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 Center, 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 most non-malignant cell types frequently observed in the micro-environment of human tumors. We then integrate these data into the EPIC deconvolution framework (Racle *et al.*, 2017) to quantify cell-type heterogeneity in bulk ATAC-Seq data. Our EPIC-ATAC tool accurately predicts non-malignant and

malignant cell fractions in tumor samples. When applied to a human breast cancer cohort, EPIC-ATAC accurately infers the immune contexture of the main breast cancer subtypes.

Introduction

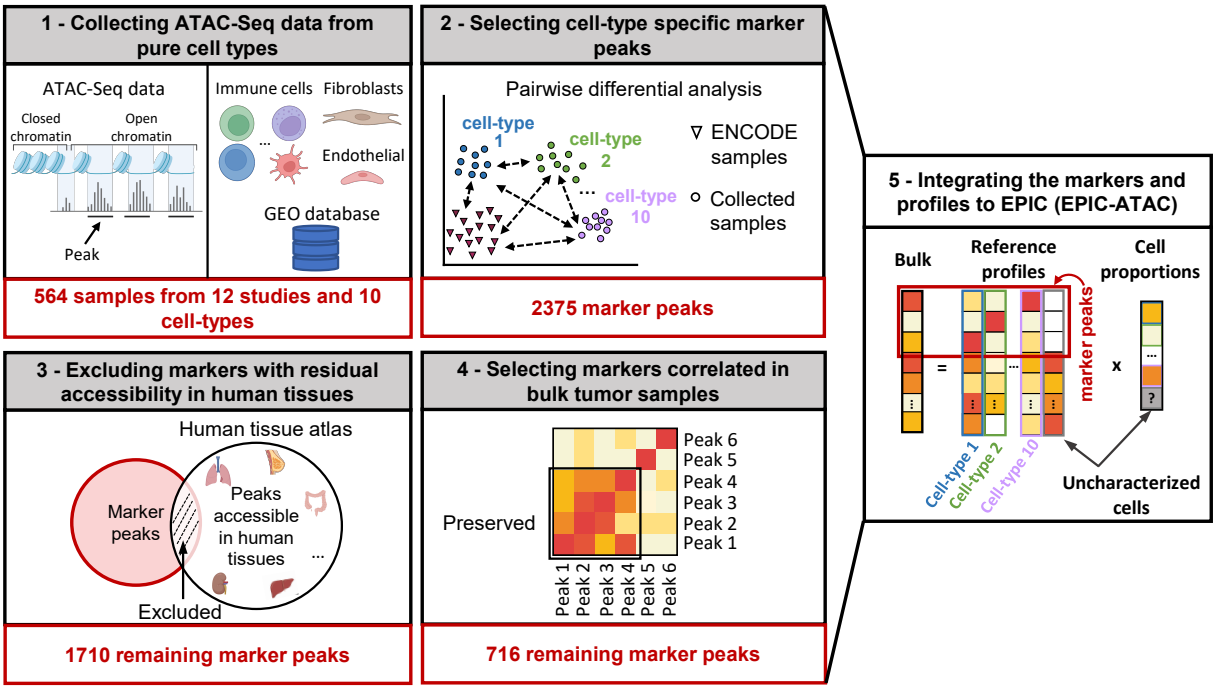
Gene regulation is a dynamic process largely determined by the physical access of chromatin-binding factors such as transcription factors (TFs) to regulatory regions of the DNA (*e.g.,* enhancers and promoters) (Klemm, Shipony and Greenleaf, 2019). The genome-wide landscape of chromatin accessibility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

Most biological tissues are composed of multiple cell types. For instance, tumors are complex ecosystems including malignant and stromal cells as well as a large diversity of immune cells. This cellular heterogeneity impacts tumor progression as well as response to immunotherapy (Fridman *et al.*, 2012, 2017; de Visser and Joyce, 2023). Most existing ATAC-Seq data from tumors were performed on bulk samples, thereby mixing the signal from both cancer and non-malignant cells. Precisely quantifying the proportions of different cell types in such samples represents a promising way to

51 explore the immune contexture and the composition of the tumor micro-environment (TME) across
52 large cohorts. Carefully assessing cell-type heterogeneity is also important to handle confounding
53 factors in genomic analyses in which samples with different cellular compositions are compared.
54 Recently, single-cell ATAC-Seq (scATAC-Seq) has been developed to explore cellular heterogeneity with
55 high resolution in complex biological systems (Cusanovich *et al.*, 2015; Lareau *et al.*, 2019; Satpathy *et*
56 *al.*, 2019). However, the resulting data are sensitive to technical noise and such experiments require
57 important resources, which so far limits the use of scATAC-Seq in contrast to bulk ATAC-Seq in the
58 context of large cohorts.
59
60 In the past decade, computational deconvolution tools have been developed to predict the proportion
61 of diverse cell types from bulk genomic data obtained from tumor samples (Becht *et al.*, 2016; Racle
62 *et al.*, 2017; Avila Cobos *et al.*, 2018, 2020; Finotello *et al.*, 2019; Monaco *et al.*, 2019; Newman *et al.*,
63 2019; Sturm *et al.*, 2019; H. Li *et al.*, 2020). As described in more details elsewhere (Avila Cobos *et al.*,
64 2018; Sturm *et al.*, 2019), many of these tools model bulk data as a mixture of reference profiles either
65 coming from purified cell populations or inferred from single-cell genomic data for each cell type. The
66 accuracy of cell-type proportion predictions relies on the quality of these reference profiles as well as
67 on the specificity of cell-type markers (Avila Cobos *et al.*, 2018). A limitation of most deconvolution
68 algorithms is that they do not predict the proportion of cell types that are not present in the reference
69 profiles (hereafter referred to as 'uncharacterized' cells). In the context of tumor samples, these
70 uncharacterized cell populations include malignant cells whose molecular profiles differ not only from
71 one cancer type to another, but also from one patient to another even within the same tumor type
72 (Corces *et al.*, 2018). To address this limitation, a few tools developed for bulk RNA-Seq data consider
73 uncharacterized cells in their deconvolution framework by using cell-type specific markers not
74 expressed in the uncharacterized cells (Gosink, Petrie and Tsinoremas, 2007; Clarke, Seol and Clarke,
75 2010; Racle *et al.*, 2017; Finotello *et al.*, 2019). These tools include EPIC (Estimating the Proportion of

76 Immune and Cancer cells) which simultaneously quantifies immune, stromal, vascular as well as
77 uncharacterized cells from bulk tumor samples (Racle *et al.*, 2017; Racle and Gfeller, 2020).
78
79 Most deconvolution algorithms have been developed for transcriptomic data (RNA-Seq data) (Gong
80 and Szustakowski, 2013; Newman *et al.*, 2015, 2019; Racle *et al.*, 2017; Finotello *et al.*, 2019; Jimenez-
81 Sanchez, Cast and Miller, 2019; Monaco *et al.*, 2019; T. Li *et al.*, 2020). More recently they have been
82 proposed for other omics layers such as methylation (Aryee *et al.*, 2014; Teschendorff *et al.*, 2017,
83 2020; Chakravarthy *et al.*, 2018; Hicks and Irizarry, 2019; Rahmani *et al.*, 2019; Arneson, Yang and
84 Wang, 2020; H. Zhang *et al.*, 2021; Salas *et al.*, 2022) proteomics (Feng *et al.*, 2023) or chromatin
85 accessibility (H. Li *et al.*, 2020). For the latter, a specific framework called DeconPeaker (H. Li *et al.*,
86 2020) was developed to estimate cell-type proportions from bulk ATAC-Seq samples. Deconvolution
87 tools developed initially for RNA-Seq data can also be applied on ATAC-Seq. For example, the popular
88 deconvolution tool, CIBERSORT (Newman *et al.*, 2015), in combination with ATAC-Seq profiles, was
89 used to deconvolve leukemic ATAC-Seq samples (Corces *et al.*, 2016).
90 Other methods have been proposed to decompose ATAC-Seq bulk profiles into subpopulation-specific
91 profiles (Burdziak *et al.*, 2019; Zeng *et al.*, 2019) or compartments (Peng *et al.*, 2019). However, these
92 methods have more requisites: (i) the integration of the ATAC-Seq data with single-cell or bulk RNA-
93 Seq (Burdziak *et al.*, 2019; Zeng *et al.*, 2019) and HIChIP data (Zeng *et al.*, 2019) or, (ii) subsequent
94 feature annotation to associate compartments with cell types or biological processes (Peng *et al.*,
95 2019).
96 The application of existing bulk ATAC-Seq data deconvolution tools to solid tumors faces some
97 limitations. First, current computational frameworks do not quantify populations of uncharacterized
98 cell types. Second, ATAC-Seq based markers (*i.e.,* chromatin accessible regions called peaks) and
99 reference profiles generated so far have been derived in the context of hematopoietic cell mixtures
100 (Corces *et al.*, 2016; H. Li *et al.*, 2020). Markers and profiles for major populations of the TME (*e.g.,*
101 stromal and vascular cells) are thus missing. While cell-type specific markers have been identified from

102 scATAC-Seq data (K. Zhang *et al.*, 2021), not all TME-relevant cell types are covered (*e.g.*, lack of
103 scATAC-Seq data from neutrophils due to extracellular traps formation). Many of these markers have
104 also not been curated to fulfill the requirements of tools such as EPIC to quantify uncharacterized cells
105 (*i.e.*, markers of a cell-type should not be accessible in other human tissues).
106
107 In this study, we collected ATAC-Seq data from pure cell types to identify cell-type specific
108 marker peaks and to build reference profiles from most major non-malignant cell types typically
109 observed in tumors. These data were integrated in the EPIC (Racle *et al.*, 2017) framework to perform
110 bulk ATAC-Seq samples deconvolution (Figure 1). Applied on peripheral blood mononuclear cells
111 (PBMCs) and tumor samples, the EPIC-ATAC framework showed accurate predictions of the
112 proportions of non-malignant and malignant cells.
113



114

115 ***Figure 1: Graphical description of the identification of cell-type specific marker peaks and reference***
116 ***ATAC-Seq profiles included in the EPIC-ATAC framework.*** *1) 564 pure ATAC-Seq data of sorted cells*
117 *were collected to build reference profiles for non-malignant cell types observed in the tumor*
118 *microenvironment. 2) Cell-type specific marker peaks were identified using differential accessibility*

analysis. 3) Markers with previously observed chromatin accessibility in human healthy tissues were excluded. 4) For tumor bulk deconvolution, the set of remaining marker peaks was refined by selecting markers with correlated behavior in tumor bulk samples. 5) The cell-type specific marker peaks and reference profiles were integrated in the EPIC-ATAC framework to perform bulk ATAC-Seq deconvolution. Partly created with BioRender.com.

Results

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

A key determinant for accurate predictions of cell-type proportions by most deconvolution tools is the availability of reliable cell-type specific markers and reference profiles. To identify robust chromatin accessibility marker peaks of cell types observed in the tumor microenvironment, we collected 564 samples of sorted cell populations from twelve studies including B cells (Corces *et al.*, 2016; Calderon *et al.*, 2019; P. Zhang *et al.*, 2022), CD4+ T cells (Corces *et al.*, 2016; Mumbach *et al.*, 2017; Liu *et al.*, 2020; Giles *et al.*, 2022; P. Zhang *et al.*, 2022), CD8+ T cells (Corces *et al.*, 2016; Calderon *et al.*, 2019; Liu *et al.*, 2020; Giles *et al.*, 2022; P. Zhang *et al.*, 2022), natural killer (NK) cells (Corces *et al.*, 2016; Calderon *et al.*, 2019), dendritic cells (DCs) (mDCs and pDCs are grouped in this cell-type category) (Calderon *et al.*, 2019; Leylek *et al.*, 2020; Liu *et al.*, 2020), macrophages (Liu *et al.*, 2020; P. Zhang *et al.*, 2022), monocytes (Corces *et al.*, 2016; Calderon *et al.*, 2019; Leylek *et al.*, 2020; Trizzino *et al.*, 2021; P. Zhang *et al.*, 2022), neutrophils (Perez *et al.*, 2020; Ram-Mohan *et al.*, 2021), fibroblasts (Liu *et al.*, 2020; Ge *et al.*, 2021) and endothelial cells (Liu *et al.*, 2020; Xin *et al.*, 2020) (Figure 1 box 1, Figure 2A, Supplementary file 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 observed across samples and studies, *i.e.*, for each study, peaks detected in at least half of the samples were considered, and for each cell type, only peaks detected jointly in all studies were kept (see Materials and Methods, section 1).

145 These peaks were then used to perform pairwise differential analysis to identify marker peaks for each
146 cell type (Figure 1, box 2). To ensure that the cell-type specific marker peaks are not accessible in other
147 human tissues, we included in the differential analysis ATAC-Seq samples from diverse human tissues
148 from the ENCODE data (The ENCODE Project Consortium *et al.*, 2020; Rozowsky *et al.*, 2023) (Figure 1-
149 figure supplement 1). To select a sufficient number of peaks prior to peak filtering, the top 200 peaks
150 recurrently differentially accessible across all cell-type pairs were selected as cell-type specific markers
151 (see Materials and Methods, section 2). To filter out markers that could be accessible in other human
152 cell types than those included in our reference profiles, we used the human atlas study (K. Zhang *et*
153 *al.*, 2021), which identified modules of open chromatin regions accessible in a comprehensive set of
154 human tissues, and we excluded from our marker list the markers overlapping these modules (Figure
155 1, box 3, see Materials and Methods section 2). The resulting marker peaks specific only to the immune
156 cell types were considered for the deconvolution of PBMC samples (PBMC markers). For the
157 deconvolution of tumor bulk samples, the lists of marker peaks specific to fibroblasts and endothelial
158 cells were added to the PBMC markers. This extended set of markers was further refined based on the
159 correlation patterns of the markers in tumor bulk samples from the diverse solid cancer types from
160 The Cancer Genome Atlas (TCGA) (Corces *et al.*, 2018), *i.e.*, markers exhibiting the highest correlation
161 patterns in the tumor bulk samples were selected using the *findCorrelation* function from the caret R
162 package (Kuhn, 2008) (Figure 1, box 4, see the Material and methods, section 2). The latter filtering
163 ensures the relevance of the markers in the TME context since cell-type specific TME markers are
164 expected to be correlated in tumor bulk ATAC-Seq measurements (Qiu *et al.*, 2021). 716 markers of
165 immune, fibroblasts and endothelial cell types remained after the last filtering (defined as TME
166 markers). Considering the difference in cell types and the different filtering steps applied on the PBMC
167 and TME markers, we recommend to use the TME markers and profiles to deconvolve bulk samples
168 from tumor samples and the PBMC markers and profiles to deconvolve PBMC samples.
169 To assess the quality and reproducibility of these markers, we first performed principal component
170 analysis (PCA) based on each set of marker peaks. Computing silhouette coefficients based on the cell-

171 type classification and on the study of origin showed that samples clustered by cell type and not by
172 study of origin (averaged silhouette coefficients above 0.45 for cell type and around 0 for study of
173 origin). The representation of the samples based on the first axes of the PCA confirmed this observation
174 (Figure 2B and Figure 2-figure supplement 1). These results indicate limited remaining batch effects
175 after data processing and marker selection.
176 We then used the collected samples to generate chromatin accessibility profiles by computing the
177 average of the normalized counts for each peak in each cell type as well as peak variability in each cell
178 type (see Material and methods) since EPIC uses features variability for the estimation of the cell-type
179 proportions (Racle *et al.*, 2017). Figure 2C represents the average chromatin accessibility of each
180 marker peak in each cell type of the reference dataset and highlights, as expected, the cell-type
181 specificity of the selected markers (see also Supplementary files 2 and 3), which was confirmed in
182 independent ATAC-Seq data from sorted cells and single-cell ATAC-Seq samples from blood and diverse
183 human tissues (Figure 2D and 2E, see Materials and methods).
184
185
186

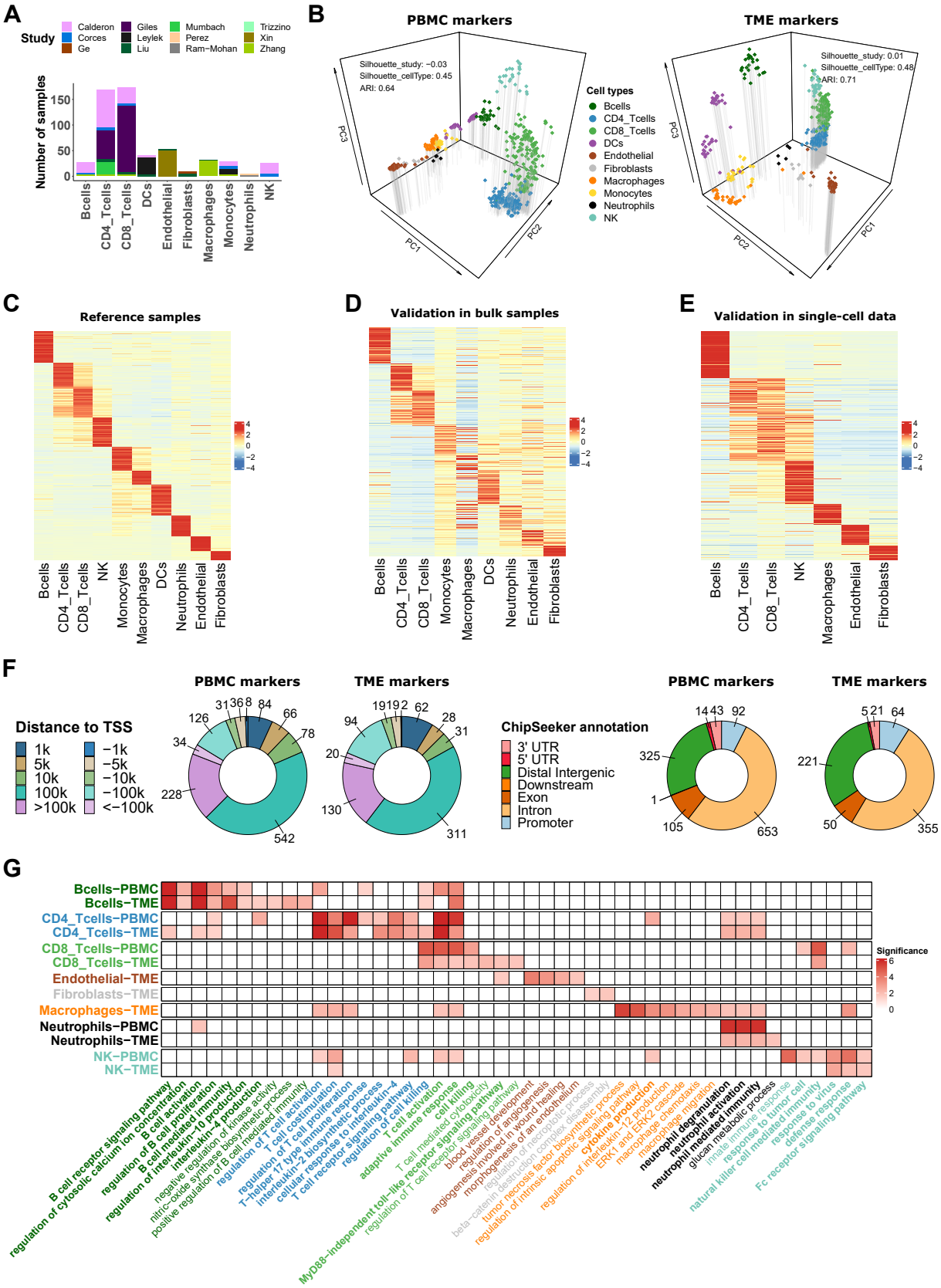


Figure 2: ATAC-Seq data from sorted cell populations reveal cell-type specific marker peaks and reference profiles. A) *Number of samples collected for each cell type. The colors correspond to the*

190 *different studies of origin.* ***B)*** *Representation of the collected samples using the first three components*
191 *of the PCA based on the PBMC markers (left) and TME markers (right). Colors correspond to cell types.*
192 ***C)*** *Scaled averaged chromatin accessibility of all the cell-type specific marker peaks (PBMC and TME*
193 *markers) (rows) in each cell type (columns) in the ATAC-Seq reference samples used to identify the*
194 *marker peaks.* ***D)*** *Scaled averaged chromatin accessibility of all the marker peaks in external ATAC-Seq*
195 *data from samples of pure cell types excluded from the reference samples (see Material and Methods).*
196 ***E)*** *Scaled averaged chromatin accessibility of all the marker peaks in an external scATAC-Seq dataset*
197 *(Human Atlas* (K. Zhang *et al.*, 2021)*).* ***F)*** *Distribution of the marker peak distances to the nearest*
198 *transcription start site (TSS) (left panel) and the ChIPseeker annotations (right panel).* ***G)*** *Significance (-*
199 *log10(q.value)) of pathways (columns) enrichment test obtained using ChIP-Enrich on each set of cell-*
200 *type specific marker peaks (rows). A subset of relevant enriched pathways is represented. Colors of the*
201 *names of the pathways correspond to cell types where the pathways were found to be enriched. When*
202 *pathways were significantly enriched in more than one set of peaks, pathways names are written in*
203 *bold.*
204

205 **Annotations of the marker peaks highlight their biological relevance**

206 To characterize the different marker peaks, we annotated them using ChIPseeker (Yu, Wang and He,
207 2015). We observed that most of the markers are in distal and intergenic regions (Figure 2F), which is
208 expected considering the large proportion of distal regions in the human genome and the fact that
209 such regions have been previously described as highly cell-type specific (Corces *et al.*, 2016). We also
210 noticed that 7% of the PBMC and TME marker peaks are in promoter regions in contrast to 4% when
211 considering matched genomic regions randomly selected in the set of peaks identified prior to the
212 differential analysis (see Material and methods), which suggests enrichment in our marker peaks for
213 important regulatory regions.
214 To assess the biological relevance of the marker peaks, we associated each marker peak to its nearest
215 gene using ChIP-Enrich based on the "nearest transcription start site (TSS)" locus definition (Welch *et*

216 *al.*, 2014) (Supplementary files 4 and 5). Nearest genes reported as known marker genes in public
217 databases of gene markers (*i.e.,* PanglaoDB (Franzén, Gan and Björkegren, 2019) and CellMarker (Hu
218 *et al.*, 2023)) are listed in Table 1.

