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

Aurélie AG Gabriel[1,2,3,4], Julien Racle[1,2,3,4], Maryline Falquet[3,5,6,7], Camilla Jandus[3,5,6,7], David Gfeller[1,2,3,4,*]

Affiliations:

[1] Department of Oncology, Ludwig Institute for Cancer Research, University of Lausanne, Lausanne, Switzerland

[2] Agora Cancer Research Centre, Lausanne, Switzerland

[3] Swiss Cancer Center Leman (SCCL), Switzerland

[4] Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland.

[5] Ludwig Institute for Cancer Research, Lausanne Branch, Lausanne, Switzerland

[6] Department of Pathology and Immunology Faculty of Medicine, University of Geneva, Geneva, Switzerland

[7] Geneva Center for Inflammation Research, Geneva, Switzerland

* Corresponding author: david.gfeller@unil.ch

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 all major cancer-relevant cell types. We then capitalize on the EPIC deconvolution framework (Racle et al. 2017) previously shown to accurately predict cell-type composition in tumor bulk RNA-Seq data and integrate our markers and reference profiles to EPIC 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.

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 (K. Zhang et al. 2021; Klemm, Shipony, and Greenleaf 2019). 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. The reasonable cost and technical advantages of these protocols foreshadow an increased usage of ATAC-Seq in cancer studies.

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, in particular the presence of specific immune cell types, 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 including information from both cancer and non-malignant cells. Precisely quantifying the proportions of different cell types in such samples represents therefore 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 sequencing 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 (Avila Cobos et al. 2018; 2020; Sturm et al. 2019; Racle et al. 2017; Monaco et al. 2019; Newman et al. 2019; H. Li et al. 2020; Finotello et al. 2019; Becht et al. 2016). A large number of these tools model bulk data as a mixture of reference profiles identified in purified cell populations for each cell type. The accuracy of the predictions of cell-type proportions relies on the quality of these reference profiles as well as on the use of cell-type specific 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 (here referred to as 'uncharacterized' cells). In the context of cancer 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). A few tools consider uncharacterized cells in their deconvolution framework by using cell-type specific markers not expressed in the uncharacterized cells (Clarke, Seol, and Clarke 2010; Gosink, Petrie, and Tsinoremas 2007; 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 initially developed for transcriptomic data (RNA-Seq data) (Newman et al. 2015; Racle et al. 2017; Finotello et al. 2019; Monaco et al. 2019; Newman et al. 2019; T. Li et al. 2020; Jimenez-Sanchez, Cast, and Miller 2019; Gong and Szustakowski 2013). More recently they have been adapted for other omics layers such as methylation (Chakravarthy et al. 2018; Teschendorff et al. 2020; Arneson, Yang, and Wang 2020; H. Zhang et al. 2021) and proteomics (Feng et al. 2023) or chromatin accessibility. For the latter, a specific framework called DeconPeaker (H. Li et al. 2020) was developed to estimate cell-type proportions from bulk samples. Deconvolution tools developed initially for other omics modalities, such as RNA-Seq, can also be applied on ATAC-Seq if appropriate ATAC-Seq profiles are provided to the tool. For example, the popular deconvolution tool, CIBERSORT (Newman et al. 2015), was used to deconvolve leukemic ATAC-Seq samples (Corces et al. 2016). Other methods have been proposed to decompose ATAC-Seq bulk profiles into subpopulation-specific profiles (Zeng et al. 2019; Burdziak et al. 2019) or compartments (Peng et al. 2019). However, these methods have more requisites: (i) the integration of the ATAC-Seq data with single-cell or bulk RNA-Seq (Zeng et al. 2019; Burdziak et al. 2019) and HIChIP data (Zeng et al. 2019) or, (ii) subsequent feature annotation to associate compartments with cell types or biological processes (Peng et al. 2019).

The application of existing bulk ATAC-Seq data deconvolution tools to solid tumors is limited. 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; H. Li et al. 2020). 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), not all TME-relevant cell types are covered (*e.g.*, lack of scATAC-Seq data from neutrophils due to extracellular traps formation). Also, these markers have 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) framework to perform bulk ATAC-Seq samples deconvolution (Figure 1). 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 with similar or higher performances than other existing tools.

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 cancer relevant cell types, we collected 564 samples of sorted cell populations from twelve studies including eight immune cell types (B cells (Calderon et al. 2019; Corces et al. 2016; P. Zhang et al. 2022), CD4+ T cells (Corces et al. 2016; Liu et al. 2020; P. Zhang et al. 2022; Mumbach et al. 2017; Giles et al. 2022), CD8+ T cells (Calderon et al. 2019; Corces et al. 2016; Liu et al. 2020; P. Zhang et al. 2022; Giles et al. 2022), natural killer (NK) cells (Calderon et al. 2019; Corces et al. 2016), dendritic cells (DCs) (Calderon et al. 2019; Leylek et al. 2020; Liu et al. 2020), macrophages (Liu et al. 2020; P. Zhang et al. 2022), monocytes (Calderon et al. 2019; Corces et al. 2016; Leylek et al. 2020; P. Zhang et al. 2022; Trizzino et al. 2021) and neutrophils (Ram-Mohan et al. 2021; Perez et al. 2020), as well as fibroblasts (Ge et al. 2021; Liu et al. 2020) and endothelial (Liu et al. 2020; Xin et al. 2020) cells (Figure 1 box 1, Figure 2A, 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 of stable peaks, *i.e.,* peaks observed across samples and studies (see Materials and Methods).

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 1). 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). Using the human atlas study (K. Zhang et al. 2021), markers with potential residual accessibility in human tissues were then filtered out (Figure 1, box 3, see Materials and Methods). The resulting marker peaks specific to the immune cell types were considered for the deconvolution of PBMC samples (PBMC markers). For tumor bulk sample deconvolution, 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 (TCGA) (Corces et al. 2018) (Figure 1, box 4, see the Material and methods). 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 later filtering and were considered for the deconvolution of bulk tumor samples (TME markers).

To assess the quality and reproducibility of these markers, we 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). Two-dimensional UMAP representations of the samples confirmed this observation (Figure 2B). 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 (Racle et al. 2017) (see Material and methods). 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 suggest 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.

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 SLCO2B1 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

Table 1: *List of nearest genes and enriched CBPs reported in the PanglaoDB or CellMarker databases.*

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

The cell-type specific marker peaks and profiles derived from the reference samples were integrated to 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 test the accuracy of EPIC-ATAC predictions, we first collected PBMCs from five healthy donors. In 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).

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 (Satpathy et al. 2019; Granja 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. DeconPeaker and CIBERSORTx include the option to define cell-type specific markers and profiles from a set of reference samples. We thus 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 2A-C). The fact that our marker peaks and reference profiles could be used with EPIC-ATAC and other existing tools demonstrates their broad applicability.

Predictions accuracies were also evaluated in each cell type separately. Since the number of samples was low in each dataset, samples from both datasets were combined for this analysis. EPIC-ATAC demonstrated good accuracies across cell types with RMSE values ranging from 0.02 for B cells to 0.13 for NK cells (Supplementary Figure 3). As expected, predictions with all tools were more accurate for frequent cell types with well-characterized markers (*e.g.,* CD8/CD4 T cells, B cells) compared to less frequent cell types (*e.g.,* NK cells, dendritic cells) (Supplementary Figure 2 and 3). Note that MCPcounter is a marker-based method that derives cell-type specific scores which cannot be compared between cell types. This method was thus only included in the benchmark considering each cell type separately.

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). We generated two pseudobulk datasets by averaging the chromatin accessibility signal across all cells of each sample (see Material and methods). Applying EPIC-ATAC to both datasets shows that this framework is able to simultaneously predict the proportions of both uncharacterized cells and immune, stromal and vascular cells (Figure 4A). In these cancer samples, the proportion of uncharacterized cells can be seen as a proxy of the proportion of cancer cells.

As for the PBMC datasets, we compared EPIC-ATAC performances to other existing deconvolution tools. For both datasets, EPIC-ATAC led to the highest performances and was the only method to accurately predict the proportion of uncharacterized cells (Figure 4B, Supplementary Figure 4 and 5). 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 Figure 4).

In the EPIC-ATAC and quanTIseq frameworks, predictions correspond to absolute cell-type fraction, *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). We thus 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). EPIC-ATAC outperformed most of the other methods also when excluding the uncharacterized cells (Figure 4C, Supplementary Figure 4 and 5).

Supplementary Figure 6 reports the performances of each tool when considering each cell type separately. Overall, EPIC-ATAC showed comparable or higher correlation and lower RMSE values when compared to the other deconvolution tools.

T cell subtypes quantification reveals the ATAC-Seq deconvolution limits for closely related cell types.

To explore the limitations of ATAC-Seq deconvolution, we next 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 naïve, helper/memory and T regulatory CD4+ T cells. 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 7A) and observed that the markers were conserved in external data (Supplementary Figure 7B). 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 prediction 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 5A). In particular, low accuracies were obtained for helper/memory CD4+ and naïve T-cell subtypes (Figure 5B). Similar results were obtained using other deconvolution tools (Supplementary Figure 8).

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-) samples and 7 triple negative (TN) tumors (Kumegawa et al. 2023). No cell sorting was performed in parallel to the chromatin accessibility sequencing. We thus used EPIC-ATAC to estimate cell-type proportions. We observed a higher proportion of T cells, B cells, NK cells and macrophages in the TN samples in comparison to ER+/HER2- samples (Figure 6A).

