.94833.2.sa3>

Introduction

Gene regulation is a dynamic process largely determined by the physical access of chromatin-binding factors such as transcription factors (TFs) to regulatory regions of the DNA (e.g., enhancers and promoters) (Klemm, Shipony and Greenleaf, 2019 ). The genome-wide landscape of

chromatin accessibility is essential in the control of cellular identity and cell fate and thus varies in different cell types (Klemm, Shipony and Greenleaf, 2019; K. Zhang *et al.*, 2021). Over the last decade, Assay for Transposase-Accessible Chromatin (ATAC-Seq) (Buenrostro *et al.*, 2013) has become a reference epigenomic technique to profile chromatin accessibility and the activity of gene regulatory elements in diverse biological contexts including cancer (Luo, Gribskov and Wang, 2022) and across large cohorts (Corces *et al.*, 2018). Several optimized ATAC-seq protocols have been developed to improve the quality of ATAC-Seq data and expand its usage to different tissue types. These include the OMNI-ATAC protocol, which leads to cleaner signal and is applicable to frozen samples (Corces *et al.*, 2017; Grandi *et al.*, 2022), as well as the formalin-fixed paraffin-embedded (FFPE)-ATAC protocol adapted to FFPE samples (H. Zhang *et al.*, 2022). The reasonable cost and technical advantages of these protocols foreshadow an increased usage of ATAC-Seq in cancer studies (Grandi *et al.*, 2022; Luo, Gribskov and Wang, 2022).

Most biological tissues are composed of multiple cell types. For instance, tumors are complex ecosystems including malignant and stromal cells as well as a large diversity of immune cells. This cellular heterogeneity impacts tumor progression as well as response to immunotherapy (Fridman *et al.*, 2012, 2017; de Visser and Joyce, 2023). Most existing ATAC-Seq data from tumors were performed on bulk samples, thereby mixing the signal from both cancer and non-malignant cells. Precisely quantifying the proportions of different cell types in such samples represents a promising way to explore the immune contexture and the composition of the tumor micro-environment (TME) across large cohorts. Carefully assessing cell-type heterogeneity is also important to handle confounding factors in genomic analyses in which samples with different cellular compositions are compared. Recently, single-cell ATAC-Seq (scATAC-Seq) has been developed to explore cellular heterogeneity with high resolution in complex biological systems (Cusanovich *et al.*, 2015; Lareau *et al.*, 2019; Satpathy *et al.*, 2019). However, the resulting data are sensitive to technical noise and such experiments require important resources, which so far limits the use of scATAC-Seq in contrast to bulk ATAC-Seq in the context of large cohorts.

In the past decade, computational deconvolution tools have been developed to predict the proportion of diverse cell types from bulk genomic data obtained from tumor samples (Becht *et al.*, 2016; Racle *et al.*, 2017; Avila Cobos *et al.*, 2018, 2020; Finotello *et al.*, 2019; Monaco *et al.*, 2019; Newman *et al.*, 2019; Sturm *et al.*, 2019; H. Li *et al.*, 2020). Many of these tools model bulk data as a mixture of reference profiles either coming from purified cell populations or inferred from single-cell genomic data for each cell type. The accuracy of cell-type proportion predictions relies on the quality of these reference profiles as well as on the specificity of cell-type markers (Avila Cobos *et al.*, 2018). A limitation of most deconvolution algorithms is that they do not predict the proportion of cell types that are not present in the reference profiles (hereafter referred to as ‘uncharacterized’ cells). In the context of tumor samples, these uncharacterized cell populations include malignant cells whose molecular profiles differ not only from one cancer type to another, but also from one patient to another even within the same tumor type (Corces *et al.*, 2018). To address this limitation, a few tools developed for bulk RNA-Seq data consider uncharacterized cells in their deconvolution framework by using cell-type specific markers not expressed in the uncharacterized cells (Gosink, Petrie and Tsinoremas, 2007; Clarke, Seol and Clarke, 2010; Racle *et al.*, 2017; Finotello *et al.*, 2019). These tools include EPIC (Estimating the Proportion of Immune and Cancer cells) which simultaneously quantifies immune, stromal, vascular as well as uncharacterized cells from bulk tumor samples (Racle *et al.*, 2017; Racle and Gfeller, 2020).

Most deconvolution algorithms have been developed for transcriptomic data (RNA-Seq data) (Gong and Szustakowski, 2013; Newman *et al.*, 2015, 2019; Racle *et al.*, 2017; Finotello *et al.*, 2019; Jimenez-Sanchez, Cast and Miller, 2019; Monaco *et al.*, 2019; T. Li *et al.*, 2020). More recently they have been proposed for other omics layers such as methylation (Aryee *et al.*, 2014; Teschendorff *et al.*, 2017, 2020; Chakravarthy *et al.*, 2018; Hicks and Irizarry, 2019; Rahmani *et al.*, 2019; Arneson, Yang and Wang, 2020; H. Zhang *et al.*, 2021; Salas *et al.*,

2022 [\[1\]](#)) proteomics (Feng *et al.*, 2023 [\[2\]](#)) or chromatin accessibility (H. Li *et al.*, 2020 [\[3\]](#)). For the latter, a specific framework called DeconPeaker (H. Li *et al.*, 2020 [\[3\]](#)) was developed to estimate cell-type proportions from bulk ATAC-Seq samples. Deconvolution tools developed initially for RNA-Seq data can also be applied on ATAC-Seq. For example, the popular deconvolution tool, CIBERSORT (Newman *et al.*, 2015 [\[4\]](#)), in combination with ATAC-Seq profiles, was used to deconvolve leukemic ATAC-Seq samples (Corces *et al.*, 2016 [\[5\]](#)).

Other methods have been proposed to decompose ATAC-Seq bulk profiles into subpopulation-specific profiles (Burdziak *et al.*, 2019 [\[6\]](#); Zeng *et al.*, 2019 [\[7\]](#)) or compartments (Peng *et al.*, 2019 [\[8\]](#)). However, these methods have more requisites: (i) the integration of the ATAC-Seq data with single-cell or bulk RNA-Seq (Burdziak *et al.*, 2019 [\[6\]](#); Zeng *et al.*, 2019 [\[7\]](#)) and HiChIP data (Zeng *et al.*, 2019 [\[8\]](#)) or, (ii) subsequent feature annotation to associate compartments with cell types or biological processes (Peng *et al.*, 2019 [\[8\]](#)).

The application of existing bulk ATAC-Seq data deconvolution tools to solid tumors faces some limitations. First, current computational frameworks do not quantify populations of uncharacterized cell types. Second, ATAC-Seq based markers (*i.e.*, chromatin accessible regions called peaks) and reference profiles generated so far have been derived in the context of hematopoietic cell mixtures (Corces *et al.*, 2016 [\[5\]](#); H. Li *et al.*, 2020 [\[3\]](#)). Markers and profiles for major populations of the TME (*e.g.*, stromal and vascular cells) are thus missing. While cell-type specific markers have been identified from scATAC-Seq data (K. Zhang *et al.*, 2021 [\[9\]](#)), not all TME-relevant cell types are covered (*e.g.*, lack of scATAC-Seq data from neutrophils due to extracellular traps formation). Many of these markers have also not been curated to fulfill the requirements of tools such as EPIC to quantify uncharacterized cells (*i.e.*, markers of a cell-type should not be accessible in other human tissues).

In this study, we collected ATAC-Seq data from pure cell types to identify cell-type specific marker peaks and to build reference profiles from most major non-malignant cell types typically observed in tumors. These data were integrated in the EPIC (Racle *et al.*, 2017 [\[10\]](#)) framework to perform bulk ATAC-Seq samples deconvolution (**Figure 1** [\[11\]](#)). Applied on peripheral blood mononuclear cells (PBMCs) and tumor samples, the EPIC-ATAC framework showed accurate predictions of the proportions of non-malignant and malignant cells.

Results

ATAC-Seq data from sorted cell populations reveal cell-type specific marker peaks and reference profiles

A key determinant for accurate predictions of cell-type proportions by most deconvolution tools is the availability of reliable cell-type specific markers and reference profiles. To identify robust chromatin accessibility marker peaks of cell types observed in the tumor microenvironment, we collected 564 samples of sorted cell populations from twelve studies including B cells (Corces *et al.*, 2016 [\[5\]](#); Calderon *et al.*, 2019 [\[12\]](#); P. Zhang *et al.*, 2022 [\[13\]](#)), CD4+ T cells (Corces *et al.*, 2016 [\[5\]](#); Mumbach *et al.*, 2017 [\[14\]](#); Liu *et al.*, 2020 [\[15\]](#); Giles *et al.*, 2022 [\[16\]](#); P. Zhang *et al.*, 2022 [\[13\]](#)), CD8+ T cells (Corces *et al.*, 2016 [\[5\]](#); Calderon *et al.*, 2019 [\[12\]](#); Liu *et al.*, 2020 [\[15\]](#); Giles *et al.*, 2022 [\[16\]](#); P. Zhang *et al.*, 2022 [\[13\]](#)), natural killer (NK) cells (Corces *et al.*, 2016 [\[5\]](#); Calderon *et al.*, 2019 [\[12\]](#)), dendritic cells (DCs) (mDCs and pDCs are grouped in this cell-type category) (Calderon *et al.*, 2019 [\[12\]](#); Leylek *et al.*, 2020 [\[17\]](#); Liu *et al.*, 2020 [\[15\]](#)), macrophages (Liu *et al.*, 2020 [\[15\]](#); P. Zhang *et al.*, 2022 [\[13\]](#)), monocytes (Corces *et al.*, 2016 [\[5\]](#); Calderon *et al.*, 2019 [\[12\]](#); Leylek *et al.*, 2020 [\[17\]](#); Trizzino *et al.*, 2021 [\[18\]](#); P. Zhang *et al.*, 2022 [\[13\]](#)), neutrophils (Perez *et al.*, 2020 [\[19\]](#); Ram-Mohan *et al.*, 2021 [\[20\]](#)), fibroblasts (Liu *et al.*, 2020 [\[15\]](#); Ge *et al.*, 2021 [\[21\]](#)) and endothelial cells (Liu *et al.*, 2020 [\[15\]](#); Xin *et al.*, 2020 [\[22\]](#)) (**Figure 1** [\[11\]](#) box 1, **Figure 2A** [\[11\]](#), Supplementary Table 1). To limit batch effects, the collected samples were homogeneously processed from read alignment to peak calling. For each cell type, we derived a set

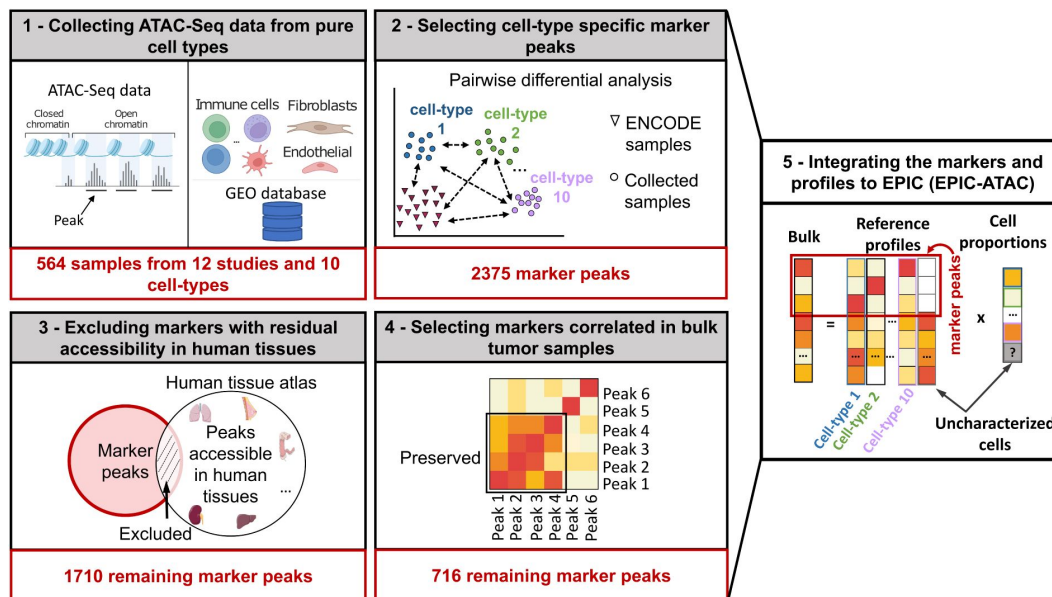


Figure 1

Graphical description of the identification of cell-type specific marker peaks and reference ATAC-Seq profiles included in the EPIC-ATAC framework.

1) 564 pure ATAC-Seq data of sorted cells were collected to build reference profiles for non-malignant cell types observed in the tumor microenvironment. 2) Cell-type specific marker peaks were identified using differential accessibility analysis. 3) Markers with previously observed chromatin accessibility in human healthy tissues were excluded. 4) For tumor bulk deconvolution, the set of remaining marker peaks was refined by selecting markers with correlated behavior in tumor bulk samples. 5) The cell-type specific marker peaks and reference profiles were integrated in the EPIC-ATAC framework to perform bulk ATAC-Seq deconvolution.

© 2024, BioRender Inc. Any parts of this image created with *BioRender* are not made available under the same license as the Reviewed Preprint, and are © 2024, BioRender Inc.

of stable peaks observed across samples and studies, *i.e.*, for each study, peaks detected in at least half of the samples were considered, and for each cell type, only peaks detected jointly in all studies were kept (see Materials and Methods, section 11).

These peaks were then used to perform pairwise differential analysis to identify marker peaks for each cell type (**Figure 1** [↗](#), box 2). To ensure that the cell-type specific marker peaks are not accessible in other human tissues, we included in the differential analysis ATAC-Seq samples from diverse human tissues from the ENCODE data (The ENCODE Project Consortium *et al.*, 2020; Rozowsky *et al.*, 2023 [↗](#)) (Supplementary Figure 1A). To select a sufficient number of peaks prior to peak filtering, the top 200 peaks recurrently differentially accessible across all cell-type pairs were selected as cell-type specific markers (see Materials and Methods, section 2). To filter out markers that could be accessible in other human cell types than those included in our reference profiles, we used the human atlas study (K. Zhang *et al.*, 2021 [↗](#)), which identified modules of open chromatin regions accessible in a comprehensive set of human tissues, and we excluded from our marker list the markers overlapping these modules (**Figure 1** [↗](#), box 3, see Materials and Methods section 2). The resulting marker peaks specific only to the immune cell types were considered for the deconvolution of PBMC samples (PBMC markers). For the deconvolution of tumor bulk samples, the lists of marker peaks specific to fibroblasts and endothelial cells were added to the PBMC markers. This extended set of markers was further refined based on the correlation patterns of the markers in tumor bulk samples from the diverse solid cancer types from The Cancer Genome Atlas (TCGA) (Corces *et al.*, 2018 [↗](#)), *i.e.*, markers exhibiting the highest correlation patterns in the tumor bulk samples were selected using the *findCorrelation* function from the caret R package (Kuhn, 2008 [↗](#)) (**Figure 1** [↗](#), box 4, see the Material and methods, section 2). The latter filtering ensures the relevance of the markers in the TME context since cell-type specific TME markers are expected to be correlated in tumor bulk ATAC-Seq measurements (Qiu *et al.*, 2021 [↗](#)). 716 markers of immune, fibroblasts and endothelial cell types remained after the last filtering (defined as TME markers). Considering the difference in cell types and the different filtering steps applied on the PBMC and TME markers, we recommend to use the TME markers and profiles to deconvolve bulk samples from tumor samples and the PBMC markers and profiles to deconvolve PBMC samples.

To assess the quality and reproducibility of these markers, we first performed principal component analysis (PCA) based on each set of marker peaks. Computing silhouette coefficients based on the cell-type classification and on the study of origin showed that samples clustered by cell type and not by study of origin (averaged silhouette coefficients above 0.45 for cell type and around 0 for study of origin). The representation of the samples based on the first axes of the PCA confirmed this observation (**Figure 2B** [↗](#) and Supplementary Figure 1B). These results indicate limited remaining batch effects after data processing and marker selection.

We then used the collected samples to generate chromatin accessibility profiles by computing the average of the normalized counts for each peak in each cell type as well as peak variability in each cell type (see Material and methods) since EPIC uses features variability for the estimation of the cell-type proportions (Racle *et al.*, 2017 [↗](#)). **Figure 2C** [↗](#) represents the average chromatin accessibility of each marker peak in each cell type of the reference dataset and highlights, as expected, the cell-type specificity of the selected markers (see also Supplementary Tables 2 and 3), which was confirmed in independent ATAC-Seq data from sorted cells and single-cell ATAC-Seq samples from blood and diverse human tissues (**Figure 2D** [↗](#) and **2E** [↗](#), see Materials and methods).

Annotations of the marker peaks highlight their biological relevance

To characterize the different marker peaks, we annotated them using ChIPseeker (Yu, Wang and He, 2015 [↗](#)). We observed that most of the markers are in distal and intergenic regions (**Figure 2F** [↗](#)), which is expected considering the large proportion of distal regions in the human genome and the fact that such regions have been previously described as highly cell-type specific (Corces *et al.*, 2016 [↗](#)). We also noticed that 7% of the PBMC and TME marker peaks are in promoter regions in contrast to 4% when considering matched genomic regions randomly selected in the set of peaks identified prior to the differential analysis (see Material and methods), which suggests enrichment in our marker peaks for important regulatory regions.

To assess the biological relevance of the marker peaks, we associated each marker peak to its nearest gene using ChIP-Enrich based on the “nearest transcription start site (TSS)” locus definition (Welch *et al.*, 2014 [↗](#)) (Supplementary Tables 4 and 5). Nearest genes reported as known marker genes in public databases of gene markers (*i.e.*, PanglaoDB (Franzén, Gan and Björkegren, 2019 [↗](#)) and CellMarker (Hu *et al.*, 2023 [↗](#))) are listed in **Table 1** [↗](#).

In each set of cell-type specific peaks, we observed an overrepresentation of chromatin binding proteins (CBPs) reported in the JASPAR2022 database (Castro-Mondragon *et al.*, 2022 [↗](#)) (using Signac (Stuart *et al.*, 2021 [↗](#)) and MonaLisa (Machlab *et al.*, 2022 [↗](#)) for assessing the overrepresentation) and the ReMap catalog (Hammal *et al.*, 2022 [↗](#)) (using RemapEnrich, see Material and Methods). Overrepresented CBPs also reported as known marker genes in the PanglaoDB and CellMarker databases are listed in **Table 1** [↗](#). Detailed peaks annotations are summarized in Supplementary Tables 4 and 5.

Based on the “nearest TSS” annotation, we tested, using ChIP-Enrich (Welch *et al.*, 2014 [↗](#)), whether each set of cell-type specific marker peaks was enriched for regions linked to specific biological pathways (GO pathways). **Figure 2G** [↗](#) highlights a subset of the enriched pathways that are consistent with prior knowledge on each cell type. Some of these pathways are known to be characteristic of immune responses to inflammatory or tumoral environments. The complete list of enriched pathways is listed in the Supplementary Tables 6 and 7. Overall, these analyses demonstrate that the proposed cell-type specific marker peaks capture some of the known biological properties associated to each cell type.

EPIC-ATAC integrates the marker peaks and profiles into EPIC to perform bulk ATAC-Seq data deconvolution

The cell-type specific marker peaks and profiles derived from the reference samples were integrated into the EPIC deconvolution tool (Racle *et al.*, 2017 [↗](#); Racle and Gfeller, 2020 [↗](#)). We will refer to this ATAC-Seq deconvolution framework as EPIC-ATAC. To ensure the compatibility of any input bulk ATAC-Seq dataset with the EPIC-ATAC marker peaks and reference profiles, we provide an option to lift over hg19 datasets to hg38 (using the liftOver R package) as the reference profiles are based on the hg38 reference genome. Subsequently, the features of the input bulk matrix are matched to our reference profiles’ features. To match both sets of features, we determine for each peak of the input bulk matrix the distance to the nearest peak in the reference profiles peaks. Overlapping regions are retained and the feature IDs are matched to their associated nearest peaks. If multiple features are matched to the same reference peak, the counts are summed. Before the estimation of the cell-type proportions, we transform the data following an approach similar to the transcripts per million (TPM) transformation which has been shown to be appropriate to estimate cell fractions from bulk mixtures in RNA-Seq data (Racle *et al.*, 2017 [↗](#); Sturm *et al.*, 2019 [↗](#)). We normalize the ATAC-Seq counts by dividing counts by the peak lengths as well as samples depth and rescaling counts so that the counts of each sample sum to 10^6 . In RNA-

Table 1

List of nearest genes and enriched CBPs reported in the PanglaoDB or CellMarker databases.

Cell type	Nearest genes	Enriched CPBs
Bcells	DHTKD1 LHPP WDFY4 ARID5B HHEX SIDT2 CD82 MS4A1 FCHSD2 USP8 RHCG ATF7IP2 CIITA GGA2 SNX29P2 C16orf74 CBFA2T3 CD79B BCL2 GNG7 CD22 FCER2 FCRL1 LY9 PTPRC LAPTM5 IGLL5 VPREB3 CENPM AFF3 SP100 INPP5D DTNB CD86 RFTN1 ST6GAL1 NGLY1 OSBPL10 TLR9 CD38 SMIM14 ARHGAP24 ADAM19 EBF1 BASP1 CD83 PLEKHG1 CCR6 CCND3 HDAC9 CDCA7L BLK MTSS1 LYN PLEKHF2 MOB3B PAX5	SPIB POU2F2 TCF4 EBF1 TCF3 NFKB1 STAT1 NFKB2 IKZF1 FOXO1 FOXP1 BCL6 POU2AF1 STAT3 BACH2 IKZF3 FLI1 TBX21 JUNB MITF NKX6-2 RBPJ
CD4_Tcells	IL2RA CD6 CD5 CD4 RORA PTPRC CTLA4 ICOS SLC9A9 FHIT TCF7 FYB1 ATXN1 CD40LG	TCF7 RUNX3 SOHLH2 IRF9 GATA3 TBX21 MAF STAT3 RORA BATF CREM
CD8_Tcells	MKI67 JAML MAML2 KLRD1 NELL2 LAG3 PPP1R13B PTPRC LYST CASP8 CD8A CD8B CD96 BTLA GZMA THEMIS ETV1	ETV1 FOXP3 TBX21 FOXP1 EOMES CREM IRF4 ZEB1 ARNT JUNB TCF7
NK	PRF1 ZBTB16 KLRD1 SPN CD226 SH2D1B CD247 IL2RB CXCR4 NMUR1 GNLY ZAP70 TXK	EOMES TBX21 NFIL3 FOS JUN
DCs	C12orf75 LYZ APP CD8A RIOX2 NFKB1 QDPR ABCG2 PRELID2 DST CD36 IDO2 PCMTD1	SPIB IRF8 MYB NR4A1 REL CUX2 FOXO1 ETV6 IRF5 BATF3 RUNX2
Neutrophils	TLE3 CA4 CYP4F3 CEACAM8 PGLYRP1 FPR1 CTSS ALPL PI3 MMP9 CXCR1 DRC1 ASPRV1 LTF MGAM SLC25A37	FOS
Monocytes	VENTX GLT1D1 CLEC4E CARS2 SLC24A4 C16orf74 FFAR2 STXBP2 NLRP3 CYRIA CMTM7 TGFBI DIAPH1 VCAN MCTP1 IFNGR1 STX11 CAPZA2 CD36 MTSS1 DENND3 ASAH1 TNFRSF10B BNIP3L NACC2 MAMDC2 FBP1	CEBPA CEBPD CEBPB CEBPE SPI1 VENTX JUND RXRA TCF7L2
Macrophages	CXCL12 PSAP P2RY6 SLC02B1 CMKLR1 MMP19 LGMN CLEC10A C5AR1 FPR3 LILRB4 RGL1 SIGLEC1 MMP9 CD80	STAT1 SPI1 FOSL2 FOS SPIC
Endothelial	FAM107B ROBO4 FLI1 ACVRL1 FLT1 DOCK9 ABCC1 S1PR1 ELOVL1 PLPP3 ASAP2 SNRK ECSCR ARAP3 LAMA4 BMP6 SERPINE1 LAMB1 DOCK4 NOS3	ETV2 ELF1 FLI1 ELK3 FOSB ETS1 ERG GATA2 ZEB1 ETS2 FOXC1 SOX18
Fibroblasts	LOX CAV1 COL15A1	FOSL2 FOSB FLI1 HIF1A PBX1

Seq based deconvolution, EPIC uses an estimation of the amount of mRNA in each reference cell type to derive cell proportions while correcting for cell-type-specific mRNA bias. For the ATAC-Seq based deconvolution these values were set to 1 to give similar weights to all cell-types quantifications. Indeed ATAC-Seq measures signal at the DNA level, hence the quantity of DNA within each reference cell type is similar.