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

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

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

Cell type	Nearest genes	Enriched CPBs
Bcells	DHTKD1 LHPP WDFY4 ARID5B HHEX SIDT2 CD82 MS4A1 FCHSD2 USP8 RHCG ATF7IP2 CIITA GGA2 SNX29P2 C16orf74 CBFA2T3 CD79B BCL2 GNG7 CD22 FCER2 FCRL1 LY9 PTPRC LAPTM5 IGLL5 VPREB3 CENPM AFF3 SP100 INPP5D DTNB CD86 RFTN1 ST6GAL1 NGLY1 OSBPL10 TLR9 CD38 SMIM14 ARHGAP24 ADAM19 EBF1 BASP1 CD83 PLEKHG1 CCR6 CCND3 HDAC9 CDCA7L BLK MTSS1 LYN PLEKHF2 MOB3B PAX5	SPIB POU2F2 TCF4 EBF1 TCF3 NFKB1 STAT1 NFKB2 IKZF1 FOXO1 FOXP1 BCL6 POU2AF1 STAT3 BACH2 IKZF3 FLI1 TBX21 JUNB MITF NKX6-2 RBPJ
CD4_Tcells	IL2RA CD6 CD5 CD4 RORA PTPRC CTLA4 ICOS SLC9A9 FHIT TCF7 FYB1 ATXN1 CD40LG	TCF7 RUNX3 SOHLH2 IRF9 GATA3 TBX21 MAF 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 EGR1 FLI1 HIF1A PBX1

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

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

252 10^6. In RNA-Seq based deconvolution, EPIC uses an estimation of the amount of mRNA in each
253 reference cell type to derive cell proportions while correcting for cell-type-specific mRNA bias. For the
254 ATAC-Seq based deconvolution these values were set to 1 to give similar weights to all cell-types
255 quantifications. Indeed ATAC-Seq measures signal at the DNA level, hence the quantity of DNA within
256 each reference cell type is similar.
257

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

259 To test the accuracy of EPIC-ATAC predictions, we first collected PBMCs from five healthy donors. For
260 each donor, half of the cells was used to generate a bulk ATAC-Seq dataset and the other half was used
261 to determine the cellular composition of each sample, *i.e.,* the proportions of monocytes, B cells, CD4+
262 T cells, CD8+ T cells, NK cells and dendritic cells, by multiparametric flow cytometry (Figure 3A, see
263 Materials and methods). We then applied EPIC-ATAC to the bulk ATAC-Seq data. The predicted cell
264 fractions are consistent with the cell fractions obtained by flow cytometry (Figure 3B, Pearson
265 correlation coefficient of 0.78 and root mean squared error (RMSE) of 0.10). However, the accuracy of
266 the estimations is variable depending on the samples. In particular, one sample (*Sample4*, star shape
267 on Figure 3B) exhibits larger discrepancies between EPIC-ATAC predictions and the ground truth.
268 Visualizing our marker peaks in the five PBMC samples (Figure 3-figure supplement 1) showed that this
269 sample might be an outlier considering that its cellular composition is similar to that of Samples 2 and
270 5 but this sample shows particularly high ATAC-Seq accessibility at the monocytes and dendritic marker
271 peaks. This could explain why EPIC-ATAC overestimates the proportions of the two populations in this
272 case.
273 As a second validation, we applied EPIC-ATAC to pseudo-bulk PBMC samples (referred to as the PBMC
274 pseudobulk dataset), generated using three publicly available PBMC scATAC-Seq datasets (Granja *et*
275 *al.*, 2019; Satpathy *et al.*, 2019; 10x Genomics, 2021) (see Material and methods). A high correlation
276 (0.91) between EPIC-ATAC predictions and true cell-type proportions and a low RMSE (0.05) were
277 observed for this dataset (Figure 3C).

278 The accuracy of the predictions obtained with EPIC-ATAC was then compared with the accuracy of
279 other deconvolution approaches which could be used with our reference profiles and marker peaks
280 (Figure 3D-E). To this end, we considered both the DeconPeaker method (H. Li *et al.*, 2020) originally
281 developed for bulk ATAC-Seq as well as several algorithms developed for bulk RNA-Seq (CIBERSORTx
282 (Newman *et al.*, 2019), quanTIseq (Finotello *et al.*, 2019), ABIS (Monaco *et al.*, 2019), and MCPcounter
283 (Becht *et al.*, 2016)). To enable meaningful comparison across the cell types considered in this work
284 and use the method initially developed for bulk RNA-Seq deconvolution, the marker peaks and profiles
285 derived in this work were used in each of these methods (See Material and methods, section 4). As in
286 EPIC-ATAC, the features of the input bulk matrices were matched to our reference profiles' features.
287 Having reference profiles/markers and an ATAC-Seq bulk query with matched features was the only
288 requirement of the different deconvolution models to run on ATAC-Seq data with the default methods
289 parameters. DeconPeaker and CIBERSORTx also include the option to define cell-type specific markers
290 and profiles from a set of reference samples. We thus additionally fed our ATAC-Seq samples collection
291 to both algorithms and used the resulting profiles and marker peaks to perform bulk ATAC-Seq
292 deconvolution. The resulting predictions are referred to as DeconPeaker-Custom and CIBERSORTx-
293 Custom.
294 Many tools displayed high correlation and low RMSE values, similar to those of EPIC-ATAC, and no
295 single tool consistently outperformed the others (Figure 3D-E, Figure 3-figure supplement 2). The fact
296 that our marker peaks and reference profiles could be used with EPIC-ATAC and other existing tools
297 demonstrates their broad applicability.
298

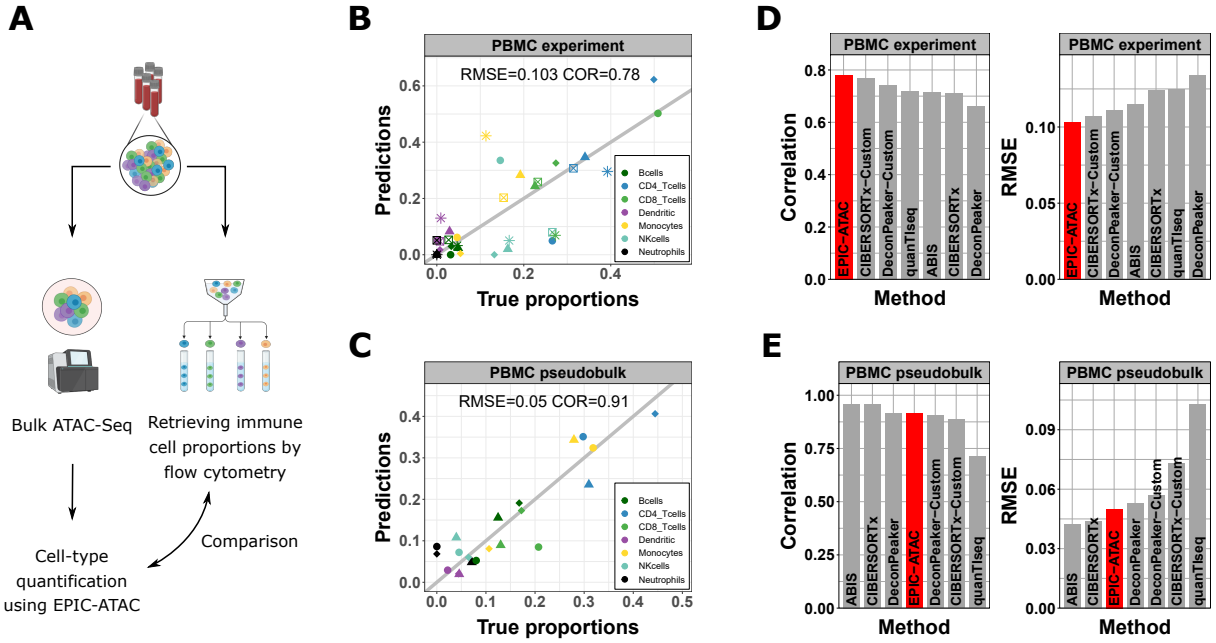


Figure 3: EPIC-ATAC accurately estimates immune cell fractions in PBMC ATAC-Seq samples. A) *Schematic description of the experiment designed to validate the ATAC-Seq deconvolution on PBMC samples (Created with BioRender.com).* ***B)*** *Comparison between cell-type proportions predicted by EPIC-ATAC 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.*

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

We evaluated the ability of the EPIC-ATAC framework to predict not only immune and stromal cells proportions but also the proportion of cells for which reference profiles are not available (*i.e.,* uncharacterized cells). For this purpose, we considered two previously published scATAC-Seq datasets containing basal cell carcinoma and gynecological cancer samples (Satpathy *et al.*, 2019; Regner *et al.*, 2021) as well as the samples from the Human Tumor Atlas Network (HTAN) single-cell multiome dataset composed of samples from diverse cancer types (Terekhanova *et al.*, 2023). For each dataset,

317 we generated pseudobulks by averaging the chromatin accessibility signal across all cells of each
318 sample (see Material and methods, section 3). Applying EPIC-ATAC to each dataset shows that this
319 framework can simultaneously predict the proportions of both uncharacterized cells and immune,
320 stromal and vascular cells (Figure 4A). In basal cell carcinoma and gynecological cancer samples, the
321 proportion of uncharacterized cells can be seen as a proxy of the proportion of cancer cells. In the
322 HTAN dataset, some of the samples also contain cell-types that are neither immune, fibroblasts or
323 endothelial cells nor malignant cells. Hence, the uncharacterized cells in these samples group cancer
324 and normal cells from the tumor site. EPIC-ATAC, in this context, was able to not only estimate the
325 proportion of cell types included in the TME reference profiles but also the proportion of
326 uncharacterized cells in most of the cancer types (Figure 4A, Figure 4-figure supplement 1).
327 As for the PBMC datasets, we compared EPIC-ATAC performances to other existing deconvolution tools
328 (see Materials and methods, section 4). For each dataset, EPIC-ATAC led to the highest performances
329 and was the only method to accurately predict the proportion of uncharacterized cells (Figure 4B,
330 Figure 4-figure supplements 2, 3 and 4). Although quanTIseq also allows users to perform such
331 predictions, the method resulted in lower correlation and higher RMSE values when comparing the
332 estimated and true proportions of the uncharacterized cells (Figure 4B, Figure 4-figure supplements 2
333 and 4).
334 In the EPIC-ATAC and quanTIseq frameworks, predictions correspond to absolute cell-type fractions
335 which include the proportion of uncharacterized cells, *i.e.,* proportions of all cells present in the bulk,
336 while the estimations obtained from the other tools correspond to relative cell fractions, *i.e.,*
337 proportions of cells present in the reference profiles (CIBERSORTx, DeconPeaker) or to scores with
338 arbitrary units (ABIS, MCPcounter) without considering the presence of uncharacterized cells. For the
339 latter group of tools, we fixed the uncharacterized cells estimations to 0. This approach provides a clear
340 and significant advantage to EPIC-ATAC and to quanTIseq (Figure 4B). For this reason, we conducted a
341 second benchmark excluding the predictions of uncharacterized cell fractions and rescaling both
342 estimations and true proportions to sum to 1 (see Material and methods, section 4). EPIC-ATAC

outperformed most of the other methods also when excluding the uncharacterized cells (Figure 4C, Figure 4-figure supplements 2, 3 and 4). Note that in terms of computational resources, all the tools require a limited amount of time and memory to run, *i.e.*, less than 20 seconds in average (Figure 4-figure supplement 5).

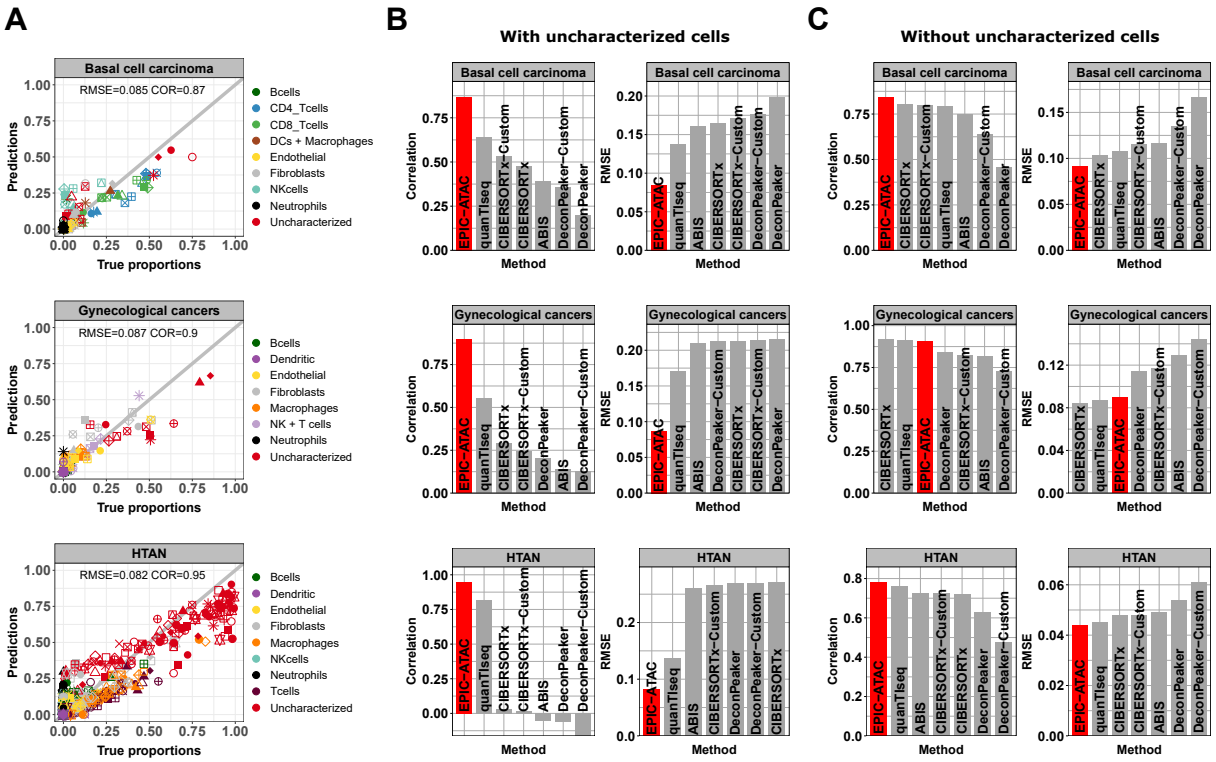


Figure 4: EPIC-ATAC accurately predicts fractions of cancer and non-malignant cells in tumor samples. A) *Comparison between cell-type proportions estimated by EPIC-ATAC and true proportions for the basal cell carcinoma (top), gynecological (middle) and HTAN (bottom) pseudobulk datasets. Symbols correspond to pseudobulks.* ***B)*** *Pearson's correlation and RMSE values obtained for the deconvolution tools included in the benchmark. EPIC-ATAC is highlighted in red.* ***C)*** *Same analyses as in panels B, with the uncharacterized cell population excluded for the evaluation of the prediction accuracy. The predicted and true proportions of the immune, stromal and vascular cell types were rescaled to sum to 1.*

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

360 To investigate the impact of cell type abundance on the accuracy of ATAC-Seq deconvolution, we
361 evaluated EPIC-ATAC predictions in each major cell type separately in the different benchmarking
362 datasets (Figure 5A). NK cells, endothelial cells, neutrophils or dendritic cells showed lower correlation
363 values. These values can be explained by the fact that these cell types are low-abundant in our
364 benchmarking datasets (Figure 5A). For the endothelial cells and dendritic cells, the RMSE values
365 associated to these cell types remain low. This suggests that while the predictions of EPIC-ATAC might
366 not be precise enough to compare these cell-type proportions between different samples, the cell-
367 type quantification within each sample is reliable. For the NK cells and the neutrophils, we observed
368 more variability with higher RMSE values in some datasets which suggests that the markers and
369 profiles for these cell types might be improved. Figure 5-figure supplements 1 and 2 detail the
370 performances of each tool when considering each cell type separately in the PBMC and the cancer
371 datasets. As for EPIC-ATAC, the predictions from the other deconvolution tools are more reliable for
372 the frequent cell types.

373

374 To explore the accuracy of ATAC-Seq deconvolution for more related cell types, we evaluated whether
375 EPIC-ATAC could predict the proportions of T-cell subtypes. To this end, we considered naive and non-
376 naive CD8+ as well as naive, helper/memory and regulatory CD4+ T cells (Tregs). We redefined our list
377 of cell-type specific marker peaks and reference profiles including also these five T-cell subtypes
378 (Supplementary files 8-9, Figure 5-figure supplement 3A) and observed that the markers were
379 conserved in external data (Figure 5-figure supplement 3B). The annotations of the markers associated
380 to the T-cell subtypes are available in Supplementary files 10-13.
381 We capitalized on the more detailed cell-type annotation of the PBMC datasets as well as the basal cell
382 carcinoma dataset to evaluate the EPIC-ATAC predictions of cell-subtype fractions using these updated
383 markers and profiles. Overall, the correlations observed between the predictions and true proportions

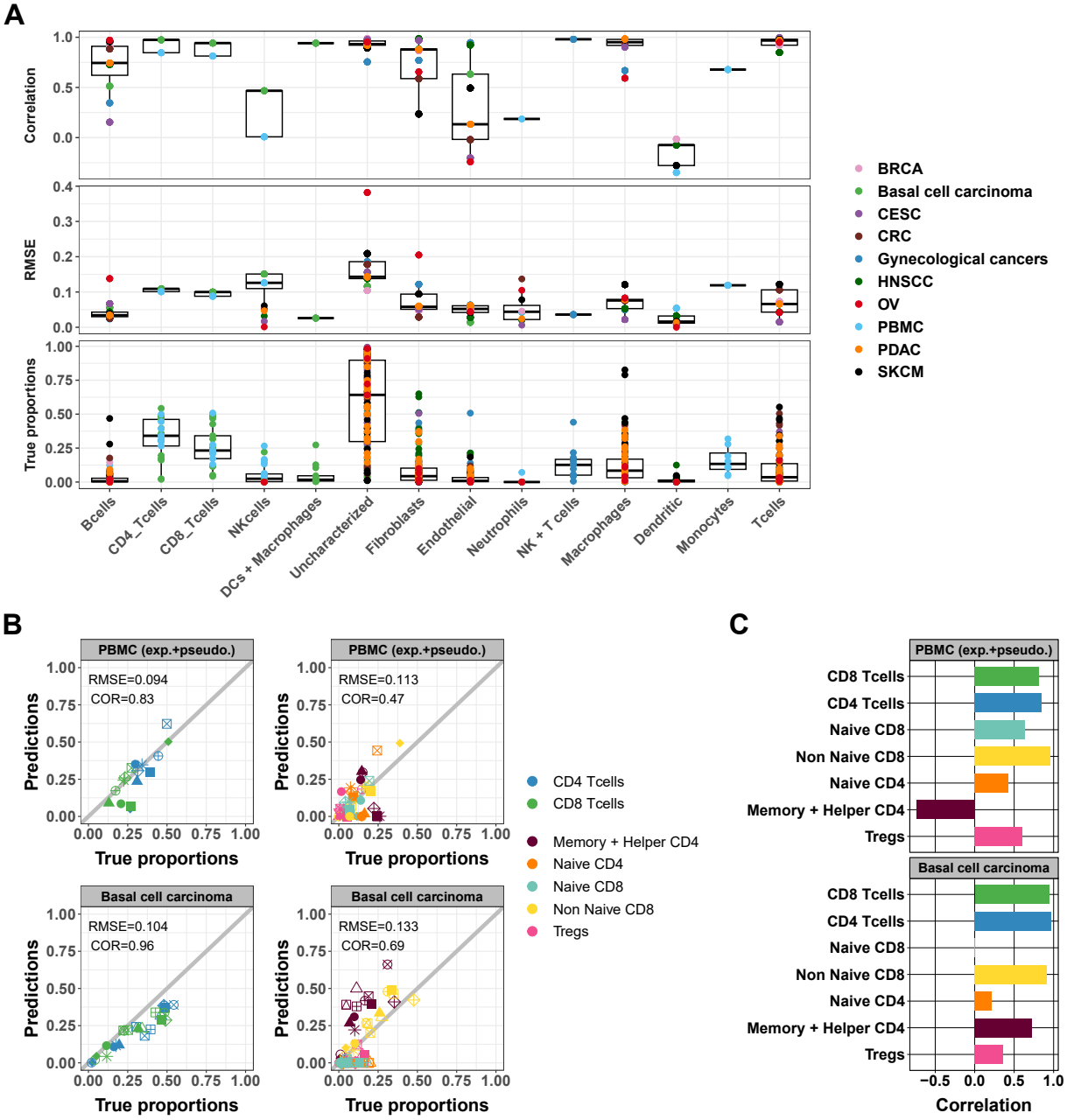
384 of T cells decreased when considering T-cell subtypes rather than CD4+ and CD8+ cell types only (Figure

385 5B). In particular, low accuracies were obtained for helper/memory CD4+ and naive T-cell subtypes

386 (Figure 5C). Similar results were obtained using other deconvolution tools (Figure 5-figure supplement

387 4).