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 predictions are consistent with the infiltration level estimations reported in the original publication, although no differences in macrophages infiltration was observed between the ER+/HER2- subgroups in our case (Kumegawa et al. 2023).

EPIC-ATAC performs similarly to EPIC RNA-seq based deconvolution and better than gene activity based deconvolution

We finally compared the accuracy of EPIC when applied on ATAC-Seq data and on RNA-Seq data. For this purpose, we used the 10X multiome PBMC dataset (10x Genomics 2021) which provides for each cell both its chromatin accessibility profile and its gene expression profile and simulated 100 pseudobulks with diverse cellular compositions (see Material and methods). 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, based on peak distances to each gene, into gene activity (GA) variables as proxy for gene expression. We thus applied the GA transformation to the 10x multiome PBMC dataset and performed GA-based RNA deconvolution using the original EPIC tool (See Material and methods).

Figure 7 shows 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.

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 cancer-relevant cell types from diseased, stimulated and healthy samples. 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. To our knowledge this work provides the first benchmark of deconvolution tools adapted to ATAC-Seq data in the context of solid tumor samples. 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 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.

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 deviates from true bulk data and errors in the true cell-type proportions. For this reason, we anticipate that the newly generated benchmarking PBMCs dataset with ground truth cell proportions obtained by flow cytometry will nicely complement pseudobulk from scATAC-Seq data in future benchmarks of ATAC-Seq deconvolution. The qualitative evaluation of our method on true bulk ATAC-Seq samples from breast cancer patients and the observation of similar immune compositions in TN and ER+/HER2- samples as the ones identified in the original paper (Figure 6) further support the accuracy of EPIC-ATAC to deconvolve bulk ATAC-Seq data, without requiring additional scATAC-Seq data which are not always available for all cancer types.

Overall the evaluation of the EPIC-ATAC deconvolution resulted in an average absolute error of 7% 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). While the estimated proportions of these populations by EPIC-ATAC are low (*e.g.,* dendritic cells), comparing such estimations across samples should be performed with care due to the uncertainty of the predictions.

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 (Figure 5 and Supplementary figure 8). We thus recommend to use the EPIC-ATAC framework using the markers and reference profiles based on the major cell 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 (GSEA, (Subramanian et al. 2005)). Such 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).

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; Javaid 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 (Sanseviero 2019; Bonnema et al. 1994). 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 all major cancer-relevant cell types. 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

Generation of an ATAC-Seq reference dataset of cancer relevant cell types.

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 (Calderon et al. 2019; Corces et al. 2016; P. Zhang et al. 2022), CD4+ (Corces et al. 2016; Liu et al. 2020; P. Zhang et al. 2022; Mumbach et al. 2017; Giles et al. 2022), CD8+ (Calderon et al. 2019; Corces et al. 2016; Liu et al. 2020; P. Zhang et al. 2022; Giles et al. 2022), natural killer (NK) (Calderon et al. 2019; Corces et al. 2016), 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 (Calderon et al. 2019; Corces et al. 2016; Leylek et al. 2020; P. Zhang et al. 2022; Trizzino et al. 2021) and neutrophils (Ram-Mohan et al. 2021; Perez et al. 2020) as well as fibroblasts (Ge et al. 2021; Liu et al. 2020) 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) was used to process the raw fastq files based on the following tool: 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 (Y. Zhang et al. 2008) (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).

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)). 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), 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), 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) to transform the data and model the meanvariance 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). 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 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 two-dimensional space running Uniform Manifold Approximation (UMAP) based on the PBMC and TME markers (Figure 2B). We also run PCA and 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.

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 ChipEnrich 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. ChipEnrich 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 Remap 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.

Running EPIC-ATAC on bulk ATAC-Seq data

The samples used to generate the reference profiles were aligned using the hg38 reference genome. To assure 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 (use of the liftOver R package). 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. In RNA-Seq based deconvolution, EPIC uses an estimation of the amount of mRNA in each reference cell type to derive cell proportions. For the ATAC-Seq based deconvolution these values were set to 1 to give similar weights to all cell-types quantifications.

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, Biolgend), 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 Celc9a (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 5 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) (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) (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). 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). 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) (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.

For Basal cell carcinoma and Gynecological cancer 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), 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 and Gynecological cancer 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. 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 and Gynecological cancer 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). 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). As for the previously mentioned datasets, raw counts were normalized using the TPM-like transformation prior to bulk deconvolution.

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) 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 (parameters: scaling set to 1 for all cell types and method set to "lsei").
- DeconPeaker (H. Li et al. 2020) 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_cust." 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 3 and 5. Deconvolution was run using the deconvolution module deconPeaker (using findctsps 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_cust." 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 immunedeconv R package (Sturm et al. 2019) was used 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.

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. 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 sampling 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 form the 10X dataset 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.

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 (doi: 10.5281/zenodo.8431792). The other data related to this work are available in the supplementary tables and on the Zenodo deposit (doi: 10.5281/zenodo.8431792).

Competing interests:

The authors declare that they have no competing interests.

Acknowledgement:

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.

Authors contributions:

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

Figure legends:

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 cancer-relevant cell populations. 2) Cell-type specific marker peaks were identified using differential accessibility analysis. 3) Markers with previously observed chromatin accessibility in human healthy tissues were then 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*

finally integrated in the EPIC-ATAC framework to perform bulk ATAC-Seq deconvolution. Parts of this figure were created with BioRender.com.

Figure 2: ATAC-Seq data from sorted cell populations reveal cell-type specific marker peaks and reference profiles. A) *Number of samples collected for each cell type. The colors correspond to the different studies of origin.* ***B)*** *Representation of the collected samples in 2D using UMAP based on the PBMC markers (left) and TME markers (right). Colors correspond to cell types.* ***C)*** *Scaled averaged chromatin accessibility of the cell-type specific marker peaks (rows) in each cell type (columns) in the ATAC-Seq reference samples used to identify the marker peaks.* ***D)*** *Scaled averaged chromatin accessibility of the marker peaks in external ATAC-Seq data from samples of pure cell types excluded from the reference samples (see Material and Methods).* ***E)*** *Scaled averaged chromatin accessibility of the marker peaks in an external scATAC-Seq dataset (Human Atlas* (K. Zhang et al. 2021)*).* ***F)*** *Distribution of the marker peak distances to the nearest transcription start site (TSS) (left panel) and the ChiPSeeker annotations (right panel).* ***G)*** *Significance (-log10(q.value)) of pathways (columns) enrichment test obtained using ChIP-Enrich on each set of cell-type specific marker peaks (rows). A subset of relevant enriched pathways is represented. Colors of the names of the pathways correspond to cell types where the pathways were found to be enriched. When pathways were significantly enriched in more than one set of peaks, pathways names are written in bold.*

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. Parts of this figure (panel 1) were created with BioRender.com.*

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) and gynecological (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 predictions accuracy. The predicted and true proportions of the immune, stromal and vascular cell types were rescaled to sum to 1.*

Figure 5: T cell subtypes quantification reveals the ATAC-Seq deconvolution limits for closely related cell types. A) *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 subtype predictions using the reference profiles including the T-cell subtypes.* ***B)*** *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 the samples stratified based on two breast cancer subtypes.* ***B)*** *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.*

Figure 7: EPIC-ATAC performs similarly to EPIC RNA-seq based deconvolution and better than gene activity based deconvolution. *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 the 100 pseudobulks simulated form the 10x multiome PBMC dataset* (10x Genomics 2021). *Dots correspond to outlier pseudobulks.*

Supplementary Figures: Additional file named Supplementary_figures.pdf

Supplementary Tables:

Sup. Table 1: Metadata of the ATAC-Seq samples used in the study

Sup. Table 2: Averaged chromatin accessibility of the PBMC marker peaks in each cell-type.

Sup. Table 3: Averaged chromatin accessibility of the TME marker peaks in each cell-type.

Sup. Table 4: Annotations of the cell-type specific PBMC marker peaks

Sup. Table 5: Annotations of the cell-type specific TME marker peaks

Sup. Table 6: GO pathways enriched in each set of cell-type specific PBMC marker peaks

Sup. Table 7: GO pathways enriched in each set of cell-type specific TME marker peaks

Sup. Table 8: Averaged chromatin accessibility of the PBMC marker peaks in each cell-type (T cells subtypes included).

Sup. Table 9: Averaged chromatin accessibility of the TME marker peaks in each cell-type (T cells subtypes included).

Sup. Table 10: Annotations of the cell-type specific PBMC marker peaks (T cells subtypes included).

Sup. Table 11: Annotations of the cell-type specific TME marker peaks (T cells subtypes included).

Sup. Table 12: GO pathways enriched in each set of cell-type specific PBMC marker peaks (T cell subtypes).

Sup. Table 13: GO pathways enriched in each set of cell-type specific TME marker peaks (T cell subtypes).

References:

10x Genomics. 2021. “PBMC from a Healthy Donor - Granulocytes Removed Through Cell Sorting (10k).” 2021. https://www.10xgenomics.com/resources/datasets/pbmc-from-a-healthy-donor-granulocytes-removed-through-cell-sorting-10-k-1-standard-2-0-0.

Arneson, Douglas, Xia Yang, and Kai Wang. 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.