EPIC-ATAC accurately estimates immune cell fractions in PBMC ATAC-Seq samples

To test the accuracy of EPIC-ATAC predictions, we first collected PBMCs from five healthy donors. For each donor, half of the cells was used to generate a bulk ATAC-Seq dataset and the other half was used to determine the cellular composition of each sample, *i.e.*, the proportions of monocytes, B cells, CD4+ T cells, CD8+ T cells, NK cells and dendritic cells, by multiparametric flow cytometry (**Figure 3A** [↗](#), see Materials and methods). We then applied EPIC-ATAC to the bulk ATAC-Seq data. The predicted cell fractions are consistent with the cell fractions obtained by flow cytometry (**Figure 3B** [↗](#), Pearson correlation coefficient of 0.78 and root mean squared error (RMSE) of 0.10). However, the accuracy of the estimations is variable depending on the samples. In particular, one sample (*Sample4*, star shape on **Figure 3B** [↗](#)) exhibits larger discrepancies between EPIC-ATAC predictions and the ground truth. Visualizing our marker peaks in the five PBMC samples (Supplementary Figure 2) showed that this sample might be an outlier considering that its cellular composition is similar to that of Samples 2 and 5 but this sample shows particularly high ATAC-Seq accessibility at the monocytes and dendritic marker peaks. This could explain why EPIC-ATAC overestimates the proportions of the two populations in this case.

As a second validation, we applied EPIC-ATAC to pseudo-bulk PBMC samples (referred to as the PBMC pseudobulk dataset), generated using three publicly available PBMC scATAC-Seq datasets (Granja *et al.*, 2019 [↗](#); Satpathy *et al.*, 2019 [↗](#); 10x Genomics, 2021 [↗](#)) (see Material and methods). A high correlation (0.91) between EPIC-ATAC predictions and true cell-type proportions and a low RMSE (0.05) were observed for this dataset (**Figure 3C** [↗](#)).

The accuracy of the predictions obtained with EPIC-ATAC was then compared with the accuracy of other deconvolution approaches which could be used with our reference profiles and marker peaks (**Figure 3D-E** [↗](#)). To this end, we considered both the DeconPeaker method (H. Li *et al.*, 2020 [↗](#)) originally developed for bulk ATAC-Seq as well as several algorithms developed for bulk RNA-Seq (CIBERSORTx (Newman *et al.*, 2019 [↗](#)), quanTiseq (Finotello *et al.*, 2019 [↗](#)), ABIS (Monaco *et al.*, 2019 [↗](#)), and MCPcounter (Becht *et al.*, 2016 [↗](#))). To enable meaningful comparison across the cell types considered in this work and use the method initially developed for bulk RNA-Seq deconvolution, the marker peaks and profiles derived in this work were used in each of these methods (See Material and methods, section 4). As in EPIC-ATAC, the features of the input bulk matrices were matched to our reference profiles' features. Having reference profiles/markers and an ATAC-Seq bulk query with matched features was the only requirement of the different deconvolution models to run on ATAC-Seq data with the default methods parameters. DeconPeaker and CIBERSORTx also include the option to define cell-type specific markers and profiles from a set of reference samples. We thus additionally fed our ATAC-Seq samples collection to both algorithms and used the resulting profiles and marker peaks to perform bulk ATAC-Seq deconvolution. The resulting predictions are referred to as DeconPeaker-Custom and CIBERSORTx-Custom.

Many tools displayed high correlation and low RMSE values, similar to those of EPIC-ATAC, and no single tool consistently outperformed the others (**Figure 3D-E** [↗](#), Supplementary Figure 3A-C). The fact that our marker peaks and reference profiles could be used with EPIC-ATAC and other existing tools demonstrates their broad applicability.

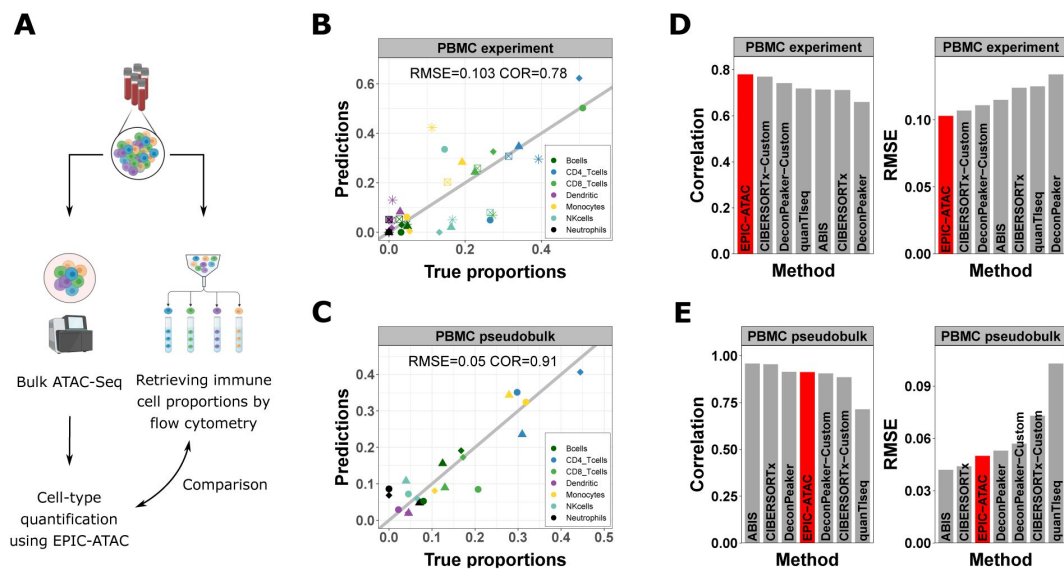


Figure 3

EPIC-ATAC accurately estimates immune cell fractions in PBMC ATAC-Seq samples.

A) Schematic description of the experiment designed to validate the ATAC-Seq deconvolution on PBMC samples. **B)** Comparison between cell-type proportions predicted by EPIC-ATAC and the true proportions in the PBMC bulk dataset. Symbols correspond to donors. **C)** Comparison between the proportions of cell-types predicted by EPIC-ATAC and the true proportions in the PBMC pseudobulk dataset. Symbols correspond to pseudobulks. **D)** Pearson correlation (left) and RMSE (right) values obtained by each deconvolution tool on the PBMC bulk dataset. The EPIC-ATAC results are highlighted in red. **E)** Pearson correlation (left) and RMSE (right) values obtained by each deconvolution tool on the PBMC pseudobulk dataset. © 2024, BioRender Inc. Any parts of this image created with *BioRender* are not made available under the same license as the Reviewed Preprint, and are © 2024, BioRender Inc.

EPIC-ATAC accurately predicts fractions of cancer and non-malignant cells in tumor samples

We evaluated the ability of the EPIC-ATAC framework to predict not only immune and stromal cells proportions but also the proportion of cells for which reference profiles are not available (*i.e.*, uncharacterized cells). For this purpose, we considered two previously published scATAC-Seq datasets containing basal cell carcinoma and gynecological cancer samples (Satpathy *et al.*, 2019 [↗](#); Regner *et al.*, 2021 [↗](#)) as well as the samples from the Human Tumor Atlas Network (HTAN) single-cell multiome dataset composed of samples from diverse cancer types (Terekhanova *et al.*, 2023 [↗](#)). For each dataset, we generated pseudobulks by averaging the chromatin accessibility signal across all cells of each sample (see Material and methods, section 3). Applying EPIC-ATAC to each dataset shows that this framework can simultaneously predict the proportions of both uncharacterized cells and immune, stromal and vascular cells (**Figure 4A** [↗](#)). In basal cell carcinoma and gynecological cancer samples, the proportion of uncharacterized cells can be seen as a proxy of the proportion of cancer cells. In the HTAN dataset, some of the samples also contain cell-types that are neither immune, fibroblasts or endothelial cells nor malignant cells. Hence, the uncharacterized cells in these samples group cancer and normal cells from the tumor site. EPIC-ATAC, in this context, was able to not only estimate the proportion of cell types included in the TME reference profiles but also the proportion of uncharacterized cells in most of the cancer types (**Figure 4A** [↗](#), Supplementary Figure 4).

As for the PBMC datasets, we compared EPIC-ATAC performances to other existing deconvolution tools (see Materials and methods, section 4). For each dataset, EPIC-ATAC led to the highest performances and was the only method to accurately predict the proportion of uncharacterized cells (**Figure 4B** [↗](#), Supplementary Figures 5, 6 and 7). Although quanTIseq also allows users to perform such predictions, the method resulted in lower correlation and higher RMSE values when comparing the estimated and true proportions of the uncharacterized cells (**Figure 4B** [↗](#), Supplementary Figures 5 and 7).

In the EPIC-ATAC and quanTIseq frameworks, predictions correspond to absolute cell-type fractions which include the proportion of uncharacterized cells, *i.e.*, proportions of all cells present in the bulk, while the estimations obtained from the other tools correspond to relative cell fractions, *i.e.*, proportions of cells present in the reference profiles (CIBERSORTx, DeconPeaker) or to scores with arbitrary units (ABIS, MCPcounter) without considering the presence of uncharacterized cells. For the latter group of tools, we fixed the uncharacterized cells estimations to 0. This approach provides a clear and significant advantage to EPIC-ATAC and to quanTIseq (**Figure 4B** [↗](#)). For this reason, we conducted a second benchmark excluding the predictions of uncharacterized cell fractions and rescaling both estimations and true proportions to sum to 1 (see Material and methods, section 4). EPIC-ATAC outperformed most of the other methods also when excluding the uncharacterized cells (**Figure 4C** [↗](#), Supplementary Figures 5, 6 and 7). Note that in terms of computational resources, all the tools require a limited amount of time and memory to run, *i.e.*, less than 20 seconds in average (Supplementary figure 8).

Accuracy of ATAC-Seq deconvolution is determined by the abundance and specificity of each cell type

To investigate the impact of cell type abundance on the accuracy of ATAC-Seq deconvolution, we evaluated EPIC-ATAC predictions in each major cell type separately in the different benchmarking datasets (**Figure 5A** [↗](#)). NK cells, endothelial cells, neutrophils or dendritic cells showed lower correlation values. These values can be explained by the fact that these cell types are low-abundant in our benchmarking datasets (**Figure 5A** [↗](#)). For the endothelial cells and dendritic cells, the RMSE values associated to these cell types remain low. This suggests that while the predictions of EPIC-ATAC might not be precise enough to compare these cell-type proportions

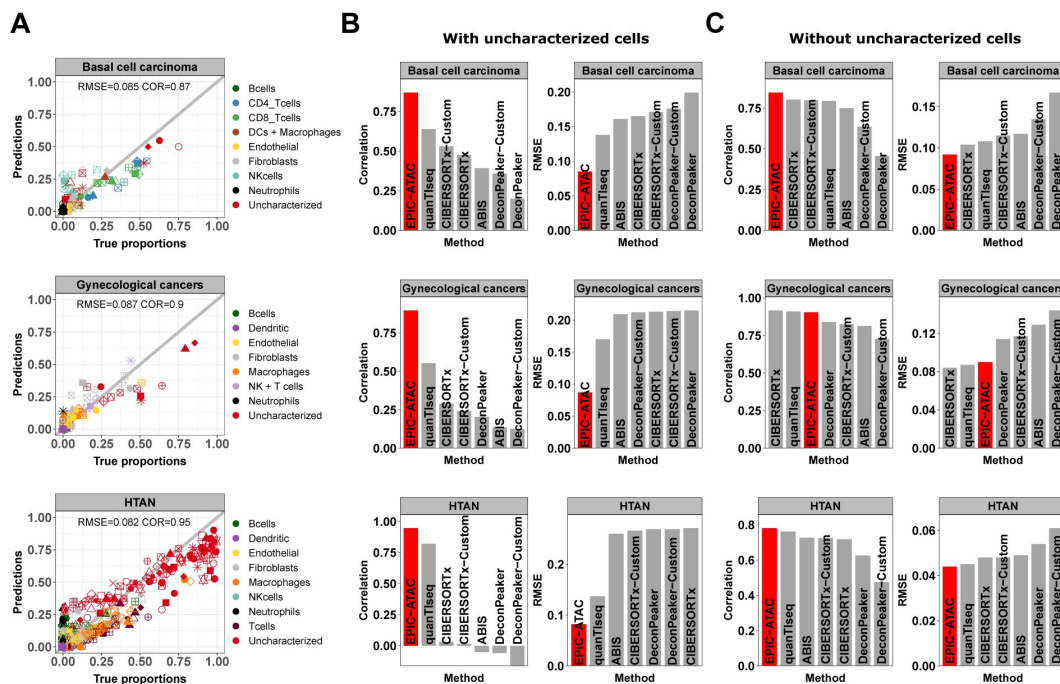


Figure 4

EPIC-ATAC accurately predicts fractions of cancer and non-malignant cells in tumor samples.

A) Comparison between cell-type proportions estimated by EPIC-ATAC and true proportions for the basal cell carcinoma (top), gynecological (middle) and HTAN (bottom) pseudobulk datasets. Symbols correspond to pseudobulks. **B)** Pearson's correlation and RMSE values obtained for the deconvolution tools included in the benchmark. EPIC-ATAC is highlighted in red. **C)** Same analyses as in panels B, with the uncharacterized cell population excluded for the evaluation of the prediction accuracy. The predicted and true proportions of the immune, stromal and vascular cell types were rescaled to sum to 1.

between different samples, the cell-type quantification within each sample is reliable. For the NK cells and the neutrophils, we observed more variability with higher RMSE values in some datasets which suggests that the markers and profiles for these cell types might be improved. Supplementary Figures 9 and 10 detail the performances of each tool when considering each cell type separately in the PBMC and the cancer datasets. As for EPIC-ATAC, the predictions from the other deconvolution tools are more reliable for the frequent cell types.

To explore the accuracy of ATAC-Seq deconvolution for more related cell types, we evaluated whether EPIC-ATAC could predict the proportions of T-cell subtypes. To this end, we considered naive and non-naive CD8+ as well as naive, helper/memory and regulatory CD4+ T cells (Tregs). We redefined our list of cell-type specific marker peaks and reference profiles including also these five T-cell subtypes (Supplementary Tables 8-9, Supplementary Figure 11A) and observed that the markers were conserved in external data (Supplementary Figure 11B). The annotations of the markers associated to the T-cell subtypes are available in Supplementary Tables 10-13.

We capitalized on the more detailed cell-type annotation of the PBMC datasets as well as the basal cell carcinoma dataset to evaluate the EPIC-ATAC predictions of cell-subtype fractions using these updated markers and profiles. Overall, the correlations observed between the predictions and true proportions of T cells decreased when considering T-cell subtypes rather than CD4+ and CD8+ cell types only (**Figure 5B** [↗](#)). In particular, low accuracies were obtained for helper/memory CD4+ and naive T-cell subtypes (**Figure 5C** [↗](#)). Similar results were obtained using other deconvolution tools (Supplementary Figure 12).

EPIC-ATAC accurately infers the immune contexture in a bulk ATAC-Seq breast cancer cohort

We applied EPIC-ATAC to a breast cancer cohort of 42 breast ATAC-Seq samples including samples from two breast cancer subtypes, *i.e.*, 35 oestrogen receptor (ER)-positive human epidermal growth factor receptor 2 (HER2)-negative (ER+/HER2-) breast cancer samples and 7 triple negative breast cancer (TNBC) samples (Kumegawa *et al.*, 2023 [↗](#)). No cell sorting was performed in parallel to the chromatin accessibility sequencing. For this reason, the authors used a set of cell-type-specific cis-regulatory elements (CREs) identified in scATAC-Seq data from similar breast cancer samples (Kumegawa *et al.*, 2022 [↗](#)) and estimated the amount of infiltration of each cell type by averaging the ATAC-Seq signal of each set of cell-type-specific CREs in their samples. We used EPIC-ATAC to estimate the proportions of different cell types of the TME. These predictions were then compared to the metric used by Kumegawa and colleagues in their study to infer levels of infiltration. A high correlation between the two metrics was observed for each cell type (Pearson's correlation coefficient from 0.5 for myeloid cells to 0.94 for T cells, **Figure 6A** [↗](#)).

We observed a higher proportion of T cells, B cells and NK cells in the TNBC samples in comparison to ER+/HER2- samples (**Figure 6B** [↗](#)). We then compared the cellular composition of ER+/HER2- subgroups identified in the original study (clusters CA-A, CA-B and CA-C). A higher infiltration of T and B cells was observed in cluster CA-C and higher proportions of endothelial cells and fibroblasts were observed in cluster CA-B (**Figure 6B** [↗](#)). These differences recapitulate those reported in the Kumegawa study except for the difference in myeloid infiltration observed in the original study between the different breast cancer subgroups (Kumegawa *et al.*, 2023 [↗](#)). However, when considering each myeloid cell type present in our reference profiles separately, a higher infiltration of macrophages was observed in the TNBC samples in comparison to the ER+/HER2- samples (Supplementary Figure 13). Also, we observed a difference in the levels of NK cells infiltration between TNBC and ER+/HER2- samples while no NK infiltration estimation was provided in the original paper for this cell type.

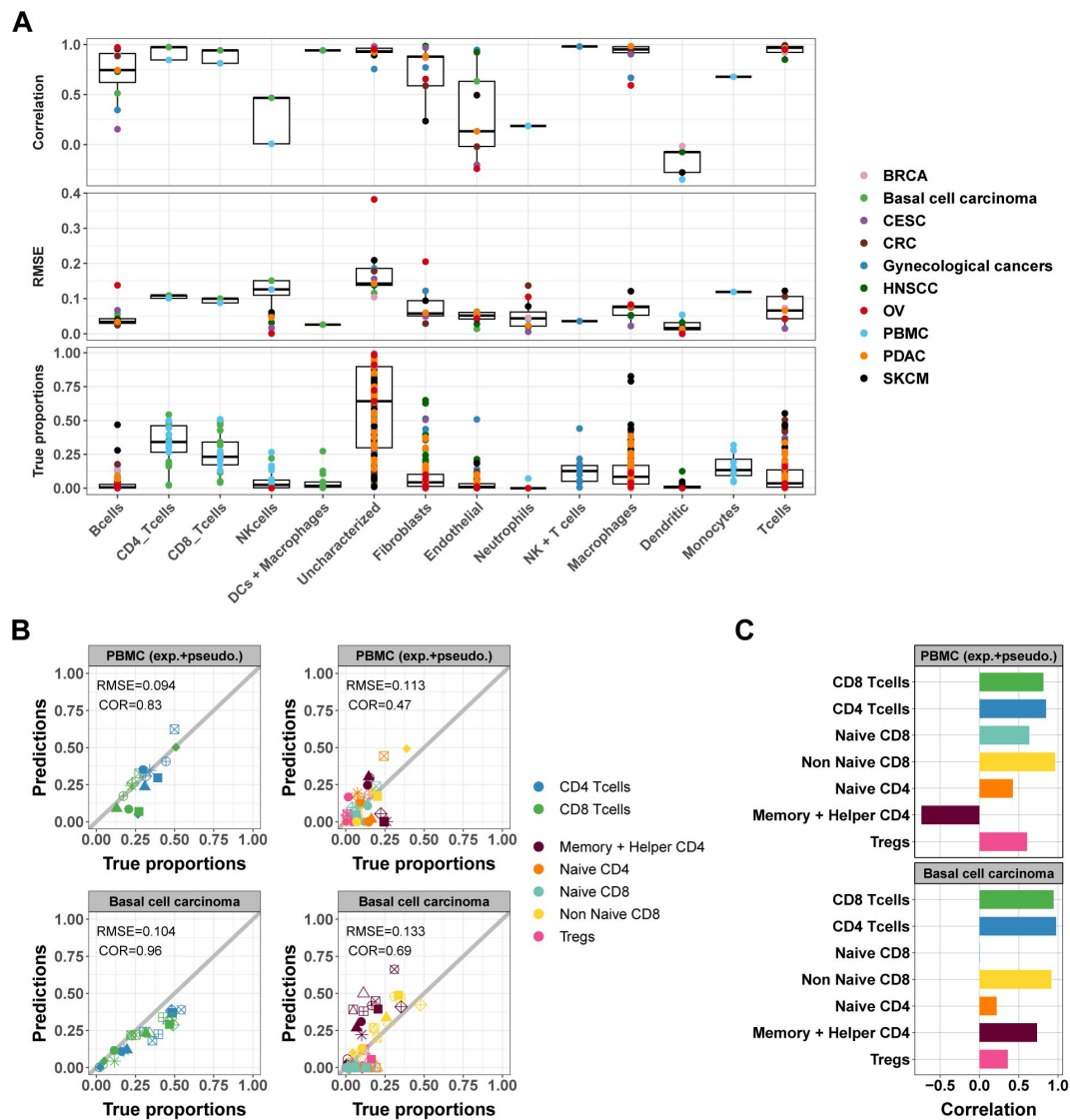


Figure 5

Accuracy of ATAC-Seq deconvolution is determined by the abundance and specificity of each cell type.

A) Correlations (top) and RMSE (middle) between EPIC-ATAC predictions and true cell-type proportions in each cell-type. True proportions are also shown for each cell type (bottom). Colors correspond to different datasets. **B**) Comparison of the proportions estimated by EPIC-ATAC and the true proportions for PBMC samples (PBMC experiment and PBMC pseudobulk samples combined) (top) and the basal cell carcinoma pseudobulks (bottom). Predictions of the proportions of CD4+ and CD8+ T-cells were obtained using the reference profiles based on the major cell types and subtypes predictions using the reference profiles including the T-cell subtypes. **C**) Pearson's correlation values obtained by EPIC-ATAC in each cell type.

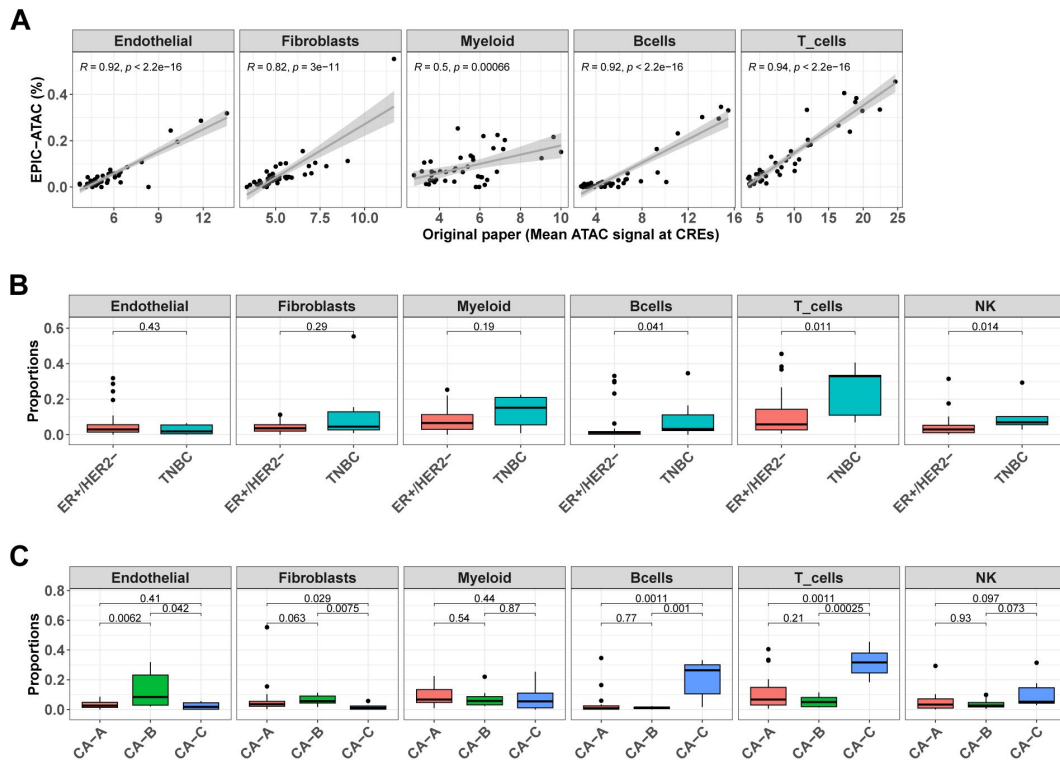


Figure 6

EPIC-ATAC accurately infers the immune contexture in a bulk ATAC-Seq breast cancer cohort.