388



389

390 ***Figure 5: Accuracy of ATAC-Seq deconvolution is determined by the abundance and specificity of each***

391 ***cell type. A)*** *Correlations (top) and RMSE (middle) between EPIC-ATAC predictions and true cell-type*

392 *proportions in each cell-type. True proportions are also shown for each cell type (bottom). Colors*

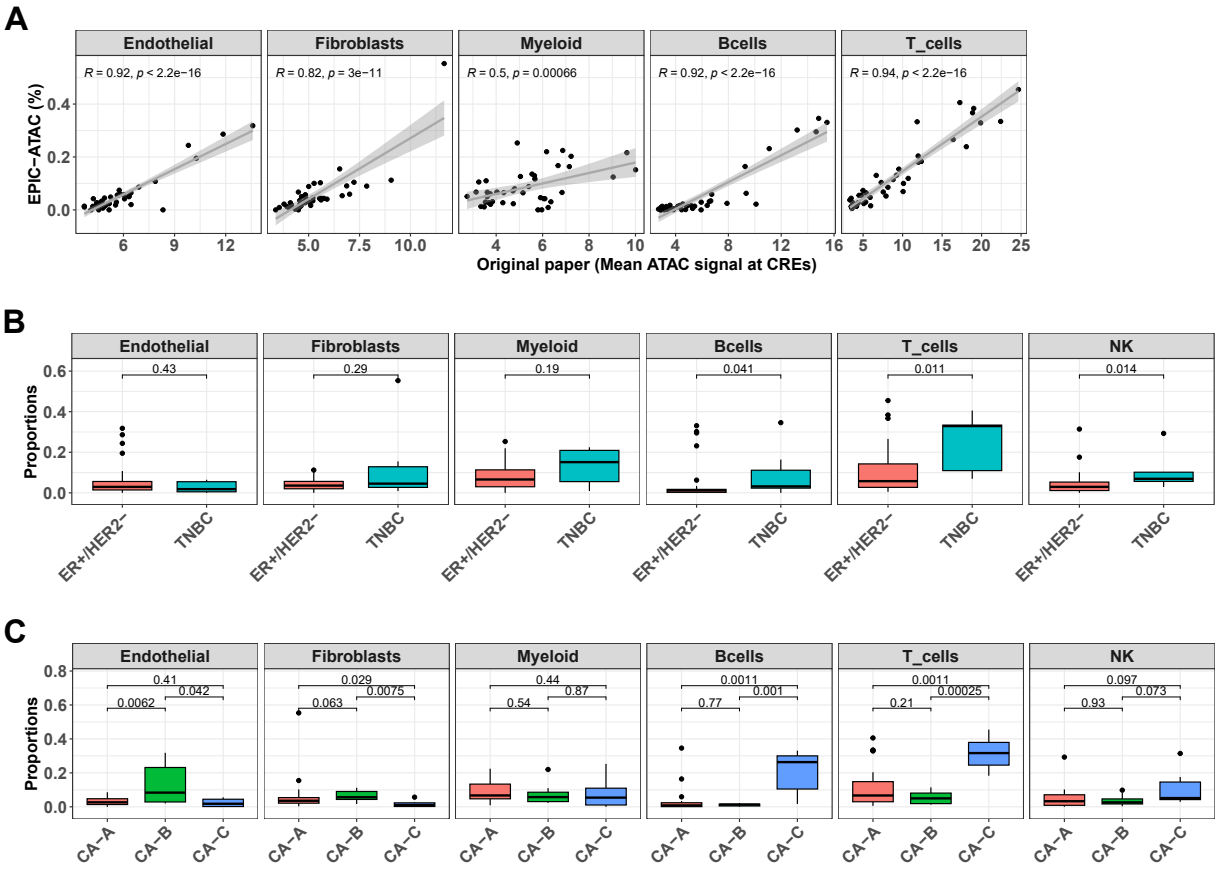
393 *correspond to different datasets.* ***B)*** *Comparison of the proportions estimated by EPIC-ATAC and the*
394 *true proportions for PBMC samples (PBMC experiment and PBMC pseudobulk samples combined) (top)*
395 *and the basal cell carcinoma pseudobulks (bottom). Predictions of the proportions of CD4+ and CD8+*
396 *T-cells were obtained using the reference profiles based on the major cell types and subtypes*
397 *predictions using the reference profiles including the T-cell subtypes.* ***C)*** *Pearson's correlation values*
398 *obtained by EPIC-ATAC in each cell type.*
399

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

401 We applied EPIC-ATAC to a breast cancer cohort of 42 breast ATAC-Seq samples including samples
402 from two breast cancer subtypes, *i.e.,* 35 oestrogen receptor (ER)-positive human epidermal growth
403 factor receptor 2 (HER2)-negative (ER+/HER2-) breast cancer samples and 7 triple negative breast
404 cancer (TNBC) samples (Kumegawa *et al.*, 2023). No cell sorting was performed in parallel to the
405 chromatin accessibility sequencing. For this reason, the authors used a set of cell-type-specific cis-
406 regulatory elements (CREs) identified in scATAC-Seq data from similar breast cancer samples
407 (Kumegawa *et al.*, 2022) and estimated the amount of infiltration of each cell type by averaging the
408 ATAC-Seq signal of each set of cell-type-specific CREs in their samples. We used EPIC-ATAC to estimate
409 the proportions of different cell types of the TME. These predictions were then compared to the metric
410 used by Kumegawa and colleagues in their study to infer levels of infiltration. A high correlation
411 between the two metrics was observed for each cell type (Pearson's correlation coefficient from 0.5
412 for myeloid cells to 0.94 for T cells, Figure 6A).
413 We observed a higher proportion of T cells, B cells and NK cells in the TNBC samples in comparison to
414 ER+/HER2- samples (Figure 6B). We then compared the cellular composition of ER+/HER2- subgroups
415 identified in the original study (clusters CA-A, CA-B and CA-C). A higher infiltration of T and B cells was
416 observed in cluster CA-C and higher proportions of endothelial cells and fibroblasts were observed in
417 cluster CA-B (Figure 6C). These differences recapitulate those reported in the Kumegawa study except
418 for the difference in myeloid infiltration observed in the original study between the different breast

419 cancer subgroups (Kumegawa *et al.*, 2023). However, when considering each myeloid cell type present
420 in our reference profiles separately, a higher infiltration of macrophages was observed in the TNBC
421 samples in comparison to the ER+/HER2- samples (Figure 6-figure supplement 1). Also, we observed a
422 difference in the levels of NK cells infiltration between TNBC and ER+/HER2- samples while no NK
423 infiltration estimation was provided in the original paper for this cell type.

424



425

426 ***Figure 6: EPIC-ATAC accurately infers the immune contexture in a bulk ATAC-Seq breast cancer***
427 ***cohort. A)*** *Proportions of different cell types predicted by EPIC-ATAC in each sample as a function of*
428 *the average ATAC signal at the cell-type specific CREs used by Kumegawa et al.* (Kumegawa *et al.*, 2023)
429 *to infer the level of cell-type infiltration in the tumor samples.* ***B)*** *Proportions of different cell types*
430 *predicted by EPIC-ATAC in the samples stratified based on two breast cancer subtypes.* ***C)*** *Proportions*
431 *of different cell types predicted by EPIC-ATAC in the samples stratified based on three ER+/HER2-*
432 *subgroups. Wilcoxon test p-values are represented at the top of the boxplots.*

433

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

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

449 We also applied EPIC-ATAC and the original version of EPIC (EPIC-RNA) on a bulk dataset composed of
450 PBMC samples with matched bulk RNA-Seq, bulk ATAC-Seq and flow cytometry data (Morandini *et al.*,
451 2024) (Figure 7C). We compared the predictions obtained using each modality to the flow cytometry
452 cell-type quantifications. EPIC-ATAC predictions were better correlated with the flow cytometry
453 measures for some cell types (*e.g.*, CD8+, CD4+ T cells, NK cells) while this trend was observed with the
454 EPIC-RNA predictions in other cell types (B cells, neutrophils, monocytes) (Figure 7C). We then tested
455 whether the predictions obtained from both modalities could be combined to improve the accuracy of
456 each cell-type quantification. Averaging the predictions obtained from both modalities shows a
457 moderate improvement (Figure 7C), suggesting that the two modalities can complement each other.
458

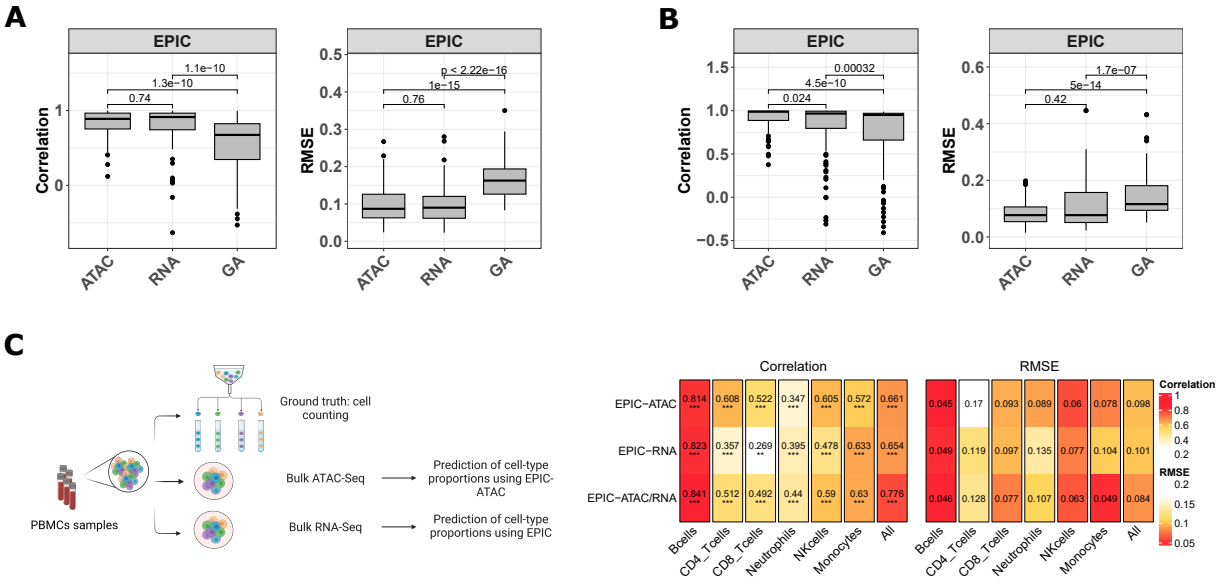


Figure 7: EPIC-ATAC and EPIC RNA-seq based deconvolution have similar accuracy and can complement each other. A-B) *Pearson's correlation (left) and RMSE (right) values comparing the proportions predicted by the ATAC-Seq deconvolution, the RNA-Seq deconvolution and the GA-based RNA deconvolution and true cell-type proportions in 100 pseudobulks simulated form the 10x multiome PBMC dataset* (10x Genomics, 2021) *(panel A) and in the pseudobulks generated from the HTAN cohorts (panel B). Dots correspond to outlier pseudobulks.* ***C)*** *Left panel: Schematic description of the dataset from Morandini et al.* (Morandini *et al.*, 2024) *composed of matched bulk ATAC-Seq, RNA-Seq and flow cytometry data from PBMC samples (Created with BioRender.com). Right panel: Pearson's correlation (left) and RMSE values (right) obtained by EPIC-ATAC, EPIC-RNA and EPIC-ATAC/RNA (i.e., averaged predictions of EPIC-ATAC and EPIC-RNA) in each cell type separately or all cell types together (columns).*

Discussion

Bulk chromatin accessibility profiling of biological tissues like tumors represents a reliable and affordable technology to map the activity of gene regulatory elements across multiple samples in different conditions. Here, we collected ATAC-Seq data from pure cell populations covering major immune and non-immune cell types found in the tumor micro-environment. This enabled us to identify

477 reliable cell-type specific marker peaks and chromatin accessibility profiles for both PBMC and solid
478 tumor sample deconvolution. We integrated these data in the EPIC deconvolution framework to
479 accurately predict the fraction of both malignant and non-malignant cell types from bulk tumor ATAC-
480 Seq samples.

481 Although not tested in this work, the TME marker peaks and profiles could be used on normal
482 tissues where immune cells are expected to be present. In cases where specific cell types are expected
483 in a sample but are not part of our list of reference profiles (*e.g.,* neuronal cells in brain tumors or
484 tissues other than human PBMCs or tumor samples), custom marker peaks and reference profiles can
485 be provided to EPIC-ATAC to perform cell-type deconvolution. To this end, users should select markers
486 that are cell-type specific, which could be identified using pairwise differential analysis performed on
487 ATAC-Seq data from sorted cells from the populations of interest, following the approach developed
488 in this work (Figure 1, see Code availability).

489 Solid tumors contain large and heterogeneous fractions of cancer cells for which it is
490 challenging to build reference profiles. We show that the EPIC-ATAC framework, in contrast to other
491 existing tools, allows users to accurately predict the proportion of cells not included in the reference
492 profiles (Figure 4 and Supplementary Figure 4). These uncharacterized cells can include cancer cells
493 but also other non-malignant cells. Since the major cell types composing TMEs were included in our
494 reference profiles, the proportion of uncharacterized cells approximates the proportion of the cancer
495 cells in most cases.

496 For our benchmarking, we provided our reference profiles and markers to each tool that do
497 not provide the option to automatically build new profiles from a set of reference samples. While this
498 allowed us to run deconvolution tools previously developed on RNA-Seq data with their default
499 parameters, we cannot exclude that hyper parametrization could improve the performance of these
500 tools, and we anticipate that the cell-type specific ATAC-Seq profiles and markers curated in this work
501 could help improving bulk ATAC-Seq deconvolution with different tools. Also, for RNA-Seq data
502 deconvolution, some of the methods followed specific pre-processing steps, *e.g.,* the quanTIseq

503 framework removes a manually curated list of noisy genes as well as aberrant immune genes identified
504 in the TCGA data, and ABIS uses immune-specific housekeeping genes. Additional filtering could be
505 explored to improve the performance of deconvolution tools in the context of ATAC-Seq data.

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

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

521 Another limitation of cell-type deconvolution is often reached when closely related cell types
522 are considered. In the reference-based methods used in this study, this limit was reached when
523 considering T-cell subtypes in the reference profiles (Figures 5B-C and Supplementary Figure 12). We
524 thus recommend to use the EPIC-ATAC framework with the markers and reference profiles based on
525 the major cell-type populations. We additionally provide the marker peaks of the T-cell subtypes which
526 could be used to build cell-type specific chromatin accessibility signatures or perform "peak set
527 enrichment analysis" similarly to gene set enrichment analysis (Subramanian *et al.*, 2005). Such an

528 application could be useful for the annotation of scATAC-Seq data, which often relies on matched RNA-
529 Seq data and for which there is a lack of markers at the peak level (Jiang *et al.*, 2023).

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

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

554

555 ## Conclusion

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

564

565 ## Materials and methods

566 **1- <u>Generation of an ATAC-Seq reference dataset of non-malignant cell types frequently</u>**
567 **<u>observed in the tumor microenvironment.</u>**

568 **Pre-processing of the sorted ATAC-Seq datasets**

569 We collected pure ATAC-Seq samples from 12 studies. The data include samples from (i) ten major
570 immune, stromal and vascular cell types (B (Corces *et al.*, 2016; Calderon *et al.*, 2019; P. Zhang *et al.*,
571 2022), CD4+ (Corces *et al.*, 2016; Mumbach *et al.*, 2017; Liu *et al.*, 2020; Giles *et al.*, 2022; P. Zhang *et*
572 *al.*, 2022), CD8+ (Corces *et al.*, 2016; Calderon *et al.*, 2019; Liu *et al.*, 2020; Giles *et al.*, 2022; P. Zhang
573 *et al.*, 2022), natural killer (NK) (Corces *et al.*, 2016; Calderon *et al.*, 2019), dendritic (DCs) cells
574 (Calderon *et al.*, 2019; Leylek *et al.*, 2020; Liu *et al.*, 2020), macrophages (Liu *et al.*, 2020; P. Zhang *et*
575 *al.*, 2022), monocytes (Corces *et al.*, 2016; Calderon *et al.*, 2019; Leylek *et al.*, 2020; Trizzino *et al.*,
576 2021; P. Zhang *et al.*, 2022) and neutrophils (Perez *et al.*, 2020; Ram-Mohan *et al.*, 2021) as well as
577 fibroblasts (Liu *et al.*, 2020; Ge *et al.*, 2021) and endothelial (Liu *et al.*, 2020; Xin *et al.*, 2020) cells (See
578 Figure 2A), and (ii) eight tissues from distinct organs (*i.e.* bladder, breast, colon, liver, lung, ovary,

579 pancreas and thyroid) from the ENCODE data (The ENCODE Project Consortium *et al.*, 2020; Rozowsky
580 *et al.*, 2023). The list of the samples and their associated metadata (including cell types and accession
581 number of the study of origin) is provided in Supplementary file 1. To limit batch effects, the samples
582 were reprocessed homogeneously from the raw data (fastq files) processing to the peak calling. For
583 that purpose, raw fastq files were collected from GEO using the SRA toolkit and the PEPATAC
584 framework (Smith *et al.*, 2021) was used to process the raw fastq files based on the following tools:
585 trimmomatic for adapter trimming, bowtie2 (with the PEPATAC default parameters) for reads pre-
586 alignment on human repeats and human mitochondrial reference genome, bowtie2 (with the default
587 PEPATAC parameters: *--very-sensitive -X 2000*) for alignment on the human genome (hg38), samtools
588 (PEPATAC default parameters: *-q 10*) for duplicates removal and MACS2 (Zhang *et al.*, 2008) (PEPATAC
589 default parameters: *--shift -75 --extsize 150 --nomodel --call-summits --nolambda --keep-dup all -p*
590 *0.01*) for peak calling in each sample. After alignment, reads mapping on chromosome M were
591 excluded. TSS enrichment scores were computed for each sample and used to filter out samples with
592 low quality (criteria of exclusion: TSS score < 5) (See Supplementary file 1 containing the TSS score of
593 each sample). 789 samples (including 564 from our ten reference cell-types) had a TSS score > 5.
594

595 **Generation of a consensus set of peaks**

596 Peak calling was performed in each sample individually. Peaks were then iteratively collapsed to
597 generate a set of reproducible peaks. For each cell type, peaks collapse was performed adapting the
598 iterative overlap peak merging approach proposed in the PEPATAC framework. A first peaks collapse
599 was performed at the level of each study of origin, *i.e.,* if peaks identified in distinct samples overlapped
600 (minimum overlap of 1bp between peaks), only the peak with the highest peak calling score was kept.
601 Also, only peaks detected in at least half of the samples of each study were considered for the next
602 step. If a study had only two samples, only peaks detected in both samples were considered. After this
603 first selection, a second round of peaks collapse was performed at the cell-type level to limit batch
604 effects in downstream analyses. For each cell type, only peaks detected in all the studies of origin were

605 considered. The final list of peaks was then generated by merging each set of reproducible peaks. Peaks
606 located on chromosome Y were excluded from the rest of the analyses. ATAC-Seq counts were
607 retrieved for each sample and each peak using featureCounts (Liao, Smyth and Shi, 2014).
608