Avila Cobos, Francisco, José Alquicira-Hernandez, Joseph E. Powell, Pieter Mestdagh, and Katleen De Preter. 2020. “Benchmarking of Cell Type Deconvolution Pipelines for Transcriptomics Data.” *Nature Communications* 11 (1): 1–14. https://doi.org/10.1038/s41467-020-19015-1.

Avila Cobos, Francisco, Jo Vandesompele, Pieter Mestdagh, and Katleen De Preter. 2018. “Computational Deconvolution of Transcriptomics Data from Mixed Cell Populations.” *Bioinformatics* 34 (11): 1969–79. https://doi.org/10.1093/BIOINFORMATICS/BTY019.

Becht, Etienne, Nicolas A. Giraldo, Laetitia Lacroix, Bénédicte Buttard, Nabila Elarouci, Florent Petitprez, Janick Selves, et al. 2016. “Estimating the Population Abundance of Tissue-Infiltrating Immune and Stromal Cell Populations Using Gene Expression.” *Genome Biology* 17 (October): 218. https://doi.org/10.1186/S13059-016-1070-5/TABLES/4.

Bonnema, Joy D., Larry M. Karnitz, Renee A. Schoon, Robert T. Abraham, and Paul J. Leibson. 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 (4): 1427–35. https://doi.org/10.1084/JEM.180.4.1427.

Buenrostro, Jason D., Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf. 2013. “Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position.” *Nature Methods* 10 (12): 1213–18. https://doi.org/10.1038/nmeth.2688.

Burdziak, Cassandra, Elham Azizi, Sandhya Prabhakaran, and Dana Pe'er. 2019. "A Nonparametric Multi-View Model for Estimating Cell Type-Specific Gene Regulatory Networks." *ArXiv*.

Calderon, Diego, Michelle L.T. Nguyen, Anja Mezger, Arwa Kathiria, Fabian Müller, Vinh Nguyen, Ninnia Lescano, et al. 2019. "Landscape of Stimulation-Responsive Chromatin across Diverse Human Immune Cells." *Nature Genetics* 51 (10): 1494–1505. https://doi.org/10.1038/s41588-019-0505-9.

Carvalho, Klebea, Elisabeth Rebboah, Camden Jansen, Katherine Williams, Andrew Dowey, Cassandra McGill, and Ali Mortazavi. 2021. "Uncovering the Gene Regulatory Networks Underlying Macrophage Polarization Through Comparative Analysis of Bulk and Single-Cell Data." *BioRxiv*, January. https://doi.org/10.1101/2021.01.20.427499.

Castro-Mondragon, Jaime A., Rafael Riudavets-Puig, Ieva Rauluseviciute, Roza Berhanu Lemma, Laura Turchi, Romain Blanc-Mathieu, Jeremy Lucas, et al. 2022. "JASPAR 2022: The 9th Release of the Open-Access Database of Transcription Factor Binding Profiles." *Nucleic Acids Research* 50 (D1): D165–73. https://doi.org/10.1093/NAR/GKAB1113.

Chakravarthy, Ankur, Andrew Furness, Kroopa Joshi, Ehsan Ghorani, Kirsty Ford, Matthew J. Ward, Emma V. King, et al. 2018. "Pan-Cancer Deconvolution of Tumour Composition Using DNA Methylation." *Nature Communications* 9 (August). https://doi.org/10.1038/s41467-018-05570-1.

Clarke, Jennifer, Pearl Seol, and Bertrand Clarke. 2010. "Statistical Expression Deconvolution from Mixed Tissue Samples." *Bioinformatics* 26 (8): 1043–49. https://doi.org/10.1093/BIOINFORMATICS/BTQ097.

Corces, M. Ryan, Jason D. Buenrostro, Beijing Wu, Peyton G. Greenside, Steven M. Chan, Julie L. Koenig, Michael P. Snyder, et al. 2016. "Lineage-Specific and Single-Cell Chromatin Accessibility Charts Human Hematopoiesis and Leukemia Evolution." *Nature Genetics* 48 (August): 1193–1203. https://doi.org/10.1038/ng.3646.

Corces, M. Ryan, Jeffrey M. Granja, Shadi Shams, Bryan H. Louie, Jose A. Seoane, Wanding Zhou, Tiago C. Silva, et al. 2018. “The Chromatin Accessibility Landscape of Primary Human Cancers.” *Science* 362 (6413). https://doi.org/10.1126/science.aav1898.

Corces, M. Ryan, Alexandro E. Trevino, Emily G. Hamilton, Peyton G. Greenside, Nicholas A. Sinnott-Armstrong, Sam Vesuna, Ansuman T. Satpathy, et al. 2017. “An Improved ATAC-Seq Protocol Reduces Background and Enables Interrogation of Frozen Tissues.” *Nature Methods* 14 (10): 959–62. https://doi.org/10.1038/nmeth.4396.

Cusanovich, Darren A., Riza Daza, Andrew Adey, Hannah A. Pliner, Lena Christiansen, Kevin L. Gunderson, Frank J. Steemers, Cole Trapnell, and Jay Shendure. 2015. “Multiplex Single-Cell Profiling of Chromatin Accessibility by Combinatorial Cellular Indexing.” *Science* 348 (6237): 910–14. https://doi.org/10.1126/SCIENCE.AAB1601/SUPPL_FILE/PAP.PDF.

Feng, Song, Anna Calinawan, Pietro Pugliese, Pei Wang, Michele Ceccarelli, Francesca Petralia, and Sara J C Gosline. 2023. “Decomprolute : A Benchmarking Platform Designed for Multiomics-Based Tumor Deconvolution.” *BioRxiv*.

Finotello, Francesca, Clemens Mayer, Christina Plattner, Gerhard Laschober, DIetmar Rieder, Hubert Hackl, Anne Krogsdam, et al. 2019. “Molecular and Pharmacological Modulators of the Tumor Immune Contexture Revealed by Deconvolution of RNA-Seq Data.” *Genome Medicine* 11 (May): 34. https://doi.org/10.1186/s13073-019-0638-6.

Franzén, Oscar, Li Ming Gan, and Johan L.M. Björkegren. 2019. “PanglaoDB: A Web Server for Exploration of Mouse and Human Single-Cell RNA Sequencing Data.” *Database* 2019: baz046. https://doi.org/10.1093/DATABASE/BAZ046.

Fridman, Wolf H., Franck Pagès, Catherine Sautès-Fridman, and Ǹrôme Galon. 2012. “The Immune Contexture in Human Tumours: Impact on Clinical Outcome.” *Nature Reviews Cancer* 12 (4): 298–306. https://doi.org/10.1038/nrc3245.

Fridman, Wolf H., Laurence Zitvogel, Catherine Sautès-Fridman, and Guido Kroemer. 2017. “The Immune Contexture in Cancer Prognosis and Treatment.” *Nature Reviews Clinical Oncology*. Nature Publishing Group. https://doi.org/10.1038/nrclinonc.2017.101.

Ge, Xiangyu, Mojca Frank-Bertoncelj, Kerstin Klein, Amanda McGovern, Tadeja Kuret, Miranda Houtman, Blaž Burja, et al. 2021. “Functional Genomics Atlas of Synovial Fibroblasts Defining Rheumatoid Arthritis Heritability.” *Genome Biology* 22 (1): 247. https://doi.org/10.1186/S13059-021-02460-6/FIGURES/7.

Geng, Degui, Liqin Zheng, Ratika Srivastava, Nicole Asprodites, Cruz Velasco-Gonzalez, and Eduardo Davila. 2010. “When Toll-like Receptor and T-Cell Receptor Signals Collide: A Mechanism for Enhanced CD8 T-Cell Effector Function.” *Blood* 116 (18): 3494–3504. https://doi.org/10.1182/BLOOD-2010-02-268169.

Giles, Josephine R., Sasikanth Manne, Elizabeth Freilich, Derek A. Oldridge, Amy E. Baxter, Sangeeth George, Zeyu Chen, et al. 2022. “Human Epigenetic and Transcriptional T Cell Differentiation Atlas for Identifying Functional T Cell-Specific Enhancers.” *Immunity* 55 (3): 557-574.e7. https://doi.org/10.1016/J.IMMUNI.2022.02.004.

Gong, Ting, and Joseph D. Szustakowski. 2013. “DeconRNASeq: A Statistical Framework for Deconvolution of Heterogeneous Tissue Samples Based on MRNA-Seq Data.” *Bioinformatics* 29 (8): 1083–85. https://doi.org/10.1093/BIOINFORMATICS/BTT090.

Gosink, Mark M., Howard T. Petrie, and Nicholas F. Tsinoremas. 2007. “Electronically Subtracting Expression Patterns from a Mixed Cell Population.” *Bioinformatics* 23 (24): 3328–34. https://doi.org/10.1093/BIOINFORMATICS/BTM508.

Grandi, Fiorella C., Hailey Modi, Lucas Kampman, and M. Ryan Corces. 2022. “Chromatin Accessibility Profiling by ATAC-Seq.” *Nature Protocols*, April, 1518–52. https://doi.org/10.1038/s41596-022-00692-9.

Granja, Jeffrey M., M. Ryan Corces, Sarah E. Pierce, S. Tansu Bagdatli, Hani Choudhry, Howard Y. Chang, and William J. Greenleaf. 2021. “ArchR Is a Scalable Software Package for Integrative Single-Cell Chromatin Accessibility Analysis.” *Nature Genetics* 53 (February): 403–11. https://doi.org/10.1038/s41588-021-00790-6.