A) Proportions of different cell types predicted by EPIC-ATAC in each sample as a function of the average ATAC signal at the cell-type specific CREs used by Kumegawa et al. (Kumegawa et al., 2023) to infer the level of cell-type infiltration in the tumor samples. **B)** Proportions of different cell types predicted by EPIC-ATAC in the samples stratified based on two breast cancer subtypes. **C)** Proportions of different cell types predicted by EPIC-ATAC in the samples stratified based on three ER+/HER2- subgroups. Wilcoxon test p-values are represented at the top of the boxplots.

EPIC-ATAC and EPIC RNA-seq based deconvolution have similar accuracy and can complement each other

We compared the accuracy of EPIC when applied on ATAC-Seq data and on RNA-Seq data. For this purpose, we first used single-cell multiomic data which provides for each cell both its chromatin accessibility profile and its gene expression profile. These data were retrieved from the 10X multiome PBMC dataset (10x Genomics, 2021 [↗](#)) and the HTAN dataset (Terekhanova *et al.*, 2023 [↗](#)) (see Material and methods section 3). We used EPIC-ATAC to perform ATAC-Seq based deconvolution on the chromatin accessibility levels of the peaks and the original EPIC tool to perform standard RNA-seq deconvolution on the gene expression levels. ATAC-Seq peaks can also be aggregated into gene activity (GA) variables as proxy for gene expression, based on peak distances to each gene. We applied the GA transformation to the pseudobulk data and performed GA-based RNA deconvolution using the original EPIC tool (See Material and methods section 5). **Figures 7A and B** [↗](#) show that EPIC-ATAC performs similarly to the EPIC RNA-seq based deconvolution and outperforms the GA-based RNA deconvolution. The lower performances of GA based RNA deconvolution could be explained by the fact that GA features, by construction, do not perfectly match the transcriptomic data.

We also applied EPIC-ATAC and the original version of EPIC (EPIC-RNA) on a bulk dataset composed of PBMC samples with matched bulk RNA-Seq, bulk ATAC-Seq and flow cytometry data (Morandini *et al.*, 2024 [↗](#)) (**Figure 7C** [↗](#)). We compared the predictions obtained using each modality to the flow cytometry cell-type quantifications. EPIC-ATAC predictions were better correlated with the flow cytometry measures for some cell types (*e.g.*, CD8+, CD4+ T cells, NK cells) while this trend was observed with the EPIC-RNA predictions in other cell types (B cells, neutrophils, monocytes) (**Figure 7C** [↗](#)). We then tested whether the predictions obtained from both modalities could be combined to improve the accuracy of each cell-type quantification. Averaging the predictions obtained from both modalities shows a moderate improvement (**Figure 7C** [↗](#)), suggesting that the two modalities can complement each other.

Discussion

Bulk chromatin accessibility profiling of biological tissues like tumors represents a reliable and affordable technology to map the activity of gene regulatory elements across multiple samples in different conditions. Here, we collected ATAC-Seq data from pure cell populations covering major immune and non-immune cell types found in the tumor micro-environment. This enabled us to identify reliable cell-type specific marker peaks and chromatin accessibility profiles for both PBMC and solid tumor sample deconvolution. We integrated these data in the EPIC deconvolution framework to accurately predict the fraction of both malignant and non-malignant cell types from bulk tumor ATAC-Seq samples.

In cases where specific cell types are expected in a sample but are not part of our list of reference profiles (*e.g.*, neuronal cells in brain tumors), custom marker peaks and reference profiles can be provided to EPIC-ATAC to perform cell-type deconvolution and we provide the code to generate such markers and profiles based on ATAC-Seq data from sorted cells, following the approach developed in this work (**Figure 1** [↗](#), see Code availability).

Solid tumors contain large and heterogeneous fractions of cancer cells for which it is challenging to build reference profiles. We show that the EPIC-ATAC framework, in contrast to other existing tools, allows users to accurately predict the proportion of cells not included in the reference profiles (**Figure 4** [↗](#) and Supplementary Figure 4). These uncharacterized cells can include cancer

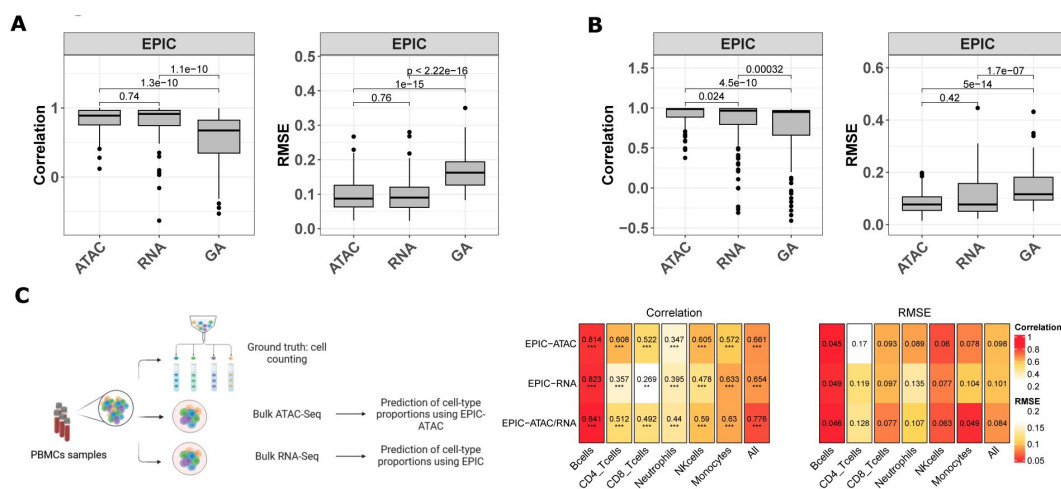


Figure 7

EPIC-ATAC and EPIC RNA-seq based deconvolution have similar accuracy and can complement each other.

A-B) Pearson's correlation (left) and RMSE (right) values comparing the proportions predicted by the ATAC-Seq deconvolution, the RNA-Seq deconvolution and the GA-based RNA deconvolution and true cell-type proportions in 100 pseudobulks simulated from the 10x multiome PBMC dataset (10x Genomics, 2021) (panel A) and in the pseudobulks generated from the HTAN cohorts (panel B). Dots correspond to outlier pseudobulks. **C)** Left panel: Schematic description of the dataset from Morandini et al. (Morandini et al., 2024) composed of matched bulk ATAC-Seq, RNA-Seq and flow cytometry data from PBMC samples. Right panel: Pearson's correlation (left) and RMSE values (right) obtained by EPIC-ATAC, EPIC-RNA and EPIC-ATAC/RNA (i.e., averaged predictions of EPIC-ATAC and EPIC-RNA) in each cell type separately or all cell types together (columns).

© 2024, BioRender Inc. Any parts of this image created with *BioRender* are not made available under the same license as the Reviewed Preprint, and are © 2024, BioRender Inc.

cells but also other non-malignant cells. Since the major cell types composing TMEs were included in our reference profiles, the proportion of uncharacterized cells approximates the proportion of the cancer cells in most cases.

For our benchmarking, we provided our reference profiles and markers to each tool that do not provide the option to automatically build new profiles from a set of reference samples. While this allowed us to run deconvolution tools previously developed on RNA-Seq data with their default parameters, we cannot exclude that hyper parametrization could improve the performance of these tools, and we anticipate that the cell-type specific ATAC-Seq profiles and markers curated in this work could help improving bulk ATAC-Seq deconvolution with different tools. Also, for RNA-Seq data deconvolution, some of the methods followed specific pre-processing steps, *e.g.*, the *quantiseq* framework removes a manually curated list of noisy genes as well as aberrant immune genes identified in the TCGA data, and ABIS uses immune-specific housekeeping genes. Additional filtering could be explored to improve the performance of deconvolution tools in the context of ATAC-Seq data.

The pseudobulk approach provides unique opportunities to design benchmarks with known cell-type proportions but also comes with some limitations. Indeed, pseudobulks are generated from single-cell data which are noisy and whose cell-type annotation is challenging in particular for closely related cell types. These limitations might lead to chromatin accessibility profiles that deviate from true bulk data and errors in the true cell-type proportions. The evaluation of our method on true bulk ATAC-Seq samples from PBMCs and breast cancer samples supported the accuracy of EPIC-ATAC to deconvolve bulk ATAC-Seq data. In the breast cancer application, the observation of similar immune compositions in TNBC and ER+/HER2- samples as the ones identified in the original paper ([Figure 6](#)) highlighted that the use of scATAC-Seq data, which are not always available for all cancer types, could be avoided for the estimation of different cell-type infiltration in bulk samples.

The evaluation of the EPIC-ATAC deconvolution resulted in an average absolute error of 8% across cell types. This number is consistent with previous observations in RNA-Seq data deconvolution ([Racle *et al.*, 2017](#)). Considering this uncertainty, the quantification of low frequency populations remains challenging ([Jin and Liu, 2021](#)). Comparing such estimations across samples should be performed with care due to the uncertainty of the predictions ([Figure 5A](#)).

Another limitation of cell-type deconvolution is often reached when closely related cell types are considered. In the reference-based methods used in this study, this limit was reached when considering T-cell subtypes in the reference profiles ([Figures 5B-C](#) and Supplementary Figure 12). We thus recommend to use the EPIC-ATAC framework with the markers and reference profiles based on the major cell-type populations. We additionally provide the marker peaks of the T-cell subtypes which could be used to build cell-type specific chromatin accessibility signatures or perform “peak set enrichment analysis” similarly to gene set enrichment analysis ([Subramanian *et al.*, 2005](#)). Such an application could be useful for the annotation of scATAC-Seq data, which often relies on matched RNA-Seq data and for which there is a lack of markers at the peak level ([Jiang *et al.*, 2023](#)).

The comparison of EPIC-ATAC applied on ATAC-Seq data with EPIC applied on RNA-Seq data has shown that both modalities led to similar performances and that they could complement each other. Another modality that has been frequently used in the context of bulk sample deconvolution is methylation. Methylation profiling techniques such as methylation arrays are cost effective ([Kaur *et al.*, 2023](#)) and DNA methylation signal is highly cell-type specific ([Kaur *et al.*, 2023](#); [Loyfer *et al.*, 2023](#)). Considering that methylation and chromatin accessibility measure different features of the epigenome, additional analyses comparing and/or complementing ATAC-seq based deconvolution with methylation-based deconvolution could be of interest as future datasets profiling both modalities in the same samples become available.

Another possible application of our marker peaks relies on their annotation (**Figure 2G** [↗](#), Supplementary Tables 4-5), which could be used to expand the list of genes and CBPs associated to each cell type or subtype. For example, the neutrophils marker peaks were enriched for motifs of TFs such as SPI1 (Supplementary Table 4), which was not listed in the neutrophil genes in the databases used for annotation but has been reported in previous studies as involved in neutrophils development (Watt *et al.*, 2021 [↗](#)). The annotations related to the set of major cell types and T-cell subtypes are provided in Supplementary Tables 4-5 and 10-11. Finally, the annotation of marker peaks highlighted pathways involved in immune responses to tumoral environments (**Figure 2G** [↗](#)). Examples of these pathways are the toll-like receptor signaling pathway involved in pathogen-associated and recognition of damage-associated molecular patterns in diverse cell types including B and T cells (Geng *et al.*, 2010 [↗](#); Javadi and Choi, 2020 [↗](#)), glucan metabolic processes which are known to be related to trained immunity which can lead to anti-tumor phenotype in neutrophils (Kalafati *et al.*, 2020 [↗](#)) or the Fc-receptor signaling observed in NK cells (Bonnema *et al.*, 1994 [↗](#); Sanseviero, 2019 [↗](#)). These observations suggest that our marker peaks contain regulatory regions not only specific to cell types but also adapted to the biological context of solid tumors.

Conclusion

In this work, we identified biologically relevant cell-type specific chromatin accessibility markers and profiles for most non-malignant cell types frequently observed in the tumor micro-environment. We capitalized on these markers and profiles to predict cell-type proportions from bulk PBMC and solid tumor ATAC-Seq data (<https://github.com/GfellerLab/EPIC-ATAC> [↗](#)). Evaluated on diverse tissues, EPIC-ATAC shows reliable predictions of immune, stromal, vascular and cancer cell proportions. With the expected increase of ATAC-Seq studies in cancer, the EPIC-ATAC framework will enable researchers to deconvolve bulk ATAC-Seq data from tumor samples to support the analysis of regulatory processes underlying tumor development, and correlate the TME composition with clinical variables.

Materials and methods

1- Generation of an ATAC-Seq reference dataset of non-malignant cell types frequently observed in the tumor microenvironment

Pre-processing of the sorted ATAC-Seq datasets

We collected pure ATAC-Seq samples from 12 studies. The data include samples from (i) ten major immune, stromal and vascular cell types (B (Corces *et al.*, 2016 [↗](#); Calderon *et al.*, 2019 [↗](#); P. Zhang *et al.*, 2022 [↗](#)), CD4+ (Corces *et al.*, 2016 [↗](#); Mumbach *et al.*, 2017 [↗](#); Liu *et al.*, 2020 [↗](#); Giles *et al.*, 2022 [↗](#); P. Zhang *et al.*, 2022 [↗](#)), CD8+ (Corces *et al.*, 2016 [↗](#); Calderon *et al.*, 2019 [↗](#); Liu *et al.*, 2020 [↗](#); Giles *et al.*, 2022 [↗](#); P. Zhang *et al.*, 2022 [↗](#)), natural killer (NK) (Corces *et al.*, 2016 [↗](#); Calderon *et al.*, 2019 [↗](#)), dendritic (DCs) cells (Calderon *et al.*, 2019 [↗](#); Leylek *et al.*, 2020 [↗](#); Liu *et al.*, 2020 [↗](#)), macrophages (Liu *et al.*, 2020 [↗](#); P. Zhang *et al.*, 2022 [↗](#)), monocytes (Corces *et al.*, 2016 [↗](#); Calderon *et al.*, 2019 [↗](#); Leylek *et al.*, 2020 [↗](#); Trizzino *et al.*, 2021 [↗](#); P. Zhang *et al.*, 2022 [↗](#)) and neutrophils (Perez *et al.*, 2020 [↗](#); Ram-Mohan *et al.*, 2021 [↗](#)) as well as fibroblasts (Liu *et al.*, 2020 [↗](#); Ge *et al.*, 2021 [↗](#)) and endothelial (Liu *et al.*, 2020 [↗](#); Xin *et al.*, 2020 [↗](#)) cells (See **Figure 2A** [↗](#)), and (ii) eight tissues from distinct organs (*i.e.* bladder, breast, colon, liver, lung, ovary, pancreas and thyroid) from the ENCODE data (The ENCODE Project Consortium *et al.*, 2020; Rozowsky *et al.*, 2023 [↗](#)). The list of the samples and their associated metadata (including cell types and accession number of the study of origin) is provided in Supplementary Table 1. To limit batch

effects, the samples were reprocessed homogeneously from the raw data (fastq files) processing to the peak calling. For that purpose, raw fastq files were collected from GEO using the SRA toolkit and the PEPATAC framework (Smith *et al.*, 2021 [DOI](#)) was used to process the raw fastq files based on the following tools: trimmomatic for adapter trimming, bowtie2 (with the PEPATAC default parameters) for reads pre-alignment on human repeats and human mitochondrial reference genome, bowtie2 (with the default PEPATAC parameters: `--very-sensitive -X 2000`) for alignment on the human genome (hg38), samtools (PEPATAC default parameters: `-q 10`) for duplicates removal and MACS2 (Zhang *et al.*, 2008 [DOI](#)) (PEPATAC default parameters: `--shift -75 --extsize 150 --nomodel --call-summits --nolambda --keep-dup all -p 0.01`) for peak calling in each sample. After alignment, reads mapping on chromosome M were excluded. TSS enrichment scores were computed for each sample and used to filter out samples with low quality (criteria of exclusion: TSS score < 5) (See Supplementary Table 1 containing the TSS score of each sample). 789 samples (including 564 from our ten reference cell-types) had a TSS score > 5.

Generation of a consensus set of peaks

Peak calling was performed in each sample individually. Peaks were then iteratively collapsed to generate a set of reproducible peaks. For each cell type, peaks collapse was performed adapting the iterative overlap peak merging approach proposed in the PEPATAC framework. A first peaks collapse was performed at the level of each study of origin, *i.e.*, if peaks identified in distinct samples overlapped (minimum overlap of 1bp between peaks), only the peak with the highest peak calling score was kept. Also, only peaks detected in at least half of the samples of each study were considered for the next step. If a study had only two samples, only peaks detected in both samples were considered. After this first selection, a second round of peaks collapse was performed at the cell-type level to limit batch effects in downstream analyses. For each cell type, only peaks detected in all the studies of origin were considered. The final list of peaks was then generated by merging each set of reproducible peaks. Peaks located on chromosome Y were excluded from the rest of the analyses. ATAC-Seq counts were retrieved for each sample and each peak using featureCounts (Liao, Smyth and Shi, 2014 [DOI](#)).

2- Identification of cell-type specific markers

Differential accessibility analysis

To identify cell-type specific markers, we split the samples collection in ten folds (created with the *create_folds* function from the R package splitTools (Mayer, 2023 [DOI](#))). For each fold, we performed pairwise differential accessibility analysis across the ten cell types considered in the reference samples as well as the ENCODE samples from diverse organs. The differential analysis was performed using limma (Ritchie *et al.*, 2015 [DOI](#)), version 3.56.2). Effective library sizes were computed using the method of trimmed mean of M-values (TMM) from the edgeR package in R (Robinson, McCarthy and Smyth, 2010 [DOI](#)), version 3.42.4). Due to differences of library size across all samples collected, we used voom from the limma package (Law *et al.*, 2014 [DOI](#)) to transform the data and model the mean-variance relationship. Finally, a linear model was fitted to the data to assess the differential accessibility of each peak across each pair of cell types. To identify our marker peaks, all peaks with log2 fold change higher than 0.2 were selected and ranked by their maximum adjusted *p*-value across all pairwise comparisons. The top 200 features (with the lowest maximum adjusted *p*-value) were considered as cell-type specific marker peaks. The marker peaks identified in at least three folds were considered in the final list of marker peaks.

Marker peaks filtering

Modules of open chromatin regions accessible in all (universal modules) or in specific human tissues have been identified in the study Zhang *et al.* (K. Zhang *et al.*, 2021 [DOI](#)). These regions were used to refine the set of marker peaks and exclude peaks with residual accessibility in other cell

types than those considered for deconvolution. More precisely, for immune, endothelial and fibroblasts specific peaks, we filtered out the peaks overlapping the universal modules as well as the tissue specific modules except the immune (modules 8 to 25), endothelial (modules 26 to 35) and stromal related modules (modules 41 to 49 and 139-150) respectively. As a second filtering step, we retained markers exhibiting the highest correlation patterns in tumor bulk samples from different solid cancer types, *i.e.*, The Cancer Genome Atlas (TCGA) samples (Corces *et al.*, 2018 [↗](#)). We used the Cancer Genomics Cloud (CGC) (Lau *et al.*, 2017 [↗](#)) to retrieve the ATAC-Seq counts for each marker peaks in each TCGA sample (using *featureCounts*). For each set of cell-type specific peaks, we identified the most correlated peaks using the *findCorrelation* function of the caret R package (Kuhn, 2008 [↗](#)), version 6.0-94 with a correlation cutoff value corresponding to the 90th percentile of pairwise Pearson correlation values.

Evaluation of the study of origin batch effect

To identify potential batch effect issues, we run principal component analysis (PCA) based on the cell-type specific peaks after normalizing ATAC-Seq counts using full quantile normalization (FQ-FQ) implemented in the EDASeq R package (Risso *et al.*, 2011 [↗](#)) to correct for depth and GC biases. These data were used to visualize the data in three dimensions using the first axes of the PCA based on the PBMC and TME markers (Figure 2B [↗](#) and Supplementary Figure 1B). We also used the ten first principal components to evaluate distances between samples and compute silhouette coefficients based on the cell type and study of origin classifications as well as to compute the ARI metric comparing the cell-type annotation and the clustering obtained using model based clustering.

Building the reference profiles

It has been previously demonstrated in the context of RNA-Seq based deconvolution approaches (Racle *et al.*, 2017 [↗](#); Sturm *et al.*, 2019 [↗](#)) that the transcripts per million (TPM) transformation is appropriate to estimate cell fractions from bulk mixtures. We thus normalized the ATAC-Seq counts of the reference samples using a TPM-like transformation, *i.e.*, dividing counts by peak length, correcting samples counts for depth and rescaling counts so that the counts of each sample sum to 10^6 . We then computed for each peak the median of the TPM-like counts across all samples from each cell type to build the reference profiles of the ten cell types considered in the EPIC-ATAC framework (Figure 2C [↗](#)). In the EPIC algorithm, weights reflecting the variability of each feature of the reference profile can be considered in the constrained least square optimization. We thus also computed the inter-quartile range of the TPM-like counts for each feature in each cell type. Two ATAC-Seq reference profiles are available in the EPIC-ATAC framework: (i) a reference profile containing profiles for B cells, CD4⁺ T cells, CD8⁺ T cells, NK, monocytes, dendritic cells and neutrophils to deconvolve PBMC samples, and (ii) a reference profile containing profiles for B cells, CD4⁺ T cells, CD8⁺ T cells, NK, dendritic cells, macrophages, neutrophils, fibroblasts and endothelial cells to deconvolve tumor samples. The reference profiles are available in the EPICATAC R package and the reference profiles restricted to our cell-type specific marker peaks are available in the Supplementary Tables 2 and 3.

Assessing the reproducibility of the marker peaks signal in independent samples

We evaluated the chromatin accessibility level of the marker peaks in samples that were not included in the peak calling step. Firstly, we considered samples from two independent studies (Ucar *et al.*, 2017 [↗](#); Carvalho *et al.*, 2021 [↗](#)) providing pure ATAC-Seq data for five immune cell types (*i.e.*, B, CD4⁺ T cells, CD8⁺ T cells, Monocytes, Macrophages) (Figure 2D [↗](#)). To consider the other cell types, samples that were excluded from the reference dataset due to a low TSS enrichment score were also considered in this validation dataset (Supplementary Table 1). Secondly, we collected the data from a single-cell atlas chromatin accessibility from human tissues and considered the cell types included in our reference data (K. Zhang *et al.*, 2021 [↗](#)) (Figure

2E [↗](#)). We used the cell-type annotations provided in the original study (GEO accession number: GSE184462). The Signac R package ((Stuart *et al.*, 2021 [↗](#)), 1.9.0) was used to extract fragments counts for each cell and each marker peak and the ATAC-Seq signal of each marker peak was averaged across all cells of each cell type.