609 ## 2- Identification of cell-type specific markers

610 **Differential accessibility analysis**

611 To identify cell-type specific markers, we split the samples collection in ten folds (created with the
612 *create_folds* function from the R package splitTools (Mayer, 2023)). For each fold, we performed
613 pairwise differential accessibility analysis across the ten cell types considered in the reference samples
614 as well as the ENCODE samples from diverse organs. The differential analysis was performed using
615 limma ((Ritchie *et al.*, 2015), version 3.56.2). Effective library sizes were computed using the method
616 of trimmed mean of M-values (TMM) from the edgeR package in R ((Robinson, McCarthy and Smyth,
617 2010), version 3.42.4). Due to differences of library size across all samples collected, we used voom
618 from the limma package (Law *et al.*, 2014) to transform the data and model the mean-variance
619 relationship. Finally, a linear model was fitted to the data to assess the differential accessibility of each
620 peak across each pair of cell types. To identify our marker peaks, all peaks with log2 fold change higher
621 than 0.2 were selected and ranked by their maximum adjusted *p*-value across all pairwise comparisons.
622 The top 200 features (with the lowest maximum adjusted *p*-value) were considered as cell-type specific
623 marker peaks. The marker peaks identified in at least three folds were considered in the final list of
624 marker peaks.
625

626 **Marker peaks filtering**

627 Modules of open chromatin regions accessible in all (universal modules) or in specific human tissues
628 have been identified in the study Zhang *et al.* (K. Zhang *et al.*, 2021). These regions were used to refine
629 the set of marker peaks and exclude peaks with residual accessibility in other cell types than those
630 considered for deconvolution. More precisely, for immune, endothelial and fibroblasts specific peaks,

631 we filtered out the peaks overlapping the universal modules as well as the tissue specific modules
632 except the immune (modules 8 to 25), endothelial (modules 26 to 35) and stromal related modules
633 (modules 41 to 49 and 139-150) respectively. As a second filtering step, we retained markers exhibiting
634 the highest correlation patterns in tumor bulk samples from different solid cancer types, *i.e.,* The
635 Cancer Genome Atlas (TCGA) samples (Corces *et al.*, 2018). We used the Cancer Genomics Cloud (CGC)
636 (Lau *et al.*, 2017) to retrieve the ATAC-Seq counts for each marker peaks in each TCGA sample (using
637 *featureCounts*). For each set of cell-type specific peaks, we identified the most correlated peaks using
638 the *findCorrelation* function of the caret R package ((Kuhn, 2008), version 6.0-94) with a correlation
639 cutoff value corresponding to the 90th percentile of pairwise Pearson correlation values.
640

641 **Evaluation of the study of origin batch effect**

642 To identify potential batch effect issues, we run principal component analysis (PCA) based on the cell-
643 type specific peaks after normalizing ATAC-Seq counts using full quantile normalization (FQ-FQ)
644 implemented in the EDASeq R package (Risso *et al.*, 2011) to correct for depth and GC biases. These
645 data were used to visualize the data in three dimensions using the first axes of the PCA based on the
646 PBMC and TME markers (Figure 2B and Figure 2-figure supplement 1). We also used the ten first
647 principal components to evaluate distances between samples and compute silhouette coefficients
648 based on the cell type and study of origin classifications as well as to compute the ARI metric comparing
649 the cell-type annotation and the clustering obtained using model based clustering.
650

651 **Building the reference profiles**

652 It has been previously demonstrated in the context of RNA-Seq based deconvolution approaches
653 (Racle *et al.*, 2017; Sturm *et al.*, 2019) that the transcripts per million (TPM) transformation is
654 appropriate to estimate cell fractions from bulk mixtures. We thus normalized the ATAC-Seq counts of
655 the reference samples using a TPM-like transformation, *i.e.,* dividing counts by peak length, correcting
656 samples counts for depth and rescaling counts so that the counts of each sample sum to 10^6. We then

657 computed for each peak the median of the TPM-like counts across all samples from each cell type to
658 build the reference profiles of the ten cell types considered in the EPIC-ATAC framework (Figure 2C).
659 In the EPIC algorithm, weights reflecting the variability of each feature of the reference profile can be
660 considered in the constrained least square optimization. We thus also computed the inter-quartile
661 range of the TPM-like counts for each feature in each cell type. Two ATAC-Seq reference profiles are
662 available in the EPIC-ATAC framework: (i) a reference profile containing profiles for B cells, CD4+ T cells,
663 CD8+ T cells, NK, monocytes, dendritic cells and neutrophils to deconvolve PBMC samples, and (ii) a
664 reference profile containing profiles for B cells, CD4+ T cells, CD8+ T cells, NK, dendritic cells,
665 macrophages, neutrophils, fibroblasts and endothelial cells to deconvolve tumor samples. The
666 reference profiles are available in the EPICATAC R package and the reference profiles restricted to our
667 cell-type specific marker peaks are available in the Supplementary files 2 and 3.
668
669 **Assessing the reproducibility of the marker peaks signal in independent samples**
670 We evaluated the chromatin accessibility level of the marker peaks in samples that were not included
671 in the peak calling step. Firstly, we considered samples from two independent studies (Ucar *et al.*,
672 2017; Carvalho *et al.*, 2021) providing pure ATAC-Seq data for five immune cell types (*i.e.,* B, CD4+ T
673 cells, CD8+ T cells, Monocytes, Macrophages) (Figure 2D). To consider the other cell types, samples
674 that were excluded from the reference dataset due to a low TSS enrichment score were also
675 considered in this validation dataset (Supplementary file 1). Secondly, we collected the data from a
676 single-cell atlas chromatin accessibility from human tissues and considered the cell types included in
677 our reference data (K. Zhang *et al.*, 2021) (Figure 2E). We used the cell-type annotations provided in
678 the original study (GEO accession number: GSE184462). The Signac R package ((Stuart *et al.*, 2021),
679 1.9.0) was used to extract fragments counts for each cell and each marker peak and the ATAC-Seq
680 signal of each marker peak was averaged across all cells of each cell type.
681
682 **Annotation of the marker peaks**

683 The cell-type specific markers were annotated using ChIPseeker R package ((Yu, Wang and He, 2015),
684 version 1.34.1) and the annotation from *TxDb.Hsapiens.UCSC.hg38.knownGene* in R to identify the
685 regions in which the marker peaks are (*i.e.,* promoter, intronic regions, etc.) and ChIP-Enrich to
686 associate each peak to the nearest gene TSS (Welch *et al.*, 2014). The nearest genes identified were
687 then compared to cell-type marker genes listed in the PanglaoDB (Franzén, Gan and Björkegren, 2019)
688 and CellMarker databases (Hu *et al.*, 2023). PanglaoDB provides an online interface to explore a large
689 collection on single-cell RNA-Seq data as well as a community-curated list of cell-type marker genes.
690 CellMarker is a database providing a large set of curated cell-type markers for more than 400 cell types
691 in human tissues retrieved from a large collection of single-cell studies and flow cytometry,
692 immunostaining or experimental studies. ChIP-Enrich was also used to perform gene set enrichment
693 and identify for each set of cell-type specific peaks potential biological pathways regulated by the
694 marker peaks. The enrichment analysis was performed using the *chipenrich* function (*genesets =*
695 *"GOBP" , locusdef = "nearest_tss"*) from the chipenrich R package (v2.22.0).
696 Chromatin accessibility peaks can also be annotated for chromatin binding proteins (CBPs) such as
697 transcription factors (TFs), whose potential binding in the peak region is reported in databases. In our
698 study we chose the JASPAR2022 (Castro-Mondragon *et al.*, 2022) database and the ReMap database
699 (Hammal *et al.*, 2022).
700 Using the JASPAR2022 database, we assessed, for each cell type, whether the cell-type specific marker
701 peaks were enriched in specific TFs motifs using two TFs enrichment analysis frameworks: Signac
702 (Stuart *et al.*, 2021) and MonaLisa (Machlab *et al.*, 2022). For the MonaLisa analysis, the cell-types
703 specific markers peaks were categorized in bins of sequences, one bin per cell type (use of the
704 *calcBinnedMotifEnrR* function). To test for an enrichment of motifs, the sequences of each bin were
705 compared to a set of background peaks with similar average size and GC composition obtained by
706 randomly sampling regions in all the peaks identified from the reference dataset. The enrichment test
707 was based on a binomial test. For the Signac analysis, we used the *FindMotif* function to identify over-
708 represented TF motifs in each set of cell-type specific marker peaks (query). This function used a

709 hypergeometric test to compare the number of query peaks containing the motif with the total
710 number of peaks containing the motif in the background regions (matched to have similar GC content,
711 region length and dinucleotide frequencies as the query regions), corresponding in our case to the
712 peaks called in the reference dataset.
713 The ReMap database associates chromatin binding proteins (CBPs), including TFs, transcriptional
714 coactivators and chromatin-remodeling factors, to their DNA binding regions based on DNA-binding
715 experiments such as chromatin immunoprecipitation followed by sequencing (ChIP-seq). For each
716 association of a CBP to its binding region, the cell type in which the binding has been observed is
717 reported in the ReMap database (biotype). We used the ReMapEnrich R package (version 0.99) to test
718 if the cell-type specific marker peaks are significantly enriched in CBPs-binding regions listed in the
719 ReMap 2022 catalog. We considered the non-redundant peaks catalog from ReMmap 2022, containing
720 non-redundant binding regions for each CBP in each biotype. Similarly to the previously mentioned
721 enrichment methods, we chose the consensus peaks called in the reference samples as universe for
722 the enrichment test. Note that, for each cell type, an enrichment was retained only if the biotype in
723 which the CBP-regions were identified matched the correct cell-type.
724
725
726 **3- <u>Datasets used for the evaluation of ATAC-Seq deconvolution</u>**
727 **PBMCs ATAC-Seq data from healthy donors**
728 ***Peripheral blood mononuclear cell (PBMC) isolation***
729 Venous blood from five healthy donors was collected at the local blood transfusion center of Geneva
730 in Switzerland, under the approval of the Geneva University Hospital's Institute Review Board, upon
731 written informed consent and in accordance with the Declaration of Helsinki. PBMCs were freshly
732 isolated by Lymphoprep (Promega) centrifugation (1800 rpm, 20 minutes, without break, room
733 temperature). Red blood cell lysis was performed using red blood lysis buffer (Qiagen) and platelets

734 were removed by centrifugation (1000 rpm, 10 minutes without break, room temperature). Cells were
735 counted and immediately used.
736
737 ***Flow cytometry***
738 Immune cell populations were identified using multiparameter flow cytometry and the following
739 antibodies: FITC anti-human CD45RA (HI100, Biolegend), PerCP-Cyanine5.5 anti-human CD19 (H1B19,
740 Biolegend), PE anti-human CD3 (SK7, Biolegend), PE-Dazzle anti-human CD14 (M0P9, BD Biosciences),
741 PE-Cyanine7 anti-human CD56 (HCD56, Biolegend), APC anti-human CD4 (RPA-T4, Biolgend), APC-
742 Cyanine7 anti-human CCR7 (G043H7, Biolegend), Brilliant Violet 421 anti-human CD8 (RPA-T8,
743 Biolegend), Brilliant Violet 510 anti-human CD25 (BC96, Biolegend), Brilliant Violet 711 anti-human
744 CD16 (3G8, Biolegend), Brilliant Violet 786 anti-human CD127 (A019D5, Biolegend), Ultra-Brilliant
745 Violet anti-human CD45 (HI30, BD Biosciences), FITC anti-human Celc9a (8F9, Miltenyi) , PE anti-human
746 XCR1 (S15046E, Biolegend), PE-Dazzle anti-human BDCA-2 (201A, Biolegend), APC anti-human BDCA-
747 3 (AD5-14H12, Miltenyi), Brilliant Violet 421 anti-human CD3 (UCHT1, Biolegend), Brilliant Violet 421
748 anti-human CD14 (M5E2, BD Pharmingen), Brilliant Violet 421 anti-human CD19 (SJ25C1, Biolegend),
749 Brilliant Violet 510 anti-human BDCA-1 (L161, Biolegend), Brilliant Violet 650 anti-human CD11c (3.9,
750 Biolegend), Brilliant Violet 711 anti-human CD11c (N418, Biolegend) and Brilliant Violet 711 anti-
751 human HLA-DR (L243, Biolegend). Dead cells were excluded using the Zombie UV™ Fixable Viability Kit
752 (Biolegend). Intracellular staining was performed after fixation and permeabilization of the cells with
753 the FoxP3 Transcription Factor Staining Buffer Set (00-5523-00, Invitrogen) using Alexa 700 anti-human
754 FoxP3 antibody (259D/C7, BD Biosciences). Data were acquired on LSRFortessa flow cytometer and
755 analysed using FlowJo software (v10.7.1).
756
757 ***Cell preparation for ATAC-Sequencing***
758 50000 CD45+ cells were sorted from total PBMCs using anti-human Ultra-Brilliant Violet (BUV395)
759 CD45 (HI30, BD Biosciences) with a FACSAria II (Becton Dickinson) and were collected in PBS with 10%

760 Foetal Bovine Serum (FBS). Cell pellets were resuspended in cold lysis buffer (10mM Tris-Cl pH 7.4,
761 10mM NaCl, 3mM MgCl2, 0,1% NP40 and water) and immediately centrifuged at 600g for 30min at
762 4°C. Transposition reaction was performed using the Illumina Tagment DNA Enzyme and Buffer kit
763 (20034210, Illumina) and transposed DNA was eluted using the MinElute PCR Purification Kit (Qiagen).
764 Libraries were generated by PCR amplification using indexing primers and NEBNext High-Fidelity
765 Master Mix (New England BioLabs) and were purified using AMPure XP beads (A63880, Beckman
766 Coulter). Libraries were quantified by a fluorometric method (QubIT, Life Technologies) and their
767 quality assessed on a Fragment Analyzer (Agilent Technologies). Sequencing was performed as a paired
768 end 50 cycles run on an Illumina NovaSeq 6000 (v1.5 reagents) at the Genomic Technologies Facility
769 (GTF) in Lausanne, Switzerland. Raw sequencing data were demultiplexed using the bcl2fastq2
770 Conversion Software (version 2.20, Illumina).
771
772 ***Data processing***
773 The same steps as for the processing of the reference ATAC-Seq samples were followed. (See Pre-
774 processing of the ATAC-Seq datasets).
775
776 **ATAC-Seq pseudobulk data from PBMCs and cancer samples**
777 To evaluate the accuracy of our ATAC-Seq deconvolution framework, we generated pseudo-bulk
778 datasets from 6 single-cell datasets:
779 - **PBMC pseudobulk dataset:** combination of three single-cell datasets for PBMCs.
780 - Dataset 1 corresponds to a scATAC-Seq dataset obtained from Satpathy et al.
781 (Satpathy *et al.*, 2019) (GEO accession number: GSE129785). This dataset contains
782 FACS-sorted populations of PBMCs. Since the cells of some cell types came from a
783 unique donor, all the cells of this dataset were aggregated to form one pseudobulk.
784 Ground truth cell fractions were obtained by dividing the number of cells in each cell
785 type by the total number of cells.

786 ○ Dataset 2 (included in the PBMC pseudobulk dataset) was retrieved from Granja *et al.*
787 (Granja *et al.*, 2019) (GEO accession number GSE139369). B cells, monocytes,
788 dendritic, CD8+, CD4+ T, NK cells, neutrophils from healthy donors were considered.
789 The neutrophil cells came from a single donor. As for dataset 1, we thus aggregated
790 all the cells to generate one pseudobulk. Ground truth cell fractions were obtained by
791 dividing the number of cells in each cell type by the total number of cells.
792 ○ Dataset 3 (included in the PBMC pseudobulk dataset) corresponds to the 10X
793 multiome dataset of PBMC cells (10x Genomics, 2021). Since these data come from
794 one donor, one pseudobulk sample was generated for this dataset. The pseudobulk
795 was generated by averaging the ATAC-Seq signal from all cells from the following cell
796 types: B cells, CD4+ T cells , CD8+ T cells, NK cells, Dendritic cells and monocytes.
797 • **Basal cell carcinoma dataset:** obtained from the study of Satpathy *et al.* (Satpathy *et al.*, 2019).
798 This dataset is a scATAC-Seq dataset composed of 13 basal cell carcinoma samples composed
799 of immune (B cells, plasma cells, CD4+ T cells, CD8+ T cells, NK cells, myeloid cells), stromal
800 (endothelial and fibroblasts) and cancer cells. Plasma cells and cancer cells were both
801 considered as uncharacterized cells (*i.e.,* cell types not included in the reference profiles). Cell
802 annotations were retrieved from the original study.
803 • **Gynecological cancer dataset:** obtained from the study of Regner *et al.* (Regner *et al.*, 2021)
804 (GEO accession number GSE173682). In this study, the authors performed scATAC-Seq on 11
805 gynecological cancer samples from two tumor sites (*i.e.,* endometrium and ovary) and
806 composed of immune (B cells, NK and T cells grouped under the same cell-type annotation,
807 macrophages, mast cells), stromal (fibroblast, endothelial, smooth muscle) and cancer cells.
808 Mast cells, smooth muscle and cancer cells were considered as uncharacterized cells. Cell
809 annotations were retrieved from the original study.
810 • **HTAN:** obtained from the study of Terekhanova et al. (Terekhanova *et al.*, 2023) from the HTAN
811 DCC Portal. In this study the authors generated an atlas of single-cell multiomic data (RNA-Seq

812 and ATAC-Seq profiling for the same cell) for diverse cancer types. We considered the
813 annotation provided by the HTAN consortium. This cell annotation was composed of immune
814 cells (B cells, T cells, dendritic cells, macrophages), fibroblasts, endothelial cells, cancer cells
815 and normal cells. The two latter cell groups were considered as part of the uncharacterized
816 cell population.

817 For Basal cell carcinoma, Gynecological cancer and the HTAN datasets, one pseudobulk per sample
818 was generated and ground truth cell fractions were obtained for each sample by dividing the number
819 of cells in each cell type by the total number of cells in the sample.

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

836

837 **Bulk ATAC-Seq data from a breast cancer cohort**

838 Bulk ATAC-Seq samples from a breast cancer cohort was obtained from Kumegawa *et al.* (Kumegawa
839 *et al.*, 2023). These data include 42 breast cancer samples which can be classified based on two
840 features: (i) the breast cancer subtype ER+/HER2- or triple negative, and (ii) the molecular classification
841 provided by the original study (CA-A, CA-B and CA-C). The ATAC-Seq raw counts and the samples
842 metadata were retrieved from figshare (Kumegawa, 2023). As for the previously mentioned datasets,
843 raw counts were normalized using the TPM-like transformation prior to bulk deconvolution.
844

845 **4- <u>Benchmarking of the EPIC-ATAC framework against other existing deconvolution tools</u>**

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

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

analyses). The results of DeconPeaker obtained using its original markers and profiles are also provided for the cell types in common with the cell types considered in this work in Figure 5-figure supplements 1 and 2. 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_Custom" in our analyses). To run CIBERSORTx, we used the docker container provided by the authors of CIBERSORTx on their website. The algorithm was run using the default options (i.e --absolute FALSE, --rmbatchBmode FALSE and –rmbatchSmode FALSE), which results in cell-type proportions relative to the total amount of cells from the reference cell types.
- ABIS (Monaco *et al.*, 2019) uses robust linear modeling to estimate cell-type proportions in bulk RNA-Seq samples. No ATAC-Seq reference profiles are available in the deconvolution framework. ABIS was run using the EPIC-ATAC reference profiles by using the *rlm* function from the MASS R package (as performed in the deconvolute_abis function from the immunedeconv R package (Sturm *et al.*, 2019)) to quantify each cell type from the reference profiles. The cell-types quantifications returned by this approach are in arbitrary units. To compare the estimations and the true cell proportions, we scaled the estimations of each sample between 0 and 1 to obtained relative proportions.

889 • MCPcounter (Becht *et al.*, 2016): MCPcounter returns scores instead of cell type proportions.
890 The scores were obtained using the *appendSignatures* function from the MCPcounter R
891 package by providing the list of marker peaks specific to each cell type. The cell-type scores
892 are not comparable between cell type, MCPcounter was thus included only in the evaluation
893 of the performances in each cell type separately.
894
895 For all the tools, TPM-like data were used as input bulk samples for the deconvolution.
896 Since CIBERSORTx, ABIS and DeconPeaker do not predict proportions of uncharacterized cells, we
897 performed two benchmarking analyses: (i) including all cell types and (ii) excluding the cell types that
898 are absent from the reference profiles (uncharacterized cells) and rescaling the estimated and true
899 proportions of the immune cells, endothelial cells and fibroblasts so that their sum equals 1.
900
901 **Benchmarking of the running time of each tool**
902 We compared the running time of each tool on the different datasets considered in our benchmarking
903 and used the bash command *time* to retrieve the CPU time of each run. The measured running time is
904 composed of the following steps: i) matching the bulk features to the reference profiles' features and
905 ii) running the deconvolution algorithm.
906
907 **5- <u>Comparing deconvolution based on RNA-Seq, gene activity or peaks features.</u>**
908 100 pseudobulks were generated from the 10X PBMC multiome dataset (10x Genomics, 2021) based
909 on 3000 cells for each pseudobulk. Cell fractions were defined using the *rdirichlet* function from the
910 *gtools* R package. We also considered pseudobulks built from the HTAN datasets by averaging the
911 signal in all cells belonging to a sample. Three sets of features were extracted from the data, *i.e.,* gene
912 expression features extracted from the RNA-Seq layer, ATAC-Seq peaks and gene activity derived from
913 the ATAC-Seq layer. The same cells were considered for each modality.