Granja, Jeffrey M., Sandy Klemm, Lisa M. McGinnis, Arwa S. Kathiria, Anja Mezger, M. Ryan Corces, Benjamin Parks, et al. 2019. “Single-Cell Multiomic Analysis Identifies Regulatory Programs in Mixed-Phenotype Acute Leukemia.” *Nature Biotechnology* 37 (12): 1458–65. https://doi.org/10.1038/s41587-019-0332-7.

Hammal, Fayrouz, Pierre De Langen, Aurelie Bergon, Fabrice Lopez, and Benoit Ballester. 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 (D1): D316–25. https://doi.org/10.1093/NAR/GKAB996.

Hu, Congxue, Tengyue Li, Yingqi Xu, Xinxin Zhang, Feng Li, Jing Bai, Jing Chen, 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 (D1): D870–76. https://doi.org/10.1093/NAR/GKAC947.

Javaid, Nasir, and Sangdun Choi. 2020. “Toll-like Receptors from the Perspective of Cancer Treatment.” *Cancers* 12 (2): 297. https://doi.org/10.3390/CANCERS12020297.

Jiang, Yijia, Zhirui Hu, Junchen Jiang, Alexander Zhu, Yi Zhang, Allen W. Lynch, Yingtian Xie, et al. 2023. “ScATAnno: Automated Cell Type Annotation for Single-Cell ATAC Sequencing Data.” *BioRxiv*, June. https://doi.org/10.1101/2023.06.01.543296.

Jimenez-Sanchez, Alejandro, Oliver Cast, and Martin L. Miller. 2019. “Comprehensive Benchmarking and Integration of Tumor Microenvironment Cell Estimation Methods.” *Cancer Research* 79 (24): 6238–46. https://doi.org/10.1158/0008-5472.CAN-18-3560.

Jin, Haijing, and Zhandong Liu. 2021. “A Benchmark for RNA-Seq Deconvolution Analysis under Dynamic Testing Environments.” *Genome Biology* 22 (April): 102. https://doi.org/10.1186/s13059-021-02290-6.

Kalafati, Lydia, Ioannis Kourtzelis, Jonas Schulte-Schrepping, Xiaofei Li, Aikaterini Hatzioannou, Tatyana Grinenko, Eman Hagag, et al. 2020. “Innate Immune Training of Granulopoiesis Promotes Anti-Tumor Activity.” *Cell* 183 (3): 771-785.e12. https://doi.org/10.1016/J.CELL.2020.09.058.

Klemm, Sandy L., Zohar Shipony, and William J. Greenleaf. 2019. "Chromatin Accessibility and the Regulatory Epigenome." *Nature Reviews Genetics* 20 (4): 207–20. https://doi.org/10.1038/s41576-018-0089-8.

Kuhn, Max. 2008. "Building Predictive Models in R Using the Caret Package." *Journal of Statistical Software* 28 (5): 1–26. https://doi.org/10.18637/JSS.V028.I05.

Kumegawa, Kohei. 2023. "ATAC-Seq Data of 42 BC Samples as SummarizedExperiment Object with Count Matrix, Normalized Count Matrix, Peak Info, and Clinical Info." 2023. https://doi.org/10.6084/m9.figshare.21992609.v1.

Kumegawa, Kohei, Sumito Saeki, Yoko Takahashi, Liying Yang, Tomo Osako, Tomoyoshi Nakadai, Sayuri Amino, 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 (7): 1208–22. https://doi.org/10.1038/s41416-023-02178-1.

Lareau, Caleb A., Fabiana M. Duarte, Jennifer G. Chew, Vinay K. Kartha, Zach D. Burkett, Andrew S. Kohlway, Dmitry Pokholok, et al. 2019. "Droplet-Based Combinatorial Indexing for Massive-Scale Single-Cell Chromatin Accessibility." *Nature Biotechnology* 37 (8): 916–24. https://doi.org/10.1038/s41587-019-0147-6.

Lau, Jessica W., Erik Lehnert, Anurag Sethi, Raunaq Malhotra, Gaurav Kaushik, Zeynep Onder, Nick Groves-Kirkby, et al. 2017. "The Cancer Genomics Cloud: Collaborative, Reproducible, and Democratized - A New Paradigm in Large-Scale Computational Research." *Cancer Research* 77 (21): e3–6. https://doi.org/10.1158/0008-5472.CAN-17-0387.

Law, Charity W., Yunshun Chen, Wei Shi, and Gordon K. Smyth. 2014. "Voom: Precision Weights Unlock Linear Model Analysis Tools for RNA-Seq Read Counts." *Genome Biology* 15 (February): R29. https://doi.org/10.1186/GB-2014-15-2-R29/FIGURES/11.

Leylek, Rebecca, Marcela Alcántara-Hernández, Jeffrey M. Granja, Michael Chavez, Kimberly Perez, Oscar R. Diaz, Rui Li, Ansuman T. Satpathy, Howard Y. Chang, and Juliana Idoyaga. 2020. "Chromatin Landscape Underpinning Human Dendritic Cell Heterogeneity." *Cell Reports* 32 (12):

108180. https://doi.org/10.1016/J.CELREP.2020.108180.

Li, Huamei, Amit Sharma, Kun Luo, Zhaohui S. Qin, Xiao Sun, and Hongde Liu. 2020. “DeconPeaker, a Deconvolution Model to Identify Cell Types Based on Chromatin Accessibility in ATAC-Seq Data of Mixture Samples.” *Frontiers in Genetics* 11 (June). https://doi.org/10.3389/fgene.2020.00392.

Li, Taiwen, Jingxin Fu, Zexian Zeng, David Cohen, Jing Li, Qianming Chen, Bo Li, and X. Shirley Liu. 2020. “TIMER2.0 for Analysis of Tumor-Infiltrating Immune Cells.” *Nucleic Acids Research* 48 (May): W509–14. https://doi.org/10.1093/nar/gkaa407.

Liao, Yang, Gordon K. Smyth, and Wei Shi. 2014. “FeatureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features.” *Bioinformatics* 30 (7): 923–30. https://doi.org/10.1093/BIOINFORMATICS/BTT656.

Liu, Qian, Lisa Zaba, Ansuman T. Satpathy, Michelle Longmire, Wen Zhang, Kun Li, Jeffrey Granja, et al. 2020. “Chromatin Accessibility Landscapes of Skin Cells in Systemic Sclerosis Nominate Dendritic Cells in Disease Pathogenesis.” *Nature Communications* 11 (1): 5843. https://doi.org/10.1038/s41467-020-19702-z.

Luo, Liheng, Michael Gribskov, and Sufang Wang. 2022. “Bibliometric Review of ATAC-Seq and Its Application in Gene Expression.” *Briefings in Bioinformatics*, March. https://doi.org/10.1093/BIB/BBAC061.

Machlab, Dania, Lukas Burger, Charlotte Soneson, Filippo M. Rijli, Dirk Schübeler, and Michael B. Stadler. 2022. “MonaLisa: An R/Bioconductor Package for Identifying Regulatory Motifs.” *Bioinformatics* 38 (9): 2624–25. https://doi.org/10.1093/BIOINFORMATICS/BTAC102.

Mayer, Michael. 2023. “R Package ‘ SplitTools ’: Tools for Data Splitting. R Package Version 1.0.1. Https://Cran.r-Project.Org/Web/Packages/SplitTools/Index.Html.”

Monaco, Gianni, Bernett Lee, Weili Xu, Seri Mustafah, You Yi Hwang, Christophe Carré, Nicolas Burdin, et al. 2019. “RNA-Seq Signatures Normalized by MRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types.” *Cell Reports* 26 (6): 1627-1640.e7.

https://doi.org/10.1016/J.CELREP.2019.01.041.

Mumbach, Maxwell R., Ansuman T. Satpathy, Evan A. Boyle, Chao Dai, Benjamin G. Gowen, Seung Woo Cho, Michelle L. Nguyen, et al. 2017. “Enhancer Connectome in Primary Human Cells Identifies Target Genes of Disease-Associated DNA Elements.” *Nature Genetics* 49 (11): 1602–12. https://doi.org/10.1038/ng.3963.

Newman, Aaron M., Chih Long Liu, Michael R. Green, Andrew J. Gentles, Weiguo Feng, Yue Xu, Chuong D. Hoang, Maximilian Diehn, and Ash A. Alizadeh. 2015. “Robust Enumeration of Cell Subsets from Tissue Expression Profiles.” *Nature Methods* 12 (5): 453–57. https://doi.org/10.1038/nmeth.3337.

Newman, Aaron M., Chloé B. Steen, Chih Long Liu, Andrew J. Gentles, Aadel A. Chaudhuri, Florian Scherer, Michael S. Khodadoust, et al. 2019. “Determining Cell Type Abundance and Expression from Bulk Tissues with Digital Cytometry.” *Nature Biotechnology* 37 (7): 773–82. https://doi.org/10.1038/s41587-019-0114-2.

Peng, Xianlu Laura, Richard A. Moffitt, Robert J. Torphy, Keith E. Volmar, and Jen Jen Yeh. 2019. “De Novo Compartment Deconvolution and Weight Estimation of Tumor Samples Using DECODER.” *Nature Communications* 10 (1): 4729. https://doi.org/10.1038/s41467-019-12517-7.