Annotation of the marker peaks

The cell-type specific markers were annotated using ChIPseeker R package (Yu, Wang and He, 2015 [↗](#)), version 1.34.1) and the annotation from *TxDb.Hsapiens.UCSC.hg38.knownGene* in R to identify the regions in which the marker peaks are (*i.e.*, promoter, intronic regions, etc.) and ChIP-Enrich to associate each peak to the nearest gene TSS (Welch *et al.*, 2014 [↗](#)). The nearest genes identified were then compared to cell-type marker genes listed in the PanglaoDB (Franzén, Gan and Björkegren, 2019 [↗](#)) and CellMarker databases (Hu *et al.*, 2023 [↗](#)). PanglaoDB provides an online interface to explore a large collection on single-cell RNA-Seq data as well as a community-curated list of cell-type marker genes. CellMarker is a database providing a large set of curated cell-type markers for more than 400 cell types in human tissues retrieved from a large collection of single-cell studies and flow cytometry, immunostaining or experimental studies. ChIP-Enrich was also used to perform gene set enrichment and identify for each set of cell-type specific peaks potential biological pathways regulated by the marker peaks. The enrichment analysis was performed using the *chipenrich* function (*genesets* = “GOBP”, *locusdef* = “nearest_tss”) from the *chipenrich* R package (v2.22.0).

Chromatin accessibility peaks can also be annotated for chromatin binding proteins (CBPs) such as transcription factors (TFs), whose potential binding in the peak region is reported in databases. In our study we chose the JASPAR2022 (Castro-Mondragon *et al.*, 2022 [↗](#)) database and the ReMap database (Hammal *et al.*, 2022 [↗](#)).

Using the JASPAR2022 database, we assessed, for each cell type, whether the cell-type specific marker peaks were enriched in specific TFs motifs using two TFs enrichment analysis frameworks: Signac (Stuart *et al.*, 2021 [↗](#)) and MonaLisa (Machlab *et al.*, 2022 [↗](#)). For the MonaLisa analysis, the cell-types specific markers peaks were categorized in bins of sequences, one bin per cell type (use of the *calcBinnedMotifEnrR* function). To test for an enrichment of motifs, the sequences of each bin were compared to a set of background peaks with similar average size and GC composition obtained by randomly sampling regions in all the peaks identified from the reference dataset. The enrichment test was based on a binomial test. For the Signac analysis, we used the *FindMotif* function to identify over-represented TF motifs in each set of cell-type specific marker peaks (query). This function used a hypergeometric test to compare the number of query peaks containing the motif with the total number of peaks containing the motif in the background regions (matched to have similar GC content, region length and dinucleotide frequencies as the query regions), corresponding in our case to the peaks called in the reference dataset.

The ReMap database associates chromatin binding proteins (CBPs), including TFs, transcriptional coactivators and chromatin-remodeling factors, to their DNA binding regions based on DNA-binding experiments such as chromatin immunoprecipitation followed by sequencing (ChIP-seq). For each association of a CBP to its binding region, the cell type in which the binding has been observed is reported in the ReMap database (biotype). We used the ReMapEnrich R package (version 0.99) to test if the cell-type specific marker peaks are significantly enriched in CBPs-binding regions listed in the ReMap 2022 catalog. We considered the non-redundant peaks catalog from ReMmap 2022, containing non-redundant binding regions for each CBP in each biotype. Similarly to the previously mentioned enrichment methods, we chose the consensus peaks called in the reference samples as universe for the enrichment test. Note that, for each cell type, an enrichment was retained only if the biotype in which the CBP-regions were identified matched the correct cell-type.

3- Datasets used for the evaluation of ATAC-Seq deconvolution

PBMCs ATAC-Seq data from healthy donors

Peripheral blood mononuclear cell (PBMC) isolation

Venous blood from five healthy donors was collected at the local blood transfusion center of Geneva in Switzerland, under the approval of the Geneva University Hospital's Institute Review Board, upon written informed consent and in accordance with the Declaration of Helsinki. PBMCs were freshly isolated by Lymphoprep (Promega) centrifugation (1800 rpm, 20 minutes, without break, room temperature). Red blood cell lysis was performed using red blood lysis buffer (Qiagen) and platelets were removed by centrifugation (1000 rpm, 10 minutes without break, room temperature). Cells were counted and immediately used.

Flow cytometry

Immune cell populations were identified using multiparameter flow cytometry and the following antibodies: FITC anti-human CD45RA (HI100, Biolegend), PerCP-Cyanine5.5 anti-human CD19 (H1B19, Biolegend), PE anti-human CD3 (SK7, Biolegend), PE-Dazzle anti-human CD14 (M0P9, BD Biosciences), PE-Cyanine7 anti-human CD56 (HCD56, Biolegend), APC anti-human CD4 (RPA-T4, Biolegend), APC-Cyanine7 anti-human CCR7 (G043H7, Biolegend), Brilliant Violet 421 anti-human CD8 (RPA-T8, Biolegend), Brilliant Violet 510 anti-human CD25 (BC96, Biolegend), Brilliant Violet 711 anti-human CD16 (3G8, Biolegend), Brilliant Violet 786 anti-human CD127 (A019D5, Biolegend), Ultra-Brilliant Violet anti-human CD45 (HI30, BD Biosciences), FITC anti-human Ccl9a (8F9, Miltenyi), PE anti-human XCR1 (S15046E, Biolegend), PE-Dazzle anti-human BDCA-2 (201A, Biolegend), APC anti-human BDCA-3 (AD5-14H12, Miltenyi), Brilliant Violet 421 anti-human CD3 (UCHT1, Biolegend), Brilliant Violet 421 anti-human CD14 (M5E2, BD Pharmingen), Brilliant Violet 421 anti-human CD19 (SJ25C1, Biolegend), Brilliant Violet 510 anti-human BDCA-1 (L161, Biolegend), Brilliant Violet 650 anti-human CD11c (3.9, Biolegend), Brilliant Violet 711 anti-human CD11c (N418, Biolegend) and Brilliant Violet 711 anti-human HLA-DR (L243, Biolegend). Dead cells were excluded using the Zombie UV™ Fixable Viability Kit (Biolegend). Intracellular staining was performed after fixation and permeabilization of the cells with the FoxP3 Transcription Factor Staining Buffer Set (00-5523-00, Invitrogen) using Alexa 700 anti-human FoxP3 antibody (259D/C7, BD Biosciences). Data were acquired on LSRFortessa flow cytometer and analysed using FlowJo software (v10.7.1).

Cell preparation for ATAC-Sequencing

50000 CD45+ cells were sorted from total PBMCs using anti-human Ultra-Brilliant Violet (BUV395) CD45 (HI30, BD Biosciences) with a FACSARIA II (Becton Dickinson) and were collected in PBS with 10% Foetal Bovine Serum (FBS). Cell pellets were resuspended in cold lysis buffer (10mM Tris-Cl pH 7.4, 10mM NaCl, 3mM MgCl₂, 0.1% NP40 and water) and immediately centrifuged at 600g for 30min at 4°C. Transposition reaction was performed using the Illumina Tagment DNA Enzyme and Buffer kit (20034210, Illumina) and transposed DNA was eluted using the MinElute PCR Purification Kit (Qiagen). Libraries were generated by PCR amplification using indexing primers and NEBNext High-Fidelity Master Mix (New England BioLabs) and were purified using AMPure XP beads (A63880, Beckman Coulter). Libraries were quantified by a fluorometric method (Qubit, Life Technologies) and their quality assessed on a Fragment Analyzer (Agilent Technologies). Sequencing was performed as a paired end 50 cycles run on an Illumina NovaSeq 6000 (v1.5 reagents) at the Genomic Technologies Facility (GTF) in Lausanne, Switzerland. Raw sequencing data were demultiplexed using the bcl2fastq2 Conversion Software (version 2.20, Illumina).

Data processing

The same steps as for the processing of the reference ATAC-Seq samples were followed. (See Pre-processing of the ATAC-Seq datasets).

ATAC-Seq pseudobulk data from PBMCs and cancer samples

To evaluate the accuracy of our ATAC-Seq deconvolution framework, we generated pseudo-bulk datasets from 6 single-cell datasets:

- **PBMC pseudobulk dataset:** combination of three single-cell datasets for PBMCs.
 - Dataset 1 corresponds to a scATAC-Seq dataset obtained from Satpathy *et al.* (Satpathy *et al.*, 2019 [link](#)) (GEO accession number: GSE129785). This dataset contains FACS-sorted populations of PBMCs. Since the cells of some cell types came from a unique donor, all the cells of this dataset were aggregated to form one pseudobulk. Ground truth cell fractions were obtained by dividing the number of cells in each cell type by the total number of cells.
 - Dataset 2 (included in the PBMC pseudobulk dataset) was retrieved from Granja *et al.* (Granja *et al.*, 2019 [link](#)) (GEO accession number GSE139369). B cells, monocytes, dendritic, CD8+, CD4+ T, NK cells, neutrophils from healthy donors were considered. The neutrophil cells came from a single donor. As for dataset 1, we thus aggregated all the cells to generate one pseudobulk. Ground truth cell fractions were obtained by dividing the number of cells in each cell type by the total number of cells.
 - Dataset 3 (included in the PBMC pseudobulk dataset) corresponds to the 10X multiome dataset of PBMC cells (10x Genomics, 2021 [link](#)). Since these data come from one donor, one pseudobulk sample was generated for this dataset. The pseudobulk was generated by averaging the ATAC-Seq signal from all cells from the following cell types: B cells, CD4+ T cells, CD8+ T cells, NK cells, Dendritic cells and monocytes.
- **Basal cell carcinoma dataset:** obtained from the study of Satpathy *et al.* (Satpathy *et al.*, 2019 [link](#)).

This dataset is a scATAC-Seq dataset composed of 13 basal cell carcinoma samples composed of immune (B cells, plasma cells, CD4+ T cells, CD8+ T cells, NK cells, myeloid cells), stromal (endothelial and fibroblasts) and cancer cells. Plasma cells and cancer cells were both considered as uncharacterized cells (*i.e.*, cell types not included in the reference profiles). Cell annotations were retrieved from the original study.

- **Gynecological cancer dataset:** obtained from the study of Regner *et al.* (Regner *et al.*, 2021 [link](#)) (GEO accession number GSE173682). In this study, the authors performed scATAC-Seq on 11 gynecological cancer samples from two tumor sites (*i.e.*, endometrium and ovary) and composed of immune (B cells, NK and T cells grouped under the same cell-type annotation, macrophages, mast cells), stromal (fibroblast, endothelial, smooth muscle) and cancer cells. Mast cells, smooth muscle and cancer cells were considered as uncharacterized cells. Cell annotations were retrieved from the original study.
- **HTAN:** obtained from the study of Terekhanova *et al.* (Terekhanova *et al.*, 2023 [link](#)) from the HTAN DCC Portal. In this study the authors generated an atlas of single-cell multiomic data (RNA-Seq and ATAC-Seq profiling for the same cell) for diverse cancer types. We considered the annotation provided by the HTAN consortium. This cell annotation was composed of immune cells (B cells, T cells, dendritic cells, macrophages), fibroblasts, endothelial cells, cancer cells and normal cells. The two latter cell groups were considered as part of the uncharacterized cell population. For Basal cell carcinoma, Gynecological cancer and the

HTAN datasets, one pseudobulk per sample was generated and ground truth cell fractions were obtained for each sample by dividing the number of cells in each cell type by the total number of cells in the sample.

For each dataset, raw fragments files were downloaded from the respective GEO accession numbers and data were preprocessed using ArchR ((Granja *et al.*, 2021 [link](#)), ArchR R package 1.0.2). Cells with TSS score below four were removed. Doublets removal was performed using the *doubletsRemoval* function from ArchR. To match as much as possible real bulk ATAC-seq data processing, peak calling was not performed on each cell type or cell cluster as usually done in scATAC-Seq studies but using all cells for each dataset from the PBMC pseudobulk data or grouping cells by sample for the Basal cell carcinoma, Gynecological cancer and HTAN datasets. Peak calling was performed using MACS2 within the ArchR framework. Fragments counts were extracted using ArchR for each peak called to generate single-cell peak counts matrices. Note that for the HTAN dataset, in some cancer types, less than 15 peaks were in common with our reference profiles peaks. These cancer types were thus not considered in the analysis (GBM, UCEC, CEAD). These matrices were normalized using a TPM-like transformation, *i.e.*, dividing counts by peak length and correcting samples counts for depth. Finally, for each peak, the average of the normalized counts was computed across all the cells for each dataset from the PBMC pseudobulk data and across all the cells of each sample for the Basal cell carcinoma, Gynecological cancer and HTAN datasets. Averaged data were then rescaled so that the sum of counts of each sample sum to 10^6 .

Bulk ATAC-Seq data from a breast cancer cohort

Bulk ATAC-Seq samples from a breast cancer cohort was obtained from Kumegawa *et al.* (Kumegawa *et al.*, 2023 [link](#)). These data include 42 breast cancer samples which can be classified based on two features: (i) the breast cancer subtype ER+/HER2- or triple negative, and (ii) the molecular classification provided by the original study (CA-A, CA-B and CA-C). The ATAC-Seq raw counts and the samples metadata were retrieved from figshare (Kumegawa, 2023 [link](#)). As for the previously mentioned datasets, raw counts were normalized using the TPM-like transformation prior to bulk deconvolution.

4- Benchmarking of the EPIC-ATAC framework against other existing deconvolution tools

The performances of the EPIC-ATAC framework were benchmarked against the following deconvolution tools:

- *quantIseq* (Finotello *et al.*, 2019 [link](#)) is a deconvolution tool using constrained least square regression to deconvolve RNA-Seq bulk samples. No reference profiles are available in this framework to perform ATAC-Seq deconvolution and *quantIseq* does not provide the option to automatically build reference profiles from pure bulk samples. *quantIseq* was thus run using the reference profiles derived in this work for the EPIC-ATAC framework and the *quantIseq* function from the *quantiseqr* R package. Note that the correction for cell-type-specific mRNA content bias was disabled for ATAC-Seq data deconvolution (parameters used in *quantiseqr*: scaling set to 1 for all cell types and method set to “lsei”).
- *DeconPeaker* (H. Li *et al.*, 2020 [link](#)) relies on SIMPLS, a variant of partial least square regression to perform bulk RNA-Seq and bulk ATAC-Seq deconvolution. ATAC-Seq reference profiles are available in this deconvolution framework however not all cell types considered in the EPIC- ATAC framework are included in the *DeconPeaker* reference profiles. This tool was thus run using different reference profiles: (i) the reference profiles derived in this work for the EPIC- ATAC framework (corresponds to “*DeconPeaker*” or “*DeconPeaker_ourmarkers*” in our analyses), and (ii) reference profiles automatically generated by *DeconPeaker* from the sorted reference samples collected in this work (corresponds to “*DeconPeaker_Custom*” in our analyses). The results of *DeconPeaker*

obtained using its original markers and profiles are also provided for the cell types in common with the cell types considered in this work in Supplementary Figures 9 and 10. Deconvolution was run using the deconvolution module *deconPeaker* (using *findctsp*s with the following parameter: `--lib-strategy=ATAC-Seq`). *DeconPeaker* outputs cell-type proportions relative to the total amount of cells from the reference cell types.

- *CIBERSORTx* (Newman *et al.*, 2019 [↗](#)) is a deconvolution algorithm based on linear support vector regression. *CIBERSORTx* does not provide ATAC-Seq reference profiles, however it is possible to automatically generate new profiles from a set of pure bulk samples. This tool was thus run using different reference profiles: i) the reference profiles derived in this work for the EPIC- ATAC framework (corresponds to “*CIBERSORTx*” or “*CIBERSORTx_ourMarkers*” in our analyses), and ii) reference profiles automatically generated by *CIBERSORTx* from the sorted reference samples collected in this work (corresponds to “*CIBERSORTx_Custom*” in our analyses). To run *CIBERSORTx*, we used the docker container provided by the authors of *CIBERSORTx* on their website. The algorithm was run using the default options (i.e `--absolute FALSE`, `--rmbatchBmode FALSE` and `--rmbatchSmode FALSE`), which results in cell-type proportions relative to the total amount of cells from the reference cell types.
- *ABIS* (Monaco *et al.*, 2019 [↗](#)) uses robust linear modeling to estimate cell-type proportions in bulk RNA-Seq samples. No ATAC-Seq reference profiles are available in the deconvolution framework. *ABIS* was run using the EPIC-ATAC reference profiles by using the *rlm* function from the MASS R package (as performed in the *deconvolute_abis* function from the *immunedecov* R package (Sturm *et al.*, 2019 [↗](#))) to quantify each cell type from the reference profiles. The cell-types quantifications returned by this approach are in arbitrary units. To compare the estimations and the true cell proportions, we scaled the estimations of each sample between 0 and 1 to obtained relative proportions.
- *MCPcounter* (Becht *et al.*, 2016 [↗](#)): *MCPcounter* returns scores instead of cell type proportions. The scores were obtained using the *appendSignatures* function from the *MCPcounter* R package by providing the list of marker peaks specific to each cell type. The cell-type scores are not comparable between cell type, *MCPcounter* was thus included only in the evaluation of the performances in each cell type separately.

For all the tools, TPM-like data were used as input bulk samples for the deconvolution.

Since *CIBERSORTx*, *ABIS* and *DeconPeaker* do not predict proportions of uncharacterized cells, we performed two benchmarking analyses: (i) including all cell types and (ii) excluding the cell types that are absent from the reference profiles (uncharacterized cells) and rescaling the estimated and true proportions of the immune cells, endothelial cells and fibroblasts so that their sum equals 1.

Benchmarking of the running time of each tool

We compared the running time of each tool on the different datasets considered in our benchmarking and used the bash command *time* to retrieve the CPU time of each run. The measured running time is composed of the following steps: i) matching the bulk features to the reference profiles’ features and ii) running the deconvolution algorithm.

5- Comparing deconvolution based on RNA-Seq, gene activity or peaks features

100 pseudobulks were generated from the 10X PBMC multiome dataset (10x Genomics, 2021 [↗](#)) based on 3000 cells for each pseudobulk. Cell fractions were defined using the *rdirichlet* function from the *gtools* R package. We also considered pseudobulks built from the HTAN datasets by averaging the signal in all cells belonging to a sample. Three sets of features were extracted from the data, i.e., gene expression features extracted from the RNA-Seq layer, ATAC-Seq peaks and gene activity derived from the ATAC-Seq layer. The same cells was considered for each modality.

Gene activity features were extracted from the single-cell data using ArchR (1.0.2), which considers distal elements and adjusts for large differences in gene size in the gene activity score calculation. Gene activity pseudobulks were built by averaging the gene activity scores across all cells belonging to the pseudobulk. For ATAC-Seq pseudobulk, peaks called using ArchR on all cells from the single-cell ATAC data were considered (see the method section “ATAC-Seq pseudobulk data from PBMCs and cancer samples”) and counts were averaged across all cells of each pseudobulk. For RNA-Seq pseudobulks, counts were also averaged across all cells of each pseudobulk. All aggregated data were depth normalized across each features to 10^6 . Cell-type deconvolution was performed on each pseudobulk using EPIC-ATAC on the peak matrix using our ATAC-Seq marker peaks and reference profiles. The RNA-Seq and gene activity pseudobulks were deconvolved with EPIC.

Bulk ATAC-Seq samples from a PBMC cohort were also obtained from Morandini *et al.* (Morandini *et al.*, 2024 [↗](#)). These data include more than 100 samples with matched bulk RNA-Seq, bulk ATAC-Seq as well as matched flow cytometry data. The ATAC-Seq and RNA-Seq raw counts were obtained from GEO using the following GEO accession numbers: GSE193140 and GSE193141. As for the previously mentioned datasets, raw counts were normalized using the TPM-like transformation prior to bulk deconvolution. The flow cytometry data were obtained from the supplementary file 7 from the original publication and the data were rescaled as follows: CD4⁺ and CD8⁺ T cells proportions were rescaled to sum up to the proportion of T cells and the T cells, B cells and NK cells proportions were rescaled to sum up to the proportion of lymphocytes. All proportions were then rescaled to sum up to 1. EPIC-ATAC was applied on the ATAC-Seq data and EPIC to the RNA-Seq data and the performances of each set of predictions were compared to each other.

Code availability

The code to download and preprocess publicly available ATAC-Seq samples as well as the code used to identify our cell-type specific marker peaks and generate the reference profiles is available on GitHub (https://github.com/GfellerLab/EPIC-ATAC_Manuscript [↗](#)). A README file is provided on the GitHub repository with more details on how to use the code.

The code to perform ATAC-Seq deconvolution using the EPIC-ATAC framework is available as an R package called EPICATAC and is available on GitHub (<https://github.com/GfellerLab/EPIC-ATAC> [↗](#)).

Data availability

The newly generated ATAC-Seq data have been deposited on Zenodo (<https://zenodo.org/records/12571947> [↗](#)). The other data related to this work are available in the supplementary tables and on the Zenodo deposit (<https://zenodo.org/records/12571947> [↗](#)).

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

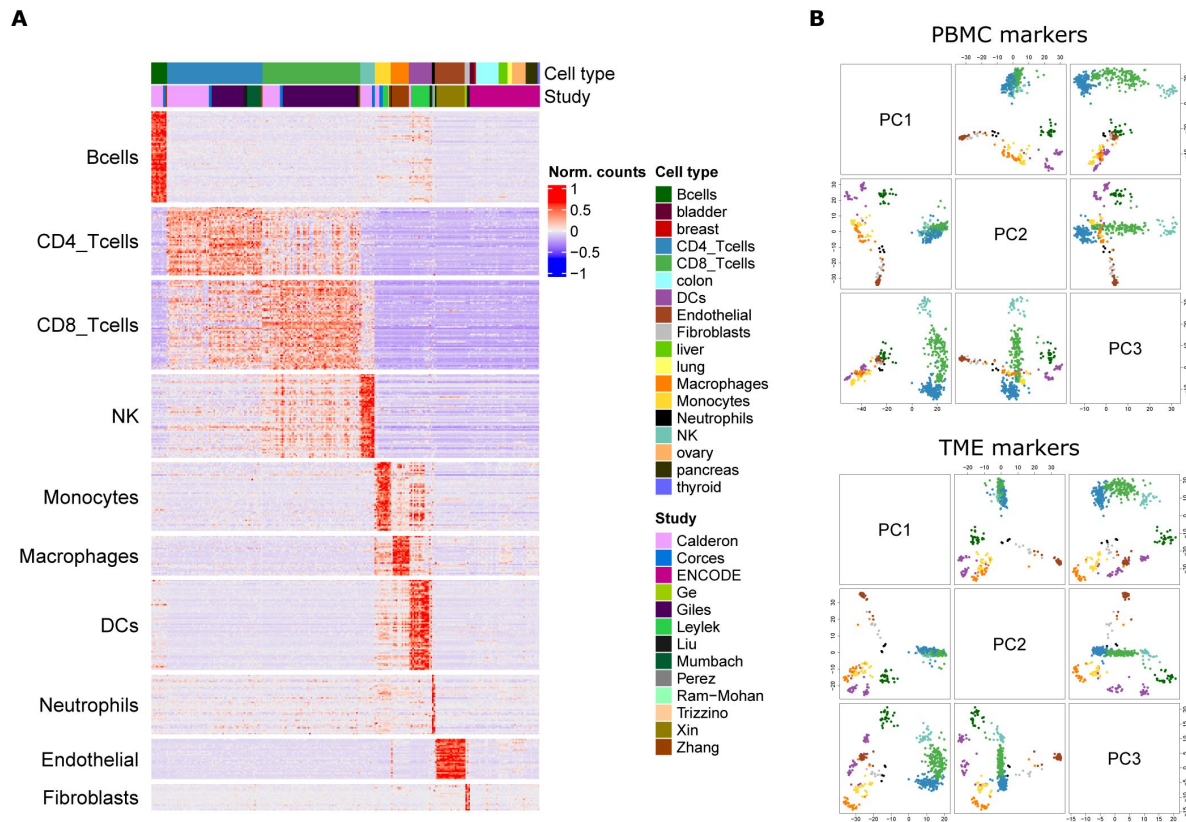
We thank the Lausanne Genomic Technologies Facility, University of Lausanne, Switzerland (<https://www.unil.ch/gtf/en/home.html>) for the sequencing of the PBMC samples as well as Yan Liu, Dana Moreno and Matei Teleman for testing the EPICATAC R package. Some of the illustrations were created with BioRender.com (**Figures 1**, 3A, 7C).