914 Gene activity features were extracted from the single-cell data using ArchR (1.0.2), which considers
915 distal elements and adjusts for large differences in gene size in the gene activity score calculation. Gene
916 activity pseudobulks were built by averaging the gene activity scores across all cells belonging to the
917 pseudobulk. For ATAC-Seq pseudobulk, peaks called using ArchR on all cells form the single-cell ATAC
918 data were considered (see the method section "ATAC-Seq pseudobulk data from PBMCs and cancer
919 samples") and counts were averaged across all cells of each pseudobulk. For RNA-Seq pseudobulks,
920 counts were also averaged across all cells of each pseudobulk. All aggregated data were depth
921 normalized across each features to 10^6. Cell-type deconvolution was performed on each pseudobulk
922 using EPIC-ATAC on the peak matrix using our ATAC-Seq marker peaks and reference profiles. The RNA-
923 Seq and gene activity pseudobulks were deconvolved with EPIC.
924 Bulk ATAC-Seq samples from a PBMC cohort were also obtained from Morandini *et al.* (Morandini *et*
925 *al.*, 2024). These data include more than 100 samples with matched bulk RNA-Seq, bulk ATAC-Seq as
926 well as matched flow cytometry data. The ATAC-Seq and RNA-Seq raw counts were obtained from GEO
927 using the following GEO accession numbers: GSE193140 and GSE193141. As for the previously
928 mentioned datasets, raw counts were normalized using the TPM-like transformation prior to bulk
929 deconvolution. The flow cytometry data were obtained from the supplementary file 7 from the original
930 publication and the data were rescaled as follows: CD4+ and CD8+ T cells proportions were rescaled to
931 sum up to the proportion of T cells and the T cells, B cells and NK cells proportions were rescaled to
932 sum up to the proportion of lymphocytes. All proportions were then rescaled to sum up to 1. EPIC-
933 ATAC was applied on the ATAC-Seq data and EPIC to the RNA-Seq data and the performances of each
934 set of predictions were compared to each other.
935
936 **Code availability**
937 The code to download and preprocess publicly available ATAC-Seq samples as well as the code used to
938 identify our cell-type specific marker peaks, generate the reference profiles and perform the main

939 analyses of the paper is available on GitHub (https://github.com/GfellerLab/EPIC-ATAC_Manuscript).
940 A README file is provided on the GitHub repository with more details on how to use the code.
941 The code to perform ATAC-Seq deconvolution using the EPIC-ATAC framework is available as an R
942 package called EPICATAC and is available on GitHub (https://github.com/GfellerLab/EPIC-ATAC).
943
944 **Data availability**
945 The newly generated ATAC-Seq data have been deposited on Zenodo
946 (https://zenodo.org/records/13132868). The other data related to this work are available in the
947 supplementary files and on the Zenodo deposit (https://zenodo.org/records/13132868).
948
949 **Competing interests:**
950 The authors declare that they have no competing interests.
951
952 **Acknowledgement:**
953 We thank the Lausanne Genomic Technologies Facility, University of Lausanne, Switzerland
954 (https://www.unil.ch/gtf/en/home.html) for the sequencing of the PBMC samples as well as Yan Liu,
955 Dana Moreno and Matei Teleman for testing the EPICATAC R package. Some of the illustrations were
956 created with BioRender.com (Figures 1, 3A, 7C).
957
958 **Authors contributions:**
959 Conceptualization: AAGG, JR, DG; Data curation: AAGG; Software: AAGG, JR; Experiments: MF, CJ;
960 Visualization: AAGG, JR, DG; Methodology: AAGG, JR, DG; Writing—original draft: AAGG, DG; Writing—
961 review and editing: all authors.
962
963 **Figure supplements grouped in one file:** Supplementary_figures.pdf
964 **Supplementary files:**

965 **Supplementary file 1:** Metadata of the ATAC-Seq samples used in the study

966 **Supplementary file 2:** Averaged chromatin accessibility of the PBMC marker peaks in each cell-type.

967 **Supplementary file 3:** Averaged chromatin accessibility of the TME marker peaks in each cell-type.

968 **Supplementary file 4:** Annotations of the cell-type specific PBMC marker peaks

969 **Supplementary file 5:** Annotations of the cell-type specific TME marker peaks

970 **Supplementary file 6:** GO pathways enriched in each set of cell-type specific PBMC marker peaks

971 **Supplementary file 7:** GO pathways enriched in each set of cell-type specific TME marker peaks

972 **Supplementary file 8:** Averaged chromatin accessibility of the PBMC marker peaks in each cell-type (T

973 cells subtypes included).

974 **Supplementary file 9:** Averaged chromatin accessibility of the TME marker peaks in each cell-type (T

975 cells subtypes included).

976 **Supplementary file 10:** Annotations of the cell-type specific PBMC marker peaks (T cells subtypes

977 included).

978 **Supplementary file 11:** Annotations of the cell-type specific TME marker peaks (T cells subtypes

979 included).

980 **Supplementary file 12:** GO pathways enriched in each set of cell-type specific PBMC marker peaks (T

981 cell subtypes).

982 **Supplementary file 13:** GO pathways enriched in each set of cell-type specific TME marker peaks (T

983 cell subtypes).

984

985 **References:**

986

987 10x Genomics (2021) *PBMC from a Healthy Donor - Granulocytes Removed Through Cell Sorting*

988 *(10k)*. Available at: https://www.10xgenomics.com/resources/datasets/pbmc-from-a-healthy-donor-

989 granulocytes-removed-through-cell-sorting-10-k-1-standard-2-0-0 (Accessed: 6 September 2023).

990 Arneson, D., Yang, X. and Wang, K. (2020) 'MethylResolver—a method for deconvoluting bulk DNA

991 methylation profiles into known and unknown cell contents', *Communications Biology*, 3. doi:
992 10.1038/s42003-020-01146-2.
993 Aryee, M. J. *et al.* (2014) 'Minfi: a flexible and comprehensive Bioconductor package for the analysis
994 of Infinium DNA methylation microarrays', *Bioinformatics*, 30(10), pp. 1363–1369. doi:
995 10.1093/bioinformatics/btu049.
996 Avila Cobos, F. *et al.* (2018) 'Computational deconvolution of transcriptomics data from mixed cell
997 populations', *Bioinformatics*, 34(11), pp. 1969–1979. doi: 10.1093/BIOINFORMATICS/BTY019.
998 Avila Cobos, F. *et al.* (2020) 'Benchmarking of cell type deconvolution pipelines for transcriptomics
999 data', *Nature Communications*, 11(1), pp. 1–14. doi: 10.1038/s41467-020-19015-1.
1000 Becht, E. *et al.* (2016) 'Estimating the population abundance of tissue-infiltrating immune and
1001 stromal cell populations using gene expression', *Genome Biology*, 17, p. 218. doi: 10.1186/S13059-
1002 016-1070-5.
1003 Bonnema, J. D. *et al.* (1994) 'Fc receptor stimulation of phosphatidylinositol 3-kinase in natural killer
1004 cells is associated with protein kinase C-independent granule release and cell-mediated cytotoxicity.',
1005 *Journal of Experimental Medicine*, 180(4), pp. 1427–1435. doi: 10.1084/JEM.180.4.1427.
1006 Buenrostro, J. D. *et al.* (2013) 'Transposition of native chromatin for fast and sensitive epigenomic
1007 profiling of open chromatin, DNA-binding proteins and nucleosome position', *Nature Methods*,
1008 10(12), pp. 1213–1218. doi: 10.1038/nmeth.2688.
1009 Burdziak, C. *et al.* (2019) 'A Nonparametric Multi-view Model for Estimating Cell Type-Specific Gene
1010 Regulatory Networks', *ArXiv*.
1011 Calderon, D. *et al.* (2019) 'Landscape of stimulation-responsive chromatin across diverse human
1012 immune cells', *Nature Genetics*, 51(10), pp. 1494–1505. doi: 10.1038/s41588-019-0505-9.
1013 Carvalho, K. *et al.* (2021) 'Uncovering the Gene Regulatory Networks Underlying Macrophage
1014 Polarization Through Comparative Analysis of Bulk and Single-Cell Data', *bioRxiv*. doi:
1015 10.1101/2021.01.20.427499.
1016 Castro-Mondragon, J. A. *et al.* (2022) 'JASPAR 2022: the 9th release of the open-access database of

1017 transcription factor binding profiles', *Nucleic Acids Research*, 50(D1), pp. D165–D173. doi:
1018 10.1093/NAR/GKAB1113.
1019 Chakravarthy, A. *et al.* (2018) 'Pan-cancer deconvolution of tumour composition using DNA
1020 methylation', *Nature Communications*, 9. doi: 10.1038/s41467-018-05570-1.
1021 Clarke, J., Seol, P. and Clarke, B. (2010) 'Statistical expression deconvolution from mixed tissue
1022 samples', *Bioinformatics*, 26(8), pp. 1043–1049. doi: 10.1093/BIOINFORMATICS/BTQ097.
1023 Corces, M. R. *et al.* (2016) 'Lineage-specific and single-cell chromatin accessibility charts human
1024 hematopoiesis and leukemia evolution', *Nature Genetics*, 48, pp. 1193–1203. doi: 10.1038/ng.3646.
1025 Corces, M. R. *et al.* (2017) 'An improved ATAC-seq protocol reduces background and enables
1026 interrogation of frozen tissues', *Nature Methods*, 14(10), pp. 959–962. doi: 10.1038/nmeth.4396.
1027 Corces, M. R. *et al.* (2018) 'The chromatin accessibility landscape of primary human cancers', *Science*,
1028 362(6413). doi: 10.1126/science.aav1898.
1029 Cusanovich, D. A. *et al.* (2015) 'Multiplex single-cell profiling of chromatin accessibility by
1030 combinatorial cellular indexing', *Science*, 348(6237), pp. 910–914. doi: 10.1126/SCIENCE.AAB1601.
1031 Feng, S. *et al.* (2023) 'Decomprolute : A benchmarking platform designed for multiomics-based tumor
1032 deconvolution', *bioRxiv*.
1033 Finotello, F. *et al.* (2019) 'Molecular and pharmacological modulators of the tumor immune
1034 contexture revealed by deconvolution of RNA-seq data', *Genome Medicine*, 11, p. 34. doi:
1035 10.1186/s13073-019-0638-6.
1036 Franzén, O., Gan, L. M. and Björkegren, J. L. M. (2019) 'PanglaoDB: a web server for exploration of
1037 mouse and human single-cell RNA sequencing data', *Database*, 2019, p. baz046. doi:
1038 10.1093/DATABASE/BAZ046.
1039 Fridman, W. H. *et al.* (2012) 'The immune contexture in human tumours: impact on clinical outcome',
1040 *Nature Reviews Cancer*, 12(4), pp. 298–306. doi: 10.1038/nrc3245.
1041 Fridman, W. H. *et al.* (2017) 'The immune contexture in cancer prognosis and treatment', *Nature*
1042 *Reviews Clinical Oncology*. Nature Publishing Group, pp. 717–734. doi: 10.1038/nrclinonc.2017.101.

1043 Ge, X. *et al.* (2021) 'Functional genomics atlas of synovial fibroblasts defining rheumatoid arthritis
1044 heritability', *Genome Biology*, 22(1), p. 247. doi: 10.1186/S13059-021-02460-6.
1045 Geng, D. *et al.* (2010) 'When Toll-like receptor and T-cell receptor signals collide: a mechanism for
1046 enhanced CD8 T-cell effector function', *Blood*, 116(18), pp. 3494–3504. doi: 10.1182/BLOOD-2010-
1047 02-268169.
1048 Giles, J. R. *et al.* (2022) 'Human epigenetic and transcriptional T cell differentiation atlas for
1049 identifying functional T cell-specific enhancers', *Immunity*, 55(3), pp. 557-574.e7. doi:
1050 10.1016/J.IMMUNI.2022.02.004.
1051 Gong, T. and Szustakowski, J. D. (2013) 'DeconRNASeq: a statistical framework for deconvolution of
1052 heterogeneous tissue samples based on mRNA-Seq data', *Bioinformatics*, 29(8), pp. 1083–1085. doi:
1053 10.1093/BIOINFORMATICS/BTT090.
1054 Gosink, M. M., Petrie, H. T. and Tsinoremas, N. F. (2007) 'Electronically subtracting expression
1055 patterns from a mixed cell population', *Bioinformatics*, 23(24), pp. 3328–3334. doi:
1056 10.1093/BIOINFORMATICS/BTM508.
1057 Grandi, F. C. *et al.* (2022) 'Chromatin accessibility profiling by ATAC-seq', *Nature protocols*, pp. 1518–
1058 1552. doi: 10.1038/s41596-022-00692-9.
1059 Granja, J. M. *et al.* (2019) 'Single-cell multiomic analysis identifies regulatory programs in mixed-
1060 phenotype acute leukemia', *Nature Biotechnology*, 37(12), pp. 1458–1465. doi: 10.1038/s41587-019-
1061 0332-7.
1062 Granja, J. M. *et al.* (2021) 'ArchR is a scalable software package for integrative single-cell chromatin
1063 accessibility analysis', *Nature Genetics*, 53, pp. 403–411. doi: 10.1038/s41588-021-00790-6.
1064 Hammal, F. *et al.* (2022) 'ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis
1065 regulatory regions from an integrative analysis of DNA-binding sequencing experiments', *Nucleic
1066 Acids Research*, 50(D1), pp. D316–D325. doi: 10.1093/NAR/GKAB996.
1067 Hicks, S. C. and Irizarry, R. A. (2019) 'MethylCC: Technology-independent estimation of cell type
1068 composition using differentially methylated regions', *Genome Biology*, 20(1), pp. 1–13. doi:

1069 10.1186/S13059-019-1827-8.

1070 Hu, C. *et al.* (2023) 'CellMarker 2.0: an updated database of manually curated cell markers in

1071 human/mouse and web tools based on scRNA-seq data', *Nucleic Acids Research*, 51(D1), pp. D870–

1072 D876. doi: 10.1093/NAR/GKAC947.

1073 Javaid, N. and Choi, S. (2020) 'Toll-like Receptors from the Perspective of Cancer Treatment', *Cancers*,

1074 12(2), p. 297. doi: 10.3390/CANCERS12020297.

1075 Jiang, Y. *et al.* (2023) 'scATAnno: Automated Cell Type Annotation for single-cell ATAC Sequencing

1076 Data', *bioRxiv*. doi: 10.1101/2023.06.01.543296.

1077 Jimenez-Sanchez, A., Cast, O. and Miller, M. L. (2019) 'Comprehensive benchmarking and integration

1078 of tumor microenvironment cell estimation methods', *Cancer Research*, 79(24), pp. 6238–6246. doi:

1079 10.1158/0008-5472.CAN-18-3560.

1080 Jin, H. and Liu, Z. (2021) 'A benchmark for RNA-seq deconvolution analysis under dynamic testing

1081 environments', *Genome Biology*, 22, p. 102. doi: 10.1186/s13059-021-02290-6.

1082 Kalafati, L. *et al.* (2020) 'Innate Immune Training of Granulopoiesis Promotes Anti-tumor Activity',

1083 *Cell*, 183(3), pp. 771-785.e12. doi: 10.1016/J.CELL.2020.09.058.

1084 Kaur, D. *et al.* (2023) 'Comprehensive Evaluation of The Infinium Human MethylationEPIC v2

1085 BeadChip', *Epigenetics communications*, 3(1). doi: 10.1186/S43682-023-00021-5.

1086 Klemm, S. L., Shipony, Z. and Greenleaf, W. J. (2019) 'Chromatin accessibility and the regulatory

1087 epigenome', *Nature Reviews Genetics*, 20(4), pp. 207–220. doi: 10.1038/s41576-018-0089-8.

1088 Kuhn, M. (2008) 'Building Predictive Models in R Using the caret Package', *Journal of Statistical

1089 Software*, 28(5), pp. 1–26. doi: 10.18637/JSS.V028.I05.

1090 Kumegawa, K. *et al.* (2022) 'GRHL2 motif is associated with intratumor heterogeneity of cis-

1091 regulatory elements in luminal breast cancer', *npj Breast Cancer 2022 8:1*, 8(1), pp. 1–12. doi:

1092 10.1038/s41523-022-00438-6.

1093 Kumegawa, K. (2023) *ATAC-seq data of 42 BC samples as SummarizedExperiment object with count

1094 matrix, normalized count matrix, peak info, and clinical info*. Available at:

1095 https://doi.org/10.6084/m9.figshare.21992609.v1.

1096 Kumegawa, K. *et al.* (2023) 'Chromatin profile-based identification of a novel ER-positive breast
1097 cancer subgroup with reduced ER-responsive element accessibility', *British Journal of Cancer*, 128(7),
1098 pp. 1208–1222. doi: 10.1038/s41416-023-02178-1.

1099 Lareau, C. A. *et al.* (2019) 'Droplet-based combinatorial indexing for massive-scale single-cell
1100 chromatin accessibility', *Nature Biotechnology*, 37(8), pp. 916–924. doi: 10.1038/s41587-019-0147-6.

1101 Lau, J. W. *et al.* (2017) 'The cancer genomics cloud: Collaborative, reproducible, and democratized - A
1102 new paradigm in large-scale computational research', *Cancer Research*, 77(21), pp. e3–e6. doi:
1103 10.1158/0008-5472.CAN-17-0387.

1104 Law, C. W. *et al.* (2014) 'Voom: Precision weights unlock linear model analysis tools for RNA-seq read
1105 counts', *Genome Biology*, 15, p. R29. doi: 10.1186/GB-2014-15-2-R29.

1106 Leylek, R. *et al.* (2020) 'Chromatin Landscape Underpinning Human Dendritic Cell Heterogeneity', *Cell*
1107 *Reports*, 32(12), p. 108180. doi: 10.1016/J.CELREP.2020.108180.

1108 Li, H. *et al.* (2020) 'DeconPeaker, a Deconvolution Model to Identify Cell Types Based on Chromatin
1109 Accessibility in ATAC-Seq Data of Mixture Samples', *Frontiers in Genetics*, 11. doi:
1110 10.3389/fgene.2020.00392.

1111 Li, T. *et al.* (2020) 'TIMER2.0 for analysis of tumor-infiltrating immune cells', *Nucleic acids research*,
1112 48, pp. W509–W514. doi: 10.1093/nar/gkaa407.

1113 Liao, Y., Smyth, G. K. and Shi, W. (2014) 'featureCounts: an efficient general purpose program for
1114 assigning sequence reads to genomic features', *Bioinformatics*, 30(7), pp. 923–930. doi:
1115 10.1093/BIOINFORMATICS/BTT656.

1116 Liu, Q. *et al.* (2020) 'Chromatin accessibility landscapes of skin cells in systemic sclerosis nominate
1117 dendritic cells in disease pathogenesis', *Nature Communications*, 11(1), p. 5843. doi:
1118 10.1038/s41467-020-19702-z.

1119 Loyfer, N. *et al.* (2023) 'A DNA methylation atlas of normal human cell types', *Nature*, 613(7943), pp.
1120 355–364. doi: 10.1038/s41586-022-05580-6.

1121 Luo, L., Gribskov, M. and Wang, S. (2022) 'Bibliometric review of ATAC-Seq and its application in gene
1122 expression', *Briefings in Bioinformatics*. doi: 10.1093/BIB/BBAC061.
1123 Machlab, D. *et al.* (2022) 'monaLisa: an R/Bioconductor package for identifying regulatory motifs',
1124 *Bioinformatics*, 38(9), pp. 2624–2625. doi: 10.1093/BIOINFORMATICS/BTAC102.
1125 Mayer, M. (2023) 'R package " splitTools ": Tools for Data Splitting. R package version 1.0.1.
1126 https://cran.r-project.org/web/packages/splitTools/index.html'.
1127 Monaco, G. *et al.* (2019) 'RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute
1128 Deconvolution of Human Immune Cell Types', *Cell Reports*, 26(6), pp. 1627-1640.e7. doi:
1129 10.1016/J.CELREP.2019.01.041.
1130 Morandini, F. *et al.* (2024) 'ATAC-clock: An aging clock based on chromatin accessibility', *GeroScience*,
1131 46(2), pp. 1789–1806. doi: 10.1007/S11357-023-00986-0.
1132 Mumbach, M. R. *et al.* (2017) 'Enhancer connectome in primary human cells identifies target genes
1133 of disease-associated DNA elements', *Nature Genetics*, 49(11), pp. 1602–1612. doi: 10.1038/ng.3963.
1134 Newman, A. M. *et al.* (2015) 'Robust enumeration of cell subsets from tissue expression profiles',
1135 *Nature Methods*, 12(5), pp. 453–457. doi: 10.1038/nmeth.3337.
1136 Newman, A. M. *et al.* (2019) 'Determining cell type abundance and expression from bulk tissues with
1137 digital cytometry', *Nature Biotechnology*, 37(7), pp. 773–782. doi: 10.1038/s41587-019-0114-2.
1138 Peng, X. L. *et al.* (2019) 'De novo compartment deconvolution and weight estimation of tumor
1139 samples using DECODER', *Nature Communications*, 10(1), p. 4729. doi: 10.1038/s41467-019-12517-7.
1140 Perez, C. *et al.* (2020) 'Immunogenomic identification and characterization of granulocytic myeloid-
1141 derived suppressor cells in multiple myeloma', *Blood*, 136(2), pp. 199–209. doi:
1142 10.1182/BLOOD.2019004537.
1143 Qiu, Y. *et al.* (2021) 'Identification of cell-type-specific marker genes from co-expression patterns in
1144 tissue samples', *Bioinformatics*, 37(19), pp. 3228–3234. doi: 10.1093/BIOINFORMATICS/BTAB257.
1145 Racle, J. *et al.* (2017) 'Simultaneous enumeration of cancer and immune cell types from bulk tumor
1146 gene expression data', *eLife*, 6. doi: 10.7554/eLife.26476.