Perez, Cristina, Cirino Botta, Aintzane Zabaleta, Noemi Puig, Maria-Teresa Cedena, Ibai Goicoechea, Daniel Alameda, et al. 2020. “Immunogenomic Identification and Characterization of Granulocytic Myeloid-Derived Suppressor Cells in Multiple Myeloma.” *Blood* 136 (2): 199–209. https://doi.org/10.1182/BLOOD.2019004537.

Qiu, Yixuan, Jiebiao Wang, Jing Lei, and Kathryn Roeder. 2021. “Identification of Cell-Type-Specific Marker Genes from Co-Expression Patterns in Tissue Samples.” *Bioinformatics* 37 (19): 3228–34. https://doi.org/10.1093/BIOINFORMATICS/BTAB257.

Racle, Julien, and David Gfeller. 2020. “EPIC: A Tool to Estimate the Proportions of Different Cell Types from Bulk Gene Expression Data.” In *Methods in Molecular Biology*, 2120:233–48. Humana Press Inc. https://doi.org/10.1007/978-1-0716-0327-7_17.

Racle, Julien, Kaat de Jonge, Petra Baumgaertner, Daniel E. Speiser, and David Gfeller. 2017. "Simultaneous Enumeration of Cancer and Immune Cell Types from Bulk Tumor Gene Expression Data." *ELife* 6 (November). https://doi.org/10.7554/eLife.26476.

Ram-Mohan, Nikhil, Simone A Thair, Ulrike M Litzenburger, Steven Cogill, Nadya Andini, Xi Yang, Howard Y Chang, and Samuel Yang. 2021. "Profiling Chromatin Accessibility Responses in Human Neutrophils with Sensitive Pathogen Detection." *Life Science Alliance* 4 (8). https://doi.org/10.26508/LSA.202000976.

Regner, Matthew J., Kamila Wisniewska, Susana Garcia-Recio, Aatish Thennavan, Raul Mendez-Giraldez, Venkat S. Malladi, Gabrielle Hawkins, et al. 2021. "A Multi-Omic Single-Cell Landscape of Human Gynecologic Malignancies." *Molecular Cell* 81 (23): 4924-4941.e10. https://doi.org/10.1016/j.molcel.2021.10.013.

Risso, Davide, Katja Schwartz, Gavin Sherlock, and Sandrine Dudoit. 2011. "GC-Content Normalization for RNA-Seq Data." *BMC Bioinformatics* 12 (December): 480. https://doi.org/10.1186/1471-2105-12-480/FIGURES/7.

Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. 2015. "Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies." *Nucleic Acids Research* 43 (7): e47. https://doi.org/10.1093/nar/gkv007.

Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2010. "EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." *Bioinformatics* 26 (1): 139–40. https://doi.org/10.1093/BIOINFORMATICS/BTP616.

Rozowsky, Joel, Jiahao Gao, Beatrice Borsari, Roderic Guigó, Thomas R Gingeras, Mark Gerstein, Yucheng T Yang, et al. 2023. "The EN-TEx Resource of Multi-Tissue Personal Epigenomes & Variant-Impact Models." *Cell* 186 (7): 1493-1511.e40. https://doi.org/10.1016/j.cell.2023.02.018.

Sanseviero, Emilio. 2019. "NK Cell-Fc Receptors Advance Tumor Immunotherapy." *Journal of Clinical Medicine* 8 (10): 1667. https://doi.org/10.3390/JCM8101667.

Satpathy, Ansuman T., Jeffrey M. Granja, Kathryn E. Yost, Yanyan Qi, Francesca Meschi, Geoffrey P. McDermott, Brett N. Olsen, et al. 2019. “Massively Parallel Single-Cell Chromatin Landscapes of Human Immune Cell Development and Intratumoral T Cell Exhaustion.” *Nature Biotechnology* 37 (8): 925–36. https://doi.org/10.1038/s41587-019-0206-z.

Smith, Jason P, M Ryan Corces, Jin Xu, Vincent P Reuter, Howard Y Chang, and Nathan C Sheffield. 2021. “PEPATAC: An Optimized Pipeline for ATAC-Seq Data Analysis with Serial Alignments.” *NAR Genomics and Bioinformatics* 3 (4). https://doi.org/10.1093/NARGAB/LQAB101.

Stuart, Tim, Avi Srivastava, Shaista Madad, Caleb A. Lareau, and Rahul Satija. 2021. “Single-Cell Chromatin State Analysis with Signac.” *Nature Methods* 18 (November): 1333–41. https://doi.org/10.1038/s41592-021-01282-5.

Sturm, Gregor, Francesca Finotello, Florent Petitprez, Jitao David Zhang, Jan Baumbach, Wolf H Fridman, Markus List, and Tatsiana Aneichyk. 2019. “Comprehensive Evaluation of Transcriptome-Based Cell-Type Quantification Methods for Immuno-Oncology.” *Bioinformatics* 35 (14): i436–45. https://doi.org/10.1093/bioinformatics/btz363.

Subramanian, Aravind, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, et al. 2005. “Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles.” *PNAS* 102 (43): 15545–50. https://doi.org/10.1073/pnas.0506580102.

Teschendorff, Andrew E., Tianyu Zhu, Charles E. Breeze, and Stephan Beck. 2020. “EPISCORE: Cell Type Deconvolution of Bulk Tissue DNA Methylomes from Single-Cell RNA-Seq Data.” *Genome Biology* 21 (September). https://doi.org/10.1186/s13059-020-02126-9.

The ENCODE Project Consortium, Jill E. Moore, Michael J. Purcaro, Henry E. Pratt, Charles B. Epstein, Noam Shoresh, and Jessika Adrian. 2020. “Expanded Encyclopaedias of DNA Elements in the Human and Mouse Genomes.” *Nature* 583 (July): 699–710. https://doi.org/10.1038/s41586-020-2493-4.

Trizzino, Marco, Avery Zucco, Sandra Deliard, Fang Wang, Elisa Barbieri, Filippo Veglia, Dmitry

Gabrilovich, and Alessandro Gardini. 2021. “EGR1 Is a Gatekeeper of Inflammatory Enhancers in Human Macrophages.” *Science Advances* 7 (3). https://doi.org/10.1126/SCIADV.AAZ8836/SUPPL_FILE/AAZ8836_TABLE_S7.XLSX.

Ucar, Duygu, Eladio J. Márquez, Cheng Han Chung, Radu Marches, Robert J. Rossi, Asli Uyar, Te Chia Wu, et al. 2017. “The Chromatin Accessibility Signature of Human Immune Aging Stems from CD8+ T Cells.” *Journal of Experimental Medicine* 214 (10): 3123–44. https://doi.org/10.1084/jem.20170416.

Visser, Karin E. de, and Johanna A. Joyce. 2023. “The Evolving Tumor Microenvironment: From Cancer Initiation to Metastatic Outgrowth.” *Cancer Cell* 41 (3): 374–403. https://doi.org/10.1016/J.CCELL.2023.02.016.

Watt, Stephen, Louella Vasquez, Klaudia Walter, Alice L. Mann, Kousik Kundu, Lu Chen, Ying Sims, et al. 2021. “Genetic Perturbation of PU.1 Binding and Chromatin Looping at Neutrophil Enhancers Associates with Autoimmune Disease.” *Nature Communications* 12 (April): 2298. https://doi.org/10.1038/S41467-021-22548-8.

Welch, Ryan P., Chee Lee, Paul M. Imbriano, Snehal Patil, Terry E. Weymouth, R. Alex Smith, Laura J. Scott, and Maureen A. Sartor. 2014. “ChIP-Enrich: Gene Set Enrichment Testing for ChIP-Seq Data.” *Nucleic Acids Research* 42 (13): e105. https://doi.org/10.1093/NAR/GKU463.

Xin, Jingxue, Hui Zhang, Yaoxi He, Zhana Duren, Caijuan Bai, Lang Chen, Xin Luo, et al. 2020. “Chromatin Accessibility Landscape and Regulatory Network of High-Altitude Hypoxia Adaptation.” *Nature Communications* 11 (1): 4928. https://doi.org/10.1038/S41467-020-18638-8.

Yu, Guangchuang, Li Gen Wang, and Qing Yu He. 2015. “ChIPseeker: An R/Bioconductor Package for ChIP Peak Annotation, Comparison and Visualization.” *Bioinformatics* 31 (14): 2382–83. https://doi.org/10.1093/BIOINFORMATICS/BTV145.

Zeng, Wanwen, Xi Chen, Zhana Duren, Yong Wang, Rui Jiang, and Wing Hung Wong. 2019. “DC3 Is a Method for Deconvolution and Coupled Clustering from Bulk and Single-Cell Genomics Data.”

Nature Communications 10 (October): 4613. https://doi.org/10.1038/s41467-019-12547-1.

Zhang, Hanyu, Ruoyi Cai, James Dai, and Wei Sun. 2021. “EMeth: An EM Algorithm for Cell Type Decomposition Based on DNA Methylation Data.” *Scientific Reports* 11 (March): 5717. https://doi.org/10.1038/s41598-021-84864-9.