Authors contributions

Conceptualization: AAGG, JR, DG; Data curation: AAGG; Software: AAGG, JR; Experiments: MF, CJ; Visualization: AAGG, JR, DG; Methodology: AAGG, JR, DG; Writing—original draft: AAGG, DG; Writing—review and editing: all authors.

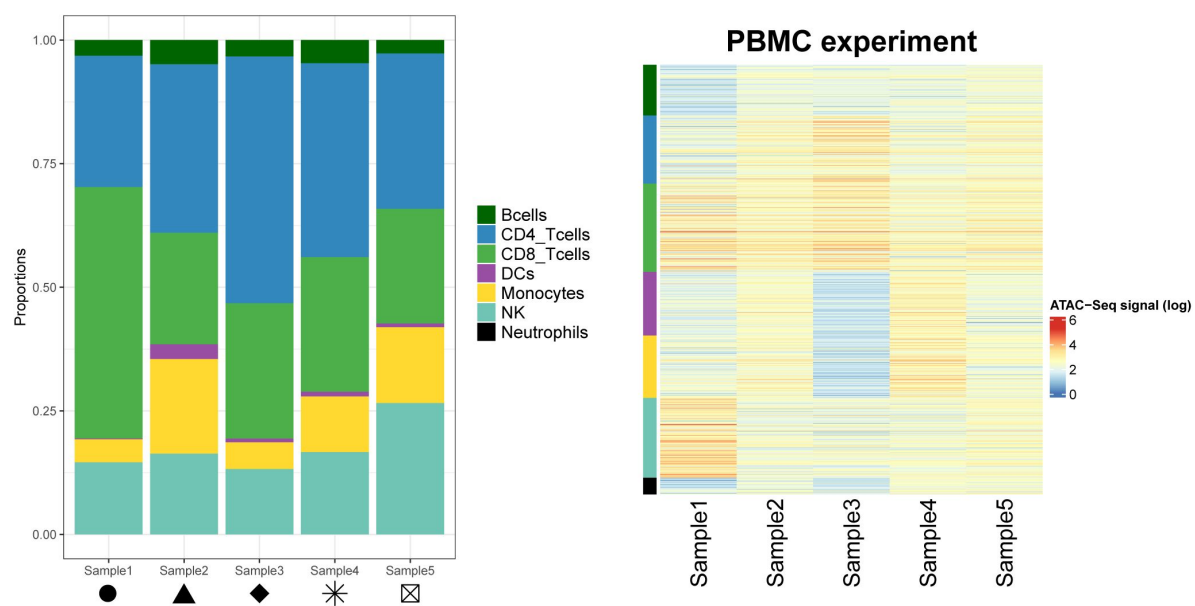
List of abbreviations

- ATAC: Assay for Transposase-Accessible chromatin
- CBP(s): chromatin binding protein(s)
- ChIP-seq: chromatin immunoprecipitation followed by sequencing
- CRE: cis-regulatory element
- DC: dendritic cells
- HTAN: Human Tumor Atlas Network
- NK: natural killer cells
- PCA: principal component analysis
- RMSE: root mean squared error
- TCGA: The Cancer Genome Atlas
- TF(s): transcription factor(s)
- TNBC: triple negative breast cancer
- TSS: transcription start site



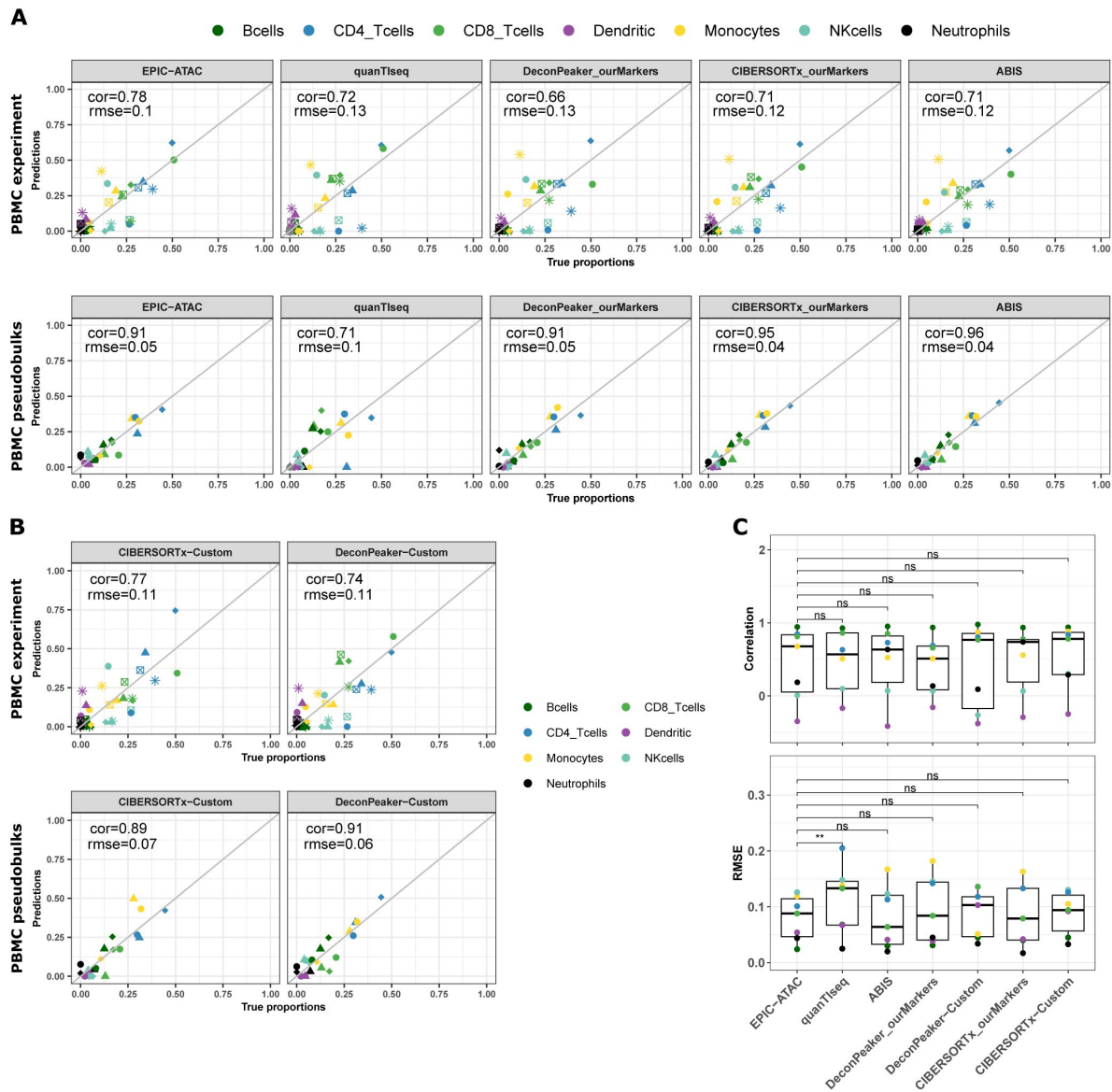
Supplementary Figure 1

A) Chromatin accessibility signal of each marker peak (rows) in each reference sample and in the ENCODE samples from diverse tissues (columns). **B)** Pairwise scatterplots of the first 3 axes of the PCA run on the PBMC markers (top) and TME markers (bottom).



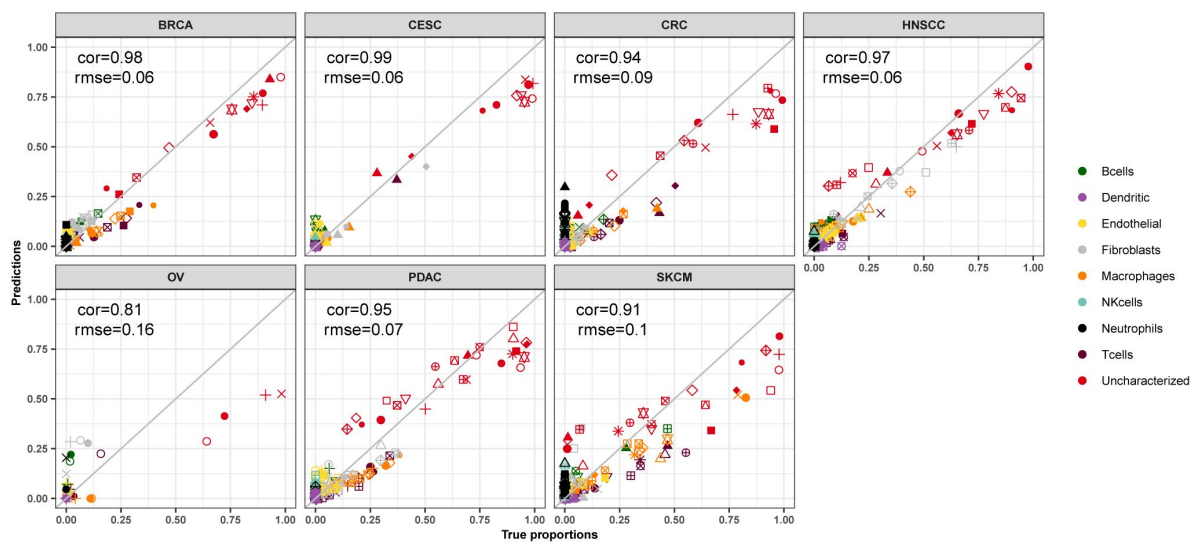
Supplementary Figure 2

Left panel: Proportions of each cell type in the PBMC experiment samples. Shapes match the points shapes from [Figure 3B](#) in the main text. Right panel: Chromatin accessibility signal (log scale) of the cell-type specific marker peaks (rows) in each sample (columns) of the PBMC experiment dataset.



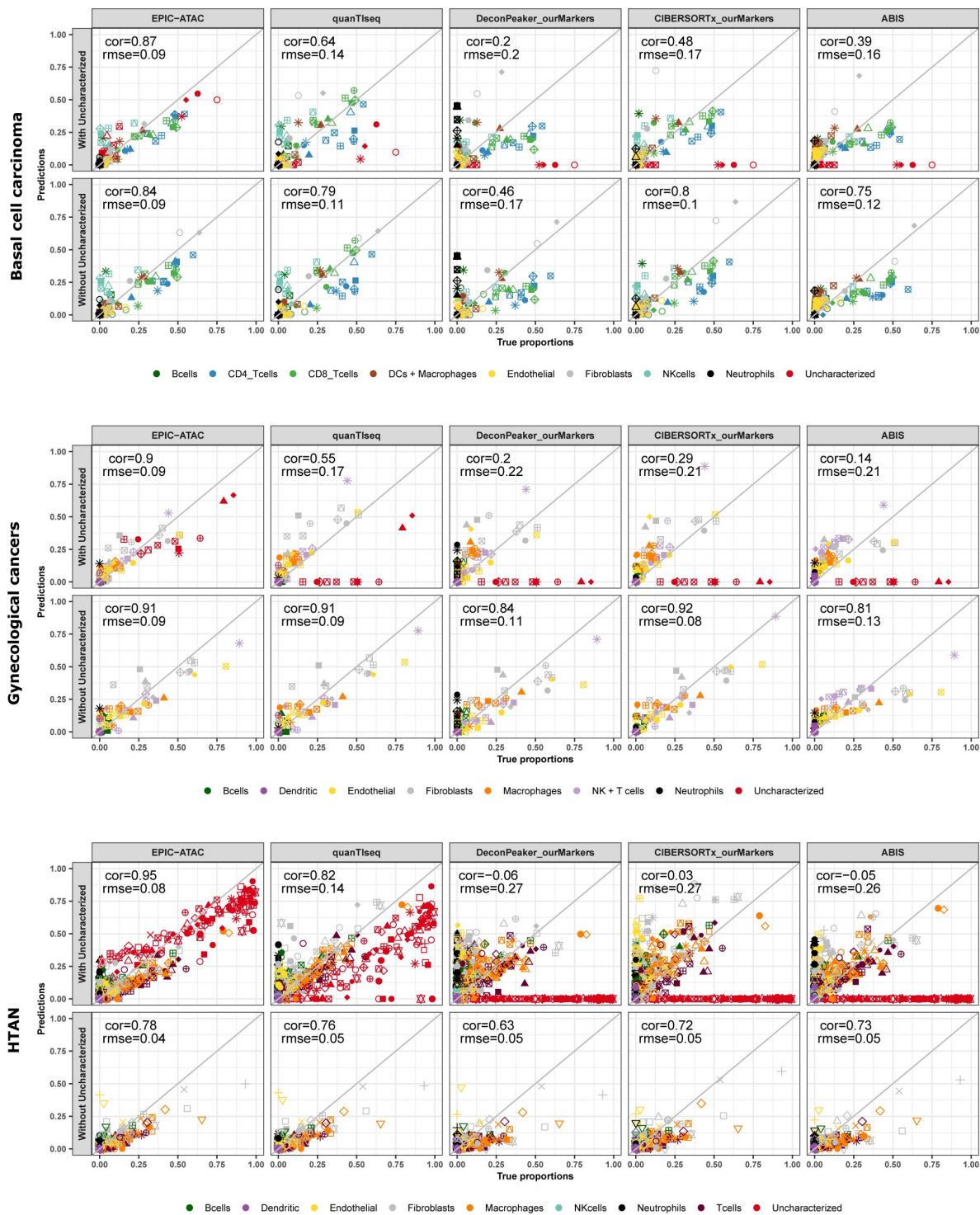
Supplementary Figure 3

A) Comparison of the cell-type proportions estimated by the different deconvolution methods included in the benchmark (y axis) and the true proportions (x axis) in the PBMC experiment samples (top panel) and in the PBMC pseudobulk samples (bottom panel). **B)** Comparison of the cell-type proportions estimated by CIBERSORTx and DeconPeaker using reference profiles built by each tool using our collection of pure ATAC-Seq samples and the true proportions in the PBMC experiment samples (top panel) and in the PBMC pseudobulk samples (bottom panel). **C)** Pearson's correlation coefficient and RMSE values associated to the comparison of the cell-type proportions estimated by each deconvolution tool in each cell type of the PBMC datasets, colors correspond to cell types. Note that for quantIseq, gray dots correspond to uncharacterized cells since the prediction of an uncharacterized cell type can not be turned off using the quantIseq function from the quantiseqr R package. *P*-values resulting from a paired wilcoxon test are represented for each pair of comparison between tools. Significance levels are represented as follows: *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.



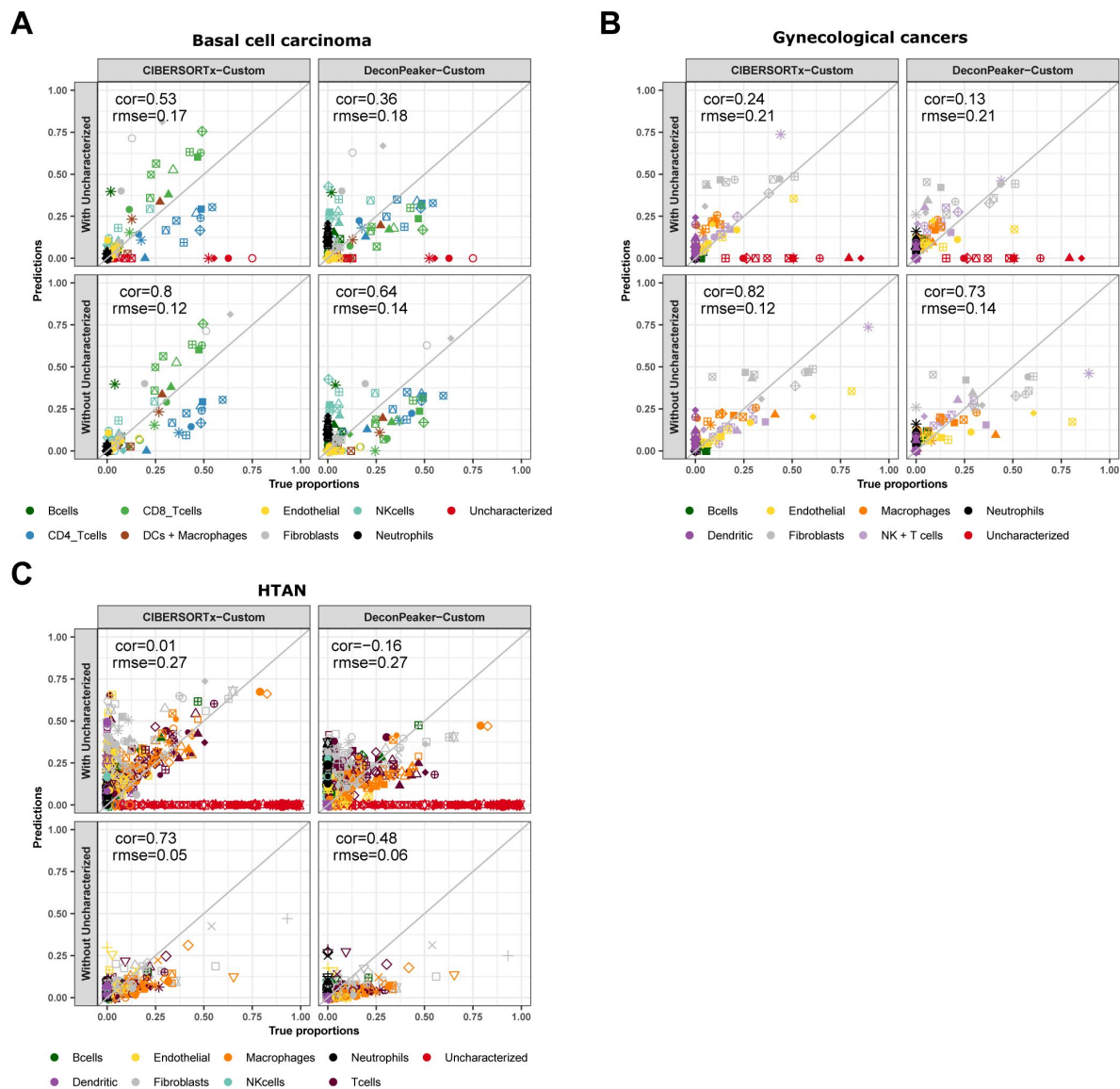
Supplementary Figure 4

Comparison of the cell-type proportions estimated by EPIC-ATAC (y axis) and the true proportions (x axis) in the HTAN dataset.



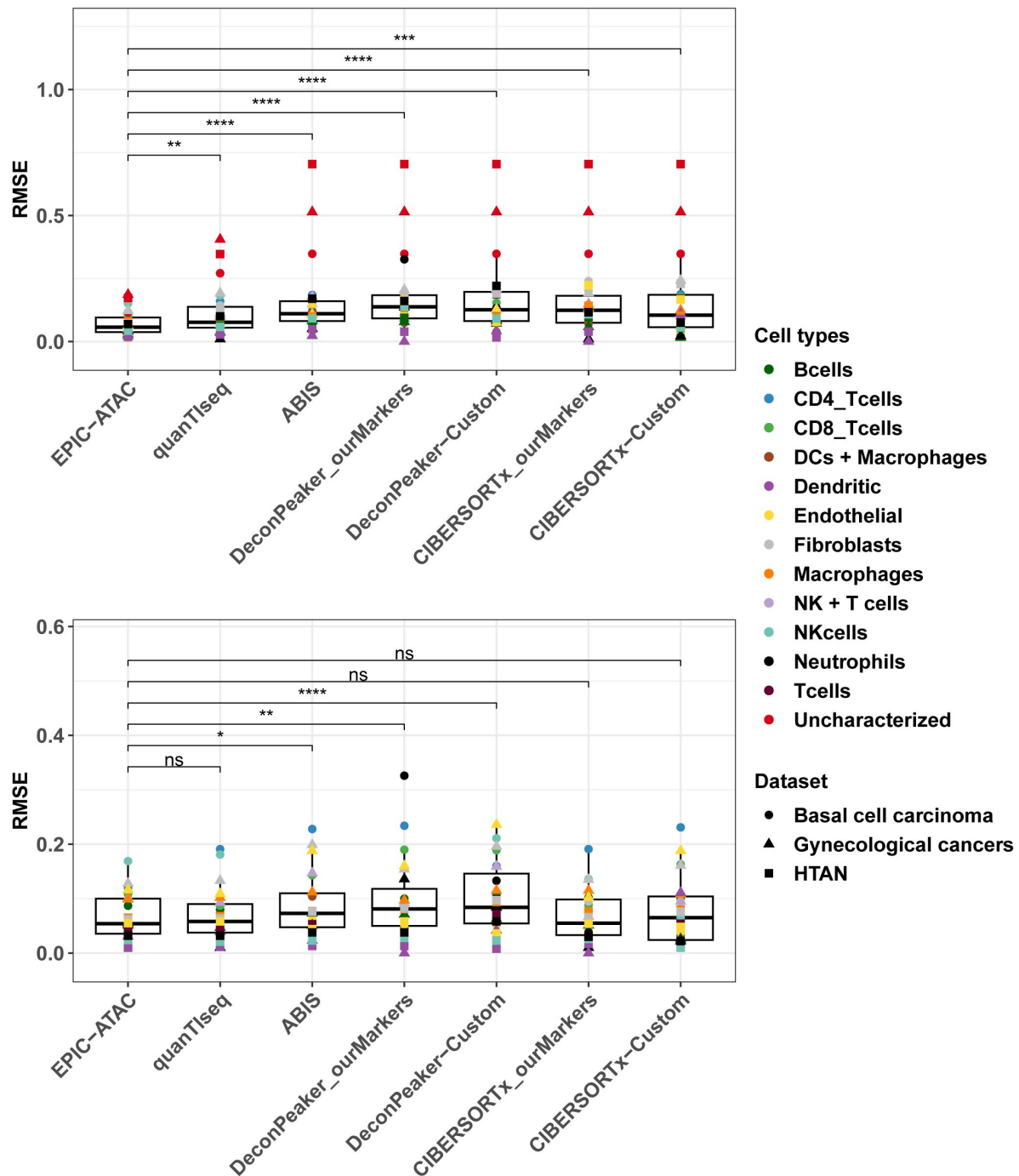
Supplementary Figure 5

Comparison of the cell-type proportions estimated by the different deconvolution methods using our reference profiles and the true proportions in the basal cell carcinoma samples (top panel), in the gynecological cancer samples (middle panel) and in the HTAN samples (bottom panel). In each panel, the top row includes the uncharacterized cells while the bottom row excludes them.