1147 Racle, J. and Gfeller, D. (2020) 'EPIC: A tool to estimate the proportions of different cell types from
1148 bulk gene expression data', in *Methods in Molecular Biology*. Humana Press Inc., pp. 233–248. doi:
1149 10.1007/978-1-0716-0327-7_17.
1150 Rahmani, E. *et al.* (2019) 'Cell-type-specific resolution epigenetics without the need for cell sorting or
1151 single-cell biology', *Nature Communications 2019 10:1*, 10(1), pp. 1–11. doi: 10.1038/s41467-019-
1152 11052-9.
1153 Ram-Mohan, N. *et al.* (2021) 'Profiling chromatin accessibility responses in human neutrophils with
1154 sensitive pathogen detection', *Life Science Alliance*, 4(8). doi: 10.26508/LSA.202000976.
1155 Regner, M. J. *et al.* (2021) 'A multi-omic single-cell landscape of human gynecologic malignancies',
1156 *Molecular Cell*, 81(23), pp. 4924-4941.e10. doi: 10.1016/j.molcel.2021.10.013.
1157 Risso, D. *et al.* (2011) 'GC-Content Normalization for RNA-Seq Data', *BMC Bioinformatics*, 12, p. 480.
1158 doi: 10.1186/1471-2105-12-480.
1159 Ritchie, M. E. *et al.* (2015) 'limma powers differential expression analyses for RNA-sequencing and
1160 microarray studies', *Nucleic Acids Research*, 43(7), p. e47. doi: 10.1093/nar/gkv007.
1161 Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2010) 'edgeR: a Bioconductor package for
1162 differential expression analysis of digital gene expression data', *Bioinformatics*, 26(1), pp. 139–140.
1163 doi: 10.1093/BIOINFORMATICS/BTP616.
1164 Rozowsky, J. *et al.* (2023) 'The EN-TEx resource of multi-tissue personal epigenomes & variant-impact
1165 models', *Cell*, 186(7), pp. 1493-1511.e40. doi: 10.1016/j.cell.2023.02.018.
1166 Salas, L. A. *et al.* (2022) 'Enhanced cell deconvolution of peripheral blood using DNA methylation for
1167 high-resolution immune profiling', *Nature Communications 2022 13:1*, 13(1), pp. 1–13. doi:
1168 10.1038/s41467-021-27864-7.
1169 Sanseviero, E. (2019) 'NK Cell-Fc Receptors Advance Tumor Immunotherapy', *Journal of Clinical*
1170 *Medicine*, 8(10), p. 1667. doi: 10.3390/JCM8101667.
1171 Satpathy, A. T. *et al.* (2019) 'Massively parallel single-cell chromatin landscapes of human immune
1172 cell development and intratumoral T cell exhaustion', *Nature Biotechnology*, 37(8), pp. 925–936. doi:

1173 10.1038/s41587-019-0206-z.

1174 Smith, J. P. *et al.* (2021) 'PEPATAC: an optimized pipeline for ATAC-seq data analysis with serial

1175 alignments', *NAR Genomics and Bioinformatics*, 3(4). doi: 10.1093/NARGAB/LQAB101.

1176 Stuart, T. *et al.* (2021) 'Single-cell chromatin state analysis with Signac', *Nature Methods*, 18, pp.

1177 1333–1341. doi: 10.1038/s41592-021-01282-5.

1178 Sturm, G. *et al.* (2019) 'Comprehensive evaluation of transcriptome-based cell-type quantification

1179 methods for immuno-oncology', *Bioinformatics*, 35(14), pp. i436–i445. doi:

1180 10.1093/bioinformatics/btz363.

1181 Subramanian, A. *et al.* (2005) 'Gene set enrichment analysis: a knowledge-based approach for

1182 interpreting genome-wide expression profiles.', *PNAS*, 102(43), pp. 15545–50. doi:

1183 10.1073/pnas.0506580102.

1184 Terekhanova, N. V *et al.* (2023) 'Epigenetic regulation during cancer transitions across 11 tumour

1185 types', *Nature 2023*, 623, pp. 1–10. doi: 10.1038/s41586-023-06682-5.

1186 Teschendorff, A. E. *et al.* (2017) 'A comparison of reference-based algorithms for correcting cell-type

1187 heterogeneity in Epigenome-Wide Association Studies', *BMC bioinformatics*, 18(1). doi:

1188 10.1186/S12859-017-1511-5.

1189 Teschendorff, A. E. *et al.* (2020) 'EPISCORE: Cell type deconvolution of bulk tissue DNA methylomes

1190 from single-cell RNA-Seq data', *Genome Biology*, 21. doi: 10.1186/s13059-020-02126-9.

1191 The ENCODE Project Consortium *et al.* (2020) 'Expanded encyclopaedias of DNA elements in the

1192 human and mouse genomes', *Nature*, 583, pp. 699–710. doi: 10.1038/s41586-020-2493-4.

1193 Trizzino, M. *et al.* (2021) 'EGR1 is a gatekeeper of inflammatory enhancers in human macrophages',

1194 *Science Advances*, 7(3). doi: 10.1126/SCIADV.AAZ8836.

1195 Ucar, D. *et al.* (2017) 'The chromatin accessibility signature of human immune aging stems from CD8+

1196 T cells', *Journal of Experimental Medicine*, 214(10), pp. 3123–3144. doi: 10.1084/jem.20170416.

1197 de Visser, K. E. and Joyce, J. A. (2023) 'The evolving tumor microenvironment: From cancer initiation

1198 to metastatic outgrowth', *Cancer Cell*, 41(3), pp. 374–403. doi: 10.1016/J.CCELL.2023.02.016.

1199 Watt, S. *et al.* (2021) 'Genetic perturbation of PU.1 binding and chromatin looping at neutrophil
1200 enhancers associates with autoimmune disease', *Nature Communications*, 12, p. 2298. doi:
1201 10.1038/S41467-021-22548-8.
1202 Welch, R. P. *et al.* (2014) 'ChIP-Enrich: gene set enrichment testing for ChIP-seq data', *Nucleic Acids*
1203 *Research*, 42(13), p. e105. doi: 10.1093/NAR/GKU463.
1204 Xin, J. *et al.* (2020) 'Chromatin accessibility landscape and regulatory network of high-altitude
1205 hypoxia adaptation', *Nature Communications*, 11(1), p. 4928. doi: 10.1038/S41467-020-18638-8.
1206 Yu, G., Wang, L. G. and He, Q. Y. (2015) 'ChIPseeker: an R/Bioconductor package for ChIP peak
1207 annotation, comparison and visualization', *Bioinformatics*, 31(14), pp. 2382–2383. doi:
1208 10.1093/BIOINFORMATICS/BTV145.
1209 Zeng, W. *et al.* (2019) 'DC3 is a method for deconvolution and coupled clustering from bulk and
1210 single-cell genomics data', *Nature Communications*, 10, p. 4613. doi: 10.1038/s41467-019-12547-1.
1211 Zhang, H. *et al.* (2021) 'EMeth: An EM algorithm for cell type decomposition based on DNA
1212 methylation data', *Scientific Reports*, 11, p. 5717. doi: 10.1038/s41598-021-84864-9.
1213 Zhang, H. *et al.* (2022) 'Profiling chromatin accessibility in formalin-fixed paraffin-embedded
1214 samples', *Genome Research*, 32(1), pp. 150–161. doi: 10.1101/GR.275269.121.
1215 Zhang, K. *et al.* (2021) 'A single-cell atlas of chromatin accessibility in the human genome.', *Cell*,
1216 184(24), pp. 5985-6001.e19. doi: 10.1016/j.cell.2021.10.024.
1217 Zhang, P. *et al.* (2022) 'Epigenomic analysis reveals a dynamic and context-specific macrophage
1218 enhancer landscape associated with innate immune activation and tolerance.', *Genome Biology*,
1219 23(1), p. 136. doi: 10.1186/S13059-022-02702-1.
1220 Zhang, Y. *et al.* (2008) 'Model-based analysis of ChIP-Seq (MACS)', *Genome Biology*, 9(9), p. R137. doi:
1221 10.1186/GB-2008-9-9-R137.
1222
1223 **List of abbreviations:**
1224 **ATAC:** Assay for Transposase-Accessible chromatin

1225 **CBP(s):** chromatin binding protein(s)

1226 **ChIP-seq:** chromatin immunoprecipitation followed by sequencing

1227 **CRE:** cis-regulatory element

1228 **DC:** dendritic cells

1229 **HTAN**: Human Tumor Atlas Network

1230 **NK:** natural killer cells

1231 **PCA:** principal component analysis

1232 **RMSE:** root mean squared error

1233 **TCGA :** The Cancer Genome Atlas

1234 **TF(s):** transcription factor(s)

1235 **TNBC:** triple negative breast cancer

1236 **TSS:** transcription start site

1237

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,*]

1 - Department of Oncology, Ludwig Institute for Cancer Research, University of Lausanne, Lausanne, Switzerland
2 - Agora Cancer Research Center, Lausanne, Switzerland
3 - Swiss Cancer Center Leman (SCCL), Switzerland
4 - Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland.
5 - Department of Pathology and Immunology Faculty of Medicine, University of Geneva, Geneva, Switzerland
6 - Ludwig Institute for Cancer Research, Lausanne Branch, Lausanne, Switzerland
7 - Geneva Center for Inflammation Research, Geneva, Switzerland

* Corresponding author: david.gfeller@unil.ch

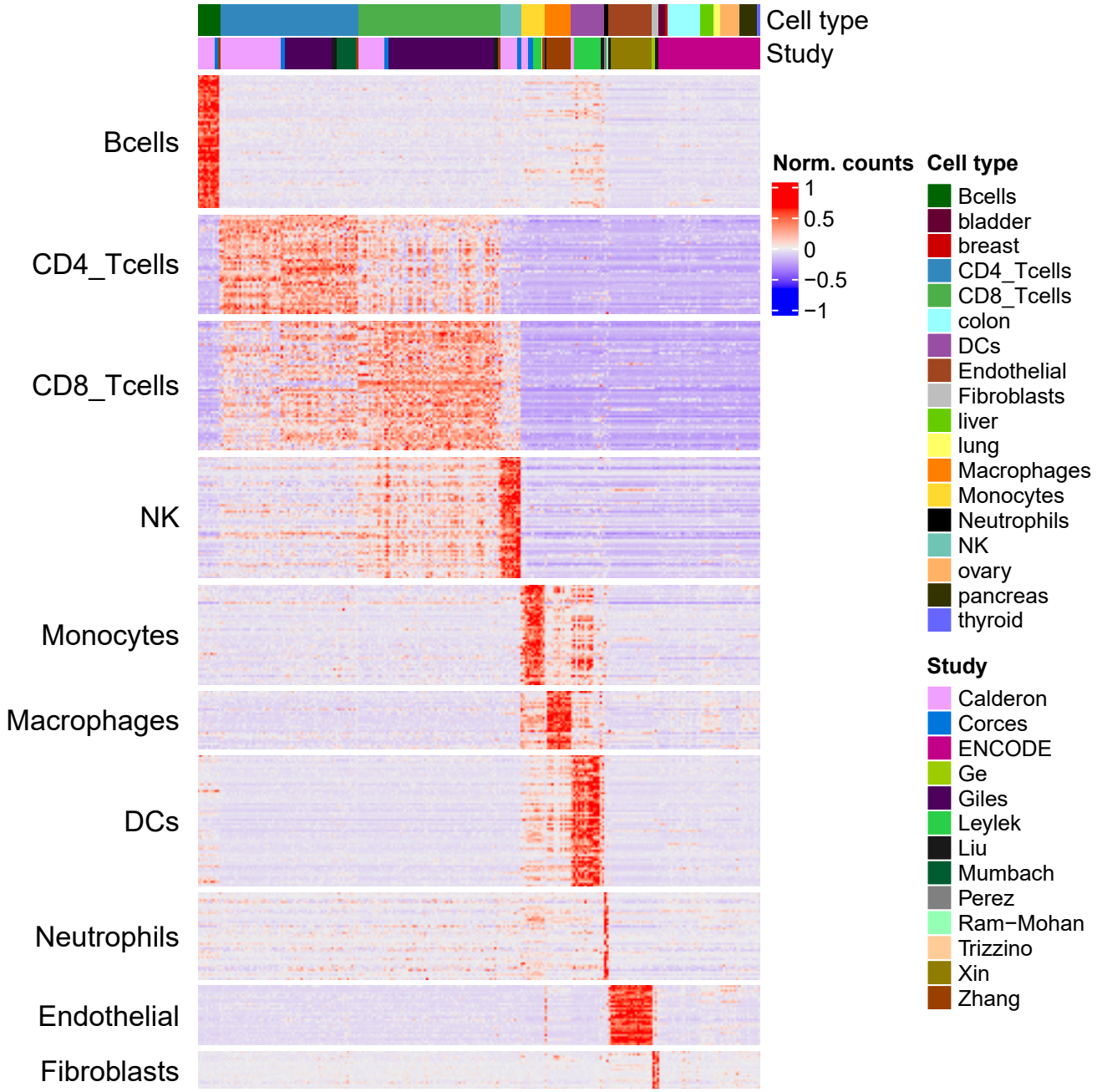


Figure 1-figure supplement 1: Chromatin accessibility signal of each marker peak (rows) in each reference sample and in the ENCODE samples from diverse tissues (columns).

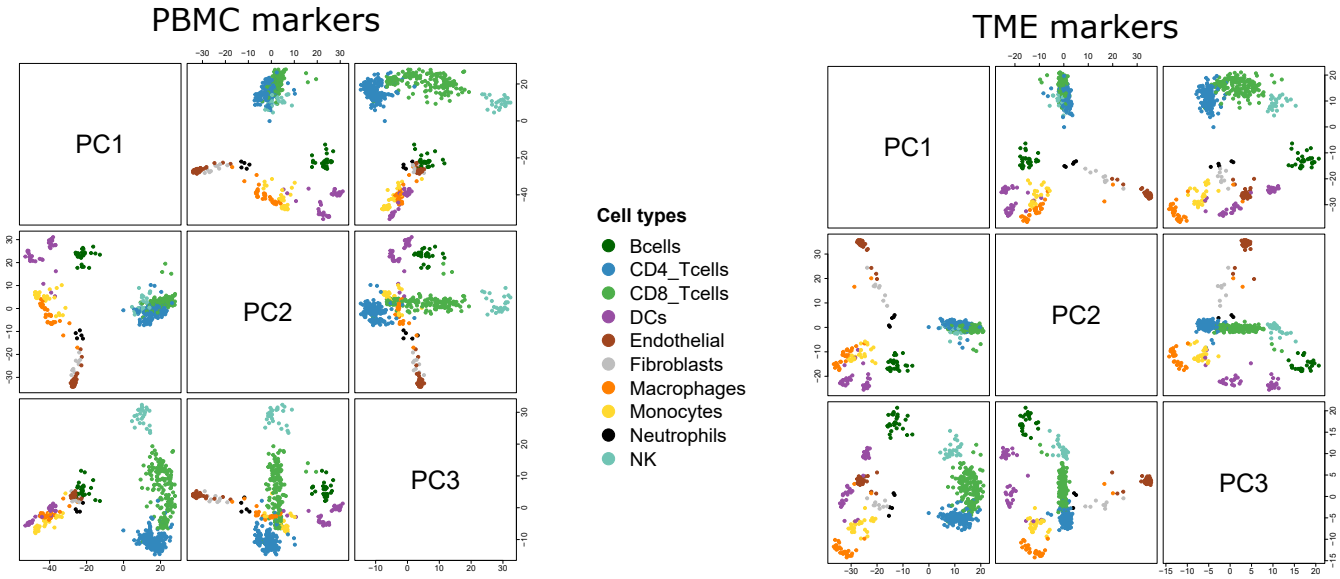


Figure 2-figure supplement 1: Pairwise scatterplots of the first 3 axes of the PCA run on the PBMC markers (left) and TME markers (right).

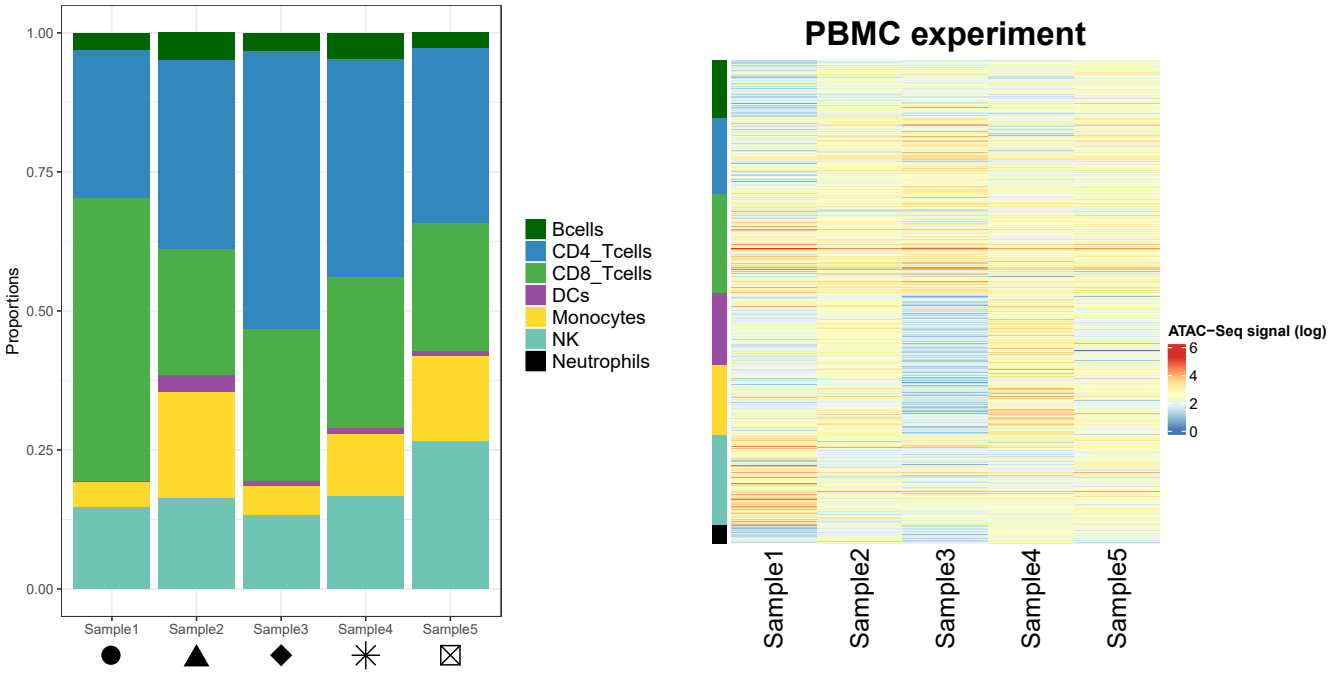


Figure 3-figure supplement 1: Left panel: Proportions of each cell type in the PBMC experiment samples. Shapes match the points shapes from Figure 3B in the main text. Right panel: Chromatin accessibility signal (log scale) of the cell-type specific marker peaks (rows) in each sample (columns) of the PBMC experiment dataset.