Zhang, Kai, James D Hocker, Michael Miller, Xiaomeng Hou, Joshua Chiou, Olivier B Poirion, Yunjiang Qiu, et al. 2021. “A Single-Cell Atlas of Chromatin Accessibility in the Human Genome.” *Cell* 184 (24): 5985-6001.e19. https://doi.org/10.1016/j.cell.2021.10.024.

Zhang, Ping, Harindra E. Amarasinghe, Justin P. Whalley, Chwen Tay, Hai Fang, Gabriele Migliorini, Andrew C. Brown, et al. 2022. “Epigenomic Analysis Reveals a Dynamic and Context-Specific Macrophage Enhancer Landscape Associated with Innate Immune Activation and Tolerance.” *Genome Biology* 23 (1): 136. https://doi.org/10.1186/S13059-022-02702-1.

Zhang, Yong, Tao Liu, Clifford A. Meyer, Jérôme Eeckhoute, David S. Johnson, Bradley E. Bernstein, Chad Nussbaum, et al. 2008. “Model-Based Analysis of ChIP-Seq (MACS).” *Genome Biology* 9 (9): R137. https://doi.org/10.1186/GB-2008-9-9-R137/FIGURES/3.

List of abbreviations:

ATAC: Assay for Transposase-Accessible chromatin

CBP(s): chromatin binding protein(s)

ChIP-seq: chromatin immunoprecipitation followed by sequencing

DC: dendritic cells

NK: natural killer cells

PCA: principal component analysis

RMSE: root mean squared error

TCGA : The Cancer Genome Atlas

TF(s): transcription factor(s)

TSS: transcription start site

UMAP: Uniform Manifold Approximation

Main Figures.

Figure 1

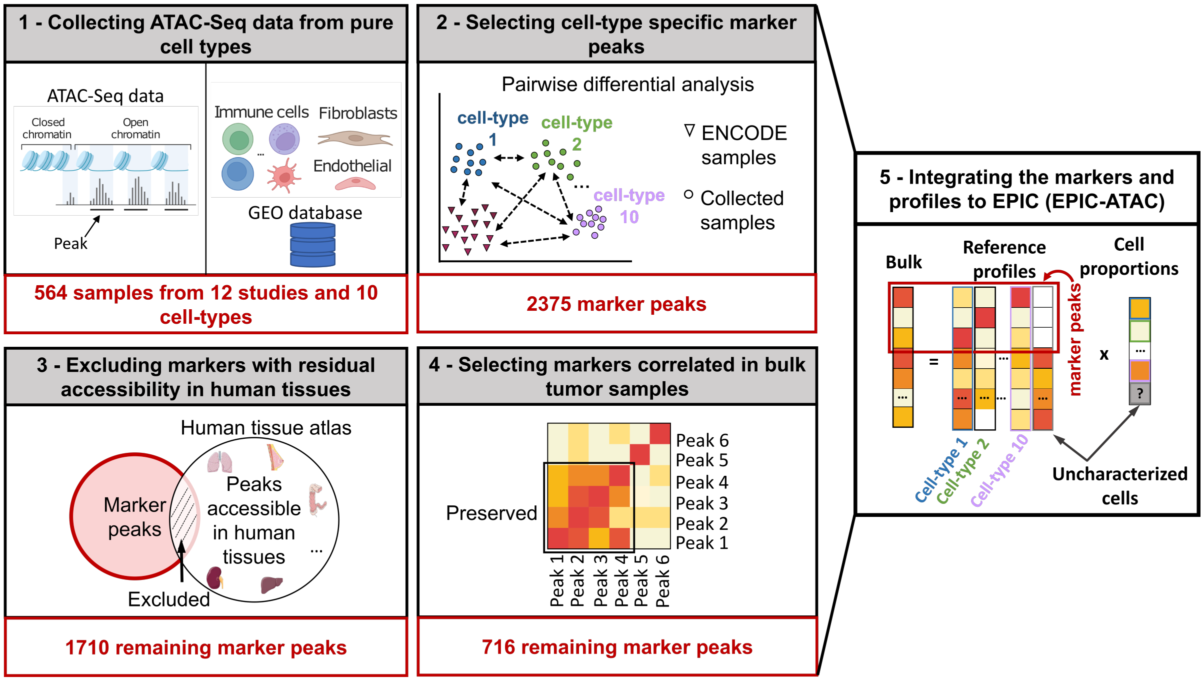

1 - Collecting ATAC-Seq data from pure cell types
ATAC-Seq data
Closed chromatin
Open chromatin
Peak
Immune cells
Fibroblasts
Endothelial
GEO database
564 samples from 12 studies and 10 cell-types
2 - Selecting cell-type specific marker peaks
Pairwise differential analysis
cell-type 1
cell-type 2
cell-type 10
▽ ENCODE samples
○ Collected samples
2375 marker peaks
3 - Excluding markers with residual accessibility in human tissues
Human tissue atlas
Marker peaks
Peaks accessible in human tissues
Excluded
1710 remaining marker peaks
4 - Selecting markers correlated in bulk tumor samples
Preserved
Peak 6
Peak 5
Peak 4
Peak 3
Peak 2
Peak 1
Peak 1
Peak 2
Peak 3
Peak 4
Peak 5
Peak 6
716 remaining marker peaks
5 - Integrating the markers and profiles to EPIC (EPIC-ATAC)
Bulk
Reference profiles
Cell proportions
marker peaks
=
x
?
Cell-type 1
Cell-type 2
Cell-type 10
Uncharacterized cells