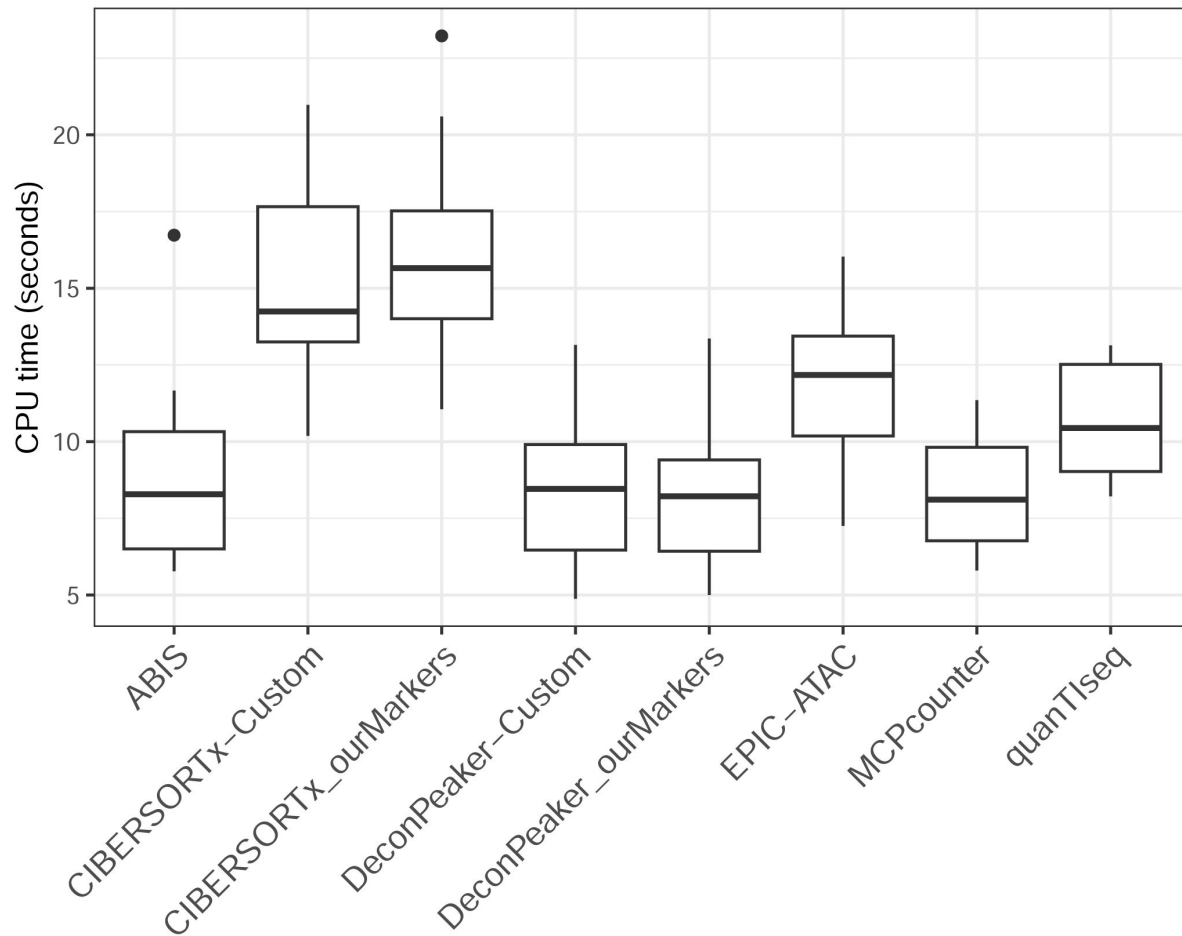
Supplementary Figure 6

Comparison of the cell-type proportions estimated by CIBERSORTx and Decon-Peaker using reference profiles built by each tool using our collection of pure ATAC-Seq samples and the true proportions in the basal cell carcinoma samples (panel A), in the gynecological cancer samples (panel B) and in the HTAN samples (panel C). In each panel, the top row includes the uncharacterized cells while the bottom row excludes them.



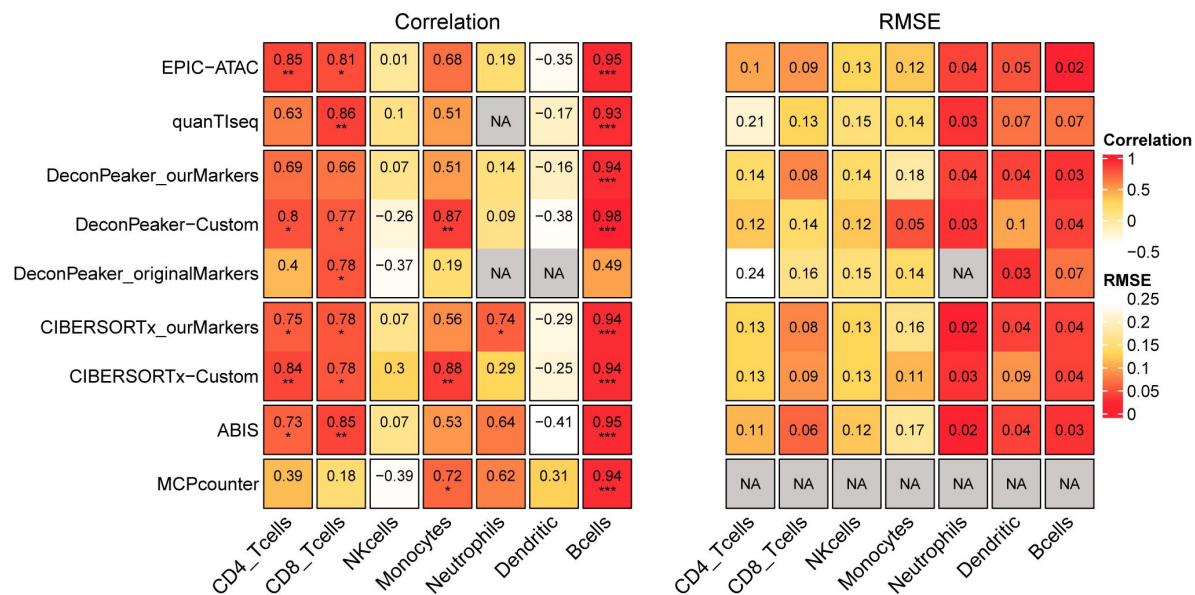
Supplementary Figure 7

Top panel: RMSE values associated to the comparison of the true proportions and the cell-type proportions estimated by each deconvolution tool in each cell type. For each cell type, all samples from each cancer pseudobulk dataset were considered, *i.e.*, colors correspond to cell types and shapes to datasets. *P*-values resulting from a paired wilcoxon test are represented for each pair of comparison between tools. Significance levels are represented as follows: *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$. Bottom panel: Considering that not all deconvolution tools are able to predict the proportion of uncharacterized cells, the same analysis as in the top panel was performed without considering the uncharacterized cells predictions.



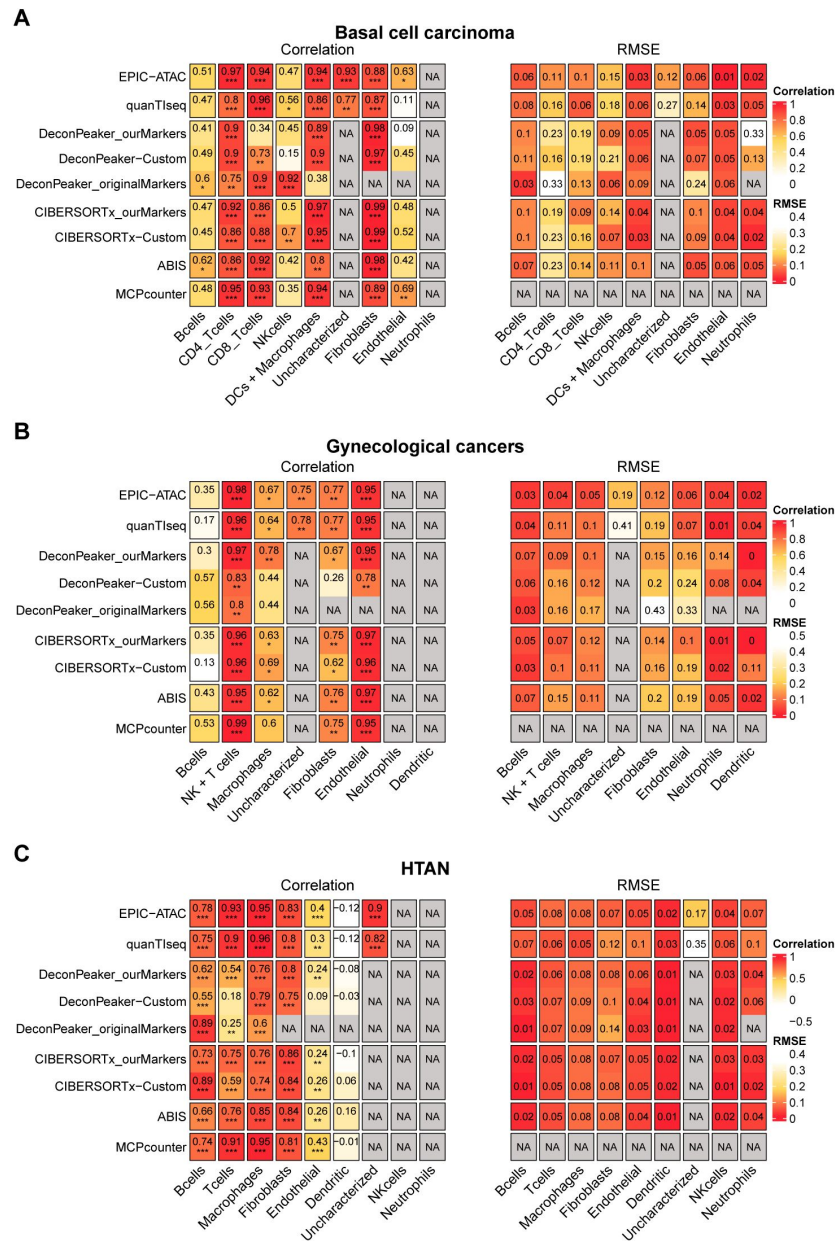
Supplementary Figure 8

CPU time in seconds needed to run each tool on the benchmarking datasets.



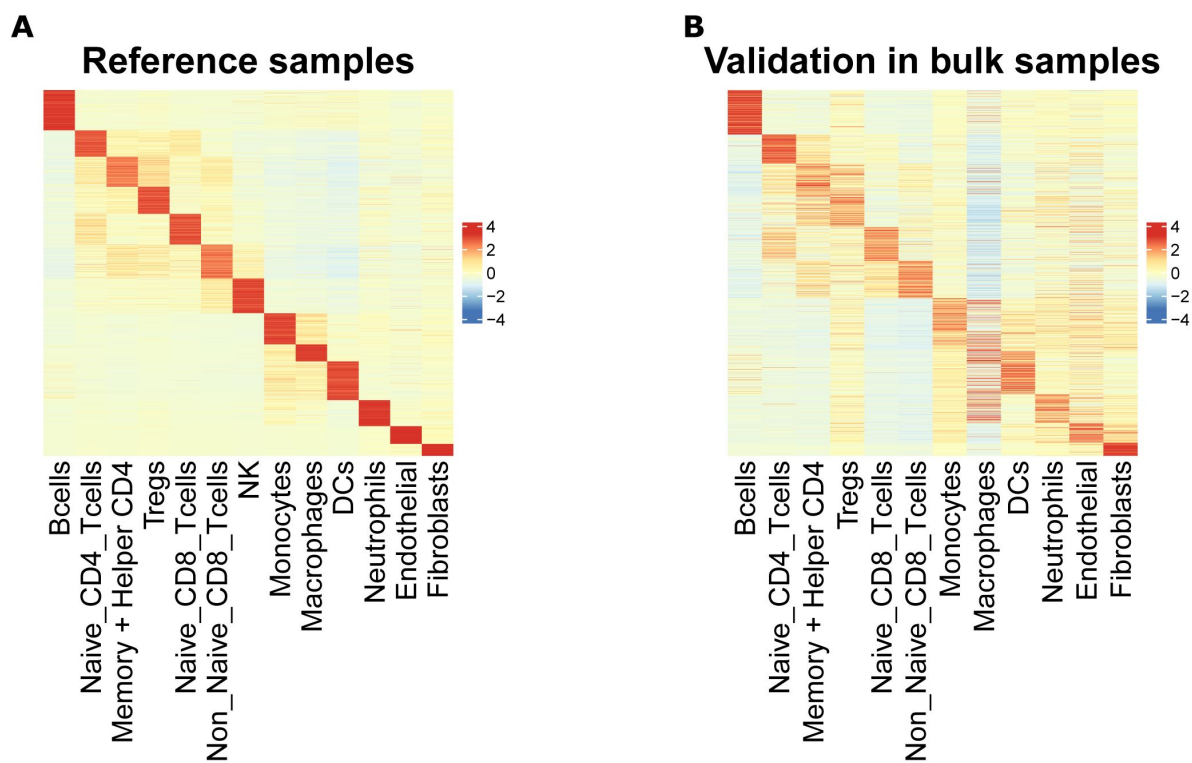
Supplementary Figure 9

Pearson's correlation (left panel) and RMSE values (right panel) associated to the comparison of the cell-type proportions estimated by each deconvolution tool (rows) included in the benchmark and the true cell-type proportions in the PBMC samples. The comparisons were made in each cell-type (columns) separately. Significance level of the correlations are represented as follows: *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.



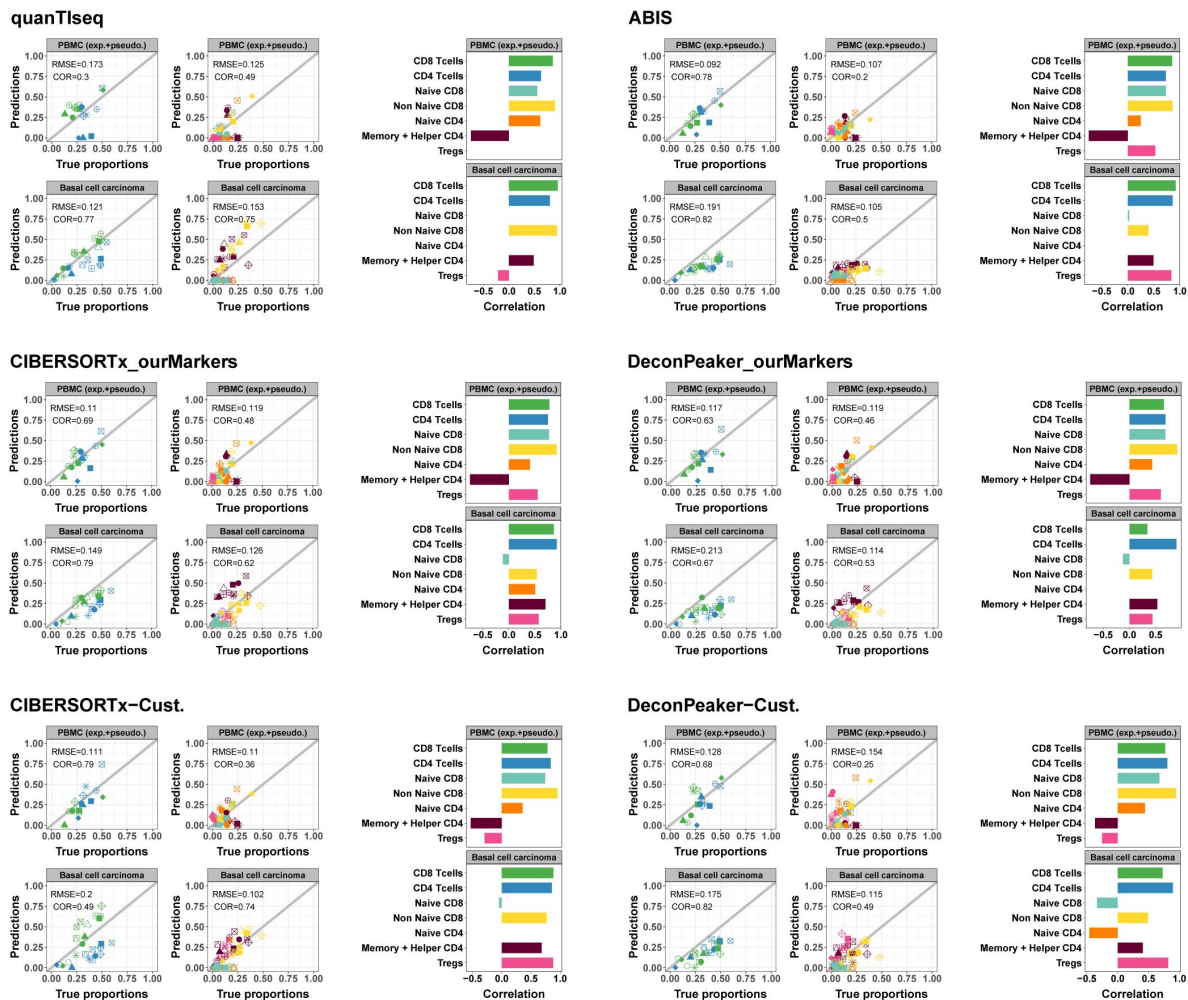
Supplementary Figure 10

Pearson's correlation (left panel) and RMSE values (right panel) associated to the comparison of the cell-type proportions estimated by each deconvolution tool (rows) included in the benchmark and the true cell-type proportions. The comparisons were made in each cell-type (columns) separately. Panel **A**, **B** and **C** are associated to the benchmark performed on the basal cell carcinoma dataset (A), the gynecological cancer dataset (B) and the HTAN dataset (C). Significance level of the correlations are represented as follows: *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.



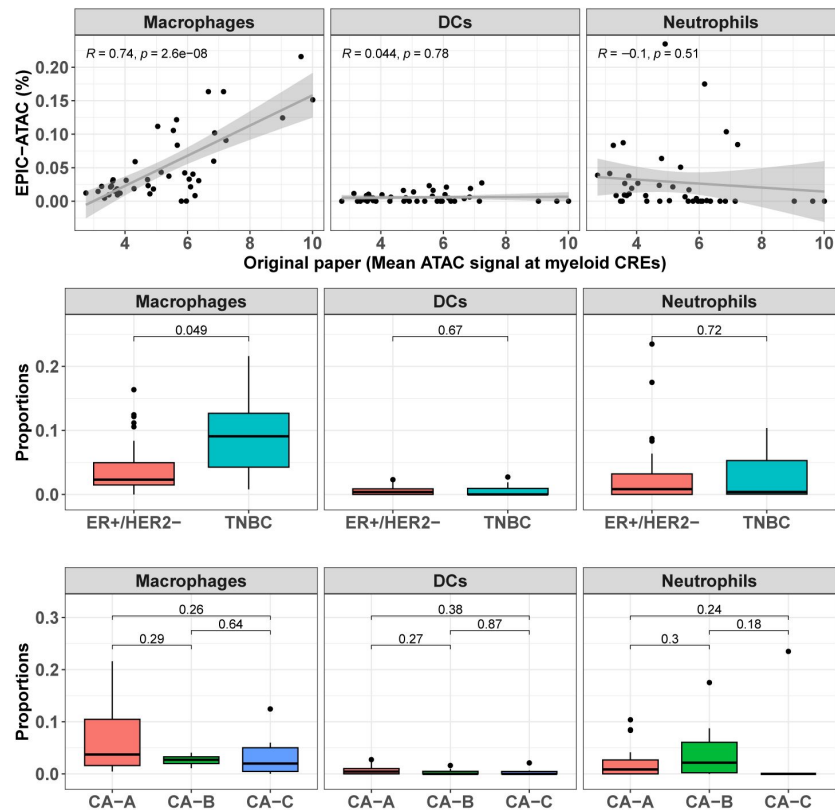
Supplementary Figure 11

A) Scaled averaged chromatin accessibility of each cell-type specific marker peak in each cell-type obtained based on the ATAC-Seq reference samples used to identify the marker peaks. **B)** Scaled averaged chromatin accessibility of each cell-type specific marker peak in each cell-type from the sorted samples excluded from the reference samples (See Material and Methods).



Supplementary Figure 12

Left panels: Scatter plots representing proportions estimated by the different deconvolution tools included in the benchmark (y-axis) as a function of the true proportions (x-axis) for PBMC samples (top row) and the basal cell carcinoma pseudobulks (bottom row). CD4+ and CD8+ T-cells predictions were obtained using the reference profiles based on the major cell-types and the CD4+ and CD8+ T-cell subtypes predictions using the reference profiles including the T-cells subtypes. Right panels: Barplots representing Pearson's correlation values (x-axis) obtained based on the comparison of the predictions and true proportions in each cell-type (y-axis).



Supplementary Figure 13

Top panel: Proportions of each myeloid cell type predicted by EPIC-ATAC in each sample as a function of the average ATAC signal at the myeloid specific CREs used by Kumegawa and colleagues to infer the level of infiltration in the tumor samples. Middle panel: Proportions of different cell types predicted by EPIC-ATAC in the samples stratified based on two breast cancer subtypes. Bottom panel: Proportions of different cell types predicted by EPIC-ATAC in the samples stratified based on three ER+/HER2- subgroups. Wilcoxon test p-values are represented at the top of the boxplots

References

- 10x Genomics (2021) **10x Genomics (2021) PBMC from a Healthy Donor - Granulocytes Removed Through Cell Sorting (10k)**. Available at: <https://www.10xgenomics.com/resources/datasets/pbmc-from-a-healthy-donorgranulocytes-removed-through-cell-sorting-10-k-1-standard-2-0-0> (Accessed: 6 September 2023).
- Arneson D., Yang X., Wang K (2020) **MethylResolver—a method for deconvoluting bulk DNA methylation profiles into known and unknown cell contents** *Communications Biology* **3** <https://doi.org/10.1038/s42003-020-01146-2>
- Aryee M. J., et al. (2014) **Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays** *Bioinformatics* **30**:1363–1369 <https://doi.org/10.1093/bioinformatics/btu049>
- Avila Cobos F., et al. (2018) **Computational deconvolution of transcriptomics data from mixed cell populations** *Bioinformatics* **34**:1969–1979 <https://doi.org/10.1093/BIOINFORMATICS/BTY019>
- Avila Cobos F., et al. (2020) **Benchmarking of cell type deconvolution pipelines for transcriptomics data** *Nature Communications* **11**:1–14 <https://doi.org/10.1038/s41467-020-19015-1>
- Becht E., et al. (2016) **Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression** *Genome Biology* **17** <https://doi.org/10.1186/S13059-016-1070-5/TABLES/4>
- Bonnema J. D., et al. (1994) **Fc receptor stimulation of phosphatidylinositol 3-kinase in natural killer cells is associated with protein kinase C-independent granule release and cell-mediated cytotoxicity** *Journal of Experimental Medicine* **180**:1427–1435 <https://doi.org/10.1084/JEM.180.4.1427>
- Buenrostro J. D., et al. (2013) **Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position** *Nature Methods* **10**:1213–1218 <https://doi.org/10.1038/nmeth.2688>
- Burdziak C., et al. (2019) **A Nonparametric Multi-view Model for Estimating Cell Type-Specific Gene Regulatory Networks** *ArXiv*
- Calderon D., et al. (2019) **Landscape of stimulation-responsive chromatin across diverse human immune cells** *Nature Genetics* **51**:1494–1505 <https://doi.org/10.1038/s41588-019-0505-9>
- Carvalho K., et al. (2021) **Uncovering the Gene Regulatory Networks Underlying Macrophage Polarization Through Comparative Analysis of Bulk and Single-Cell Data** *bioRxiv* <https://doi.org/10.1101/2021.01.20.427499>