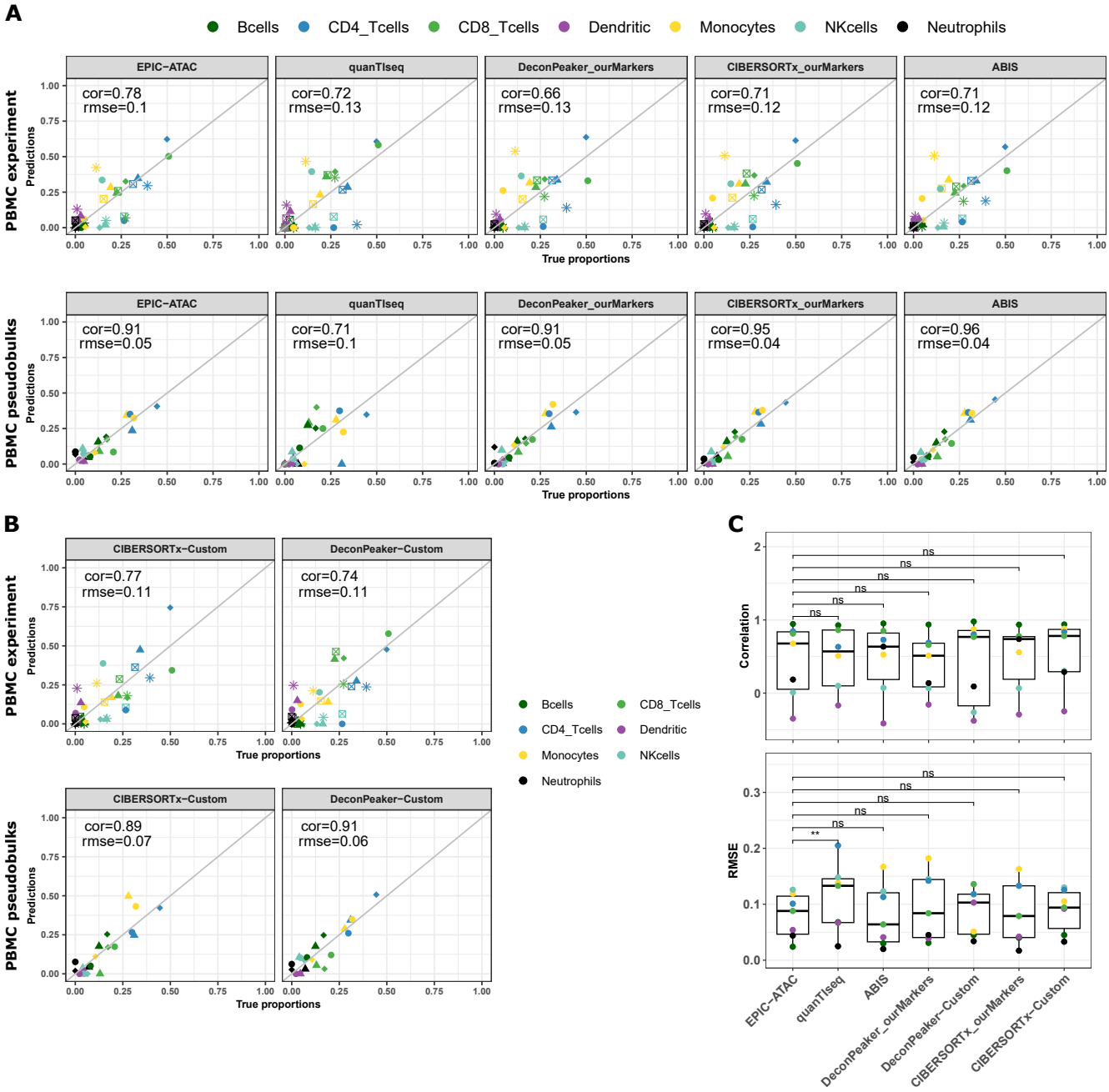


Figure 3-figure supplement 2: A) Comparison of the cell-type proportions estimated by the different deconvolution methods included in the benchmark (y axis) and the true proportions (x axis) in the PBMC experiment samples (top panel) and in the PBMC pseudobulk samples (bottom panel). **B)** Comparison of the cell-type proportions estimated by CIBERSORTx and DeconPeaker using reference profiles built by each tool using our collection of pure ATAC-Seq samples and the true proportions in the PBMC experiment samples (top panel) and in the PBMC pseudobulk samples (bottom panel). **C)** Pearson's correlation coefficient and RMSE values associated to the comparison of the cell-type proportions estimated by each deconvolution tool in each cell type of the PBMC datasets, colors correspond to cell types. Note that for quanTIseq, gray dots correspond to uncharacterized cells since the prediction of an uncharacterized cell type cannot be turned off using the quanTIseq function from the quantiseqr R package. *P*-values resulting from a paired wilcoxon test are represented for each pair of comparison between tools. Significance levels are represented as follows: *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

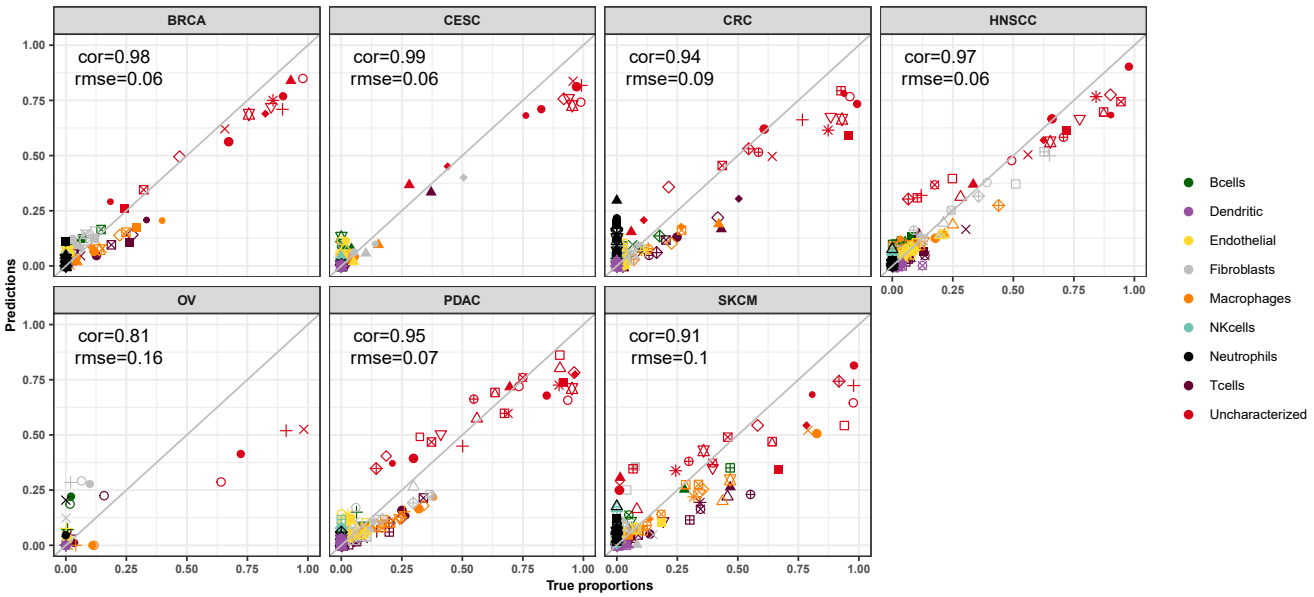


Figure 4-figure supplement 1: Comparison of the cell-type proportions estimated by EPIC-ATAC (y axis) and the true proportions (x axis) in the HTAN dataset.

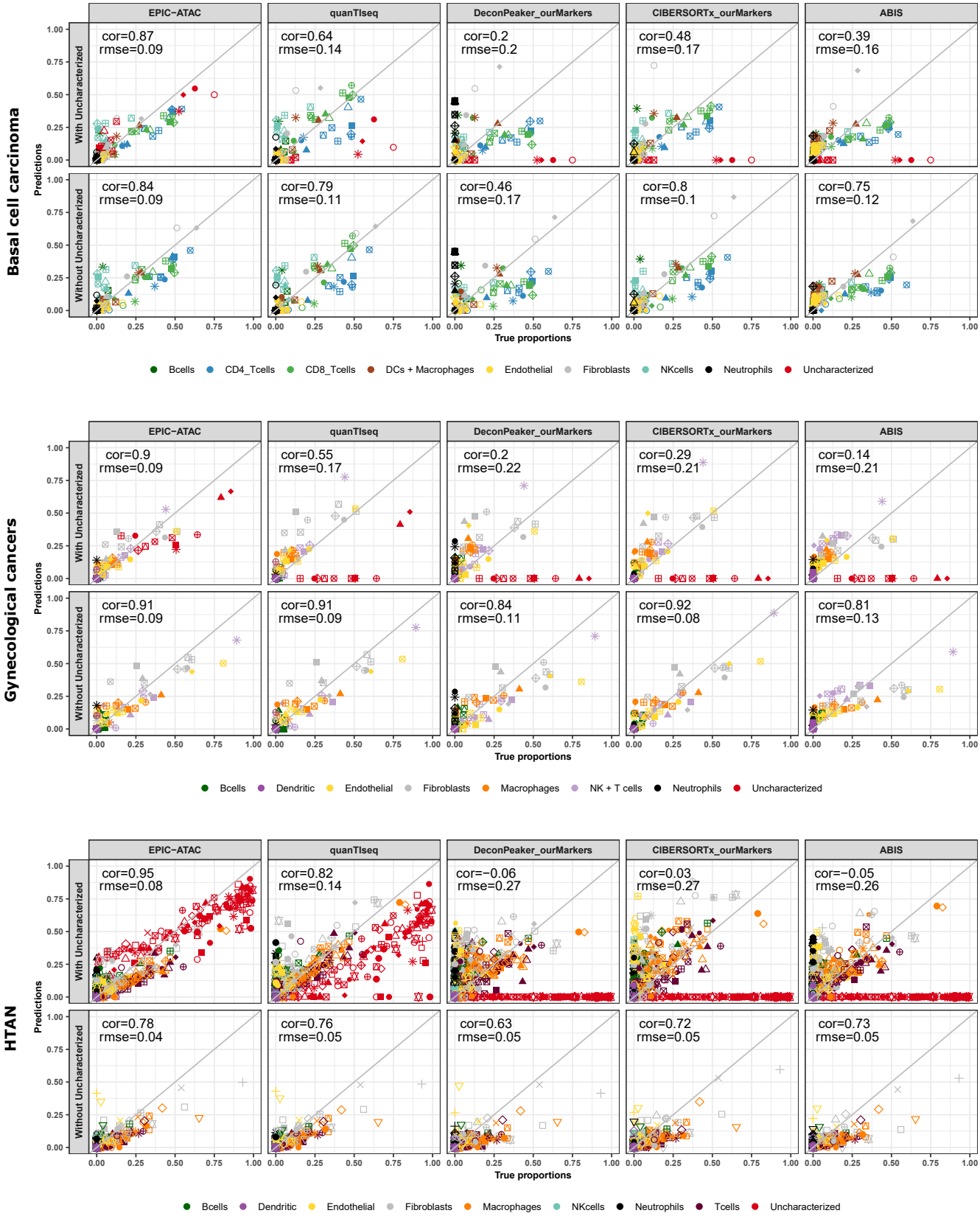


Figure 4-figure supplement 2: Comparison of the cell-type proportions estimated by the different deconvolution methods using our reference profiles and the true proportions in the basal cell carcinoma samples (top panel), in the gynecological cancer samples (middle panel) and in the HTAN samples (bottom panel). In each panel, the top row includes the uncharacterized cells while the bottom row excludes them.

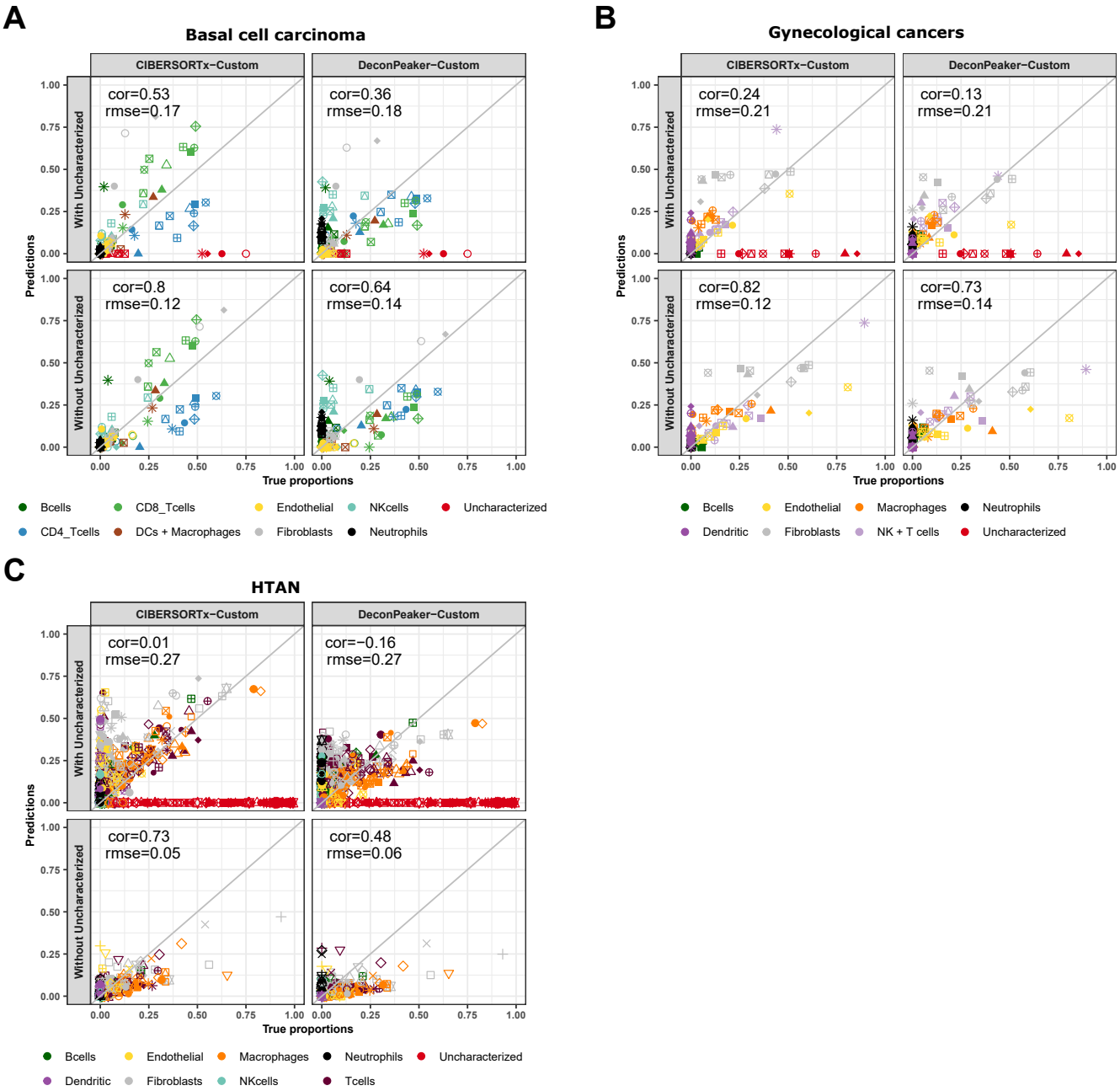


Figure 4-figure supplement 3: Comparison of the cell-type proportions estimated by CIBERSORTx and DeconPeaker using reference profiles built by each tool using our collection of pure ATAC-Seq samples and the true proportions in the basal cell carcinoma samples (panel A), in the gynecological cancer samples (panel B) and in the HTAN samples (panel C). In each panel, the top row includes the uncharacterized cells while the bottom row excludes them.

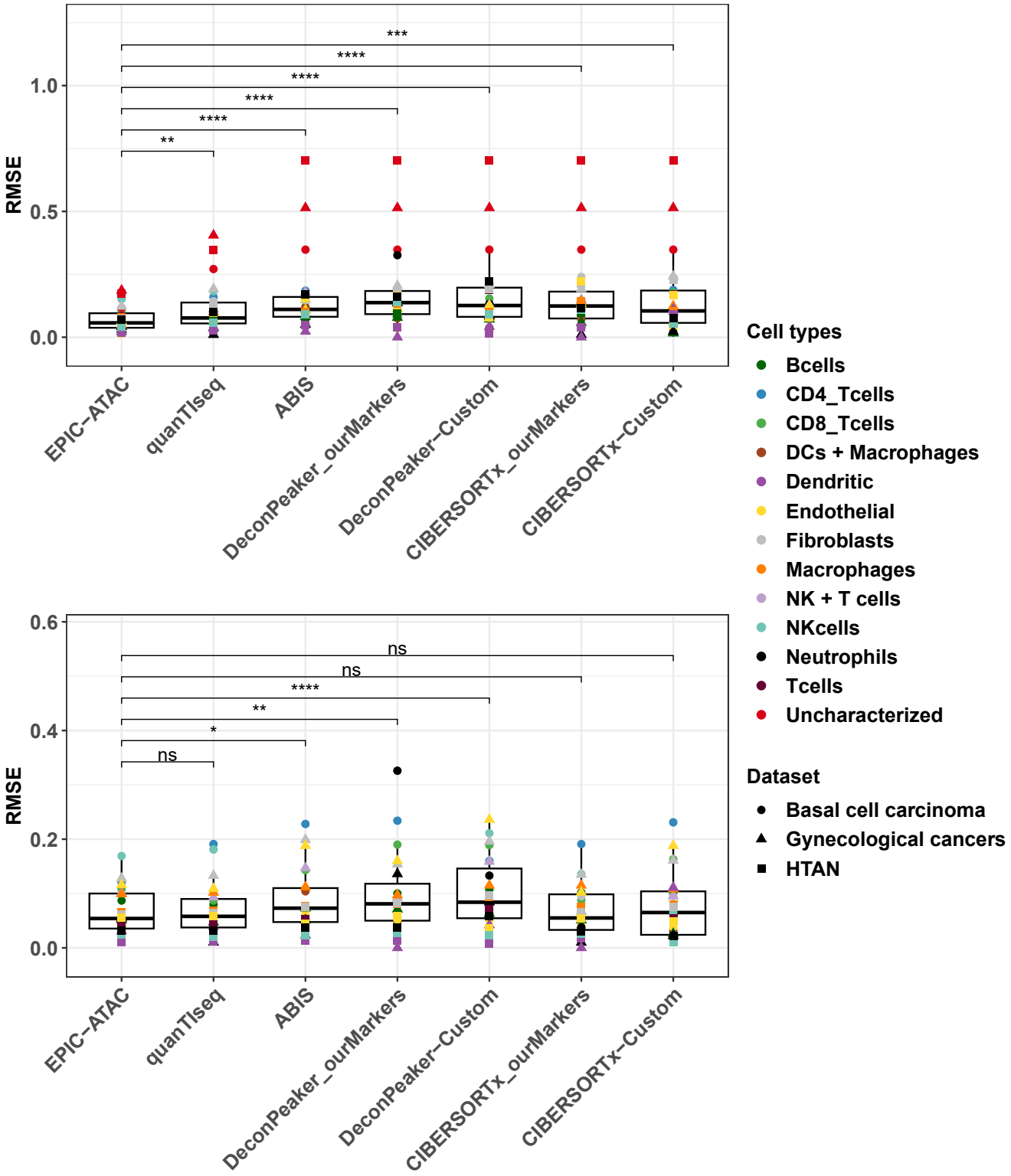


Figure 4-figure supplement 4: Top panel: RMSE values associated to the comparison of the true proportions and the cell-type proportions estimated by each deconvolution tool in each cell type. For each cell type, all samples from each cancer pseudobulk dataset were considered, *i.e.*, colors correspond to cell types and shapes to datasets. *P*-values resulting from a paired wilcoxon test are represented for each pair of comparison between tools. Significance levels are represented as follows: *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$. Bottom panel: Considering that not all deconvolution tools are able to predict the proportion of uncharacterized cells, the same analysis as in the top panel was performed without considering the uncharacterized cells predictions.