Figure 2

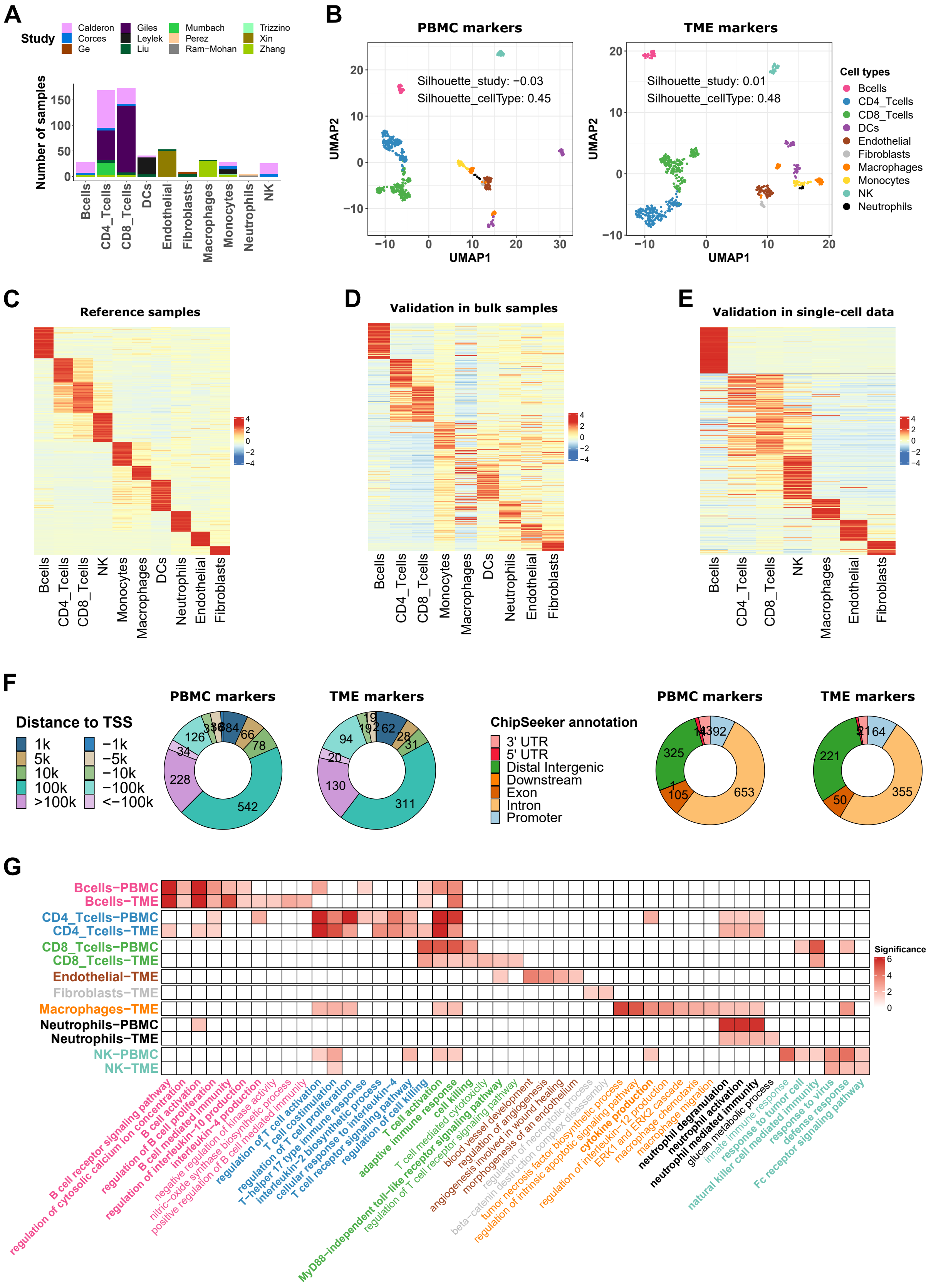

A
Study
Calderon
Corces
Ge
Giles
Leylek
Liu
Mumbach
Perez
Ram−Mohan
Trizzino
Xin
Zhang
Number of samples
0
50
100
150
Bcells
CD4_Tcells
CD8_Tcells
DCs
Endothelial
Fibroblasts
Macrophages
Monocytes
Neutrophils
NK
B
PBMC markers
Silhouette_study: −0.03
Silhouette_cellType: 0.45
TME markers
Silhouette_study: 0.01
Silhouette_cellType: 0.48
UMAP1
UMAP2
−10
0
10
20
30
Cell types
Bcells
CD4_Tcells
CD8_Tcells
DCs
Endothelial
Fibroblasts
Macrophages
Monocytes
NK
Neutrophils
C
Reference samples
Bcells
CD4_Tcells
CD8_Tcells
NK
Monocytes
Macrophages
DCs
Neutrophils
Endothelial
Fibroblasts
4
2
0
−2
−4
D
Validation in bulk samples
Bcells
CD4_Tcells
CD8_Tcells
Monocytes
Macrophages
DCs
Neutrophils
Endothelial
Fibroblasts
E
Validation in single-cell data
Bcells
CD4_Tcells
CD8_Tcells
NK
Macrophages
Endothelial
Fibroblasts
F
Distance to TSS
1k
5k
10k
100k
>100k
−1k
−5k
−10k
−100k
<−100k
PBMC markers
3 3 8 84
66
78
542
228
34
126
TME markers
19
1 9 2
62
28
31
311
130
20
94
ChipSeeker annotation
3' UTR
5' UTR
Distal Intergenic
Downstream
Exon
Intron
Promoter
PBMC markers
14 43 92
653
105
1
325
TME markers
3 21 64
355
50
221
G
Bcells−PBMC
Bcells−TME
CD4_Tcells−PBMC
CD4_Tcells−TME
CD8_Tcells−PBMC
CD8_Tcells−TME
Endothelial−TME
Fibroblasts−TME
Macrophages−TME
Neutrophils−PBMC
Neutrophils−TME
NK−PBMC
NK−TME
Significance
6
4
2
0
B cell receptor signaling pathway
regulation of cytosolic calcium ion concentration
B cell activation
regulation of B cell proliferation
B cell mediated immunity
regulation of interleukin−10 production
interleukin−4 production
negative regulation of kinase activity
nitric−oxide synthase biosynthetic process
positive regulation of B cell mediated immunity
regulation of T cell activation
T cell costimulation
regulation of T cell proliferation
T−helper 17 type immune response
interleukin−2 biosynthetic process
cellular response to interleukin−4
T cell receptor signaling pathway
regulation of cell killing
T cell activation
adaptive immune response
cell killing
T cell mediated cytotoxicity
MyD88−independent toll−like receptor signaling pathway
regulation of T cell receptor signaling pathway
blood vessel development
regulation of angiogenesis
angiogenesis involved in wound healing
morphogenesis of an endothelium
regulation of necroptotic process
beta−catenin destruction complex disassembly
tumor necrosis factor biosynthetic process
regulation of intrinsic apoptotic signaling pathway
cytokine production
regulation of interleukin−12 production
ERK1 and ERK2 cascade
macrophage chemotaxis
macrophage migration
neutrophil degranulation
neutrophil activation
neutrophil mediated immunity
glucan metabolic process
innate immune response
response to tumor cell
natural killer cell mediated immunity
response to virus
defense response
Fc receptor signaling pathway

Figure 3

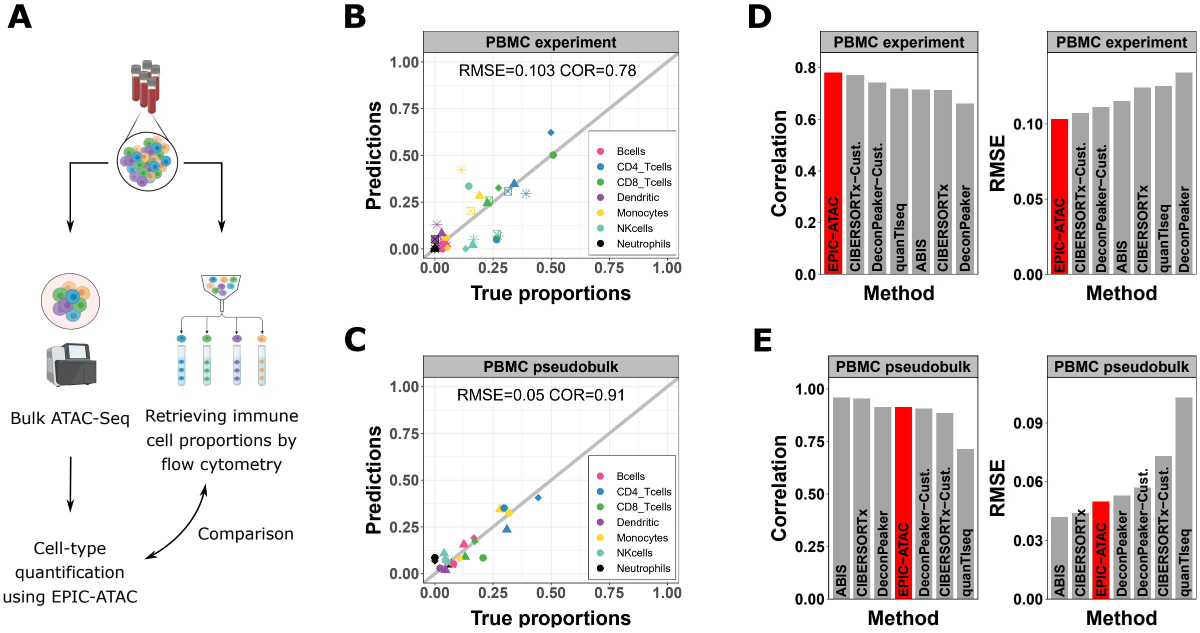

A
Bulk ATAC-Seq
Retrieving immune cell proportions by flow cytometry
Cell-type quantification using EPIC-ATAC
Comparison
B
PBMC experiment
RMSE=0.103 COR=0.78
Predictions
True proportions
0.00
0.25
0.50
0.75
1.00
Bcells
CD4_Tcells
CD8_Tcells
Dendritic
Monocytes
NKcells
Neutrophils
C
PBMC pseudobulk
RMSE=0.05 COR=0.91
Predictions
True proportions
D
PBMC experiment
Correlation
0.0
0.2
0.4
0.6
0.8
EPIC−ATAC
CIBERSORTx−Cust.
DeconPeaker−Cust.
quanTIseq
ABIS
CIBERSORTx
DeconPeaker
Method
PBMC experiment
RMSE
0.00
0.05
0.10
EPIC−ATAC
CIBERSORTx−Cust.
DeconPeaker−Cust.
ABIS
CIBERSORTx
quanTIseq
DeconPeaker
Method
E
PBMC pseudobulk
Correlation
0.00
0.25
0.50
0.75
1.00
ABIS
CIBERSORTx
DeconPeaker
EPIC−ATAC
DeconPeaker−Cust.
CIBERSORTx−Cust.
quanTIseq
Method
PBMC pseudobulk
RMSE
0.00
0.03
0.06
0.09
ABIS
CIBERSORTx
EPIC−ATAC
DeconPeaker
DeconPeaker−Cust.
CIBERSORTx−Cust.
quanTIseq
Method

Figure 4

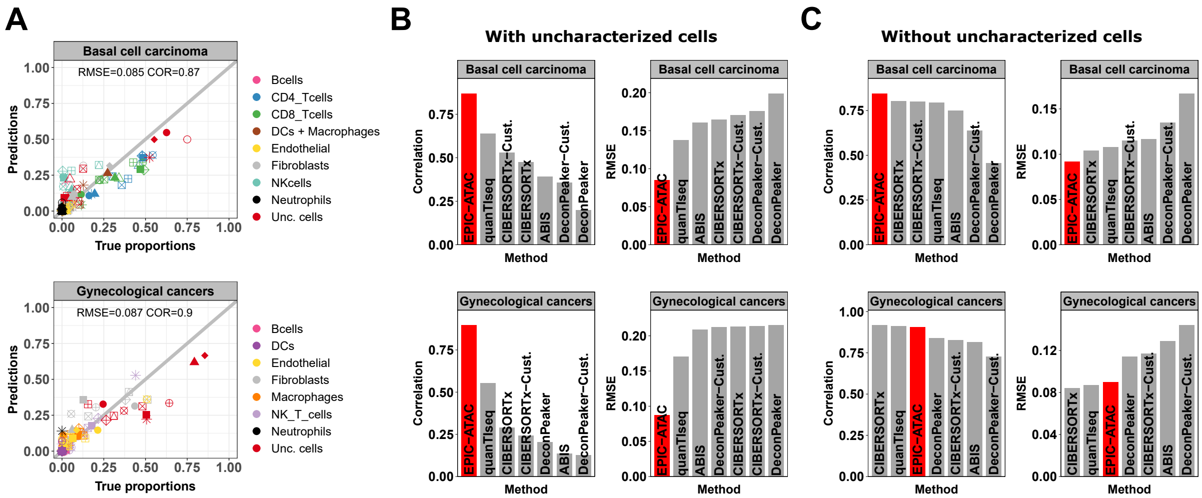

A
B
C
With uncharacterized cells
Without uncharacterized cells
Basal cell carcinoma
RMSE=0.085 COR=0.87
Predictions
True proportions
0.00
0.25
0.50
0.75
1.00
Bcells
CD4_Tcells
CD8_Tcells
DCs + Macrophages
Endothelial
Fibroblasts
NKcells
Neutrophils
Unc. cells
Gynecological cancers
RMSE=0.087 COR=0.9
Bcells
DCs
Endothelial
Fibroblasts
Macrophages
NK_T_cells
Neutrophils
Unc. cells
Correlation
RMSE
Method
EPIC−ATAC
quanTIseq
CIBERSORTx−Cust.
CIBERSORTx
ABIS
DeconPeaker−Cust.
DeconPeaker
0.00
0.05
0.10
0.15
0.20
0.04
0.08
0.12