- Castro-Mondragon J. A., et al. (2022) **JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles** *Nucleic Acids Research* **50**:D165–D173 <https://doi.org/10.1093/NAR/GKAB1113>
- Chakravarthy A., et al. (2018) **Pan-cancer deconvolution of tumour composition using DNA methylation** *Nature Communications* **9** <https://doi.org/10.1038/s41467-018-05570-1>
- Clarke J., Seol P., Clarke B (2010) **Statistical expression deconvolution from mixed tissue samples** *Bioinformatics* **26**:1043–1049 <https://doi.org/10.1093/BIOINFORMATICS/BTQ097>
- Corces M. R., et al. (2016) **Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution** *Nature Genetics* **48**:1193–1203 <https://doi.org/10.1038/ng.3646>
- Corces M. R., et al. (2017) **An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues** *Nature Methods* **14**:959–962 <https://doi.org/10.1038/nmeth.4396>
- Corces M. R., et al. (2018) **The chromatin accessibility landscape of primary human cancers** *Science* **362** <https://doi.org/10.1126/science.aav1898>
- Cusanovich D. A., et al. (2015) **Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing** *Science* **348**:910–914 https://doi.org/10.1126/SCIENCE.AAB1601/SUPPL_FILE/PAP.PDF
- Feng S., et al. (2023) **Decomprolute : A benchmarking platform designed for multiomics-based tumor deconvolution** *bioRxiv*
- Finotello F., et al. (2019) **Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data** *Genome Medicine* **11** <https://doi.org/10.1186/s13073-019-0638-6>
- Franzén O., Gan L. M., Björkegren J. L. M (2019) **PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data** *Database* **2019** <https://doi.org/10.1093/DATABASE/BAZ046>
- Fridman W. H., et al. (2012) **The immune contexture in human tumours: impact on clinical outcome** *Nature Reviews Cancer* **12**:298–306 <https://doi.org/10.1038/nrc3245>
- Fridman W. H., et al. (2017) **The immune contexture in cancer prognosis and treatment** *Nature Reviews Clinical Oncology*. Nature Publishing Group :717–734 <https://doi.org/10.1038/nrclinonc.2017.101>
- Ge X., et al. (2021) **Functional genomics atlas of synovial fibroblasts defining rheumatoid arthritis heritability** *Genome Biology* **22** <https://doi.org/10.1186/S13059-021-02460-6/FIGURES/7>
- Geng D., et al. (2010) **When Toll-like receptor and T-cell receptor signals collide: a mechanism for enhanced CD8 T-cell effector function** *Blood* **116**:3494–3504 <https://doi.org/10.1182/BLOOD-2010-02-268169>
- Giles J. R., et al. (2022) **Human epigenetic and transcriptional T cell differentiation atlas for identifying functional T cell-specific enhancers** *Immunity* **55**:557–574 <https://doi.org/10.1016/J.IMMUNI.2022.02.004>

- Gong T., Szustakowski J. D (2013) **DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data** *Bioinformatics* **29**:1083–1085 <https://doi.org/10.1093/BIOINFORMATICS/BTT090>
- Gosink M. M., Petrie H. T., Tsinoremas N. F (2007) **Electronically subtracting expression patterns from a mixed cell population** *Bioinformatics* **23**:3328–3334 <https://doi.org/10.1093/BIOINFORMATICS/BTM508>
- Grandi F. C., et al. (2022) **Chromatin accessibility profiling by ATAC-seq** *Nature protocols* :1518–1552 <https://doi.org/10.1038/s41596-022-00692-9>
- Granja J. M., et al. (2019) **Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia** *Nature Biotechnology* **37**:1458–1465 <https://doi.org/10.1038/s41587-019-0332-7>
- Granja J. M., et al. (2021) **ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis** *Nature Genetics* **53**:403–411 <https://doi.org/10.1038/s41588-021-00790-6>
- Hammal F., et al. (2022) **ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments** *Nucleic Acids Research* **50**:D316–D325 <https://doi.org/10.1093/NAR/GKAB996>
- Hicks S. C., Irizarry R. A (2019) **MethylCC: Technology-independent estimation of cell type composition using differentially methylated regions** *Genome Biology* **20**:1–13 <https://doi.org/10.1186/S13059-019-1827-8/FIGURES/5>
- Hu C., et al. (2023) **CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data** *Nucleic Acids Research* **51**:D870–D876 <https://doi.org/10.1093/NAR/GKAC947>
- Javaid N., Choi S (2020) **Toll-like Receptors from the Perspective of Cancer Treatment** *Cancers* **12** <https://doi.org/10.3390/CANCERS12020297>
- Jiang Y., et al. (2023) **scATAnno: Automated Cell Type Annotation for single-cell ATAC Sequencing Data** *bioRxiv* <https://doi.org/10.1101/2023.06.01.543296>
- Jimenez-Sanchez A., Cast O., Miller M. L (2019) **Comprehensive benchmarking and integration of tumor microenvironment cell estimation methods** *Cancer Research* **79**:6238–6246 <https://doi.org/10.1158/0008-5472.CAN-18-3560>
- Jin H., Liu Z (2021) **A benchmark for RNA-seq deconvolution analysis under dynamic testing environments** *Genome Biology* **22** <https://doi.org/10.1186/s13059-021-02290-6>
- Kalafati L., et al. (2020) **Innate Immune Training of Granulopoiesis Promotes Anti-tumor Activity** *Cell* **183**:771–785 <https://doi.org/10.1016/J.CELL.2020.09.058>
- Kaur D., et al. (2023) **Comprehensive Evaluation of The Infinium Human MethylationEPIC v2 BeadChip** *Epigenetics communications* **3** <https://doi.org/10.1186/S43682-023-00021-5>
- Klemm S. L., Shipony Z., Greenleaf W. J (2019) **Chromatin accessibility and the regulatory epigenome** *Nature Reviews Genetics* **20**:207–220 <https://doi.org/10.1038/s41576-018-0089-8>

- Kuhn M (2008) **Building Predictive Models in R Using the caret Package** *Journal of Statistical Software* **28**:1–26 <https://doi.org/10.18637/JSS.V028.I05>
- Kumegawa K., et al. (2022) **‘GRHL2 motif is associated with intratumor heterogeneity of cis-regulatory elements in luminal breast cancer** *npj Breast Cancer* **8**:1–12 <https://doi.org/10.1038/s41523-022-00438-6>
- Kumegawa K (2023) **ATAC-seq data of 42 BC samples as SummarizedExperiment object with count matrix, normalized count matrix, peak info, and clinical info** *figshare* <https://doi.org/10.6084/m9.figshare.21992609.v1>
- Kumegawa K., et al. (2023) **Chromatin profile-based identification of a novel ER-positive breast cancer subgroup with reduced ER-responsive element accessibility** *British Journal of Cancer* **128**:1208–1222 <https://doi.org/10.1038/s41416-023-02178-1>
- Lareau C. A., et al. (2019) **Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility** *Nature Biotechnology* **37**:916–924 <https://doi.org/10.1038/s41587-019-0147-6>
- Lau J. W., et al. (2017) **The cancer genomics cloud: Collaborative, reproducible, and democratized - A new paradigm in large-scale computational research** *Cancer Research* **77**:e3–e6 <https://doi.org/10.1158/0008-5472.CAN-17-0387>
- Law C. W., et al. (2014) **Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts** *Genome Biology* **15** <https://doi.org/10.1186/GB-2014-15-2-R29/FIGURES/11>
- Leylek R., et al. (2020) **Chromatin Landscape Underpinning Human Dendritic Cell Heterogeneity** *Cell Reports* **32** <https://doi.org/10.1016/j.CELREP.2020.108180>
- Li H., et al. (2020) **DeconPeaker, a Deconvolution Model to Identify Cell Types Based on Chromatin Accessibility in ATAC-Seq Data of Mixture Samples** *Frontiers in Genetics* **11** <https://doi.org/10.3389/fgene.2020.00392>
- Li T., et al. (2020) **TIMER2.0 for analysis of tumor-infiltrating immune cells** *Nucleic acids research* **48**:W509–W514 <https://doi.org/10.1093/nar/gkaa407>
- Liao Y., Smyth G. K., Shi W (2014) **featureCounts: an efficient general purpose program for assigning sequence reads to genomic features** *Bioinformatics* **30**:923–930 <https://doi.org/10.1093/BIOINFORMATICS/BTT656>
- Liu Q., et al. (2020) **Chromatin accessibility landscapes of skin cells in systemic sclerosis nominate dendritic cells in disease pathogenesis** *Nature Communications* **11** <https://doi.org/10.1038/s41467-020-19702-z>
- Loyfer N., et al. (2023) **A DNA methylation atlas of normal human cell types** *Nature* **613**:355–364 <https://doi.org/10.1038/s41586-022-05580-6>
- Luo L., Gribskov M., Wang S (2022) **Bibliometric review of ATAC-Seq and its application in gene expression** *Briefings in Bioinformatics* <https://doi.org/10.1093/BIB/BBAC061>
- Machlab D., et al. (2022) **monaLisa: an R/Bioconductor package for identifying regulatory motifs** *Bioinformatics* **38**:2624–2625 <https://doi.org/10.1093/BIOINFORMATICS/BTAC102>

Mayer M. (2023) **package “ splitTools ”**

Monaco G., et al. (2019) **RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types** *Cell Reports* **26**:1627–1640 <https://doi.org/10.1016/j.CELREP.2019.01.041>

Morandini F., et al. (2024) **ATAC-clock: An aging clock based on chromatin accessibility** *GeroScience* **46**:1789–1806 <https://doi.org/10.1007/S11357-023-00986-0/FIGURES/4>

Mumbach M. R., et al. (2017) **Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements** *Nature Genetics* **49**:1602–1612 <https://doi.org/10.1038/ng.3963>

Newman A. M., et al. (2015) **Robust enumeration of cell subsets from tissue expression profiles** *Nature Methods* **12**:453–457 <https://doi.org/10.1038/nmeth.3337>

Newman A. M., et al. (2019) **Determining cell type abundance and expression from bulk tissues with digital cytometry** *Nature Biotechnology* **37**:773–782 <https://doi.org/10.1038/s41587-019-0114-2>

Peng X. L., et al. (2019) **De novo compartment deconvolution and weight estimation of tumor samples using DECODER** *Nature Communications* **10** <https://doi.org/10.1038/s41467-019-12517-7>

Perez C., et al. (2020) **Immunogenomic identification and characterization of granulocytic myeloid-derived suppressor cells in multiple myeloma** *Blood* **136**:199–209 <https://doi.org/10.1182/BLOOD.2019004537>

Qiu Y., et al. (2021) **Identification of cell-type-specific marker genes from co-expression patterns in tissue samples** *Bioinformatics* **37**:3228–3234 <https://doi.org/10.1093/BIOINFORMATICS/BTAB257>

Racle J., et al. (2017) **Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data** *eLife* **6** <https://doi.org/10.7554/eLife.26476>

Racle J., Gfeller D (2020) **‘EPIC: A tool to estimate the proportions of different cell types from bulk gene expression data’**, in *Methods in Molecular Biology* Humana Press Inc :233–248 https://doi.org/10.1007/978-1-0716-0327-7_17

Rahmani E., et al. (2019) **Cell-type-specific resolution epigenetics without the need for cell sorting or single-cell biology** *Nature Communications* **10**:1–11 <https://doi.org/10.1038/s41467-019-11052-9>

Ram-Mohan N., et al. (2021) **Profiling chromatin accessibility responses in human neutrophils with sensitive pathogen detection** *Life Science Alliance* **4** <https://doi.org/10.26508/LSA.202000976>

Regner M. J., et al. (2021) **A multi-omic single-cell landscape of human gynecologic malignancies** *Molecular Cell* **81**:4924–4941 <https://doi.org/10.1016/j.molcel.2021.10.013>

Risso D., et al. (2011) **GC-Content Normalization for RNA-Seq Data** *BMC Bioinformatics* **12** <https://doi.org/10.1186/1471-2105-12-480/FIGURES/7>

- Ritchie M. E., et al. (2015) **limma powers differential expression analyses for RNA-sequencing and microarray studies** *Nucleic Acids Research* **43** <https://doi.org/10.1093/nar/gkv007>
- Robinson M. D., McCarthy D. J., Smyth G. K (2010) **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data** *Bioinformatics* **26**:139–140 <https://doi.org/10.1093/BIOINFORMATICS/BTP616>
- Rozowsky J., et al. (2023) **The EN-TEx resource of multi-tissue personal epigenomes & variant-impact models** *Cell* **186**:1493–1511 <https://doi.org/10.1016/j.cell.2023.02.018>
- Salas L. A., et al. (2022) **Enhanced cell deconvolution of peripheral blood using DNA methylation for high-resolution immune profiling** *Nature Communications* **13**:1–13 <https://doi.org/10.1038/s41467-021-27864-7>
- Sanseviero E (2019) **NK Cell-Fc Receptors Advance Tumor Immunotherapy** *Journal of Clinical Medicine* **8** <https://doi.org/10.3390/JCM8101667>
- Satpathy A. T., et al. (2019) **Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion** *Nature Biotechnology* **37**:925–936 <https://doi.org/10.1038/s41587-019-0206-z>
- Smith J. P., et al. (2021) **PEPATAC: an optimized pipeline for ATAC-seq data analysis with serial alignments** *NAR Genomics and Bioinformatics* **3** <https://doi.org/10.1093/NARGAB/LQAB101>
- Stuart T., et al. (2021) **Single-cell chromatin state analysis with Signac** *Nature Methods* **18**:1333–1341 <https://doi.org/10.1038/s41592-021-01282-5>
- Sturm G., et al. (2019) **Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology** *Bioinformatics* **35**:i436–i445 <https://doi.org/10.1093/bioinformatics/btz363>
- Subramanian A., et al. (2005) **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles** *PNAS* **102**:15545–50 <https://doi.org/10.1073/pnas.0506580102>
- Terekhanova N. V, et al. (2023) **Epigenetic regulation during cancer transitions across 11 tumour types** *Nature* **2023**:1–10 <https://doi.org/10.1038/s41586-023-06682-5>
- Teschendorff A. E., et al. (2017) **A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies** *BMC bioinformatics* **18** <https://doi.org/10.1186/S12859-017-1511-5>
- Teschendorff A. E., et al. (2020) **EPISCORE: Cell type deconvolution of bulk tissue DNA methylomes from single-cell RNA-Seq data** *Genome Biology* **21** <https://doi.org/10.1186/s13059-020-02126-9>
- The ENCODE Project Consortium, et al. (2020) **Expanded encyclopaedias of DNA elements in the human and mouse genomes** *Nature* **583**:699–710 <https://doi.org/10.1038/s41586-020-2493-4>

- Trizzino M., et al. (2021) **EGR1 is a gatekeeper of inflammatory enhancers in human macrophages** *Science Advances* **7** https://doi.org/10.1126/SCIADV.AAZ8836/SUPPL_FILE/AAZ8836_TABLE_S7.XLSX
- Ucar D., et al. (2017) **The chromatin accessibility signature of human immune aging stems from CD8+ T cells** *Journal of Experimental Medicine* **214**:3123–3144 <https://doi.org/10.1084/jem.20170416>
- de Visser K. E., Joyce J. A. (2023) **The evolving tumor microenvironment: From cancer initiation to metastatic outgrowth** *Cancer Cell* **41**:374–403 <https://doi.org/10.1016/j.CCELL.2023.02.016>
- Watt S., et al. (2021) **Genetic perturbation of PU.1 binding and chromatin looping at neutrophil enhancers associates with autoimmune disease** *Nature Communications* **12** <https://doi.org/10.1038/S41467-021-22548-8>
- Welch R. P., et al. (2014) **ChIP-Enrich: gene set enrichment testing for ChIP-seq data** *Nucleic Acids Research* **42** <https://doi.org/10.1093/NAR/GKU463>
- Xin J., et al. (2020) **Chromatin accessibility landscape and regulatory network of high-altitude hypoxia adaptation** *Nature Communications* **11** <https://doi.org/10.1038/S41467-020-18638-8>
- Yu G., Wang L. G., He Q. Y (2015) **ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization** *Bioinformatics* **31**:2382–2383 <https://doi.org/10.1093/BIOINFORMATICS/BTV145>
- Zeng W., et al. (2019) **DC3 is a method for deconvolution and coupled clustering from bulk and single-cell genomics data** *Nature Communications* **10** <https://doi.org/10.1038/s41467-019-12547-1>
- Zhang H., et al. (2021) **EMeth: An EM algorithm for cell type decomposition based on DNA methylation data** *Scientific Reports* **11** <https://doi.org/10.1038/s41598-021-84864-9>
- Zhang H., et al. (2022) **Profiling chromatin accessibility in formalin-fixed paraffin-embedded samples** *Genome Research* **32**:150–161 <https://doi.org/10.1101/GR.275269.121>
- Zhang K., et al. (2021) **A single-cell atlas of chromatin accessibility in the human genome** *Cell* **184**:5985–6001 <https://doi.org/10.1016/j.cell.2021.10.024>
- Zhang P., et al. (2022) **Epigenomic analysis reveals a dynamic and context-specific macrophage enhancer landscape associated with innate immune activation and tolerance** *Genome Biology* **23** <https://doi.org/10.1186/S13059-022-02702-1>
- Zhang Y., et al. (2008) **Model-based analysis of ChIP-Seq (MACS)** *Genome Biology* **9** <https://doi.org/10.1186/GB-2008-9-9-R137/FIGURES/3>

Editors

Reviewing Editor

Siming Zhao

Dartmouth College, Lebanon, United States of America

Senior Editor

Richard White

University of Oxford, Oxford, United Kingdom

Reviewer #1 (Public Review):

Summary:

Building upon their famous tool for the deconvolution of human transcriptomics data (EPIC), Gabriel et al. implemented a new methodology for the quantification of the cellular composition of samples profiled with Assay for Transposase-Accessible Chromatin sequencing (ATAC-seq). To build a signature for ATAC-seq deconvolution, they first created a compendium of ATAC-seq data and derived chromatin accessibility marker peaks and reference profiles for 12 cell types, encompassing immune cells, endothelial cells, and fibroblasts. Then, they coupled this novel signature with the EPIC deconvolution framework based on constrained least-square regression to derive a dedicated tool called EPIC-ATAC. The method was then assessed using real and pseudo-bulk ATAC-seq data from human peripheral blood mononuclear cells (PBMC) and, finally, applied to ATAC-seq data from breast cancer tumors to show it accurately quantifies their immune contexture.

Strengths:

Overall, the work is of very high quality. The proposed tool is timely; its implementation, characterization, and validation are based on rigorous methodologies and results in robust estimates. The newly-generated, validation data and the code are publicly available and well-documented. Therefore, I believe this work and the associated resources will greatly benefit the scientific community.

Weaknesses:

In the benchmarking analysis, EPIC-ATAC was compared also to deconvolution methods that were originally developed for transcriptomics and not for ATAC-seq data. However, the authors described in detail the specific settings used to analyze this different data modality as robustly as possible, and they discussed possible limitations and ideas for future improvement.

<https://doi.org/10.7554/eLife.94833.2.sa2>

Reviewer #2 (Public Review):

Summary:

The manuscript expands the current bulk sequencing data deconvolution toolkit to include ATAC-seq. The EPIC-ATAC tool successfully predicts accurate proportions of immune cells in bulk tumour samples and EPIC-ATAC seems to perform well in benchmarking analyses. The authors achieve their aim of developing a new bulk ATAC-seq deconvolution tool.

Strengths:

The manuscript describes simple and understandable experiments to demonstrate the accuracy of EPIC-ATAC. They have also been incredibly thorough with their reference dataset collections and have been robust in their benchmarking endeavours and measured EPIC-ATAC against multiple datasets and tools. This tool will be valuable to the community it serves.

<https://doi.org/10.7554/eLife.94833.2.sa1>

Author response:

The following is the authors' response to the original reviews.

Public Reviews:

Reviewer #1 (Public Review):

Summary:

Building upon their famous tool for the deconvolution of human transcriptomics data (EPIC), Gabriel et al. implemented a new methodology for the quantification of the cellular composition of samples profiled with Assay for Transposase-Accessible Chromatin sequencing (ATAC-Seq). To build a signature for ATAC-seq deconvolution, they first created a compendium of ATAC-seq data and derived chromatin accessibility marker peaks and reference profiles for 21 cell types, encompassing immune cells, endothelial cells, and fibroblasts. They then coupled this novel signature with the EPIC deconvolution framework based on constrained least-square regression to derive a dedicated tool called EPIC-ATAC. The method was then assessed using real and pseudo-bulk RNA-seq data from human peripheral blood mononuclear cells (PBMC) and, finally, applied to ATAC-seq data from breast cancer tumors to show it accurately quantifies their immune contexture.

Strengths:

Overall, the work is of very high quality. The proposed tool is timely; its implementation, characterization, and validation are based on rigorous methodologies and resulted in robust results. The newly-generated, validation data and the code are publicly available and well-documented. Therefore, I believe this work and the associated resources will greatly benefit the scientific community.

Weaknesses:

CA few aspects can be improved to clarify the value and applicability of the EPIC-ATAC and the transparency of the benchmarking analysis.

(1) Most of the validation results in the main text assess the methods on all cell types together, by showing the correlation, RMSE, and scatterplots of the estimated vs. true cell fractions. This approach is valuable for showing the overall method performance and for detecting systematic biases and noisy estimates. However, it provides very limited insights regarding the capability of the methods to estimate the individual cell types, which is the ultimate aim of deconvolution analysis. This limitation is exacerbated for rare cell types, which could even have a negative correlation with the ground truth fractions, but not weigh much on the overall RMSE and correlation. I would suggest integrating into the main text and figures an in-depth assessment of the individual cell types. In particular, it should be shown and discussed which cell types can be accurately quantified and which ones are less reliable.

We thank the reviewer for raising this important point. Discussing the accuracy of EPIC-ATAC in predicting individual cell-type proportions would indeed be valuable in the main text. We have updated the text as follows.

In the first version of our manuscript, we had a section called “T cell subtypes quantification reveals the ATAC-Seq deconvolution limits for closely related cell types” which highlighted that EPIC-ATAC shows low performances when predicting the proportions of cell types that are closely related, e.g., CD4⁺ T cell or CD8⁺ T cell subtypes. The section is now named

“Accuracy of ATAC-Seq deconvolution is determined by the abundance and specificity of each cell type” and has been expanded to discuss the accuracy of EPIC-ATAC predictions within each major cell type.

To do so, we represented in Figure 5A the performances of EPIC-ATAC in each cell type present in the benchmarking datasets from Figures 3 and 4. Additionally, we have kept in the supplementary figures the details of the correlation values and RMSE values within each cell type and for each tool (Supplementary Figures 9 and 10). The following text has been added in the main text to describe these analyses:

“Accuracy of ATAC-Seq deconvolution is determined by the abundance and specificity of each cell type

To investigate the impact of cell type abundance on the accuracy of ATAC-Seq deconvolution, we evaluated EPIC-ATAC predictions in each major cell type separately in the different benchmarking datasets (Figure 5A). NK cells, endothelial cells, neutrophils or dendritic cells showed lower correlation values. These values can be explained by the fact that these cell types are low-abundant in our benchmarking datasets (Figure 5A). For the endothelial cells and dendritic cells, the RMSE values associated to these cell types remain low. This suggests that while the predictions of EPIC-ATAC might not be precise enough to compare these cell-type proportions between different samples, the cell-type quantification within each sample is reliable. For the NK cells and the neutrophils, we observed more variability with higher RMSE values in some datasets which suggests that the markers and profiles for these cell types might be improved. Supplementary Figures 9 and 10 detail the performances of each tool when considering each cell type separately in the PBMC and the cancer datasets. As for EPIC-ATAC, the predictions from the other deconvolution tools are more reliable for the frequent cell types.”

(2) In the benchmarking analysis, EPIC-ATAC is compared to several deconvolution methods, most of which were originally developed for transcriptomics data. This comparison is not completely fair unless their peculiarities and the limitations of tweaking them to work with ATAC-seq data are discussed. For instance, some methods (including the original EPIC) correct for cell-type-specific mRNA bias, which is not present in ATAC-seq data and might, thus, result in systematic errors.

We thank the reviewer for this comment and have updated the results and methods sections as follows:

We provide in the Materials and methods section, the paragraph “Benchmarking of the EPIC-ATAC framework against other existing deconvolution tools” which describes how each tool included in the benchmark was used in the ATAC-Seq context. We have added a reference to this section in the main text when introducing the first benchmarking analysis.

For each tool, the main changes consisted in: (i) replacing the initial RNA-Seq profiles and markers by the EPIC-ATAC reference profiles and markers and (ii) providing as input a bulk ATAC-Seq dataset with matched ATAC-Seq features (the same approach as the one used in EPIC-ATAC was considered, see answer to the next comment). Having reference profiles/markers and an ATAC-Seq bulk query with matched features was the only requirement of the different deconvolution models to be able to run on ATAC-Seq data with the default methods parameters, except for quanTIseq. Indeed, this method, like EPIC, corrects its estimations for cell-type-specific mRNA content bias. We have disabled this option for the bulk ATAC-Seq deconvolution.