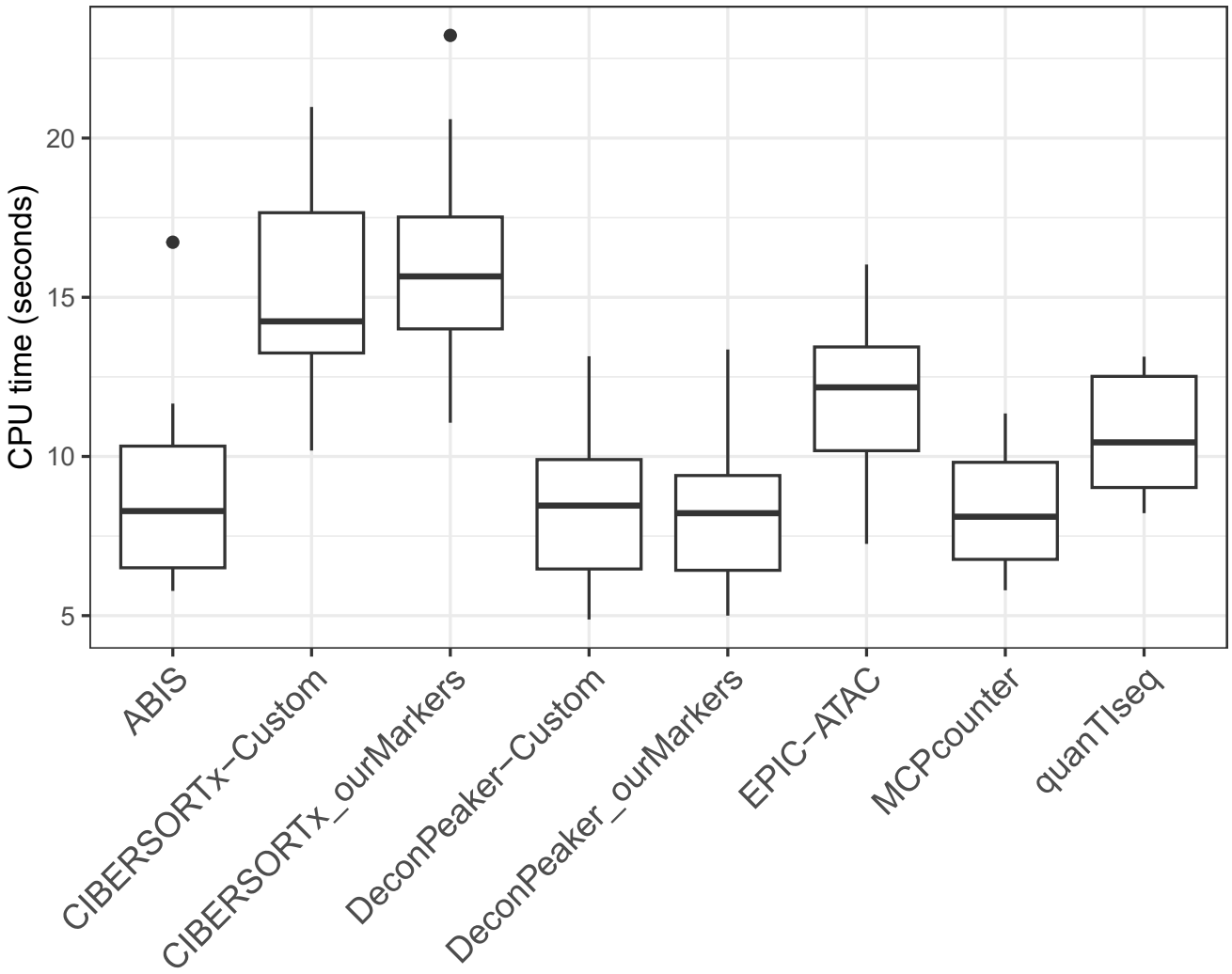


Figure 4-figure supplement 5: CPU time in seconds needed to run each tool on the benchmarking datasets.

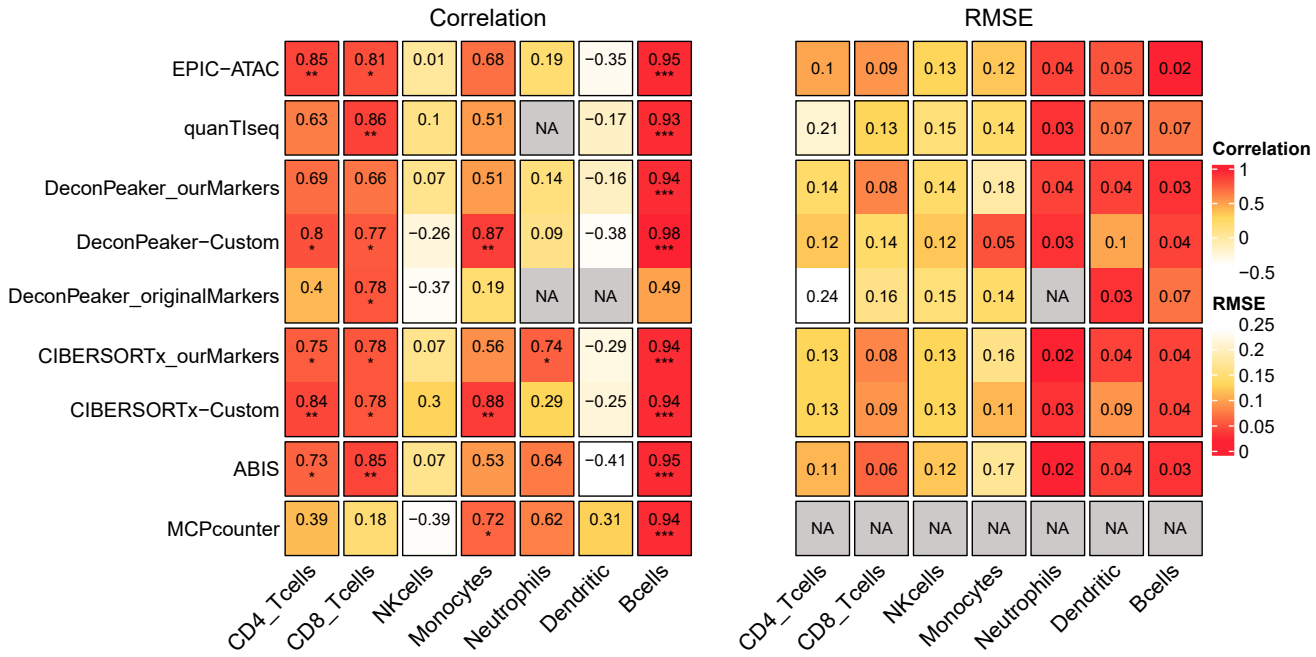


Figure 5-figure supplement 1: Pearson's correlation (left panel) and RMSE values (right panel) associated to the comparison of the cell-type proportions estimated by each deconvolution tool (rows) included in the benchmark and the true cell-type proportions in the PBMC samples. The comparisons were made in each cell-type (columns) separately. Significance level of the correlations are represented as follows: *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

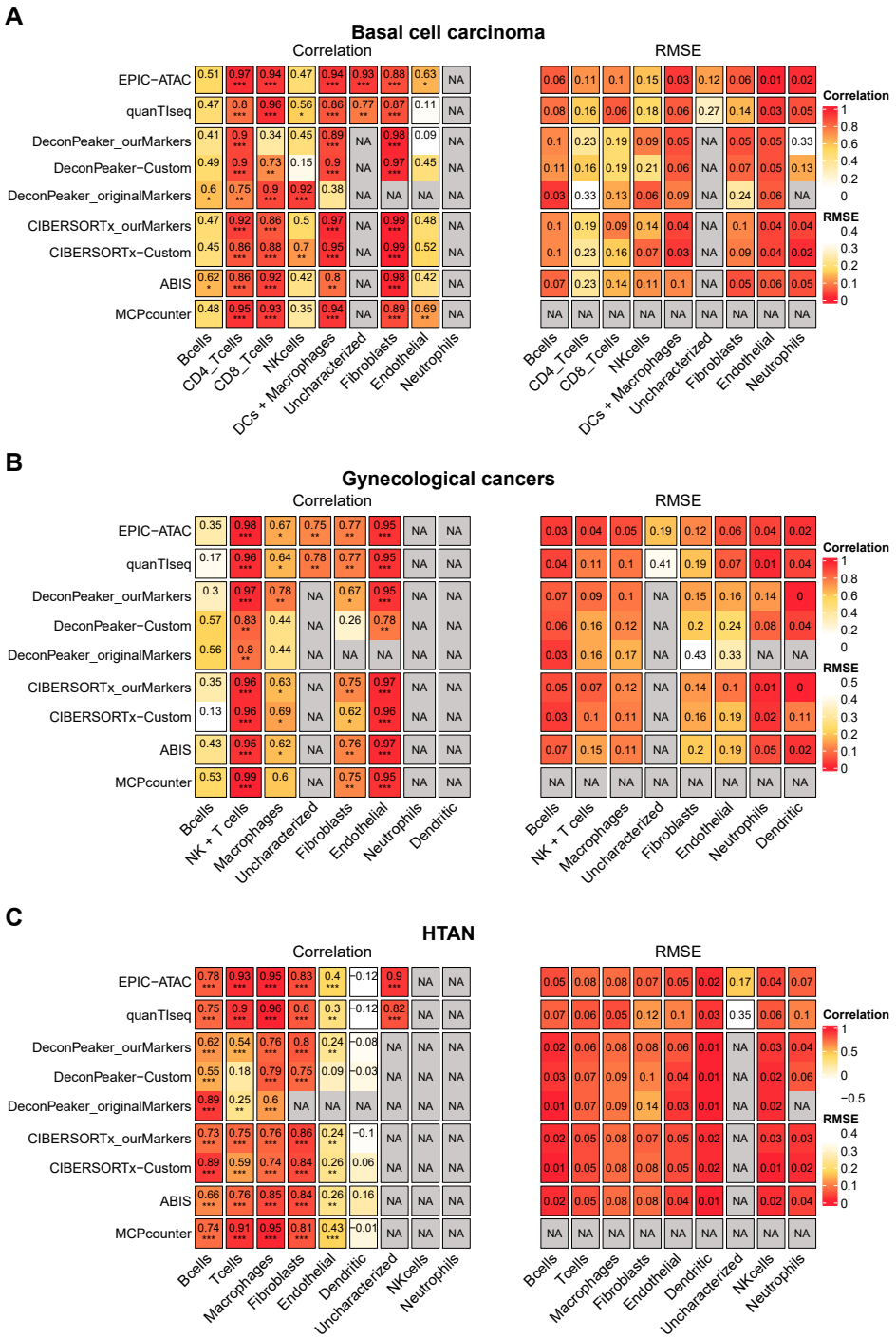


Figure 5-figure supplement 2: Pearson's correlation (left panel) and RMSE values (right panel) associated to the comparison of the cell-type proportions estimated by each deconvolution tool (rows) included in the benchmark and the true cell-type proportions. The comparisons were made in each cell-type (columns) separately. Panel **A**, **B** and **C** are associated to the benchmark performed on the basal cell carcinoma dataset (A), the gynecological cancer dataset (B) and the HTAN dataset (C). Significance level of the correlations are represented as follows: *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

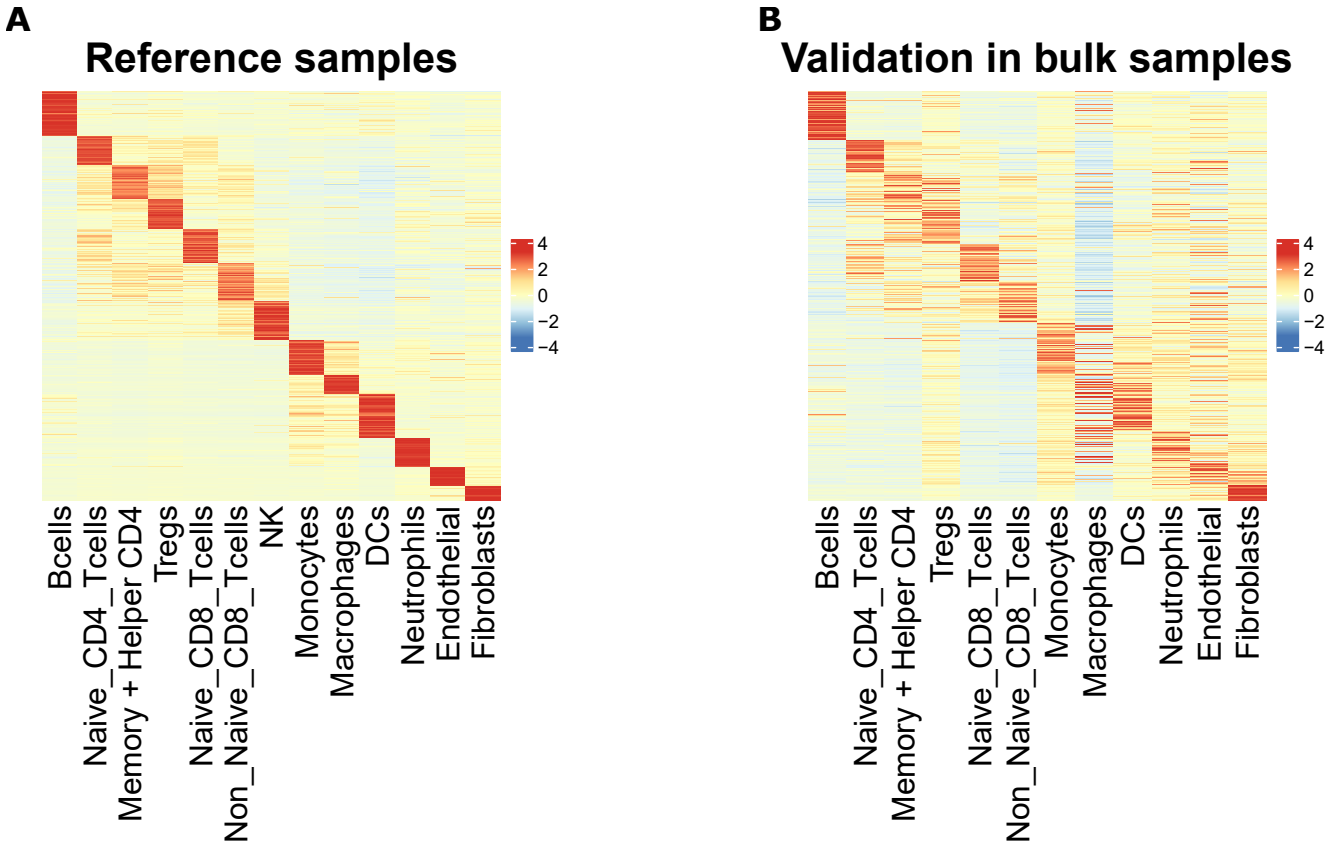


Figure 5-figure supplement 3: A) Scaled averaged chromatin accessibility of each cell-type specific marker peak in each cell-type obtained based on the ATAC-Seq reference samples used to identify the marker peaks. **B)** Scaled averaged chromatin accessibility of each cell-type specific marker peak in each cell-type from the sorted samples excluded from the reference samples (See Material and Methods).

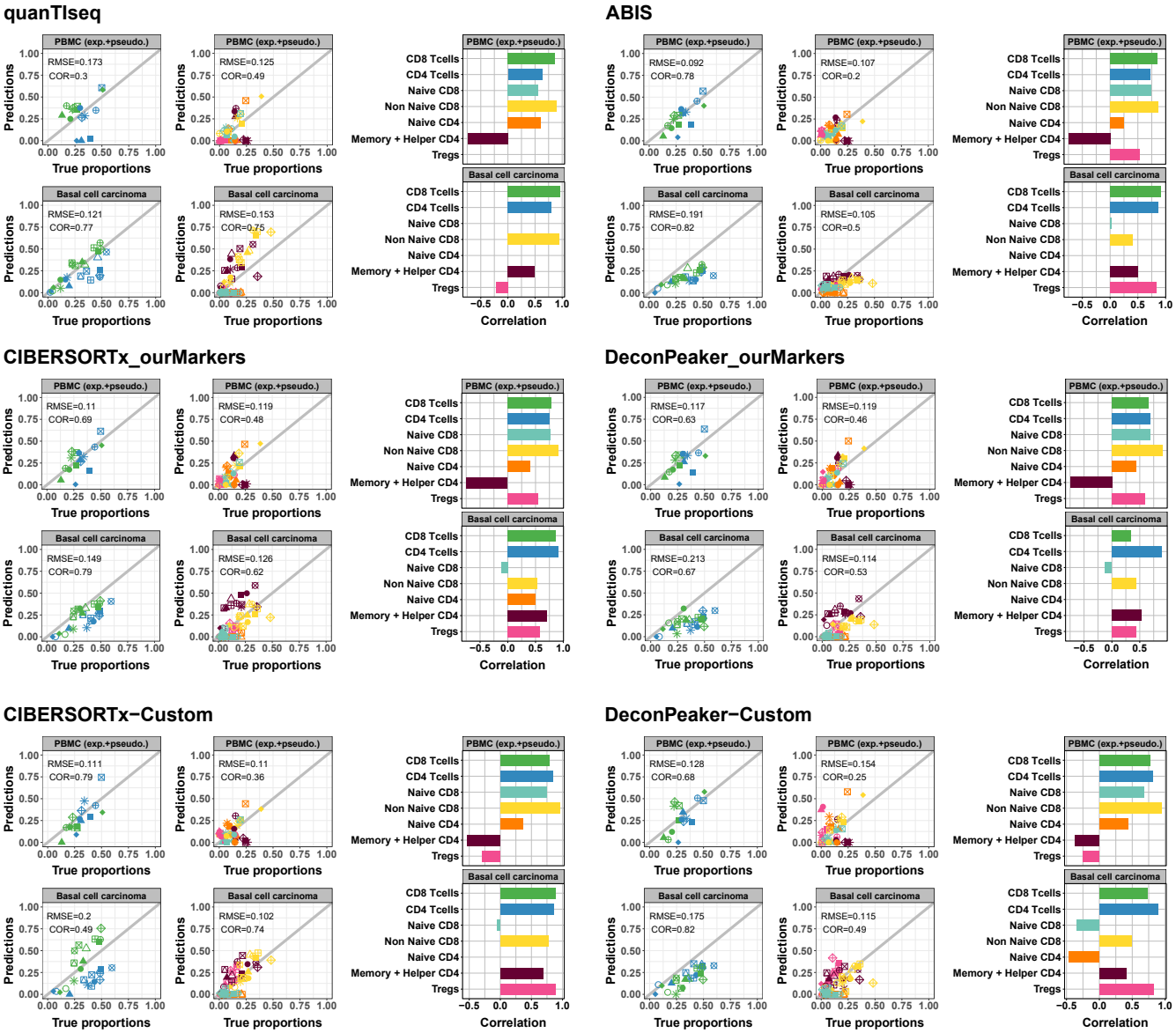


Figure 5-figure supplement 4: Left panels: Scatter plots representing proportions estimated by the different deconvolution tools included in the benchmark (y-axis) as a function of the true proportions (x-axis) for PBMC samples (top row) and the basal cell carcinoma pseudobulks (bottom row). CD4+ and CD8+ T-cells predictions were obtained using the reference profiles based on the major cell-types and the CD4+ and CD8+ T-cell subtypes predictions using the reference profiles including the T-cells subtypes. Right panels: Barplots representing Pearson's correlation values (x-axis) obtained based on the comparison of the predictions and true proportions in each cell-type (y-axis).

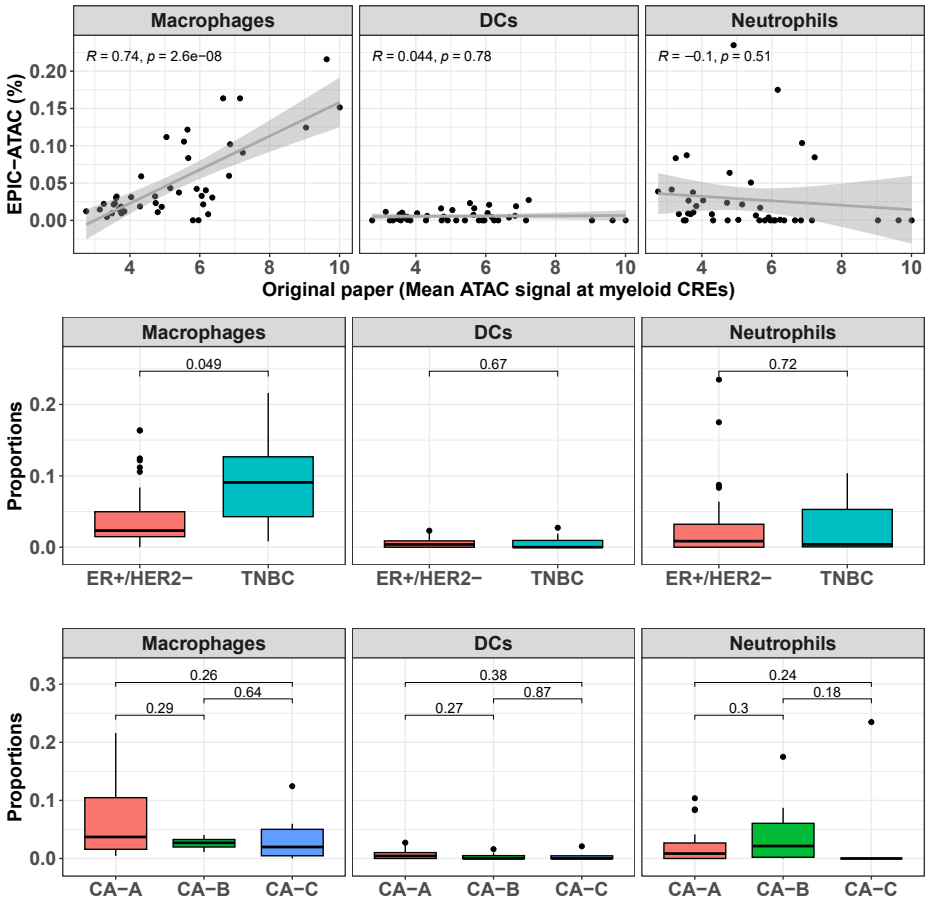


Figure 6-figure supplement 1: Top panel: Proportions of each myeloid cell type predicted by EPIC-ATAC in each sample as a function of the average ATAC signal at the myeloid specific CREs used by Kumegawa and colleagues to infer the level of infiltration in the tumor samples. Middle panel: Proportions of different cell types predicted by EPIC-ATAC in the samples stratified based on two breast cancer subtypes. Bottom panel: Proportions of different cell types predicted by EPIC-ATAC in the samples stratified based on three ER+/HER2- subgroups. Wilcoxon test p-values are represented at the top of the boxplots.