Figure 5

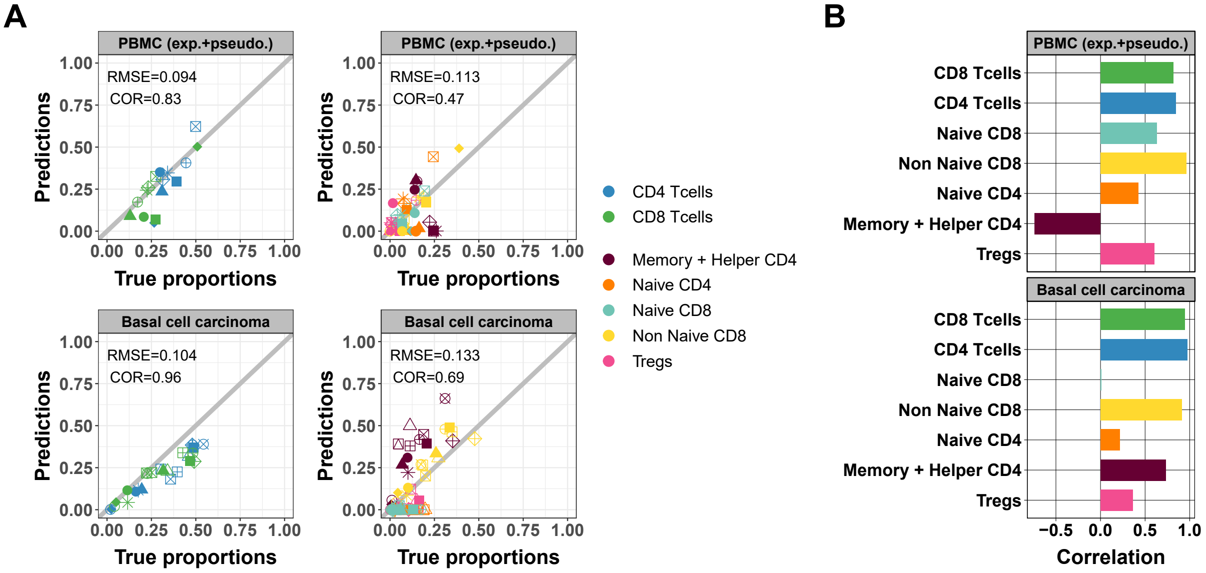

A
PBMC (exp.+pseudo.)
RMSE=0.094
COR=0.83
PBMC (exp.+pseudo.)
RMSE=0.113
COR=0.47
Basal cell carcinoma
RMSE=0.104
COR=0.96
Basal cell carcinoma
RMSE=0.133
COR=0.69
Predictions
True proportions
0.00
0.25
0.50
0.75
1.00
CD4 Tcells
CD8 Tcells
Memory + Helper CD4
Naive CD4
Naive CD8
Non Naive CD8
Tregs
B
PBMC (exp.+pseudo.)
Basal cell carcinoma
CD8 Tcells
CD4 Tcells
Naive CD8
Non Naive CD8
Naive CD4
Memory + Helper CD4
Tregs
−0.5
0.0
0.5
1.0
Correlation

Figure 6

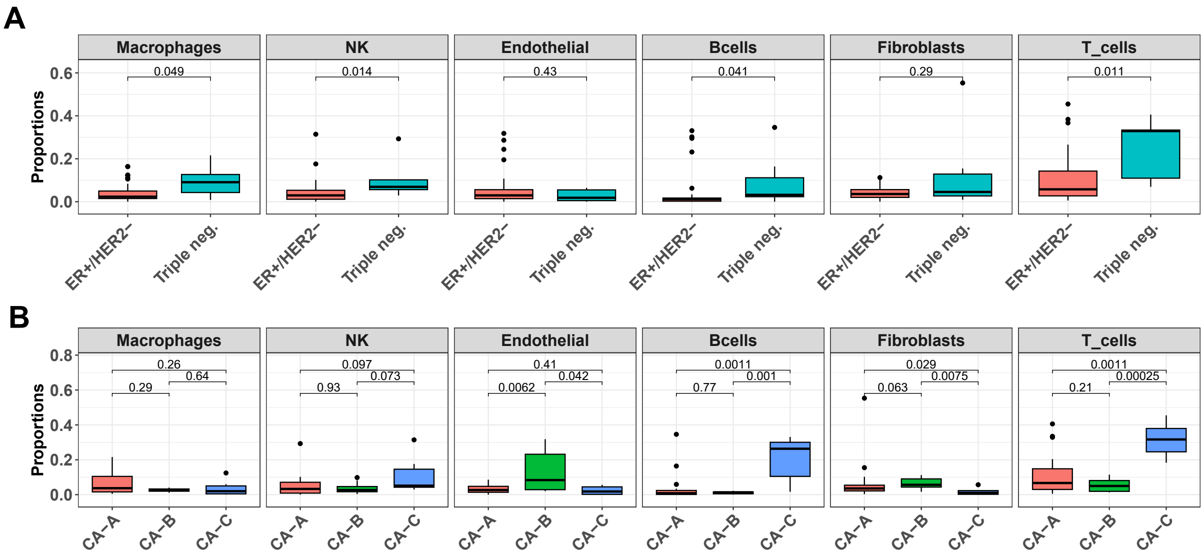
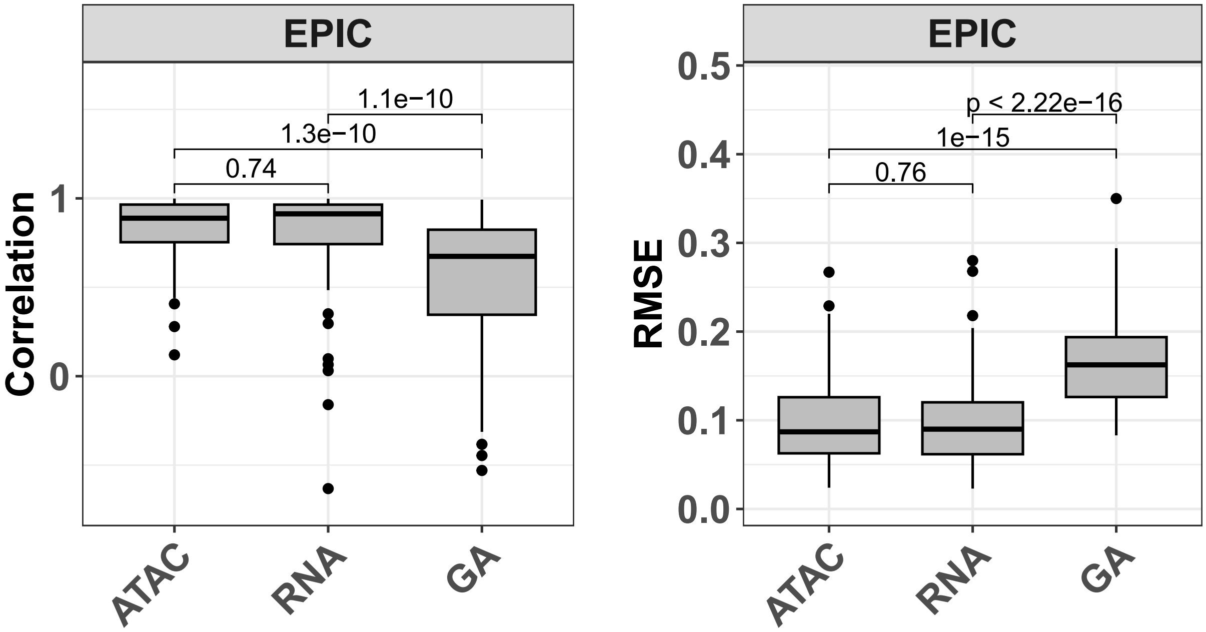

A
Macrophages
NK
Endothelial
Bcells
Fibroblasts
T_cells
Proportions
0.6
0.4
0.2
0.0
0.049
0.014
0.43
0.041
0.29
0.011
ER+/HER2−
Triple neg.
B
Macrophages
NK
Endothelial
Bcells
Fibroblasts
T_cells
Proportions
0.8
0.6
0.4
0.2
0.0
0.26
0.64
0.29
0.097
0.073
0.93
0.41
0.042
0.0062
0.0011
0.001
0.77
0.029
0.0075
0.063
0.0011
0.00025
0.21
CA−A
CA−B
CA−C

Figure 7


EPIC
1.1e−10
1.3e−10
0.74
1
0
Correlation
ATAC
RNA
GA
EPIC
0.5
0.4
0.3
0.2
0.1
0.0
p < 2.22e−16
1e−15
0.76
RMSE
ATAC
RNA
GA