We can however not exclude that a hyper parametrization of each tool could have helped to improve their current performances. Also, for RNA-Seq data deconvolution, some of the methods followed specific features filtering, e.g., the quanTIseq framework removes a

manually curated list of noisy genes as well as aberrant immune genes identified in the TCGA data and ABIS uses immune-specific housekeeping genes. We can hypothesize that additional filtering could be explored for the ATAC-Seq deconvolution to improve the performance of the tools.

We have clarified these points in the results section when introducing the benchmarking, in the methods and in the discussion section.

(3) On a similar note, it could be made more explicit which adaptations were introduced in EPIC, besides the ad-hoc ATAC-seq signature, to make it applicable to this type of data.

In the first version of the manuscript, we described the changes brought to EPIC to perform bulk ATAC-Seq deconvolution in the Material and methods section in the paragraph “Running EPIC-ATAC on bulk ATAC-Seq data”. We have moved and completed this paragraph in the results section before the description of the evaluation of EPIC-ATAC in different datasets. The paragraph is the following:

“EPIC-ATAC integrates the marker peaks and profiles into EPIC to perform bulk ATAC-Seq data deconvolution

The cell-type specific marker peaks and profiles derived from the reference samples were integrated into the EPIC deconvolution tool (Racle et al., 2017; Racle and Gfeller, 2020). We will refer to this ATAC-Seq deconvolution framework as EPIC-ATAC. To ensure the compatibility of any input bulk ATAC-Seq dataset with the EPIC-ATAC marker peaks and reference profiles, we provide an option to lift over hg19 datasets to hg38 (using the liftOver R package) as the reference profiles are based on the hg38 reference genome. Subsequently, the features of the input bulk matrix are matched to our reference profiles’ features. To match both sets of features, we determine for each peak of the input bulk matrix the distance to the nearest peak in the reference profiles peaks. Overlapping regions are retained and the feature IDs are matched to their associated nearest peaks. If multiple features are matched to the same reference peak, the counts are summed. Before the estimation of the cell-type proportions, we transform the data following an approach similar to the transcripts per million (TPM) transformation which has been shown to be appropriate to estimate cell fractions from bulk mixtures in RNA-Seq data (Racle et al., 2017; Sturm et al., 2019). We normalize the ATAC-Seq counts by dividing counts by the peak lengths as well as samples depth and rescaling counts so that the counts of each sample sum to 106. In RNA-Seq based deconvolution, EPIC uses an estimation of the amount of mRNA in each reference cell type to derive cell proportions while correcting for cell-type-specific mRNA bias. For the ATAC-Seq based deconvolution these values were set to 1 to give similar weights to all cell-types quantifications. Indeed ATAC-Seq measures signal at the DNA level, hence the quantity of DNA within each reference cell type is similar.”

(4) Given that the final applicability of EPIC-ATAC is on real bulk RNA-seq data, whose characteristics might not be completely recapitulated by pseudo-bulk samples, it would be interesting to see EPIC and EPIC-ATAC compared on a dataset with matched, real bulk RNA-seq and ATAC-seq, respectively. It would nicely complement the analysis of Figure 7 and could be used to dissect the commonalities and peculiarities of these two approaches.

We thank the reviewer for raising this important point. EPIC-ATAC will be applied to real bulk ATAC-Seq data and pseudobulk data cannot indeed fully recapitulate the bulk signals. Recently, a dataset composed of more than 100 samples with matched bulk RNA-Seq, bulk ATAC-Seq as well as matched flow cytometry data has been published by Morandini and colleagues in GeroScience in November 2023. We thus retrieved these data to compare the predictions obtained by EPIC-ATAC on the bulk ATAC-Seq data and the predictions of the

original version of EPIC on the bulk RNA-Seq data to the cell-type quantification obtained by flow cytometry. We also assessed whether both modalities could be complementary using a simple approach averaging the predictions obtained from both modalities. The results of these analyses have been summarized in the Figure 7C and are described in the main text in the last paragraph of the paper:

“We compared the predictions obtained using each modality to the flow cytometry cell-type quantifications. EPIC-ATAC predictions were better correlated with the flow cytometry measures for some cell types (e.g., CD8+, CD4+ T cells, NK cells) while this trend was observed with the EPIC-RNA predictions in other cell types (B cells, neutrophils, monocytes) (Figure 7C). We then tested whether the predictions obtained from both modalities could be combined to improve the accuracy of each cell-type quantification. Averaging the predictions obtained from both modalities shows a moderate improvement (Figure 7C), suggesting that the two modalities can complement each other.”

Reviewer #2 (Public Review):

Summary:

The manuscript expands the current bulk sequencing data deconvolution toolkit to include ATAC-seq. The EPIC-ATAC tool successfully predicts accurate proportions of immune cells in bulk tumour samples and EPIC-ATAC seems to perform well in benchmarking analyses. The authors achieve their aim of developing a new bulk ATAC-seq deconvolution tool.

Strengths:

The manuscript describes simple and understandable experiments to demonstrate the accuracy of EPIC-ATAC. They have also been incredibly thorough with their reference dataset collections. The authors have been robust in their benchmarking endeavours and measured EPIC-ATAC against multiple datasets and tools.

Weaknesses:

Currently, the tool has a narrow applicability in that it estimates the percentage of immune cells in a bulk ATAC-seq experiment.

Comments:

(1) Has any benchmarking been done on the runtime of the tool? Although EPIC-ATAC seems to "win" in benchmarking metrics, sometimes the differences are quite small. If EPIC-ATAC takes forever to run, compared to another tool that is a lot quicker, might some people prefer to sacrifice 0.01 in correlation for a quicker running tool?

We thank the reviewer for raising this point that was not addressed in the manuscript. We have added a supplementary figure (Supplementary Figure 8) which represents the CPU time used by each tool. The figure shows that all the tools could be run in less than 20 seconds in average. This figure has been mentioned at the end of the benchmarking paragraphs.

(2) In Figure 3B the data points look a bit squashed in the bottom-left corner. Could the plot be replotted with the data point spread out? There also seems to be some inter-patient variability. Could the authors comment on that?

We have updated Figure 3B to increase the visibility of the dots in the bottom-left corner. To do so, we have limited the x and y axes to the maximum of the predicted proportions for the y axis and true proportions for the x axis.

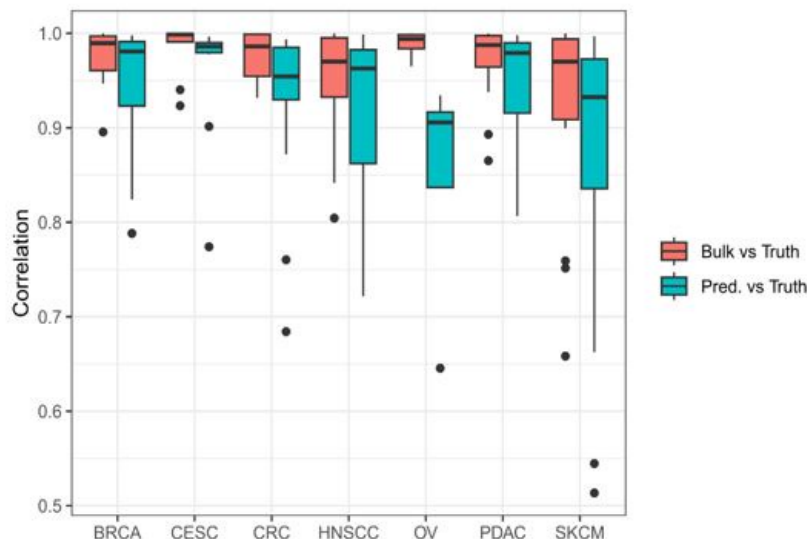
We also acknowledge that the accuracy of the predictions varies across samples. In particular, one sample (Sample4, star shape on Figure 3B) exhibits larger discrepancies between EPIC-ATAC predictions and the ground truth. To understand the lower performance, we have visualized our marker peaks in the five PBMC samples (Figure below). Based on this visualization, we can see that Sample4 might be an outlier sample considering that its cellular composition is similar to that of Sample2 and Sample5, however this sample shows particularly high ATAC-Seq accessibility at the monocytes and dendritic markers. This can explain why EPIC-ATAC overestimates the proportions of the two populations in this case. We have added the previously mentioned figures as a Supplementary Figure (Supplementary Figure 2) and have described it in the results section in the paragraph “EPIC-ATAC accurately estimates immune cell fractions in PBMC ATAC-Seq samples”.

(3) Could the authors comment on the possibility of expanding EPIC-ATAC into more than a percentage prediction tool? Perhaps EPIC-ATAC could remove the immune cell signal from the bulk ATAC-seq data to "purify" the uncharacterised cells in silico, or generate pseudo-ATAC-seq tracks of the identified cell types.

We thank the reviewer for this interesting question. As suggested by the reviewer, one approach to purify bulk genomics data using the cell-type proportions estimated by a cell-type deconvolution tool is to subtract the weighted sum of the signal observed in the reference data, weights corresponding to the predicted proportions. We used this approach on the EPIC-ATAC predictions obtained from pseudobulks built from scATAC-Seq data from diverse cancer types coming from the Human Tumor Atlas Network (HTAN) (See also the answer of the first recommendation of Reviewer 1). This dataset allows us to compare for a relatively large number of samples (a maximum of 25 samples in a cancer type cohort) the purified signal to the true signal derived from the single-cell data. The results are presented in the figure below which shows that the correlations between the predicted and true signals are relatively good in most of the cancer types (blue boxplots). However, these correlation levels are lower than the ones obtained when comparing the signal obtained from the entire pseudobulk (red boxplots) with the true signal. This suggests that this purification approach leads to a signal that is less precise and accurate than the signal resulting from all cells mixtures.

Author response image 1.

Boxplots of the correlation values obtained from the comparison of the bulk signal and the ground truth signal from the uncharacterized cells in each sample (red) and from the comparison of the predicted signal and the ground truth signal from the uncharacterized cells in each sample (blue).



Also, note that in our simple approach, negative values can be obtained. The predicted signal will thus be difficult to interpret and to use in downstream analyses. Methods claiming to perform purification of bulk samples use more complex and dedicated algorithms. For example, Symphony (Burdziak et al., 2019) (cited in our introduction) uses single-cell RNA-Seq data in addition to the bulk chromatin accessibility data to infer cluster-specific accessibility profiles. Considering that EPIC was not designed for purification purposes, we decided not to include this analysis in the updated version of the manuscript.

Recommendations For The Authors:

Reviewer #1 (Recommendations For The Authors):

(1) The original EPIC had two different signatures for application to blood or tumor RNA-seq. It is not clear instead if EPIC-ATAC applies with the same signature and framework to any tissue and disease context. This aspect should be clarified in the text.

We thank the reviewer for raising this point which was not clear in the previous version of the manuscript. As in the original version of EPIC, in EPIC-ATAC two reference profiles and sets of markers are available, the PBMC reference and the TME reference. We used the PBMC reference profiles and markers to deconvolve the PBMC samples and the TME reference profiles and markers to deconvolve the cancer samples. We have clarified this point in the result section of the main text in the paragraph “ATAC-Seq data from sorted cell populations reveal cell-type specific marker peaks and reference profiles” as follows (added text underlined):

“The resulting marker peaks specific only to the immune cell types were considered for the deconvolution of PBMC samples (PBMC markers). For the deconvolution of tumor bulk samples, the lists of marker peaks specific to fibroblasts and endothelial cells were added to the PBMC markers. This extended set of markers was further refined based on the correlation patterns of the markers in tumor bulk samples from the diverse solid cancer types from The Cancer Genome Atlas (TCGA) (Corces et al., 2018), i.e., markers exhibiting the highest correlation patterns in the tumor bulk samples were selected using the findCorrelation function from the caret R package (Kuhn, 2008) (Figure 1, box 4, see the Material and methods, section 2). The latter filtering ensures the relevance of the markers in the TME context since cell-type specific TME markers are expected to be correlated in tumor bulk

ATAC-Seq measurements (Qiu et al., 2021). 716 markers of immune, fibroblasts and endothelial cell types remained after the last filtering (defined as TME markers). Considering the difference in cell types and the different filtering steps applied on the PBMC and TME markers, we recommend to use the TME markers and profiles to deconvolve bulk samples from tumor samples and the PBMC markers and profiles to deconvolve PBMC samples.”

We also note that when running EPIC-ATAC using the PBMC markers and the TME markers independently to perform the deconvolution of the cancer datasets, we see that overall the use of the TME markers leads to a better performance (Figure below).

Figure legend: Correlation and RMSE values obtained when running EPIC-ATAC on each cancer dataset (points) using the PBMC (red) and the TME (blue) markers.

To demonstrate that the TME markers can be applied to different cancer types, we have completed the evaluation of EPIC-ATAC on tumor samples by considering an additional dataset: the Human Tumor Atlas Network (HTAN) single-cell multiomic (scRNA-Seq and scATAC-Seq) dataset. We have processed this dataset and built scATAC-Seq pseudobulks for 7 cancer types on which EPIC-ATAC was applied to. This analysis has been summarized in Figure 4 and Supplementary Figure 4 and shows that EPIC-ATAC is applicable in a diverse set of tissues.

(2) EPIC and EPIC-ATAC have a valuable feature, which is absent from most deconvolution methods: the estimation of unknown content. It would be informative for the users to understand from the benchmarking analysis whether this feature gives an advantage to EPIC-ATAC with respect to the other approaches.

Indeed, among the tools that we included in our benchmarking analysis, only EPIC-ATAC and quanTIseq enable users to predict the proportions of cells that are not present in the reference profiles, i.e., the uncharacterized cells. For the other tools we thus fixed the estimated proportions of uncharacterized cells to 0. This approach provides a clear and significant advantage to EPIC-ATAC and to quanTIseq. For this reason, we also provide a version of the benchmarking in which we exclude the uncharacterized cells and rescale the true and estimated cell-type proportions to sum to 1. In this second benchmarking approach, EPIC-ATAC still outperforms some of the other deconvolution tools.

We have clarified this point in the results section, in the paragraph “EPIC-ATAC accurately predicts fractions of cancer and non-malignant cells in tumor samples”.

(3) The selection of the most discriminative markers is very well described in the text and beautifully illustrated in Figure 2. However, it is unclear why UMAP plots are used to represent cell-type similarities and dissimilarities. Would a linear dimensionality reduction approach like PCA be already sufficient to show these groups, especially considering the not-so-extreme dimensionality of the underlying data? In addition, a statistic that could be also considered to compare clusters to the cell type labels in the two scenarios is the Adjusted Rand Index (ARI).

We thank the reviewer for this relevant comment. We initially used UMAP to facilitate the visualization of the different cell-type groups. However, it is true that the three first axes of the principal component analyses performed based on each set of marker peaks already capture most of the structure in the data and that the use of UMAP can lead to an artificial enhancement of separation between the different groups of cells. We have updated Figure 2B by replacing the UMAP scatter plots by 3D representations of the first three principal components of the PCA and have added in Supplementary Figure 1B the pairwise scatter plots of these first 3 principal components. On the main figures, we have also added the ARI

metric comparing the cell-type annotation and the clustering obtained using the first 10 axes of the PCA and model based clustering.

(4) In the introduction, it is stated that "the reasonable cost and technical advantages of these protocols foreshadow an increased usage of ATAC-Seq in cancer studies". I would suggest adding a reference to justify this trend. Also, it should be discussed how ATAC-seq deconvolution compares to other types of deconvolution approaches applied to cheaper epigenetic data like methylation one (e.g. EPIdish, methylCC, tca, minfi).

We have complemented this sentence with two references to justify the assertion: (i) a review published by Luo, Gribskov and Wang in 2022 showing the increasing number of ATAC-Seq studies in the field of cancer research, and (ii) a protocol paper from Grandi et al. published in 2022 on the state-of-the-art Omni-ATAC protocol for ATAC-seq which discusses the broad applicability and the technical advantages of ATAC-seq. Also in the preceding sentence, a recent ATAC-Seq protocol that can be applied to FFPE samples has been mentioned, FFPE samples being the most common samples in clinical cancer research.

We agree with the reviewer on the fact that other epigenetic assays such as methylation assays are cost effective. However, ATAC-seq provides additional information on the epigenetic landscape of a sample's genome and some questions regarding regulatory regions and transcription factor activity cannot be answered with methylation data. Methods that can be applied on ATAC-Seq data specifically are thus needed. Most of the cell-type deconvolution algorithms existing so far are applicable on RNA-Seq or methylation data. These algorithms often use similar methodological concepts, e.g., linear combination of the reference profiles for reference-based methods, which could be used in different modalities. However, methylation-based deconvolution tools often take as input a data format that is specific to methylation data, e.g., two color micro array data (RGChannelSet R object) for the minfi deconvolution function (estimatesCellCounts) or leverage methylation-specific information to perform the deconvolution. For example, methylCC uses a model based on latent variables representing a binarized measures of the methylation status of cell-type specific regions (1 or 0 for clearly methylated or unmethylated regions). Such methods are more difficult to adapt than tools based on RNA-Seq data where the signal is quantified using read counts similarly to ATAC-Seq data.

Nevertheless, some methods such as EPIdish or MethylCIBERSORT have proposed new methylation reference profiles and have used existing models that are not specific to methylation data to deconvolve the bulk data. In our work, we followed a similar approach where we propose new reference profiles specific to chromatin accessibility data, integrate them to an existing method EPIC as well as test them in other existing tools. Note that methylation reference profiles cannot be directly used for ATAC-Seq data deconvolution considering that methylation measures methylation status at CpG sites (dinucleotides) and ATAC-Seq measures the accessibility of regions of hundreds base pairs.

An analysis comparing the performance of methylation-based deconvolution and ATAC-Seq based deconvolution would be informative. However, such analysis is beyond the scope of our paper considering that none of the datasets used for our benchmarking provide these two modalities for the same samples.

In the manuscript, we have completed the references associated to the methylation-based deconvolution tools with the ones mentioned in the previous paragraphs and by the reviewer and have completed the discussion as follows:

"The comparison of EPIC-ATAC applied on ATAC-Seq data with EPIC applied on RNA-Seq data has shown that both modalities led to similar performances and that they could complement each other. Another modality that has been frequently used in the context of bulk sample deconvolution is methylation. Methylation profiling techniques such as methylation arrays

are cost effective (Kaur et al., 2023) and DNA methylation signal is highly cell-type specific (Kaur et al., 2023; Loyfer et al., 2023). Considering that methylation and chromatin accessibility measure different features of the epigenome, additional analyses comparing and/or complementing ATAC-seq based deconvolution with methylation-based deconvolution could be of interest as future datasets profiling both modalities in the same samples become available.”

(5) In the Results section, some methodological steps could be phrased in a bit more extensive way to let the reader understand the rationale and the actual approach. I recognize there is also a reference to the Methods section, where all methodologies are reported in detail, but some of the sentences are hard to understand due to their synthetic format, e.g.: "markers with potential residual accessibility in human tissues were then filtered out".

We thank the reviewer for this comment and we have followed his recommendation to expand sentences with a synthetic format. Text changes and additions are underlined below:

“To limit batch effects, the collected samples were homogeneously processed from read alignment to peak calling. For each cell type, we derived a set of stable peaks observed across samples and studies, i.e. for each study, peaks detected in at least half of the samples were considered, and for each cell type, only peaks detected jointly in all studies were kept (see Materials and Methods, section 1).”

“To filter out markers that could be accessible in other human cell-types than those included in our reference profiles, we used the human atlas study (K. Zhang et al., 2021), which identified modules of open chromatin regions accessible in a comprehensive set of human tissues, and we excluded from our marker list the markers overlapping these modules (Figure 1, box 3, see Materials and Methods section 2).”

“For the deconvolution of tumor bulk samples, the lists of marker peaks specific to fibroblasts and endothelial cells were added to the PBMC markers. This extended set of markers was further refined based on the correlation patterns of the markers in tumor bulk samples from the diverse solid cancer types from The Cancer Genome Atlas (TCGA) (Corces et al., 2018), i.e., markers exhibiting the highest correlation patterns in the tumor bulk samples were selected using the findCorrelation function from the caret R package (Kuhn, 2008) (Figure 1, box 4, see the Material and methods, section 2).”

Also, following the comments and recommendations of the Reviewer 1, we have: (i) moved the method section describing the adaptation of EPIC to ATACseq data to provide more details in the results section (see answer to the third comment of Reviewer 1), (ii) clarified how the existing tools used in the benchmarking analyses were adapted for ATAC-Seq deconvolution (see answer to the second comment of Reviewer 1), and (iii) detailed how the comparison between our estimations of the infiltration levels in the samples from Kumegawa et al. and the estimations from the original study was performed (see answer to the seventh recommendation of Reviewer 1).

(6) In the main text, it is stated that "the list of markers was further refined based on the correlation patterns of the markers in tumor bulk samples from diverse cancer types from The Cancer Genome Atlas". It should be clarified if these are only solid cancers, or if blood cancers were also used.

We have considered only the solid cancers and have clarified this point in the results section: “This extended set of markers was further refined based on the correlation patterns of the markers in tumor bulk samples from the diverse solid cancer types from The Cancer Genome Atlas”.

(7) When reporting that "these predictions are consistent with the infiltration level estimations reported in the original publication", it should be mentioned how the infiltration levels were quantified in this publication and how this agreement was quantified. This would be important also to claim in the abstract that "EPIC-ATAC accurately infers the immune contexture of the main breast cancer subtypes".

We thank the reviewer for this comment, we acknowledge that the agreement between the EPIC-ATAC predictions and the infiltration levels quantified in the original publication should be further described in the paper. We have expanded the text in the results section in the paragraph "EPIC-ATAC accurately infers the immune contexture in a bulk ATAC-Seq breast cancer cohort" to clarify this point. Additionally, we have added a panel in Figure 6 (panel A) which shows a good agreement between EPIC-ATAC predictions and the metric used in the original paper to evaluate the infiltration levels of different cell types.

The added text is underlined below:

"We applied EPIC-ATAC to a breast cancer cohort of 42 breast ATAC-Seq samples including samples from two breast cancer subtypes, i.e., 35 oestrogen receptor (ER)-positive human epidermal growth factor receptor 2 (HER2)-negative (ER+/HER2-) samples and 7 triple negative (TNBC) tumors (Kumegawa et al., 2023). No cell sorting was performed in parallel to the chromatin accessibility sequencing. For this reason, the authors used a set of cell-type-specific cis-regulatory elements (CREs) identified in scATAC-Seq data from similar breast cancer samples (Kumegawa et al., 2022) and estimated the amount of infiltration of each cell type by averaging the ATAC-Seq signal of each set of cell-type-specific CREs in their samples. We used EPIC-ATAC to estimate the proportions of different cell types of the TME. These predictions were then compared to the metric used by Kumegawa and colleagues in their study to infer levels of infiltration. A high correlation between the two metrics was observed for each cell type (Pearson's correlation coefficient from 0.5 for myeloid cells to 0.94 for T cells, Figure 6A)."

(8) It should be made explicit if EPIC-ATAC quantifies mDC, pDC, or their sum.

In our collection of reference ATAC-Seq samples from which the markers and profiles have been derived, mDCs and pDCs were both included in the dendritic cells. EPIC-ATAC thus quantifies the total amount of dendritic cells, i.e., mDCs and pDCs included. We have added a sentence in the main text to clarify this point:

To identify robust chromatin accessibility marker peaks of cancer relevant cell types, we collected 564 samples of sorted cell populations from twelve studies including eight immune cell types (B cells [...] dendritic cells (DCs) (mDCs and pDCs are grouped in this cell-type category) [...] and endothelial (Liu et al., 2020; Xin et al., 2020) cells (Figure 1 box 1, Figure 2A, Supplementary Table 1).

Reviewer #2 (Recommendations For The Authors):

The authors should double-check the naming of tools is done correctly e.g. ChIPSeeker has been spelled incorrectly in some instances throughout the manuscript.

We thank the reviewer for pointing out this mistake and have corrected the mistake in the main text.

<https://doi.org/10.7554/eLife.94833.2.sa0>